Laboratoire d'Informatique
Fondamentale de Lille

THESE

Nouveau régime

Présentée à

L'UNIVERSITE DES SCIENCES ET TECHNOLOGIES DE LILLE

Pour obtenir le titre de

DOCTEUR EN INFORMATIQUE

Par

Jean SIMON

MODELES FORMELS DE L'APPRENTISSAGE ET CATEGORISATION

Contribution à une étude comparée

Thèse soutenue le 10 décembre 1999 devant le jury composé de :

Président et rapporteur :	Françoise CORDIER	Université de Poitiers
Rapporteur :	Antoine CORNUEJOLS	Université de Paris Sud
Examinateurs :	Isabelle BONNOTTE	Université de Lille 3
	Max DAUCHET	Université de Lille 1
	François DENIS	Université de Lille 3
	Rémi GILLERON	Université de Lille 3

UNIVERSITE DES SCIENCES ET TECHNOLOGIES DE LILLE
U.F.R. d'I.E.E.A. Bâtiment M3. 59655 Villeneuve d'Ascq CEDEX
Tél. 03.20.43.47.24 Fax.03.20.43.65.66

A mon frère,
merci d'être.

Remerciements

Je tiens d'abord à remercier les rapporteurs. La version finale de ce travail tient compte de leur remarques.
Ainsi, je tiens à remercier vivement Françoise Cordier d'avoir accepté de présider le jury. Ce sont ses ouvrages qui m'ont poussé dans cette recherche. Durant son cours sur les représentations cognitives et le langage, j'ai pu constater sa rigueur scientifique, rigueur qu'elle a manifesté aussi dans la lecture de ce travail. Je tiens également à la remercier pour ses conseils tant bibliographiques que concernant les expériences.
Je tiens à remercier vivement aussi Antoine Cornuéjols pour l'intérêt qu'il a manifesté pour ce travail et la lecture exhaustive et fine qu'il en a faite. Ses nombreuses remarques m'ont permis d'appréhender de manière encore plus approfondie les liens qui peuvent exister entre le modèle PAC et la catégorisation. J'ai particulièrement apprécié son ouverture d'esprit et sa réelle curiosité pour les sciences cognitives.

Je souhaite également exprimer ma profonde gratitude aux personnes qui ont dirigé ce travail :
- Max Dauchet qui a accepté d'en prendre officiellement la direction, (Max semble toujours en quête d'absolu et le chercher dans les mathématiques),
- François Denis qui m'a permis de comprendre le sens d'un modèle formel et à quel point les mathématiques pouvaient être belles (vu sa manière de penser, François me fait parfois douter que tous les êtres humains ont un cerveau qui obéit aux mêmes règles de fonctionnement),
- Rémi Gilleron, pédagogue patient, qui a su m'aider et m'encourager lorsque cela était nécessaire (vrai modeste, Rémi, semble marquer un certain détachement Zen),
François et Rémi m'ont ainsi fait découvrir la rigueur indispensable à la recherche et, en même temps, la nécessaire distance à prendre vis-à-vis de cette même recherche. En me canalisant, ils ont permis que ce travail aboutisse.

J'ai aussi bénéficié d'un directeur officieux en psychologie. Je veux ainsi remercier Isabelle Bonnotte dont la lecture de la toute première version de ce travail m'a permis d'éviter de nombreux quiproquos, la remercier aussi pour son aide dans la mise en place des expériences et l'analyse des résultats. Néanmoins, en aucun cas, les erreurs contenues dans la partie psychologie de ce travail ne peuvent lui être imputées, ni d'ailleurs à aucun psychologue cité ici.

Je tiens ensuite à remercier tous ceux qui de près ou de loin m'ont permis de mener ce travail à son terme :
- A. Dubus et A. Hénau qui ont créé les conditions objectives me permettant de reprendre des études,
- F. Terryn qui m'a fait gagner un temps précieux en me trouvant des sujets qui acceptent de se prêter aux expériences,
- F. Anceaux qui m'a orienté, dès le début de cette recherche, vers les psychologues qui travaillaient sur le sujet,
- S. Chemlal qui a su écouter patiemment mes diverses élucubrations concernant la catégorisation, me trouver les sujets qui me manquaient pour ma première expérience et qui avait toujours sous le coude l'article qu'il me fallait concernant le problème qui me préoccupait,
- tous les membres du LIFL qui m'ont accepté dans leur locaux et m'ont aidé dans mes démêlés administratifs, les locataires du bureau 316, Cyrille, Francis, Yves mais aussi ceux

des bureaux voisins Bruno, Francesco, Yves et les habitants de la petite maison de Lille 3, Alain, Alain, Dominique, Isabelle, Jean, Marc.

Enfin, je tiens à remercier tous ceux sans qui ma vie ne serait pas une vie : Marie, Raoul, Luc, Philippe G., Rolande, Jean-Claude, Sylvain, Stéphane, Myrianne, Guy, Gaëlle, Yann, Marie-Anne, Annie, Franck, Anthony, Martine, Philippe L., Marlène, Julien, Pauline, Sophie, Manon, Françoise, Marianne. Qu'ils devinent l'ordre quelque peu égocentrique dans lequel ils apparaissent ici et qu'ils me pardonnent de leur avoir fait subir les prétests..

SOMMAIRE

Introduction générale

Préambule

Un théorème peut-il aider à comprendre comment l'individu se forge ses concepts ? A partir de quels présupposés psychologiques, l'informatique théorique établit-elle ses modélisations de l'apprentissage ? Ces questions en amènent une autre plus générale : existe-t-il des points communs en psychologie et en informatique concernant l'apprentissage qu'une étude interdisciplinaire permettrait de découvrir ?

En 86, Berwick [Berwick, 86] va ainsi se confronter au problème suivant :

> *Exemple :* Supposons que nous ayons un ensemble de blocs qui peuvent avoir comme forme 'carré' ou 'rond', comme taille 'petit' ou 'grand' et comme couleur 'rouge' ou 'vert'. Supposons aussi que nous voulons faire apprendre le concept «petit carré» à un apprenant et que pour cela nous lui présentions des exemples de ce concept.

Ce type d'apprentissage est qualifié d'*inductif.*

> Supposons de plus que l'on ne présente à l'apprenant que des exemples relevant du concept, [petit carré rouge, oui][1] et [petit carré vert, oui], et aucun contre-exemple.

On parle alors d'apprentissage inductif par *exemples positifs* uniquement car on ne présente aucun contre-exemple.

> Cet apprenant peut alors formuler l'hypothèse «carré» car un petit carré rouge et un petit carré vert sont bien des exemples du concept de «carré». Le problème est que son hypothèse est trop générale : il a *surgénéralisé*. Même si nous continuons à lui présenter des exemples positifs, son hypothèse ne sera jamais démentie, elle restera trop générale. Il faudrait pouvoir lui présenter un contre-exemple, un exemple négatif comme [grand carré vert, non], mais ceci n'est pas autorisé dans l'apprentissage inductif par exemples positifs.

Ainsi le problème en apprentissage inductif par exemples positifs est le risque de *surgénéralisation* : si l'apprenant émet une hypothèse trop générale aucun exemple positif ultérieur ne pourra infirmer cette hypothèse.

Berwick aborde cette difficulté d'un triple point de vue. D'abord pour éviter ce problème de surgénéralisation dans l'apprentissage par exemples positifs, Berwick propose l'heuristique suivante : à la présentation de chaque exemple, l'apprenant propose l'hypothèse *la plus spécifique* qui rende compte de tous les exemples vus jusque-là.

> Comme l'apprenant doit faire l'hypothèse *la plus spécifique*, au premier exemple, il émet l'hypothèse «petit carré rouge ». Au second il doit «laisser tomber»[2] l'attribut couleur et émet l'hypothèse «petit carré» qui est l'hypothèse correcte.

Ensuite, et c'est ce qu'il y a d'intéressant dans le travail de Berwick, il établit des liaisons avec les travaux d'autres disciplines.

---

[1] Dans l'exemple [xxx, oui], le premier terme xxx décrit l'objet, le second indique s'il appartient (oui) ou n'appartient pas (non) au concept.

[2] [Michalski, 83]

D'une part, Berwick démontre que son heuristique est valide en s'appuyant sur un théorème proposé par [Angluin, 80] dans le cadre d'une modélisation informatique de l'apprentissage.

D'autre part, en utilisant les travaux de [Sommers, 71] et [Keil, 79], en psychologie, il pense que les enfants apprennent, conceptualisent, de cette façon. Il interprète leurs travaux ainsi : l'enfant formule des hypothèses sur le monde qu'il ne généralise que petit à petit.

Ce dernier point est discutable mais ce qui nous importe dans le travail de Berwick réside dans la démarche. On y voit *qu'un théorème permet de fonder une heuristique qui à son tour permet, peut-être, de décrire un apprentissage naturel [Simon, 95].*

Problématique générale

L'article de Berwick illustre l'idée suivante : *les modèles formels de l'apprentissage proposés en informatique doivent pouvoir rendre compte de la réalité pour être légitimés. Réciproquement, si les approches empiriques de l'apprentissage naturel, en psychologie notamment, correspondent à un modèle formel, elles sont par-là même confortées [Simon, Denis et Gilleron, 97].*

Notre travail se place donc dans *le cadre des sciences cognitives.* Dans cette introduction, nous n'allons pas refaire l'historique de celles-ci, nous renvoyons le lecteur intéressé aux ouvrages suivants [Piattelli-Palmarini, 79], [Andler, 92], [Varela, 89] qui permettent de se faire une idée assez complète du statut de celles-ci dans le monde scientifique et de leur évolution. Par contre, nous allons préciser *notre* position vis-à-vis d'elles, ce qui amène inévitablement à définir ce que nous entendons par *représentations.* Nous expliquerons ensuite les *difficultés méthodologiques* que cela pose d'évoluer dans un tel cadre.

L'objet de ce travail concerne *l'apprentissage inductif.* Nous allons exposer rapidement comment l'informatique, d'une part, et la psychologie, de l'autre, appréhendent l'apprentissage en général. Ceci permettra de montrer que le terme «d'apprentissage» que l'on trouve dans la dénomination «*théories formelles de l'apprentissage* » est trop général, et qu'en psychologie le concept le plus proche de l'apprentissage inductif tel que l'envisage l'informatique est, en fait, la *catégorisation.*

Enfin, nous donnerons le *plan d'ensemble* de ce travail, ce sera l'occasion de préciser la problématique générale ci-dessus et de décrire la *démarche employée.*

Un travail dans le cadre des sciences cognitives

Si le point de départ de notre travail est l'informatique, il se situe dans le cadre plus large des sciences cognitives. Voici comment Andler les définit : « Elles ont pour objet de décrire, d'expliquer et le cas échéant de simuler les principales dispositions et capacités de l'esprit humain - langage, raisonnement, perception, coordination motrice, planification... » [Andler, 92]. Les sciences cognitives recouvrent de nombreuses disciplines : de l'informatique à la psychologie en passant par les neurosciences, l'anthropologie, etc. [Varela, 89]. Ce qui fédèrent ces disciplines dans les sciences cognitives, c'est le rapport qu'elles entretiennent avec l'informatique.

Ce rapport à l'informatique est cependant l'objet d'un premier clivage au sein de ces sciences. Pour certains chercheurs, ce rapport relève de la métaphore, ils utilisent les recherches effectuées en informatique pour décrire les processus cognitifs humains. Pour d'autres, il y a identité entre les processus cognitifs humains et ceux supportés par un système artificiel. Pour notre part, nous nous plaçons dans le cadre de la métaphore : nous utilisons les recherches effectuées en informatique sur l'apprentissage pour étudier celles effectuées en psychologie sur le même domaine.

Il existe un second clivage au sein des sciences cognitives. Pour les *cognitivistes «purs»*, une activité mentale est un calcul sur des représentations, sur des symboles. Leur démarche est aussi appelée computo-symbolique. Au symbole du niveau mental correspond une place précise au niveau physique. Les cellules du cerveau et les circuits électroniques sont des supports différents mais ils sont à un niveau d'abstraction approprié, les supports de systèmes *formels* équivalents [Haugeland, 89] : la cognition est ramenée à un système formel. Les symboles sont à la fois signifiant et matériel, et l'ordinateur est une machine qui respecte le sens des symboles tout en ne manipulant que leur forme physique. Cette approche a un point faible que révèle l'étude plus approfondie des phénomènes cognitifs : elle ne permet pas de rendre compte de la fluidité, de la robustesse et de la faillibilité caractéristiques de la cognition humaine [Andler, 92].

De ce fait, à côté de ce courant cognitiviste pur, est apparu un courant *connexionniste*. Pour les connexionnistes, les représentations internes sont d'un grain plus fin que le symbole. Certains chercheurs parlent du connexionnisme comme du 'paradigme sub-symbolique'. Au niveau sub-symbolique, les descriptions sont construites à partir de constituants plus petits que le symbole. Le sens, toutefois, ne réside pas dans ces constituants en soi, mais dans les schémas d'activités complexes émergeant d'une interaction entre plusieurs d'entre eux. Le domaine sub-symbolique, est de plus haut niveau que le biologique mais s'en rapproche davantage que le niveau symbolique [Varela, 89].

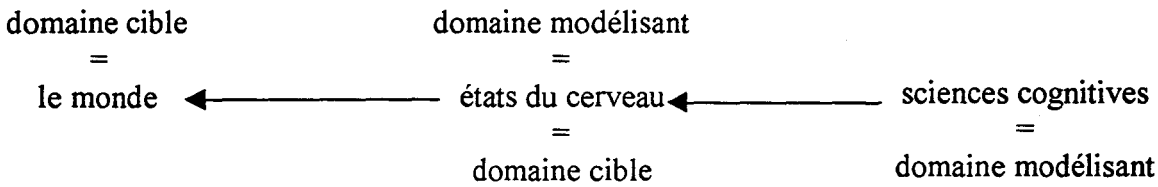
Ces deux approches, computo-symbolique et connexionniste, posent la question de la relation entre les différents niveaux d'explication dans l'étude de la cognition. Comment l'émergence sub-symbolique et la computation symbolique peuvent-elles être reliées ? Selon [Varela, 89], la réponse est que ces deux approches sont complémentaires, l'une

ascendante et l'autre descendante. Néanmoins, lorsqu'on lit les recherches en psychologie actuellement, on s'aperçoit que les chercheurs utilisent de moins en moins la métaphore computo-symbolique et de plus en plus celle connexionniste [Barsalou, 92] : « Les réseaux connexionnistes ont un potentiel énorme en tant que modèles de la catégorisation. Malgré de nombreuses différences, leur architecture simple ressemble en partie à celle du cerveau. »

Notre travail se situe ainsi dans le cadre des sciences cognitives élargies au connexionnisme, ceci exige de préciser notre position relativement aux représentations car nous serons fréquemment amenés à parler de celles-ci.

Les représentations

Si le cognitivisme pur est remis en question, c'est sur la question centrale des représentations. Lorsque nous parlons de représentations, nous les envisageons selon la proposition faite par [Barsalou, 92] : *étant donné un domaine cible et un domaine 'modélisant', il existe une correspondance systématique entre la structure pertinente dans le domaine modélisant et la structure pertinente du domaine cible.* Cette façon d'envisager les représentations permet d'englober aussi bien les représentations symboliques que les représentations subsymboliques et peut même être étendue aux sciences :



Symbolique, sub-symbolique, différents types de domaines 'modélisant' peuvent alors correspondre à un même domaine cible. Pour l'être humain : « La structure neuronale du cerveau est capable d'établir des correspondances systématiques entre des sous-ensembles de sa structure et des sous-ensembles de la structure de l'environnement physique. » [Barsalou, 92]. Autrement dit, lorsque nous parlerons de la représentation d'un objet du monde chez l'être humain nous ferons simplement référence à l'activation d'une population de neurones de son cerveau en relation avec cet objet. Cependant, le rôle de la psychologie n'est pas de donner le détail de cette activation neuronale, c'est celui de la neurobiologie. La psychologie, en tant que domaine modélisant du domaine cible qu'est l'ensemble des états du cerveau, essaie de décrire ces états en termes de processus cognitifs, elle travaille au niveau des caractéristiques du traitement de l'information [Barsalou, 92].

Disciplines concernées et problèmes de méthodologie

Ce qui motive ce travail est l'idée commune selon laquelle, lorsqu'on aborde un problème par deux points de vue différents, la totalité de l'information que l'on obtient sur le problème est supérieure à la somme des informations de chacun des points de vue. Bateson, dans *La nature et la pensée* [Bateson, 84], illustrait ceci en faisant remarquer que la stéréovision apporte une information en plus de celles apportées par chaque œil : la profondeur. Nous souhaitons ainsi montrer qu'une double approche d'un même phénomène peut être fructueuse. Cette double approche s'appuiera, comme nous venons de le dire, sur deux domaines des sciences cognitives : l'informatique et la psychologie.

Le fait de travailler sur deux disciplines pose des problèmes de méthodologie. Ces problèmes sont accrus lorsque, comme c'est le cas ici, l'une d'entre elles s'apparente aux mathématiques et l'autre relève des sciences humaines. Cette absence de méthodologie commune rend difficile la validation du travail. Néanmoins, étant donné qu'en informatique la recherche considérée ici porte sur la modélisation et, qu'en psychologie, la recherche porte sur le réel, *si le modèle informatique est capable de rendre compte des concepts produits et des résultats obtenus en psychologie, nous pourrons considérer que les rapprochements que nous aurons effectués sont justifiés.*

En plus des problèmes méthodologiques, les disciplines se distinguent aussi en fonction de leur vocabulaire et de ce que l'on peut appeler leurs us et coutumes. Aussi, est-il nécessaire de vérifier que les deux disciplines font bien le même usage des mots et, quand ce n'est pas le cas, de préciser le sens qui leur est donné dans chaque domaine. C'est d'ailleurs ce que nous sommes amenés à faire dès cette introduction avec le terme même «d'apprentissage ». Concernant les us et coutumes, nous avons essayé de les respecter. Par exemple, en informatique, il est fréquent lorsque l'on cite un chercheur, de le faire dans sa langue, tout au moins s'il est anglo-saxon, alors qu'en psychologie, on le traduira en français. Le premier choix s'explique par la peur de trahir la pensée de l'auteur par une traduction malencontreuse, le second, par le souhait de pouvoir être compris de tout lecteur. Aussi gardons-nous les règles propres à chaque discipline et traduisons lorsque l'auteur est un psychologue et gardons la citation dans sa langue lorsque l'auteur est un informaticien.

Enfin, si nous travaillons sur deux disciplines, le point de départ est quand même l'informatique. Cela nous amène parfois en psychologie à rapprocher des sous-champs disciplinaires qui d'habitude font l'objet d'études bien séparées. A chaque fois, nous le signalons. De la même façon, *les articles en psychologie que nous sommes amenés à présenter sont choisis parce qu'ils permettent d'illustrer tel ou tel point que nous souhaitons mettre en exergue* et non en raison de leur caractère récent ou définitif dans le domaine. Nous espérons que le lecteur psychologue ne trouvera pas trop iconoclaste cette façon de faire.

L'apprentissage

Le cadre étant défini, précisons maintenant l'objet de l'étude, l'apprentissage et plus exactement l'apprentissage inductif. Comme nous l'avons expliqué, dans ce type d'apprentissage, on présente à l'apprenant des exemples en spécifiant pour chacun de ces exemples s'il est positif (représentant le concept) ou négatif (contre-exemple) et l'apprenant doit retourner un concept qui correspond à ces exemples. Pour illustrer ceci, on peut imaginer l'enfant qui se construit le concept de chien au travers des différents animaux qu'il rencontre et qui seront étiquetés par son entourage en «chien » (exemples positifs) ou «non-chien » (contre-exemples : « chat », «vache », etc.)[3].

Aussi bien en informatique qu'en psychologie, l'apprentissage inductif n'est qu'une partie de ce que chaque discipline définit comme l'apprentissage. En outre, ce qui s'apparente le plus à l'apprentissage tel que l'envisagent les théories formelles en informatique porte un autre nom en psychologie : la catégorisation. C'est pourquoi nous allons brosser à grands traits ce qu'est l'apprentissage dans chacune de ces disciplines pour que le lecteur puisse plus aisément cerner le sens qu'elles donnent à ce mot.

L'apprentissage en informatique[4]

L'idée de faire apprendre les machines apparaît quasiment en même temps que l'informatique [Mitchell, 97] et, dès les années 50, Samuel propose des stratégies d'apprentissage pour un logiciel de jeu de dames [Samuel, 59]. Cependant, c'est avec l'apparition d'immenses bases de données que se présente la *nécessité* de l'apprentissage en informatique [Haton et Haton, 93] ; il suffit de voir l'essor que prend le Data Mining actuellement. Pour aider l'individu dans ses recherches, il faut que la machine puisse trier, cataloguer, ordonner ces données. Les données étant de toutes formes, il n'est pas possible d'user d'algorithme précis, il faut que ce soit l'application qui «trouve» la démarche appropriée.

De manière un peu caricaturale, on peut classer les recherches sur l'apprentissage en informatique essentiellement en deux courants : l'*apprentissage automatique* («machine learning ») et *les théories formelles de l'apprentissage* («computational learning theory »)[5]. L'apprentissage automatique consiste à *construire des systèmes apprenants*, des machines capables d'apprendre. Les théories formelles de l'apprentissage ont pour objet d'essayer de caractériser l'*apprenabilité* : quels sont les objets apprenables et sous quelles conditions le sont-ils ? Notre travail se place dans le cadre de ces dernières, il nous paraît toutefois utile de présenter aussi l'apprentissage automatique afin que le

---

[3] Bien évidemment, comme nous le verrons, cette analogie relève de la caricature, mais elle permet de se faire une première idée.

[4] Le lecteur intéressé trouvera dans [Mitchell, 97] une vision complète et très bien architecturée de ce que peut être l'apprentissage en informatique actuellement.

[5] La distinction que l'on fait ici est un peu arbitraire, car le « computational learning theory » relève du « machine learning » mais elle permet de situer le travail. Elle est empruntée à Boucheron qui l'utilise implicitement dans son livre : « Théorie de l'apprentissage : de l'approche formelle aux enjeux cognitifs », à l'opposé Natarajan lui a intitulé son livre «Machine Learning : a theoretical approach. »

lecteur puisse se faire une intuition assez complète de ce que peut être l'apprentissage en informatique

L'apprentissage automatique

Outre l'analyse statistique, qui peut être apparentée à un apprentissage dans le sens où l'on cherche à trouver des lois qui rendent compte d'un ensemble de faits, on peut distinguer deux types majeurs d'apprentissage automatique [Michalski, 86], [Mitchell, 97] : l'apprentissage symbolique et l'apprentissage adaptatif. Nous développons, en annexe 1, un exemple simple de chacun de ces deux types d'apprentissage.

L'apprentissage symbolique

On trouvera dans [Carbonell, Michalski, Mitchell, 83], [Dietterich, Michalski, 83] [Michalski, Kodratoff, 90] [Mitchell, 97] différents classements de l'apprentissage symbolique. Sommairement, on peut différencier les différents systèmes apprenants selon le type d'inférence qu'ils utilisent pour apprendre, et/ou selon le type de représentation (on parle de *schéma de représentations*) employé pour décrire le monde.

Les principaux types d'inférence suivants sont ainsi exploités [Haton et Haton, 93] :
-l'induction : trouver des règles à partir d'exemples et de contre-exemples,
-la déduction : détermination de traits spécifiques à partir de règles générales, construction de nouvelles règles générales à partir d'anciennes, affinement de stratégies, etc.
-l'analogie : utilisation de la connaissance d'un problème et de la façon de le résoudre pour résoudre un problème voisin.
Bien souvent, un système apprenant n'est pas construit sur un type unique d'inférence mais sur une composition des différents types proposés ci-dessus. Ainsi, et comme on peut le voir ici, l'apprentissage symbolique ne se résume pas à l'apprentissage inductif.

Les *schémas de représentations symboliques* utilisés pour décrire les connaissances sont divers : les grammaires formelles, les systèmes à production de règles, les expressions basées sur la logique formelle, les graphes, les réseaux, les frames, ... Parmi ces schémas, l'un des plus courants est *l'arbre de décisions* (cf. annexe 1) qui permet de répartir les objets en différentes catégories. Pour classer un exemple, on lui fait descendre l'arbre de nœud en nœud jusqu'à ce qu'il atteigne une feuille. Chaque nœud est un test qui l'oriente vers une branche ou une autre, chaque feuille représente une classe. L'apprentissage des arbres de décisions est utilisé pour repérer des régularités dans de grandes bases de données [Mitchell, 97], la NASA, notamment, l'utilise pour classifier automatiquement des objets célestes [Fayyad et al, 1995]. Dans le domaine des jeux, TD-GAMMON [Tesauro, 1995] est capable de jouer au backgammon à un niveau international.

L'apprentissage adaptatif (connexionnisme et génétique)

Parallèlement aux précédents paradigmes, d'autres recherches ont lieu qui, elles, relèvent davantage de l'apprentissage inductif. Sous le terme d'adaptatif, on classe deux types d'apprentissage : l'apprentissage connexionniste et l'apprentissage génétique. Ces deux types d'apprentissage sont relativement récents. On parle d'apprentissage *adaptatif*, terme quelque peu pléonastique, car il s'agit d'un apprentissage basé sur l'inhibition et le renforcement.

Le *connexionnisme* s'est inspiré de la structure du cerveau pour construire des machines qui s'y apparentent. Son élément de base est le neurone formel. L'apprentissage (voir annexe 1) consiste à présenter les exemples aux neurones d'entrée du réseau l'un après l'autre. A chaque exemple, on observe la réponse du réseau et on essaie de réduire l'écart entre celle-ci et la réponse que l'on souhaite en obtenir. Ceci se fait soit en intervenant sur les coefficients synaptiques des neurones donnant une réponse incorrecte soit en modifiant les coefficients de tous les neurones, proportionnellement à la différence entre réponse obtenue et réponse espérée. Dans des cas de figures simples, au bout d'un certain nombre de présentations, les efficacités synaptiques finissent par se stabiliser : l'algorithme 'converge' et l'apprentissage est réalisé. L'une des motivations de l'apprentissage de réseaux de neurones artificiels (RNA) est d'essayer de capturer le type de calcul hautement parallèle effectué par le cerveau sur des représentations distribuées [Mitchell, 97]. Ainsi, l'apprentissage de RNA est particulièrement bien approprié à des problèmes où l'échantillon d'entraînement est constitué de données issues de senseurs, de capteurs tels que des caméras ou des microphones. Par ailleurs, l'apprentissage de RNA est robuste aux erreurs dans l'échantillon. Cela fait qu'il est bien adapté à des problèmes du monde réel. Il est ainsi utilisé dans la reconnaissance de caractères [LeCun et al, 89] dans la reconnaissance de la parole [Lang et al, 90] dans la reconnaissance de visages [Cottrell, 90].

Un autre type d'apprentissage adaptatif est celui des *algorithmes génétiques*. Ceux-ci sont issus d'un certain type d'algorithmes appelés algorithmes probabilistes. Le principe général de ces méthodes consiste à parcourir, d'une manière plus ou moins aléatoire, l'espace des hypothèses pour en extraire un ensemble appelé population. Ces hypothèses subissent aléatoirement des opérations dites génétiques, qui donnent naissance à d'autres hypothèses qui remplacent les hypothèses non sélectionnées jusqu'à ce que l'adéquation soit satisfaisante vis-à-vis de l'échantillon. Les opérations sont du type mutation d'une hypothèse ou croisement de deux hypothèses. En se représentant une hypothèse comme un ensemble d'instructions, une mutation consiste à insérer, supprimer ou transformer des instructions, alors que le croisement de deux hypothèses revient à échanger des sous-séquences. Ainsi l'évolution d'une population s'effectue par mutation ou croisement de ses éléments les plus forts pour remplacer les éléments les plus faibles, d'où l'appellation d'algorithmes génétiques.

Les théories formelles de l'apprentissage

Parallèle à ces études sur l'apprentissage automatique, il existe en informatique un domaine de recherches appelé *«théories formelles de l'apprentissage»*. Il est important de bien faire la distinction entre les deux. Le premier se place surtout au niveau de la technique et a pour objectif immédiat de proposer des systèmes apprenants qui peuvent être implantés dans des machines et testés. Le critère de validité peut alors être l'efficacité. Tandis que les secondes, comme leur nom l'indique, relèvent de la théorie, d'ailleurs, pour un néophyte, elles s'apparentent davantage aux mathématiques qu'à l'informatique. Le critère de validité, outre la nécessaire cohérence interne, consiste surtout dans l'adéquation des modèles proposés à la réalité qu'ils sont sensés appréhender.
Alors que la plupart des travaux mettant en parallèle informatique et psychologie se situent dans le cadre de l'apprentissage automatique (e.g. [Kant, 96], [Borderie, 97]...), nous nous plaçons dans le cadre des théories formelles.

Les théories formelles de l'apprentissage ont pour objectif premier de poser des bornes théoriques sur l'apprentissage automatique vu précédemment. Pour cela, *leur démarche va consister à définir un modèle qui essaie d'appréhender d'une manière aussi rigoureuse que possible cet apprentissage.* Les modèles proposés actuellement par les théories formelles se préoccupent surtout de l'apprentissage inductif: on suppose que l'apprenant doit découvrir un concept sur la base d'une présentation qu'on lui fait d'exemples de ce concept.

Ces théories proposent ainsi un cadre d'apprentissage et cherchent, pour chaque classe de concepts, un algorithme capable d'apprendre tous les concepts de cette classe dans les conditions définies par le cadre. Un néologisme a été créé pour définir cette problématique : *l'apprenabilité*. On cherche à démontrer que telle classe de concepts est *apprenable* dans tel modèle et la classe est dite apprenable si on est capable de proposer un apprenant/algorithme qui sur la base d'une présentation d'un échantillon d'exemples d'un concept est capable de retourner le concept. Ainsi, il a été démontré que la classe des conjonctions est apprenable car il existe un algorithme qui, lorsqu'on lui présente en entrée des exemples (ex : [carré petit rouge, oui],...), et des contre-exemples (ex : [rond petit rouge, non],...) d'une conjonction est capable de retourner la conjonction («carré et rouge »).

Les modèles proposés se distinguent selon les aspects qu'ils prennent en compte pour décrire l'apprentissage :
- le type d'apprentissage : exact, approximatif, etc.,
- le protocole d'apprentissage : présentation d'exemples positifs uniquement, ou positifs et négatifs, séquentiellement ou en un seul bloc, etc.,
...

Actuellement, deux modèles et leur dérivés constituent principalement le domaine des théories formelles de l'apprentissage : celui de [Gold, 67] *d'identification à la limite* et celui de [Valiant, 84] d'*apprentissage Probablement Approximativement Correct.* Nous

les présentons brièvement ci-dessous. Nous reprendrons plus longuement la description du modèle de Valiant dans le chapitre 1 puisque c'est celui que nous avons adopté.

Le modèle de Gold d'identification à la limite [Gold, 67]

Dans le modèle de Gold d'identification à la limite, on présente à l'apprenant un flux *infini* d'exemples. A chaque exemple, l'algorithme doit conjecturer une hypothèse sur le concept à découvrir. *L'algorithme converge à la limite, si ses hypothèses convergent vers le concept inconnu après un nombre fini d'exemples.* Supposons que les exemples soient [24, oui], [123456, oui], [123, non]... il faut qu'à un certain moment l'algorithme finisse par conjecturer «ensemble des nombres pairs » et ne change plus d'hypothèse par la suite. Cette convergence est réclamée sur tous les flux d'exemples possibles, c'est-à-dire quel que soit leur ordre de présentation. Il est supposé aussi que la présentation des exemples est complète, c'est-à-dire que l'apprenant est amené à rencontrer *l'ensemble* des exemples et contre-exemples du concept.

Cette définition de l'apprentissage capte assez bien l'aspect toujours incomplet de l'apprentissage naturel. Personne ne peut prétendre posséder parfaitement sa langue ou un concept. Ainsi, pour un enfant, la baleine pourra être d'abord un poisson, ce n'est que plus tard qu'il considérera que c'est un mammifère. Ses trois hypothèses («baleine», «poisson», «mammifère») sont donc erronées à un moment donné : les hypothèses ne sont pas encore identiques aux concepts respectifs. Cependant dans le modèle de Gold, on ne considérera qu'il a appris que lorsqu'il ne fera plus d'erreur, que ses hypothèses se seront stabilisées sur les concepts exacts.
Cette définition permet aussi de rendre compte de la «théorisation» du monde : une théorie reste valable tant qu'un contre-exemple ne la contredit pas.

Le problème est que cette définition permet difficilement *d'évaluer l'apprentissage.* Si on reprend l'exemple de la langue, avec le modèle Gold on considérera que ni l'enfant *ni l'adulte* ne l'ont apprise. Dans ce modèle, si l'on voulait démontrer que l'être humain est capable de dominer une langue, il faudrait supposer qu'il est éternel et prouver qu'un jour il en aurait une connaissance parfaite.

Cette quasi-impossibilité d'évaluer tient aux caractéristiques suivantes du modèle :
- présentation des exemples *complète*,
- apprentissage *exact*,
- *pas de critère d'arrêt* qui garantisse qu'à un certain moment l'apprentissage se terminera,
- *pas de considération de coûts* tel que le temps que l'algorithme prendra pour apprendre ou l'espace mémoire dont il aura besoin.
Cette critique que l'on peut faire du modèle de Gold amène au modèle de Valiant.

Le modèle de Valiant d'Apprentissage Probablement Approximativement Correct [Valiant, 84]

Le modèle de [Valiant, 84] d'apprentissage Probablement Approximativement Correct, apprentissage PAC, va permettre de lever la plupart des problèmes posés par le modèle de Gold.
D'abord, à l'opposé du modèle de Gold qui demande une présentation *complète*, dans l'apprentissage PAC, on suppose que l'apprenant ne voit qu'un certain nombre d'exemples. Cela oblige à accepter que l'apprentissage ne puisse être qu'approximatif. En effet, il est toujours possible qu'un exemple rare, que l'apprenant n'a pas vu, soit mal classé. On ne demande donc pas que l'hypothèse soit identique au concept mais seulement qu'elle en soit une bonne approximation. Cette approximation doit tenir compte des exemples vus par l'apprenant et donc des probabilités de présentation de ces exemples.
Par ailleurs, on ne suppose plus un flux infini d'exemples, on y réclame au contraire que l'apprentissage s'arrête au bout d'un certain temps. Ce temps doit être fonction des exemples et du concept à apprendre mais aussi de l'approximation tolérée. Ce critère d'arrêt permet d'évaluer l'apprentissage. Cependant, de même que le modèle de Gold réclamait que l'apprentissage ait lieu sur tous les flux d'exemples possibles, le modèle de Valiant exige que l'apprentissage ait lieu pour toutes distributions de probabilités possibles sur les exemples.

Comme nous le verrons, cette modélisation appréhende assez bien l'apprentissage naturel. La différence entre les différents environnements d'apprentissage possibles est prise en compte au travers des distributions de probabilités sur les exemples. Comme dans l'apprentissage naturel, l'hypothèse apprise est rarement exacte. C'est pourquoi on peut considérer que le côté «approximatif» de l'apprentissage de Valiant rejoint le côté «à la limite» du modèle de Gold, en ce sens que l'hypothèse courante dans le modèle de Gold qui tend vers le concept à apprendre, correspond au concept appris dans le modèle de Valiant. Si on reprend le concept de «mammifère», le fait que l'enfant ne considère pas la baleine comme un mammifère relève de cette approximation, la baleine fait partie de l'erreur de l'hypothèse. Cependant à la différence du modèle de Gold, il existe ici un critère d'arrêt. Dans le modèle de Gold, nous ne pouvions pas évaluer l'apprentissage car il ne se terminait jamais. En mettant ce critère d'approximation en place, le modèle de Valiant permet d'évaluer un apprentissage délimité dans le temps.

Ce que nous voulons montrer dans ce travail, c'est que le modèle de Valiant permet bien de décrire un processus cognitif que nous avons appelé «apprentissage naturel» dans les paragraphes précédants. Cependant, comme nous allons le voir maintenant, il s'agit d'un abus de langage car ce processus n'est pas ce que la psychologie appelle «apprentissage».

L'apprentissage en psychologie

Il est impossible de présenter ce qu'est l'apprentissage pour la psychologie en quelques paragraphes. Néanmoins, les quelques indications données ci-dessous permettent de montrer que ce terme recouvre en psychologie un domaine beaucoup plus vaste que celui envisagé par les théories formelles.

Que ce soit dans le Larousse, le Petit Robert, ou l'Encyclopédia Universalis, les définitions données de l'apprentissage sont «une modification durable du comportement ». Une définition plus précise est proposée par Berbaum : «Apprentissage (Learning) : Acquisition d'une conduite nouvelle, capacité de pratiquer un comportement nouveau ou une manière d'être nouvelle.» [Berbaum, 94]. Houdé pour sa part le définit ainsi : « L'apprentissage est une modification de la capacité à réaliser une tâche sous l'effet d'une interaction avec l'environnement.» [Houdé et al 98].

Cependant, ces définitions restent encore assez vagues. C'est avec les diverses taxonomies proposées par les chercheurs que l'on approche plus précisément ce que peut être l'apprentissage considéré en psychologie. Selon Annick Weil-Barais [Weil-Barais, 93] l'apprentissage se définit selon :
- les béhavioristes, ou comportementalistes, comme une modification des comportements, on parle alors d'apprentissage par conditionnement,
- les cognitivistes, comme une modification des connaissances, on parle alors d'apprentissage par construction de la réponse.

On distingue deux types principaux d'apprentissage par *conditionnement* [Berbaum J., 94] :
- conditionnement répondant (Pavlov),
- conditionnement opérant ou instrumental (Skinner, béhaviorisme).

Dans le conditionnement répondant, il y a établissement d'une connexion entre le stimulus et la réponse. L'expérience la plus connue est celle effectuée par Pavlov sur un chien. Dans cette expérience, chaque fois que l'on présente de la nourriture au chien, ce qui le fait saliver, on fait simultanément tinter une sonnette. Par la suite, on ne fait plus que tinter la sonnette et à cette seule sonnerie (stimulus), le chien se met à saliver (réponse).
Dans le conditionnement opérant ou instrumental, l'expérience est différente. Le conditionnement instrumental est basé sur le stimulus consécutif à une action spécifique du sujet : le stimulus peut-être positif et il y a renforcement, l'action a alors une plus forte probabilité de réapparaître, ou être négatif et il y a inhibition, l'action a alors une plus faible probabilité de réapparaître. L'illustration la plus courante qui en est donnée est celle du pigeon. L'animal enfermé dans une cage, agissant de manière aléatoire, finit par taper du bec sur une cible (action du sujet), ce qui libère de la nourriture (stimulus de renforcement). L'action du pigeon consistant à taper sur la cible a alors une plus forte probabilité de se répéter.

Dans l'apprentissage par *construction*, le sujet apprenant doit être capable de saisir des informations nouvelles, de les traiter et de les mémoriser afin de parvenir à un comportement nouveau. Il y aura dans un premier temps construction d'un nouveau comportement par un travail d'analyse et de synthèse, plus ou moins conscientes, dans un second temps, il y aura intégration de ce nouveau comportement au registre des comportements disponibles [Berbaum, 94].

On retrouve plus ou moins cette distinction entre les deux types d'apprentissage (conditionnement ou construction) chez Houdé qui parle lui d'*apprentissages élémentaires* par opposition aux *apprentissages complexes* médiatisés par des représentations symboliques [Houdé et al, 98]. Ainsi, parmi les divers types d'apprentissage suivants que propose Weil-Barais :
- l'empreinte (éthologie) : comportement inné déclenché par des objets spécifiques,
- l'habituation : capacité d'apprendre à ne pas réagir à certains stimulus connus non-porteurs d'information nouvelle,
- l'apprentissage associatif :
  par essais-erreurs : la bonne réponse est de plus en plus rapidement trouvée,
  par conditionnement : pavlovien, skinnérien (instrumental, opérant) voir ci-dessus,
- l'apprentissage par l'action : action du sujet comme source d'information (dont le traitement de l'information)
- l'apprentissage par observation et imitation de l'autre,
- l'apprentissage coactif : c'est en agissant avec l'autre que l'on apprend,
- l'apprentissage par tutorat et par instruction tel l'enseignement scolaire,

Houdé regroupe les trois premiers types dans les apprentissages élémentaires et les quatre suivants dans les apprentissages complexes.

Quelle que soit la façon de classifier l'apprentissage en psychologie, nous pouvons constater que ces descriptions sont assez éloignées de ce que les théories formelles envisagent. *De fait, le terme d'apprentissage est quelque peu usurpé par celles-ci, car elles ne considèrent principalement que l'apprentissage inductif. En psychologie, ce qui approche le plus, alors, de ce type d'apprentissage est la catégorisation (cf.[Pitt, 97], [Boucheron, 92]),* mécanisme qui comme on le verra au chapitre 2, est à la base de la plupart des processus cognitifs, y compris l'apprentissage tel qu'il est défini ci-dessus.

Théories formelles de l'apprentissage et catégorisation

Nous pouvons distinguer deux objectifs à l'origine des théories formelles. D'une part, comme nous l'avons dit, essayer de poser des bornes théoriques en apprentissage automatique car il est toujours intéressant de connaître a priori les limites que l'on va rencontrer lorsque l'on construit un système apprenant. De l'autre, essayer d'appréhender ce qu'est l'apprentissage inductif, y compris l'apprentissage inductif

naturel[6]. C'est sur ce deuxième aspect que nous allons travailler ici : nous allons étudier les rapports qui peuvent exister entre le modèle d'apprentissage Probablement Approximativement Correct (PAC) de Valiant et la catégorisation en psychologie.

Comme nous l'avons dit, il n'existe pas de méthodologie pour ce type de recherche. Néanmoins si nous montrons, aussi bien au niveau des concepts que des résultats obtenus, que les deux disciplines convergent, nous pourrons estimer que les rapprochements que nous effectuons sont fondés. Cependant il ne suffit pas d'établir des parallèles, encore faut-il que ces parallèles soient utiles à chaque discipline, c'est le principe même des sciences cognitives. Ainsi quelques idées clefs ont orienté ce travail.

Pour l'informaticien, par le fait qu'il permet une lecture des recherches en psychologie sur la catégorisation, le modèle PAC est légitimé. Dans le même temps, cette lecture invite à proposer quelques modifications du modèle et pose quelques questions. Parmi celles-ci, sous-jacente au modèle, est la question de savoir ce qu'est l'extension d'un concept lorsque l'on fait intervenir des distributions de probabilités sur ses éléments. Par exemple, quelle est la différence entre l'ensemble vide et un singleton $\{a\}$ tel que la probabilité de a est nulle ?

Pour le psychologue, le modèle, étant légitimé, peut éventuellement l'intéresser par les notions qui lui sont inhérentes. Il invite d'abord à ne pas dissocier le processus (apprentissage/catégorisation) et le résultat de ce processus (la représentation de la catégorie). Il invite aussi à s'interroger sur la nature de ce processus et les conditions dans lesquelles il se déroule, en suggérant qu'au moins cinq aspects les caractérisent :
- la catégorisation est un apprentissage, une adaptation de l'individu à l'environnement,
- l'environnement impose à l'individu les catégories qu'il doit former,
- l'adaptation, l'apprentissage n'est pas optimal, il est approximatif,
- l'environnement est complexe et l'individu a des capacités finies ; de là l'économie cognitive : apprendre c'est comprimer très fortement l'information, ce que montrent les théories formelles,
- l'environnement n'est pas quelconque ; un individu ne rencontre pas tous les objets de ce monde, de plus certains sont plus fréquents que d'autres dans le milieu où il vit. Son environnement obéit donc à une distribution de probabilités donnée ; de là, la typicalité qui est, entre autres, une expression de cette distribution. La typicalité trouve ainsi sa raison d'être dans le fait qu'elle permet une meilleure adaptation de l'individu à son environnement.

Les trois premiers aspects sont étudiés dans le chapitre 2, les deux suivants dans le chapitre 3. Le chapitre 5 part alors de l'idée que puisque ces principes sont inhérents au modèle PAC, il serait intéressant de le transformer en protocole expérimental.

Ainsi, dans le chapitre 1, nous présentons le modèle de Valiant d'apprentissage PAC (Probablement Approximativement Correct) que nous avons esquissé ci-dessus. La qualité d'un modèle formel tient autant de ses résultats, les théorèmes, que de la qualité des différentes définitions qu'il propose. C'est pourquoi dans ce chapitre nous

[6] Dans son article [Gold, 67] visait explicitement l'apprentissage de la syntaxe de la langue.

décortiquons la définition du modèle PAC, afin que chacun des concepts qu'il sous-tend soit bien appréhendé. Préalablement, pour que le lecteur puisse fixer ces différents concepts, nous présentons un exemple simple d'apprentissage sur lequel s'appuie cette analyse. Puis, nous expliquons les limites de notre présentation, les variations du modèle et les résultats obtenus dans celui-ci. Enfin, nous en listons les notions clefs que nous devons retrouver dans les recherches sur la catégorisation en psychologie pour pouvoir affirmer que les deux disciplines tentent d'approcher le même phénomène.

Le chapitre 2 consiste alors en une lecture, filtrée par le modèle PAC, des recherches en psychologie sur la catégorisation. Nous commençons par des généralités : définitions d'une catégorie et de la catégorisation, rapide historique des recherches concernant celles-ci et présentation de trois expériences que nous pensons caractéristiques des époques et des problématiques. Nous continuons en nous intéressant plus particulièrement au pourquoi de la catégorisation, quelle utilité celle-ci a-t-elle pour l'individu ? Une autre question, celle du comment, nous entraîne dans l'étude des représentations : comment sont représentés chez l'individu les objets du monde, comment sont représentées les catégories ? Enfin, le modèle d'apprentissage PAC, en plus d'un processus, définit aussi les conditions dans lesquelles se déroule ce processus, cela nous amène à nous préoccuper des aspects méthodologiques de ces recherches.

Ainsi les chapitres 1 et 2, outre le fait de présenter les recherches dans les deux domaines, sont l'occasion de montrer que l'on retrouve des concepts semblables dans les deux disciplines. Le chapitre 3, lui, est davantage concerné par les résultats. Dans ce chapitre, nous étudions deux points de convergence entre psychologie et théories formelles concernant l'apprentissage / la catégorisation.
Le premier est la notion d'économie que l'on retrouve en psychologie par le biais de l'économie cognitive et en informatique par celui du rasoir d'Occam.
Le second est la notion de typicalité en psychologie que l'on met en parallèle avec celle de représentativité en informatique. Ce point est développé plus longuement car il nous amène à revenir en psychologie sur les notions de prototype et de similarité. En informatique, la notion de représentativité, nous conduit à présenter le modèle d'apprentissage PAC avec distributions bienveillantes et sa source, celui d'apprentissage PAC-simple de Li et Vitanyi. Nous concluons le chapitre en mettant en évidence les similitudes et les différences qui existent entre typicalité en psychologie et représentativité en informatique et le lien qui pourrait exister entre les deux.

Le chapitre 4, est un chapitre technique. Il donne, de manière plus formelle que dans le chapitre 3, la définition du modèle d'apprentissage PAC avec distributions bienveillantes, les théorèmes obtenus dans ce modèle et leurs démonstrations. Les distributions bienveillantes facilitent l'apprentissage en favorisant l'apparition des exemples représentatifs du concept. Par ailleurs, le modèle y est mis en parallèle avec d'autres modèles, ce qui permet ainsi d'importer leurs résultats. Le survol de ce chapitre par les non-initiés pourra leur donner une idée de la démarche employée en théories formelles.

Enfin, dans le chapitre 5, nous proposons d'opérationnaliser le modèle PAC c'est-à-dire de le transformer en protocole d'expérimentation en psychologie. Nous exposons dans un premier temps l'intérêt d'une telle démarche aussi bien pour l'informatique que pour la psychologie. Puis nous analysons la définition de l'apprentissage PAC pour y retrouver les variables expérimentales sous-jacentes. Nous proposons ensuite une définition de ce que peut être «un concept appris de manière Approximativement Correcte» en psychologie et un algorithme d'expérimentation dérivé du modèle PAC. Nous terminons par le compte-rendu de deux expériences construites sur le protocole défini précédemment. La première est destinée à montrer que les distributions de probabilités, selon lesquelles on tire les exemples que l'on présente à l'apprenant, jouent sur la facilité de l'apprentissage (problème relevant des théories formelles). La seconde, était prévue pour mettre en évidence que la fréquence d'apparition des valeurs d'attributs intervient dans le poids de celles-ci pour définir le prototype (problème relevant de l'étude de la catégorisation en psychologie). De fait, elle montre surtout les lacunes de l'opérationnalisation et les questions que celles-ci posent quant à la nature du processus de catégorisation.

Dans la conclusion, nous présentons les limites de ce travail et les nombreuses pistes de recherche qu'il ouvre.

Au niveau de l'écriture, nous avons souhaité que chacune des disciplines puisse avoir une idée de ce qui se fait dans l'autre. Toutefois, le point de départ est l'informatique et c'est par le biais de la modélisation de l'apprentissage proposée par Valiant que sont présentées les recherches en psychologie. Sachant que toute discipline a une structuration propre, cela amènera parfois à transgresser cette structuration. Nous espérons que lorsque nous le faisons, nous ne trahissons pas pour autant la pensée, les principes de celle-ci.

Chapitre 1

Présentation du modèle d'apprentissage Probablement Approximativement Correct

1 Introduction

Selon Natarajan [Natarajan, 91], les définitions habituelles de l'apprentissage (psychologiques, pédagogiques, etc.) manquent de rigueur[5] *lorsqu'il s'agit de définir dans quelles conditions on apprend.* Définir les conditions consiste à définir l'environnement mais aussi les hypothèses que l'on fait sur l'apprenant. L'objet des théories formelles de l'apprentissage est de définir ces conditions. Un modèle formel d'apprentissage définit un cadre d'apprentissage qui décrit les conditions précises dans lesquelles se déroule l'apprentissage, et essaie de trouver, pour chaque classe de concepts, un apprenant susceptible d'apprendre cette classe, dans ces conditions. Comme nous l'avons vu dans l'introduction, dans la majeure partie des travaux en théories formelles, l'apprentissage se limite à l'inférence inductive : étant donné une séquence d'exemples présentée au système apprenant, celui-ci doit retourner le concept dont ces exemples sont des instances.

---

[7] Nous ne sommes pas sûr que Natarajan soit conscient du côté provocateur de son assertion. Nous l'avons reprise non pas pour poursuivre la provocation mais parce qu'elle mettait en évidence quel était le principal objet d'investigation des théories formelles.

Dans l'introduction, nous avons brièvement présenté le modèle d'apprentissage à la limite de [Gold, 67], les critiques que l'on pouvait lui faire et comment ces critiques amènent au modèle de Valiant. Le modèle que nous décrivons dans ce chapitre est donc celui de l'apprentissage PAC (apprentissage probablement approximativement correct) défini par [Valiant, 84]. Dans cette modélisation, on considère que l'apprentissage peut échouer (apprentissage probable mais non sûr) et que l'hypothèse apprise ne correspond pas exactement au concept (apprentissage approximatif).

Dans un premier temps nous proposons un exemple d'apprentissage qui permettra de se faire une intuition du modèle. Ensuite nous donnerons la définition du modèle d'apprentissage PAC et nous l'analyserons. Nous verrons ainsi se préciser les différentes notions impliquées dans le modèle. Puis nous présenterons des variantes du modèle. Nous terminerons en exposant ce qui, de notre point de vue, est important de retenir dans le modèle PAC.

1.1 Un exemple d'apprentissage

Avant d'aborder un décodage du formalisme du modèle d'apprentissage PAC et afin d'ancrer les notions clefs de ce modèle, nous proposons un exemple simple d'apprentissage. Nous allons considérer un système cognitif extrêmement simplifié, que nous appellerons RobSimp (pour robot simple ou simplet, le lecteur choisira) que nous allons faire évoluer dans un environnement tout aussi simplifié. RobSimp devra s'adapter à cet environnement et notamment apprendre à reconnaître les objets de cet environnement qui lui sont nocifs.

1.1.1 Comment RobSimp perçoit le monde

RobSimp ne connaît le monde et ses objets qu'au travers des perceptions qu'il en a. Pour connaître ce monde, RobSimp n'a à sa disposition que 4 capteurs sensoriels : le premier lui permet de percevoir la forme, le second la taille, le troisième la couleur et le quatrième les flèches. A ces quatre capteurs sensoriels nous en ajoutons un cinquième, un peu à part, que l'on va appeler « étiquette », qui pourra être considéré par RobSimp comme un attribut du genre «nocif». En plus d'être limité au niveau du nombre de capteurs, RobSimp est aussi limité quant aux valeurs que donnent ces capteurs. Son capteur forme indique à RobSimp si les objets sont ronds ou carrés, son capteur taille, s'ils sont grands ou petits, son capteur couleur s'ils sont noirs ou blancs et enfin son capteur flèche lui indique l'orientation de la flèche. Le capteur « étiquette », lui, indique pour sa part si les objets sont nocifs ou non pour lui[8]. Le tableau 1.1.1 page suivante résume les perceptions possibles de RobSimp.

[8] Avouons-le, RobSimp en plus d'être primaire, est un peu binaire, la raison en est dans le manque de crédits du laboratoire. Mais, par ailleurs, nous, êtres humains, qui sommes des êtres hautement perfectionnés nous ne percevons pas les ultrasons, ni les infrarouges. Alors ne jetons pas la pierre à RobSimp/Marie Madeleine.

attribut	forme	taille	couleur	flèche
Valeur 1	Carré	Petit	Blanc	Intérieure
Valeur 0	Rond	Grand	Noir	Extérieure

Tableau 1.1.1
Les attributs et leurs valeurs qui permettent à RobSimp d'appréhender le monde.

Nous notons que l'attribut « nocif» n'apparaît pas dans le tableau car cet attribut ne dépend pas des objets en eux-mêmes, mais de l'interaction de RobSimp avec ces objets.

1.1.2 L'environnement de RobSimp

Nous voyons que le monde tel que RobSimp le perçoit n'est constitué que de 16 (2^4) objets *différents possibles* (voir la figure 1.1.1, page suivante), mais RobSimp n'est pas amené à tous les rencontrer. Certains objets sont plus fréquents dans son environnement et d'autres plus rares, au même titre qu'il est plus rare de croiser une vipère dans le Nord de la France que dans le Sud. La probabilité qu'il a de les rencontrer est indiquée au-dessous de chaque objet dans la figure 1.1.1.

1.1.3 L'espace d'hypothèses de RobSimp

L'objectif pour RobSimp va être de survivre dans cet environnement, de s'y adapter. Certains des objets qu'il va rencontrer lui sont nocifs et d'autres non. Il doit donc très rapidement être capable de catégoriser cet environnement en fonction de la catégorie « nocif », s'il veut pouvoir survivre.

Une première possibilité pour lui pourrait consister à mettre en mémoire, au fur et à mesure qu'il les rencontre, une représentation de chacun des exemples qui lui sont nocifs. Chaque fois qu'il rencontre un élément, il va vérifier s'il en a ou non une représentation dans sa mémoire ; si c'est le cas, il ne s'approche pas de l'objet s'il est nocif. Le problème avec cette démarche, c'est d'abord que, comme le laboratoire est vraiment très pauvre, nous n'avons pas pu munir RobSimp d'assez d'espace mémoire. Ensuite, RobSimp n'est pas masochiste, il aimerait bien être capable d'anticiper et de *prédire* si un *nouvel* élément qu'il n'a jamais vu précédemment est nocif ou pas.

C'est pourquoi nous avons muni RobSimp d'un *espace d'hypothèses*. Il est ainsi capable de formuler des hypothèses, mais encore une fois en raison de la restriction de crédits, cet espace est très rudimentaire. Les seules hypothèses que RobSimp est capable de faire sont des conjonctions de valeurs d'attributs, autrement dit des « et » ($\wedge$), le « ou » ($\vee$) étant hors de prix.

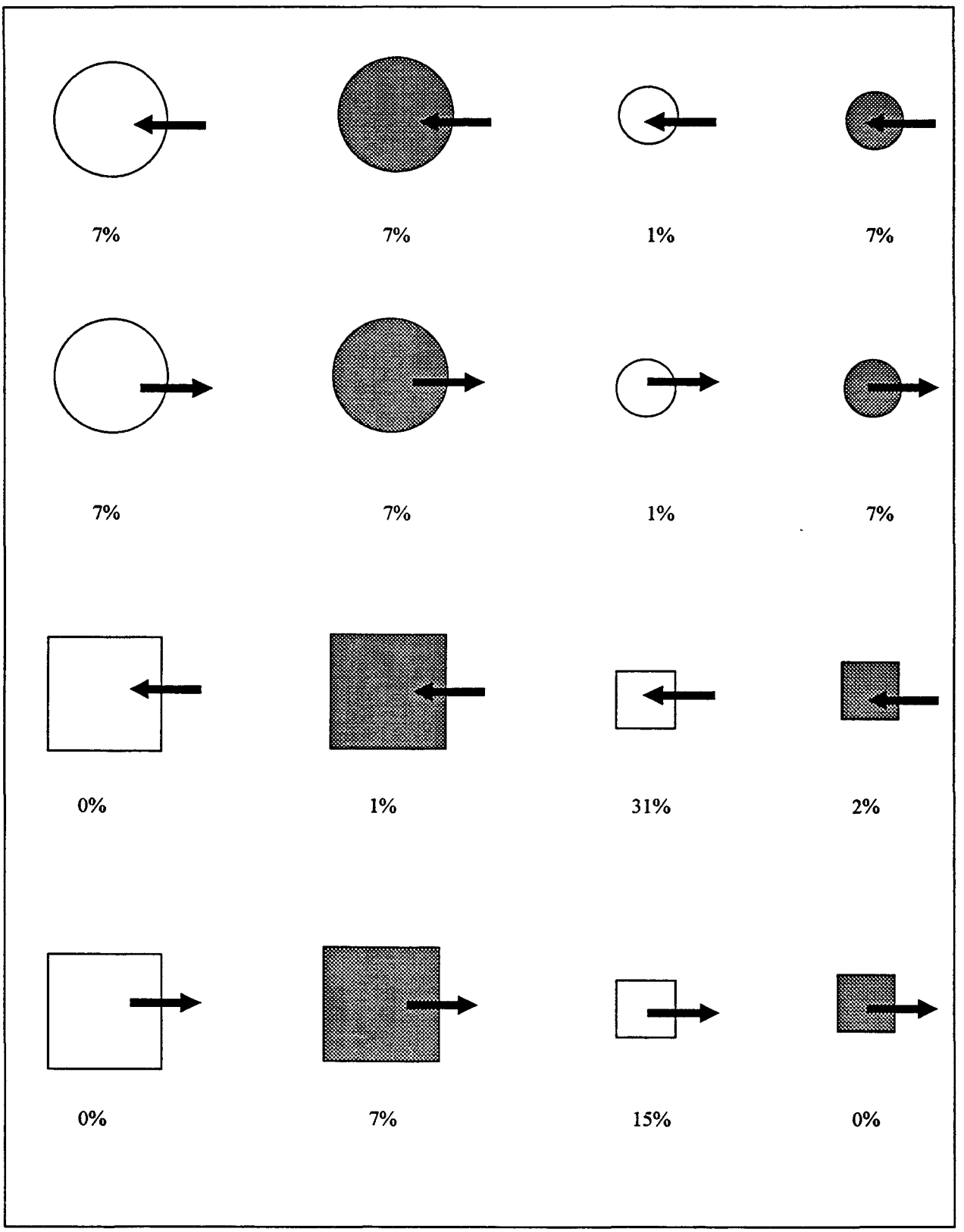


Figure 1.1.1

Les objets possibles de l'environnement de RobSimp tels qu'il peut les percevoir. Les pourcentages en dessous indiquent les probabilités qu'il a de les rencontrer.

1.1.4 La classe de concepts cible

Nous voudrions que RobSimp soit capable de s'adapter à son environnement, c'est-à-dire de distinguer les éléments nocifs de cet environnement des autres. Il faut donc définir quels sont les éléments nocifs. Les possibilités sont très nombreuses, autant qu'il y a de sous-ensembles possibles du monde de RobSimp. Nous ne sommes cependant pas mesquins et nous savons que RobSimp est limité aux conjonctions de valeurs d'attributs pour ses hypothèses, aussi nous choisissons un concept qui est une conjonction de deux valeurs d'attributs. Ce qui nous laisse le choix entre 24 concepts possibles (8 x 6 / 2) : R∧P, R∧G, R∧B, R∧N, R∧I, R∧E, C∧P, C∧G, C∧B, C∧N, C∧I, C∧E, P∧B, P∧N, P∧I, P∧E, G∧B, G∧N, G∧I, G∧E, B∧I, B∧E, N∧I, N∧E [9]. Choisissons C∧P, « carré et petit », (la deuxième colonne du tableau 1.1.2, page suivante indique quels sont parmi les 16 éléments ceux qui sont nocifs). Comme nous sommes sûrs que RobSimp est capable de formuler toutes les conjonctions possibles de valeurs d'attribut, nous savons qu'il est donc capable de former chacun des 24 concepts possibles et donc le concept cible. La classe de concepts cible est incluse dans son espace d'hypothèses.

1.1.5 L'échantillon que RobSimp rencontre

Il s'agit maintenant de voir comment RobSimp va s'adapter à son environnement et d'imaginer les interactions entre l'un et l'autre. Il faut donc définir les éléments qui participeront à son apprentissage, ceux qui lui permettront de forger son hypothèse. Pour cela nous allons tirer selon la distribution de probabilités 13 éléments que RobSimp va rencontrer. A l'issue de sa rencontre, il devra avoir formulé une hypothèse qui lui permette de distinguer entre les éléments qui lui sont nocifs et ceux qui ne le sont pas. Le tableau 1.1.2 indique les éléments qui ont été tirés, on les retrouve dans la figure 1.1.2 (page 37).

On note que certains exemples sont présentés plusieurs fois, que d'autres ne le sont qu'une fois et que d'autres enfin ne le sont pas du tout. Cela provient du fait que pour établir cet échantillon, nous avons dû le faire selon la distribution de probabilités que nous retrouvons dans la troisième colonne du tableau. On constate que certains exemples qui ont une probabilité non nulle d'être tirés n'apparaissent pas pour autant dans l'échantillon, cela est dû à la taille de l'échantillon qui est de 13 exemples. Par exemple le carré petit noir avec une flèche intérieure (CPNI) a une probabilité de 2% et n'apparaît pas dans les 13 exemples.

---

[9] Une hypothèse est une conjonction (∧) de deux valeurs d'attribut décrites par leur initiale : R pour rond, C pour Carré, P pour petit, G pour Grand, N pour Noir, B pour Blanc, E pour flèche Extérieure, et I pour Intérieure.

figure	Etiquetage selon le concept C∧P « + » = « nocif »	distribution de probabilités	nombre d'occurrences dans l'échantillon
RGNE	-	7%	1/13
RGNI	-	7%	1/13
RGBE	-	7%	1/13
RGBI	-	7%	1/13
RPNE	-	7%	1/13
RPNI	-	7%	1/13
RPBE	-	1%	0/13
RPBI	-	1%	0/13
CGNE	-	7%	1/13
CGNI	-	1%	0/13
CGBE	-	0%	0/13
CGBI	-	0%	0/13
CPNE	+	0%	0/13
CPNI	+	2%	0/13
CPBE	+	15%	2/13
CPBI	+	31%	4/13

Tableau 1.1.2 : La première colonne indique l'exemple par ses initiales (RGNE : Rond Grand Noir flèche Extérieure). La seconde colonne donne son étiquetage selon le concept C∧P. La troisième donne sa probabilité d'être tiré. La dernière indique le tirage réel qui a lieu pour présenter l'échantillon.

Il faut noter aussi que l'échantillon « correspond » relativement bien à la distribution de probabilités, ce n'est pas toujours le cas. Si vous tirez 6 fois à pile ou face avec une pièce non biaisée, il est probable que vous n'obteniez pas 3 piles et 3 faces, mais par exemple 2 piles et 4 faces. Pourtant la probabilité d'apparition pour chacune des faces dans un tirage avec une pièce non biaisée est de 1 possibilité sur 2. Si l'échantillon avait été de 100 exemplaires, il est possible alors que la figure CPNI en fasse partie, de la même manière que plus le nombre de tirages à pile ou face est grand plus la fréquence de piles et la fréquence de faces approche de ½. Notons, enfin, que RobSimp n'a pas accès à cette distribution de probabilité. La seule chose à laquelle il ait accès, c'est l'échantillon.

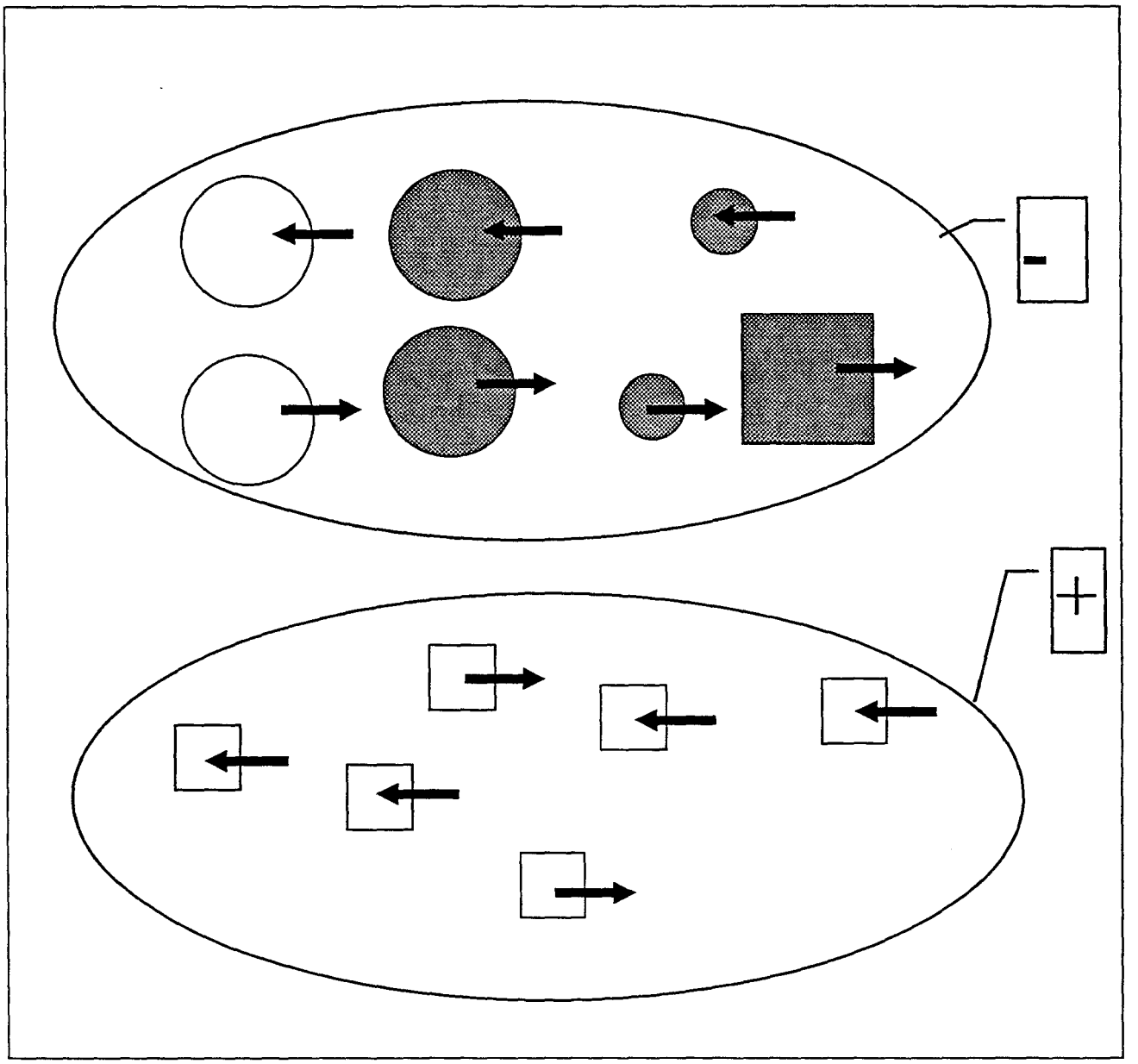


Figure 1.1.2
Ci-dessus, l'échantillon d'exemples étiquetés selon le concept cible. Le concept cible est une conjonction de deux valeurs d'attributs C∧P. Les éléments marqués « + » sont nocifs à RobSimp. L'apprentissage pour RobSimp consiste sur la base de cet échantillon à retrouver une hypothèse proche du concept.

1.1.6 Le calcul de l'erreur

Vu les limites de RobSimp nous n'allons pas lui demander d'apprendre exactement le concept, mais seulement approximativement. C'est-à-dire que son hypothèse pourra être partiellement erronée, son adaptation à l'environnement ne sera pas optimale. Considérons que nous acceptons que s'il ne se trompe qu'au plus 5% du temps son hypothèse est approximativement correcte. Mais que veut dire une erreur d'au plus 5% ? La réponse est que lorsqu'il rencontre 100 éléments, il n'est autorisé à en classer mal que 5 au plus. *Sa marge d'erreur est donc dépendante des probabilités avec lesquelles il rencontre les éléments de son monde. Le poids de l'erreur n'est donc pas le nombre d'éléments, parmi les 16 possibles, qui seront mal classés mais la somme des probabilités de chacun de ces éléments.* Il est peu important pour un habitant des villes du Nord d'être capable de distinguer entre une couleuvre et une vipère, il n'en rencontre jamais.

Nous avons vu qu'il y avait 24 conjonctions de 2 valeurs d'attributs. Pour chacune de ces hypothèses, il est possible de calculer le poids de son erreur avec le concept cible. L'erreur entre cette hypothèse et le concept cible est constituée des exemples qui ne sont pas étiquetés de la même façon par l'une et par l'autre. Le poids de l'erreur est la somme des probabilités des éléments qui seront classés différemment.

figure	distribution de probabilités	**concept C∧P**	hypot. C∧I	Erreur de C∧I	poids de l'erreur de C∧I	Hypot. C∧B	Erreur de C∧B	poids de l'erreur de C∧B
RGNE	7%	-	-	non		-	non	
RGNI	7%	-	-	non		-	non	
RGBE	7%	-	-	non		-	non	
RGBI	7%	-	-	non		-	non	
RPNE	7%	-	-	non		-	non	
RPNI	7%	-	-	non		-	non	
RPBE	1%	-	-	non		-	non	
RPBI	1%	-	-	non		-	non	
CGNE	7%	-	-	non		-	non	
CGNI	**1%**	-	+	**oui**	**1%**	-	non	
CGBE	*0%*	-	-	non		+	*oui*	*0%*
CGBI	**0%**	-	+	**oui**	**0%**	+	*oui*	*0%*
CPNE	**0%**	+	-	**oui**	**0%**	-	*oui*	*0%*
CPNI	2%	+	+	non		-	*oui*	*2%*
CPBE	**15%**	+	-	**oui**	**15%**	+	non	
CPBI	31%	+	+	non		+	non	
			erreur	**totale**	**16%**	**erreur**	**totale**	**2%**

Tableau 1.1.3
Calcul du poids de l'erreur de 2 hypothèses, «carré et flèche intérieure » (C∧I) et «carré et blanc » (C∧B) par rapport au concept cible «carré et petit » (C∧P).

RobSimp va devoir apprendre le concept «carré et petit » (C∧P). Le tableau 1.1.3 calcule le poids de l'erreur de 2 hypothèses qu'il pourrait émettre, l'hypothèse «carré et flèche intérieure » (C∧I) et l'hypothèse «carré et blanc » (C∧B), en fonction de la distribution de probabilités que nous avons définie.

Si RobSimp émet l'hypothèse C∧I nous considérerons qu'il a échoué dans son apprentissage car cette hypothèse fait 16% d'erreur (supérieure aux 5%), par contre s'il émet l'hypothèse C∧B nous considérerons qu'il a appris. Pourtant ces deux hypothèses sont erronées sur le même nombre d'éléments : 4. Il faut comprendre que si nous agissons ainsi, c'est pour le bien de RobSimp. En apprenant l'hypothèse C∧B, RobSimp ne fera des rencontres désagréables que 2 fois sur 100, tandis qu'avec C∧I, ce sera 16 fois sur 100, ce qui pourrait lui être fatal. La différence entre ces 2 hypothèses, C∧I et C∧B, repose sur la probabilité des éléments mal classés. En retenant C∧I, chaque fois que RobSimp rencontrerait l'exemple CPBE, il le classerait comme négatif, comme ne relevant pas du concept cible, (n'étant pas nocif) alors qu'il l'est. Cet exemple a une probabilité de 15%, cela veut donc dire que RobSimp se tromperait 15 fois sur cent. A l'opposé en retenant C∧B, RobSimp classerait mal l'exemple CPNI, mais celui-ci n'a une probabilité que de 2%, c'est-à-dire que RobSimp ne se tromperait que 2 fois sur cent, ce qui est tolérable. Il faut noter que, de la même manière, il y a des erreurs qui ne prêtent pas à conséquence, ce sont tous les éléments qui ont une probabilité de 0% ce qui signifie que RobSimp ne peut pas être amené à les rencontrer, et donc à se tromper en les classant.

1.1.7 L'algorithme d'apprentissage

Nous avons décrit l'environnement de RobSimp (l'ensemble des éléments qui constituent cet environnement et leurs probabilités d'apparition). Nous avons décidé quels étaient, dans cet environnement, les éléments qui lui étaient nocifs (le concept cible). Nous avons précisé les éléments de cet environnement qu'il va rencontrer (l'échantillon). Il reste maintenant à RobSimp à s'adapter à cet environnement, à formuler une hypothèse qui lui permette de classer les éléments de cet environnement en « nocifs » ou « non–nocifs ». Pour formuler cette hypothèse les seules informations dont bénéficie RobSimp sont l'échantillon d'exemples étiquetés selon le concept cible. Il n'a directement accès ni au concept cible lui-même, ni aux distributions de probabilités qui régissent son univers. Pour l'instant tout ce qu'il peut faire c'est percevoir les objets de cet environnement avec leur étiquette (nocif ou non) et formuler une hypothèse qui soit une conjonction. Lorsque nous disons qu'il est capable de formuler une hypothèse cela veut dire qu'il est seulement capable, à partir de cette hypothèse, de classer tous les éléments de son environnement selon cette hypothèse. Il reste encore à le rendre capable, sur la seule information donnée par les exemples de trouver une hypothèse consistante avec ces exemples, qui les étiquette de la même manière que le concept. Cette capacité est l'algorithme d'apprentissage.

Un algorithme d'apprentissage possible pour RobSimp va consister à démarrer avec l'hypothèse, la conjonction, la plus spécifique qui contient toutes les valeurs des

variables possibles. Ici ce sera C∧R∧P∧G∧B∧N∧I∧E [10]. Au fur et à mesure de la présentation des exemples *positifs*, RobSimp élimine de son hypothèse les valeurs d'attributs qui ne sont pas compatibles avec ces exemples, c'est-à-dire qu'il généralise cette hypothèse.
Le premier exemple est CPBI, il élimine toutes les valeurs d'attributs de l'hypothèse incompatibles avec cet exemple et il obtient : C∧~~R~~∧P∧~~G~~∧B∧~~N~~∧I∧~~E~~ soit C∧P∧B∧I. Cette élimination correspond en français au fait qu'un exemple carré n'est pas compatible avec une hypothèse qui suppose que les exemples soient ronds, qu'un exemple petit n'est pas compatible avec une hypothèse qui suppose que les exemples soient grands, etc.
Il prend un second exemple positif, différent du premier, CPBE et élimine de la même manière les valeurs d'attribut incompatibles. Il n'y en a qu'une, « I », ce qui donne C∧P∧B∧I ce qui lui laisse comme hypothèse C∧P∧B.
Comme il a épuisé tous les exemples positifs différents, son hypothèse est donc C∧P∧B. On peut vérifier qu'elle est compatible avec les exemples négatifs, c'est-à-dire qu'elle les étiquette, elle aussi, tous comme négatifs.

Le tableau 1.1.4, indique que son hypothèse est approximativement correcte, car le poids de son erreur est de 2% et donc inférieur à 5%

figure	distribution de probabilités	**concept C∧P**	hypot. C∧P∧B	Erreur de C∧P∧B	poids de l'erreur de C∧P∧B
RGNE	7%	-	-	Non	
RGNI	7%	-	-	Non	
RGBE	7%	-	-	Non	
RGBI	7%	-	-	Non	
RPNE	7%	-	-	Non	
RPNI	7%	-	-	Non	
RPBE	1%	-	-	Non	
RPBI	1%	-	-	Non	
CGNE	7%	-	-	Non	
CGNI	1%	-	-	Non	
CGBE	0%	-	-	Non	
CGBI	0%	-	-	Non	
CPNE	**0%**	+	-	**Oui**	**0%**
CPNI	**2%**	+	-	**Oui**	**2%**
CPBE	15%	+	+	Non	
CPBI	31%	+	+	Non	
			Erreur totale		**2%**

Tableau 1.1.4
Calcul du poids de l'erreur de l'hypothèse apprise par RobSimp, «carré et petit et blanc » (C∧P∧B) par rapport au concept cible «carré et petit » (C∧P).

[10]Notons que cette hypothèse est vide, c'est-à-dire qu'elle ne peut contenir aucun élément. En effet aucun objet ne peut être à la fois rond ET carré, grand ET petit, etc.

En formulant l'hypothèse C∧P∧B, RobSimp a donc été capable d'apprendre approximativement le concept cible C∧P. Cela nous rassure, il est apte à survivre dans cet univers impitoyable.

1.1.8 Question métaphysique : mais quel concept a réellement appris RobSimp ?

RobSimp a formulé l'hypothèse C∧P∧B et nous considérons qu'il a appris approximativement le concept C∧P. Mais ne peut-on pas considérer qu'il a tout aussi bien appris le concept C∧B. Si nous observons l'échantillon que nous avons soumis à RobSimp précédemment, nous verrons que C∧P et C∧B l'étiquettent de la même manière. De plus si nous reprenons le précédent tableau et que nous retirons de ce tableau tous les éléments qui ont une probabilité nulle d'être tirés nous constatons qu'un seul élément distingue C∧P de C∧B (voir tableau 1.1.5). Cet élément CPNI a une probabilité de 2%, c'est pourquoi il n'apparaît pas dans l'échantillon. Nous pouvons donc tout aussi bien considérer que RobSimp a appris *approximativement* le concept C∧B plutôt que C∧P.

figure	distribution de probabilités	**concept C∧P**	**concept C∧B**
RGNE	7%	-	-
RGNI	7%	-	-
RGBE	7%	-	-
RGBI	7%	-	-
RPNE	7%	-	-
RPNI	7%	-	-
RPBE	1%	-	-
RPBI	1%	-	-
CGNE	7%	-	-
CGNI	1%	-	-
CPNI	**2%**	+	-
CPBE	15%	+	+
CPBI	31%	+	+

Tableau 1.1.5
Comparaison de deux concepts : «carré et petit » (C∧P) et «carré et blanc » (C∧B). Tous les éléments de probabilité 0% ont été retirés. Un seul élément CPNI distingue alors ces deux concepts mais il ne fait que 2%. On peut donc considérer que ces deux concepts sont approximativement identiques.

La distribution de probabilités a ainsi « déformé » les concepts, au point que C∧P et C∧B sont approximativement semblables. Il aurait d'ailleurs suffi que CPNI ait une probabilité de 0% pour qu'ils soient totalement identiques. Cette déformation se retrouve dans l'échantillon.

Ceci nous amène à dire que *RobSimp n'apprend pas un concept, mais un concept « déformé » par sa distribution de probabilités. Autrement dit RobSimp apprend en même temps le concept et une partie de l'environnement dans lequel il a rencontré ce concept*[11].

1.1.9 Les limites de cet exemple

Nous avons présenté cet exemple pour mettre en place les notions principales de l'apprentissage PAC telles que celles de classe de concept cible, de classe d'hypothèses de l'apprenant, de distributions de probabilité, de calcul de poids de l'erreur et d'algorithme d'apprentissage. Il ne constitue pas une démonstration de l'apprenabilité d'une classe de concepts telle que cela se fait dans le cadre du modèle PAC[12]. Cet exemple est surtout destiné à se faire une intuition de la manière qu'a le modèle PAC d'*envisager* l'apprentissage. Il faut ainsi en signaler les limites.

La première d'entre elles est que pour pouvoir affirmer qu'une classe de concepts est apprenable dans l'apprentissage PAC, il faut démontrer que l'apprentissage a lieu pour *tous* les concepts de la classe, alors qu'ici nous ne l'avons fait que pour l'un d'entre eux. De la même façon le modèle réclame qu'il y ait apprentissage pour *toute* distribution de probabilité, alors qu'ici nous n'en avons envisagé qu'une seule.

La seconde est que nous avons choisi comme classe de concepts cibles, une sous-classe des fonctions booléennes, celle des conjonctions, alors que l'apprentissage PAC ne s'intéresse pas seulement aux fonctions booléennes, encore moins à cette seule sous-classe des conjonctions, ainsi certains chercheurs ([Shapire, 1990], [Pitt, 1989],...) étudient l'apprentissage de langages. Dans la suite et par souci de lisibilité, nous continuerons de présenter le modèle PAC à partir des classes de fonctions booléennes, mais il faut savoir qu'il existe d'autres classes de concepts étudiées dans le cadre PAC. En réalité, dans ce cadre, il est possible de rechercher l'apprenabilité de toute classe clairement définie.

La dernière est que, si nous avons montré en quoi l'apprentissage PAC est *approximatif*, nous n'avons pas montré en quoi il est seulement probable et non sûr. Ce sera fait dans la section 1.2 lorsque nous parlerons du paramètre δ.

Cet exemple d'apprentissage a d'autres limitations que nous verrons par la suite, mais, malgré celles-ci, il va permettre d'illustrer une présentation plus formelle du modèle d'apprentissage PAC.

---

[11] Nous reviendrons sur cet aspect au chapitre 3.
[12] On trouvera une telle démonstration en annexe 2.

1.2 L'apprentissage PAC

Voici comment [Kearns, Vazirani 94] définit formellement l'apprentissage :

Définition de l'apprentissage PAC [13]
Soit C_n *une classe de concepts sur* X_n *(où* X_n *est soit* $\{0,1\}^n$ *soit l'espace euclidien de dimension n* $\Re^n$*), et* H_n *une classe de représentations, soit* $X=\cup_{n\geq 1} X_n$, $C=\cup_{n\geq 1} C_n$ *et* $H=\cup_{n\geq 1} H_n$. *On dit que* C_n *est PAC apprenable s'il existe un algorithme* L *(déterministe ou non) avec la propriété suivante : pour tout concept* $c\in C_n$, *pour toute distribution de probabilité* D *sur* X, *et pour tout* $0<\varepsilon<1/2$ *et* $0<\delta<1/2$ *si L a accès à EX(c,D) et aux entrées* ε *et* δ,*alors avec une probabilité d'au moins* $1-\delta$, L *retourne une hypothèse* $h \in H_n$ *satisfaisant erreur(h)*$\leq\varepsilon$.
Si L *tourne en temps polynomial en n, taille(c),* $1/\varepsilon$ *et* $1/\delta$ *on dit que C est efficacement PAC apprenable.*

L'objectif de cette section est d'aider le lecteur à s'approprier cette définition. Dans le modèle PAC, l'apprentissage est considéré comme l'interaction entre un apprenant et son environnement. Nous allons exposer comment cet environnement est modélisé, comment il peut être découpé en concepts. Nous présenterons aussi ce qu'est l'apprenant, quelles sont ses hypothèses. Nous décrirons comment sont envisagées les interactions entre l'apprenant et l'environnement. Nous serons amenés à faire la distinction entre apprentissage et apprenabilité. Enfin le modèle PAC est avant tout un modèle qui se veut quantificateur, nous montrerons les diverses mesures utilisées.

1.2.1 L'environnement, les concepts, les hypothèses

1.2.1.1 Le monde n'est connu qu'au travers de descriptions

Dans la définition, **X** est l'ensemble des représentations ou des encodages des objets du monde et non les objets eux-mêmes. « We think of X as being a set of encodings of instances or objects in the learner's world » [Kearns, Vazirani, 94]. Quel que soit le système cognitif (naturel ou artificiel) celui-ci n'apprend pas à partir des objets du monde mais sur des *représentations* de ces objets, la représentation d'un objet pouvant être la perception sensorielle qu'a le système cognitif de cet objet[14]. Dans le modèle PAC, on considère que l'apprentissage est un travail dans un espace de descriptions : on part de la représentation des exemples pour trouver une représentation du concept.

---

[13] Cette définition est celle proposée par Kearns qui considère un ensemble de représentations des objets du monde où X_n est soit $\{0,1\}^n$ soit l'espace euclidien de dimension *n* $\Re^n$, on en trouvera une autre à caractère plus général dans [Natarajan, 91].

[14] Dans la suite nous parlerons d'objets du monde pour alléger le discours mais il faudra interpréter en représentations d'objets.

Cette idée est constante dans les travaux en apprentissage automatique (voir entre autres [Dietterich, Michalski, 83])

Dans cette définition-ci [15], les objets sont décrits par des valeurs d'attributs. Lorsqu'on parle de X_n cela signifie que l'on ne considère que les représentations utilisant *n* attributs. Cela aura son importance lorsqu'il s'agira de mesurer l'apprentissage car une représentation de longueur n=2 prend moins de temps à traiter qu'une représentation de longueur n=100. X est alors égal à l'union de tous X_n pour tous les n supérieurs ou égaux à 1 ($X=\cup_{n\geq 1} X_n$).

Si on ne s'intéresse qu'à des attributs binaires, qui ne peuvent avoir que deux valeurs 1 ou 0, alors $X_n = \{0,1\}^n$. Dans l'exemple de la section 1.1, le monde était composé de 16 éléments {RGNE, RGNI, RGBE, RGBI, RPNE, RPNI, RPBE, RPBI, CGNE, CGNI, CGBE, CGBI, CPNE, CPNI, CPBE, CPBI } décrits par 4 attributs binaires : forme (rond/carré), taille (petit/grand), couleur(blanc/noir), flèche(intérieure/extérieure). Si les attributs prennent leurs valeurs sur $\Re$, X_n correspond à l'espace euclidien de dimension *n* : $\Re^n$. C'est le cas lorsque l'on souhaite décrire un objet par sa taille en centimètres ou par sa température en degrés. Les attributs sont alors ce que les psychologues appellent des dimensions. Le modèle peut aussi travailler avec d'autres attributs que ceux décrits ici tels que les attributs à valeurs discrètes : couleur (bleu, blanc, rouge), temps (pluvieux, ensoleillé, brumeux, nuageux)...

Le fait de choisir de travailler avec des attributs binaires, à valeurs discrètes ou prenant leurs valeurs sur $\Re$ dépend de l'utilisateur du modèle et non pas du modèle lui-même. De la même façon, le modèle ne prend pas en charge la sémantique des attributs. Par exemple, si pour le modèle il est indifférent que l'attribut « flèche » utilisé dans l'exemple 1.1 ait les valeurs « gauche/droite » plutôt que les valeurs « présence/absence », il n'en va pas de même pour un sujet. Cela se comprend aisément lorsque l'on observe le schéma suivant. Pour un sujet les dessins du A ont tous les deux 4 valeurs d'attributs tandis que, parmi les dessins du B, celui de gauche a 4 attributs alors que celui de droite n'en a plus que 3. Il n'est pas considéré dans le cas du B que les deux dessins ont l'attribut flèche avec comme valeur « présence/absence », mais que seul le premier bénéficie de cet attribut tandis que le second n'en bénéficie pas.

[15]Cette définition de l'apprentissage PAC est fortement dépendante de la classe de concept cible qui est l'ensemble des fonctions booléennes. Avec d'autres classes de concepts telles que les langages réguliers ou d'autres on ne parle plus d'attributs mais plutôt de mots du langage qui sont définis dans un alphabet.

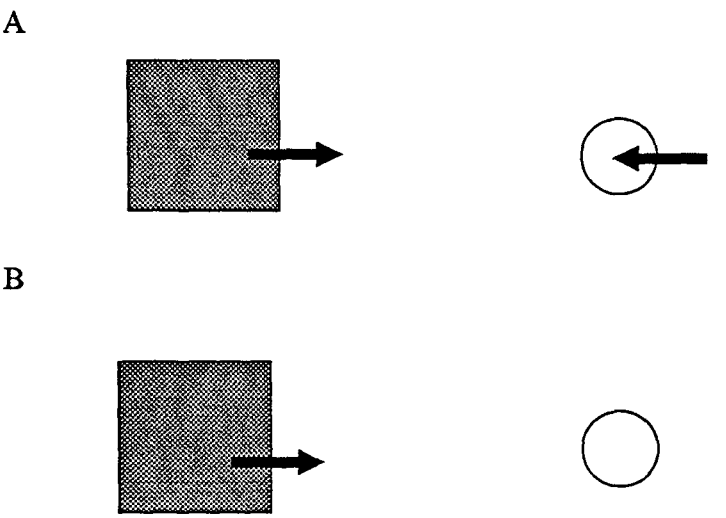


Figure 1.2.1

Pour les figures du A, les sujets considèrent que les deux figures ont 4 attributs binaires (grand/petit, rond/carré, blanc/noir, flèche droite/flèche gauche). En revanche, pour les figures du B, certains sujets considèrent que celle de gauche a quatre attributs alors que celle de droite n'en a que trois : l'absence de flèche n'est pas considérée comme la valeur « absence » de l'attribut flèche mais comme une absence d'attribut. Nous voyons donc que des variables booléennes peuvent avoir des sémantiques très différentes.

1.2.1.2 Le monde de l'apprenant obéit à une distribution de probabilités

Quelle est la probabilité pour le petit Africain de rencontrer un ours polaire ou pour le petit Inuit de rencontrer un baobab ? Un environnement est un ensemble d'objets mais aussi de probabilités sur ces objets. Le petit Inuit et le petit Africain ne rencontrent ni les mêmes arbres, ni les mêmes animaux : la probabilité d'apparition d'un baobab dans l'environnement du petit Inuit est de 0%, de même celle de l'ours polaire dans l'environnement du petit Africain. *L'introduction des distributions de probabilités (D) dans son modèle est un des apports fondamentaux de Valiant à la modélisation de l'apprentissage. De plus, comme nous le verrons, en liant l'évaluation de l'apprentissage à la distribution de probabilités, Valiant réussit à modéliser le fait que l'apprentissage est une adaptation à l'environnement.* C'est un point sur lequel nous reviendrons à plusieurs reprises. Dans l'exemple de la section 1.1, nous donnions les probabilités suivantes à chacun des éléments du micromonde : RGNE, 7%, RGNI, 7%, RGBE , 7%,RGBI, 7%, RPNE, 7%, RPNI, 7%, RPBE, 1%,RPBI, 1%, CGNE, 7%, CGNI, 1%, CGBE , 0%, CGBI, 0%, CPNE, 0%, CPNI, 2%, CPBE, 15%, CPBI, 31%.

1.2.2 Le concept

1.2.2.1 Un concept est un sous-ensemble des objets du monde

Pour les théories formelles, **un concept** est d'abord un sous ensemble des objets du monde, le concept est alors défini en extension. Dans l'exemple de la section 1.1, c'était {CPNI, CPNE, CPBI, CPBE}.

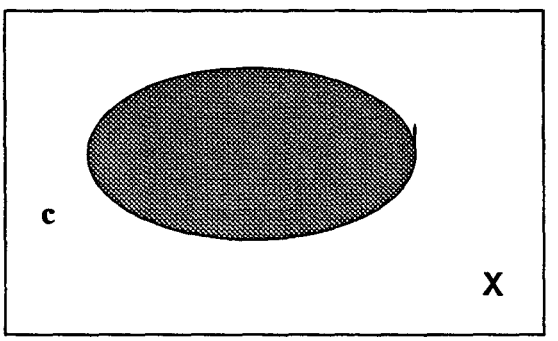


Schéma 1.2.1
X est l'ensemble des objets du monde.
***c* est le concept à apprendre.**

Un concept peut aussi être pensé comme l'ensemble des éléments de X qui vérifient une règle simple ou intéressante. La **représentation** de cette règle constitue la définition en compréhension ou en intension du concept. Dans l'exemple du 1.1, c'était C∧P.

1.2.2.2 Le concept comme représentation

Pour se faire une première idée de ce qu'est la **représentation d'un concept**, il est possible de l'envisager comme la définition qu'en donne un dictionnaire quelconque. Si cette définition est correctement faite, on peut dire si un exemple relève ou non du concept : la définition nous permet de classer les exemples.

Toute représentation repose sur *un schéma de représentation*. Lorsque l'on considère une définition dans un dictionnaire, cette définition repose sur le schéma de représentation qu'est la langue. [Kearns, Vazirani, 94] précise « Formally speaking, a representation scheme for a concept class C is a function R : $\Sigma^* \rightarrow C$ where Σ is a finite alphabet of symbols. [...] We call any string $\sigma \in \Sigma^*$ such that $R(\sigma)=c$ a representation of c (under R). Note that there may be many representations of a concept c under the representation scheme R. ». Un schéma de représentation est une fonction qui associe à une chaîne de caractères un concept. Dans l'exemple présenté dans le 1.1, nous envisageons deux types d'*alphabet*. Le premier est constitué des mots du français :

carré, rond, blanc, noir, etc. Le deuxième est constitué de l'initiale de ces mots : {C, R, G, P, N, B, I, E}. Le schéma de représentations est constitué de la conjonction « et » («∧»). A toute chaîne de caractères construite avec ce schéma de représentations correspond un concept : « carré et petit » (C∧P), « rond et blanc » (R∧B), etc.

Si nous reprenons notre analogie avec la langue, ce qui est défini, ici, comme un alphabet correspond à l'ensemble des mots de la langue. Ce qui est défini comme une chaîne, est toute phrase correctement construite dans cette langue, c'est-à-dire respectant la syntaxe de cette langue. Une représentation d'un concept est alors une définition, une phrase, qui décrit ce concept. Pour un même concept, il est possible d'avoir plusieurs définitions toutes écrites dans la même langue, il suffit de passer du Larousse au Robert. Il faut noter, ici, qu'une représentation de l'animal chat n'est pas le mot «chat », mais une définition qui permet de reconnaître l'animal.

En informatique, il existe différents types de schémas de représentation, de langages : les formules booléennes, les arbres de décisions (voir introduction générale), les listes de décision[16], les automates, etc. Parmi ceux-ci, les réseaux neuronaux (voir introduction générale) illustrent bien la manière dont nous envisageons les représentations. Le réseau de neurones peut être considéré à la fois comme un encodage symbolique et comme une réalité physique (il est possible de créer physiquement un réseau de neurones artificiel). Dans le second cas, le réseau de neurones n'est plus un symbole mais une réalité matérielle, il en va alors de même pour les représentations qu'il supporte. De plus, dans un réseau de neurones, la représentation d'un objet n'est pas supportée par un neurone isolé mais par un groupe de neurones, ou par l'ensemble du réseau, ce n'est plus une petite case en mémoire. Envisagée ainsi, la notion de représentation peut difficilement être rejetée, il est possible de nier la notion subjective, que tout sujet ressent, de représentation, on ne peut nier que l'environnement modifie l'état d'activité des neurones du cerveau.

Le PAC apprentissage dans son aspect quantitatif prend en considération, la plus petite représentation du concept comme il prend en compte, nous venons de le voir la taille des exemples (n). En effet, pour un même concept, il est possible de trouver des définitions plus ou moins longues qui le décrivent de la même manière. Aussi lorsque les théories formelles parlent de la taille d'un concept (taille(c)) elles ne prennent en considération que la plus petite de ses représentations. Dans le 1.1, nous avons d'emblée considéré la plus petite représentation du concept, une représentation à deux éléments («carré et petit »). Nous aurions pu en considérer de plus longues qui n'auraient rien modifié au concept : «carré et non rond et petit » ou encore «carré et non rond et petit et non grand ».

[16]On pourrait qualifier ces différents langages de langages à attributs.

1.2.2.3 C_n une classe de représentations sur X_n

Le modèle PAC considère l'apprentissage d'une **classe de concepts** et non d'un concept seul. Une classe de concepts est un sous-ensemble des concepts que l'on peut former sur un monde à l'aide d'un schéma de représentation. Par exemple dans le 1.1, on aurait pu considérer la classe de concepts que l'on pouvait former avec une conjonction de 2 valeurs d'attributs, la classe aurait alors été {R∧P, R∧G, R∧B, R∧N, R∧I, R∧E, C∧P, C∧G, C∧B, C∧N, C∧I, C∧E, P∧B, P∧N, P∧I, P∧E, G∧B, G∧N, G∧I, G∧E, B∧I, B∧E, N∧I, N∧E}. On aurait tout aussi bien pu considérer la classe des conjonctions formées avec trois valeurs d'attributs {G∧N∧E, G∧N∧I ...} ou la classe des disjonctions de deux valeurs d'attributs {G∨N, G∨N,...} ou des classes plus complexes mêlant les conjonctions et les disjonctions { (G∧N∧E) ∨ (P∧B), (G∧N∧E) ∨ (P∧B), ...}.

Le fait de considérer une classe de concepts plutôt qu'un concept isolé est légitime du point de vue de l'apprentissage naturel. Comme nous le verrons, en apprentissage naturel, le psychologue va rarement s'intéresser à l'apprentissage d'un concept isolé, par exemple « le chien de monsieur Dupont», mais plutôt à une classe de concepts, « le mobilier ».

Du point de vue quantitatif, comme le modèle PAC considère des exemples d'une taille *n* donnée, il considère également les classes de concepts que l'on peut former sur ces exemples ; c'est pourquoi on parle de « C_n une classe de représentations sur X_n ». Le *n* ici n'est en rien indicatif de la taille du concept, il signifie seulement que la classe de concepts est associée à X_n. De la même façon que pour le monde des objets, C est alors égal à l'union de tous C_n pour tous les *n* supérieurs ou égaux à 1 ($\mathbf{C=\cup_{n\geq 1} C_n}$).

1.2.3 Les hypothèses de l'apprenant

1.2.3.1 Formalisation de la classe des hypothèses de l'apprenant

A côté de la classe de concepts cible, le modèle PAC définit **une classe d'hypothèses de l'apprenant, H**. Du point de vue de la *formalisation,* rien ne distingue la classe d'hypothèses de la classe de concepts. La classe d'hypothèses de l'apprenant est simplement la classe de concepts utilisée par celui-ci. Comme un concept, une hypothèse de l'apprenant peut se définir en extension comme un sous-ensemble des objets du monde. Comme un concept, une hypothèse de l'apprenant peut se définir en compréhension et repose alors sur un schéma de représentations. Comme pour les classes de concepts, le *n* de H_n ne fait que décrire la classe de concept se rapportant à X_n. Enfin, de la même façon que pour les classes de concepts, H est égal à l'union de tous H_n pour tous les *n* supérieurs ou égaux à 1 ($\mathbf{H=\cup_{n\geq 1} H_n}$).

Le fait que l'apprenant travaille dans un schéma de représentations donné suppose que cet apprenant est capable :
- premièrement, de formuler toute hypothèse dans ce schéma de représentation *et*

- deuxièmement, étant donné une hypothèse et un exemple, de dire si l'exemple relève de l'hypothèse ou non.

En définissant la classe d'hypothèses de l'apprenant, le modèle commence ainsi à définir cet apprenant. Mais ce n'est pas parce que l'apprenant est capable de formuler une hypothèse et de vérifier pour tout exemple que celui-ci relève ou non de l'hypothèse qu'il est pour autant capable d'apprendre. Si nous faisons un parallèle avec l'apprentissage naturel, ce n'est pas parce qu'un être humain est capable

a) de formuler des descriptions *et*,

b) pour une description donnée et un exemple donné, de vérifier que l'exemple vérifie ou non la description *qu'il est capable*

c) sur la base d'un ensemble d'exemples de donner une description globale approximative de tous ces exemples.

Par contre pour être capable de faire le c) il est nécessaire qu'il soit capable de faire le a) et le b).

Souvent dans les modèles d'apprentissage, cette définition de la classe d'hypothèses est considérée comme une *connaissance a priori* dont bénéficierait l'apprenant. On voit ici qu'il s'agit moins d'une *connaissance* que d'une *compétence* au sens chomskyen du terme, c'est-à-dire une certaine capacité qu'a l'apprenant. En effet, si on considère qu'une connaissance est un ensemble d'hypothèses *déjà formées*, ce n'est pas ce qui est formalisé par la classe d'hypothèses puisque celle-ci consiste dans l'ensemble des hypothèses que l'apprenant *peut former*. Par contre, on peut dire que cette compétence est une «compétence a priori » dans le sens où le modèle suppose a priori que l'apprenant en bénéficie.

1.2.3.2 Une classe de concepts cible pour l'observateur et une classe d'hypothèses pour l'apprenant

Distinguer entre classe de concepts cible et classe d'hypothèses de l'apprenant permet de faire la différence entre observateur et apprenant. La classe de concepts cible est celle qu'utilise l'expérimentateur en psychologie ou l'informaticien en apprentissage automatique, et la classe d'hypothèses, celle qu'utilise l'apprenant. L'observateur peut se définir un concept cible à partir d'un schéma de représentation. L'apprenant peut de son côté formuler ses hypothèses dans un *autre* schéma de représentations. En psychologie cela correspond à l'expérimentateur qui prend comme classe de concepts cible celle des mammifères marins, clairement définie dans un dictionnaire, et qui va chercher à appréhender la classe des hypothèses du sujet, les représentations qu'a le sujet des mammifères marins.

Si nous reprenons le micromonde du 1.1, nous pourrions ainsi formuler un concept dans le schéma de représentations des 2-termes-DNF (disjonction de deux termes qui sont des conjonctions d'un nombre quelconque de valeurs d'attribut), par exemple «(grand *et* noir) *ou* (carré *et* petit) » qui nous permettrait d'étiqueter l'exemple 'rond petit noir flèche extérieure' comme négatif, et l'exemple 'carré grand noir flèche

intérieure' comme positif. De son côté, l'apprenant, pourrait émettre ses hypothèses dans un autre schéma de représentations, celui des 2-CNF (conjonction d'un nombre quelconque de termes qui sont des disjonctions de deux valeurs d'attributs) et l'hypothèse suivante, «(grand *ou* carré) *et* (noir *ou* carré) *et* (noir *ou* petit)», étiquetterait de la même façon les exemples précédents.

Dans le modèle PAC, l'introduction d'une classe d'hypothèses à côté d'une classe de concepts cible n'a pas pour seule raison de coller au plus près à la réalité. Une autre raison est qu'elle peut amener des résultats différents. Il a ainsi été démontré [Pitt, Valiant, 88] pour l'apprentissage de la classe cible des 3-termes-DNF (disjonction de trois termes qui sont des conjonctions d'un nombre quelconque de valeurs d'attribut) que, si on prend comme classe d'hypothèses la même classe des 3-termes-DNF, alors cette classe n'est pas apprenable. Par contre si on prend comme classe d'hypothèses la classe des 3-CNF (conjonction d'un nombre quelconque de termes qui sont des disjonctions de trois valeurs d'attributs) alors elle devient apprenable. Autrement dit le simple fait d'autoriser l'apprenant à formuler ses hypothèses dans une autre classe que la classe de concepts cible a changé le résultat.
L'idée qu'une classe de concepts puisse être apprenable dans un schéma de représentations et non dans un autre n'est pas contre-intuitive. Tout enseignant a un jour rencontré un élève qui était capable de résoudre un problème sans pour autant être capable d'expliquer comment il obtenait cette solution. On peut alors faire l'hypothèse que l'élève a résolu le problème dans un certain schéma de représentations autre que le schéma de représentations qu'est la langue.

Cette séparation entre les deux classes de représentations pose le problème de *l'expressivité* de la classe d'hypothèses de l'apprenant par rapport à celle de la classe de concepts cible. L'expressivité d'une classe est sa capacité plus ou moins grande à représenter tout sous-ensemble des objets du monde[17]. Certains schémas de représentation ont un pouvoir d'expression plus grand que d'autres. Ainsi les arbres de décisions que l'on a vus dans l'introduction ont un pouvoir plus grand que celui des simples conjonctions de valeurs d'attributs. Aucune conjonction de valeurs d'attributs ne permet de représenter le concept correspondant au sous-ensemble {CPNE, RGBI} alors qu'il est aisé de le représenter par un arbre de décisions.

Dans la plupart des travaux en théories formelles, on va supposer que H est aussi expressive que C et qu'ainsi il y a une représentation dans H de chaque concept de C. Il paraît évident que si H ne permet pas d'exprimer la classe de concepts cible, cette dernière ne sera pas apprenable dans le schéma de représentation produisant H. Par exemple, les 3-termes-DNF et les 3-CNF (ou les 2-termes-DNF et les 2-CNF) ont le même pouvoir d'expression, dans le sens où pour tout concept représenté par une représentation au format 3-termes-DNF, on pourra trouver une représentation au format 3-CNF qui exprimera le même concept. Ainsi, la représentation «(grand *et* noir) *ou* (carré *et* petit) » de format 2-termes-DNF recouvre exactement le même concept (les

[17] On peut trouver une formalisation de ce « pouvoir d'expression » dans [Haussler, 87] et [Haussler, 90] avec $\Pi_H(I)$ ensemble de toutes les dichotomies de I (ensemble de représentations d'objets) induites par les hypothèses de H.

mêmes éléments du micromonde) que la formule «(grand *ou* carré) *et* (noir *ou* carré) *et* (noir *ou* petit)» qui est au format 2-CNF soit les éléments {RGNE, RGNI, CGNE, CGNI, CPNE, CPNI, CPBE, CPBI }

Nous verrons plus loin avec les variantes du modèle PAC que cette dissociation entre classe de concepts cible C et classe des hypothèses de l'apprenant H permet de lever beaucoup de conditions sur la classe d'hypothèses. Ainsi dans les modèles prédictifs, on ne posera comme seule condition sur H que H soit aussi expressive que C sans définir expressément le schéma de représentation qu'H utilise. Cette position est alors très proche de celle du psychologue qui ne peut définir H puisque c'est l'objet de sa recherche. Le fait que l'on dissocie le schéma de représentation qui sert à définir les concepts cibles de celui qu'utilise l'apprenant pour formuler ses hypothèses rapproche ainsi le modèle PAC de l'apprentissage naturel. Dans les expériences en psychologie, le chercheur définit, en général, le concept cible en conditions nécessaires et suffisantes et cherche à comprendre le schéma de représentation de l'apprenant qui, lui, ne se définira pas obligatoirement en termes de conditions nécessaires et suffisantes.

Le fait qu'H puisse contenir C, c'est-à-dire, que le schéma de représentations de l'apprenant lui permette de représenter un ensemble plus vaste de concepts que la classe cible, va parfois amener à ce que les informaticiens appellent un biais. Ce biais consiste dans une information sur la classe de concepts cible donné à l'apprenant. Cette information est alors réellement une *connaissance a priori* donnée à l'algorithme. Cette connaissance a priori lui permet de restreindre son espace d'hypothèses (restricted hypothesis space [Haussler, 87])[18]. En apprentissage naturel ce biais peut correspondre au contexte qui permet de délimiter l'apprentissage. Si un élève suit un cours d'histoire, il sait que les concepts à apprendre seront plutôt des concepts historiques que des concepts mathématiques.

Si nous reprenons l'exemple du 1.1, nous aurions pu « informer » RobSimp que la cible était une conjonction de *2* valeurs d'attributs. Un autre algorithme aurait alors été possible : construire toutes les conjonctions de 2 valeurs d'attributs, éliminer toutes celles incompatibles avec l'échantillon.

1.2.4 L'apprentissage

L'apprentissage vu par le modèle PAC est une interaction entre l'apprenant et l'environnement. Nous avons déjà décrit partiellement l'apprenant au travers de son espace d'hypothèses H. Pour le compléter, il convient de préciser qu'il s'agit d'un algorithme. Cet **algorithme L** doit être capable sur la base d'un ensemble d'exemples étiquetés selon un concept *c* de C de retourner une hypothèse *h* de H proche de ce concept *c*. L'apprentissage considéré est donc un apprentissage inductif

[18] Ce problème de la connaissance a priori est à relier à celui de la collusion : à partir de quel point y a-t-il *trop* d'information donnée à l'apprenant concernant le concept. [Jackson, Tomkins, 92] essaient de quantifier cet apport d'information. Ainsi si l'on suppose que l'apprenant travaille dans les k-DNF l'information a priori donnée à l'apprenant sera la valeur précise de k : la classe de concepts cible sera par exemple les 3-DNF et l'apprenant sera informé que k=3.

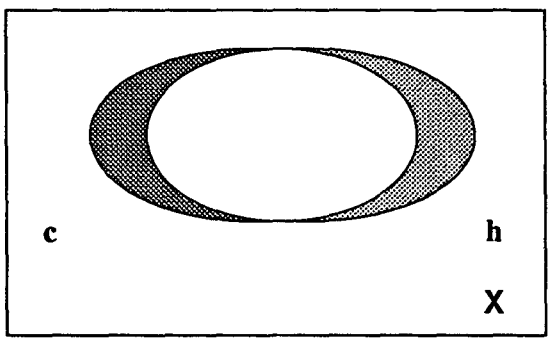


Schéma 1.2.2
X est l'ensemble des objets du monde. *c* est le concept à apprendre. *h* est l'hypothèse apprise par l'apprenant. Dans le schéma les parties grisées correspondent à l'erreur (les éléments de X qui ne sont pas classés de la même manière par *c* et *h*).

Bien que l'algorithme soit en général qualifié d'apprenant (« L » pour «learner »), nous pouvons le considérer comme un processus d'apprentissage particulier qui peut en côtoyer d'autres dans un système apprenant plus vaste. Si nous voulons établir un parallèle avec l'apprentissage automatique ou l'apprentissage naturel, cet apprenant ne modélise pas tout le système cognitif mais, au mieux, une partie de celui-ci.
L'apprenant, l'algorithme L, peut être **déterministe ou non.** Lorsque l'apprenant est déterministe, si on lui présente plusieurs fois le *même* échantillon, il forme toujours la *même* hypothèse, tandis que s'il est non déterministe, il peut fournir une hypothèse *différente*[19]. On notera sans peine que ceci reste valable en apprentissage naturel et que l'on aura même tendance à considérer que l'apprenant est plutôt non déterministe.

Dans le modèle PAC, l'environnement est caractérisé par **EX(c,D)**. EX(c,D) est la procédure chargée de fournir les exemples à l'apprenant. Elle présente à l'apprenant un échantillon d'exemples qu'elle étiquette selon le concept *c* et qu'elle tire selon une distribution de probabilité *D*. Les exemples donnés sont positifs lorsqu'ils relèvent du concept, ils sont alors étiquetés 1, ou négatifs dans le cas de contre-exemples, ils sont alors étiquetés 0. L'échantillon présenté au système cognitif simpliste du 1.1 pourrait être ainsi décrit : ((RGBI,0), (RGBI,0), (RGNI,0), (RGNE,0), (CGNE,0), (RPNE,0), (RPNI,0), (CPBI, 1), (CPBI,1), (CPBE,1), (CPBE,1), (CPBE,1), (CPBE,1)). On

[19] Pour qu'un algorithme soit non déterministe, il suffit qu'à une étape quelconque de son processus il soit amené à choisir entre deux possibilités en tirant à pile ou face.

constate que, de par la distribution de probabilités, un même exemple peut être présenté plusieurs fois tandis que d'autres ne sont pas présentés du tout.

EX(c,D) est généralement appelée Oracle. On peut l'imaginer comme le professeur qui donne les exemples d'un concept ou comme un parent qui nomme à son enfant les objets qu'il rencontre ou d'une manière plus générale comme l'environnement qui étiquette les objets au travers, par exemple, de la douleur ou du bien-être. Nous y reviendrons dans le chapitre suivant. Notons cependant la pertinence de la notion de distribution de probabilités : d'un professeur à l'autre ce ne seront pas les mêmes exemples qui seront présentés ; selon l'endroit où il vit, l'enfant ne se verra pas nommer les mêmes objets par ses parents, ou ne rencontrera pas les mêmes situations de danger ou de bien-être.

L'apprentissage requis n'est pas un apprentissage exact, mais un apprentissage approximatif, l'apprenant peut se tromper. Si on assimile l'apprentissage à une adaptation à l'environnement, le fait de ne réclamer qu'un apprentissage approximatif est pertinent car cette adaptation est rarement optimale.
C'est le paramètre ε qui borne le poids de l'erreur entre l'hypothèse apprise et le concept cible. L'apprenant a accès à cette borne d'erreur afin de pouvoir calculer le nombre d'exemples qui lui est nécessaire pour que l'hypothèse qu'il forme soit assez proche du concept[20]. Le fait que le sujet connaisse le taux d'erreur qu'on lui tolère pourrait correspondre en pédagogie à *l'évaluation formatrice*, dans le cadre de la pédagogie par objectifs. Dans celle-ci, on informe l'apprenant sur les buts à atteindre et la marge d'erreur qui lui est accordée, l'apprenant est alors amené à évaluer lui-même son erreur pour se corriger. Le fait que ces informations lui soient données, l'invite à prendre davantage en charge sa formation et à calculer en quelque sorte le nombre d'exemples qui lui sont nécessaires. Ce courant s'oppose à celui plus traditionnel où c'est le professeur qui évalue l'erreur de l'élève et qui correspond, lui, au modèle fonctionnel (cf. note ci-dessus).
Une façon moins scolaire d'illustrer, en apprentissage naturel, cette connaissance de la borne d'erreur pourrait être la cueillette des champignons. Etes-vous suffisamment sûr de votre connaissance des champignons pour aller en cueillir et les manger. ε correspond alors dans la borne d'erreur de votre hypothèse «champignons comestibles ». Si vous pensez que votre hypothèse a une trop grande marge d'erreur ($>\varepsilon$), il est probable que vous reprendrez cet apprentissage avant d'aller à la cueillette.

Dans l'exemple du 1.1, nous avons joué le rôle de l'Oracle pour RobSimp en définissant l'échantillon d'exemples. Nous n'avons pas étiqueté les exemples en 1 ou 0 qui est le codage binaire traditionnel mais en «plus » ou en «moins ». Nous avons imposé le nombre d'exemples car nous savions qu'il était suffisant pour trouver une hypothèse approximativement correcte, mais normalement c'est à l'apprenant qu'il revient de calculer ce nombre d'exemples nécessaire. Comme dans la définition, nous les avons

[20]Il existe des variantes du modèle PAC où ces paramètres ne sont pas livrés à l'apprenant. Ces variantes sont appelées modèles fonctionnels (voir [Haussler, Kearns, Littlestone et Warmuth, 91])

tirés selon la distribution de probabilités et nous avons défini ε en imposant un apprentissage qui fasse moins de 5% d'erreurs.

L'algorithme L a aussi accès à δ, appelé le paramètre de confiance. Ce paramètre est plus technique, il exprime la distorsion qui peut arriver exceptionnellement entre la distribution de probabilités et le tirage réel. Supposons une pièce non biaisée pour tirer à pile ou face. A priori on peut s'attendre à ce que sur 100 tirages on obtienne un rapport plus près de 50 piles/50 faces que de 90 piles/10 faces. δ est là pour le cas où le tirage réel correspondrait malheureusement à ce dernier rapport. Dans l'exemple du 1.1, il n'apparaît donc pas puisque le tirage réel n'était pas trop « éloigné » de la distribution de probabilités.

Comme nous l'avons vu, les théories formelles considèrent l'apprentissage d'une classe de concepts et non celui d'un concept particulier (**L doit être capable d'apprendre tout concept** $c \in C$). Etudier l'apprentissage d'un seul concept n'a pas grand sens car cela supposerait un algorithme particulier pour chaque concept. Pour des raisons d'économie cognitive, nous pouvons raisonnablement penser que l'homme n'utilise pas des mécanismes différents pour apprendre le concept de «chat » ou le concept de «chien ». C'est une des limites de l'exemple 1.1 où nous n'avons appris qu'un seul concept. Pour vérifier que l'algorithme était correct, il aurait fallu le tester avec tous les concepts.

Une condition plus forte du modèle PAC est que **L doit être capable d'apprendre pour toutes distributions de probabilités D sur X.** Cela revient à demander que l'apprentissage réussisse quel que soit le professeur, quel que soit l'environnement. Nous reviendrons sur cet aspect dans le chapitre 3. Cette condition constitue une autre limite de l'exemple 1.1 puisque nous n'avons réalisé l'apprentissage qu'avec une seule distribution de probabilités.

1.2.5 Mesurer l'apprentissage

Le modèle PAC ne définit pas l'apprentissage de manière seulement qualitative, il le définit aussi quantitativement.

1.2.5.1 L'hypothèse h retournée par L doit satisfaire erreur(h)≤ε, pour tout ε, paramètre d'erreur, $0<\varepsilon<1/2$

Sur le schéma 1.2.2, page 52, nous voyons que l'hypothèse *h* et le concept *c* ne se superposent pas, qu'il y a une zone en grisé. Cette zone en grisé correspond à l'ensemble des éléments de X qui ne sont pas classés de la même manière par *h* et par *c*, ce que l'on peut écrire par $c\Delta h$: la différence symétrique de *c* et de *h*. $c\Delta h$ constitue l'erreur de *h* par rapport à *c* : les éléments qui appartiennent à *c* qui n'appartiennent pas à *h* et les éléments qui appartiennent à *h* qui n'appartiennent pas à *c*. Ces éléments «erronés » ont une certaine probabilité selon D d'être tirés. La somme des probabilités de tous ces éléments est ce que l'on appelle le poids de l'erreur. Il est souhaité que le

poids de cette erreur soit inférieur ou égal à ε. Ainsi *le poids de l'erreur d'une hypothèse n'est pas le nombre d'exemples mal classés mais la somme des probabilités de ces exemples mal classés, on la note* P_D *(cΔh).*

Si on appelle $P_D(x)$ la probabilité selon D de l'élément x, si l'erreur $c\Delta h = \{x_1,x_2,....x_k\}$ c'est-à-dire que $x_1,x_2,....x_n$ sont des éléments de X classés différemment par *c* et *h*, alors on souhaite que :

$$P_D(x_1)+ P_D(x_2)+....+ P_D(x_k) \leq \varepsilon.$$

Ce que l'on peut écrire plus brièvement par

$$\sum_{c(x)\neq h(x)} P_D(x) \leq \varepsilon ,$$

ou

$$P_D(c\Delta h)\leq\varepsilon,$$

Il est donc souhaité qu'à l'issue de l'apprentissage, l'algorithme retourne une hypothèse *h* telle que par rapport au concept *c* la probabilité selon D que l'on tire un élément mal étiqueté par *h* soit inférieure à ε. Il faut noter qu'ε n'est pas l'erreur, mais le paramètre qui borne le poids de l'erreur tolérable. C'est ε qui exprime le fait que l'apprentissage est *approximatif.*

C'est un point clef du modèle de ne pas évaluer le poids de l'erreur de l'hypothèse par le nombre d'éléments qui seraient mal classés par celle-ci, mais par la somme des probabilités de ces éléments. En liant évaluation de l'hypothèse et distribution de probabilité par le biais du poids de l'erreur, Valiant exprime l'idée que l'apprentissage est une adaptation à un environnement particulier. Pour une personne habitant dans la métropole lilloise, il n'est pas très utile que sa catégorie « serpent » lui permette de distinguer entre une couleuvre et une vipère, ceci est par contre beaucoup moins évident pour un viticulteur du sud de la France. Dans le même temps le Lillois a rencontré beaucoup moins de serpents que le viticulteur : la même distribution de probabilités sert à l'apprentissage et à l'évaluation.

Dans l'exemple 1.1, le tableau 1.1.3, page 38, présentait le calcul du poids de l'erreur de deux hypothèses. Ce poids de l'erreur était constitué de la somme des poids (probabilités) de tous les exemples mal classés.

En ne tolérant qu'une erreur strictement inférieure à ½ ($\varepsilon < 1/2$) , on évite d'accepter une hypothèse aléatoire (qui classerait les exemples en tirant à pile ou face). De plus ε est une borne qui sert à l'algorithme pour trouver le nombre d'exemples qui lui sont nécessaires. La plupart du temps, plus on souhaitera que ε soit petit plus le nombre d'exemples devra être grand. Il semble légitime que si l'on souhaite que l'apprentissage soit le plus exact possible, il faille plus d'exemples.

1.2.5.2 L doit retourner cette hypothèse h avec une probabilité d'au moins 1-δ, pour tout δ, paramètre de confiance, 0<δ<1/2

Le rôle de δ est moins évident à comprendre que celui de ε. Alors que ε entre dans le calcul du nombre d'exemples pour être sûr que l'erreur est petite, δ entre dans le calcul du nombre d'exemples pour que l'apprentissage réussisse avec une probabilité suffisante. δ peut ainsi être considéré comme un critère d'arrêt : à partir de combien d'exemples la probabilité que l'erreur soit suffisamment petite ($P_D(c\Delta h)<\varepsilon$) est supérieure à un certain seuil (1-δ).
δ est un paramètre technique qui a sa raison d'être dans le fait que l'on travaille avec des distributions de probabilités et qu'il est toujours possible que l'on ait un tirage « pathologique ». Reprenons le cas de tirages à pile ou face avec une pièce non biaisée. Nous savons que pile ou face ont chacun 1 chance sur 2 d'apparaître, un tirage « pathologique »consisterait en 90 piles 10 faces.

La distorsion, exprimée par δ, entre distribution de probabilités et tirage réel, peut ainsi surgir lors du choix de l'échantillon et donc influencer l'apprentissage. Reprenons l'exemple 1 .1 et supposons un tirage pathologique : nous allons imaginer que parmi les exemples positifs CPBE n'apparaît plus dans l'échantillon malgré sa probabilité de 15% et qu'à l'opposé CPNI apparaisse malgré ses 2%. L'échantillon positif sera alors
E+ ={ CPBI, CPBI, CPBI, CPBI, CPBI, CPBI, CPNI}
Parmi les exemples négatifs, nous supposons que RPNI n'apparaît plus non plus malgré ces 7% et qu'à l'opposé RGNE apparaît 2 fois. L'échantillon négatif serait alors
E- ={RGNE, RGNI, RGBE, RPNE, CGNE, RGNE}.
Une telle possibilité de distorsion entre distribution de probabilités théorique et tirage réel est rare mais peut arriver.

Les trois conjonctions suivantes, C∧P, C∧I et P∧I seront alors consistantes avec l'échantillon et pourront être apprises. Le problème est que les deux hypothèses C∧I et P∧I auront une erreur supérieure à 15%. Ainsi 2 des trois hypothèses retournées par l'algorithme seront fausses alors que l'algorithme est correct. Le problème provient de ce que le tirage réel n'est pas vraiment aléatoire selon la distribution de probabilités. Alors que l'exemple CPBE a une probabilité de 15%, le tirage ne le fait pas apparaître dans l'échantillon. De ce fait, les hypothèses consistantes avec l'échantillon ne le sont pas obligatoirement avec cet exemple. De là, l'erreur supérieure à 15%.

La probabilité qu'un tel événement, qu'une telle distorsion entre distribution de probabilités et tirage réel se produise est exprimé par δ. Une façon d'éviter qu'un tel événement ait une trop grande probabilité (que δ soit trop grand) est d'augmenter le nombre d'exemples. Si vous tirez 4 fois à pile ou face avec une pièce non biaisée, il peut arriver que vous obteniez 4 piles, par contre si vous tirez 100 fois à pile ou face, il est peu probable que vous obteniez 100 piles.

Il faut noter que l'apparition de δ est lié à la problématique des théories formelles qui est l'*apprenabilité*. Comme les chercheurs en théories formelles travaillent avec les

distributions de probabilités, ils ne peuvent ignorer cette possible distorsion entre tirage réel et distribution. C'est ce qui fait du modèle de Valiant un modèle d'apprentissage *Probable*. Du point de vue de l'apprentissage naturel ce paramètre peut être ignoré, ne serait-ce que parce que le plus souvent l'expérimentateur peut vérifier qu'il n'y a pas eu tirage « pathologique »

1.2.5.3 Si L tourne en temps polynomial en n, taille(c), 1/ε et 1/δ on dit que C est efficacement PAC apprenable.

« Si L tourne en temps polynomial en n, taille(c), 1/ε et 1/δ on dit que C est efficacement PAC apprenable. » . Cette notion de *polynomialité* a envahi tout le champ de l'informatique. Avec la naissance de celle-ci, les chercheurs se sont très vite préoccupés de savoir si tous les problèmes ont une solution, s'ils sont *décidables*. Un problème est décidable, si l'on est capable d'exhiber un algorithme qui le résout. Ils se sont cependant très vite aperçus que ceci était insuffisant. En effet, il est très facile *en théorie* de proposer un logiciel de jeu d'échecs qui battrait n'importe quel joueur. Il «suffit» de lui faire parcourir l'ensemble de tous les mouvements possibles et de lui faire choisir l'enchaînement de ces mouvements qui l'amène à la victoire. Le problème est que le calcul de tous ces mouvements amène à une explosion combinatoire. Le logiciel n'aura ni le temps, ni l'espace mémoire suffisant pour un tel calcul. Ainsi, à la question de la décidabilité d'un problème, (Y a-t-il un algorithme capable de le résoudre ?), les informaticiens enchaînent la question de la *complexité* de ce problème (Quel est le temps et l'espace mémoire que réclame cet algorithme ?). Habituellement, il est considéré qu'un temps polynomial selon les paramètres d'entrée est raisonnable et qu'un temps exponentiel ne l'est pas. Cela se comprend aisément, il suffit pour cela de supposer que l'on ait deux algorithmes d'apprentissage qui prennent en entrée des exemples de taille n et que l'un tourne en un temps égal à n^2 et l'autre en 2^n et de calculer ces temps lorsque n vaut 10 et lorsque n vaut 10000.

Ainsi, la définition réclame ici que l'apprentissage soit réaliste et qu'il se termine en un temps raisonnable. Elle fait dépendre ce temps de $1/\varepsilon$ et de $1/\delta$ car plus on souhaite que la probabilité de réussite de l'apprentissage soit grande et que l'hypothèse retournée soit proche du concept cible, plus il faut de temps. Ce temps doit aussi dépendre de la taille des exemples et de celle de la plus petite représentation du concept. De la taille des exemples car l'algorithme doit «prendre le temps» de les lire, de la taille de la plus petite représentation du concept car il faudra bien qu'il «prenne le temps» de la retourner.

Avec cette borne sur le temps d'apprentissage apparaît le côté réaliste du modèle d'apprentissage PAC. C'est avec cette condition que le modèle passe d'un aspect purement qualitatif et descriptif à un aspect quantitatif. Le passage du modèle de Gold d'apprentissage à la limite, à celui de Valiant d'apprentissage Probablement Approximativement Correct peut être caractérisé par cette condition car cette borne n'apparaît pas dans le modèle de Gold. Dans ce modèle, il n'y a pas de critère d'arrêt. Cette condition supplémentaire apportée au modèle oblige à tenir compte des distributions de probabilités car il n'est plus possible de considérer que l'apprenant

verra l'échantillon complet du concept. A son tour, ceci implique que le modèle passe d'un apprentissage exact à un apprentissage approximatif puisque cet apprentissage, du fait des distributions de probabilités, sera très dépendant des exemples vus.

Dans l'exemple du 1.1, il n'est pas fait mention du temps d'apprentissage, il était implicite. Dans les expérimentations, comme dans la vie, nous supposons toujours que le temps d'apprentissage est borné ne serait-ce que par la patience du sujet ou la durée de vie.
Ce temps d'apprentissage dépend de la taille des exemples et de celle du concept. Dans l'exemple 1.1, le nombre d'attributs utilisé est de 4, nous aurions pu en définir 5, il aurait suffi d'ajouter, par exemple, une flèche à gauche. A partir de là, l'algorithme que nous avons proposé aurait pris plus de temps puisque l'hypothèse la plus spécifique (conjonction de tous les attributs) aurait été de longueur 10 au lieu de 8. De la même façon, le concept cible est de taille 2 car il est défini sur la valeur de 2 attributs, s'il passait à trois, il y aurait alors davantage d'hypothèses à envisager et l'apprentissage en serait d'autant rallongé.

1.2.6 L'apprenabilité ou comment «fonctionne » cette modélisation

L'objet des théories formelles en général et du modèle PAC en particulier n'est cependant pas d'étudier l'apprentissage mais *l'apprenabilité* « **On dit que *C* est PAC *apprenable* s'il existe un algorithme L ».** Cette phrase est à la base de ce qui distingue la problématique des théories formelles de l'apprentissage de celle de la psychologie. Les théories formelles étudient l'*apprenabilité* : sous quelles conditions une classe de concepts est-elle *apprenable*. Elles considèrent qu'une classe de concepts est apprenable *s'il existe* un processus, un algorithme qui, partant de la description des exemples, propose une représentation d'une hypothèse qui soit consistante avec ces exemples. Sa question première est donc « existe-t-il un algorithme ? », alors qu'en psychologie *on sait qu'il existe* un algorithme chez le sujet, l'objectif est de le cerner, de le comprendre.

Pour démontrer qu'un algorithme existe, il suffit de le décrire, et montrer ensuite que cet algorithme respecte les paramètres d'erreur et de confiance et tourne en temps polynomial, en utilisant un nombre polynomial d'exemples. Ce nombre d'exemples est calculé en fonction des paramètres d'erreur, ε, de confiance, δ, de la taille des exemples, n, et de la taille de la cible, taille(c). On trouvera en annexe 2 un tel type de démonstration pour la classe des conjonctions de valeurs d'attributs.
A l'inverse, il peut aussi être démontré que, pour certaines classes de concepts, de tels algorithmes respectant ces conditions d'approximation, de confiance et de polynomialité *ne peuvent pas exister*. Dans ce cas, on a démontré que ces classes de concepts ne sont pas apprenables.

De telles démonstrations permettent de mettre des limites *théoriques* en apprentissage automatique. En effet lorsqu'on cherche à construire un système apprenant pour une classe donnée de concepts, il peut être intéressant de connaître les bornes théoriques

relativement à cet apprentissage. Mais les conditions posées par le modèle PAC sur l'apprentissage sont très fortes, notamment celle réclamant un apprentissage *pour toutes distributions* de probabilités. Il n'est pas sûr qu'elle soit toujours nécessaire. De ce fait, certaines heuristiques en apprentissage automatique fonctionnent avec des résultats corrects et le modèle PAC est incapable d'en rendre compte.

La démonstration dans le modèle PAC qu'une classe de concept est apprenable va donc toujours être amenée à exhiber un algorithme. *Elle ne prétend pas pour autant que tout système cognitif qui apprend cette classe de concepts utilise cet algorithme.* L'objet des théories formelles n'est pas d'affirmer que le processus, l'algorithme présenté, est celui qu'utilise l'être humain. Mais, à l'opposé, si l'homme est capable d'utiliser ce processus alors il est capable d'apprendre cette classe de concepts. Il pourra donc arriver qu'incidemment le processus décrit corresponde bien à celui qu'utilise l'homme ou qu'un des aspects de ce processus soit utilisé par l'homme.

1.3 Limites de cette présentation, variantes du modèle PAC et quelques résultats

1.3.1 Les limites de cette présentation du modèle PAC

Les limites de cette présentation portent essentiellement sur le fait que, pour exposer le modèle PAC, nous avons utilisé comme classe de concepts cibles des fonctions booléennes. L'avantage d'une telle présentation est qu'elle est plus accessible. Le danger est qu'elle puisse donner à penser que c'est le seul type de classes de concepts dont l'apprentissage puisse être modélisé par le modèle PAC. En fait, le modèle PAC peut être utilisé pour l'apprentissage de classes de concepts les plus diverses. Cela peut amener quelques modifications, notamment l'introduction de paramètres supplémentaires dans le calcul de la polynomialité.

Par exemple, le modèle PAC permet de modéliser l'apprentissage de langage défini par un automate. Supposons un langage défini sur un alphabet Σ possédant 2 lettres, $\Sigma=\{a,b\}$. L'automate suivant (voir schéma 1.3.1 page suivante) décrit ce langage. Tous les mots qui sont acceptés par cet automate sont des exemples positifs du concept, ou des mots acceptés par le langage, et ceux qui ne le sont pas sont des exemples négatifs.
Un automate est défini par des états et des transitions entre ces états. On entre dans l'automate par l'état initial et on en ressort par l'état final. Pour passer d'un état au suivant, il faut lire une lettre. Il est possible de boucler un nombre indéfini de fois. Un mot du langage est un mot accepté par l'automate.
Ainsi, avec le langage défini par le schéma 1.3.1, 'aba', 'abaaba', 'abaabaabaaba' sont des mots acceptés par l'automate. Tandis que 'abaab' ne l'est pas car il s'arrête à l'état 2

qui n'est pas un état de sortie, de même que 'aab' restera bloqué à l'état 1, car sur cet état il n'y a pas de transition avec a.
L'apprentissage va consister sur la base d'un échantillon de mots à retrouver l'automate. On présente à l'apprenant un échantillon d'exemples étiquetés tel que {(aba, 1), (abaaba, 1), (abaabaabaaba, 1), (abaab, 0), (aab, 0)}et l'apprenant doit construire l'automate correspondant. La différence majeure qu'il y a avec les fonctions booléennes est que les exemples peuvent varier en taille. C'est pourquoi la taille de l'exemple entre dans le calcul de la polynomialité.

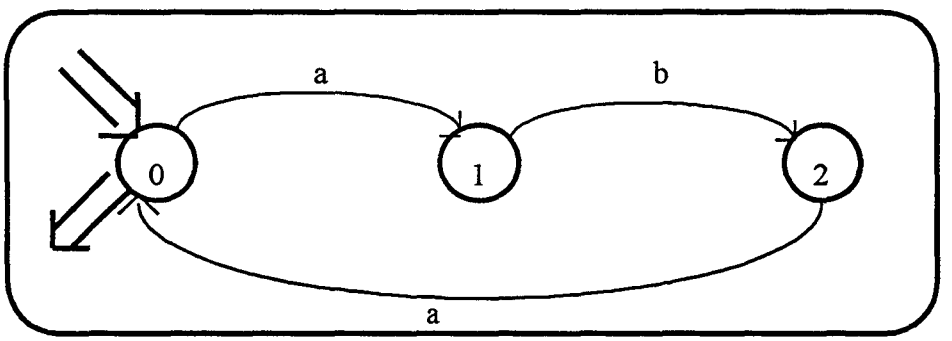


Schéma 1.3.1

Un automate décrit un langage. Cet automate possède 3 états {0, 1, 2}. L'état 0 est en même temps l'état initial et l'état final. Pour passer d'un état au suivant, il faut lire une lettre. Pour passer de l'état 0 à l'état 1, il faut lire un a. Pour passer de l'état 1 à l'état 2, il faut lire un b, pour passer de l'état 2 à l'état 0, il faut lire à nouveau un a. Il est possible de boucler un nombre indéfini de fois. Un mot du langage est un mot accepté par l'automate. Ainsi aba, abaaba, abaabaabaaba sont des mots acceptés par l'automate. Tandis que abaab ne l'est pas car il s'arrête à l'état 2 qui n'est pas un état de sortie, de même que aab restera bloqué à l'état 1 car sur cet état il n'y a pas de transition avec a.

Dans la présentation que nous avons faite, nous n'avons envisagé que l'apprentissage de concepts binaires, de deux concepts en quelque sorte (le concept et sa négation), alors que rien dans le modèle n'y oblige. On pourrait ainsi envisager l'apprentissage simultané de plusieurs concepts.

Nous n'avons parlé que d'apprentissage non bruité, alors que le modèle PAC permet aussi de modéliser l'apprentissage en présence de bruits. Le bruit peut porter sur l'étiquetage, c'est-à-dire que l'étiquetage de quelques exemples peut être erroné. Il peut porter aussi sur la description des exemples.

Enfin nous avons toujours présenté des algorithmes déterministes, alors que nous avons vu que la définition envisage aussi que les algorithmes puissent être non déterministes.

1.3.2 Les variantes du modèle PAC

Il existe une multitude de variantes du modèle PAC. Quelques-unes de ces variantes portent sur les conditions de l'apprentissage et notamment celle imposant que l'apprentissage ait lieu pour toute distribution de probabilités. Nous y reviendrons dans le chapitre 3.

Les variantes dont nous voudrions parler maintenant concernent les rapports réciproques entre la classe de concepts cible C et la classe d'hypothèses de l'apprenant H.
Le modèle PAC «standard » correspond à la version que nous avons donnée, c'est-à-dire l'apprentissage de C dans H où H est au moins aussi expressif que C, ce qui correspond à $C \subseteq H$. $C = H$, où l'on utilise le même schéma de représentation pour décrire C et décrire H, n'est plus alors qu'un cas particulier du précédent.
Lorsque $C \subset H$, C et H ne sont plus décrits par le même schéma de description. S'il est toujours nécessaire de connaître précisément celui de C pour étiqueter les exemples, il est possible d'envisager le cas où l'on ne sait rien de H. Cela pose un problème car pour évaluer l'hypothèse de l'apprenant, il faut bien que celui-ci puisse la communiquer. Comme il n'y a pas de schéma de représentation défini, il faut qu'il le fasse par un autre moyen. La solution consiste alors à demander à l'apprenant d'étiqueter de nouveaux exemples, de prédire l'appartenance de nouveaux exemples. Dans ce cas, on parlera d'apprentissage *prédictif*.

Avec $C \subseteq H$, nous avons toujours la garantie qu'il existe bien dans H une hypothèse correcte (i.e. qui étiquette les exemples de la même manière que le concept). C'est en ce sens que H est plus expressive que C. Dans ce type d'apprentissage, l'apprenant «connaît » cette inclusion. Il existe d'autres modèles où l'apprenant n'a aucune information sur les rapports entre C et H. Dans ces modèles, l'inclusion peut notamment être inversée et l'on a alors $H \subseteq C$. Ces modèles sont baptisés modèles *agnostiques* car l'apprenant est agnostique, dans le sens où il ne fait aucune supposition sur les rapports de C et H [Mitchell, 97].

Notons que si $H \subseteq C$, il est possible qu'il n'y ait pas dans H d'hypothèse identique au concept c de C, mais seulement une *meilleure* hypothèse. Parmi toutes les hypothèses h de H, ce sera celle qui est la plus proche du concept cible c de C. Cette hypothèse est appelée h_{opt}, pour hypothèse optimale, c'est-à-dire celle dont l'erreur avec le concept cible est minimale. Ainsi, face à un échantillon, il est possible que l'apprenant, ne puisse pas fournir une hypothèse consistante avec cet échantillon mais simplement une hypothèse qui fasse le moins d'erreurs possibles. L'approximation portera alors sur la différence entre l'hypothèse de l'apprenant et cette hypothèse optimale.

1.3.3 Quelques résultats

Les recherches en théories formelles ont produit des résultats susceptibles d'intéresser des chercheurs d'autres disciplines.

1.3.3.1 L'apprentissage dans un espace d'hypothèses autre que l'espace de concepts cible.

Nous avons vu que la classe de concepts des 3-termes-DNF n'est pas apprenable si nous considérons comme espace d'hypothèses la même classe des 3-termes-DNF. Par contre, cette classe devient apprenable si on prend comme espace d'hypothèses la classe des 3-CNF. Ces deux classes de concepts, 3-termes-DNF et 3-CNF, recouvrent pourtant le même ensemble de concepts, c'est-à-dire les mêmes sous-ensembles du monde. Elles ont la même expressivité.

Cette distinction entre classe de concepts cible et classe d'hypothèses de l'apprenant, permet d'établir un parallèle que l'on développera dans le chapitre 2 avec la psychologie. Dans de nombreuses expériences le psychologue est amené à définir une classe de concepts cible, par exemple le mobilier, les mammifères marins etc. Il cherche ensuite à cerner les hypothèses de l'apprenant relativement à ces concepts. Pour éviter que l'apprenant ne soit forcé de restituer son hypothèse dans le même schéma de représentation, la langue, l'expérimentateur va mettre en place toute une série de protocoles : étude des temps de réponse, etc.

1.3.3.2 Le weak learning

Dans l'apprentissage *faible* (*weak* learning), on va supposer que l'algorithme est tout juste capable de ne se tromper qu'un tout petit peu moins qu'une fois sur deux (ε à peine inférieur à ½) et ceci pas très souvent (δ proche de 1). L'apprentissage est très approximatif et sa probabilité de réussite est très faible. [Shapire, 90] a démontré que s'il existe un tel algorithme, alors cet algorithme est capable d'apprendre dans le modèle PAC classique (qui est alors appelé apprentissage PAC *fort*).

Ce résultat est important car il invite à penser que tout organisme qui est capable de faire un peu mieux que le hasard est capable d'apprendre.

1.3.3.3 L'apprentissage avec requêtes

Le modèle PAC suppose comme interactions entre l'apprenant et l'Oracle, que l'Oracle soumette à l'apprenant des exemples qu'il a tirés selon une distribution de probabilités. [Angluin, 87], [Angluin, 88] a étudié d'autres types d'interactions possibles. Par exemple, au lieu d'imposer à l'apprenant les exemples étiquetés, on lui laisse la possibilité de les choisir et de les présenter à l'oracle qui lui indique en retour si l'exemple relève ou non du concept. Un autre type d'interaction possible consiste en ce que l'apprenant soumette son hypothèse à l'oracle et celui-ci dans le cas où l'hypothèse est erronée lui retourne un contre-exemple. Angluin met ainsi en évidence que certaines classes de concepts qui ne sont pas apprenables lorsque l'on impose à l'apprenant les exemples le deviennent lorsqu'on laisse à celui-ci la possibilité de les choisir. La démarche d'apprentissage ne se situe pas exactement dans le modèle de Valiant parce

que, par le jeu du questionnement de l'apprenant, les exemples proposés ne sont plus tirés selon la distribution de probabilités. Néanmoins, elle n'est pas sans écho en sciences humaines, ne serait-ce que parce qu'elle permet de décrire le scientifique qui teste ses hypothèses. On verra par ailleurs dans le chapitre 2 que ces différences dans le protocole d'apprentissage ont aussi été étudiées en psychologie.

1.3.3.4 La VCD : Vapnik-Chervonenkis Dimension

La dimension de Vapnik-Chervonenkis a été proposée par [Vapnik et Chervonenkis,71] et introduite dans la modélisation PAC par [Blumer, Ehrenfurt, Haussler et Warmuth, 87]. Cette dimension mesure, en quelque sorte, l'expressivité d'une classe de concepts. Elle est égale à la taille du plus grand sous-ensemble d'objets de X tel que la classe de concepts permet de décrire chacune des parties de ce sous-ensemble.

Techniquement, la VCD est souvent utilisée en apprentissage PAC car elle permet de calculer le nombre d'exemples dont a besoin un algorithme d'apprentissage consistant[21] pour apprendre un concept inconnu. Autrement dit elle permet de vérifier qu'il y a suffisamment d'informations dans l'échantillon pour apprendre le concept.

1.3.3.5 Le rasoir d'Occam

Nous présenterons plus longuement le rasoir d'Occam dans le chapitre 3 lorsque nous parlerons d'économie cognitive. Le rasoir d'Occam stipule qu'entre plusieurs hypothèses, il convient de choisir la plus simple. Un théorème garantit qu'en adoptant cette démarche on apprend. Cela revient à dire que toute compression *très grande* de l'information est un apprentissage. Ce théorème est fréquemment utilisé dans les théories formelles car il permet de faciliter les démonstrations.

1.4 Les notions clefs du modèle PAC

L'apprentissage PAC tente de modéliser l'apprentissage inductif afin de pouvoir procéder à une analyse quantitative de celui-ci [Haussler, 87]. Ce que nous souhaitons dans ce travail c'est repérer les aspects de l'apprentissage naturel qui sont effectivement appréhendés par ce modèle PAC afin d'étudier les apports mutuels possibles entre le modèle PAC et le domaine de la psychologie concerné par l'apprentissage. Pour cela, nous résumons ici les concepts/notions clefs du modèle PAC. C'est à partir de ces notions clefs que nous présenterons dans le chapitre suivant comment la psychologie envisage l'apprentissage inductif. Préalablement, nous proposons un schéma qui synthétise ces différentes notions.

---

[21] Un algorithme est consistant s'il retourne une hypothèse qui étiquette les exemples de *l'échantillon* sans erreur.

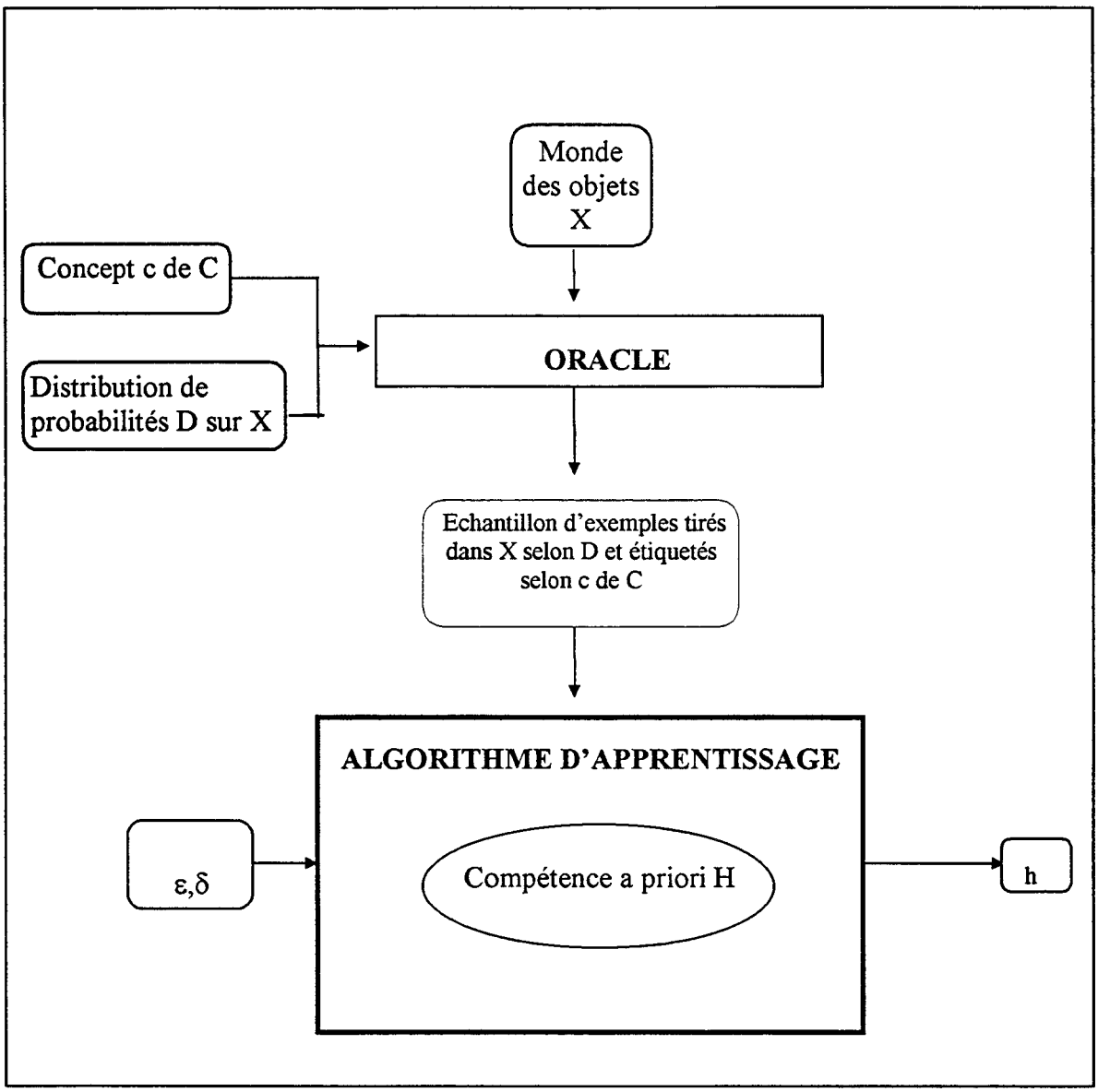


Schéma 1.4.1 : L'apprentissage décrit par le modèle PAC.
L'environnement (l'oracle) présente à l'apprenant (l'algorithme d'apprentissage) un échantillon d'exemples tirés dans le monde X selon une distribution de probabilités D et étiquetés selon un concept c de C. Sur la base de cet échantillon l'apprenant doit retourner avec une probabilité de 1-δ une hypothèse h de H qui soit proche du concept c, c'est-à-dire dont le poids de l'erreur est inférieure à ε. H est la classe des hypothèses que l'apprenant peut formuler.

1.4.1 Le PAC apprentissage tente de modéliser l'apprentissage inductif

Le modèle PAC considère l'apprentissage inductif supervisé

Dans l'apprentissage inductif, l'apprenant doit formuler une représentation d'un concept sur la base d'un échantillon d'exemples étiquetés selon ce concept. Un concept est envisagé comme le sous-ensemble de tous les objets du monde qui vérifient une règle.

Les exemples sont présentés à l'apprenant par un Oracle qui les étiquette selon le concept. L'Oracle peut être considéré comme le professeur, ou plus généralement comme l'environnement. Dans ce dernier cas, comme l'apprentissage est supervisé, c'est l'environnement qui impose au sujet les catégories qu'il doit former.

Les exemples présentés peuvent être des exemples positifs (vérifiant le concept) ou négatifs (contre-exemples) ou uniquement des exemples positifs.

Le modèle PAC considère qu'une classe de concepts est apprenable dès qu'il existe un apprenant pour cette classe de concepts

Dans le modèle PAC, une classe de concepts est dite apprenable dès qu'il existe, c'est-à-dire dès que l'on peut décrire, un processus efficace qui retourne une représentation de ce concept sur la base d'un échantillon d'exemples étiquetés de ce concept. C'est ce qui distingue les théories formelles de l'apprentissage de la psychologie. Dans cette dernière, l'apprenant est «donné », c'est l'être humain et celui-ci est l'objet de l'étude.

Le modèle PAC considère l'apprentissage comme un travail au niveau des représentations : l'apprenant part des représentations des exemples pour en arriver à une représentation du concept.

L'apprenant n'apprend pas à partir des objets du monde réel, mais à partir d'une représentation de ceux-ci. Dans la présentation que nous avons faite ici, les objets sont représentés par *des valeurs d'attributs*. Les concepts et les hypothèses sont représentés *par des attributs avec leurs valeurs possibles.* Ces attributs sont reliés entre eux selon *un schéma de représentation* pour donner la représentation du concept ou de l'hypothèse.
Pour fixer les idées, il est possible de considérer la langue comme un schéma de représentation, les attributs et leurs valeurs étant les mots. Les règles du schéma étant données par la syntaxe. Une même classe de concepts peut être apprenable dans un schéma et pas dans un autre.

Le modèle PAC standard considère l'apprentissage de C, la classe de concepts, dans H, la classe d'hypothèses. C'est-à-dire qu'il existe deux schémas de représentations.
Le premier permet de décrire les concepts. Ainsi, dans le modèle PAC il est implicitement supposé que pour tout concept, il existe une représentation *exacte* de ce concept et *connue* de l'environnement, l'Oracle, qui permettra à celui-ci d'étiqueter les instances en exemples positifs ou négatifs.
Le second schéma de représentation est celui utilisé par l'apprenant pour représenter ses hypothèses. Sur la base de ce schéma et de l'ensemble d'attributs avec leurs valeurs, l'apprenant peut formuler tout un ensemble d'hypothèses appelé *la classe d'hypothèses de l'apprenant*. On suppose que cette classe d'hypothèses est suffisamment expressive pour permettre de représenter chaque concept de la classe de concepts ($C \subseteq H$).

Si nous continuons le parallèle avec la langue, la représentation d'un concept correspond à la définition que nous pouvons trouver de ce concept dans un dictionnaire. Celle-ci permet, si elle est bien faite, d'étiqueter les exemples. Les hypothèses de l'apprenant sont les représentations que se fait l'apprenant des concepts, on peut supposer que la langue n'est pas son seul schéma de représentations. C'est un des objets de la psychologie d'essayer de cerner cette question.

Le modèle PAC suppose de l'apprenant qu'il bénéficie d'une certaine compétence et a accès à une certaine connaissance a priori

La *compétence* (au sens chomskyen du terme) que réclame le modèle PAC est la *capacité de l'apprenant à pouvoir formuler toutes les hypothèses possibles dans son schéma de représentations et pour chacune de ces hypothèses la capacité à vérifier si un exemple en est une instance ou non.* Si l'apprenant ne bénéficie pas de ces compétences, il ne sera pas capable d'apprendre, puisqu'il ne sera pas capable de contrôler la validité de ses hypothèses. Mais ces compétences, à elles-seules, ne suffisent pas pour garantir un apprentissage, car elles ne permettent pas, à elles-seules, de passer d'une description des exemples à une description du concept.

La *connaissance a priori* dont pourra bénéficier l'apprenant consiste dans *une certaine information sur la classe de concepts* à apprendre. Cette connaissance, lui permet de réduire son espace d'hypothèses.

Le modèle PAC suppose que l'hypothèse apprise peut correspondre approximativement au concept cible et non exactement

Si le modèle PAC suppose que la représentation du concept est exacte[22], par contre, il autorise l'apprenant à fournir une hypothèse qui ne corresponde pas exactement au concept. C'est-à-dire que certains objets pourront être classés différemment par le concept et l'hypothèse. Cette différence entre l'hypothèse apprise et le concept cible

[22]Cette affirmation peut paraître triviale pour des informaticiens. Nous verrons qu'en psychologie, cette exactitude du concept cible n'est pas toujours posée.

constitue l'erreur de l'hypothèse. Le modèle PAC propose de borner le poids de cette erreur, c'est-à-dire de préciser dans quelles limites le poids de cette erreur est tolérable. Le poids de cette erreur ne sera pas calculé en fonction du nombre d'exemples mal classés mais en fonction de leur probabilité. Il est ainsi peu important que l'apprenant classe mal des objets qu'il rencontre très rarement. Par contre, il est capital qu'il classe correctement un objet qu'il voit fréquemment.

Le modèle PAC considère que les exemples sont présentés selon une certaine distribution de probabilités

Dans le modèle PAC, on considère que les objets n'ont pas tous la même probabilité d'être présentés en tant qu'exemple à l'apprenant, que cette probabilité dépend d'une distribution donnée. L'Oracle s'appuie sur cette distribution pour tirer les exemples.
Pour fixer les idées, on peut considérer la distribution de probabilités sur les objets comme une caractérisation de l'environnement de l'apprenant. Les objets qu'il est amené à rencontrer dépendent du lieu où il vit. Une autre manière possible d'envisager la distribution de probabilités est de considérer que les exemples sont présentés par un professeur et que d'un professeur à l'autre, les exemples choisis ne sont pas les mêmes.

Le modèle PAC suppose que l'apprentissage doit avoir lieu quelle que soit la distribution de probabilités

Le modèle demande néanmoins que l'apprentissage ait lieu quelle que soit la distribution de probabilités sur les objets. Cela revient à demander qu'il y ait apprentissage quel que soit l'environnement de l'apprenant ou quel que soit le professeur. Cette contrainte très forte est néanmoins 'adoucie' par le fait que l'évaluation de l'hypothèse tient compte de cette distribution de probabilités (voir ci-dessus concernant l'erreur).
C'est le professeur qui a participé à l'apprentissage qui évalue. Le sujet ne change pas d'environnement lors de l'évaluation, ainsi un objet rare lors de l'apprentissage aura peu de chances d'être présenté lors de l'évaluation.

Le modèle PAC suppose que l'apprentissage peut échouer

Il est envisagé dans le modèle PAC que le tirage réel qui a servi à la présentation des exemples soit assez éloigné de la distribution de probabilité théorique et qu'ainsi l'apprentissage échoue. Le modèle PAC impose alors une borne à l'éventualité d'un tel événement.
Replacé dans le cadre de l'apprentissage naturel, un tel événement pourrait correspondre à une série de rencontres *inhabituelles* de l'apprenant avec les objets, comme par exemple de ne rencontrer que des chats sans queue, ce qui l'amènerait à penser qu'un chat n'a pas de queue.

Le PAC apprentissage suppose que l'apprentissage ait lieu en un temps 'raisonnable'

Le modèle PAC suppose que l'apprentissage se déroule en un temps fini et que ce temps ne soit pas trop long. Cette borne sur le temps constitue un critère d'arrêt. Elle définit le moment où l'on considère que l'apprentissage a probablement réussi.

Le PAC apprentissage tient compte de toutes les conditions ci-dessus lors de l'évaluation de l'apprentissage

Toutes les conditions ci-dessus interviennent lors de l'évaluation de l'apprentissage. Nous avons vu que l'erreur n'est pas définie en fonction du nombre d'exemples mal classés, mais en fonction de leur probabilité. Ainsi, la première des conditions est la distribution de probabilités. C'est sur la base de celle-ci que le PAC apprentissage bornera l'erreur tolérable et la probabilité de réussite de l'apprentissage. Mais le PAC apprentissage tient compte aussi de la taille des exemples présentés ainsi que celle du concept à apprendre. De ces quatre paramètres dépendront le nombre d'exemples nécessaires à l'apprentissage et le temps d'apprentissage.

Le PAC apprentissage considère un apprenant vierge de tout passé

Il faut noter que l'apprenant, tel que l'envisagent la plupart du temps les théories formelles, est privé de passé. Il aborde chaque concept comme étant le premier concept à apprendre. Ceci n'est pas inhérent au modèle. Nous pouvons très bien envisager, bien que cela n'ait pas encore été fait, que le passé de l'apprenant lui soit « donné» par le biais de la connaissance a priori, ou que l'algorithme « se construise un passé » au cours de l'apprentissage.

1.4.2 Le modèle PAC procède à une analyse quantitative de l'apprentissage.

Une définition plus formelle de tout ce qui a été décrit précédemment permet de mettre à jour les différents aspects quantitatifs du modèle PAC.

X est l'ensemble des représentations des objets. C est une classe de concepts sur cet ensemble X. H est une classe de représentations aussi expressive que C, c'est-à-dire qu'il y a une représentation h dans H de chaque concept c de C. H est la classe d'hypothèses de l'algorithme d'apprentissage PAC.

Les paramètres introduits pour quantifier l'apprentissage sont les suivants :

n : le nombre d'attributs. X_n est l'ensemble des représentations des objets utilisant au plus n attributs, C_n est la classe de concepts sur X_n et H_n est la classe de représentations utilisée par l'apprenant[23].

taille(c) : la taille de la plus petite représentation du concept c de C_n dans H_n.

D : la distribution de probabilités des exemples

m : la taille de l'échantillon

Erreur (h) : P_D (cΔh) qui correspond à la somme des probabilités selon D des éléments qui ne sont pas étiquetés de la même manière par c et par h,

ε : le paramètre d'erreur qui sert à borner le poids de l'erreur admissible,

δ : le paramètre de confiance qui sert à indiquer avec quelle probabilité on veut que l'apprentissage réussisse

Définition de l'apprentissage PAC
Soit C_n *une classe de concepts sur* X_n *(où* X_n *est soit* $\{0,1\}^n$ *soit l'espace euclidien de dimension n* $\Re^n$*), et* H_n *une classe de représentations, soit* $X=\cup_{n\geq 1} X_n$., $C=\cup_{n\geq 1} C_n$ *et* $H=\cup_{n\geq 1} H_n$. *On dit que* C_n *est PAC apprenable s'il existe un algorithme* L *(déterministe ou non) avec la propriété suivante : pour tout concept* $c\in C_n$, *pour toute distribution de probabilité* D *sur* X, *et pour tout* $0<\varepsilon<1/2$ *et* $0<\delta<1/2$ *si L a accès à EX(*c,D*) et aux entrées* ε *et* δ,*alors avec une probabilité d'au moins* $1-\delta$, L *retourne une hypothèse* $h \in H_n$ *satisfaisant erreur(*h*)*$\leq\varepsilon$.
Si L *tourne en temps polynomial en n, taille(c),* $1/\varepsilon$ *et* $1/\delta$ *on dit que C est efficacement PAC apprenable.*

Comme nous l'avons vu, l'erreur est quantifiée en termes de *probabilités* des éléments mal classés par l'hypothèse apprise et non en fonction du *nombre* d'éléments mal classés. Cela implique que même s'il y a de nombreux éléments mal classés par l'hypothèse, à partir du moment où ces éléments ont une faible probabilité d'apparaître, on considère que l'erreur de l'hypothèse est faible. C'est par ce biais que, lors de l'évaluation de l'hypothèse apprise, on prend en compte l'échantillon qui a servi à l'apprentissage. *Autrement dit le PAC apprentissage postule que toute évaluation d'une hypothèse doit tenir compte des conditions dans lesquelles cette hypothèse a été formée.*

Le fait de réclamer que l'algorithme tourne en temps polynomial exprime le côté réaliste de l'apprentissage : un temps polynomial est assimilable à un temps raisonnable. Le fait de demander que l'algorithme tourne en un temps polynomial, implique

[23]Rappelons que ce premier paramètre *n* est lié à la classe de concept cible, celle des fonctions booléennes. Qu'avec d'autres classes de concepts, il pourrait être remplacé par la taille des exemples (voir 1.7)

naturellement que le nombre d'exemples présentés soit lui aussi polynomial. Une grande partie des démonstrations d'apprenabilité PAC va d'ailleurs consister à calculer ce nombre d'exemples qui doit être un polynôme en n, taille(c), $1/\varepsilon$ et $1/\delta$.

1 Conclusion

Ce chapitre a exposé ce qu'est le modèle d'apprentissage Probablement Approximativement Correct proposé par Valiant en 1984. Nous avons d'abord présenté un exemple d'apprentissage, qui ne constitue pas une démonstration d'apprentissage PAC, mais permet de se faire une intuition des principales notions qui sont inhérentes au modèle. Nous avons, ensuite décrit le modèle et analysé chacune des notions qu'il intègre. Puis, nous avons présenté les limites de cette présentation et quelques variantes et résultats du modèle PAC. Nous avons terminé en faisant un récapitulatif des points principaux.

Le modèle d'apprentissage PAC a pour objet la recherche de *l'apprenabilité* de certaines classes de concepts. Il considère qu'une classe est apprenable lorsqu'il existe un apprenant capable de l'apprendre. La psychologie, pour sa part, a déjà l'apprenant, c'est l'individu ; ce qu'elle souhaite c'est essayer de le comprendre et, pour cela connaître, les classes de concepts qu'il forme. A priori, la problématique est symétrique.

Pourtant, le fait que le modèle PAC étudie l'apprenabilité, l'oblige à définir l'apprentissage, son critère d'arrêt, les conditions dans lesquelles il opère, etc. Il définit ainsi tout un ensemble d'objets et de conditions qui font l'un comme l'autre l'objet de mesures. Rechercher ces objets et ces conditions dans les travaux des psychologues sur l'apprentissage inductif peut permettre d'envisager ces travaux d'une autre manière. Par ailleurs, leur analyse peut en retour amener à proposer, ou à justifier certaines modifications du modèle. C'est pourquoi dans le chapitre suivant nous essaierons de retrouver en psychologie les objets et les conditions suivants :

Les objets :
-une classe de concepts cible,
-une classe d'hypothèses de l'apprenant qui n'est pas nécessairement identique à la classe de concepts cible,
-un ensemble d'objets du monde,
-un échantillon d'objets étiquetés en exemples positifs ou négatifs relativement à un concept cible,
-un ensemble d'attributs permettant de décrire les objets et les exemples,
-un schéma de représentation permettant de décrire les concepts,
-un schéma de représentation permettant de décrire la classe d'hypothèses de l'apprenant,
-un oracle qui propose des exemples étiquetés d'un concept cible à l'apprenant selon une distribution de probabilités,

-un apprenant qui est un processus qui, sur la présentation d'un échantillon, retourne une hypothèse compatible avec cet échantillon.

Les conditions :
-une distribution de probabilité sur les représentations des objets du monde,
-une borne concernant le poids de l'erreur tolérable de l'hypothèse,
-une borne concernant l'échec possible,
-la manière avec laquelle sont présentés les exemples,
-l'information donnée à l'apprenant.

Chapitre 2

La catégorisation

2 Introduction

Le modèle PAC considère l'apprentissage inductif : sur la base d'un ensemble d'exemples étiquetés selon un concept l'apprenant doit être capable de se former une représentation proche de ce concept. En psychologie, ceci correspond aux études sur l'apprentissage de concepts et sur la catégorisation. L'objet général de ce travail est de vérifier s'il est pertinent d'étudier conjointement ce type d'apprentissage du double point de vue des théories formelles et de la psychologie.
Etudier la pertinence d'une approche conjointe consiste à contrôler plus précisément que les deux disciplines tentent bien d'approcher le même phénomène. *L'objectif principal de ce chapitre est ainsi de montrer que l'on peut utiliser les concepts inhérents au modèle d'apprentissage PAC pour exposer les études en psychologie. Ce faisant, nous validons du point de vue cognitif ce modèle. L'objectif second est de présenter ces études à des lecteurs non familiers du domaine.* Cette présentation ne prétend pas pour autant à l'exhaustivité. Plus particulièrement, comme le point de départ est le modèle PAC où la formation de la représentation d'une catégorie s'effectue essentiellement sur la base de l'information contenue dans les exemples, toutes les recherches relatives à l'inscription d'un concept dans un réseau de concepts sont ignorées. Ainsi en est-il de l'étude de la relation d'inclusion chère à Piaget («Y a-t-il plus de fleurs que de roses dans le vase ? »).

Etudier la pertinence consiste aussi à montrer que la double approche est fructueuse, c'est-à-dire que chacune des disciplines apporte quelque chose à l'autre. En établissant les parallèles entre les concepts du modèle PAC et ceux de l'étude de la catégorisation en psychologie nous serons amenés à présenter d'une manière un peu différente les problématiques abordées par cette dernière. En retour, l'analyse des expériences en psychologie invitera à questionner la pertinence de certaines conditions imposées dans le modèle PAC.

De ces objectifs, il découle que la démarche que nous suivons ici n'est pas une démarche usuelle en psychologie. Habituellement, dans ce domaine, un chercheur, s'intéressant à un problème donné, va d'abord exposer l'ensemble des résultats déjà obtenus relativement à ce problème puis présenter les fruits de ses propres recherches. Notre objectif étant de montrer la pertinence cognitive du modèle PAC, nous ne pouvons employer une telle démarche. *Ce que nous faisons, c'est partir des notions du modèle PAC et rechercher leur pendant en psychologie*. C'est pourquoi les expériences que nous présentons sont choisies non pas en raison de leur caractère récent ou définitif mais parce qu'elles permettent d'illustrer tel ou tel point que nous souhaitons aborder. Pour les mêmes raisons, nous sommes parfois conduits à exposer dans une même section des recherches qui sont habituellement présentées dans des champs bien séparés en psychologie. Cette démarche permet néanmoins de présenter les controverses qui divisent les psychologues et nous amène parfois à faire des propositions.

Dans un premier temps, nous faisons un rapide historique des recherches relatives à la catégorisation. Cet historique permet de montrer comment les chercheurs sont passés de l'étude de l'apprentissage de concepts à celle de la catégorisation.
Dans un second temps, nous nous intéressons à ce que peut recouvrir en apprentissage naturel la notion d'apprentissage *supervisé* du modèle PAC. Ce sera l'occasion de s'interroger sur le type de catégories que forme l'individu et pourquoi il les forme.
Dans le modèle PAC, il est supposé que l'apprenant ne travaille pas directement à partir des objets du monde mais à partir de représentations de ces objets. La catégorisation opérée par l'individu ne s'effectue pas non plus directement à partir des objets du monde, mais à partir de la perception qu'il en a, c'est pourquoi dans un troisième temps nous nous intéressons à la notion de traits et d'attributs en psychologie.
Dans un quatrième temps, nous décrivons ce qui, en psychologie, correspond à ce que l'on appelle «schéma de représentations » et «espace d'hypothèses de l'apprenant » dans le modèle PAC. C'est un sujet pour lequel la littérature en psychologie est particulièrement abondante et les controverses nombreuses.
Dans un cinquième temps, nous abordons le fait que, dans le modèle PAC, l'hypothèse fournie par l'apprenant est indissociable du processus qui l'a produite, ce qui mène à l'opposition en psychologie entre l'approche écologique et l'approche en laboratoire.
Enfin, le modèle PAC est aussi un modèle quantitatif. C'est un aspect qui est moins pris en compte en psychologie, nous en présentons quelques points.
Ces deux derniers points seront l'occasion d'insister sur les aspects méthodologiques des expériences.

2.1 Généralités

Cette section est surtout destinée au lecteur non familier de la psychologie. Nous proposons, d'abord, une définition de la catégorisation et de la catégorie et montrons en quoi elle s'oppose au concept. Puis, nous faisons un rapide historique et présentons trois expériences qui l'illustrent

2.1.1 Catégorisation, catégories, concepts

2.1.1.1 La catégorisation, une conduite adaptative fondamentale

Pour [Houdé, 92] la catégorisation est la conduite adaptative fondamentale par laquelle les systèmes cognitifs, biologiques ou artificiels «découpent » le réel physique et social. Cette conduite est *adaptative* car elle correspond à une dialectique entre système cognitif et environnement. Si comme le pense [Vignaux, 91], les psychologues cognitifs considèrent que les catégories psychologiques correspondent aux régularités dans le monde, aux invariants, elles ne se limitent pas à cela[24]. Une liste exhaustive de tous ces invariants ne permet pas, en elle-même, de comprendre comment l'homme catégorise[25]. Pour comprendre la catégorisation, il est nécessaire de prendre en compte le système cognitif qui les perçoit.
La catégorisation est *fondamentale* car elle est le pont entre perception et cognition, elle est à la base de la majorité des processus cognitifs [Barsalou, 92]. Ainsi en est-il de la langue : « La capacité de catégoriser les objets, les événements... devient très vite une capacité cognitive critique pour l'activité langagière... Si cette capacité à catégoriser n'existe pas, chaque occurrence d'un objet, d'un événement... doit être considérée comme un exemplaire unique, et la communication devient rapidement impossible. » [Cordier, 94].

2.1.1.2 Catégories

Cette catégorisation se fait sur la base de représentations. Nous rappelons que lorsque nous parlons de représentations nous ne faisons pas pour autant référence au symbolisme classique (voir l'introduction générale). Nous suivons en cela [Barsalou, 92] selon lequel le cerveau «représente » la chaise par des états du cerveau définis sur de larges populations de neurones et non par le mot «chaise ». [Smith et Medin, 81] donne la définition suivante d'une catégorie : c'est «la représentation d'une classe d'objets qui se trouvent regroupés sur la base de propriétés communes. Cette classe peut être dénommée par un mot correspondant à une unité sémantique ». Le défaut de cette définition, de notre point de vue, est, d'une part, qu'elle peut donner à penser que tous les éléments d'une catégorie partagent les mêmes propriétés, d'autre part, qu'elle n'insiste pas sur le fait qu'une catégorie est formée par l'individu pour s'adapter. C'est

[24] Vignaux fait référence à l'approche écologique de la psychologie cognitive selon laquelle il existe des invariants dans le monde qu'il s'agit de repérer.
[25] [LeNy, 89] «Il n'existe dans le réel rien qui soit en lui même abstrait et général, rien qui ne soit particulier et concret. »

pourquoi nous préférons partir de la définition de [Neisser, 1987] «catégoriser c'est traiter une série d'objets comme quelque peu équivalents : les mettre dans le même ensemble, les appeler du même nom, ou y répondre de la même manière.» pour proposer la définition suivante : *une catégorie est la représentation mentale chez l'individu d'un ensemble d'objets qu'il appelle du même nom ou qui induisent de sa part un comportement identique.* La seule propriété commune à tous les objets d'une même catégorie dont on soit sûr, est le rapport semblable que le sujet entretient avec eux.

2.1.1.3 Catégories vs concepts

Pour le modèle PAC, un concept est un sous-ensemble des objets du monde. Ceci n'est pas repris en psychologie où l'on distingue entre catégorie et concept bien que cette opposition ne soit pas toujours très bien explicitée. Ainsi [Hampton et Dubois, 93] proposent la distinction suivante selon laquelle, le concept de chaise, par exemple, est défini comme «l'état psychologique au moyen duquel une personne est capable de comprendre qu'un objet particulier...peut être considéré comme un type de chaise. Comprendre...signifie être capable de faire la liaison avec des connaissances précédentes à partir desquelles quelques inférences plausibles peuvent être faites. D'un autre côté la catégorie des chaises réfère à l'ensemble des entités qui peuvent être catégorisées comme une chaise en cela que le concept de chaise peut être utilisé pour les comprendre. » Il semble qu'ils considèrent la catégorie comme l'extension d'un ensemble d'objets et le concept comme sa définition en compréhension.

Nous préférons pour notre part nous appuyer sur une autre opposition : empirique/relationnel. Annick Weil-Barais [Weil-Barais, 93] parle ainsi de la conception empiriste des concepts : extraire des invariants des exemples pour former le concept, qu'elle oppose aux concepts relationnels. Ces derniers tels que les concepts mathématiques ne sont pas réductibles à une approche empiriste[26], ils proviennent d'une réflexion de la pensée sur elle-même qui s'impose des règles de cohérence, d'économie, de non-contradiction. Ils ne peuvent être formés de manière isolée. Cette distinction entre concepts empiriques et concepts relationnels se retrouve chez nombre de psychologues, mais rarement sous la même terminologie, Piaget (abstraction empirique, abstraction réfléchissante), Vygotsky (concept spontané, concept scientifique),...néanmoins l'idée est bien que :
- le concept spontané ou empirique est basé sur les exemples et il se définit sur la base des propriétés des exemples,
- le concept scientifique, logique se définit par ses relations aux autres concepts.

Les psychologues ont tendance maintenant à utiliser le terme de «catégorie » pour les concepts empiriques et de «concept » pour les concepts scientifiques ou relationnels. *C'est dans ce sens que [Pitt, 97] est amené à dire que le PAC apprentissage modélise la*

[26] Cependant même les concepts mathématiques peuvent exhiber des phénomènes de typicalité liés aux exemples (voir les effets de typicalité constatés par [Cordier, 93] dans l'apprentissage du théorème de Thalès).

catégorisation et non l'apprentissage de concepts.[27] Dans ce qui suit nous utiliserons les termes du modèle PAC de concepts cibles pour désigner les catégories de référence utilisées par l'expérimentateur et d'hypothèses de l'apprenant pour désigner les catégories formées par l'apprenant. Lorsque le chercheur utilise une cible, celle-ci est, pour lui, un concept clairement défini par ses relations aux autres concepts et non simplement une représentation qu'il a induite des exemples. A l'opposé ce qu'il étudie chez le sujet est une catégorie. F. Cordier distingue clairement entre les deux (cibles et hypothèses) ainsi dans [Cordier, 93] : «Nous soulignons ainsi qu'il s'agit bien, pour nos sujets, de «catégorisation naturelle» et non pas de l'étude d'une catégorie conceptuelle. »

Cette opposition entre concept et catégorie peut mieux se comprendre aussi au travers d'un bref historique de la recherche en psychologie

2.1.1.4 Rapide historique

Lorsque l'on fait une recherche concernant la catégorisation dans une base d'articles de psychologie, on distingue nettement deux époques. La séparation entre ces deux époques est ce que Neisser a appelé la «révolution roschienne» [Pacherie, 93]. C'est dans les années 70[28] que Rosch publie ses premiers travaux qui vont influencer toute la recherche ultérieure dans le domaine de la catégorisation. Ils remettent en question la possibilité que les catégories soient représentées en conditions nécessaires et suffisantes et ceci pour trois raisons [Pitt, 97] :
- l'impossibilité pour certaines catégories intuitives d'être définies par une règle adéquate,
- le problème des cas frontières : le mobilier est une catégorie mal définie, une photo en fait-elle partie,
- l'observation d'effets de typicalité : plus un élément a les caractéristiques jugées pertinentes pour l'appartenance à la catégorie, plus il est typique. Ainsi, le berger allemand est plus typique de la catégorie chien que le pékinois.

On peut ainsi caricaturer l'opposition entre ces deux époques par le passage de la définition d'un concept en conditions nécessaires et suffisantes à la définition d'une catégorie en air de famille. D'une époque à l'autre, le type et la forme des recherches ne sont plus les mêmes. Dans la première époque les articles portent essentiellement sur les difficultés relatives à apprendre selon la complexité des concepts : unidimensionnel, conjonctions, disjonctions, etc. [Bruner, Goodnow and Austin, 56], [Bourne,70], [Dominowski,73], [Neisser et Weene, 62].... Les tâches sont alors des tâches

[27] Historiquement, il est facile de comprendre pourquoi dans le PAC apprentissage on parle de concepts et non de catégorie. D'abord le modèle PAC (1984) suit de peu les recherches de Rosch, de plus il est en quelque sorte une continuation du modèle d'identification à la limite de Gold (1967), qui visait lui-même à modéliser (on peut le supposer) les tâches d'identification de concepts et de grammaires que l'on expérimentait en psychologie à l'époque.

[28] L'article le plus souvent cité étant de 75, « Family resemblances : studies in the internal structures of categories » [Rosch, Mervis 75] mais dès 71 elle publie des articles sur la typicalité

d'identification de concepts : on présente au sujet des exemples et il doit retourner une représentation du concept (le concept en compréhension ou intension).

Dans la seconde époque, avec la révolution roschienne, la question que se posent les psychologues concerne la manière dont les catégories sont encodées en mémoire à long terme. Les tâches données aux sujets doivent permettre à l'expérimentateur de connaître la structure de la représentation de la catégorie chez le sujet [Rosch et Mervis, 75], [Dubois, 93],[Medin , Wattenmaker et Hampson, 87]..... Pour cela on demande aux sujets de classifier des exemples, ou de donner le degré de typicalité des exemples, etc.. [Cordier, 93]

La psychologie est passée d'une recherche où le but était de savoir si les concepts d'une structure donnée pouvaient être induits par l'homme à partir d'un ensemble d'exemples, à une recherche où le but est de connaître selon quelle structure sont représentées les catégories que l'homme a induites à partir des objets qu'il a rencontrés dans sa vie quotidienne. On est passé de la question : «Est-ce que l'homme est capable d'apprendre cette catégorie ? » à la question «Quelles sont les catégories que l'homme se forme et quelle structure ont-elles ? ».

Les trois expériences que nous présentons ci-dessous, permettent d'illustrer cette évolution dans l'étude de la catégorisation.

2.1.2 Description de trois expériences

Les expériences décrites ici ont été choisies non pas en raison de leur valeur intrinsèque, bien que les deux premières aient été souvent citées, mais parce qu'elles permettent d'éclairer plusieurs points que nous serons amenés à évoquer dans ce chapitre. Elles montrent notamment l'évolution des recherches et l'impact des études de Rosch dans ces recherches. Cette évolution a pour mot d'ordre «toujours plus d'écologie » ce qui influe sur la forme que prennent les expériences[29]. On voit au travers de ces trois expériences comment on passe de l'étude de catégorie inventée par le chercheur pour l'expérience, à l'étude des catégories qui ont été formées en dehors de l'expérience par l'individu dans sa vie quotidienne et de cette dernière à la question : quelles sont les catégories que l'individu se forme ?
Cette présentation a aussi pour but de montrer au lecteur non initié la démarche qu'utilisent les chercheurs en psychologie. C'est pourquoi, contrairement à l'usage dans ce domaine, nous reprenons le même plan pour les détailler que celui utilisé par les auteurs dans leurs articles.

2.1.2.1 Hiérarchies dans l'acquisition de concepts, [Neisser & Weene, 62][30]

L'expérience décrite ici fait suite à celle de [Bruner, Goodnow and Austin, 56] qui étudient l'apprentissage des conjonctions et des disjonctions. Elle s'inscrit dans le cadre

[29] Cet aspect est développé dans la section 2.5 lorsque nous traitons des conditions de la catégorisation.
[30] « Hierarchies in concept attainment », [Neisser & Weene 62]

d'une définition des concepts *en conditions nécessaires et suffisantes (CNS)*. L'idée générale est que les concepts peuvent être organisés selon une hiérarchie de construction et que pour être capable d'apprendre des concepts de niveau 2 (par exemple les conjonctions), il faut avoir maîtrisé l'apprentissage de concepts de niveau 1 (les concepts unidimensionnels). Ce qui fait qu'un concept tel que «noir » est pensé être plus simple à apprendre qu'un concept tel que «blanc ou carré » lui-même plus simple que «(blanc ou carré) et (noir ou rond) ».

Ainsi [Neisser et Weene, 62] procèdent à une étude comparative des dix classes de concepts suivantes rangées en trois niveaux, chaque niveau étant construit sur le précédent. L'hypothèse est que *les concepts de plus haut niveau sont plus difficiles à apprendre que ceux des niveaux inférieurs.*
- niveau 1 : A, non A
- niveau 2 : (A et B), (A ou B), (A et non B), (A ou non B), (non A et non B), (non A ou non B)
- niveau 3 : ((A et non B) ou (non A et B)), ((A et B) ou (nonA et nonB))

Méthode

Sujets : les sujets sont 20 étudiants qui travaillent par groupe de 5, chaque matin durant 3 heures.

Matériel : les stimuli sont des cartes sur lesquelles est écrite une série de 4 lettres choisies parmi 5 : J, Q, V, X, Z. Une carte est de la forme QJVZ ou ZQJV : l'ordre n'a pas d'importance. Les sujets sont avertis que la redondance d'une lettre n'a pas d'importance non plus mais que son absence en a : la carte JJJV équivaut à JV nonQ non X non Z. Les sujets doivent donc se rappeler qu'il y a 5 attributs (les 5 lettres) qui peuvent prendre deux valeurs (présence / absence). Les concepts sont du type «V », «X ou J », «Q et non Z », etc.

Procédure : Préalablement à l'expérimentation les sujets sont informés sur les concepts cibles. Après une période d'entraînement avec trois concepts, les sujets sont testés sur 2 cycles de 10 concepts. L'ordre des concepts varie d'un groupe à l'autre.
L'expérimentateur présente les exemples un par un. Pour chaque exemple le sujet répond «plus » s'il pense que l'exemple est une instance positive du concept et «moins » sinon. Après chaque exemple, lorsque tous les sujets ont répondu, l'expérimentateur donne la réponse correcte. Le sujet a au plus 15 secondes pour répondre. Les auteurs considèrent que le sujet a appris correctement le concept lorsqu'il a étiqueté 25 exemples successifs en faisant au plus une erreur. L'apprentissage d'un concept s'arrête lorsque tous les sujets ont trouvé ou au bout de 100 exemples. Les exemples sont choisis de telle manière qu'il y ait autant de positifs que de négatifs.

Résultats : les résultats sont établis sur la base d'une comparaison deux à deux des dix concepts. La synthèse indique que les concepts de niveau 1 sont plus

facilement appris que ceux de niveau 2 et ces derniers plus facilement appris que ceux de niveau 3. Parmi les concepts de niveau 2 la disjonction (A ou B) est plus difficilement apprise que la conjonction (A et B) et le concept le plus difficile du niveau 2 est (non A ou non B)

Commentaires

Nous avons choisi de présenter cette expérience, d'une part, parce que la plupart des chercheurs qui travailleront par la suite sur le domaine la citent : [Conant et Trabasso, 64], [Dominowski, 73], [Richard, 75], [Bourne, 70 et 82]... mais surtout parce qu'elle est caractéristique de l'époque :
- le concept à apprendre est *artificiel,* il n'a d'existence que dans le cadre de l'expérience,
- le concept se définit en *conditions nécessaires et suffisantes.*

Si nous faisons un parallèle avec le modèle d'apprentissage PAC, nous voyons que le protocole d'expérimentation le suit exactement. Nous avons une classe de concepts cibles clairement définie sur des attributs clairement définis aussi qui permet à l'expérimentateur d'étiqueter sans ambiguïté les exemples. Il s'agit d'un apprentissage prédictif dans le sens où l'on ne demande pas à l'apprenant de définir son hypothèse en compréhension («c'est le concept «X ou J » ») mais d'étiqueter correctement les exemples. Il s'agit d'un apprentissage approximatif puisque le sujet est autorisé à faire une erreur sur 25.

2.1.2.2 Air de famille : études de la structure interne des catégories [Rosch et Mervis, 1975][31]

Dans les années 70, les travaux de Rosch, dont nous présentons une expérience ci-dessous, vont amener les psychologues à amorcer un virage dans la recherche. On ne parle plus de *concept* mais de *catégorie*. Par ailleurs, il est supposé que les représentations qu'ont les sujets des catégories ne sont pas construites en conditions nécessaires et suffisantes mais en *air de famille* : les membres de la catégorie ressemblent les uns aux autres. De plus, certains items d'une catégorie sont plus typiques que d'autres (la vache est un mammifère plus typique que la baleine). Enfin, l'accent est mis sur l'aspect écologique de la catégorisation.

Ainsi, dans l'expérience 1 de leur article «Air de famille : études de la structure interne des catégories », [Rosch et Mervis, 1975] se proposent d'étudier la relation existant entre le degré de liaison entre les membres d'une catégorie («*l'air de famille* ») et le taux de typicalité de ces membres (un marteau est un outil plus *typique* qu'une bêche, il a donc un degré de typicalité plus élevé). L'hypothèse est que la mesure du degré de l'air de famille qui relie un item aux autres membres de la catégorie est corrélée à son degré de typicalité.

[31] Family resemblance : studies in the internal structure of categories [Rosch et Mervis, 1975]

Méthode

Sujets : 400 étudiants en psychologie

Matériel : six catégories pour lesquelles les degrés de typicalité de 50 à 60 items ont déjà été établis par Rosch au cours d'une précédente expérimentation (voir le tableau 2.1 page suivante). Pour chaque catégorie, 20 items ont été choisis qui couvrent l'ensemble des degrés de typicalité.

Procédure : chaque sujet a 6 items (1 par catégorie). Les items sont différents d'un sujet à l'autre et tous les items sont vus par 20 sujets. Chacun des 6 items est écrit en haut d'une page. Pour chaque item, le sujet a une minute trente pour donner tous les attributs qui lui viennent à l'esprit.

Mesure de l'air de famille : pour chaque item, on établit la liste de tous les attributs définis par tous les sujets. Des juges éliminent les attributs manifestement faux ou dont ils doutent. Par ailleurs, s'ils considèrent qu'un attribut relevant d'un item peut aussi relever d'un autre item, ils l'ajoutent à la liste des attributs de ce dernier.
Chaque attribut reçoit un score de 1 à 20 correspondant au nombre d'items dans la catégorie crédités de cet attribut. A partir de ce score on calcule la mesure du degré d'air de famille en faisant la somme des scores pondérés de chacun des attributs dont l'item a été gratifié.

Les résultats : pour chaque catégorie, il y a très peu d'attributs partagés par les 20 items. De fait, le nombre d'attributs commun à plusieurs items d'une même catégorie décroît au fur et à mesure que croît le nombre d'items considérés. Ainsi, pour quatre catégories, seulement un attribut est partagé par tous les items, ce qui n'est pas le cas pour les deux autres. Il apparaît donc qu'une catégorie ne se définit pas par le fait que quelques attributs soient vrais pour tous ses membres mais par le fait qu'un grand nombre d'attributs soient vrais pour certains.
Par ailleurs la mesure de l'air de famille d'un item et son degré de typicalité sont corrélés : plus un item a d'attributs en commun avec les autres membres de la catégorie plus son degré de typicalité est élevé.

Commentaires
Ici, le protocole expérimental ne suit plus du tout le modèle PAC. Si on considère que la représentation d'une catégorie est le résultat d'un apprentissage, l'apprentissage ici a eu lieu en dehors de l'expérience. Lorsque le sujet arrive à l'expérimentation, son hypothèse est déjà construite. Le rôle du psychologue va consister à essayer de définir son schéma de représentation (au sens du chapitre 1). Comme l'apprentissage n'a pas lieu durant l'expérience nous n'avons aucune indication sur les exemples qui ont servi à

la formation de cette hypothèse. Il y a bien toujours un concept cible[32] mais ce concept cible n'est pas clairement défini et il faut faire appel à des «juges» pour déterminer si oui ou non un attribut est vrai pour l'item. Enfin la notion même d'attribut est ambiguë, s'agit-il d'un attribut ou d'une valeur d'attribut.

Catégories surordonnées et les items utilisés dans l'expérience 1						
			catégorie			
Item	Mobilier	Véhicule	Fruit	Arme	Légume	Vêtement
1	Chaise	Voiture	Orange	Revolver	Petits pois	Pantalon
2	Canapé	Camion	Pomme	Couteau	Carottes	Chemise
3	Table	Bus	Banane	Epée	Haricot vert	Robe
4	Vaisselier	Moto	Pêche	Bombe	Epinard	Jupe
5	Bureau	Train	Poire	Grenade	Brocoli	Veston
6	Lit	Tramway	Abricot	Lance	Asperge	Manteau
7	Bibliothèque	Bicyclette	Prune	Canon	Blé	Pull-over
8	Tabouret	Avion	Raisin	Arc et flèche	Chou fleur	Slip
9	Lampe	Bateau	Fraise	Matraque	Chou de Bruxelles	Chaussette
10	Piano	Tracteur	Pamplemousse	Tank	Laitue	Pyjama
11	Coussin	Charrette	Ananas	Gaz lacrymogène	Betterave	Costume de bain
12	Miroir	Fauteuil roulant	Myrtille	Fouet	Tomate	Chaussure
13	Tapis	Tank	Citron	Pic à glace	Haricot de Lima	Veste
14	Radio	Radeau	Pastèque	Poing	Aubergine	Cravate
15	Poêle	Traîneau	Melon	Roquette	Oignon	Moufle
16	Pendule	Cheval	Grenade	Poison	Pomme de terre	Chapeau
17	Photo	Dirigeable	Date	Ciseaux	Patate douce	Tablier
18	Placard	Patins	Noix de coco	Mots	Champignon	Porte-monnaie
19	Vase	Brouette	Tomate	Pied	Potiron	Montre
20	Téléphone	Ascenseur	Olive	Tournevis	Riz	Collier

Tableau 2.1 Catégories surordonnées et les items utilisés dans [Rosch and Mervis, 75]

2.1.2.3 Classification de photos de sections de route [Mazet, 93]

Il est reproché à Rosch de n'être pas encore assez écologique. [Dubois, 93] remarque ainsi «Le caractère ambivalent des conceptualisations de Rosch : se réclamant d'une conception 'écologique' des représentations cognitives, c'est-à-dire de celles qui se trouvent opératoires dans la vie de tous les jours, elle utilise constamment la référence à une organisation savante, normative, scientifique des connaissances. ». Alors que dans la première expérience, le concept cible est défini par l'expérimentateur en conditions nécessaires et suffisantes, alors que dans la seconde, il s'agit de catégories définies par un mot, ici, le concept de référence est lui-même l'objet d'investigation. Les chercheurs se posent la question de savoir quelles sont les catégories que l'homme forme et pourquoi il les forme.

[32]Dans l'expérience, il y a en fait 6 concepts cibles mais on pourrait assimiler cette expérience à 6 expériences différentes ayant chacune un concept cible.

Ainsi, [Mazet, 93] s'intéresse aux fonctionnalités des représentations. Elle considère que les recherches sur les représentations sont trop axées sur les fonctions cognitives, c'est-à-dire que les représentations sont envisagées comme un découpage du monde qui n'a pour but que la connaissance. «Les catégories étudiées ne sont ni construites pour répondre aux objectifs d'une tâche, ni même élaborées pour servir l'action. »

Le matériel : 50 photographies de sections de routes de campagne

Les sujets : 40 experts, ce sont des personnes qui conduisent depuis plus de 5 ans

Procédure : on demande aux sujets de classer les différentes photos en tenant compte de leur expérience. Ils sont libres de créer autant de classes qu'ils le souhaitent. Après le classement les sujets sont invités à indiquer verbalement les critères qu'ils pensent avoir utilisés.

Résultats 1 Les photographies des sections de routes sont réparties en 9 classes de tailles inégales (de 3 à 11 photos), au sein de chaque classe certaines photographies sont considérées comme plus ou moins typiques en fonction de la dangerosité.

Résultats 2 : L'analyse des déclarations verbales fournies par les sujets montre que :
- les classes sont décrites principalement en termes de problèmes potentiels (visibilité, niveau de vitesse, etc.)
- secondairement des caractères perceptifs de l'environnement interviennent (route droite, présence d'arbres sur les bas côtés, etc.)

Ainsi, le classement s'opère en fonction de la dangerosité de la route : les sujets forment les catégories qui leur sont utiles.

Commentaires
Nous sommes encore plus éloignés du modèle PAC dans cette expérience-ci. Comme précédemment la catégorisation s'est effectuée en dehors de l'expérimentation, au cours de la conduite quotidienne de leur véhicule par les sujets. L'expérimentateur présente bien aux sujets des exemples d'un concept (celui de section de route) mais ce qui l'intéresse, c'est la manière avec laquelle les sujets vont établir différentes sous-catégories. Si on se réfère aux dénominations définies dans le PAC, le concept section de route n'est pas vraiment le concept cible, il sert plutôt à définir l'ensemble des exemples. Les concepts cibles seraient plutôt les différentes sous-catégories que les sujets ont construites. Ces sous-catégories correspondraient au degré de dangerosité des différentes sections routes.
Le modèle PAC décrit l'apprentissage supervisé, l'observateur étiquette les exemples que l'on présente à l'apprenant. Selon la manière d'envisager le problème on peut considérer que la précédente expérimentation entre encore ou non dans ce cadre. Elle entre encore dans le cadre parce que l'étiquetage d'une section de route en plus ou moins dangereux est imposé par l'environnement au sujet : s'il n'en tient pas compte, il

peut avoir un accident. Elle n'entre plus dans le cadre de l'apprentissage supervisé car cet étiquetage par l'environnement n'est pas explicite.

2.1.2.4 PAC et protocole expérimental

Ces trois expériences ont été choisies car elles montrent l'évolution de la problématique en psychologie. Il convient, cependant, de ne pas se méprendre sur leur exemplarité. Toutes les expériences actuelles sur la catégorisation ne consistent pas à chercher quelles sont les catégories que le sujet forme. On continue encore à faire des expériences similaires à la première sur des catégories artificielles pour étudier tel ou tel aspect[33] de la catégorisation ([Bourne, 82], [Pazzani, 91]....), ou similaires à la seconde sur des catégories bien connues ([Cordier, 93]...).

Si nous avons choisi de les présenter, c'est aussi parce qu'elles permettent de montrer la distance qui peut exister entre elles et le modèle PAC lorsqu'on envisage celui-ci en tant que protocole expérimental. De ce point de vue, seule la première expérience correspond. Dans les deux autres, *le processus de catégorisation s'étant effectué en dehors du laboratoire*, il n'y a plus de correspondance terme à terme entre le protocole et le modèle.
Ceci est évident au niveau du rôle que joue l'échantillon dans l'expérience. Dans la première, l'échantillon d'exemples est un échantillon d'apprentissage comme dans le modèle PAC ; il est présenté à l'apprenant pour que celui-ci se forge le concept. Dans les deux suivantes, l'échantillon n'est là que pour évoquer chez le sujet des représentations qu'il a déjà formées.
De la même façon, la classe de concepts de référence n'a pas toujours un rôle conforme dans les expériences à celui que joue la classe de concepts cible dans le modèle PAC. Elle est utilisée de manière identique lors des expériences préroschiennes d'identification de concept, (première expérience, voir aussi [Bruner, Goodnow and Austin, 56]...., et l'apprentissage de catégories artificielles [Cordier, 83]) ou lorsqu'elle permet de rapprocher les représentations de l'apprenant d'une représentation standard, ce rapprochement n'étant destiné qu'à comprendre la représentation du sujet (deuxième expérience, voir aussi [Cordier, 93]). Elle a un tout autre rôle lorsqu'elle sert à définir un «monde » pour l'apprenant, monde que celui-ci sera amené à éclater en classes. Ainsi dans la troisième expérience [Mazet, 93], le travail consiste, non pas à comparer la représentation du sujet avec une représentation standard dans la culture puisqu'il n'y en a pas, mais à repérer les attributs que le sujet a jugés pertinents pour établir différentes sous-classes de routes. Enfin, il existe encore d'autres types de concepts référents tel que ceux construits sur une composition de concepts (conjonction, disjonction, etc. par exemple «fruits ou légumes ».)[Hampton, 88,a], [Hampton, 88,b], [Egeth, Virzi, Garbart, 84]. Dans ce cas l'étude consiste à comparer la représentation qu'a le sujet de chacune des catégories avec la représentation de la composition de ces catégories.

[33] Les tâches d'identification de concepts sont un sous-ensemble des tâches d'identification de règles, et même après les travaux de Rosch ce type de recherches a continué mais concernant les identifications de grammaire [Reber, 76][Reber et Allen, 78]...

Ainsi, envisager le modèle PAC comme protocole d'expérimentation n'a de sens que sur des catégories artificielles, c'est ce qui sera fait au chapitre 5. Ce n'est pas pour autant que le modèle PAC ne permette pas de «lire » les résultats de la psychologie car dans les trois expériences, *il n'est pas considéré que l'homme ne puisse pas catégoriser sur la base d'un échantillon d'exemples.*

2.2 Quelles catégories forme-t-on et pourquoi ?

La problématique de l'apprentissage PAC est de connaître les classes de concepts *apprenables*. De cette problématique, il découle que l'apprentissage PAC est un apprentissage *supervisé* : un oracle va *imposer* à l'apprenant les exemples étiquetés selon un concept pour vérifier que ce concept est apprenable
Un pendant de cette problématique en psychologie est de découvrir *quelles sont les catégories apprises* par l'être humain. Une problématique sous-jacente est de comprendre *pourquoi* l'individu se forme tel type de catégorie plutôt que telle autre. Si on prend le même cheminement que pour le modèle PAC, la réponse à cette question est que *l'apprenant se voit imposer par l'environnement les catégories qu'il forme*. En d'autres termes, la catégorisation naturelle est un apprentissage supervisé. Cette affirmation est assez forte et il est nécessaire de l'étayer c'est ce que nous allons faire dans cette section.

Il convient préalablement de préciser ce que le modèle d'apprentissage PAC considère comme *apprentissage supervisé* car le terme peut prêter à confusion. Derrière ce terme, il y a une notion de feed-back qui n'existe pas dans le modèle PAC. Lorsque l'on pense «apprentissage supervisé », il y a l'idée d'un professeur qui est amené à corriger les erreurs de l'élève[34]. Ce n'est pas ainsi que le modèle PAC l'entend. Dans ce modèle, «supervisé» signifie uniquement que les exemples sont étiquetés selon un concept. En fait, le terme d'apprentissage *«imposé»* correspondrait mieux, dans le sens où l'on impose la catégorie à apprendre.

Dans le modèle, cette notion d'apprentissage supervisé entraîne celles subséquentes de :
- classe de concepts cible,
- schéma de représentation permettant de décrire ces concepts,
- échantillon d'objets étiquetés en exemples positifs ou négatifs relativement à un de ces concepts,
- oracle qui propose des exemples étiquetés d'un concept cible à l'apprenant.

Si on reprend la première expérience du 2.1.2, on retrouve chacune de ces notions. On peut d'ailleurs généraliser à toutes les expériences utilisant des catégories artificielles.

[34] Ce type de démarche correspond davantage au Minimal Adequat Teacher proposé par Angluin, décrit au chapitre 1, et ne relève pas à proprement parler du modèle PAC, pour la simple raison qu'il fait disparaître la condition d'apprentissage pour toutes distributions de probabilités.

Pour la seconde expérience, où la catégorie est caractérisée par un mot, nous allons montrer que nous les retrouvons aussi mais de façon moins immédiate. Ce sera l'occasion de présenter les recherches sur les catégories lexicalisées et les problèmes qui s'y posent. Enfin, la troisième expérience pose la question même de savoir si la catégorisation est un apprentissage supervisé. Nous allons montrer en utilisant les résultats de diverses recherches que toute catégorisation y compris celle de catégorie non lexicalisée est un apprentissage supervisé.

2.2.1 Les catégories lexicalisées, les catégories artificielles

Les catégories *lexicalisées* sont des catégories qui correspondent à des lexèmes dans la langue. [Barsalou, 85] parle de catégories taxonomiques. En psychologie on distingue entre catégorie naturelle, catégorie artefact ou manufacturées (ensemble des objets fabriqués par l'homme), catégorie de procès (action), catégorie sociale. Pour les catégories naturelles, les champs lexicaux les plus fréquemment utilisés dans les expériences relèvent des taxonomies de plantes et d'animaux. Pour les catégories artefacts, ce sont le plus souvent le mobilier, les vêtements. Dans les catégories de procès, on considère les changements qui font passer d'une situation initiale à une situation finale et on examine des verbes tels que 'tomber'...[Cordier, 93]. L'analyse en catégories sociales consiste à étudier le découpage effectué par le sujet de la société en groupes sociaux. Cette liste de catégories n'est pas exhaustive, il existe d'autres types de catégories étudiées selon les expériences. Elle n'est pas non plus arbitraire, les psychologues ont constaté, par exemple, que les catégorisations d'objets naturels ne sont pas du même type que celles des objets artefacts [Keil, 89].

Le travail du psychologue avec ces catégories consiste à étudier la différence entre la représentation qu'a le sujet de la catégorie et la définition de cette catégorie (traduit en «langage» PAC : la différence entre l'hypothèse de l'apprenant et le concept cible). Ainsi dans le cadre de la typicalité, si au niveau des définitions la vache et l'ours sont l'une et l'autre, à titre égal, des mammifères, il n'en sera pas de même dans la représentation qu'en a le sujet où la vache est un mammifère plus typique que l'ours.

Le problème avec ce type de catégorie c'est que l'on ne connaît pas précisément les conditions dans lesquels s'est effectué la catégorisation. C'est pourquoi : «...Le choix de catégories naturelles ne permet pas toujours de contrôler avec suffisamment de rigueur certaines variables comme la familiarité, la fréquence d'usage des mots et donc d'écarter une influence possible dans l'interprétation des effets facilitateurs. Par contre le choix de catégories artificielles telles qu'ensembles de points, bonhommes stylisés, suite de lettres permet à l'expérimentateur d'en cerner précisément l'organisation. » [Cordier, 93]. Les catégories *artificielles* ne sont pas à confondre avec les catégories artefacts vues plus haut. Ces catégories artificielles sont définies uniquement pour l'expérience. Ainsi les expériences d'identification de concepts (conjonctifs, disjonctifs, etc.) entrent dans ce type de catégorie. L'avantage pour l'expérimentateur est que les conditions dans lesquelles s'opère la catégorisation sont plus nettement définies et permettent mieux d'isoler les caractéristiques pertinentes du processus.

Aussi bien pour les catégories artificielles que pour les catégories lexicalisées, on peut considérer que la catégorisation est bien un apprentissage *supervisé/imposé*. Dans le cas des catégories artificielles, l'expérimentateur correspond à l'oracle, c'est lui qui *étiquette les exemples* pour l'apprenant. Dans le cas de catégories lexicalisées, c'est l'environnement social/humain de l'individu qui lui *nomme les objets* qu'il rencontre. Par le biais des mots, la société impose à l'individu de regrouper dans une même catégorie certains éléments du monde.
Dans les deux cas aussi, il existe *un schéma de représentations* permettant de représenter le concept cible : la *langue*. La représentation d'un concept dans le cas de catégorie artificielle pourra être toute règle que le chercheur souhaite que le sujet apprenne (une disjonction, une conjonction,....). Dans le cas d'une catégorie lexicalisée, la représentation d'un concept *cible*[35] sera la définition que la langue donne du mot.

Du point de vue méthodologique, cette définition du mot par la langue ne va pas toujours sans poser problèmes aux chercheurs. En informatique, le concept cible permet un étiquetage sans ambiguïté[36], ce n'est pas systématiquement le cas en psychologie. Si certaines catégories cibles sont définies scientifiquement, telles celles des «oiseaux», d'autres sont plutôt floues. Lorsqu'on s'intéresse notamment aux catégories artefacts telles que «le mobilier», l'appartenance commence à devenir problématique (une photographie encadrée fait-elle partie du mobilier). Cela devient réellement un problème lorsque l'on passe aux catégories sociales. Dans ce dernier cas, les psychologues font faire, préalablement à l'expérience, une expertise par des «juges» pour définir si tel ou tel élément relève de la catégorie ou non. [Hutteau, 93] décrit ainsi comment des juges à partir de la catégorie «croyance» doivent définir si «l'athéisme» et «l'évolution» en sont des exemples. Le résultat est que, selon la procédure utilisée, «l'évolution» passe d'exemple atypique à contre-exemple et «l'athéisme» d'exemple négatif à exemple positif. Ceci l'amène à dire que «la comparaison de ces résultats nous en apprend sans doute davantage sur la variabilité des préférences idéologiques que sur la structure de la catégorie 'croyance'.»
Ainsi le choix de la catégorie cible lors des expérimentations et sa définition plus ou moins nette est un réel problème pour la psychologie. Lorsque l'on reste dans le domaine des catégories définies scientifiquement les problèmes méthodologiques ne sont pas trop criants. Mais dès que l'on quitte ce type de catégorie, les problèmes ont tendance à s'amalgamer. Ainsi les phénomènes de typicalité qui ne se situent normalement qu'au niveau des représentations du sujet vont se confondre avec des problèmes d'appartenance mal définis qui eux concernent la classe de concepts cible. Comme le fait remarquer Hutteau, on ne sait plus si ce sont les classes de concepts cibles qui sont l'objet de l'étude ou les classes d'hypothèses de l'apprenant. C'est ce qui amène F. Cordier [Cordier, 93] à utiliser comme catégories cibles, les catégories naturelles pour lesquelles l'appartenance ne fait pas problème.

---

[35] Par opposition à l'hypothèse de l'apprenant.
[36] Il existe aussi des recherches sur l'apprentissage bruité, c'est-à-dire un apprentissage où l'étiquetage des exemples n'est pas toujours correct. Mais l'erreur sur l'étiquetage ne provient pas de la définition du concept en soi mais par exemple de l'oracle qui étiquetterait mal.

Avec ces deux types de catégories (artificielle et lexicalisée), la catégorisation consiste bien en un apprentissage supervisé. On retrouve le fait que l'étiquetage des exemples est imposé à l'apprenant par une entité extérieure. Les notions de catégories cibles (lexème ou catégorie artificielle) et de schéma de représentations pour définir ces catégories (la langue) sont présentes aussi. Mais ces catégories ne sont pas les seules que forment l'être humain. Dans la troisième expérience que nous avons décrite, il apparaît que les sections de route sont catégorisées naturellement par le sujet selon leur dangerosité. La question qui peut se poser est de savoir si systématiquement toute catégorisation y compris celles qui ne sont pas lexicalisées est apprentissage supervisé.

2.2.2 L'individu forme les catégories qui lui sont utiles pour s'adapter à l'environnement

Dans cette section nous voulons montrer que toute catégorisation est apprentissage supervisé. Le raisonnement part de la définition de Houdé vue plus haut selon laquelle la catégorisation est la conduite adaptative fondamentale par laquelle l'individu découpe son environnement. Les recherches que nous allons présenter montrent toutes que les individus regroupent dans une même catégorie les objets du monde qui entraînent de leur part un même comportement, autrement dit qui ont la même utilité/fonctionnalité pour eux. Cette utilité n'est jamais que l'expression de cette nécessaire adaptation à l'environnement. Comme l'environnement est premier, l'individu ne l'a pas choisi, c'est donc lui qui impose les catégories à former. Il s'agit donc bien d'apprentissage supervisé. Toutes les catégories lexicalisées entrent dans ce schéma, faute de les construire l'individu ne sera pas capable de communiquer, de s'adapter à son environnement social. Mais ce ne sont pas les seules, le classement des sections de route en dangereuses ou non est imposé au conducteur par l'environnement, si ce n'est pas fait l'accident guette. L'enfant qui se brûle aura vite fait de catégoriser le feu comme objet dangereux sans que sa mère ait besoin de le lui dire.

Les recherches que nous présentons maintenant montrent donc que l'individu place dans une même catégorie des objets qui entraînent de sa part un même comportement. Ces recherches sont d'habitude l'objet d'études bien séparées et poursuivent un autre but que celui que nous leur assignons présentement. Nous espérons ne pas trop sombrer dans l'hérésie en agissant ainsi.

2.2.2.1 Les catégories ad hoc de Barsalou

Les *catégories ad hoc* ont été proposées par [Barsalou, 83]. L'exemple le plus fréquemment cité ([Pitt, 97], [Thibaut, 97]...) est celui-ci : dans quel cas mettrez-vous dans la même catégorie vos enfants, vos bijoux, vos photos de famille, votre chien, etc. ? Réponse : dans le cas où votre maison brûle ; ce sera la catégorie des objets à sauver. Cette catégorie a la particularité d'être définie par le contexte, c'est le contexte qui impose à l'individu la façon de regrouper les objets. Néanmoins, on peut se poser la question de savoir si on peut réellement parler de catégorie dans ce cas-ci car la démarche semble s'apparenter davantage à de la résolution de problème plutôt qu'à de

l'induction sur la base d'exemples. Ainsi, la catégorie n'est pas formée avant l'événement et il est probable qu'il n'y ait pas de mémorisation de celle-ci par la suite.

2.2.2.2 Les catégories slot filler de Nelson

Les *catégories slot filler* ont été proposées par Nelson. De notre point de vue, ces recherches, tout en étant très différentes, ont des points communs avec celles de Barsalou. Pour Nelson, [Nelson, 85], les connaissances en mémoire chez le jeune enfant sont organisées en scripts. Un script est la représentation d'un événement ou d'une séquence d'événements organisée selon des relations causales ou temporelles. L'enfant se formera d'abord des scripts élémentaires représentant des événements routiniers, le script « boire », le script « manger », puis, plus tard dans le développement des scripts plus complexes tels que celui du petit-déjeuner. *La catégorie est alors constituée des éléments qui peuvent occuper la même place dans le script.* Ainsi l'enfant mettra dans une même catégorie le lait et le jus de fruit parce qu'il les boit. Dans le script « boire », le lait et le jus de fruit ont *la même fonction* : ils sont *substituables.* Ici comme précédemment, c'est le contexte qui impose le regroupement dans une même catégorie. *Les objets sont regroupés en fonction de leur utilité/fonctionnalité pour l'enfant.*

2.2.2.3 Les catégories de base

C'est avec les travaux de Rosch que le terme de catégorie de base est apparu en psychologie. Cette notion est largement acceptée par les psychologues mais aussi par des spécialistes d'autres domaines : [Barth, 87] en pédagogie, [Edelman, 92] en neurobiologie.
F.Cordier [Cordier, 93] en donne la définition suivante : le niveau de base est le niveau le plus général, donc le plus abstrait pour lequel les sous-catégories[37] appartenant à une même catégorie :
- ont en commun un nombre important de propriétés (tous les chats miaulent, ont des griffes)
- suscitent un même comportement de la part de l'observateur (on caresse les chats, on les nourrit)
- possèdent des caractéristiques perceptives (par l'observateur) semblables aisément figurables.

Dans la littérature psychologique, les catégories de base sont souvent interprétées en termes de niveau. Ainsi il est considéré qu'il y a trois niveaux d'abstraction : de base, sur-ordonné, sous-ordonné. Le niveau de base est celui qui permet une différenciation maximale entre catégories de niveaux d'abstractions semblables. Comme exemples de catégories de niveau de base nous pouvons citer le «chien », le «chat », la «vache », la «chaise », la «table », etc. Le niveau de base a des caractéristiques concrètes (structures, composants), alors que le niveau sur-ordonné des caractéristiques plus abstraites. Si nous reprenons les catégories précédemment citées, les catégories sur-ordonnées

[37] On utilise le terme de sous-catégorie d'une catégorie plutôt que celui d'objet. En effet, lorsqu'on dit qu'un «marteau» relève de la catégorie «outil», on ne pense pas à un marteau particulier mais à la sous-catégorie «marteau».

pourront être les «mammifères» et le «mobilier ». Les catégories sous-ordonnées se distinguent souvent par leurs fonctions, la «chaise de la cuisine », la «chaise du salon », et elles apportent peu d'information supplémentaire par rapport au niveau de base, un «berger allemand », un «labrador ».

Ce n'est pas tant l'interprétation en termes de niveaux qui nous intéresse ici, que le fait que les éléments des catégories de bases sont ceux qui induisent un même comportement de la part du sujet, ce que les comportementalistes interpréteraient en termes de réponse identique. Nous mettons dans une même catégorie des éléments qui induisent de notre part un comportement semblable (cf. la définition de Neisser au début de ce chapitre).

2.2.2.4 L'individu regroupe dans une même catégorie les éléments qui ont une même fonction pour lui

Ces recherches, bien que fort différentes, ont comme point commun de considérer que l'individu regroupe dans une même catégorie les éléments qui ont une même *fonction pour lui*. Autrement dit, l'environnement lui impose la façon de le catégoriser pour s'y adapter, la catégorisation est un apprentissage supervisé. Cette idée est partagée par [Mazet, 93] «Nous avons commencé de montrer que les fondements des structures catégorielles n'étaient plus assurés par les seules propriétés intrinsèques du monde (les corrélats d'attributs perceptifs), mais par les régularités des comportements en œuvre dans une activité déterminé. Nous rejoignons sur ce point les positions de Dubois (1990) qui insiste sur le fait que «ce n'est plus la perception comme processus «reflétant » des structures du monde réel, mais les activités finalisées des sujets (ou opérateurs) qui doivent être pris en compte comme principe organisateur des catégories. »
Nous constatons ainsi que les catégories précédentes (lexicalisées, artificielles) ne sont alors qu'un cas particulier de cette règle. Les catégories étiquetées par un lexème sont des catégories formées par le sujet devant la nécessité (l'utilité) pour lui de communiquer avec ses semblables. Les catégories artificielles ne sont jamais qu'une réponse à la demande d'un expérimentateur. Il est d'ailleurs probable que le sujet fasse beaucoup plus de catégories que celles décrites ici.

Si l'environnement impose les catégories à former, se pose alors le problème de l'étiquetage pour les catégories non lexicalisées. La réponse est dans l'interaction de l'individu avec les objets. La première fois, il est possible qu'un enfant empoigne à pleines mains des orties, il est alors probable que la sensation d'irritation (étiquetage) qu'il en recueillera, fera qu'il ne renouvellera plus ce comportement.

2.2.2.5 Intérêt d'une telle question pour la psychologie

Lorsque l'on s'interroge pour savoir si la catégorisation est un apprentissage supervisé, nous essayons de répondre à la question du *pourquoi* de la catégorie. Pourquoi l'individu forme-t-il telle catégorie plutôt que telle autre ? Cette question est actuellement l'objet de recherche (cf. l'expérience de Mazet ci-dessus). Interprété dans

les termes du modèle PAC, nous dirions que les psychologues cherchent à définir quels peuvent être les concepts cibles. Longtemps cette question a été confondue avec celle du *comment*, c'est-à-dire par quels moyens l'individu se construit sa représentation de la catégorie. Ceci est particulièrement évident lorsque l'on affirme que l'individu assemble dans une même catégorie des objets qui se ressemblent. Cette affirmation se voit généralement réfutée par le fait qu'elle pose le problème du choix des attributs pertinents [Thibaut, 97] : un chien et une maison ont comme attributs communs d'être à des milliers de kilomètres du soleil, de mesurer l'un et l'autre moins de 500m de haut, etc., ce n'est pas pour autant qu'on les met dans la même catégorie. Par contre, une fois que l'environnement a imposé à l'individu les éléments à regrouper dans une même catégorie, il est probable que l'individu utilise la similarité pour se construire une représentation de cette catégorie, comme on le verra au chapitre 3.

2.3 Représentation des objets du monde

Dans le modèle PAC, il est supposé que l'apprenant ne travaille pas directement à partir des objets du monde mais à partir *de représentations de ces objets.* Ces représentations sont construites sur la base de descripteurs : attributs (binaires, ternaires...), mots d'un alphabet. Le pendant de ceci en psychologie est la notion d'attributs, de traits, etc.

2.3.1 Trait/propriété physique

En général les psychologues font la différence entre les propriétés des objets du monde physique (les invariants) et la perception de ces propriétés par un système cognitif. C'est dans ce sens que Le Ny [Le Ny, 89] fait remarquer que «le concept de chien n'aboie pas. La propriété relève du monde réel, le trait de l'intellect. » Pour Le Ny un trait est un support de discrimination. Les psychologues appellent généralement «discrimination » l'activité psychologique par laquelle un individu distingue une entité A d'une entité B. Le trait est une représentation composante et dans le meilleur des cas une connaissance élémentaire. Etre marron ne peut être instancié que par des réalités qui sont porteuses de marron : une représentation comme moineau «contient » une certaine représentation de la propriété, une représentation de la couleur marron. Certains traits sont lexicalisés et explicitables : c'est le cas de marron par rapport à moineau. Dans d'autres cas *l'existence d'un trait particulier, à l'intérieur d'une représentation conceptuelle, peut n'être pas consciente ni explicite pour le locuteur lui-même* ; elle n'est alors mise à jour que par une analyse extérieure. Il existe des combinaisons de traits séparables (ex : forme, taille) et des combinaisons de traits dont la séparation n'est psychologiquement pas possible, bien qu'elle le soit objectivement (exemple : dans la perception des objets colorés, la brillance et la tonalité chromatique). Ce type de trait permettra à un individu de remarquer que deux objets sont différents mais il sera incapable de définir le trait qui marque cette différence.

2.3.2 Trait/attribut

Il faut distinguer entre un attribut et sa valeur. Le trait peut être assimilé à la valeur d'attribut, ainsi dans le couple attribut-valeur : couleur-marron. *On perçoit immédiatement une valeur d'attribut plutôt que l'attribut* : rouge plutôt que couleur. «L'existence de détecteurs sensoriels primitifs de traits élémentaires est bien établie depuis les travaux de Hubel et Wiesel (1962) : ces traits sont, en l'occurrence, des valeurs d'attribut. » [Le Ny, 89]. Toutefois, la notion d'attribut est bien celle qui cognitivement, sert de support à celle de valeurs d'attributs. Même sous l'angle neurobiologique, il y a des configurations de cellules qui «s'occupent de » l'orientation spatiale, de l'angularité ou de contours, et qui fonctionnent en synergies [Le Ny, 89]. Les humains ont ainsi de nombreux systèmes de détections innés : vue (couleur, forme globale, solides simples (géons)), audition, etc. [Barsalou, 92]. Il n'y a pas que des attributs émanant de la perception, il y a aussi des attributs liés à la culture de l'individu qui ont pour caractéristique d'être inscrits en mémoire à long terme. Ainsi une orange se voit associée systématiquement l'attribut fruit. Mais si pour les occidentaux les bovidés sont étroitement liés à comestible, il n'en va pas de même pour d'autres cultures [Barsalou, 92]. Des attributs peuvent être binaires, multivalents, certains ont des statuts de dimensions et d'autres sont du type «partie-tout ».

2.3.3 Disponibilité et pertinence des attributs

Un des problèmes de la psychologie consiste à expliquer pourquoi le sujet utilise certains attributs et non d'autres, quels sont les attributs pertinents pour le sujet. Le Ny [Le Ny, 89] illustre ainsi le problème : «La couleur est un trait du concept d'animal mais pas de celui de haut fonctionnaire. »
Barsalou [Barsalou, 92] relie la pertinence des attributs à leur disponibilité. Cette disponibilité dépend d'un ensemble de contraintes. Parmi celles-ci, il y a d'abord celles du niveau biologique : les processus de la perception rendent automatiquement disponibles certains attributs (couleur, forme et taille). Ces attributs ont, selon lui, le plus haut niveau de disponibilité. Il y a aussi les attributs si fortement ancrés en mémoire à long terme que leur activation est automatique (orange et fruit ci-dessus). Pour les attributs moins tranchés leur pertinence est définie par le contexte. [Barsalou, 92] prend comme exemple «le journal ». Si l'on veut allumer un feu, un des attributs du journal retenu est qu'il est «inflammable » tandis que si l'on veut essuyer des carreaux, ce sera qu'il est «absorbant », mais dans tous les cas, dit Barsalou, l'attribut «noir et blanc » apparaît. Néanmoins, Barsalou remarque que distinguer pertinence et disponibilité est insuffisant et qu'il reste beaucoup à apprendre relativement à la pertinence.

2.3.4 Descripteurs : apprentissage PAC et psychologie

Nous pouvons noter que le fait que l'apprentissage PAC parte d'une représentation des exemples basés sur des descripteurs trouve une certaine pertinence en psychologie.

Dans le chapitre 1, nous avons présenté une version simplifiée de l'apprentissage PAC où les attributs sont binaires. Nous avons vu que les attributs pouvaient avoir plusieurs valeurs, ou être définis sur $\Re$, nous retrouvons ainsi des attributs qui sont ce que les psychologues appellent des dimensions. Dans l'apprentissage des langages tels que les langages réguliers, on ne travaille plus avec des attributs mais avec un alphabet. Au niveau des descripteurs, il n'y a donc pas de différence fondamentale entre le modèle PAC et la façon qu'a la psychologie d'envisager la catégorisation.

Cependant, le modèle PAC ne se préoccupe pas de la sémantique qui peut être mise sur les attributs. Ainsi des attributs relationnels tels que «partie de» peuvent être traduits en termes d'attributs binaires : l'attribut caractérisant le tout, l'étiquetage positif signifiant que l'objet est une partie de ce tout et l'étiquetage négatif l'inverse[38]. De la même manière, l'opposition que l'on retrouve en psychologie entre attributs perceptifs (par exemple grand/ petit) ou fonctionnels (sert à se laver : oui/non) ne concerne pas directement le modèle. Dans le cadre PAC, on considère que les attributs perceptifs et les attributs fonctionnels font partie de l'espace des attributs disponibles, le fait de savoir si le concept se décrit selon les uns ou les autres relève de la pertinence des attributs.

2.3.5 Divergence sur l'espace des descripteurs

Lorsque l'on présente le modèle PAC à des psychologues, le principal reproche qui lui est adressé est qu'il considère que ces attributs sont donnés, que l'espace des attributs est fixe alors que comme nous venons de le voir le problème de savoir pourquoi ce sont certains attributs et non d'autres qui sont pris en compte par le sujet est l'un des problèmes majeurs de la psychologie[39]. Il est nécessaire, ici, de reprendre la distinction faite plus haut entre attributs pertinents et attributs disponibles. Tandis que «noir et blanc », «inflammable », «absorbant » sont des attributs disponibles pour caractériser le journal, seul «inflammable » est un attribut pertinent quand il s'agit d'allumer un feu. Si on fait la distinction entre les deux, le modèle PAC correspond à l'approche psychologique. L'espace des attributs tel qu'il le définit correspond à l'ensemble des attributs disponibles pour le sujet au moment de l'apprentissage. Les attributs pertinents sont ceux qui sont utilisés dans la description du concept. Dans le cas du haut fonctionnaire, la couleur de son costume est disponible mais non pertinente. Le modèle PAC ne diffère pas de la psychologie sur ce point : une des difficultés majeures dans les démonstrations d'apprenabilité est de trouver les attributs pertinents.

Cependant le modèle PAC considère comme *fini* l'espace des attributs disponibles alors que la psychologie considère que cet espace est *potentiellement infini*. La preuve est du genre énumératif : tous les journaux sont à x millions de kilomètres du soleil, tous les journaux mesurent moins d'un décamètre carré, tous les journaux ne sont pas des

[38] Mais comme le signale [Barsalou, 92] le procédé est peu élégant.
[39] voir notamment les travaux de Thibaut [Thibaut, 95], [Thibaut, & Schyns, 95], [Thibaut, 97], qui étudie comment le sujet définit son espace d'attributs

chevaux, etc.[40] et effectivement cette liste d'attributs peut être infinie. L'un des problèmes des psychologues est de comprendre pourquoi, dans cette infinité, seuls quelques-uns sont pris en compte. Décider, comme le fait le PAC apprentissage, que l'espace des attributs est fini est donc inacceptable pour eux. Nous pouvons nous interroger sur le fait que cette liste est infinie lorsque le sujet s'adapte à son environnement. Barsalou définit la disponibilité des attributs par leur côté ou non actif. Cette activité des attributs est plus ou moins forte et, bien sûre, dépendante du contexte : l'attribut «absorbant » du journal est très peu actif lorsqu'il s'agit d'allumer un feu. Il est difficile d'envisager des situations de catégorisation où le sujet activera une infinité d'attributs. Il n'est donc pas trop fort de considérer que ce que le PAC définit comme l'espace des attributs est l'ensemble de tous les attributs que le sujet peut activer. Cet ensemble *théoriquement* infini, est fini dans la réalité, même s'il peut être très grand[41]. Il reste que, même si on considère que l'espace d'attributs du sujet est l'ensemble des attributs qui sont disponibles pour lui au moment de l'apprentissage, ce n'est pas pour autant que l'expérimentateur, lui, connaisse cet espace d'attributs. C'est une des difficultés de l'approche écologique[42].

Le fait que le modèle PAC considère comme donné un ensemble d'attributs ou de descripteurs n'est donc pas un handicap pour l'utiliser en psychologie. C'est plus évident encore lorsque l'on considère la catégorisation comme un apprentissage de C (la classe de concepts cible) dans H (la classe d'hypothèses de l'apprenant). Nous avons vu que, dans ce cas, la seule condition que pose le modèle PAC sur l'ensemble des hypothèses de l'apprenant (H) est que H contienne une hypothèse proche du concept, on ne définit pas alors l'ensemble des descripteurs utilisés par H.

Il reste cependant un problème, c'est celui de la taille du concept qui intervient dans le calcul du temps d'apprentissage. Dans les théories formelles, celui-ci est en général égal au nombre de valeurs d'attributs utilisées pour décrire le concept. Il est ainsi difficile de définir qu'elle est la plus courte représentation d'un concept naturel tel que celui de «chat ». Sur quels critères ou quels attributs peut-on la définir ? Il semble que l'on puisse passer outre avec les catégories artificielles, que dans celles-ci, les attributs sont clairement repérés. Le concept «triangle ou jaune » est clairement défini avec la valeur de deux attributs, on peut donc considérer qu'il est de longueur deux. En est-il vraiment ainsi ? Est-ce que la forme constitue à elle seule un attribut ou n'est-elle pas elle-même une composition d'attributs ? Si c'est le cas, est-ce que le concept «rectangle ou bleu » a la même longueur que «triangle ou jaune » car la décomposition de «rectangle » en attributs plus petits est peut-être plus longue que celle de «triangle ». Cette question se ramène à une autre que [Schyns et Rodet, 97] ont appelé «la quête des quanta de la cognition » : quelle est la plus petite unité d'information sur laquelle s'appuie la

[40] on retrouve l'objection faite à la similarité comme principe explicatif de la catégorisation (voir plus haut)

[41] [Littlestone, 88] a étudié l'apprenabilité dans le cas où le nombre d'attributs non pertinents est élevé.

[42] Nous sommes conscients qu'ainsi nous ne faisons que déplacer le problème de la sélection des attributs pertinents à la sélection des attributs disponibles, mais en opérant ce déplacement nous amenons peut-être à modifier le problème.

cognition humaine. Il paraît donc difficile de définir la plus courte représentation d'un concept. Si le problème existe, il ne faut pas, pour autant, se focaliser dessus. Si nous reprenons le concept de forme, considérer que les représentations du triangle et du rectangle ont la même longueur est acceptable dans notre civilisation car les valeurs «rectangle » et «triangle » sont si souvent activées dans notre culture que l'on peut les considérer comme des valeurs de base [Barsalou, 92].

2.4 Schéma[43] de représentation utilisé par l'individu pour ses catégories

Rappelons que, dans les théories formelles, un schéma de représentation est ce qui permet à l'oracle, pour la classe de concepts cible, ou à l'apprenant, pour sa classe d'hypothèses, de dire si un élément relève ou non du concept, de la catégorie. Ainsi, dans l'apprentissage PAC, la classe de concepts cible est définie par une composition d'attributs ou de descripteurs selon un schéma de représentation : les fonctions booléennes, les langages, etc. La classe d'hypothèses de l'apprenant peut, elle-aussi, être définie de cette manière ou ne pas être définie du tout. Dans l'apprentissage prédictif, on se contente de vérifier que l'apprenant est capable d'étiqueter correctement un certain nombre d'exemples. On suppose simplement que la classe d'hypothèses de l'apprenant contient une hypothèse proche du concept cible sans s'interroger sur le schéma de représentation de cette hypothèse.

Pour les psychologues, connaître l'espace d'hypothèses de l'apprenant et le schéma de représentations qui le sous-tend est une de leur quête principale. Cette question se confond d'ailleurs avec celle de concept cible lorsque, comme dans la troisième expérience décrite (celle de [Mazet, 93]), le psychologue cherche à connaître les catégories que le sujet forme. Derrière cette notion de schéma de représentation de la classe d'hypothèses de l'apprenant, on trouve ainsi plusieurs débats en psychologie. L'un d'entre eux consiste dans la façon dont la représentation de la catégorie est encodée en mémoire à long terme, nous le présentons rapidement ici car nous y reviendrons plus longuement dans le chapitre 3. Nous regroupons par ailleurs, dans un second temps, toute une série de controverses, de façon encore une fois quelque peu hérétique, dans ce que nous appelons l'opposition entre verbalisable/ non verbalisable.

2.4.1 Les catégories considérées au travers de la structure de leur représentation

Une des questions que se posent les psychologues concerne la structure de représentation des catégories en mémoire à long terme. Selon [Barsalou, 92], ils envisagent différents types de représentations pour modéliser la catégorisation humaine : les modèles à base d'exemplaires, à base de prototypes, en conditions nécessaires et suffisantes.

[43] Le terme de schéma doit se comprendre dans le sens des théories formelles et tel qu'il est défini au chapitre 1 et non dans le sens où la psychologie l'utilise habituellement.

Les modèles à base d'exemplaires. Les exemplaires spécifiques de la catégorie sont stockés en mémoire. Pour un exemple donné, la recherche de l'exemplaire correspondant se fait en parallèle. La première critique que l'on peut formuler relativement à ce modèle est qu'il réclame trop de mémoire même s'il y a toujours la possibilité d'oublier. La seconde critique est que les gens font des abstractions sur les catégories et ce modèle n'en tient pas compte[Barsalou, 92].

Les modèles à base de prototypes. « Un prototype est une représentation de la catégorie unique et centralisée. Selon la plupart des modèles à prototypes, le système cognitif abstrait les propriétés qui sont représentatives des exemples de la catégorie et les intègre dans un prototype de la catégorie » [Barsalou, 92]. Les prototypes peuvent être construits à partir de la moyenne de dimensions particulières, la fréquence la plus grande d'attributs et la moyenne ou la fréquence de relation interattributs des exemples. Le prototype se construit ainsi sur la base de similarités. La critique formulée à l'égard des modèles à base de prototypes est qu'ils échouent lorsqu'il s'agit de spécifier les contraintes sur le processus d'abstraction. En effet, le système cognitif peut considérer un nombre infini de propriétés (cf. ci-dessus concernant le nombre infini d'attributs) lors de l'abstraction et le modèle devrait normalement permettre d'expliquer pourquoi seules les propriétés pertinentes relativement au but sont prises en compte.

Les modèles classiques ou aristotéliciens. Ce sont les modèles à base de règles ou de conditions nécessaires et suffisantes dont on a vu la critique plus haut : les catégories sont rarement clairement définies et ces modèles ne prennent pas en compte les phénomènes de typicalité.

Selon Barsalou l'être humain utilise probablement les trois types de représentations. Bourne [Bourne 82] pour sa part considère qu'un concept a deux modes de représentation : d'une part, un noyau (a core) qui s'exprimerait en conditions nécessaires et suffisantes, ce serait une représentation en intension de la catégorie, et, d'autre part, un prototype qui permettrait de reconnaître des instances de la catégorie à partir de ses caractéristiques de surface (voir aussi [Osherson et Smith, 90]).

Les psychologues caractérisent donc les représentations des catégories qu'encode le sujet en fonction de leur structure. C'est pourquoi Pitt [Pitt, 97] propose d'étudier, dans le cadre de l'apprentissage PAC, la classe de concepts de type prototypique dont il définit ainsi la structure : «Le modèle prototypique habituel est basé sur une mesure de similarité que l'on peut assimiler pour la simplicité de l'analyse, à une simple fonction linéaire des attributs. L'attribut a peut prendre les valeurs $v_1,...v_k$, chacune d'elle ayant un poids $t_a(v_1),....t_a(v_k)$ indiquant à quel point la valeur v_i est typique pour l'attribut a. De plus chaque attribut a peut avoir un poids w_a reflétant son importance. La valeur de similarité que l'exemple x reçoit sera la somme des poids avec l'attribut a contribuant pour $w_a.t_a(v_i)$, si x a la valeur v_i sur l'attribut a. Plus haute est la valeur , plus x est prototypique. »

Nous reviendrons sur l'opposition entre une représentation par exemplaires et une représentation par prototype dans le chapitre 3 lorsque nous parlerons d'économie cognitive.

2.4.2 Verbalisable vs non verbalisable

Diverses recherches en psychologie concernent une autre façon d'envisager le schéma de représentation, non en tant que structure comme précédemment, mais en tant que langage. Nous les signalons car de notre point de vue elles correspondent à l'opposition qui peut exister, en apprentissage automatique, entre apprentissage symbolique et apprentissage connexionniste, et qu'elles sont susceptibles d'intéresser les informaticiens. Toutes ces recherches tournent autour d'une opposition entre phénomène de bas niveau / phénomène de haut niveau et invitent à penser que l'individu bénéficie de deux schémas de représentation.

2.4.2.1 Traitement automatique vs stratégique

[Barsalou, 92] fait une distinction entre le traitement automatique et le traitement stratégique : «Alors que certaines opérations cognitives apparaissent automatiquement sans avoir été souhaitées, d'autres apparaissent stratégiquement comme des tentatives délibérées d'atteindre un but». Parmi les traitements automatiques, il distingue ceux qui sont innés de ceux qui sont appris. Tandis qu'il considère que les traitements stratégiques sont tous appris. Le traitement automatique est du type stimulus-réponse. Le meilleur exemple du *traitement automatique inné* est le réflexe moteur (clignement de paupière). La perception bénéficie de ce type de traitement (les couleurs, la différenciation d'un objet de son contexte), de même que la mémoire (comptage inconscient de la fréquence, de la location, et du temps d'un événement). Le *traitement automatique appris* permet une plus grande adaptation à l'environnement. Si un stimulus et une réponse apparaissent souvent simultanément, il y a automatisation. La conduite de voiture, le laçage de chaussures sont d'abord appris et avec l'habitude deviennent automatiques (voir aussi l'effet Stroop [Stroop, 1935][44]). Le *traitement stratégique* est la poursuite de buts conscients. Selon [Barsalou, 92], l'état courant de l'environnement et du système cognitif peuvent déterminer quel est le comportement à déclencher. L'expérience subjective de contrôle et de liberté accompagnent souvent ces déclenchements sans en être la cause. Nous retrouvons ici l'idée que le sujet s'adapte à son environnement et que c'est l'environnement qui lui impose son comportement.

Pour [Lecoq, 94], il est difficile de différencier entre traitement automatique et traitement stratégique : «divers auteurs ont proposé 4 critères pour appréhender les processus de traitement automatique :
- les processus automatiques sont rapides

[44] L'expérience consiste à demander aux sujets de donner la couleur de mots écrits dans des couleurs différentes, *et non de les lire*. Lorsque le mot signifie une couleur et que cette couleur est différente de celle avec lequel le mot a été écrit (e.g. le mot «vert » écrit en rouge) le temps de réponse est plus long que lorsque le mot ne décrit pas une couleur (e.g. le mot «voiture » écrit en rouge), ce qui implique qu'il y a systématiquement lecture.

- on ne peut les éviter ni les arrêter
- ils ne sont pas disponibles à la conscience
- ils ne réduisent pas la capacité de réaliser d'autres tâches

Toutefois, là encore, ces critères ne sont pas absolus et ne sont pas toujours satisfaits. » Cependant, dans tous les cas de traitement automatique, les sujets ont peu de capacité à décrire ce qu'ils savent et à répondre à des questions concernant les règles qu'ils suivent ou les stratégies qu'ils appliquent.

2.4.2.2 Apprentissage implicite vs explicite

Lecoq considère que l'apprentissage implicite relève du traitement automatique et l'explicite du stratégique. Derrière la notion d'implicite, de manière caricaturale, il y a deux idées : l'absence de conscience de la part du sujet qu'il apprend, qu'il s'adapte, et l'absence de conscience de la structure de la situation. Dans sa revue de question, [Perruchet, 98] montre la difficulté qu'il y a à définir l'apprentissage implicite[45]. Il propose la définition suivante «Nous définirons l'apprentissage implicite comme un mode d'adaptation dans lequel le comportement d'un sujet apparaît sensible à la structure d'une situation sans que cette adaptation ne soit imputable à l'exploitation intentionnelle de la connaissance explicite de cette structure. » Par ailleurs, il montre qu'il n'est pas simple de délimiter clairement la frontière entre les deux apprentissages. Les deux types d'apprentissages peuvent interférer : selon [Reber, 76] lorsque les sujets reçoivent l'information explicite selon laquelle il s'agit de découvrir une règle, l'apprentissage est moins bien réussi. De plus, une recherche explicite pour des règles produit chez les sujets une forte tendance à induire ou inventer des règles qui ne sont pas une représentation correcte des stimuli, tendance que l'on n'observe pas chez les sujets qui ont reçu des instructions implicites.

2.4.2.3 Structure componentielle de la catégorie vs holistique

Concernant plus précisément les catégories, l'apprentissage PAC considère que la représentation d'un concept est une composition de descripteurs comme le faisait [Bruner, Goodnow and Austin, 56], cette vision n'est pas systématiquement partagée par tous les psychologues. On retrouve ainsi une opposition entre une vision componentielle du signifié [Clark, 73] et une vision holistique [Nelson, 74] (voir [Cordier, 94]) La distinction entre l'une et l'autre est la part accordée à l'analyse. Alors que, dans la vision componentielle, comme son nom l'indique, le concept est envisagé comme une composition de descripteurs, dans la vision holistique la discrimination (/ regroupement) entre deux objets se fait sur la base d'une différence (/similitude) globale non analytique. Si nous voulons nous faire une intuition de cette appréhension holistique, il suffit de se rappeler cette expérience : on rencontre un collègue et on observe que quelque chose a changé dans son visage, on pense qu'il a changé de lunettes alors qu'il s'est rasé la barbe. On a perçu globalement la différence mais on n'a pas été capable de mettre à jour le descripteur source de la différence. Cette opposition componentielle/holistique concerne surtout les catégories formées par le jeune enfant.

[45] Dans son article, il met aussi en évidence que le concept appris, l'hypothèse de l'apprenant, n'est pas obligatoirement le même que le concept cible mais plutôt une approximation de celui-ci.

Nous retrouvons ici l'opposition que mettait [Vigotsky, 85] entre pseudo-concept et préconcept (ou concept empirique). Dans le cas des pseudo-concepts l'enfant range dans une même catégorie les triangles et dans une autre les rectangles parce qu'ils se «ressemblent » (jugement global), dans le cas du préconcept parce qu'ils «ont le même nombre de côtés » (jugement analytique).

Nous pouvons nous interroger sur cette notion de perception globale non analytique. Nous avons vu ci-dessus qu'il existe des combinaisons de traits dont la séparation n'est psychologiquement pas possible. La conséquence est qu'un changement de l'un des traits semble modifier la perception globale plutôt qu'une de ses caractéristiques [Le Ny, 89]. Néanmoins la distinction entre les objets se fait bien sur la base de la distinction entre traits. La perception globale a été modifiée parce qu'un trait a changé. L'analyse n'est peut-être pas verbalisable mais elle a néanmoins eu lieu. Plutôt que d'opposer traitement analytique de haut niveau à traitement holistique de bas niveau, comme le font [Medin, Wattenmaker et Hampson, 87] ne faudrait-il pas opposer traitement verbalisable (haut niveau) à traitement non verbalisable (bas niveau). Dans les deux cas il y aurait bien une représentation en composants du concept (représentation componentielle) mais dans un cas le sujet pourrait la décrire dans l'autre non. Il est bien évident que, dans le cas où le sujet ne peut décrire les attributs qu'il utilise, cela pose un problème au psychologue pour définir la structure componentielle de la représentation qu'a le sujet, mais ce n'est pas pour autant qu'il peut affirmer qu'elle n'est pas componentielle.

2.4.2.4 Deux types de schémas de représentations utilisés par l'apprenant

Ainsi, en psychologie, on distingue entre structure componentielle et structure holistique, entre traitement automatique et traitement stratégique, entre apprentissage implicite et apprentissage explicite. Si nous nous plaçons dans le cadre de l'apprentissage PAC, nous pouvons considérer que ces différentes oppositions correspondent à celle qui peut exister entre deux schémas de représentation utilisés par l'apprenant[46]. L'un serait de haut niveau et la langue en serait une des caractéristiques et l'autre de bas niveau. Dans un cas, le sujet peut communiquer aisément les représentations de ses concepts. Dans l'autre, la représentation n'étant pas «communicable », l'expérimentateur doit, pour vérifier qu'elle est correcte, interroger le sujet sur des exemples. L'interprétation PAC de ces deux types d'expérimentation n'est pas la même. Dans le cas où le sujet retourne une hypothèse décrite en français («rond ou vert »), cela correspond à l'apprentissage de la classe de concepts dans la classe de concepts, (i.e. le sujet formule ses hypothèses dans la classe de concepts cible). Lorsque l'on demande au sujet de définir l'appartenance ou non des exemples cela correspond à l'apprentissage d'une classe de concepts cible dans une classe d'hypothèses que l'on suppose plus vaste que la classe de concepts cibles (apprentissage prédictif).
Ces deux schémas de représentations ne sont pas disjoints, de là, la difficulté pour les chercheurs d'exhiber des phénomènes d'apprentissage qui seraient purement implicites ou purement explicites. Nous pouvons alors envisager que des observations réalisées,

[46] Cette notion de deux schémas de représentation rejoint la notion de symbolisme et de subsymbolisme tel que l'envisage [Smolensky, 88]

telles que l'air de famille, viennent de ce que la catégorisation se réalise dans un schéma de représentation de bas niveau. C'est ce qu'envisagent entre autres [Medin, Wattenmaker et Hampson, 87], [Bideaud et Houdé, 93]. [Dubois, 93] propose : «Dans les termes récents des sciences cognitives ne peut-on considérer que la typicalité constitue une forme de représentation subsymbolique, 'présymbolique', 'proto symbolique' à partir de laquelle, des processus cognitifs, en particulier des opérations langagières, peuvent faire émerger diverses formes de représentations symboliques telles les prototypes conduisant aux symboles, signifiés et concepts ». Le subsymbolique, ici, renvoie à ce que nous appelons «bas-niveau ».

2.5 Les conditions de la catégorisation

Le modèle PAC lie de manière indissociable le *processus* de catégorisation, l'apprentissage inductif, et le *résultat* de ce processus, la représentation de la catégorie. Cela suppose implicitement que les conditions dans lesquelles s'est déroulé le processus influent sur le résultat. Cette question est au centre d'un débat en psychologie qui oppose les tenants d'une approche en laboratoire et ceux d'une approche écologique[47]. Ce sera le premier point abordé ici au travers des notions du modèle PAC d'oracle et de protocole de présentation des exemples. Le second point consiste à rechercher en psychologie ce qui correspond au calcul de l'erreur en théories formelles, c'est-à-dire la différence qui existe entre le concept cible et l'hypothèse de l'apprenant. Enfin, nous présentons ce qui correspond dans les expériences en psychologie à ce que l'on considère comme une information apportée à l'apprenant sur la classe de concepts cibles dans le modèle PAC.

Ces diverses questions nous amènent ainsi à nous intéresser davantage aux problèmes de *méthode* utilisée dans les expériences en psychologie. Nous avons vu dans le 2.1.2.4 qu'envisager le modèle PAC comme un protocole d'expérimentation n'a de sens qu'avec des catégories artificielles, cependant il est aussi possible de l'utiliser pour décrire les protocoles utilisés dans des expériences avec d'autres catégories. Ainsi à la fin de la section nous proposons une classification des expériences selon le rôle qu'y joue la classe de concepts de référence.

2.5.1 Oracle et protocole de présentation des exemples

Le PAC apprentissage considère que les exemples étiquetés du concept cible sont présentés à l'apprenant par un oracle. Cet oracle correspond en psychologie à l'environnement. Comme nous l'avons vu, nous pouvons distinguer deux cas qui seront repris ci-dessous. Dans le premier cas, la rencontre du sujet avec les exemples et le processus de catégorisation se sont déroulés avant l'expérience, l'expérience étant destinée alors à étudier les représentations résultant de cet apprentissage. Dans le

[47]Nous parlons d'approche écologique, non pas parce que les expériences se feraient en dehors d'un laboratoire, mais parce que la catégorisation, elle, s'est faite en dehors du laboratoire.

second, c'est l'expérimentateur qui joue le rôle d'oracle et qui présente les exemples étiquetés à l'apprenant. Dans ce cas, tout le processus d'apprentissage se déroule durant l'expérience.
Dans le modèle PAC les exemples sont présentés selon un protocole, exemples uniquement positifs ou exemples positifs et négatifs, présentation séquentielle ou en un bloc, etc. Les exemples choisis sont tirés selon une distribution de probabilités. Cette distribution de probabilité correspond à ce que les psychologues appellent familiarité ou fréquence d'instantiation [Barsalou, 85].
Ce qui distingue le modèle PAC de l'étude de la catégorisation en psychologie, c'est que, pour le premier, ces conditions sont définies, alors que pour la seconde cela dépend du type de recherches : écologiques ou en laboratoire. Ici aussi la révolution roschienne a joué un rôle en préconisant une approche davantage écologique.

2.5.1.1 L'approche écologique et l'approche en laboratoire

Selon [Barsalou, 92], les tenants de l'approche écologique considèrent que les psychologues doivent comprendre l'environnement physique s'ils veulent comprendre la cognition. Les hommes s'étant développés sur la terre, le cerveau est un résultat de l'adaptation à l'environnement physique. Il faut d'abord repérer les informations qui sont importantes dans l'environnement pour cerner de plus près les mécanismes dans le cerveau qui les traite. Autrement dit, l'étude de l'environnement permet de limiter la recherche sur la cognition à des domaines plus ciblés. Le reproche qui est adressée à la démarche écologique par les tenants de la recherche en laboratoire, est qu'elle est trop descriptive. Dans l'approche en laboratoire, le psychologue, après avoir identifié un phénomène présentant de l'intérêt, développera un protocole d'expérimentation très précis pour l'étudier. Pour cela, il est nécessaire qu'il contrôle toutes les variables qui peuvent concerner le phénomène. Le reproche que leur adressent les écologistes est que si le phénomène est bien identifié ce n'est pas pour autant que l'on peut le généraliser. On retrouve ces deux approches dans l'étude de la catégorisation. D'un côté, toutes les tâches d'identification de concept s'apparentent à la démarche en laboratoire. De l'autre, toutes les expérimentations qui visent à connaître les catégories formées par l'homme s'apparentent à la démarche écologique, car ces catégories ont été construites par lui au travers de son adaptation perpétuelle à l'environnement.

2.5.1.2 L'approche écologique

Dans l'approche écologique, l'expérimentation ne consiste pas à étudier chez le sujet sa capacité à catégoriser mais les catégories qu'il a construites dans sa vie quotidienne. Les exemples proposés lors de l'expérience ne servent pas, pour le sujet, à la formation d'une catégorie mais à lui rappeler une catégorie déjà formée. Le but est de définir la structure de cette catégorie. Par exemple, suite aux travaux de Rosch, il est généralement admis que la typicalité intervient dans cette structure. La mise en évidence de celle-ci se fait lors des expériences en étudiant [Cordier, 93] :
- la fréquence de citations : il sera demandé, pour une catégorie donnée, de produire une liste ordonnée de sous-catégories les plus représentatives.

- un jugement moyen : on donne une liste de sous-catégories et les sujets doivent estimer le degré de typicalité sur une échelle 1 à 7.
- un classement : classer des dessins renvoyant à des sous-catégories en fonction de leur représentativité par rapport à une catégorie donnée.
- le temps de réponse à des questions d'appartenance : plus la sous-catégorie est typique, plus le temps est court, etc.

Dans ces expériences, l'expérimentateur n'a pas d'information sur les exemples qui ont servi à la catégorisation et d'une manière plus générale sur les conditions dans lesquelles s'est effectuée cette catégorisation. L'expérience permet de faire un constat. C'est pourquoi il arrive qu'elle soit doublée par d'autres qui permettent, soit, de mieux cibler les causes des phénomènes constatés par l'approche écologique (expériences en laboratoire avec des catégories artificielles, voir ci-dessous) soit de mieux cerner les conditions de cette catégorisation. Dans ces dernières, on cherchera, par exemple, à connaître comment les mères de famille s'adressent à leurs enfants [White, 82], ou l'on interrogera le sujet sur sa 'familiarité' avec la catégorie, [Barsalou, 85].

2.5.1.3 L'approche en laboratoire

Selon [Barsalou, 92] cette approche est encore largement dominante. Ici, aussi nous pouvons faire la distinction entre deux époques : avant et après Rosch.

Nous rappelons qu'avant les travaux de Rosch, les expériences proches de l'apprentissage inductif consistent surtout en tâches d'identification de concept. Dans ce type de tâches, il est demandé au sujet de formuler une hypothèse qui correspond aux exemples qu'on lui a présentés. Comme nous l'avons vu, ce type de tâches suit un protocole d'expérimentation qui correspond assez étroitement aux conditions définies dans le PAC. Dans les articles suivants [Bruner,Goodnow and Austin, 56], [Neisser et Weene, 62], [Conant &Trabasso, 64], [Bourne, 70], [Dominowski, 73], [Richard, 75], [Bourne, 82], [Reber, 76], [Reber et Allen, 78]... on retrouve ainsi des constantes. Le *concept cible* est défini par sa structure (unidimensionnelle, conjonction, disjonction, grammaire...) sur des attributs très simples (figures géométriques, lettres,...). L'accent est parfois mis sur la nature, le schéma de représentation, de l'hypothèse de l'apprenant : apprentissage implicite ou explicite. [Reber, 76][Reber et Allen, 78]... Pour connaître l'hypothèse formée par le sujet, l'expérimentateur lui demande selon le cas, de décrire cette hypothèse («rond ou vert »), ou d'effectuer un certain nombre de prédictions sans erreur. Dans le premier cas, l'expérimentateur réclame que l'hypothèse soit exacte, dans le second, il est toléré en général une erreur sur vingt ou vingt-cinq. Dans tous ces articles, *la présentation des exemples* est séquentielle, jamais par blocs. Deux démarches sont utilisées : l'apprenant propose un exemple et demande son appartenance (sélection : oracle d'appartenance) ou l'exemple étiqueté lui est imposé (réception) [Dominowski, 73]. Dans tous les articles, on travaille avec des exemples positifs et négatifs. Lorsque la présentation des exemples est en réception, le choix des exemples par l'expérimentateur est fait selon une distribution que l'on peut assimiler à la *distribution uniforme ou équitable,* tous les exemples ont une même probabilité d'être présentés, ou selon une distribution telle que le nombre d'exemples négatifs soit égal au

nombre d'exemples positifs. Enfin dans ce type d'expériences, on vérifie que le sujet a bien repéré les différents attributs utilisés.

Depuis les travaux de Rosch ce type d'expérience n'a pas été abandonné (voir notamment [Pazzani, 91], [Roberts et Horowitz, 86] et les travaux d'identification de grammaire) mais Rosch plaide pour une démarche plus écologique[48]. L'approche par identification de concepts est considérée comme trop simpliste. Une des critiques qui lui est adressée est qu'elle ne tient pas compte de certains facteurs, entre autres la manière selon laquelle le sujet choisit l'espace d'attributs pertinents. Ces reproches vont entraîner un changement dans le protocole d'expérimentation notamment dans le choix de la classe de concepts cibles (voir 2.1.2.4).

La classe cible sera une classe naturelle (par exemple les pinnipèdes [Cordier, 93]), ou une classe artificielle. L'expérimentateur choisit des classes artificielles pour pouvoir mieux contrôler certaines variables et notamment les attributs, mais il n'indique plus au sujet les attributs à prendre en compte : «..Les catégories artificielles supplantent les catégories naturelles pour que la preuve soit faite que les résultats obtenus ne sont pas les reflets d'une organisation catégorielle préexistante. » [Cordier, 93]. Les exemples utilisés pour ces classes artificielles sont en général des dessins dont la décomposition en attributs n'est plus aussi évidente que précédemment (des visages stylisés [Tversky, 77], des bonhommes, [Cordier, 93], des animaux fictifs [Medin, Wattenmaker et Hampson, 87] voir figure 2.5 page suivante). Les tâches demandées aux sujets auront pour objet de vérifier si un phénomène donné apparaît : la plupart du temps il s'agit de vérifier certaines hypothèses concernant la typicalité ou l'air de famille.

Parmi ces tâches, certaines peuvent encore être apparentées à des tâches d'identification de concept notamment celles qui consistent à classer des échantillons d'exemples en deux catégories. Ainsi [Medin, Wattenmaker et Hampson, 87] cherchent à vérifier si l'échantillon présenté ci-dessous sera classé en 2 catégories selon une structure en «air de famille » ou selon une structure en conditions nécessaires et suffisantes.

Cette approche en laboratoire fait l'objet de critiques : [Dubois, 93] «Ces paradigmes expérimentaux ont plusieurs conséquences :
1)ils ne peuvent pas traiter la question pourtant critique de l'établissement des frontières naturelles entre catégories puisqu'elles sont imposées au sujet ou trop explicitement 'lisibles' dans la construction du matériel expérimental ; en d'autres termes l'appartenance est donnée et l'articulation entre appartenance et typicalité non affrontée empiriquement,
2) ils éliminent le jeu des effets corrélés des processus antagonistes de discrimination et de généralisation
3) ils se situent toujours en référence à des catégories définies a priori dans le cadre expérimental et de plus de catégorie 'vraies' et/ou socialement bien normées. Donc de manière générale, restent dans le cadre d'une conception de la connaissance exhaustive

[48] L'article de [Bourne, 82] cité ci-dessus marque assez bien la rupture. C'est un plaidoyer pour que les protocoles précédents soient maintenus. Il cherche à montrer que l'on peut mettre en évidence des phénomènes de typicalité dans les expérimentations habituelles d'identification de concepts.

et vraie qui nous semble contradictoire avec les principes fondateurs de la catégorie que Rosch avait posés. »

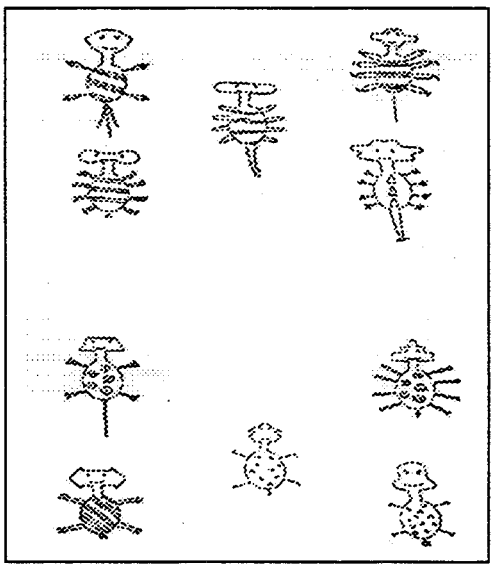


Figure 2.5 : stimuli utilisés par [Medin, Wattenmaker et Hampson, 87]

Derrière cette critique de D. Dubois, on retrouve le problème traité plus haut d'apprentissage supervisé. Ce que Dubois reproche c'est que l'on ne s'intéresse qu'à des catégories de type lexical ou artificiel alors que l'environnement en impose d'autres qu'il s'agit de découvrir (voir l'opposition entre l'expérience de [Rosch et Mervis, 75] et celle de [Mazet, 93] présentées au début). Cette critique permet de poser le problème de l'écart entre concept cible et hypothèse de l'apprenant.

2.5.2 L'écart entre la cible et l'hypothèse

Envisagées sous un certain angle, nous pouvons considérer que les recherches concernant le décalage entre la définition d'un concept de référence et la représentation qu'en a le sujet correspond au calcul d'ε dans l'apprentissage PAC. Ceci repose le problème de la distinction à faire entre la catégorie cible du modèle PAC et la catégorie de référence selon les expériences en psychologie (voir 2.1.2.4). Ce n'est que lorsque la catégorie de référence est une catégorie artificielle ou une norme (catégorie lexicalisée) que la question a du sens

Contrairement au modèle PAC, l'analyse du décalage, en psychologie, entre concept de référence et hypothèse de l'apprenant est de type qualitatif et non quantitatif. Ainsi en est-il de tous les travaux qui consistent à étudier l'écart qu'il y a entre une représentation du concept et l'information fournie jusque-là par le biais des exemples (recherches surtout préroschiennes). La question que se posent les chercheurs porte sur la consistance de l'hypothèse de l'apprenant relativement aux exemples vus.

> Ainsi dans [Bruner, Goodnow and Austin, 56] on présente au sujet un ensemble de cartes où sont dessinées des figures géométriques variant sur plusieurs dimensions, la forme, la couleur et la taille. Les sujets sont avisés qu'ils doivent trouver un concept connu de l'expérimentateur. Pour cela, ils proposent des exemples (des cartes) avec leur étiquette (exemple ou contre-exemple). L'expérimentateur confirme ou infirme l'étiquette. Ce type de protocole permet au chercheur d'analyser, au travers du choix des cartes que fait le sujet, si celui-ci a utilisé complètement ou non l'information qu'il a eue jusque là par le biais des exemples précédents. Supposons que le concept soit «triangle rouge », que le sujet ait déjà rencontré les exemples suivants : (grand triangle vert, non), (petit triangle rouge, oui), (grand triangle rouge, oui) il a alors assez d'information pour fournir l'hypothèse correcte sachant qu'il s'agit d'une conjonction à deux éléments. Si le sujet repropose une carte, l'expérimentateur peut légitimement supposer que son hypothèse est incorrecte.

Un autre type d'analyse de l'écart consiste dans les études sur la surgénéralisation ou la surspécialisation avec notamment l'étude de la «conjunction fallacy » [Tversky, 77], [Hampton,88,a] ou «la disjunction fallacy » [Hampton,88,b]. La «conjunction fallacy » est l'erreur qui consiste à croire que la probabilité d'un élément possédant conjointement deux propriétés peut être plus grande que celle d'un élément n'en possédant qu'une.

> [Tversky et Kahneman, 83] soumettent le problème suivant à 89 étudiants :
> «Linda est âgée de 31 ans. Elle est extravertie. Elle a suivi des études de philosophie. Alors qu'elle était étudiante elle se sentait profondément concernée par les questions de discrimination et de justice sociale. Elle participait aux manifestations antinucléaires. »
> Les étudiants sont ensuite invités à ranger par ordre décroissant de probabilité, les assertions suivantes :
> a- Linda est enseignante dans une école élémentaire
> b- Linda travaille dans une librairie et prend des cours de yoga
> c- Linda est militante d'un mouvement féministe
> d- Linda est une assistante sociale
> e- Linda est un membre de la «League of Women Voters »
> f- Linda est caissière dans une banque
> g- Linda vend des assurances
> h- Linda est caissière dans une banque et militante d'un mouvement féministe.

> 90% des étudiants ont alors classé l'assertion h comme plus probable que f, c'est ce qui est appelé la «conjunction fallacy », car il est évident que l'on ne peut avoir h sans f mais par contre que l'on peut avoir f sans h. h ne peut donc être plus probable que f.

La typicalité est parfois utilisée pour expliquer des erreurs de catégorisation, comme dans le cas précédent. Ainsi, au cours du développement de l'enfant, les sous-catégories jugées non typiques sont d'abord classées par lui comme n'appartenant pas à la catégorie. D'un autre côté, les sous-catégories typiques peuvent être une source possible de sur-extensions [Cordier, 93].

Il convient, cependant, d'éviter de considérer la typicalité simplement comme une erreur : «Les phénomènes que Rosch a empiriquement dégagés et qui se trouvent maintenant incontestés, ne s'avéreront productifs que si on leur assigne un autre statut que celui de déviation, de décalage, voire 'd'erreur' par rapport à une conception normative d'un savoir unifié sinon vrai » [Dubois, 93] Nous retrouvons la même idée dans le côté «approximatif » de l'apprentissage PAC. Le modèle PAC ne réclame pas que l'apprentissage soit exact, l'hypothèse de l'apprenant peut être approximative. Cette approximation peut être assimilée à une adaptation suffisante mais non optimale. Tant que la différence entre concept et hypothèse est inférieure à un certain seuil ε, l'hypothèse du sujet est suffisamment efficace pour que l'on ne parle pas d'erreur. Si la différence dépasse ce seuil, alors il n'y aura pas adaptation, et il s'agira vraiment d'erreur. A l'opposé, dans le modèle PAC l'approximation est basée sur une erreur d'appartenance, tandis que dans la typicalité l'appartenance ne pose pas problème, la typicalité exprimant plutôt un degré d'appartenance. De la même manière que dans la ferme d'Orwell où tous les animaux sont égaux, mais où certains sont «plus égaux» que d'autres, le marteau et le tournevis sont tous les deux des outils, mais le marteau «est plus un outil » que le tournevis.

2.5.3 L'information apportée à l'apprenant, la connaissance a priori

Dans les expériences préroschiennes, on peut distinguer entre les tâches d'identification de concepts selon l'information apportée à l'apprenant. [Bourne, 70] :
- apprentissage complet : trouver la règle et les attributs pertinents (complete learning),
- apprentissage de règles : trouver seulement, la règle les attributs pertinents étant donnés (rule learning)
- ou la réciproque (attribute identification).

La plupart des articles ([Bruner, Goodnow and Austin, 56], [Neisser et Weene, 62]...) se placent dans le cadre de l'identification d'attributs : *l'expérimentateur signale aux sujets la règle* et leur demande de trouver les valeurs pertinentes des attributs pertinents. Dans le cas d'un concept disjonctif, les sujets seront avertis que le concept est de la forme «x ou y ». A l'inverse, [Bourne, 70] travaille en identification de règle, *les attributs pertinents étant donnés*. Seul l'article de [Richard, 75] s'intéresse à l'apprentissage complet, c'est-à-dire que les sujets ne savent pas, a priori, si le concept est unidimensionnel, une conjonction de deux littéraux ou une disjonction de deux

littéraux. Il faut signaler que dans [Bourne, 82] en plus de donner aux sujets l'information sur le type de concept, il s'agit d'une disjonction, on leur propose différents algorithmes permettant de la trouver.

Ces différents types d'information apportés à l'apprenant appellent plusieurs remarques dans le cadre d'un rapprochement avec le modèle PAC. La première est qu'elles réduisent singulièrement l'espace d'hypothèses de l'apprenant, lui facilitant ainsi le travail. Par exemple le plus grand ensemble d'éléments décrits par des attributs binaires proposé par ces articles, est de 64 éléments (un élément étant décrit par 6 attributs binaires, il y a donc 2^6 éléments différents possibles). Si l'on considère qu'à chaque partie de cet ensemble correspond une hypothèse, comme il y a 2^{64} parties, le nombre possible d'hypothèses est donc de 2^{64}. En informant le sujet sur le fait qu'il s'agit d'une disjonction à deux valeurs d'attributs, on descend à 60 hypothèses possibles (12 x 10 / 2). Cette information réduit donc singulièrement la complexité de la tâche. La seconde est l'interprétation que l'on peut faire dans le modèle PAC de cet apport d'information car il pose le problème du type d'apprentissage effectué : s'agit-il de l'apprentissage de C dans C ou de C dans H. Le fait que les sujets soient informés sur la classe de concepts invite à répondre qu'il s'agit de l'apprentissage de C dans C. A l'opposé, lorsque l'on voit que pour connaître l'hypothèse des sujets, on leur demande de donner l'étiquette des exemples, on pense alors plutôt à l'apprentissage de C dans H (modèle prédictif). Ce problème rejoint celui des deux schémas de représentations signalés plus haut (explicite/implicite). Une manière de le résoudre est de considérer que lorsque l'on ne demande pas au sujet de formuler explicitement son hypothèse, on doit considérer qu'il s'agit d'un apprentissage de C dans H. Cette deuxième remarque montre l'ambiguïté de la formulation «apprentissage de C dans H» utilisée par le modèle PAC. Cette formulation permet en même temps de définir le schéma de représentations de l'espace d'hypothèses de l'apprenant et l'information qui lui est donnée sur l'espace de solutions possibles. Elle ne pose pas de problème pour les théories formelles où les deux choses peuvent être confondues. Ce n'est pas le cas en apprentissage naturel où, si nous parlons d'apprentissage de C dans H, nous devons cependant préciser les informations données à l'apprenant sur C car toute information apportée à l'apprenant sur la classe de concept cible réduit son espace d'hypothèses H.

La plupart des expériences préroschiennes attiraient l'attention des sujets sur les attributs disponibles. C'est d'ailleurs un des reproches que leur fait l'approche écologique. Aussi, dans les expérimentations postroschiennes avec des catégories artificielles, les attributs ne sont plus signalés à l'apprenant et encore moins les attributs pertinents. Les expériences avec les catégories artificielles ne sont toutefois pas les plus nombreuses de cette période et, parmi les autres, l'information apportée à l'apprenant n'a plus le même rôle. Cette information supplémentaire permet de définir un contexte. Ainsi nous ne donnerons pas le même exemple de poisson le plus typique selon que l'on nous dit juste avant le test : «le plongeur photographiait les poissons de toutes les couleurs» ou «le poisson attaquait le plongeur».

Parmi les expériences postroschiennes, il faut signaler celle de [Pazzani, 91] qui s'intéresse à l'impact que peut avoir une information *sémantique* supplémentaire sur le

choix du concept. Plus précisément, il étudie dans quelle mesure une information donnée en plus de celle contenue dans l'échantillon peut aider ou au contraire parasiter l'apprentissage.

> La tâche est une tâche d'identification de concepts. Les exemples sont des photos représentant une personne avec un ballon. Ils correspondent à un stimulus à 4 attributs : la couleur du ballon (jaune ou violet), la taille du ballon (grand ou petit), l'âge de la personne (adulte ou enfant) et enfin l'action sur le ballon (l'étirer ou le plonger dans l'eau)
>
> On présente séquentiellement les photos aux sujets. Après chaque exemple, les sujets doivent dire si l'exemple relève du concept ou non. On considère que le sujet a trouvé le concept lorsqu'il fait 6 prédictions successives sans erreur. La difficulté d'apprentissage d'une hypothèse est alors mesurée par le nombre d'exemples présentés avant que l'apprentissage ne réussisse. Tous les exemples du concept sont présentés (échantillon complet) mais la distribution est telle que le nombre d'exemples positifs est égal au nombre de négatifs.
>
> Il y a 4 groupes de sujets. 2 groupes bénéficient d'une connaissance a priori, les 2 autres non. L'information donnée a priori est qu'il faut gonfler le ballon. Au sein de chacun de ces 2 groupes, le concept cible est soit une conjonction (le ballon est petit et jaune) soit une disjonction (adulte ou étirer le ballon). Il faut noter que la connaissance a priori est pertinente avec la disjonction et non pertinente avec la conjonction.
>
> Les résultats montrent que les sujets apprennent mieux la conjonction que la disjonction s'ils n'ont pas d'information a priori. C'est-à-dire que l'apprentissage de la conjonction nécessite moins d'exemples. A l'opposé, ils apprennent mieux la disjonction que la conjonction lorsque l'information a priori est pertinente avec la disjonction et non pertinente avec la conjonction. L'explication que donne Pazzani de ces résultats est qu'il existe moins d'hypothèses (voir ci-dessus) qui soient à la fois cohérentes avec les exemples et avec la connaissance a priori et que l'apprentissage en est facilité. La réduction de l'espace d'hypothèse est due notamment au fait que la connaissance a priori intervient dans la sélection des attributs pertinents.

Cette expérimentation nous intéresse à plus d'un titre. D'abord, elle permet d'illustrer l'importance de la connaissance délivrée a priori. Deuxièmement, Pazzani ne demande pas aux sujets de retourner explicitement leur hypothèse mais leur demande d'étiqueter des exemples, la raison en est que : « .. réclamer une règle explicite ou informer le sujet que le concept à apprendre peut être représenté par une règle logique d'une forme donnée augmente l'attention consciente sur le processus d'apprentissage et peut gêner la détection inconsciente de covariations ». Nous retrouvons ici le problème des schémas de représentations possibles. Enfin les stimuli ne sont pas des dessins géométriques où les attributs sont facilement identifiables mais des photos. Concernant ce dernier point,

nous pouvons nous demander dans quelle mesure les attributs pertinents présentés au début sont réellement ceux pris en compte par le sujet.

2.5.4 Classification des expériences

Il est possible de proposer une classification des expériences selon le rôle qu'y joue la classe de concepts de référence par opposition au rôle que joue le concept cible dans le modèle PAC. Nous obtenons ainsi un arbre (schéma 2.5 page suivante) qui croît selon deux dimensions : de gauche à droite, nous avons la dimension écologique, du plus écologique au moins écologique, de bas en haut la connaissance du contexte du plus défini au moins défini.
Nous trouvons ainsi en haut et à gauche, les expériences où le concept référent ne correspond pas au concept cible du modèle PAC, mais à une définition des objets du monde que l'on va demander au sujet de classer. C'est dans ces expérimentations que l'on peut classer les travaux de [Dubois, 93], ceux de [Mazet, 93] sur les photographies de sections de route. L'objectif est de comprendre comment l'environnement impose à l'apprenant les catégories qu'il doit former.
Puis viennent les catégories ad hoc de [Barsalou, 83] [49], où l'on définit le concept référent au travers d'un contexte «votre maison brûle », «vous faites régime ». L'expérimentateur s'intéresse surtout au choix des attributs pertinents par le sujet. En cas d'incendie, ce sera {irremplaçable, transportable,...}, en cas de conduite {dangerosité, visibilité...}.
Ces deux types de recherches insistent sur le côté écologique de la catégorisation. Soit la catégorie de référence n'est pas une catégorie cible mais un monde d'objets, soit-elle est une catégorie cible mais elle n'est pas une norme, elle n'a pas une structure clairement définie et acceptée.

Dans les autres cas, la catégorie de référence est bien une catégorie cible. Celle-ci sera plus ou moins écologique selon qu'elle aura été formée en dehors du laboratoire au travers des interactions avec l'environnement (catégorie naturelle, catégorie artefact, sociale...), ou dans le laboratoire (artificielle). Parmi les catégories artificielles on insistera plus ou moins sur la structure (dessins de bonhomme, définies par une grammaire, une disjonction).
La profondeur indique le degré de connaissance du chercheur du contexte dans lequel s'est opérée la catégorisation. Plus on est en haut et à gauche, moins le psychologue a d'informations et plus le nombre d'inconnues est grand : quelle est la catégorie formée, quel est le schéma de représentation de l'apprenant et notamment est-il implicite, explicite, est-il de la forme prototypale, par exemplaires, quel est l'espace d'attributs disponibles, quels sont les attributs pertinents. En bas et à l'extrême droite, la cible est précisément définie, les attributs disponibles aussi, toute la recherche porte sur le fait de savoir si l'apprentissage est implicite ou explicite [Reber, 76][Perruchet, 97], ou si le sujet est capable d'identifier des concepts d'une structure donnée [Bruner, Goodnow and Austin, 56]... On peut ainsi caricaturer la révolution roschienne par le fait qu'elle a fait croître l'arbre vers le haut et vers la gauche.

---

[49] Rappelons néanmoins notre réticence à parler de «catégorie » dans le cas des catégories ad hoc

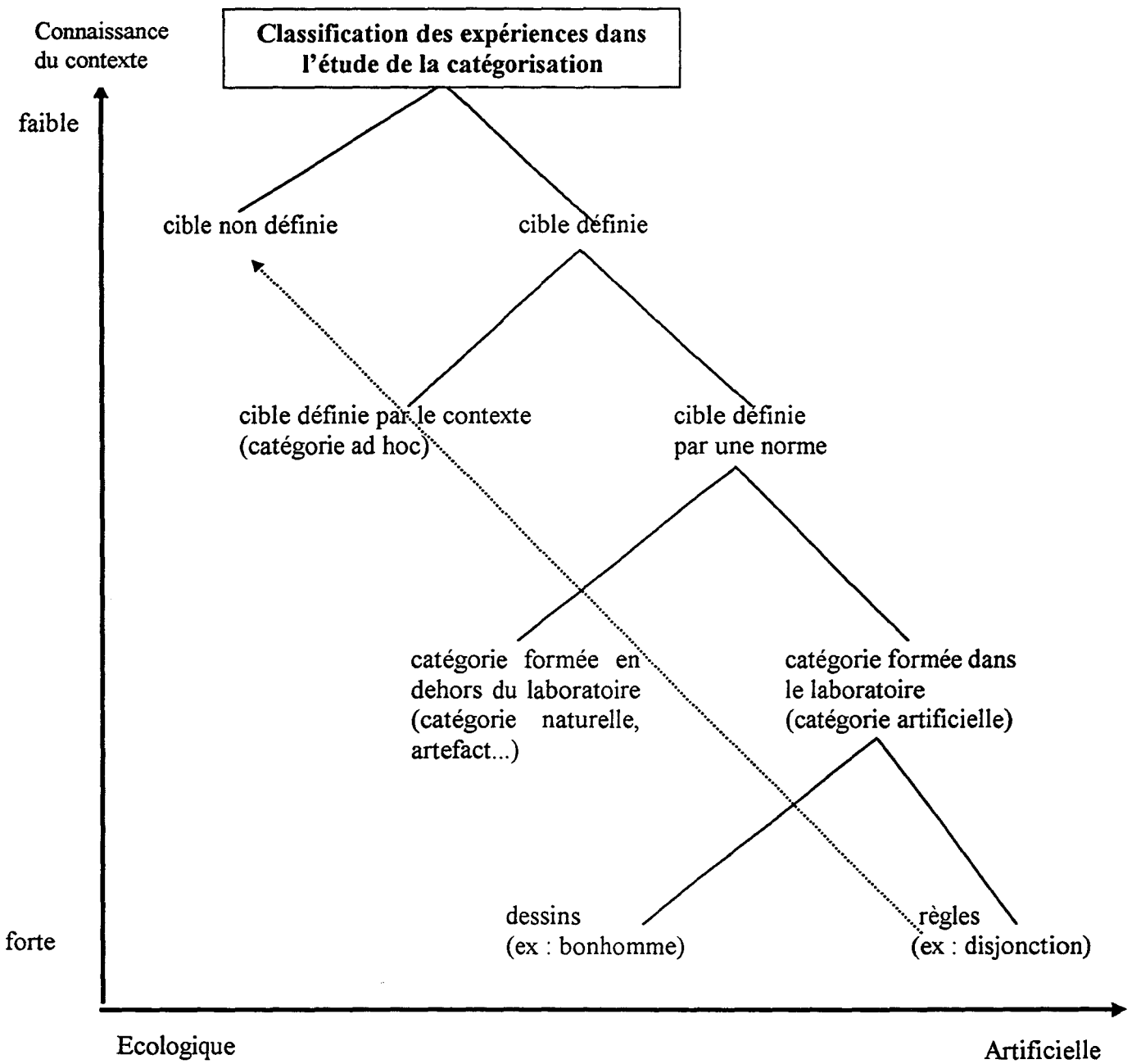


Schéma 2.5 : Une caractérisation par le concept de référence de la recherche en psychologie sur la catégorisation. En pointillés, la direction qu'a imprimée à ces recherches la révolution roschienne

Selon la position du chercheur dans l'arbre, il lui est adressé des reproches provenant de ses collègues se situant ailleurs. En haut à gauche, le contexte est si difficile à cerner que l'on reproche aux résultats d'être trop descriptifs, en bas à droite le contexte est tellement défini, que l'on reproche aux résultats de ne pas pouvoir être généralisés. Ces reproches ont la même origine : la difficulté à pouvoir établir des liens entre catégorisation et conditions de la catégorisation

2.6 Analyse quantitative de la catégorisation

Les notions de mesures sont inhérentes au modèle PAC. Ce n'est pas par caprice de mathématicien qu'il en est ainsi mais par souci de réalisme. Peut-on dire qu'un concept est apprenable s'il faut une éternité pour l'apprendre, ou s'il faut une infinité d'exemples ? Peut-on appréhender des exemples dont le nombre d'attributs est infini ? Dans cette section, nous passons en revue les différentes mesures que l'on trouve dans le PAC et établissons un parallèle avec les études en psychologie[50]. Dans ce parallèle nous ferons encore une fois la distinction entre la période préroschienne et postroschienne car il est plus facile d'identifier les mesures dans la première que dans la seconde.

Nous rappelons les mesures qu'utilise le modèle PAC : le nombre d'attributs, la taille de la plus petite représentation du concept, la distribution de probabilité des exemples, la taille de l'échantillon, le poids de l'erreur de l'hypothèse, le paramètre d'erreur ε qui sert à borner le poids de l'erreur admissible, le paramètre δ de confiance dans le succès de l'apprentissage.

2.6.1 Le nombre d'attributs et la taille du concept

Parmi toutes ces mesures les seules qui apparaissent en psychologie sont les quatre premières et souvent de manière implicite. Ainsi dans les travaux d'identification de concepts, le *nombre moyen d'attributs* varie souvent entre 3 et 6. Dans ces articles, les exemples sont des cartes sur lesquelles sont dessinées des figures géométriques variant sur les attributs suivants : formes, couleur, taille, nombre, figure ouverte/fermée, gauche/droite. Les attributs peuvent être binaires [Bruner, Goodnow and Austin, 56], ... ou ternaires [Bourne, 70 et 82]. La plupart du temps, les *cibles* sont majoritairement d'une longueur de deux termes. Selon [Neisser et Weene, 62], il semble que les sujets ne soient pas capables de traiter plus de deux variables. D'un point de vue théorique nous pouvons constater que l'espace des hypothèses est ainsi réduit et nous avons vu qu'il l'est davantage encore avec l'information donnée aux sujets. Le nombre de

[50]Nous rappelons que nous avons utilisé pour illustrer le PAC l'apprentissage des langages à attributs par souci de lisibilité. Dans ce type de langage le nombre d'attributs permet de fixer la longueur des exemples. Si nous avions choisi des langages tels que les langages réguliers (ou d'autres) nous aurions été amenés à considérer d'autres mesures telles que la longueur du plus grand exemple. Cela aurait simplement compliqué la lecture sans apporter d'information supplémentaire.

variables totales, et donc le cardinal de l'ensemble des exemples, varie lui aussi sans que les auteurs ne tiennent compte de cette variation lors de la comparaison de leurs résultats. On passe ainsi de 8 (3 variables binaires : 2^3)[Richard, 75] à 81 [Bourne, 70 et 82] (4 variables ternaires : 3^4). Que la complexité variable des situations n'ait pas été prise en compte par les auteurs peut s'expliquer par le fait que leurs résultats concordent dans l'ensemble, mais aussi par le fait que cette variation de la complexité est relativement faible, la taille de l'ensemble des exemples ne dépassant jamais 100 éléments. Il serait ainsi intéressant de vérifier si les résultats seraient identiques si l'on passait à 1000 éléments.

L'article de Pazzani [Pazzani, 91] permet de faire la liaison avec l'époque postroschienne. Comme nous venons de le voir, dans son article les exemples varient théoriquement selon 4 paramètres binaires (16 exemples possibles). Cet article est écrit après la révolution roschienne, les stimuli choisis sont plus «écologiques» et consistent dans des photos. Rien ne garantit donc que d'autres variables, non pensées par l'expérimentateur, n'aient pu être repérées par le sujet. Après la révolution roschienne ce n'est qu'avec les expérimentations sur la base de catégories artificielles que l'on peut espérer retrouver une mesure sur le nombre d'attributs : [Cordier, 93] «En règle générale, les catégories sont caractérisées par n dimensions, 3 le plus souvent (x, y, z) sur lesquelles l'expérimentateur sélectionne 5 valeurs (1, 2, 3, 4, 5). Toutes les caractéristiques autres que les n dimensions pertinentes demeurent invariables dans tous les dessins». Avec les dessins de bonhommes ronds (voir figure 2.5 ci-dessous), l'expérience fait varier la longueur de la tête, la largeur de la tête et l'écart bouche-yeux [Cordier, 83]. Dans ce type d'expérimentation on vérifie comment le sujet classe, et bien souvent le classement fait intervenir toutes les variables. Cela revient à dire que les variables pertinentes sont aussi les variables disponibles. On peut considérer que les concepts sont de longueur trois. En essayant de reproduire le phénomène d'air de famille en laboratoire [Medin, Wattenmaker et Hampson, 87], sont amenés à faire varier ce nombre d'attributs mais ils constatent que cette variation n'est pas significative[51].

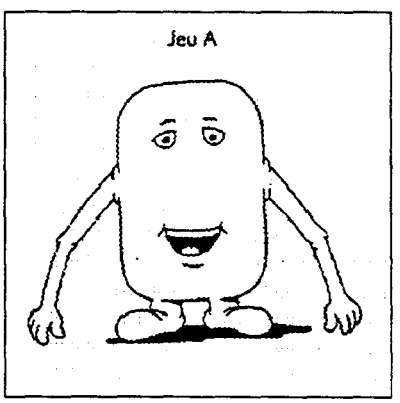


Figure 2.6 : stimulus en forme de bonhomme utilisé par [Cordier, 83] (voir [Cordier, 93])

[51] On peut se demander si cette variation n'a pas eu d'effets significatifs simplement par ce qu'elle n'était pas non plus vraiment significative (passage de 4 à 6 attributs).

2.6.2 La distribution de probabilités

Dans les tâches d'identification de concepts, la distribution de probabilité des exemples n'est pas réellement prise en compte, en tant que telle. Dans tous ces articles les exemples sont présentés séquentiellement. Les auteurs n'indiquent pas pour autant comment sont tirés les exemples. Lorsque c'est le sujet qui choisit l'exemple, c'est son choix qui définit, en quelque sorte, la distribution. Par contre, lorsque c'est l'expérimentateur qui impose l'exemple au sujet, l'information que l'on a sur la distribution est mince. La plupart du temps il est mentionné que le nombre d'exemples positifs est égal au nombre d'exemples négatifs[52]. Il est aussi indiqué que tous les exemples sont présentés, ce que les informaticiens appellent une présentation par échantillon complet. Comme certaines expérimentations se font jusqu'à ce que le sujet ait trouvé le concept, celui-ci peut être amené à voir l'échantillon plusieurs fois. Dans certaines expérimentations [Bruner, Goodnow and Austin, 56], le sujet a constamment l'échantillon complet sous les yeux mais sans que les exemples soient étiquetés. Il n'y a pas, dans ces articles, d'étude de l'impact de la distribution sur l'apprentissage. Le seul type de recherche qui s'en approche est celui qui analyse les rôles respectifs des exemples positifs et négatifs dans la réussite de l'identification. La plupart des articles constatent ainsi la difficulté des sujets à exploiter les exemples négatifs.

Par contre l'approche postroschienne s'intéresse davantage aux distributions de probabilités. Nous pouvons les assimiler à ce que les psychologues appellent la «familiarité » ou la «fréquence d'instantiation »[Ashcraft, 78], [Malt et Smith, 1982], [Hampton et Gardiner, 1983], [Barsalou, 85]. [Barsalou, 85] fait la distinction suivante entre fréquence d'instantiation et familiarité : la familiarité consiste dans le nombre de contacts entre le sujet et un exemple sans qu'il soit fait référence lors du contact à la catégorie de cet exemple (un exemple non étiqueté), la fréquence d'instantiation correspond au nombre de rapports du sujet avec un exemple étant donné sa catégorie (exemple étiqueté). La familiarité correspond ainsi au nombre de fois où le sujet a rencontré un chat et la fréquence d'instantiation au nombre de fois où il a rencontré un chat *étiqueté en tant que tel* «Regarde, le chat est encore en train de griffer les meubles ! »[53]. L'objectif de ces recherches est de déterminer le poids de la familiarité et/ou de la fréquence d'instantiation dans la typicalité. Barsalou définit ainsi trois déterminants de la typicalité, dont la fréquence d'instantiation et va étudier le poids relatif de chacun dans le jugement de typicalité[54]. Pour cela, il va distinguer entre deux types de concepts cibles : les catégories taxonomiques et les catégories ad hoc.

---

[52]Le seul article qui fasse une allusion à un groupe de sujets ayant travaillé avec des exemples uniquement positifs est celui de [Neisser et Weene, 62] mais les auteurs ont globalisé les résultats du groupe avec ceux des autres groupes ayant constaté qu'il n'y avait pas de différence majeure.

[53] On note qu'ainsi définie la fréquence d'instantiation correspond exactement à la distribution sur les exemples positifs et que la familiarité correspond partiellement à un oracle statistique tel que STATINFO défini par [Denis, 98].

[54] Les deux autres déterminants sont la «tendance centrale » et l' «idéal » mais il serait trop long de les définir, aussi nous ne décrivons ici que la partie de l'expérience se référant aux fréquences d'instanciation.

L'expérience 1 D'abord, on demande à un premier groupe de sujets de donner des items de 9 catégories ad hoc (cadeau d'anniversaire, équipement de camping, transport entre San Francisco et New York, ...) et de 9 catégories taxonomiques (véhicules, vêtements, oiseaux, armes...). Puis il présente à un deuxième groupe ces items rangés par catégorie et lui demande d'indiquer sur une échelle graduée de 1 à 9 la typicalité de chacun des items de chaque catégorie, et sa fréquence d'instantiation sur une échelle allant aussi de 1 à 9.

Les résultats montrent que la fréquence d'instantiation est très fortement corrélée avec la typicalité dans le cas des catégories ad hoc et beaucoup moins dans le cas des catégories taxonomiques. Par contre, lorsque l'on compare le nombre de fois où un item est donné pour une catégorie par le premier groupe avec la fréquence d'instantiation définie par le deuxième, on trouve une très forte corrélation et ceci pour les deux catégories. Selon Barsalou ces derniers résultats correspondent aux recherches concernant la mémoire selon lesquelles, plus souvent deux objets sont rencontrés simultanément, plus souvent l'un amènera l'autre dans des tâches de production.

Barsalou veut distinguer ensuite entre familiarité et fréquence d'instantiation. Pour cela tous les items produits par le premier groupe sont mélangés (non classés par catégorie) et on demande à 10 autres sujets de donner toujours sur une échelle de 1 à 9 leur degré de familiarité avec chacun des items. La corrélation entre familiarité et fréquence d'instantiation est plus élevée pour les catégories taxonomiques que pour les catégories ad hoc, ce qui indique que ces deux mesures ne sont pas identiques puisqu'il y a une différence selon le type de catégorie. De plus, la corrélation entre familiarité et typicalité est plus faible que celle entre fréquence d'instantiation et typicalité.

Commentaires
Si la notion de fréquence d'instantiation recouvre bien celle de distributions de probabilités sur les exemples positifs, il faut noter qu'il existe une différence notable entre les deux. Dans le cas des distributions de probabilités la mesure est objective tandis que dans le cas de la fréquence d'instantiation elle est subjective, c'est le sujet qui la donne.

2.6.3 La taille de l'échantillon

Nous y avons déjà fait mention en parlant des probabilités. Comme on présente généralement à l'apprenant l'échantillon complet et même plusieurs fois, nous pouvons considérer que la taille de l'échantillon correspond à un multiple de la taille de l'espace des exemples. Dans les expérimentations d'identification de concepts, cette taille sert assez souvent de mesure de difficulté de la tâche (voir l'expérience de Pazzani décrite plus haut). Plus l'apprenant demande d'exemples avant d'identifier le concept, plus l'apprentissage est considéré difficile.

2.6.4 La taille de l'erreur

Nous avons déjà décrit le statut de l'erreur dans les expérimentations et notamment l'hésitation qu'ont certaines recherches postroschiennes à considérer l'écart entre hypothèse de l'apprenant et concept cible comme une erreur. Dans les recherches préroschiennes, il n'y a pas non plus vraiment d'interrogation sur ce sujet. En général ces recherches considèrent l'apprentissage exact. Même lorsque l'on demande au sujet de classer avec au plus une erreur sur vingt, il n'y a pas d'interrogation sur le statut de cette erreur et sur le fait qu'il s'agit alors d'un apprentissage approximatif.

Ce parallèle entre l'aspect quantitatif du modèle PAC et la psychologie se révèle assez pauvre dès que l'on aborde les recherches postroschiennes, ce qui paraît normal. Supposons que nous souhaitions étudier chez l'individu la représentation d'une catégorie naturelle, par exemple celle de «chat ». Premier écueil : quelle est la plus petite représentation du concept cible de chat ? Peut-on se contenter de la plus petite définition que l'on trouvera dans les dictionnaires ? Deuxième écueil : quels sont les attributs que le sujet utilise ? On retrouve la difficulté à définir l'espace des attributs disponibles. Le fait que le chat soit un mammifère n'est probablement pas un trait pertinent de la représentation que se fait le sujet pour reconnaître un chat. Il est donc très malaisé de procéder à une analyse quantitative lorsque les représentations des catégories visées proviennent d'un apprentissage écologique.

2 Conclusion

Ce chapitre avait deux ambitions, d'une part, montrer que le modèle PAC est pertinent du point de vue des sciences cognitives, d'autre part, présenter les recherches sur la catégorisation en psychologie à des lecteurs non familiers de cette discipline. En raison de ces deux objectifs, nous n'avons pas suivi une démarche propre à la psychologie, mais sommes parti des concepts constitutifs du modèle PAC pour présenter la catégorisation. Cela a pu entraîner quelques fois une impression de «catalogue» dont nous prions le lecteur de bien vouloir nous pardonner.

Dans un premier temps, nous avons repris la définition de la catégorisation de Houdé pour qui *la catégorisation est la conduite adaptative fondamentale par laquelle l'individu découpe le réel physique et social.* A partir de celle de Neisser de ce qu'est catégoriser, nous avons proposé de définir *une catégorie comme étant une représentation mentale chez l'individu d'un ensemble d'objets qu'il appelle du même nom ou qui induisent de sa part un comportement identique*. Après un bref historique des recherches qui a permis de présenter la «révolution roschienne», nous avons décrit trois expériences destinées à illustrer cet historique et à étayer ce qui a suivi.

Le premier point que nous avons abordé ensuite concerne le fait que le modèle PAC décrit l'apprentissage supervisé. Il s'est agi de montrer que la catégorisation est aussi un apprentissage supervisé dans le sens où l'environnement impose à l'individu les

catégories qu'il doit former, s'il veut s'y adapter. Cela a été l'occasion de présenter les divers types de catégories étudiés par les psychologues.

Puisque le modèle PAC considère l'apprentissage comme un travail au niveau des représentations (on part de la représentation des exemples pour arriver à la représentation de la catégorie), nous avons présenté ce qui correspond à la notion de descripteurs dans ce modèle. Cela nous a amené à distinguer entre propriétés et attributs, entre attributs et traits. Ce qui correspond à l'ensemble des attributs dans le modèle PAC correspond à ce que Barsalou définit par l'ensemble des attributs disponibles. Un des problèmes de la psychologie est d'arriver à définir quel est cet ensemble d'attributs disponibles chez l'individu.

Cela nous a conduit à aborder la notion de schéma de représentations de l'hypothèse de l'apprenant. Derrière ce concept du modèle PAC, nous trouvons en psychologie les problématiques suivantes : traitement de l'information automatique vs stratégique, apprentissage implicite vs explicite, structure de la catégorie holistique vs componentielle. Autant d'oppositions qui s'apparentent à celle que l'on trouve, en apprentissage automatique, entre apprentissage symbolique et apprentissage connexionniste. Autant d'oppositions qui invitent aussi à penser que l'individu bénéficie de deux schémas de représentations qui probablement s'entrelacent et que l'on pourrait caractériser en ce que l'un est verbalisable et l'autre non. Derrière cette notion de schéma de représentation, on retrouve aussi, en psychologie, les controverses concernant la façon dont sont encodées les représentations des catégories : en conditions nécessaires et suffisantes, par prototypes, par exemplaires.

Puis, les conditions définies par le modèle PAC nous ont amené à nous intéresser à des aspects davantage méthodologiques de l'étude la catégorisation en psychologie et notamment à l'antagonisme entre les tenants de l'approche écologique et ceux de l'approche en laboratoire. Nous avons conclu cette section en proposant de caractériser les expériences selon le rôle qu'y joue la classe de concepts de référence. Nous obtenons ainsi deux axes : du plus au moins écologique, connaissance plus ou moins grande du contexte

Enfin, le modèle PAC étant aussi un modèle quantitatif, nous avons présenté les différentes mesures de la catégorisation quoique celles-ci, de par leur nature, soient plutôt rares.

Ainsi, il est possible d'utiliser les concepts du modèle PAC pour présenter les études relatives à la catégorisation en psychologie, cela le légitime au niveau cognitif. Mais peut-il, pour autant, intéresser le chercheur en psychologie ? Nous pensons que oui.
Par exemple, en l'utilisant comme grille de décodage de ses propres expériences. Ainsi, la dichotomie qu'il introduit entre concept cible et hypothèse de l'apprenant permet d'éviter quelques faux problèmes tel celui de l'appartenance que nous avons évoqué. Le problème de l'appartenance n'est pas le même selon qu'il se situe au niveau du concept cible (choisi par le chercheur) ou au niveau de l'hypothèse de l'apprenant. S'il se situe, au niveau du concept cible, c'est que celui-ci est mal choisi. Lorsqu'il n'y a pas de règle

établie pour définir si une photographie relève ou non du mobilier, il ne faut pas s'étonner que la représentation qu'en a le sujet soit plus ou moins floue et le chercheur ne peut rien en déduire. Par contre lorsque l'appartenance d'un item à la catégorie cible est clairement définie alors toute variation constatée chez le sujet de cette appartenance devient indicatrice.
Toujours en tant qu'outil de décodage, un autre aspect du modèle PAC qui nous semble être intéressant, est celui d'apprentissage supervisé et la notion concomitante d'étiquetage. *Ils amènent à distinguer entre le pourquoi d'une catégorie et le comment.* L'étiquette (l'appartenance ou non à une catégorie) donne le pourquoi. D'un certain point de vue, l'étiquette est un attribut particulier de l'item distinct des autres attributs. Cet attribut est défini par le rapport qu'entretient l'individu avec l'objet, rapport qui est identique pour tous les objets de la catégorie et qui est significatif de l'adaptation de l'individu à son environnement. De ce fait, cet attribut est nécessairement un attribut fonctionnel. Le terme de « fonctionnel » est à prendre dans un sens assez large : si le bol permet de boire, l'ortie «permet » de déclencher une démangeaison, le chat «permet » de le caresser. Cet attribut est le seul dont on ait la garantie qu'il se retrouvera parmi tous les items de la catégorie. La question de savoir si la représentation de la catégorie est construite sur la base d'une certaine similarité entre ses éléments, autre que celle liée à l'étiquette, relève, alors, plus du comment que du pourquoi. Ce n'est pas la similarité entre les divers attributs des items qui détermine que des objets distincts sont regroupés dans une même catégorie par l'individu. Cette similarité n'intervient qu'après et, à ce moment-là, elle peut porter aussi bien sur des attributs perceptifs que fonctionnels.

En plus d'une grille de lecture, nous pensons que le modèle PAC peut être utile parce qu'il lie de manière indissociable le processus de catégorisation (et les conditions dans lequel il se déroule), et la représentation produite de la catégorie. En cela, il permet de combler partiellement le fossé entre tenant d'une approche en laboratoire et ceux d'une approche écologique. Ces conditions, outre le protocole de présentation des exemples, sont définies par le biais des distributions de probabilités. Les introduire en tant que facteur dans des expériences sur des catégories artificielles nous semble utile. C'est un point que nous développerons au chapitre 5.

Enfin le modèle PAC est un modèle quantitatif. Nous avons vu les difficultés qu'il y a dans la catégorisation naturelle à pouvoir placer des mesures, néanmoins c'est un paramètre qui semble intéressant aussi d'introduire dans les recherches sur les catégories artificielles. C'est ce qui amène [Pitt, 97] à dire : « Les questions de complexité de l'échantillon et sa relation aux rythmes d'apprentissage et à la complexité de la catégorie n'ont pas été prises en compte dans la communauté des sciences cognitives et nous croyons qu'il est d'un grand intérêt d'étudier si les bornes analytiques proposées par la théorie de l'apprentissage permettent de prédire des différences dans les rythmes d'apprentissage humain. »[55] . Nous pensons comme lui que cet aspect de la catégorisation doit être pris en compte, non pas tant pour vérifier des bornes analytiques, que parce qu'il peut être à l'origine de phénomènes qualitatifs. Nous pensons par exemple aux travaux de [Medin, Wattenmaker et Hampson, 87] et à leur

[55] Il faut noter cependant que ces bornes sont très dépendantes de l'algorithme utilisé dans l'apprentissage et rien ne garantit que le-dit algorithme soit celui utilisé par un quelconque processus du cerveau.

difficulté à reproduire en laboratoire le phénomène de catégorisation par «air de famille ». Dans leur conclusion, ils proposent que le phénomène peut n'apparaître que si la structure de la représentation est *complexe*. Cette notion de complexité est elle-aussi une dimension écologique.

Ce qui précède relève de ce que pourrait apporter éventuellement le modèle PAC à la psychologie. Dans l'autre sens, nous avons vu que les expérimentations en laboratoire, ne se faisaient pas sur des distributions de probabilités aberrantes. Dans aucune expérience, l'expérimentateur ne s'est amusé à présenter à l'apprenant les exemples les plus bizarres d'un concept, ou des exemples uniquement négatifs. Bien souvent, la présentation des exemples se fait selon une distribution qui s'apparente à la distribution uniforme. Cela amène à se poser la question de savoir si la condition très forte dans le PAC apprentissage, qui réclame que l'apprentissage ait lieu pour toutes distributions de probabilités, est totalement pertinente. Lorsqu'on envisage d'utiliser le modèle PAC pour caractériser l'apprentissage naturel, il semble plus utile de modifier cette condition et de rechercher les distributions avec lesquelles apprennent les humains. Quelles sont les distributions «naturelles » qui président à l'apprentissage ? Quelle détermination jouent-elles dans la représentation que se fait un être humain d'une catégorie ?
La condition reste-t-elle valide lorsqu'il s'agit de l'apprentissage par une machine ? Si on utilise le modèle PAC pour définir des bornes absolues, il semble que la réponse est oui. Mais on déplace ainsi le problème. Est-il nécessaire d'avoir des bornes absolues en apprentissage automatique, ou celui-ci comme l'apprentissage humain n'opère-t-il pas dans un échantillon restreint de distributions de probabilités ?

Pour essayer de répondre à quelques-unes de ces questions, il sera nécessaire d'opérationnaliser le modèle PAC, c'est-à-dire de le transformer en protocole d'expérimentation. Avant cela, nous voudrions montrer plus encore que le modèle PAC et la recherche sur la catégorisation en psychologie étudient bien le même phénomène en présentant la manière qu'ont l'un et l'autre de traiter deux problèmes : l'économie cognitive et la typicalité.

Chapitre 3

Economie cognitive, typicalité
contribution à une étude comparée en psychologie et en théories formelles de l'apprentissage de ces deux phénomènes.

3 Introduction

Dans ce chapitre, nous voulons montrer que l'étude de la catégorisation et les théories formelles de l'apprentissage ne se rejoignent pas seulement sur le fait que l'on retrouve des concepts similaires dans les deux domaines, mais aussi par le fait qu'elles sont amenées à traiter des problèmes similaires. Nous allons plus particulièrement nous intéresser à deux d'entre eux, *l'économie cognitive*, d'une part, *la typicalité* de l'autre.

Comme nous l'avons vu, avant les travaux de Rosch, une catégorie naturelle était envisagée comme une catégorie conceptuelle, de type aristotélicien, qui se définit en conditions nécessaires et suffisantes et dans laquelle la catégorie sous-ordonnée a toutes les caractéristiques de la catégorie sur-ordonnée : un caniche est un chien qui est lui-même un mammifère qui est lui-même un animal... Cette vision de la catégorie a été partiellement abandonnée car elle ne tient pas compte du

- *niveau d'abstraction des catégories sémantiques,*
- *degré de typicalité.*

Nous avons déjà abordé ces deux notions dans le chapitre 2. Le niveau d'abstraction fait référence aux catégories de base, leurs sur-ordonnées et leurs sous-ordonnées. Alors qu'une vache est en même temps un ruminant, un mammifère, un animal, c'est par le terme de «vache» que l'on y fait référence et non par l'un des trois autres. Les catégories de base ont leur raison d'être en fonction d'une certaine *économie cognitive*. Cette notion d'économie sera le premier point abordé ici.

Le degré de typicalité fait référence au fait que certains éléments sont des représentants plus typiques que d'autres d'une catégorie. Ainsi le berger allemand est plus typique de la catégorie «chien» que le pékinois. Cette notion de *typicalité*, de *représentativité* est le deuxième point abordé ici.

Nous souhaitons montrer dans ce chapitre que ces notions d'économie et de typicalité constituent des points de convergence entre l'étude de la catégorisation en psychologie et les théories formelles de l'apprentissage en informatique. Nous les abordons de manière un peu différente en psychologie et en informatique. Alors qu'en psychologie nous faisons état des recherches concernant les représentations des catégories, en informatique nous parlons davantage du processus de catégorisation, d'apprentissage. L'idée, qui a déjà été émise précédemment, est que la forme des représentations trouve son explication dans le processus qui a généré ces représentations[56].

3.1 Notion d'économie

A l'inverse de ce que nous avons fait jusqu'à présent, nous allons présenter dans un premier temps l'approche psychologique de la notion d'économie, puis l'approche informatique. Nous montrons que ce qui est appelé économie cognitive en psychologie, relève en partie[57] de la taille de l'espace disponible en informatique. En théories formelles de l'apprentissage, une expression possible de cette taille de l'espace se trouve dans les algorithmes d'Occam. Ce sont des algorithmes qui retournent une hypothèse dont la représentation est plus courte que la somme des représentations des exemples contenus dans l'échantillon. Nous présentons un théorème montrant que, s'il existe un algorithme d'Occam pour une classe de concepts, alors cette classe est PAC apprenable et réciproquement que pour toute classe PAC apprenable il existe un algorithme d'Occam. *Ce théorème amène à proposer que toute compression **très forte** de l'information est apprentissage et que tout apprentissage est compression très forte de l'information.*

3.1.1 Economie cognitive en psychologie

3.1.1.1 Les catégories du niveau de base : meilleur compromis entre économie cognitive et contenu en information

[Rosch, 76] explique : «le rôle du système catégoriel est de fournir le maximum d'information pour le moindre effort cognitif (...) catégoriser un stimulus signifie le considérer dans la finalité de cette catégorisation, non seulement comme équivalent à des autres stimuli de la même catégorie, mais également différent des stimuli qui n'appartiennent pas à cette catégorie. D'un côté, il apparaît avantageux pour l'organisme

[56] Il existe tout un courant en informatique concernant la représentation des connaissances mais il n'est ni possible, ni utile, d'en faire état ici.

[57] En partie, car un autre aspect de l'économie cognitive consiste dans la rapidité du traitement de l'information.

de prédire toute propriété à partir d'une propriété quelconque, principe qui conduirait à la formation d'un très grand nombre de catégories les plus finement discriminées que possible. D'un autre côté, le but de la catégorisation est de réduire les différences infinies entre les stimuli à des proportions cognitivement et comportementalement utilisables. »

Toute catégorie est un compromis entre économie cognitive et contenu en informations (informativeness) [Komatsu, 92]. D'un côté, l'être humain catégorise le monde pour des raisons d'économie cognitive. Il ne lui est pas possible de mémoriser et de traiter chaque chose de son environnement comme unique, cela lui demanderait des capacités cognitives énormes. Le gain de l'économie cognitive résulte de ce que les caractéristiques qui distinguent deux choses peuvent être ignorées. De l'autre côté, la catégorisation a aussi sa raison d'être dans l'information qu'elle apporte. Lorsqu'une personne sait qu'un item particulier appartient à une catégorie donnée, elle peut supposer qu'il partage la plupart des propriétés de cette catégorie. Ainsi plus les catégories seront nombreuses et spécifiques et plus l'individu aura d'informations sur leurs membres. A l'opposé, moins elles seront nombreuses et plus elles seront générales et moins il aura d'informations.

Ce sont les catégories du niveau de base (voir 2.2.2.3) qui réalisent le meilleur compromis entre économie cognitive et contenu en information[Komatsu, 92], [Cordier, 93]. Lorsque l'on demande à des sujets de donner une liste de propriétés, le nombre de propriétés citées augmente très fortement lorsque l'on passe d'une catégorie sur-ordonnée à une catégorie de base. Par contre, il augmente très peu lorsque l'on passe d'une catégorie de base à ses sous-ordonnées [Cordier, 93]. Ainsi, lorsque l'on nous parle sans autre indication d'un «meuble» (catégorie sur-ordonnée), nous avons beaucoup moins d'informations sur l'objet que lorsque l'on nous parle d'une «chaise» ou d'une «table» (catégories de base). A l'opposé nous n'avons pas beaucoup plus d'informations lorsque l'on précise qu'il s'agit d'une «chaise de jardin» ou d'une «chaise de cuisine» (catégories sous-ordonnées).

3.1.1.2 L'économie cognitive correspond à la taille de l'espace disponible

Cette notion d'économie cognitive va servir de critère pour analyser les différentes structures possibles des représentations (voir 2.4.1). Komatsu oppose ainsi, d'une part, les représentations en conditions nécessaires et suffisantes et celles en air de famille, aux représentations par exemplaires, de l'autre. Tandis que les deux premiers types de représentation respectent le critère d'économie (une seule représentation pour toute la catégorie), il n'en est pas de même pour le troisième. Pour ce dernier, il existe différentes versions de ce que peut être la représentation en exemplaires qui vont d'un exemplaire par exemple, au multiprototype. Dans tous les cas, on stocke plusieurs représentations pour une même catégorie, ce qui réclame une grande capacité cognitive et pose un problème d'économie. La réponse apportée à cette objection par les tenants de ce type de représentation, est que certains items ou certains attributs non spécifiques sont oubliés au bout d'un certain temps, mais ils n'expliquent pas quels items ou quels attributs doivent être oubliés.

La notion d'économie cognitive sert aux psychologues pour tenter de définir comment sont encodées les représentations. Ainsi, le schéma ci-dessous emprunté à [Barsalou, 92] propose deux façons de le faire. Dans le panel A, la représentation est très économique car, il n'y a pas de redondance de l'information à l'opposé du panel B.

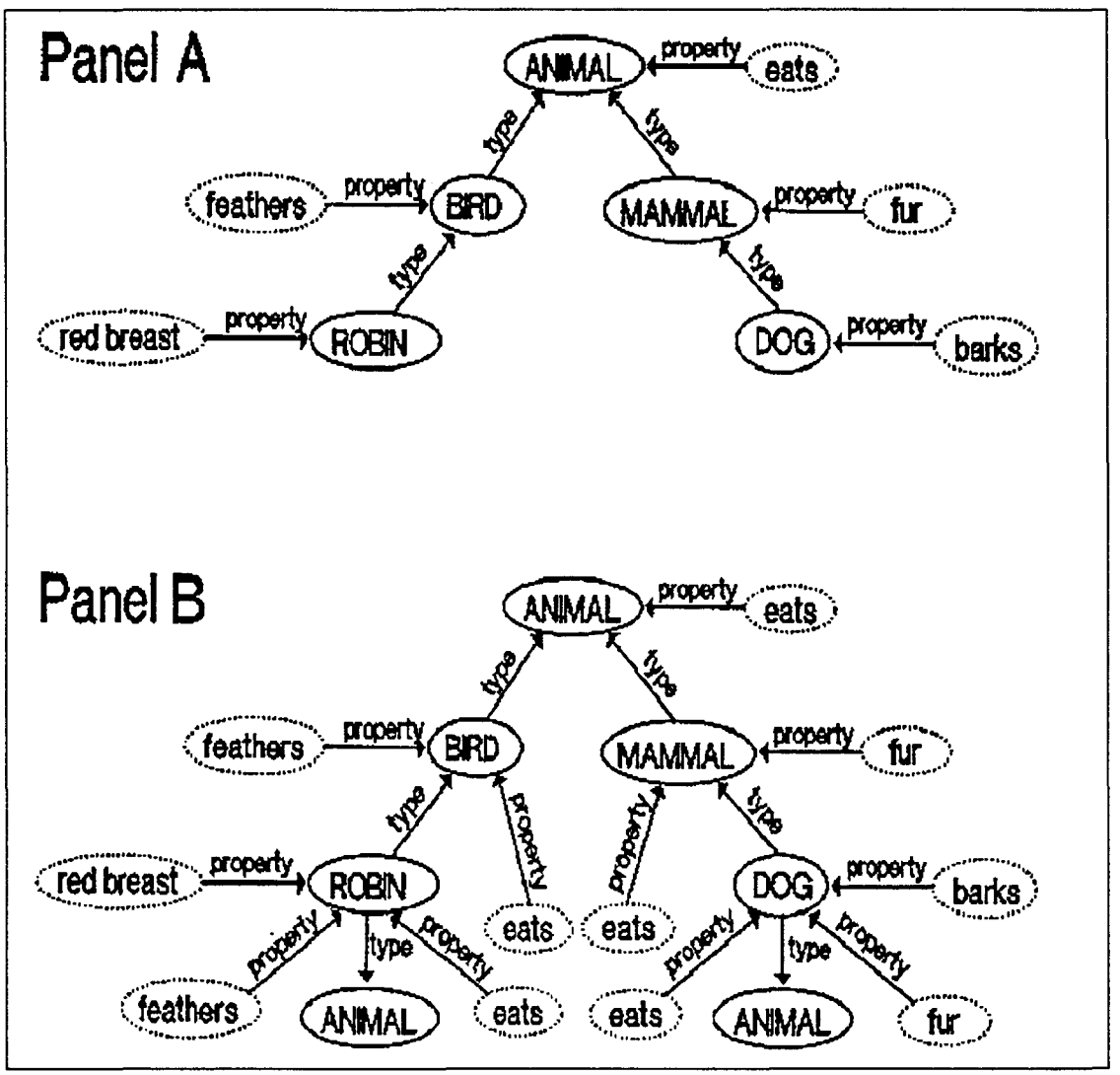


Figure 3.1.1 Exemple de taxonomie avec économie cognitive (panel A) et sans économie cognitive (panel B) ([Barsalou, 92])

La première, panel A, est basée sur des représentations en conditions nécessaires et suffisantes. Une des conditions «nécessaires » dans la définition du concept de chien est

d'être un mammifère, et tout mammifère est un animal. De la même manière, le rouge gorge est un oiseau, et tout oiseau est aussi un animal.
[Le Ny, 89] explique que lorsqu'on manipule un réseau sémantique, qui prend appui sur la notion d'héritage (tel que le A), on a très souvent affaire à deux catégories de représentations, des représentations d'individus et des représentations de traits, souvent interprétés comme des propriétés. La redondance de l'information est évitée par le biais de l'héritage, l'individu hérite de toutes les propriétés de son ascendant. Pour retrouver les propriétés du chien et du rouge-gorge, il suffit de remonter les chaînes d'héritage. Le chien hérite de toutes les propriétés des mammifères et des animaux, le rouge-gorge de celles des oiseaux et des animaux.
Conséquemment, un sujet devrait mettre plus de temps à retrouver pour le chien les propriétés qui relèvent du niveau animal que celles qui relèvent du niveau chien. Il est montré que c'est le cas. Chez l'homme, il y a une rapidité moindre dans la vérification de «un a est un c » («un moineau est un animal ») en comparaison de «un a est un b » («un moineau est un oiseau ») [Le Ny, 89].
Cependant, ce temps de traitement plus grand ne s'explique pas par cette structure en réseau, mais par des raisons de similarité. Le fait que le temps soit plus grand tient de la nécessité de faire abstraction de, c'est-à-dire d'inhiber, un certain nombre de valeurs d'attributs ou d'attributs eux-mêmes. C'est cela qui peut comporter un certain *coût cognitif* [Le Ny, 89] car, à mesure que l'on s'élève dans la hiérarchie conceptuelle, on voit s'appauvrir le nombre de traits. Il y a moins de traits en commun entre le rouge-gorge et l'animal qu'entre le rouge gorge et l'oiseau. Par ailleurs, d'autres expériences ont montré que les sujets mettent plus de temps pour confirmer l'assertion «le cantaloup est un melon », que celle «le cantaloup est un fruit » [Barsalou, 92].
Il est donc supposé qu'il y a probablement redondance des informations dans l'encodage comme indiqué dans le panel B. Barsalou oppose ainsi l'économie cognitive qui ne se situe qu'au niveau du stockage et le traitement (processing). La plupart des informations utiles, relativement au chien sont stockées avec le chien, ce qui permet de les retrouver rapidement. L'héritage peut cependant encore jouer un rôle, pour retrouver des informations qui sont rarement utilisées en relation avec la catégorie. Par exemple, le fait que le chien a des poumons sera inféré de ce qu'il est un mammifère.

Nous n'avons pas trouvé dans la littérature de définition stricte de ce qu'est l'économie cognitive. L'économie cognitive telle qu'on la voit ici est considérée comme une limitation de l'espace disponible pour stocker les représentations[58]. Cette économie cognitive consiste alors en une compression des données. On ne stocke pas les diverses représentations des exemples, mais une représentation de la catégorie à laquelle appartiennent ces exemples. La représentation de la catégorie est économique car elle est plus courte que la somme des représentations des exemples. Ceci nous amène naturellement au rasoir d'Occam en informatique.

[58] Encore une fois nous renvoyons à l'introduction générale, pour ce qui est de la manière d'envisager les représentations : lorsque nous parlons d'espace plus ou moins grand de stockage pour les informations nous faisons, par exemple, référence à des populations de neurones plus ou moins grandes.

3.1.2 Le théorème du rasoir d'Occam en informatique

Nous avons vu dans le chapitre 1 que, dans le modèle d'apprentissage PAC, une classe de concepts est apprenable dès qu'il existe un algorithme capable de l'apprendre. Il existe ainsi une variété d'algorithmes qui permettent d'apprendre, ce sont les algorithmes d'Occam dont la caractéristique est de comprimer très fortement l'information.
Alors que dans l'apprentissage PAC, l'algorithme doit retourner une hypothèse qui approche une cible, sur la base des exemples que l'Oracle lui a présentés, un algorithme d'Occam[59] cherche seulement à retourner une hypothèse qui explique, qui rende compte des exemples. «La différence cruciale qui existe entre l'apprentissage PAC et la nouvelle définition [Occam] est que dans l'apprentissage PAC l'échantillon tiré par l'algorithme d'apprentissage est destiné à n'être qu'une aide pour atteindre un modèle adéquat d'un processus externe (le concept cible et la distribution) tandis qu'avec la nouvelle définition, nous ne sommes concernés que par l'échantillon particulier devant nous et non par un processus externe.» [Kearns et Vazzirani,94]

L'évaluation du premier apprentissage (le PAC) est la capacité de l'hypothèse trouvée à classifier de nouveaux exemples sans faire trop d'erreurs. Une hypothèse n'est bonne que si elle a un bon pouvoir prédictif, que si elle est proche de la cible.
L'évaluation du second apprentissage (Occam) se fait sur la taille de l'hypothèse proposée : celle-ci doit être courte. L'apprentissage est alors synonyme de compression de données. L'idée est qu'en ayant les exemples non étiquetés d'un côté, l'hypothèse de l'autre, on puisse retrouver, grâce à cette hypothèse, l'étiquette de chaque exemple. La définition ci-dessous formalise cette idée de compression, de réduction de taille

3.1.2.1 Définition d'un algorithme d'Occam

Définition d'un algorithme d'Occam (définition 6 de [Kearns et Vazirani, 94])

$X=\cup_{n\geq 1}X_n$ *est l'espace d'exemples,*
$C=\cup_{n\geq 1}C_n$ *est la classe de concepts cibles,*
$H=\cup_{n\geq 1}H_n$ *est la classe de représentations des hypothèses.*
$c\in C_n$ *est le concept cible*
et $S=\{<x_1,c(x_1)>,\ldots..,<x_m,c(x_m)>\}$ *un échantillon de m exemples étiquetés selon c.*

Soient $\alpha\geq 0$ *et* $0\leq\beta<1$ *des constantes. L est un algorithme* (α,β)*-Occam pour C en utilisant H si, sur l'échantillon S de cardinalité m étiqueté suivant* $c\in C_n$ *, L retourne une hypothèse h telle que :*

---

[59]William of Ockham (1285-1349) à qui on attribue le «rasoir d'Occam » (autre orthographe) selon lequel «il ne faut pas multiplier les entités inutilement», ce qui est interprété en «l'explication la plus simple d'un phénomène est probablement l'explication correcte».

-h soit consistante avec les exemples,

-taille(h)≤ (n.taille(c))$^{\alpha}m^{\beta}$ (taille(c) est la taille de la plus courte représentation de c dans H)

L est un algorithme (α,β)-Occam efficace s'il tourne en temps polynomial en n, m, taille(c)

On pourra se reporter au chapitre 1 pour ce qui est de la définition de la classe de concepts, de la classe des représentations des hypothèses et le fait de les considérer en fonction de X_n. L'échantillon S est un ensemble de couples constitués d'un exemple et de son étiquette. L'étiquette $c(x_i)$ indique si l'exemple x_i relève (1) du concept *c* ou non (0). Nous ne revenons pas non plus sur le fait que L doit être un algorithme efficace qui doit tourner en temps polynomial en n, m, taille(c).

La définition d'un algorithme d'Occam réclame deux choses : que l'hypothèse soit consistante avec les exemples et que la taille de l'hypothèse soit beaucoup plus courte que la somme des étiquetages de ces exemples. La consistance de l'hypothèse avec les exemples implique que $h(x_i)=c(x_i)$ quel que soit i, c'est-à-dire qu'hypothèse et concept étiquettent les exemples de l'échantillon de la même façon. La seconde que l'hypothèse retournée doit être courte. C'est-à-dire que la taille de la représentation de l'hypothèse doit être inférieure à un polynôme en n, m et size(c).
Si la première condition (la consistance) est assez simple, la seconde par contre appelle quelques remarques. Nous pouvons envisager deux cas en fonction de *m* et *n*. Dans le premier, le nombre d'attributs, *n*, est beaucoup plus grand que le nombre d'exemples, *m*. C'est le cas le moins commun, il pourrait correspondre à l'apprentissage d'un concept avec *très peu* d'exemples qui comprennent eux-mêmes de *nombreux* attributs. Dans ce cas la condition est assez souple, elle réclame que la taille de la représentation de l'hypothèse soit un polynôme en la taille de la cible et du nombre d'attributs. Autrement dit, on ne réclame pas que la taille de l'hypothèse soit la taille de la plus petite représentation du concept (c'est un problème généralement difficile de trouver une telle hypothèse) mais une hypothèse un peu plus grande que le concept cible.

Le second cas où le nombre d'exemples, *m*, est beaucoup plus grand que le nombre d'attributs, *n*, est le plus courant. Nous pouvons alors simplifier *taille(h)≤ (n.taille(c))$^{\alpha}m^{\beta}$* en *taille(h)≤ m^{β}*. Autrement dit, on réclame que la taille de l'hypothèse soit beaucoup plus petite que le nombre d'exemples car $0\leq\beta<1$.
Pour avoir une idée nous pouvons proposer l'exemple suivant : supposons que l'on demande d'indiquer quels sont les nombres entre 0 et 10000 qui sont pairs. Une première possibilité pourrait consister à renvoyer l'étiquette de chacun des 10000 nombres qui indiquerait s'il est pair ou non. On aurait alors une hypothèse qui ferait 10000 caractères, un par nombre. Une autre consisterait à renvoyer l'hypothèse : « divisez le nombre par 2, si le reste est 0 alors le nombre est pair ». Cette hypothèse fait moins de 100 caractères. Cette hypothèse est beaucoup plus courte que l'hypothèse

définie par le nombre d'étiquettes. En posant $\beta<1$, la définition interdit la première solution et autorise la seconde[60].

Le fait que β soit un exposant plutôt qu'un coefficient multiplicateur garantit que l'hypothèse sera *beaucoup* plus courte. Si nous avons 10000 exemples et que l'on prend $\beta=0.5$, taille(h) doit alors être inférieure à 100. Si β avait été un coefficient multiplicateur, taille(h) n'aurait dû être inférieure qu'à 5000. Le fait que β soit un exposant empêche ainsi d'affirmer que tout logiciel de compression de données est un système apprenant.

On constate que dans la définition d'un algorithme d'Occam, il n'y a aucune référence au fait que l'hypothèse doit être proche du concept, ni aucune référence aux distributions de probabilités qui ont servi à tirer les exemples. On ne demande pas davantage que l'hypothèse obtenue étiquette à peu près correctement de nouveaux exemples. *Tout ce qu'on demande à l'hypothèse est qu'elle soit consistante avec les exemples vus et qu'elle soit beaucoup plus courte que le nombre d'exemples. Pourtant ces deux conditions suffisent pour qu'on obtienne une hypothèse approximativement correcte.* C'est ce que démontre le théorème du rasoir d'Occam.

3.1.2.2 Théorème du Rasoir d'Occam

Théorème du Rasoir d'Occam

Une classe de concepts C est PAC apprenable si et seulement s'il existe un algorithme d'Occam pour C

La démonstration que « s'il existe un algorithme d'Occam pour une classe de concepts C alors cette classe est PAC apprenable » est due à [Blumer, Ehrenfeucht, Haussler et Warmuth, 87]. La réciproque selon laquelle « si une classe est PAC apprenable alors il existe un algorithme d'Occam pour elle » est due à [Board et Pitt, 92].

Le théorème du rasoir d'Occam est un outil qui permet de faciliter les démonstrations d'apprenabilité dans le modèle PAC. Pour démontrer qu'un algorithme est un algorithme $(\alpha\text{-}\beta)$Occam, il suffit de démontrer qu'il produit une hypothèse consistante avec l'échantillon et que la représentation de cette hypothèse est beaucoup plus courte que celle de l'échantillon. Ainsi, les distributions de probabilités n'interviennent plus dans les démonstrations ce qui les simplifie singulièrement.

---

[60] Par ailleurs, si on demande par la suite si le nombre 10002 est pair, la première hypothèse ne permettra pas d'y répondre alors que la seconde le fera. Une mémorisation simple de l'échantillon n'a aucun pouvoir prédictif.

Le théorème d'Occam est à la base de plusieurs heuristiques efficaces en apprentissage automatique où l'on cherche à construire des systèmes apprenants. Une de ces heuristiques est *le principe MDL (minimum description length)*, énoncé par [Rissanen, 78]. Il propose que l'on choisisse l'hypothèse de telle manière que la longueur de la description de l'hypothèse et celle des exemples connaissant l'hypothèse soit la plus courte. Ce principe a une interprétation Bayésienne dans laquelle la longueur de la représentation détermine la probabilité à priori [Kearns et Vazirani, 94], [Li et Vitànyi, 95].

Cela peut sembler un peu «magique » qu'il suffise de retourner une hypothèse qui soit beaucoup plus petite que l'échantillon pour qu'il y ait apprentissage, cela l'est moins lorsque l'on étudie la démonstration de ce théorème. Le fait de réclamer une hypothèse petite, limite le nombre d'hypothèses possibles et facilite d'autant la recherche d'une hypothèse approximativement correcte. Pour s'en convaincre, il suffit de reprendre le nombre d'hypothèses possibles dans l'univers de RobSimp : nous y avons limité les conjonctions à deux valeurs d'attributs, cela fait 24 conjonctions possibles, si nous passons à trois valeurs d'attributs le nombre de conjonctions différentes est alors de 32. Plus la longueur des hypothèses est grande et plus il y a d'hypothèses différentes possibles.

3.1.3 Economie cognitive et rasoir d'Occam

Dans le cadre du modèle d'apprentissage PAC, le théorème du rasoir d'Occam est d'une «autre nature » que celui qui démontre, par exemple, l'apprenabilité des conjonctions. De par sa généralité, nous avons envie de dire qu'il est un niveau au-dessus. Cette généralité peut s'exprimer de la façon suivante : *toute compression très grande de l'information est un apprentissage et tout apprentissage est une compression très grande de l'information.*

Qu'en est-il pour l'apprentissage naturel ? Quel que soit le système cognitif, ses ressources en espace sont limitées. Si le nombre de neurones du cerveau et celui de leurs connexions est très grand, il n'est pas pour autant infini[61]. Nous pouvons donc supposer que l'adaptation par le système cognitif à son environnement *doit passer par une compression des données nécessitée par des raisons d'économie cognitive.* Il n'est pas économique/réaliste que chaque fois que l'enfant rencontre un chat il doive se *rappeler toutes les expériences* qu'il a eues précédemment avec des chats pour savoir l'attitude qu'il doit adopter vis-à-vis de celui-ci[62]. *Nous pouvons alors considérer qu'un algorithme d'Occam formalise cette capacité de l'individu à comprimer les données.*

---

[61] Notons cependant que dans ce domaine rien n'a encore n'a pu être démontré [Barsalou, 92] et notamment, nous ne savons toujours pas si la taille de l'espace en mémoire à long terme est limitée. Néanmoins voir la citation de Smith qui suit.

[62]Ceci va bien entendu à l'encontre d'une représentation de la catégorie en exemplaires.

Nous avons vu comment le modèle d'apprentissage PAC formalise la catégorisation : la catégorisation est vue comme la rencontre d'un apprenant avec des exemples (et contre-exemples) d'une catégorie. Cet apprenant doit se former une représentation de cette catégorie sur l'information délivrée par ces seuls exemples. Le modèle définit les critères de réussite (approximative) de cette catégorisation. Un de ces critères ε, la borne sur l'approximation, relie la réussite à la distribution de probabilités, ce faisant il prend en compte l'environnement dans lequel s'est effectuée la catégorisation.

Nous avons donc deux formalisations : la première décrit la capacité à comprimer l'information, la seconde décrit la catégorisation. *Ce que démontrent le théorème d'Occam et sa réciproque est que toute catégorisation est compression très forte de l'information, et que toute compression très forte de l'information est catégorisation.* Il confirme ainsi une thèse dominante en psychologie : «Coder par catégorie est fondamental pour la vie mentale, parce que cela réduit fortement les demandes sur les processus perceptuels, **l'espace de stockage,** et les processus de raisonnement, qui sont tous connus comme étant limités.» [Smith, 90]

Nous avons vu que, selon [Houdé, 92], la catégorisation est une *conduite adaptative fondamentale* par laquelle les systèmes cognitifs découpent le réel. La catégorisation est donc une condition nécessaire de l'adaptation à l'environnement. Nous pouvons alors généraliser l'une des deux implications : *toute adaptation à l'environnement par un système cognitif suppose de celui-ci qu'il soit capable de comprimer très fortement l'information donnée par cet environnement.*

Une conséquence en psychologie du théorème d'Occam consiste à rejeter une variante, au moins, du modèle de représentation par exemplaire. Dans cette variante, il est supposé que l'individu stocke une représentation de chaque élément de la catégorie qu'il a rencontré, nous pouvons alors estimer qu'il n'y a pas catégorisation car nous nous retrouvons dans le cas où $\beta=1$. Ceci équivaut au rejet de ce modèle par les psychologues en raison de l'économie cognitive.

Ce premier parallèle entre théories formelles de l'apprentissage et étude de la catégorisation concernant l'économie cognitive donne à penser que les rapprochements que l'on peut établir entre les deux disciplines peuvent être utiles. Malheureusement, il est difficile de l'approfondir. La raison en est que, bien que l'économie cognitive soit souvent invoquée en psychologie comme principe explicatif, cette notion n'est pas réellement définie et n'a pas fait l'objet d'étude particulière[63]. Cela se comprend aisément dans le sens où, si on assimile économie cognitive à taille de l'espace disponible, il faudrait pouvoir définir la nature de cet espace. Nous avons parlé de populations de neurones, car toute représentation est supportée par le cerveau. Si actuellement il est possible de définir des cartes des populations de neurones plus ou moins concernées par la vue ou par l'ouïe [Changeux, 83], [Edelman, 92], les recherches

[63]Il faut quand même signaler toutes les études relatives à la capacité de mémoire de travail, ou mémoire à court terme, mais ce qui nous intéresse ici relève davantage de la mémoire à long terme.

dans le domaine sont encore loin de pouvoir définir les populations concernées par une représentation précise.
Dire que «comprimer très fortement l'information c'est apprendre » amène à réfléchir à la compression de *quelle information*. A la fin de l'exemple 1.1 nous posions la question de savoir ce qu'avait réellement appris RobSimp. Pour les théories formelles de l'apprentissage, si l'hypothèse formulée par l'apprenant répond aux critères d'approximation alors on peut considérer que le concept est appris. Pourtant l'hypothèse peut être différente de ce concept, elle ne lui est qu'approximativement semblable et cette approximation est fonction de la distribution de probabilités. Il suffit que la distribution de probabilités soit autre pour que l'hypothèse précédente qui était approximativement correcte devienne erronée.
Prenons un exemple simple, supposons un monde où les objets sont décrits par deux attributs binaires, la taille (petit/grand), la couleur (blanc/noir), supposons que le concept cible soit «petit et blanc » et supposons que l'hypothèse apprise soit simplement «petit ». Si la distribution de probabilité est telle que tous les éléments «noirs » ont une probabilité de 0, aucun élément ne permettra de distinguer l'hypothèse du concept. Par contre si elle est telle que se sont tous les éléments «grands » qui ont une probabilité de 0 alors l'hypothèse et le concept seront très fortement différents puisque l'hypothèse étiquettera comme positifs tous les éléments de ce monde alors que le concept n'étiquettera comme positifs que les éléments «blancs ».
Ainsi, l'hypothèse apprise, lorsqu'elle est approximativement correcte, est le reflet de deux types d'informations : le concept et la distribution de probabilités. Notons que ces deux types d'informations sont ceux qui président au choix de l'échantillon. Il n'est donc pas étonnant qu'en comprimant l'échantillon (algorithme d'Occam), l'hypothèse apprise soit approximativement correcte (apprentissage PAC).
Nous pouvons donc considérer que l'apprenant apprend un concept et/avec une distribution de probabilités et cela pose la question, en psychologie, de savoir si la représentation que se fait un individu d'une catégorie ne reflète pas non plus la distribution de probabilités selon laquelle l'individu a rencontré les éléments de cette catégorie. Ce que les psychologues appellent la familiarité et/ou la fréquence d'instantiation. Cela amène à étudier de plus près le rôle des distributions de probabilité en théories formelles et la typicalité en psychologie. C'est l'objet des sections suivantes.

3.2 La typicalité en psychologie

La typicalité est un fait admis en psychologie. Il est néanmoins difficile d'en trouver une définition qui ne se ramène aussitôt à l'extension : « La typicalité ? Par exemple, la vache est un mammifère typique, la baleine ne l'est pas ». [Le Ny, 89] propose de la voir comme une *propriété des représentations*, comme un prédicat binaire : typique_de(«outil », «marteau ») qui se lirait 'le «marteau » est un exemple typique de «outil »'. Il est aussi possible de l'envisager comme un prédicat à trois termes : typique_de(«outil », «marteau », X). X représentant le degré de typicalité de «marteau » dans la catégorie «outil ». La typicalité a alors le statut de variable à valeurs multiples, c'est une échelle.

Il faut noter que le prédicat reçoit deux représentations en entrée, il est généralement admis qu'il s'agit d'une catégorie, «outil », et d'une sous-catégorie, «marteau ». Comme nous l'avons déjà expliqué, on parle de «sous-catégorie » pour le «marteau » et non pas d'exemple car lorsque l'on dit qu'un marteau est un outil typique, on ne pense pas à un marteau particulier mais à n'importe quel marteau. Une sous-catégorie étant elle-même une catégorie, il nous arrivera dans la suite de parler de «catégorie typique », il faudra le comprendre comme une ellipse de «sous-catégorie typique d'une catégorie sur-ordonnée ».

Pour cerner cette notion de typicalité nous allons dans un premier temps en présenter les différentes manifestations et, dans un second temps, essayer d'approcher une définition en compréhension qui nous obligera à passer par les schémas de représentations utilisés par l'individu.

3.2.1 Mise en évidence de la notion de la typicalité

Pour étudier la typicalité, les psychologues doivent pouvoir distinguer entre catégories typiques et catégories non-typiques. Faute d'une définition, puisque c'est l'objet de leur étude, ils recherchent des normes qui leur permettent de faire la distinction. Ainsi, dans le 2.1.2.2, dans la relation d'une des expériences de [Rosch et Mervis, 75], nous y montrions que les auteurs, préalablement à l'expérience, ont fait établir des listes de 50 à 60 items rangés par degré de typicalité pour 6 catégories. Pour établir ces normes, ils se basent sur différentes expériences que l'on a brièvement relatées dans le chapitre 2.

3.2.1.1 Normes de typicalité [Cordier, 93]

Le degré de typicalité d'une sous-catégorie peut être défini à partir de la fréquence de citations. On demande aux sujets de produire une liste de sous-catégories qui sont les meilleurs exemples d'une catégorie donnée. Les catégories cibles sont en général les animaux, les fleurs, les fruits, les oiseaux. [Cordier, 80] montre ainsi un large accord sur les animaux (85%), les fleurs (82%), les fruits et les légumes (73%) et un accord moindre sur les oiseaux (37%), où le moineau est en compétition avec l'aigle. Par exemple, pour les animaux, 238 sujets sur 280 citeront le «chien », 210 le «chat ».

Le degré de typicalité peut aussi être fourni par un jugement moyen. On donne une liste de sous-catégories et les sujets doivent estimer le degré de typicalité sur une échelle 1 à 7 (voir 2.1.2.2, l'expérience de [Rosch et Mervis, 75]), l'expérimentateur fait ensuite la moyenne obtenue par chaque sous-catégorie. Ce qui importe ce n'est pas tant la note que l'ordonnancement des sous-catégories en fonction de leur typicalité. Ce type d'études ayant été fait de nombreuses fois, il permet de montrer qu'il y a un large accord sur le degré de typicalité, ceci, bien évidemment, au sein d'une même culture.

Enfin le degré de typicalité peut être fonction d'un classement. On demande aux sujets de classer des dessins renvoyant à des sous-catégories en fonction de leur représentativité

par rapport à une catégorie donnée. Ce type de démarches est surtout utilisé avec des enfants qui ne peuvent comprendre les échelles d'intervalles et la numération.

Comme le fait remarquer, F. Cordier, chacune de ses méthodes a ses défauts. Pour la première, la fréquence d'usage des mots peut interférer avec la typicalité, tandis que, pour les deux dernières, le biais réside dans le fait que c'est l'expérimentateur qui sélectionne a priori les sous-catégories dont les sujets évalueront le degré de typicalité.

Le degré de typicalité des sous-catégories les plus typiques est relativement stable selon les âges. Par ailleurs, on peut supposer, bien que cela ne soit toujours pas clairement exprimé, que le degré de typicalité d'une sous-catégorie n'a pas valeur d'universel. Il est peu probable qu'un indien totonaque donne le même degré de typicalité à certaines sous-catégories de fleurs qu'un occidental, ne serait-ce que parce qu'ils ne connaissent pas les mêmes fleurs.

Ce n'est pas parce qu'il y a consensus au sein d'un même groupe concernant la typicalité de certaines catégories que cette typicalité est pour autant une réelle propriété de la représentation. Si la typicalité constitue une *variable* de la représentation de la catégorie, alors un changement de valeur de cette variable devrait affecter le traitement lors de l'utilisation de cette représentation. C'est ce que l'on peut voir dans ce qui suit.

3.2.1.2 Typicalité et traitement de l'information [Cordier, 93]

Dans divers traitements de l'information, la typicalité intervient d'une façon ou d'une autre. Le premier type de traitement qui montre le rôle de la typicalité est le jugement d'appartenance. On demande à un sujet si telle sous-catégorie appartient «oui» ou «non» à telle catégorie et on mesure son temps de réponse. Toutes les études s'accordent pour constater que le temps de réponse est plus court pour les sous-catégories typiques que pour les non-typiques.

La typicalité intervient aussi dans les erreurs de catégorisation chez les enfants. Au cours du développement de l'enfant, les sous-catégories jugées non typiques sont d'abord classées par lui comme n'appartenant pas à la catégorie. Les sous-catégories typiques sont une source possible de sur-extensions : un «lit» étant tout endroit où l'on peut s'allonger.

Dans les tâches de raisonnement, la typicalité joue dans la rapidité de traitement. [Cordier, 93] cite ainsi l'expérience de Cherniak qui demande à des sujets de dire si un syllogisme est correct ou non. Pour les deux syllogismes suivants le temps de réponse pour le premier est plus court que pour le second

Tous les A sont des moineaux
Tous les moineaux sont des oiseaux
Donc tous les A sont des oiseaux.
et

Tous les A sont des dindes
Toutes les dindes sont des oiseaux
Donc tous les A sont des oiseaux.

La typicalité peut aussi expliquer des erreurs de raisonnement. De l'assertion «tous les moineaux sont des oiseaux », le sujet passe à «tous les oiseaux sont des moineaux ». Chose qui n'arrive pas si l'assertion est «toutes les dindes sont des oiseaux ». Dans le même registre, on trouve la «conjunction fallacy » décrite dans le 2.5.2, avec l'expérience de [Tversky et Kahneman, 83]. Pour expliquer certaines erreurs, les psychologues émettent l'idée que le sujet attribue à la catégorie des propriétés qui n'appartiennent qu'à la sous-catégorie typique.

Les représentations sémantiques privilégiées ont des effets facilitateurs sur l'accessibilité lexicale : les sous-catégories typiques sont parmi les premières nommées par les enfants.

Enfin, les sous-catégories typiques se rappellent plus facilement et l'apprentissage est plus rapide lorsqu'il s'effectue avec de telles sous-catégories. «On observe un apprentissage plus rapide et plus stable des catégories lorsque les enfants sont confrontés exclusivement aux sous-catégories typiques par rapport à une situation où leur sont présentées des sous-catégories de typicalité variable. L'apprentissage de sous-catégories typiques précède l'apprentissage des sous-catégories non typiques. Cette efficacité peut être interprétée sur la base d'un processus de généralisation des réponses fondée sur la similitude, une similitude maximale entre bons exemples et minimale par rapport aux mauvais exemples». [Cordier, 93]

Nous venons de voir que la typicalité a une réelle existence. Il s'agit maintenant d'essayer de cerner de plus près sa nature.

3.2.2 Une approche de la définition de la typicalité

Dès que les psychologues essaient de définir ce qu'est la typicalité, ils sont très vite amenés à s'intéresser à la structure de la catégorie en «air de famille ». La typicalité est un épiphénomène de cette structure. Si on considère comme Le Ny que la typicalité est une propriété d'une représentation, il est naturel d'essayer d'appréhender la représentation pour comprendre la propriété.
Les psychologues vont alors s'efforcer de déterminer plus précisément cette structure. Nous avons vu dans le chapitre 2 les différentes structures envisagées : en conditions nécessaires et suffisantes, par exemplaires, par prototype. Nous allons brièvement les passer en revue puis nous intéresser plus particulièrement à la structure en prototypes.
Un autre centre d'intérêt des psychologues consiste dans le fait que les items d'une catégorie partagent un «air de famille », ils se ressemblent. Les psychologues tentent de définir des mesures qui appréhendent cette similarité. Ce sera le troisième point que nous aborderons.

3.2.2.1 les différentes représentations

Nous avons vu qu'une définition *en conditions nécessaires et suffisantes* n'est pas une bonne candidate à exprimer les représentations des catégories. Bien qu'elle permette d'exprimer l' «air de famille » puisque les membres de la catégorie doivent tous respecter les conditions nécessaires et suffisantes, elle ne permet pas de rendre compte des phénomènes de typicalité décrits ci-dessus.

Nous avons vu qu'une représentation *par exemplaire* était envisagée de deux façons. La première considère que l'individu encoderait autant d'exemplaires qu'il y a d'exemples, ce qui revient à dire qu'il mémorise tous les exemples. Comme nous l'avons expliqué, ce type de représentation contredit le principe d'économie cognitive, de plus il pose un problème : qu'est-ce qu'un exemple ? Est-ce un objet ou une apparition de cet objet ? Du point de vue perceptif, nous ne connaissons un objet particulier qu'au travers des diverses perceptions successives que nous avons de cet objet. Nous pouvons supposer qu'une représentation par exemplaire devrait mémoriser chaque apparition de l'objet. Nous voyons tout de suite le problème que cela pose du point de vue économique. L'idée dans ces modèles n'est pas que l'individu stocke chaque apparition d'un objet mais plutôt une représentation de cet objet. Cela implique que l'individu regroupe les diverses perceptions comme étant différents états du même objet. Ceci est déjà de la catégorisation. Nous pouvons alors nous demander pourquoi cette catégorisation s'arrêterait à l'objet particulier et non pas à la classe des objets.
Dans le second modèle de représentation par exemplaires, ceux-ci s'apparentent à des multiprototypes c'est-à-dire qu'une catégorie donnée est représentée au travers de plusieurs prototypes. Cette vision des choses est tout à fait envisageable du point de vue de l'économie cognitive, mais nous ne voyons pas en quoi elle diffère du modèle par prototype, le fait qu'il y en ait plusieurs pour une catégorie ne changeant pas fondamentalement les choses.

Dans le chapitre 2, nous avons vu qu'il existe un troisième type de schéma de représentation envisagé : le prototype. Nous le détaillons dans la section suivante.

3.2.2.2 Les représentations par prototypes

Les modèles à prototype connaissent un certain succès car ils permettent de rendre compte des phénomènes de typicalité et d' «air de famille ». Les catégories ne sont pas définies en conditions nécessaires et suffisantes mais par un ensemble de caractéristiques typiques. L'appartenance d'un item à une catégorie se définit par comparaison de l'instance au prototype de la catégorie. Ces prototypes sont des représentations

relativement stables en mémoire. Ainsi [Cordier, 93] étudie «l'organisation des représentations sémantiques en mémoire à long terme. »[64]

Deux types de prototypes envisagés :

Il est envisagé deux types de prototypes possibles :
-l'exemple prototypique : le prototype est le membre le plus typique de la catégorie, la «vache » pour représenter les «mammifères »
-le prototype comme résumé de la catégorie dans son entier.

Il y a peu d'auteurs qui défendent encore la première conception du prototype, le problème principal étant d'expliquer pourquoi, alors que le premier et le deuxième élément d'une catégorie sont très proches sur l'échelle de typicalité, l'un serait privilégié par rapport à l'autre.
Le prototype est plutôt considéré comme le point de référence de la catégorie, une entité abstraite que le sujet compose à partir des exemples rencontrés. [Mervis et Rosch, 81] «il existe au cours de ces formations de représentations, un mécanisme qui établit une sorte de 'centre de gravité' de la catégorie. ». Le prototype est alors ce que [Barsalou, 85] appelle les «tendances centrales » (central tendancies).
Dans ce qui suit, lorsque nous parlerons de «prototype », nous ferons implicitement référence à cette seconde version, c'est-à-dire un «prototype théorique » qui est une abstraction, un résumé de la catégorie. Pour être plus exact encore, les modèles à prototypes sont pour nous des schémas de représentations, dans le sens défini au chapitre 1[65], un prototype étant une représentation particulière dans ce schéma.

Rôle des valeurs d'attributs dans la définition du prototype

Si le prototype est une entité abstraite, il faut pouvoir expliquer comment elle se constitue. Deux théories sont envisagées. Dans l'une, la représentation du prototype est celle qui réunit les valeurs des attributs les plus fréquemment présentées sur des dimensions définies, elle est dite «modale »[66]. Dans l'autre, c'est la représentation qui réunit les moyennes des valeurs sur les dimensions définies, elle est dite «moyenne ». Pour faire la distinction entre les deux, il suffit d'imaginer qu'il faille apprendre un certain type de figures dont un des attributs est la taille. Pour cela nous avons 10 figures d'une taille de 10 cm, et 3 figures d'une taille de 3 cm. Dans le cas de prototype modal, la taille sera de 10 cm, (10 figures de 10 cm contre 3 de 3 cm) tandis que dans le cas du prototype moyen, il sera de 8 cm environ (109/13). [Cordier, 93] constate que selon que la présentation propose des exemples très discriminables ou non la sous-catégorie sera modale ou moyenne.

---

[64] F. Cordier n'utilise cependant par le terme de prototype pour désigner ces représentations mais préfère parler de sous-catégorie typique.
[65] Un schéma de représentations doit permettre, étant donnée une représentation et un exemple, de dire si l'exemple relève de l'hypothèse ou non (voir 1.2).
[66]On parle de prototype *modal* car en psychologie les valeurs d'un attribut sont appelées ses «modalités ».

Il faut noter que la théorie du prototype moyen ne peut se défendre que dans le cas où les attributs sont des dimensions. Il est plus difficile d'imaginer ce que peut être la valeur moyenne de l'attribut «forme » d'un prototype d'une catégorie qui comprend des carrés et des ronds.

Le poids des valeurs d'attributs dans la définition du prototype théorique dépend de la familiarité

Que la représentation du prototype soit modale ou moyenne, la familiarité des exemples dans l'environnement intervient dans sa genèse : « On fait l'hypothèse que le sujet a à sa disposition une certaine représentation de la probabilité d'apparition des propriétés (éventuellement de leurs corrélations) et une représentation statistique de la distribution des valeurs pour chacune d'elles »[Cordier, 93]. « On trouve dans l'intellect des individus, une représentation relativement valide des fréquences des valeurs de traits » [Mervis et Rosch, 81] Cette fréquence de la valeur d'attribut intervient ainsi dans la définition du prototype. On peut parler de son poids ou de son relief. Certains traits ont ainsi un poids tellement grand qu'ils peuvent en devenir quasi *définitoires.* Ainsi en est-il du trait «voler » dans le prototype d'oiseau.
Mais le relief d'un trait ne s'explique pas toujours par sa seule fréquence. Alors que le morse est un pinnipède, ses défenses le font considérer comme atypique de cette catégorie [Cordier, 93]. Le trait bénéficie, en quelque sorte, d'un poids «affectif ».

Le prototype ne se définit pas seulement en interne

Tel que nous l'avons décrit jusqu'ici le prototype est une abstraction qui est proche des éléments de la catégorie. C'est aussi une abstraction qui est loin des éléments des autres catégories, des autres prototypes. [Mervis et Rosch, 81] «les intellects humains maximisent spontanément les proximités cognitives intracatégories, ainsi que les distances cognitives intercatégories ». Selon [Cordier, 93] il semble que l'homogénéité intra-catégorielle soit plus forte que l'hétérogénéité inter-catégorielle pour l'enfant : dans l'évolution de l'enfant celui-ci mentionne d'abord les descripteurs qui soulignent le mieux l'air de famille avant d'énumérer ceux qui maximisent à la fois l'information dans la catégorie et le contraste entre catégories.

Flexibilité de la représentation par prototype

Nous avons dit que le prototype est une représentation stable stockée dans la mémoire à long terme. Cela ne signifie pas qu'elle ne soit pas flexible. Cette flexibilité va dépendre du contexte. Comme nous le disions dans le chapitre 2, le poisson le plus typique que l'on citera ne sera pas le même si on définit le contexte par : « le pécheur attendait que le poisson morde à l'hameçon. » ou «l'homme grenouille est attaqué par le poisson ». Cette flexibilité joue aussi sur le relief des traits.

> [Le Ny, 89] propose l'expérience suivante. Le matériel est composé de 16 items : une phrase suivie d'un mot
> Par exemple :

A «Un homme mendie près de l'église » a «porche »
B «En arrivant au village, on voyait l'église » b «clocher »
Les sujets doivent répondre aussi vite que possible, en appuyant sur un bouton «oui » ou un bouton «non » selon que le mot constitue ou non un détail de la scène représentée par la phrase. La mesure de l'importance du relief se fait par la mesure du temps de réponse. Il est constaté que Aa et Bb sont plus courts que Ba et Ab. Autrement dit selon le contexte, le clocher ou le porche de l'église sont des traits de l'église qui ont plus ou moins de relief. Le Ny explique que selon le contexte, on active des représentations transitoires de la catégorie.

L'appartenance d'un élément à la catégorie

L'appartenance d'un élément à la catégorie se définit par sa distance au prototype, sa ressemblance avec le prototype. Cette ressemblance pourra être plus ou moins grande selon que le nombre de traits qu'ils ont en commun est grand ou que le poids de ces traits est élevé.
Il arrive que ces *traits doivent être corrélés*. Ce n'est pas parce que la chauve-souris vole qu'elle est pour autant un oiseau. Ainsi pour les oiseaux, il faudra que les traits «bec » *et* «plumes » ou «bec » *et* «vole » soient présents simultanément dans l'item pour qu'il soit considéré comme relevant de la catégorie.
Cette distance de l'item au prototype définit son degré de typicalité, plus la distance est courte et plus l'élément est typique. La distance aux prototypes des autres catégories peut aussi intervenir mais de façon inverse. Ainsi un élément sera d'autant plus typique d'une catégorie qu'il sera proche du prototype de cette catégorie et éloigné des prototypes des autres catégories. Cette notion de distance ou de similarité suppose des mesures. C'est ce qui sera étudié dans la section suivante.

Critique de la formalisation d'un prototype proposée par [Pitt, 97]

Nous avons vu dans le chapitre 2, comment Pitt formalise la représentation en prototypes. Nous la reprenons ici et la mettons en parallèle avec ce qui vient d'être dit.
Pitt considère qu'un prototype est un ensemble pondéré d'attributs. Il rejoint ce que Barsalou et Komatsu appellent les «tendances centrales ». Le prototype est donc décrit au travers de ces attributs : « L'attribut a peut prendre les valeurs $v_1,...v_k$, chacune d'elle ayant un poids $t_a(v_1),....t_a(v_k)$ indiquant à quel point la valeur v_i est typique pour l'attribut a. De plus chaque attribut a peut avoir un poids w_a reflétant son importance. »
Ici, le relief d'un trait est exprimé à deux niveaux. Si nous considérons qu'un trait est une valeur d'attribut, Pitt donne un poids à la valeur parmi les autres valeurs mais aussi un poids à l'attribut. Ainsi le fait qu'une des caractéristiques du merle est d'être noir s'exprimera par le poids donné à «noir », $t_{couleur}(noir)$, parmi les autres valeurs de l'attribut couleur, mais aussi par le poids donné à l'attribut «couleur », $w_{couleur}$, parmi les autres attributs. Nous pouvons nous interroger sur la nécessité de ce double système de poids et si le poids au niveau des attributs est utile. Pour distinguer un oiseau des autres animaux, la couleur n'est pas une caractéristique fondamentale, on peut lui donner alors

un poids très faible *ou simplement l'ignorer*. Par contre pour distinguer un merle des autres oiseaux le fait qu'il soit noir est important, mais dans ce cas il faut prendre en considération le poids de la valeur d'attribut et non celui de l'attribut car ce dernier est valable de manière équivalente pour tous les oiseaux.
Pitt ne tranche pas entre représentation modale ou moyenne car il ne définit pas comment sont calculés les poids.

« La valeur de similarité que l'exemple x reçoit, sera la *somme* des poids des différents attributs avec l'attribut a contribuant pour $w_a.t_a(v_i)$, si x a la valeur v_i sur l'attribut a. Plus haute est la valeur, plus x est prototypique. » La typicalité est alors la distance de l'exemple au prototype exprimée par la somme des poids des traits. Le poids d'un trait se calcule en multipliant le poids de l'attribut par le poids de la valeur de cet attribut : pour le merle nous aurions $w_{couleur}$ multiplié par $t_{couleur}(noir)$.

Le défaut de cette modélisation est qu'elle ne tient pas compte du fait que les traits peuvent être corrélés. Il est possible d'y remédier en proposant que le poids de certaines conjonctions d'attributs ou de valeurs d'attributs soit plus élevé que la somme des poids respectifs. Si nous reprenons notre exemple du merle : le fait qu'il soit noir *avec* un bec jaune est plus important que la somme respective des deux traits :
$(w_{couleur} \cdot t_{couleur}(noir)) + (w_{couleur\text{-}bec} \cdot t_{couleur\text{-}bec}(jaune)) < w_{couleur\ et\ couleur\text{-}bec} \cdot t_{couleur\ et\ couleur\text{-}bec}(noir, jaune)$.

Cette modification complique, quelque peu, le calcul de la similarité de x. A l'ensemble d'attributs avec leurs poids et le poids de leurs valeurs qui constitue le prototype, il faut ajouter les corrélations d'attributs avec leur poids. Le fait de mettre des poids aux deux niveaux (attribut et valeur d'attribut) ne simplifie rien. Il semblerait donc plus cohérent de ne mettre les poids qu'au niveau des valeurs d'attributs, car c'est le trait qui est caractéristique de l'objet, et non l'attribut. Nous aurions alors des attributs a qui peuvent prendre les valeurs $v_1,...v_k$, chacune d'elle ayant un poids $t_a(v_1),....t_a(v_k)$ indiquant à quel point la valeur v_i est typique pour la classe c. Nous aurions aussi des poids pour les conjonctions d'attributs : $t_{a1\ et\ a2}(v_1, v_1),....,t_{a1\ et\ a2}(v_k, v_l)$. La similarité serait alors comme précédemment définie par la somme de ces poids.

Illustration de cette proposition

Pour illustrer, caricaturer ceci, nous pouvons reprendre notre exemple proposé en 1.1. Nous ne considérons que les exemples positifs de la catégorie et comme l'apprentissage se fait sur la base d'un échantillon et non sur la base d'une distribution théorique, ce sont les occurrences de l'exemple dans l'échantillon que nous prenons en compte. Le concept était C∧P et nous avions 6 exemples positifs de ce concept qui se répartissait comme suit : CPBE, 2 occurrences, CPBI, 4 occurrences. Les exemples CPNE et CPNI n'apparaissaient pas dans l'échantillon car leur probabilité était trop faible.

Les caractéristiques se voient alors affectées des poids suivants :

C (carré)	6,
P (petit)	6,
B (blanc)	6,
I (flèche intérieure)	4,
E (flèche extérieure)	2,
autres caractéristiques	0.

Si nous suivons la proposition de Pitt, nous obtenons que l'exemple CPBE est le plus typique puisque son poids total est de 6+6+6+4 soit 22. CPBE sera moins typique : 20. Un exemple comme CPNE se verra affecté du poids 16 bien qu'il n'apparaisse pas dans l'échantillon. De la même manière, RGNE sera affecté d'une typicalité de 2.
Concernant ce dernier exemple, notons que RGNE est un exemple négatif du concept, il n'est donc pas étonnant que sa typicalité soit faible. Il peut paraître contradictoire de parler de la typicalité d'un item relativement à une catégorie à laquelle il n'appartient pas. Pour le comprendre, on peut dire qu'un chat est plus typique d'un chien, qu'il ne l'est d'un oiseau. Nous retrouvons ici le fait que typicalité et similarité se recouvrent. Le chat n'est pas un chien mais il est plus similaire à un chien qu'à un oiseau.

Nous pouvons aussi favoriser les cooccurrences en donnant un poids plus fort aux exemples de la forme C∧P∧B puisque ces trois caractéristiques apparaissent simultanément dans tous les exemples positifs. Si nous doublons alors le poids lorsqu'il y a cooccurrence dans un exemple nous aurions CPBE qui aurait une typicalité encore plus forte que CPNE : 40 contre 16.

Notons que Pitt laisse en suspens le problème de l'appartenance. A partir de quel point la typicalité est-elle tellement faible qu'elle est le signe d'une non-appartenance. La réponse tient dans le fait que l'appartenance se définit par la typicalité de l'exemple avec les divers prototypes existants. L'exemple relève de la catégorie dont il est le plus typique.

Limites de la représentation par prototype

Le principal reproche fait à une représentation de la catégorie par prototype est qu'elle n'explique pas pourquoi certaines valeurs d'attributs sont retenues et d'autres non. Ceci renvoie à l'opposition que nous faisions dans le chapitre 2, entre attributs disponibles et attributs pertinents. Les traits qui sont retenus pour définir le prototype sont les traits pertinents. Leur pertinence peut s'expliquer par leur haute fréquence parmi les représentants de la catégorie. Nous voudrions, ici, citer [Le Ny, 89] «il n'existe dans le réel rien qui ne soit en lui-même abstrait et général, rien qui ne soit particulier et concret. Mais la pensée, ou l'activité cognitive, peut analyser chaque entité particulière en un faisceau de caractéristiques (ou de traits) et «décider » (explicitement et implicitement) de négliger certaines d'entre elles pour s'en tenir à un nombre plus limité de caractéristiques essentielles.
Seront conservées, c'est à dire cognitivement prises en considération, les caractéristiques que :

1) les individus ont en commun avec d'autres
2)l'activité cognitive a identifiées comme importantes, c'est-à-dire comme ne pouvant ou ne devant pas être négligées.
Ce processus d'abstraction, c'est-à-dire de *suppression* de traits - de 'propriétés'- est empiriquement premier ».
Cette citation permet d'illustrer l'opposition entre attributs disponibles et attributs pertinents : on élague dans les attributs disponibles pour ne garder que les attributs pertinents. Par ailleurs, elle rejoint ce que nous disions concernant les représentations par exemplaire.

Limites de cette présentation

Il n'existe pas dans la littérature de définition aussi tranchée de ce que peut être un prototype théorique que celle que l'on vient de présenter. La notion même de prototype théorique est rarement évoquée et pas toujours avec le sens donné ici. La raison en est que la plupart des travaux peuvent parler de la catégorie sans devoir faire référence à une définition en intension, en compréhension, de celle-ci. On peut très bien comparer la typicalité de deux éléments *a* et *b* d'une catégorie C sans pour autant décrire la représentation de cette catégorie. On peut ainsi tester les sujets pour savoir s'ils répondent plus vite à la question «le moineau est un oiseau ? » qu'à la question «l'autruche est un oiseau ? » sans pour autant devoir définir la représentation de la catégorie «oiseau ».

Par contre et systématiquement, dès que les auteurs doivent faire référence à la catégorie en intension, ils utilisent une représentation proche de celle que l'on vient de décrire. Barsalou, Komatsu parleront des « tendances centrales », et *de la représentation mentale de ces tendances centrales* [Komatsu, 92]. F.Cordier parlera de «sous-catégorie typique » : «la sous-catégorie typique peut très bien ne correspondre à aucune des catégories existantes. Il s'agit de toute manière d'une construction abstraite. » ou de «schéma : représentation mentale du prototype». [Le Ny, 89] pour sa part distingue entre *prototype observé*, *prototype théorique hypothétique* et *schéma*. Le prototype théorique dont nous parlons ici s'apparente alors davantage à ce que lui décrit comme schéma[67].

3.2.3 Les mesures de similarité

Que ce soit dans le modèle par exemplaires ou dans le modèle par prototype, il est toujours considéré que la typicalité d'un élément *x* est fonction de sa ressemblance aux autres membres de la catégorie ; plus l'objet partage l'air de famille plus il est typique.

[67] Concernant Le Ny, nous avons du mal à faire la distinction entre ce qu'il définit comme *prototype observé* et ce qu'il définit comme *prototype théorique hypothétique*, car toutes les objections qu'il formule à l'encontre d'une représentation en termes de prototype (p126 et suivantes) ne concernent que les prototypes observés. La réponse se situe peut-être p 109. : effectivement pour nous un prototype théorique est une représentation dans un schéma de représentation.

Dans le cas d'un modèle par exemplaire, il sera calculé la similarité de *x* à *tous* les exemplaires de la catégorie tandis que dans l'autre il sera calculé la similarité au prototype, aux tendances centrales.

Il existe différentes propositions pour définir la similarité en termes de distances. La première d'entre elles envisage la distance dans un espace multidimensionnel.

3.2.3.1 L'approche géométrique multidimensionnelle

Dans cette approche, un stimulus est décrit par les valeurs de *n* variables qui sont autant de dimensions. Par exemple les animaux seront décrits, entre autres, par les attributs «taille », «férocité », «proximité de l'homme ». Chaque dimension peut prendre ses valeurs sur un intervalle continu.
Un stimulus devient lors un point dans cet espace à *n* dimensions. La similarité entre deux stimuli est la distance qui les sépare dans cet espace. Dans la figure ci-dessous empruntée à [Osherson et Smith, 90] la pomme est plus similaire au raisin qu'à la tomate.

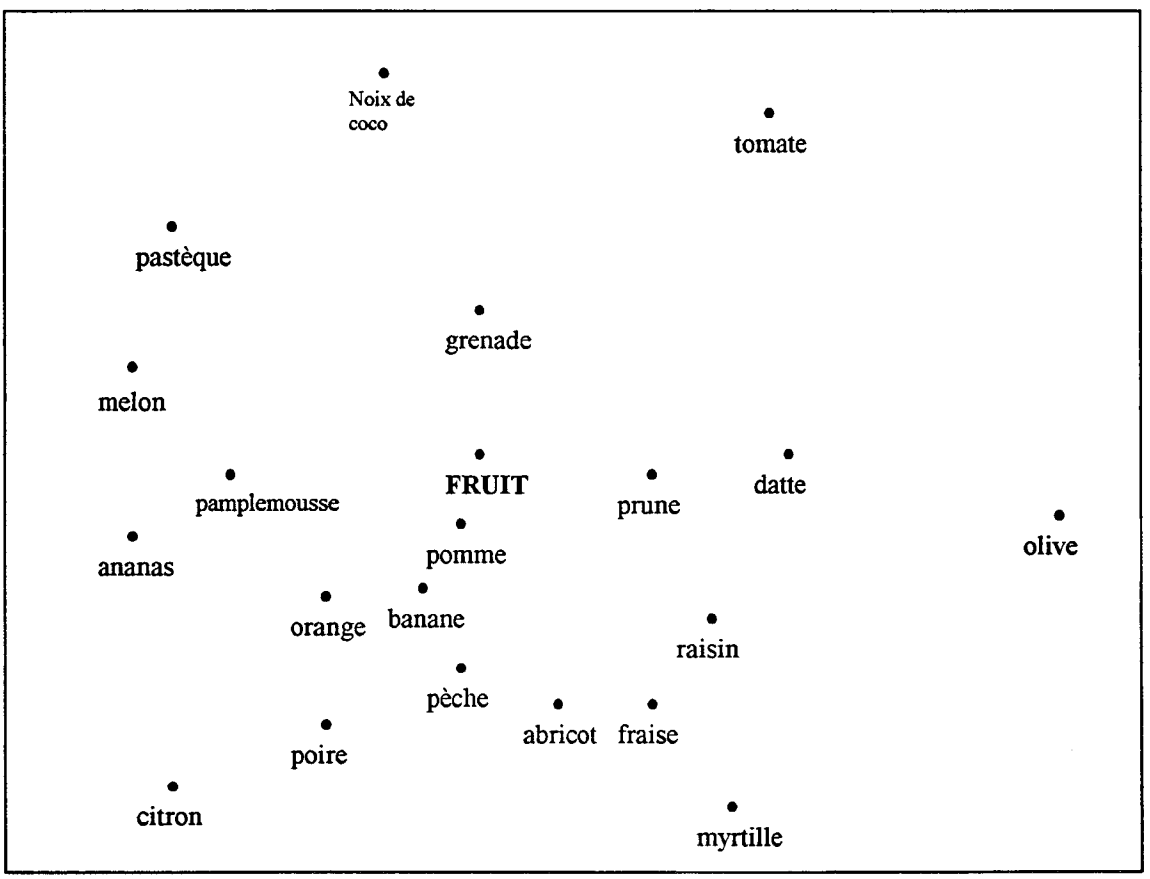


Figure 3.2.3 Un espace à deux dimensions pour représenter les relations de similarité parmi 20 exemples de la catégorie fruit et la catégorie «fruit » elle-même. [Osherson et Smith, 90].

Pour construire une telle représentation, l'expérimentateur demande au sujet de définir le degré de similarité entre les items, deux à deux. Ces données sont ensuite traitées par un logiciel [Shepard, 62] qui sur la base d'un espace dont le nombre de dimensions est prédéterminé (dans la figure cet espace est de 2 dimensions) retourne une représentation géométrique qui respecte le degré de similarité inter-items proposé par le sujet. Il reste alors au psychologue à essayer de définir ce que recouvrent ces dimensions. Ainsi Henley cité par [Thibaut, 97]en étudiant la catégorie des «animaux», a obtenu une représentation adéquate de la similarité des animaux dans un espace à 3 dimensions. Il a pu interpréter deux d'entre elles comme étant la «taille » et la «férocité ». Par contre, il n'a pu donner d'interprétation de la troisième. Selon Thibaut, cette troisième dimension est la résultante d'une série de composantes difficiles à réduire à une seule dimension sémantique.

Dans cette approche multidimensionnelle, la distance entre deux éléments est métrique, elle doit donc obéir à trois axiomes.
La *minimalité* selon laquelle la distance entre un point *a* et lui-même est égale à 0 et la distance de tout autre point *b* à ce point *a* est supérieure à la distance de ce point à lui-même

$$d(a,b) \geq d(a,a)=d(b,b)=0$$

La *symétrie* selon laquelle la distance d'un point *a* à un point *b* est égale à la distance d'un point *b* à un point *a*

$$d(a,b) = d(b,a)$$

L'inégalité triangulaire enfin selon laquelle, la distance de *a* à *c*, doit être inférieure à la somme des distances de *a* à *b* et de *b* à *c*.

$$d(a,c) \leq d(a,b) + d(b,c)$$

Le problème est que les jugements de similarité émis par les sujets ne respectent pas ces axiomes [Thibaut, 97], [Osherson et Smith, 90]. Le président des Etats-Unis est ainsi jugé plus semblable à lui-même qu'un sénateur quelconque, ce qui est contraire à l'axiome de minimalité ($d(a,a)\neq d(b,b)$). De la même façon l'axiome de symétrie est bafoué par des jugements qui considèrent que la Corée du Nord est plus semblable à la Chine que la Chine à la Corée du Nord, ou que la grenade est plus semblable à la pomme que la pomme à la grenade. Enfin, l'inégalité triangulaire n'est pas non plus respectée. Les sujets considèrent que Cuba est proche de la Jamaïque et de l'ex-URSS mais que l'ex-URSS est très dissemblable de la Jamaïque

$$d(\text{Jamaïque,URSS}) >> d(\text{Jamaïque,Cuba}) + d(\text{Cuba,URSS})$$

Nosofsky explique ces irrégularités par le degré d'activation des stimuli. La force de la représentation en mémoire d'un stimulus dépend de la fréquence de présentation de ce stimulus [Thibaut, 97]. Ainsi la représentation de «pomme » est plus souvent activée que celle de «grenade », elle aurait donc un poids plus grand. De même pour la Chine relativement à la Corée du Nord. Par ailleurs, selon l'interrogation, l'attention portée aux éléments n'est pas la même, le premier terme proposé a plus de poids que le second.

Le *modèle contrasté* de Tversky [Tversky, 77] permet de rendre compte de ces manquements aux axiomes. Osherson et Smith considèrent que son modèle relève d'une approche par caractéristiques.

3.2.3.2 L'approche de la similarité par caractéristiques

Dans l'approche par caractéristiques, un stimulus est représenté par des caractéristiques c'est-à-dire les valeurs d'attribut définies sur un espace discret (par opposition à continu), tel que «rouge », «rond » etc. La similarité entre deux objets sera fonction du nombre de caractéristiques communes et du nombre de caractéristiques différentes.
Dans le modèle contrasté de Tversky [Tversky, 77] deux objets sont d'autant plus similaires que le nombre de leurs caractéristiques communes est grand et que le nombre de caractéristiques distinctives est faible. Plus formellement :

$$sim(x,y)= k_c F(C)- k_x F(D_x)- k_y F(D_y)$$

C représente le nombre de caractéristiques communes aux stimulus x et y. D_x représente le nombre de caractéristiques distinctives de x, c'est-à-dire qui appartiennent à x et n'appartiennent pas à y (réciproquement pour D_y). La fonction F donne un poids particulier à chaque caractéristique selon sa pertinence. Les constantes k donnent un poids particulier aux caractéristiques communes ou distinctives selon le type de jugement effectué. Ce poids prend des valeurs différentes selon qu'il s'agit d'un jugement de similarité ou de dissimilarité.

Le tableau page suivante tiré de [Osherson et Smith, 90] indique comment ce modèle peut rendre compte des irrégularités observées relativement aux axiomes. Dans ce tableau, on considère que F attribue 1 à chaque caractéristique, que $k_c=a$, $k_x=b$, $k_y=c$.
Le premier tableau explique comment une pomme peut être plus similaire à elle-même qu'une grenade ne l'est à elle-même. Cela vient du fait qu'il y a plus caractéristiques pour décrire la pomme que la grenade.
Le second explique comment si l'on prend un paramètre *b* plus grand que *c*, on obtient que l'axiome de symétrie ne soit pas respecté, et qu'une grenade est plus similaire à une pomme que l'inverse
Enfin le troisième est relatif à l'inégalité triangulaire et l'on voit qu'un citron peut être jugé similaire à une orange, que celle-ci peut, elle-même, être similaire à un abricot sans que le citron soit jugé similaire à l'abricot.

Le problème avec le modèle contrasté est qu'il n'explique pas les caractéristiques à prendre en compte. La plupart du temps les chercheurs font établir par un groupe de sujets préalablement à l'expérience une liste de caractéristiques pour chaque catégorie. Le second problème est qu'il n'explique pas non plus comment calculer la fonction F et les paramètres *a, b, c* (ou k_c, k_x, k_y) qui permettent de donner plus ou moins de poids aux différentes valeurs d'attributs.

Illustration du modèle à contraste

Pomme	Pomme	Grenade	Grenade
Rouge	Rouge	Rouge	Rouge
Ronde	Ronde	Ronde	Ronde
Dure	Dure		
Sucrée	Sucrée		
Arbre	Arbre		
Sim(P,P)=a(5)-b(0)-c(0)		Sim(G,G)=a(2)-b(0)-c(0)	

Grenade	Pomme	Pomme	Grenade
Rouge	Rouge	Rouge	Rouge
Ronde	Ronde	Ronde	Ronde
	Dure	Dure	
	Sucrée	Sucrée	
	Arbre	Arbre	
Sim(G,P)=a(2)-b(0)-c(3)		Sim(P,G)=a(2)-b(3)-c(0)	

Citron	Orange	Orange	Abricot	Citron	Abricot
Jaune	Orange	Orange	Rouge	Jaune	Rouge
Ovale	Ronde	Ronde	Rond	Ovale	Rond
Sur	Sucrée	Sucrée	Sucrée	Sur	Sucrée
Arbre	Arbre	Arbre	Arbre	Arbre	Arbre
Agrume	Agrume	Agrume		Agrume	
Ade	Ade	Ade		Ade	
Sim(C,O)=a(3)-b(3)-c(3)		Sim(O,A)=a(3)-b(3)-c(1)		Sim(C, A)=a(1)-b(5)-c(3)	

Tableau 3.2.3 : illustration de l'utilisation du modèle à contraste [Osherson et Smith, 90]

Les constantes a,b,c donnent un poids particulier aux caractéristiques communes ou distinctives selon le type de jugement effectué. Dans le premier tableau, comme il y a plus caractéristiques pour décrire la pomme que la grenade, la pomme est plus similaire à elle-même qu'une grenade ne l'est à elle-même. Le second explique comment si l'on prend un paramètre *b* plus grand que *c*, on obtient qu'une grenade est plus similaire à une pomme que l'inverse. Enfin le troisième montre qu'un citron peut être jugé similaire à une orange, que celle-ci peut, elle-même, être similaire à un abricot sans que le citron soit jugé similaire à l'abricot.

3.2.4 Remarques sur la typicalité en psychologie

3.2.4.1 La démarche de la psychologie

Du point de vue expérimental, la typicalité est d'abord un phénomène observé. C'est un phénomène constaté par les chercheurs qui travaillent sur l'*extension* de la catégorie. Nous voulons dire par-là que les psychologues étudient la typicalité d'un item selon sa relation aux autres items de la catégorie (par opposition à la relation de cet item à une représentation de cette catégorie). Cette étude de la relation de l'item aux autres items a mis en évidence deux points :
-la typicalité introduit une relation d'ordre sur les éléments d'une catégorie : la pomme est un fruit plus typique que la grenade.
- la typicalité d'un élément est fonction de sa similarité à tous les éléments de la catégorie.

A partir de ce constat, les psychologues essaient de proposer des schémas de représentations qui soient cohérents avec ces observations. Une modélisation des représentations de l'individu doit pouvoir expliquer ce phénomène de typicalité, expliquer, entre autres, pourquoi le sujet met plus de temps à confirmer que «l'autruche est un oiseau », que «le moineau est un oiseau ». Le modèle en conditions nécessaires et suffisantes ne le permet pas. Une autre proposition, le modèle à prototypes, lui, l'autorise. C'est dans ce sens que nous disions que la typicalité est une propriété d'une représentation, il serait plus juste de dire d'un modèle de représentations. Pour illustrer, nous avons présenté, à partir de [Pitt, 97], une approche un peu formalisée de ce que pourrait être un schéma de représentations par prototypes. La typicalité s'y exprime par le poids attribué aux valeurs d'attributs.

Ainsi, si la typicalité en psychologie est envisagée comme une relation entre l'item et sa catégorie (la vache est un mammifère typique), dans les expérimentations, elle est mise à jour au travers des relations d'un item avec les autres items de la catégorie (la vache est un mammifère plus typique que la baleine). Le modèle de représentation proposé de la catégorie doit alors tenir compte des faits observés. Il doit permettre d'expliquer pourquoi une vache «appartient *plus*» à la catégorie mammifère que la baleine.

En psychologie, nous avons donc d'un côté des observations qui mettent en évidence la typicalité, de l'autre des modèles de représentations des catégories qui tiennent compte de ces observations. Il reste à expliquer pourquoi les individus en arrivent à former des représentations structurées selon ces modèles. Ainsi, si les études portent surtout sur les représentations *déjà* formées, quelques-unes cherchent à comprendre les représentations au travers de leur genèse.

3.2.4.2 La distribution de probabilités comme un des facteurs possibles de la typicalité

Ces représentations des catégories sont celles qui sont *portées* par l'individu. Elles correspondent à son espace d'hypothèses. Un exemple est alors *typique de l'hypothèse que se fait l'individu* de la catégorie. L'individu s'est forgé une hypothèse d'une catégorie particulière au travers des diverses rencontres qu'il a eues avec des exemples de cette catégorie, au travers d'un échantillon d'exemples. Monsieur Dupont n'a pas rencontré les mêmes exemples de la catégorie «chien » que Monsieur Durand. Mais l'échantillon que Monsieur Dupont a vu est probablement beaucoup plus proche de l'échantillon de Monsieur Durand que de celui d'un indien d'Amazonie. Ainsi, si Monsieur Dupont et Monsieur Durand sont d'accord pour dire que le berger allemand est un chien typique, nous pouvons émettre un doute quant à penser que l'indien d'Amazonie fait de même. Nous pouvons donc supposer que la typicalité du berger allemand affirmée par Messieurs Dupont et Durand provient des exemples qu'ils ont rencontrés.

Les exemples présentés, l'échantillon étant le point de rencontre entre un concept et une distribution de probabilité, et le concept étant le même pour ces trois personnes[68], on peut supposer qu'un des facteurs de la typicalité est donc la distribution de probabilités. Cela rejoint ainsi ce que nous écrivions à la fin du 3.1, selon quoi une hypothèse est un résumé du concept *et* de la distribution de probabilités qui a présidé à la rencontre de l'individu avec les exemples. La typicalité serait *en partie* ce qui dans l'hypothèse reflète la distribution de probabilités.

Allant dans le sens d'une telle hypothèse, nous trouvons les observations faites plus haut sur le rôle de la fréquence des traits dans le poids, le relief de ces traits, et les études de [Barsalou, 85] et de [Malt et Smith, 82] relatives au rôle de la familiarité et de la fréquence d'instantiation comme facteurs de la typicalité. A l'opposé de cette hypothèse, on trouvera les travaux de [Rosch-Heider, 71] et [Heider-Rosch et Olivier, 72] qui proposent qu'il puisse exister des sous-catégories typiques universelles. Ainsi il existerait un «rouge » reconnu universellement comme plus typique. Mais [Dubois, Resche-Rigon et Tenin, 97] montrent que les matériaux utilisés dans leurs expériences sont trop entachés d'ethnocentrisme pour valider un tel principe d'universalité[69]. Toujours contre ce principe d'une universalité des sous-catégories typiques, on peut aussi se référer à l'article de [Lammel, 97] qui montre que les représentations sémantiques du maïs chez les Totonaques du Mexique ne sont pas les mêmes que chez les Hongrois.

Ainsi, la typicalité est une propriété des représentations, des hypothèses, d'un individu qui a grandi dans un milieu donné. Ce n'est pas une propriété du concept qui est défini dans les dictionnaires, de la norme établie par la culture.

---

[68] En tant que norme qui leur est extérieure.

[69] Lorsque nous disons qu'il n'existe pas de sous-catégorie typique universelle, cela ne signifie pas que nous pensons que le principe de la typicalité, lui, n'est pas universel.

Avant de montrer comment est envisagée la typicalité en informatique, nous voudrions donc retenir de cette présentation que :
1) la typicalité d'un exemple en psychologie est une relation entre l'exemple et *l'hypothèse* du sujet et non le concept lui-même, que c'est une propriété de l'hypothèse du sujet,
2) les distributions de probabilités interviennent comme facteur dans cette typicalité.

3.3 La représentativité dans l'apprentissage PAC

3.3.1 Le modèle d'apprentissage PAC simple de [Li et Vitànyi, 91]

Avant de présenter comment est envisagée la représentativité dans l'apprentissage PAC, il nous paraît nécessaire de faire un détour par une variante du modèle PAC : le modèle d'apprentissage PAC simple de Li et Vitànyi. La présentation de ce modèle sera l'occasion de présenter quelques notions clefs qui serviront par la suite. Nous avons expliqué dans le chapitre 1 que certaines variantes du modèle PAC portaient sur la condition relative à l'exigence d'apprentissage sous toutes distributions de probabilité. Le modèle décrit ici remet en question cette condition et va proposer que l'apprentissage n'ait lieu que pour des distributions qui garantissent que tous les exemples simples font partie de l'échantillon présenté à l'apprenant. Il va donc falloir décrire dans un premier temps ce que peuvent être des exemples simples, dans un second temps les distributions de probabilités qui garantissent que ces exemples apparaissent dans l'échantillon, dans un troisième temps le modèle et enfin les résultats et les limites de ce modèle.

3.3.1.1 La complexité de Kolmogorov

De manière informelle

Pour appréhender la notion de simplicité, Li et Vitànyi reprennent la définition de la complexité de Kolmogorov selon laquelle un objet est d'autant plus simple que le programme capable de l'engendrer est plus court :

$$K(x)=min\{|p| \ tq \ U(p)=x\}$$

La complexité de Kolmogorov d'un objet x, que l'on note K(x), est la longueur du plus court programme qui, fourni à une machine de Turing universelle[70] préfixe[71] U, sort x.
Ainsi des objets très longs peuvent être très simples : 10.....0 avec un million de 0 sera écrit très simplement par un programme très court :

[70]Une machine de Turing universelle peut être assimilée à un ordinateur
[71]Pour des raisons techniques, on choisit de travailler avec des machines *préfixes*, c'est-à-dire pour lesquelles aucun programme n'est préfixe d'un autre, c'est le cas le plus courant.

```
écrire 1
pour i=1 à 1000000
        écrire 0
fin de pour
```

La complexité de Kolmogorov est de type algorithmique, c'est la taille du plus court programme qui permet d'engendrer l'objet. [Delahaye, 96] considère que la complexité de Kolmogorov permet d'appréhender la complexité aléatoire qu'il oppose à la complexité organisée. Ainsi une suite de 0 et de 1 obtenus en tirant à pile ou face est une chaîne aléatoire, il n'est pas possible de la comprimer. Le programme qui permet de reproduire une telle suite d'un million de tirages, ne peut être qu'un programme qui l'affiche (' Print « 10110...1 »'). Il ne pourrait pas être plus court que la chaîne elle-même, par opposition au programme ci-dessus qui affiche lui aussi une chaîne d'un million de caractères.

Pour [Delahaye, 96] la complexité de Kolmogorov d'un cristal est faible, car c'est la même structure qui se répète, le cristal peut donc être décrit par un programme court. Celle d'un être vivant est déjà plus grande mais n'est pas extrêmement longue. Le programme doit contenir la spécification de certaines lois physiques, l'indication des branches à emprunter si des choix non sélectifs nombreux doivent être faits dans une simulation de l'évolution. Ainsi la complexité de Kolmogorov d'une amibe est plus petite que celle d'un être humain car le programme permettant de décrire une amibe est certainement plus court que celui permettant de décrire un être humain. La complexité de Kolmogorov d'un gaz est très grande car le programme doit comporter l'ensemble des descriptions de *chacune* des molécules son emplacement et sa vitesse. Lorsqu'il écrit ceci, il envisage chacun de ces objets comme une collection d'atomes et le programme décrit leur position.

Plus formellement[72]

Pour définir la complexité de Kolmogorov, on considère une machine de Turing à trois bandes : un ruban d'entrée unidirectionnel ne pouvant contenir que des 0 et des 1, un ruban de travail bidirectionnel pouvant contenir des 0, des 1, des espaces et un ruban de sortie pouvant contenir des 0, des 1 et des espaces.
Si la machine de Turing T s'arrête avec la chaîne x sur le ruban de sortie, alors la chaîne finie p qu'elle a lue sur le ruban d'entrée est le programme qui a permis d'engendrer x : T(p)=x. On ne considère que les programmes préfixes, c'est-à-dire tous ceux qui ne sont pas le début d'un autre programme. On parle alors de programmes autodélimités, car ils indiquent eux-mêmes leur fin. Par exemple, des programmes écrits en Pascal sont des programmes préfixes, car ils se terminent tous par un marqueur de fin de programme, il

[72]Il est possible de trouver dans [Li et Vitànyi, 93] les définitions de la complexité de Kolmogorov et les preuves des théorèmes. Ce qui suit est un peu technique, la lecture n'en est pas indispensable pour comprendre la suite

n'est donc pas possible de trouver un programme donné écrit en Pascal qui soit le début d'un autre.

La complexité de Kolmogorov pour la machine de Turing T de la chaîne x est la longueur de la plus courte chaîne p qui permet d'engendrer x sur cette machine.

Définition de la complexité de Kolmogorov
La complexité de Kolmogorov préfixe de la chaîne x pour la machine de Turing T, notée $K_T(x)$*, est égale à min* $\{|p|\ tq\ T(p)=x\}$ *s'il existe un tel p, et à* ∞ *sinon*

Telle qu'elle est définie la complexité de Kolmogorov est dépendante de la machine, il conviendrait d'avoir une définition plus générale. Celle-ci est obtenue grâce au théorème d'invariance et aux machines de Turing universelles. Une machine de Turing universelle est une machine capable de simuler le comportement de n'importe quelle machine de Turing.

Théorème d'invariance
Soit U une machine de Turing universelle. Quelle que soit T une machine de Turing, $\exists\ C_{U,T} \in \mathrm{N}$ *tel que* $\forall\ x \in \{0,1\}^*$, $K_U(x) \leq K_T(x) + C_{U,T}$

De ce théorème d'invariance il est possible de tirer le corollaire suivant :

Corollaire
*Soit U et U' deux machines de Turing universelles, a*lors $\exists\ C_{U,U'} \in \mathrm{N}$ *tel que* $\forall\ x \in \{0,1\}^*$, $|\ K_U(x) - K_{U'}(x)\ | \leq C_{U,U'}$

Grâce à ce théorème on peut définir la complexité de Kolmogorov d'un objet x de manière quasi absolue, c'est-à-dire de manière indépendante de la machine universelle utilisée, puisqu'à une constante près les résultats sont les mêmes quelle que soit la machine. Ce corollaire est important car il libère la définition de la complexité de Kolmogorov, K(x), d'un objet x de la machine sur laquelle elle repose. C'est pourquoi, lorsque nous parlerons de la complexité de Kolmogorov d'un objet x, nous ferons référence à la complexité de Kolmogorov relative à une machine universelle U donnée $K_U(x)$, le corollaire rendant indifférent le choix de cette machine (à une constante près).

3.3.1.2 La distribution de probabilité «universelle » de Solomonoff-Levin

Pour introduire cette complexité dans la distribution des exemples Li et Vitànyi reprennent la définition de la distribution de probabilité «universelle »[73] proposée par Solomonoff-Levin :

[73]Le terme d' «universelle » utilisé pour définir cette distribution de probabilités, peut éventuellement se justifier dans le cadre théorique des distributions de probabilités, car la distribution de probabilités universelle *m* domine toute distribution de probabilités μ calculable (pour laquelle il existe un programme qui donne la valeur de probabilité pour tout entier *n*).

$\forall \mu, \exists$ c, une constante, telle que pour toute chaîne x, $m(x) > c\ \mu(x)$

$$m(x) = \sum_{p \in \{0,1\}^* \ tq \ U(p)=x} 2^{-|p|}$$

Ce que l'on pourrait traduire en : *la probabilité universelle d'un objet x d'apparaître est la probabilité qu'une machine de Turing le produise alors qu'elle a reçu comme programme une suite aléatoire de 0 et de 1.* Autrement dit, on tire à pile (0) ou face (1) chacun des bits qui constitueront la chaîne qui sera donné comme programme p à la machine de Turing. Si on considère que la longueur de la chaîne p est |p|, la probabilité d'avoir une chaîne particulière p en tirant chacun de ses bits de manière aléatoire est de $2^{-|p|}$ car elle est égale à (1/2 x 1/2 x.... 1/2) |p| fois.

Les seules chaînes, programmes, qui nous intéressent sont celles qui, données à la machine de Turing universelle U, produisent x en sortie, U(p)=x. Comme il existe plusieurs programmes différents qui peuvent produire x, on fait la somme des probabilités d'apparition de ces programmes.

Il a été démontré la propriété suivante :

$$m(x) \approx 2^{-K(x)}$$

Un objet simple donc engendré par des programmes courts aura une probabilité plus grande d'apparaître qu'un objet complexe donc engendré par des programmes longs. Par la définition de la complexité de Kolmogorov, dire qu'un objet est simple revient à dire que le plus petit programme capable de l'engendrer est court, donc plus K(x) est petit plus la probabilité qu'il apparaisse, $2^{-K(x)}$, est grande.

3.3.1.3 Modèle de Li et Vitànyi : apprentissage PAC-simple

Maintenant que nous avons montré comment nous pouvions introduire dans le modèle de Valiant la notion de simplicité des exemples, il est possible de compléter le modèle d'apprentissage PAC de Valiant pour obtenir le modèle d'apprentissage PAC simple de Li et Vitànyi. Pour cela il suffit de remplacer la distribution de probabilité D, quelconque, par la distribution de probabilités universelle m.

Cependant le terme paraît abusif sorti de son contexte. Nous le gardons néanmoins, pour des raisons de cohérence avec la littérature sur le sujet.

Définition de l'apprentissage simple PAC de Li et Vitànyi [74]

Soit C *une classe de concept sur* X *et* H *une classe de représentations. On dit que C est* ***simplement*** *PAC-apprenable s'il existe un algorithme* L *avec la propriété suivante : pour tout concept* $c \in C$, *pour la* ***distribution de probabilités universelle*** **m** *sur* X, *et pour tout* $0<\varepsilon<1/2$ *et* $0<\delta<1/2$ *si L a accès à EX*(c,m) *et aux entrées* ε *et* δ, *alors avec une probabilité d'au moins* $1-\delta$, L *retourne une hypothèse* $h \in H$ *satisfaisant erreur*(h)$\leq\varepsilon$.

Si L *tourne en temps polynomial en taille(c),* $1/\varepsilon$, *et* $1/\delta$, *on dit que C est efficacement* ***simplement*** *PAC apprenable.*

Nous pouvons constater que la seule différence avec le modèle PAC standard tient donc à ce que l'on demande ici que la distribution de probabilités soit universelle. Cette distribution de probabilités universelle garantit que les objets simples ont une plus grande probabilité d'être présentés à l'apprenant. De ce fait l'apprenant ayant des exemples simples à traiter, on peut supposer que l'apprentissage en sera facilité.

3.3.1.4 Les résultats et limites du modèle d'apprentissage simple PAC de Li et Vitànyi

En ne demandant plus que l'apprentissage se fasse pour toute distribution de probabilités comme dans le modèle standard, le modèle est en quelque sorte adouci. Li et Vitànyi ont ainsi démontré que certaines classes de concepts sont apprenables dans leur modèle telle celle des log-n-DNF. Une log-n-DNF est une DNF (disjonction de conjonctions) dont chaque monôme (conjonction) de littéraux a une complexité de Kolmogorov en $O(\log(n))$. Le problème est qu'une telle classe de concepts n'avait jamais été envisagée avant eux. C'est un peu comme si, ayant construit un bel outil, le chercheur se demandait ensuite à quoi il pourrait bien servir. Un autre problème est un problème technique : la complexité de Kolmogorov n'est pas calculable. Dans le cas de la classe des log-n-DNF, cela revient à dire qu'il n'existe pas d'algorithme capable de construire l'ensemble des concepts appartenant à celle-ci.

Il est difficile d'établir des parallèles entre ce modèle et l'apprentissage naturel. Ce que l'on peut retenir de cette section c'est une définition formelle de la simplicité qui s'appuie sur la théorie de la complexité de Kolmogorov, comment cette simplicité est prise en compte dans les distributions de probabilités et le fait que la condition réclamée dans le modèle PAC standard d'un apprentissage pour toute distribution de probabilités est remise en question. Seules les distributions permettant l'apparition d'exemples simples sont retenues.

---

[74] Dans cette définition et dans celles qui suivront nous ne faisons plus référence à *n* le nombre de variables pour en faciliter la lecture. Ceci est parfaitement acceptable dans le cadre de l'apprentissage de fonctions booléennes et sous certaines conditions qu'il n'est pas nécessaire de développer ici.

La notion de simplicité est ici en quelque sorte absolue dans le sens où elle ne dépend pas du concept à apprendre. Quel que soit le type de concept, ce sont toujours les mêmes exemples qui sont les plus simples. Il semble plus adéquat de relier la simplicité des exemples au concept dont il relève, c'est ce qui est fait dans le modèle suivant d'apprentissage avec distributions bienveillantes.

3.3.2 Le modèle d'apprentissage PAC avec distributions bienveillantes de Denis et Gilleron

Le modèle que nous allons présenter maintenant est décrit dans [Denis, D'Halluin et Gilleron, 96], [Denis, Gilleron et Simon, 97] et [Denis et Gilleron, 97, a]. Il modélise deux choses, d'une part, qu'il peut exister dans l'environnement de l'apprenant un enseignant, de l'autre, qu'il existe des exemples plus ou moins représentatifs d'un concept. Nous rejoignons, ici, le problème de la typicalité.

3.3.2.1 les distributions bienveillantes intègrent un enseignant dans l'environnement

Lorsque nous avons décrit le modèle PAC standard, nous avons dit qu'il s'agissait d'un apprentissage supervisé dans le sens où l'environnement *imposait* à l'apprenant les catégories qu'il devait former. L'étiquetage des exemples lui était fixé. Ce n'est pas le conducteur qui décide de classer arbitrairement les sections de routes en dangereuses ou non, c'est l'environnement qui le lui impose. De la même manière, c'est la communauté dans laquelle il vit qui lui impose d'appeler un chat, «un chat ». Cependant quand on parle d' «apprentissage supervisé » on pense plutôt au deuxième exemple qu'au premier. On imagine une «conscience » qui étiquette *sciemment* les exemples, on imagine un enseignant[75]. Le modèle PAC standard ne fait pas la distinction entre les deux. Il ne permet pas de faire la différence entre l'apprentissage par un enfant *sauvage* dans un environnement sans être humain et un enfant élevé dans une communauté. Le modèle PAC avec distribution bienveillante pallie ce manque en introduisant la notion de distribution bienveillante. C'est pourquoi on peut dire qu'il dépasse en l'englobant le modèle d'apprentissage PAC standard de Valiant. Avec ce modèle, la notion d'apprentissage *supervisé* prend tout son sens.

Nous allons présenter trois définitions que nous explicitons, nous invitons le lecteur à retourner à l'article [Denis et Gilleron, 97, a] pour les développements. Par ailleurs, certains des résultats obtenus sont décrits au chapitre 4 qui reprend l'article de [Denis, Gilleron et Simon, 97].

---

[75] Il faut prendre ce terme d'enseignant au sens large, cela peut être tout aussi bien les parents qu'un enseignant stricto sensu.

3.3.2.2 Définition d'un enseignant et d'un ensemble d'enseignement

Définition d'un ensemble d'enseignement

Soit C une classe de concepts. Un ensemble d'enseignement pour c de C est un échantillon de c. Un enseignant pour C est une fonction T qui associe à chaque concept c, un ensemble d'enseignement T(c), et telle qu'il existe une constante k telle que pour tout concept c, Card(T(c))≤|c|^k. Un enseignant est calculable s'il existe un algorithme qui prend en entrée un concept c de C et retourne un ensemble d'enseignement T(c).

Cette première définition modélise l'existence d'un enseignant. Cet enseignant a pour tâche de choisir un ensemble d'exemples en fonction de la cible. Cet ensemble d'enseignement doit être polynomial en la taille de la cible, c'est-à-dire qu'il doit être d'une taille réaliste. De la même manière, l'enseignant doit être calculable, c'est-à-dire que l'on veut être sûr que l'ensemble d'enseignement peut être créé. Ces questions de polynomialité et de calculabilité ne se posent pas en apprentissage naturel où l'on suppose que l'enseignant est capable de fournir un échantillon et qu'il fournit un échantillon d'une taille raisonnable.

Il faut faire attention à éviter le faux-sens : *cette définition ne donne aucune caractéristique «pédagogique » de l'ensemble d'enseignement.* Autrement dit, elle ne garantit en rien que l'enseignant ne soit pas mauvais et qu'il ne fournisse systématiquement le même ensemble d'enseignement quel que soit le concept. Elle ne garantit pas non plus que l'enseignant ne puisse être absent et que l'ensemble d'enseignement soit vide.

3.3.2.3 Définition des distributions de probabilités bienveillantes

La définition qui suit décrit des distributions de probabilités qui garantissent que l'ensemble d'enseignement est bien dans l'échantillon présenté à l'apprenant. Autrement dit que tout élément de l'ensemble d'enseignement a une probabilité non nulle d'être tiré.

Définition d'une distribution de probabilités bienveillante

Soit C une classe de concepts et T un enseignant pour C. Soit c un concept cible sur X_n et P une distribution de probabilités donnée sur X_n.
On définit

$$P_{\min}(c) = \begin{cases} \min\{P(x) \mid (x, c(x)) \in T(c)\} & \text{si } T(c) \neq \varnothing \\ 1 & \text{sinon} \end{cases}$$

Une distribution P est bienveillante pour c et T si $P_{min}(c)\neq 0$.

Cette définition divise l'ensemble des distributions de probabilité en deux classes : les distributions bienveillantes telles que $P_{min}(c)\neq 0$ et les autres telles que $P_{min}(c)=0$. La bienveillance d'une distribution dépend de $P_{min}(c)$ qui lui-même dépend de T. Ainsi d'un T à l'autre ce ne seront plus les mêmes distributions qui seront bienveillantes.
Pour qu'une probabilité P soit bienveillante, on calcule $P_{min}(c)$ ainsi : pour chaque élément de l'ensemble d'enseignement, on regarde sa probabilité selon P. $P_{min}(c)$ est alors la probabilité de l'élément de T(c) qui a la plus petite probabilité selon P. Si $P_{min}(c)=0$, cela signifie que la distribution P n'est pas bienveillante car il existe un élément de T(c) qui a une probabilité nulle et ne sera donc pas présenté à l'apprenant. Naturellement, si l'ensemble d'enseignement est vide ($T(c)=\varnothing$), $P_{min}(c)=1$, et toutes les distributions de probabilités sont bienveillantes.

Cette définition est au centre du nouveau modèle, elle permet d'y intégrer l'existence possible d'un enseignant dans l'environnement de l'apprenant. Elle exprime le fait que l'apprenant puisse rencontrer tous les exemples proposés par l'enseignant. Il faut cependant remarquer que si l'échantillon peut contenir tous les exemples de l'ensemble d'enseignement, il peut aussi en contenir d'autres, de la même façon que l'élève ne rencontre pas seulement des exemples lors du cours mais aussi en dehors de celui-ci dans sa vie quotidienne.

Encore une fois, il faut faire attention au faux-sens, la distribution est *bienveillante* parce qu'elle garantit que s'il existe un ensemble d'enseignement, alors ses éléments peuvent apparaître dans l'échantillon. Par contre, elle ne garantit pas que l'ensemble d'enseignement soit bienveillant, ni même qu'il existe, il peut être vide. La bienveillance de la distribution indique simplement qu'il y a un enseignant dans l'environnement de l'apprenant. Elle ne décrit pas les qualités pédagogiques de l'enseignant.

Ainsi la première définition «crée» l'enseignant et l'ensemble d'enseignement, la deuxième définition garantit que cet enseignant est dans l'environnement de l'apprenant. Il reste alors à définir un modèle d'apprentissage dans lequel l'enseignant apparaît. C'est la définition suivante.

3.3.2.4 Définition de l'apprentissage PAC avec distributions bienveillantes

Définition de l'apprentissage PAC avec distributions bienveillantes

Soit C une classe de concepts, H une classe de représentations et T un enseignant pour C. On dit que C est PAC-apprenable avec ***distribution bienveillante*** *s'il existe un algorithme* L *avec la propriété suivante : pour tout concept* $c\in C$, *pour toute* ***distribution de probabilités bienveillante P*** *sur* X, *et pour tout* $0<\varepsilon<1/2$ *et* $0<\delta<1/2$ *si L a accès à EX*(c,P) *et aux entrées* ε *et* δ, *alors avec une probabilité d'au moins* $1-\delta$, L *retourne une hypothèse* $h \in H$ *satisfaisant erreur*(h)$\leq\varepsilon$ et L *tourne en temps polynomial en taille(c),* $1/\varepsilon$, $1/\delta$ *et* $1/P_{min}(c)$.

Relativement au modèle PAC standard, ce modèle introduit deux différences. D'abord il ne considère que les distributions bienveillantes, ensuite il réclame que l'algorithme d'apprentissage tourne en temps polynomial en $1/ P_{min}(c)$. Il faut noter que si le paramètre $1/ P_{min}(c)$ est polynomial en $|c|$, la contrainte sur le temps est la même que dans le modèle PAC classique. Cette polynomialité en $1/ P_{min}(c)$ amène à élargir un peu l'interprétation que nous avons donnée de ce modèle. Pour l'instant nous avons toujours supposé qu'il modélisait l'existence d'un enseignant dans l'environnement. Une autre façon de l'envisager est de considérer que pour un concept donné, il existe un ensemble caractéristique d'éléments de ce concept (l'ensemble d'enseignement). Pour que le concept soit appris, il faut que cet ensemble caractéristique fasse partie de l'échantillon d'apprentissage. Dans cet ensemble caractéristique, il peut y avoir des éléments rares, avec une probabilité faible. Ce sont ces éléments qui déterminent $P_{min}(c)$. Il est alors judicieux de demander que le temps d'apprentissage soit fonction de l'apparition de ces éléments car étant rares, ils mettront plus de temps à apparaître. De là, la polynomialité en $1/ P_{min}(c)$.

Du point de vue informatique ce modèle généralise le modèle PAC classique[76] car une classe de concepts C apprenable dans le modèle PAC classique est apprenable dans ce modèle-ci. Il suffit pour cela que l'ensemble d'enseignement soit vide, ainsi toutes les distributions de probabilités deviennent bienveillantes. C'est pourquoi aussi, tel qu'il est défini, le modèle est peu opérant. Il ressemble à une coquille vide, dans le sens où il y a peu de contraintes sur l'enseignant ou sur l'ensemble d'enseignement. Tout va donc dépendre de la définition de l'enseignant qui sera donnée lors des démonstrations.

Nous voudrions maintenant présenter comment peut être abordée la représentativité des exemples. C'est ici que nous allons utiliser les notions vues de simplicité de l'échantillon et de distribution bienveillante.

3.3.3 L'apprentissage PAC avec un enseignant simple [Denis et Gilleron, 97]

Précisons tout de suite qu'un enseignant simple, n'est pas un enseignant dont le QI laisse à désirer mais un enseignant qui propose des exemples simples du concept. Avant de présenter le modèle, il est donc nécessaire de revenir sur cette notion de simplicité et sur la complexité de Kolmogorov.

3.3.3.1 La complexité de Kolmogorov conditionnelle

La complexité conditionnelle de Kolmogorov de x connaissant y pour exprimer la représentativité

[76] Comme dans le modèle PAC classique, un algorithme d'Occam et un théorème d'Occam pour ce modèle d'apprentissage sont définis. Nous les décrivons plus précisément dans le chapitre 4.

Nous avons défini dans le 3.3.1 ce qu'était la complexité de Kolmogorov d'un objet x. Il est nécessaire maintenant d'envisager la description d'un objet connaissant celle d'un autre objet. On peut imaginer qu'un programme qui décrit un carré puisse s'appuyer sur la description d'un rectangle. Le programme qui engendre la description du carré sera d'autant plus court. C'est ce qu'exprime la complexité conditionnelle de Kolmogorov de x connaissant y qui se note K(x|y).

Définition de la complexité conditionnelle de Kolmogorov [77]

La complexité conditionnelle de Kolmogorov préfixe de la chaîne x connaissant y pour la machine de Turing universelle U, notée $K_U(x|y)$, est égale à min $\{|p|$ tq $U(p,y)=x\}$ s'il existe un tel p, et à ∞ sinon

En apprentissage naturel la complexité de Kolmogorov de x connaissant y permet d'approcher aussi bien l'analogie[78] que l'induction. C'est dans le cadre de cette dernière que nous l'utilisons ici. Dans la suite ce qui sera envisagé, c'est la complexité de l'échantillon E connaissant le concept C : K(E|C). Cette complexité de K(E|C) exprime en quelque sorte la représentativité de l'échantillon relativement au concept. Si nous considérons deux échantillons E et E' dire que K(E|C) est inférieur à K(E'|C) revient à dire qu'il est plus facile de donner une représentation de l'échantillon E à partir de la description du concept C que de donner une représentation E' à partir de cette même représentation de C. Nous avons l'inégalité suivante

$$\textit{Soit x, y, z trois chaînes de } \Sigma^* : K(x|y) \leq K(x) + O(1)$$

L'inégalité montre qu'il est moins complexe de décrire la chaîne x à partir de la chaîne y qu'en ne partant de rien. Si nous reprenons notre exemple de description du carré, il est plus facile d'en faire la description en partant du rectangle que de la faire en ne partant de rien. Bien évidemment, si l'on part de la description d'un moteur à injection (y) pour arriver à celle d'un chat (x), l'information contenue dans la première ne sera pas d'une grande utilité pour la seconde : K(x|y) sera voisine de K(x).

C'est sur la base de la complexité conditionnelle de Kolmogorov que Denis et Gilleron proposent leur enseignant simple que nous présentons dans la section suivante, mais préalablement nous voudrions faire un aparté car les travaux de Kolmogorov amènent à une équation quelque peu fascinante.

---

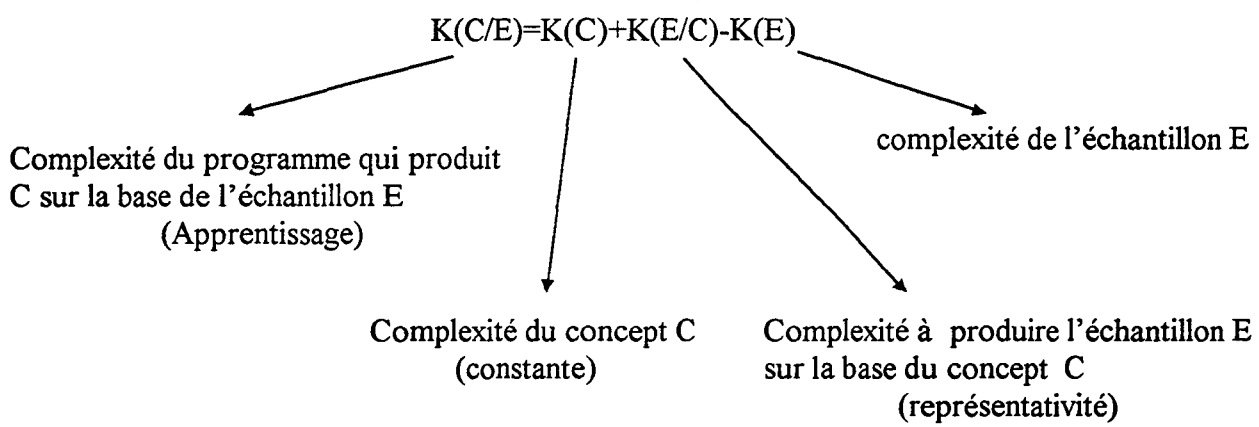
[77] Nous donnons tout de suite la définition avec une machine de Turing universelle car comme précédemment, il existe un théorème d'invariance et son corollaire. De la même manière nous abrégeons en K(x|y). Comme dans le 3.3.1 nous renvoyons à [Li et Vitànyi, 93] pour les preuves et les définitions plus complètes.

[78] voir notamment le travail de [Cornuejols, 95] qui utilise la complexité de Kolmogorov dans un apprentissage par analogie et les rapports qu'il établit avec les travaux de Cordier.

Aparté

Soit C et E deux chaînes de Σ^* : $K(C/E)+K(E)=K(C)+K(E/C)$ [79]

De cette équation, il est possible de déduire la suivante : K(C/E)=K(C)+K(E/C)-K(E). Si l'on considère que C est la description du concept et E la description des exemples, on peut se demander si cette équation ne décrirait pas la complexité de l'apprentissage. En effet, l'apprentissage tel qu'on le décrit, ici, consiste à partir d'une représentation des exemples pour arriver à une représentation du concept (C|E). L'apprenant est alors un algorithme qui passe de l'une à l'autre. K(C|E) exprime la taille du plus court de ces algorithmes.



Une interprétation possible de cette égalité est que, pour un concept C donné, plus la complexité de l'échantillon (K(E)) est grande plus l'apprentissage est facilité (K(C|E) petit), sous réserve que K(E|C) ne varie pas dans le même temps. De la même manière plus l'échantillon est représentatif, (K(E|C) petit) et plus l'apprentissage est aisé, sous réserve que K(E) ne varie pas dans le même temps.

3.3.3.2 L'apprentissage PAC avec un enseignant simple

Nous pouvons maintenant présenter le modèle d'apprentissage PAC avec un enseignant simple.

Définition d'un enseignant simple

Soit C une classe de concepts, T est un enseignant simple s'il existe une constante $k > 0$ *satisfaisant :*

$\forall c \in C, \forall (x,(c(x)) \in T(c), K(x|c) \leq k \log(|c|)$

[79] A O(log(K(E))+log(K(C))) près.

Autrement dit, l'enseignant sera dit simple si les exemples qu'il propose dans son échantillon ont une complexité conditionnelle de Kolmogorov petite relativement au concept. Si l'on accepte d'assimiler la faible complexité conditionnelle de Kolmogorov d'un exemple relativement au concept à la représentativité de cet exemple, cela revient à dire que l'enseignant propose un échantillon représentatif. Nous voyons, ici, que le terme d'*enseignant simple* est un raccourci pour exprimer le fait que les exemples proposés par l'enseignant sont simples.

Avec d'un côté, un modèle qui intègre les distributions de probabilités facilitantes, c'est-à-dire qui incluent un enseignant dans l'environnement, de l'autre, cet enseignant simple, il est possible de définir ce qu'est un apprentissage PAC avec un enseignant simple.

Définition de l'apprentissage PAC avec un enseignant simple

Soit C une classe de concepts. C est PAC apprenable avec un enseignant simple si C est PAC apprenable sous des distributions de probabilités bienveillantes pour un enseignant simple.

Nous avions dit que le modèle d'apprentissage était une coquille vide que pour la remplir, il fallait définir l'enseignant. Avec cette définition, Denis et Gilleron intègrent un enseignant possible. On notera la pertinence cognitive de leur enseignant qui a pour objectif de présenter des exemples simples du concept.

Ainsi dans un premier temps, Denis et Gilleron ont défini un modèle d'apprentissage PAC qui tient compte de l'existence possible d'un enseignant dans l'environnement, dans un second temps, ils définissent un enseignant qui propose des exemples destinés à aider l'apprenant à se former le concept.

3.3.3.3 La représentativité de l'échantillon est aussi fonction de l'apprenant

Dans le chapitre 4, nous verrons cependant que, dans les démonstrations, la représentativité ne se définit pas seulement en fonction du concept cible mais aussi en tenant compte de l'algorithme qu'utilise l'apprenant pour apprendre. C'est-à-dire qu'un échantillon n'est réellement représentatif que si l'algorithme d'apprentissage tient compte en quelque sorte de cette représentativité. Ainsi, si l'on suppose que l'échantillon contient tous les exemples simples relativement au concept, l'algorithme tiendra compte de ce fait pour apprendre. Il «saura» que tous les exemples simples figurent dans l'échantillon. Autrement dit, il bénéficie d'une information *supplémentaire*. En théories formelles de l'apprentissage, ceci pose le problème de la *collusion* entre enseignant et apprenant : dans quelle mesure l'information que l'on donne à l'apprenant en plus de celle contenue dans les exemples de l'échantillon ne biaise-t-elle pas l'apprentissage ?

Une autre façon d'envisager ceci, qui a une certaine pertinence cognitive, est que l'enseignant choisit un ensemble d'exemples représentatifs non pas seulement en fonction du concept mais aussi de l'apprenant. Personne ne s'est jamais interrogé sur le fait qu'en classe de cours préparatoire, l'instituteur ne donne pas du Proust à lire aux enfants. Si l'enseignant, en théories formelles, propose des exemples simples à l'apprenant, c'est qu'il sait que l'apprenant est surtout capable de traiter ceux-là.
D'une connaissance supplémentaire de l'apprenant sur le concept, on passe à une connaissance de l'enseignant sur l'apprenant.

3.3.4 Interprétation et limites du modèle PAC avec distributions bienveillantes

Résumons cette section 3.3. Nous sommes partis du modèle de Li et Vitànyi, qui propose de ne considérer que des distributions de probabilités qui ne présentent que des exemples simples. Cette simplicité est définie par la théorie de la complexité de Kolmogorov. Denis et Gilleron élargissent cette vision en distinguant deux choses qui sont confondues dans le modèle de Li et Vitànyi. La première est l'apprentissage pour *certaines* classes de distributions de probabilités et non pour toutes. La seconde est la notion de *simplicité* des exemples. De plus concernant la simplicité des exemples, ils ne considèrent pas la simplicité brute des exemples, mais la simplicité relativement au concept. Sous certaines conditions, leur modèle généralise celui de Valiant.

Du point de vue informatique, le modèle avec enseignant simple a un gros défaut dû à la complexité de Kolmogorov : celle-ci n'est pas calculable. Elle est donc peu opérationnelle. Du point de vue cognitif, le défaut est que toute distribution calculable est simple [Denis et Gilleron, 97]. Comme on peut considérer que les distributions selon lesquelles sont tirés les exemples que rencontre l'individu sont calculables, elles sont donc toutes simples. Le critère $K(x|c) \leq k \log(|c|)$ n'est donc pas exploitable pour exprimer la simplicité, la représentativité d'un exemple relativement à un concept en apprentissage naturel. Ce n'est pas pour autant qu'il faille rejeter l'ensemble. L'idée de base de la théorie de Kolmogorov selon laquelle la représentativité d'un exemple relativement à un concept est fonction de la taille du programme qui, connaissant le concept, retourne l'exemple, n'est pas sans intérêt.

Mais l'aspect essentiel du modèle PAC avec distributions bienveillantes est de considérer que toutes les distributions de probabilités ne sont pas équivalentes, qu'il y en a de bonnes et de mauvaises. Les bonnes sont celles qui facilitent l'apprentissage et les mauvaises sont celles qui le freinent. En ce sens, ce modèle constitue un réel dépassement du modèle PAC standard de Valiant. Pas seulement dans le sens précédent, où tout ce que l'on peut démontrer apprenable dans le modèle standard est apprenable dans ce modèle-ci, mais dans le sens où il amène aussi bien informaticiens que psychologues à s'interroger sur le rôle des distributions de probabilités et non plus seulement sur celui des classes de concepts. ***Il invite à envisager l'apprenabilité des classes de concepts dans des classes de distributions.***

Cependant, dire qu'il existe des classes de distributions qui sont bienveillantes, qui facilitent l'apprentissage, ne permet pas pour autant de comprendre ce qui les rend bienveillantes. Nous l'avons expliqué, leur bienveillance tient dans le fait qu'elles garantissent que l'ensemble d'enseignement apparaîtra dans l'échantillon. Ainsi, si l'ensemble d'enseignement est un «mauvais» ensemble, de bienveillantes, les distributions deviennent malveillantes. Toute la bienveillance tient donc dans la définition de cet ensemble d'enseignement. Cette bienveillance est fonction du concept mais aussi de l'algorithme apprenant. Ceci, peut poser, du point de vue des théories formelles, un problème de collusion, mais dans le même temps il accroît la pertinence cognitive du modèle. Un enseignant est une personne qui, normalement, cherche à faciliter l'apprentissage et donc à présenter des exemples simples en fonction du concept mais aussi en fonction de l'élève. On trouvera dans [Barth, 87] tout un ensemble de conseils pédagogiques sur la manière de choisir les exemples relativement à un concept cible pour faciliter l'apprentissage. Barth, pour cela, s'appuie sur la connaissance qu'elle a de l'enfant.

Cependant, ne voir dans ce modèle que l'addition d'un enseignant dans l'environnement, c'est le réduire. Nous proposons, dans la conclusion, une autre interprétation en termes seulement d'exemples représentatifs.

3 Conclusion

Dans ce chapitre, nous avons voulu mettre en exergue la convergence entre informatique et psychologie sur deux points particuliers : la notion d'économie et la typicalité. Ce que la psychologie appelle économie cognitive s'apparente à la notion de taille d'espace disponible en informatique. En psychologie, les catégories de base sont le meilleur compromis entre économie cognitive et contenu en information. En informatique, le théorème d'Occam montre qu'un algorithme capable de compresser très fortement l'information est un algorithme d'apprentissage. Dans les deux domaines, nous retrouvons l'idée que *toute compression très forte de l'information est apprentissage et que tout apprentissage est compression très forte de l'information*. Ceci va donc à l'encontre d'une représentation de la catégorie par exemplaire.
Un échantillon est le résultat de la rencontre d'un concept avec une distribution de probabilités. Une hypothèse qui comprime cet échantillon reflète donc les deux sources d'information. Ce qui nous amène à dire que l'apprenant apprend un concept et/avec sa distribution de probabilités. Il ne faut pas faire de faux sens sur cette dernière phrase. Ce que nous voulons dire c'est que l'hypothèse que l'apprenant fournit est un compromis entre le concept cible et la distribution de probabilités. Le concept est *apprenable* quelle que soit la distribution de probabilités mais l'hypothèse *apprise* dépend, elle, de cette distribution.

Dans le 3.2, nous avons parlé de la typicalité en psychologie. Nous avons expliqué comment les psychologues ont mis en évidence cette typicalité au travers d'une

comparaison des éléments d'une même catégorie : certains éléments sont plus typiques de la catégorie que d'autres. Cette typicalité les a amenés à rechercher d'autres schémas de représentation que ceux en conditions nécessaires et suffisantes car ces derniers ne peuvent rendre compte de ce phénomène. Parmi ceux-ci, nous avons surtout décrit le schéma de représentation que nous appelons modèle par prototype théorique[80] où l'individu stocke en mémoire une représentation de la catégorie qui est un résumé de celle-ci.
La typicalité d'un élément est fonction de sa similarité aux autres éléments de la catégorie et même de sa dissimilarité aux éléments des autres catégories. Autrement dit le calcul de la typicalité d'un élément se fait relativement à l'extension de la catégorie. Le calcul dont nous parlons ici est celui effectué par les psychologues pour comprendre la typicalité, il ne s'agit pas du processus interne de l'individu. Concernant ce processus interne, les versions des psychologues divergent selon qu'ils sont tenants d'une représentation par exemplaires ou d'une représentation par prototype. Les premiers considèrent qu'un individu «calcule » la typicalité d'un élément en évaluant sa similarité à tous les exemplaires de la catégorie stockés en mémoire. Les seconds considèrent que le «calcul » de la typicalité est fonction de sa similarité au prototype théorique qui contient les tendances centrales de la catégorie.
Il est supposé que les caractéristiques de ce prototype sont affectées d'un poids indiquant l'importance de cette caractéristique dans la représentation de la catégorie. Ainsi la couleur noire du merle sera une caractéristique affectée d'un poids assez grand. Certains chercheurs pensent que les caractéristiques sont corrélées, c'est-à-dire que ce n'est pas la caractéristique en elle-même qui est importante mais sa cooccurrence avec une autre caractéristique. Pour définir le merle, c'est la cooccurrence d'un bec jaune et d'un plumage noir qui peut être définitoire.

La fréquence des occurrences d'une caractéristique parmi les exemples de la catégorie intervient dans la définition de son poids. Dire que la fréquence des occurrences d'une caractéristique intervient dans le poids de celle-ci revient à dire que la fréquence des occurrences des *items supportant cette caractéristique* intervient dans son poids. *Autrement dit la typicalité est dépendante de la distribution de probabilité qui sous-tend les exemples que l'individu a rencontrés.*
On ne peut pas alors dire, d'un côté,
- que la fréquence des occurrences des caractéristiques intervient dans la définition du prototype théorique que se fait un individu d'une catégorie et que la typicalité attribuée par l'individu à un item est la distance qui sépare cet item de ce prototype théorique,
et, de l'autre,
- que la familiarité d'un individu avec les exemples de la catégorie n'intervient pas dans la typicalité que l'individu attribue à un item,

[80]Dans le champ de la recherche sur les représentations de l'individu, la polysémie fleurit (voir [Le Ny, 89]) et le terme de prototype théorique recouvre selon les auteurs des modèles de représentations différentes, alors que dans le même temps le modèle de représentation qu'il recouvre ici est nommé par d'autres termes ailleurs.

puisque de la familiarité de l'individu avec les items dépend la fréquence des occurrences dont dépend à son tour le prototype. C'est pourtant ce que fait Rosch [Rosch, Simpson et Miller, 76].

Ce que l'on peut, par contre, admettre c'est que la fréquence des occurrences d'une caractéristique n'est pas le seul déterminant du poids de cette caractéristique. La caractéristique «férocité » pour décrire un animal peut avoir un poids très fort pour un individu, non pas parce qu'il a rencontré beaucoup d'animaux féroces, mais parce que la rencontre avec un seul animal féroce peut l'avoir marqué très profondément[81].

Dans la section 3.3 nous avons présenté comment est envisagée la représentativité d'un exemple en informatique. Denis et Gilleron proposent de définir la représentativité d'un exemple relativement au concept par la complexité conditionnelle de Kolmogorov K(E|C) ou E est l'ensemble des exemples caractéristiques et C le concept. La complexité conditionnelle de Kolmogorov K(E|C) correspond à la taille du plus court programme qui recevant en entrée une description du concept permet d'engendrer les exemples. Pour que ces exemples apparaissent dans l'échantillon présenté à l'apprenant, ils définissent alors ce qu'ils appellent des distributions bienveillantes. Ces distributions garantissent que tout exemple de l'ensemble représentatif à une probabilité non nulle d'apparaître. De là, ils proposent un modèle d'apprentissage PAC avec distributions bienveillantes.
La complexité de Kolmogorov n'étant pas calculable, il est toujours possible de conserver l'idée que la typicalité d'un exemple est fonction de la taille du programme qui engendre cet exemple sur la base de la description du concept.

Quels parallèles peut-on établir entre la représentativité en informatique et la typicalité en psychologie ? Nous avons envie de dire que la représentativité en informatique intervient en amont de l'apprentissage pour déterminer l'échantillon d'apprentissage, c'est une liaison entre le concept cible et les exemples, tandis que la typicalité intervient en aval, c'est une liaison entre les exemples et l'hypothèse de l'apprenant. Nous pourrions caricaturer par : en informatique l'échantillon est représentatif du concept, en psychologie l'hypothèse est typique de l'échantillon.
Mais la représentativité comme la typicalité se rejoignent sur le fait que le 'programme' qui permet d'engendrer l'exemple à partir de la représentation de la catégorie est d'autant plus court que l'exemple est représentatif /typique de cette représentation. Dans le cas de la représentativité ceci est formalisé par la complexité conditionnelle de Kolmogorov[82], dans le cas de la typicalité cela est montré par diverses expérimentations et notamment celle où l'on observe que les sous-catégories les plus typiques sont les premières nommées. La représentativité et la typicalité diffèrent par contre sur l'individu

[81] Nous pourrions cependant encore rapporter ceci à la fréquence d'occurrences au travers de la compulsion de répétition. La compulsion de répétition est le phénomène par lequel un individu revit mentalement de nombreuses fois une scène qui l'a profondément marqué.
[82] Avec un bémol cependant car le programme le plus court n'est pas nécessairement le plus rapide. Voir [Delahaye, 96], pour l'opposition entre complexité de Kolmogorov et profondeur de Bennett.

qui supporte le programme, dans la représentativité c'est l'enseignant, dans la typicalité c'est l'apprenant.
Le second point sur lequel se rejoignent représentativité et typicalité sont les distributions de probabilités. Nous avons expliqué dans le 3.3 qu'un ensemble d'exemples représentatifs n'a de sens que si une distribution de probabilités est là qui garantisse son apparition dans l'échantillon, tandis que la typicalité permet, entre autres, de refléter la distribution de probabilités sous-jacente à l'échantillon qui a permis à l'apprenant de se former le concept.
En résumé, représentativité en informatique et typicalité en psychologie sont semblables mais en quelque sorte symétriques, cette symétrie les empêche de les assimiler l'un à l'autre. La question est de savoir si les deux ne sont pas liés, la représentativité en amont entraînant la typicalité en aval.

Une proposition qu'il est possible de faire mais qui demanderait pour être validée un travail que nous ne pouvons entreprendre ici est celle-ci : *en apprentissage naturel, un échantillon représentatif d'un concept est un échantillon tel que la similarité interne entre ses éléments est très grande. Cela implique que les fréquences de certains traits (ou de leurs cooccurrences) sont elles aussi très grandes. La typicalité refléterait alors ces fréquences.*
Pour caricaturer ceci, supposons les deux échantillons suivants définis sur quatre attributs. Le premier attribut étant l'attribut qui définit l'appartenance (l'étiquette : voir chapitre 2), il est donc toujours à 1.

1111	1000
1110	1111
1101	1010
1011	1101
Echantillon	Echantillon
représentatif	non représentatif

On constate que le premier échantillon a une plus grande similarité interne que le second car deux éléments quelconques du premier ont une similarité minimale de 2 (un minimum de deux valeurs communes), tandis que dans le second la similarité minimale est de 1 (l'étiquette). Tandis que dans le premier, il est facile de dégager un prototype (1111 en prenant les attributs de plus grande fréquence), ce n'est pas le cas dans le second.
Ainsi, avant l'apprentissage, un échantillon est représentatif car son degré de similarité interne est grand. A l'issue de l'apprentissage, un exemple est typique, car il est fortement similaire au prototype. Une forte similarité interne entraîne une plus grande fréquence de certains traits qui se retrouve elle-même exprimée dans la typicalité. A l'appui de ceci nous pouvons reprendre ce qui a été dit dans le 3.2.1.2, notamment la citation de F. Cordier, selon laquelle l'apprentissage est plus rapide lorsque l'on présente des sous-catégories typiques. On peut faire l'hypothèse qu'un échantillon d'apprentissage constitué de sous-catégories typiques présente une plus grande similarité interne.

Par ailleurs, en théories formelles de l'apprentissage, cette proposition invite à réfléchir sur une représentativité de l'échantillon qui ne se résumerait pas à la représentativité de chacun de ses exemples (ou leur somme) mais qui prendrait en compte la relation entre ses exemples.

Avant de clore ce chapitre, nous voudrions faire une remarque relative aux mécanismes de la catégorisation. Nous avons dit que la catégorisation était nécessaire pour que l'individu s'adapte à son environnement. Les capacités de l'individu ne sont pas illimitées, l'économie cognitive permet de rendre compte de cet aspect. Par ailleurs l'environnement n'est pas aléatoire. La typicalité, en tant que reflet des distributions de probabilités qui lui sont sous-jacentes, est un outil efficace, car elle maximise l'adaptation de l'individu à cet environnement. Plus un événement est probable plus la réponse de l'individu à cet événement est rapide et ceci grâce à la typicalité. Ainsi de la définition de la catégorisation en tant qu'adaptation découle deux caractéristiques : typicalité et économie cognitive.
Nous retrouvons sous une forme similaire ces deux caractéristiques dans la formalisation PAC[83] de l'apprentissage : cela le légitime. Nous pouvons aussi noter la cohérence du modèle PAC dans lequel, l'évaluation de l'erreur se fait en fonction de la distribution de probabilités utilisée pour l'apprentissage. En termes d'apprentissage naturel cela correspond au fait que le sujet apprend dans un environnement, pour survivre dans *cet* environnement.

[83]au sens large c'est-à-dire incluant les variantes du modèle standard

Chapitre 4

Modèle d'apprentissage PAC avec distributions bienveillantes

4 Introduction

Ce chapitre est un chapitre technique. Les principales définitions ont été présentées au chapitre 3 ainsi que leur « interprétation cognitive ». La lecture n'en est donc pas indispensable à un lecteur non familier des théories formelles. Néanmoins, un survol de ce chapitre pourra lui permettre de se faire une idée de la démarche adoptée par ces dernières.

Dans le modèle PAC standard de nombreuses classes de concepts sont démontrées non apprenables. Nous avons expliqué au chapitre 1 que cela peut être dû au fait que le modèle réclame que l'apprentissage se fasse selon toutes distributions de probabilités, même en tenant compte du fait que l'évaluation de l'apprentissage se fait avec ces

mêmes distributions. Par ailleurs, comme nous l'avons vu dans les chapitres 2 et 3, dans la plupart des situations d'apprentissage naturel, celui-ci ne se fait pas n'importe comment. Plus les exemples servant à l'apprentissage sont typiques et plus l'apprentissage est rapide. Les parents ne parlent pas à leur enfant en utilisant des phrases à la syntaxe alambiquée ou en utilisant un vocabulaire abscons. L'enseignant choisit ses exemples en fonction du concept qu'il veut faire apprendre.

Autrement dit, à chaque concept correspondent des exemples plus ou moins représentatifs. Cette notion d'ensemble caractéristique a été étudiée en théories formelles par différents auteurs mais plutôt dans le cadre du modèle de Gold d'identification à la limite (voir [Gold, 78], [Goldman et Kearns, 95], [Goldman et Mathias, 96] [Shinohara et Miyano, 91]...). Introduire dans le cadre du modèle de Valiant cet ensemble d'enseignement, nécessite d'avoir la garantie qu'il apparaisse dans l'échantillon présenté à l'apprenant. C'est ce qui est à la base de la définition du modèle d'apprentissage PAC avec distributions bienveillantes. *Une distribution bienveillante est une distribution telle que tout élément de l'ensemble d'enseignement a une probabilité non nulle d'apparaître.*

Cependant, si l'on restreint la classe des distributions, il peut y avoir *collusion* entre enseignant et apprenant. Un exemple de collusion possible consiste à coder au travers des exemples le concept à apprendre, l'apprenant n'ayant plus qu'à décoder les exemples pour trouver le concept cible. Ainsi, si nous supposons que les concepts peuvent être ordonnés, il suffirait de ne donner qu'un échantillon dont le cardinal correspondrait au numéro d'ordre du concept. L'apprenant n'aurait plus alors qu'à compter les exemples pour retourner le concept adéquat. Aussi, pour éviter cette possible collusion, le modèle réclame que l'apprentissage ait lieu pour *toutes* distributions bienveillantes.

Le fait d'introduire un ensemble d'enseignement et des distributions de probabilités bienveillantes dans le modèle oblige dans le même temps à reconsidérer le temps d'apprentissage. En effet, on suppose qu'il n'y aura apprentissage que si l'apprenant rencontre tous les éléments de l'ensemble d'enseignement, même celui dont la probabilité est la plus petite. Il est donc normal que le temps d'apprentissage dépende de la probabilité de ce dernier.

Dans ce chapitre nous présentons le modèle d'apprentissage avec distributions bienveillantes, nous y démontrons un théorème d'Occam et sa réciproque et prouvons que les listes de décisions sont apprenables dans ce modèle. Puis nous établissons des correspondances avec d'autres modèles ce qui permet de récupérer les résultats obtenus dans ces modèles.

4.1 Modèle d'apprentissage PAC avec distributions bienveillantes

4.1.1 Définitions et notations

Les définitions et notations que nous donnons ci-dessous sont légèrement différentes de celles que nous avons utilisées jusqu'à présent. En effet, pour simplifier la lecture ultérieure nous identifions la classe de concepts à sa classe de représentations.

Soit B_n l'ensemble des fonctions booléennes de l'ensemble $X_n = \{0,1\}^n$ dans $\{0,1\}$ et soit B l'ensemble défini par $B = \cup_{n\geq 1} B_n$. Une classe F de fonctions booléennes est un sous-ensemble de B. Un schéma de représentations pour une classe de fonctions booléennes F est une fonction $R : F \rightarrow 2^{\Sigma^*}$ où Σ est un alphabet fini telle que pour toutes fonctions f et f' dans F, $R(f)$ soit non vide et si $f \neq f'$ alors $R(f) \cap R(f') = \varnothing$. Nous supposons que R est calculable en temps polynomial, c'est-à-dire qu'il existe un algorithme déterministe polynomial qui, avec en entrée deux mots x et c, retourne 1 si $f(x)=1$ lorsque $c \in R(f)$, et 0 sinon. Une *classe de concepts* C est définie comme un ensemble de représentations d'un ensemble de fonctions booléennes, i.e. $C = \cup_{f\in F} R(f)$. La *taille* $|c|$ d'un concept c est sa longueur. Nous supposons que $|c| \geq n$. Dans la suite, nous identifions un concept c de C avec la fonction f qu'il représente.
Soit C une classe de concepts, un *exemple* d'un concept c est un couple (x,c(x)) dans lequel x appartient au domaine de c. Un exemple (x,c(x)) est *positif* si c(x)=1 et *négatif* sinon. Nous notons EX(c) l'ensemble des exemples d'un concept c. Un *échantillon* de c est un sous-ensemble de EX(c). Soit c un concept sur X_n, soit P une distribution de probabilités quelconque sur X_n, EX(c, P) est une procédure qui à chaque appel, retourne un exemple (x,c(x)) en une unité de temps, où x est tiré selon la distribution de probabilités P (les tirages sont supposés indépendants). Un concept cible c étant choisi, pour toute hypothèse h, l'erreur de h est définie par : $erreur(h) = P(\{x \in X_n \mid c(x) \neq h(x)\})$.

4.1.2 Définition du modèle [84]

4.1.2.1 Définition 1 : Enseignant, ensemble d'enseignement

Soit C une classe de concepts. Un ensemble d'enseignement pour c de C est un échantillon de c. Un enseignant pour C est une fonction T qui associe à chaque concept c, un ensemble d'enseignement T(c), et telle qu'il existe une constante k telle que pour tout concept de c, $Card(T(c)) \leq |c|^k$. Un enseignant est calculable s'il existe un algorithme qui prend en entrée un concept c de C et retourne un ensemble d'enseignement T(c).

Cette première définition modélise un enseignant qui a pour tâche de choisir un échantillon d'exemples en fonction de la cible. Il est réclamé que cet ensemble

[84] Nous reprenons ici, les définitions déjà proposées au chapitre 3 afin d'éviter au lecteur un aller retour fastidieux. Néanmoins nous ne reprenons plus les « justifications cognitives » de ces définitions.

d'enseignement soit polynomial en la taille de la cible, c'est-à-dire qu'il soit d'une taille réaliste. De la même manière, l'enseignant doit être calculable.

4.1.2.2 Définition 2 : Distribution bienveillante

Soit C une classe de concepts et T un enseignant pour C. Soit c un concept cible sur X_n et P une distribution de probabilité donnée sur X_n.
On définit

$$P_{min}(c) = \begin{cases} \min\{P(x) \mid (x, c(x)) \in T(c)\} & \text{si } T(c) \neq \varnothing \\ 1 & \text{sinon} \end{cases}$$

Une distribution P est bienveillante pour c et T si $P_{min}(c) \neq 0$.

Une distribution de probabilités est bienveillante à partir du moment où elle garantit que l'ensemble d'enseignement apparaîtra dans l'échantillon qui sera soumis à l'apprenant. Pour cela, la distribution ne sera dite bienveillante que si tous les éléments de l'ensemble d'enseignement ont une probabilité non nulle d'être présentés.

4.1.2.3 Définition 3 : Apprentissage PAC avec distributions bienveillantes

Soit C une classe de concepts, H une classe de représentations et T un enseignant pour C. On dit que C est PAC-apprenable ***avec distributions bienveillantes*** *s'il existe un algorithme L avec la propriété suivante : pour tout concept* $c \in C$, *pour toute* ***distribution de probabilités bienveillante P*** *sur* X, *et pour tout* $0<\varepsilon<1/2$ *et* $0<\delta<1/2$ *si L a accès à EX(c,P) et aux entrées* ε, δ, *et taille de c alors avec une probabilité d'au moins* $1-\delta$, *L retourne une hypothèse* $h \in H$ *satisfaisant erreur(h)*$\leq\varepsilon$ *et L tourne en temps polynomial en taille(c),* $1/\varepsilon$, $1/\delta$ *et* $1/P_{min}(c)$.

Par rapport au modèle PAC standard, le modèle ici restreint les distributions de probabilités aux distributions bienveillantes. Par ailleurs, il demande, en plus des paramètres usuels, que le temps soit polynomial en $1/P_{min}(c)$, ce qui fait dépendre ce temps de la plus petite probabilité d'un élément caractéristique. De façon évidente ce modèle généralise le modèle PAC classique car une classe est PAC apprenable si et seulement si elle est apprenable avec un ensemble d'enseignement vide.

Remarque : en demandant que le temps d'apprentissage soit polynomial en $1/P_{min}(c)$ nous pouvons en arriver à un temps d'apprentissage exponentiel en n. Ce sera notamment le cas avec la distribution uniforme où $1/P_{min}(c)$ sera égal à 2^n et le temps d'apprentissage sera alors polynomial en 2^n et donc exponentiel en n.

4.1.2.4 Définition 4 : Apprentissage PAC avec distributions bienveillantes en temps usuellement polynomial

Pour énoncer un théorème d'Occam, il est nécessaire de définir également la notion d'algorithme en temps *usuellement* polynomial

Soit C une classe de concepts et T un enseignant pour C. C est PAC apprenable en temps usuellement polynomial avec des distributions bienveillantes s'il y a un algorithme A de PAC apprentissage avec distributions bienveillantes pour C tel qu'avec une probabilité d'au moins 1-δ, A s'arrête en temps polynomial en taille(c), 1/ε, 1/δ, et $1/P_{min}(c)$

Comme dans le modèle PAC classique, si *A* est un algorithme d'apprentissage en temps usuellement polynomial pour l'enseignant T qui prend en entrée taille(c), ε, δ alors il existe un algorithme d'apprentissage en temps usuellement polynomial qui prend en entrée ε et δ seulement.

4.1.3 Un théorème d'Occam pour l'apprentissage PAC avec distributions bienveillantes

Un algorithme d'Occam est basé sur l'heuristique selon laquelle une hypothèse courte consistante avec un échantillon est probablement une bonne hypothèse. Un théorème d'Occam démontre que s'il existe un algorithme d'Occam alors celui-ci peut-être transformé en un algorithme d'apprentissage (voir chapitre 3).

Nous proposons la définition d'un algorithme d'Occam et démontrons que si un tel algorithme existe pour une classe de classe concepts C avec un enseignant T alors cette classe de concept est PAC apprenable avec distributions bienveillantes.

4.1.3.1 Définition 5 : un multi-échantillon

Pour lever toute ambiguïté, nous définissons ici un multi-échantillon. Bien souvent dans la littérature sur le domaine, un échantillon est défini comme un ensemble. Ceci peut entraîner certaines confusions car dans le même temps on suppose qu'un même élément peut être représenté plusieurs fois dans cet échantillon, ce qui est contraire à la définition mathématique d'un ensemble

Soient C une classe de concepts et c un concept de C, un multi-échantillon de c est un multi-ensemble d'exemples de c, i.e. une application de EX(c) dans $\aleph$

Soit un multi-échantillon S d'un concept c de C, soit (x,c(x)) un exemple dans EX(c), S(x,c(x)) est le nombre d'occurrences de (x,c(x)) dans S. Un échantillon S_c de c est

inclus dans un multi-échantillon S de c si, pour tout (x,c(x)) dans S_c, S(x,c(x)) ≥1. Le cardinal d'un multi-échantillon S est défini par

$$Card(S) = \sum_{(x,c(x))\in EX(c)} S(x,c(x))$$

4.1.3.2 Définition 6 : fréquence minimale d'apparition

Soit C une classe de concepts et T un enseignant pour C. Soit S un multi-échantillon non vide d'un concept c de C. On définit :

$$f_{min}(S,c) = \begin{cases} \min\{\dfrac{S(x,c(x))}{Card(S)} \mid (x,c(x)) \in T(c)\} & \text{si } T(c) \neq \varnothing \\ 1 & \text{sinon} \end{cases}$$

où S(x,c(x)) est le nombre d'occurrences de (x,c(x)) dans S.

$f_{min}(S,c)$ est la fréquence de l'élément de l'ensemble d'enseignement qui est le moins fréquent dans l'échantillon. Si P est la distribution de probabilités uniforme sur S alors $P_{min}(c) = f_{min}(S,c)$. Par ailleurs, lorsque $T(c) \subseteq S$ alors $1/f_{min}(S,c) \leq Card(S)$.

4.1.3.3 Définition 7 : algorithme d'Occam pour l'apprentissage PAC avec distributions bienveillantes

Soit C une classe de concepts et T un enseignant pour C, B est un algorithme d'Occam pour C,T, s'il existe des constantes $a\geq 0$, $b\geq 0$ et $0\leq\alpha<1$ telles qu'avec en entrée le multiéchantillon S de c dans X_n tel que $T(c)\subseteq S$, B retourne une hypothèse h telle que :
-h est consistante avec S,
-$|h| \leq a(|c|/f_{min}(S,c))^b (Card(S))^\alpha$
-B tourne en temps polynomial en |c|, Card(S)

Par rapport à l'algorithme d'Occam standard du modèle PAC classique il y a deux différences. La première est que l'on réclame que l'ensemble d'enseignement fasse partie du multi-échantillon, la seconde est que l'on demande que la taille de l'hypothèse retournée soit courte par rapport à la taille de la cible et la taille de l'échantillon (comme dans le modèle classique) mais aussi qu'elle soit courte si $1/f_{min}(S,c)$ n'est pas trop petit. Autrement dit, plus les exemples de l'ensemble d'enseignement seront bien représentés dans le multi-échantillon et plus l'hypothèse sera courte. Si l'on suppose que le multi-échantillon est tiré selon une distribution de probabilités bienveillante P, ceci se produira dès que la probabilité minimale d'apparition d'un exemple caractéristique ($P_{min}(c)$) n'est pas trop petite.

4.1.3.4 Théorème 1 : théorème d'Occam pour l'apprentissage PAC avec distributions bienveillantes

Soit C une classe de concepts et T un enseignant pour C. S'il existe un algorithme d'Occam pour C,T alors C est PAC apprenable avec distributions bienveillantes en temps usuellement polynomial.

Preuve du Théorème d'Occam pour l'apprentissage PAC avec distributions bienveillantes

Soit C une classe de concepts et T un enseignant pour cette classe. Soit k une constante telle que pour tout concept c, Card(T(c)) ≤ ($|c|^k$). Soit B un algorithme d'Occam pour C, T avec les constantes (a, b, α). Soit c un concept de C. Soit q le polynôme tel que B est d'une complexité en temps de q(|c|, Card(S)). Soit EX(c, P) un oracle tel que P est une distribution de probabilités bienveillante pour c et T.

Nous prouvons d'abord quelques lemmes techniques.

Le premier lemme établit que, pour un multi-échantillon suffisamment large, nous pouvons estimer le paramètre $P_{min}(c)$ et avoir ainsi la garantie (supérieure à 1-δ) que des exemples rares de l'ensemble d'enseignement font bien partie de l'échantillon.

Lemme 1
Soit p un entier tel que $p \geq 1/P_{min}(c)$. $N_1(\delta,|c|,p)=\lceil 8\, p\, \log(|c|^k/\delta) \rceil$ appels à EX(c,P). Ceci définit un multi-échantillon S. Alors la probabilité que $f_{min}(S,c) \geq P_{min}(c)/2$ est au moins de 1-δ.

La preuve est basée sur les bornes de Chernoff.

Théorème 2 des bornes de Chernoff (voir [Kearns et Vazirani, 1994] p190-192)
Soit $X_1,\ldots,X_m$ une séquence de m tirages indépendants de Bernouilli, chaque tirage ayant une probabilité de succès de $E[X_i] = \phi$. Soit $S=X_1+\ldots+X_m$ la variable aléatoire indiquant le nombre total de succès, ainsi $E(S) = \phi m$, alors pour $0 \leq \gamma \leq 1$ nous obtenons les bornes suivantes :

$$\text{borne 1 : } Pr[S > (1+\gamma)\phi m] \leq e^{-m\phi\gamma^2/3}$$
$$\text{borne 2 : } Pr[S < (1-\gamma)\phi m] \leq e^{-m\phi\gamma^2/2}$$

Dans ce qui suit nous utilisons la borne 2. m étant positif elle équivaut à

$$\text{borne 2 : } Pr[S/m < (1-\gamma)\phi] \leq e^{-m\phi\gamma^2/2}$$

si nous prenons $\gamma=1/2$ nous obtenons

$$\text{borne 2 : } Pr[S/m < \phi/2] \leq e^{-m\phi/8}$$

Preuve du lemme 1 :
Pour avoir : $f_{min}(S,c) \geq P_{min}(c)/2$

de par la définition de $f_{min}(S,c)$ et de $P_{min}(c)$, il suffit que :

$$\forall x \in T(c), f(x,S) \geq P(x)/2$$

où f(x,S) est la fréquence de x dans S.

en utilisant la borne 2 nous avons

$$\forall x \in T(c), Pr[f(x,S) < P(x)/2] \leq e^{-mP(x)/8}$$

sachant que, de par la définition de $P_{min}(c)$, $P(x) > P_{min}(c)$ et comme $P_{min}(c) > 1/p$, nous avons $P(x) > 1/p$ et nous obtenons donc

$$\forall x \in T(c), Pr[f(x,S) < P(x)/2] \leq e^{-m/8p}$$

comme $Card(T(c)) \leq |c|^k$ nous obtenons

$$Pr[\forall x \in T(c), f(x,S) < P(x)/2] \leq |c|^k e^{-m/8p}$$

pour qu'une telle probabilité soit inférieure à δ, il suffit que :

$$|c|^k e^{-m/8p} < \delta$$

en prenant le logarithme népérien nous obtenons

$$\ln(|c|^k e^{-m/8p}) < \ln(\delta)$$

$$\ln(|c|^k) + \ln(e^{-m/8p}) < \ln(\delta)$$

$$\ln(|c|^k) - m/8p < \ln(\delta)$$

$$-m < 8p(\ln(\delta) - \ln(|c|^k))$$

$$m \geq 8p(\ln(|c|^k) - \ln(\delta))$$

$$m \geq 8p \ln(|c|^k / \delta)$$

sachant que log(a)>ln(a), pour que la précédente inégalité soit satisfaite, il suffit que

$$m \geq 8p \log(|c|^k / \delta)^*$$

Fin de la preuve du lemme 1

Lemme 2

$N_2(\varepsilon,\delta,|c|,p)=\lceil((\log(1/\delta) + a(2p|c|)^b + 1) /\varepsilon)^{1/1-\alpha}\rceil$ appels à EX(c,P). Ceci définit un multi-échantillon S. Supposons que $f_{min}(S,c)\geq 1/2p$. Soit h l'hypothèse retournée par l'algorithme d'Occam B sur l'entrée S. Alors la probabilité que erreur(h)>ε est d'au plus δ.

Ce lemme garantit (avec une probabilité supérieure à 1-δ) que l'hypothèse fournie par l'algorithme d'Occam est une hypothèse approximativement correcte si le nombre d'exemples de l'échantillon est suffisamment grand.

Preuve du lemme 2

Dans la démonstration nous allons proposer un ensemble d'hypothèses courtes et erronées (erreur(h)>ε) et montrer qu'un algorithme d'Occam ne peut retourner une hypothèse de cet ensemble qu'avec une probabilité inférieure à δ si l'échantillon est suffisamment grand. Autrement dit le fait que l'algorithme d'Occam retourne une hypothèse courte implique qu'il est peu probable (<δ) que cette hypothèse ne soit pas approximativement correcte à partir du moment où l'échantillon est suffisamment grand.

Soit S un multi-échantillon de cardinalité N et considérons l'ensemble

$$H_c= \{h\in C \mid \text{erreur}(h)>\varepsilon \text{ et } |h| \leq a(2p|c|)^b N^\alpha\}$$

Soit $h\in H_c$. La probabilité qu'un appel à EX(c,P) retourne un exemple consistant avec h est inférieure à 1-ε. De là, la probabilité que N appels à EX(c,P) retourne un multi-échantillon S consistant avec h est inférieure à $(1-\varepsilon)^N$. De plus le cardinal de H est inférieur à $2^{a(2p|c|)^b N^\alpha +1}$. Alors la probabilité que le multi-échantillon S de cardinalité N soit consistant avec le concept h de H est inférieure à $2^{a(2p|c|)^b N^\alpha +1}(1-\varepsilon)^N$. On peut vérifier que $2^{a(2p|c|)^b N^\alpha +1}(1-\varepsilon)^N \leq\delta$ si $N\geq N_2(\varepsilon, \delta,|c|,p)$ ainsi :

Nous voulons obtenir :

$$2^{N^\alpha a(2p|c|)^b + 1}(1-\varepsilon)^N \leq \delta$$

en prenant le log nous obtenons

$$N^\alpha a(2p|c|)^b+ 1+ \log((1-\varepsilon)^N) \leq \log(\delta)$$

$$\log((1-\varepsilon)^N)+N^\alpha a(2p|c|)^b+ 1 \leq \log(\delta)$$

$$N\log(1-\varepsilon)+N^\alpha a(2p|c|)^b+ 1 \leq \log(\delta)$$

$$\log(\delta) \geq N\log(1-\varepsilon)+N^\alpha a(2p|c|)^b+ 1$$

$$-\log(1/\delta) \geq N\log(1-\varepsilon)+N^\alpha a(2p|c|)^b+1$$

$$-N\log(1-\varepsilon) \geq \log(1/\delta) + N^{\alpha} a(2p|c|)^{b} + 1$$

sachant que $\alpha \geq 0$ et que N est un entier naturel nous obtenons $N^{\alpha} \geq 1$ et donc l'inégalité précédente est vérifiée si

$$-N\log(1-\varepsilon) \geq N^{\alpha} \log(1/\delta) + N^{\alpha} a(2p|c|)^{b} + N^{\alpha}$$

$$\text{inégalité 1 } -N\log(1-\varepsilon) \geq N^{\alpha} (\log(1/\delta) + a(2p|c|)^{b} + 1)$$

Sachant que $\ln(1+x) \leq x$
nous obtenons $\ln(1-\varepsilon) \leq -\varepsilon$

$$\ln(1-\varepsilon)/\ln(2) \leq -\varepsilon / \ln(2)$$
$$\log(1-\varepsilon) \leq -\varepsilon / \ln(2)$$
$$-\log(1-\varepsilon) > \varepsilon / \ln(2))$$
$$-\log(1-\varepsilon) \geq \varepsilon / \ln(2)$$
$$-N\log(1-\varepsilon) \geq N\varepsilon / \ln(2)$$

l'inégalité 1 est alors satisfaite si :

$$N\varepsilon/\ln(2) \geq N^{\alpha} (\log(1/\delta) + a(2p|c|)^{b} + 1)$$

sachant que $\ln(2) \leq 1$ l'inégalité précédente est satisfaite si :

$$N\varepsilon \geq N^{\alpha} (\log(1/\delta) + a(2p|c|)^{b} + 1)$$

$$N\varepsilon \geq N^{\alpha} (\log(1/\delta) + a(2p|c|)^{b} + 1)$$

$$N \geq N^{\alpha} ((\log(1/\delta) + a(2p|c|)^{b} + 1)/\varepsilon)$$

$$N^{1-\alpha} \geq (\log(1/\delta) + a(2p|c|)^{b} + 1)/\varepsilon$$

$$N \geq ((\log(1/\delta) + a(2p|c|)^{b} + 1)/\varepsilon)^{1/1-\alpha}$$

$$N = \lceil ((\log(1/\delta) + a(2p|c|)^{b} + 1)/\varepsilon)^{1/1-\alpha} \rceil$$

Supposons maintenant que $N \geq N_2(\varepsilon, \delta, |c|, p)$ et que $f_{min}(S,c) \geq 1/2p$. $f_{min}(S,c) \geq 0$, ainsi $T(c) \subseteq S$. Sur l'entrée S l'algorithme B retourne une hypothèse h consistante avec S telle que :

$$|h| \leq a(|c|/ f_{min}(S,c))^{b} (\text{Card}(S))^{\alpha} \leq a(2p|c|)^{b} N^{\alpha}$$

Finalement si erreur(h)>ε, $h \in H_c$.

Fin de la preuve du lemme 2.

L'algorithme TEST

Dans le domaine de l'apprentissage PAC, on trouve un algorithme TEST qui teste si une hypothèse est une bonne approximation du concept cible. Cet algorithme TEST, qui prend en paramètres ε, δ, i et h (voir [Haussler, Kearns, Littlestone, et Warmuth, 91]), fait $\lceil (32/\varepsilon)(i\ \ln(2)+\ln(2/\delta)) \rceil$ appels à EX(c, P) pour tester l'hypothèse h. Il accepte l'hypothèse si celle-ci ne se trompe au plus que sur 3ε/4 des exemples retournés par l'oracle et la rejette autrement. TEST(ε, δ, i , h) est polynomial en 1/ε, 1/δ, i et |h|.

Lemme 3 [Haussler, Kearns, Littlestone, et Warmuth, 88]
Le test TEST(ε, δ, i , h) a la propriété que :
-si erreur(h) ≥ ε, la probabilité est au plus de $\delta / 2^{i+1}$ *que le test accepte h,*
-si erreur(h) ≤ ε /2, la probabilité est au plus de $\delta / 2^{i+1}$ *que le test rejette h.*

Ainsi TEST peut accepter une mauvaise hypothèse ou en rejeter une bonne mais avec une probabilité bien inférieure à δ.

L'algorithme d'apprentissage PAC avec distributions bienveillantes pour C et T

L'algorithme d'apprentissage PAC avec distributions bienveillantes pour C et T, basé sur l'algorithme d'Occam B est alors :

1. **Algorithme A d'apprentissage avec distributions bienveillantes pour C,T**
2. **Entrée** ε,δ,|c|
3. **Début**
   4. soit S = ∅ /* S est le multi-échantillon*/
   5. soit p = 1 /*p est l'hypothèse courante pour $1/P_{min}(c)$*/
   6. **boucle**
      7. soit $N = \sup\{ \lceil 8p \log(|c|^k/\delta) \rceil , \lceil ((\log(1/\delta)+a(2pl)^b+1)/\varepsilon)^{1/1-\alpha} \rceil \}$
      8. faire N appel à Ex(c,P)
      9. faire tourner au plus q(|c|,N,2p) pas de B sur l'entrée S
      10. **si** B retourne une hypothèse h **et** h consistante avec S
      11. **et** $|h| \leq a(p|c|)^b N^\alpha$ **et** TEST(ε, δ/3, p, h)
          12. **alors** retourner h et s'arrêter
      13. **fin de si**
      14. soit p = p+1
   15. **fin de boucle**
16. **fin**

Avant de présenter la preuve l'algorithme d'apprentissage, il est utile de rappeler que $P_{min}(c)$, n'est pas connu et qu'il est donc nécessaire d'approximer $1/ P_{min}(c)$ c'est le rôle de la boucle basée sur p (lignes 5, 6, 14 et 15). De ce fait nous n'avons aucune garantie que la taille de l'échantillon (lignes 7 et 8) soit suffisamment grande pour

que l'ensemble d'enseignement en fasse partie et donc nous n'avons aucune garantie que l'algorithme fonctionne correctement (lignes 10 et 11) ni que l'hypothèse qu'il fournit soit approximativement correcte. C'est pour cela que nous devons utiliser un algorithme TEST (ligne 11) qui vérifie cette hypothèse. Comme cet algorithme TEST ne fonctionne correctement qu'avec une certaine probabilité, nous obtenons un apprentissage PAC en temps usuellement polynomial.

Preuve que A est un algorithme d'apprentissage avec distributions bienveillantes pour C et T.

Pour démontrer que l'algorithme est un algorithme d'apprentissage en temps usuellement polynomial, il faut montrer
1)qu'il s'arrête en temps polynomial avec une probabilité inférieur ou égale à 1-δ,
2)que lorsqu'il s'arrête, la probabilité que erreur(h) ≥ε est au plus de δ

L'algorithme s'arrête avec une probabilité inférieur ou égale à 1-δ

Il faut prouver, qu'avec une probabilité de 1-δ, A s'arrête en temps polynomial en 1/ε, 1/δ, |c| et $1/P_{min}(c)$.
Soit $p = \lceil 1/P_{min}(c) \rceil$ et $N \geq N_1$ (δ/3, |c|, p). Alors la probabilité que $f_{min}(S,c) \geq P_{min}(c)/2$ est au moins de 1-δ/3 (lemme 1).
Supposons que $f_{min}(S,c) \geq P_{min}(c)/2$ et $N \geq N_2(\varepsilon/2, \delta/3, |c|, p)$, alors la probabilité que erreur(h) > ε /2 est au plus de δ/3 (Lemme 2).
Supposons maintenant que erreur(h) ≤ ε /2, la probabilité que le test TEST(ε, δ, i , h) rejette h est au plus $\delta/(3 \times 2^{p+1})$ (Lemme 3).
Donc la probabilité que l'algorithme d'apprentissage ne s'arrête pas à l'étape p égal ou plus petit que $\lceil 1/P_{min}(c) \rceil$ est au plus de δ.
On peut aisément vérifier que si l'algorithme A s'arrête avant $p = \lceil 1/P_{min}(c) \rceil$, alors le temps d'exécution est borné par un polynôme en 1/ε, 1/δ, |c| et $1/P_{min}(c)$.

Lorsque l'algorithme s'arrête, la probabilité que erreur(h) ≥ε est au plus de δ

Quand l'algorithme A s'arrête à une étape p, la probabilité que erreur(h)≥ε est au plus $\delta / (3 \times 2^{p+1})$. Ceci est dû à la condition d'arrêt et au lemme 3. Aussi, quand l'algorithme d'apprentissage A s'arrête la probabilité que erreur(h) ≥ε est au plus de δ/3.

4.1.3.5 Réciproque du théorème d'Occam

Comme dans le modèle PAC classique, il existe une réciproque pour le théorème d'Occam pour *les classes de concepts fortement fermées sous exception.* Une classe de concepts est *fermée sous exception* si en incorporant les exceptions d'un ensemble fini dans un concept de la classe, on obtient encore un concept de la classe. Cette propriété

est intéressante à partir du moment où elle est calculable. Cela conduit à la notion de forte fermeture sous exception, que nous définissons formellement de la façon suivante.

***Définition 8** [Board et Pitt, 90] [Natarajan, 91]*
Une classe de concepts C est fortement fermée sous exception s'il existe un algorithme F et des constantes α et β tels que :
(1)F prend en entrée un concept c de C et un échantillon fini S de c et retourne un concept c' de C tel que pour tout $(x,c(x))\in S$, $c'(x)\neq c(x)$ et pour tout $(x,c'(x))\notin S$, $c'(x)=c(x)$,
(2)$|c'| \leq \alpha(|c|+|S|) \log(|c|+|S|)+\beta$
(3) F tourne en temps polynomial en $|c|$ et $|S|$

Ainsi, la classe des DNF est fermée sous exception car lorsque l'on incorpore un échantillon quelconque à une DNF quelconque avec on obtient encore une DNF.

Théorème 3 : réciproque du théorème d'Occam

Soit C une classe de concepts et T un enseignant pour C. Si C est PAC apprenable avec distributions bienveillantes en temps usuellement polynomial et si C est fortement fermée sous exception alors il existe un algorithme aléatoire B tel qu'avec les entrée δ et un multi-échantillon S de c tel que $T(c) \subseteq S$, alors avec une probabilité d'au moins 1-δ, B retourne une hypothèse h telle que :
- h est consistante avec S,
- il existe des constantes $a\geq 0$ et $b\geq 0$ et $\alpha<1$ qui ne dépendent ni de S ni de c et telles que $|h| \leq a(|c|/f_{min}(S,c))^b |S|^{\alpha}$
- B tourne en temps polynomial en $\log(1/\delta)$, Card(S) et $|c|$

Preuve
Soit S un multi-échantillon de c tel que $T(c)\subseteq S$. Soit P la distribution uniforme sur S [85]. Rappelons qu'alors $P_{min}(c) = f_{min}(S,c)$. Soit F un algorithme qui calcule la fermeture sous exception de la classe C. Soit A un algorithme d'apprentissage PAC avec distributions bienveillantes qui prend en entrée ε et δ. Il existe une constante k tel que A s'arrête en temps polynomial $(|c|/ (\varepsilon\delta P_{min}(c))^k$ avec une probabilité d'au moins 1-δ. L'algorithme B est le suivant

1. **Algorithme d'Occam aléatoire B pour C et T**
2. **Entrées** : δ, S multi-échantillon
3. **Début**
   4. ε reçoit $|S|^{-1/(k+1)}$
   5. exécuter en parallèle $\lceil \log(1/\delta) \rceil$ fois l'algorithme $A(\varepsilon,1/4)$ et s'arrêter dès que l'un des algorithmes A retourne une hypothèse h telle que
      $$|\{x|\ (x,c(x))\in S,\ c(x)\neq h(x)\}| \leq \varepsilon |S|$$

[85] Supposons que S=(a,a,a,b,b) alors P(a)=3/5 et P(b)=2/5

/* chacun des algorithmes A demande des exemples qui sont tirés dans le multi-échantillon S selon la distribution P */

6. retourner F(h, {x| (x,c(x))∈S, c(x)≠h(x)}) et arrêter
7. **fin**

Avant de montrer que B est bien un algorithme d'Occam nous voudrions faire quelques commentaires. En ligne 2, δ est le paramètre qui indique la probabilité de succès de l'algorithme d'Occam B et non celui de l'algorithme A qui est fixé, lui, à ¼. En ligne 5 on propose d'exécuter en parallèle plusieurs fois l'algorithme A. Chacune des exécutions en parallèle de A demande des exemples à l'oracle qui les tire dans le multi-échantillon S selon la distribution P. S est le multi-échantillon de l'algorithme B et non celui de A : l'oracle de A puise dans S selon la probabilité P des exemples étiquetés. On passe à la ligne 6 dès que l'une des exécutions de l'algorithme s'arrête en retournant une hypothèse acceptable. En ligne 6, on garantit grâce à l'algorithme de fermeture F que l'hypothèse retournée soit bien consistante avec l'échantillon.

Quand A prend en entrée ε et ¼ , A s'arrête en un temps plus petit que $t=(4|c|/(\varepsilon P_{min}(c))^k$ avec une probabilité de ¾ parce que A est un algorithme d'apprentissage PAC en temps usuellement polynomial avec distributions bienveillantes.
P est la distribution uniforme sur S, aussi quand A prend en entrée ε et 1/4, A retourne un concept h tel que le nombre d'exceptions de $h = |\{x| (x,c(x))\in S, c(x)\neq h(x)\}| \leq \varepsilon|S|$ avec une probabilité d'au moins ¾.
Une exécution de A(ε,1/4) *s'arrête dans un temps inférieur à t et retourne un concept h* tel que le nombre d'exceptions de $h \leq \varepsilon|S|$ avec une probabilité d'au moins ½.
La probabilité qu'aucune des $\lceil \log(1/\delta) \rceil$ exécutions de A(ε,1/4) s'arrête dans un temps inférieur à t et retourne un concept h tel que le nombre d'exceptions de $h \leq \varepsilon|S|$ est inférieure à $(1/2)^{\lceil \log(1/\delta) \rceil} \leq \delta$.
Conséquemment, l'algorithme calcule une hypothèse h telle que le nombre d'exceptions de $h \leq \varepsilon|S|$ avec une probabilité d'au moins 1-δ.

Il faut maintenant incorporer les exceptions du concept h, c'est-à-dire calculer c' = F(h, {x| (x,c(x))∈S, c(x)≠h(x)}). En utilisant la propriété de forte fermeture sous exceptions nous obtenons :

$$|c'| \leq \alpha(|h|+\varepsilon|S|) \log(|h|+\varepsilon|S|)+\beta$$

Il existe a tel que :

$$|c'| \leq a(|h|+\varepsilon|S|)^{(1+1/2k)}$$

La taille de h est bornée par le temps de calcul de h. En effet le temps de calcul de l'algorithme ne peut être plus petit que la taille de l'hypothèse qu'il retourne. Et comme $\varepsilon = |S|^{-1/(k+1)}$ nous obtenons :

$$|c'| \leq a((\frac{4|c|}{\varepsilon\ P_{min}(c)})^k + |S|^{k/k+1})^{(1+1/2k)}$$

$$|c'| \leq a|S|^{k/k+1}((\frac{4|c|}{\varepsilon\ P_{min}(c)})^k |S|^{-k/k+1} + 1)^{(1+1/2k)}$$

$$|c'| \leq a|S|^{k/k+1}((\frac{4|c|}{|S|^{-1/k+1} P_{min}(c)})^k |S|^{-k/k+1} + 1)^{(1+1/2k)}$$

$$|c'| \leq a|S|^{k/k+1}((\frac{4|c||S|^{-1/k+1}}{|S|^{-1/k+1} P_{min}(c)})^k + 1)^{(1+1/2k)}$$

$$|c'| \leq a|S|^{(k/k+1)\times(1+1/2k)}(1+(\frac{4|c|}{P_{min}(c)})^k)^{(1+1/2k)}$$

Finalement, notons que (k/k+1) x (1 +1/2 k)=(2k+1)/(2k+2)<1. Ainsi nous obtenons la borne attendue sur |c'|. Puisque le temps d'exécution de notre algorithme est polynomial selon les différents paramètres, nous avons montré que B est un algorithme d'Occam aléatoire pour C et T.

4.1.4 Apprentissage des listes de décisions dans l'apprentissage PAC avec distributions bienveillantes

4.1.4.1 Les listes de décisions

Une liste de décisions sur $x_1,\ldots,x_n$ est une liste ordonnée de termes

$$L= (m_1,b_1),\ldots,(m_p,b_p)$$

dans laquelle chacun des m_i est un monôme sur $x_1,\ldots,x_n$ et chacun des b_i est un élément de {0,1}. Le monôme du dernier terme de la liste est toujours défini par m_p = vrai =1. Pour tout n-uplet a$\in$ $\{0,1\}^n$, L(a) est égal à b_i, où i est le plus petit indice pour lequel $m_i(a)$ est vrai.

Exemple
Soit la liste L = (m_1,b_1), $(m_2,b_2),(m_3,b_3) = (x_1,1),(\neg x_2x_3,0),(vrai,1)$

L(10001)=1 car $m_1(10001)$ est vrai et $b_1 = 1$

L(00111)=0 car $m_1(00111)$ est faux donc on passe à m_2 et $m_2(00111)$est vrai et $b_2 = 0$

L(01111)=1car $m_1(01111)$ est faux, $m_2(01111)$est faux, on passe à m_3 et m_3 est toujours vrai avec $b_3 = 1$

Nous considérons une représentation des listes de décision telle que la taille d'une liste de décision c de p termes sur n variables vérifie $|c|=O(np)$. La classe des k-listes de décision (tout monôme contient au plus k littéraux) est PAC apprenable [Rivest, 1987]. On ne sait pas si les listes de décisions sont PAC apprenables. Nous démontrons que les listes de décisions sont PAC apprenables avec distributions bienveillantes à l'aide du théorème d'Occam.

4.1.4.2 Ensemble d'enseignement d'une liste de décisions

Pour démontrer cela il est d'abord nécessaire de définir un enseignant T. Soit m un monôme sur $x_1,\ldots,x_n$. Nous définissons 0_m comme le plus petit n-uplet (dans l'ordre lexicographique) satisfaisant m. De même, 1_m est le plus grand n-uplet satisfaisant m [Li et Vitànyi, 1991] .

Exemple
Si n=5, $m=\neg x_2x_3$, alors 0_m=00100, 1_m=10111

Soit c la représentation d'une liste de décision $L=(m_1,b_1),\ldots,(m_p,b_p)$, on définit l'ensemble d'enseignement de c par :

$$T(c)=\{(0_{mi}, c(0_{mi})) \mid 1\leq i \leq p\} \cup \{(1_{mi}, c(1_{mi})) \mid 1\leq i \leq p\}$$

Le but d'un tel ensemble d'enseignement pour une liste c est de permettre la reconstruction des monômes de c. En effet, soient x et x' deux n-uplets dans X_n, on définit $x \oplus x'$ comme étant le monôme sur $x_1,\ldots x_n$ qui contient x_i si x et x' ont la valeur 1 pour l'indice i, contient $\neg x_i$ si x et x' ont la valeur 0 pour l'indice i, et qui ne contient pas la variable x_i sinon ($1\leq i \leq p$).

Exemple
x=10010, x'=01010, $x \oplus x'=\neg x_3x_4\neg x_5$

Etant donné l'ensemble caractéristique T(c) d'une liste de décision c, l'ensemble $\{m= x \oplus x' \mid (x, c(x)),(x', c(x'))\in T(c)\}$contient tous les monômes de c.

4.1.4.3 Proposition 1

Les listes de décisions sont PAC apprenables avec distributions bienveillantes en temps usuellement polynomial.

Démonstration

Soit D la classe des listes de décision. Nous donnons un algorithme d'Occam pour cette classe. Cet algorithme est basé sur l'algorithme de Rivest pour la classe des k-listes de décision. L'algorithme utilise les exemples dans le multi-échantillon S par nombre d'occurrences décroissant jusqu'à ce que tous les exemples représentatifs soient utilisés.

1. **Algorithme d'Occam B**
2. **Entrée** : un multi-échantillon S du concept cible c
3. **Début**
   4. S' :=S
   5. i :=1
   6. h := liste vide /* l'hypothèse courante h est initialisée à vide*/
   7. **tant que** S'≠∅
      8. S_i :={(x,c(x)) ∈S | S(x,c(x))≥ Card(S)/i}
      9. M_i := {m= x ⊕ x' | (x, c(x)) ∈S_i,(x', c(x'))∈S_i}
      10. M_i := M_i /{m | m est dans h}
      11. **Tant qu'**il existe un monôme m dans M_i satisfait par un exemple (y,b) de S et par aucun exemple (z,¬b) de S
          12. h :=h+(m,b)
          13. S' :=S'-{(y,b) ∈ S' | m(y)=1}
      14. **Fin tant que**
      15. i :=i+1
   16. **fin tant que**
   17. retourner l'hypothèse h
18. **fin**

Il est facile de démontrer que B est un algorithme d'Occam pour D et T. Soit c la liste de décision cible et supposons que T(c) ⊆ S. Soit j = $\lceil 1/f_{min}(S,c) \rceil$, l'algorithme s'arrête dans le pire des cas pour la valeur i = j car l'ensemble M_j en ligne 9 contient tous les monômes de la liste cible c.

L'hypothèse h est consistante avec S et la longueur de cette hypothèse h est bornée supérieurement par Card(M_j) = ($\lceil 1/f_{min}(S,c) \rceil$)^2. Comme $1/f_{min}(S,c) \leq$ Card(S), il est facile de prouver que l'algorithme B tourne en temps polynomial en |c|, Card(S). Par conséquent le théorème 1 peut être appliqué.

4.2 Relation entre le modèle PAC avec distributions bienveillantes et les modèles d'enseignabilité

Dans le cadre du modèle d'identification à la limite de Gold, l'idée que l'enseignant puisse aider à l'apprentissage en utilisant un échantillon caractéristique n'est pas neuve. On parle alors d'enseignabilité.

4.2.1 Modèles d'enseignabilité

Ainsi Gold [Gold, 78] a proposé un modèle d'identification par données fixées qui formalise cette idée. En suivant [De La Higuera, 1996], et en utilisant les définitions des sections précédentes, en particulier celle d'un enseignant, nous pouvons formuler ce modèle de la manière suivante :

4.2.1.1 Définition 9 : polynomialement identifiable

Une classe de concepts C est polynomialement identifiable à partir de données fixées si et seulement si il existe un enseignant T, un polynôme p et un algorithme A tel que :
1) pour tout concept c de C et tout échantillon S de c de taille m, l'algorithme A avec S pour entrée retourne une hypothèse h consistante avec S en un temps p(m)
2)pour tout concept c de C et tout échantillon S de c contenant T(c), A avec S pour entrée retourne un concept h tel que h≡c.

Plus récemment, Goldman et Mathias [Goldman et Mathias, 96] ont proposé un modèle d'enseignabilité basé sur la paire Enseignant/Elève.

4.2.1.2 Définition 10 : semi-polynomialement enseignable

Une classe de concept C est semi-polynomialement enseignable si et seulement si il existe un enseignant T, un élève L, un polynôme p tels que pour tout adversaire Adv :
1) l'adversaire Adv choisit un concept c dans C et le transmet à l'enseignant T,
2)l'enseignant T choisit un échantillon T(c) d'exemples de c et le transmet à l'adversaire,
3)l'adversaire Adv complète l'échantillon T(c) pour former un échantillon S de c de taille m et le transmet à l'élève.
4)l'élève L retourne en un temps inférieur à p(m) une hypothèse h telle que h≡c.

Le rôle de l'adversaire est d'éviter la collusion entre enseignant et élève. Son intervention au point 1) de la définition équivaut à une quantification universelle sur les concepts de la classe considérée et au point 3), à une quantification universelle sur les échantillons contenant l'ensemble d'enseignement. Goldman et Mathias parlent de classes *semi*-polynomialement enseignables car il n'est pas réclamé ici que l'enseignant soit polynomial, seul l'apprenant doit l'être.

Il est démontré le théorème suivant :

4.2.1.3 Théorème 4 [De La Higuera, 1996]

Une classe de concepts C est polynomialement identifiable à partir de données fixées si et seulement si elle est semi-polynomialement enseignable.

Par ailleurs, il existe un autre théorème démontré dans [Goldman et Mathias, 96]

4.2.1.4 Théorème 5 [Goldman et Mathias, 96]

Si une classe de concepts C est apprenable exactement par requêtes[86] *en temps polynomial alors elle est semi-polynomialement enseignable*

On déduit de ce théorème que les k-DNF, les k-CNF, les k-listes de décision sont semi-poly enseignables.

4.2.2 Comparaison des modèles d'enseignabilité

Nous comparons notre modèle d'apprentissage aux modèles d'apprentissage définis précédemment. Nous démontrons tout d'abord le résultat suivant :

4.2.2.1 Théorème 6

Si une classe de concepts C est semi-polynomialement enseignable alors elle est PAC apprenable avec distributions bienveillantes en temps usuellement polynomial.

Démonstration

Soit C une classe de concepts semi-polynomialement enseignable et soient T l'enseignant et L l'élève correspondant. L'élève L constitue alors un algorithme d'Occam car T(c) est inclus dans l'échantillon. L retourne l'hypothèse *exacte,* qui est donc nécessairement consistante avec l'échantillon, et L tourne en temps polynomial en |c| et card(S). L étant un algorithme d'Occam, de par le théorème 1 , la classe C est apprenable avec distributions bienveillantes en temps usuellement polynomial.

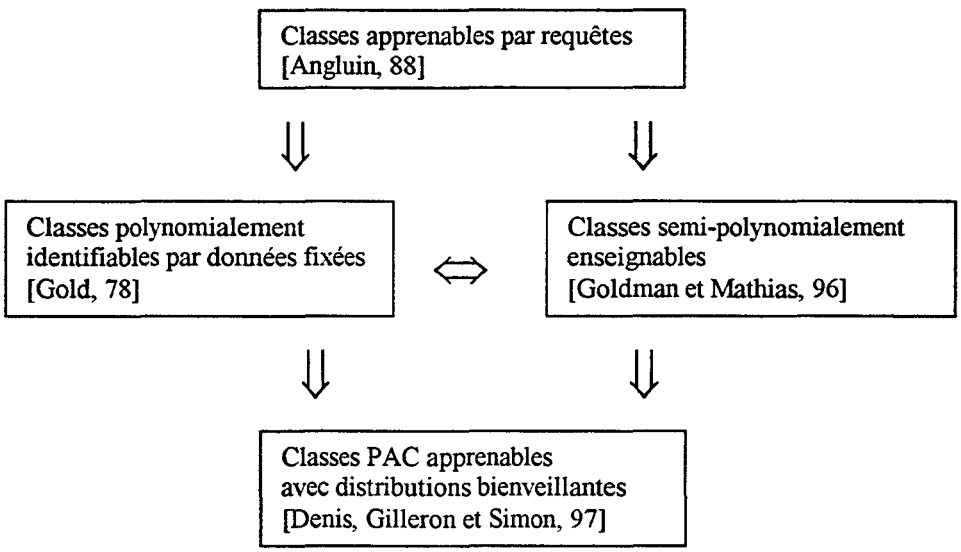
On peut donc déduire de ce résultat et du théorème 5 que toute classe apprenable exactement par requêtes est PAC apprenable avec distributions bienveillantes.

---

[86] On trouvera les différents types de requêtes autorisées dans [Angluin, 88] et [Goldman et Mathias, 96]

4.2.2.2 Comparaison des modèles d'enseignabilité

Nous obtenons ainsi la série d'équivalences et d'implications suivante :



4 Conclusion

Dans ce chapitre, nous avons défini ce qu'est un ensemble d'enseignement et ce que sont des distributions bienveillantes. Nous avons présenté un modèle d'apprentissage PAC avec distributions bienveillantes et un théorème d'Occam avec sa réciproque pour ce modèle. Nous avons démontré l'apprenabilité des listes de décisions dans ce modèle. Nous avons mis en évidence les relations de ce modèle avec d'autres modèles, ce qui permet de récupérer les résultats d'apprenabilité obtenus dans ceux-ci.

Le modèle d'apprentissage avec distributions bienveillantes formalise l'idée naturelle suivante : en situation d'apprentissage, les exemples du concept cible qui sont fournis à l'apprenant ne sont pas complètement arbitraires, à tout concept correspondent des exemples plus ou moins représentatifs de ce concept qui constituent l'ensemble d'enseignement. Les distributions bienveillantes garantissent que tout élément de cet ensemble d'enseignement a une probabilité non nulle d'apparaître dans l'échantillon.

Cependant le choix de ces exemples ne dépend par seulement du concept cible mais aussi de la connaissance qu'a l'enseignant de l'apprenant. C'est ainsi que l'on rejoint le modèle d'enseignabilité de [Goldman et Mathias, 96] pour lesquels : « this T/L pair [paire Enseignant/Elève] relies on the power of the teacher's knowledge of the learner

to help the learner to solve a hard computational problem ». Dans l'apprentissage des listes de décisions (4.1.4) si l'apprenant ne « reconstruit » pas les monômes en croisant avec $\oplus$ tous les exemples d'une certaine fréquence, l'ensemble d'enseignement perd sa raison d'être. Dès que l'on introduit la notion d'ensemble d'enseignement dans le modèle, on suppose alors que l'enseignant a une certaine connaissance de l'apprenant.

Ceci pose le problème de la collusion entre enseignant et apprenant : dans quelle mesure l'enseignant ne fournit pas trop d'information à l'apprenant. Pour éviter ce problème de collusion [Goldman et Mathias, 96] introduisent dans leur modèle un « adversaire ». On a vu que cet adversaire correspond dans notre modèle aux différents quantificateurs universels sur le concept cible et sur les distributions bienveillantes qui servent au tirage de l'échantillon soumis à l'apprenant. De ce fait, le modèle avec distributions bienveillantes évite lui aussi une possible collusion.

Dans le 4.2.1.3, en donnant la définition de l'apprentissage PAC avec distributions bienveillantes, nous faisions remarquer que le fait que le temps d'apprentissage soit polynomial en $1/P_{min}(c)$ implique que si ce paramètre est exponentiel en n alors le temps d'apprentissage est lui-aussi exponentiel en n. Il pourrait alors être intéressant de restreindre encore la classe des distributions possibles aux seules distributions bienveillantes « polynomiales » où $1/P_{min}(c)$ serait polynomial en n. Peut-être qu'alors de nouvelles classes seraient apprenables. La distribution uniforme, ne pourrait appartenir à une telle classe de distributions puisqu'avec celle-ci $1/P_{min}(c)=2^n$. Ceci semble être une hypothèse plausible au niveau cognitif car, comme on le verra dans le prochain chapitre, il est rare qu'elle soit utilisée lors des expériences.

Chapitre 5

Opérationnalisation du modèle PAC

5 Introduction

«Aujourd'hui, il est légitime de se demander si les théories formelles ne sont pas uniquement des constructions esthétiques destinées à nourrir les fantasmes cognitifs de quelques informaticiens ou à satisfaire les tendances formalistes de quelques chercheurs en sciences cognitives. » [Boucheron, 92][87]

Le travail que nous avons entrepris est en quelque sorte une réponse à cette interrogation de Boucheron. Si le modèle PAC correspond à un fantasme cognitif, alors c'est un fantasme très proche de la réalité. La lecture des expérimentations en psychologie filtrée par le modèle PAC du chapitre 2 a permis de montrer que chaque concept qui sous-tend ce modèle (hypothèse de l'apprenant, distributions de probabilités, etc.) trouve son correspondant dans l'étude de la catégorisation en psychologie. De la même manière l'étude comparée de la typicalité et de l'économie cognitive au chapitre 3 montre qu'il

[87]A notre connaissance, Boucheron est une des rares personnes qui se soit préoccupée de l'intérêt des théories formelles dans le cadre des sciences cognitives. Dans la première partie de ce chapitre nous « débattons » avec lui c'est pourquoi nous sommes amené à le citer abondamment.

est possible, au niveau des résultats, de trouver des points de convergence. Ainsi, le modèle PAC permet de décrire, au moins partiellement, la catégorisation (en tant que *processus*) et, même s'il a besoin d'être amendé, d'appréhender de nombreux aspects de celle-ci.

Nous souhaiterions maintenant aller plus loin et opérationnaliser le modèle PAC. Opérationnaliser le modèle PAC, consiste à transformer la définition de l'apprentissage PAC en un protocole d'expérience en psychologie. L'idée n'est pas neuve car, dès 92, Boucheron note que le modèle PAC s'y prête relativement bien : «Mais c'est sur les questions épistémologique et méthodologique, que la théorie de l'apprentissage PAC s'avère la plus novatrice. En temps polynomial, il n'est certes pas possible de décider si l'hypothèse produite par un apprenti est correcte, mais il est possible d'estimer le taux d'erreur de cette hypothèse avec une bonne précision et une forte probabilité. En adoptant la définition du réalisable qui sous-tend la théorie (ce qui est calculable en temps probabiliste polynomial), le succès devient donc effectivement vérifiable, et ce, à partir d'un comportement publiquement observable ! [...] *Ceci permet idéalement tout au moins, de concevoir des protocoles expérimentaux pour étudier un processus d'apprentissage qui suivrait la définition de l'apprentissage PAC*[88]. Ce travail n'a, certes, jamais été effectué, mais il est tout de même encourageant de constater qu'une approche purement formelle par son origine, normative dans son ambition, puisse prétendre intervenir au niveau descriptif. » [Boucheron, 92]

Comme nous le verrons dans les deux prochaines sections, l'intérêt d'un tel travail est de permettre de lier plus étroitement encore étude de la catégorisation en psychologie et théories formelles de l'apprentissage et de voir si cela ne permet pas de lever certains problèmes qui apparaissent dans l'un ou l'autre domaine.
Ainsi, dans le 5.1, nous verrons l'utilité d'une telle démarche pour les théories formelles : mieux comprendre les mécanismes de la catégorisation naturelle permet en retour de proposer des variantes du modèle qui seront plus proches de la réalité.
Dans le 5.2, nous verrons que dans certaines expériences en psychologie sur des catégories artificielles, plusieurs aspects de la catégorisation que nous qualifions d'«écologiques», inhérents au modèle PAC, ne sont pas pris en compte et que ceci explique peut-être les résultats négatifs obtenus.
Nous passerons ensuite, dans le 5.3, à l'opérationnalisation proprement dite et analyserons la définition de l'apprentissage PAC point par point ce qui nous amènera à mettre à jour les différentes variables expérimentales que le modèle suppose.
Dans le 5.4 nous définirons ce que peut être un concept appris de manière Approximativement Correcte et décrirons un protocole d'expérimentation issu de cette définition.
Enfin, dans le 5.5 nous présenterons le compte-rendu de deux expériences. Dans la seconde, l'absence de résultats positifs amène à une réflexion sur la distance qui existe entre le modèle PAC opérationnalisé et la catégorisation.

[88]souligné par nous

5.1 L'opérationnalisation du modèle PAC pour justifier le choix de certains modèles

Il y a deux raisons au moins pour opérationnaliser le modèle PAC. La première tient à la nature même d'un modèle. Un modèle a, peu ou prou, l'ambition de capter certains aspects du monde réel. Si ce n'est le cas, il perd sa raison d'être et devient, comme le dit Boucheron, une construction théorique quelque peu gratuite. Dans le cas des théories formelles de l'apprentissage, nous pouvons distinguer deux objectifs au modèle : la modélisation de la catégorisation et celle de l'apprentissage automatique.
En permettant d'approcher davantage encore la catégorisation, l'opérationnalisation devrait aider à justifier ou à invalider les choix théoriques opérés par les différentes déclinaisons du modèle standard (apprentissage PAC simple, apprentissage PAC avec distributions bienveillantes, ...). Elle devrait permettre de repérer celles qui sont les plus proches de la réalité.
Dans le cas de l'apprentissage automatique, les théories formelles sont déjà utilisées et donc les modèles proposés remplissent déjà plus ou moins leur tâche. Cependant, l'apprentissage automatique a pour ambition de simuler autant que possible l'apprentissage naturel car, comme le disent aussi bien Valiant que Pitt, le meilleur apprenant est quand même l'homme. C'est ici que nous trouvons une seconde raison d'opérationnaliser le modèle PAC : cette opérationnalisation devrait permettre de mettre plus facilement à jour les heuristiques utilisées par l'homme qui pourraient être simulées ensuite en apprentissage automatique.

Dans cette section, nous montrons en premier lieu que les théories formelles de l'apprentissage ont pour objectif de modéliser l'apprentissage automatique mais aussi l'apprentissage naturel. Si, dans les débuts de la recherche, ce second objectif était clairement exprimé, il l'est beaucoup moins maintenant et, à l'heure actuelle, les avis sont partagés. En second lieu nous présentons les résultats de ces théories relativement à l'apprentissage naturel. Nous verrons que ces résultats «formels », lorsqu'ils existent, ne rendent pas vraiment compte des résultats observés, ce qui amène une réflexion en retour sur les modèles.

5.1.1 Une modélisation de l'apprentissage naturel

Nous pouvons distinguer trois qualités à un modèle théorique :
-qu'il soit cohérent,
-qu'il soit pertinent du point de vue cognitif,
-qu'il soit compatible avec les données expérimentales.

Le premier point va de soi, un modèle qui n'est pas cohérent n'a pas de réelle valeur scientifique, de là l'appel de D. Angluin pour que les chercheurs appuient leurs affirmations par des théorèmes «Rigorous : theorems, please » [Angluin, 92].

Les deux points suivants sont indissociables. Pour montrer la pertinence cognitive d'un modèle, il faut que celui-ci soit compatible avec les données expérimentales. Mais alors que le premier point fait surtout référence à l'apprentissage naturel, le second pourrait se contenter des résultats obtenus en apprentissage automatique.

Dans l'introduction générale, nous avons brièvement présenté le modèle de Gold. Ce modèle en tant que référence en apprentissage automatique, est de moins en moins utilisé au profit de celui de Valiant. Concernant l'apprentissage naturel, ce modèle avait à l'origine pour ambition plus ou moins déclarée d'essayer d'appréhender le langage naturel. «Si, aujourd'hui, un article se réclamant d'une théorie formelle de l'apprentissage traite généralement son domaine comme une pure élaboration formelle, voire comme une discipline d'ingénieur, l'article fondateur de Gold, *Language identification in the limit*, formulait un programme de recherche fortement marqué par les sciences cognitives. Ce programme s'articulait autour de deux volets : l'étude de la structure linguistique et *la possibilité d'acquérir cette structure.* » [Boucheron, 92]. Pour appuyer les dires de Boucheron, il suffit d'examiner la bibliographie proposée par Gold à la fin de son article : elle renvoie à Chomsky via Harris. «Le fait même qu'il [Gold] évoque une structure linguistique, un dispositif stable qui permet d'exercer la faculté de langage, qu'il réduise - pour des raisons méthodologiques - à ses aspects grammaticaux, indique que la pensée de Chomsky a fortement influencé la construction du paradigme de l'identification à la limite. » [Boucheron, 92]
En 67, l'année où Gold propose son modèle d'identification à la limite, Chomsky est un élément incontournable dans le champ des sciences cognitives. De plus, l'étude que Chomsky fait des langages se prête particulièrement bien à une approche informatique de ceux-ci. C'est pourquoi le modèle de Gold est fortement inspiré de ses travaux. Gold va rechercher les conditions formelles de la maîtrise du langage, en espérant que cette étude sera pertinente vis-à-vis de l'apprentissage naturel. Comme le fait remarquer Boucheron, ce qui est original dans la démarche de Gold ce n'est pas que la faculté de langage soit étudiée formellement, mais que le *processus* de la mise en place de cette faculté le soit.

Concernant le modèle PAC de Valiant, avant de l'envisager comme modélisation de l'apprentissage naturel, nous pouvons constater qu'il est relativement bien implanté en ce qui concerne l'apprentissage automatique. Boucheron, toutefois, fait une distinction entre les systèmes d'apprentissage symboliques et les autres. Concernant les systèmes d'apprentissage symboliques : «les articles les plus connus sur la relation entre apprentissage PAC et l'apprentissage symbolique[...] présentent une formalisation d'une technique classique de l'intelligence artificielle (versions spaces de Mitchell et AQ de Michalsky, conceptual clustering de Michalsky...) et montrent l'un la pertinence de l'heuristique en question vis-à-vis de l'apprentissage PAC, l'autre, la complexité de la technique générale des versions spaces. » [Boucheron, 92]. Boucheron marque ainsi un certain scepticisme quant à l'utilisation du modèle PAC en apprentissage symbolique, c'est omettre que des résultats tels que la VCD ou l'équivalence entre apprentissage faible et apprentissage fort (voir chapitre 1) sont parfois utilisés dans ce type d'apprentissage. Son scepticisme peut s'expliquer par le fait que la plupart de ces articles ont été écrits après la parution de son livre. Par contre en ce qui concerne l'apprentissage

automatique non-symbolique : «La confrontation avec la reconnaissance de formes et avec le connexionnisme est heureusement plus féconde » [Boucheron, 92].

De fait, dans le domaine de l'apprentissage automatique, le modèle de Valiant devient bien un repère qui permet aux praticiens, préoccupés de créer des systèmes apprenants, de confronter leurs résultats empiriques aux résultats théoriques. Nous avons déjà expliqué qu'au rasoir d'Occam correspondait l'heuristique du Minimum Description Length Principle proposée par Rissanen et fréquemment utilisée en apprentissage automatique. Par ailleurs, il suffit de prendre les trois volumes de «Computational learning theory and natural learning systems » édités en 94, et l'on constate que sur 61 articles consacrés à l'apprentissage automatique, 14 font référence à l'apprentissage PAC. Certains de ces articles se préoccupent d'adapter le modèle de Valiant à d'autres types d'apprentissage : apprentissage analogique [Cook, 94], apprentissage de concepts flous [Kearns et Shapire, 94], révision de la théorie [Mooney, 94] ; d'autres l'utilisent comme cadre pour résoudre des problèmes théoriques relativement à leur pratique ([Baum, 94]) ; ou confirment empiriquement des résultats théoriques ([Drucker, Shapire, Simard, 92] relativement à l'apprentissage faible (voir chapitre 1)), etc. Dans tous ces articles, il y a une constante : le modèle PAC est une référence, un cadre qui permet de positionner les recherches propres aux auteurs.

Pour l'apprentissage naturel, la filiation vis-à-vis de la psychologie est moins évidente chez Valiant que chez Gold. Cette filiation est beaucoup plus implicite. Ainsi nous avons expliqué en note page 77, comment le modèle de Valiant suivait d'assez près les protocoles expérimentaux pré-roschiens tels ceux développés par Neisser et Weene ou Richard. Par ailleurs dans son article fondateur «A theory of the learnable », [Valiant, 84], Valiant fait référence à plusieurs reprises à l'apprentissage naturel. Le résumé de l'article commence ainsi : «Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventionnal sense. » Tout au long de l'article, Valiant fait référence à cette capacité humaine d'apprendre de nouveaux concepts, sans que cette acquisition du concept soit explicitement programmée.
Mais alors que Gold cherchait à modéliser un apprentissage naturel pour mieux l'appréhender, Valiant cherche à le modéliser afin de pouvoir plus facilement le simuler par des machines. «These skills[89] often have the additionnal property that, although we have learned them, we find it difficult to articulate what algorithm we are really using. In these cases it would be especially significant if machines could be made to acquire them by learning .»

Alors que Gold affichait assez clairement que son modèle visait l'apprentissage de grammaires, nous ne trouvons pas chez Valiant d'affirmation équivalente. Valiant ne répond pas à la question de savoir quel processus humain son modèle formalise, la réponse ci-dessus est plutôt évasive. N'y a-t-il pas alors de champ défini en psychologie concerné par le modèle de Valiant ? [Boucheron, 92] apporte la réponse : « Si ces disciplines formelles abordent un phénomène naturel d'apprentissage, il s'agit de

[89]« Skills » fait référence, ici, aux talents appris et non programmés

l'apprentissage des catégories. » Mais, il va se rétracter aussitôt : «Cette intuition appelle deux séries d'interrogations : la première porte sur cette notion de catégorie, elle recouvre en fait des entités qui semblent posséder des statuts psychologiques différents ; la seconde porte sur le caractère effectivement appris de ces catégories, et donc sur la pertinence de la modélisation de cet apprentissage. » Après l'avoir abondamment cité depuis le début de ce chapitre, nous devons maintenant nous démarquer de Boucheron. Le raisonnement de Boucheron pour affirmer que la catégorisation naturelle n'est pas modélisée par l'apprentissage PAC est assez simple et tient en deux points :
1)les catégories dont le modèle PAC pourrait modéliser l'apprentissage sont les catégories perceptives,
2) ces catégories ne sont pas *apprises* par l'être humain.
Si nous sommes d'accord sur le premier point, nous divergeons sur le second.

Boucheron distingue entre deux classes de catégories. D'une part, les catégories perceptives, telles que les couleurs, les directions, et les catégories perceptives techniques, telles que, les triangles, les cercles. Il appelle ces catégories, les catégories perceptives naturelles. De l'autre, les catégories peu tributaires des sens et beaucoup du système linguistique telles que les coniques. Notons déjà qu'il est difficile de distinguer entre les catégories perceptives «techniques » (?) et celles définies essentiellement par le langage. Nous admettons, cependant, que l'apprentissage PAC ne s'intéresse qu'aux catégories perceptives, si nous définissons celles-ci par le fait que l'information qui a permis de les former provient essentiellement des exemples[90].
Selon Boucheron, ces catégories perceptives naturelles ne seraient pas apprises mais résulteraient du développement des systèmes perceptifs et cognitifs. Ainsi : «En ce sens on n'apprend pas les couleurs, on apprend tout au plus à les nommer. » [Boucheron, 92] Pour conforter son opinion, il ne cite que deux psychologues Melher et Dupoux auteurs de «Naître humain ». Nous pouvons penser que c'est aller un peu vite en besogne. Certaines représentations résulteraient du développement et d'autres de l'apprentissage, et les deux termes seraient antinomiques. Nous pourrions caricaturer sa position en disant que l'enfant n'apprend pas, il se développe, nous pourrions étendre le raisonnement à l'adulte. Cette opposition nous paraît arbitraire et pour justifier de notre désaccord, nous n'utiliserons qu'un seul argument : la typicalité. Si la reconnaissance des couleurs relève du développement et l'acquisition du théorème de Thalès, par exemple, relève de l'apprentissage, il faudrait pouvoir expliquer pourquoi nous trouvons chez le sujet des représentations qui sont jugées plus typiques dans un cas comme dans l'autre. Si la typicalité relève du développement, elle ne devrait pas apparaître dans l'apprentissage et réciproquement. A moins que nous ne considérions qu'elle puisse apparaître dans les deux, mais, en prenant cette position, nous en arriverions vite à constater qu'il n'y a plus grand chose qui les distingue[91] [92].

---

[90]Ce genre de catégorie correspond au niveau 1 de Blewitt [Blewitt, 94]. Cela pose de réelles limitations à la modélisation PAC de l'apprentissage qui seront discutées dans la conclusion générale.

[91] Cette dichotomie entre apprentissage et développement va permettre à Boucheron d'affirmer ensuite que les théories formelles sont nativistes (innéistes), position que nous ne partageons pas non plus.

[92] Ceci ne signifie pas que nous considérons que le développement se résume à l'apprentissage, mais nous pensons qu'il existe un processus de catégorisation présent chez l'enfant qui perdure chez l'adulte.

Pour Pitt [Pitt, 97], comme pour Boucheron, il est évident que ce qui est modélisé par les théories formelles, en psychologie, est la catégorisation mais, à l'opposé de Boucheron, il considère que les catégories s'apprennent. En fait, toute une partie de son article va consister à passer en revue les recherches en psychologie concernant celle-ci pour améliorer les modèles informatiques.

5.1.2 Un modèle doit pouvoir rendre compte des réalités observées

L'ambition de Gold était l'apprentissage du langage naturel. Malheureusement les résultats qu'il obtient sont négatifs[93]. Il suppose, d'une part, que les langues naturelles sont incluses dans la classe des langages context-sensitive et, de l'autre, que les langues naturelles ne s'apprennent que par exemples positifs. Pour ce dernier point, il s'appuie sur les travaux de Mac Neill selon lesquels on ne présente pas à l'enfant des phrases à la syntaxe défectueuse en lui expliquant qu'elles sont incorrectes. Le problème est que Gold a démontré que, dans son modèle d'identification à la limite, les langages context-sensitive ne sont pas apprenables par exemples positifs [Gold, 67]. Conséquemment les langages naturels ne sont pas apprenables, ce qui est en contradiction avec la réalité. Pour lui, il y a trois explications possibles à cette contradiction :
1) soit la classe des langages naturels est une classe beaucoup plus petite que celle des langages context-sensitive, ou, autre possibilité, l'enfant démarre l'apprentissage avec plus d'information que celle contenue dans les seuls exemples,
2) soit l'enfant reçoit aussi des exemples négatifs mais ces exemples négatifs sont présentés de manière telle que les psychologues ne l'ont pas encore repérée.
3) soit les exemples positifs n'apparaissent pas n'importe comment. Par exemple, les exemples pourraient suivre un certain ordre de présentation.
Ce dernier point est repris par [Osherson, Stob et Weinstein, 86]. Dans l'introduction, nous expliquions que l'apprentissage dans le modèle de Gold devait avoir lieu pour tout flux d'exemples. Ces auteurs considèrent que cela n'est pas le cas dans le cas de l'apprentissage de la langue. Utilise-t-on des structures syntaxiques Proustiennes lorsque l'on s'adresse aux enfants ?

A notre connaissance, pour le modèle PAC, il n'existe pas de résultats relatifs à l'apprentissage naturel. Cela s'explique par la réticence des chercheurs à envisager le modèle PAC comme une modélisation de cet apprentissage. Les résultats y sont plus généraux. Ainsi pour [Pitt, 97], si dans le cadre de l'apprentissage PAC, seules des classes de concepts simples ont pu être démontrés apprenables cela n'est pas outre mesure choquant car «people don't do to well either.» Pour affirmer cela, il s'appuie sur les résultats de [Neisser et Weene, 62] selon lesquels les disjonctions de deux conjonctions sont difficilement apprenables (voir chapitre 2). La question reste donc ouverte pour ce qui concerne le modèle PAC car le principe d'utiliser celui-ci pour modéliser l'apprentissage naturel n'a pas vraiment été exploité.

---

[93] Notons cependant que l'article de Berwick présenté dans l'introduction montrait que le théorème d'Angluin (qui s'inscrit dans le modèle de Gold) pouvait avoir une interprétation en apprentissage naturel.

5.1.3 L'intérêt d'opérationnaliser le modèle PAC pour les théories formelles de l'apprentissage

Dans ce contexte, quel est l'intérêt d'opérationnaliser le modèle PAC ? Si nous reprenons ce que nous disions au début de ce chapitre, c'est d'abord pour lier davantage encore les études sur la catégorisation et celles concernant le modèle PAC. Un modèle est légitime s'il permet de rendre compte, au moins partiellement de la réalité. Dans ce sens, que ce soit celui de Gold ou celui de Valiant, ces modèles sont déjà validés par l'usage qui en est fait en apprentissage automatique. Ceci semble omis par [Boucheron, 92] qui ne semble considérer que l'apprentissage naturel lorsqu'il dit : «Si les théories formelles de l'apprentissage s'avèrent incapables [...]de susciter des développements dans les spécialités expérimentales, il faudra admettre qu'elles n'ont constitué qu'un divertissement, une manière - pas nécessairement futile - de reformuler des problèmes de calculabilité, de probabilités, d'optimisation et de complexité abstraite, une manière déguisée de faire de l'analyse asymptotique. »

Mais il est souhaitable d'aller plus loin que la seule légitimation par l'apprentissage automatique. L'intérêt selon [Pitt,97] est que plus le modèle sera proche de l'apprentissage naturel plus il sera utile. Il permettra de mieux comprendre ce qu'est cet apprentissage naturel, ce qui est l'ambition de Gold, mais il permettra aussi de mettre à jour des heuristiques qui seront ensuite implantées dans des machines, ce qui est l'ambition de Valiant. Ainsi essayer d'appréhender l'apprentissage humain par le biais des théories formelles n'a pas qu'un aspect cognitif « qu'est-ce qu'apprendre ? » mais aussi un aspect utilitaire « comprendre ce qu'est apprendre pour mieux le simuler par des machines ».

L'opérationnalisation d'un modèle consiste à le transformer en protocole expérimental, ce qui permet de construire des expériences qui suivent de près sa définition. Pour des raisons de calculabilité en temps polynomial, citées plus haut par Boucheron, le modèle de Gold ne se prête pas à cette opérationnalisation ce qui n'est pas le cas de celui de Valiant.
Opérationnaliser un modèle doit permettre de voir dans quelle mesure celui-ci est capable d'appréhender l'apprentissage naturel. Il est possible ensuite en expérimentant selon ce protocole de privilégier telle ou telle version du modèle qui serre la réalité au plus près. L'opérationnalisation du modèle PAC devrait permettre de choisir parmi les différentes déclinaisons qui existent de ce modèle (PAC standard, PAC simple, PAC avec distributions bienveillantes,...) celle qui correspond le mieux à la catégorisation.

Cependant, tel quel, le modèle PAC n'est pas opérationnel. Il est relativement facile de l'utiliser en apprentissage automatique où les principes de base sont similaires à ceux des théories formelles, il est beaucoup plus difficile de le mettre en application en psychologie. Les chapitres 2 et 3, s'ils ont permis de montrer que les deux domaines, théories formelles et étude de la catégorisation, avaient de nombreux points communs, ont aussi montré que l'adéquation de l'un à l'autre n'était pas immédiate. Citons Boucheron une dernière fois : «On peut certes chercher à illustrer les théories formelles

de l'apprentissage, à montrer que tel problème concret, naturel, d'apprentissage gagne à être examiné à leur lumière, mais [...] il s'agit d'une entreprise délicate. » [Boucheron, 92]. C'est l'objet des sections consacrées à l'opérationnalisation d'essayer de lever les difficultés. Avant cela, nous voudrions montrer en quoi elle peut être utile à l'étude de la catégorisation en psychologie.

5.2 Intérêt d'opérationnaliser le modèle PAC du point de vue de la psychologie

Pour justifier de l'utilité de l'opérationnalisation du modèle PAC en psychologie, nous allons présenter trois articles[94]. Comme précédemment, ces articles ne sont pas choisis en fonction de leur caractère définitif mais parce qu'ils permettent d'illustrer notre propos. L'opposition entre les deux premiers, [Medin, Wattenmaker et Hampson, 87] et [Kemler-Nelson, 84] permet de mettre à jour certaines caractéristiques de la catégorisation qui sont pris en compte dans le modèle PAC et qu'il semblerait intéressant d'inclure dans l'étude de la catégorisation avec des catégories artificielles. Le troisième [Flannagan, Fried et Holyoak, 86] s'intéresse plus particulièrement à l'impact des distributions de probabilités.

5.2.1 Air de famille, cohérence conceptuelle et construction de catégorie [Medin, Wattenmaker et Hampson 87]

Nous avons déjà effleuré le premier article dans le chapitre 2. Il s'agit de celui de [Medin, Wattenmaker et Hampson, 87] *Air de famille, cohérence conceptuelle et construction de catégorie*. Utiliser l'apprentissage PAC comme modèle de la catégorisation revient d'emblée à considérer la catégorisation comme un *processus*. C'est à ce titre que leur article est intéressant car ces chercheurs partent du principe que la *représentation* de la catégorie est le résultat d'un *processus* qui a généré cette représentation. *Ils ont donc essayé de reproduire ce processus en laboratoire avec des catégories artificielles.* Le problème est qu'ils échouent : ils s'attendaient à obtenir des représentations des catégories en «air de famille» alors que leurs sujets ont établi des représentations en conditions nécessaires et suffisantes (CNS). L'analyse de leur «échec» et l'opposition que nous allons faire avec les expériences de Kemler-Nelson permet d'illustrer plusieurs points d'ordre *méthodologique* qui nous paraissent justifier une opérationnalisation de l'apprentissage PAC. Ces différents points sont tous issus du fait que *la catégorisation, en tant que processus, est une conduite qui permet à l'individu de s'adapter à son environnement.*

Signalons que, parmi leurs sept expériences, nous ne nous intéressons qu'aux quatre premières. En effet, nous considérons que les expériences 5, 6 et 7 ne sont pas concluantes, dans le sens où, lors de celles-ci, l'expérimentateur évoque pour le sujet

[94] La présentation que nous faisons ici est schématique, le lecteur intéressé pourra se reporter à l'annexe 8 où nous faisons un compte rendu plus développé de ces expériences. Les remarques restent néanmoins identiques.

des catégories qu'il peut s'être déjà formé : des individus décrits selon cinq symptômes de maladies, des descriptions de personnes introverties/extraverties, des dessins d'animaux à classer en animaux volants ou non. Il ne s'agit pas pour nous de catégories réellement artificielles. Il n'y a donc pas étude du processus de catégorisation mais de catégories déjà formées.

Dans les quatre premières expériences, il est demandé aux sujets de classer 10 items en deux catégories de taille égale. Ces items sont des dessins (voir figures 5.2 ci-dessous et 2.5 p 104) ou des descriptions de personnes. Selon les expériences, ces items sont construits sur un nombre d'attributs variant de 4 à 6, ces attributs pouvant prendre deux ou trois valeurs.

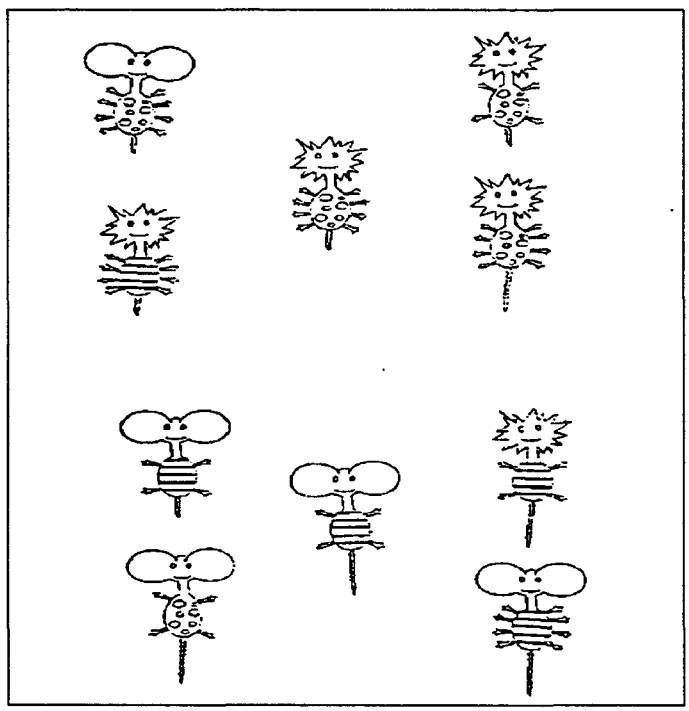


Figure 5.2 : stimuli utilisés par [Medin, Wattenmaker et Hampson, 87]

Dans chaque expérience, l'expérimentateur explique aux sujets qu'il n'y a pas de règle prédéterminée pour classer et que le sujet peut effectuer le classement de son choix.
A l'issue de l'expérience, le chercheur observe si les catégories formées sont en CNS ou en air de famille. Nous pouvons décrire ces classement en utilisant 4 variables binaires[95] (ou 6 selon les expériences).

[95] Le codage binaire est arbitraire mais il permet une plus grande lisibilité. Les attributs pris en compte sont la tête (ronde ou pointue), le corps (rayé ou à pois) , le nombre de pattes (4 ou 8), la queue (courte ou longue). Les exemples 1111 et 0000 en tête de colonne, correspondent aux dessins qui sont au centre des deux groupes de cinq de la figure 5.2, ce sont les prototypes. Il aurait été possible de les coder autrement «1010» et «0101» mais l'air de famille des autres figures auraient été moins facilement lisible

Classement en «air de famille »		Classement selon une dimension	
Catégorie A	Catégorie B	Catégorie A	Catégorie B
1111	0000	1000	0100
1110	0001	1111	0111
1101	0010	1110	0000
1011	0100	1101	0010
0111	1000	1011	0001

Le classement en «air de famille » est caractérisé par le fait que chaque exemple d'une catégorie ne diffère d'un autre exemple de la même catégorie que par la valeur de 2 attributs et que tous les exemples ne diffèrent relativement à l'exemplaire prototypique (en tête de colonne) que par la valeur d'un attribut. Le classement en CNS est basé sur le premier attribut (x1) : selon la valeur de cet attribut, l'item est classé en A (x1=1) ou en B (x1=0).

Tous les classement opérés par les sujets dans les quatre premières expériences proposent des catégories en CNS. La question que nous nous posons alors est celle-ci : est-ce que, dans ces expériences, il y a bien reproduction du processus de catégorisation ?

Le premier point sur lequel nous voudrions nous appesantir réside dans la démarche même de leurs expériences. Dans chacune de celles-ci, les chercheurs demandent aux sujets de classer, *à leur gré*, les 10 items. *La catégorisation y est assimilée à une simple opération de classement.* Ceci appelle deux remarques pouvant expliquer partiellement leur «échec» à obtenir des catégories en air de famille.
La première est que les sujets sont des étudiants en psychologie donc des adultes qui se sont formés dans la culture occidentale. Dans cette culture, lorsque l'on demande à un individu de classer, il y a toujours implicitement l'idée que ce classement doit obéir à des conditions nécessaires et suffisantes. Même une consigne donnée aux sujets expliquant qu'il n'existe pas de règle sous-jacente, qu'il n'y a pas de norme au classement, n'est pas suffisante, de notre point de vue, pour aller à l'encontre de cette idée.
La seconde, qui nous intéresse davantage, est que cette procédure revient à assimiler la catégorisation à un classement des items, ce qui est exact, dans des catégories totalement arbitraires, ce qui est faux. L'individu ne classe pas arbitrairement les éléments de son environnement, il les classe de façon à pouvoir s'adapter à cet environnement (voir 2.2). *C'est pourquoi nous disons que l'environnement impose au sujet les catégories qu'il doit établir. La catégorisation n'est pas simplement un processus de classement mais un processus d'apprentissage.* Cette pression de l'environnement n'apparaît pas dans leurs expériences.

---

qu'ici. Bien évidemment les sujets n'ont pas connaissance de ce codage, ils n'ont connaissance que des figures.

Le second point concerne la complexité de la situation qui peut aussi être une explication de leurs résultats. Sur la base de 4 variables binaires, l'espace des exemples n'est que de taille 16. De plus, seuls 10 de ces 16 exemples sont retenus. *Cette situation n'est donc pas assez complexe.* Cette insuffisance dans la complexité rejaillit sur l'absence d'opposition nette entre les deux options de classement : un classement par air de famille ne diffère d'un classement unidimensionnel que par l'appartenance de deux exemples : 0111 et 1000. Si nous permutons ces deux éléments, en les faisant changer de catégorie, nous passons de catégories en CNS à des catégories en «air de famille ». Cette absence de complexité se vérifie aussi au travers des mesures de similarité calculées par les auteurs : la similarité interne à une catégorie est de 9,6 dans le premier classement et de 9,2 dans le second[96]. La différence entre les deux n'est peut-être pas significative. Il est tout aussi possible de considérer que le second classement respecte un «air de famille ». Dans une autre expérience, les auteurs prêteront attention à ce problème en passant à 6 variables ou en proposant des variables ternaires, mais n'obtiendront pas de changement : le classement opéré par les sujets reste en CNS. Ces accroissements sont sans doute insuffisants pour permettre une modification des résultats.

Ainsi, l'incapacité à reproduire en laboratoire une catégorisation en air de famille peut-elle s'expliquer par les raisons que ces chercheurs donnent à la fin de leur article, et notamment par le fait que la catégorisation est sans doute un processus de bas niveau, mais aussi par le fait qu'ils n'ont peut-être pas correctement reproduit le processus de catégorisation en question car :
- la situation n'est probablement pas assez complexe,
- l'environnement n'impose pas au sujet les catégories qu'il doit établir,
- la catégorisation est envisagée comme un simple processus de classement et non pas comme un processus d'apprentissage.

On notera que ces trois remarques ne sont jamais qu'une lecture des expériences de Medin et al au travers du modèle PAC. La question qui peut se poser maintenant est de savoir si elles sont fondées : si les trois conditions sont respectées, nous devrions obtenir une catégorisation en «air de famille ». La réponse est affirmative et pour le montrer nous allons relater les expériences de Kemler-Nelson.

5.2.2 Les effets de l'intention sur le type de concepts acquis [Kemler-Nelson, 84]

La raison principale de l'article de Nelson est *l'apprentissage incident, celui que le sujet effectue à son insu opposé à l'apprentissage intentionnel dans lequel le sujet sait qu'il apprend.* Elle va proposer des expériences qui opposent ces deux types d'apprentissage. L'idée est qu'en *apprentissage incident, les sujets formeront les catégories en «air de famille », tandis qu'en apprentissage intentionnel, ils formeront des catégories en CNS,* unidimensionnelles.

---

[96] Le calcul est simple. Par exemple, pour la catégorie A du classement par «air de famille », le premier élément 1111 a 3 fois les valeurs de toutes ses dimensions répétées dans les autres exemples de la même catégorie, il a donc une similarité de 12, en calculant de la même façon les autres ont une similarité de 9, la moyenne pour la catégorie est de 9,6 ((12+9+9+9+9)/5)

Les stimuli dans son expérience sont des caricatures de visage variant sur 4 dimensions binaires : le type de chevelure, le type de moustache, la couleur des yeux et la forme du nez. Les exemples, comme dans Medin et al, sont construits autour de 2 prototypes[97] :

Catégorie «docteur»	Catégorie «policier»
0000	1111
0100	1011
0010	1101
0001	1110

Items critiques
0111
1000

Les 8 stimuli présentés sous les catégories sont ceux qui servent à l'apprentissage. Les deux stimuli sous «items critiques» servent à vérifier le type de catégories qu'ont formées les sujets. On aura tout de suite compris que si le sujet classe le premier item critique dans la catégorie «docteur» et le second dans la catégorie «policier», c'est qu'il aura formé des catégories unidimensionnelles, en fonction de la valeur du premier attribut. Tandis que s'il fait l'inverse, c'est qu'il aura formé des catégories en «air de famille ». En effet, le premier item critique a une plus grande similarité avec les éléments de la catégorie «policier» qu'avec ceux de la catégorie «docteur». Il partage 3 valeurs d'attributs avec le prototype de la catégorie «policier» (en tête de colonne) contre 1 avec celui de la catégorie «docteur».

Les sujets, des étudiants, sont répartis en 2 groupes. Un groupe est en apprentissage intentionnel, l'autre en apprentissage incident. L'apprentissage pour le sujet consiste à être capable de classer en 2 catégories, docteur ou policier, les visages que l'on va lui présenter. Durant la phase d'apprentissage, on ne lui présente pas les deux items critiques, ceux-ci ne servent que dans la phase de test et selon le classement qu'en fera le sujet, le chercheur considérera qu'il a effectué une catégorisation en CNS ou en air de famille.

En apprentissage intentionnel, l'expérimentateur signale au sujet qu'il doit apprendre à distinguer entre les deux catégories et qu'il lui demandera ensuite de classer des éléments. L'apprentissage se déroule de la manière suivante : on présente au sujet une série de dessins de visage et on lui demande pour chacun de ces items à quelle catégorie, «docteur» ou «policier», il appartient. Pour répondre, le sujet place le visage au-dessus de l'uniforme d'un policier ou de celui d'un médecin. L'expérimentateur corrige la réponse du sujet lorsqu'elle est erronée. Le test consiste ensuite, en un classement, par le

[97] voir note précédente concernant le codage. Par ailleurs les catégories «docteur» ou «policier» sont bien des catégories artificielles, on aurait aussi bien pu parler de «garagistes» et de «pompiers», ou de «X» et de «Y».

sujet, d'un certain nombre de visages, dont les deux items critiques. Il les classe en les plaçant au-dessus de l'un des deux uniformes.
En apprentissage incident, la démarche est plus «sournoise» : il n'est pas dit au sujet qu'il doit apprendre. Au contraire, l'expérimentateur, lui présente une série de visage en les plaçant au-dessus de l'un des deux uniformes et lui demande simplement de dire s'il a déjà vu le visage en question. Comme l'expérimentateur est amené à présenter plusieurs fois le même visage, le sujet n'est pas étonné et croit qu'il s'agit d'une expérience sur la mémoire. Lors du test, l'expérimentateur lui précise alors que les visages qu'il a vu appartenaient soit à des docteurs soit à des policiers selon l'uniforme sur lequel il les posait. Le sujet doit donc classer les visages qu'il va lui présenter maintenant dans l'une ou l'autre de ces classes en les plaçant au-dessus de l'uniforme adéquat.

Ainsi dans les deux types d'apprentissages, les tests sont identiques. Ce sont les phases d'apprentissage qui diffèrent. En apprentissage intentionnel le sujet sait qu'il doit apprendre, en apprentissage incident, le sujet apprend à son insu.

Pour les résultats, Kemler-Nelson considère que les sujets ont appris lorsqu'ils ont 20 réponses correctes sur les 30. 70% du groupe de sujets en apprentissage incident a ainsi appris contre 94% du groupe en apprentissage intentionnel. Sur les deux groupes, Nelson a pu obtenir 16 sujets ayant appris.
Lors du test, les 2 items critiques apparaissent chacun 5 fois. Le nombre moyen de classements unidimensionnels est de 7,9/10 pour le groupe intentionnel et de 3,5/10 pour le groupe incident. Le nombre moyen de classements par «air de famille » est donc de 2,1/10 pour le groupe intentionnel et de 6,5/10 pour le groupe incident.

Ainsi, contrairement à Medin et al, Kemler-Nelson arrive à reproduire une catégorisation en «air de famille ». Kemler-Nelson respecte deux des points que nous avions signalés plus haut concernant la reproduction du processus de catégorisation en laboratoire : l'environnement impose au sujet les catégories qu'il doit former et le processus de catégorisation est un processus d'apprentissage.

Il faut cependant noter que la catégorisation en air de famille est nettement plus forte pour le groupe en apprentissage incident que pour le groupe en apprentissage intentionnel. Nous serons amené à revenir sur ce point lorsque nous relaterons l'échec de notre deuxième expérience à la fin de ce chapitre.

Si comme dans Medin et al, les stimuli ne sont pas très complexes, nous pouvons cependant constater que c'est dans la plus difficile des deux situations (apprentissage incident) que l' «air de famille » apparaît le plus. Par ailleurs, la procédure employée par Kemler-Nelson est moins critique relativement à cette complexité car le fait de mettre les exemples tests dans l'une ou l'autre catégorie détermine le type de catégorisation. Il aurait néanmoins été intéressant de voir ce que cela aurait donné si la situation avait été beaucoup plus complexe. En effet, on peut noter que même dans le cas de

l'apprentissage intentionnel, certains sujets classent par «air de famille », il est possible que si la situation avait été plus complexe, ce nombre de sujets aurait été plus grand.

5.2.3 Fréquence d'instantiation et distributions de probabilités

Dans Medin et al comme dans Kemler-Nelson, il y a un facteur implicite dans leur expérimentation, il s'agit de la fréquence d'instantiation, c'est-à-dire, la présentation ou non de l'exemple avec son appartenance à la classe. Dans le cas de Medin et al, cela a moins de sens puisqu'il n'y pas apprentissage, néanmoins les exemples proposés ne sont pas quelconques. Chez les deux chercheurs, les exemples de leurs premières expériences sont construits sur 4 variables binaires ce qui autorise 16 exemples : 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111

Cependant, seulement 10 de ces 16 exemples seront présentés :

Catégorie I : 0000,0100,0010,0001 Catégorie II : 1111,1011,1101,1110

Items critiques : 0111, 1000

Durant l'apprentissage, Kemler-Nelson propose ainsi 24 exemples au sujet qui sont les 8 exemples des deux catégories présentés 3 fois. Cela correspond à une distribution qui donne une probabilité de 0% aux exemples 1100, 1010, 0011, 0101, 0110, 1001. Les exemples critiques ont un statut un peu particulier puisqu'ils n'apparaissent pas dans l'apprentissage.

Ainsi on ne présente pas tous les exemples possibles définis sur les valeurs des variables. Cela s'explique par la construction des exemples autour d'un prototype. Certains exemples ne peuvent apparaître car ils sont trop éloignés du prototype. Les exemples 1100, 1010, 0011, 0101, 0110, 1001 n'ont que deux valeurs communes avec l'un ou l'autre des prototypes. Le fait de ne pas inclure ces exemples, revient à admettre implicitement qu'ils ne permettraient pas une construction en «air de famille ». Autrement dit, à reconnaître que la fréquence d'instantiation intervient dans la formation des catégories.

Nous avons vu dans le 2.6.2, que ce problème a été étudié, entre autres, par [Barsalou, 85]. Barsalou cherche à savoir si la fréquence d'instantiation est un déterminant (parmi d'autres) de la typicalité. Pour cela, il demande au sujet d'indiquer sur une échelle si les items de telle catégorie lui paraissent plus ou moins fréquents. La question que pose cette démarche est celle-ci : le critère subjectif de fréquence d'intantiation correspond-il bien à la réalité ? Dans quelle mesure les rencontres les plus récentes ne jouent pas dans l'estimation de l'individu lorsqu'il affirme qu'il a rencontré plus de bergers allemands que de labradors dans sa vie ? Pour démontrer que son estimation est correcte, il faudrait pouvoir redérouler sa vie. Une étude du facteur fréquence d'instantiation par interrogation du sujet est donc sujette à caution.

Si la fréquence d'instantiation, la distribution de probabilités sous-jacente aux exemples rencontrés par l'individu, est un facteur de la catégorisation, il semble normal de l'intégrer comme tel lorsque l'on étudie le processus de catégorisation en laboratoire. Cela a été fait, mais d'une manière autre, par Flannagan M.J., Fried L.S, Holyoak K.J. en 1986 dans *Attentes sur les distributions et l'induction de structure de catégorie* [Flannagan, Fried, et Holyoak, 1986]. Leur article vise à vérifier si les sujets ont des attentes relativement aux distributions de probabilités. Il surprend car *la catégorie ne s'y définit que par sa distribution de probabilités.*

Ainsi, dans leur expérience, la catégorie à apprendre consiste dans les tableaux d'un peintre abstrait appelé Vango. Ces tableaux varient selon trois attributs (x, y, z) correspondant à des figures dans les tableaux. Comme pour chaque attribut, il y a dix valeurs possibles, cela fait qu'il peut y avoir alors 1000 tableaux différents possibles. Avec ces trois attributs, les auteurs construisent une fonction f(x,y,z) qui associe à chacun des mille tableaux un nombre entre 1 et 10 qu'ils appellent «stimulus dimension». Ainsi plusieurs tableaux peuvent avoir le même « stimulus dimension» selon cette fonction. C'est à partir de cette fonction que va se définir, en terme de probabilités, l'appartenance au concept. Ils vont ainsi tester 3 groupes de sujets sur 3 distributions de probabilités différentes NL, NH, U («normal low-mean », «normal high-mean » et «U-Shaped ») que l'on voit dans la figure 5.2.3.1 ci-dessous. Ces distributions correspondent à trois définitions probabilistes du concept « tableaux de Vango».

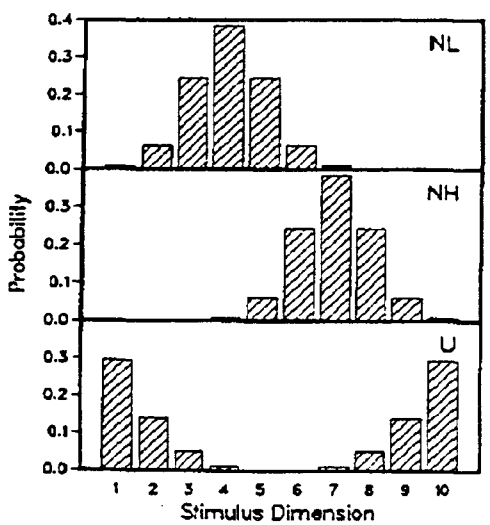


Figure 5.2.3.1 les fonctions de probabilité utilisées pour définir les catégories NL, NH et U [Flannagan, Fried, et Holyoak, 1986]

Chacun des trois groupes affronte donc une distribution (concept) particulière. Pour un premier groupe les tableaux de Vango seront définis par la distribution NL, pour un second par NH et pour le troisième par U.

Lors de l'apprentissage, à chaque groupe on présente un certain nombre de tableaux de Vango. Ainsi, pour le groupe avec la distribution NL, près de 40% des exemples sont tels que f(x,y,z)=4 alors qu'il n'y a aucun exemple tel que f(x,y,z)=8 ou 9 ou 10. En terme d'appartenance, nous pourrions traduire ceci en disant que les tableaux de la forme f(x,y,z)=8 ou 9 ou 10 ne sont pas peints par Vango, mais, et c'est ici que cela se complique, qu'un tableau tel que f(x,y,z)=4 n'a qu'une certaine probabilité d'avoir été peint par Vango : on ne peut pas en être sûr.

Lors du test, on tire 125 tableaux cette fois-ci selon la distribution uniforme, chaque tableau a ainsi une chance égale d'être présenté. On demande aux sujets de les étiqueter en exemples positifs ou négatifs (peints par Vango ou non). Bien entendu, comme le concept est défini par sa distribution de probabilités, ce n'est pas la manière qu'ont les sujets d'étiqueter *chaque* tableau qui compte mais la manière qu'ils ont d'étiqueter *tout l'échantillon.* Par exemple avec la distribution NL, on s'attend à ce que parmi les exemples étiquetés comme positifs, relevant du concept, on trouve 40% de tableaux tels que f(x,y,z)=4 et aucun des tableaux tels que f(x,y,z)=8 ou 9 ou 10. Autrement dit, pour un f(x,y,z) donné peu importe que ce soit tel tableau plutôt que tel autre qui soit étiqueté positif, l'important c'est que la probabilité soit respectée.

Ils obtiennent ainsi les résultats suivants :

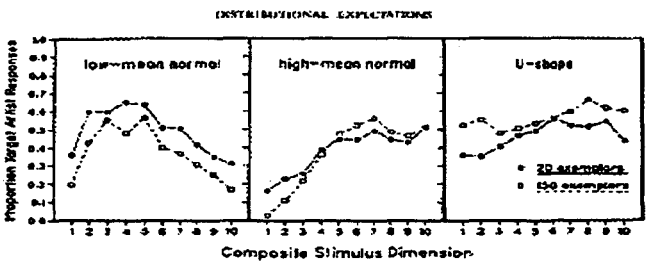


Figure 5.2.3.2 catégorisation effectuée par les sujets [Flannagan, Fried, et Holyoak, 1986]

Flannagan et al, comparent ces courbes-ci avec les courbes précédentes. Ils constatent que les courbes «normales» sont relativement bien apprises tandis que les courbes en U ne le sont pas. Ils en déduisent que les sujets ont une certaine attente des distributions de probabilités.

Dans cet article, nous sommes bien dans une situation d'apprentissage où l'environnement impose à l'apprenant la catégorie à former. Il s'agit même d'un type d'apprentissage particulier car *seuls* les tableaux du peintre sont présentés pendant la

phase d'apprentissage, ce que l'on appelle apprentissage par exemples positifs dans le modèle PAC.

Comme nous l'avons dit, ce qui déroute dans leur article est la définition de la catégorie exclusivement par sa distribution de probabilités. Pour un tableau donné, il est impossible de définir son appartenance au concept, c'est l'étiquetage complet de l'échantillon qui doit respecter la distribution de probabilités.
Alors que jusqu'à présent, nous pouvions reprocher aux articles le fait que la distribution de probabilités n'était pas réellement prise en compte, c'est maintenant le concept qui semble un peu omis. Les sujets, ici, apprennent une distribution de probabilités plutôt qu'un concept

5.2.4 L'intérêt d'opérationnaliser le modèle PAC pour la psychologie

La catégorisation est un processus et un résultat. Le processus est l'apprentissage, c'est-à-dire une adaptation de l'individu à son environnement. Le résultat est la représentation que l'individu se fait des catégories, des «découpages » que son adaptation l'a obligé à faire dans le monde. Les psychologues s'intéressent d'abord aux résultats, aux *représentations* qu'a un individu des catégories. Néanmoins, pour vérifier une hypothèse, ils peuvent être amenés à se préoccuper du *processus* et, soit testent le sujet sur des catégories artificielles, soit l'interrogent sur son vécu. Ces deux démarches peuvent poser problème.

Ainsi Medin et al, ont voulu vérifier en laboratoire sur des catégories artificielles qu'effectivement lorsque l'on demandait à des sujets de classer des objets, ils le faisaient selon un air de famille et non selon des conditions nécessaires et suffisantes. L'échec de leurs expériences s'explique partiellement par les raisons qu'ils donnent, mais aussi parce qu'ils ont assimilé le processus de catégorisation à un simple processus de classement. Ils n'ont pas tenu compte que la catégorie est *imposée* par l'environnement à l'apprenant et que la représentation que s'en fait le sujet est le résultat d'un *apprentissage*. Par ailleurs les situations qu'ils proposent sont insuffisamment *complexes*. Les résultats positifs de Kemler-Nelson proviennent à notre avis de ce qu'elle a pris en compte les deux premiers points.

Kemler-Nelson comme Medin et al, ne considèrent pas, en tant que *facteur* dans leur expérimentation, *les distributions de probabilités* sous-jacentes à leur présentation d'exemples. Cela a été fait par d'autres auteurs (cf. [Barsalou, 85] entre autres) au travers de l'interrogation du sujet sur la familiarité qu'il a avec les items d'une catégorie. Cette démarche est trop empreinte de subjectivisme. D'autres chercheurs ont étudié les distributions de probabilités sur des catégories artificielles mais malheureusement dans leurs travaux, les catégories ne se définissent que par ces seules distributions.

Essayer donc de mettre à jour le processus de catégorisation au travers de l'interrogation du sujet sur son passé paraît trop subjectif. A l'opposé, étudier le processus sur des catégories artificielles pose problème. Nous savons depuis les travaux de Rosch,

l'importance des dimensions écologiques de la catégorisation. Le problème est que ces dimensions sont rarement introduites dans les expérimentations sur les catégories artificielles. En général, pour que celles-ci ne paraissent pas trop artificielles, on donne aux stimuli un semblant de forme qui permet d'apparenter la catégorie artificielle à des catégories déjà connues, par exemple des animaux fictifs, des visages stylisés. Cette liaison entre une catégorie artificielle et une catégorie naturelle sert alors de caution écologique à l'expérience.

Ce qu'il faut donc pouvoir faire, c'est introduire dans l'étude sur des catégories artificielles de réelles dimensions écologiques. Il faut pouvoir tenir compte que l'apprentissage humain ne se déroule pas n'importe comment. L'être humain est le produit de son histoire, de sa rencontre avec les différents objets qu'il a catégorisés. D'un individu à l'autre, les objets ne sont pas les mêmes et si, dans un groupe socioculturel donné, on trouve chez les individus des représentations similaires cela est dû à la similarité de leur histoire. Par ailleurs, les catégories que l'individu a établies, l'ont été sous la contrainte de l'environnement. Ce n'est pas l'individu qui décide d'appeler un chat un «chat », c'est son environnement qui l'y contraint. De manière identique, c'est par une nécessaire adaptation à son environnement que l'individu décide de classer des sections de routes en dangereuses ou non. Enfin l'environnement est complexe, il ne se résume pas à une dizaine d'exemples pour une catégorie donnée.

Nous pouvons ainsi distinguer cinq aspects principaux, dans la catégorisation écologique :
- le processus de catégorisation est un apprentissage, une adaptation à l'environnement,
- l'environnement impose à l'apprenant le regroupement des objets en certaines catégories,
- l'apprentissage est approximatif, l'adaptation n'est pas optimale,
- l'apprenant ne rencontre pas n'importe quels objets d'une catégorie, ni tous les objets de celle-ci,
- l'environnement est complexe.

Ces cinq aspects sont inhérents au modèle PAC. Celui-ci considère l'apprentissage supervisé, l'étiquetage d'un objet dans une catégorie donnée est imposé à l'apprenant. L'apprentissage est approximatif, nous pouvons considérer la borne ε qui marque les limites de cette approximation comme celle marquant l'adéquation de l'adaptation à l'environnement. Au travers des distributions de probabilités, le modèle PAC reproduit la fréquence de rencontres du sujet avec les différents objets de la catégorie. Enfin, il permet des calculs de complexité.
L'opérationnaliser, c'est-à-dire le transformer en modèle d'expérimentation, devrait donc permettre d'introduire ces cinq dimensions écologiques dans les expérimentations sur des catégories artificielles.

5.3 Opérationnalisation

Comme nous l'avons expliqué, l'opérationnalisation du modèle PAC en psychologie consiste à transformer la définition du PAC apprentissage en un algorithme d'expérimentation en psychologie.

En informatique, un des objectifs est de montrer la plus grande pertinence cognitive d'une modélisation de l'apprentissage sur une autre. Par exemple, l'opérationnalisation devra permettre de mettre en place un protocole expérimental qui montrera que selon les distributions de probabilités certaines classes de concepts sont plus facilement apprenables par les sujets.
En psychologie, l'opérationnalisation a pour but de recréer en laboratoire le processus de catégorisation, *la formation* de catégorie, en introduisant certaines dimensions écologiques, de proposer un cadre assez général où ces dimensions seront clairement définies. Par exemple, elle devra permettre d'étudier l'impact de la distribution de probabilités, de la fréquence d'instantiation, sur les représentations que se font les sujets d'une catégorie et plus particulièrement sur le poids des caractéristiques utilisées dans cette représentation.

Nous allons maintenant analyser la définition du modèle d'apprentissage PAC de façon à transformer ses principaux concepts en ce que l'on appelle, en psychologie, les variables expérimentales.

Définition de l'apprentissage PAC
Soit C une classe de concepts sur X et H une classe de représentations. On dit que C est PAC apprenable s'il existe un algorithme L avec la propriété suivante : pour tout concept $c \in C$, pour toute distribution de probabilités D sur X, et pour tout $0<\varepsilon<1/2$ et $0<\delta<1/2$ si L a accès à EX(c,D) et aux entrées ε et δ, alors avec une probabilité d'au moins $1-\delta$, L retourne une hypothèse $h \in H$ satisfaisant erreur(h)$\leq\varepsilon$.
Si L tourne en temps polynomial en taille(c), $1/\varepsilon$ et $1/\delta$ on dit que C est efficacement PAC apprenable.

5.3.1 Un espace d'hypothèses H et un algorithme L : le Sujet

En premier lieu, nous considérons conjointement les deux variables expérimentales H et L, car elles relèvent de ce qui est au centre de l'expérimentation en psychologie : le sujet. Concrètement, nous envisageons le sujet comme un espace d'hypothèses et un processus/algorithme.

Comme nous l'avons expliqué dans le chapitre 1, le «il existe un algorithme capable de... » est au cœur des théories formelles de l'apprentissage. Dès que les informaticiens ont trouvé un tel algorithme pour une classe de concepts, ils peuvent affirmer que cette classe est apprenable dans les conditions définies par le modèle.
La problématique est autre en psychologie. Cet algorithme est une description partielle de l'apprenant. Il correspond à un processus cognitif, à une compétence au sens Chomskyen du terme. Pour que les psychologues puissent généraliser à tout individu, il faut que cette compétence apparaisse chez tous les sujets. La traduction du «il existe un algorithme capable de... » du modèle PAC devient alors en psychologie «chez *tout* sujet, il existe une compétence telle que... ».
Notons qu'en psychologie, *l'existence* de l'algorithme ne pose pas problème. Il est admis que l'homme est capable de catégoriser. Ce qui est inconnu, c'est l'étendue de cette compétence et la manière selon laquelle elle fonctionne : apprentissage implicite/explicite, intentionnel/incident, etc.

Une autre de leurs quêtes est *l'espace d'hypothèses H, c'est-à-dire le type de représentations que se construit l'apprenant*. Le parti que nous avons pris ici est de faire le moins de suppositions possibles sur cet espace. La définition du modèle PAC, proposée ci-dessus, le permet puisqu'elle correspond au modèle prédictif : apprentissage de C dans H. Pour opérationnaliser le modèle PAC, la seule condition que nous devons poser est qu'il existe dans H une hypothèse proche du concept cible. Par contre, comprendre la nature des représentations que se font les sujets est primordiale pour les psychologues ; c'est ici que se situent les problématiques sur les structures de ces représentations : en conditions nécessaires et suffisantes, par exemplaires, par prototypes, etc. L'opérationnalisation du modèle PAC n'obligeant pas le psychologue à opter, a priori, pour un schéma de représentation ou un autre, elle lui laisse ainsi toute latitude pour expérimenter dans le sens qu'il souhaite.

Ce n'est pas parce que le sujet en psychologie est assimilé à l'espace d'hypothèses et l'algorithme dans le modèle PAC que l'adéquation est totale. Cela est évident avec la traduction du «il existe un algorithme capable de... » en «chez *tout* sujet, il existe une compétence telle que... ». Nous devons «ajouter » le «sujet » au modèle PAC.
Pour que les résultats puissent être généralisés et ne pas s'expliquer par les caractéristiques particulières de l'individu qui a passé l'expérience, les expérimentations se font avec des groupes de sujets. Il faut que les variations de la variable dépendante se retrouvent chez tous les sujets du groupe (ou une majorité d'entre eux) ayant subi les mêmes variations du facteur.

5.3.2 pour tout concept $c \in C$

La classe de concepts

Une des caractéristiques des théories formelles est de pouvoir définir la classe de concepts : l'appartenance n'y pose pas problème. La psychologie n'a malheureusement pas cette chance. Le «haricot vert» est généralement considéré comme un «légume » et

non comme un «fruit ». Pourtant, lorsque l'on se réfère au dictionnaire, le «haricot vert » est d'abord un «fruit », certains fruits étant aussi des «légumes ». Ainsi l'étude de la catégorie «fruit ou légume » peut s'en trouver biaiser ; quelle définition doit-on prendre : celle du dictionnaire ou celle qui domine dans la culture ? L'erreur pourrait consister à supposer que l'approximation de la représentation vienne du sujet alors qu'elle relève de la culture dans laquelle il baigne. En étudiant le *processus* de la catégorisation en laboratoire, ce problème est évacué car la catégorie doit être *nouvelle* pour le sujet. Ainsi, avec *les catégories artificielles*, il n'y a pas d'ambiguïté concernant l'appartenance, étant donné que c'est l'expérimentateur qui les définit. L'expérimentation de Kemler-Nelson ci-dessus en est un exemple, la catégorie y est clairement décrite par son extension.

Pourtant, il convient, dans le cadre d'une opérationnalisation du modèle PAC, de définir des catégories et des stimuli qui permettent un *accroissement de la complexité* sans pour autant changer la nature de ceux-ci. Si l'on prend les caricatures de visages de Nelson ceux-ci sont définis sur 4 attributs : le nez, les cheveux, la moustache, la couleur des yeux. Si nous voulons établir une comparaison avec des catégories définies sur 30 attributs, il sera difficile de trouver pour un visage 30 attributs qui aient la même pertinence et la comparaison s'en trouvera faussée.

Par ailleurs, il est souhaitable que toute l'information provienne des exemples, que la catégorie soit aussi artificielle que possible. Nous avons vu ci-dessus, avec les expériences de Medin et al, que l'introduction de variables «culturelles » pouvaient entraîner un biais à l'expérimentation car l'on n'étudiait plus des catégories formées durant l'expérimentation, mais des catégories établies avant cette expérimentation. Prôner davantage d'artifice dans la catégorie ne revient pas à nier les dimensions écologiques du processus de catégorisation mais au contraire doit permettre de repérer plus facilement quelles sont ces dimensions.

Pour tout concept c de C

L'apprentissage doit réussir pour *tout* concept de C. Cette exigence, tout à fait légitime dans le cadre des théories formelles, est impossible à tenir dans le cadre d'expérimentations en psychologie, elle pose ainsi un problème méthodologique. Si on supprime cette condition les résultats perdent en généralisation. Il convient donc de mettre le concept en paramètre de l'expérience et de tenir compte, dans les résultats, de ce paramètre. Le concept devient alors un facteur de l'expérimentation.
Par ailleurs, ceci est moyennement gênant dans le cadre de l'étude de la catégorisation car la généralisation n'est pas de même nature en psychologie. En psychologie, la généralisation porte moins sur les *concepts* d'une classe donnée que sur les représentations *des individus* relativement à cette classe. La classe de concepts cible est souvent un prétexte.[98]

[98] Ceci ne signifie pas que les chercheurs ignorent l'importance du *type de classe* de concepts : ainsi en va-t-il de la différence entre catégories artefacts opposées aux catégories naturelles (voir [Keil, 89]). Ils n'ignorent pas non plus que la catégorie « baleine » n'a probablement pas la même représentation que la

5.3.3 pour toute distribution de probabilité D sur X : la fréquence d'instantiation

Le problème du quantificateur universel se retrouve aussi avec les distributions de probabilités : la définition réclame que l'apprentissage réussisse pour *toute* distribution de probabilités. Nous avons vu dans les chapitres 3 et 4, avec le modèle d'apprentissage PAC avec distributions bienveillantes, que cette condition était remise en question. Cependant, même dans ce modèle, il est supposé que l'apprentissage ait lieu pour *toutes* distributions bienveillantes. Le quantificateur universel est déplacé d'un degré, il n'est pas éliminé. Comme pour les concepts, la solution va consister à mettre les distributions de probabilité en paramètre de l'expérimentation et à tenir compte de ce paramètre dans les résultats. Ce sera d'ailleurs l'objet des expérimentations décrites plus loin d'étudier l'impact de ce paramètre sur l'apprentissage.

5.3.4 S a accès à EX(c, D) et aux entrées ε et δ,

L'accès à EX(c,D) est naturel, il décrit l'accès de l'apprenant à l'environnement qui lui soumet des exemples. EX(c,D) constitue une autre des dimensions écologiques de l'apprentissage puisque la procédure impose à l'apprenant l'étiquetage des exemples, elle lui impose la catégorie qu'il doit former. Dans le cas d'une opérationnalisation du modèle PAC, l'expérimentateur jouera le rôle de cette procédure. D'un côté, il définit le concept cible et peut étiqueter sans erreur les exemples, de l'autre, il définit la distribution de probabilités avec laquelle il tire les exemples.

Concernant les paramètres ε et δ, nous avons vu dans le chapitre 1 qu'ils sont à la base de la définition du modèle PAC : probablement et approximativement correct. Alors qu'ε exprime le côté approximatif de l'apprentissage, δ exprime son côté probable. Nous avons vu, dans ce même chapitre, que le paramètre δ est nécessaire car il exprime la possibilité que le tirage réel des exemples ne soit pas aléatoire relativement à la distribution de probabilités théorique qui a présidé à ce tirage. A partir du moment où l'on opérationnalise le modèle PAC, l'expérimentateur peut toujours contrôler que le tirage réel n'est pas pathologique. A partir de là, il n'est plus nécessaire de conserver ce paramètre δ dans les expérimentations. Le paramètre d'approximation ε, ramené à l'apprentissage naturel, exprime le fait que l'hypothèse de l'apprenant puisse être proche du concept cible sans être identique à ce concept. A ce titre, il représente la dimension écologique selon laquelle un apprentissage approximatif est suffisant pour s'adapter à l'environnement : tout individu qui considère que le «haricot vert » n'est pas un «fruit » ne se verra pas handicapé par cette méconnaissance dans sa vie. L'introduction de ce paramètre dans les expérimentations est donc indispensable pour permettre une certaine latitude dans la représentation que se fait un individu d'une catégorie.

---

catégorie « vache », ces deux catégories appartenant à la classe des mammifères. En mettant le concept en paramètre, on permet justement d'étudier les différences.

Dans le modèle PAC, l'apprenant a connaissance de ce paramètre ε, cette connaissance lui permettant de savoir quand il doit arrêter l'apprentissage et le nombre d'exemples nécessaires. Nous avons vu, en note dans le chapitre 1, qu'il existe des variantes du modèle (les modèles fonctionnels) où l'apprenant n'a pas accès à ce paramètre, où c'est l'oracle qui calcule le nombre d'exemples nécessaires. Une telle démarche n'est pas possible dans le cas de l'opérationnalisation car le nombre d'exemples sera utilisé comme variable dépendante, celle que l'expérimentateur utilisera pour mesurer l'impact sur l'apprentissage des variations du facteur qu'il aura choisi. Supposer qu'il connaisse a priori ce nombre revient à vider l'expérience de sa raison d'être. Cependant faire supporter par l'apprenant le poids du calcul du nombre d'exemples qui lui sont nécessaires pour apprendre sans trop d'erreurs pourra interférer avec l'apprentissage. La solution est donc que le sujet ait accès à une procédure TEST qui lui soumet des exemples qu'il doit étiqueter. Cette procédure calcule alors le pourcentage d'erreurs et indique à l'apprenant s'il doit continuer ou non l'apprentissage sans donner toutefois d'information sur les erreurs qu'il a commises.

5.3.5 Le sujet retourne une hypothèse h satisfaisant erreur(h)≤ε

«Le sujet retourne une hypothèse h satisfaisant erreur(h)≤ε » pose le problème de l'évaluation de l'hypothèse. Comme nous travaillons en prédiction, l'hypothèse est évaluée en demandant à l'apprenant de classer des exemples, de prédire l'appartenance de ces exemples, la procédure TEST peut alors aussi servir d'évaluateur. Cette procédure devra tenir compte de la distribution de probabilités car c'est l'apport principal du modèle PAC de lier le poids de l'erreur à la distribution de probabilité. La question est alors de connaître le nombre d'exemples qui est nécessaire à cette procédure. Nous devons envisager cette question d'un double point de vue. Du point de vue théorique, il faut que la procédure nous garantisse avec une assez grande fiabilité que l'apprentissage a réussi ou a échoué. Du point de vue psychologique, pour éviter la fatigue du sujet qui peut être amené à rencontrer plusieurs fois la procédure TEST, il faut que le nombre d'exemples utilisés pour le test ne soit pas trop grand. Nous verrons plus loin qu'en s'appuyant sur la loi binomiale, il est possible d'atteindre ces deux objectifs.

5.3.6 L'apprentissage doit se dérouler en un temps polynomial en taille(c), 1/ε et 1/δ.

La notion de temps polynomial en taille(c), 1/ε *et* 1/δ et d'échantillon polynomial est difficile à transposer dans une expérimentation : tout temps d'expérimentation est polynomial. Il faut donc remplacer cette condition par la condition en un temps «raisonnable » et un nombre «raisonnable » d'exemples. Mais raisonnable signifie nécessairement arbitraire car l'expérimentateur ne peut attendre que le sujet ait la gentillesse de bien vouloir décéder pour pouvoir affirmer qu'il a échoué. De ce fait, le temps et le nombre d'exemples doivent être indiqués dans les résultats pour limiter leur

portée. Par ailleurs, cela va dans le sens d'une étude *quantitative* de l'apprentissage conformément au souhait de [Pitt, 97].

5.3.7 avec une probabilité d'au moins 1-δ,

La condition *avec une probabilité d'au moins* 1-δ disparaît de l'expérimentation en même temps que l'on fait disparaître δ. Puisque l'expérimentateur peut contrôler que les tirages des exemples pour l'apprentissage et pour le test sont aléatoires selon la distribution de probabilités, il n'y a plus de raison que l'apprentissage puisse échouer suite à une possible distorsion entre les deux. Nous ne nous attendons plus à un apprentissage probablement approximativement correct mais à un apprentissage approximativement correct.

5.3.8 la procédure TEST

Nous avons vu dans le 5.3.4 que le modèle PAC considère que le sujet est informé de la borne maximale d'erreur tolérée pour son hypothèse. Cette information doit lui permettre de calculer le nombre d'exemples nécessaires pour fournir une hypothèse approximativement correcte. Dans le cadre d'expériences, il ne faut pas que ce calcul interfère avec l'apprentissage, il est donc nécessaire de proposer au sujet une procédure qui lui indique si son hypothèse est approximativement correcte. Cette procédure est aussi utile dans le cas d'apprentissage prédictif où l'on vérifie que le sujet a trouvé la bonne hypothèse en lui demandant de classifier des exemples. Le problème qui se pose alors est celui-ci : *quelle confiance peut-on avoir dans la procédure TEST* sachant que celle-ci ne peut proposer un nombre très grand d'exemples tests ?

De fait TEST est sujette à deux types d'erreurs possibles. Rappelons que cette procédure soumet à l'apprenant un certain nombre d'exemple et lui demande de les étiqueter. Si l'étiquetage a un nombre d'erreurs inférieur à un certain pourcentage, l'hypothèse est acceptée, sinon elle est rejetée. Les deux types d'erreurs possibles[99] sont alors que la procédure :
- rejette une hypothèse approximativement correcte,
- accepte une hypothèse erronée (telle que l'erreur est supérieure à la borne d'erreur tolérée)

Si TEST peut accepter des hypothèses erronées ou rejeter des hypothèses approximativement correctes, cela est dû à ce qu'elle n'a droit qu'à un nombre limité de tirages. Plus la procédure TEST tire d'exemples et plus il y a de chances que le tirage soit proche de la distribution théorique, et que l'erreur constatée soit proche de l'erreur réelle. Le problème est que l'on ne peut pas soumettre le sujet à un nombre très grand d'exemples. La question est alors de savoir combien il faut d'exemples pour que nous puissions avoir une certaine confiance dans l'erreur constatée.

---

[99] On trouvera en annexe 3 une illustration de ces types d'erreurs et une présentation plus générale.

C'est ici qu'intervient la loi binomiale (voir annexe 4). Elle garantit que si l'on teste le sujet sur 25 exemples :

0 erreur observée correspond à une erreur réelle comprise entre 0 et 12%
1 erreur observée correspond à une erreur réelle comprise entre 0 et 18%
2 erreurs observées correspondent à une erreur réelle comprise entre 1 et 24%

et ceci avec *une confiance de 90%.*

Choisir ainsi 1 erreur sur 25 semble un bon compromis. Nous avons la garantie, avec une confiance de 90%, que l'erreur de l'hypothèse du sujet ne dépassera pas 18%. L'application de la loi binomiale garantit que l'on évite un des deux types d'erreurs que peut commettre TEST : nous obtenons (avec une garantie assez grande, 90%) que TEST n'accepte pas une hypothèse dont l'erreur est supérieure à 18%.

Par contre, nous n'avons aucune garantie qu'une hypothèse approximativement correcte ne sera pas rejetée. Notons que TEST ne rejettera jamais une hypothèse correcte puisque celle-ci étiquettera les exemples de la même manière que le concept. Le problème se pose quand l'hypothèse est approximativement correcte. Toutefois, le fait qu'il s'agisse, ici, d'expérience avec des *groupes* de sujets permet d'ignorer partiellement le problème. Pour expliquer ceci, il faut anticiper sur la description de l'expérience. Dans celle-ci, le sujet apprend tant qu'il n'a pas atteint le temps limite ou tant qu'il n'a pas passé le test avec succès. S'il échoue au test, l'apprentissage reprend. Dans le cas où TEST refuse une hypothèse approximativement correcte, cela rallonge le temps d'apprentissage et le nombre d'exemples nécessaires. Il faut noter que TEST ne refusera pas systématiquement toutes les hypothèses approximativement correctes du sujet, mais qu'il y a une certaine probabilité qu'elle en rejette une. Il est donc probable que le sujet réussira l'apprentissage mais, dans le cas où TEST rejette son hypothèse approximativement correcte, son apprentissage prendra plus de temps et d'exemples.
Par ailleurs, nous travaillons avec des groupes de sujets, et non avec un sujet isolé, nous pouvons donc calculer la moyenne dans le groupe, du nombre d'exemples d'apprentissage utilisés. Cela atténue encore l'erreur de TEST
Enfin et surtout, nous ne nous contentons pas des résultats sur *une* modalité mais nous étudions et *comparons* les résultats sur les *diverses* modalités du facteur et pour chacune de ces modalités il y a un groupe de sujets. On peut, ainsi, supposer que pour chaque modalité, TEST a probablement refusé le même nombre d'hypothèses approximativement correctes et fait augmenter de manière similaire le nombre d'exemples nécessaires. Les résultats *relatifs, comparatifs,* restent alors valables[100].

[100] Cependant, il est toujours possible que, pour un sujet particulier, le refus d'une hypothèse approximativement correcte le déstabilise, l'amène à quitter une piste de recherche correcte et à échouer.

5.4 Une définition et un algorithme d'expérimentation

A partir de ce qui précède, nous pouvons envisager de proposer une définition *possible* de l'apprentissage PAC en psychologie.

5.4.1 Une définition possible de l'apprentissage Approximativement Correct en psychologie

Soit c un concept d'une classe de concepts C définie sur X, D une distribution de probabilités, S un groupe de sujets et T un temps limite.
On dit que c est Approximativement Correctement appris avec la distribution de probabilités D et avec une erreur maximale ε dans un temps inférieur à T, si pour tout s de S,
s ayant accès à :
-EX(c, D) qui lui présente des exemples étiquetés selon le concept c et tirés selon la distribution D,
-TEST qui lui présente N exemples à étiqueter, tirés selon la distribution D, arrête l'apprentissage si le sujet a fait au plus M erreurs et relance l'apprentissage sinon.
s est capable, dans un temps t inférieur ou égal à T, de trouver une hypothèse qui lui permette de satisfaire TEST.

Nous notons que la définition a éliminé le «P » de l'apprentissage PAC. Nous ne parlons plus maintenant d'apprentissage Probablement Approximativement Correct mais seulement d'apprentissage Approximativement Correct. Nous avons expliqué dans le 5.3.3 la raison de ceci : l'expérimentateur peut contrôler que le tirage des exemples d'apprentissage n'est pas pathologique en regard de la distribution. Avec la procédure TEST nous réintroduisons une certaine probabilité mais elle n'est pas de même nature. Alors que le «Probablement » dans le modèle PAC exprime la probabilité que l'apprentissage réussisse, la probabilité dans la procédure TEST exprime le fait que l'erreur puisse éventuellement dépasser la borne d'erreur ε. Nous proposons de choisir N égal à 25 exemples et M à 1 exemple ainsi ε est égal à 18%, et la probabilité d'un tel dépassement sera de 10%. Il est en effet difficile d'obtenir moins avec un petit nombre d'exemples de test. Même en prenant 30 exemples de test et en réclamant du sujet qu'il ne se trompe pas une seule fois, l'erreur pourra encore être de 10%. De plus, demander à l'apprenant qu'il ne se trompe pas revient à le pousser vers un apprentissage exact et donc davantage à une représentation en conditions nécessaires et suffisantes.

Nous avons dû abandonner les quantificateurs universels (∀) sur les concepts et les distributions de probabilités. De ce fait nous perdons en généralisation, mais cette perte n'est que relative. Alors que l'objet du modèle PAC est d'étudier l'apprenabilité de classes de concepts, l'opérationnalisation du modèle a pour but l'étude des conditions d'apprentissage et leur impact sur le temps d'apprentissage d'un point de vue expérimental avec des sujets humains. Alors que le premier vise des résultats universels,

la seconde vise des résultats relatifs : que se passe-t-il si on change le concept, la distribution de probabilité.

Nous avons, par contre, ajouté le groupe Sujets qui n'existe pas dans l'apprentissage PAC. Sur ce groupe nous avons placé un quantificateur universel. Demander que l'apprentissage ait lieu pour tous les sujets est peut être trop sévère. En psychologie, on se contente en général qu'une donnée soit vraie pour une majorité significative de sujets.

La définition supporte de tester aussi bien l'apprentissage implicite que l'apprentissage explicite et ceci grâce à la procédure TEST qui permet de ne pas demander explicitement son hypothèse à l'apprenant. Cette condition permet de ne pas «forcer » l'apprenant à former une hypothèse en conditions nécessaires et suffisantes.

Dans cette définition, nous ne parlons pas du protocole de présentations des exemples. D'un point de vue formel, que l'apprenant ait accès à tous les exemples en une seule fois ou que l'Oracle les lui présente séquentiellement dépend des modèles. Du point de vue psychologique, cela peut faire une grosse différence car une présentation séquentielle joue sur les capacités mnésiques de l'apprenant. Il est ainsi possible que l'apprenant au bout du centième exemple propose une hypothèse qui ne soit pas consistante avec le premier exemple présenté simplement parce qu'il a oublié ce premier exemple. Toutes les expérimentations en induction de règle proposent cependant une présentation séquentielle. Cela se comprend car, d'un point de vue écologique, l'apprenant ne rencontre pas en une seule fois les objets d'une même catégorie. C'est le type de présentation que nous adopterons.
Toujours concernant le protocole de présentation des exemples, en théories formelles, il existe différents types d'Oracle (voir 1.3.3.3). Par exemple, plutôt que ce soit l'Oracle qui impose l'exemple à l'apprenant, c'est l'apprenant qui propose un exemple à l'Oracle et celui-ci en retour lui donne l'étiquette. Il est alors précisé dans les résultats le type d'Oracle utilisé. Nous pouvons faire de même, ici, il suffit de préciser le type d'interaction autorisée entre l'expérimentateur et le sujet. Si la définition que nous proposons correspond davantage à l'Oracle «standard », c'est que les interactions précédemment décrites, influencent en quelque sorte les distributions de probabilités sous-jacentes à la présentation des exemples et ce qui nous intéresse est justement d'essayer de cerner le rôle de ces distributions. Si nous nous référons à ce qui a été écrit dans le 2.5.1.3 nous travaillons donc «en réception » et non en «sélection ».
Enfin, nombre d'articles en théories formelles ([Angluin, 80], [Berwick, 86], [Denis, 98]) s'intéressent à l'apprentissage par exemples positifs uniquement, c'est-à-dire que seuls des exemples du concept sont présentés à l'apprenant et aucun contre-exemple. Naturellement lors du test, l'apprenant est interrogé sur des exemples positifs et négatifs. Le fait d'utiliser TEST *pendant* l'apprentissage fausse alors le type d'apprentissage. En effet, dans la procédure TEST, on présente à l'apprenant des exemples positifs et négatifs et, bien qu'ils ne soient pas étiquetés, l'apprenant est donc amené à rencontrer des exemples négatifs.

Nous avons dit que l'opérationnalisation avait pour objectif d'étudier le processus de catégorisation. Ainsi l'intérêt de la définition que nous avons présentée tient dans ce qu'elle intègre dans l'étude de ce processus quatre des cinq dimensions écologiques définies dans le 5.2 :
1) le processus de catégorisation est un apprentissage, une adaptation à l'environnement,
2) l'environnement impose à l'apprenant le regroupement des objets en certaines catégories,
3) l'apprentissage est approximatif, l'adaptation n'est pas optimale,
4) l'apprenant ne rencontre pas n'importe quels objets d'une catégorie ni tous les objets de celle-ci,

Le 1) est dans la définition même : sur la base d'un ensemble de stimuli étiquetés, l'apprenant doit être capable de se former une représentation qui lui permette d'étiqueter d'autres stimuli. Le 2) apparaît avec EX(c,D) qui impose au sujet la catégorie à apprendre. Le 3) se retrouve dans les 18% d'erreur acceptée, ce chiffre précis étant dû à des raisons techniques. Le 4) est obtenu par la distribution de probabilités D. La cinquième dimension consistait dans le fait que l'environnement est complexe. Elle n'apparaît pas directement dans la définition car elle est dépendante du type d'exemples, de stimuli, proposés.

5.4.2 L'algorithme d'expérimentation

Nous proposons ci-dessous un algorithme d'expérimentation qui tient compte de toutes les remarques que nous avons faites précédemment. Cet algorithme est le résultat de l'opérationnalisation du modèle PAC. Nous avons proposé dans le 5.3 que la distribution de probabilité D, le concept cible c, le temps limite T soient mis en paramètres.

Avant de présenter l'algorithme général, il convient de présenter les deux procédures qu'il va utiliser. La première, Ex(c,D), est extrêmement simple : elle tire un exemple de X selon la distribution D, l'étiquette selon c et le retourne à la procédure appelante. Notons que c'est au niveau de cette procédure que l'expérimentateur vérifie que les exemples proposés sont bien tirés aléatoirement selon la distribution de probabilités, ce qui permet d'éliminer le 'probablement' du modèle PAC.

Ex(c,D)
    retourner un exemple de X tiré selon D et étiqueté selon c
fin de Ex

La seconde procédure est TEST que l'on a étudiée dans le 5.3.2. Cette procédure boucle sur 25 exemples (ligne 2). Elle demande ainsi successivement 25 exemples étiquetés à la procédure Ex(c,D) (ligne 3). Elle propose l'exemple, sans son étiquette, à l'apprenant et lui demande de l'étiqueter (ligne 4). Elle compare les étiquettes. Si l'apprenant s'est trompé, elle incrémente la variable erreur (ligne 5 à 7). Si l'erreur est strictement

supérieure à 1 la procédure retourne faux (échec) sinon elle retourne vrai (réussite) (ligne 10 à 14). Nous avons donc la garantie, à 90%, que l'erreur de l'hypothèse de l'apprenant est d'au plus de 18%. Par ailleurs, si nous n'arrêtons pas le test dès que l'apprenant a fait deux erreurs, c'est que cela pourrait lui permettre de repérer les exemples sur lesquels il s'est trompé et de bénéficier ainsi d'une information supplémentaire.

TEST

1. erreur=0; n=1
2. **tant que** n≤25
    3. demander un exemple (x,c(x)) à Ex(c,D)
    4. étiquette(x)=étiquetage de x par le sujet
    5. **si** étiquette(x)≠c(x) **alors**
        6. erreur=erreur+1
    7. **fin de si**
    8. n=n+1
9. **fin tant que**
10. **si** erreur > 1 **alors**
    11. retourner faux
12. **sinon**
    13. retourner vrai
14. **fin de si**

fin de TEST

L'algorithme d'expérimentation est simple lui aussi. Il est constitué de deux boucles. La plus grande (ligne 4 à 21) concerne les sujets, elle ne fait qu'exprimer l'évidence selon laquelle tous les sujets du groupe doivent faire l'expérience. La seconde incluse dans la première (ligne 8 à 15) est l'apprentissage par un sujet du groupe. Cet apprentissage boucle tant que le sujet n'a pas réussi le test (ligne 12) et tant qu'il n'a pas atteint le temps limite. On notera que c'est le sujet qui demande à passer le test (ligne 11), il peut ainsi évaluer son apprentissage. Cela correspond à la connaissance qu'a l'apprenant de ε dans le modèle PAC. L'algorithme stocke le nombre d'exemples nécessaires au sujet s pour son apprentissage ainsi que le nombre total d'exemples utilisé par le groupe et le nombre total d'exemples utilisés par ceux qui ont réussi. Le nombre total d'exemples utilisés par le groupe n'est pas réellement pertinent si certains sujets ont échoué car, pour eux, le nombre d'exemples correspond à ceux qu'ils ont vus dans le temps limite. Le nombre total d'exemples utilisés par ceux qui ont réussi a plus de valeur, il permet de calculer une moyenne par individu. C'est cette moyenne qui pourra être prise en compte dans l'analyse des résultats car elle annule partiellement comme nous l'avons vu, l'erreur possible de TEST. L'algorithme enregistre aussi le nombre de sujets qui ont réussi l'apprentissage.

Algorithme d'expérimentation **EXPSYPAC**(c,D,T,S)
/*c est le concept cible, D la distribution de probabilités, T le temps limite et S le groupe de sujets*/

1. nombre_d_apprentissages_réussis=0
2. nb_exemples_total=0
3. nb_exemples_total_réussi=0
4. **pour** tout s de S /*S est l'ensemble des sujets*/
    5. trouvé=faux;
    6. temps=0,
    7. nb_exemples(s)=0
    8. **Tant que** temps <T **et** non trouvé
        9. nb_exemples(s)= nb_exemples(s) + 1
        10. Présenter 1 exemple étiqueté avec Ex(c,D) /* apprentissage*/
        11. **si** demande du sujet **alors** /*test ?*/
            12. trouvé=TEST
        13. **fin de si**
        14. incrémenter temps
    15. **fin de tant que**
    16. **si trouvé alors**
        17. nombre_d_apprentissages_réussis= nombre_d_apprentissages_réussis + 1
        18. nb_exemples_total_réussi = nb_exemples_total_réussi + nb_exemples(s)
    19. **fin de si**
    20. nb_exemples_total= nb_exemples_total + nb_exemples(s)
21. **fin de pour**
22. n = nombre_d_apprentissages_réussis
23. retourner «il y a n sujets sur card(S) qui ont réussi l'apprentissage du concept c avec la distribution D avec un taux d'erreur maximal de 18% en un nombre total d'exemples de nb_exemples_total_réussi et un nombre moyen d'exemples de nb_exemples_total_réussi /card(S)»

fin de EXPSYPAC

Cet algorithme ne définit pas précisément une expérience, mais plutôt un cadre d'expérience. Il convient que l'expérimentateur fixe les paramètres. Ces paramètres sont les facteurs de l'expérience, fixer ces paramètres, c'est-à-dire leur donner une valeur, revient à fixer la modalité des facteurs. Les variables dépendantes sont alors la réussite ou l'échec et, en cas de réussite, le nombre d'exemples utilisés par les sujets.
Cet algorithme ne concerne qu'*un* groupe de sujets pour *une* modalité de chaque facteur. Ce n'est qu'en le faisant «tourner » plusieurs fois avec différentes modalités qu'il prend du sens. Ainsi, l'algorithme ne répond pas à la question de savoir s'il faut établir un plan d'expérimentation en groupes appareillés ou en groupes indépendants. Supposons que ce soit la variable D, la distribution de probabilités, qui soit étudiée et que l'on envisage

d'étudier trois de ses modalités D1, D2 et D3. Il est possible d'étudier les trois modalités avec le même groupe sujet, ce que l'on pourrait écrire :

Expérimentation 1 étude de D
EXPSYPAC(c,D1,T,S)
EXPSYPAC(c,D2,T,S)
EXPSYPAC(c,D3,T,S)
étude des résultats
fin d'expérimentation 1 étude de D

Le problème avec ce plan d'expérimentation c'est que le groupe sujet bénéficie de l'apprentissage effectué avec la distribution D1 lorsqu'il aborde l'apprentissage avec la distribution D2, ce plan pose des problèmes de transfert. Il est alors possible de lui préférer un plan avec des groupes indépendants de sujets S1, S2 et S3 affectés à chacune des modalités de la variable distribution. Alors que le précédent correspondait à un déroulement en séquence, celui-ci correspond à un déroulement en parallèle.

Expérimentation 2 étude de D
EXPSYPAC(c,D1,T,S1)//EXPSYPAC(c,D2,T,S2) //EXPSYPAC(c,D3,T,S3)
étude des résultats
fin d'expérimentation 2 étude de D

De ce qui précède, il ne faut pas s'étonner que la ligne 23 de l'algorithme corresponde à un constat et non à une analyse des résultats. En psychologie, les chercheurs font varier le(s) facteur(s) et observent les conséquences de cette variation sur la variable dépendante. Ce n'est qu'en comparant les différents constats obtenus en faisant «tourner » plusieurs fois l'algorithme EXPSYPAC avec des paramètres modifiés que les psychologues obtiendront des résultats qui seront pertinents pour eux. Il est plus facile de comprendre pourquoi nous pouvions abandonner sans trop de remords les quantificateurs universels sur la classe de concept et les distributions de probabilité.

Comme pour la définition que nous avons proposée, l'algorithme d'expérimentation intègre les cinq dimensions écologiques que nous avons repérées. La complexité de l'environnement apparaît au travers du paramètre c, le concept. En faisant varier ce paramètre selon des mesures de complexité à définir, il est possible d'en étudier l'impact sur l'apprentissage. Si nous assimilons, par exemple, cette notion de complexité au nombre d'attributs, il faut, pour que seule la complexité puisse être étudiée, que l'accroissement du nombre de variables n'entraîne pas parallèlement l'introduction de variables qui n'aient pas le même degré de pertinence (voir 5.2).

5.5 Deux expériences

L'opérationnalisation décrite précédemment doit permettre de mieux relier les théories formelles de l'apprentissage et notamment le modèle PAC et les études sur la catégorisation en psychologie. Il s'agit maintenant de la mettre en œuvre. Pour cela nous présentons deux expériences basées sur le protocole issu de l'opérationnalisation. *Dans les deux expériences, l'accent est mis sur le rôle des distributions de probabilités dans l'apprentissage*. Mais ces deux expériences sont bien différentes quant à leur objectif.
La première doit permettre de justifier la déclinaison du modèle PAC proposé au chapitre 4. Dans cette déclinaison, nous supposons que l'apprentissage n'a pas lieu pour toutes distributions de probabilités des exemples mais seulement pour les distributions bienveillantes. Cela suppose que selon le type de distribution proposée l'apprentissage s'en trouve ou non facilité. L'objectif, ici, est donc davantage un objectif informatique, dans le sens où l'expérience doit permettre de légitimer un modèle.
La seconde expérience, si elle utilise aussi le protocole issu de l'opérationnalisation, a quant à elle un objectif davantage psychologique : étudier le rôle de la fréquence d'instanciation des items d'une catégorie dans le poids des valeurs d'attributs (au sens de [Pitt, 97]) dans la représentation de cette catégorie chez le sujet.
Nous verrons que ces deux expériences ne vont pas sans poser problème, *on ne s'improvise pas psychologue*. Mais dans le même temps elles sont riches d'interrogations sur ce qu'est la catégorisation et les limitations qu'il faudrait apporter encore au modèle PAC pour que celui-ci en soit bien une expression formelle.

5.5.1 Expérience n°1

Dans un premier temps nous présentons les hypothèses générales et opérationnelles et les variables concernées, dans un second temps nous décrivons la méthode enfin nous donnons les résultats de l'expérimentation. La présentation que nous faisons ne suit pas exactement la présentation standard en psychologie et peut dérouter. Cela est dû à ce que nous avons souhaité reprendre les différents concepts présentés lors de l'opérationnalisation. Cela est dû aussi à l'insistance que nous mettons à décrire les différentes distributions de probabilités utilisées, ce que l'on appelle les différentes modalités du facteur en psychologie.

5.5.1.1 Les hypothèses

Hypothèse générale

Le modèle d'apprentissage PAC standard proposé par Valiant pose comme condition que l'apprentissage ait lieu pour toutes distributions de probabilités des exemples, même des distributions qui n'ont aucun rapport avec la catégorie cible. Cette condition paraît trop forte car elle réclame que l'apprenant soit capable d'apprendre une catégorie

même lorsqu'on ne lui présente que les exemples les plus bizarres de cette catégorie ou que des contre-exemples, etc.
En psychologie, la question des distributions de probabilités des exemples est rarement posée lorsque l'on étudie la catégorisation, mais on ne suppose pas que celle-ci se fait pour toutes distributions et on n'utilise jamais dans les expérimentations des distributions aberrantes.
Si le modèle PAC doit permettre de caractériser l'apprentissage naturel (la catégorisation), il convient donc de s'interroger sur la pertinence de cette condition. Est-il en effet nécessaire de réclamer que l'apprentissage ait lieu pour toutes distributions de probabilités si l'on sait qu'en apprentissage naturel l'apprenant ne le fait que pour quelques-unes ? Par ailleurs, nous avons vu, au chapitre 4, un modèle qui ne réclame l'apprentissage que pour certaines classes de distributions et que ce modèle permet d'obtenir de nouvelles classes de concepts apprenables. C'est pourquoi, il paraît utile de montrer que l'apprentissage naturel est dépendant des distributions de probabilités ce qui légitimerait ce nouveau modèle.
De là, l'hypothèse générale : *l'apprentissage naturel est dépendant des distributions de probabilités qui président au tirage des exemples présentés à l'apprenant.*

Hypothèse opérationnelle

L'expérience suit le protocole proposé en 5.4.2. Sur un écran d'ordinateur, on présente au sujet des exemples d'apprentissage étiquetés selon un concept. L'objectif pour lui est d'être capable d'étiqueter 25 exemples de test, en faisant au plus une erreur, après avoir vu un minimum d'exemples d'apprentissage. Il doit demander à passer le test dès qu'il pense être capable de le réussir. En cas d'échec au test, il reprend l'apprentissage. Il peut ainsi repasser plusieurs fois le test mais à chaque fois les exemples de tests changent. Le sujet a échoué s'il n'arrive pas à réussir le test avant le temps limite.
Les sujets sont répartis en trois groupes. Ces groupes se distinguent en fonction de la distribution de probabilités qui sert à tirer les exemples d'apprentissage et de test. *Nous nous attendons à ce que le nombre d'apprentissages réussis soit dépendant de la distribution de probabilités.*

5.5.1.2 Les variables expérimentales

Nous reprenons, ici, les différents paramètres envisagés lors de l'opérationnalisation et précisons comment nous les traitons.

La classe de concepts cible

La classe de concepts cible est celle des conjonctions de deux valeurs de variables booléennes. Cette classe est choisie en raison de sa simplicité. Il est connu ([Bruner, Goodnow and Austin, 56], [Neisser et Weene, 62], [Richard, 75]...) qu'elle est apprenable lorsque les exemples sont de petite taille (inférieure ou égale à 6 variables) avec une distribution équitable (*tous* les exemples sont présentés et l'on présente autant d'exemples positifs que de négatifs). Bien que nous prenions des exemples de taille 16 et

que nous fassions varier les distributions de probabilités, nous pensons pouvoir obtenir encore des résultats positifs.

La taille des exemples

La taille des exemples est de 16. C'est-à-dire qu'un exemple est défini par les valeurs de 16 variables booléennes. Cela permet d'avoir un espace d'exemples de 2^{16} (65536 exemples possibles) qui se démarque des recherches précédentes (habituellement de 2^4 soit 16 exemples). La complexité de l'espace reproduit partiellement la complexité de l'environnement écologique dans lequel les sujets apprennent habituellement.

Le type d'oracle

On présente les exemples avec leur étiquette à l'apprenant. L'exemple est dit positif si c'est un exemple appartenant au concept et négatif si c'est un contre-exemple. Ce type d'oracle est ce que [Dominowski, 73] appelle une présentation en réception, le sujet n'ayant aucun contrôle sur la séquence d'exemples. Ceci est nécessaire car nous voulons tester l'impact des distributions de probabilités, ce qui ne serait pas possible si c'était l'apprenant qui proposait les exemples.

Présentation séquentielle des exemples

La présentation des exemples se fait séquentiellement et non en un bloc. Cela pourrait être considéré comme un biais dans le sens où le modèle PAC ne réclame pas un type de présentation plutôt qu'un autre. Cependant, en apprentissage naturel, le sujet n'est jamais amené à rencontrer tous les exemples du concept en une fois, mais plutôt de manière séquentielle c'est d'ailleurs pourquoi la plupart des recherches procèdent ainsi [Neisser et Weene, 62], [Kemler-Nelson, 84], etc. Par ailleurs, il est difficile de présenter en un bloc plusieurs centaines d'exemples.

Apprentissage implicite vs explicite

Nous voulons laisser la possibilité à l'apprenant de se placer dans un type d'apprentissage ou dans un autre (implicite ou explicite), c'est pourquoi nous ne lui demandons pas de définir explicitement son hypothèse mais d'étiqueter de nouveaux exemples comme le fait [Pazzani, 91]. Du point de vue des théories formelles cela correspond à un apprentissage de la classe de concepts (C) dans une classe d'hypothèses (H).

Facteur et variables dépendantes

Le facteur est la distribution de probabilité qui préside au tirage des exemples. Les variables dépendantes sont la réussite/échec et, dans le cas de réussite, le nombre d'exemples nécessaires au sujet, nous suivons ainsi assez étroitement le modèle PAC.

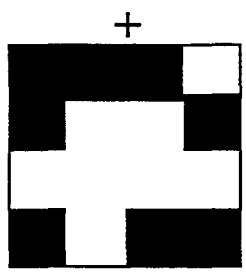
5.5.1.3 Méthode

Sujets

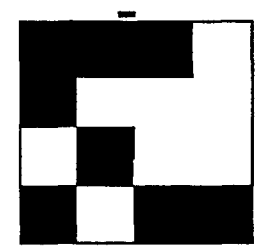
L'échantillon est composé de 39 sujets, étudiants en 1ère ou 2ème année de DEUG à l'université de Lille 3. Pour chacune des trois modalités du facteur distribution de probabilités, il y a 13 sujets : S13<D3>.

Matériel

Un *stimulus* est une grille de 4x4 cases. Chaque case peut-être blanche ou noire et constitue une variable binaire. On considère que les variables sont numérotées de gauche à droite et de haut en bas. Ainsi la variable x1 correspond à la case en haut à gauche et la variable x16 à la case en bas à droite. L'étiquette («+ » ou «- ») est placée au-dessus de la grille afin que le sujet puisse voir rapidement l'ensemble du stimulus (exemple et étiquette).



Un exemple positif du concept «x8 nonx14»
écriture binaire 1110100100001011



Un exemple négatif du concept «x8 nonx14 »
écriture binaire 1110100001001011

Ce type de stimulus a été choisi de manière à introduire aisément un grand nombre de variables et afin que toutes ces variables soient également disponibles (dans le sens de [Barsalou, 92]). Deux variables (ou cases) ne diffèrent que par leur emplacement et leur couleur.

Le *concept cible* est une conjonction définie sur les valeurs de 2 variables. Le concept choisi pour l'expérience est «x8 nonx14 », il correspond à l'ensemble de toutes les grilles dont la 8ème case est noire *et* la 14ème blanche. Toutes les grilles dont la case 8 n'est pas noire *ou* la case 14 blanche n'appartiennent pas au concept, sont les exemples négatifs.

Facteur : les distributions de probabilités

Pour présenter les exemples au sujet, l'expérimentateur va être amené à les tirer selon trois distributions de probabilités différentes : le facteur «distributions de probabilités » a

trois modalités. L'hypothèse est que l'on observera une variation dans le succès à l'apprentissage selon ces modalités.
Le problème est de définir des distributions plus ou moins facilitantes. Comme on constate en psychologie que l'homme catégorise en utilisant la similarité, il est fait l'hypothèse que si la similarité inter-exemples est parallèle au concept, c'est-à-dire que les exemples positifs sont plus semblables entre eux qu'avec les exemples négatifs, cela facilitera l'apprentissage (modalité A). Par contre lorsque la similarité inter-exemples est orthogonale au concept cela handicapera l'apprentissage (modalité C). Nous avons décidé d'ajouter une troisième distribution aux deux précédentes, la distribution uniforme (modalité B), car cette distribution est, en quelque sorte, une distribution «standard ».

Distribution de probabilités A

Dans cette distribution, les exemples choisis sont organisés autour de deux exemples centraux, l'un positif et l'autre négatif. On peut assimiler ces exemples centraux à des prototypes. On peut voir les prototypes choisis sur la figure 5.5.1.1, page 225 : le négatif est exactement l'inverse du positif. On constate que le nombre de cases noires est égal au nombre de cases blanches. En équilibrant ainsi le nombre de cases blanches et noires, on ne permet pas à l'apprenant de bâtir une stratégie basée uniquement sur le nombre de cases d'une couleur donnée.

Les exemples sont construits de la manière suivante : chaque exemple est une variation du prototype, c'est-à-dire que toutes les variables de l'exemple ont les mêmes valeurs que celles du prototype sauf 2 choisies aléatoirement[101]. On voit ainsi, sur la figure 5.5.1.1 que les exemples positifs ne varient du prototype positif que par la couleur de deux cases et de la même manière les exemples négatifs ne varient du prototype négatif que par la couleur de deux cases. On retrouve ainsi le même type de construction des exemples que celui utilisé par Kemler-Nelson ou Medin et al (voir 5.2).

Pour le tirage d'un exemple qui sera présenté à l'apprenant, la sélection se fait alors ainsi :
1>sélection aléatoire de l'étiquette de l'exemple : positif ou négatif. 25% des exemples seront positifs et 75% négatifs.
2>sélection du prototype correspondant à l'étiquette
3>sélection aléatoire de deux variables de ce prototype (hormis celles définissant le concept x8 et x14)
4>inversion des valeurs de ces deux variables.

Tous les exemples positifs sont similaires au prototype positif, puisqu'ils ne sont différents de celui-ci que par la couleur de deux cases. De la même manière tous les exemples négatifs sont similaires au prototype négatif. Il faut noter que, de par la

[101] Dans toute cette partie lorsque l'on parle de choix aléatoire il s'agit de choix aléatoire selon la distribution uniforme

construction des exemples, ni le prototype positif ni le prototype négatif n'apparaîtront jamais parmi ceux-ci.

Distribution de probabilités B

Avec la distribution B, toutes les valeurs des variables sont tirées aléatoirement selon la distribution uniforme. Chacune des variables peut prendre la valeur 1 ou 0. La probabilité que l'exemple soit positif est de 25% et négatif de 75% comme précédemment. Dans la figure 5.5.1.2, page 226, nous présentons quelques exemples possibles avec cette distribution.

Distribution de probabilités C

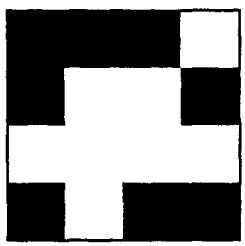
Cette distribution se définit de manière similaire à la distribution facilitante, si ce n'est qu'il n'y a qu'un seul exemple central valable aussi bien pour les positifs que pour les négatifs. Comme pour les distributions facilitantes, cet exemple central a un nombre de cases blanches égal au nombre de cases noires.
Chaque exemple (positif ou négatif) est une variation de ce prototype, c'est-à-dire que toutes les variables de l'exemple (hormis celles définissant l'appartenance) ont les mêmes valeurs que celles du prototype sauf deux choisies aléatoirement.

Comme les exemples positifs et les exemples négatifs sont des variations du même prototype, ils sont tous très semblables et on peut constater sur la figure 5.5.1.3, page 227, qu'un exemple positif peut ressembler davantage à un exemple négatif qu'à un autre exemple positif.

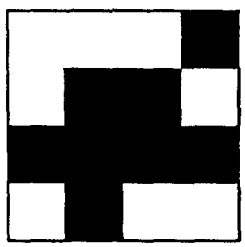
Pour le tirage d'un exemple qui sera présenté à l'apprenant, la sélection d'un exemple se fait ainsi :
1>sélection aléatoire de la valeur de la première variable du concept
2>sélection aléatoire de la valeur de la deuxième variable du concept
3>sélection aléatoire des deux variables du prototype qui varieront (hormis celles définissant le concept)
4>inversion des valeurs de ces deux variables.
La probabilité que l'exemple soit positif est ainsi de 25% et négatif de 75%.

Figure 5.5.1.1 Distribution A
Présentation de quelques exemples

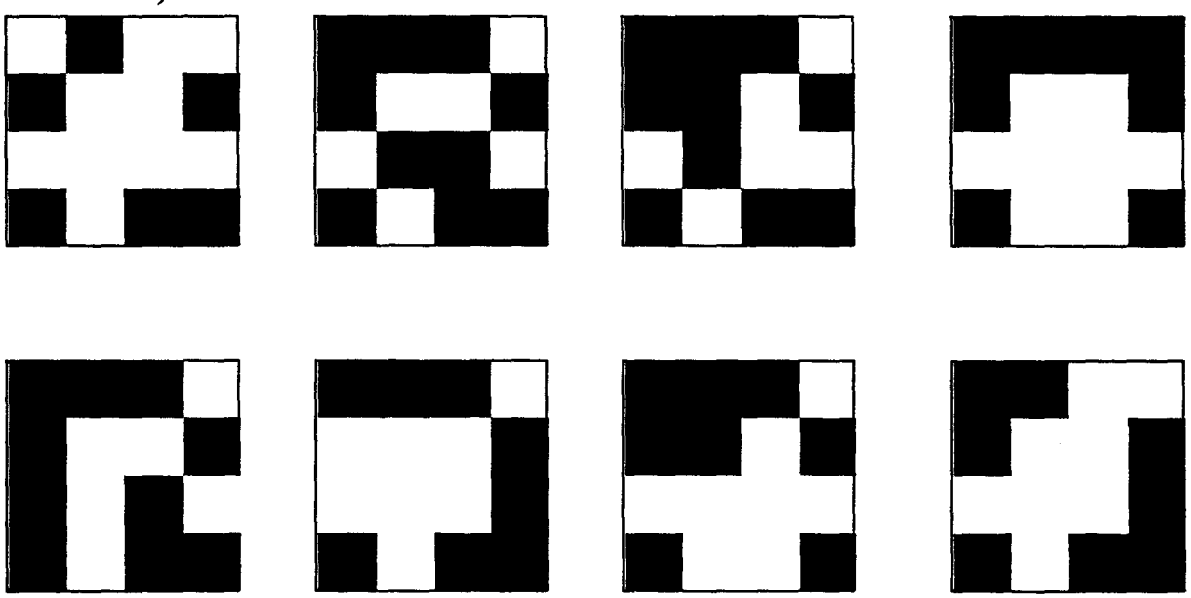
Prototype positif



Prototype négatif



Huit exemples positifs : on a fait varier la couleur de deux cases du prototype positif (hormis la 8 et la 14)



Huit exemples négatifs : on a fait varier la couleur de deux cases du prototype négatif (hormis la 8 et la 14)

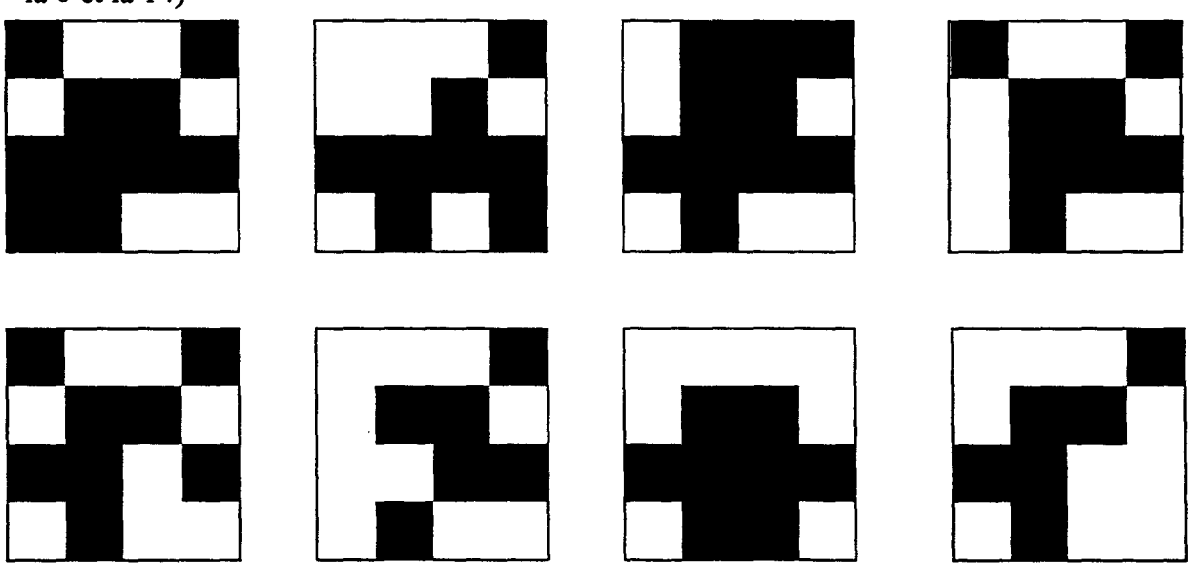
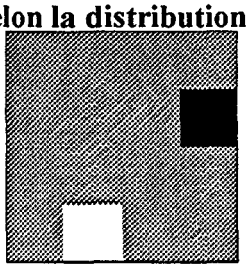
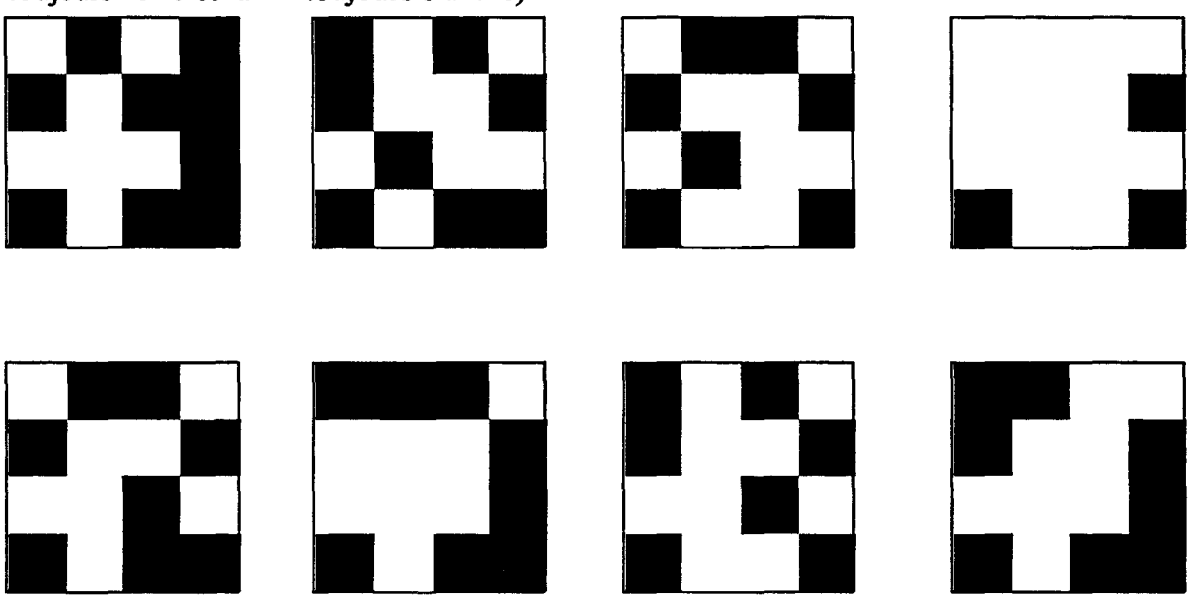


Figure 5.5.1.2 Distribution B
Présentation de quelques exemples

Les cases en grisé sont tiré aléatoirement selon la distribution uniforme



Huit exemples positifs : on choisit aléatoirement la couleur de chaque case (hormis la 8 toujours noire et la 14 toujours blanche)



Huit exemples négatifs : on choisit aléatoirement la couleur des cases. Les cases 8 et 14 peuvent prendre comme couleur : 8 blanche-14 blanche, 8 blanche-14 noire, 8 noire-14 noire

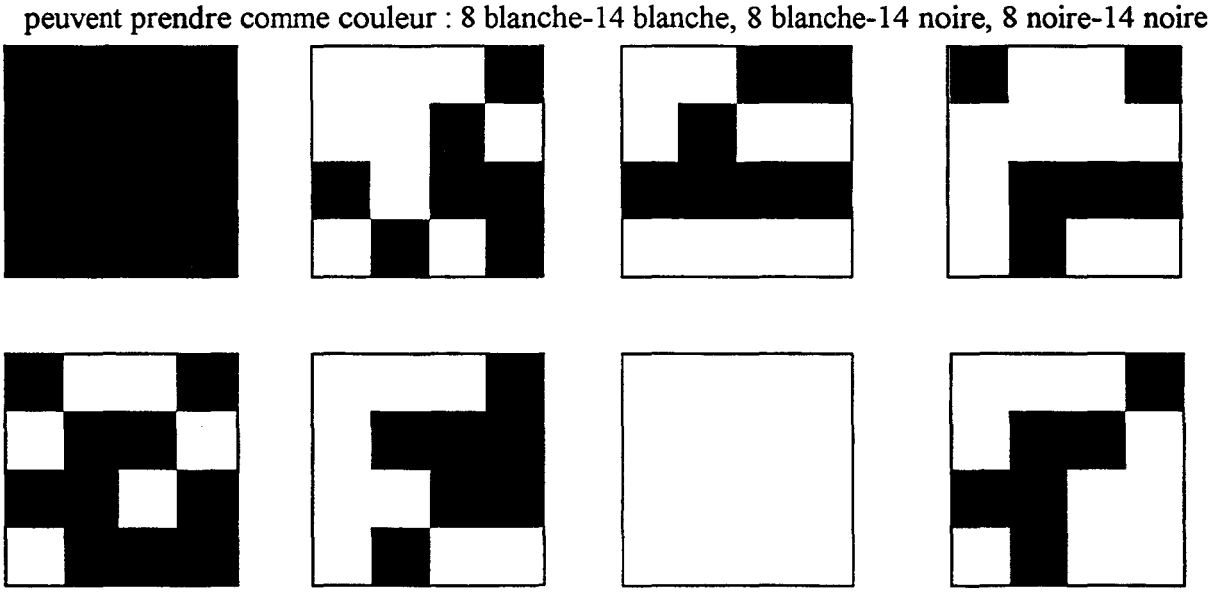
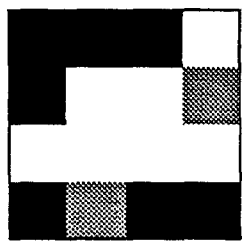
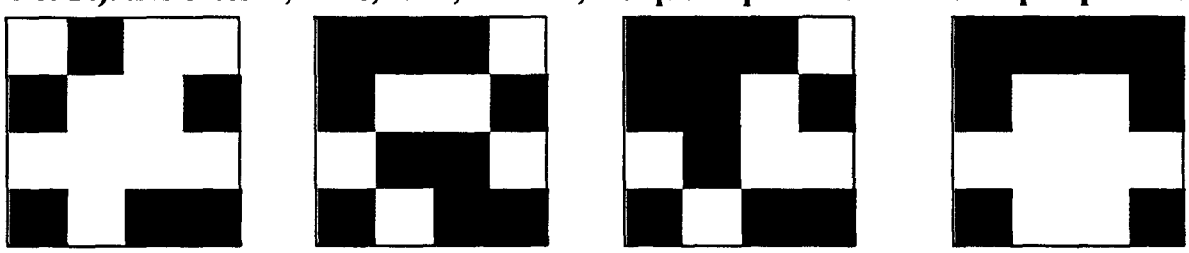


Figure 5.5.1.3 Distribution C
Présentation de quelques exemples

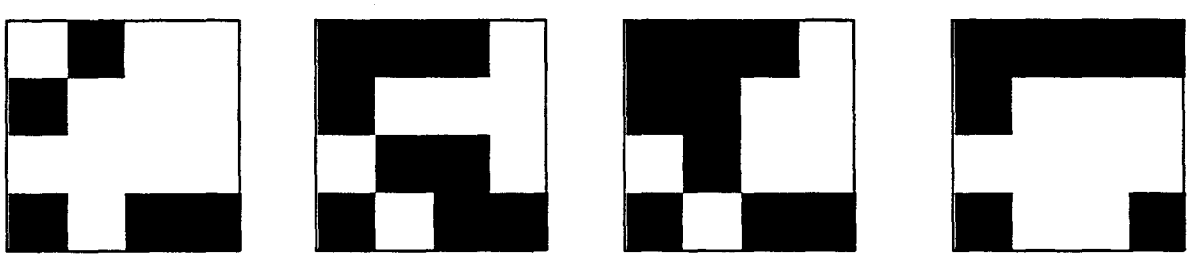
Exemple de base



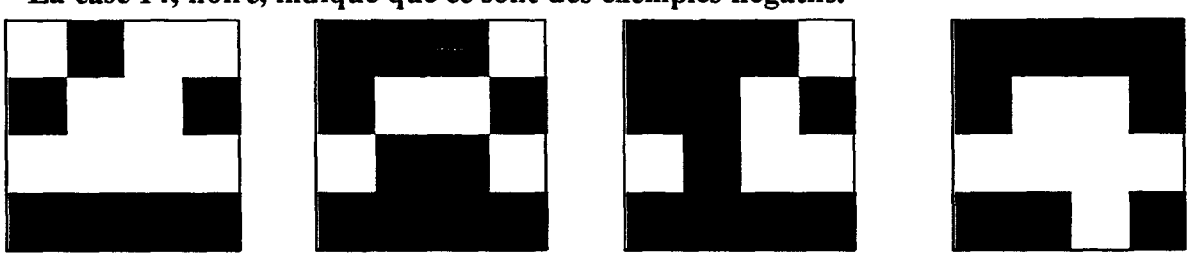
4 exemples positifs : on a fait varier la couleur de deux cases de l'exemple de base(sauf 8 et 14). Les cases 8, noire, et 14, blanche, indiquent que ce sont des exemples positifs.



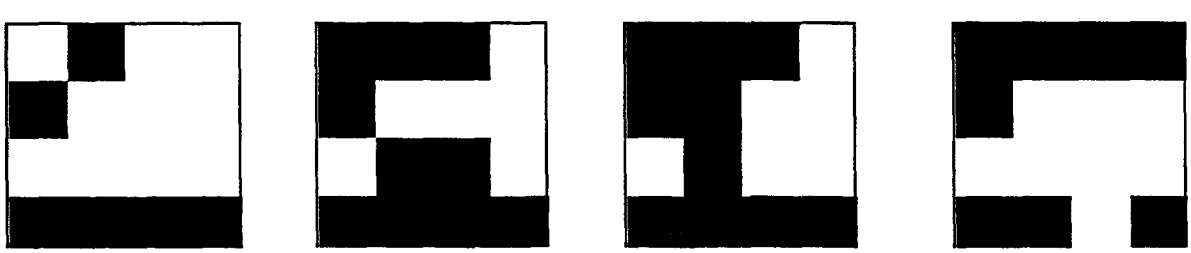
12 exemples négatifs : on a fait varier la couleur de deux cases de l'exemple de base(sauf 8 et 14). La case 8, blanche, indique que ce sont des exemples négatifs.



La case 14, noire, indique que ce sont des exemples négatifs.



Les cases 8, blanche, et 14, noire, indiquent que ce sont des exemples négatifs.



Remarques sur ces distributions de probabilités

On constate que la façon de construire les différentes distributions de probabilités a un impact sur la similarité inter-exemples ce qui est normal puisque celle-ci est à la base de leur construction mais aussi sur le nombre d'exemples possibles.

Distance minimale et maximale entre les exemples

Le tableau ci-dessous indique la dissimilarité (ou distance) minimum et maximum entre deux exemples pour chacune des distributions. Chaque exemple est construit sur 16 variables. La distance entre deux exemples est le nombre de valeurs de variables différentes.
Ainsi

d(0111011010111010, 0111011010111010)=0

et

d(0111011010111010, 1000100101000101)=16

Tableau 5.5.1.1 : distance selon les distributions (minimale-maximale)	A	B	C
d(pos,pos)	0-4	0-14	0-4
d(neg,neg)	0-4	0-16	0-6
d(pos,neg)	13-16	1-16	1-6

Pour la distribution A, on constate que la distance entre les exemples positifs est inférieure ou égale à quatre, de même pour les négatifs. C'est-à-dire que deux exemples positifs ou deux exemples négatifs ont au plus entre eux 4 cases de couleurs différentes. Tandis que la distance est supérieure ou égale à treize entre positif et négatif. Il y a donc maximum de similarité intra-catégorielle et minimum inter-catégorielle.
Pour la distribution B, la similarité entre les exemples ne dépend pas de leur appartenance à la catégorie et la distance (dissimilarité) entre deux exemples positifs peut être très grande (14) et être supérieure éventuellement à celle entre un exemple positif et un exemple négatif qui est au minimum de 1.
Pour la distribution C, il en va de même si ce n'est que la distance entre deux exemples est bornée à 6.
Du point de vue de la similarité, les distributions neutre et handicapante se ressemblent donc puisque dans les deux cas un exemple positif peut être plus similaire à un exemple négatif qu'à un autre exemple positif.

Nombre d'exemples possibles

De par la construction même des distributions de probabilités, le nombre d'exemples qui ont une probabilité non nulle d'apparaître dans l'échantillon n'est pas le même. Le tableau ci-dessous indique le nombre d'exemples possibles pour chacune d'elle.

Tableau 5.5.1.2 nombre d'exemples possibles selon les distributions	A	B	C
Positifs	91	16384	91
Négatifs	91	49152	273
Total	182	65536	364

Ainsi, si la distribution B (uniforme) permet à tous les exemples d'apparaître éventuellement dans l'échantillon, il n'en va pas de même avec les distributions A et C où tous les exemples qui sont trop éloignés du prototype n'ont aucune chance d'être tirés.

Procédure

Les stimuli sont présentés séquentiellement sur un écran d'ordinateur compatible IBM. La taille de chaque case correspond à celle de deux caractères soit un carré de 7mm sur 7mm sur un écran de 14 pouces.

Information sur le concept et consigne

Préalablement à l'expérience, chaque groupe de sujets est informé sur le type de concept cible : on vérifie qu'ils ont bien compris ce qu'est une conjonction (on trouve en annexe 5 la feuille donnée aux sujets). Chaque sujet est aussi informé sur la façon dont se déroulera l'expérience.

« On vous présentera des grilles qui sont des exemples positifs et négatifs d'une catégorie alpha. Les exemples positifs sont indiqués par un «+ » et les exemples négatifs sont indiqués par un «- ». Une catégorie alpha est constituée des grilles qui ont toutes les mêmes valeurs pour deux cases particulières. Vous pouvez voir sur la feuille un exemple de catégorie alpha : la deuxième case est noire *et* la quinzième blanche, mais il peut exister des catégories alpha avec deux cases blanches et d'autres avec deux cases noires. Ici, avec cette catégorie alpha, tous les exemples *positifs* ont la deuxième case noire *et* la quinzième blanche ». Les sujets passent en revue les exemples positifs présentés sur la feuille. « Les exemples négatifs ont *soit* la deuxième case blanche *soit* la quinzième noire *soit* la deuxième blanche *et* la quinzième noire ». Les sujets passent en revue les exemples négatifs. « Exemple test : est-ce que les exemples ci-dessous appartiennent ou non à la catégorie alpha ? ».

« On vous présentera ces exemples sur un ordinateur. Vous devrez ensuite en classer vous-même 25 en «+ » ou «- » en ne faisant au maximum qu'une seule erreur. *Dès que* vous pensez pouvoir classer de nouveaux exemples *sans trop vous tromper,* demandez à faire le test en appuyant sur la lettre «t ». Ce que *l'on demande c'est que vous réussissiez après avoir vu le minimum d'exemples possibles.* Cependant si vous vous trompez lors du test, on vous présentera à nouveau des exemples jusqu'à ce que vous réussissiez ou que vous ayez vu plus de 900 exemples.» Les sujets sont invités à manipuler le clavier avant que l'expérience ne commence.

Passation

Les sujets passent par groupes de 6 à 12, à raison d'un sujet par ordinateur. Au sein de chaque groupe, il y a un nombre égal de sujets par modalité. Dans le cas d'un groupe de 6 : 2 sujets apprendront avec la modalité A, 2 avec la modalité B et 2 avec la modalité C.
Le sujet voit défiler des exemples avec leur étiquette. Dès qu'il pense qu'il peut passer le test, il appuie sur la touche «t ». S'il se trompe plus d'une fois sur 25, les exemples étiquetés recommencent à défiler. L'ordinateur enregistre le nombre d'exemples étiquetés que le sujet a vu avant de réussir (et pour information le nombre d'exemples test ainsi que le nombre d'erreurs).
Le temps d'affichage de chaque exemple est de 3 secondes. Ce temps est suffisant pour permettre de balayer visuellement la figure. Il est nettement supérieur à celui indiqué par [Brunel, Ninio, 97] concernant la discrimination entre deux grilles. Le temps de réponse lors du test n'est pas borné.
L'expérience cesse pour un sujet dès qu'il a appris. Dans le cas d'échec, le temps maximal est de 45 mn.

Analyse statistique

Les données nombre de réussite/échec et nombre d'exemples en cas de réussite ont été soumises à une analyse de variances de type S13<D3>

Résultats

On trouve dans le tableau en annexe 6 les résultats obtenus par chacun des sujets, tous les sujets ayant étiqueté correctement les quatre exemples sur papier. On constate que si 100% des sujets ont réussi l'apprentissage avec la distribution A seulement 76,92% (10/13) l'ont fait avec la distribution B et 61,5% (8/13) avec la distribution C. *Le type de distribution de probabilités joue bien en ce qui concerne la réussite ou l'échec à l'apprentissage de deux conjonctions de valeurs de variables ($F(2,36)=3.257$, $p=0.05$).*

Les sujets qui ont réussi l'apprentissage avec la distribution facilitante ont utilisé en *moyenne* 213 exemples, contre 323 pour la distribution uniforme et 365 pour la distribution handicapante. Cependant, pour une même distribution de probabilités, l'écart entre les résultats des sujets est trop grand et on ne peut tenir compte de ces moyennes ($F(2,28)=2.575$, $p=0.094>0.05$). *On ne peut donc affirmer que le nombre d'exemples*

nécessaires à l'apprentissage varie selon la modalité de la variable «distribution de probabilités».

5.5.1.5 Discussion

Il convient, ici, de bien appréhender l'objectif de cette expérience. *Celle-ci avait pour but d'étayer une thèse en théories formelles et de montrer que l'apprentissage se trouve plus ou moins facilité selon les distributions de probabilités.* Pour cela nous avons construit trois distributions en nous appuyant sur la similarité interexemples, en supposant que cette similarité peut aider ou non l'apprentissage, mais l'objectif n'était pas de démontrer que la distribution A est plus facile que la distribution B et celle-ci plus facile que la distribution C. L'important était de montrer que, selon la distribution, l'apprentissage n'est pas le même. Il suffisait qu'une seule de ces distributions se distingue des deux autres pour que l'objectif soit atteint. Le résultat obtenu avec la variable dépendante «réussite vs échec» permet d'affirmer que c'est le cas ici. Cependant, il semble que ce soit surtout la distribution A qui se distingue des deux autres : elle est plus facile. Pour les deux autres, il n'y a pas de résultats significatifs permettant de les différencier.

Une analyse des «propriétés» de ces distributions permet d'expliquer ces résultats. Nous avons vu plus haut qu'il n'y avait que 182 exemples susceptibles d'apparaître avec la distribution A contre 65536 pour la distribution B et 364 pour la distribution C. Ce n'est donc pas le nombre d'exemples susceptibles d'être tirés qui peut expliquer la différence entre les résultats, sinon nous aurions dû obtenir de bien plus mauvais résultats avec la distribution B qu'avec la distribution C alors qu'elles se «valent» plus ou moins.

Par contre, le calcul de similarité présenté plus haut peut constituer une première explication car aussi bien dans la distribution neutre que dans la distribution handicapante un exemple positif peut être plus similaire à un exemple négatif qu'à un autre exemple positif, ce qui n'est pas le cas avec la distribution facilitante ; du point de vue de la similarité, la distribution, B est aussi difficile que la distribution C.

Mais c'est surtout lorsque l'on calcule le nombre d'hypothèses approximativement correctes qui peuvent être acceptées par le test que l'on obtient l'explication. Le calcul théorique que l'on trouvera en annexe 7 donne le tableau suivant :

Tableau 5.5.1.4 Nombre d'hypothèses (conjonction) selon les distributions	A	B	C
Total	480	480	480
Approximativement correctes (taux d'erreur<18%)	120	1	1
Ratio nb. hyp. approx.correctes/nb. total	1/4	1/480	1/480

Ainsi avec la distribution A, une hypothèse sur quatre semble approximativement correcte *en théorie*, il convient de voir *en pratique* ce qu'il en est et plus exactement

combien d'hypothèses sont à chaque fois capables de passer le test. Ceci a été fait par programme : nous avons construit l'ensemble des 480 hypothèses de la forme x'y' et, chaque fois qu'une de ces hypothèses était infirmée plus d'une fois par un exemple du test, elle était retirée de l'ensemble. Il restait ainsi un certain nombre d'hypothèses qui ne faisaient qu'une seule erreur ou moins sur le test. Le résultat figure dans le tableau suivant :

Tableau 5.5.1.5 Nbre d'hypothèses approximativement correctes pour le test n° selon la distribution	1	2	3	4	5	6	7	8	9	10	moyenne
A	69	91	49	50	42	82	69	47	65	37	60,1
B	2	1	2	1	1	1	1	1	1	1	1,2
C	1	1	7	1	1	1	1	1	2	1	1,7

On constate ainsi qu'il y a un nombre plus grand d'hypothèses approximativement correctes avec la distribution A et que 69 hypothèses sur 480 permettent de passer le premier test avec succès alors qu'il n'y en a que deux pour la distribution B et une pour la distribution C. Ainsi seule la distribution A permet un apprentissage approximatif. Encore une fois, on constate que la distribution B ne se distingue pas de la distribution C.

Du point de vue des théories formelles, si on considère l'hypothèse : « les distributions de probabilités jouent sur la facilité de l'apprentissage », on peut alors considérer que le résultat positif obtenu ici est biaisé dans le sens où la distribution facilitante permet d'obtenir un nombre d'hypothèses approximativement correctes plus grand que pour les deux autres distributions, et donc de prévoir le résultat obtenu. Néanmoins, adopter ce point de vue revient à admettre que toutes les distributions de probabilités ne sont pas les mêmes que certaines sont facilitantes, ce qui revient à valider le modèle PAC avec distributions bienveillantes. Par ailleurs, cette expérimentation invite à s'interroger sur les propriétés des distributions de probabilités. En théories formelles, il serait notamment intéressant de vérifier l'impact des distributions de probabilités sur l'apprentissage lorsque celles-ci permettent un nombre plus grand d'hypothèses approximativement correctes.

Du point de vue de l'étude de la catégorisation, il semble remarquable que, tandis que nous avons construit la distribution A sur la seule similarité des exemples à deux prototypes, nous constatons que cela joue sur le nombre d'hypothèses approximativement correctes. Cela amène à s'interroger si le fait de catégoriser en utilisant la similarité ne permet pas de faciliter l'adaptation dans le sens où le nombre d'hypothèses différentes qui permettent l'adaptation est plus grand.

Une autre question en psychologie que pose cette expérimentation est celle de découvrir ce qu'est une distribution naturelle bienveillante, c'est-à-dire qui facilite l'apprentissage ? Nous avons fait des propositions, ici, en construisant les distributions sur la similarité inter-exemples. Une autre réponse possible à cette question est celle apportée par [Fried et Holyoak, 84], [Flannagan, Fried, et Holyoak, 1986], il s'agirait alors des distributions apparentées à la distribution normale. Mais nous avons vu que, dans leur article, le concept était en quelque sorte «omis». Il conviendrait donc de monter une expérience selon le modèle PAC en utilisant comme une des modalités du facteur «distribution», la distribution normale.

Il reste à signaler que, bien que l'on n'ait pas demandé aux sujets de découvrir une règle explicite, c'est pourtant ce qu'on fait la plupart d'entre eux d'après ce qu'ils en ont dit à l'issue de l'expérience. Celle-ci s'apparente alors davantage à une expérience d'induction de règle telle les expériences préroschiennes que nous avons décrites au chapitre 2 et il n'est pas évident que nous puissions assimiler celles-ci à un processus de catégorisation, ainsi que le montre l'expérience de Kemler-Nelson décrite dans le 5.2. Peut-être faut-il envisager la catégorisation comme un processus de bas niveau ? Comme nous allons le voir, l'expérience suivante reposera ce problème.

Enfin, il reste à signaler les étonnements et les erreurs du néophyte en psychologie. L'étonnement est apparu devant la difficulté à trouver un nombre adéquat de sujets volontaires. L'erreur a consisté dans la difficulté de l'expérience pour les sujets et sa durée (45 mn), ce qui n'a pas été sans déclencher des réactions quelque peu agressives chez ceux qui échouaient.

5.5.2 Expérience n°2

Comme nous l'avons vu dans la conclusion du chapitre 3, le rôle de la fréquence d'instanciation des items dans la définition du poids des valeurs d'attributs dans la représentation qu'a le sujet d'une catégorie n'est pas accepté par tous les chercheurs. L'expérience n°2 avait pour but de montrer que les représentations que se fait un individu d'une catégorie prennent en compte la fréquence des valeurs d'attributs par le biais de la fréquence d'instanciation des exemplaires.

En réutilisant le même type de stimuli que précédemment (des grilles de 4x4 cases), nous avons construit deux distributions de probabilités. La première était la distribution uniforme où tous les exemples possibles avaient une chance égale d'apparaître, et où chaque case avait une chance sur deux d'être de couleur noire. L'autre distribution de probabilités était conçue de telle manière que la fréquence de valeurs pour deux attributs étaient plus élevée que pour les autres (2 cases étaient noires 90 fois sur cent alors que les autres ne l'était que 50 fois sur cent). Nous l'avons appelée, de manière un peu abusive, «prototypique ». Un groupe de sujets apprenait avec la distribution uniforme et l'autre avec la distribution prototypique. A l'issue de l'apprentissage, nous demandions aux sujets de choisir entre deux grilles, A et B, celle qui ressemblait le plus à celles

observées durant l'apprentissage. Dans la grille A, les deux cases les plus fréquentes apparaissaient ce qui n'était pas le cas dans la B. Nous nous attendions alors à ce que les sujets apprenant avec la distribution uniforme choisissent de manière égale la grille A ou la B, tandis que pour ceux apprenant avec la distribution prototypique, la case A soit davantage sélectionnée. Les résultats obtenus sont mitigés et invitent à une réflexion sur la manière de recréer le processus de catégorisation en laboratoire.

5.5.2.1 Méthode

Sujets Les sujets étaient 40 aides-éducateurs de la région de Roubaix-Tourcoing dont la formation variait entre bac +2 et bac+3. Il y avait deux groupes d'une vingtaine de sujets : le premier voyait des dessins tirés selon une distribution de probabilités uniforme, le second une distribution de probabilité prototypique. Le plan était donc de type S20<D2>

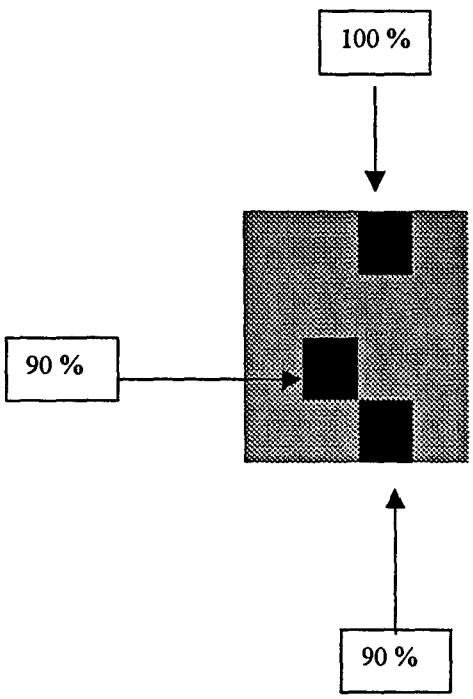
Matériel

Un stimulus étant une grille de 4x4 cases qui peuvent prendre les valeurs blanc ou noir, un stimulus est un exemple positif de la catégorie si sa 3ème case est noire (on numérote de gauche à droite et de haut en bas), il est négatif sinon. Il s'agit donc du concept unidimensionnel «x3».

Dans la distribution de probabilité uniforme, la valeur de chaque case est tirée à pile ou face. C'est-à-dire qu'une case donnée de la grille sera blanche dans 50% des exemplaires et noire dans les 50% restant et ceci quelle que soit la case. Il y a autant d'exemples positifs que d'exemples négatifs.

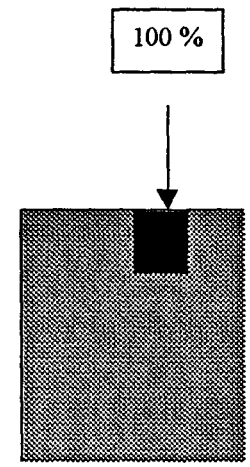
Dans la distribution «prototypique », pour les *exemplaires positifs* la case x10 sera noire pour 90% d'entre eux et blanche pour les 10% restant, de la même façon la case x15 sera noire pour 90% d'entre eux et blanche pour les 10% restant. Toutes les autres cases seront blanches pour 50% des exemplaires positifs et noires pour les 50% restant, hormis la case x3 qui est toujours noire puisque les exemplaires sont positifs. Pour les *exemplaires négatifs* toutes les cases sont tirées à pile ou face, c'est-à-dire qu'elles seront blanches pour 50% des exemplaires et noires pour les 50% restant, hormis la case x3 qui est toujours blanche puisque les exemplaires sont négatifs. Il y a autant d'exemples positifs que d'exemples négatifs.

A : Distribution prototypique



Dans la distribution prototypique : pour les exemples positifs les couleurs de toutes les cases, hormis les trois indiquées sont tirées à pile ou face. Les cases 3 (noire), 10 (noire) et 15 (noire) définissent le prototype.
La case 3 sert à définir l'appartenance, c'est pourquoi elle a la même couleur pour tous les exemples positifs (on a choisi noire). La case10 sera noire pour 90% des exemples positifs et pour 10% blanche. La case 15 sera noire pour 90% des exemples positifs et blanche pour 10%.
Les exemples négatifs ont tous la case 3 blanche. Les couleurs des autres cases sont tirées à pile ou face (y compris les cases 10 et 15).
Il y a autant d'exemples positifs que négatifs.

B : Distribution uniforme

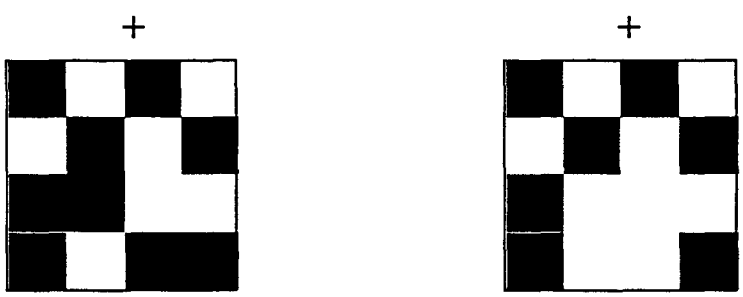


Dans la distribution uniforme les couleurs de toutes les cases des exemples positifs et négatifs (hormis la 3) sont tirées à pile ou face. La case 3 définit l'appartenance, elle sera noire pour les exemples positifs et blanche pour les exemples négatifs.
Il y a autant d'exemples positifs que négatifs.

Procédure

Les 40 sujets sont passés en quatre vagues dans une salle informatique. Chaque vague était séparée en deux groupes : l'un affrontant la distribution uniforme et l'autre la distribution prototypique. Naturellement, chaque sujet était seul face à l'ordinateur. On demandait aux sujets d'observer attentivement les exemples en distinguant bien entre les exemples positifs (marqués '+') et les exemples négatifs (marqués '-'). Le sujet visionnait 200 exemples (à 3 secondes par exemple cela fait 10 mn d'apprentissage). A l'issue de cet apprentissage les sujets étaient invités à choisir entre 2 grilles qui étaient des exemples positifs, celle qui ressemblait le plus aux exemples positifs vus durant l'apprentissage

Choisissez parmi ces deux exemples celui qui ressemble le plus aux exemples positifs vus durant l'apprentissage.



Analyse statistique

Nous avons enregistré, pour chaque groupe, le nombre de A et de B et avons établi un χ^2.

Résultats

Distribution\ grille	A	B
Prototypique	17	3
Uniforme	12	8

Si effectivement les sujets soumis à la distribution «prototypique » choisissent davantage la grille A par rapport à B que les sujets soumis à la distribution uniforme, les résultats ne sont pas significatifs (χ^2= 3,13479624,ddl=1) $p<0,1$ mais $p>0,05$.

5.5.2.2 Discussion

Rappelons que l'hypothèse était : sachant qu'aucune des deux grilles présentées à la fin n'aura été vue durant l'apprentissage, on s'attend à ce que, pour le groupe de sujets affrontant la distribution uniforme, le nombre de sujets choisissant la grille A devrait être approximativement égal au nombre de sujets choisissant la grille B. Par contre pour le groupe de sujets ayant affronté la distribution «prototypique » le nombre de sujets choisissant la grille A devrait être supérieur au nombre de sujets choisissant la grille B car dans la grille A les cases 10 et 15 sont noires (correspondant donc à la plus grande fréquence des couleurs de ces cases parmi les exemples vus) alors que dans la grille B c'est l'inverse.

Les résultats obtenus ne sont pas probants. Nous avons alors vérifié que le tirage réel des exemples pour chaque distribution de probabilités étant bien aléatoire selon la distribution et c'était bien le cas. Nous sommes donc revenu sur la démarche. Cela a amené à constater des erreurs de débutant. Ainsi la figure «prototypique » est toujours la A et elle est toujours placée à gauche sur l'écran, ce qui peut expliquer le nombre important de A avec la distribution uniforme. Mais même si nous avions inversé, l'écart relatif entre les deux distributions serait probablement resté le même : nous aurions obtenu moins de grilles prototypiques avec la distribution uniforme mais aussi avec la distribution prototypique.

Ce retour sur la démarche amène à se demander s'il existe un seuil *minimal* de fréquence des valeurs d'attribut pour que celles-ci soit prises en compte, seuil qui serait alors supérieur à 90%, ou dans le cas de cooccurrence de deux variables, supérieur à 80%, les cases 10 et 15 n'étant simultanément noires que 80% du temps.

Par ailleurs, rien ne garantit que le sujet se soit constitué une représentation pour les exemples positifs et une autre pour les exemples négatifs. Il a pu très bien ne se constituer qu'une seule représentation pour l'ensemble des grilles. Pour le sujet, toutes les grilles deviennent alors des exemples positifs d'une catégorie que l'on pourrait baptisée : « ensemble des grilles de l'expérimentation ». Dans ce cas, la fréquence des cases 10 et 15 noires tombent à 70% et l'on peut se demander si ce seuil est suffisant.

Mais c'est en comparant ce que nous avons fait avec ce que des psychologues nous avaient proposé de faire que l'on est amené à plus de discernement concernant ce que peut être une étude du processus de catégorisation en laboratoire avec des catégories artificielles. Apparaissent ainsi deux points :
- l'opposition entre catégorisation et induction de règle,
- la catégorisation comme apprentissage par exemple positifs.

Une première proposition qui nous avait été faite était de ne pas solliciter de la part du sujet de trouver la règle qui permettait de distinguer les exemples positifs des exemples négatifs. Ainsi, à l'opposé de ce qui s'était passé dans l'expérience précédente, nous n'avons pas demandé explicitement aux sujets d'être capables d'étiqueter de nouveaux

exemples. Néanmoins, nous avons quand même attiré leur attention sur la distinction entre exemples positifs et négatifs. De ce fait, il est possible qu'ils se soient quand même mis en situation de recherche de règles. Comme nous le signalions dans l'expérience précédente, ce type de démarche ne correspond peut être pas au processus de catégorisation. Nous rejoindrions ainsi les résultats de [Kemler-Nelson, 84], les conclusions de [Medin, Wattenmaker et Hampson 87] et les remarques de [Dubois, 93,b] qui supposent que la catégorisation est un processus de bas niveau, de niveau subsymbolique. Se pose alors le problème de la recréation d'un tel processus en laboratoire : aviser le sujet qu'il devra classifier de nouveaux exemples l'amène automatiquement en situation de recherche de règle et à la mise en place de processus de haut niveau. Seules des techniques, telles celles de l'apprentissage incident de Kemler-Nelson, permettent d'éviter cet écueil, mais celles-ci ne permettent pas d'utiliser la procédure TEST puisque celle-ci invite, à nouveau, le sujet à utiliser des processus de haut niveau. Ainsi, si d'autres techniques ne sont pas trouvées, l'opérationnalisation du modèle PAC ne permettrait que d'étudier l'induction de règle et non la catégorisation.

Une autre proposition des psychologues était que le sujet ne se voit présenter que des exemples positifs du concept et aucun exemple négatif. Cela revient à dire que la catégorisation est un apprentissage par exemples positifs uniquement. Cela va dans le sens des études faites en induction où il est mis en évidence la difficulté des sujets à traiter les exemples négatifs [Oakhill et Johnson-Laird, 85], [Dijkstra et Dekker, 82], [Evans et Newstead, 80]... Ceci ne pose pas trop de problèmes du point de vue du modèle d'apprentissage PAC, l'apprentissage par exemples positifs a été relativement bien étudié. Cependant, encore une fois, l'utilisation dans une expérience de la procédure TEST peut biaiser l'apprentissage, car lorsque le sujet l'utilise, il rencontre des exemples non étiquetés et obtient ainsi une information statistique sur les exemples. Ce type d'apprentissage a été étudié dans [Denis, 98].

Ces deux propositions, interprétées en termes de modélisation de l'apprentissage PAC, peuvent s'interpréter ainsi : la catégorisation est un apprentissage qui utilise comme schéma de représentations[102] un schéma de représentation subsymbolique et, comme protocole d'apprentissage, un apprentissage par exemples positifs avec information statistique. Du point de vue de l'opérationnalisation, si le fait de ne présenter que des exemples positifs ne pose pas problème, le fait par contre d'obliger le sujet à ne travailler qu'à un niveau subsymbolique ne paraît pas évident de par l'utilisation de la procédure TEST.

[102] Encore une fois, le terme de schéma de représentation est à interpréter dans le sens donné au chapitre 1 et non dans celui utilisé parfois en psychologie.

5 Conclusion

Dans ce chapitre nous avons présenté l'opérationnalisation de l'apprentissage PAC, c'est-à-dire la traduction sous forme d'algorithme d'expérimentation en psychologie de la définition du modèle d'apprentissage. Dans un premier temps, nous avons montré l'intérêt d'une telle démarche pour les théories formelles de l'apprentissage. L'opérationnalisation doit permettre d'aider à valider les choix théoriques qui sont opérés dans les différentes variantes du modèle standard. Pour l'apprentissage automatique, elle devrait faciliter la mise à jour des heuristiques utilisées par l'être humain qui pourraient alors être implantées en machines.
En psychologie, au travers de la description de diverses expérimentations, nous avons montré que l'opérationnalisation permet d'introduire davantage d'écologie dans l'étude du *processus* de catégorisation sur des catégories artificielles. Les cinq dimensions écologiques que nous avons repérées sont :
1) le processus de catégorisation est un *apprentissage*, une adaptation à l'environnement,
2) l'environnement *impose* à l'apprenant le regroupement des objets en certaines catégories,
3) l'apprentissage est *approximatif*, l'adaptation n'est pas optimale
4) l'apprenant ne rencontre pas n'importe quels objets d'une catégorie ni tous les objets de celle-ci,
5) l'environnement est *complexe*.
Ces cinq dimensions sont inhérentes au modèle d'apprentissage PAC.

Nous avons ensuite analysé la définition du modèle PAC de manière à pouvoir produire une définition de ce que pourrait être un concept appris en psychologie. Cela nous a obligé à abandonner les quantificateurs universels sur la classe de concepts et les distributions de probabilités. Mais cet abandon se justifie par le changement de problématique. En psychologie, l'objectif n'est pas tant d'étudier l'apprenabilité des concepts que de comprendre quels sont les facteurs qui interviennent dans cet apprentissage. En mettant en paramètres le concept et la distribution de probabilités, nous les transformons ainsi en facteurs. L'analyse nous a aussi amené à abandonner le «probablement» de l'apprentissage «Probablement Approximativement Correct» puisque l'expérimentateur peut contrôler que les tirages des exemples ne sont pas pathologiques en regard de la distribution de probabilités. Par contre, il a fallu intégrer le Sujet. Ceci, aussi, est dû au changement de problématique. La psychologie ne cherche pas à trouver un algorithme théorique désincarné mais étudie un processus bien réel chez l'être humain.
Le fait que dans le modèle PAC, l'apprenant ait connaissance de la borne maximale d'erreur tolérée nous a entraîné à définir une procédure TEST. Cette procédure demande à l'apprenant d'étiqueter un certain nombre d'exemples. Si les réponses sont approximativement correctes, elle arrête l'apprentissage, sinon, elle invite le sujet à le continuer. Outre que cette procédure simule la connaissance par l'apprenant de ε, elle permet aussi d'éviter que le sujet ne soit obligé de donner explicitement son hypothèse et de le laisser, ainsi, travailler aussi bien au niveau implicite qu'au niveau explicite. Nous ne pouvons, cependant, demander au sujet de tester un trop grand nombre d'exemples,

ce point joue sur la garantie que nous pouvons avoir concernant l'erreur de son hypothèse. Ainsi si TEST propose 25 exemples et ne tolère au plus qu'une erreur, nous sommes sûrs à 90% que l'erreur de l'hypothèse de l'apprenant est d'au plus 18%, ce qui est une marge d'erreur relativement grande.

Cette analyse a permis de déboucher sur une définition de ce que nous pouvons envisager comme un «concept appris de manière approximativement correcte » et sur la présentation d'un algorithme d'expérimentation. Cet algorithme ne recouvre pas toute l'expérimentation mais seulement la partie qui concerne un groupe de sujets. Il ne définit pas un plan d'expérimentation mais peut être utilisé au sein de tout plan. Il permet d'intégrer dans les expérimentations en laboratoire sur la catégorisation, les cinq aspects écologiques évoqués ci-dessus.

Nous avons ensuite testé cette opérationnalisation et monté deux expériences qui suivaient le protocole obtenu. La première avait un objectif davantage informatique et devait permettre de justifier l'hypothèse développée dans le modèle PAC avec distributions bienveillantes selon laquelle l'apprentissage est fortement dépendant des distributions de probabilités. Le résultat obtenu va dans ce sens.

La seconde concernait les recherches sur la catégorisation en psychologie et devait permettre de montrer que la fréquence d'instantiation des items d'une catégorie joue sur le poids des valeurs d'attributs utilisées dans la représentation de cette catégorie. Les résultats obtenus ne sont pas probants et ont amené les réflexions suivantes concernant la catégorisation :
- la première concerne la nature même de ce qu'est catégoriser. D'une part, il semble que ce pourrait être un apprentissage par exemples *uniquement* positifs. D'autre part catégoriser est peut-être essentiellement un processus de bas niveau ou, de manière moins tranchée, certains processus en œuvre dans la catégorisation sont de bas niveau.
- la seconde d'ordre méthodologique découle de cette dernière remarque. Est-ce que l'opérationnalisation du modèle PAC ne permet d'obtenir qu'un protocole d'expérience pour l'induction de concepts et de règles (processus de haut niveau) ou permet-il aussi de monter des expériences recréant la catégorisation en laboratoire avec ses processus de bas niveau ?

Ceci ne remet pas en question le modèle PAC comme modélisation de la catégorisation. Par contre, l'opérationnalisation du modèle PAC débouchant sur un protocole d'expérience est incomplète. Si cette opérationnalisation permet d'introduire certaines dimensions écologiques, elle ne prend pas en compte le problème du schéma de représentation utilisé par le sujet. Plus encore, en utilisant la procédure TEST, elle l'invite à n'utiliser que des processus de haut niveau. De la même manière qu'il est possible de faire apprendre le sujet à son insu (voir l'apprentissage incident de Kemler-Nelson), il faudrait pouvoir le tester à son insu.

Conclusion générale

Ce travail s'est situé dans le cadre des sciences cognitives et plus particulièrement de l'informatique et de la psychologie. Son objet est l'apprentissage inductif où l'on considère que l'apprenant doit, sur la base d'exemples qu'on lui soumet, retrouver le concept dont ces exemples sont des instances. En informatique, et plus précisément en théories formelles de l'apprentissage, Valiant a proposé une modélisation de cet apprentissage avec le modèle d'apprentissage Probablement Approximativement Correct (apprentissage PAC). Dans ce modèle, il n'est pas réclamé que l'hypothèse que retourne l'apprenant soit exacte mais seulement approximativement correcte. En psychologie, le concept le plus voisin de l'apprentissage inductif, est la catégorisation. Cette catégorisation, Houdé la définit comme étant la conduite adaptative fondamentale par laquelle les systèmes cognitifs, biologiques ou artificiels «découpent» le réel physique et social. Nous avons voulu exposer aux chercheurs de chacune des deux disciplines les études effectuées dans la discipline voisine mais aussi montrer la légitimité du modèle PAC. De ce modèle, nous avons extrait les principales notions qui nous ont servi pour présenter les recherches sur la catégorisation en psychologie. *Ceci a permis de montrer que l'apprentissage PAC peut être utilisé pour décrire la catégorisation. Ce faisant, d'une part, le modèle est légitimé et, de l'autre, la définition qu'Houdé donne de la catégorisation est confortée.*

Légitimité du modèle PAC au niveau de ses concepts

Dans les trois premiers chapitres nous avons montré la légitimité du point de vue cognitif du modèle PAC. Tant au niveau des concepts qu'au niveau des résultats, on trouve des éléments similaires entre les recherches effectuées dans le cadre de ce modèle et les études sur la catégorisation en psychologie.

Le modèle PAC décrit le processus de catégorisation

Ce modèle lie de manière indissociable le processus, l'apprentissage ou la catégorisation, et le résultat de ce processus, la représentation de la catégorie. L'adaptation à un environnement qu'est la catégorisation est modélisée dans l'apprentissage PAC par les relations entre l'Oracle et l'apprenant. De la même façon que l'environnement impose à l'individu les catégories qu'il doit former pour s'adapter (cf. les études sur les catégories de base, catégories ad hoc [Barsalou, 83], catégories slot-filler [Nelson, 85]), dans le modèle PAC, l'Oracle impose à l'apprenant le concept à apprendre. *Ainsi tous les éléments d'une catégorie ont-ils un point commun, c'est le rapport identique qu'entretient l'individu avec ces éléments.*

Ce qui correspond, dans la catégorisation, à la *fréquence d'instanciation*, à la familiarité de l'individu avec les objets qui lui ont servi à se former une représentation de la catégorie est représenté dans le modèle par la *distribution de probabilités* selon laquelle l'Oracle tire les exemples.

Enfin l'individu catégorise son environnement pour survivre[103] *dans cet environnement* et *cette survie ne réclame pas une adaptation optimale*. L'hypothèse de l'apprenant dans le modèle PAC ne doit pas non plus être exacte, seulement approximative, et l'évaluation de cette hypothèse se fait en fonction de la distribution de probabilités qui a servi pour l'apprentissage.

Le modèle PAC caractérise partiellement l'apprenant

Le modèle PAC permet aussi de décrire l'apprenant, ou plus modestement certains aspects de celui-ci. D'abord, dans le modèle PAC, *l'algorithme d'apprentissage correspond à la capacité de catégoriser chez l'être humain.*

Ensuite, on suppose que *l'apprenant bénéficie d'un ensemble de descripteurs pour représenter les exemples.* En apprentissage naturel, l'individu doit être capable de percevoir le monde pour pouvoir le catégoriser, il bénéficie des informations que lui fournissent ses sens. Dans les deux cas, l'apprenant, l'individu ne travaille pas au niveau des objets réels mais au niveau d'une représentation de ces objets.

[103] Le terme de survie peut paraître un peu fort mais si, comme on le suppose, la catégorisation est inhérente à l'espèce humaine, aux débuts de celle-ci il s'agissait bien de survie.

Enfin, le modèle considère que l'apprentissage est un processus qui part de la représentation des objets d'une catégorie pour en arriver à une représentation de cette catégorie, représentation qui a une certaine structure. Ainsi, dans le modèle, l'apprenant est aussi un espace d'hypothèses construites selon un/des schéma(s) de représentations. Le modèle en lui-même ne privilégie pas tel ou tel schéma de représentations, il le suppose donné, mais les recherches en informatique ont amené deux types majeurs de schémas de représentations : *le schéma symbolique et le schéma subsymbolique* ou connexionniste. A ces deux types correspond en psychologie toute une série d'oppositions telles celles entre apprentissage explicite vs implicite, apprentissage intentionnel vs incident, catégorie componentielle vs holistique, traitement de l'information stratégique vs automatique.

Par ailleurs, le schéma de représentations en informatique est aussi ce qui permet, étant données la représentation d'un objet et la représentation de la catégorie, d'affirmer si l'objet relève ou non de la catégorie. Le pendant de ceci en psychologie est la représentation des catégories en conditions nécessaires et suffisantes, en exemplaires ou en prototypes et les mécanismes, souvent basés sur la similarité, qui permettent de calculer l'appartenance de l'objet à la catégorie.

Les chapitres 1 et 2 ont ainsi permis de montrer que l'apprentissage PAC peut être utilisé comme modélisation de la catégorisation car on retrouve dans le modèle la plupart des caractéristiques de celle-ci : le modèle s'en trouve légitimé.

Légitimité du modèle PAC au niveau des résultats

Dans le chapitre 3 nous avons cherché si, au niveau des résultats des études effectuées, il en allait de même. Ceci nous a conduit à nous intéresser plus particulièrement à deux aspects caractéristiques de la catégorisation découverts par les psychologues : *l'économie cognitive et la typicalité*.

La notion d'économie dans les deux domaines

La contrepartie, en informatique, de ce que la psychologie appelle *économie cognitive* est la notion de *taille de l'espace disponible*. En psychologie, on considère que l'environnement est complexe et que l'individu a des capacités finies[104], de ces prémisses découle la notion d'économie cognitive. Ainsi les catégories de base sont-elles le meilleur compromis entre économie cognitive et contenu en information. En informatique, le théorème d'Occam montre qu'un algorithme capable de compresser très fortement l'information est un algorithme d'apprentissage et que réciproquement tout algorithme d'apprentissage comprime très fortement l'information. Dans les deux domaines, nous retrouvons ainsi l'idée *que tout apprentissage doit consister en une compression très forte de l'information.*

[104] Ce dernier point n'est toutefois pas partagé par tous les chercheurs, notamment par les tenants de la représentation par exemplaire.

Typicalité vs représentativité

Une autre caractéristique de la catégorisation très étudiée en psychologie est celle de *typicalité* : la vache est une sous-catégorie plus typique de la catégorie mammifère que l'ours. Cette notion de typicalité a amené les psychologues à l'abandon d'une représentation de la catégorie en conditions nécessaires et suffisantes au profit d'une représentation par prototypes, dans laquelle on considère que le prototype est un résumé de la catégorie[105]. Une sous-catégorie est alors d'autant plus typique d'une catégorie que sa similarité au prototype de la catégorie est grande. La fréquence élevée d'apparition de certains traits où de leurs cooccurrences parmi les exemplaires de la catégorie amène à intégrer ces traits dans la définition du prototype[106]. Par conséquent, la fréquence des items supportant ces traits intervient-elle dans la définition du prototype. *La typicalité peut alors être considérée comme reflétant en partie la distribution de probabilités selon laquelle l'individu a rencontré les divers éléments d'une catégorie.*

En parallèle à cette notion de typicalité, nous trouvons en informatique celle de *représentativité*. Dans le *modèle d'apprentissage PAC avec distributions bienveillantes*, on considère qu'une distribution est bienveillante si elle facilite l'apprentissage, et *elle facilite l'apprentissage en garantissant que les exemples représentatifs du concept ont une probabilité non nulle d'apparaître*. Ce modèle fait suite à celui de Li et Vitànyi «d'apprentissage PAC simple » dans lequel ce sont les exemples simples qui ont une probabilité non nulle d'apparaître. Une expression possible de la représentativité d'un exemple relativement au concept peut être donnée par la complexité conditionnelle de Kolmogorov, c'est-à-dire la taille du plus court programme capable d'engendrer l'exemple lorsqu'on lui donne en entrée le concept. Le problème est que cette complexité de Kolmogorov n'est pas calculable et n'est donc pas opérationnelle.

Ainsi, alors qu'en psychologie le prototype peut être considéré, en quelque sorte, comme une hypothèse de l'apprenant typique de l'échantillon, en informatique c'est l'échantillon qui est représentatif du concept à apprendre. La représentativité intervient en amont de l'apprentissage dans la définition de l'échantillon, la typicalité intervient en aval dans la construction de l'hypothèse de l'apprenant. Les deux notions semblent ainsi symétriques mais reposent l'une et l'autre sur la distribution de probabilités des exemples.

En confortant la définition qu'Houdé propose, le modèle PAC voit sa légitimité renforcée

On peut noter que d'une définition de la catégorisation et de deux prémisses découlent deux caractéristiques de cette catégorisation. La catégorisation est un processus nécessaire à l'adaptation de l'individu à son environnement. Les capacités de l'individu pour s'adapter à cet environnement complexe ne sont pas illimitées[107], l'économie

[105] Tous les chercheurs ne sont pas partisans d'une représentation par prototype. Certains proposent qu'une représentation par exemplaires permet aussi d'expliquer les phénomènes de typicalité.
[106] Ici, aussi, les avis sont partagés.
[107] Cf. note supra

cognitive permet de rendre compte de cet aspect. Par ailleurs cet environnement n'est pas quelconque, la typicalité, en tant que reflet des distributions de probabilités qui lui sont sous-jacentes, est un outil efficace, car elle maximise l'adaptation de l'individu à cet environnement. Plus un événement est probable, plus la réponse de l'individu à cet événement est rapide et ceci grâce à la typicalité. Le fait que l'on retrouve sous une forme similaire ces caractéristiques dans le modèle d'apprentissage PAC le légitime davantage encore.

Le modèle PAC avec distributions bienveillantes

En ne réclamant plus, comme le fait le modèle PAC standard de Valiant, un apprentissage pour toutes distributions de probabilités mais seulement pour les distributions de probabilités bienveillantes, nous avons obtenu certains résultats en informatique. Le premier est d'avoir montré que, comme dans le modèle PAC standard, il existe aussi un théorème d'Occam et sa réciproque dans ce modèle-ci. Le second est que des classes de concepts dont on ne sait si elles sont apprenables dans le modèle PAC standard le sont ici (e.g. les listes de décisions). Enfin ce modèle peut être mis en correspondance avec d'autres modèles d'apprentissage, tel que celui d'Angluin d'apprentissage par requêtes, ou des modèles d'enseignabilité tel celui de Goldman et Mathias. Nous pouvons alors importer les résultats obtenus dans ces derniers. Cependant, toute notion de représentativité d'un exemple relativement à un concept introduit obligatoirement une certaine connaissance de l'apprenant. Si celui-ci n'utilise pas cette représentativité pour découvrir le concept, elle est inutile. La représentativité est donc aussi fonction de l'apprenant. Ce point peut entraîner un risque de collusion : dans quelle mesure l'Oracle n'aide-t-il pas trop l'apprenant ? En réclamant que l'apprentissage ait lieu pour toutes distributions bienveillantes et tout concept de la classe, ce danger est évité.

Opérationnaliser le modèle PAC pour l'étude de catégories artificielles

A partir du moment où le modèle d'apprentissage PAC permet de modéliser la catégorisation, il paraît naturel de vouloir l'opérationnaliser, c'est-à-dire de le transformer en protocole d'expérience.

Intérêt de l'opérationnalisation

Du point de vue des théories formelles de l'apprentissage, *cette opérationnalisation doit permettre de justifier des choix théoriques* qui sont faits.

Du point de vue de la recherche en psychologie, cinq dimensions qui sont inhérentes au modèle peuvent éventuellement intéresser le chercheur :
1) le processus de catégorisation est un *apprentissage*, une adaptation à l'environnement,
2) l'environnement *impose* à l'apprenant le regroupement des objets en certaines catégories,
3) l'apprentissage est *approximatif*, l'adaptation n'est pas optimale

4) l'apprenant ne rencontre pas n'importe quels objets d'une catégorie ni tous les objets de celle-ci,
5) l'environnement est *complexe*.

En effet, ces cinq aspects de la catégorisation ne sont pas systématiquement considérés lors de l'étude de celle-ci sur des catégories artificielles. Les deux premiers points vont notamment à l'encontre de l'assimilation de la catégorisation à une simple opération de classement d'items. Le point trois est souvent implicite dans les expériences où l'on tolère que l'apprenant puisse se tromper. Le point quatre est rarement pris en compte, pourtant, lorsque l'on présente des items à un apprenant, il y a toujours une distribution de probabilités sous-jacente à cette présentation. Enfin le cinquième point invite à s'interroger si un changement quantitatif dans l'expérience ne peut introduire un changement qualitatif dans les résultats.

Une définition possible d'un concept appris de façon approximativement correcte :

Pour cette opérationnalisation, il a été nécessaire d'analyser la définition de l'apprentissage pour en extraire les différentes variables expérimentales sous-jacentes au modèle. Nous avons obtenu alors une définition de ce que peut être un concept appris de façon approximativement correcte :

Soit c un concept d'une classe de concepts C définie sur X, D une distribution de probabilités, S un groupe de sujets et T un temps limite.
On dit que c est Approximativement Correctement appris avec la distribution de probabilités D et avec une erreur maximale ε dans un temps inférieur à T, si pour tout s de S,
s ayant accès à :
*-EX(*c, D*) qui lui présente des exemples étiquetés selon le concept c et tirés selon la distribution D,*
-TEST qui lui présente N exemples à étiqueter, tirés selon la distribution D, arrête l'apprentissage si le sujet a fait au plus M erreurs et relance l'apprentissage sinon.
s est capable, dans un temps t inférieur ou égal à T, de trouver une hypothèse qui lui permette de satisfaire TEST.

Deux expériences

Il est alors aisé de tirer de cette définition un protocole d'expériences pour l'apprentissage de concepts artificiels. C'est ce qui a été fait. Deux expériences ont ainsi été mises en place. *La première montre qu'effectivement les distributions de probabilités jouent sur la facilité d'apprentissage*, ce qui valide du point de vue cognitif le modèle avec distributions bienveillantes. *La seconde n'a pas permis d'avoir des résultats significatifs*. Elle avait pour objectif de montrer que la fréquence d'apparition des traits joue dans la représentation que l'individu se fait de la catégorie. L'échec de cette seconde expérience pousse à s'interroger sur la nature même du processus de catégorisation comme on le verra plus loin.

Les limites de cette recherche

Cet échec invite à présenter les limites de cette recherche et les pistes de travail qu'elle ouvre.

Limites de la présentation du modèle PAC et des recherches en psychologie

La nature même de ce travail nous a amené à présenter de manière succincte les recherches en apprentissage en informatique et les études sur la catégorisation en psychologie.
Au niveau informatique nous n'avons travaillé qu'avec des fonctions booléennes, l'idée étant de présenter à des non initiés ce qu'est le modèle PAC. Néanmoins une présentation, plus complète que celle que nous avons faite à la fin du chapitre 1, de l'apprentissage PAC des langages devrait intéresser aussi les psychologues, notamment tout ce qui relève de la reconnaissance des formes.
En psychologie, nous n'avons abordé la catégorisation qu'au travers du filtre du modèle PAC qui ne considère que l'apprentissage inductif. Ceci n'est qu'un sous-champ de la catégorisation. Par ailleurs, développer tout ce qui concerne l'apprentissage implicite devrait aussi intéresser les informaticiens.

Limites du modèle PAC en tant que modélisation de la catégorisation

Ainsi, c'est surtout en tant qu'utilisation de l'apprentissage PAC comme modèle de la catégorisation que l'on rencontre le plus de limites. Dans l'apprentissage PAC, tel qu'il est étudié actuellement en théories formelles, l'apprenant est vierge de toute connaissance et l'information dont il bénéficie provient essentiellement des exemples. *Ainsi, le modèle ne se préoccupe pas de l'intégration d'un concept dans un réseau de concepts* et en quoi cette intégration peut en retour faciliter l'apprentissage. Cette manière d'envisager la catégorisation correspond en quelque sorte au niveau I de [Blewitt, 94] où le très jeune enfant est capable de catégoriser mais n'intègre pas encore les catégories formées dans un réseau hiérarchique. Cependant, même dans ce cas, nous pouvons supposer que très vite l'enfant s'appuie sur les catégories qu'il s'est déjà formé pour s'en former d'autres. De plus, lorsqu'il commence à maîtriser la langue, la plus grande partie de l'information qu'il reçoit ne provient plus des exemples. L'enfant bénéficie alors de ce que nous avons appelé deux schémas de représentations qui sont probablement entrelacés : un schéma de représentation davantage subsymbolique et un schéma de représentations symbolique caractérisé essentiellement par la langue. Ce qui est vrai pour l'enfant l'est encore plus pour l'adulte.

La possibilité de dépasser ces limites

Si, à notre connaissance, aucune de ces critiques n'a été prise en compte dans les recherches en théories formelles, rien pourtant ne s'oppose à ce qu'on les intègre dans le modèle PAC.
Ainsi, il est très possible d'envisager que l'apprenant puisse aborder l'apprentissage d'un second concept d'une classe après en avoir déjà appris un premier et que

l'apprentissage du second s'appuie sur le premier. Toutefois, une telle démarche suppose en elle-même que tous les concepts de la classe ne sont pas équivalents. En effet, l'apprentissage PAC réclame que l'apprentissage ait lieu pour tous les concepts d'une classe donnée. S'il est nécessaire, pour que l'apprenant puisse apprendre un concept c' d'une classe, qu'il doive s'appuyer sur la connaissance d'un concept c qu'il a déjà appris, cela implique que les concepts c et c' ne sont pas identiquement faciles à apprendre. Cela introduit alors une relation d'ordre sur les concepts de la classe : c étant avant c' en terme de facilité d'apprentissage. Nous nous retrouvons ainsi avec des classes de concepts un peu particulières qui s'apparentent assez à celles proposées par [Berwick, 86] et [Angluin, 80] cités dans le préambule de l'introduction générale.

Concernant le fait que l'individu bénéficie éventuellement de deux schémas de représentations, nous pouvons considérer que ceci est déjà partiellement intégré dans le modèle PAC. Nous sommes alors dans le cas de l'apprentissage de C (la classe de concepts cibles) dans H (la classe d'hypothèses de l'apprenant), H pouvant inclure C. Dans ce cas-ci, une première difficulté consiste à déterminer le type d'informations dont bénéficie l'apprenant ; l'information ne provient plus essentiellement de la perception immédiate de l'objet mais aussi des divers renseignements que l'environnement linguistique de l'apprenant lui a apporté sur cet objet. Des travaux en théories formelles de l'apprentissage, ont été réalisés dans cette direction notamment ceux de [Jackson, Tomkins, 92] qui essaient de quantifier l'information apportée en plus de celle provenant des exemples (cf. note chapitre 1). En apprentissage automatique, des articles tels que ceux de [Ragavan et Rendell, 94] cherchent à évaluer l'aide apportée à l'apprentissage par une connaissance du domaine.

Une procédure TEST pour travailler au niveau subsymbolique

Une seconde difficulté apparaît avec l'opérationnalisation, comme nous avons pu l'observer lors de la deuxième expérience. Si l'on veut travailler au niveau subsymbolique, il est difficile alors d'employer la procédure TEST car celle-ci invite l'individu à se placer de manière quasi-systématique en position de recherche de règle qui se situe plutôt au niveau symbolique. Ceci serait relativement peu embarrassant si nous n'utilisions cette procédure qu'en fin d'apprentissage comme le fait [Kemler-Nelson, 84] : l'individu découvre a posteriori qu'il était sensé apprendre. Par contre, lorsque nous devons l'utiliser en cours d'apprentissage, comme nous le proposons, cela peut faire passer le sujet d'un type d'apprentissage à un autre, d'apprentissage incident à apprentissage intentionnel. S'il n'est pas possible de contourner cette difficulté, alors la seule solution qui reste est de ne faire passer le test qu'au bout d'un certain temps d'apprentissage et de ne conserver que les résultats des sujets qui l'ont réussi, ce qui oblige à travailler avec des populations de sujets assez nombreuses.

Une présentation écologique des exemples

Une autre limite de ce travail qui, elle, tient moins à l'opérationnalisation qu'à la mise en pratique qui en a été faite, est le mode de présentation des exemples. Ceux-ci étaient présentés de manière séquentielle à raison de trois secondes par exemple. Pour définir

cette façon de procéder, nous nous étions appuyé sur les présentations d'exemples proposées dans les diverses expériences en reconnaissance de règle. Ce type d'expérience n'est toutefois pas équivalent à l'étude de la catégorisation sur des catégories artificielles. Un des points de divergence pourrait être justement celui évoqué ci-dessus : la découverte de règles étant un travail qui relève davantage du niveau symbolique et la catégorisation du niveau subsymbolique. Peut-être conviendrait-il mieux alors de laisser à l'apprenant le choix du moment où il souhaite passer à l'exemple suivant. Ce type de présentation serait alors plus écologique, correspondrait mieux à ce qui se passe dans la catégorisation «naturelle ».

La catégorisation : un apprentissage par exemples positifs ?

Enfin, toujours concernant la présentation des exemples, *la seconde expérience pose la question de savoir si la catégorisation ne correspond pas à un apprentissage par exemples positifs uniquement.* [Kemler-Nelson, 84] dans son expérience ne présente pas des exemples positifs et négatifs du concept de «docteur » mais des exemples positifs de deux concepts celui de «docteur » et celui de «policier ». Ainsi, il ne s'agit pas chez elle de l'apprentissage d'un concept par exemples et contre-exemples mais de l'apprentissage de deux concepts et, même si le concept de policier est la négation de celui de docteur, il est très possible que cela fasse une différence importante au niveau de la perception qu'en a l'individu.

Les pistes de recherche

Ainsi, ce travail laisse en suspens de nombreuses questions qui sont autant de pistes de recherche.

En informatique

Représentativité des exemples ou de l'échantillon

En informatique, la première pourrait concerner la représentativité des exemples relativement au concept. Nous avons vu qu'un bon candidat pour formaliser cette représentativité est la complexité de Kolmogorov, malheureusement, cette complexité n'étant pas calculable, elle n'est donc pas non plus opératoire et ne peut être utilisée en apprentissage automatique[108]. Un autre exemple possible de représentativité est celle utilisée dans l'apprentissage des listes de décisions où le croisement des n-uplets de types 0_m et 1_m permet de retrouver la liste de décision. Pourtant, si nous supposons, comme nous l'avons fait dans la conclusion du chapitre 3, qu'un sous-ensemble représentatif en apprentissage naturel se définit par une forte similarité interne, cela invite à réfléchir en informatique, non pas à la représentativité d'*un exemple* relativement au concept, mais à la représentativité d'un *sous-ensemble* d'exemples

[108] Voir toutefois comment [Cornuéjols, 96] propose de contourner cette difficulté.

relativement au concept. Ce n'est plus l'exemple en lui-même qui est représentatif du concept mais le sous-ensemble. Autrement dit, *ce n'est qu'au travers de ses relations inter-exemples que le sous-ensemble est représentatif*, de la même manière que les n-uplets 0_m ne sont représentatifs que si nous pouvons les croiser avec les n-uplets 1_m.

Ce type de démarche pose néanmoins un problème car les distributions bienveillantes garantissent seulement que les exemples représentatifs du concept ont une probabilité non nulle. Elles ne garantissent pas que *seuls* les exemples représentatifs ont une probabilité non nulle. Aussi, si l'on considère que ce n'est plus un exemple qui est représentatif mais un sous-ensemble d'exemples au travers de ses relations inter-exemples, ce sous-ensemble pourra très bien se retrouver «noyé» parmi d'autres exemples dans l'échantillon. Si on suppose que ce sont les relations inter-exemples qui sont importantes, cela signifie qu'il faille étudier la relation d'un exemple à tous les exemples. Ainsi, dans le cadre de l'apprentissage des listes de décisions, a-t-il fallu croiser tous les n-uplets et pas seulement les 0_m et les 1_m. Il n'est pas certain alors que le sous-ensemble d'exemples reste représentatif du concept lorsqu'il se trouve ainsi inclus dans un échantillon plus vaste. Si l'on envisage, par exemple, qu'un sous-ensemble est représentatif parce que ses éléments présentent une très forte similarité entre eux, lorsque l'on plonge ce sous-ensemble dans un échantillon plus grand, il est probable que l'on ne retrouve plus cette très forte similarité au sein de l'échantillon.

Des distributions bienveillantes telles que $1/P_{min(c)}$ soit polynomial en n

Nous pouvons alors proposer que les éléments d'un sous-ensemble caractéristique aient une plus forte probabilité que les autres exemples, ce qui nous amènerait à une deuxième proposition de recherche que nous avons esquissée dans la conclusion du chapitre 4 et qui consisterait à limiter les distributions bienveillantes aux seules distributions bienveillantes «polynomiales» où $1/P_{min(c)}$ serait polynomial en n, le nombre d'attributs. Nous avons vu que, dans le modèle d'apprentissage PAC avec distributions bienveillantes, on demande que le temps d'apprentissage soit polynomial, entre autres, en $1/P_{min(c)}$ et que si ce dernier paramètre est exponentiel en n, comme dans le cadre de la distribution uniforme, alors le temps d'apprentissage devient exponentiel. Il semble donc réaliste de réclamer un temps d'apprentissage qui soit toujours polynomial et par conséquent que $1/P_{min(c)}$ soit polynomial.

Pour illustrer ceci supposons que le sous-ensemble caractéristique soit 1111, 1110, 1101,1011, la première valeur d'attribut définissant l'appartenance au concept (cf. chapitre 3). Ce sous-ensemble est construit autour du prototype, 1111, en ne faisant varier qu'une seule valeur d'attribut et se caractérise par une très forte similarité interne. Si nous «noyons» ce sous-ensemble parmi les autres éléments du concept qui sont 1100,1001,1010,1000 nous obtenons 1100,1001,1010,1000, 1111, 1110, 1101,1011 ; il n'y a plus de similarité interne. En demandant que la distribution soit bienveillante avec $1/P_{min(c)}$ polynomial, ce cas de figure a alors une plus faible probabilité d'apparaître.

Des candidats pour la représentativité

Quel peut être le type de relation inter-exemple au sein du sous-ensemble représentatif qui rende justement ce sous-ensemble représentatif ? La *similarité* qui semble jouer un si grand rôle dans la catégorisation semble être une possibilité et elle est souvent utilisée en apprentissage automatique ([Michalski, 83], [Michalski et Stepp, 83], [Rendell et Seshu, 94]...).

Une autre possibilité serait qu'il existe *une relation d'ordre sur les exemples* (voir la présentation de l'apprentissage de Winston dans [Dietterich, Michalski, 83]). En apprentissage naturel supervisé (e.g. l'enseignement à l'école), si on ne présente pas n'importe pas quels exemples pour apprendre un concept, on ne les présente pas non plus dans n'importe quel ordre. En général, on part de l'exemple le plus simple qui est d'ailleurs bien souvent proche du prototype, le plus typique, pour aller vers les exemples les plus complexes qui sont souvent ceux les plus éloignés du prototype [Barth, 87]. Il semblerait intéressant d'exploiter ceci en théories formelles.

Collusion

Par cette démarche, on approcherait néanmoins dangereusement d'un problème de collusion. *Le danger de collusion apparaît dès que l'on commence à parler de sous-ensemble représentatif d'un concept.* Comme nous l'expliquions en conclusion du chapitre 4, un sous-ensemble représentatif n'a de sens que si l'apprenant exploite cette représentativité. Autrement dit, l'apprenant bénéficie d'une information supplémentaire sur l'échantillon en plus de celle contenue dans les exemples. Ainsi dans [Denis et Gilleron, 97, a] l'apprenant exploite le fait qu'il «sait» que les exemples les plus simples relativement au concept ont une probabilité non nulle d'être dans l'échantillon. Une autre manière de présenter la représentativité est de considérer comme le font [Goldman et Mathias, 96] que l'enseignant possède une certaine connaissance de l'apprenant et qu'il utilise cette connaissance dans le choix des exemples. Si on reprend l'exemple précédent, c'est parce que l'enseignant sait que l'apprenant va utiliser les exemples les plus simples relativement au concept, qu'il va s'arranger pour que ces exemples soient dans l'échantillon. Ainsi dans un sens comme dans l'autre, il y a de l'information en plus, soit celle que possède l'apprenant sur l'échantillon, soit celle que possède l'enseignant sur l'apprenant. Il y a collusion lorsque cette information est «trop» importante. Pour l'éviter [Goldman et Mathias, 96] proposent un Adversaire qui «enfouit» dans l'échantillon, sous d'autres exemples, les exemples du sous-ensemble représentatif proposé par l'enseignant. Cet Adversaire correspond dans le modèle PAC avec distributions bienveillantes au quantificateur universel sur ces distributions. Ce quantificateur semble peu compatible avec une présentation ordonnée des exemples.

Etudier les distributions de probabilités

Enfin, en informatique, reste la question de ce qu'est l'extension d'un concept lorsque l'on fait intervenir les distributions de probabilités. Une autre manière de poser cette question est de se demander face à un ensemble d'objets relevant du même concept ce

qui, dans cet ensemble, reflète le concept et ce qui, dans cet ensemble, reflète la distribution de probabilités qui l'a fait apparaître. Comme nous le disions dans l'introduction générale, qu'est ce qui distingue l'ensemble vide du singleton $\{a\}$ si la probabilité de l'élément a est nulle ? Cette question est au centre du modèle d'apprentissage PAC avec distributions bienveillantes. En théories formelles de l'apprentissage, l'accent jusqu'à présent a toujours été mis sur la classe de concepts, *il conviendrait peut-être maintenant de s'intéresser aussi aux distributions de probabilités*. Par exemple, le fait de réclamer, comme nous l'avons proposé ci-dessus, que $1/P_{min(c)}$ soit polynomial en n introduit une dichotomie parmi les distributions de probabilité qui est pertinente cognitivement : est-il utile d'étudier l'apprenabilité de concepts avec des distributions qui impliquent un temps d'apprentissage exponentiel et donc non réaliste ?

En psychologie

En psychologie aussi, nous avons évoqué diverses pistes de recherche au cours de ce travail.

L'étiquetage des exemples ou quelles sont les catégories que forme l'individu ?

La première concerne ce qui, dans le modèle PAC, correspond à l'étiquetage des exemples par l'Oracle. Cette notion d'étiquetage invite à distinguer entre la raison d'être d'une catégorie pour l'individu (le pourquoi de la catégorie) et la façon dont l'individu se forme une représentation de cette catégorie (le comment). L'individu place dans une même catégorie tous les objets vis-à-vis desquels il a un même comportement. C'est pourquoi nous disons que c'est l'environnement qui impose à l'individu les catégories qu'il doit former. Ce n'est pas l'individu qui *choisit* qu'un objet est dangereux pour lui, cet objet est dangereux pour lui indépendamment de sa volonté. De la même façon ce n'est pas lui qui *décide* d'appeler un chat «un chat », c'est son environnement linguistique qui lui impose. En ce sens, l'étiquette est un attribut particulier de l'objet qui exprime le rapport de l'individu à cet objet, et comme nous l'avons dit, c'est le seul attribut dont nous ayons la garantie qu'il soit partagé par tous les objets de la catégorie.

Le «comment » est très étudié en psychologie, le «pourquoi » commence à l'être. Ceci est compréhensible car, pour la plupart des catégories étudiées, le «pourquoi » est la plupart du temps donné par l'environnement linguistique : étude de la catégorie «oiseaux », de la catégorie «fleurs». L'individu néanmoins ne se forme pas *uniquement* les catégories qui lui sont imposées par la langue. L'expérience de [Mazet, 93] le montre bien lorsqu'elle met en évidence que le conducteur se construit des catégories relativement aux sections de route : dangereuses ou non. Ce type de recherche devrait mettre ainsi à jour les différents étiquetages observés par l'individu autres que ceux imposés par la langue. Parmi ceux-ci, nous devrions très vite pouvoir en repérer quelques principaux tel celui de mal-être/danger comme dans l'expérience de Mazet et son pendant, celui de bien-être/sécurité. L'un et l'autre sont issus du fait que la catégorisation est avant tout un processus qui permet de s'adapter à l'environnement et ces étiquetages ne sont qu'une traduction de cette adaptation.

Typicalité en aval de l'apprentissage et représentativité en amont

Le deuxième type de recherche possible concerne le «comment ». Elle nous intéresse tout particulièrement car il s'agirait de vérifier s'il existe *bien un lien entre représentativité et typicalité*. Une fois que l'environnement a imposé à l'individu les catégories qu'il doit former, celui-ci doit pouvoir se construire une représentation de chacune de ces catégories, c'est ici qu'intervient le «comment ». Ce que nous avons vu au chapitre 3 invite à penser que l'individu se construit une représentation en utilisant la similarité inter-exemples de la catégorie. Ceci peut alors l'amener à éclater un ensemble d'objets qui pourraient relever d'une même catégorie en plusieurs catégories si la similarité interne est trop faible. Si on considère le feu et l'ortie, l'un et l'autre «brûlent » la peau mais sont cependant trop dissemblables pour être placés dans une même catégorie ou, dit autrement, les placer dans une même catégorie ne serait pas cognitivement économiques[109]. La représentativité d'un échantillon consisterait alors dans sa forte similarité interne, et pourrait expliquer la typicalité si l'on considère celle-ci comme exprimant, en partie, la fréquence de valeurs d'attributs.

C'est d'ailleurs ce qui apparaît dans les expériences de [Kemler-Nelson, 84] lorsqu'elle propose les exemples d'apprentissage suivants :

Catégorie «docteur»	Catégorie «policier»
0000	1111
0100	1011
0010	1101
0001	1110

On peut constater que les exemples représentent une forte similarité interne (représentativité) mais dans le même temps, certaines valeurs d'attributs sont sur-représentées ce qui pourrait expliquer une typicalité ultérieure[110]. Il n'y a aucun exemple de la forme 1001 ou 0110. Cet exemple est simpliste, il permet néanmoins de montrer comment la typicalité, en aval de la catégorisation, peut avoir comme origine la représentativité, en amont de l'apprentissage. Le fait que la similarité soit très forte au sein de l'échantillon implique que la fréquence de certaines valeurs d'attributs soit plus élevée, dégageant ainsi des éléments plus ou moins typiques. Comme nous l'avons dit dans le chapitre 3, ce type de recherche pourrait expliquer pourquoi l'apprentissage est plus rapide lorsque l'on présente des sous-catégories typiques. Cela supposerait qu'un échantillon d'apprentissage constitué de sous-catégories typiques présente une plus grande similarité interne.

---

[109] Néanmoins ce qui est premier reste la pression de l'environnement à placer dans une même catégorie certains éléments et non la similarité. Ainsi deux objets très similaires seront placés dans des catégories différentes par l'individu car ils entraînent de la part de celui-ci des comportements différents : l'enfant ne place pas dans la catégorie «canard» son canard en plastique.

[110] Kemler-Nelson ne parle que de l'air de famille cependant cette notion a des liens très étroits avec la typicalité. On le voit ici où l'air de famille est construit autour du prototype en tête de colonne.

Etudier le rôle des distributions de probabilités

Ce type d'interrogation nous amène naturellement au rôle des distributions de probabilités dans la catégorisation. Cet aspect est peu étudié en psychologie si l'on excepte les travaux d'Holyoak (cf. [Flannagan, Fried, et Holyoak, 1986]. [Fried et Holyoak, 84]...). Si nous considérons la typicalité de la manière dont nous l'avons envisagée dans les exemples ci-dessus, nous constatons qu'au travers de la fréquence des valeurs d'attributs, celle-ci reflète la distribution de probabilités sous-jacente à la présentation des exemples. Les travaux d'Holyoak sont donc très intéressants de ce point de vue mais, comme nous l'avons signalé dans le chapitre 5, ce qui nous semble erroné dans leur démarche est qu'une catégorie ne se définit *que* par sa distribution de probabilités et pris individuellement un objet donné n'appartient à une catégorie qu'avec une certaine probabilité. Il pourrait donc être intéressant de reprendre leurs travaux mais en y intégrant la notion d'étiquetage des objets imposé par l'environnement. Nous en arriverions peut-être alors à montrer que si les distributions normales sont plus faciles à apprendre, comme le constatent ces auteurs, c'est parce qu'elles sont construites sur la base d'une certaine similarité entre les éléments.

Des calculs de complexité

Un autre point qu'il nous semble intéressant de développer consiste dans les calculs de complexité : complexité en temps, complexité en espace de travail. Nous avons expliqué que le modèle PAC est aussi un modèle *quantitatif* : un concept n'est apprenable que si cet apprentissage se déroule en un temps polynomial avec un nombre polynomial d'exemples. Cet aspect quantitatif de l'apprentissage obéit donc à un principe de réalité. En psychologie on retrouve, sous ces calculs de complexité, des notions telle que l'économie cognitive. Nous supposons que les ressources d'un individu sont limitées mais comment mettre à jour ces limites ? La difficulté en psychologie est de pouvoir définir des unités : s'il est aisé de découper le temps en tranches, qu'en est-il de l'espace de travail d'un individu, de sa mémoire, sur quelle unité se baser ? Il est possible que la réponse à cette question ne puisse provenir que d'une étude au niveau neurobiologique.

Deux schémas de représentation pour l'individu

Enfin, une dernière question est l'opposition qui peut exister entre les deux schémas de représentation, entre processus de bas niveau et processus de haut niveau, entre processus subsymboliques et processus symboliques. Nous avons déjà dit que des chercheurs tel que [Medin, Wattenmaker et Hampson, 87], [Bideaud et Houdé, 93]. [Dubois, 93]... se demandent si des phénomènes telle que la typicalité ne proviennent pas de processus davantage subsymboliques, ce que confirme en quelque sorte [Kemler-Nelson, 84] avec les résultats qu'elle a obtenus sur l'apprentissage incident. Nous allons dans ce sens et nous nous demandons même si cette opposition ne concerne pas aussi les catégories de base : dans quelle mesure celles-ci ne relèveraient-elles pas davantage du niveau subsymbolique alors que les catégories sur et sous-ordonnées relèveraient davantage du niveau symbolique ? Ceci pourrait expliquer pourquoi c'est avec ces catégories que la typicalité est le plus facilement mise en évidence.

Les problèmes qui restent posés

Derrière toutes ces pistes de recherches, restent latents certains problèmes. L'un d'entre eux vient juste d'être évoqué, comment définir ce que peut être l'unité de base qui permettrait de quantifier l'espace de travail disponible d'un individu ? Cette question est très proche de celle que [Schyns et Rodet, 98] ont appelé «la quête des quanta de la cognition » : quelle est la plus petite unité d'information sur laquelle s'appuie la cognition humaine ?

Une autre assez similaire concerne l'ensemble des attributs disponibles lors de la catégorisation. Nous avons fait la distinction lors du chapitre 2 entre l'ensemble des attributs pertinents, qui interviennent dans la représentation de la catégorie, et l'ensemble des attributs disponibles lors de la catégorisation dans lequel seront sélectionnés les attributs pertinents. Théoriquement, l'espace des attributs disponibles est infini mais dans la pratique l'individu n'en dispose que d'un nombre fini lorsqu'il catégorise. Il faudrait maintenant pouvoir définir quels sont ces attributs disponibles. Autrement comment parler de similarité dans l'échantillon ou de typicalité reflétant les fréquences de valeur d'attributs si l'on ne connaît pas ces attributs ? Comment parler de la complexité d'un concept s'il n'est pas possible de définir l'espace d'attributs sur lequel est construit le concept.

Là encore il est à craindre que la réponse ne puisse être que du niveau neurobiologique et qu'au niveau psychologique, on ne puisse travailler que par approximation. C'est ce que nous avons fait lors des deux expériences quand nous avons supposé que les seuls attributs qui seraient pris en compte par les sujets pour percevoir les grilles seraient les différentes cases de ces grilles. En a-t-il réellement été ainsi ?

Quelques remarques concernant un travail dans le cadre des sciences cognitives

Pour conclure nous souhaiterions faire quelques remarques concernant les disciplines, informatique et psychologie, dans lesquelles nous avons travaillé.

La première concerne l'opposition que l'on fait habituellement entre sciences dures (sciences exactes) et sciences molles (sciences humaines). En jouant sur le mot «dur», un chercheur nous a dit un jour que les sciences dures sont certainement moins dures/difficiles que les sciences molles. Nous rejoignons son point de vue, outre la difficulté pour la mise en place de ses expériences[111], le chercheur en sciences humaines doit constamment se frotter à une réalité mouvante dans laquelle la reproductibilité de l'expérience ne va pas toujours de soi, ce qui n'est pas le cas en sciences exactes. C'est ce qui a amené un autre chercheur en sciences humaines à nous parler de celles-ci

[111] Nous avons pu constater par exemple qu'il n'est pas si évident que cela de recruter une vingtaine de volontaires pour une expérience.

comme des sciences du doute. Alors que l'informaticien, lui, a la garantie que si sa démonstration est bonne alors seule une remise en question de ses axiomes de départ pourra infirmer ses résultats, le psychologue risque toujours de s'apercevoir qu'une variable parasite traînait dans son expérience qui a biaisé ses résultats.

Si nous avons été impressionné par les techniques utilisées par les psychologues pour affronter ce réel, pour le découper, nous l'avons aussi été par la capacité des informaticiens/mathématiciens à modéliser ce même réel. Voir comment Valiant en une définition de quelques lignes a pu appréhender un phénomène tel que l'apprentissage/la catégorisation, voir comment, dans ces quelques lignes, l'apprentissage y est d'emblée considéré comme une adaptation à l'environnement, voir comment Denis et Gilleron, n'en déplaise à leur modestie, invitent à s'interroger sur ce qu'est l'extension d'un concept, nous a fasciné.

Une dernière remarque enfin concerne le travail dans le cadre des sciences cognitives. Nous sommes persuadés de la validité d'un tel travail car il permet d'avoir plusieurs points de vue sur un même problème ce qui peut amener à des solutions qui ne seraient pas apparues en envisageant ce problème à partir d'une seule discipline. Cependant ce type de travail, pour être correctement réalisé, doit être l'œuvre d'une équipe pluridisciplinaire, il ne peut être l'œuvre d'une seule personne qui serait à la fois spécialiste dans les deux domaines[112]. De plus, ce type de travail nécessite que les membres de cette équipe arrivent à communiquer, qu'ils ne soient pas trop enfermés dans leur spécialité[113] et qu'ils aient donc une certaine connaissance de ce qui se passe dans la discipline voisine. Ainsi, cette recherche a été pensée comme une contribution éventuelle à l'amélioration de cette communication.

[112] S'il invite à s'interroger sur la nature de la catégorisation, l'échec de la seconde expérience montre surtout les limites du psychologue néophyte que nous sommes.

[113] Il nous est arrivé de débattre pendant plus d'une heure avec des psychologues pour nous apercevoir à la fin que nous étions d'accord et que seule la barrière du langage propre à chaque discipline faisait que nous ne nous étions pas compris.

Table des tableaux, figures et schémas

Annexe 1

Deux exemples d'apprentissage automatique

Dans cette annexe nous présentons deux exemples d'apprentissage automatique, celui des arbres de décisions et celui des réseaux de neurones artificiels. Ces exemples permettent de se faire une intuition sur ce que peuvent être en informatique des systèmes apprenants.

Les arbres de décision

Les arbres de décision sont un des moyens les plus largement utilisés en apprentissage automatique symbolique. Dans un arbre de décision les exemples sont classifiés en les faisant descendre l'arbre, un peu paradoxalement, de la racine vers les feuilles. Chaque nœud de l'arbre est un test qui oriente l'exemple vers une des branches subalternes, chaque feuille de l'arbre correspond à une classe possible.

Supposons que Monsieur X aille voir son banquier pour obtenir un prêt. Celui-ci lui demande sa profession, son état civil, s'il a des enfants et s'il est propriétaire ou locataire de son logement. Monsieur X indique qu'il est ingénieur informatique dans le secteur privé, qu'il est marié, qu'il a des enfants et qu'il est propriétaire. Le banquier entre tous ces renseignements dans son ordinateur de bureau où un arbre de décision a déjà été construit (voir ci-dessous) par le programme ID3 [Quinlan,79][Quinlan, 83] et celui-ci indique que le client est probablement non solvable. Monsieur X a alors le désagrément de se voir refuser son prêt par le banquier.

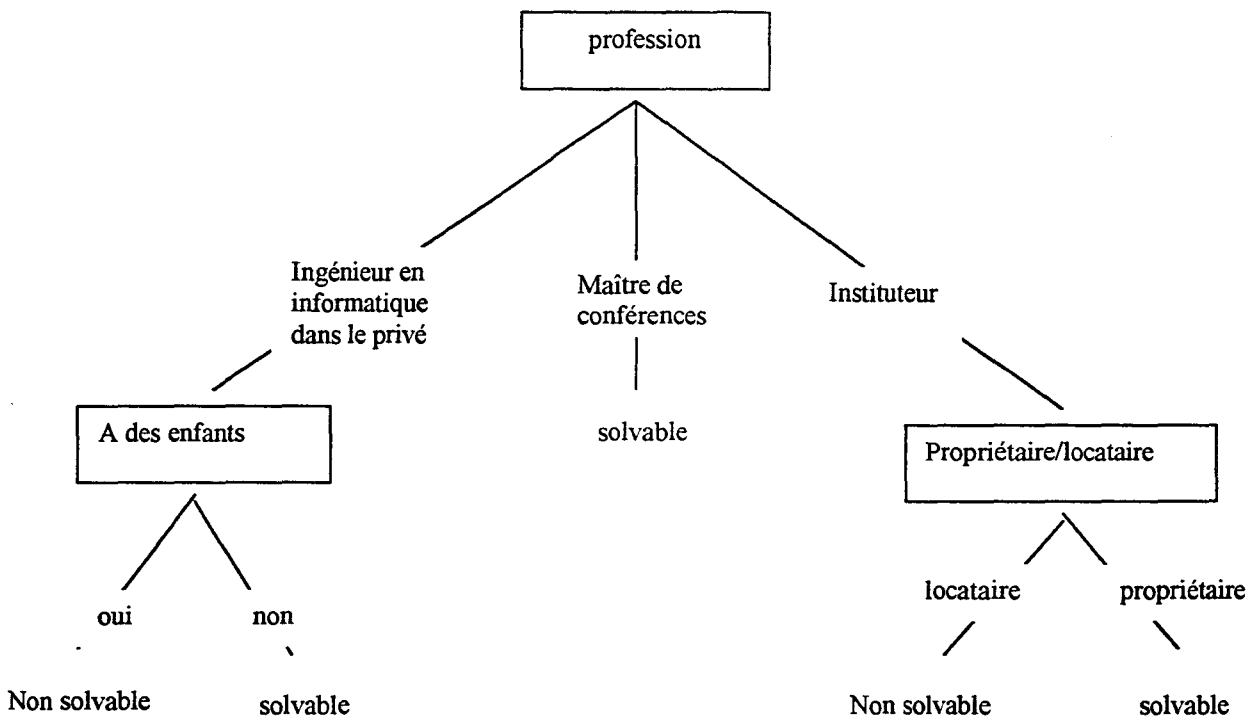


Schéma AN1 un arbre de décision dont la cible est « client solvable ou non » selon [Mitchell, 97][114]

[114] Ce schéma comme le tableau ci-après est une adaptation de celui de [Mitchell, 97] et n'a aucune réalité sociologique

Pour décider que le client est non solvable, le programme lui a fait parcourir l'arbre de haut en bas. Monsieur X étant ingénieur en informatique, le premier test « profession » l'oriente sur la branche gauche, le second test « a des enfants » l'oriente sur la feuille « non solvable ».

L'apprentissage dans les arbres de décision

Un des premiers programmes d'apprentissage dans les arbres de décision a été ID3 de Quinlan. Ce programme construit des arbres de décisions sur la base d'un échantillon d'exemples étiquetés. Le tableau ci-dessous décrit l'échantillon qui a permis de construire l'arbre de décision vu plus haut.

Client	Profession	Etat civil	a des enfants	Propriétaire/ Locataire	solvable
C1	Ingénieur informatique dans le privé	Marié	Oui	Locataire	Non
C2	Ingénieur informatique dans le privé	Marié	Oui	Propriétaire	Non
C3	Maître de conférences	Marié	Oui	Locataire	Oui
C4	Instituteur	Célibataire	Oui	Locataire	Oui
C5	Instituteur	Divorcé	Non	Locataire	Oui
C6	Instituteur	Divorcé	Non	Propriétaire	Non
C7	Maître de conférences	Divorcé	Non	Propriétaire	Oui
C8	Ingénieur informatique dans le privé	Célibataire	Oui	Locataire	Non
C9	Ingénieur informatique dans le privé	Divorcé	Non	Locataire	Oui
C10	Instituteur	Célibataire	Non	Locataire	Oui
C11	Ingénieur informatique dans le privé	Célibataire	Non	Propriétaire	Oui
C12	Maître de conférences	Célibataire	Oui	Propriétaire	Oui
C13	Maître de conférences	Marié	Non	Locataire	Oui
C14	Instituteur	Célibataire	Oui	Propriétaire	Non

Tableau AN1
Exemples d'entraînement pour construire un arbre de décision dont la cible est « client solvable ou non », selon [Mitchell, 97]

Le problème principal pour construire un arbre de décision est de choisir les « meilleurs » nœuds, c'est-à-dire de trouver quels sont les « meilleurs » attributs qui doivent servir de tests. La difficulté est de trouver une mesure qui permette de dire qu'un attribut est meilleur qu'un autre. Dans ID3, on utilise une mesure de gain d'information qui est une propriété statistique basée sur une autre mesure appelée entropie.

$$\text{Entropie}(E) = -(p_+ \log_2 p_+) - (p_- \log_2 p_-)$$

Où E représente l'échantillon (les 14 clients du banquier), p_+ le nombre d'exemples positifs sur le nombre d'exemples total (le nombre de clients solvables sur le nombre de clients total) et p_- le nombre d'exemples négatifs sur le nombre d'exemples total. Si on calcule ainsi l'entropie de notre échantillon où nous avons 9 exemples positifs et 5 négatifs elle sera de :

$$\text{Entropie}[9+,5-] = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0{,}940$$

Le gain d'information, lui, est donné par la formule :

$$Gain\ (E, A)\ = Entropie\ (E) - \sum_{v \in\ valeurs\ (A)} \frac{|Ev|}{|E|}\quad Entropie\ (Ev)$$

Où E est l'échantillon, A un attribut, gain(E,A) le gain d'information que l'on réalise en choisissant l'attribut A et Ev l'ensemble des exemples de E qui sont vrais pour la valeur v de l'attribut A.

Pour chaque nœud de l'arbre ID3 va calculer le gain d'information apporté par chacun des attributs. Ainsi au démarrage ID3 peut choisir entre « profession », « état civil », « a des enfants », « propriétaire /locataire » :

Gain(E, profession)=0,246
Gain(E, a des enfants)=0,151
Gain(E, propriétaire/locataire)=0,048
Gain(E, Etat civil)=0,029

Comme l'attribut profession apporte le plus d'informations, c'est celui-ci qui sera choisi comme racine ce qui donne le début d'arbre suivant :

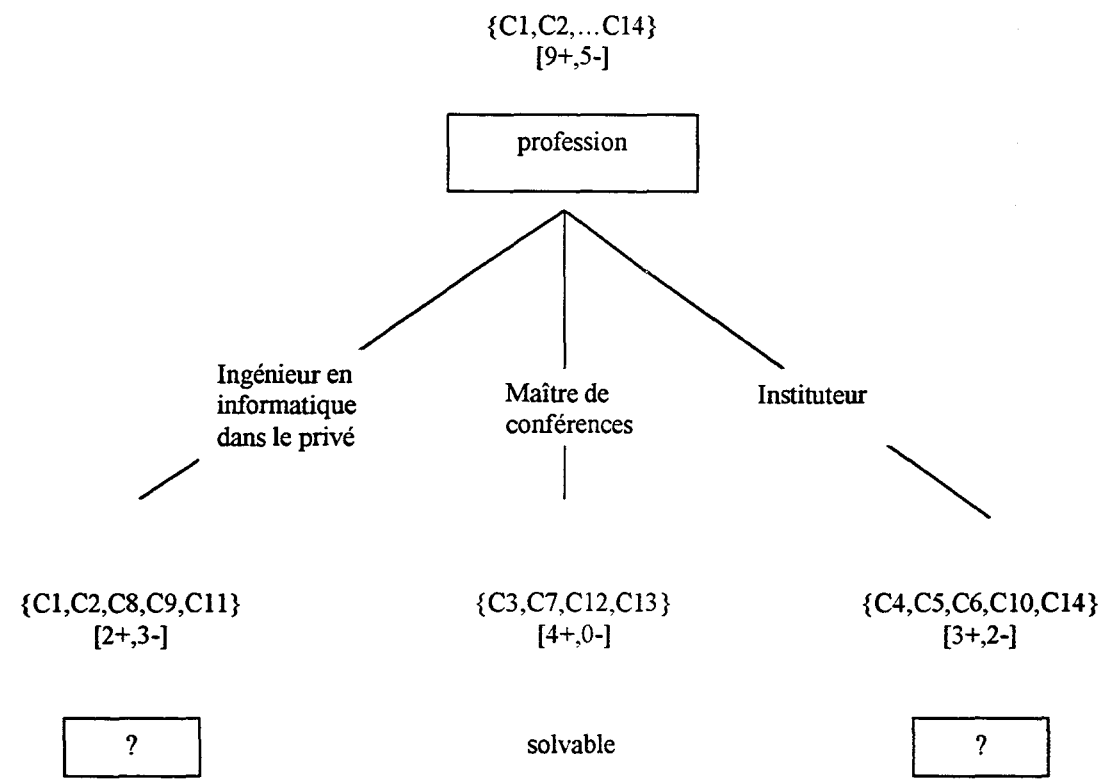


Schéma AN2
Choix du premier nœud pour l'arbre de décision dont la cible est « client solvable ou non » [Mitchell, 97]

Avec ce premier nœud, certains exemples sont définitivement classés, il s'agit des « maîtres de conférences » qui sont tous solvables. Si on calcule l'entropie sur ce sous-ensemble {C3,C7,C12,C13} elle est de 0.
Pour les deux autres branches, il faut à nouveau choisir des tests, car les sous-ensembles obtenus contiennent des éléments positifs et des éléments négatifs. Pour cela ID3 utilise à nouveau le gain d'entropie mais cette fois-ci par rapport aux nouveaux sous-ensembles {C1,C2,C8,C9,C11} et {C4,C5,C6,C10,C14}. L'attribut « profession » ne peut plus être utilisé car il est déjà dans les ascendants. Le résultat donne l'arbre proposé au début de cette section.

Une des difficultés de l'apprentissage des arbres de décision est la surspécialisation. Supposons qu'un des exemples de l'échantillon soit mal étiqueté, l'arbre de décision se développera de manière à prendre en compte cet exemple erroné. Il risque alors d'étiqueter de nouveaux exemples de façon erronée. Une autre cause possible de surspécialisation est que l'échantillon soit insuffisamment grand et que l'arbre de décision prenne en compte des régularités qui ne sont là que parce que l'échantillon est trop petit. Est-ce que tous les maîtres de conférence sont solvables ? Sur la base de l'échantillon qui lui a été donné, c'est ce qu'a « décidé » ID3. Il existe diverses techniques pour éviter cela [Mitchell,97]. L'une d'entre elles est l'élagage qui va consister à supprimer des branches de façon à ce que l'arbre ne soit pas trop spécifique à l'échantillon proposé.

Commentaires

Il est nécessaire de distinguer entre la phase d'apprentissage et la phase d'utilisation. Dans la phase d'apprentissage on présente à ID3 des exemples étiquetés qui lui permettent de construire l'arbre de décision. Dans la phase d'utilisation, on fournit des exemples non étiquetés et c'est ID3 qui va les étiqueter à partir de l'arbre de décision qu'il a formé durant l'apprentissage.

Les mesures d'entropie décrites ici sont celles utilisées par ID3, il en existe d'autres (par exemple une mesure de distance [Lopez de Mantaras, 1991]), mais elles jouent globalement le même rôle.

L'espace des hypothèses dans ID3 est complet. C'est-à-dire qu'ID3 est capable de former toutes les hypothèses possibles sur la base des attributs et de leurs valeurs, En d'autres termes, à toute partition de l'ensemble de tous les exemples que l'on peut décrire avec ces attributs peut correspondre une hypothèse définie par un arbre de décision. Cependant durant l'apprentissage, étant donné un échantillon particulier, ID3 ne maintient qu'une seule hypothèse, plutôt que les considérer toutes, et ne fait pas de backtracking, c'est-à-dire que le choix d'un nœud est définitif, il ne sera plus reconsidéré. Le problème avec cette démarche est qu'elle peut s'arrêter sur un minimum local, tel que la minimisation de l'erreur peut ne pas être optimale.
Le fait qu'ID3 ne fait qu'une seule hypothèse à la fois s'appuie sur un biais d'apprentissage qui consiste à choisir l'hypothèse la plus courte possible : le moins de nœuds possibles. Un tel biais se justifie par le rasoir d'Occam (cf. chapitre 3) qui préconise de choisir la plus petite hypothèse consistante avec les exemples.

Les problèmes appropriés pour les arbres de décisions sont ceux dans lesquels :

- les exemples sont représentés par des attributs sur des valeurs discrètes mais il existe des techniques pour discrétiser des valeurs continues [Fayyad, 1991].
- les cibles sont des fonctions sur des valeurs discrètes elles-aussi. Ici la cible était binaire (solvable : oui/non), il est néanmoins possible de travailler avec des cibles à classe multiples.
- pour lesquels des représentations disjonctives peuvent être demandées, ce qui n'est pas systématiquement le cas pour tous les systèmes d'apprentissage,
- pour lesquels les données peuvent éventuellement contenir des erreurs ou n'ont pas de valeurs, il existe ainsi des techniques où l'on donnera par défaut la valeur majoritaire [Mingers, 1989] etc...

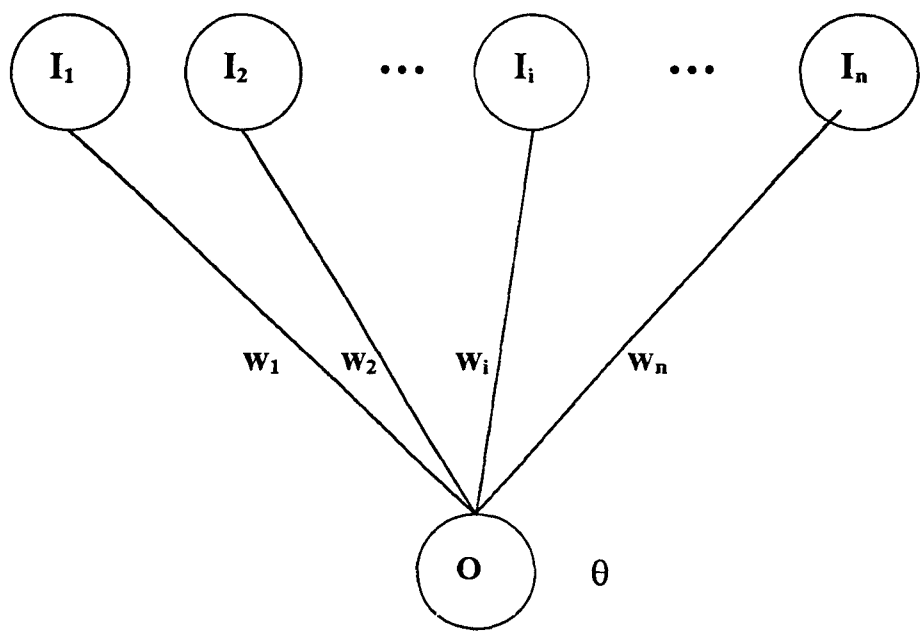
L'apprentissage des arbres de décisions est utilisé pour repérer des régularités dans de grandes bases de données. La NASA l'utilise pour classifier automatiquement des objets célestes [Fayyad et al, 1995]. Dans le domaine des jeux, TD-GAMMON de [Tesauro, 1995] basé aussi sur les arbres de décisions est capable de jouer au backgammon à un niveau international. Toutes les stratégies utilisées par l'application ont été apprises en la faisant jouer contre elle-même.

Apprentissage des réseaux de neurones artificiels [Denis et Gilleron, 97]

Les recherches sur les réseaux de neurones artificiels ne sont pas récentes. Dès 1943, McCulloch et Pitts proposent de simuler le fonctionnement du cerveau au travers de celui des neurones artificiels. Bien entendu la simulation est grossière.

Le perceptron

Un des premiers modèles proposés a été le perceptron. Dans certains modèles de perceptron nous avons *n* neurones d'entrée ($I_1,\ldots,I_n$) et un seul neurone de sortie (O). Tous les neurones d'entrée sont reliés au neurone de sortie par des connexions synaptiques. Chaque connexion synaptique est affectée d'un coefficient synaptique($w_1,\ldots,w_n$) ou poids synaptique.

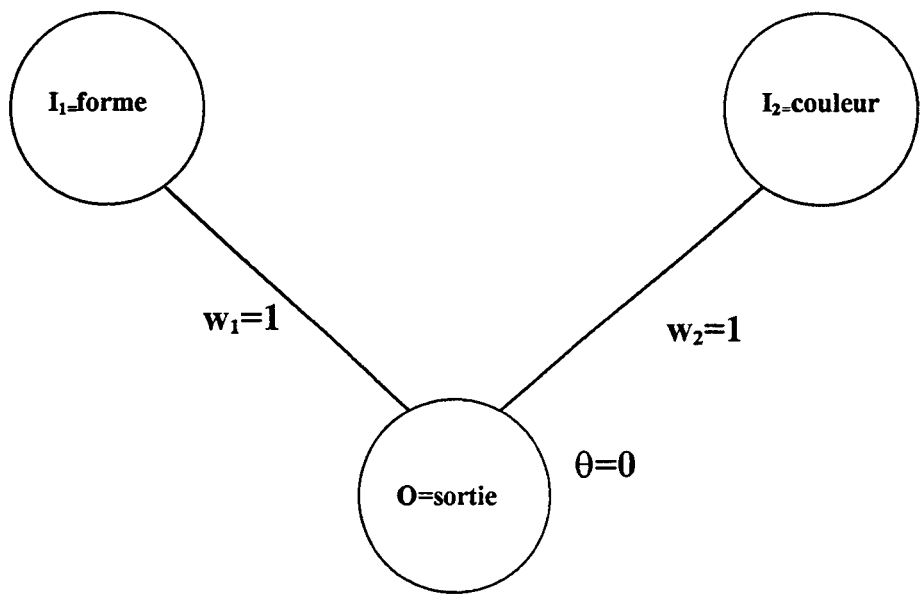


On présente un exemple aux neurones d'entrée et le neurone de sortie propose une classification de cet exemple. Les neurones d'entrée correspondent en quelque sorte aux sens qui perçoivent l'exemple et le neurone de sortie donne la catégorie de l'exemple.
La sortie est calculée ainsi : si la somme des entrées multipliées par les poids synaptiques est au-dessus d'un certain seuil θ alors la sortie est 1 sinon c'est 0.

$$O = \begin{cases} 1 \ si \ \sum_i w_i I_i > \theta \\ 0 \ \text{sinon} \end{cases}$$

Exemple :

On suppose un ensemble d'exemples décrits par deux attributs qui peuvent prendre deux valeurs : la couleur (blanc ou noir), la forme (carré ou rond). Le « neurone de la forme » prend la valeur 1 quand l'exemple est un carré et la valeur 0 quand c'est un rond, de la même manière le « neurone de la couleur » prend la valeur 1 quand l'exemple est noir et 0 quand l'exemple est blanc. Le perceptron ci-dessous permet de classifier tous les exemples du concept « carré ou noir ». Il suffit pour cela de prendre $w_i=1$ $w_2=1$ et $\theta=0$.



Le tableau suivant indique les calculs effectués pour chacun des exemples. On constate que tous les exemples positifs du concept « carré ou noir » sont classés comme tels par le perceptron.

exemple	codage	I_1	I_2	$w_1 I_1 + w_2 I_2$	$w_1 I_1 + w_2 I_2 > \theta$	Sortie
Carré noir	11	1	1	2	Oui	1
Carré blanc	10	1	0	1	Oui	1
Rond noir	01	0	1	1	Oui	1
Rond blanc	00	0	0	0	Non	0

Cette description est celle d'un perceptron qui a déjà appris à classifier, il s'agit de voir maintenant comment se déroule cet apprentissage

L'apprentissage dans un perceptron

L'apprentissage dans un perceptron va consister à corriger les erreurs. Il y a erreur quand la réponse donnée par le perceptron est différente de la réponse attendue. L'algorithme d'apprentissage peut être décrit de la manière suivante. On initialise les poids et le seuil du perceptron à des valeurs quelconques . A chaque fois que l'on présente un nouvel exemple, on ajuste les poids et le seuil selon que le perceptron l'a correctement classé ou non. L'algorithme s'arrête lorsque tous les exemples ont été présentés sans qu'il y ait modification d'aucun poids ni du seuil. Dans l'algorithme ci-dessous le paramètre (t) (le temps) permet de repérer les étapes successives de l'apprentissage.

Algorithme
Entrée un échantillon S
Initialisation du paramètre t à O.
Initialisation des poids $w_i(t)$ et du seuil θ à des valeurs quelconques.
Répéter
Choisir un exemple s de S
Présenter cet exemple au perceptron
Soient T la sortie attendue et O la sortie calculée par le perceptron
Si $T \neq 0$ alors
Modifier les poids et le seuil selon les formules suivantes :
$w_i(t+1)=w_i(t)+\Delta w_i(t)$ où $\Delta w_i(t)=(T-O)I_i$
$\theta(t+1)= \theta(t)+ \Delta\theta(t)$ où $\Delta\theta(t)=-(T-O)$
fin de si
$t=t+1$
jusqu'à ce que tous les exemples de S aient été présentés au perceptron sans modification de celui-ci
Sortie : un perceptron qui discrimine S

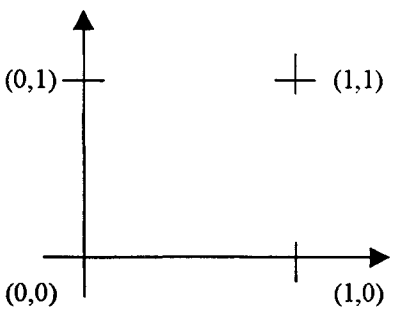
On note que :si O=0 et T=1 cela signifie que le perceptron n'a pas assez pris en compte les neurones actifs de l'entrée : dans ce cas l'algorithme ajoute la valeur de ces entrées aux poids synaptiques (renforcement), et si O=1 et T=0, l'algorithme retranche ces valeurs aux poids synaptique (inhibition).

Exemple : apprentissage du OU vu précédemment. On initialise $w_1(0)=w_2(0)= \theta(0)=1$. Le tableau ci-dessous indique les différentes étapes de l'apprentissage. On note que l'on en arrive à $w_1(0)=w_2(0)=1$ $\theta(0)=0$ qui ne sont pas les mêmes valeurs que celles décrites plus haut

t	Entrée	$w_1(t)$	$w_2(t)$	$\Sigma w_i I_i$	θ	O	T	$\Delta w_1(t)$	$\Delta w_2(t)$	$\Delta\theta$
0	00	1	1	0	1	0	0	0	0	0
1	01	1	1	1	1	0	1	0	1	-1
2	10	1	2	1	0	1	1	0	0	0
3	11	1	2	3	0	1	1	0	0	0
4	00	1	2	0	0	0	0	0	0	0
5	01	1	2	2	0	1	1	0	0	0

Il faut bien distinguer entre la phase d'apprentissage et la phase « d'utilisation ». En phase d'apprentissage, ce sont les poids synaptiques et le seuil qui sont les variables alors que les entrées de l'échantillon S apparaissent comme des constantes. En phase de calcul les poids synaptiques et le seuil sont devenus des constantes.

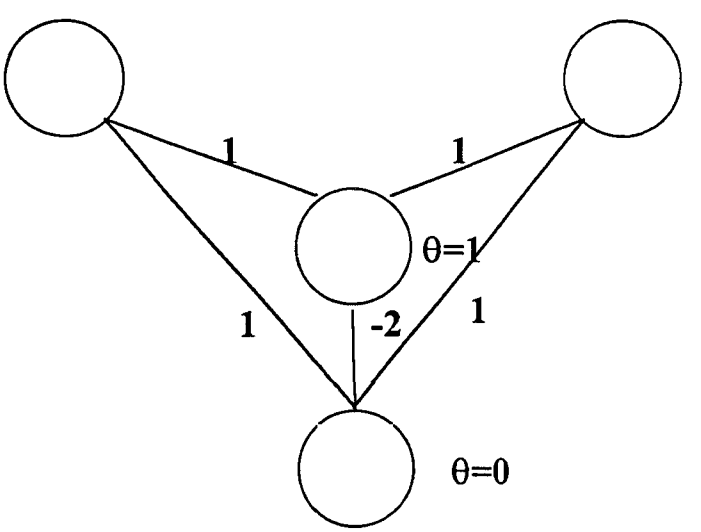
Les travaux de recherche sur le perceptron vont continuer jusque dans les années 60 : Widrow et Hoff (1960) proposent Adaline et la règle Delta, Rosenblatt (1962) prouve la convergence de la règle d'entraînement du perceptron. Mais il devient clair que les perceptrons à un seul niveau ont des capacités représentationnelles limitées. Le problème avec le perceptron est qu'il ne permet de classifier que des éléments linéairement séparables. Ainsi, si un 'OU' peut être calculé par un perceptron, il n'en va pas de même du XOR (« OU exclusif »), car les éléments d'un XOR ne sont pas linéairement séparables : on ne peut trouver une droite qui sépare les éléments (0,1) et (1,0) (positifs) des éléments (0,0) et (1,1) éléments négatifs. Comme on peut le voir dans le schéma ci-dessous.



Suite au livre « Perceptrons » de [Minsky et Papert, 69] qui met en lumière les limites du perceptron, la recherche sur les réseaux de neurones régresse dans les années 70. C'est dans le milieu des années 80 avec l'invention de la rétropropagation pour l'entraînement des réseaux multicouches [Rumelhart et McClelland, 86],[Parker, 1985] que les recherches reprennent.

Les réseaux multicouches

Les réseaux multicouches permettent de calculer plus de fonctions booléennes que le perceptron ainsi le XOR peut-être calculé par le réseau suivant où l'on a simplement rajouté un neurone entre les neurones d'entrée et celui de la sortie.



Dans les réseaux multicouches on ajoute entre la couche d'entrée et celle de sortie des couches de neurones cachés. Dans ces réseaux, tous les neurones des couches intermédiaires sont reliés en entrée avec tous ceux de la couche précédente et en sortie avec tous ceux de la couche suivante. (Ce n'est pas le cas du XOR ci-dessus mais il est aisé d'en construire un qui corresponde à cette définition).

Pour l'apprentissage des réseaux multicouches, il est nécessaire de définir d'abord l'architecture du réseau, le nombre de couches et le nombre de neurones dans chacune de ces couches, puis de trouver un algorithme qui calcule à partir de l'échantillon d'apprentissage les valeurs des poids synaptiques et des seuils. Le deuxième problème est en général résolu par l'algorithme de rétropropagation du gradient. Celui-ci repose sur le même principe de correction d'erreur que celui que l'on vient de voir. L'idée est de faire remonter l'erreur parmi les différentes couches et de la corriger à chaque niveau en modifiant les poids et les seuils.

Contrairement au perceptron il n'est pas nécessaire ici de présenter tous les exemples, et de chercher à obtenir une erreur nulle, c'est d'ailleurs une des raisons qui font que les réseaux multicouches sont résistants aux bruits . La difficulté avec l'algorithme de rétropropagation est de trouver une condition de terminaison, (quand l'algorithme doit-il s'arrêter ?). Celle-ci pourrait être d'attendre que le réseau descende en dessous d'un certain seuil d'erreur sur les exemples mais c'est une mauvaise stratégie car elle peut amener à une surspécialisation sur les exemples qui diminue l'adéquation sur les exemples non vus. Un moyen utilisé pour éviter cela est d'utiliser un échantillon test et, avec certaines précautions, d'arrêter l'apprentissage dès que les résultats sur l'échantillon test ne montrent plus une baisse de l'erreur mais une remontée de celle-ci.

L'intérêt des réseaux de neurones

L'une des motivations dans l'apprentissage des réseaux de neurones artificiels (RNA) est d'essayer de capturer le genre de calcul hautement parallèle effectué par le cerveau sur des représentations distribuées. L'apprentissage par RNA est particulièrement bien approprié à des problèmes où l'échantillon d'entraînement est constitué de données issues de senseurs, de capteurs tels que des caméras ou des microphones. Par ailleurs, l'apprentissage RNA est robuste aux erreurs dans l'échantillon. Cela fait qu'il est particulièrement bien adapté à des problèmes du monde réel. Il est ainsi utilisé dans la reconnaissance de caractères [LeCun et al, 1989] dans la reconnaissance de la parole [Lang et al, 1990] dans la reconnaissance de visages [Cottrell, 1990].
La plupart des RNA sont émulés sur des machines séquentielles, telles que celles que l'on retrouve sur la plupart des bureaux, quelques uns ont été implémentés sur des machines parallèles où il y a plusieurs processeurs interconnectés, ou, enfin, sur des machines conçues pour eux.

Il faut noter que dans un réseau de neurones toute affectation possible des poids et du seuil représente une hypothèse différente. En d'autres mots l'espace d'hypothèse est l'espace euclidien de dimensions n des n poids du réseau. Cet espace d'hypothèses est continu à l'opposé de l'espace d'hypothèses des arbres de décision basés sur des représentations discrètes.

Annexe 2

Apprentissage des conjonctions

La démonstration de l'apprentissage PAC des conjonctions a été proposé pour la première fois par Valiant dans son article fondateur du modèle [Valiant, 84]. Elle est reprise dans [Kearns, Vazirani, 94].

La classe des concepts cible est celle des conjonctions plus exactement des 1-CNF, c'est-à-dire des conjonctions de littéraux. A la différence des conjonctions que nous avons proposées dans l'exemple 1.1, elles ne sont pas limitées, ici, par le nombre de littéraux. Ces conjonctions sont définies sur un monde d'objets X_n à n variables. Pour illustrer la démonstration, nous prendrons n=5 et comme concept cible $c=x_1 \wedge \neg x_4 \wedge x_5$

x_1, $\neg x_4$, x_5 sont des exemples de littéraux z. Si z correspond à x_i, dire que z=0 signifie que $x_i=0$ (et donc $\neg x_i=1$) et réciproquement pour z=1 signifie que $x_i=1$ (et donc $\neg x_i=0$). A l'opposé si z correspond $\neg x_i$, dire que z=0 signifie que $\neg x_i=0$ (et donc $x_i=1$) et réciproquement pour z=1 signifie que $\neg x_i=1$ (et donc $x_i=0$).

Dans le 1.1, les littéraux z correspondent à «carré » , «rond », «petit », «grand » etc. Si l'on code «carré » par x_1 alors «rond » est codé par $\neg x_1$ (non «rond »). Si z correspond à «carré » il correspond donc à x_1 . Si la figure est carrée alors z=1 ($x_1=1$ et donc $\neg x_1=0$) si la figure est ronde alors z=0 ($x_1=0$ et donc $\neg x_1=1$).

Nous travaillons dans X_n, les exemples a sont écrits avec n « 0 » ou « 1 ». Ainsi dans X_5, les exemples sont de la forme a=10011, ce qui veut dire que $x_1=1, x_2=0, \ldots x_5=1$ et qu'ils sont décrits par 5 valeurs d'attributs.

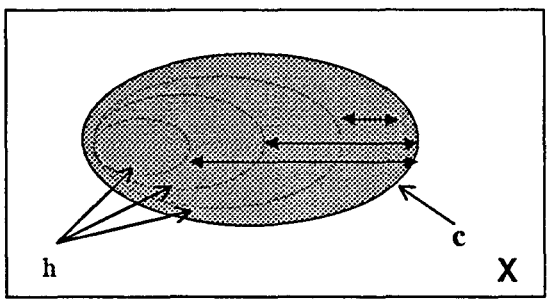
Dire qu'un exemple a vérifie une conjonction c ($c(a)=1$) signifie que pour tout i, z_i de la conjonction est égal à 1. Si on reprend la conjonction $c=x_1 \wedge \neg x_4 \wedge x_5$ et l'exemple a=10011 on voit que cet exemple ne vérifie pas la conjonction car pour que $z_4=1$ il faut que $x_4=0$ (car $z_4=\neg x_4$) et dans l'exemple $x_4=1$. Par contre a=10001 vérifie la conjonction puisque nous avons $x_1=1$, $x_4=0$ donc $\neg x_4=1$ et $x_5=1$. Les trois littéraux du concept sont vrais (=1) dans a.

Algorithme

L'algorithme correspond à l'algorithme présenté en 1.1. Rappelons brièvement son fonctionnement. L'apprenant débute avec h= C∧R∧P∧G∧B∧N∧I∧E (carré et rond et petit et grand et blanc et noir et flèche intérieure et flèche extérieure).
Avec le premier exemple positif CPBI, il élimine tous les littéraux de l'hypothèse incompatibles avec cet exemple. Ce qui donne : C∧~~R~~∧P∧~~G~~∧B∧~~N~~∧I∧~~E~~ soit C∧P∧B∧I.
Il recommence avec le second exemple positif, CPBE ce qui donne C∧P∧B∧~~I~~ et ce qui laisse comme hypothèse C∧P∧B alors que le concept cible était C∧P. Comme tous les exemples positifs différents sont épuisés nous gardons C∧P∧B comme hypothèse. Nous pouvons faire 2 remarques. La première est que l'hypothèse C∧P∧B est incluse dans le concept C∧P, elle est plus spécifique que le concept, tous les carrés petits et blancs sont des carrés petits. La

seconde est que l'hypothèse contient un littéral «erroné » B qui l'amène à refuser de classer comme positif les exemples CPNI et CPNE, mais ce littéral n'est pas «*mauvais*» car les erreurs qu'il recouvre ont un poids inférieur à la borne tolérée ε que l'on avait fixée à 5% : 0% pour CPNE et 2% pour CPNI, soit 2% d'erreur en tout (voir tableau 2).

Ainsi, le principe de l'algorithme est de commencer avec l'hypothèse la plus spécifique, celle qui contient le moins d'exemples possibles, donc l'hypothèse vide, et de la généraliser, petit à petit, au fur et à mesure de la présentation d'exemples positifs. L'hypothèse sera donc toujours plus spécifique que le concept, c'est-à-dire qu'elle sera toujours incluse dans le concept et ne fera donc pas d'erreur sur les exemples négatifs du concept. Elle ne classera pas comme positifs des exemples que le concept classera comme négatifs. Par contre étant plus spécifique que le concept, elle classera comme négatifs des exemples que le concept classera comme positifs (voir schéma 3 ci-dessous).



Schéma

X est l'ensemble des objets du monde. *c* est le concept à apprendre : l'ensemble des éléments de X qui vérifient le concept. En gris nous avons donc les exemples positifs du concept et en blanc les exemples négatifs du concept. Les ellipses en pointillés indiquent les généralisations successives de *h*. Les doubles-flèches indiquent les zones d'erreur successives de ces différentes hypothèses h. Nous voyons que les erreurs ne portent que sur des exemples positifs de c. Plus h se généralise, plus elle se rapproche de c et plus la zone d'erreur diminue.

Pour apprendre les conjonctions, *l'apprenant démarre avec une hypothèse qui contient tous les littéraux* :

dans le cas de X_5, ce sera $h = x_1 \wedge \neg x_1 \wedge x_2 \wedge \neg x_2 \wedge x_3 \wedge \neg x_3 \wedge x_4 \wedge \neg x_4 \wedge x_5 \wedge \neg x_5$

Cette hypothèse est l'ensemble vide car on ne peut avoir $x_i \wedge \neg x_i$ simultanément (on ne peut avoir une figure à la fois carrée et ronde). Elle ne comporte donc aucun exemple, elle est bien la plus spécifique.

L'apprenant *ne travaille qu'avec les positifs* et ignore tous les exemples négatifs.

A chaque exemple positif rencontré *il élimine les littéraux de l'hypothèse h qui ne sont pas vrais (=1) pour cet exemple.* Ainsi, on généralise petit à petit une hypothèse trop spécifique. Par exemple l'hypothèse «carré et petit et blanc et flèche intérieur » est plus spécifique que l'hypothèse «carré et petit et blanc», on a éliminé le littéral «flèche intérieure », on généralise

l'hypothèse car elle peut contenir plus d'éléments : les carrés petits blancs flèche intérieure et les carrés petits blancs et flèche extérieure.

11101 est un exemple positif
Après cet exemple h est égal à $x_1 \wedge x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$

Une hypothèse erronée est donc une hypothèse h telle que z est un littéral de h et n'est pas un littéral de c. Nous dirons que de tels littéraux z sont «erronés »,

h= $x_1 \wedge x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$ est une hypothèse «erronée » car elle contient les littéraux «erronés » x_2 et x_3 qui ne sont pas des littéraux de c=$x_1 \wedge \neg x_4 \wedge x_5$.

L'ensemble des éléments *a* de X_n qui font partie de $c \Delta h$, (l'ensemble *des contre-exemples* de h, ou l'ensemble des exemples qui constituent l'erreur de h) sont tels que : $c(a)=1$ (*a* est un exemple positif de c) et $z=0$ dans *a* (il y a un littéral de h qui est faux dans a et donc h refuse *a* alors que c l'accepte).

Ainsi h= $x_1 \wedge x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$ est encore trop spécifique car elle n'accepte pas 10001 alors que c le fera.
10001 est donc *un contre-exemple* de h, il fait partie de l'erreur ($c \Delta h$). En effet les littéraux x_2 et x_3 de h sont faux (=0) dans 10001

Démonstration

Ce qui précède décrit l'algorithme. Il faut maintenant *démontrer* que cet algorithme fournira *probablement* une hypothèse *approximativement correcte*. Dans ce qui suit, nous allons calculer le nombre d'exemples nécessaires en fonction de la distribution de probabilités, pour que l'algorithme retourne avec une probabilité de $1-\delta$ une hypothèse d'erreur de poids au plus ε.

Pour que l'apprentissage réussisse avec une probabilité d'au moins 1-δ, il faut que la probabilité qu'il puisse rester un z «mauvais » *dans h soit inférieure à* δ.

1-La probabilité qu'il reste un z «erroné » dans h est égale à la probabilité de ne pas tirer un contre-exemple *a* de h puisque si l'on tire un tel contre-exemple le z sera éliminé de h. Si on appelle p(z) cette probabilité qu'il reste un z «erroné » on a alors

$$p(z)=P_D[c(a)=1 \text{ et } z=0 \text{ dans a}]$$

p(z) est en quelque sorte le poids de l'erreur de z. C'est la probabilité de ne tirer aucun contre-exemple de h qui puisse éliminer le z «erroné ». C'est-à-dire la probabilité de ne pas tirer un exemple positif *a* de c (tel que $c(a)=1$) qui soit, en quelque sorte, un exemple négatif de z ($z=0$ dans a) et donc un exemple négatif de h. Une telle probabilité est égale à la somme des probabilités de tous les *a* tels que $c(a)=1$ et $z=0$ dans *a*.

Si nous reprenons notre hypothèse h= $x_1 \wedge x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$ et notre concept c=$x_1 \wedge \neg x_4 \wedge x_5$. Les littéraux x_2 et x_3 de h sont faux. Si l'on ne considère pour le moment que le littéral «erroné » x_2 , celui-ci «empêche » que les exemples 10001 et 10101 soient reconnus comme relevant du concept. Autrement dit, h les considère

comme faux (puisque x_2 =0 dans ces exemples) alors que c les considère comme vrais. Si on suppose que 10001 a une probabilité d'être tiré de 6% et 10101 une probabilité de 5%, la probabilité d'éliminer x_2 est de 11% : $p(x_2)$=11%. Cela correspond à la probabilité de tirer 10001 *ou* 10101.

Nous remarquons alors que si nous considérons tous les z «erronés» possibles de h nous avons :

$$\text{erreur}(h) \leq \Sigma_{z \text{ dans } h} p(z)$$

Dans notre exemple, $\text{erreur}(h) \leq p(x_1)+p(x_2)+p(x_3)+p(\neg x_4)+ p(x_5)$ puisque

$$\text{erreur}(h) \leq p(x_2)+p(x_3)$$

Comme h a au plus 2n littéraux, (2 fois le nombre de variables),

$$\text{erreur}(h) \leq 2n\, p(z)$$

Au départ h avait 10 littéraux ($h= x_1 \wedge \neg x_1 \wedge x_2 \wedge \neg x_2 \wedge x_3 \wedge \neg x_3 \wedge x_4 \wedge \neg x_4 \wedge x_5 \wedge \neg x_5$) Donc l'erreur était d'*au plus* la somme des poids de tous ces littéraux. Bien évidemment, tous les littéraux ne sont pas faux, («erronés»), mais en prenant en compte tous les littéraux nous obtenons une borne *maximale* de l'erreur.

2- Nous distinguons maintenant parmi les littéraux z «erronés» ceux qui sont «mauvais» et ceux qui ne le sont pas. Il faut comprendre que notre hypothèse peut très bien contenir des littéraux «erronés» si ceux-ci ne provoquent pas une erreur de plus de ε. Un «mauvais» littéral sera tel que

$$p(z) \geq = \varepsilon/2n$$

car ainsi nous sommes sûrs *que si notre hypothèse ne contient aucun «mauvais» littéral, son erreur est acceptable car inférieure à* ε

$$\text{erreur}(h) \leq 2n(\varepsilon/2n) = \varepsilon$$

Il faut donc éliminer tous les «mauvais» littéraux. La démonstration va maintenant consister à calculer combien il faut tirer d'exemples pour être sûr d'en tirer suffisamment pour éliminer tous ces «mauvais» littéraux z de h.

3-La probabilité, en 1 tirage, de *ne pas* éliminer un «mauvais» z de h est donc égale à la probabilité de *ne pas* tirer un *a* qui élimine ce z, et elle est d'*au plus* 1-ε/2n puisque p(z) est d'au moins ε/2n lorsque z est «mauvais».

4-La probabilité, en m tirages, de ne pas éliminer un «mauvais» z de h est donc au plus égale $(1-\varepsilon/2n)^m$

5-Mais comme il peut y avoir *au plus* 2n «mauvais» z à éliminer, la probabilité de ne pas *tous* les éliminer est *au plus* de $2n\,(1-\varepsilon/2n)^m$

6- Nous voulons que cette probabilité de ne pas éliminer tous les «mauvais» z soit inférieure ou égale à δ. Nous voulons donc que :

$$2n\,(1-\varepsilon/2n)^m \leq \delta$$

7-à partir de cette inégalité nous pouvons donc borner m :

on sait que $1-x \leq e^{-x}$
donc $2n\, e^{-\varepsilon m/2n} \leq \delta$
donc $e^{-\varepsilon m/2n} \leq \delta/2n$
on prend le log népérien : $-\varepsilon m/2n \leq \ln(\delta/2n)$
donc $-m \leq (2n/\varepsilon) \ln(\delta/2n)$
donc $m \geq (2n/\varepsilon) \ln(2n/\delta)$
donc $m \geq (2n/\varepsilon) [\ln(2n)+\ln(1/\delta)]$

m, le nombre de tirages correspond au nombre d'exemples présentés à l'apprenant. Il faut donc présenter m exemples à l'apprenant pour être sûr avec une probabilité de $1-\delta$ que l'erreur sera au plus de ε.

8- L'algorithme traite chaque exemple en un temps linéaire et comme le nombre d'exemples est un polynôme en n, $1/\varepsilon$ et $1/\delta$ *l'apprentissage se fera en temps polynomial.*

Annexe 3

Un exemple de dysfonctionnement de la procédure TEST

Nous rappelons que la procédure TEST soumet à l'apprenant un certain nombre d'exemple et lui demande de les étiqueter. Si l'étiquetage est inférieur à un certain pourcentage d'erreurs, l'hypothèse est acceptée. Le problème est que cette procédure TEST est sujette à deux types d'erreurs possibles :
- elle peut rejeter une hypothèse approximativement correcte,
- elle peut accepter une hypothèse erronée (telle que l'erreur est supérieure à la borne d'erreur tolérée)

Pour illustrer ceci, nous prenons un exemple très similaire à celui proposé en 1.1 (voir tableau A.2.1 ci-dessous). Nous gardons comme concept cible C∧P. Nous allons supposer que la procédure TEST propose 5 exemples tirés selon la distribution de probabilité que l'on retrouve dans le tableau. TEST ne demande pas explicitement son hypothèse au sujet mais lui fait étiqueter 5 exemples. Elle considère son hypothèse comme «mauvaise» s'il fait plus d'une erreur, ce qui revient à tolérer une erreur de 20%.

figure	distribution de probabilités	nombre d'occurrences dans l'échantillon de test 1	nombre d'occurrences dans l'échantillon de test 2
RGNE	2%	0/5	0/5
RGNI	7%	0/5	0/5
RGBE	0%	0/5	0/5
RGBI	7%	1/5	0/5
RPNE	1%	0/5	0/5
RPNI	1%	0/5	0/5
RPBE	7%	0/5	0/5
RPBI	7%	0/5	0/5
CGNE	7%	1/5	0/5
CGNI	1%	0/5	0/5
CGBE	0%	0/5	0/5
CGBI	7%	0/5	1/5
CPNE	0%	0/5	0/5
CPNI	7%	0/5	1/5
CPBE	15%	1/5	1/5
CPBI	31%	2/5	2/5

Tableau A.2.1 : La première colonne indique l'exemple par ses initiales (RGNE : rond grand noir flèche extérieure). La seconde colonne donne sa probabilité d'être tiré. La troisième et la quatrième colonne indiquent 2 tirages possibles de l'échantillon de test.

Pour tirer les exemples TEST obéit à la loi de distribution indiquée en colonne 2 du tableau. Dans le premier tirage les exemples RGBI, CGNE, CPBE, CPBI, apparaissent (voir colonne 3), le dernier exemple étant présenté 2 fois puisqu'il a une probabilité de 31%. Ce tirage est aléatoire selon la distribution de probabilité, cependant tous les exemples qui ont 7% de probabilité n'ont pu être tirés puisqu'il n'y a que 5 tirages. Dans le second tirage, ce sont les

exemples CGBI, CPNI, CPBE, CPBI, qui apparaissent (voir colonne 4), le dernier exemple étant toujours présenté 2 fois. Comme pour le premier tirage, ce second tirage est aléatoire selon la distribution de probabilités avec le même problème pour les exemples qui ont 7% de probabilité.

1er cas : hypothèse P∧B et tirage test 1

Nous allons supposer que le sujet formule l'hypothèse P∧B et que la procédure TEST opère le tirage n°1. En étiquetant ces exemples tests selon cette hypothèse, le sujet donne le même étiquetage que celui donné par le concept cible C∧P. Les exemples RGBI et CGNE seront étiquetés négatifs et CPBE et CPBI seront étiquetés positifs. La procédure TEST considérera que l'hypothèse du sujet est approximativement correcte puisqu'il n'y a pas de divergence dans l'étiquetage des exemples qu'elle a proposés. Le problème est que cette hypothèse à une erreur de 21%. Les exemples sur lesquels P∧B et C∧P ne donnent pas le même étiquetage sont RPBE, RPBI, CPNE, CPNI ce qui fait une erreur de 7 +7 +0 +7 soit 21% : La procédure TEST a donc accepté une hypothèse avec une erreur trop grande.

2ème cas : hypothèse C∧B et tirage 2

Nous allons supposer, maintenant que le sujet formule l'hypothèse C∧B et que la procédure TEST opère le tirage n°2. En étiquetant ces exemples tests selon cette hypothèse, le sujet donne un étiquetage différent sur 2 exemples de celui donné par le concept cible C∧P. Alors que les exemples CPBE et CPBI seront étiquetés positifs par les deux, l'exemple CGBI sera étiqueté positif par l'hypothèse et négatif par le concept, tandis que CPNI sera étiqueté négatif par l'hypothèse et positif par le concept. La procédure TEST considérera que l'hypothèse du sujet est erronée puisqu'il y a plus d'une erreur d'étiquetage. Le problème est que cette hypothèse n'a que 14% d'erreur, elle est donc approximativement correcte puisque l'erreur tolérée est de 20%. Les exemples sur lesquels C∧B et C∧P ne donnent pas le même étiquetage sont CGBE, CGBI, CPNE, CPNI ce qui fait une erreur de 0 +7 +0 +7 soit 14% : La procédure TEST a donc rejeté une hypothèse avec une erreur acceptable.

Plus généralement

De manière un peu plus générale, on constate que TEST peut faire deux types d'erreurs. Supposons que TEST tire un nombre x d'exemples (par exemple 100) selon la distribution D puis soumet ces x exemples à l'apprenant pour qu'il les étiquette. Elle compare les étiquettes et pour que l'apprentissage soit réussi il faut que le pourcentage d'erreurs soit inférieur à ε. Il y a alors 2 types d'erreurs possibles :

1) TEST retourne faux car le pourcentage d'exemples tirés de cΔh est supérieur au poids de cΔh alors que le poids de cΔh est inférieur à ε, TEST rejette ainsi une hypothèse approximativement correcte.

> Exemple : supposons que l'hypothèse h a une erreur avec le concept c de 4% ($P_D(c\Delta h)=0,04$) et que la borne d'erreur tolérée est de 5% ($\varepsilon=0,05$). Il est possible que TEST tire 6 exemples de cΔh sur 100[115]. TEST constatera pour l'hypothèse h un taux

[115] Nous retrouvons toujours le même problème d'écart en distribution de probabilité et tirage réel. Mais, ici, il n'est pas possible à l'expérimentateur de corriger cet écart car il ne connaît pas l'hypothèse de l'apprenant, par conséquent il ne peut évaluer $P_D(c\Delta h)$

d'erreur de 6% et la rejettera puisque la borne est de 5%, alors que l'erreur de h est en fait de 4%. h est approximativement correcte mais TEST l'a rejetée.

2>TEST retourne vrai car le pourcentage d'exemples tirés de $c\Delta h$ est inférieur au poids de $c\Delta h$, alors que le poids de $c\Delta h$ est supérieur à ε, TEST accepte ainsi une hypothèse qui n'est pas approximativement correcte.

Exemple : supposons que l'hypothèse h a une erreur avec le concept c de 7% ($P_D(c\Delta h)=0,07$) et que la borne d'erreur tolérée est de 5% ($\varepsilon=0,05$). Il est possible que TEST ne tire que 4 exemples de $c\Delta h$ sur 100. TEST constatera pour l'hypothèse h un taux d'erreur de 4% et l'acceptera puisque la borne est de 5%, alors que l'erreur de h est en fait de 7%. h n'est pas approximativement correcte mais TEST l'a acceptée.

Annexe 4

Utilisation de la loi binomiale

Supposons que nous tirions à pile ou face avec une pièce, mais que cette fois-ci la pièce soit biaisée et qu'elle donne une probabilité de 30% à pile et de 70% à face. Sur 40 tirages nous devrions avoir théoriquement 12 piles et 28 faces. En réalité, nous savons qu'il est possible d'avoir 5 piles et 35 faces ou 19 piles et 21 faces. Cependant, intuitivement, nous admettons qu'il est plus probable d'avoir 12 piles sur 40 que 5 ou 19. Plus nous nous éloignons du nombre de 12 piles moins la probabilité d'un tel événement est grande. C'est ce que traduit la loi binomiale.

Si nous appelons r le nombre de piles, n le nombre de tirages indépendants et p la probabilité de l'événement pile, la loi binomiale nous indique que :

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

Le P(r) correspond à la probabilité qu'en n tirages nous puissions avoir r piles alors que la probabilité de «pile » est de p.

Le schéma ci-dessous décrit la loi Binomiale pour p=0.3 et n=40.

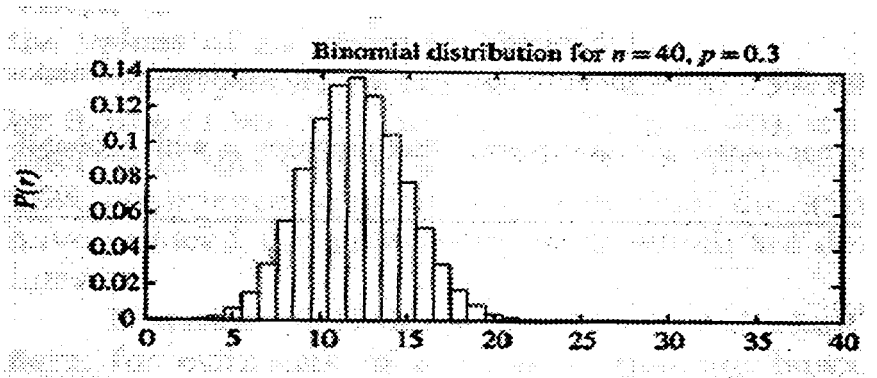


Figure 10 : la distribution binomiale pour n=40 et p=0,3 [Mitchell, 97]

Si nous prenons r=5, n=40 et p=0,3

$$P(r) = [\ 40! / (5!\ 35!)]\ 0{,}3^5\ 0{,}7^{35}$$

$$P(r) = [\ (40 \times 39 \times 38 \times 37 \times 36\ ) / (5 \times 4 \times 3 \times 2 \times 1)]\ 0{,}3^5\ 0{,}7^{35}$$

$$P(r) = 658008 \times 0{,}00243 \times 0{,}000003788$$

$$P(r) = 0{,}006$$

Autrement dit, il y a 6 chances sur 1000, lorsque la probabilité d'avoir pile est de 30%, que nous puissions avoir 5 piles sur 40. De la même manière si nous nous reportons au schéma ci-dessus, il y a 14 chances sur 100 environ d'avoir 12 piles sur 40.

Si nous convertissons ceci en termes d'erreur et supposons que l'erreur de l'apprenant est de 30% ($P_D(c\Delta h)=0.3$), il est possible avec une probabilité de 6 chances sur 1000 que TEST ne tire que 5 exemples sur 40 de la zone d'erreur, ce qui, ramené en pourcentages, donne 12,5% d'erreur apparente. La distorsion entre l'erreur réelle 30% et l'erreur apparente 12,5% est donc de 17,5% mais une telle distorsion n'a que 6 chances sur 1000 d'apparaître. Il est ainsi possible de calculer avec une certaine probabilité un intervalle de confiance. Ainsi dans le cas de p=0,3 et n=40 nous sommes sûrs à 95% que l'erreur apparente ne déviera de l'erreur réelle que d'au plus 14%. Nous sommes ainsi sûrs à 95% que l'erreur observée sera comprise entre 16% et 44%. Il est possible à partir du schéma de Mitchell de voir comment on peut approximativement obtenir de tels chiffres. Une distorsion de 14% correspond à + ou - 5 exemples, c'est-à-dire à l'intervalle entre 7 et 17 exemples. Si nous faisons la somme des probabilités sur cet intervalle nous obtenons environ 95%.

Ce qui est valable dans un sens est valable dans l'autre : il est possible à partir de l'erreur observée d'obtenir un intervalle de confiance dans lequel se situe l'erreur réelle avec une certaine probabilité. Ainsi si nous observons une erreur de 12 exemples sur 40 (30%), nous pouvons être sûrs à 95% que l'erreur réelle est dans l'intervalle 30% + ou - 14%. De plus, ce qui nous intéresse ici n'est pas l'intervalle mais seulement l'erreur maximale. Dans ce cas nous avons alors la garantie à 97,5% que l'erreur réelle est d'au plus de 44% (30% + 14%).

La loi binomiale est quelque peu lourde à calculer, c'est pourquoi, elle est approximée par la loi Normale. Mais cette approximation n'est valable qu'au-dessus de 30 tirages. Avec le développement de la puissance de calcul des machines, il est possible, maintenant, de la calculer exactement, ce qui permet d'obtenir des chiffres en dessous de 30 tirages. Les formules sont les suivantes

$$p_{sup}(k)= \sup \{p \mid Pr_p(X\leq k)> \alpha/2\}$$

$$p_{inf}(k)= \inf \{p \mid Pr_p(X>k)> \alpha/2\}$$

$$\text{avec } Pr_p(X\leq k)=\Sigma_{i=0 \text{ à } k}\, C^i_n\, p^i\, (1-p)^{n-i}$$

$$\text{et } Pr_p(X>k)=1- Pr_p(X\leq k)$$

Où k est le nombre d'erreurs constaté, $p_{sup}(k)$ et $p_{inf}(k)$ les bornes supérieure et inférieure de l'intervalle de confiance, p les probabilités à tester et α la confiance.

Supposons que n=25 et que k=1 et α=10% ($\alpha/2$=0,05) et que nous voulions connaître la borne supérieure, nous obtenons en prenant :

$$p=0,17 ; Pr_p(X\leq k)=0,058>0,05$$

$$p=0,18 ; Pr_p(X\leq k)=0,045 <0,05$$

Il faut donc prendre 18% comme borne supérieure de l'erreur.

Ainsi si nous faisons un calcul exact avec la loi binomiale de l'intervalle de confiance, nous obtenons les chiffres suivants avec **une confiance de 90%** :
si l'on teste le sujet sur 20 exemples :

0 erreur observée correspond à une erreur réelle comprise entre 0 et 14%
1 erreur observée correspond à une erreur réelle comprise entre 0 et 22%
2 erreurs observées correspondent à une erreur réelle comprise entre 1 et 29%

si l'on teste le sujet sur 25 exemples :

0 erreur observée correspond à une erreur réelle comprise entre 0 et 12%
1 erreur observée correspond à une erreur réelle comprise entre 0 et 18%
2 erreurs observées correspondent à une erreur réelle comprise entre 1 et 24%

si l'on teste le sujet sur 30 exemples :

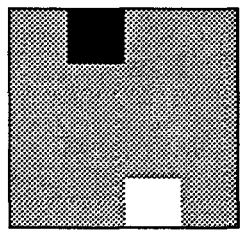
0 erreur observée correspond à une erreur réelle comprise entre 0 et 10%
1 erreur observée correspond à une erreur réelle comprise entre 0 et 15%
2 erreurs observées correspondent à une erreur réelle comprise entre 1 et 20%

Un bon compromis semble donc de n'accepter qu'une erreur sur 25, ce qui nous garantit à 90% que l'erreur réelle sera au plus de 18%. Pour obtenir la garantie d'une erreur réelle plus faible il serait nécessaire de tester le sujet sur beaucoup plus d'exemples, cela entraînerait alors des problèmes de fatigue qui biaiseraient les résultats.

L'application de la loi binomiale permet ainsi d'éviter un des deux types d'erreurs que peut commettre TEST. Elle donne avec une garantie assez grande, 90%, que TEST n'acceptera pas une hypothèse dont l'erreur est supérieure à 18%. Elle ne garantit pas par contre qu'une hypothèse approximativement correcte ne sera pas rejetée.

Annexe 5 : présentation du test aux sujets

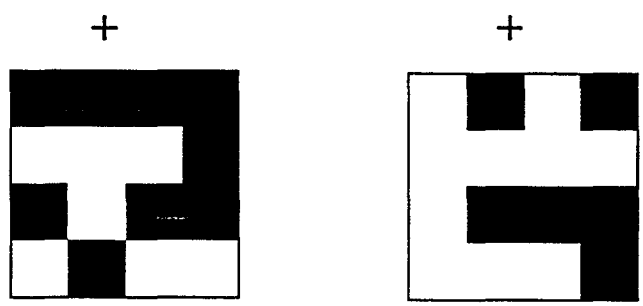
Un concept alpha est défini par les valeurs de 2 cases particulières. Ci-dessous ce sont les cases 2 et 15. alpha



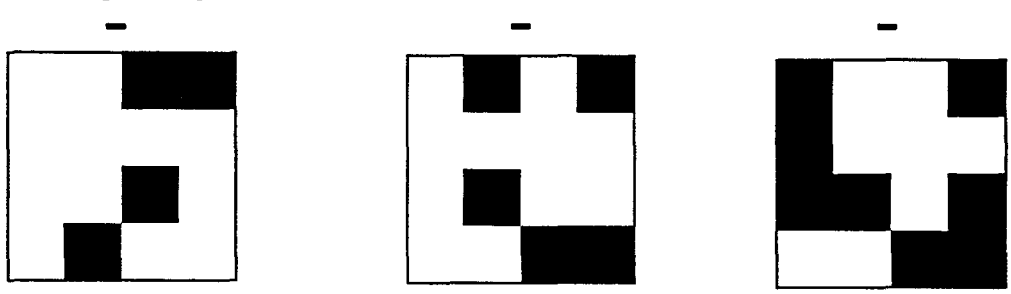
Un exemple de concept alpha

Dans ce cas-ci, tous les exemples positifs du concept doivent avoir la 2ème case noire et la 15ème blanche.

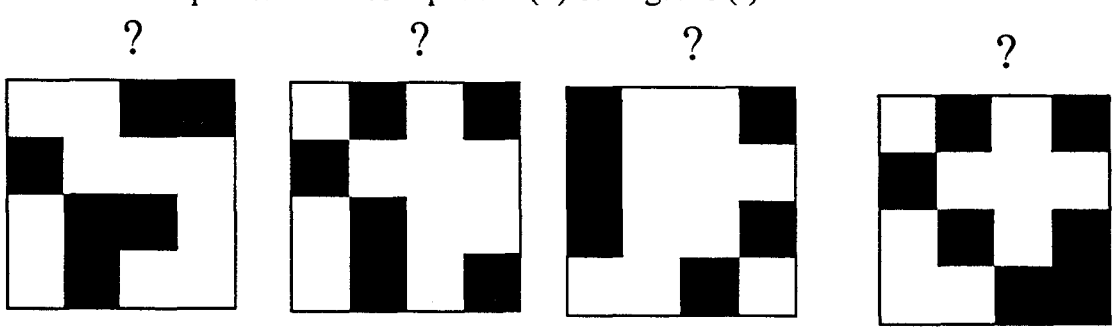
Exemples d'exemples positifs



Exemples d'exemples négatifs



Indiquez si les exemples suivants sont positifs (+) ou négatifs (-)



Annexe 6 : résultats des 39 sujets à l'expérience n°1

Le tableau indique la distribution, l'âge et le sexe des sujets, s'ils ont réussi ou non l'apprentissage, dans le cas de réussite le nombre d'exemples qui leur ont été nécessaires pour apprendre, le nombre de fois où ils ont passé le test avant de réussir, le nombre d'exemples qu'ils ont eu à affronter lors de tous ces tests et le nombre d'exemples qu'ils ont mal étiquetés.

Distri	Age	Sexe	Résul.	nb ex	nb appel à TEST	nb exemples. de test	mal étiquetés
F	18	F	O	28	3	41	2
F	21	M	O	53	1	25	0
F	22	F	O	89	1	25	0
F	21	F	O	97	1	25	1
F	22	F	O	125	1	25	0
F	18	F	O	132	2	28	1
F	24	M	O	164	1	25	0
F	17	F	O	214	2	50	4
F	18	F	O	222	1	25	0
F	20	M	O	261	2	50	2
F	21	F	O	387	1	25	1
F	18	F	O	445	6	133	56
F	19	F	O	555	3	51	5
U	18	F	N		6	150	59
U	19	F	N		8	67	38
U	18	F	N		4	77	38
U	21	M	O	122	1	25	0
U	21	F	O	188	1	25	0
U	44	F	O	221	1	25	0
U	19	F	O	229	1	25	0
U	18	F	O	332	1	25	0
U	19	F	O	405	1	25	0
U	19	F	O	418	1	25	0
U	20	F	O	424	1	25	0
U	22	M	O	438	2	50	6
U	18	F	O	461	3	75	19
D	23	M	N		11	236	101
D	19	F	N		2	50	15
D	18	F	N		6	126	52
D	17	F	N		11	251	75
D	20	F	N		5	125	44
D	21	F	O	132	3	75	14
D	22	F	O	155	1	25	0
D	22	F	O	198	2	50	2
D	22	M	O	292	5	63	32
D	21	F	O	438	1	25	0
D	18	F	O	462	1	25	0
D	18	F	O	605	3	75	19
D	18	F	O	641	3	51	8

Annexe 7

1>Etude de l'erreur théorique possible cΔh selon les différentes distributions

Dans cette annexe nous étudions les erreurs des différentes hypothèses possibles selon les différentes distributions de probabilité. Nous n'étudions que les hypothèses qui sont des conjonctions (monômes) de deux littéraux.

Nous rappelons le nombre d'éléments possibles (c'est-à-dire qui ont une possibilité d'apparaître dans l'échantillon).

nombre d'exemples possibles selon les distributions	facilitante	Uniforme/ neutre	handicapante
positifs	91	16384	91
négatifs	91	49152	273
total	182	65536	364

Exemple du calcul pour la distribution facilitante : pour les positifs, 2 variables servent au concept, il n'y a donc que 14 variables qui peuvent changer de valeur. Pour construire un exemple on choisit aléatoirement 2 de ces variables et on inverse leur valeur. Le nombre de combinaisons possibles est alors de C^2_{14}= (14 x 13) /2=91. Pour les négatifs de la forme nonx et non y même chose soit 91.
Le principe est le même pour la distribution handicapante, et pour la distribution uniforme tous les exemples ont une possibilité égale d'apparaître.

Dans ce qui suit, le concept cible est xy et l'hypothèse est x'y'. Lorsque nous écrivons x=x' cela signifie que nous supposons que le concept et l'hypothèse ont un litttéral en commun. Lorsque nous écrivons que x' appartient au prototype (ou l'exemple central) cela signife que l'hypothèse et la définition du prototype ont un litttéral en commun

Si on appelle xy le concept et x'y' l'hypothèse :
-l'erreur sur les positifs est l'ensemble des éléments qui sont étiqutés positifs par le concept et étiquetés négatifs par l'hypothèse, autrement dit les éléments de la forme xy¬y' ou xy¬x'. L'ensemble est alors (xy¬y' ∪ xy¬x')= (xy¬x' ∪ xyx'¬y'). Nous rappelons que l'on tire 25% de positifs.
-l'erreur sur les négatifs est l'ensemble des éléments qui sont étiquetés positif par l'hypothèse (donc de la forme x'y' et étiquetés négatifs par le concept donc de la forme ¬x¬y ou x¬y ou ¬xy soit (x'y'¬x¬y) ∪(x'y'x¬y) ∪(x'y'¬xy). Nous rappelons que l'on tire 75% de négatifs.

Distribution facile

1er cas : x=x' y≠y' et y' appartient au prototype positif

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y'). Comme x'=x (xy¬x' ∪ xyx'¬y')= xyx'¬y'= xy¬y'. Comme y' appartient au prototype, ¬y' n'apparaît que 2 fois sur 14 dans l'échantillon positif. Comme il y a 91 positifs cela fait 13 erreurs sur les positifs (91*2/14).

L'erreur sur les négatifs est égale à (x'y'¬x¬y) ∪(x'y'x¬y) ∪(x'y'¬xy). Comme les négatifs sont construits autour d'un prototype négatif qui est l'inverse du positif il n'y a pas d'exemples de la forme : x¬y ou ¬xy. L'ensemble d'erreur est donc (x'y'¬x¬y) et comme x=x' cet ensemble est vide.

Comme il n'y a que 25% de positifs, l'erreur totale est donc de (13/91)x25% soit 3,57%.

2ème cas : x≠x' et y≠y' et x' et y' appartiennent au prototype positif

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y')
Le nombre d'éléments de xy¬x' est de 13 éléments (2 éléments sur 14 parmi les 91 positifs)
Le nombre d'éléments de xyx'¬y' est de 12 éléments (2 éléments sur 13 parmi les 78 positifs de la forme xyx')
Le nombre d'éléments correspondant à l'erreur sur les positifs est donc de 25.

L'erreur sur les négatifs est de la forme (x'y'¬x¬y) ∪(x'y'x¬y) ∪(x'y'¬xy). Comme les négatifs sont construits autour d'un prototype négatif qui est l'inverse du positif il n'y a pas d'exemples de la forme : x¬y ou ¬xy. L'ensemble d'erreur est donc (x'y'¬x¬y).

Le nombre d'éléments de (x'y'¬x¬y) est de 1 élément (2 éléments sur 182 parmi les 91 négatifs de la forme ¬x¬y) car le prototype négatif est de la forme ¬x'¬y'¬x¬y. On a donc 2 possibilités sur 14 d'avoir x' et 1 possibilité sur 13 d'avoir ensuite y' soit 2 possibilité sur 182 (soit 1/91) d'avoir x'y'.

L'erreur totale est de (25/91)x25% +(1/91)x75% soit 6,86%+0,82% soit 7,68%. On note que même avec une telle erreur théorique et bien que l'on n'autorise qu'au plus 1 erreur sur 25 lors du test, cette hypothèse pourra éventuellement le passer avec succès puisqu'il est possible que le test accepte une hypothèse avec 18% d'erreur mais avec une probabilité très faible.

3ème cas : x≠x' et y≠y' et x' appartient et y' n'appartient pas au prototype positif

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y')
Le nombre d'éléments de xy¬x' est de 13 éléments (2 éléments sur 14 parmi les 91 positifs)
Le nombre d'éléments de xyx'¬y' est de 66 éléments (11 éléments sur 13 parmi les 78 positifs de la forme xyx'). En effet y' n'appartient pas au prototype positif donc ¬y' appartient au prototype, il a 11 chances sur 13 d'apparaître.
Le nombre d'éléments correspondant à l'erreur sur les positifs est donc de 79 sur 91.

L'erreur sur les négatifs est égale à (x'y'¬x¬y). Comme y' n'appartient pas au prototype positif il appartient au prototype négatif
Le nombre d'éléments de (x'y'¬x¬y) est de 12 éléments (24 éléments sur 182 parmi les 91 négatifs de la forme ¬x¬y). On a donc 2 possibilités sur 14 d'avoir x' et 12 chances sur 13 d'avoir ensuite y' soit 24 possibilités sur 182 d'avoir x'y'.
L'erreur sur les négatifs est donc de 12

L'erreur totale est de (79/91)x25% +(12/91)x75% soit 21,70%+9,89% soit 31,59%

4ème cas : x≠x' et y≠y' et x' et y' n'appartiennent pas au prototype positif

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y')
Le nombre d'éléments de xy¬x' est de 78 éléments (12 éléments sur 14 parmi les 91 positifs). En effet x' n'appartient pas au prototype donc ¬x' appartient au prototype et a 12 chances sur 14 d'apparaître. Le nombre d'éléments de xyx'¬y' est de 12 éléments (12 éléments sur 13 parmi les 13 positifs de la forme xyx'). En effet y' n'appartient pas au prototype donc ¬y' appartient et a 12 chances sur 13 d'apparaître.
Le nombre d'éléments correspondant à l'erreur sur les positifs est donc de 90 sur 91.

L'erreur sur les négatifs est égale à (x'y'¬x¬y)
Le nombre d'éléments de (x'y'¬x¬y) est de 66 éléments (132 éléments sur 182 parmi les 91 négatifs de la forme ¬x¬y) car le prototype négatif, ici, est de la forme x'y'¬x¬y. Comme x' et y' n'appartiennent pas au prototype positif , ils appartiennent aux prototypes négatifs. On a donc 12 possibilités sur 14 d'avoir x' et 11 possibilités sur 13 d'avoir ensuite y' soit 132 possibilités sur 182 d'avoir x'y'.

L'erreur totale est de (90/91)x25% +(66/91)x75% soit 24,72%+54,39% soit 79,11%

5ème cas : x=x' et y≠y' et y' n'appartient pas au prototype positif

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y'). Comme x'=x, xy¬x' ∪ xyx'¬y'= xyx'¬y'= xy¬y'. Comme y' n'appartient au prototype, ¬y' lui appartient et apparaît 12 fois sur 14 dans l'échantillon positif. Comme il y a 91 positifs cela fait 78 erreurs sur les positifs.

L'erreur sur les négatifs est égale à (x'y'¬x¬y). Comme x=x' cet ensemble est vide.

L'erreur totale est de (78/91)x25% 0x75% soit 21,42%

Conclusion

En testant sur 25 exemples et en ne tolérant qu'une erreur au plus sur 25 (soit 4%), nous obtenons 29 hypothèses sur 480 qui sont légitimement correctes : ce sont toutes les conjonctions qui comportent 1 des deux littéraux de la cible, l'autre littéral étant un littéral du prototype positif..

Cependant ce type de test (avec une tolérance de 1) n'offre une garantie de rejeter que les hypothèses qui ont au moins 18% d'erreur. Il existe ainsi tout un ensemble d'hypothèses qui pourront éventuellement (avec une certaine probabilité) être acceptées par le test, il s'agit de celles constituées par deux éléments du prototype positif qui ont 7,7% d'erreur. Il faut noter qu'il est aisé de trouver une telle hypothèse. Il suffit de sélectionner deux littéraux qui ne varient pas sur 3 exemples positifs. La probabilité qu'un seul de ces deux littéraux n'appartienne pas au prototype positif est de $(1/7)^3$ soit 1 possibilité sur 343. Cependant pour qu'un sujet adopte cette stratégie il faudrait qu'il connaisse la distribution de probabilité.

Cas de la distribution neutre (uniforme)

cas x'=x et y'≠y

L'erreur sur les positifs est égale à $(xy\neg x' \cup xy\neg y')$ ce qui est équivalent à $(xy\neg x' \cup xyx'\neg y')$. Comme x'=x, $xy\neg x' \cup xyx'\neg y' = xyx'\neg y' = xy\neg y'$. Comme il y a 2^{14} éléments de type xy et parmi ceux-ci 1 sur 2 est de type $\neg y'$, cela fait une erreur de 2^{13} sur les positifs
Rappelons qu'avec la distribution uniforme, il y a 3 types de négatifs $x\neg y, \neg xy, \neg x\neg y$. Comme x'=x, l'erreur sur les négatifs ne concerne que les négatifs de type $x\neg y$ soit 2^{14} éléments, l'erreur sur les négatifs est alors $x\neg yy'$ soit aussi une erreur de 2^{13}
L'erreur totale est donc de 2^{14} sur 2^{16} soit 25%.
Autre façon de calculer : $(2^{13}/2^{14})x25\% + (2^{13}/(2^{16}-2^{14}))x75\% = (1/2)x25\% + (1/2(4-1))x75\%$
$= 12,5\% + 1/6x75\% = 12,5\% + 12,5\% = 25\%$

cas x'≠x et y'≠y

L'erreur sur les positifs est égale à $(xy\neg x' \cup xyx'\neg y')$
L'erreur $xy\neg x'$ représente 2^{13} exemples et $xyx'\neg y'$ 2^{12} exemples. L'erreur sur les positifs représente en tout $2^{13} + 2^{12}$ exemples
L'erreur sur les négatifs est de la forme $(x'y'\neg x\neg y) \cup (x'y'x\neg y) \cup (x'y'\neg xy)$ soit 3 fois 2^{12} exemples
L'erreur totale est $2^{13} + 4 \times 2^{12}$ exemples soit $2^{14} + 2^{13}$ exemples sur 2^{16} exemples soit 37,5%.

Conclusion

Avec la distribution uniforme, la seule hypothèse de type monôme de deux littéraux d'erreur inférieure à 18% est l'hypothèse exacte c'est-à-dire le concept.

Cas de la distribution difficile

cas x=x' y'≠y et y' appartient à l'exemple central

L'erreur sur les positifs est égale à $(xy\neg x' \cup xyx'\neg y')$. Comme x'=x, $xy\neg x' \cup xyx'\neg y' = xyx'\neg y' = xy\neg y'$. Parmi les positifs on ne voit apparaître $\neg y'$, 2 fois sur 14 ce qui fait 13 éléments car il y a 91 positifs.

L'erreur sur les négatifs est égale à $(x'y'\neg x\neg y) \cup (x'y'x\neg y) \cup (x'y'\neg xy)$ comme x=x' l'erreur est égale à $x'y'x\neg y$ et donc à $x\neg y\, y'$. Comme y' appartient à l'exemple central il apparaît 12 fois sur 14. Comme il y a 91 exemples de la forme $x\neg y$ cela fait 78 exemples pour l'erreur sur les négatifs.

L'erreur totale est de (13/ 91)x25% + (78/273)x75% soit 3,57% + 21,43% soit 25%

cas x≠x' y'≠y et x' et y' appartiennent à l'exemple central

L'erreur sur les positifs est égale à $(xy\neg x' \cup xyx'\neg y')$

L'erreur xy¬x' est 13 (91 x 2/14), l'erreur xyx'¬y' est de 12 exemples car il y a 78 exemples de la forme xyx' et seulement 2 sur 13 de ces exemples est de la forme xyx'¬y'. L'erreur sur les positifs est donc de 25 exemples sur 91.

L'erreur sur les négatifs est de la forme (x'y'¬x¬y) ∪(x'y'x¬y) ∪(x'y'¬xy)
L'erreur (x'y'¬x¬y) est de 66 car il y a 91 exemples de la forme ¬x¬y, 12/14 sont de la forme x'¬x¬y soit 78 et parmi ceux-ci 11/13 de la forme x'y'¬x¬y soit 66. Le calcul étant le même pour x'y'x¬y et x'y'¬xy on obtient une erreur de 66 x3 soit 198 exemples sur les négatifs.

L'erreur totale est donc de 223 exemples sur 364 soit 61,3%.

cas x'≠x y'≠y et x' et y' n'appartiennent pas à l'exemple central

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y')
L'erreur xy¬x' est de 78 car il y a 91 exemples de la forme xy et 12/14 de ces exemples sont de la forme xy¬x'.L'erreur xyx'¬y' est de 12 exemples car il y a 13 exemples de la forme xyx' et 12 sur 13 de ces exemples sont de la forme xyx'¬y'. L'erreur sur les positifs est donc de 90.

L'erreur sur les négatifs est de la forme (x'y'¬x¬y) ∪(x'y'x¬y) ∪(x'y'¬xy)
L'erreur (x'y'¬x¬y) est de 1 car il y a 91 exemples de la forme ¬x¬y, 2/14 sont de la forme x'¬x¬y soit 13 et parmi ceux-ci 1/13 de la forme x'y'¬x¬y . Le calcul étant le même pour x'y'x¬y et x'y'¬xy on obtient une erreur de 3 exemples sur les négatifs.

L'erreur totale est donc de 93 exemples sur 364 soit 25,5%.

cas x'=x y'≠y et y' n'appartient pas à l'exemple central

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y')
L'erreur xy¬x'= xy¬x= 0. L'erreur xyx'¬y' est égale à xy¬y' puisque x'=x. Elle est de 78 exemples car il y a 91 exemples de la forme xy et 12 sur 14 de ces exemples sont de la forme xy¬y'. L'erreur sur les positifs est donc de 78.

L'erreur sur les négatifs est de la forme (x'y'¬x¬y) ∪(x'y'x¬y) ∪(x'y'¬xy). Puisque x=x' l'erreur se réduit à x'y'x¬y soit x¬y y'. L'erreur ici est de 13 exemples car il y a 91 exemples de la forme x¬y et 2 sur 14 de ces exemples est de la forme x¬yy'. L'erreur sur les négatifs est donc de 13.

L'erreur totale est donc de 91 exemples sur 364 soit 25%.

cas x'≠x y'≠y et x' appartient et y' n'appartient pas à l'exemple central

L'erreur sur les positifs est égale à (xy¬x' ∪ xyx'¬y')
L'erreur xy¬x'=13 éléments (2/14x91). L'erreur xyx'¬y'= 66 éléments (11/13x78). L'erreur totale est donc de 79 sur 91.

L'erreur sur les négatifs est de la forme (x'y'¬x¬y) ∪(x'y'x¬y) ∪(x'y'¬xy).
L'erreur x'y'¬x¬y= (12/14 x 1/13) x 91=6 éléments, De la même façon pour (x'y'x¬y) et (x'y'¬xy) on botient aussi 6 éléments soit 18 éléments en tout.

L'erreur totale est donc de 97 exemples sur 364 soit 26,6%.

Conclusion

Comme pour la distribution neutre (uniforme), la seule hypothèse de type monôme de deux littéraux d'erreur inférieure à 21 % est l'hypothèse exacte c'est-à-dire le concept.

2> Calcul du nombre d'hypothèses approximativement correcte selon les différentes distributions

Nous avons vu dans la section précédente que pour la distribution facile, il existe plusieurs hypothèses qui peuvent passer le test avec succès, celles de la forme x'y' avec x'=x et y'≠y mais y' appartenant à la définition du prototype. Il y a ainsi 29 hypothèses qui ont une erreur de moins de 4% (14 de la forme x=x' et y≠y' mais y' appartient au prototype, 14 de la forme y=y' et x≠x' mais x' appartient au prototype et enfin l'hypothèse identique au concept x=x' et y=y') . Un autre type d'hypothèses qui peuvent passer le test avec succès mais avec une probabilité plus faible car elles ont une erreur de 7,7% sont de la forme x'≠x et y'≠y mais x' et y' appartenant à la définition du prototype positif, il y en a 91. Cela fait alors 120 hypothèses susceptible de passer le test donc 1 sur 4.

Annexe 8 : 3 expériences

Nous présentons 3 expériences pour justifier de l'utilité de l'opérationnalisation du modèle PAC en psychologie. L'opposition entre celle de [Medin, Wattenmaker et Hampson, 87] et [Kemler-Nelson, 84] permet de mettre à jour certains aspects écologiques de la catégorisation qui sont pris en compte dans le modèle PAC et qu'il semblerait intéressant d'inclure dans l'étude de la catégorisation avec des catégories artificielles. Le troisième [Flannagan, Fried et Holyoak, 86] s'intéresse plus particulièrement à l'impact des distributions de probabilités.

A.8.1 Air de famille, cohérence conceptuelle et construction de catégorie [Medin, Wattenmaker et Hampson 87]

Medin et al définissent leur objectif dans le résumé de leur article «Cet article est le compte-rendu de 7 expériences utilisant des *tâches de classement*[116] pour évaluer les conditions sous lesquelles les individus préfèrent construire des catégories selon le principe de l'air de famille.» Ils opposent ainsi des catégories en conditions nécessaires et suffisantes à des catégories en air de famille.

Les stimuli Pour cela, ils vont utiliser, dans 4 de leurs 7 expériences, des exemples décrits par 4 variables binaires.

Exemples\Dimension	D1	D2	D3	D4
1	1	1	1	1
2	1	1	1	0
3	1	0	0	0
4	0	1	0	0
5	0	1	1	1
6	0	0	0	0
7	0	0	1	0
8	1	0	1	1
9	0	0	0	1
10	1	1	0	1

Tableau A.8.1 Codage des stimuli sur 4 dimensions binaires dans [Medin, Wattenmaker et Hampson 87][117]

[116] souligné par nous

[117] Ce tableau correspond, entre autres, aux animaux présentés dans la figure A.8.1 (page suivante). Le codage binaire est arbitraire mais il permet une plus grande lisibilité. Les exemples 1 et 6 correspondent aux dessins qui sont au centre des deux groupes de cinq de la figure A.8.1, ce sont les prototypes. Il aurait été possible de coder la figure 1 « 1010 » et la figure 6 « 0101 » mais ensuite l'air de famille des autres figures auraient été moins facilement repérables qu'ici. Bien évidemment les sujets n'ont pas connaissance de ce codage, ils n'ont connaissance que des figures.

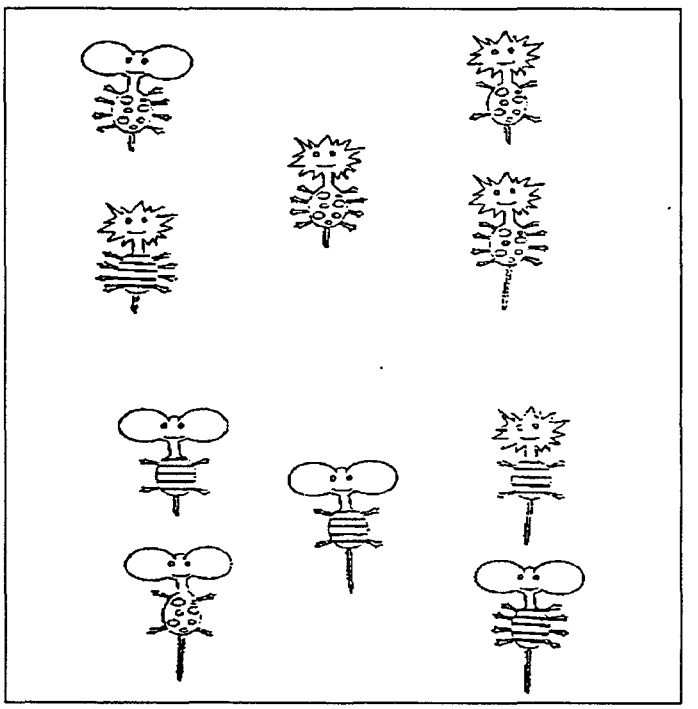


Figure A.8. 1 : stimuli utilisés par [Medin, Wattenmaker et Hampson, 87]

Selon que l'on classe ces exemples en «air de famille » ou en conditions nécessaires et suffisantes il est possible d'obtenir les classements suivants :

Classement en «air de famille »		Classement selon une dimension	
Catégorie A	Catégorie B	Catégorie A	Catégorie B
1111	0000	1000	0100
1110	0001	1111	0111
1101	0010	1110	0000
1011	0100	1101	0010
0111	1000	1011	0001

Le classement en «air de famille » est caractérisé par le fait que chaque exemple d'une catégorie ne diffère d'un autre exemple de la même catégorie que par la valeur de 2 dimensions et que tous les exemples ne diffèrent relativement à l'exemplaire prototypique (en tête de colonne) que par la valeur d'une variable. Le classement unidimensionnel est basé sur la dimension D1, selon les valeurs de cette variable, l'item est classé en A (D1=1) ou en B (D1=0).
Nous pouvons déjà faire deux remarques. La première consiste dans la complexité de l'échantillon. Sur la base de 4 variables binaires, l'espace des exemples n'est que de taille 16. De plus, seuls 10 de ces exemples sont retenus.
La seconde remarque est une conséquence de ce premier constat. Elle consiste dans les différences entre les deux types de catégories : un classement par «air de famille » ne diffère d'un classement unidimensionnel que par l'appartenance de deux exemples : 0111 et 1000. Si nous permutons ces deux éléments, en les faisant changer de catégorie, nous passons de catégories en CNS à des catégories en «air de famille ». Cette absence de complexité se vérifie aussi au travers des mesures de similarité calculées par les auteurs : la similarité interne à une catégorie est de 9,6 dans le

premier classement et de 9,2 dans le second[118]. La différence entre les deux n'est pas significative. Il est tout aussi possible de considérer que le second classement respecte un «air de famille ».

Cette situation n'est donc pas assez complexe. Cette insuffisance dans la complexité rejaillit sur l'absence d'opposition nette entre les deux options de classement. Les auteurs prêteront attention à ce problème en passant à 6 variables ou en proposant des variables ternaires. Mais cet accroissement peut être toujours insuffisant pour permettre une modification des résultats.

Les stimuli sont construits autour des deux prototypes symbolisés par 1111 ou 0000. Ce seront soient des dessins représentant des animaux (voir les dessins de la figure A.8.1), soient des individus définis par 4 ou 5 variables décrivant leur personnalité : extraverti/introverti, agréable/désagréable, consciencieux ou non, stable ou non émotionnellement, cultivé ou non. Notons qu'avec le second type de stimuli, les auteurs introduisent des catégories qui ont *déjà* été construites par les sujets.

La procédure générale

Celle-ci consiste à demander aux sujets de classer les 10 items en 2 classes de taille égale, en leur expliquant qu'il n'y a pas de normes au classement, que l'on peut en envisager de diverses sortes. Dans quelques-unes des 7 expériences, des consignes supplémentaires seront introduites de façon à essayer d'obtenir des classements en «air de famille ». Cette procédure explique pourquoi nous avons mis en italique, ci-dessus, le fait qu'ils utilisent des tâches de classement. Pour nous, cela pose un second problème : cette procédure revient à assimiler la catégorisation à un classement des items, ce qui est exact, dans des catégories totalement arbitraires, ce qui est faux. L'individu ne classe pas arbitrairement les éléments de son environnement, il les classe de façon utilitaire, de manière à pouvoir s'adapter à cet environnement (voir 2.2), *c'est pourquoi nous disons que l'environnement impose au sujet les catégories qu'il doit établir. Par ailleurs, la catégorisation n'est pas simplement un processus de classement mais un processus d'apprentissage*. Le classement n'apparaît qu'au bout de l'apprentissage. De quelques exemples étiquetés d'une catégorie que lui fournit l'environnement, le sujet doit pouvoir en inférer une représentation de la catégorie qui lui permette d'étiqueter d'autres exemples. Enfin, les sujets sont des étudiants en psychologie donc des adultes qui se sont formés dans la culture occidentale. Dans cette culture, lorsque l'on demande à un individu de classer, il y a toujours implicitement l'idée que ce classement doit obéir à des conditions nécessaires et suffisantes.

Les expériences

Dans l'expérience 1, les stimuli sont construits sur 4 variables et consistent dans des dessins d'animaux fictifs. Echec : les sujets classent de manière unidimensionnelle.

L'expérience 2 sera déclinée en plusieurs versions, la différence étant dans les stimuli proposés.

---

[118] Le calcul est simple. Par exemple, pour la catégorie A du classement par «air de famille » : le premier élément 1111 a 3 fois les valeurs de toutes ses dimensions répétées dans les autres exemples de la même catégorie, il a donc une similarité de 12, en calculant de la même façon les autres ont une similarité de 9, la moyenne pour la catégorie est de 9,6 ((12+9+9+9+9)/5)

Dans l'expérience 2a, les auteurs proposent les mêmes figures mais en passant de 4 variables à 6 (sourire ou non, des antennes) : échec pour obtenir une catégorisation en «air de famille »
Dans l'expérience 2 b, les figures proposées sont moins «similaires ». Les têtes, les queues, les pattes ne sont plus totalement identiques. Mais il y a toujours 2 dimensions essentielles pour chacune des variables. Ainsi si les pattes ne sont pas toutes les mêmes, les items en ont toujours soit 4 soit 6. Résultat : échec pour obtenir une catégorisation en «air de famille »

Dans l'expérience 2 c et 2d, les stimuli sont des descriptions de personnes selon 4 caractéristiques (stabilité émotionnelle, consciencieux (oui ou non), cultivé (oui/non), sympathique/antipathique). Comme nous le disions ci-dessus, les stimuli deviennent alors beaucoup moins artificiels. Les résultats sont néanmoins identiques : échec pour obtenir une catégorisation en «air de famille ». Notons cependant que pour ces dernières expériences les sujets ne classent pas non plus toujours selon une seule dimension. Il est possible que le fait que les items ne soient pas artificiels amènent les sujets, dans le classement présent, à utiliser les catégories qu'ils ont déjà formées.

L'expérience 3 est plus intéressante dans le sens où les auteurs donnent au sujet une information sur les animaux. Dans l'expérience 3a on leur dit que les animaux ont une même racine génétique, dans l'expérience 3b qu'ils vivent sur ou sous l'eau : cela commence à s'apparenter à un étiquetage. Le classement proposé par les sujets reste néanmoins unidimensionnel[119].

L'expérience 4 se fait avec des dessins mais on passe de variables binaires à des variables ternaires, en s'arrangeant pour qu'un classement de type unidimensionnel ne soit plus possible. Les auteurs indiquent que sur 18 personnes 4 ont classé en «air de famille ». Cependant les explications que les sujets donnent de leur classement ne correspondent pas à un classement en air de famille mais à un classement en conditions nécessaires et suffisantes : disjonctions ou conjonction de disjonctions.

Dans toutes ces expériences les classifications étaient faites selon une dimension. Dans ces expériences les variables étaient indépendantes. Dans les suivantes, les auteurs veulent vérifier l'impact de la corrélation des valeurs des variables sur le classement. Aussi, dans l'expérience 5, les stimuli se répartissent-ils ainsi :

Exemples\Dimensions	D1	D2	D3	D4	D5
1	1	0	1	0	1
2	1	0	0	0	0
3	1	1	1	1	1
4	1	1	0	1	0
5	0	1	1	1	1
6	0	1	0	1	0
7	0	0	1	0	1
8	0	0	0	0	0

Tableau A.8.2 Codage binaire des stimuli à 5 dimensions dans [Medin, Wattenmaker et Hampson 87]. Les dimensions 2 et 4 sont corrélées, de même pour les dimensions 3 et 5

[119] Il y a une certaine ironie à constater que même lorsque l'on dit aux sujets que les animaux ont la «même racine génétique » on ne puisse obtenir un classement en «air de famille ».

On peut noter que les variables D2 et D4 varient simultanément, de même que les variables D3 et D5.
Le problème est que dans le même temps, Medin et al proposent d'autres types de stimuli, ce qui, de notre point de vue, ne permet pas d'établir des comparaisons avec les précédentes expérimentations et surtout remet en question les résultats des expériences 6 et 7.

Dans l'expérimentation 5a, les items seront des individus décrits selon 5 symptômes de maladies. Par exemple : «perte de sommeil », «muscles ankylosés »... Dans l'expérience 5b ce seront des descriptions d'animaux : «vit dans les arbres », «coloré », etc. Résultats : les sujets classent majoritairement selon la corrélation d'attributs.

Dans les deux dernières expériences, 6 et 7, les auteurs testent l'idée selon laquelle si la corrélation d'attributs est évidente les sujets classeront en «air de famille ».
Dans l'expérimentation 6, les stimuli sont des descriptions de personnes correspondant à deux types : introverti/extraverti. Alors que dans l'expérience 6a on ne précise rien, dans l'expérience 6b on avise les sujets que le classement doit correspondre à ces deux classes (introverti/extraverti). Les sujets classent majoritairement par «air de famille ». Ceux de l'expérience 6a expliquent même leur classement en fonction de termes similaires à introverti et extraverti : «agréable d'être avec » ou «le genre de personne avec laquelle il vaut mieux ne pas sortir ». Le problème pour nous, c'est que les catégories n'ont pas été formées au cours de l'expérience. Au contraire, les sujets ont classé les items selon des catégories qu'ils avaient formées *avant l'expérience*.

Nous adressons le même reproche à la dernière expérimentation. Dans l'expérience 7, les stimuli sont des dessins d'animaux comportant des plumes, des poils des becs etc. et l'on demande aux sujets de classer en animaux volants ou non.

Commentaires

Nous considérons que les expériences 5, 6 et 7 ne sont pas concluantes, dans le sens où ces expériences n'étudient pas le processus de catégorisation mais des catégories déjà formées. Il est donc possible que ce ne soit pas l'évocation d'une éventuelle corrélation d'attributs qui peut expliquer le classement en «air de famille » mais plus simplement l'évocation de représentations de catégories déjà établies.
Les expériences 1 à 4 sont, elles, plus proches d'une étude du processus de catégorisation. Le problème est de savoir si les auteurs ont correctement reproduit le processus en question, ce qui pourrait expliquer pourquoi leurs résultats sont négatifs. Les remarques que nous avons formulées sont :
-la situation n'est pas assez complexe,
-l'environnement n'impose pas au sujet les catégories qu'il doit établir,
-la catégorisation est envisagée comme un simple processus de classement et non pas comme un processus d'apprentissage

On notera que ces trois remarques ne sont jamais qu'une lecture des expérimentations de Medin et al au travers du modèle PAC. La question qui peut se poser maintenant est de savoir si elles sont fondées : si les trois conditions sont respectées nous devrions obtenir une

catégorisation en «air de famille ». La réponse est affirmative et pour le montrer nous allons, plus brièvement, relater les expériences de Kemler-Nelson.

A.8.2 Les effets de l'intention sur le type de concepts acquis [Kemler-Nelson, 84]

La raison principale de l'article de Nelson est l'apprentissage incident, celui que le sujet effectue à son insu opposé à l'apprentissage intentionnel dans lequel le sujet sait qu'il apprend. Elle va donc proposer des expériences qui opposent ces deux types d'apprentissage. L'idée est qu'en apprentissage incident, les sujets formeront les catégories en «air de famille », tandis qu'en apprentissage intentionnel, ils formeront des catégories unidimensionnelles.

Les stimuli sont des caricatures de visage variant sur 4 dimensions binaires : le type de chevelure, le type de moustache, la couleur des yeux et la forme du nez. Les exemples, comme dans Medin et al, sont construits autour de 2 prototypes[120] :

Catégorie I	Catégorie II
0000	1111
0100	1011
0010	1101
0001	1110

Test
0111
1000

Les 8 stimuli présentés sous les catégories sont ceux qui servent à l'apprentissage. Les deux stimuli sous «test » servent à vérifier le type de catégories qu'ont formées les sujets. On aura tout de suite compris que si le sujet classe le premier exemple test dans la catégorie I et le second dans la catégorie II c'est que les représentations de ses catégories sont unidimensionnelles (en fonction de la valeur de la première variable), tandis que, s'il fait l'inverse, c'est qu'il aura formé des catégories en «air de famille ». Le premier exemple a en effet une plus grande similarité avec les éléments de la catégorie II qu'avec ceux de la catégorie I. Il partage 3 valeurs d'attributs avec le prototype de la catégorie II contre 1 avec celui de la catégorie I.

Les sujets, des étudiants, sont répartis en 2 groupes. Un groupe est en apprentissage intentionnel, l'autre en apprentissage incident. Les sujets du groupe en apprentissage intentionnel sont informés qu'ils devront reconnaître quels sont les visages qui appartiennent aux docteurs et ceux qui appartiennent aux policiers. Notons, ici, que la référence à des catégories déjà formées n'a aucune importance, Nelson aurait tout aussi bien pu dire aux sujets qu'ils devraient repérer si les visages sont des visages de X ou de Y.
L'apprentissage se déroule ainsi : à chaque essai, l'expérimentateur présente à l'apprenant un visage et lui demande de l'étiqueter. Dès que le sujet a répondu, l'expérimentateur apporte la solution. Tous les sujets voient ainsi défiler 24 exemples

[120] voir note précédente concernant le codage.

(3 fois les 8 exemples d'apprentissage) nonobstant l'adéquation de leurs précédentes réponses.
Pour faire la liaison avec l'autre groupe, il faut préciser que les sujets ne répondent pas oralement, mais en plaçant le dessin du visage sur un uniforme de docteur ou un uniforme de policier.

Pour les sujets du groupe d'apprentissage incident, la démarche est un peu plus complexe. L'expérimentateur explique au sujet qu'il devra dire s'il a déjà vu le dessin de visage qu'il lui présente. Bien entendu, au fur et à mesure de la présentation le sujet sera amené à dire qu'il a déjà vu les visages. Pour présenter chaque dessin l'expérimentateur, le place au-dessus de l'uniforme de docteur ou celui de policier.

Dans la phase de test, pour les deux groupes, la procédure sera la même que celle utilisée pour l'apprentissage intentionnel. C'est-à-dire que l'on présente aux sujets des visages qu'ils doivent classer en docteur ou policier, mais cette fois-ci l'expérimentateur ne leur donne plus la réponse correcte. Préalablement au test, pour le groupe d'apprentissage incident, l'expérimentateur a fait remarquer aux sujets que parfois il plaçait le visage au-dessus de l'uniforme de policier et parfois au-dessus de celui de docteur et qu'ils doivent maintenant retrouver les policiers et les docteurs. Le test se déroule en trois phases, la première est constituée de 10 essais au cours desquels on présente tous les exemples qui ont servi à l'apprentissage, les prototypes étant présentés 2 fois. Dans la seconde phase, il y a 18 essais, chaque exemple prototype est présenté une fois, chaque exemple test est présenté 4 fois et le reste est constitué de la moitié des exemples d'apprentissage. Dans la troisième, il y a 18 essais aussi, tous les exemples pouvant être générés par les 4 variables binaires sont présentés 1 fois et les prototypes 2 fois. Ainsi parmi les 46 exemples de test, 30 sont des exemples déjà vus lors de l'apprentissage

Les résultats

Kemler-Nelson considère que les sujets ont appris lorsqu'ils ont 20 réponses correctes sur les 30 exemples vus lors de l'apprentissage. 70% du groupe de sujets en apprentissage incident a ainsi appris contre 94% du groupe en apprentissage intentionnel. Sur les 2 groupes, Nelson a pu obtenir 16 sujets ayant appris.
Concernant les 10 exemples test, les 2 exemples apparaissant chacun 5 fois, le nombre moyen de classements unidimensionnels est de 7,9 pour le groupe intentionnel et de 3,5 pour le groupe incident. Le nombre moyen de classements par «air de famille » est de 2,1 pour le groupe intentionnel et de 6,5 pour le groupe incident.

Commentaires

Contrairement à Medin et al, Kemler-Nelson reproduit une catégorisation en «air de famille ». Celle-ci est plus forte pour le groupe d'apprentissage incident que pour le groupe intentionnel. Kemler-Nelson respecte deux des points que nous avions signalés plus haut concernant la reproduction du processus de catégorisation en laboratoire : l'environnement impose au sujet les catégories qu'il doit former et le processus de catégorisation est un processus d'apprentissage. Il est étonnant que, lorsque Medin et al se sont interrogés sur l'échec de leurs expérimentations, ces points ne soient pas apparus, alors qu'ils connaissaient les expériences de Kemler-Nelson.

Si comme dans Medin et al, les stimuli ne sont pas très complexes, nous pouvons cependant constater que c'est dans la plus difficile des deux situations (apprentissage incident) que l' «air de famille » apparaît le plus. Par ailleurs, la procédure employée par Kemler-Nelson est moins critique relativement à cette complexité car le fait de mettre les exemples tests dans l'une ou l'autre catégorie détermine le type de catégorisation[121]. Il aurait néanmoins été intéressant de voir ce que cela aurait donné si la situation avait été beaucoup plus complexe : on peut noter que même dans le cas de l'apprentissage intentionnel, certains sujets classent par «air de famille », il est possible que si la situation avait été plus complexe, ce nombre de sujets aurait été plus grand.

A.8.3 Attentes sur les distributions et l'induction de structure de catégorie [Flannagan, Fried, et Holyoak, 1986]

Dans Medin et al comme dans Kemler-Nelson, il y a un facteur implicite dans leur expérimentation, il s'agit des distributions de probabilités. Ainsi, les exemples 1100, 1010, 0011, 0101, 0110, 1001 n'apparaissent pas car les exemples présentés ont été construits autour de deux prototypes et ces exemples en sont trop élopignés. Le fait de ne pas inclure ces exemples, revient à admettre implicitement qu'ils ne permettraient pas une construction en «air de famille ». Autrement dit, à reconnaître que la distribution de probabilités intervient dans la formation des catégories.

L'étude des distributions de probabilités a été faite, mais d'une manière autre, par Flannagan M.J., Fried L.S, Holyoak K.J. en 1986 dans *Attentes sur les distributions et l'induction de structure de catégorie* [Flannagan, Fried, et Holyoak, 1986]. Leur article vise à vérifier si les sujets ont des attentes relativement aux distributions de probabilités. Il surprend car la catégorie ne s'y définit *que* par sa distribution de probabilités. Pour cette raison, nous présentons un exemple simplifié de leur type de stimuli et de concept qui permettra de mieux comprendre leur expérience décrite ensuite.

Supposons que les stimuli soient des tableaux représentant des croix. Chacune des deux branches (x ou y) de ces croix varient entre 55 cm et 100 cm par intervalle de 5cm. Les stimuli sont donc constitués de 2 attributs qui peuvent avoir 10 valeurs. Il y a donc 100 stimuli différents possibles.
Supposons maintenant une fonction f= int[(x+y-100)/10], avec cette fonction à chaque stimulus correspond un nombre entre 1 et 10 et à chaque nombre 1 et 10 correspondent 10 stimuli. Par exemple si la croix a une branche qui fait 75cm et l'autre 70cm nous obtenons f(x,y)=int(4,5)=4.
L'originalité chez Flannagan et al est qu'un concept va être défini par sa distribution de probabilité « the objective definition of the target was probabilistic ». Si nous reprenons notre exemple et que nous choisissons la distribution NL de la figure A.8.3.1 (ci-dessous) nous voyons que les tableaux tels que f(x,y)=4 ont une probabilité proche de 40%, ceux tels que f(x,y)=10 une probabilité de 0.

Lorsque l'on présente des exemples du concept aux sujets cette présentation correspond à la distribution de probabilités. Ainsi près de 40% des exemples sont tels que f(x,y)=4 alors qu'il n'y a aucun exemple tel que f(x,y)=8 ou 9 ou 10.

[121] Il aurait été intéressant de savoir si les sujets étaient «constants » dans le sens où, pour un sujet donné, celui-ci optait systématiquement pour un type de classement ou l'autre, mais Kemler-Nelson n'en parle pas.

Lors du test par contre on prend la distribution uniforme pour présenter les exemples de test, les 100 exemples auront la même chance d'apparaître, et l'on demande aux sujets de les étiqueter en exemples positifs ou négatifs. Bien entendu comme le concept est défini par sa distribution de probabilités, ce n'est pas la manière qu'ont les sujets d'étiqueter *chaque* tableau qui compte mais la manière qu'ils ont d'étiqueter *tout l'échantillon.* On s'attend par exemple à ce que parmi les exemples étiquetés comme positifs, relevant du concept, on trouve 40% de tableaux tels que f(x,y)=4 et aucun des tableaux tels que f(x,y)=8 ou 9 ou 10. Autrement dit, pour un f(x,y) donné peu importe que ce soit tel tableau plutôt que tel autre qui soit étiqueté positif, l'important c'est que la probabilité soit respectée.

Flannagan et al testent trois groupes de sujets sur trois distributions de probabilités différentes.

L'expérimentateur présente les stimuli aux sujets comme étant des tableaux abstraits. Ces tableaux sont composés de trois rectangles, chacun de ces rectangles ne peut varier que sur une dimension, soit la hauteur soit la largeur. Les stimuli varient donc sur trois dimensions, appelons x, y, z chacune de ces dimensions. Comme il y 10 degrés de variation possible pour chacune de ces trois dimensions, cela fait 10x10x10=1000 exemples possibles.

Les auteurs définissent une fonction f(x,y,z) qui à chaque exemple affecte un nombre entre 0 et 10. Cette fonction permettra de définir le concept, les tableaux d'un peintre appelé Vango, selon l'une des trois distributions de probabilités NL, NH, U («normal high-mean », «normal low-mean » et «U-Shaped ») que l'on voit dans la figure A.8.3.1

Chacun des trois groupes affrontera donc une distribution (concept) particulière. Pour un premier groupe les tableaux de Vango seront définis par la distribution NL, pour un second par NH et pour le troisième par U. Chacun de ces groupes est scindé en deux : le premier sous-groupe ne voit que 20 exemples tandis que le second en voit 150.

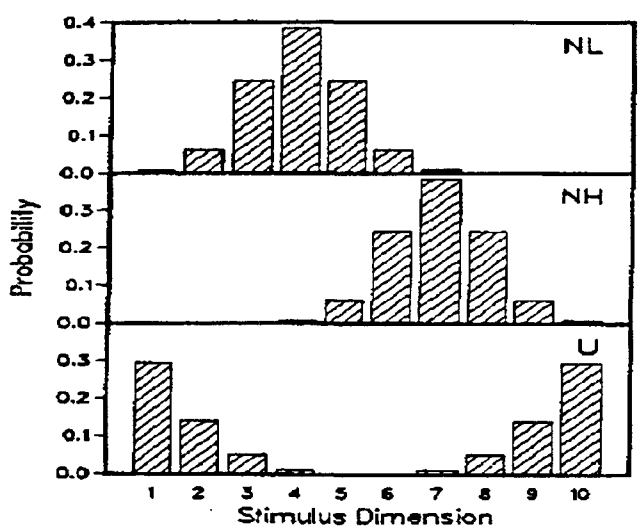


Figure A.8.3.1 les fonctions de probabilité utilisées pour définir les catégories NL, NH et U [Flannagan, Fried, et Holyoak, 1986]

Dans la phase d'apprentissage, durant 3 secondes, selon les groupes l'expérimentateur montre chacun des 20 ou 150 tableaux différents de Vango.
Les exemples de test, eux sont choisis aléatoirement selon la distribution *uniforme*. Les sujets doivent ainsi en classer 125 selon qu'ils pensent que c'est l'artiste en question qui les a peints ou non. Lors de la phase de test, l'expérimentateur regarde selon quelle distribution les sujets ont classé les exemples qu'ils pensent relever de cette même catégorie (voir figure ci-dessous).

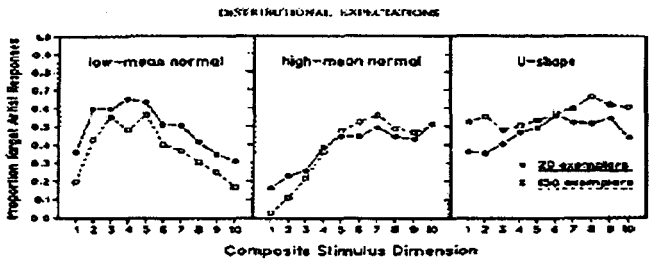


Figure A.8.3.2 catégorisation effectuée par les sujets [Flannagan, Fried, et Holyoak, 1986]

Flannagan et al, comparent ces courbes-ci avec les courbes précédentes. Ils constatent que les courbes «normales» sont relativement bien apprises tandis que les courbes en U ne le sont pas. Ils en déduisent que les sujets ont une certaine attente des distributions de probabilités.

Commentaires

Dans cet article, nous sommes bien dans une situation d'apprentissage où l'environnement impose à l'apprenant la catégorie à former. Il s'agit même d'un type d'apprentissage particulier car *seuls* les tableaux du peintre sont présentés au cours de l'apprentissage, ce que l'on appelle apprentissage par exemples positifs dans le modèle PAC.

Comme nous l'avons dit, ce qui déroute dans leur article est la définition de la catégorie exclusivement par sa distribution de probabilités. Pour un tableau donné, il est impossible de définir son appartenance ou non au concept, c'est l'étiquetage complet de l'échantillon qui doit respecter la distribution de probabilités.

Alors que jusqu'à présent, nous pouvions reprocher aux articles le fait que la distribution de probabilités n'était pas réellement prise en compte, c'est maintenant le concept qui semble un peu omis. Les sujets, ici, apprennent une distribution de probabilités plutôt qu'un concept[122].

[122] Ceci ne constitue pas une critique vis-à-vis à l'article qui est totalement cohérent avec lui-même puisque son objectif est de vérifier si les sujets ont des attentes relativement aux distributions de probabilités.

BIBLIOGRAPHIE

Andler D., 1992, *Calcul et représentation,* Introduction aux sciences cognitives, Gallimard, Paris
Angluin D, 1980, *Inductive inference of formal languages from positive data,* Information and control, 45, 117-135,
Angluin D, 1983, *Inductive inference theory and methods,* Computing Surveys, vol. n°15, ,
Angluin D, 1987, *Learning Regular sets from queries and counterexamples,* Information and computation, 75, 87-106,
Angluin D, 1988, *Queries and concept learning,* Machine learning, 2, 319-342, Kluwer academic publisher, Boston,
Angluin D, 1992, *Computational learning theory : survey and selected bibliography,* 24 th annual ACM STOC,
Ashcraft M-H., 1978, *Property dominance and typicality effects in property statement verification,* Journal of verbal Learning and verbal Behavior, 17, 155-164,
Barsalou L.W., 1983, *Ad hoc categories* , Memory and Cognition, 11, p211-227,
Barsalou L.W., 1985, *Ideals, central tendancy, and frequency of instantiation as determinants of graded structure in categories,* Journal of experimental psychology : learning, memory and cognition vol 11, n°4, p629-654,
Barsalou L.W., 1992, *Cognitive psychology an overview for cognitive scientists,* Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey
Barsalou L.W., 1993, *Challenging about concepts,* Cognitive development, 8, 169-180,
Barth B-M, 1987, *L'apprentissage de l'abstraction* , Editions Retz, Paris
Bateson G., 1984, *La nature et la pensée,* Editions du Seuil,
Baum E.B., 1994, *When are k-nearest neighbor and backpropagation accurate for feasible-sized sets of examples,* Computational learning theory and natural learning systems, vol 1, ch 14,
Bechtel W. Abrahamsen A., 1993, *Le connexionnisme et l'esprit introduction au traitement parallèle par réseaux,* Edition La découverte, Paris
Berbaum J., 1994, *Apprentissage,* Dictionnaire encyclopédique de l'éducation et de la formation, Nathan Université,
Berwick R.C, 1986, *Learning from positive-only examples : the subset principle and three cases studies,* Machine Learning, Tome 2, Chapitre 21, Morgan Kaufmann,
Bettman J.R., Johnson E.J., Luce M.F., Payne J.W., 1993, *Correlation, conflict and choice,* Journal of experimental psychology : learning, memory and cognition vol 19, n°4, p931-951,
Bideaud J., Houdé O., 1993, *Catégorisation, logique et prototypicalité : aspects developpementaux,* Sémantique et cognition, D.Dubois, Editions du CNRS, Paris
Blewitt P., 1993, *Taxonomic structure in lexical memory : the nature of developmental change,* Annals of child development, vol 9, 103-132,
Blewitt P., 1994, *Understanding categorical hierarchies : the ealiest levels of skill,* Annals of child development, vol 65, 1279-1298,
Blum A.L. et Langley P., 1998, (in press),*Selection of relevant features and examples in machine learning,* , Artificial Intelligence

Blumer A., Ehrenfeucht A., Haussler D., Warmuth M., 1986, *Classifying learnable geometric concept with the VCD,* Proc. 18th ACM symposium on theory of computing, p 273-282 ACM press,
Blumer A., Ehrenfeucht A., Haussler D., Warmuth M., 1987, *Occam's razor,* Inform. Proc. Lett, 24, p 377-380,
Board, Pitt, 1990, *On the neccesity of Occam algorithms,* Proc. 22 TH ACM Symposium on Theory of Computing, p 54-63,
Borderie D., 1997, *Conception et implémentation d'un système de simulation du processus de catégorisation chez l'enfant : les système ROCE.,* Thèse de Doctorat en Productique et Informatique, Université de Droit, d'Economie et des Science d'Aix Marseille III,
Boucheron S., 1992, *Théories formelles de l'apprentissage : de l'approche formelle aux enjeux cognitifs,* Hermès,
Bourne L.E., 1970, *Knowing and using concepts,* Psychological review, 77, 546-556,
Bourne, -Lyle-E. , 1982, *Typicality effects in logically defined categories. ,* Memory-and-Cognition; 1982 Jan Vol 10(1) 3-9 ,
Brunel N., Ninio J., 1997, *Time to detect the difference between two images presented side by side,* Cognitive Brain Research 5, 273-282,
Bruner, Goodnow et Austin, 1956, *A study of thinking,* Wiley, New York
Bshouty N H, 1993, *Exact learning via the monotone theory,* IEEE,
Caraux G. Lechevallier Y., *Règles de décision de Bayes et méthodes statistiques de discrimination, ,*
Carbonell J.G., Michalski R.S., Mitchell M, 1983, *An Overview of Machine Learning,* Machine Learning, Tome 1, Chapitre 1, Morgan Kaufmann,
Changeux J-P., 1983, *L'homme neuronal,* Librairie Fayard,
Chater N., Lyon K., Myers T., 1990, *Why are conjunctive categories overextended ?,* Journal of experimental psychology : learning, memory and cognition, 16, 497-508,
Clark E.V., 1973, *What's in a word ? On the child's acquisition of semantics in his first language,* Cognitive development and the acquisition of language T.E. Moore (ed), New York, Academic Press,
Conant M.B., Trabasso T., 1964, *Conjunctive and disjunctive concept formation under equal condition information,* Journal of experimental psychology, 67, 250-255,
Cook D.J, 1994, *Defining the limits of analogical planning,* Computational learning theory and natural learning systems, Ed. Hanson SJ, Drastal G.A., Rivest RL, vol 1, ch 4, MIT
Cordier F., 1980, *Echelles de typicalité pour 5 catégories sémantiques,* Psychologie Française, 27, p 213-221,
Cordier F., 1983, *Abstraction d'une information typique chez des enfants,* Cahiers de psychologie cognitive, 3, 461-474,
Cordier F., 1993, *Les représentations cognitives privilégiées typiclaité et niveau de base,* PUL, Lille
Cordier F., 1994, *Représentation cognitive et langage: une conquête progressive,* Armand Colin, Paris
Cordier F., Dubois D., 1981, *Typicalité et représentation cognitive,* Cahiers de psychologie cognitive 1:299-333,

Cordier F., Spitz E., 1998, *Nature des catégories et typiclaité : une étude développementale,* ENFANCE, N°4/1998, p 189-202,
Cornuéjols A, 1995, *Analogie, principe d'économie et complexité algorithmique,* Actes des Journées Françaises de l'Apprentissage, JFA96,
Cottrell G.W., 1990, *Extracting features from faces using compression networks : faces, identity, emotion and gender recognition using holons,* Connection Models : proceedings of the 1990 summer school, D.Touretsky editor, Morgan Kaufmann, San Mateo CA
Courrieu P., 1994, *Connexionisme et fonctions symboliques,* Psychologie française N°39-2, 231-236,
Craven M.W., Shavlik J.W., 1994, *Investigating the value of a good input representation,* Computational learning theory and natural learning systems, vol 3, ch 16,
De la Higuera C, 1996, *Ensembles caractéristiques en inférence grammaticale,* Actes des Journées Françaises de l'Apprentissage, JFA96,
Delahaye J-P., 1996, *Information complexité et hasard,* Langue-Raisonnement-Calcul, Hermès,
Delepoulle S. Preux P. Darcheville J-C, 1998, *Partage des tâches et apprentissage par renforcement,* Journées Françaises de l'Apprentissage, JFA' 98,
Denis F., 1998, *PAC learning with positive examples,* ALT98, 9 th International Workshop on Algorithmic Learning Theory
Denis F., 1998, *Apprentissage PAC par exemples positifs,* Journées Françaises de l'Apprentissage, JFA' 98,
Denis F., D'Halluin C., Gilleron R., 1996, *Pac Learning with simple examples,* In STACS96, 13th Annual Symposium on Theoretical Aspects of Computer Science, vol . 1046 of lecture Notes in Computer Science, p231-242,
Denis F., Gilleron R., 1997, a, *PAC learning under helpful distributions,* ALT 97, 8 th International Workshop on Algorithmic Learning Theory, vol 1316 of lecture notes in Artificial Intelligence, p 132-145,
Denis F., Gilleron R., 1997, b, *Notes de cours sur l'apprentissage automatique,* cours optionnel de la maîtrise d'informatique de l'université de Lille 1
Denis F., Gilleron R., Simon J., 1997, *Apprentissage PAC avec enseignant,* In JFA 97, pages 175-186, 1997,
Dietterich T.G., Michalski R.S., , 1983, *A Comparative Review of Selected Methods for Learning from Examples,* Machine Learning, Tome 1, Chapitre 3, Morgan Kaufmann,
Dijkstra S., Dekker P.H., 1982, *Inference processes in learning well-defined concepts,* Acta-psychologica, vol 51(3) 181-205,
Dominowski R.L., 1973, *Requiring hypotheses and the identifcation of unidimensional, conjunctive and disjunctive concept,* Journal of experimental psychology, 100, 387-394,
Drastal G., 1994, *Learning in abstraction space,* Computational learning theory and natural learning systems, Ed. Hanson SJ, Drastal G.A., Rivest RL, vol 1, ch 7, MIT
Drucker H., Schapire R., Simard P., 1992, *Improving performance in neural networks using a boosting algorithm,* In S.J. Hanson, J.D Cowan and C.L. Giles Edcitor, Advances in Neural Information Processing Systems, p 42-49, Morgan Kaufmann, San-Mateo, CA,

Dubois D., 1993, a, *Catégorisation et cognition "10 ans après" :une évaluation des concepts de Rosch,* Sémantique et cognition, D.Dubois, Editions du CNRS, Paris
Dubois D., 1993, b, *Introduction,* Sémantique et cognition, D.Dubois, Editions du CNRS, Paris
Dubois D., 1994, *From classifications to cognitive representations : the exemple of road lexicons,* New approaches in classifications and data analyses, E.Diday Springer Verlag, Paris
Dubois D., 1997, *Catégorisation et cognition : de la perception au discours,* Sémantique et cognition, Editions Kimé, Paris
Dubois D., Resche-Rigon P., Tenin A., 1997, *Des couleurs et des formes : catégories perceptives ou constructions cognitives,* Sémantique et cognition, Editions Kimé, Paris
Edelman G.M., 1992, *Biologie de la conscience,* Edition Odile Jacob, Paris
Egeth H.E., Virzi R.A., Garbart H., 1984, *Searching for conjunctively defined target,* Journal of experimental psychology : human perception and performance, 10, 32-39,
Evans J.S.B, Newstead S.E, 1980, *A study of disjunctive reasonning,* Psychological-Research Vol 41(4) 373-388,
Fayyad U.M., 1991, *On the induction of decision trees for multiple concept learning,* Ph. D. Dissertation, EECS Ddepartment, University of Michigan
Fayyad U.M.,Weir N., Djorgovski S., 1993, *SKICAT: a machine learning system for automated cataloging of large scale sky surveys,* Proceedings of the Tenth International Conference on Machine Learning, 112-129, Morgan Kaufmann,
Finton J.F, Hu Y.H., 1994, *Importance-based feature extraction for reinforcement learning,* Computational learning theory and natural learning systems, vol 3, ch 5,
Flannagan M.J., Fried L.S, Holyoak K.J., 1986, *Distributionnals expectations and the induction of category structure,* Journal of experimental psychology : learning, memory and cognition, 12, 241-256,
Freivalds R., Kinber E.B., Wiehagen R., 1993, *On the power of inductive inference from good examples,* Theoretical Computer Science 110, p131-144, Elsevier,
Fried L.S, Holyoak K.J., 1984, *Induction of category distributions : a framework for classification learning,* Journal of experimental psychology : learning, memory and cognition vol 10, n°2, p 234-257,
Gallinari P. Gascuel O., *Statistique, apprentissage et généralisation ; application aux réseaux de neurones,* Intelligence Artificielle, n°10,
Gascuel O., 1993, *Aspects statistiques de l'apprentissage inductif,* Support de cours Ecole sur l'apprentissage automatique, St-Raphaël,
Gold E M, 1967, *Language identification in the limit,* Information and control, 10, 447-474,
Gold E M, 1978, *Complexity of automaton identification from given data,* Information and control, 37, 302-320,
Goldman S., Kearns M., 1995, *On the complexity of teachiing,* Journal of Computer and System Science, 50, p 20-31,
Goldman S.A., Kearns M.J., 1991, *On the complexity of teaching,* Proc. Of the 4th International Workshop on Computational Learning Theory, p303-314, Morgan Kaufmann Publishing, Inc., San-Mateo,
Goldman S.A., Mathias H.D., 1996, *Teaching a smarter learner,* Journal of Computer and Sytem Science, 52, p 255-267,

Goldstone R.L, Barsalou L.W, 1996, *Reuniting perception and conception : the perceptual bases of bases of similarity and rules,* ? sur le net,
Hampton J., Dubois D., 1993, *Psychological models of concepts : introduction,* Categories and concepts : theoretical views and inductive data analysis, Mechelen, Hampton, Michalski and Theuns (Eds), Cognitive Science Series, Academic Press,
Hampton J.A., 1988, *Overextension of conjunctive concepts : evidence for a unitary model of concept typicality and classe inclusion,* Journal of experimental psychology : learning, memory and cognition, 14, 12-32,
Hampton J.A., Gardiner M.M., 1983, *Measures of internal category structure : a correlational analysis of normative data,* British journal of psychology, 74, p 491-516,
Hampton, -James-A. , 1988, *Disjunction of natural concepts. ,* Memory-and-Cognition; 1988 Nov Vol 16(6) 579-591 ,
Haton J.P., Haton M.C., 1993, *L'Intelligence Artificielle,* PUF, Paris, 3e édition corrigée,
Haugeland J., 1989, *L'esprit dans la machine, fondements de l'intelligence artificielle,* Editions Odile Jacob, Paris
Haussler D., 1987, *Bias, Version Spaces and Valiant's learning framework,* Proc. Of the 4th International Workshop on Machine learning, p 324-336, Irvine, California, Morgan Kaufmann,
Haussler D., 1990, *Applying Valiant's learning framework to AI-concept-learning problems,* Machine Learning, Tome 1, Chapitre 22, Morgan Kaufmann,
Haussler D., Kearns M., Littlestone N., Warmuth M.K., 1991, *Equivalence of models for polynomial learnability,* Inform. Comput., 95(2), p129-161,
Hayes Roth B, Hayes Roth F., 1977, *Concept learning and the recognition and classification of exemplars,* Journal of verbal learning and verbal behavior , 16, 321-338,
Heider-Rosch E., Olivier D., 1972, *The structure of the color space in naming and memory for two languages,* Cognitive Psychology, 3, p334-354,
Hinton G., 1990, *Connectionnist learning procedures,* Machine Learning, Tome 3, Chapitre 20, Morgan Kaufmann,
Hinton G., 1992, *Apprentissage et réseaux de neurones,* Pour la science n°181,
Houdé O, 1992, *Catégorisation et développement cognitif,* PUF, Paris
Houdé O., Kayser D., Koenig O., Proust J., et Rastier F., 1998, *Apprentissage,* Vocabulaire de sciences cognitives, Presses Universitaires de France,
Hubel D.H., Wiesel T.H., 1962, *Receptive fields, binocular interaction, and functionnal architecture in the cat's visual cortex,* Journal of physiology, 166, 106-154
Huteau M., 1993, *Organisation catégorielle des objets sociaux : portée et limites des conceptualisations de E. Rosch,* Sémantique et cognition, D.Dubois, Editions du CNRS, Paris
Jackson J., Tomkins A., 1992, *A computational model of teaching,* Proc. Of the 5th International Workshop on Computational Learning Theory, p 319-326, ACM press,
Jacob P., 1992, *Le problème du rapport du corps et de l'esprit aujourd'hui,* Introduction aux sciences cognitives, Gallimard, Paris
Jozefowiez J., Darchevile J-C., Preux P., 1998, *L'émergence de comportements de contrôle chez des agents sélectionnistes leur permet de résoudre le dilemme du prisonnier,* Journées Françaises de l'Apprentissage, JFA' 98,

Juhel J., 1989, *Analyse des aptitudes intellectuelles : revue de quelques travaux récents,* Annee-Psychologique; 1989 Vol 89, 63-86 ,
Kant J.-D., 1996, *Modélisation et mise en œuvre de processus cognitifs de catégorisation à l'aide d'un réseau connexionniste.*, Thèse de Doctorat en Informatique, Université de Rennes 1,
Kearns M., Valiant L., 1994, *Cryptographic Limitations on learning boolean formulae and finite automata,* Journal of the Association for Computing Machinery, Vol 41, n°1, pp 67-95,
Kearns M.J., Li M., Pitt L., Valiant L.G., 1987, *Recent results on boolean concept learning,* Proc. Of the 4th International Workshop on Machine learning, p337-352, Irvine, California, Morgan Kaufmann,
Kearns M.J., Schapire R.E., 1994, *Efficient distribution-free learning of probabilistic concepts,* Computational learning theory and natural learning systems, vol 1, ch 10,
Kearns M.J., Schapire R.E., Sellie L.M., 1992, *Toward Efficient Agnostic Learning,* Proceedings of COLT'92,
Kearns M.J., Vazirani U.V., 1994, *An introduction to Computational Learning Theory,* MIT Press,
Keil F., 1979, *Semantic and conceptual development : an ontological perspective,* Harvard University Press, Cambridge,
Keil F.C., 1989, *Concepts, kinds and cognitive development,* Cambridge, MA : MIT Press,
Kemler-Nelson D.G., 1984, *The effect of intention on what concepts are acquired,* Journal of verbal learning and verbal behavior , 23, 734-759,
Kleiber G, 1993, *Prototype et prototypes : encore une affaire de famille,* Sémantique et cognition, D.Dubois, Editions du CNRS, Paris
Kleiber G., 1990, *La sémantique du prototype,* PUF, Paris
Komatsu L.K, 1992, *Recent views of conceptual structure,* Psychological bulletin, vol 112, n°3, 500-526,
Lammel A., 1997, *Mots, catégories conceptuelles, processus de catégorisation,* Sémantique et cognition, Editions Kimé, Paris
Lang K.J., Waibel A.H., Hinton G.E, 1990, *A time-delay neural network architecture for isolated word recognition* , Neural Network vol. 3, 33-43,
Lecoq P., 1994, *Apprentissage implicite et explicite : propos sur l'inconscient cognitif* , Support de cours, LABACOLIL (1994),
LeCun Y., Boser B., Denker J.S., Henderson D., Howard J.E., Hubbard W., Jackel L.D., 1989, *Backpropagation applied to handwritten zip code recognition,* Neural computation Vol. 1(4),
Lee W.S., Bartlett P.L., Williamson R.C., 1995, *On efficient agnostic learning of linear combination of basis functions,* Proceedings of COLT'95,
LeNy J-F., 1989, *Science cognitive et compréhension du langage,* PUF, Paris
Lewicki, 1986, *Processing information about covariations that cannot be articulated,* Journal of experimental psychology : learning, memory and cognition, 12, 135-146,
Li M, Vitànyi P., 1993, *An introduction to Kolmogorov complexity and its applications,* Text and monographs in Computer-Science. Springer-Verlag,
Li M., Vitanyi P., 1991, *Learning simple concepts under simple distributions,* SIAM, J. Comput., vol 20, 911-935,

Li M., Vitanyi P., 1995, *Computational machine learning in theory and praxis,* NeuroCOLT Technical Report Series, NC-TR-95-052,
Lindsay P.H., NormanD.A., 1980, *Traitement de l'information et comportement humain une introduction à la psychologie,* Edition Vigot, Québec
Littlestone N., 1988, *Learning when irrelevant attributes abound : a new linear-threshold algorithm* , Machine learning, 2, p 285-318,
Lopez de Mantaras R., 1991, *A distance-based attribute selection measure for decision tree induction,* Machine learning, 6(1),81-92
Lucariello J., Nelson K., 1985, *Slot-filler categories as memory organizers for young children,* Developpemental psychology vol 21, n°2, p 272-282,
Malt B.C, Smith E.E., 1982, *The role of familiarity in determining typicality,* Memory and cognition, vol 10(1), p 69-75,
Mazet C., 1993, *Fonctionnalité dans l'organisation catégorielle,* Sémantique et cognition, D.Dubois, Editions du CNRS, Paris
Medin D.L., Schwanenflugel P.J., 1981, *Linear separability in classification learning* , Journal of experimental psychology : human learning and, memory 7, 355-368,
Medin D.L., Wattenmaker W.D., Michalski R.S., 1987, *Constraints and preference in inductive reasonning : an empirical study of human and machine performance,* Cognitive science , 11, 299-339,
Medin D.L., Wattenmaker W.D.et Hampson S.E., 1987, *Family resemblance, conceptual cohesiveness and category construction,* Cognitive psychology, 19, 242-279,
Mervis C.B., Pani J.R., 1980, *Acquisition of basic objects category,* Cognitive psychology, 12, 496-522,
Mervis C-B., Rosch E., 1981, *Categorization of natural objects,* Annual Review of psychology, 32, p 89-115,
Meulemans T., 1998, *Apprentissage implicite, mémoire implicite et développement,* Psychologie française, n°43-1, p 27-37,
Michalski R.S., 1983, *A Theory and Methodology of Inductive Learning,* Machine Learning, Tome 1, Chapitre 4, Morgan Kaufmann,
Michalski R.S., 1986, *Understanding the Nature of Learning : Issues and Research Directions,* Machine Learning, Tome 2, Chapitre 1, Morgan Kaufmann,
Michalski R.S. and Stepp R.E, 1983, *Learning form observation : conceptual clustering,* Machine Learning, Tome 1, Chapitre 11, Morgan Kaufmann,
Michalski R.S., Kodratoff Y., 1990, *Research in Machine Learning : recent progress, classificiation of methodes and future directions,* Machine Learning, Tome 3, Chapitre 1, Morgan Kaufmann,
Mingers J, 1989, *An emprirical comparison of pruning methods for decision-tree induction,* Machine Learning, 4(2), 227-243
Minsky M., Papert S., 1969, *Perceptrons,* Cambridge M.A., MIT Press
Mitchell T.M., 1997, *Machine learning* , McGraw Hill International Editions,
Mooney R.J., 1994, *A preliminary PAC analysis of theory revision,* Computational learning theory and natural learning systems, Ed. Petsche T., Hanson S.J., Shavlik J., vol 3, ch 3, MIT
Myllymäki P., Tirri H., 1994, *Learning in neural networks with Bayesian prototypes,* Proceedings of SOUTHCON'94, 60-64,

Natarajan B.K., 1991, *Machine learning : a theoretical approach,* Morgan Kauffman,
Neisser U., 1987, *Concepts and conceptual development. Ecological and intellectual factors in categorization,* Memory Symposia in cognition Cambridge University Press,
Neisser U., Weene P., 1962, *Hierarchies in concept attainment,* Journal of experimental psychology, vol 64, n°6, 640-645,
Nelson K., 1974, *Concepts, word and sentences : interrelations in acquisition and development,* Psycholgical Review, 81, n°4, p 267-285,
Nelson K., 1985, *Le développement de la représentation sémantique chez l'enfant,* Psychologie Française, 30, p 261-268,
Neumann P.G., 1977, *Visual protype formation with discontinuous representation of dimensions of variability,* Memory and cognition, vol 5, 187-197,
Nicolas S., 1998, *Effet de l'apprentissage intentionnel en mémoire implicite et en mémoire explicite,* Psychologie française, n°43-1, p 89-96,
Nosofsky R.M., 1984, *Choice, similarity, and the context theory of classification,* Journal of experimental psychology : learning, memory and cognition, 10, 104-114,
Nosofsky R.M., 1986, *Attention , similarity and the identification-categorization relation-ship,* Journal of experimental psychology :general, vol 115, p 39-57,
Nosofsky R.M., 1987, *Attention and learning processes in the identification and categorization of integral stimuli,* Journal of experimental psychology : learning, memory and cognition, 13, 87-109,
Oakhill J.V, Johnson P.N., 1985, *Rationality, memory and the search of counterexamples,* Cognition Vol 20(1) 79-94,
Osherson D.N., Smith E.E, 1990, *Thinking An Invitation to Cogntive Science, vol. 3,* MIT Press, Cambridge, Massacussetts,
Osherson D.N., Stob M., Weinstein S., 1986, *Systems that learns* , MIT Press Press, Cambridge, MA,
Pacherie E., 1993, *Aristote et Rosch : un air de famille ?,* Sémantique et cognition, D.Dubois, Editions du CNRS, Paris
Parker D, 1985, *Learning logic,* MIT Technical report TR-47, MIT Center for Research in Computational Economics and Management Science
Pazzani M.J., 1991, *Influence of prior knowledge : on concept acquisition : experimental and computationnal result,* Journal of experimental psychology : learning, memory and cognition, 17, 416-432,
Perruchet P., 1998, *L'apprentissage implicite : un débat théorique,* Psychologie française, n°43-1, p 13-25,
Piattelli-Palmarini M.,1979,*Théories du langage Théories de l'apprentissage Le débat entre Jean Piaget et Noam Chomsky,* Editions du seuil.
Pitt L, 1989, *Inductive inference, DFAs, and computational complexity,* Porc AII-89 Workshop on Analogical and Inductive Inference, Lecture Note in Artificial Intelligence, Vol 397, p 18-44, Springer-Verlag, Heidelberg,
Pitt L., 1997, *On exploiting knowledge and concept use in learning theory,* ALT 97, 8 th International Workshop on Algorithmic Learning Theory, vol 1316 of lecture notes in Artificial Intelligence, p 63-84,
Pitt L., Valiant L.G., 1988, *Computational limitations on learning from examples,* Journal of the Association for Computing Machinery, Vol 35, pp 965-984,

Pitt L., Warmuth M.K., 1990, *Prediction-Preserving Reducibility,* Journal of computer and Sytem Science, 41, p 430-467,
Quinlan J.R., 1979, *Discovering rules by induction from large collections of exemples,* D.Michie (Ed.), Expert systems in the micro electronic age, Edimburgh University Press.
Quinlan J.R., 1983, *Learning efficient classification procedures and their applications to chess end games,* Machine Learning, Tome 1, Chapitre 15, Morgan Kaufmann,
Quinlan J.R., 1994, *Comparing connectionnist and symbolic learning methods,* Computational learning theory and natural learning systems, vol 1, ch 15,
Quinlan J.R., Rivest R.L., 1989, *Inferring Decision trees using the MDLP,* Information and Computation, 80, 227-248,
Rachlin J., Kasif S., Salzberg S., Aha D.W., 1994, *Towards a better understanding of memory-based reasoning systems,* , Proc. 11th International Conference on Machine Learning, pp. 242-250, Morgan Kaufmann,
Ragavan H et Rendell L., 1994, *Learning disjunctive concepts using domain knowledge,* Computational learning theory and natural learning systems, vol 1, ch 6,
Reber A.S., 1976, *Implicit learning of synthetic langages : the rôle of instruction set,* Journal of experimental psychology : human memory and learning, 2, 88-94,
Reber A.S., Allen R., 1978, *Analogic and abstraction stratégies in synthetic grammar learning : a functionnalist interpretation,* Cognition 6, 189-221,
Rendell L. et Seshu R., 1994, *Learning hard concepts through constructive : induction : framework and rationale,* Computational learning theory and natural learning systems, vol 1, ch 5,
Richard J-F. , 1975, *Test of the hypotheses and information storage in the rule-identification task.,* Annee-Psychologique; 1975 Vol 75(2) 331-354 ,
Rissanen J., 1978, *Modeling by shortest data description* , Automatica, 14, p 465-471,
Rivest R L, 1987, *Learning decision lists,* Machine Learning, 2(3), p 229-246,
Rivest R.L., Sloan R., 1998, *Learning complicated concepts reliably and usefully,* Proc. Of the 4th International Workshop on Computational Learning Theory, p 69-79, Morgan Kaufmann Publishing, Inc., San-Mateo,
Roberts K., Horowitz D., 1986, *Basic level generalization in seven and nine-month old infants,* Journal of child language, 13, p 191-208,
Romanik K., 1992, *Approximate testing and learnability,* Proc. Of the 5th International Workshop on Computational Learning Theory, p 327-332, ACM Press,
Rosch E., 1973, *Natural categories,* Cognitive Psychology, 4 : 328-350,
Rosch E., 1976, *Classification des objets du monde réel : origines et représentations dans la cognition,* Bulletin de psychologie, numéro spécial *La mémoire sémantique* S.Ehrlich et E.Tulving (Eds) 307-313, 242-250,
Rosch E., 1978, *Principles of catégorization,* Cognition and categorization , Rosch E., Lloyd B.B, Erlbaum, Hillsdale (N.J.)
Rosch E., Mervis C.B., 1975, *Family resemblances : studies in the internal structures of categories,* Cognitive Psychology, 7 : 573-605,
Rosch E., Simpson C., Miller R.S., 1976, *Structural bases of typicality effects,* Journal of experimental psychology : human perception and performance, 2, 491-502,
Rosch-Heider E.R. , 1971, *Focal colors areas and the development of color names,* Developmental psychology, 4 , p 447-455,

Rosenblatt F., 1962, *Principles of neurodynamics*, New York, Spartan Books,
Rumelhart D.E., McClelland J.L.,1986, *Parallel distributed processing : exploration in micro-structure of cognition* (Vls 1 et 2), Cambridge M.A. :MIT Press
Salzberg S., Delcher A., Heath D., Kasif S., 1991, *Learning with a helpful teacher,* 12th International Conference on Artificial Intelligence,
Samuel A.L., 1959, *Somme studies in machine learning using the game of checkers,* IBM Journal of resarch and development 3, 211-229,
Satoh K., 1998, *Analysis of case-based representability of boolean functions by monotone theory,* , Lecture Notes in Computer Science, Vol. 1501, p. 179-196
Schofield , 1976, *The difficulty of using disjunctives hypotheses,* Journal of general psychology,
Schyns P.G., Rodet L., 1997, *Categorization creates functionnal features,* Journal of experimental psychology : learning, memory and cognition vol 23, n°3, p 681-696,
Seshu R., 1994, *Binary decision trees and an "average-case" model for concept learning : implications for feature construction and the study of bias,* Computational learning theory and natural learning systems, vol 1, ch 8,
Shapire R.E., 1990, *The strength of weak learnability,* Machine Learning, 5(2), p 197-227,
Shepard R., 1962, *The analysis of proximities : multidimensionnal scaling with an unknown distance function,* Psychometrika, 27, p 125-140,
Shepard R., 1964, *Attention and the metric structure of the stimulus space,* Journal of mathematical psychology, 1, p 54-87,
Shinohara A., Miyano S., 1991, *Teachability in computational learning,* New generation computing, 8, p 337-347,
Simon J., 1995, *Apprentissage naturel, apprentissage automatique et théories formelles de l'apprentissage,* Mémoire de DEA, LIFL de Lille, Lille
Simon J., Denis F. et Gilleron R., 1997, *Théories formelles de l'apprentissage, apprentissage automatique et apprentissage naturel,* Actes des premières journées francophones organisées par SCICOIA,
Smith E.E, 1990, *Categorization,* An invitation to cognitive science : thinking, Eds. Osherson D.N Smiths E.E.,. *Cambridge, MIT press,* 33-53
Smith E.E, Medin D.L., 1981, *Categories and concepts,* Cambridge, MA : Harvard University Press,
Smith et Osherson, 1989, *Similarity and decision making,* Similarity and analogical reasonning, S.Vosniadou et A.Ortony, Cambridge University Press, New York
Smolensky P., 1988, *On the proper treatment of connectionism,* Behavioral and brain sciences vol 11, p 1-74,
Sommers F., 1971, *Structural ontology,* Philosophia, Vol 1, p 79-85,
Stroop J.R., 1935, *Studies on interference un serial verbal reactions,* Journal of experimental psychology, 18, p 643-662,
Tellier I., 1998, *Apprentissage syntaxico-sémantique du langage naturel,* JFA'98,
Tesauro G., 1995, *Temporal difference learning and TD-Gammon,* Communication of the ACM, vol. 38, n°3, 58-68,
Thibaut, J.P., 1995, *The abstraction of relevant features by children and adults: the case of visual stimuli.* , Proceedings of the Seventeenth Annual Conference of the

Cognitive Science Society, pp. 194-199, J.D Moore & J.F. Lehman, Mahwah, NJ, Lawrence Erlb,
Thibaut, J.P. , 1997, *When children fail to learn new categories: the role of irrelevant features,* Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, pp. 745-750, In M.G. Shafto & P. Langley (Eds.) Hillsdale, NJ: Lawrence Erl,
Thibaut, J.P. , 1997, *Similarité et catégorisation.* , L'Année Psychologique, 97, 701-736.,
Thibaut, J.P. & Schyns, P.G. , 1995, *The development of feature spaces for similarity and categorization,* Psychologica Belgica, 35, 167-185.,
Tirri H., 1994, *Learning with instance-based encodings,* Computational learning theory and natural learning systems, vol 2, ch 13,
Tversky A., 1977, *Features of similarity,* Psychological review, 84, 327-352,
Tversky A., Kahneman D., 1983, *Extensional versus intuitive reasoning : the conjunction fallacy in probability judgement* , Psychology review, 90, 293-315,
Valiant L G, 1984, *A theory of the learnable,* Communication of the ACM, vol. 27, n° 11, 1134-1142,
Valiant L.G, 1985, *Learning disjunctions of conjunctions,* Proceedings of the 9thJoint Conference on Artificial Intelligence, 560-566,
Vapnik V.N., Chervonenkis A.Y., 1971, *On the uniform connvergence of relative frequencies of events to their probabilities,* Theory of probability and its applications, 16(2), p 264-280,
Varela F.J. , 1989, *Connaître les sciences cognitives- tendances et perspectives,* Seuil ,
Vignaux G., 1991, *Les sciences cognitives une introduction,* Edition La découverte, Paris
Vitgotsky, 1985, *Pensée et langage,* Terrains/Editions sociales, Paris
Weil-Barais A., 1993, *L'homme cognitif,* PUF, Paris
White T.G., 1982, *Naming practices, typicality and underextension in child language,* Journal of Experimental child psychology, 33, p 324-346,
Widrow B., Hoff M.E., 1960, *Adaptive switching circuits,* IRE WESCON Convention Record, 4, 96-104
Zhang J., 1992, *Selecting typical instances in instance-based learning,* Proceedings of the 9th International Machine Conference,