UNIVERSITE

DE

LILLE

-Φ-

FACULTE DES SCIENCES

-Φ-

L E S S Y S T E M E S

A

A U T O - A P P R E N T I S S A G E

V. CORDONNIER

UNIVERSITE DE LILLE

-0-

FACULTE DES SCIENCES

—0—

LES SYSTEMES A AUTO-APPRENTISSAGE

par

V. CORDONNIER.

Second sujet de thèse présenté en Mai 1968

PREMIERE PARTIE : LES SYSTEMES - DEFINITIONS ET PROPRIETES

INTRODUCTION.

L'importance que les systèmes de traitement et de contrôle ont prise dans tous les domaines de la science et de la technique masque la pauvreté intrinsèque qui caractérise leur fonctionnement. Les systèmes analogiques ou digitaux qui tendent à remplacer l'homme dans certaines de ses activités ne constituent qu'une imitation très approximative de son comportement.

Tels qu'ils sont, cependant, ces systèmes rendent d'importants services par la faculté dont ils disposent de travailler plus vite et avec plus de précision et de sécurité que l'homme. Leur emploi se limite à une classe bien définie de problèmes : ceux pour lesquels l'algorithme est parfaitement et totalement défini.

Il existe une grande variété de travaux qui ne peuvent être confiés aux systèmes automatiques, soit parce qu'ils sont incomplètement définis, soit parce qu'ils sont trop volumineux ; dans les deux cas, l'absence d'algorithme ou l'impossibilité pratique de le définir, empêchent l'utilisateur de profiter des avantages offerts par l'automatisme. Un exemple usuel de problème incomplètement défini est celui de la reconnaissance des structures -formes et sons- tandis que les problèmes de traduction des langues pour lesquels il existe certainement un algorithme, se heurtent à des questions de volume tant est vaste la gamme des cas que l'on peut rencontrer.

Il est donc certain que, si l'on parvient à doter certains systèmes de facultés nouvelles qui les apparenteraient mieux ou de plus près au comportement humain, le bénéfice que l'on en tirerait serait immense. Puisque la plus grande difficulté rencontrée pour élever le niveau de fonctionnement des dispositifs automatiques réside dans la définition de leur algorithme, il faut prévoir qu'ils puissent fonctionner correctement alors que leur algorithme est incomplètement défini ; mieux encore, il faut leur donner le moyen de perfectionner ou même de construire totalement cet algorithme.

L'auto-apprentissage apparaît donc comme un nouveau changement dans la hiérarchie des systèmes auxquels on ne confie plus seulement un travail à poursuivre ou une fonction à assurer mais auxquels on demande, de plus, d'apprendre par eux-mêmes la meilleure manière de le faire.

Comme le souligne STEINBUCH, cette évolution des systèmes automatiques ne peut être une simple imitation des systèmes biologiques qui constituent les meilleurs exemples d'organisations évoluées. Les moyens dont disposent ces derniers sont sans proportion avec ce qu'il est techniquement concevable de réaliser: nombre de cellules participant à l'élaboration d'une décision, nombre de positions de mémoire, richesse des moyens d'accès; d'autre part notre connaissance sur leur fonctionnement est limitée.(3) .

L'approche se fera donc simplement par la définition des buts à atteindre et non par celle d'un assemblage donné de cellules élémentaires dont les propriétés simuleraient le comportement des cellules biologiques. Cette dernière approche est tout aussi fondamentale mais n'a pas acquis, à l'heure actuelle, assez de maturité pour aborder l'étude des systèmes complets.

La définition des systèmes à auto-apprentissage, toujours de STEINBUCH, est la suivante : si une machine peut communiquer avec le milieu dans lequel se mesure la qualité de son travail, elle peut obtenir des informations permettant la modification de son action; elle acquiert de l'expérience : elle apprend .

Du simple point de vue de leur mode de fonctionnement, il est possible de classer les systèmes en quatre catégories :

A - Les systèmes rigides où l'opérateur fournit totalement l'algorithme; ce sont les servo-mécanismes classiques et les ordinateurs.

B - Les systèmes à recherche aléatoire de l'algorithme le mieux adapté. Le système reçoit uniquement des indications sur le travail à entreprendre et, sans méthode de recherche précise, essaye des solutions. Il va se produire un grand nombre d'actions inutiles, voire dangereuses, avant que l'optimum ne soit atteint.

C- Si on ajoute au système un simulateur représentant le milieu dans lequel il doit agir, il lui est possible de tester le résultat de ses actions sans passer par l'intermédiaire du milieu qu'il doit contrôler.

D - Si, enfin, on autorise le système à conserver des informations sur ses précédentes actions et à exploiter son expérience, on peut prévoir qu'il va perfectionner son algorithme et ne lui fournir qu'un simulateur imparfait.

LES SYSTEMES A AUTO-APPRENTISSAGE.

Une fois admis que les systèmes que l'on entend doter de facultés supplémentaires de décision doivent être capable d'acquérir de l'expérience, il faut préciser ce qu'ils doivent être capables d'apprendre et la manière dont ils le font. De plus, il faut prévoir comment ils font usage de cette éxpérience pour modifier leur comportement.

Ce qui caractèrise un système à auto-apprentissage, ce ne sont ni les entrées, ni les sorties qui se présentent sous la même forme que dans un système rigide. L'aspect particulier de ces système n'apparait qu'au niveau des relations qui s'établissent entre ces entrées et ces sorties. Ces relations ne sont pas figées et la nécéssité de les faire évoluer incite à choisir, de préférence, des modèles continus pour lesquels la valeur des ajustements progressifs est arbitraire.

Toute optimisation, dans quelque domaine que ce soit, doit se faire par rapport à un critère. Un système ne peut s'améliorer que s'il est jugé par un autre système extérieur à lui et qui sait ce qu'il y a lieu de faire dans chacun des cas rencontrés. La nécéssité évidente d'un moniteur appelle cependant deux remarques :

- Il n'est pas nécéssaire que le moniteur sache parfaitement ce qu'il y a lieu de faire; ses connaissances des relations à établir entre l'entrée et la sortie peuvent être limitées; les ordres ou indications qu'il fournit peuvent être entachés d'erreur. Une expérience faite à ce propos sur le PARAMETRON (27) dans laquelle on a volontairement introduit un pourcentage relativement d'erreurs de la part du moniteur, a mis ce fait en évidence. Le système qui apprend ne peut juger de la valeur de chacun des apports que lui fait le moniteur car il serait alors plus compétent que lui. Pour se perfectionner, il doit déterminer la valeur moyenne des indications, parfois contradictoires qu'il reçoit.

- La seconde remarque concerne la structure du système. Il est évident qu'il est impossible d'apprendre n'importe quoi à n'importe quel système, si qualifié que soit le moniteur. Lorsque l'on conçoit un système destiné à se perfectionner par auto-apprentissage, il faut lui donner la faculté de s'adapter à tous les cas qu'il est susceptible de rencontrer; autrement dit, il est absolument nécessaire que tous ces cas aient été virtuellement prévus, même si on l'a fait sous une forme synthétique et incomplète.

La présence obligatoire d'un moniteur plus compétent que le système suivant le travail d'un concepteur travaillant sur un cahier des charges précis font dire que le terme d'auto-apprentissage est ambigu. Dans la mesure où il sous-entend que le système assure lui-même sa propre formation par le simple contact avec le milieu dans lequel doit porter son action, le terme prête à confusion. Il serait plus normal de parler de systèmes à apprentissage dirigé.

LE FONCTIONNEMENT DES SYSTEMES A AUTO-APPRENTISSAGE.

Il existe une indétermination sur la structure définitive d'un système lorsque, soit les entrées sont imparfaitement définies, soit leur effet sur les sorties ne peut être totalement déterminé. Dans un cas comme dans l'autre, l'apprentissage doit créer ou perfectionner des relations entre les entrées et les sorties et donc associer à certaines configurations ou valeurs des unes, d'autres configurations ou valeurs des autres. La presque totalité des systèmes à auto-apprentissage comporte des sorties booléennes; le problème consiste donc d'abord à reconnaitre dans la carte des entrées quelles sont les zones qui agissent sur telle ou telle des sorties booléennes : Le fonctionnement de ce système sera donc basé sur un principe de reconnaissance de structures. Au dispositif ainsi réalisé on adjoint une possibilité de faire évoluer les frontières séparant les différentes zones dans un sens favorable sur la simple présentation par le moniteur du succès rencontré. Le rôle du moniteur pourra se limiter à la simple expression de son accord ou de son désaccord ou, mieux, à fournir une mesure de la qualité du travail fourni.

Si les études sur les systèmes à auto-apprentissage évoluaient pour aborder les machines délivrant des actions continues, fonctions ajustables des entrées, il semble que l'identification de ces dernières resterait l'opération fondamentale. Au lieu de plaquer sur la carte des entrées un découpage quantifié, on chercherait à lui superposer une carte des sorties, elle aussi capable d'évoluer à l'aide de processus d'apprentissage.

En raison de l'importance des processus d'identification, la seconde partie de ce travail leur est totalement réservée. Dans la troisième partie, on examinera alors ce qui représente réellement l'apprentissage. Une dernière partie est consacrée aux réalisations pratiques actuelles.

SECONDE PARTIE : LES PROCESSUS D'IDENTIFICATION

Un système capable de classer des objets après les avoir identifiés comporte autant de sorties ou d'états distincts communicables à l'extérieur qu'il doit discerner de classes différentes parmi les éléments qui lui sont présentés. Habituellement un seul objet est présent à la fois devant l'organe d'entrée du système et donc, une seule de ces sorties sera active. Toutes les autres seront au repos. Il en résulte que l'expression du choix est booléenne.

Au niveau de la décision, par contre, les objets n'étant pas totalement identifiables, le traitement ne peut être booléen. Il faut donc assurer une conversion. L'opérateur qui l'assure sera toujours un détecteur de maximum qui, recevant à son entrée un certain nombre N de fonctions, ne présentera à sa sortie formée, elle aussi de N voies, qu'une voie active, celle qui correspond à l'entrée ayant la valeur la plus élevée. Cet opérateur dont la réalisation pratique est relativement simple s'appelle l'unité de réponse.

Si les entrées sont P_1, P_2, ... P_N, et les voies de sortie R_1, R_2, ... R_N, l'unité de réponse fournit le résultat suivant:

$R_j = 1$ si $(P_J - P_i)$ positif pour $i = 1, 2, \ldots N$; $i \neq j$
$R_J = 0$ pour tous les autres termes.

La question de l'identification se ramène alors à l'élaboration des fonctions P_j en respectant la règle suivante; Si l'objet présenté appartient à la classe j et si donc il doit provoquer l'activation de la sortie R_j, la fonction P_j doit être supérieure à toutes les autres.

A - IDENTIFICATION PAR COMPARAISON A UN MODELE.

Les premières tentatives pour édifier une théorie des processus de reconnaissance sont naturellement partie de l'idée de comparer l'objet présenté à un modèle. Ce modèle serait la valeur moyenne des paramétres qui caractèrisent tous les objets appartenant à la classe à identifier. La mesure de la somme des différences qui existent entre l'objet et tous les modèles existants dans la "mémoire" du système doit permettre la création de la fonction P.

L'objet inconnu et les modèles étant définis par des fonctions de n variables, HIGHLEYMAN propose d'en mesurer la ressemblance par le calcul de la fonction d'intercorrélation (9).

Dans le cas de fonctions de deux variables x et y, si $I(x,y)$ est la fonction représentant l'objet et $M_j(x,y)$ $j = 1,\ldots N$, les fonctions des modèles, la fonction d'inter-correlation s'écrit :

$$\varphi_j(\sigma,\rho) = \frac{\int_x \int_y I(x+\sigma, y+\rho)\, M_j(x,y)\, dx\, dy}{\left[\int_x \int_y I^2(x,y)\, dx\, dy \int_x \int_y M_j(x,y)\, dx\, dy\right]^{1/2}} \qquad [2.1]$$

Il est possible de simplifier l'expression en remplaçant $M_j(x,y)$ par la fonction normalisée $M'_j(x,y)$:

$$M'_j(x,y) = \frac{M_j(x,y)}{\left[\int_x \int_y M_j^2(x,y)\, dx\, dy\right]^{1/2}}$$

Alors : $$\int_x \int_y M'_j(x,y)\, dx\, dy = 1$$

Comme on ne cherche pas à connaitre la valeur de la fonction de corrélation mais seulement les valeurs relatives de chacun des termes φ_j, on peut rendre le dénominateur égal à 1. Il ne dépend plus que de l'objet et reste donc le même dans chacune des expressions :

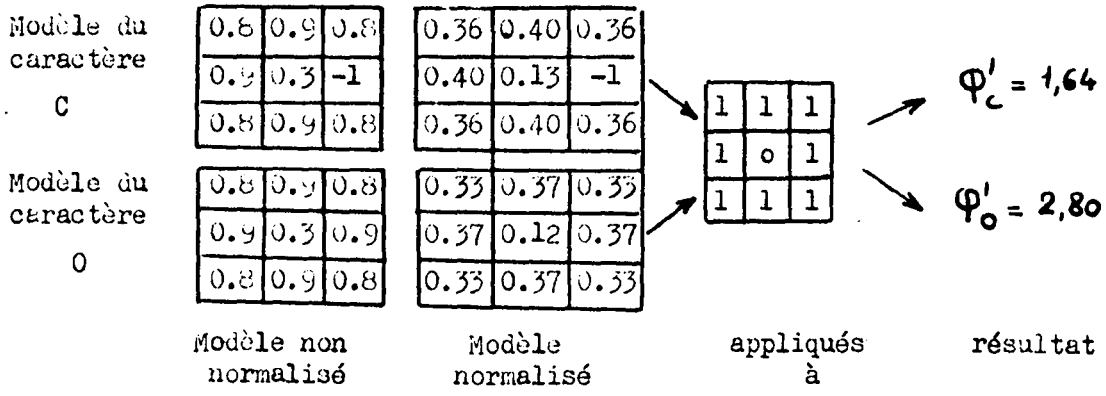
$$\varphi'_j(\sigma,\rho) = \int_x \int_y I(x+\sigma, y+\rho)\, M'_j(x,y)\, dx\, dy \qquad [2.2]$$

Pour appliquer la méthode, il faut quantifier d'une part l'espace utilisable dans le plan et d'autre part les valeurs des fonctions. En supposant que la surface utile est un rectangle de m sur n positions, la formule [2.2] devient :

$$\varphi'_j(\sigma,\rho) = \sum_{k=1}^{m} \sum_{l=1}^{n} I_{(k+\sigma)(l+\rho)}\, M'_{j\,k,l} \qquad [2.3]$$

On peut alors, pour renforcer la discrimination, déformer le modèle en lui apportant des"pénalités". L'auteur appelle ainsi des modifications des coefficients $M_{j,k,l}$ pour les positions du rectangle où la probabilité d'apparition d'une valeur élevée de $I(x,y)$ est très faible.

L'exemple de la page suivante illustre le rôle des coefficients pénalisés dans le cas de la discrimination des caractères C et O.



La méthode de comparaison à un modèle présente les avantages et les inconvénients suivants:

- Si on désire chercher un coefficient de corrélation pour toutes les valeurs possibles de σ et ρ depuis 1 à m ou n, on doit disposer d'un algorithme très lourd mais on peut tenir compte de toute translation verticale ou horizontale de l'objet par rapport au modèle.
- En contre-partie, la position de l'objet dans son cadre n'entre plus en ligne de compte et il s'introduit des risques de confusion. (le caractère 1 et les parenthèses, par exemple)
- La simulation sur machine du processus est relativement aisée mais la réalisation pratique en serait très difficile.

B - EMPLOI DES FONCTIONS DISCRIMINANTES.

L'objet, dans un espace quantifié, peut être représenté par un groupe de d termes x_1, x_2, ... x_d . Cet objet sera défini par un point X dans un espace E^d à d dimensions appelé : espace objet.

Le classement des objets consiste à établir dans l'espace E^d, une carte des domaines renfermant les points associés à chaque classe de discrimination.

Les surfaces qui séparent les domaines s'appellent : les surfaces de décision . Les domaines tels que, à l'intérieur de chacun, tous les points X soient considérés comme représentant des objets appartenant à une classe donnée sont appelés les zones de décision.

Si un point X se trouve exactement sur une surface de décision, le classement de l'objet correspondant est impossible; les surfaces de décision conduisent à un choix indéterminé.

Cette méthode, utilisée par de nombreux auteurs, (10), (11), (12) est à la fois un cas particulier de la méthode des modèles et une extension de celle-ci dans une autre direction. Elle particularise la méthode de comparaison avec pénalités en ce sens qu'elle ne peut s'appliquer qu'à un nombre fini,d,de points et pas à des fonctions continues; elle constitue par ailleurs une généralisation car elle peut s'appliquer à des cas non linéaires comme on le verra par la suite.

Soient N fonctions : $g_1(X)$, $g_2(X)$,..... $g_N(X)$. Chaque fonction est caractèristique de l'une des N classes de sortie.

On choisira ces fonctions de telle sorte que pour tout point X de E^d appartenant à la classe de réponse R_j :

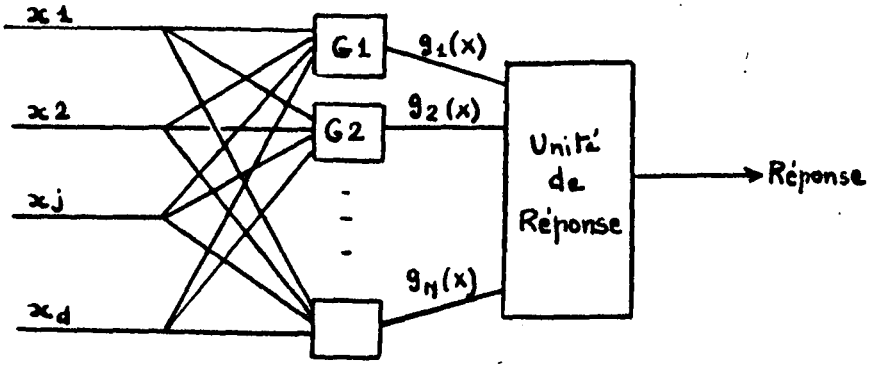
$$g_i(X) < g_j(X) \quad \text{pour } i, j = 1, 2, \ldots R \qquad i \neq j$$

En supposant les fonctions g continues, les surfaces de décision sont définies par :

$$g_i(X) - g_j(X) = 0$$

On a montré (15) que toute fonction non décroissante, appliquée aux fonctions g, redonne des fonctions g' équivalentes.

On peut, à l'aide des fonctions g, appelées <u>fonctions discriminantes</u> établir le schéma de principe d'un classificateur:



Dans le cas où N = 2, il n'existe que deux fonctions discriminantes; leur différence indique par son signe, la classe à laquelle appartient la réponse. En posant $g(X) = g_1(X) - g_2(X)$, on déduit :

- Si $g(X) > 0$, l'objet appartient à la classe 1.
- Si $g(X) < 0$, l'objet appartient à la classe 2.
- Si $g(X) = 0$, l'objet est indéterminé.

On dit que l'élément qui réalise cette fonction particulière fonctionne par dichotomie.

On simplifie la détermination des fonctions discriminantes en leur imposant d'appartenir à certaines catégories déterminées. Si une fonction dépend de m paramétres w_1, w_2, ... w_m,

$$g(X) = g(X, W_1, w_2, ... w_m) ,$$

on particularise la fonction en imposant certaines conditions sur les paramètres.

Les fonctions discriminantes les plus employées sont :

- Les fonctions linéaires;
- Les fonctions linéaires par parties;
- Les fonctions quadratiques;
- Les fonctions polynomiales encore appelées fonctions Φ qui représentent une généralisation des fonctions quadratiques.

C - LES FONCTIONS DISCRIMINANTES LINEAIRES.

On s'impose de n'employer que des fonctions de la forme :

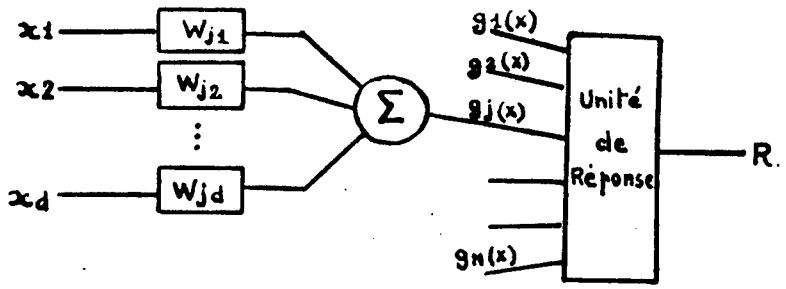
$$g(X) = w_1x_1 + w_2x_2 + ... + w_dx_d + w_{d+1}$$

Les surfaces de décision sont alors des hyperplans de l'espace E^d. Si deux domaines i et j sont contigus, l'équation du plan qui les sépare est :

$$(w_{i1}-w_{j1})x_1 + (w_{i2}-w_{j2})x_2 + ... + (w_{id}-w_{jd})x_d = 0 \qquad i \neq j$$

Il existe $R(R-1)/2$ équations telles que celle-ci donc autant de surfaces de décision. En pratique, on peut éliminer les surfaces définies sur deux domaines non contigus qui sont redondantes.

La machine linéaire correspondant à cette définition comporte, dans un premier niveau, des circuits de pondération qui affectent à chaque entrée et pour chaque fonction, un poids ou coefficient w_{ij}, dans un second niveau, un sommateur qui assure le calcul de la fonction $g_j(X)$. Enfin une unité de réponse choisit la fonction $g_j(X)$ ayant la valeur la plus élevée.



Un cas particulièrement simple de fonction discriminante linéaire est celui de la classification par la distance minimum. Soit un groupe de N points P_1, P_2, ... P_N considérés comme représentatifs de l'optimum ou de la valeur moyenne de chacune des classes à classer.

Pour un objet inconnu, X, la distance à chacun de ces points est :

$$X - P_i = \sqrt{(X - P_i)(X - P_i)}$$

Il suffit de classer les points en fonction de leurs distances aux divers points P_i et de choisir comme bonne la réponse correspondant à la distance minimum.

Comme toute fonction non décroissante de g(X) est équivalente à g(X), on prendra comme fonction le carré de cette distance comme fonction discriminante.

$$|X - P_i|^2 = (X - P_i)(X - P_i) = X.X - 2.P_i.X + P_i.P_i$$

Pour un point X donné, le produit X.X est constant; on peut le soustraire de toutes les fonctions discriminantes :

$$g_i(X) = P_iX - 1/2\ (Pi.Pi)$$

Les coefficients w_{ij} sont les coordonnées du point P_i et la constante $w_{i,d+1}$ vaut $1/2(P_i.P_i)$

Fonctions discriminantes linéaires par segments

Supposons, en reprenant la classification par la distance minimum, qu'il existe N groupes de points notés $(P_1),(P_2),\ldots (P_N)$ considérés comme caractéristiques, chacun pour la région de l'espace qu'il occupe, de la classe à identifier.

Si les points de la catégorie i sont :$P_{i,1}$, $P_{i,2}$, ... P_{i,n_i}, on appelle distance de X à (P_i), la valeur :

$$d(X/(P_i)) = \text{minimum de } \left\{ X - P_{i,j} \right\} \quad \text{pour } j = 1,2,\ldots n_i$$

Il suffit alors de reprendre alors la méthode de la distance minimum avec des discriminants linéaires. En fait, l'opération revient à appliquer deux fois le critère de la distance minimum, une première fois entre tous les points caractéristiques d'une même classe puis une seconde fois entre les distances minimum trouvées;

$$g_i(X) = \max \left\{ P_{i,j}.X - 1/2(P_{i,j}.P_{i,j}) \right\} \quad j = 1,2,\ldots n_i$$

Ce qui ne semble pas avoir été noté, c'est que le même résultat peut être obtenu de manière plus économique en considérant tous les points comme différents et en réalisant des unions logiques au niveau des réponses.

D - LES FONCTIONS DISCRIMINANTES NON LINEAIRES.

1) Les fonctions discriminantes quadratiques.

Ce sont des fonctions de la forme :

$$g_i(X) = \sum_{j=1}^{d} w_{jj} x_j^2 + \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} w_{jk} x_j x_k + \sum_{j=1}^{d} w_j x_j + w_{d+1}$$

Le nombre de poids à définir est égal à : $(d + 1).(d + 2)/2$

Les surfaces de décision dans l'espace E^d sont des hyper-sphères, hyperboloïdes, élipsoïdes ou cylindres.

Théorème fondamental (2).

Un système défini par des fonctions discriminantes quadratiques peut se ramener à un système linéaire précédé d'un transformateur quadratique. La dimension de l'espace dans lequel travaille le système linéaire correspondant est: $d' = d(d + 3)/2$

Soit X' un vecteur de $E^{d'}$ dont les composantes sont:

$$X'_1, X'_2, \ldots\ldots X'_{d'}$$

En posant que : les d premiers vecteurs $x_1^2, x_2^2, \ldots x_d^2$ sont $X'_1 \ldots X'_d$

les $d(d - 1)/2$ $x_i x_j$ sont $X'_{d+1} \ldots X'_{d(d+1)/2}$

les d derniers xj sont $X'_{d'-d} \ldots\ldots X'_{d'}$

On peut écrire :

$$g_i(X') = w_{11}X'_1 + \ldots + w_{dd}X'_d + w_{12}X'_{d+1} + \ldots + w_{d-1,d}X'_{d(d+1)/2} + \ldots$$
$$+ w_1X'_{d'-d} + \ldots + w_{d'}X'_{d'} + w_{d+1}$$

2) Les fonctions polynomiales.

Les fonctions quadratiques sont caractérisées par le fait que les poids $w_1, w_2, \ldots\ldots w_{d'}$ apparaissent linéairement. On généralise cette propriété en appelant fonction Φ toute fonction des paramètres w_1, w_2, etc. et du vecteur X linéaire par rapport à ces paramètres.

Les fonctions polynomiales d'ordre r dans l'espace E^d sont des fonctions où il apparait d+1 paramètres et telles que les coefficients de ces paramètres de poids soient de la forme:

$$x_{k1}^{n1} \cdot x_{k2}^{n2} \cdot x_{k3}^{n3} \ldots\ldots x_{kr}^{nr}$$

$k = 1, 2, \ldots\ldots d$

$n1, n2, \ldots\ldots nr = 1$ ou 0

On notera que les fonctions discriminantes, linéaires par segments ne sont pas des fonctions Φ et que les remarques qui suivent ne les concernent pas.

E - EFFICACITE DES FONCTIONS DISCRIMINANTES.

Si on considère M points dans un espace E^d, il existe 2^M classifications de ces M points en deux groupes. Posons L(M,d) le nombre de dichotomies réalisables entre ces points par une fonction discriminante g(X), linéaire.

Exemple

soit M = 4, d = 2

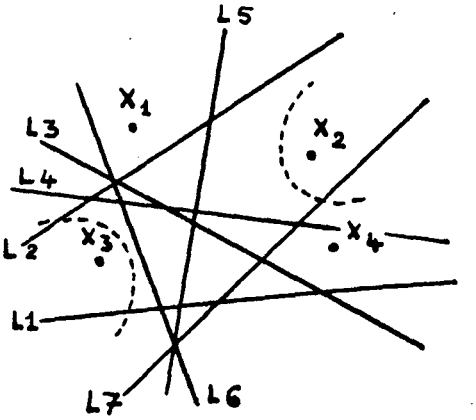
On peut tracer 7 lignes de séparation :

L_1 $X_1 X_2 X_3 X_4$ | $X_1 X_2 X_3 X_4$

L_2 X_1 | $X_2 X_3 X_4$

$X_2 X_3 X_4$ | X_1

.....



On trouve, au total L(M,d) = 2.7 = 14

Pour établir la valeur de L(M,d) on peut raisonner par récurrence: Dans le cas de M-1 points, il existe L(M-1,d) dichotomies possibles; si on ajoute un point supplémentaire, on retrouve les L(M-1,d) dichotomies précédentes plus celles que l'on peut réaliser entre le point nouveau et tous les autres, projetés sur un hyper-plan de E^d.

$$L(M, d) = L(M-1, d) + L(M-1, d-1)$$

Les valeurs initiales sont:

$$L(1, d) = 2 \quad \text{et} \quad L(M, 1) = 2M$$

Pour des fonctions polynomiales dans E^d, la correspondance avec une fonction discriminante linéaire dans $E^{d'}$ permet par la détermination de d' de déterminer le nombre de dichotomies réalisables.

Ainsi pour une fonction discriminante quadratique : $d' = d.(d+3)/2$

$$L(M, d') = L(M, d(d+3)/2)$$

On touve une efficacité supérieure à la fonction linéaire.

L'efficacité d'une machine ou de la fonction discriminante associée étant la probabilité pour que, une dichotomie étant choisie au hasard, elle puisse être réalisée est :

$$E = \frac{L(M, d)}{2^M}$$

d est la dimension de l'espace dans lequel la fonction discriminante peut être considérée comme linéaire.

TROISIEME PARTIE : LES METHODES D'APPRENTISSAGE.

L'apprentissage est possible lorsque le système a la faculté de procéder lui-même à des réajustements des fonctions discriminantes qu'il utilise en vue de perfectionner ses propres critères de classement. Il procéde à ces corrections lorsqu'il en reçoit l'ordre de l'extérieur. Cet ordre peut être soit une commande par tout ou rien indiquant si la réponse est considérée comme bonne ou mauvaise, soit une commande plus détaillée indiquant le degré de satisfaction du moniteur. D'autre part les ordres d'apprentissage peuvent être fournis au système pendant une certaine période dite " d'apprentissage " au cours de laquelle on présente à l'entrée des objets d'entrainement; pendant cette période le système perfectionne ses critères; par la suite, on le suppose apte à identifier correctement les objets qu'on lui présentera. Des solutions plus perfectionnées prévoient un apprentissage permanent du système de telle sorte qu'il reste toujours capable de s'adapter à une évolution des objets qu'on lui a présentés, à l'apparition de nouveaux objets ou à une modification des règles à respecter.

On distingue deux grandes familles de méthodes d'apprentissage: Les méthodes paramétriques et les méthodes non paramétriques. Dans les méthodes paramètriques, on suppose que les classes d'objets sont caractérisées par un groupe de paramètres dont la valeur peut être inconnue au départ mais dont la connaissance se développe progressivement et permet l'établissement des fonctions discriminantes. Dans les méthodes non paramètriques, rien ne caractérise les classes les unes par rapport aux autres; seules les indications du moniteur et la connaissance des raisons qui, à l'intérieur du système, ont motivé le choix, permet de décider des ajustements à faire. Les méthodes paramétriques sont habituellement plus simples à formuler mais la réalisation exige la mémorisation sous une forme quelconque, des paramètres. Dans les procédures d'apprentissage non paramétriques, il suffit au système, pour déterminer l'ajustement à réaliser de connaitre les valeurs actuelles de son choix d'une part, des commentaires du miniteur d'autre part. Réclamant moins d'informations, il est vraisemblable que les systèmes non paramétriques ne se perfectionneront, toutes choses égales par ailleurs, que plus lentement que les systèmes paramétriques.

A - METHODES D'ENTRAINEMENT PARAMETRIQUE.

Le paramètre le plus couramment employé est la probabilité pour qu'un objet appartienne à une classe donnée. Soit $p(X \mid i)$ la probabilité pour que X étant donné, il appartienne à la catégorie i. Soit encore $p(i)$, la probabilité d'apparition des objets de la classe i. Ces deux probabilités sont inconnues au début de la séquence d'entrainement.

La méthode consiste :

1) A exprimer les fonctions discriminantes en les définissant a partir de $p(X \mid i)$ et $p(i)$.
2) A préciser les valeurs des paramètres $p(X \mid i)$ et $p(i)$ au cours de la séquence d'entrainement.

On définit une fonction "perte" $r(i \mid j)$ pour i et j = 1, 2, .. N dans le cas de N classes à classer, exprimant la perte ressentie lorsqu'un objet appartenant à la catègorie j est classé par le système dans la catégorie i. Pour un certain dégré d'apprentissage imparfait, on peut exprimer, pour chacun des objets présentés, un taux moyen de pertes:

$$R_X(i) = \sum_{j=1}^{N} r(i \mid j)p(j \mid i) \qquad [3.1]$$

lorsque le système décide de classer l'objet en i

Le travail d'entrainement consiste à minimiser R_X .

Si pour X donné, $R_X(i_d) \quad R_X(i) \quad i = 1, 2, \ldots N$ on minimise $R_X(i_d)$ en assignant X à la catégorie i_d.

On procède alors de la manière suivante:

1) L'objet est présenté au système.
2) Le système procède au calcul de $R_X(i)$ pour toutes les catègories i = 1, 2, ... N
3) Le système choisit la catègorie i_d qui minimise $R_X(i)$

Emploi du critère de la plus grande vraisemblance.

En posant $p(X \mid j)$ Probabilité d'apparition de l'objet X étant donné qu'il appartient à la catégorie j.

$p(j)$ Probabilité d'apparition à priori d'un objet appartenant à la catégorie j.

$p(X)$ probabilité d'apparition du vecteur X représentant l'objet X.

$p(j \mid X)$ probabilité que X appartienne à la catégorie j.

le théorème de BAYES permet d'écrire :

$$p(j \mid X) = \frac{p(X \mid j).p(j)}{p(X)} \qquad [3.2]$$

En substituant [3.2] dans [3.1] , il vient :

$$R_X(i) = \frac{1}{p(X)} \sum_{j=1}^{N} r(i \mid j).p(X \mid j).p(j) \qquad [3.3]$$

La fonction de perte étant fixée par l'utilisateur, le système va poursuivre son apprentissage en calculant $p(j)$ et $p(X \mid j)$ à partir des indications exactes qui lui seront fournies lors de la présentation de chacun des objets X de la séquence d'entrainement.

Supposons que la fonction de perte soit $r(i \mid j) = 1 - \delta_{ij}$, ce qui revient à dire que la perte est égale à 1 en cas d'erreur et à 0 si la réponse est bonne.

$$R_X(i) = \frac{1}{p(X)} \quad p(X) - p(X \mid i).p(i)$$

La fonction est minimum pour $p(X \mid i).p(i)$ maximum; si, de plus la distribution des objets est homogène, $p(i) = 1/N$ et il suffit alors de chercher un maximum de $p(X \mid i)$ et de choisir la catègorie correspondante.

Exemple 1

Soit un classificateur dichotomique ayant à reconnaitre des objets dont les termes d'entrée sont soit 1, soit 0. Si on appelle 1 et 2 les reponses possibles:

$$r(i \mid j) = 1 - \delta_{ij}$$

$$r(1 \mid 1) = r(2 \mid 2) = 0$$

$$r(1 \mid 2) = r(2 \mid 1) = 1$$

La fonction discrimiante, équation de la surface séparant les domaines correspondant aux deux catégories dans E^d est :

$$g(X) = g_1(X) - g_2(X) = p(X \mid 1).p(1) - p(X \mid 2).p(2)$$

Aux fonctions $g_1(X)$ et $g_2(X)$, on peut substituer les fonctions croissantes:

$$g'_1(X) = \text{Log } p(X \mid 1) + \text{Log } p(1)$$

$$g'_2(X) = \text{Log } p(X \mid 2) + \text{Log } p(2)$$

La fonction discriminante devient:

$$g(X) = \text{Log } p(X \mid 1) + \text{Log } p(1) - \text{Log } p(X \mid 2) - \text{Log } p(2)$$

$$g(X) = \text{Log}\left[\frac{p(X \mid 1)}{p(X \mid 2)}\right] + \text{Log}\left[\frac{p(1)}{p(2)}\right] = \text{Log}\left[\frac{p(X \mid 1)}{p(X \mid 2}\right] + \text{Log}\left[\frac{p(1)}{1-p(1)}\right]$$

On peut exprimer les lois $p(X \mid 1)$ et $p(X \mid 2)$ en fonction des probabilités rapportées aux coordonnées du vecteur X. Ceci est possible en raison du caractère binaire des coordonnées.

$$p(X \mid j) = p(x_1 \mid j).p(x_2 \mid j) \ldots\ldots p(x_d \mid j) \qquad j = 1 \text{ ou } 2 .$$

$$g(X) = \sum_{i=1}^{d} \text{Log} \frac{p(x_i \mid 1)}{p(x_i \mid 2)} + \text{Log} \frac{p(1)}{1 - p(1)}$$

En posant : $p(x_i=1 \mid 1) = p_i \qquad p(x_i=0 \mid 1) = 1 - p_i$

$$p(x_i=1 \mid 2) = q_i \qquad p(x_i=0 \mid 2) = 1 - q_i$$

$$g(X) = \sum_{i=1}^{d} x_i . \text{Log} \frac{p_i(1 - q_i)}{q_i(1 - p_i)} + \sum_{i=1}^{d} \text{Log} \frac{1-pi}{1-q_i} + \text{Log} \frac{p(1)}{1-p(1)}$$

équation discriminante linéaire.

Il faut apprendre cette équation discriminante en formant dans la mémoire les valeurs de p_i et q_i c'est à dire 2d termes différents.

Exemple 2

Considérons le cas fréquent où les probabilités pour un vecteur X d'appartenir à une catégorie i sont distribuées selon une loi de GAUSS autour de la valeur optimale.

Dans le cas d'un vecteur X à deux coordonnées x_1 et x_2, on définit la distribution des probabilités par :

$$\left.\begin{array}{ll} m_1 = E(x_1) & m_2 = E(x_2) \\ \sigma_{11} = E(x_1^2) - E^2(x_1) & \\ \sigma_{22} = E(x_2^2) - E^2(x_2) & \end{array}\right\} \quad \begin{array}{l} \text{valeurs moyennes} \\ \text{eccarts type} \end{array}$$

En adoptant une notation vectorielle:

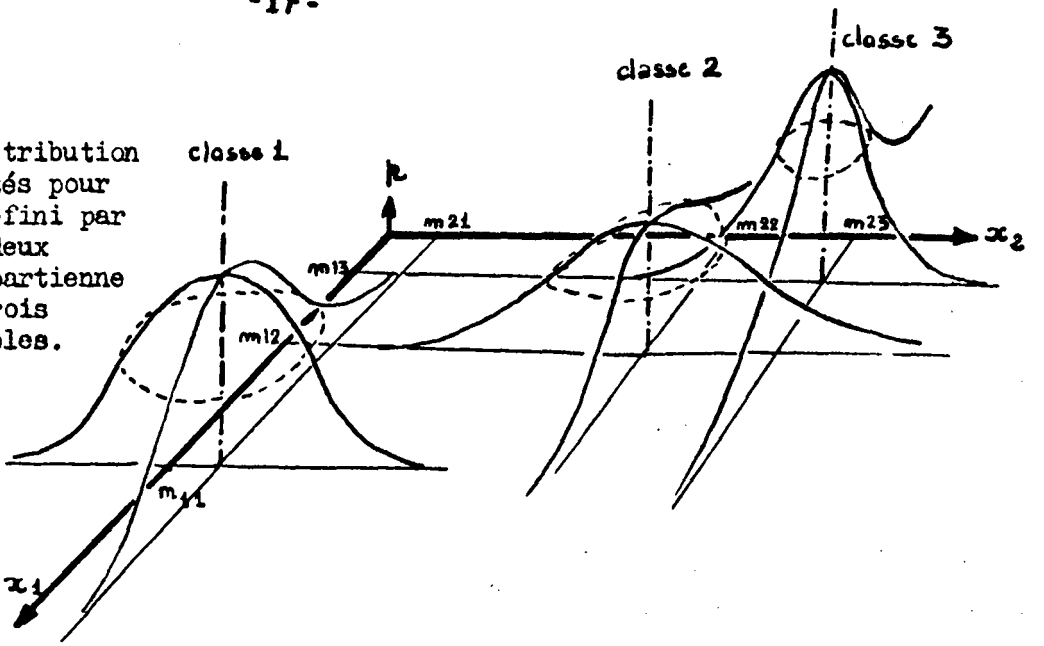
$$M = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}, \left|\sum\right| = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}$$

la loi de probabilité s'écrit :

$$p(X) = \frac{1}{2\pi \left|\sum\right|^{1/2}} \cdot e^{-\frac{1}{2}\left[(X - M)^{tr} \sum^{-1} (X - M)\right]}$$

cette loi se généralise à un espace à plus de deux dimensions. S'il existe N classes à identifier, chacune se caractérise par une distribution probabiliste dont la valeur est d'autant plus élevée que le vecteur X est plus près du point de coordonnées $(m_1, m_2 \ldots m_d)_i$ $i = 1,2,\ldots,N$.

Exemple de distribution des probabilités pour qu'un objet défini par un vecteur à deux coordonnée appartienne à l'une des trois classes possibles.



La probabilité pour qu'un vecteur X dans E^d appartienne à la classe i s'exprime par :

$$p(X \mid i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \; e^{-\frac{1}{2}\left[(X - M_i)^{tr.}\, \Sigma_i^{-1}\, (X - M_i)\right]}$$

On peut prendre cette expression comme fonction discriminante $g_i(X)$ mais il est plus simple de passer au logarithme :

$$g_i(X) = \mathrm{Log}(p(X \mid i)) + \mathrm{Log}(p_i)$$

$$g(X) = \mathrm{Log}\, p_i - \frac{d}{2}\mathrm{Log}(2\pi) - \frac{1}{2}\mathrm{Log}|\Sigma_i| - \frac{1}{2}\left[(X - M_i)^{tr}\, \Sigma_i^{-1}\, (X - M_i)\right]$$

Le premier terme est indépendant de X et le second est constant; tous deux peuvent être supprimés sans modifier les valeurs relatives des fonctions discriminantes :

$$g_i(X) = W_{i,d+1} - \frac{1}{2}\left[(X - M_i)^{tr}\, \Sigma_i^{-1} (X - M_i)\right]$$

Il s'agit d'une fonction quadratique.

L'entrainement consiste à apprendre au système à identifier les M_i et Σ_i ; Si la séquence d'entrainement comporte k_i objets appartenant à la classe i, on prendra :

$$M_i = \frac{1}{k_i} \sum_{1}^{k_i} X \quad \text{appartenant à la classe i}$$

$$\Sigma_i = \frac{1}{k_i} \sum_{1}^{k_i} (X - M_i)\,(X - M_i)^{tr.}$$

B - METHODES D'ENTRAINEMENT NON PARAMETRIQUES.

Si aucun paramètre ne caractérise la distribution des objets et de leur vecteur associé dans l'espace E^d, les seuls éléments dont on puisse disposer pour apporter des corrections au système sont les fonctions discriminantes telles qu'elles existent et la mesure de leur imperfection apportée par le moniteur dans chaque cas particulier.

Considérons, en premier lieu, le cas du classificateur dichotomique qui doit séparer deux classes d'objets notées 1 et 2. Soit (X_1) un groupe d'objets appartenant à la classe d'entrainement 1 et (X_2) un groupe d'objets appartenant à la classe d'entrainement 2. Le groupe $(X) = (X_1) + (X_2)$ s'appelle la séquence d'entrainement.

Dans l'espace E^d, l'hyper-plan qui sépare les deux classes est défini par:

$$g(X) = w_1x_1 + w_2x_2 + \ldots\ldots + w_dx_d + w_{d+1} = 0$$

Si cet hyper-plan n'est pas correctement disposé, la méthode consiste, au vu des erreurs, à apporter des corrections à sa position en apportant des corrections aux coefficients de poids $w_1, w_2, \ldots w_{d+1}$

Afin de rendre équivalents les deux espaces dans lesquels sont portés le vecteur objet X et le vecteur poids W, on ajoute une composante x_{d+1} au vecteur X. Cette composante est égale à 1.

Le produit scallaire des vecteurs

$$Y = (X, x_{d+1}) \text{ dans l'espace } E^d$$

et $$W = (w_1, w_2, \ldots\ldots w_d, w_{d+1}) \text{ dans l'espace } E^W$$

représente la fonction discriminante. L'hyper-plan de dédision est défini par :

$$g(Y) = W \cdot Y = 0$$

Au groupe d'entrainement (X_1) correspond le groupe (Y_1) et au groupe (X_2) correspond le groupe (Y_2) .

La solution est obtenue si :

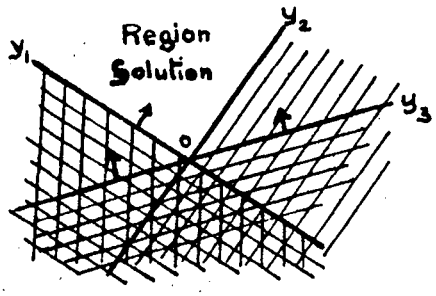
pour Y élément de (Y_1), Y . W est positif.

pour Y élément de (Y_2), Y . W est négatif.

Pour tous les vecteurs de (Y), séquence d'entrainement, on peut définir des hyper-plans qui respectent cette condition. Considérons alors l'espace E^W des poids, dual de l'espace E^d des objets. Pour un objet donné, il existe un hyper-plan séparant le demi-espace pour lequel la réponse est correcte du demi-espace où elle est fausse.

Si pour tous les vecteurs Y de la séquence d'entrainement, il exis- une région de l'espace E^W pour laquelle la réponse est toujours bonne, tout vecteur W choisi à l'intérieur de cette région sera considéré comme répondant à la condition posée.

On remarquera que le vecteur $W_0 = (0, 0, ... 0)$ répond toujours à la condition $Y.W = 0$. Le point origine de l'espace E^W se trouve sur la surface de décision de tous les cas envisagés donc tous les hyper-plans définis dans E^W passent par l'origine.



Il en résulte aussi que tout produit d'un vecteur W considéré comme correct, par un scalaire ne modifie pas les propriétés de ce vecteur.

On trouve, de plus une condition pour que la séquence d'objet (Y) soit séparable linéairement: Il faut qu'il existe dans l'espace E^W, un hyper-cone pour lequel tout vecteur W intérieur au cone fournit une réponse correcte en présence de chacun des objets de la séquence d'entrainement.

Procédure d'entrainement.

Si pour Y élément de (Y_1), le système fournit une réponse fausse : Y.W négatif dont l'objet est classé dans la catégorie 2 ou si, inversement pour Y élément de (Y_2) , Y.W est positif, cela veut dire que, dans l'es-pace E^W, le point W est du mauvais coté du plan défini pour l'objet Y en question.

Il faut déplacer le point W vers le plan Y. Le trajet le plus court étant la normale dirigée depuis W vers le plan défini par le vecteur Y. On déplace W en lui ajoutant un vecteur Y dont les cosinus directeurs sont ceux de la normale au plan. Comme il n'y a pas de métrique commune à l'espace E^d (espace ou est défini Y) et l'espace E^W (où est défini le point W) il n'est pas certain que la correction soit suffisante.

Le nouveau point des poids réajustés sera :

$$W' = W + Y$$

ou, d'une manière plus générale $W' = W + c.Y$

La règle pratique sera donc la suivante:

si la réponse est correcte : $W' = W$

sinon si Y est attribué à 2
alors qu'il appartient à 1 : $W' = W + c.Y$

Si Y est attribué à 1
alors qu'il appartient à 2 : $W' = W - c.Y$

Cette méthode est dite méthode d'entrainement à correction d'erreur.

Il existe plusieurs types de procédures de correction selon la valeur que l'on attribue au terme c.

a) Si c est une constante, la correction peut ou non apporter la modification souhaitée. Dans tous les cas, il est certain que cette modification se fera dans le sens convenable et qu'il suffit d'en appliquer un nombre fini pour obtenir le résultat cherché.

b) Pour un point W placé sur le plan Y, dans l'espace E^W, le produit W.Y est nul. De manière plus générale, ce produit exprime la mesure de la distance entre le point W à déplacer et le plan Y. Pour effectuer une correction définitive, il faut déplacer le point W en un point W' tel que:

- W'.Y = (W + c.Y) devienne positif si le produit W.Y était négatif ou nul par erreur.
- W'.Y =(W + c.Y) devienne négatif si le produit W.Y était positif ou nul par erreur.

La valeur de c assurant cette modification est à calculer dans chaque cas. On choisira pour c une valeur telle que :

$$c \gg \frac{W.Y}{Y.Y}$$

Cette régle est dite de correction absolue.

c) On peut enfin choisir c de telle sorte que la correction soit proportionnelle à la distance qui sépare le point W du plan Y. Cette régle est dite de la correction proportionnelle.

L'ancienne distance était proportionnelle à W.Y, la nouvelle sera proportionnelle à W'.Y; on choisira c pour que la correction apportée soit une fraction constante de W.Y.

Soit k la valeur de ce rapport; il vient :

$$k = \frac{W'.Y - W.Y}{W.Y}$$

$$\text{et} \quad c = k \frac{W.Y}{Y.Y}$$

Si k = 1, la correction conduit à placer W sur le plan Y.

Si K est inférieur à 1, le point W va tendre assymptotiquement vers le plan mais il faudra un nombre infini de correction pour ne parvenir qu'à une indétermination.

Si k est supérieur à 1, la correction est assurée en une seule opération. Une méthode souvent citée est celle qui prend k = 2 c'est à dire qui prend W' symétrique de W par rapport au plan Y.

Généralisation de la méthode pour des classes multiples

Si un système doit classer les objets Y qui lui sont présentés en N catégories, on lui présentera, pendant la période d'entrainement, une séquence (Y) formée des groupes $(Y_1),(Y_2) \ldots (Y_N)$.

Il existe N fonctions discriminantes de la forme :

$$g_i(X) = \sum w_{i,j}\, y_j \qquad i = 1, 2, \ldots N$$

Le choix du vecteur W initial étant arbitraire, pour chaque fonction discriminante, la méthode d'apprentissage doit prévoir de les traiter toutes.

Dans la pratique, le système ne pouvant se baser que sur le choix qu'il a fait et sur les indications du moniteur relatives à ce choix, seules deux fonctions sont concernées. Si le système a classé un objet de la catégorie i dans la catégorie k, on ne peut apporter de corrections qu'aux fonctions discriminantes $g_i(X)$ et $g_k(X)$. On déplacera le vecteur W_k de manière à le faire passer du coté du plan Y tel que le produit $W_k.Y$ diminue tandis que le vecteur W_i sera déplacé de manière à augmenter la valeur du produit $W_i.Y$.

Les deux opérations sont indépendantes et on peut considérer que la seconde est seule suffisante.

$$W_i' = W_i + c.Y$$
$$W_k' = W_k + c.Y$$

Cette méthode s'applique à tous les systèmes linéaires et, par extension, à tous les sytèmes dont les fonctions discriminantes sont des fonction polynomiales.

C - METHODE D'ENTRAINEMENT POUR DES FONCTIONS NON LINEAIRES.

Les méthodes décrites pour entrainer par des procédures non paramétriques ne sont applicables qu'à des fonctions discriminantes linéaires où pouvant s'y ramener. AMARI a proposé une méthode générale basée sur les probabilités mais non paramétrique, susceptible de s'appliquer à n'importe quelle forme de fonction.(25)

La connaissance de la loi statistique de distribution des classes permettrait de déterminer la fonction discriminante optimale pour chaque classe d'objets.

On peut définir une évolution qui ne se base que sur l'erreur actuelle et montrer que l'on parvient à une optimisation statistique des fonctions discriminantes.

Considérons un classificateur devant identifier les objets qui lui sont présentés comme appartenant à l'une des N classes C_1, C_2, ... C_N. Les fonctions discriminantes correspondantes sont $g_i(X)$. Les paramètres qui interviennent dans ces fonctions sont t_{i1}, t_{i2}, ...t_{ij} ... t_{ik}. On représente ces paramètres par un vecteur T_i.

Pour les valeurs actuelles de ces paramétres, on peut exprimer la fonction discriminante de la classe i par $g_i(X,T_i)$. Appelons T le vecteur de l'espace à k.N dimensions représentant l'ensemble des paramètres t_{ij} $i = 1, 2 \ldots N$ et $j = 1, 2, \ldots k$.

Les décisions qui seront prises par le classificateur sont entièrement déterminées par le vecteur T. L'apprentissage consiste à déterminer T afin qu'il n'entraine pas d'erreurs de classification.

Dans le cas du classificateur parfait, on peut écrire :

Si X appartient à la classe i,

$g_i(X,T)$ est supérieur à $g_j(X,T)$ $\quad j=1,2,..N \quad j \neq i$

Dans le cas contraire, soit E_i l'ensemble des classes j telles que $g_j(X,T)$ soit supérieur à $g_i(X,T)$.

Posons :

$$d_i(X,T) = \sum_{j \in E_i} s_{ij}(g_j(X,T_j) - g(X,T_i))$$

La quantité d_{ij} représente la somme pondérée par un coefficient s_{ij} des erreurs ou plus précisément, des différences existant entre la fonction discriminante qui devrait être maximale et celles qui se trouvent, en fait, être supérieure à elle.

La perte résultante sera:

$$p_i(X,T) = p(d_i(X,T))$$

Il est évident que si le vecteur T est le vecteur de décision optimale, E_i, d_i et p_i s'annulent.

Si on appelle p(i) : probabilité d'apparition d'un objet de la classe i et $p_i(X)$ la densité de probabilité des objets de la classe i, le risque moyen s'exprime par :

$$R(T) = \sum_{i=1}^{N} \int_{j \neq i} p(i).p_i(X).p(X,T)\, dX$$

Le vecteur T qui optimise la décision est donc défini par le risque minimum soit:

$$\text{grad}\ (\ R(T)\ ) = 0$$

Pour ne pas avoir à mémoriser tous les états antérieurs, la modification apportée au vecteur T_i ne doit se déterminer que sur sa valeur actuelle et le vecteur X_i.

Si $T_{i,n}$ représente le vecteur T_i à l'instant n, la régle d'apprentissage proposée est :

$$T_{i,n+1} = T_{i,n} + \delta T$$

avec δT défini par:

$$\delta T = \varepsilon C. H_i(X,T)$$

La fonction H_i est appelée la fonction d'apprentissage.

C représente une matrice positive.

ε est une constante, petite, positive dite constante d'apprentissage.

En prenant pour La fonction $H_i(X,T)$, le gradient de la fonction perte changé de signe :

$$H_i(X,T) = - \nabla p_i(X,T)$$

Il est possible de démontrer que la valeur moyenne de δT notée $\overline{\delta T}$ est:

$$\overline{\delta T} = - \varepsilon C \text{grad } R(T)$$

d'où l'on déduit:

$$\overline{\delta R} = - \varepsilon \text{grad} R^{tr} C.\text{grad} R$$: nécessairement négatif.

Cette méthode est dite d'optimisation statistique.

Il est possible d'effectuer le calcul de la vitesse de convergence et de la précision obtenues. On trouve que le choix de la matrice C détermine l'uniformité et l'isotropie des évolutions du vecteur T donc de la convergence vers l'état optimum.

La constante ε agit sur la vitesse de convergence : plus elle est grande, plus cette convergence est rapide mais ceci est obtenu au détriment de la précision.

Cette dernière remarque permet d'introduire dans un système des taux variables d'apprentissage; lorsque le vecteur T est loin de sa valeur optimale, on peut choisir relativement grand. Au contraire, lorsque T est près de sa valeur optimale, il arrivera souvent que deux corrections successives agissent dans des sens opposés; il est alors souhaitable de réduire la longueur de $\overline{\delta T}$. On peut apprendre au système une régle de modification de $\varepsilon.C$ basée sur la mesure de la valeur moyenne des corrections qu'il apporte. Ce que l'auteur appelle " l'apprentissage de l'apprentissage" permet de concilier vitesse et résolution.

QUATRIEME PARTIE : LES REALISATIONS PRATIQUES.

Il faut distinguer deux catégories de réalisations; les unes, très nombreuses sont des simulations sur ordinateurs ou, plus rarement sur calculateur analogique, des relations établies entre les objets présentés et les décisions à prendre. Les autres représentent des réalisations effectives à l'aide de circuits qui suivent et obéissent à ces relations.

Les programmes de simulation.

Plus de la moitié des travaux théoriques cités dans la bibliographie trouvent dans une simulation sur ordinateur une justification et un moyen de vérification. Il serait impossible de les décrire tous et il a semblé plus significatif de reprendre les termes d'une publication récente (29) qui, pour répondre à un problème particulier de reconnaissance et d'identification, fait le bilan des différentes méthodes possibles.

1) Apprentissage forcé.

$$\Delta w_i(j) = \frac{1}{m} X_i(j)$$ (Voir page 20 paragraphe a)

2) Correction probabiliste des poids (méthode dite de Bayes)

Il n'y a pas de correction apportée au vecteur des poids mais un recalcul permanent :

$$W_i = \text{Log} \frac{p_i}{1 - p_i} - \text{Log} \frac{q_i}{1 - q_i}$$

3) Correction d'erreur

$\Delta w_i = c. X_i(j)$ si $X(j)$ est mal classé et seulement dans de cas.

Cette méthode a été employée avec C constant.

4) Correction quadratique moyenne.

$$\Delta w_i = \frac{X_i(j)}{n} (1 \pm g(X))$$

On renforce (+) ou on diminue (-) le vecteur poids dans le sens du vecteur X la fonction qui aurait du être maximum ou celle qui a pris sa place.

5) Correction itérative.

On emploie une fonction de perte exponentielle et les corrections s'expriment sous forme d'une fonction de forme variable dans laquelle intervient la perte moyenne.

6) Madaline

Il s'agit d'une méthode dérivée de Adaline (26) mais dans laquelle on introduit deux niveaux, le premier analogique, le second digital et majoritaire.

Les objets présentés étaient des photographies de la surface lunaire dans lesquelles les systèmes devaient reconnaitre :

A - Des cratères ou des chaines circulaires.

B - Des cratères avec un sommet au centre.

C - Des chaines de montagnes alignées.

Les résultats ont été les suivants :

objets	A	B	C
pour 1000 objets par catégorie dans la séquence d'entrainement.			
Méthodes			
1	78,5	82.6	81.2
2	82.6	84.1	82.6
3	99.2	92.8	100
4	99.5	92.1	100
5	83.2	81.8	91.4
6	87.4	100	95.8

Les différences de performances ne sont pas importantes mais il parait utile de relever que les méthodes 1) et 2) entrainaient un temps d'apprentissage de l'ordre de 20 minutes sur un ordinateur moyen alors que les quatre autres réclamaient de 8 à 12 heures.

Les systèmes cablés

Il n'a été réalisé que peu de systèmes réels à auto-apprentissage pour la raison suivante : la technique de leur emploi est encore trop récente pour qu'on puisse l'envisager dans des applications réelles. Les réalisations ne peuvent être que des prototypes de laboratoire mais alors la concurence des ordinateurs est très vive. Une autre raison souvent avancée est l'inadaptation flagrante des technologies actuelles au problème posé. La cellule de base d'un système à auto-apprentissage est celle qui contient le poids w_i résultant de l'entrainement. L'information contenue doit être analogique, facilement modifiable mais doit, en l'absence de toute sollicitation extérieure, conserver exactement sa valeur.

Dans le PERCEPTRON (27) construit en 1960, on emploie des fonctions discriminantes linéaires que l'on fait évoluer par la méthode des corrections d'erreur soit constantes soit proportionnelles à l'éccart.

La mémorisation des poids est assurée par des potentiomètres et la modification de ces poids se fait par des moteurs agissant directement sur l'axe des potentiomètres. Des relais assurent la transmission des ordres émis par le moniteur au vu des réponses fournies.

Dans ADALINE (26), réalisation un peu postérieure au PERCEPTRON, la mémoire est assurée par des cuves électrolytiques. La résistance des électrodes est proportionnelle à la quantité de métal qu'elles ont reçu ou cédé. L'apprentissage consiste à faire passer un courant entre deux électrodes pour assurer les déplacements désirés du métal.

On a proposé, par la suite, un certain nombre de technologies possibles en particulier l'emploi d'éléments magnétiques à cycle d'hystérésis rectangulaire (30) mais aucune de ces mémoires analogiques ne peut rivaliser avec les mémoires pour informations quantifiées c'est à dire celle que l'on trouve dans les ordinateurs. Personne n'a, à notre connaissance, envisagé de joindre de telles mémoires à une logique de calcul elle aussi quantifié pour en faire un ordinateur spécialisé alors que les ordinateurs universels conviennent tout autant.

Une réalisation récente (28) basée sur une technologie mixte mais principalement digitale reprend une technique de corrélation. Le vecteur d'entrée, formé de 15 grandeurs binaires, est comparé à une collection de vecteurs types. Un sytème analogique choisit la catégorie qui ressemble le plus à celle qui a été présentée. Les caractéristiques les plus originales de cette réalisation, mis à part le fait qu'on a employé une technologie digitale, sont premièrement une transformation codée que subit le vecteur d'entrée avant d'être traité, deuxièmement, l'emploi d'un élément de décision à trois niveaux: L'objet appartient à la catégorie, l'objet n'appartient pas à la catégorie et enfin, l'objet est trop mal défini, il y a indétermination.

Il est possible que, dans les années qui viennent, de nouvelles technologies permettent la réalisation effective de systèmes à auto-apprentissage plus évolués et surtout plus volumineux que ceux qui ont été fabriqués jusqu'à présent. L'essort, que l'on prévoit considérable, des circuits intégrés analogiques sera peut-être le signal attendu par les laboratoires d'application pour mettre en pratique les outils des théoriciens et pour répondre aux besoins croissants qui se font jour dans tous les domaines pour de tels systèmes.

BIBLIOGRAPHIE

Une bibliographie faite en 1962 (1) fait, déjà à cette époque état de près de 500 publications sur les systèmes dont le comportement se rapproche de celui des organismes biologiques. Entre les années 1963 et 1967, le rythme des publications a considérablement diminué; par contre, au cours de cette même période sont apparus de nombreux matériels utilisant les méthodes qui avaient été mises au point; presque toutes les réalisations pratiques portaient sur la reconnaissance automatique des formes et principalement de l'écriture.

Actuellement, il semble que le sujet reprenne de l'actualité; cela provient vraisemblablement de l'apparition de nouvelles familles de besoins tels que la reconnaissance du langage, l'identification des personnes, la réduction des informations transmises par les satellites, etc. Il semble d'autre part que la technologie ait fait, depuis quelques années des progrès suffisants pour permettre la réalisation de projets nettement plus ambitieux.

GENERALITES SUR L'APPRENTISSAGE.

1 - The simulation of cognotive processes. An annotated bibliography. P.L. SIMMONS,R.F.SIMMONS - IEEE Transactions on electronic computers - Septembre 1961 et Aout 1962

2 - Learning machines - N.J.NILSON - McGraw Hill; New York; 1965 .

3 - Introductory speech on adaptative systems - STEINBUCH - Congrès de l'UNESCO sur le traitement de l'information - PARIS 1959 .

4 - Step to artificial intelligence - MINSKY - P.I.R.E. Janvier 1961.

5 - Threshold logic - DERTOUZOS - M.I.T. press Cambridge 1965

6 - Principles of neuro-dynamics - ROSENBLATT - Spratan books 1962

7 _ Pattern recognizing control systems - B.WIDROW, F.W.SMITH. Spartan books 1964 .

8 - Self-organizing systems - HAWKINS - PIRE Janvier 1961

9 - An analog method for pattern recognition - HIGHLEYMAN - IEEE transactions on electronic computers Septembre 1961

METHODES D'ENTRAINEMENT PARAMETRIQUES.

10 - Linear decision functions with application to pattern recognition - HIGHLEYMAN - PIRE Juin 1962 .

11 - Learning matrices and their applications - STEINBUCK, PISKE IEEE transactions on electronic computers Decembre 1963 ;

12 - Threshold logic in artificial intelligence - WINDER - IEEE publication on artificial intelligence . 1963

13 - Learning to recognize patterns in a random environment. ABRAMSON - Transactions of IEEE on information theory? Sept. 1962.

14 - The characteristic selection problem in recognition systems - LEWIS IEEE transactions on information theory Fevrier 1962

15 - Generalisation and information storage in network of Adaline. WIDROW - Self-organizing systems - Spartan books 1962.

16 - The use of adaptative threshold element to design a linear optimal pattern classifier. J.S.KOFORD, G.F.GRONER - Transactions of IEEE on information theory. Janvier 1966.

17 - Nearest neighbour pattern classification T.M.COVER , P.E.HART Transactions of IEEE on information theory. Janvier 1967.

18 - Sequential pattern recognition - K.S.FU, Y.T.CHIEN, G.P.CARDILLO- IEEE transactions on electronic computers. DEcembre 1967 .

19 - Statistical recognition functions and the design of pattern recognizers. T.MARILL, D.M.GREEN - PIRE decembre 1960 .

METHODES D'ENTRAINEMENT NON PARAMETRIQUES.

20 - Pattern recognition with an adaptative network L.R.ROBERTS. IRE international convention record. Mars 1960.

21 - On predicting perceptron performances. R.D.JOSEPH. IRE international convention record. Mars 1960.

22 - An empirical Bayes approach to non-parametric two-way classification. M.V.JOHNS . Studies on item analysis and prediction. Stanford university press. 1961 .

23 - Test on a cellule assembly theory of the action of brain. IEEE transactions on information theory. ROCHESTER . juin 1956.

24 - Logic systems with generalizing properties. RIDWAY . Stanford electronic laboratories technical repport. 1962 .

25 - A theory of adaptative classifiers. S. AMARI. IEEE transactions on electronic computers.

REALISATION PRATIQUES.

26 - Application of adaptative data processing systems. WIDROW Wescon convention paper 1963.

27 - The Mark 1 Perceptron - design and performances - J.C. HAY, F.C.MARTIN, C.W.WIGHTMAN. IRE international convention record Mars 1960.

28 - Machines looks, listens and learns. G.L.CLAPPER . Electronics Octobre 1967

29 - Pattern recognition from satellite altitudes E.M.DARLING, R.D.JOSEPH - IEEE transactions on systems sciences and cybernetics. Mars 1968.

30 - A magnetic intégrator for perceptron program. J.K.HAWKINS Wescon convention record on circuit theory . Mars 1960.