Numéro d'ordre : 2381

THESE

Présentée à

L'Université des Sciences et Technologies de Lille

en vue de l'obtention du grade de

DOCTEUR DE L'UNIVERSITE

Spécialité : Automatique et Informatique Industrielle

par

Christophe DESROUSSEAUX

Ingénieur I.S.E.N.

UTILISATION D'UN CRITERE ENTROPIQUE DANS LES SYSTEMES DE DETECTION

Soutenue publiquement le 9 décembre 1998 devant la commission d'examen :

Président :	J. RAGOT	Professeur à l'Institut National Polytechnique de Lorraine
Rapporteurs :	B. DUBUISSON	Professeur à l'Université de Technologie de Compiègne
	H. MAITRE	Professeur à l'ENST de Paris
Co-directeurs :	M. STAROSWIECKI	Professeur à l'Université Lille I
	D. POMORSKI	Maître de conférences à l'Université Lille I
Examinateurs :	I. NIKIFOROV	Professeur à l'Université de Technologie de Troyes
	M. LABARRERE	Directeur adjoint de l'ONERA-CERT de Toulouse

REMERCIEMENTS

Le travail présenté dans ce mémoire a été effectué au Laboratoire d'Automatique et d'Informatique Industrielle de Lille (LAIL) dans l'équipe Analyse et Surveillance des Processus Industriels Complexes (ASPIC) sous la direction de Monsieur M. Staroswiecki, et de Monsieur D. Pomorski, respectivement Professeur et Maître de Conférences à l'Université des Sciences et Technologies de Lille. Je tiens à les remercier très vivement pour l'accueil, l'encadrement et les précieux conseils dont j'ai bénéficié tout au long de ce travail.

Je suis très reconnaissant à Monsieur B. Dubuisson, Professeur à l'université de Technologie de Compiègne, et à Monsieur H. Maitre, Professeur à l'E.N.S.T. de Paris pour l'honneur qu'ils me font en examinant ce travail et en acceptant d'être les rapporteurs de cette thèse.

Je tiens également à remercier, Monsieur M. Labarrere, Directeur adjoint de l'ONERA-CRT de Toulouse, Monsieur I. Nikiforov, Professeur à l'Université de Technologie de Troyes et Monsieur J. Ragot, Professeur à l'Institut National Polytechnique de Lorraine, pour l'honneur qu'ils me font en examinant ce travail et en acceptant de participer à mon jury de thèse.

J'adresse également une pensée particulière à tous les membres de l'équipe ASPIC, et à tous les thésards pour leur sympathie et leur bonne humeur.

Je voudrais dédier ce travail à mes parents. Même si il ne semble pas leur être directement lié, ma dette envers eux est totale. Je tiens à les remercier, ici, pour le soutien qu'ils m'ont apporté tout au long de mes études.

Enfin, je tiens à remercier Céline, qui tout au long de cette thèse, m'a si souvent supporté et encouragé.

à mes parents,

à Céline,

SOMMAIRE

INTRODUCTION GENERALE

Au cours des décennies, l'évolution des technologies a permis une amélioration des équipements industriels. L'intégration de calculateurs a favorisé le développement d'algorithmes très puissants spécialisés dans la commande mais aussi dans le traitement numérique des données. Cette évolution s'est traduite par une augmentation de la complexité des installations : capteurs et actionneurs nombreux, géographiquement répartis, modes de fonctionnement multiples, interaction importante entre les opérateurs et le processus.

Dans ce contexte, nous nous trouvons souvent confrontés à un problème de décision, il nous faut choisir une ligne de conduite parmi plusieurs alternatives. Tel est le cas dans le contexte de la détection radar où une décision doit être élaborée en fonction de l'absence ou de la présence d'une cible. Dans les systèmes de communication numériques, une information codée est envoyée, celle-ci est perturbée par le canal de transmission et il s'agit à la réception de reconnaître le symbole utile qui est arrivé. En médecine, à partir d'une analyse de tissus, il faut déterminer si la tumeur opérée est cancéreuse ou non. En reconnaissance de formes, il faut par exemple reconnaître, à partir d'une photographie aérienne, quel type d'avion est basé à tel ou tel endroit. Dans toutes ces applications, le problème commun est celui de la prise de décision parmi plusieurs choix possibles. En statistique, ce type de problème est connu sous le nom de « théorie statistique de la décision ». Dans le contexte de la théorie de la communication, il est connu sous le nom de « théorie de la détection ».

C'est dans ce cadre que se situe ce travail, qui sera composé de trois parties :

- 1ère partie : Détection
- 2ème partie : Apprentissage
- 3ème partie : Apprendre afin de mieux détecter

1ère partie : Détection

Cette partie fait l'objet des deux premiers chapitres.

Chapitre 1...

Ce premier chapitre présente l'état de l'art de la détection :
L'intérêt de la « théorie de la détection » est de développer des méthodes permettant de prendre une décision, optimale selon un critère choisi, à partir d'un ensemble d'observations numériques. Il s'agit de discriminer statistiquement un nombre fini de situations. Dans le cadre très courant de la détection binaire, deux situations appelées « hypothèses » et notées H_0 et H_1 sont à discriminer. Ce problème est parfois dénommé problème du sonar ou du radar. L'hypothèse H_0 représente généralement la transmission du symbole « zéro » ou l'absence de cible. L'hypothèse H_1 correspond, quant à elle, à la transmission du symbole « un » ou la présence de la cible. En surveillance, la détection consiste à déterminer l'état de fonctionnement dans lequel se trouve le système. En général, il nous faut décider entre deux hypothèses H_0 (fonctionnement normal) et H_1 (fonctionnement anormal du système). Plus

généralement, il peut y avoir de multiples hypothèses, alors notées $H_0,H_1,...,H_M$. Nous ne considérons dans ce chapitre que le problème de la détection binaire.

La détection d'un signal à partir de plusieurs capteurs peut être envisagée de deux manières différentes. Traditionnellement, l'ensemble des capteurs communiquent leurs observations directement au détecteur central où la décision finale est prise. Cette approche, appelée « détection centralisée », nécessite souvent des lignes de communication à large bande si l'on veut obtenir une décision en temps réel. La deuxième approche, appelée « détection décentralisée », consiste à associer à chaque capteur un détecteur qui décide localement si un signal a été détecté ou non. Les décisions locales sont ensuite envoyées à un opérateur de fusion qui les combine pour prendre la décision finale. L'avantage de cette méthode est de réduire les coûts de communication. En contrepartie les performances obtenues sont moins bonnes puisque l'opérateur de fusion ne reçoit pas toute l'information nécessaire à la prise de décision finale.
Les trois architectures les plus courantes seront étudiées : la détection centralisée, et la détection décentralisée parallèle et série.

L'architecture du système de détection étant donnée, le problème est alors de déterminer une « stratégie de décision » optimale suivant un critère donné. Deux critères sont principalement utilisés : le critère de Bayes et le critère de Neyman-Pearson.

Chapitre 2...

L'optimisation des systèmes de détection centralisée et décentralisée repose sur l'établissement d'un critère. Dans le cas Bayésien, un coût est associé à chaque situation, une fonction « risque moyen » est ensuite minimisée. Dans les applications où ces coûts sont connus et ont une signification précise, l'approche Bayésienne peut être une excellente solution au problème d'optimisation. Cependant, ce n'est pas forcément le cas pour toutes les applications. Dans certaines d'entre elles, il pourrait être plus avantageux de se poser le problème de l'optimisation en s'intéressant à la quantité d'information qui peut être transmise à l'intérieur du système. Ce type d'approche pourrait notamment s'adapter aux problèmes de communication numériques où l'on s'intéresse davantage à la quantité d'information transmise, plutôt qu'à l'information elle-même. Pour de tels systèmes un critère basé sur une fonction entropique pourrait être plus approprié.

Après avoir schématisé le fonctionnement d'un système de détection binaire ainsi que le problème de la transmission d'informations binaires à travers un canal de transmission, notre objectif sera de comparer ces deux approches afin de faire émerger les avantages de chacune d'elles.

Dans le cadre du problème de la détection, deux hypothèses H_0 et H_1 doivent être discriminées. En d'autres termes, une décision (δ_0 ou δ_1) doit être prise à partir d'une observation donnée. Les probabilités de non détection $1\text{-}P_D$ et de fausse alarme P_F peuvent s'interpréter comme des probabilités d'erreur si l'on se place dans un problème de transmission d'informations binaires.

L'objectif est alors de minimiser la perte d'information sur H (H_0 ou H_1) connaissant la décision δ (δ_0 ou δ_1). Cette perte d'information peut être mesurée par l'entropie conditionnelle de Shannon $H(H/\delta)$, qui représente en fait l'incertitude sur H connaissant la décision δ.

Dans un premier temps, nous utiliserons ce critère afin d'optimiser une architecture de détection centralisée. Dans un deuxième temps, cette démarche sera étendue au cas de la

détection décentralisée parallèle, puis série. Enfin, une comparaison des résultats obtenus en utilisant la démarche classique et la démarche entropique sera entreprise.

2ème partie : Apprentissage

Cette partie fait l'objet des chapitres 3 et 4. Nous avons volontairement déconnecté cette partie de la partie traitant de la détection.

Chapitre 3...

Dans un premier temps, nous présentons le cadre général de la classification automatique afin d'introduire l'apprentissage supervisé (à base d'exemples) qui nous intéresse tout particulièrement.

Dans ce cadre, nous présentons la structure des données recueillies sur un système plus ou moins complexe. Ces données proviennent en général de capteurs qui, par définition, quantifient un signal (continu ou pas). Nous pourrons dès lors introduire la notion de finesse des variables représentant l'information fournie par les capteurs.
Nous introduisons également la notion d'incohérence des données d'apprentissage provenant d'un bruit se superposant aux données, ou provenant d'un manque d'explication dû au fait que l'expert n'aurait pas pris en compte tout le pouvoir explicatif de son système.

Enfin, pour terminer ce chapitre, nous introduisons les méthodes de construction d'arbres de décision qui permettent de traiter des variables qualitatives, mais également numériques. Ces méthodes visent à l'optimisation d'un critère global afin de discriminer les observations en différentes classes. Dans ce cadre, nous présentons l'algorithme C4.5, qui est certainement l'algorithme de construction d'arbres de décision le plus usité à l'heure actuelle. Nous discuterons alors de ses points forts, comme de ses points faibles pour lesquels nous proposerons des alternatives.

Chapitre 4...

Après un bref historique de la théorie de l'information, l'entropie de Shannon sera présentée en mettant en évidence ses propriétés les plus intéressantes. Des indices issus de la théorie de l'information seront construits, et pourront dès lors être utilisables dans des algorithmes de recherche d'informations, ou dans des algorithmes de classification.

Par définition, les outils de la théorie de l'information s'appliquent à des systèmes discrets, et nous nous proposons de les étendre à des systèmes continus.

Enfin, nous mettrons en avant l'intérêt de l'utilisation de critères entropiques dans le cadre de la théorie de la détection.

3ème partie : Apprendre afin de mieux détecter

Cette partie est composée du chapitre 5 et a pour but de fusionner la partie 1 et la partie 2.

Chapitre 5...

Le problème de la détection décentralisée reste encore aujourd'hui un problème complexe et multiforme. Le but recherché est la mise en commun de données collectées par N capteurs suivant une architecture donnée. Dans ce but, deux architectures ont été étudiées : la détection décentralisée parallèle, et la détection décentralisée série. L'optimisation de ces différentes architectures a permis de déterminer les opérateurs de traitement locaux permettant d'obtenir les meilleures performances de détection suivant un critère donné. Ces optimisations aboutissent à un difficile problème de résolution de systèmes d'équations. Dans le cas de la détection décentralisée parallèle, que l'on se place dans le cas Bayésien ou de Neyman-Pearson, l'optimisation d'un système comprenant N capteurs aboutit à un système de 2^N+N équations non linéaires couplées à résoudre. Dans le cas de la détection décentralisée série, on aboutit à un système de 2N-1 équations non linéaires couplées à résoudre. Ces systèmes n'ont pour l'instant pu être résolus que pour des cas particuliers en supposant par exemple l'indépendance des observations locales et pour des systèmes comportant peu de capteurs. De façon générale, on s'aperçoit que le nombre d'équations à résoudre simultanément croît très rapidement avec le nombre de capteurs, les calculs nécessaires à la résolution de ces équations deviennent alors très vite inextricables.

Afin de simplifier le problème d'optimisation de ces systèmes, nous proposons de limiter le nombre de capteurs à prendre en compte lors de l'optimisation du système de détection. Considérons par exemple le contexte de la surveillance d'installations industrielles complexes, où un grand nombre de capteurs observant des grandeurs physiques différentes est disponible. L'optimisation d'un système de détection décentralisée utilisant tous les capteurs s'avère très vite impossible à réaliser. Cependant, Il est possible que l'on puisse implémenter une structure de détection en ne considérant qu'un sous-ensemble des capteurs, plutôt que l'ensemble des informations disponibles. Dans certains cas, ceux-ci peuvent en effet présenter des redondances, sans améliorer les performances de l'ensemble, ou noyer le système sous un flot d'informations trop coûteux à gérer. C'est pour cette raison que nous proposons d'introduire dans les systèmes de détection décentralisée, avant toute optimisation, une étape de sélection de capteurs. Parmi tous les capteurs disponibles, nous proposons de ne faire intervenir que ceux apportant beaucoup d'information au processus de décision. Dans ce but, nous utilisons une phase d'apprentissage inspirée des problèmes de classification développés dans le chapitre 3. Nous proposons différents algorithmes de sélection basés sur le critère entropique introduit dans les chapitres 2 et 4 et qui est tout à fait adapté à ce problème de sélection.

La sélection de capteurs étant faite, on pourra optimiser le système en utilisant les résultats classiques de la théorie de la détection exposés dans le premier chapitre, ou en utilisant un critère entropique comme nous l'avons montré dans le second chapitre. Dans le premier cas, l'optimisation du système risque d'être encore très difficile, même si la complexité du système a été préalablement diminuée par notre phase de sélection de capteurs. Dans le second cas, nous montrerons que les propriétés particulières de l'entropie peuvent être mises à profit pour limiter la complexité des calculs à mettre en œuvre lors de l'optimisation du système de détection. Nous nous baserons sur les méthodes de construction d'arbres de décision développées en classification pour proposer des algorithmes basés sur le critère

entropique présenté au chapitre 2. Ces algorithmes nous permettront de nous approcher du système de détection décentralisé parallèle optimal en limitant le plus possible la complexité des calculs à mettre en œuvre.

Enfin, il paraît évident que les performances des systèmes de détection centralisée sont meilleures que les performances obtenues via des systèmes de détection décentralisée. Dans ce sens, les techniques d'optimisation précédentes seront étendues au problème de la quantification répartie afin d'obtenir un compromis entre la quantité d'information à envoyer à l'opérateur de fusion et les performances du système de détection.

1ère partie :

DETECTION

CHAPITRE 1

LA THEORIE DE LA DETECTION

-

ETAT DE L'ART

1. Introduction

Ce chapitre présente les résultats importants de la théorie de la détection. Dans un premier temps, le problème de la détection ainsi qu'un outil de comparaison des performances des systèmes de détection seront présentés.

Dans le paragraphe 4, la détection centralisée sera présentée. Dans ce cadre, l'ensemble des informations délivrées par les capteurs du système est transmis à un opérateur de décision qui élabore la décision finale. Le problème est alors de déterminer une stratégie de décision basée sur l'utilisation d'un critère d'optimisation. Deux critères sont principalement utilisés : le critère de Bayes et celui de Neyman-Pearson.

Le paragraphe 5 sera, quant à lui, consacré aux progrès accomplis ces dernières années dans le domaine de la détection décentralisée. Le principe des systèmes de détection décentralisée est d'associer à chaque capteur un détecteur qui prend une décision locale. L'ensemble de ces décisions est ensuite transmis à un opérateur de fusion qui élabore la décision finale. Dans ce cadre, nous présenterons l'étude des systèmes de détection décentralisée parallèle du point de vue Bayésien, puis du point de vue de Neyman-Pearson. Enfin, nous terminerons ce chapitre par une présentation des systèmes de détection décentralisée série.

2. Le problème de la détection

Considérons un système composé de N capteurs Y_i i=1,...,N observant le même phénomène. A partir des mesures $y=(y_1,y_2,...,y_N)$ fournies par l'ensemble des capteurs $Y=(Y_1,Y_2,\ldots,Y_N)$, deux hypothèses notées H_0 et H_1 doivent être discriminées. Le problème consiste alors à déterminer la stratégie de décision qui permettra d'associer, à chaque observation $y=(y_1,y_2,...,y_N)$ donnée, l'hypothèse H_0 ou H_1. Usuellement H_0 correspond à la situation où l'observation n'est composée que de bruit tandis que H_1 correspond à celle où le signal attendu est présent.

H_0 : y=b
H_1 : y=b+s

La densité de probabilités de la variable aléatoire Y dépend de la situation dans laquelle on se trouve :

- en présence de bruit seul - hypothèse H_0 - la densité de probabilités est $p(Y/H_0)$,
- en présence de signal et de bruit - hypothèse H_1 - la densité de probabilités est $p(Y/H_1)$.

Les probabilités *a priori* des hypothèses H_0 et H_1 sont supposées connues et sont notées respectivement P_0 et P_1 (avec $P_0+P_1=1$).

L'ensemble des observations y fournies par les capteurs constitue l'espace des observations, noté D. Décider que l'on se trouve dans la situation H_0 ou H_1 revient à diviser l'espace des observations D en deux domaines disjoints D_0 et D_1 (Figure 1) tels que si l'observation tombe dans D_0 (respectivement dans D_1) la décision prise est H_0 (respectivement H_1). Une partition de l'espace des observations D en deux classes est alors obtenue :

$$\begin{cases} \text{si } y \in D_0, \text{ on décide } H_0 \\ \text{si } y \in D_1, \text{ on décide } H_1 \end{cases} \qquad \text{avec} \quad D_0 \cup D_1 = D \quad \text{et} \quad D_0 \cap D_1 = \varnothing$$

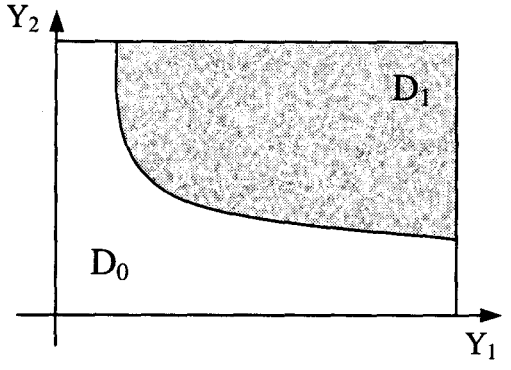


Figure 1 : Les domaines de décision dans le cas où N=2 (2 capteurs)

Quatre situations sont dès lors répertoriées :

Décision (δ)	Hypothèse vraie (H)	Intitulé
H_1	H_1	Détection vraie
H_0	H_0	Non détection vraie
H_1	H_0	Fausse alarme
H_0	H_1	Non détection fausse

En notant :

- décision δ_0 = décider que H_0 est réalisée,
- décision δ_1 = décider que H_1 est réalisée,

les probabilités conjointes de décision et d'hypothèse vraie associées aux quatre situations possibles peuvent être exprimées de la façon suivante :

$$P(\delta_i, H_j) = P(H_j)\, P(\delta_i / H_j) = P(H_j) \int_{D_i} p(y/H_j)\, dy \qquad i,j \in \{0,1\}$$

C'est une probabilité d'erreur si $i \neq j$, une probabilité conjointe de décision si $i=j$.

On peut également définir :

- la probabilité de détection $P_D = P(\delta_1/H_1) = \int_{D_1} p(y/H_1)\, dy$
- la probabilité de fausse alarme $P_F = P(\delta_1/H_0) = \int_{D_1} p(y/H_0)\, dy$

3. Comparaison des performances des systèmes de détection : les courbes C.O.R.

Afin de choisir entre plusieurs systèmes de détection, on a coutume d'évaluer leurs performances respectives en traçant leurs courbes C.O.R. (Caractéristique Opérationnelle du Récepteur) respectives, donnant les variations de la probabilité de détection en fonction de la probabilité de fausse alarme. Plus la courbe se rapproche du point (0,1), plus la qualité du détecteur est bonne. En effet, pour ce détecteur, une probabilité de fausse alarme faible sera associée à une probabilité de détection élevée (Figure 2).

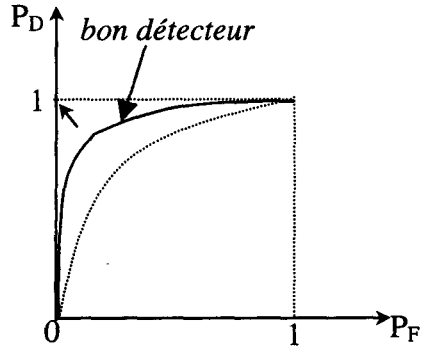


Figure 2 : Exemples de courbes C.O.R.

La construction empirique de ces courbes C.O.R. sera présentée au paragraphe 4.3.

4. La détection centralisée

Dans le cadre de la détection centralisée, l'ensemble des informations délivrées par les N capteurs Y_i (i=1,..,N) est transmis à un opérateur de décision qui élabore la décision finale (Figure 3).

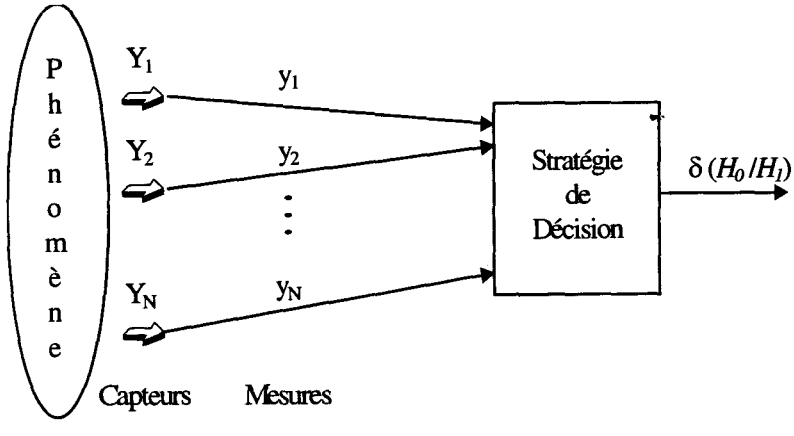


Figure 3 : La détection centralisée

Le problème est alors de déterminer une « stratégie de décision » qui nous permettra de diviser l'espace des observations D en deux domaines disjoints D_0 et D_1. Cette stratégie dépend du critère d'optimisation utilisé. Deux critères sont principalement usités :

- le critère de Bayes
- et le critère de Neyman-Pearson

4.1. Le critère de Bayes

Le problème du test d'hypothèse Bayésien a été abondamment étudié dans la littérature [MeC60] [Bar91] [Hel95a] [Poo88] [SRV95]. Les probabilités *a priori* étant connues, ainsi que les fonctions de densité de probabilités des différentes observations pour chaque hypothèse, un coût est associé à chaque situation. On appelle C_{ij} le coût correspondant à la décision δ_i lorsque H_j est vraie. Si une mauvaise décision est prise, le système est pénalisé ; si par contre une bonne décision est prise, alors il est avantagé ; dans ce sens $C_{ij}>C_{ii}$, $i\neq j$. Ces différents coûts sont en général déterminés par un expert qui prend en compte des considérations propres à chaque système. Le principe est alors de trouver la stratégie de décision qui minimise le risque moyen correspondant aux différentes situations possibles.

• Minimisation du risque moyen $\Re$

Soit $\Re$ le risque moyen :

$$\Re = \sum_{i,j\in\{0,1\}} C_{ij}\, P(\delta_i, H_j)$$

$$\Re = \sum_{i,j\in\{0,1\}} C_{ij}\, P_j\, P(\delta_i / H_j)$$

$$\Re = \sum_{i,j\in\{0,1\}} C_{ij}\, P_j \int_{D_i} p(y/H_j)\, dy \qquad (4.1.1)$$

$$\Re = P_0C_{00} \int_{D_0} p(y/H_0)\, dy + P_0C_{10} \int_{D-D_0} p(y/H_0)\, dy + P_1C_{01} \int_{D_0} p(y/H_1)\, dy + P_1C_{11} \int_{D-D_0} p(y/H_1)\, dy \qquad (4.1.2)$$

En notant que $\int_D p(y/H_j)\, dy=1 \qquad j\in\{0,1\}$

et en rassemblant les différents termes, (4.1.2) s'écrit :

$$\Re = P_0C_{10} + P_1C_{11} + \int_{D_0} \{ [P_1 (C_{01}-C_{11})\, p(y/H_1)] - [P_0 (C_{10}-C_{00})\, p(y/H_0)] \}\, dy \qquad (4.1.3)$$

La minimisation de $\Re$ permet de déterminer les régions D_0 et D_1. Nous pouvons remarquer que les deux premiers termes sont fixés et que la minimisation de $\Re$ passe par la minimisation de l'intégrale ; ce qui se traduit, pour chaque observation y, par la règle de décision suivante :

- si $[P_1 (C_{01}-C_{11})\, p(y/H_1)] - [P_0 (C_{10}-C_{00})\, p(y/H_0)] < 0$ alors on décide H_0 ($y\in D_0$)
- si $[P_1 (C_{01}-C_{11})\, p(y/H_1)] - [P_0 (C_{10}-C_{00})\, p(y/H_0)] > 0$ alors on décide H_1 ($y\in D_1$)

Ce qui s'écrit également de la façon suivante :

$$\frac{p(y/H_1)}{p(y/H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})} \tag{4.1.4}$$

La quantité de gauche est connue sous le nom de « rapport de vraisemblance », noté $\Lambda(y)$, et celle de droite est appelée « seuil » (noté en général λ).

• Le détecteur Bayésien optimal

Minimiser le risque moyen $\Re$ est équivalent à utiliser un test du rapport de vraisemblance qui peut être résumé par :

$$\Lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \qquad \text{où} \qquad \Lambda(y) = \frac{p(y/H_1)}{p(y/H_0)} \tag{4.1.5}$$

$$\text{et} \qquad \lambda = \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})}$$

Le détecteur optimal peut être schématisé de la façon suivante :

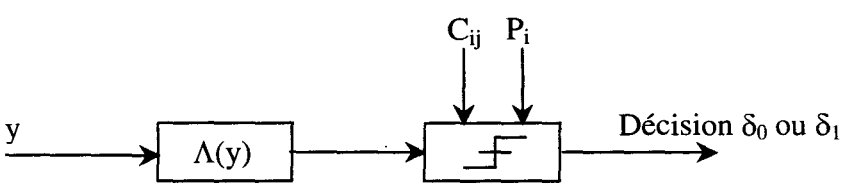


Figure 4 : Détecteur Bayésien optimal

• Cas particulier

Un cas particulier de la détection Bayésienne est le cas de la pénalisation symétrique où $C_{00}=C_{11}=0$ et $C_{10}=C_{01}=1$; le coût d'une décision correcte est fixé à « zéro » et celui d'une décision fausse à « un ». Dans ce cas le risque moyen s'écrit :

$$\Re = P_0 P_F + P_1(1-P_D) \tag{4.1.6}$$

$\Re$ représente alors une probabilité moyenne d'erreur, et le seuil λ est égal à P_0/P_1. Lorsque les deux hypothèses sont équiprobables, $\lambda=1$. Ces hypothèses sont souvent vérifiées lorsque l'on travaille sur des systèmes numériques de transmission.

• Exemple

Considérons un capteur dont les observations sont notées y. Les fonctions de densité de probabilités sous chaque hypothèse sont des gaussiennes de même écart-type et de moyennes différentes telles que :

$$p(y/H_j)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-m_j)^2}{2\sigma^2}\right) \qquad j\in\{0,1\}$$

D'après l'équation (4.1.5), la règle de décision est :

$$\frac{p(y/H_1)}{p(y/H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})}$$

Dans le cas particulier où $C_{00}=C_{11}=0$ et $C_{10}=C_{01}=1$. Nous avons tracé la valeur du seuil, la probabilité d'erreur (ou risque moyen $\Re$) en fonction de P_0 ainsi que la courbe C.O.R. en prenant P_0 comme paramètre pour $m_0=0$, $m_1=1$, et $\sigma=1$. Sur la figure 5, on peut voir que lorsque P_0 tend vers 1 (respectivement vers 0), le seuil λ prend des valeurs très grandes (respectivement très petites). Cela signifie que si un événement se produit rarement, en appliquant le critère de Bayes, on négligera complètement cet événement. Le critère de Bayes n'est donc pas du tout adapté au problème de la détection de pannes, celles-ci n'intervenant que très rarement.

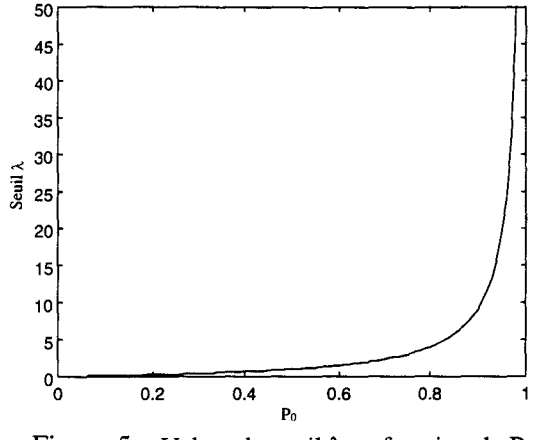


Figure 5 : Valeur du seuil λ en fonction de P_0

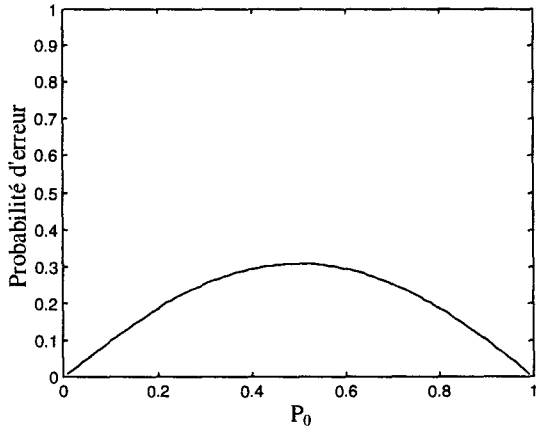


Figure 6 : Probabilité d'erreur en fonction de P_0

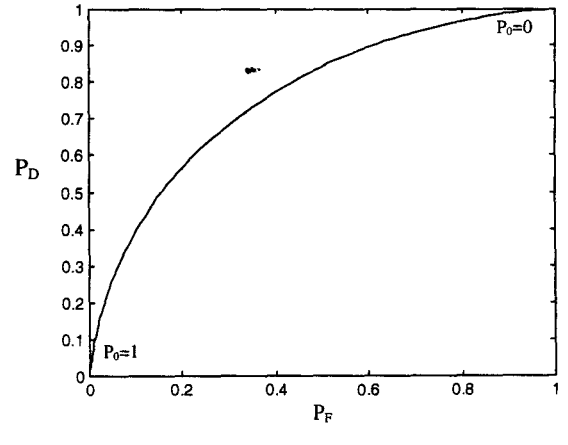


Figure 7 : Courbe C.O.R. $P_D=f(P_F)$

Dans la pratique, la fonction de coût et les probabilités *a priori* associées aux différentes hypothèses sont souvent inconnues. Dans ce cas, on pourra utiliser le critère de Neyman-Pearson.

4.2. Le critère de Neyman-Pearson

Le critère de Neyman-Pearson consiste à maximiser la probabilité de détection P_D pour une probabilité de fausse alarme donnée $P_F=\alpha$. Ce critère revient à considérer que la fausse alarme est l'erreur la plus grave ; une valeur maximale α de fausse alarme acceptable est alors fixée. A cette valeur est associée un ensemble de couples (D_0,D_1) vérifiant la contrainte. Le critère de Neyman-Pearson consiste à sélectionner dans cet ensemble le couple (D_0^*,D_1^*) auquel correspond la probabilité de détection la plus grande. Il s'agit d'un problème d'optimisation sous contrainte qui se résout par la méthode du multiplicateur de Lagrange :

• Maximisation du Lagrangien L

Le Lagrangien s'écrit :

$$\mathbf{L} = P_D - \lambda(P_F-\alpha)$$

$$\mathbf{L} = \lambda\,\alpha + \int_{D_1} p(y/H_1)\,dy - \lambda \int_{D_1} p(y/H_0)\,dy$$

$$\mathbf{L} = \lambda\,\alpha + \int_{D_1} [p(y/H_1) - \lambda\, p(y/H_0)]\,dy \qquad (4.2.1)$$

Le premier terme étant fixé, maximiser **L** revient à maximiser l'intégrale, ce qui se traduit, pour chaque observation y, par la règle de décision suivante :

- si $p(y/H_1) - \lambda\, p(y/H_0) < 0$ alors on décide H_0 $(y\in D_0)$
- si $p(y/H_1) - \lambda\, p(y/H_0) > 0$ alors on décide H_1 $(y\in D_1)$

Ce qui peut s'écrire :

$$\Lambda(y) = \frac{p(y/H_1)}{p(y/H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \qquad (4.2.2)$$

De plus, λ doit être choisi de façon à respecter la contrainte :

$$P_F = P(\delta_1/H_0) = P(\Lambda \geq \lambda/H_0) = \int_{\lambda}^{+\infty} p(\Lambda/H_0)\,d\Lambda = \alpha$$

• **Le détecteur de Neyman-Pearson optimal**

Maximiser le Lagrangien **L** revient donc à utiliser un test du rapport de vraisemblance qui peut être résumé par :

$$\Lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \qquad \text{où} \qquad \Lambda(y) = \frac{p(y/H_1)}{p(y/H_0)} \tag{4.2.3}$$

$$\text{et} \qquad P_F = \int_{\lambda}^{+\infty} p(\Lambda/H_0)\, d\Lambda = \alpha \quad \text{fixé}$$

4.3. Rapport de vraisemblance et courbes C.O.R.

• **Définition**

Les probabilités de fausse alarme et de détection s'expriment en fonction du seuil λ fixé. L'élimination de ce seuil entre les deux expressions conduit à une relation liant les probabilités de détection et de fausse alarme.

$$P_F = P(\delta_1/H_0) = P(\Lambda \geq \lambda/H_0) = \int_{\lambda}^{+\infty} p(\Lambda/H_0)\, d\Lambda = A(\lambda)$$

$$P_D = P(\delta_1/H_1) = P(\Lambda \geq \lambda/H_1) = \int_{\lambda}^{+\infty} p(\Lambda/H_1)\, d\Lambda = B(\lambda)$$

Donc $\lambda = A^{-1}(P_F)$ et $P_D = B(A^{-1}(P_F))$, par conséquent $P_D = f(P_F)$.

Les courbes C.O.R. pourront donc être tracées en utilisant λ comme paramètre.

• **Lemme** On a toujours $$\frac{P(\Lambda = \lambda/H_1)}{P(\Lambda = \lambda/H_0)} = \lambda \tag{4.3.1}$$

• **Propriétés**

Propriété 1 :

En tout point de la courbe C.O.R. d'un récepteur à rapport de vraisemblance on a $\frac{\partial P_D}{\partial P_F} = \lambda$ (4.3.2)

Démonstration :

$$P_F(\lambda) = \int_{\lambda}^{+\infty} p(\Lambda/H_0)\, d\Lambda \Rightarrow \left(\frac{\partial P_F}{\partial \Lambda}\right)_{\Lambda=\lambda} = -P(\Lambda=\lambda/H_0) \qquad \text{car} \lim_{\Lambda \to +\infty} p(\Lambda/H_0) = 0$$

de même : $P_D(\lambda) = \int_{\lambda}^{+\infty} p(\Lambda/H_1)\, d\Lambda \Rightarrow \left(\frac{\partial P_D}{\partial \Lambda}\right)_{\Lambda=\lambda} = -P(\Lambda=\lambda/H_1)$

donc d'après (4.3.1) :

$$\frac{\left(\frac{\partial P_D}{\partial \Lambda}\right)_{\Lambda=\lambda}}{\left(\frac{\partial P_F}{\partial \Lambda}\right)_{\Lambda=\lambda}} = \frac{P(\Lambda=\lambda/H_1)}{P(\Lambda=\lambda/H_0)} = \lambda$$

on en déduit que : $\frac{\partial P_D}{\partial P_F} = \lambda$ *cqfd.*

Propriété 2 :

La courbe C.O.R. $P_D=f(P_F)$ d'un récepteur à rapport de vraisemblance est à concavité négative. (4.3.4)

Démonstration : Cette propriété est liée au fait que $\frac{\partial P_D}{\partial P_F} = \lambda \geq 0$

$\Rightarrow$ Une petite variation positive de P_F entraîne une petite variation positive de P_D.

cqfd.

Propriété 3 :

On a toujours $P_D > P_F$, sinon le récepteur ne pourrait pas être optimal. (4.3.3)

Démonstration :

$$\left(\frac{\delta(P_D - P_F)}{\delta\Lambda}\right)_{\Lambda=\lambda} = -P(\Lambda=\lambda/H_1) + P(\Lambda=\lambda/H_0)$$
$$= P(\Lambda=\lambda/H_0)(1-\lambda)$$

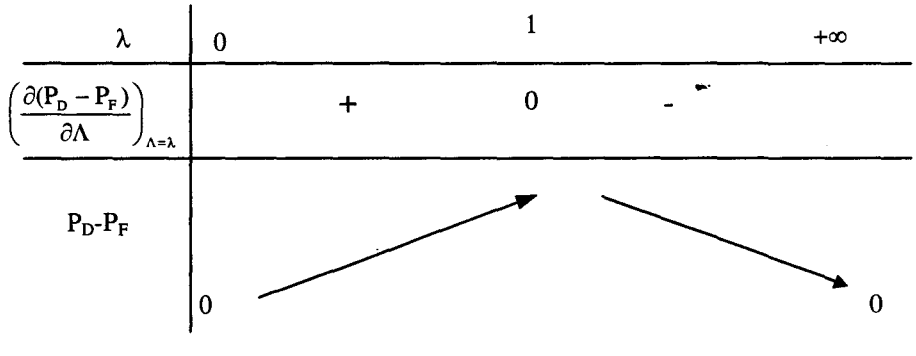


Figure 8 : Variations de P_D-P_F en fonction de λ.

cqfd.

4.4. Lien entre les critères de Bayes et de Neyman-Pearson

On constate que les critères de Bayes et de Neyman-Pearson, qui correspondent à des points de vue très différents de l'opération de détection, conduisent à la même structure optimale du récepteur qui consiste à comparer le rapport de vraisemblance à un seuil, seule la définition de ce seuil est différente.

Ainsi dans le cas Bayésien l'équation (4.1.2) peut s'écrire :

$$\Re = P_0C_{00}(1-P_F) + P_0C_{10}P_F + P_1C_{01}(1-P_D) + P_1C_{11}P_D$$
$$\Re = P_0C_{00} + P_1C_{01} + P_D(P_1C_{11} - P_1C_{01}) + P_F(P_0C_{10} - P_0C_{00})$$

ce qui entraîne que : $$P_D = -\frac{\Re - (P_0C_{00} + P_1C_{01})}{P_1(C_{01} - C_{11})} + \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} P_F$$

$$P_D = -\mu(\Re) + \eta P_F \qquad \text{où} \qquad \eta = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} > 0 \qquad \text{car } C_{ij} > C_{ii}\ i \neq j$$

et

$$\mu(\Re) = \frac{\Re - (P_0C_{00} + P_1C_{01})}{P_1(C_{01} - C_{11})} > 0$$

car $\Re - (P_0C_{00} + P_1C_{01}) = P_FP_0(C_{10}-C_{00}) + P_DP_1(C_{11}-C_{01})$

Le critère de Bayes (Figure 9) revient donc à trouver le point d'intersection M entre la courbe C.O.R. et la droite D_B d'équation $P_D=-\mu(\Re)+\eta P_F$. Tandis que le critère de Neyman-Pearson revient à trouver le point d'intersection P entre la courbe C.O.R. et la droite $D_{N\text{-}P}$ d'équation $P_F=\alpha$.

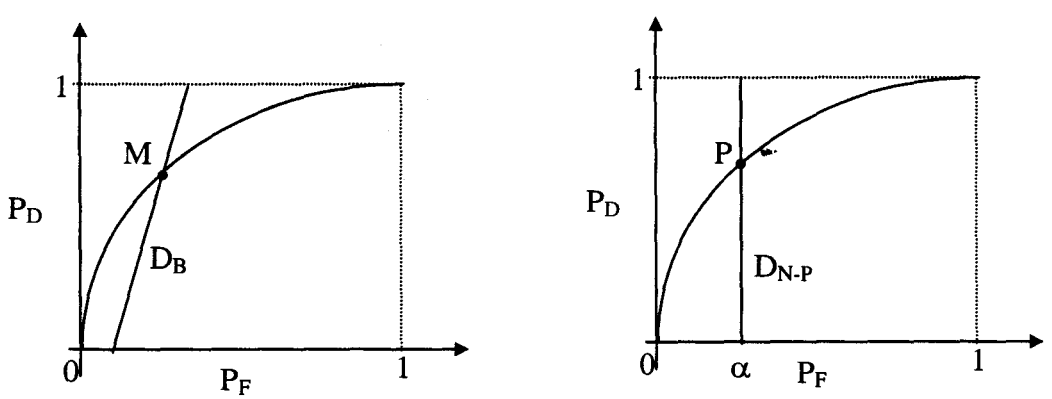


Figure 9 : Comparaison des critères de Bayes et de Neyman-Pearson

Remarquons qu'à tout point M trouvé en utilisant le critère de Bayes correspond un point P trouvé en appliquant le critère de Neyman-Pearson.

4.5. Les autres critères de détection

Il existe d'autres approches développées dans la littérature qui correspondent à des situations particulières. A titre d'exemples, citons :

• Le critère Minimax [GaL95]

Il est basé sur le critère de Bayes et est utilisé lorsque les probabilités *a priori* P_0 et P_1 ne sont pas connues. Ce critère revient à utiliser la règle de décision Bayésienne en affectant à P_0 et à P_1 les valeurs les moins favorables.

• La détection à niveau de fausse alarme constant ou détection CFAR (Constant False Alarm Rate) [FiJ68]

Cette approche est utilisée pour traiter le cas particulier de la détection radar où le signal est immergé dans un bruit de fond (bruit et encombrement dû à la présence d'objets qui n'intéressent pas l'opérateur). Lorsque l'environnement est stationnaire, les statistiques associées aux différentes hypothèses peuvent être calculées et dans ce cas le critère de Neyman-Pearson peut être utilisé. Par contre, si le bruit de fond varie, un détecteur à seuil fixe ne peut pas être utilisé car à un instant donné la probabilité de fausse alarme P_F peut être trop élevée et la probabilité de détection P_D trop faible. Dans ce cas, des techniques adaptatives sont utilisées. Dans le cas d'un bruit de fond gaussien, par exemple, la détermination d'un seuil approprié est basée sur l'estimation de la puissance moyenne du bruit de fond en fonction du temps.

• Le critère de détection séquentielle, ou critère de Wald [Wal47a] [Wal47b] [Wal48]

Il s'applique lorsque les informations fournies par le système sont collectées séquentiellement. C'est un test de type Neyman-Pearson. A chaque fois qu'une nouvelle information arrive, le rapport de vraisemblance est calculé et comparé à deux seuils λ_0 et λ_1 déterminés en fonction des probabilités P_D et P_F désirées. Si le rapport de vraisemblance est supérieur à λ_1, alors le système décide que l'hypothèse H_1 est présente. Si le rapport de vraisemblance est inférieur à λ_0, alors le système décide que l'hypothèse H_0 est présente. Dans le dernier cas le système attend d'avoir une nouvelle information afin de prendre une décision.

4.6. Conclusion

La théorie de la détection centralisée permet de prendre une décision suivant un critère précis. Que le critère soit le critère de Bayes ou celui de Neyman-Pearson, la théorie aboutit à la même structure optimale du détecteur ; l'ensemble des informations délivrées par les N capteurs Y_i (i=1,..,N) est transmis à un opérateur de décision qui consiste à comparer, pour chaque observation $y=(y_1,...y_N)$, le rapport de vraisemblance à un seuil λ. En supposant que l'information y_i délivrée par chaque capteur Y_i est codée sur m_i bits, à chaque instant le système doit transmettre à l'opérateur de décision $M=\sum_{i=1}^{N} m_i$ bits d'information.

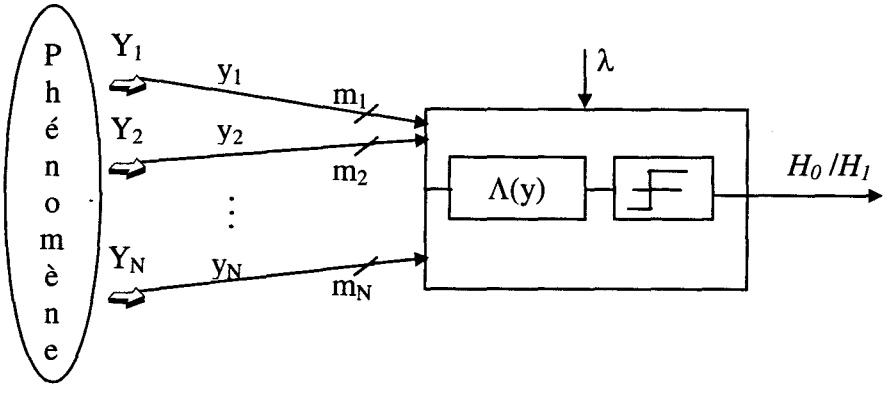


Figure 10 : Architecture d'un système de détection centralisée

Ainsi, en détection centralisée, l'opérateur de décision récupère l'ensemble de l'information délivrée par les capteurs. En terme de fiabilité, un tel système donnera donc les meilleures performances possibles, à la condition que l'information délivrée par les capteurs arrive bien au système de détection. Cependant, on voit que la quantité M d'information à transmettre peut vite devenir très importante, surtout si l'on veut travailler en temps réel et avec un grand nombre de capteurs. Considérons par exemple le contexte de la surveillance d'installations industrielles complexes, où un grand nombre de capteurs observant des grandeurs physiques différentes sont disponibles. L'installation d'un tel système de détection nécessitera la mise en place d'une ligne de transmission entre chaque capteur et le calculateur central (l'opérateur de décision). Ces lignes devront être à large bande de façon à pouvoir transmettre la totalité de l'information issue de chaque capteur, et cela, à des cadences qui peuvent être élevées. Lors de son installation, le système générera un coût proportionnel, d'une part à la qualité et à la longueur des lignes de transmission à mettre en place, et d'autre part à la dispersion géographique et au nombre de sites de mesure à intégrer.
Depuis une quinzaine d'années, une solution alternative à la détection centralisée a été développée. Appelée détection décentralisée, elle consiste à associer à chaque capteur un détecteur qui décide localement si un signal a été détecté ou non. Ces décisions locales sont ensuite envoyées à un opérateur de fusion qui les combine pour prendre la décision finale. Le but recherché est alors d'obtenir des performances se rapprochant le plus de celles de la détection centralisée (même si l'opérateur de fusion ne reçoit pas toute l'information

nécessaire à la prise de décision), en utilisant une architecture matérielle moins coûteuse à gérer en terme de flux d'informations.

5. La détection décentralisée

En 1981, Tenney et Sandell [TeS81] furent les premiers à étudier le problème de la détection décentralisée. Ils s'attachèrent au problème de la détection répartie associée à deux capteurs en parallèle. Depuis c'est un sujet qui a connu un développement exponentiel, avec la publication de plus de cent articles dans différentes revues de référence (1981-1997). Quatre grands thèmes se dégagent à la lecture de ces articles :

- l'architecture des systèmes de détection : détection parallèle [TeS81] [ChV86] [Sad86] [ReN87a] [HoV89a] [DrL91], série [PaA90] [Tan90] [Tan91b] [Swa93], ou mixte [ReN87b]
- le type de détection : détection binaire [TeS81] [ChV86] ou multi-hypothèses [Sad86] [PaA90]
- le critère d'optimisation du système de détection : critère de Bayes [TeS81] [ChV86] [Sad86] [ReN87a] [HoV89a] ou de Neyman-Pearson [TVB87] [Sri86] [TVB89] [DrL91]
- les applications [LuK89] : robotique, défense [Sri86], espace...

Après un bref historique, nous présenterons l'étude des systèmes de détection décentralisée parallèle du point de vue Bayésien, puis du point de vue de Neyman-Pearson. Enfin, nous terminerons ce chapitre par une présentation des systèmes de détection décentralisée série.

5.1. La détection décentralisée parallèle

5.1.1. Historique

Les systèmes de détection peuvent avoir de nombreuses architectures. Parmi toutes les topologies présentées dans la littérature, la détection parallèle est celle qui a été la plus étudiée. Notre but est ici de montrer, en citant différents articles clés dans le domaine, comment la problématique de la détection décentralisée parallèle a évolué au fil des années.

En 1981, Tenney et Sandell [TeS81] furent les premiers à étudier le problème de la détection décentralisée (Figure 11). Ils s'attachèrent au problème de la détection répartie associée à deux capteurs en parallèle. Leur travail consistait, en se basant sur le critère de Bayes, à déterminer la règle de décision optimale locale pour chaque capteur. Plusieurs résultats importants ressortent de leur étude :

- Le détecteur optimal local pour chaque capteur est celui qui utilise un seuillage du rapport de vraisemblance local.
- Les équations qui permettent de déterminer les différents seuils sont couplées.
- Les solutions de ces équations semblent être des optimums locaux.

Néanmoins, le problème de l'opérateur de fusion ne fut pas abordé (ils utilisent un « ou » ou un « et » logique), seules des pistes furent données pour l'extension de leurs résultats à des systèmes de détection à M-hypothèses et formés de N capteurs.

Chair et Varshney [ChV86] essayèrent de développer le problème de l'opérateur de fusion prenant comme hypothèse de travail, l'indépendance des décisions locales prises à partir de chaque capteur. Ils considèrent les règles de décision connues pour chaque détecteur local, ainsi que les probabilités de détection et de fausse alarme associées à chaque détecteur. L'opérateur de décision consiste alors à comparer une somme pondérée des décisions locales à un seuil. Ils n'étudient pas le problème d'optimisation des détecteurs locaux.

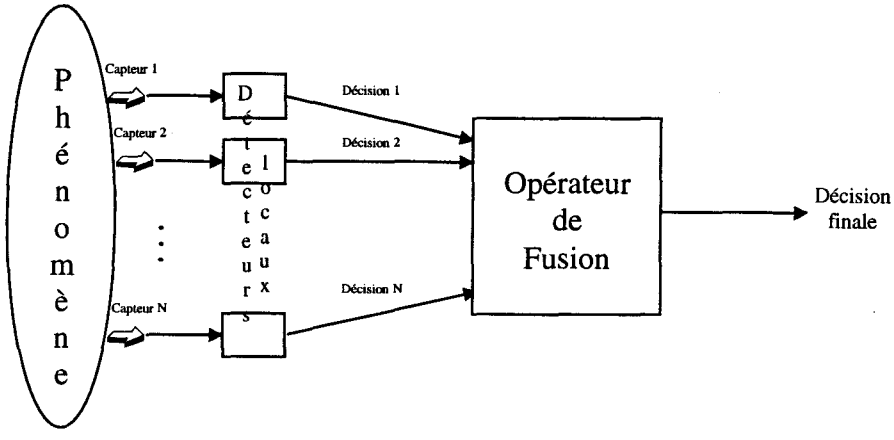


Figure 11: La détection décentralisée parallèle

Au même moment, Sadjadi [Sad86] présente une solution optimale sur le même thème de recherche que Tenney et Sandell mais en étendant leurs résultats à un système de détection à M-hypothèses et formé de N capteurs. Il définit une fonction de coût associée à chaque décision locale, et minimise un coût moyen global ce qui conduit à un ensemble d'inégalités qui l'obligent à introduire la notion de rapport de vraisemblance généralisé. Le coût des mauvaises décisions peut varier en fonction des décisions prises par les autres détecteurs locaux. Les solutions de ces inégalités conduisent à un domaine de décision optimal, qui correspond à la stratégie de fusion.

En 1987 Srinivan [Sri87] développe le problème de la détection distribuée appliquée aux radars. Il utilise pour optimiser son système le critère de Neyman-Pearson. Comme dans [TeS81], il s'intéresse à l'optimisation des règles de décision locales, l'opérateur de fusion étant un « et » ou un « ou » logique.

Début 1987, Reibman et Nolte [ReN87a] étendent les résultats précédents en optimisant simultanément les détecteurs locaux et l'opérateur de fusion. Cette optimisation conduit à un ensemble d'équations couplées qui dépendent des performances de chaque détecteur et de l'opérateur de fusion. Dans le cas où les détecteurs sont similaires (même statistique et même seuillage du rapport de vraisemblance), ils montrent que l'opérateur de fusion consiste à décider la présence d'un signal si k détecteurs parmi N l'ont eux-mêmes décidé. k est fixé en fonction de la valeur des seuils qui sont utilisés. Le cas d'un signal noyé dans un bruit non gaussien est ensuite étudié, et ils montrent l'apport de cette nouvelle approche par rapport à celles développées dans [TeS81] et [ChV86].

Dans [TVB87] Thomopoulos Viswanathan, et Bougoulias arrivent au même résultat que [ReN87a], mais montrent que l'ensemble du système de détection décentralisée a une probabilité de détection supérieure à celle de n'importe quel détecteur local pour une probabilité de fausse alarme fixée. Ils introduisent aussi une variable qualité, qui est transmise à l'opérateur de fusion en même temps que la décision locale, et qui améliore les performances de l'ensemble.

Thomopoulos, Viswanathan et Bougouglias [TVB89] démontrent en 1989 que la fonction de fusion ne peut être que monotone. Dans le cas où l'opérateur de fusion est connecté à N capteurs qui délivrent une décision binaire, il y a 2^N observations possibles et 2^{2^N} fusions possibles. En ne gardant que les fonctions de fusion monotones, on réduit donc de façon sensible le nombre de fonctions de fusion admissibles. De plus, si l'opérateur de fusion est monotone alors la probabilité de détection de l'ensemble sera toujours supérieure à la probabilité de détection associée au meilleur capteur.

La même année, Hobalah et Varshney [HoV89a] publient un article de synthèse sur le problème de la détection décentralisée et son optimisation d'un point de vue Bayésien. Pour un problème de détection binaire à N capteurs, $N+2^N$ équations non-linéaires couplées doivent être résolues. Lorsque N devient grand on se heurte donc à un important problème calculatoire. Pour simplifier le problème, ils utilisent des capteurs identiques et font l'hypothèse de l'indépendance des informations.

Au cours de l'année 1991, Drakopoulos et Lee [DrL91] présentent un article traitant de la fusion optimale de capteurs corrélés. Ils étudient un système composé d'un processeur central de fusion et de détecteurs binaires locaux qui élaborent leurs décisions à partir de leurs propres ensembles d'observations. Ces décisions sont supposées être corrélées. Ces corrélations sont caractérisées par un ensemble fini de probabilités conditionnelles et l'opérateur de fusion optimal est basé sur le critère de Neyman-Pearson. L'étude met en évidence la dégradation des performances du système en fonction de l'augmentation du degré de corrélation entre les capteurs. Intuitivement on comprend bien que si la corrélation entre les capteurs augmente, alors la quantité d'information utile pour la prise de décision diminue.

Les paragraphes suivants rappellent les principaux résultats développés dans ces articles.

5.1.2. Le point de vue Bayésien

Nous rappelons dans ce paragraphe dans quelle mesure le point de vue Bayésien a pu être étendu au problème de la détection décentralisée parallèle. Dans un premier temps, nous considérons le problème d'optimisation des différents opérateurs locaux sans prendre en compte l'opérateur de fusion. Puis, les statistiques sur les opérateurs locaux étant connues, nous rappelons de quelle façon l'opérateur de fusion peut être déterminé. Enfin, l'optimisation simultanée des détecteurs locaux et de l'opérateur de fusion sera présentée.

5.1.2.1. Optimisation des détecteurs locaux

Dans le but de privilégier la compréhension de la démarche employée, nous nous limiterons volontairement dans ce paragraphe à l'optimisation d'un système formé de deux détecteurs locaux. La généralisation à plus de deux détecteurs est tout à fait possible, mais elle compliquerait les notations.

Dans ce paragraphe, nous ne considérons que l'optimisation des détecteurs locaux sans prendre en compte le problème de la fusion [TeS81] [Sad86]. Le système étudié est composé de deux capteurs Y_1 et Y_2 en parallèle qui observent un phénomène commun. A chaque capteur est associé un détecteur qui prend une décision locale (Figure 12). Les détecteurs ne communiquent pas entre eux mais sont couplés par les coûts associés à chaque décision. Nous n'étudions dans ce paragraphe que la détection décentralisée binaire, encore appelée « détection distribuée ». H_0 et H_1 sont les deux hypothèses à discriminer, de probabilités respectives *a priori* P_0 et P_1. Les observations associées aux deux capteurs sont notées y_1 et y_2. Les fonctions de densité de probabilités conditionnelles sous chaque hypothèse sont notées $p(y_1,y_2/H_i)$, i=0,1. Les décisions prises par chaque détecteur sont notées u_i, i=1,2 :

$$u_i = \begin{cases} 0, \text{ on décide que } H_0 \text{ est vraie} \\ 1, \text{ on décide que } H_1 \text{ est vraie} \end{cases}$$

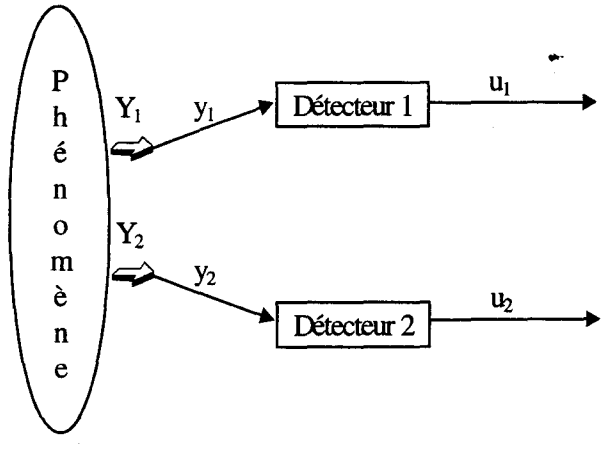


Figure 12 : Deux détecteurs en parallèle

• **Minimisation du risque moyen $\Re$**

Les coûts des différentes décisions sont notés C_{ijk}, i,j,k=0,1 où C_{ijk} représente le coût de décider H_i pour le détecteur 1, H_j pour le détecteur 2 alors que H_k est vraie. Le critère de Bayes consiste à minimiser le risque moyen $\Re=E\{C_{ijk}\}$:

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} C_{ijk}\, p(u_1,u_2,y_1,y_2,H_k)\, dy_1\, dy_2 \qquad (5.1.2.1.1)$$

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} C_{ijk}\, p(u_1,u_2,y_1,y_2/H_k)\, P_k\, dy_1\, dy_2$$

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} C_{ijk}\, p(u_1,u_2/y_1,y_2,H_k)\, p(y_1,y_2/H_k)\, P_k\, dy_1\, dy_2$$

u_1 et u_2 étant indépendants l'un de l'autre, de H_k et ne dépendant respectivement que de y_1 et de y_2, $\Re$ peut s'écrire :

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} C_{ijk}\, p(u_1/y_1)\, p(u_2/y_2)\, p(y_1,y_2/H_k)\, P_k\, dy_1\, dy_2$$

• **Optimisation du premier détecteur**

En développant suivant i, sachant que $p(u_1=1/y_1) = 1 - p(u_1=0/y_1)$, et en rassemblant les différents termes, on obtient :

$$\Re = \int_{y_1} p(u_1=0/y_1) \left(\sum_{j,k} \int_{y_2} P_k\, p(u_2/y_2)\, p(y_1,y_2/H_k)\, [C_{0jk}-C_{1jk}]\, dy_2 \right) dy_1 + C$$

où C est un terme constant indépendant de u_1 :

$$C = \sum_{j,k} \int_{y_1,y_2} C_{1jk}\, p(u_2/y_2)\, p(y_1,y_2/H_k)\, P_k\, dy_1\, dy_2$$

La minimisation de $\Re$ entraîne la minimisation de l'intégrale double, ce qui se traduit, pour chaque observation y_1, par la règle de décision suivante :

$$p(u_1=0/y_1) = \begin{cases} 0, & \text{si } \sum_{j,k} \int_{y_2} P_k\, p(u_2/y_2)\, p(y_1,y_2/H_k)\, [C_{0jk} - C_{1jk}]\, dy_2 \geq 0 \\ 1, & \text{sinon} \end{cases} \qquad (5.1.2.1.2)$$

La règle de décision est donc :

$$\sum_{j,k} \int_{y_2} P_k\, p(u_2/y_2)\, p(y_1,y_2/H_k)\, [C_{0jk}-C_{1jk}]\, dy_2 \underset{u_1=0}{\overset{u_1=1}{\gtrless}} 0$$

En développant suivant k, sachant que $p(y_1,y_2/H_k)=p(y_2/y_1,H_k)p(y_1/H_k)$ (k=0,1), et en réarrangeant les termes, la règle de décision précédente devient :

$$\Lambda(y_1) \underset{u_1=0}{\overset{u_1=1}{\gtrless}} \frac{P_0 \sum_j \int_{y2} p(u_2/y_2)\, p(y_2/y_1,H_0)\,[C_{1j0}-C_{0j0}]\,dy_2}{P_1 \sum_j \int_{y2} p(u_2/y_2)\, p(y_2/y_1,H_1)\,[C_{0j1}-C_{1j1}]\,dy_2} = \lambda_1 \qquad (5.1.2.1.3)$$

où $\Lambda(y_1) = \dfrac{p(y_1/H_1)}{p(y_1/H_0)}$

Il faut noter que le terme de droite n'est pas un seuil très simple puisqu'il dépend de y_1 (du fait du terme $p(y_2/y_1,H_k)$ k=0,1) et de la stratégie de décision de l'autre détecteur (dû au terme $p(u_2/y_2)$).

• Cas où les observations sont indépendantes

La situation se simplifie très largement si l'on suppose que y_1 et y_2 sont indépendantes. Dans ce cas, le seuil λ_1 peut s'écrire :

$$\lambda_1 = \frac{P_0 \sum_j \int_{y2} p(u_2/y_2)\, p(y_2/H_0)\,[C_{1j0}-C_{0j0}]\,dy_2}{P_1 \sum_j \int_{y2} p(u_2/y_2)\, p(y_2/H_1)\,[C_{0j1}-C_{1j1}]\,dy_2}$$

ce qui devient, en développant suivant j, en utilisant le fait que $p(u_2=1/y_2)=1-p(u_2=0/y_2)$ et en réarrangeant les différents termes :

$$\lambda_1 = \frac{P_0 \int_{y2} p(y_2/H_0)\{[C_{110}-C_{010}]+p(u_2=0/y_2)[C_{100}-C_{000}+C_{010}-C_{110}]\}dy_2}{P_1 \int_{y2} p(y_2/H_1)\{[C_{011}-C_{111}]+p(u_2=0/y_2)[C_{001}-C_{101}+C_{111}-C_{011}]\}dy_2} \qquad (5.1.2.1.4)$$

Ce seuil λ_1 est une fonction de $p(u_2=0/y_2)$ qui dépend de la stratégie de décision du deuxième détecteur qui dépend elle-même du seuil λ_2. De ce fait $\lambda_1=f_1(\lambda_2)$ et de même $\lambda_2=f_2(\lambda_1)$ où f_1 et f_2 ont une forme similaire. Pour être optimaux λ_1 et λ_2 doivent satisfaire aux deux conditions précédentes. Mais ces deux conditions ne sont pas suffisantes, elles ne permettent d'obtenir que des optimums locaux. Lorsqu'il existe plusieurs solutions, chacune d'elle doit être examinée pour déterminer l'optimum global.

Afin d'illustrer ce résultat, nous proposons d'étudier deux exemples.

• Exemple 1 [Var97]

Considérons un système formé de deux capteurs dont les observations y_1 et y_2 sont indépendantes. Les fonctions de densité de probabilités conditionnelles sous chaque hypothèse sont des gaussiennes telles que :

$$p(y_i/H_j) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_i - m_j)^2}{2\sigma^2}\right) \quad i=1,2 \quad j=0,1$$

Le rapport de vraisemblance $\Lambda(y_i)$ pour chaque capteur peut être calculé :

$$\Lambda(y_i) = \frac{p(y_i/H_1)}{p(y_i/H_0)} = \exp\left(\frac{1}{2\sigma^2}[2y_i(m_1 - m_0) - (m_1^2 - m_0^2)]\right) \quad i=1,2$$

d'après l'équation (5.1.2.1.3) la règle de décision au niveau de chaque détecteur est la suivante :

$$\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i \quad i=1,2$$

ce qui peut aussi s'écrire :

$$y_i \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \frac{2\sigma^2 \log \lambda_i + (m_1^2 - m_0^2)}{2(m_1 - m_0)} \quad i=1,2$$

Les seuils λ_1 et λ_2 peuvent être calculés en utilisant deux équations de la forme (5.1.2.1.4). A titre d'exemple, prenons les coûts suivants :

$C_{000} = C_{111} = 0$	(aucune pénalité car les deux détecteurs ont vu juste)
$C_{010} = C_{100} = C_{011} = C_{101} = 1$	(pénalité de « 1 » si l'un des deux détecteurs se trompe)
$C_{110} = C_{001} = k$	(grande pénalité car les deux détecteurs se sont trompés).

On a donc :

$$\lambda_1 = \frac{P_0 \int_{y2} p(y_2/H_0)\{[k-1] + p(u_2 = 0/y_2)[2-k]\}dy_2}{P_1 \int_{y2} p(y_2/H_1)\{[1 + p(u_2 = 0/y_2)[k-2]\}dy_2}$$

et en supposant que $P_0 = P_1$, on obtient le résultat suivant :

$$\lambda_1 = \frac{(k-1)\int_{y2} p(y_2/H_0)\,dy_2 + (2-k)\int_{y2} p(u_2 = 0/y_2)p(y_2/H_0)dy_2}{\int_{y2} p(y_2/H_1)\,dy_2 + (k-2)\int_{y2} p(u_2 = 0/y_2)p(y_2/H_1)dy_2}$$

De plus, sachant que : $\int_{y2} p(u_2 = 0/y_2)p(y_2/H_j)\,dy_2 = P(u_2=0/H_j) \quad j=0,1$

et que : $\int_{y2} p(y_2/H_j)\,dy_2 = 1 \quad j=0,1$

λ_1 peut s'écrire : $\lambda_1 = \frac{(k-1) + (2-k)\,P(u_2 = 0/H_0)}{1 + (k-2)\,P(u_2 = 0/H_1)}$

D'autre part, on a : $P(u_2=0/H_0) = \int_{-\infty}^{\frac{2\sigma^2 \log \lambda_2 + (m_1^2 - m_0^2)}{2(m_1 - m_0)}} p(y_2/H_0)\,dy_2$

$$P(u_2=0/H_0)=\int_{-\infty}^{\frac{2\sigma^2\log\lambda_2+(m_1^2-m_0^2)}{2(m_1-m_0)}}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_2-m_0)^2}{2\sigma^2}\right)dy_2$$

Et en faisant le changement de variable suivant : $y=\frac{y_2-m_0}{\sigma}$, on obtient :

$$P(u_2=0/H_0)=\int_{-\infty}^{\frac{\sigma\log\lambda_2}{m_1-m_0}+\frac{m_1-m_0}{2\sigma}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy$$

De même : $P(u_2=0/H_1)=\int_{-\infty}^{\frac{\sigma\log\lambda_2}{m_1-m_0}-\frac{m_1-m_0}{2\sigma}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy$ avec $y=\frac{y_2-m_1}{\sigma}$

Par conséquent :
$$\lambda_1=\frac{(k-1)+(2-k)\int_{-\infty}^{\frac{\sigma\log\lambda_2}{m_1-m_0}+\frac{m_1-m_0}{2\sigma}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy}{1+(k-2)\int_{-\infty}^{\frac{\sigma\log\lambda_2}{m_1-m_0}-\frac{m_1-m_0}{2\sigma}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy}$$

Le même calcul permet de déterminer λ_2. On obtient alors :

$$\lambda_2=\frac{(k-1)+(2-k)\int_{-\infty}^{\frac{\sigma\log\lambda_1}{m_1-m_0}+\frac{m_1-m_0}{2\sigma}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy}{1+(k-2)\int_{-\infty}^{\frac{\sigma\log\lambda_1}{m_1-m_0}-\frac{m_1-m_0}{2\sigma}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy}$$

Une solution commune aux deux équations non linéaires couplées précédentes permet de déterminer λ_1 et λ_2. Cette résolution peut aboutir à plusieurs solutions, ces solutions doivent ensuite être étudiées de façon à ne retenir que la meilleure. Ces solutions peuvent être calculées pour différentes valeurs de k (Figure 13). Dans le cas où $m_0=0$, $m_1=1$, et $\sigma=1$, si $1\leq k<4.528$, il y a une solution qui est $\lambda_1=\lambda_2=1$. Si $k\geq 4.528$, il y a trois solutions. La première est $\lambda_1=\lambda_2=1$, mais cette solution ne minimise pas le risque moyen $\Re$. Les solutions qui permettent de minimiser le risque moyen $\Re$ consistent à affecter aux seuils λ_1 et λ_2 deux valeurs différentes appartenant à la courbe ci-dessous.

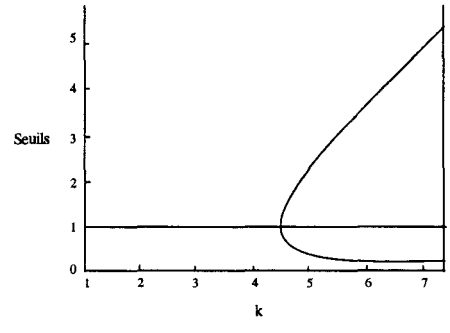


Figure 13 : Valeurs des seuils en fonction de k

• Exemple 2

Examinons maintenant le cas particulier où :

$C_{000} = C_{111} = 0$

$C_{010} = C_{100} = C_{011} = C_{101} = 1$

$C_{110} = C_{001} = 2$

(5.1.2.1.4) devient alors : $$\lambda_1 = \frac{P_0 \int_{y2} p(y_2/H_0)\{1 + p(u_2 = 0/y_2) \times 0\} dy_2}{P_1 \int_{y2} p(y_2/H_1)\{1 + p(u_2 = 0/y_2) \times 0\} dy_2}$$

$$\lambda_1 = \frac{P_0 \int_{y2} p(y_2/H_0)\, dy_2}{P_1 \int_{y2} p(y_2/H_1)\, dy_2}$$

$$\lambda_1 = \frac{P_0}{P_1} = \frac{P_0}{1-P_0} \quad \text{car} \quad \int_{y2} p(y_2/H_k)\, dy_2 = 1 \quad k=0,1 \qquad (5.1.2.1.5)$$

De même $\lambda_2 = \frac{P_0}{1-P_0}$

La règle de décision au niveau de chaque détecteur est alors la suivante :

$$\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i = \frac{P_0}{1-P_0} \qquad i=1,2$$

Les valeurs des seuils λ_1 et λ_2 peuvent être tracées en fonction de P_0 (Figure 14) :

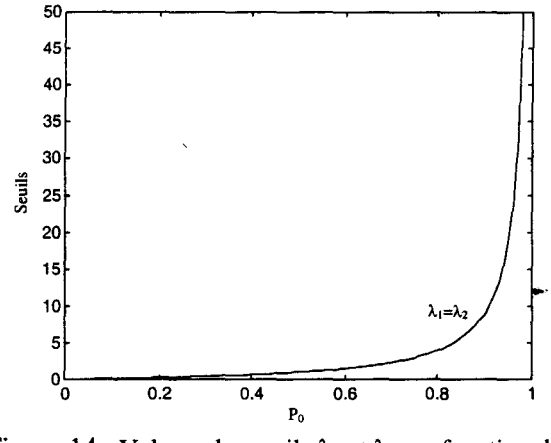


Figure 14 : Valeurs des seuils λ_1 et λ_2 en fonction de P_0

Les détecteurs locaux étant déterminés, le problème est alors d'optimiser l'opérateur de fusion qui prendra la décision finale.

5.1.2.2. Optimisation de l'opérateur de fusion

Dans le paragraphe précédent, nous avons considéré que les décisions étaient prises localement et n'étaient pas transmises à un opérateur de fusion. L'architecture de l'opérateur de fusion, qui a pour but de prendre une décision finale, n'a pas été étudiée.
Considérons à nouveau le problème de la détection décentralisée binaire avec les hypothèses H_0 et H_1 dont les probabilités *a priori* P_0 et P_1 sont connues. Le système est composé de N détecteurs locaux q_i i=1,...,N associés à N capteurs Y_i i=1,...,N tels que :

$$u_i = \begin{cases} 0, \text{ le détecteur i décide que } H_0 \text{ est vraie} \\ 1, \text{ le détecteur i décide que } H_1 \text{ est vraie} \end{cases}$$

Les décisions locales sont ensuite transmises à un opérateur de fusion qui les combine de façon à obtenir la décision finale u_0 (Figure 15).

$$u_0 = \begin{cases} 0, \text{on décide que } H_0 \text{ est vraie} \\ 1, \text{on décide que } H_1 \text{ est vraie} \end{cases}$$

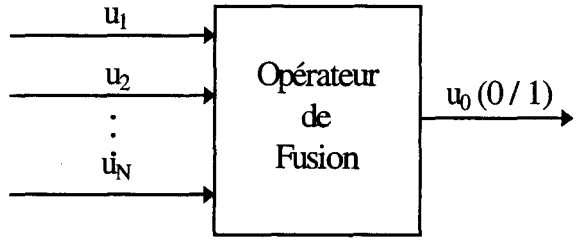


Figure 15 : L'opérateur de fusion

Chaque décision binaire transmise à l'opérateur de fusion est caractérisée par ses probabilités de fausse alarme P_{Fi} et de détection P_{Di} telles que :

$$P_{Fi} = P(u_i=1/H_0) \qquad \text{et} \qquad P_{Di} = P(u_i=1/H_1)$$

Les probabilités de fausse alarme et de détection du système complet sont notées P_F et P_D telles que :

$$P_F = P(u_0=1/H_0) \qquad \text{et} \qquad P_D = P(u_0=1/H_1)$$

Le problème est de déterminer l'opérateur de fusion f qui optimise, suivant un critère donné, la mise en commun des décisions prises par les différents détecteurs locaux [ChV86] [TVB87] [TVB89] afin d'obtenir la décision finale $u_0=f(u_1,u_2,\ldots,u_N)$. L'opérateur de fusion sera une fonction logique des N décisions binaires qui lui sont transmises. Dans le cas général, il y a 2^{2^N} opérateurs de fusion f possibles. Dans le cas particulier de deux capteurs il y en a donc 16, dont le « et logique » (f_2) et le « ou logique » (f_8) qui ne sont que des cas très particuliers (Figure 16).

Entrées		Décision finale															
u_1	u_2	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Figure 16 : Les fonctions de fusion possibles pour 2 capteurs

Quelques-unes de ces seize fonctions peuvent *a priori* être écartées. En fait, pour être acceptable il a été démontré que la fonction de fusion doit être monotone [ReN87] [TVB89], ce qui réduit considérablement la taille de l'ensemble des fonctions de fusion admissibles (Figure 17). En effet pour être optimal, en supposant que l'on a $P_{Di} \geq P_{Fi}$ (i=1,...,N), l'opérateur de fusion doit nécessairement satisfaire à la condition suivante :

$$\forall\, u^*=(u_1^*,...,u_N^*) \text{ et } \forall\, u^{*'}=(u_1^{*'},...,u_N^{*'}),\ \text{si}\ \forall\, i\in\{1,...,N\}\ u_i^* \geq u_i^{*'}\ \text{alors}\ f(u^*) \geq f(u^{*'})$$

où :
- $u^*=(u_1^*,...,u_N^*)$ est un vecteur de décisions prises par les détecteurs locaux.
- $u^{*'}=(u_1^{*'},...,u_N^{*'})$ est un autre vecteur de décisions prises par les détecteurs locaux.

Cette condition nous permet d'écarter f_3, f_5, f_7, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, et f_{15} des fonctions de fusions admissibles, pour obtenir le tableau ci-dessous.

Entrées		Décision finale					
u_1	u_2	f_1	f_2	f_4	f_6	f_8	f_{16}
0	0	0	0	0	0	0	1
0	1	0	0	0	1	1	1
1	0	0	0	1	0	1	1
1	1	0	1	1	1	1	1

Figure 17 : Les fonctions de fusion admissibles pour 2 capteurs

Considérons maintenant l'optimisation Bayésienne de l'opérateur de fusion. L'objectif est de déterminer la règle de fusion qui minimise le risque moyen $\Re$. Nous avons démontré dans le paragraphe 4.1 que minimiser ce risque moyen est équivalent à utiliser le test du rapport de vraisemblance suivant :

$$\frac{p(u_1,u_2,...,u_N/H_1)}{p(u_1,u_2,...,u_N/H_0)} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})} = \lambda \qquad (5.1.2.2.1)$$

Dans le cas où les décisions locales sont indépendantes, le terme de gauche devient :

$$\frac{p(u_1,u_2,...,u_N/H_1)}{p(u_1,u_2,...,u_N/H_0)} = \prod_{i=1}^{N} \frac{p(u_i/H_1)}{p(u_i/H_0)}$$

Et pour un vecteur de décisions locales $(u_1,u_2,\ldots, u_N)$ donné, on obtient la relation suivante :

$$\frac{p(u_1,u_2,\ldots,u_N/H_1)}{p(u_1,u_2,\ldots,u_N/H_0)}=\prod_{i=1}^{N}\frac{p(u_i/H_1)}{p(u_i/H_0)}=\prod_{S_1}\frac{P(u_i=1/H_1)}{P(u_i=1/H_0)}\prod_{S_0}\frac{P(u_i=0/H_1)}{P(u_i=0/H_0)}$$

où S_j est l'ensemble de toutes les décisions locales u_i (i=1,...,N) égales à j (j=0,1).

$$\frac{p(u_1,u_2,\ldots,u_N/H_1)}{p(u_1,u_2,\ldots,u_N/H_0)}=\prod_{S_1}\frac{P_{Di}}{P_{Fi}}\prod_{S_0}\frac{1-P_{Di}}{1-P_{Fi}} \qquad (5.1.2.2.2)$$

En prenant le logarithme de l'équation (5.1.2.2.2), la règle de fusion (5.1.2.2.1) peut s'écrire :

$$\sum_{S_1}\log\frac{P_{Di}}{P_{Fi}}+\sum_{S_0}\log\frac{1-P_{Di}}{1-P_{Fi}}\underset{u_0=0}{\overset{u_0=1}{\gtrless}}\log\lambda$$

soit : $$\sum_{i=1}^{N}\left[\,u_i\log\frac{P_{Di}}{P_{Fi}}+(1-u_i)\log\frac{1-P_{Di}}{1-P_{Fi}}\,\right]\underset{u_0=0}{\overset{u_0=1}{\gtrless}}\log\lambda$$

Et après un réarrangement, on obtient :

$$\sum_{i=1}^{N}\left[\log\frac{P_{Di}(1-P_{Fi})}{P_{Fi}(1-P_{Di})}\right]u_i\underset{u_0=0}{\overset{u_0=1}{\gtrless}}\log\left[\lambda\prod_{i=1}^{N}\frac{1-P_{Fi}}{1-P_{Di}}\right] \qquad (5.1.2.2.3)$$

La règle de fusion optimale peut donc s'écrire comme une somme pondérée des décisions locales, somme que l'on compare ensuite à un seuil qui s'exprime comme une fonction des probabilités de fausse alarme et de détection des différents détecteurs locaux, des probabilités *a priori* P_0 et P_1 et des différents coûts (voir 5.1.2.2.1).

- **Exemple 1 [Var97]**

A titre d'exemple, pour un ensemble constitué de deux capteurs, nous avons posé $P_{Fi}=0.1$ et $P_{Di}=0.9$ (i=1,2), et fixé les différents coûts en prenant $C_{00}=C_{11}=0$ et $C_{01}=C_{10}=1$. Finalement, la règle de fusion optimale est déterminée à partir de (5.1.2.2.3) pour différentes valeurs de P_0 (Figure 18).

Valeurs prises par P_0	règle de fusion optimale
$0<P_0<0.012$	f_{16}
$0.012<P_0<0.5$	f_8 (OU)
$0.5<P_0<0.988$	f_2 (ET)
$0.988<P_0<1$	f_1

Figure 18 : La règle de fusion optimale en fonction de P_0

• Exemple 2

Considérons un système formé de deux capteurs dont les observations y_1 et y_2 sont indépendantes. Les fonctions de densité de probabilités sous chaque hypothèse sont des gaussiennes. Sous l'hypothèse H_0 ces fonctions de densité de probabilité sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1 ces fonctions sont supposées être de variance 1, et respectivement de moyenne m_1 pour Y_1 et m_2 pour Y_2.

La règle de décision optimale au niveau de chaque détecteur est, d'après l'exemple 2 du paragraphe 5.1.2.1 :

$$\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i \quad i=1,2 \qquad \text{avec } \lambda_1=\lambda_2=\frac{P_0}{1-P_0}$$

Deux systèmes ont été étudiés, le premier pour lequel $m_1=m_2=1$ et le deuxième pour lequel $m_1=1$ et $m_2=1,5$. Dans le cas où $C_{00}=C_{11}=0$ et $C_{01}=C_{10}=1$, le risque moyen $\Re$ est en fait une probabilité d'erreur. Les probabilités d'erreurs ont été tracées en fonction de P_0 pour les différentes fonctions de fusion possibles (Figure 19 et Figure 21). Les figures 20 et 22 représentent les courbes C.O.R. pour chaque système en fonction de la fonction de fusion utilisée. Sur les figures 19 et 21, la règle de fusion qui minimise la probabilité d'erreur (ou le risque moyen) dépend de P_0. Dans le cas où $m_1=m_2=1$, le « ou logique » est optimal pour $0<P_0<0.5$, et le « et logique » l'est pour $0.5<P_0<1$. Dans le cas où $m_1=1$ et $m_2=1.5$, la décision prise par le capteur 2 (fonction f_6) est optimale quelque soit la valeur prise par P_0.

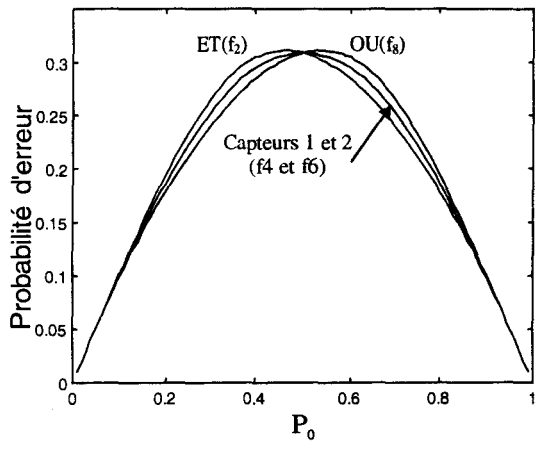


Figure 19 : Probabilités d'erreur en fonction de P_0 pour $m_1=m_2=1$

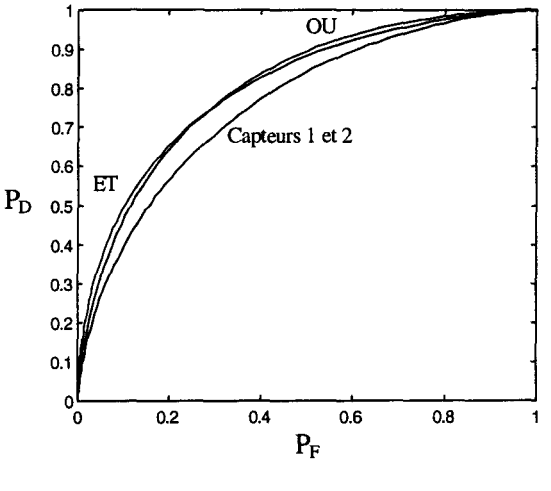


Figure 20 : Courbes COR pour $m_1=m_2=1$

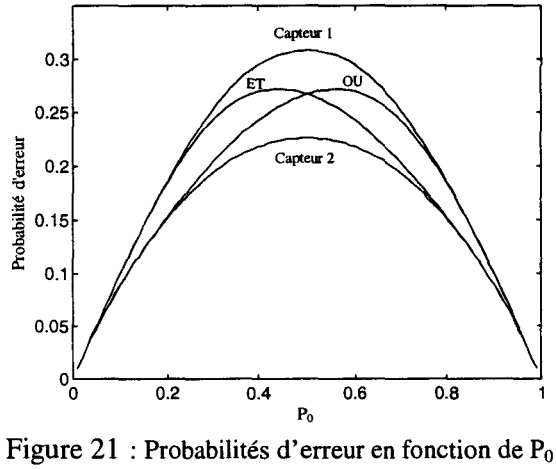


Figure 21 : Probabilités d'erreur en fonction de P_0 pour $m_1=1$, $m_2=1.5$

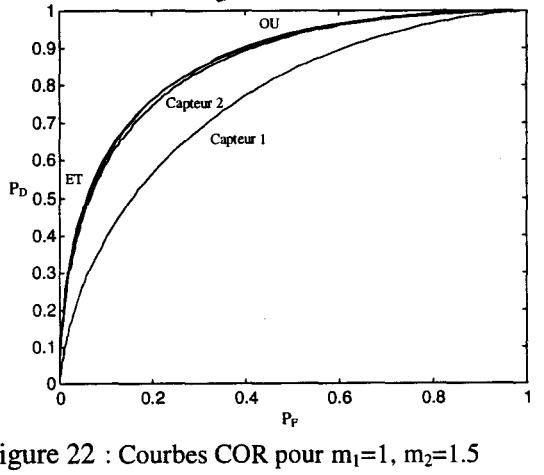


Figure 22 : Courbes COR pour $m_1=1$, $m_2=1.5$

• **Remarques**

Dans le cas où les décisions locales sont corrélées, la règle de fusion optimale peut aussi être déterminée à partir de l'équation (5.1.2.2.1), mais elle doit faire apparaître les corrélations entre les décisions locales. Drakopoulos et Lee ont montré [DrL91] que dans ce cas la règle de fusion optimale est donnée par :

$$\frac{\sum_{I\subseteq S_0}(-1)^{|I|}E_1\left(\prod_{i\in S_1\cup I}u_i\right)}{\sum_{I\subseteq S_0}(-1)^{|I|}E_0\left(\prod_{i\in S_1\cup I}u_i\right)} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \lambda, \qquad (5.1.2.2.4)$$

où :

- S_j est l'ensemble de toutes les décisions locales u_i (i=1,...,N) égales à j ; j=0,1.
- $I\subseteq\{1, 2, ..., N\}$ et $I\neq\varnothing$
- |I| est le cardinal de l'ensemble I
- E_j est l'espérance sous l'hypothèse H_j

Lorsque les décisions locales sont indépendantes, l'équation (5.1.2.2.4) se simplifie pour aboutir à l'équation (5.1.2.2.3).

Kam et Zhu ont employé une autre approche [ChK92], appelée la « généralisation de Bahadur-Lazarfeld » [DuH73] des fonctions de densité de probabilités, pour montrer que la règle de fusion optimale peut s'exprimer comme une généralisation de l'équation (5.1.2.2.3). Ils montrent ainsi que l'opérateur de fusion peut se mettre sous la forme :

$$\sum_{i=1}^{N}\left[\log\frac{P_{Di}(1-P_{Fi})}{P_{Fi}(1-P_{Di})}\right]u_i+\sum_{i=1}^{N}\log\left(\frac{1-P_{Di}}{1-P_{Fi}}\right)+\log\frac{1+\sum_{i<j}K_{ij}^1z_i^1z_j^1+\sum_{i<j<k}K_{ijk}^1z_i^1z_j^1z_k^1+...+K_{12...N}^1z_1^1z_2^1z_N^1}{1+\sum_{i<j}K_{ij}^0z_i^0z_j^0+\sum_{i<j<k}K_{ijk}^0z_i^0z_j^0z_k^0+...+K_{12...N}^0z_1^0z_2^0z_N^0}$$

$$\underset{u_0=0}{\overset{u_0=1}{\gtrless}} \log\lambda \qquad (5.1.2.2.5)$$

où :

$$z_i^h=\frac{u_i-P(u_i/H_h)}{\sqrt{P(u_i=1/H_h)[1-P(u_i=1/H_h)]}} \qquad h=0,1.$$

$$K_{ij}^h=\sum_u z_i^h z_j^h P(u/H_h)$$

$$K_{ijk}^h=\sum_u z_i^h z_j^h z_k^h P(u/H_h)$$

...

$$K_{12...N}^h=\sum_u z_1^h z_2^h ... z_N^h P(u/H_h)$$

Dans de nombreuses situations, les coefficients de corrélation deviennent nuls lorsque l'on arrive à un certain ordre et l'équation (5.1.2.2.5) devient alors un peu moins complexe.

5.1.2.3. Optimisation simultanée des détecteurs locaux et de l'opérateur de fusion

Les deux composantes du problème de la détection parallèle Bayésienne (l'optimisation des détecteurs locaux et l'optimisation de l'opérateur de fusion) ont été étudiées de façon indépendante dans les deux paragraphes précédents. Dans ce paragraphe, nous présentons l'optimisation de ces deux composantes de façon simultanée [Hob86] [Sri86] [ReN87a] [TVB87] [HoV89a] [Tan90] [Tan91a]. Le système considéré est composé de N détecteurs locaux q_i i=1,...,N associés à N capteurs Y_i i=1,...,N qui observent le même phénomène. Les observations associées à ces capteurs sont notées y_i i=1,...,N. Les probabilités conditionnelles $p(y_1,...,y_N/H_j)$ sont supposées connues. Il n'y a aucune transmission d'information entre les détecteurs. A partir de ses propres observations y_i, chaque détecteur prend une décision locale u_i. Les décisions locales sont ensuite transmises à un opérateur de fusion qui les combine de façon à obtenir la décision finale u_0 (Figure 23). La décision finale u_0 est supposée ne dépendre que des décisions locales u, $u=(u_1,...,u_N)$, et être indépendante des observations y_i faites par chaque capteur. Chaque détecteur local est caractérisé par ses probabilités de fausse alarme P_{Fi} et de détection P_{Di}. Les probabilités de fausse alarme et de détection du système complet sont, quant à elles, notées P_F et P_D.

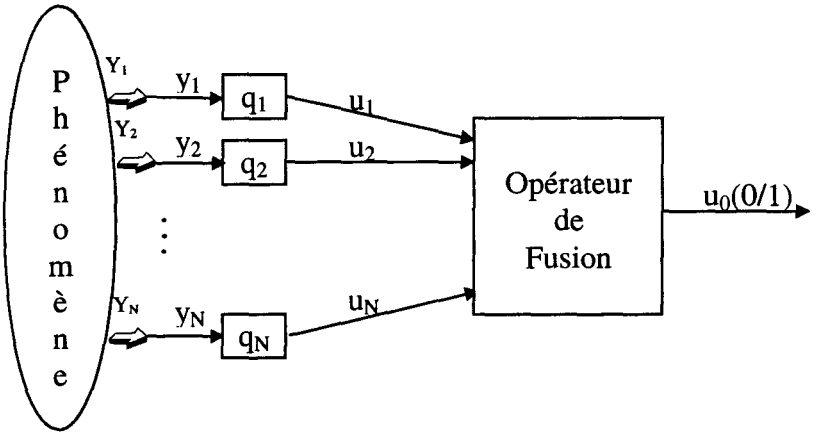


Figure 23 : Architecture de la détection décentralisée parallèle

• Le critère d'optimisation

L'approche Bayésienne pour l'optimisation de ce système consiste à déterminer l'ensemble des règles de décisions $\Gamma=\{\gamma_0,\gamma_1,...,\gamma_N\}$ qui minimise le risque moyen $\Re(\Gamma)$ associé au système global. Ainsi on passe de l'espace des observations à la décision en utilisant les règles suivantes :

$$u_i=\gamma_i(y_i), \qquad i=1,...,N$$
$$\text{et} \qquad (5.1.2.3.1)$$
$$u_0=\gamma_0(u_1,...,u_N)$$

Le risque moyen $\Re$ associé à Γ peut s'écrire :

$$\Re = \sum_{i,j\in\{0,1\}} C_{ij}\, P(u_0=i,H_j) \qquad (5.1.2.3.2)$$

$$\Re = \sum_{i,j\in\{0,1\}} C_{ij}\, P_j\, P(u_0=i/H_j)$$

Sachant que :

$$P_F=P(u_0=1/H_0)=\sum_u P(u_0=1/u)P(u/H_0) \qquad \text{et}$$

$$P_D=P(u_0=1/H_1)=\sum_u P(u_0=1/u)P(u/H_1)$$

où $\sum_u$ représente une sommation sur toutes les valeurs possibles prises par u, (5.1.2.3.2) peut s'écrire :

$$\Re = C + C_F\sum_u P(u_0=1/u)P(u/H_0) - C_D\sum_u P(u_0=1/u)P(u/H_1) \qquad (5.1.2.3.3)$$

où :

$$C_F = P_0(C_{10}-C_{00}) > 0$$
$$C_D = (1-P_0)(C_{01}-C_{11}) > 0$$
$$C = C_{01}(1-P_0) + C_{00}P_0$$

Optimiser le système de décision revient alors à déterminer les règles de fusion qui minimisent $\Re$.

• Minimisation du risque moyen $\Re$

L'optimisation de ce système se fait élément par élément. Le problème de détection décentralisée de la Figure 23 peut être vu comme une coopération entre deux sous-systèmes (Figure 24) : l'opérateur de fusion et l'ensemble des détecteurs locaux qui peut, lui aussi, être fractionné en sous-systèmes plus simples. Lorsque l'on optimise l'un des éléments du système, on considère que les autres ont déjà été optimisés et qu'ils sont fixés. L'optimisation élément par élément conduit à des équations qui sont en général des conditions nécessaires d'optimalité mais pas des conditions suffisantes.

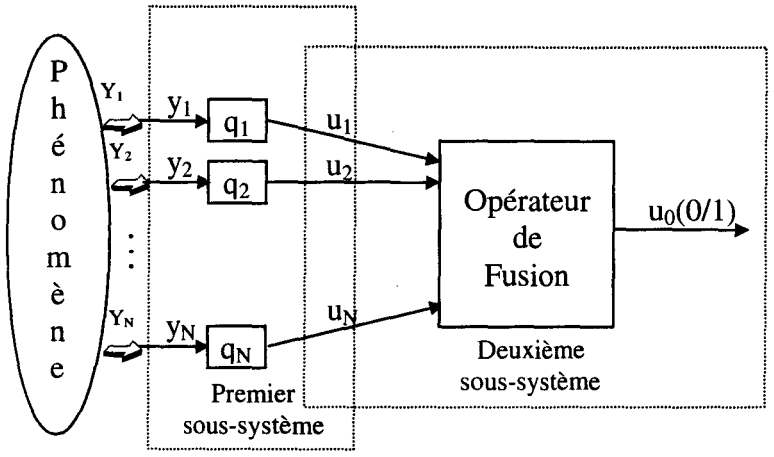


Figure 24 : Séparation du système de détection décentralisée en plusieurs sous-systèmes

• Optimisation des détecteurs locaux

Supposons fixés l'opérateur de fusion ainsi que tous les détecteurs locaux, sauf le détecteur local q_k que l'on va chercher à optimiser.

En développant suivant le détecteur k, (5.1.2.3.3) peut s'écrire :

$$\Re = C + \sum_{u^k} P(u_0=1/u^{k1}) \left[C_F P(u^{k1}/H_0) - C_D P(u^{k1}/H_1)\right] \qquad (5.1.2.3.4)$$
$$+ \sum_{u^k} P(u_0=1/u^{k0}) \left[C_F P(u^{k0}/H_0) - C_D P(u^{k0}/H_1)\right]$$

où $u^k=(u_1, ..., u_{k-1}, u_{k+1}, ..., u_N)^T$
et $u^{kj}=(u_1, ..., u_{k-1}, u_k=j, u_{k+1}, ..., u_N)^T$ $\qquad j=0,1$

sachant que $P(u^{k0}/H_j)=P(u^k/H_j)-P(u^{k1}/H_j)$, l'équation (5.1.2.3.4) peut s'écrire :

$$\Re = C_k + \sum_{u^k} A(u^k)\left[C_F P(u^{k1}/H_0) - C_D P(u^{k1}/H_1)\right] \qquad (5.1.2.3.5)$$

où :

- $C_k = C + \sum_{u^k} P(u_0=1/u^{k0}) \left[C_F P(u^k/H_0) - C_D P(u^k/H_1)\right]$
- $A(u^k) = P(u_0=1/u^{k1}) - P(u_0=1/u^{k0})$

De plus : $$P(u^{ki}/H_j) = \int_y P(u^{ki}/y)\, p(y/H_j)\, dy \quad i=0,1 \qquad (5.1.2.3.6)$$

où :

- $y = (y_1, ..., y_N)^T$
- $\int_y$ représente une intégration multi-variables suivant toutes les composantes de y.

Et puisque chaque décision locale ne dépend que de ses propres observations, on a :

$$P(u/y) = \prod_{i=1}^{N} P(u_i/y_i) = P(u_k/y_k) \,.\, P(u^k/y^k)$$

Par conséquent, on a :

$$P(u^{ki}/y) = P(u_k=i/y_k)\, P(u^k/y^k) \quad i=0,1 \tag{5.1.2.3.7}$$

où $y^k = (y_1, ..., y_{k-1}, y_{k+1}, ..., y_N)^T$

$P(u^{ki}/H_j)$ peut donc s'écrire en utilisant (5.1.2.3.6) et (5.1.2.3.7) :

$$P(u^{ki}/H_j) = \int_y P(u_k=i/y_k)\, P(u^k/y^k)\, p(y/H_j)\, dy \tag{5.1.2.3.8}$$

En utilisant (5.1.2.3.8), l'équation (5.1.2.3.5) peut donc s'écrire :

$$\Re = C_k + \int_{y_k} P(u_k = 1/y_k) dy_k \times \left\{ \sum_{u^k} \int_{y^k} A(u^k) P(u^k/y^k) \left[C_F p(y/H_0) - C_D p(y/H_1) \right] dy^k \right\}$$

Si l'on suppose que tous les détecteurs autres que q_k sont fixés, alors C_k est une constante et minimiser $\Re$ revient à décider, pour une observation y_k donnée :

$$P(u_k=1/y_k) = \begin{cases} 1, & \text{si } D(k) \le 0 \\ 0, & \text{sinon} \end{cases} \tag{5.1.2.3.9}$$

où $D(k) = \sum_{u^k} \int_{y^k} A(u^k)\, P(u^k/y^k) \left[C_F p(y/H_0) - C_D p(y/H_1) \right] dy^k$

sachant que $p(y/H_j) = p(y^k/y_k, H_j).p(y_k/H_j)$, la règle de décision pour le détecteur q_k est donc :

$$p(y_k/H_1) \sum_{u^k} \int_{y^k} A(u^k)\, C_D\, P(u^k/y^k)\, p(y^k/y_k, H_1)\, dy^k$$

$$\begin{matrix} u_k=1 \\ > \\ < \\ u_k=0 \end{matrix} \tag{5.1.2.3.10}$$

$$p(y_k/H_0) \sum_{u^k} \int_{y^k} A(u^k)\, C_F\, P(u^k/y^k)\, p(y^k/y_k, H_0)\, dy^k$$

Les autres détecteurs locaux peuvent être optimisés de la même façon : On considère que l'opérateur de fusion ainsi que tous les détecteurs locaux sont fixés, sauf le détecteur q_l et on l'optimise avec la règle précédente, en remplaçant « k » par « l ».

Il nous reste maintenant à considérer que tous les détecteurs locaux sont figés, ce qui nous permettra d'optimiser l'opérateur de fusion :

• **Optimisation de l'opérateur de fusion**

Intéressons nous maintenant à l'optimisation de l'opérateur de fusion. Comme précédemment, les détecteurs locaux sont supposés fixés. Puisque les éléments de u sont à valeurs binaires il y a 2^N valeurs possibles prises par u. Soit u^* l'une de ces valeurs. (5.1.2.3.3) peut s'écrire :

$$\Re = P(u_0=1/u^*)\,[C_F P(u^*/H_0) - C_D P(u^*/H_1)] + K(u^*) \qquad (5.1.2.3.11)$$

où :

$$K(u^*) = C + \sum_{\substack{u \\ u \neq u^*}} P(u_0=1/u)\,[C_F P(u/H_0) - C_D P(u/H_1)]$$

Pour un u^* donné, $K(u^*)$ est fixé. Minimiser $\Re$ revient donc à employer la règle de décision suivante :

$$P(u_0=1/u^*) = \begin{cases} 1, \text{ si } [C_F P(u^*/H_0) - C_D P(u^*/H_1)] \leq 0 \\ 0, \text{ sinon} \end{cases}$$

La règle de décision pour l'opérateur de fusion est donc :

$$\frac{P(u^*/H_1)}{P(u^*/H_0)} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{C_F}{C_D} \qquad (5.1.2.3.12)$$

Optimiser l'opérateur de fusion revient ainsi à résoudre 2^N équations.

En conclusion, optimiser le système revient donc à résoudre 2^N équations de la forme (5.1.2.3.12) et N équations de la forme (5.1.2.3.10). Une solution commune à ces 2^N+N équations non linéaires couplées est la solution de l'optimisation Bayésienne élément par élément du problème de détection décentralisée de la Figure 23. Le nombre d'équations à résoudre simultanément croît donc très rapidement avec le nombre de capteurs, les calculs nécessaires à la résolution de ces équations deviennent donc très vite prohibitifs. Dans certains cas, ces calculs peuvent être simplifiés. C'est le cas lorsque les observations locales sont indépendantes entre elles.

• **Cas où les observations locales sont indépendantes**

Dans ce cas, on peut écrire : $P(Y/H_j) = \prod_{i=1}^{N} p(y_i/H_j)$, j=0,1

et l'équation (5.1.2.3.10) devient, après simplification :

$$\frac{p(y_k/H_1)}{p(y_k/H_0)} \underset{u_k=0}{\overset{u_k=1}{\gtrless}} \frac{\sum_{u^k} C_F A(u^k) \prod_{i=1, i\neq k}^{N} P(u_i/H_0)}{\sum_{u^k} C_D A(u^k) \prod_{i=1, i\neq k}^{N} P(u_i/H_1)} \qquad (5.1.2.3.13)$$

Le terme de droite est une constante et par conséquent un réel seuillage pour déterminer u_k est effectué, ce qui n'était en général pas le cas avec l'équation (5.1.2.3.10). De même, l'équation (5.1.2.3.12) concernant la fusion peut être simplifiée :

$$\prod_{i=1}^{N} \frac{P(u_i/H_1)}{P(u_i/H_0)} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{C_F}{C_D} \qquad (5.1.2.3.14)$$

Dans le cas où les observations locales sont indépendantes, le nombre d'équations couplées ne varie donc pas, mais la difficulté de résolution est beaucoup moins grande.

• Exemple [Var97]

En considérant deux capteurs dont les observations sont indépendantes, une expression explicite des seuils et de l'opérateur de fusion peut être obtenue. Ainsi le seuil au niveau du premier détecteur est, d'après (5.1.2.3.13) :

$$\lambda_1 = \frac{\sum_{u_2} C_F \, A(u_2) \, P(u_2/H_0)}{\sum_{u_2} C_D \, A(u_2) \, P(u_2/H_1)}$$

De plus, $A(u_2) = P(u_0=1/u_1=1,u_2) - P(u_0=1/u_1=0,u_2)$; et en posant $P_{ijk}=P(u_0=i/u_1=j,u_2=k)$, on a :

$$\lambda_1 = \frac{C_F}{C_D} \frac{(P_{110}-P_{100})(1-P_{F2})+(P_{111}-P_{101})P_{F2}}{(P_{110}-P_{100})(1-P_{D2})+(P_{111}-P_{101})P_{D2}}$$

$$\lambda_1 = \frac{C_F}{C_D} \frac{(P_{100}-P_{110}+P_{111}-P_{101})P_{F2}+(P_{110}-P_{100})}{(P_{100}-P_{110}+P_{111}-P_{101})P_{D2}+(P_{110}-P_{100})}$$

De même : $$\lambda_2 = \frac{C_F}{C_D} \frac{(P_{100}-P_{101}+P_{111}-P_{110})P_{F1}+(P_{101}-P_{100})}{(P_{100}-P_{101}+P_{111}-P_{110})P_{D1}+(P_{101}-P_{100})}$$

De plus, dans ce cas, le vecteur u peut prendre quatre valeurs différentes suivant les valeurs de u_1 et de u_2. L'équation (5.1.2.3.10) se traduit donc par l'application d'une des quatre relations suivantes :

$$\frac{P_{D1}}{P_{F1}} \frac{P_{D2}}{P_{F2}} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{C_F}{C_D} \qquad \frac{1-P_{D1}}{1-P_{F1}} \frac{P_{D2}}{P_{F2}} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{C_F}{C_D}$$

$$\frac{P_{D1}}{P_{F1}} \frac{1-P_{D2}}{1-P_{F2}} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{C_F}{C_D} \qquad \frac{1-P_{D1}}{1-P_{F1}} \frac{1-P_{D2}}{1-P_{F2}} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{C_F}{C_D}$$

Les six équations précédentes sont couplées et leur résolution permet de trouver une solution au problème de la détection décentralisée parallèle.

Dans le cas où les fonctions de densité de probabilités conditionnelles associées aux différentes situations sont des gaussiennes, ce système d'équations a pu être résolu. Sous l'hypothèse H_0, ces fonctions sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1 ces fonctions sont supposées être de moyenne m_1 et m_2 et de variance 1. A titre d'exemple, deux systèmes ont été étudiés, le premier avec $m_1=m_2=1$ et l'autre avec $m_1=1$ et $m_2=1,5$. La règle de fusion optimale dépend alors de P_0 et est résumée dans les tableaux 25 et 26. Les seuils optimaux des deux détecteurs, ainsi que les probabilités d'erreurs qui leur sont associés sont indiqués dans les figures 27, 29, 30 et 32. Les figures 28 et 31 représentent les courbes C.O.R. pour chaque système.

Valeurs prises par P_0	règle de fusion optimale
$0< P_0<0.5$	f_8 (ou logique)
$0.5< P_0<1$	f_2 (et logique)

Figure 25 : La règle de fusion optimale en fonction de P_0 pour $m_1=m_2=1$

Valeurs prises par P_0	règle de fusion optimale
$0 < P_0< 0.185$	f_6 (décision prise par q_2)
$0.185 < P_0< 0.5$	f_8 (ou logique)
$0.5 < P_0< 0.815$	f_2 (et logique)
$0.815 < P_0< 1$	f_6 (décision prise par q_2)

Figure 26 : La règle de fusion optimale en fonction de P_0 pour $m_1=1$ et $m_2=1,5$

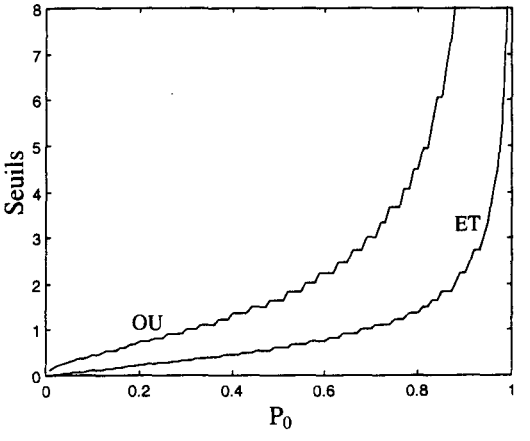


Figure 27 : Valeurs des seuils en fonction de P_0 pour $m_1=m_2=1$

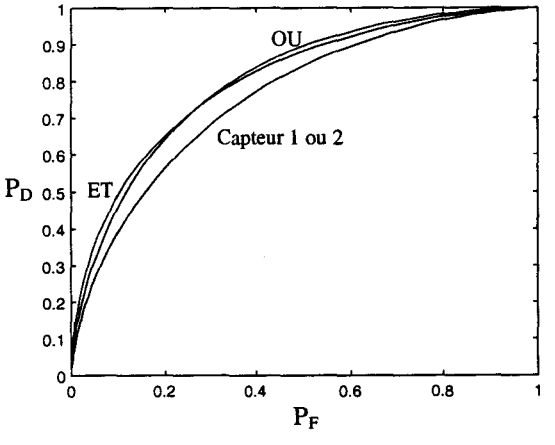


Figure 28 : Courbes C.O.R. pour $m_1=m_2=1$

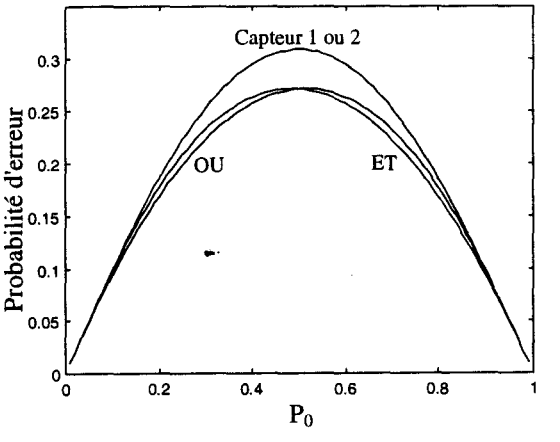


Figure 29 : Probabilités d'erreur en fonction de P_0 pour $m_1=m_2=1$

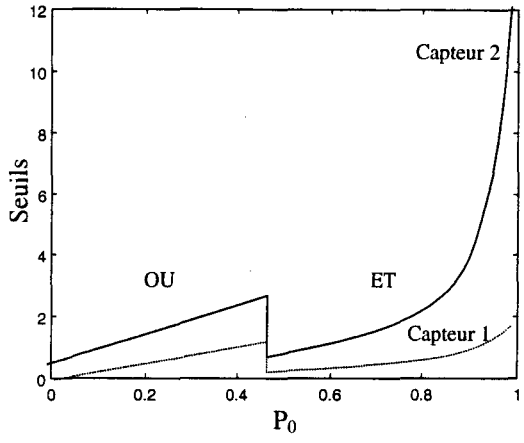


Figure 30 : Valeurs des seuils en fonction de P_0 pour $m_1=1$ et $m_2=1,5$

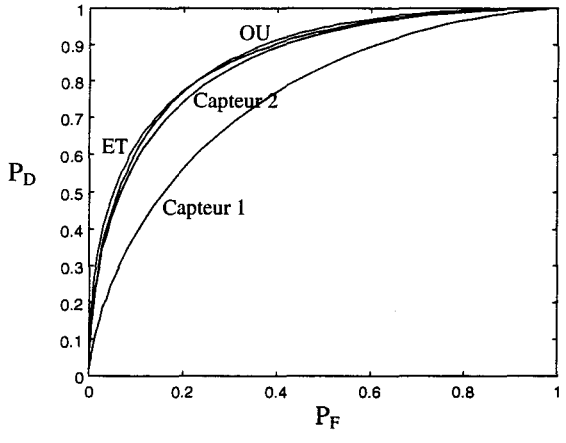


Figure 31 : Courbes C.O.R. pour $m_1=1$ et $m_2=1,5$

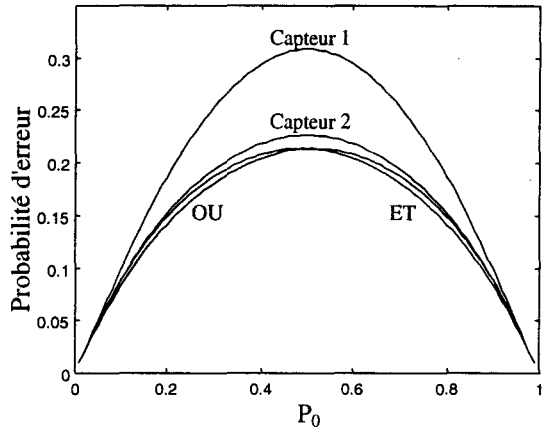


Figure 32 : Probabilités d'erreur en fonction de P_0 pour $m_1=1$ et $m_2=1,5$

Le même type de calcul peut être fait pour trois capteurs. Le cas où les fonctions de densité de probabilités conditionnelles associées aux différentes situations sont des gaussiennes a été étudié. Sous l'hypothèse H_0, ces fonctions sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1, ces fonctions sont supposées être de moyenne $m_1=m_2=m_3=1$ et de variance 1. Les seuils sont supposés être les mêmes pour tous les détecteurs, les valeurs de ces seuils sont indiqués Figure 33. La probabilité d'erreur en fonction de P_0 est indiquée Figure 34. Enfin, les performances des systèmes de détection en fonction du nombre de capteurs sont représentées Figure 35.

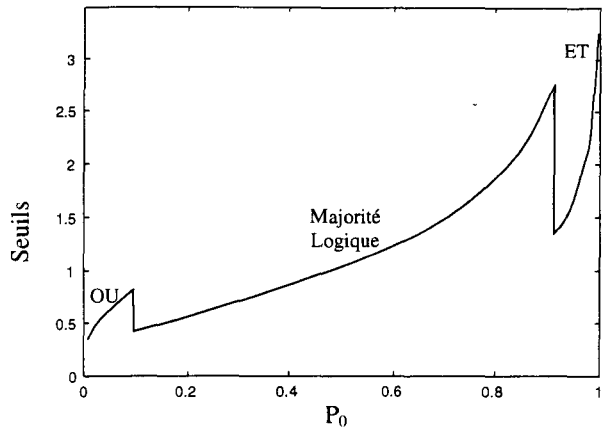


Figure 33 : Règle de fusion optimale et valeurs des seuils en fonction de P_0 pour $m_1=m_2=m_3=1$

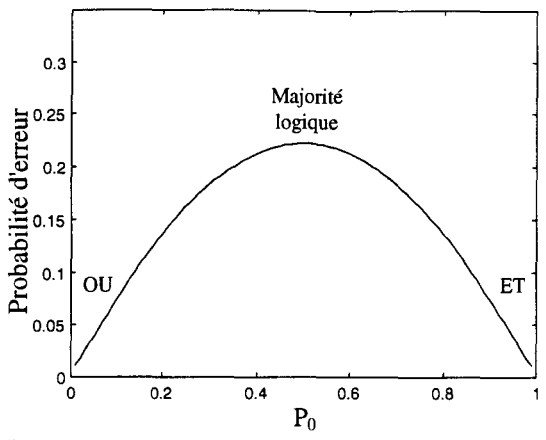


Figure 34 : Probabilités d'erreur en fonction de P_0 pour $m_1=m_2=m_3=1$

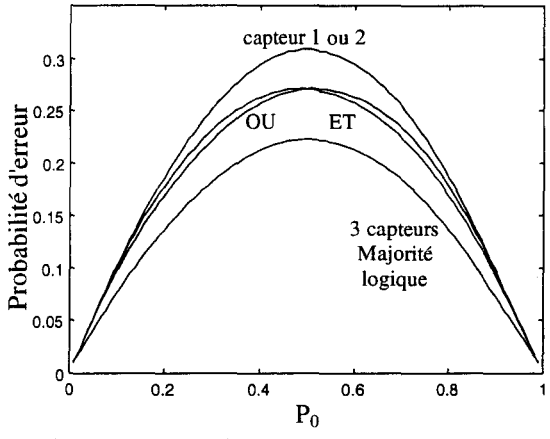


Figure 35 : Probabilités d'erreur en fonction du nombre de capteurs pour $m_1=m_2=m_3=1$

• Remarques

- La courbe C.O.R., associée au système qui résulte de l'optimisation élément par élément en utilisant la procédure décrite ci-dessus, n'est pas forcément concave.

- Lorsque les observations locales sont distribuées de façon identique, l'optimisation du système ne conduit pas forcément à des détecteurs locaux identiques. Les seuils utilisés au niveau de ces détecteurs peuvent être différents [Tsi86] [WiW91] [ChK92].

- Lorsque les observations locales sont distribuées de façon identique, et que l'on force les détecteurs locaux à être identiques ; il a été montré [ReN87a] que l'opérateur de fusion optimal se réduit à une règle pour laquelle, si K détecteurs parmi N décident 1, alors l'opérateur de fusion décide H_1. K étant fixé en fonction de P_0, des coûts associés aux différentes situations, et des probabilités de fausse alarme et de détection communes à tous les détecteurs.

- Différents algorithmes pour résoudre le problème de l'optimisation élément par élément ont été développés dans la littérature [Tan90] [Hel95b].

- Les méthodes développées ci-dessus ne sont applicables que dans le cas où les statistiques associées aux différentes observations sont connues.

5.1.3. Le point de vue de Neyman-Pearson

Pour utiliser cette approche, il n'est pas nécessaire de connaître les probabilités *a priori* associées aux différentes hypothèses, ni les différents coûts associés à chaque situation. La probabilité de détection P_D est maximisée pour une valeur de la probabilité de fausse alarme $P_F=\alpha$ fixée. Aussi, seul le système entier de la Figure 23 pourra être optimisé ici, puisque la probabilité de détection du système complet doit pouvoir être calculée. Le système considéré est composé de N détecteurs locaux q_i associés à N capteurs Y_i (i=1,...,N) qui observent le même phénomène. A partir de ses propres observations y_i, chaque détecteur prend une décision locale u_i. Les décisions locales sont ensuite transmises à un opérateur de fusion qui les combine de façon à obtenir la décision finale u_0. Nous utiliserons pour la suite, les mêmes notations qu'au paragraphe 5.1.2.3.
Comme nous l'avons vu au paragraphe 4.2 la règle de décision qui maximise P_D pour une valeur fixée de P_F constitue un seuillage (qu'il faudra se fixer) du rapport de vraisemblance. Les probabilités P_F et P_D peuvent être exprimées en fonction de P_{Fi} et P_{Di} telles que :

$$P_D = P(u_0=1/H_1) = \sum_u P(u_0=1/u)\, P(u/H_1)$$

$$P_F = P(u_0=1/H_0) = \sum_u P(u_0=1/u)\, P(u/H_0)$$

où $\sum_u$ représente une sommation sur toutes les valeurs possibles prises par u.

Les composantes de u étant indépendantes, pour chaque valeur $u^* = (u_1^*,u_2^*,\ldots,u_N^*)$ fixée prise par u, on a les relations suivantes :

$$P(u^*/H_1) = \prod_{S_0} (1-P_{Di}) \prod_{S_1} P_{Di}$$

$$P(u^*/H_0) = \prod_{S_0} (1-P_{Fi}) \prod_{S_1} P_{Fi}$$

où S_j est l'ensemble de toutes les décisions locales u_i^* égales à j, j=0,1.

• Optimisation de l'opérateur de fusion

Afin d'optimiser l'opérateur de fusion, on utilise la méthode du multiplicateur de Lagrange et on pose :

$$\mathbf{L} = P_D - \lambda.(P_F-\alpha) \qquad (5.1.3.1)$$

$$\mathbf{L} = \sum_u P(u_0=1/u).P(u/H_1) - \lambda.(\sum_u P(u_0=1/u).P(u/H_0) - \alpha)$$

$$\mathbf{L} = \lambda.\alpha + \sum_{u} P(u_0=1/u) \,.\, [P(u/H_1) - \lambda.P(u/H_0)] \qquad (5.1.3.2)$$

Le premier terme est fixé. Pour maximiser **L** il faut maximiser le terme « somme » ; ce qui se traduit, pour chaque valeur u^* fixée prise par u, par la règle de décision suivante :

$$P(u_0=1/u^*) = \begin{cases} 0, \text{ si } P(u^*/H_1) - \lambda\, P(u^*/H_0) < 0 \\ 1, \text{ si } P(u^*/H_1) - \lambda\, P(u^*/H_0) > 0 \end{cases}$$

Ce qui peut s'écrire : $$\frac{P(u^*/H_1)}{P(u^*/H_0)} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \lambda \qquad (5.1.3.3)$$

où : $$\frac{P(u^*/H_1)}{P(u^*/H_0)} = \prod_{S_1} \frac{P_{Di}}{P_{Fi}} \prod_{S_0} \frac{1-P_{Di}}{1-P_{Fi}}$$

soit : $$\frac{P(u^*/H_1)}{P(u^*/H_0)} = \prod_{i=1}^{N} \left(\frac{P_{Di}}{P_{Fi}}\right)^{u_i^*} \left(\frac{1-P_{Di}}{1-P_{Fi}}\right)^{1-u_i^*} \qquad (5.1.3.4)$$

En prenant le logarithme de l'équation (5.1.3.3) et en réarrangeant les termes, la règle de fusion (5.1.3.3) peut s'écrire :

$$\sum_{i=1}^{N} \left[\log \frac{P_{Di}(1-P_{Fi})}{P_{Fi}(1-P_{Di})}\right] u_i^* \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \log\left[\lambda \prod_{i=1}^{N} \frac{1-P_{Fi}}{1-P_{Di}}\right] \qquad (5.1.3.5)$$

La règle de fusion obtenue est donc de la même forme que celle du paragraphe 5.1.2.2. Une somme pondérée des décisions locales est calculée et comparée à un seuil.

• Optimisation des détecteurs locaux

Déterminons maintenant les règles de décision au niveau des détecteurs locaux en utilisant une optimisation élément par élément. L'équation (5.1.3.2) peut s'écrire, en développant suivant le terme u^k :

$$\mathbf{L} = \lambda\alpha + \sum_{u^k} P(u_0=1/u_k=0,u^k)\, [P(u_k=0,u^k/H_1) - \lambda\, P(u_k=0,u^k/H_0)] \qquad (5.1.3.6)$$
$$+ \sum_{u^k} P(u_0=1/u_k=1,u^k)\, [P(u_k=1,u^k/H_1) - \lambda\, P(u_k=1,u^k/H_0)]$$

où $u^k = (u_1, \ldots, u_{k-1}, u_{k+1}, \ldots, u_N)^T$

sachant que :

$P(u_k=0/H_j) = 1 - P(u_k=1/H_j)$
$P(u_k=0,u^k/H_j) = P(u_k=0/H_j)\, P(u^k/H_j)$ puisque u_k est indépendant de u^k $\quad j=0,1$
et que :

$$P(u_k/H_i)= \int_{y_k} P(u_k/y_k, H_i).p(y_k/H_i)\, dy_k \qquad i=0,1$$

$$P(u_k/H_i)= \int_{y_k} P(u_k/y_k).p(y_k/H_i)\, dy_k \qquad \text{puisque } u_k \text{ ne dépend pas de } H_i$$

L peut s'écrire :

$$\mathbf{L} = C^k + \int_{y_k} P(u_k = 1/y_k)\,[C_1^k\, p(y_k/H_1) - \lambda C_0^k\, p(y_k/H_0)]\; dy_k \qquad (5.1.3.7)$$

avec :

$$C^k = \lambda\alpha + \sum_{u^k} P(u_0=1/u_k=0,u^k)\,[P(u^k/H_1) - \lambda\, P(u^k/H_0)]$$

$$C_i^k = \sum_{u^k} [P(u_0=1/u_k=1,u^k) - P(u_0=1/u_k=0,u^k)]\, P(u^k/H_i) \qquad i=0,1$$

Puisque C^k est indépendant de la règle de décision associée au détecteur k, pour maximiser **L** il suffit de maximiser l'intégrale, ce qui se traduit, pour chaque valeur y_k, par la règle de décision suivante :

$$P(u_k=1/y_k) = \begin{cases} 0, \text{ si } C_1^k p(y_k/H_1) - \lambda C_0^k p(y_k/H_0) < 0 \\ 1, \text{ si } C_1^k p(y_k/H_1) - \lambda C_0^k p(y_k/H_0) > 0 \end{cases}$$

La règle de décision pour le détecteur q_k est donc :

$$\frac{p(y_k/H_1)}{p(y_k/H_0)} \underset{u_k=0}{\overset{u_k=1}{\gtrless}} \lambda_k \qquad \text{où} \quad \lambda_k = \lambda\, \frac{C_0^k}{C_1^k} \qquad (5.1.3.8)$$

Optimiser le système revient donc à résoudre 2^N équations de la forme (5.1.3.5) et N équations de la forme (5.1.3.8). Une solution commune à ces 2^N+N équations non linéaires couplées est une solution de l'optimisation de Neyman-Pearson élément par élément du problème de détection décentralisée de la Figure 23.

• Exemple [Var97]

En considérant deux capteurs dont les observations sont indépendantes, une expression explicite des seuils peut être obtenue en fonction de l'opérateur de fusion qui a été retenu. Dans le cas où l'opérateur de fusion est un « et logique » les seuils au niveau des deux détecteurs sont, d'après (5.1.2.3.8) :

$$\lambda_1=\lambda\ \frac{[P(u_0=1/u_1=1,u_2=0)-P(u_0=1/u_1=0,u_2=0)]P(u_2=0/H_0)+[P(u_0=1/u_1=1,u_2=1)-P(u_0=1/u_1=0,u_2=1)]P(u_2=1/H_0)}{[P(u_0=1/u_1=1,u_2=0)-P(u_0=1/u_1=0,u_2=0)]P(u_2=0/H_1)+[P(u_0=1/u_1=1,u_2=1)-P(u_0=1/u_1=0,u_2=1)]P(u_2=1/H_1)}$$

$$\lambda_1=\lambda\ \frac{P(u_2=1/H_0)}{P(u_2=1/H_1)}$$

$$\lambda_1=\lambda\ \frac{P_{F2}}{P_{D2}}$$

De même : $\lambda_2=\lambda\ \frac{P_{F1}}{P_{D1}}$

Les performances en terme de probabilités de fausse alarme P_F et de détection P_D du système complet sont données par :

$$P_D=P_{D1}.P_{D2}$$
avec $P_F=P_{F1}.P_{F2}=\alpha$ fixée

La résolution de ces équations permet de trouver la solution du problème de détection décentralisée parallèle du point de vue de Neyman-Pearson.

Dans le cas où l'opérateur de fusion est un « ou logique », les seuils associés aux deux détecteurs sont, d'après (5.1.2.3.8) :

$$\lambda_1=\lambda\ \frac{1-P_{F2}}{1-P_{D2}}$$
$$\lambda_2=\lambda\ \frac{1-P_{F1}}{1-P_{D1}}$$

Les performances en terme de probabilités de fausse alarme P_F et de détection P_D du système complet sont données par :

$$P_D=P_{D1}+P_{D2}-P_{D1}.P_{D2}$$
avec $P_F=P_{F1}+P_{F2}-P_{F1}.P_{F2}=\alpha$ fixée.

5.2. La détection décentralisée série

Nous considérons, dans ce paragraphe, le problème de la détection décentralisée série (Figure 36) encore appelée « détection décentralisée en tandem » [ReN87b] [PaA90] [Tan90] [TPK91b] [Swa93]. Chaque détecteur local reçoit une information issue d'un capteur et transmet un message binaire à son successeur. La décision du premier détecteur est basée sur les informations issues d'un seul capteur et c'est le dernier détecteur qui élabore la décision finale.

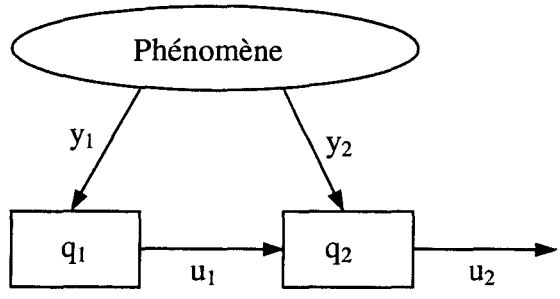


Figure 36 : Architecture de détection décentralisée série à deux détecteurs

En 1987, Reibman et Nolte [ReN87b] ont été les premiers à étudier en détail les performances de différents systèmes de détection décentralisée série. Leur but était de comparer les performances de différents systèmes en confrontant leurs courbes C.O.R.. Cinq types de systèmes de détection furent ainsi étudiés :

- un système de fusion de deux capteurs en série
- un système de fusion de deux capteurs en parallèle
- un système de trois capteurs en série
- un système de trois capteurs dont deux en série et un en parallèle
- un système de fusion de trois capteurs en parallèle.

Fin 1990, Papastavrou et Athans ont présenté une étude [PaA90] sur la détection distribuée avec plusieurs capteurs en tandem. Ils énoncent différents résultats applicables au cadre de la détection binaire et étendent ensuite ces résultats à la détection à M hypothèses.

Dans le cadre de ce travail, nous ne présenterons qu'un système en tandem formé de deux détecteurs q_1 et q_2, les systèmes en tandem formés de plus de deux détecteurs pouvant être étudiés de manière similaire. Les hypothèses possibles sont toujours H_0 et H_1 auxquelles sont associées les probabilités *a priori* P_0 et P_1. Les observations issues de chaque capteur sont notées respectivement y_1 et y_2. Les fonctions de densité de probabilités conditionnelles sous chaque hypothèse sont notées $p(y_1,y_2/H_k)$, k=0,1. Le détecteur q_1 élabore sa décision u_1 à partir de y_1. Cette décision est ensuite transmise au détecteur q_2 qui élabore la décision finale u_2 à partir de u_1 et de y_2. Chaque décision binaire est telle que :

$$u_i = \begin{cases} 0, \text{ le détecteur i décide que } H_0 \text{ est vraie} \\ 1, \text{ le détecteur i décide que } H_1 \text{ est vraie} \end{cases}$$

Remarque :
Dans ce contexte, l'architecture est imposée, c'est-à-dire que c'est le détecteur q_1 qui transmet sa décision à q_2. On pourrait tout à fait considérer l'inverse (q_2 envoie sa décision à q_1 qui élabore alors la décision finale). Ces deux architectures pourraient dès lors être comparées, et l'architecture donnant les meilleurs résultats pourrait alors être retenue.

• Minimisation du risque moyen $\Re$

Les coûts des différentes décisions sont notés C_{jk}, j,k=0,1 où C_{jk} représente le coût de décider H_j pour le détecteur 2 alors que H_k est vraie. Le problème est toujours de minimiser le risque moyen $\Re=E\{C_{jk}\}$. Lors de cette minimisation, le coût associé à $u_1=i$ lorsque H_k est vraie est pris en compte de façon indirecte. Nous supposons de plus que le coût associé à une décision incorrecte est supérieur à celui d'une décision correcte et que les observations y_1 et y_2 sont indépendantes l'une de l'autre sous chaque hypothèse. Le risque moyen $\Re=E\{C_{jk}\}$ peut s'écrire :

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} p(u_1,u_2,y_1,y_2,H_k)\, C_{jk}\, dy_1\, dy_2 \qquad (5.2.1)$$

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} p(u_1,u_2,y_1,y_2/H_k)\, P_k\, C_{jk}\, dy_1\, dy_2$$

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} p(u_2/u_1,y_1,y_2,H_k)\, p(u_1,y_1,y_2/H_k)\, P_k\, C_{jk}\, dy_1\, dy_2 \qquad (5.2.2)$$

• Optimisation du second détecteur

L'optimisation de ce système se fait élément par élément. Lors de l'optimisation du détecteur q_i (i=1,2), l'autre détecteur sera supposé fixé. u_2 étant indépendant de y_1 et de H_k, et y_1 et y_2 étant indépendantes l'un de l'autre, $\Re$ peut s'écrire :

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} P_k\, C_{jk}\, p(u_2/u_1,y_2)\, p(u_1,y_1/H_k)\, p(y_2/H_k)\, dy_1\, dy_2$$

En développant suivant j, sachant que $p(u_2=1/u_1,y_2) = 1 - p(u_2=0/u_1,y_2)$, et en rassemblant les différents termes, on obtient :

$$\Re = \sum_{i,k} \int_{y_1,y_2} P_k\, C_{1k}\, p(u_1,y_1/H_k)\, p(y_2/H_k)\, dy_1\, dy_2$$
$$+ \sum_{i} \int_{y_2} p(u_2=0/u_1,y_2) \cdot \{ \sum_{k} \int_{y_1} P_k\, [C_{0k}-C_{1k}]\, p(u_1,y_1/H_k)\, p(y_2/H_k)\, dy_1 \}\, dy_2$$

Si l'on suppose que le détecteur q_1 est fixé, le premier terme est une constante, et dans ce cas, minimiser $\Re$ revient à décider pour chaque valeur u_1^* et y_2^* prises par u_1 et y_2 :

$$p(u_2=0/u_1^*,y_2^*)=\begin{cases}0, \text{si} \sum_k \int_{y_1} P_k\,[C_{0k}-C_{1k}]\,p(u_1^*,y_1/H_k)\,p(y_2^*/H_k)\,dy_1>0\\ 1, \text{sinon}\end{cases} \quad (5.2.3)$$

En intégrant suivant y_1, en développant suivant k et en réarrangeant les termes cela devient :

$$\frac{p(y_2^*/H_1)}{p(y_2^*/H_0)} \underset{u_2=0}{\overset{u_2=1}{\gtrless}} \frac{p_0(C_{10}-C_{00})\,p(u_1^*/H_0)}{p_1(C_{01}-C_{11})\,p(u_1^*/H_1)} = \frac{C_F\,p(u_1^*/H_0)}{C_D\,p(u_1^*/H_1)} \quad \text{où} \quad C_F=P_0(C_{10}-C_{00}) \quad (5.2.4)$$

$$C_D=(1-P_0)(C_{01}-C_{11})$$

Il faut remarquer que le seuil de droite correspond en fait à deux seuils différents, l'un lorsque $u_1^*=0$ et l'autre lorsque $u_1^*=1$. Ces deux seuils sont notés λ_2^1 et λ_2^0 et dépendent des probabilités de fausse alarme P_{F1} et de détection P_{D1} du détecteur q_1 tels que :

$$\lambda_2^1 = \frac{C_F\,p(u_1=1/H_0)}{C_D\,p(u_1=1/H_1)} = \frac{C_F\,P_{F1}}{C_D\,P_{D1}} \quad (5.2.5)$$

$$\lambda_2^0 = \frac{C_F\,p(u_1=0/H_0)}{C_D\,p(u_1=0/H_1)} = \frac{C_F\,(1-P_{F1})}{C_D\,(1-P_{D1})} \quad (5.2.6)$$

• Optimisation du premier détecteur

Déterminons maintenant la règle de décision du premier détecteur. D'après (5.2.2) on a :

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} P_k\,C_{jk}\,p(u_2/u_1,y_2)\,p(u_1,y_1/H_k)\,p(y_2/H_k)\;dy_1\,dy_2$$

La décision u_1 ne dépend que de y_1 et est indépendante de H_k. $\Re$ peut donc s'écrire :

$$\Re = \sum_{i,j,k} \int_{y_1,y_2} P_k\,C_{jk}\,p(u_2/u_1,y_2)\,p(u_1/y_1)\,p(y_1/H_k)\,p(y_2/H_k)\;dy_1\,dy_2$$

En développant suivant i, sachant que $p(u_1=1/y_1) = 1 - p(u_1=0/y_1)$, et en rassemblant les différents termes, on obtient :

$$\Re = \sum_{j,k} \int_{y_1,y_2} P_k\,C_{jk}\,p(u_2/u_1=1,y_2)\,p(y_1/H_k)\,p(y_2/H_k)\;dy_1\,dy_2$$
$$+ \int_{y_1} p(u_1=0/y_1).\{\sum_{j,k} \int_{y_2} P_k.C_{jk}.p(y_1/H_k).p(y_2/H_k)[(p(u_2/u_1=0,y_2)-p(u_2/u_1=1,y_2)].dy_2\}dy_1$$

Si l'on suppose que le détecteur q_2 est fixé, alors le premier terme est une constante et minimiser $\Re$ revient à décider, pour chaque valeur y_1^* prise par y_1 :

$$p(u_1=0/y_1^*)=\begin{cases}0, \text{si} \sum_{j,k} \int_{y_2} P_k\,C_{jk}\;p(y_1^*/H_k)\,p(y_2/H_k)\,[p(u_2/u_1=0,y_2)-p(u_2/u_1=1,y_2)]\,dy_2>0\\ 1, \text{sinon}\end{cases} \quad (5.2.7)$$

En intégrant suivant y_2, en développant suivant j et k et en réarrangeant les termes, cela devient :

$$\frac{p(y_1^*/H_1)}{p(y_1^*/H_0)} \underset{u_1=0}{\overset{u_1=1}{\gtrless}} \frac{p_0(C_{10}-C_{00})\left[p(u_2=1/u_1=1,H_0)-p(u_2=1/u_1=0,H_0)\right]}{p_1(C_{01}-C_{11})\left[p(u_2=1/u_1=1,H_1)-p(u_2=1/u_1=0,H_1)\right]} \tag{5.2.8}$$

Ce qui peut s'écrire :

$$\frac{p(y_1^*/H_1)}{p(y_1^*/H_0)} \underset{u_1=0}{\overset{u_1=1}{\gtrless}} \frac{C_F P_{F2}(\lambda_2^1)-P_{F2}(\lambda_2^0)}{C_D P_{D2}(\lambda_2^1)-P_{D2}(\lambda_2^0)} = \lambda_1 \tag{5.2.9}$$

où $P_{F2}(\lambda_2^j)$ et $P_{D2}(\lambda_2^j)$ représentent respectivement les valeurs des probabilités de fausse alarme et de détection du détecteur 2 associées au seuil λ_2^j, j=0,1. Le seuil λ_1 dépend donc des deux valeurs que peut prendre le seuil λ_2.

Optimiser le système revient donc à résoudre trois équations non linéaires couplées ; deux équations de la forme (5.2.4) et une équation de la forme (5.2.9). Une solution commune à ces trois équations est une solution de l'optimisation Bayésienne élément par élément du problème de détection décentralisée de la Figure 36.

Le problème de la détection décentralisée série dans le cas où l'on a N détecteurs peut être traité de la même manière. On aboutit alors à un système de 2N-1 équations non linéaires couplées à résoudre.

• Exemple [Var97]

En considérant deux capteurs dont les observations sont indépendantes, une expression explicite des seuils λ_1, λ_2^0 et λ_2^1 peut être obtenue en résolvant les équations (5.2.5), (5.2.6) et (5.2.9). Dans le cas où les fonctions de densité de probabilités conditionnelles associées à chaque situation sont gaussiennes, ce système d'équations a pu être résolu. Sous H_0 ces fonctions sont supposées être de variance 1 et de moyenne 0. Sous H_1 elles sont supposées être de variance 1 et de moyenne $m_1=m_2=1$. Les seuils des deux détecteurs, ainsi que les probabilités d'erreurs qui leur sont associés sont donnés ci-dessous. On peut remarquer, sur la Figure 38, que dans ce cas la détection série donne de meilleurs résultats que la détection parallèle.

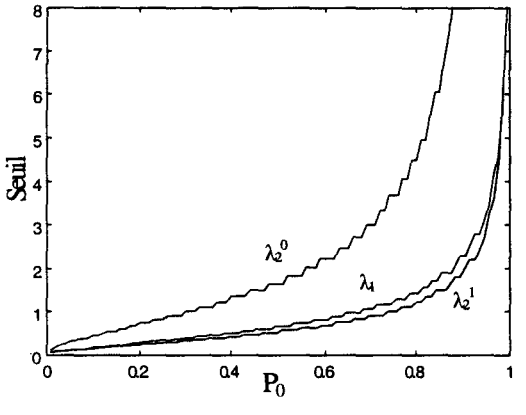


Figure 37 : Seuils en fonction de P_0, $m_1=m_2=1$

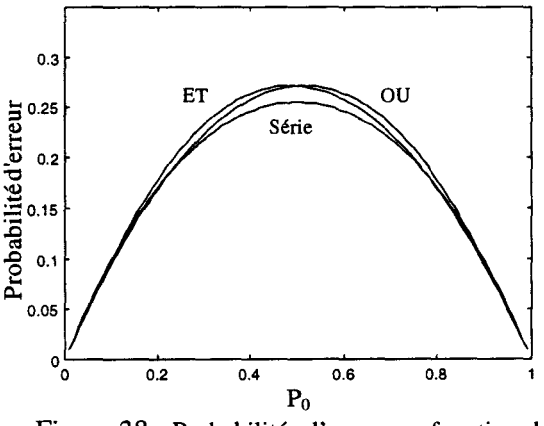


Figure 38 : Probabilités d'erreur en fonction de P_0

6. Conclusion

Pour les systèmes de détection décentralisée composés de deux détecteurs, l'architecture de détection décentralisée série donne au moins d'aussi bons résultats que l'architecture parallèle [Pap90] [Tsi93]. Cependant pour les systèmes de détection décentralisée comportant plus de deux détecteurs, il n'existe pas de résultat général. Les performances obtenues en utilisant l'une ou l'autre architecture doivent alors être évaluées au coup par coup. Dans le cas asymptotique où le nombre de détecteurs N tend vers l'infini, il a été montré que l'architecture de détection décentralisée parallèle donnait de meilleurs résultats que l'architecture série [PaA90]. On peut donc penser qu'il existe une certaine valeur de N à partir de laquelle l'architecture de détection décentralisée parallèle sera meilleure que l'architecture série. Cette valeur n'a pas encore pu être déterminée mais elle est supposée être relativement petite [Tsi93].

Un autre problème intéressant relatif à l'architecture de détection décentralisée série est le problème de l'ordre des détecteurs à l'intérieur du système. Considérons par exemple un système de détection décentralisée série composé de deux détecteurs dont l'un est meilleur que l'autre. Quel détecteur devons nous placer en premier et quel est le détecteur qui doit prendre la décision finale ? De façon intuitive on peut se dire que c'est le meilleur détecteur qui doit prendre cette décision, mais aucun résultat général n'existe dans ce domaine, et des exemples qui viennent contredire cette idée ont pu être trouvés [Pap90].

Des études ont également été menées dans le cadre de la recherche d'une architecture mixte, ou arborescente [Bal94]. Le principe est alors de trouver l'architecture optimale, suivant un critère donné, afin d'obtenir une décision aussi juste que possible.

CHAPITRE 2

L'ENTROPIE

-

UN CRITERE D'OPTIMISATION ORIGINAL

EN THEORIE DE LA DETECTION

1. Introduction

Dans le premier chapitre, nous avons rappelé différents résultats concernant l'optimisation des systèmes de détection centralisée et décentralisée. Lors de cette étude, deux critères d'optimisation ont été appliqués aux architectures les plus courantes. Dans le cas Bayésien, un coût est associé à chaque situation, une fonction risque moyen est ensuite minimisée. Dans les applications où ces coûts sont connus et ont une signification précise, l'approche Bayésienne peut être une excellente solution au problème d'optimisation. Cependant, ce n'est pas forcément le cas pour toutes les applications. Dans certaines d'entre elles, il pourrait être intéressant de se poser le problème de l'optimisation en s'intéressant à la quantité d'information pertinente pour le problème de détection. Ce type d'approche pourrait notamment s'adapter aux problèmes de communication numériques où l'on s'intéresse davantage à la quantité d'information transmise, plutôt qu'à l'information elle-même. Pour de tels systèmes un critère basé sur une fonction entropique pourrait être plus approprié. Différents travaux ont été effectués dans ce but ; Middleton [Mid60] et Gabrielle [Gab66] ont proposé une architecture de détection centralisée basée sur la minimisation de la perte d'information entre l'entrée et la sortie du système. Hoballah [HoV89b], et Warren [WWR89] ont étendu ce résultat au problème de la détection décentralisée parallèle.

Nous avons schématisé ci-dessous le fonctionnement d'un système de détection binaire ainsi que le problème de la transmission d'informations binaires à travers un canal de transmission. Notre objectif est de comparer ces deux approches.

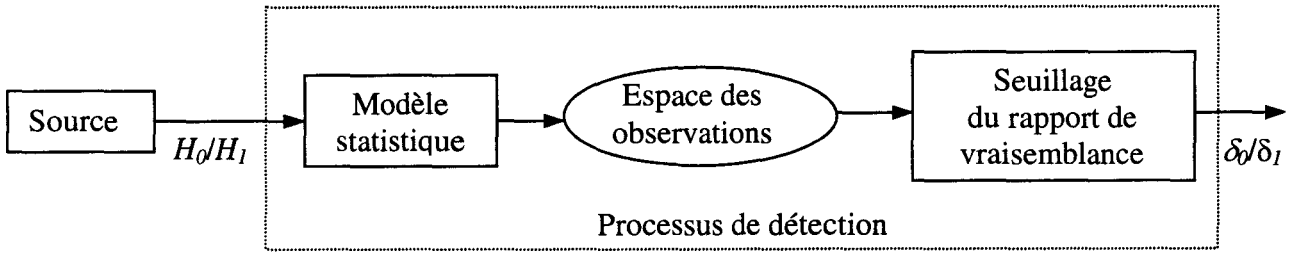


Figure 1 : Fonctionnement d'un système de détection binaire

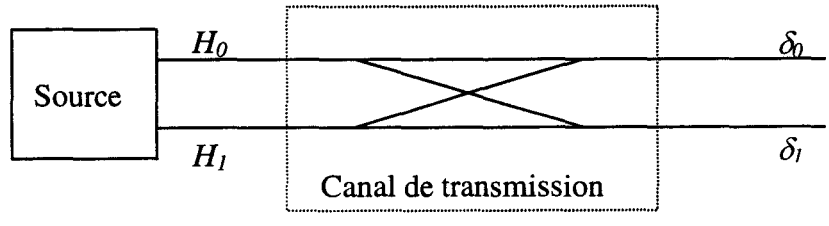


Figure 2 : Fonctionnement d'un système de transmission binaire

Dans le cadre du problème de la détection, deux hypothèses H_0 et H_1 doivent être discriminées. En d'autres termes, une décision (δ_0 ou δ_1) doit être prise à partir d'une observation donnée. Les probabilités de non détection $1\text{-}P_D$ et de fausse alarme P_F peuvent s'interpréter comme des probabilités d'erreur si on se place dans un problème de transmission d'informations binaires (Figure 3).

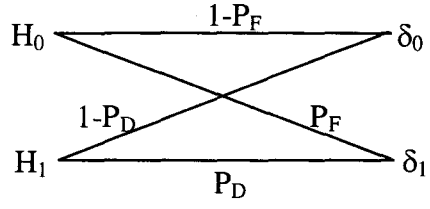


Figure 3 : Modèle de transmission appliqué au problème de la détection

L'objectif est alors de minimiser la perte d'information sur H (H_0 ou H_1) connaissant la décision δ (δ_0 ou δ_1). Cette perte d'information peut être mesurée par l'entropie conditionnelle de Shannon $H(H/\delta)$, et qui représente en fait l'incertitude sur H connaissant la décision δ. Pour plus d'explications sur ce point, le lecteur pourra se reporter au chapitre traitant de la théorie de l'information (chapitre 4).
Dans un premier temps, nous utiliserons ce critère afin d'optimiser un système de détection centralisée. Dans un deuxième temps, cette démarche sera étendue au cas de la détection décentralisée parallèle, puis série. Enfin, une comparaison des résultats obtenus en utilisant la démarche classique et la démarche entropique sera entreprise.
Ce chapitre est inspiré des travaux de Hoballah et Varshney [HoV89b] développant une approche informationnelle du problème de la détection décentralisée basée sur la minimisation de la transinformation interne $\mathring{I}(H,\delta)$ (voir chapitre 4). Notre contribution consiste en l'utilisation d'un critère aboutissant à des résultats originaux et plus complets (nous traitons le problème de l'optimisation de l'opérateur de fusion seul ainsi que le problème de l'optimisation des systèmes de détection décentralisée série) dont les formulations sont très proches de celles obtenues au chapitre 1, facilitant par la même occasion la comparaison de l'approche classique et de l'approche informationnelle.

2. Utilisation d'un critère entropique dans le cadre de la détection centralisée

Considérons un système (Figure 4) composé de N capteurs Y_i i=1, ..., N observant le même phénomène. A partir des informations $y=(y_1,y_2,...,y_N)$ fournies par l'ensemble des capteurs $Y=(Y_1, Y_2, ..., Y_N)$, une décision δ (δ_0 ou δ_1) doit être prise. Les lois de probabilités associées à chaque hypothèse sont supposées connues et on note :

- la probabilité de détection $P_D = P(\delta_1/H_1)$
- la probabilité de fausse alarme $P_F = P(\delta_1/H_0)$

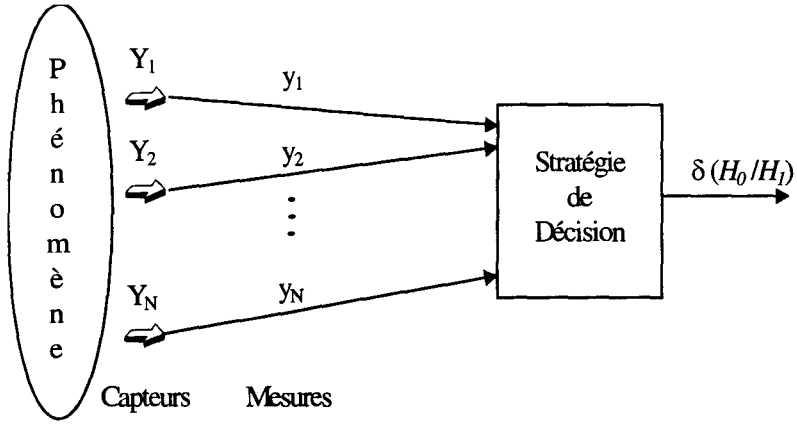


Figure 4 : La détection centralisée

Le critère entropique consiste à minimiser l'entropie conditionnelle de Shannon $H(H/\delta)$ [Mid60] [Gab66], en sachant que $H(H/\delta)=E\{\log(1/P(H/\delta)\}$. En posant $C(H,\delta)=\log(1/P(H/\delta))$, cela revient à minimiser une fonction de risque moyen pour laquelle les fonctions de coût ne sont pas constantes mais dépendent des probabilités *a posteriori*. Dans un premier temps, nous nous attacherons à déterminer l'expression de l'entropie conditionnelle $H(H/\delta)$ en fonction des probabilités de détection P_D et de fausse alarme P_F. Nous montrerons dès lors les relations existant entre le critère entropique et les critères de Bayes et de Neyman-Pearson.

• **Expression de l'entropie conditionnelle H(H/δ) en fonction des probabilités de détection P_D et de fausse alarme P_F**

L'entropie conditionnelle H(H/δ)* s'écrit :

$$\begin{aligned} H(H/\delta) = & - P(\delta_0,H_0) \log P(H_0/\delta_0) \\ & - P(\delta_1,H_0) \log P(H_0/\delta_1) \\ & - P(\delta_0,H_1) \log P(H_1/\delta_0) \\ & - P(\delta_1,H_1) \log P(H_1/\delta_1) \end{aligned} \qquad (2.1)$$

$$\begin{aligned} H(H/\delta) = & - P(\delta_0/H_0)\, P_0 \log \frac{P(\delta_0/H_0)\, P_0}{P(\delta_0)} \\ & - P(\delta_1/H_0)\, P_0 \log \frac{P(\delta_1/H_0)\, P_0}{P(\delta_1)} \\ & - P(\delta_0/H_1)\, (1-P_0) \log \frac{P(\delta_0/H_1)\,(1-P_0)}{P(\delta_0)} \\ & - P(\delta_1/H_1)\, (1-P_0) \log \frac{P(\delta_1/H_1)\,(1-P_0)}{P(\delta_1)} \end{aligned}$$

$$\begin{aligned} H(H/\delta) = & - P(\delta_0/H_0)\, P_0 \log \frac{P(\delta_0/H_0)\, P_0}{P(\delta_0/H_0)\, P_0 + P(\delta_0/H_1)\,(1-P_0)} \\ & - P(\delta_1/H_0)\, P_0 \log \frac{P(\delta_1/H_0)\, P_0}{P(\delta_1/H_0)\, P_0 + P(\delta_1/H_1)\,(1-P_0)} \\ & - P(\delta_0/H_1)\, (1-P_0) \log \frac{P(\delta_0/H_1)\,(1-P_0)}{P(\delta_0/H_0)\, P_0 + P(\delta_0/H_1)\,(1-P_0)} \\ & - P(\delta_1/H_1)\, (1-P_0) \log \frac{P(\delta_1/H_1)\,(1-P_0)}{P(\delta_1/H_0)\, P_0 + P(\delta_1/H_1)\,(1-P_0)} \end{aligned}$$

On en déduit que :

$$\begin{aligned} H(H/\delta) = & - P_0\,(1-P_F) \log \frac{P_0\,(1-P_F)}{P_0\,(1-P_F) + (1-P_0)\,(1-P_D)} \\ & - P_0\, P_F \log \frac{P_0\, P_F}{P_0\, P_F + (1-P_0)\, P_D} \\ & - (1-P_0)\,(1-P_D) \log \frac{(1-P_0)\,(1-P_D)}{P_0\,(1-P_F) + (1-P_0)\,(1-P_D)} \\ & - (1-P_0)\, P_D \log \frac{(1-P_0)\, P_D}{P_0\, P_F + (1-P_0)\, P_D} \end{aligned} \qquad (2.2)$$

L'entropie H(H/δ) peut donc s'exprimer en fonction des probabilités de fausse alarme P_F, de détection P_D, et de P_0. Nous avons tracé la courbe H(H/δ) en fonction de P_F et de P_D pour P_0=0,5.

---

* La base du logarithme peut être quelconque. Afin d'alléger les notations, nous considérons des logarithmes népériens. Dans la littérature, on utilise des logarithmes de base 2 (l'entropie est alors mesurée en bits), qui s'accorde bien avec le système binaire de représentation des informations dans les calculateurs. Le passage en base 2 revient alors à diviser l'entropie par log(2).

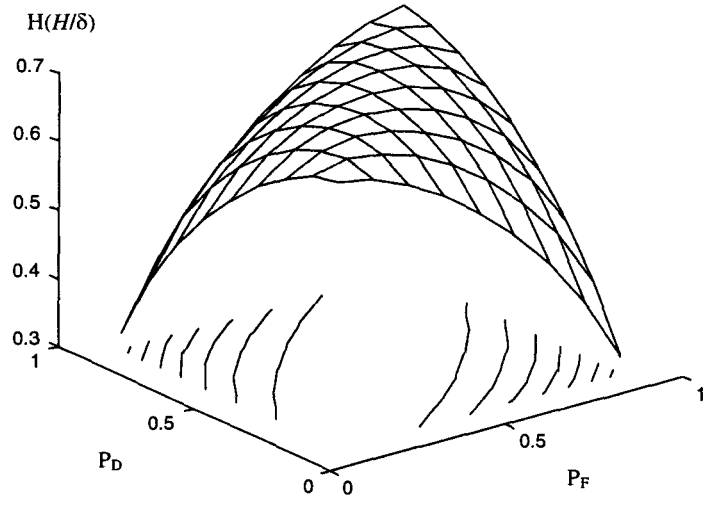


Figure 5 : $H(H/\delta) = f(P_D,P_F)$ avec $P_0=0,5$

Attachons-nous maintenant à l'étude de la fonction $H(H/\delta)=f(P_D,P_F)$. Pour cela, nous étudierons les variations de cette fonction à probabilité de fausse alarme P_F constante, puis à probabilité de détection P_D constante.

• Minimisation de l'entropie $H(H/\delta)$ à probabilité de fausse alarme P_F constante

Dérivons $H(H/\delta)$ par rapport à P_D pour une probabilité de fausse alarme P_F fixée :

$$\frac{\partial H(H/\delta)}{\partial P_D} = -(1-P_0)\log\frac{(1-P_0)P_D}{P_0 P_F+(1-P_0)P_D}+(1-P_0)\log\frac{(1-P_0)(1-P_D)}{P_0(1-P_F)+(1-P_0)(1-P_D)}$$

$$\frac{1}{(1-P_0)}\frac{\partial^2 H(H/\delta)}{\partial P_D^{\,2}}=\frac{(1-P_0)}{P_0 P_F+(1-P_0)P_D}-\frac{(1-P_0)}{(1-P_0)P_D}+\frac{(1-P_0)}{P_0(1-P_F)+(1-P_0)(1-P_D)}-\frac{(1-P_0)}{(1-P_0)(1-P_D)}$$

$$\frac{1}{(1-P_0)}\frac{\partial^2 H(H/\delta)}{\partial P_D^{\,2}}=\frac{1}{P_D}\left[\frac{1}{1+\frac{P_0 P_F}{(1-P_0)P_D}}-1\right]+\frac{1}{(1-P_D)}\left[\frac{1}{1+\frac{P_0(1-P_F)}{(1-P_0)(1-P_D)}}-1\right]$$

$$\text{donc } \frac{1}{(1-P_0)}\frac{\partial^2 H(H/\delta)}{\partial P_D^{\,2}}<0 \qquad (2.3)$$

de plus, nous avons : $\dfrac{\partial H(H/\delta)}{\partial P_D}(P_D = P_F)=0$

$\dfrac{\partial H(H/\delta)}{\partial P_D}(P_D)$ est donc une fonction décroissante de P_D, positive sur $[0,P_F]$ et négative sur $[P_F,1]$. $H(H/\delta)(P_D)$ représente donc une fonction croissante sur $[0,P_F]$ et décroissante sur $[P_F,1]$ (Figure 6). On ne retiendra ici que l'intervalle $[P_F,1]$, c'est-à-dire l'intervalle où l'on a

$P_D > P_F$ (seul cas intéressant pour la détection). Dans ce cas, $H(H/\delta)$ est une fonction décroissante de P_D.

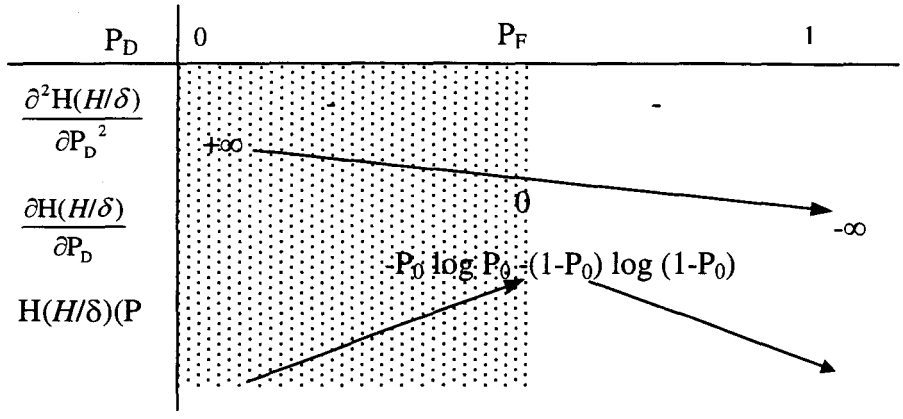


Figure 6 : Variations de $H(H/\delta)$ en fonction de le probabilité de détection P_D

On en déduit donc que pour une probabilité de fausse alarme P_F fixée, l'entropie conditionnelle $H(H/\delta)$ est une fonction décroissante de la probabilité de détection P_D. A P_F fixée, rechercher le détecteur qui minimise $H(H/\delta)$ est donc équivalent à chercher le détecteur qui maximise la probabilité de détection P_D. Minimiser $H(H/\delta)$ à probabilité de fausse alarme fixée (Figure 7) est donc équivalent à appliquer le critère de Neyman-Pearson.

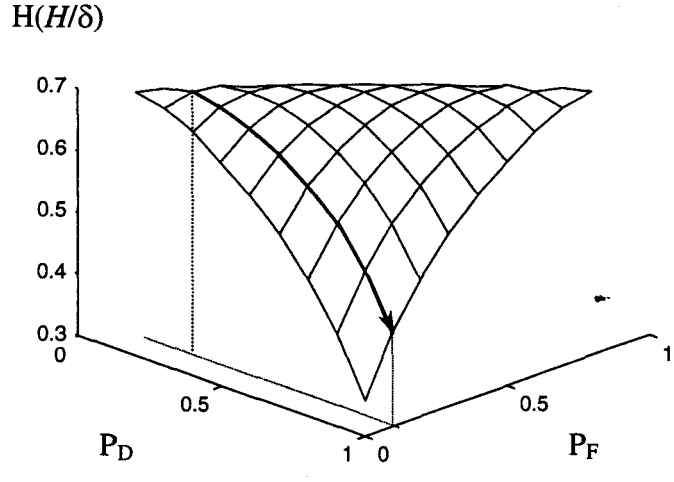


Figure 7 : Minimisation de l'entropie $H(H/\delta)$ à probabilité de fausse alarme P_F constante

• **Minimisation de l'entropie H(H/δ) à probabilité de détection P_D constante**

On dérive H(H/δ) par rapport à P_F pour une probabilité de fausse alarme P_D fixée

$$\frac{\partial H(H/\delta)}{\partial P_F} = -P_0 \log \frac{P_0 P_F}{P_0 P_F + (1-P_0) P_D} + P_0 \log \frac{P_0 (1-P_F)}{P_0 (1-P_F) + (1-P_0)(1-P_D)}$$

$$\frac{1}{P_0} \frac{\partial^2 H(H/\delta)}{\partial P_F^2} = \frac{P_0}{P_0 P_F + (1-P_0) P_D} - \frac{P_0}{P_0 P_F} + \frac{P_0}{P_0 (1-P_F) + (1-P_0)(1-P_D)} - \frac{P_0}{P_0 (1-P_F)}$$

$$\frac{1}{P_0} \frac{\partial^2 H(H/\delta)}{\partial P_F^2} = \frac{1}{P_F} \left[\frac{1}{1 + \frac{(1-P_0) P_D}{P_0 P_F}} - 1 \right] + \frac{1}{(1-P_F)} \left[\frac{1}{1 + \frac{(1-P_0)(1-P_D)}{P_0 (1-P_F)}} - 1 \right]$$

donc $\frac{1}{P_0} \frac{\partial^2 H(H/\delta)}{\partial P_F^2} < 0$ (2.3)

de plus $\frac{\partial H(H/\delta)}{\partial P_F}(P_F = P_D) = 0$

$\frac{\partial H(H/\delta)}{\partial P_F}$ est donc une fonction décroissante de P_F, positive sur $[0,P_D]$ et négative sur $[P_D,1]$.

H(H/δ)(P_F) est donc croissante sur $[0,P_D]$ et décroissante sur $[P_D,1]$ (Figure 8). Dans le cadre de la détection, on ne retiendra ici que l'intervalle $[0,P_D]$, c'est-à-dire l'intervalle dans lequel $P_F < P_D$ et dans ce cas, H(H/δ)(P_F) croissante.

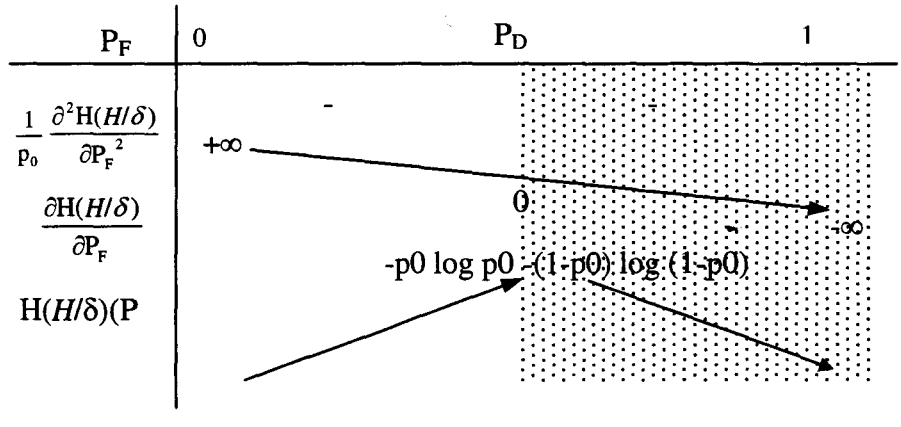


Figure 8 : Variation de H(H/δ) en fonction de le probabilité de fausse alarme P_F

On en déduit donc que pour une probabilité de détection P_D fixée, l'entropie conditionnelle H(H/δ) est une fonction croissante de la probabilité de fausse alarme P_F. A P_D fixée, rechercher le détecteur qui minimise H(H/δ) est donc équivalent à chercher le détecteur qui minimise la probabilité de fausse alarme P_F.

Minimiser $H(H/\delta)$ sans fixer ni P_F ni P_D, revient donc à résoudre le problème bi-critère $\{\min P_F, \max P_D\}$ en adoptant une fonction d'utilité non linéaire $H(H/\delta)$ (à l'inverse du critère de Bayes où l'on cherche à minimiser la fonction linéaire $\Re = \alpha.P_F+\beta.(1-P_D)$).

Montrons que minimiser $H(H/\delta)$ revient à effectuer un seuillage du rapport de vraisemblance, et que le seuil peut être déterminé à partir de la connaissance des probabilités *a posteriori* associées aux différentes situations possibles.

• **Détermination du seuil λ qui minimise** $H(H/\delta)$

Chercher le seuil λ pour lequel l'entropie conditionnelle $H(H/\delta)$ est minimale revient à chercher la valeur du rapport de vraisemblance Λ pour laquelle $\left(\frac{\partial H(H/\delta)}{\partial \Lambda}\right)_{\Lambda=\lambda} = 0$.

Rappelons que :

$$P_F = \int_{\lambda}^{+\infty} p(\Lambda/H_0)\, d\Lambda \quad \Rightarrow \quad P(\Lambda=\lambda/H_0) = -\left(\frac{\partial P_F}{\partial \Lambda}\right)_{\Lambda=\lambda}$$

$$P_D = \int_{\lambda}^{+\infty} p(\Lambda/H_1)\, d\Lambda \quad \Rightarrow \quad P(\Lambda=\lambda/H_1) = -\left(\frac{\partial P_D}{\partial \Lambda}\right)_{\Lambda=\lambda}$$

On a donc :

$$\begin{aligned}
\left(\frac{\partial H(H/\delta)}{\partial \Lambda}\right)_{\Lambda=\lambda} = \quad & + P_0\, P(\Lambda=\lambda/H_0) \log \frac{P_0 P_F}{P_0 P_F + (1-P_0) P_D} \\
& + P_0 P_F \frac{P_0\, P(\Lambda=\lambda/H_0)}{P_0 P_F} \\
& - P_0 P_F \frac{P_0\, P(\Lambda=\lambda/H_0) + (1-P_0) P(\Lambda=\lambda/H_1)}{P_0 P_F + (1-P_0) P_D} \\
& - P_0\, P(\Lambda=\lambda/H_0) \log \frac{P_0 (1-P_F)}{P_0 (1-P_F) + (1-P_0)(1-P_D)} \\
& - P_0 (1-P_F) \frac{P_0\, P(\Lambda=\lambda/H_0)}{P_0 (1-P_F)} \\
& + P_0 (1-P_F) \frac{P_0\, P(\Lambda=\lambda/H_0) + (1-P_0) P(\Lambda=\lambda/H_1)}{P_0 (1-P_F) + (1-P_0)(1-P_D)} \\
& + (1-P_0)\, P(\Lambda=\lambda/H_1) \log \frac{(1-P_0) P_D}{P_0 P_F + (1-P_0) P_D} \\
& + (1-P_0) P_D \frac{(1-P_0) P(\Lambda=\lambda/H_1)}{(1-P_0) P_D} \\
& - (1-P_0) P_D \frac{P_0\, P(\Lambda=\lambda/H_0) + (1-P_0) P(\Lambda=\lambda/H_1)}{P_0 P_F + (1-P_0) P_D} \\
& - (1-P_0)\, P(\Lambda=\lambda/H_1) \log \frac{(1-P_0)(1-P_D)}{P_0 (1-P_F) + (1-P_0)(1-P_D)}
\end{aligned}$$

$$- (1-P_0)(1-P_D)\frac{(1-P_0)P(\Lambda=\lambda/H_1)}{(1-P_0)(1-P_D)}$$
$$+ (1-P_0)(1-P_D)\frac{P_0 P(\Lambda=\lambda/H_0)+(1-P_0)P(\Lambda=\lambda/H_1)}{P_0(1-P_F)+(1-P_0)(1-P_D)}$$

L'équation précédente se simplifie pour aboutir à l'expression suivante :

$$\left(\frac{\partial H(H/\delta)}{\partial \Lambda}\right)_{\Lambda=\lambda} = + P_0\, P(\Lambda=\lambda/H_0)\log\frac{P_0 P_F}{P_0 P_F+(1-P_0)P_D}$$
$$- P_0\, P(\Lambda=\lambda/H_0)\log\frac{P_0(1-P_F)}{P_0(1-P_F)+(1-P_0)(1-P_D)}$$
$$+ (1-P_0)\, P(\Lambda=\lambda/H_1)\log\frac{(1-P_0)P_D}{P_0 P_F+(1-P_0)P_D}$$
$$- (1-P_0)\, P(\Lambda=\lambda/H_1)\log\frac{(1-P_0)(1-P_D)}{P_0(1-P_F)+(1-P_0)(1-P_D)}$$

et puisque l'on a toujours $\frac{P(\Lambda=\lambda/H_1)}{P(\Lambda=\lambda/H_0)}=\lambda$, on peut en déduire :

$$\left(\frac{\partial H(H/\delta)}{\partial \Lambda}\right)_{\Lambda=\lambda} = P(\Lambda=\lambda/H_0)\left[P_0\log\frac{P_0 P_F}{P_0 P_F+(1-P_0)P_D} - P_0\log\frac{P_0(1-P_F)}{P_0(1-P_F)+(1-P_0)(1-P_D)}\right.$$
$$\left. + (1-P_0)\lambda\log\frac{(1-P_0)P_D}{P_0 P_F+(1-P_0)P_D} - (1-P_0)\lambda\log\frac{(1-P_0)(1-P_D)}{P_0(1-P_F)+(1-P_0)(1-P_D)}\right] \quad (2.4)$$

Or, nous avons les relations suivantes :

$$P(H_0/\delta_0)=\frac{P(\delta_0/H_0)P_0}{P(\delta_0)}=\frac{(1-P_F)P_0}{P(\delta_0/H_0)P_0+P(\delta_0/H_1)(1-P_0)}=\frac{P_0(1-P_F)}{P_0(1-P_F)+(1-P_0)(1-P_D)}$$

$$P(H_0/\delta_1)=\frac{P_0 P_F}{P_0 P_F+(1-P_0)P_D};\quad P(H_1/\delta_0)=\frac{(1-P_0)(1-P_D)}{P_0(1-P_F)+(1-P_0)(1-P_D)};\quad P(H_1/\delta_1)=\frac{(1-P_0)P_D}{P_0 P_F+(1-P_0)P_D}$$

Pour que $H(H/\delta)$ soit minimale, il faut que la quantité $\left(\frac{\partial H(H/\delta)}{\partial \Lambda}\right)$ soit nulle. Ainsi l'équation (2.4) conduit tout naturellement à choisir λ tel que :

$$\lambda=\frac{P_0}{1-P_0}\,\frac{\log P(H_0/\delta_1)-\log P(H_0/\delta_0)}{\log P(H_1/\delta_0)-\log P(H_1/\delta_1)}$$

ou encore :

$$\lambda=\frac{P_0}{1-P_0}\,\frac{\log 1/P(H_0/\delta_1)-\log 1/P(H_0/\delta_0)}{\log 1/P(H_1/\delta_0)-\log 1/P(H_1/\delta_1)} \quad (2.5)$$

Fabriquer le détecteur qui minimise l'entropie $H(H/\delta)$ est équivalent à fabriquer le détecteur à rapport de vraisemblance $\Lambda(y)=\frac{p(y/H_1)}{p(y/H_0)}$ tel que :

$$\Lambda(y) \underset{\delta=H_0}{\overset{\delta=H_1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})} \qquad (2.6)$$

où C_{ij} est un coût variable défini par $C_{ij}=\log \frac{1}{P(H_j/\delta_i)}$ (i,j=0,1)

• Exemple

Reprenons un exemple traité dans le premier chapitre : le système est formé d'un capteur dont les fonctions de densité de probabilités sous chaque hypothèse sont des gaussiennes. Sous l'hypothèse H_0, la fonction de densité de probabilité est supposée être de moyenne 0 et de variance 1. Sous l'hypothèse H_1, cette fonction est supposée être de moyenne 1 et de variance 1.

La règle de décision est alors la suivante : $\Lambda(y) \underset{\delta=H_0}{\overset{\delta=H_1}{\gtrless}} \lambda$

Le seuil λ qui minimise l'entropie $H(H/\delta)$ peut être calculé. Nous avons tracé la valeur du seuil qui minimise $H(H/\delta)$ (Figure 9), la valeur de $H(H/\delta)$ (Figure 10), la probabilité d'erreur en fonction de P_0 (Figure 11) et la courbe C.O.R. en prenant P_0 comme paramètre (Figure 12). Minimiser l'entropie revient à chercher, pour chaque valeur de P_0, le point de la courbe C.O.R. qui se rapproche le plus du point (P_F=0,P_D=1). La courbe C.O.R. trouvée correspond exactement à une partie de celle trouvée en appliquant le critère de Bayes. Pour P_0=0.5, les résultats trouvés en terme de seuil, et de probabilité d'erreur sont les mêmes que ceux trouvés dans le cas Bayésien. Par contre, lorsque P_0 varie, le critère entropique ne permet pas de minimiser la probabilité d'erreur, en revanche les probabilités de détection et de fausse alarme trouvées restent dans des limites acceptables, ce qui n'est pas le cas en appliquant le critère de Bayes.

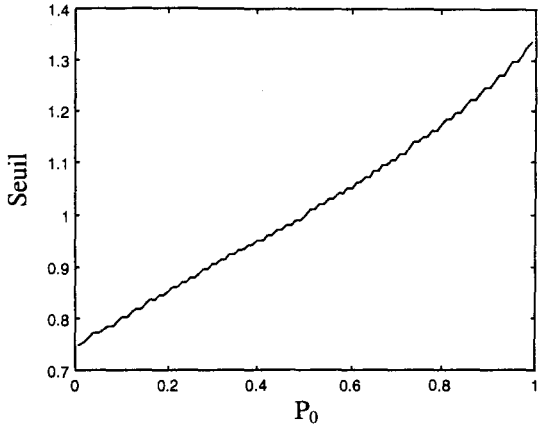


Figure 9 : Valeur du seuil λ en fonction de P_0

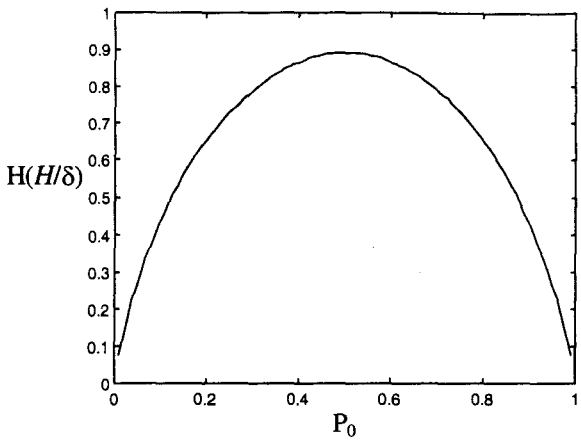


Figure 10 : Valeur de $H(H/\delta)$ en fonction de P_0

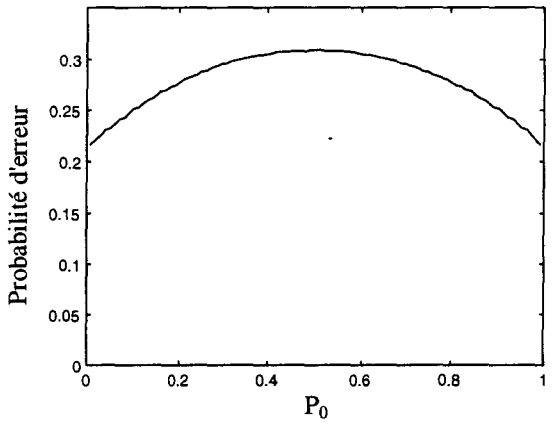


Figure 11 : Probabilité d'erreur en fonction de P_0

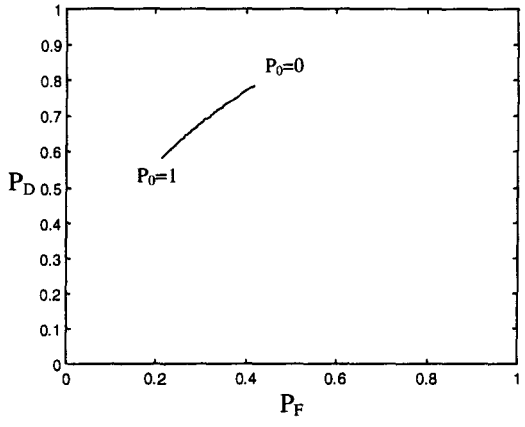


Figure 12 : Courbe C.O.R. $P_D=f(P_F)$

En conclusion, dans le cadre de la détection centralisée, utiliser le critère entropique revient à utiliser le critère de Bayes en définissant des coûts variables, qui ne sont dès lors plus définis de façon arbitraire.

3. Utilisation d'un critère entropique dans le cadre de la détection décentralisée parallèle

Dans ce paragraphe, nous montrons que l'entropie peut être utilisée comme critère d'optimisation d'une architecture de détection décentralisée parallèle. Dans un premier temps, nous considérons le problème de l'optimisation des différents opérateurs locaux sans prendre en compte l'opérateur de fusion, que nous optimisons séparément. Nous étudions ensuite l'optimisation simultanée des détecteurs locaux et de l'opérateur de fusion.

3.1. Optimisation des détecteurs locaux

3.1.1. Le cas de deux détecteurs en parallèle

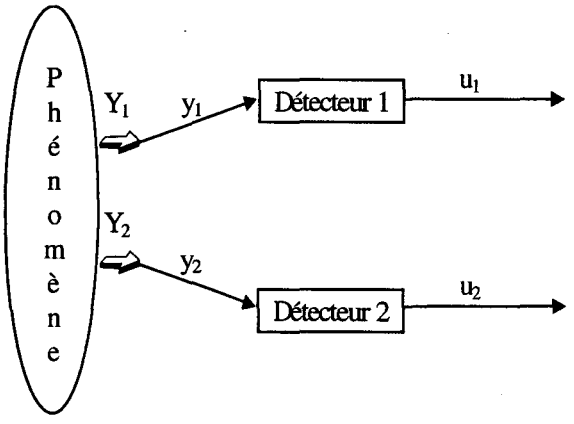


Figure 13 : Deux détecteurs en parallèle

Nous considérons dans ce paragraphe la détection parallèle sans prendre en compte le problème de la fusion. Le système étudié est composé de deux capteurs en parallèle qui observent un phénomène commun. A chaque capteur est associé un détecteur qui prend une décision locale (Figure 13). Nous n'étudions ici que la détection décentralisée binaire. H_0 et H_1 sont les deux hypothèses à discriminer de probabilités *a priori* respectives P_0 et P_1. Les observations associées aux deux capteurs sont toujours notées y_1 et y_2, et les décisions prises par chaque détecteur, u_i, i=1,2 :

$$u_i = \begin{cases} 0, \text{on décide que } H_0 \text{ est vraie} \\ 1, \text{on décide que } H_1 \text{ est vraie} \end{cases}$$

Chaque décision binaire est caractérisée par ses probabilités de fausse alarme P_{Fi} et de détection P_{Di} telles que :

$$P_{Fi}=P(u_i=1/H_0) \qquad \text{et} \qquad P_{Di}=P(u_i=1/H_1)$$

Les décisions prises par chaque détecteur sont supposées indépendantes. Le critère d'optimisation consiste à minimiser l'entropie conditionnelle de Shannon $H(H/u_1,u_2)$.

• **Expression de l'entropie conditionnelle $H(H/u_1,u_2)$ en fonction des probabilités de détection P_{Di} et de fausse alarme P_{Fi} (i=1,2)**

L'entropie conditionnelle $H(H/ u_1,u_2)$ s'écrit :

$$H(H/u_1,u_2)= - \sum_{i,j\in\{0,1\}} P(u_1=i, u_2=j, H_0) \log P(H_0/ u_1=i, u_2=j) - \sum_{i,j\in\{0,1\}} P(u_1=i, u_2=j, H_1) \log P(H_1/ u_1=i, u_2=j) \quad (3.1.1.1)$$

$$H(H/u_1,u_2)= - \sum_{i,j\in\{0,1\}} P_0\, P(u_1=i, u_2=j / H_0) \log \frac{P(u_1=i,u_2=j/H_0)\,P_0}{P(u_1=i,u_2=j)} - \sum_{i,j\in\{0,1\}} (1-P_0)\, P(u_1=i, u_2=j / H_1) \log \frac{P(u_1=i,u_2=j/H_1)(1-P_0)}{P(u_1=i,u_2=j)}$$

$$H(H/u_1,u_2)= - \sum_{i,j\in\{0,1\}} P_0\, P(u_1=i/ H_0)\, P(u_2=j / H_0) \log \frac{P(u_1=i/H_0)\,P(u_2=j/H_0)P_0}{P(u_1=i,u_2=j/H_0)P_0+P(u_1=i,u_2=j/H_1)(1-P_0)} - \sum_{i,j\in\{0,1\}} (1-P_0)\, P(u_1=i/ H_1)\, P(u_2=j / H_1) \log \frac{P(u_1=i/H_1)\,P(u_2=j/H_1)(1-P_0)}{P(u_1=i,u_2=j/H_0)P_0+P(u_1=i,u_2=j/H_1)(1-P_0)}$$

On en déduit que :

$$\begin{aligned} H(H/u_1,u_2)= & - P_0 P_{F1} P_{F2} \log \frac{P_0 P_{F1} P_{F2}}{P_0 P_{F1} P_{F2} + (1-P_0) P_{D1} P_{D2}} \\ & - P_0 (1-P_{F1}) P_{F2} \log \frac{P_0 (1-P_{F1}) P_{F2}}{P_0 (1-P_{F1}) P_{F2} + (1-P_0)(1-P_{D1}) P_{D2}} \\ & - P_0 P_{F1} (1-P_{F2}) \log \frac{P_0 P_{F1} (1-P_{F2})}{P_0 P_{F1} (1-P_{F2}) + (1-P_0) P_{D1} (1-P_{D2})} \\ & - P_0 (1-P_{F1}) (1-P_{F2}) \log \frac{P_0 (1-P_{F1})(1-P_{F2})}{P_0 (1-P_{F1})(1-P_{F2}) + (1-P_0)(1-P_{D1})(1-P_{D2})} \\ & - (1-P_0) P_{D1} P_{D2} \log \frac{(1-P_0) P_{D1} P_{D2}}{P_0 P_{F1} P_{F2} + (1-P_0) P_{D1} P_{D2}} \\ & - (1-P_0) (1-P_{D1}) P_{D2} \log \frac{(1-P_0)(1-P_{D1}) P_{D2}}{P_0 (1-P_{F1}) P_{F2} + (1-P_0)(1-P_{D1}) P_{D2}} \\ & - (1-P_0) P_{D1} (1-P_{D2}) \log \frac{(1-P_0) P_{D1} (1-P_{D2})}{P_0 P_{F1} (1-P_{F2}) + (1-P_0) P_{D1} (1-P_{D2})} \\ & - (1-P_0) (1-P_{D1}) (1-P_{D2}) \log \frac{(1-P_0)(1-P_{D1})(1-P_{D2})}{P_0 (1-P_{F1})(1-P_{F2}) + (1-P_0)(1-P_{D1})(1-P_{D2})} \end{aligned} \quad (3.1.1.2)$$

Pour simplifier cette expression, on pose :

$$\begin{aligned} \alpha_{1,1} &= P_0 P_{F1} P_{F2} \\ \alpha_{-1,1} &= P_0 (1-P_{F1}) P_{F2} \\ \alpha_{1,-1} &= P_0 P_{F1} (1-P_{F2}) \\ \alpha_{-1,-1} &= P_0 (1-P_{F1}) (1-P_{F2}) \end{aligned} \quad (3.1.1.3)$$

et

$$\begin{aligned}
\beta_{1,1} &= (1-P_0)\, P_{D1}\, P_{D2} \\
\beta_{-1,1} &= (1-P_0)\, (1-P_{D1})\, P_{D2} \\
\beta_{1,-1} &= (1-P_0)\, P_{D1}\, (1-P_{D2}) \\
\beta_{-1,-1} &= (1-P_0)\, (1-P_{D1})\, (1-P_{D2})
\end{aligned} \qquad (3.1.1.4)$$

En fonction de ces paramètres, $H(H/u_1,u_2)$ peut se mettre sous la forme :

$$\begin{aligned}
H(H/u_1,u_2) = & - \alpha_{1,1} \log \frac{\alpha_{1,1}}{\alpha_{1,1}+\beta_{1,1}} - \beta_{1,1} \log \frac{\beta_{1,1}}{\alpha_{1,1}+\beta_{1,1}} \\
& - \alpha_{-1,1} \log \frac{\alpha_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} - \beta_{-1,1} \log \frac{\beta_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} \\
& - \alpha_{1,-1} \log \frac{\alpha_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} - \beta_{1,-1} \log \frac{\beta_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} \\
& - \alpha_{-1,-1} \log \frac{\alpha_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}} - \beta_{-1,-1} \log \frac{\beta_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}}
\end{aligned} \qquad (3.1.1.5)$$

$$H(H/u_1,u_2) = - \left[\sum_{i,j\in\{-1,1\}} \alpha_{i,j} \log \frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}} + \beta_{i,j} \log \frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}} \right] \qquad (3.1.1.6)$$

$$H(H/u_1,u_2) = \sum_{i,j\in\{-1,1\}} \left[(\alpha_{i,j}+\beta_{i,j}) \log (\alpha_{i,j}+\beta_{i,j}) \right] - \sum_{i,j\in\{-1,1\}} \left[\alpha_{i,j} \log \alpha_{i,j} + \beta_{i,j} \log \beta_{i,j} \right] \qquad (3.1.1.7)$$

• Minimisation de l'entropie conditionnelle $H(H/u_1,u_2)$

On étudie ici les variations $H(H/u_1,u_2)$ en fonction de P_{D1} et on fixe P_{F1}. Les probabilités de détection P_{D2} et de fausse alarme P_{F2} du second détecteur sont indépendantes de P_{D1}.

Notons que nous avons les relations suivantes :

$$\frac{\partial \alpha_{i,j}}{\partial P_{D1}} = 0 \qquad i,j = -1,1$$

$$\frac{\partial \beta_{1,1}}{\partial P_{D1}} = \frac{\beta_{1,1}}{P_{D1}} \qquad \frac{\partial \beta_{-1,1}}{\partial P_{D1}} = - \frac{\beta_{-1,1}}{(1-P_{D1})} \qquad \frac{\partial \beta_{1,-1}}{\partial P_{D1}} = \frac{\beta_{1,-1}}{P_{D1}} \qquad \frac{\partial \beta_{-1,-1}}{\partial P_{D1}} = - \frac{\beta_{-1,-1}}{(1-P_{D1})}$$

En dérivant $H(H/u_1,u_2)$ par rapport à P_{D1} pour une probabilité de fausse alarme P_{F1} fixée, nous obtenons :

$$\begin{aligned}
\frac{\partial H(H/u_1,u_2)}{\partial P_{D1}} = & - \frac{\beta_{1,1}}{P_{D1}} \log \frac{\beta_{1,1}}{\alpha_{1,1}+\beta_{1,1}} + \frac{\beta_{-1,1}}{(1-P_{D1})} \log \frac{\beta_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} \\
& - \frac{\beta_{1,-1}}{P_{D1}} \log \frac{\beta_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} + \frac{\beta_{-1,-1}}{(1-P_{D1})} \log \frac{\beta_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}}
\end{aligned} \qquad (3.1.1.8)$$

$$\frac{\partial^2 H(H/u_1,u_2)}{\partial P_{D1}^{\,2}} = \left(\frac{\beta_{1,1}}{P_{D1}}\right)^2 \left[\frac{1}{\alpha_{1,1}+\beta_{1,1}} - \frac{1}{\beta_{1,1}}\right] + \left(\frac{\beta_{-1,1}}{1-P_{D1}}\right)^2 \left[\frac{1}{\alpha_{-1,1}+\beta_{-1,1}} - \frac{1}{\beta_{-1,1}}\right] + \left(\frac{\beta_{1,-1}}{P_{D1}}\right)^2 \left[\frac{1}{\alpha_{1,-1}+\beta_{1,-1}} - \frac{1}{\beta_{1,-1}}\right] + \left(\frac{\beta_{-1,-1}}{1-P_{D1}}\right)^2 \left[\frac{1}{\alpha_{-1,-1}+\beta_{-1,-1}} - \frac{1}{\beta_{-1,-1}}\right] \quad (3.1.1.9)$$

donc $\frac{\partial^2 H(H/u_1,u_2)}{\partial P_{D1}^{\,2}} < 0$ (3.1.1.10)

de plus $\frac{\partial H(H/u_1,u_2)}{\partial P_{D1}}(P_{D1} = P_{F1}) = 0$

$\frac{\partial H(H/u_1,u_2)}{\partial P_{D1}}(P_{D1})$ est donc décroissante négative sur $[P_{F1},1]$ (nous obtenons le même résultat que pour la détection centralisée). $H(H/u_1,u_2)(P_{D1})$ est donc décroissante sur $[P_{F1},1]$. Pour une probabilité de fausse alarme P_{F1} fixée, chercher le détecteur qui minimise $H(H/u_1,u_2)$ est donc équivalent à chercher le détecteur qui maximise la probabilité de détection P_{D1}.

De même on peut étudier la fonction $H(H/u_1,u_2)(P_{D2})$ à probabilité de fausse alarme P_{F2} fixée. On dérive $H(H/u_1,u_2)$ par rapport à P_{D2} pour une probabilité de fausse alarme P_{F2} fixée. Les résultats trouvés sont similaires aux précédents. On a :

$$\frac{\partial H(H/u_1,u_2)}{\partial P_{D2}} = -\frac{\beta_{1,1}}{P_{D2}} \log \frac{\beta_{1,1}}{\alpha_{1,1}+\beta_{1,1}} - \frac{\beta_{-1,1}}{P_{D2}} \log \frac{\beta_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} + \frac{\beta_{1,-1}}{(1-P_{D2})} \log \frac{\beta_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} + \frac{\beta_{-1,-1}}{(1-P_{D2})} \log \frac{\beta_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}} \quad (3.1.1.11)$$

$$\frac{\partial^2 H(H/u_1,u_2)}{\partial P_{D2}^{\,2}} = \left(\frac{\beta_{1,1}}{P_{D2}}\right)^2 \left[\frac{1}{\alpha_{1,1}+\beta_{1,1}} - \frac{1}{\beta_{1,1}}\right] + \left(\frac{\beta_{-1,1}}{P_{D2}}\right)^2 \left[\frac{1}{\alpha_{-1,1}+\beta_{-1,1}} - \frac{1}{\beta_{-1,1}}\right] + \left(\frac{\beta_{1,-1}}{1-P_{D2}}\right)^2 \left[\frac{1}{\alpha_{1,-1}+\beta_{1,-1}} - \frac{1}{\beta_{1,-1}}\right] + \left(\frac{\beta_{-1,-1}}{1-P_{D2}}\right)^2 \left[\frac{1}{\alpha_{-1,-1}+\beta_{-1,-1}} - \frac{1}{\beta_{-1,-1}}\right] \quad (3.1.1.12)$$

donc $\frac{\partial^2 H(H/u_1,u_2)}{\partial P_{D2}^{\ 2}}<0$ (3.1.1.13)

de plus $\frac{\partial H(H/u_1,u_2)}{\partial P_{D2}}(P_{D2}=P_{F2})=0$

On en déduit qu'il est équivalent de minimiser l'entropie conditionnelle $H(H/u_1,u_2)$ ou de maximiser la probabilité de détection P_{D2} pour une probabilité de fausse alarme P_{F2} fixée.

Les mêmes calculs peuvent être effectués pour des probabilités de détection P_{D1} et P_{D2} fixées. On montre alors que, pour une probabilité de détection P_{D1} (resp. P_{D2}) fixée, l'entropie conditionnelle $H(H/u_1,u_2)$ est une fonction croissante de la probabilité de fausse alarme P_{F1} (resp. P_{F2}). Pour une probabilité de détection P_{D1} (resp. P_{D2}) fixée, rechercher le détecteur qui minimise $H(H/u_1,u_2)$ est donc équivalent à chercher le détecteur qui minimise la probabilité de fausse alarme P_{F1} (resp. P_{F2}).

Minimiser $H(H/u_1,u_2)$ sans fixer ni les probabilités de fausse alarme, ni les probabilités de détection, revient à résoudre les problèmes bi-critères $\{\min P_{F1}, \max P_{D1}\}$ et $\{\min P_{F2}, \max P_{D2}\}$ en adoptant une fonction d'utilité non linéaire $H(H/u_1,u_2)$. En d'autres termes, cela revient à chercher les détecteurs qui offrent le meilleur compromis entre une probabilité de détection grande et une probabilité de fausse alarme petite.

Ces détecteurs peuvent être mis sous la forme d'un seuillage du rapport de vraisemblance. Les seuils λ_1 et λ_2 qui minimisent $H(H/u_1,u_2)$ peuvent être déterminés à partir de la connaissance des probabilités *a posteriori* associées aux différentes situations.

• Détermination des seuils λ_1 et λ_2 qui minimisent $H(H/u_1,u_2)$

Chercher le seuil λ_1 pour lequel l'entropie conditionnelle $H(H/u_1,u_2)$ est minimale revient à annuler la quantité $\left(\frac{\partial H(H/u_1,u_2)}{\partial \Lambda}\right)$.

Rappelons que :

$$P_{F1}=\int_{\lambda_1}^{+\infty} p(\Lambda/H_0)\,d\Lambda$$

$$P_{D1}=\int_{\lambda_1}^{+\infty} p(\Lambda/H_1)\,d\Lambda$$

on a donc les relations suivantes :

$$\left(\frac{\partial \alpha_{1,1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1}=-\frac{\alpha_{1,1}}{P_{F1}}P(\Lambda=\lambda_1/H_0) \qquad \left(\frac{\partial \alpha_{-1,1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1}=\frac{\alpha_{-1,1}}{(1-P_{F1})}P(\Lambda=\lambda_1/H_0)$$

$$\left(\frac{\partial \alpha_{1,-1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1}=-\frac{\alpha_{1,-1}}{P_{F1}}P(\Lambda=\lambda_1/H_0) \qquad \left(\frac{\partial \alpha_{-1,-1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1}=\frac{\alpha_{-1,-1}}{(1-P_{F1})}P(\Lambda=\lambda_1/H_0)$$

$$\left(\frac{\partial \beta_{1,1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1} = -\frac{\beta_{1,1}}{P_{D1}} P(\Lambda=\lambda_1/H_1) \qquad \left(\frac{\partial \beta_{-1,1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1} = \frac{\beta_{-1,1}}{(1-P_{D1})} P(\Lambda=\lambda_1/H_1) \tag{3.1.1.14}$$

$$\left(\frac{\partial \beta_{1,-1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1} = -\frac{\beta_{1,-1}}{P_{D1}} P(\Lambda=\lambda_1/H_1) \qquad \left(\frac{\partial \beta_{-1,-1}}{\partial \Lambda}\right)_{\Lambda=\lambda_1} = \frac{\beta_{-1,-1}}{(1-P_{D1})} P(\Lambda=\lambda_1/H_1)$$

on en déduit que :

$$\begin{aligned}\left(\frac{\partial H(H/u_1,u_2)}{\partial \Lambda}\right)_{\Lambda=\lambda_1} = & -\frac{\alpha_{1,1}}{P_{F1}} P(\Lambda=\lambda_1/H_0) \log\frac{\alpha_{1,1}}{\alpha_{1,1}+\beta_{1,1}} \\ & -\frac{\alpha_{1,-1}}{P_{F1}} P(\Lambda=\lambda_1/H_0) \log\frac{\alpha_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} \\ & +\frac{\alpha_{-1,1}}{(1-P_{F1})} P(\Lambda=\lambda_1/H_0) \log\frac{\alpha_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} \\ & +\frac{\alpha_{-1,-1}}{(1-P_{F1})} P(\Lambda=\lambda_1/H_0) \log\frac{\alpha_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}} \\ & -\frac{\beta_{1,1}}{P_{D1}} P(\Lambda=\lambda_1/H_1) \log\frac{\beta_{1,1}}{\alpha_{1,1}+\beta_{1,1}} \\ & -\frac{\beta_{1,-1}}{P_{D1}} P(\Lambda=\lambda_1/H_1) \log\frac{\beta_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} \\ & +\frac{\beta_{-1,1}}{(1-P_{D1})} P(\Lambda=\lambda_1/H_1) \log\frac{\beta_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} \\ & +\frac{\beta_{-1,-1}}{(1-P_{D1})} P(\Lambda=\lambda_1/H_1) \log\frac{\beta_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}}\end{aligned} \tag{3.1.1.15}$$

De plus, on a toujours $\frac{P(\Lambda=\lambda_1/H_1)}{P(\Lambda=\lambda_1/H_0)} = \lambda_1$. L'équation (3.1.1.15) conduit donc à choisir λ_1 tel que :

$$\lambda_1 = -\frac{\begin{aligned}&+\frac{\alpha_{1,1}}{P_{F1}}\log\frac{\alpha_{1,1}}{\alpha_{1,1}+\beta_{1,1}} + \frac{\alpha_{1,-1}}{P_{F1}}\log\frac{\alpha_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} - \frac{\alpha_{-1,1}}{(1-P_{F1})}\log\frac{\alpha_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} - \frac{\alpha_{-1,-1}}{(1-P_{F1})}\log\frac{\alpha_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}} \\ &+\frac{\beta_{1,1}}{P_{D1}}\log\frac{\beta_{1,1}}{\alpha_{1,1}+\beta_{1,1}} + \frac{\beta_{1,-1}}{P_{D1}}\log\frac{\beta_{1,-1}}{\alpha_{1,-1}+\beta_{1,-1}} - \frac{\beta_{-1,1}}{(1-P_{D1})}\log\frac{\beta_{-1,1}}{\alpha_{-1,1}+\beta_{-1,1}} - \frac{\beta_{-1,-1}}{(1-P_{D1})}\log\frac{\beta_{-1,-1}}{\alpha_{-1,-1}+\beta_{-1,-1}}\end{aligned}}{}$$

$$\lambda_1 = -\frac{\sum_{i,j\in\{-1,1\}} \frac{i\,\alpha_{i,j}}{P_{F1}^{\frac{1+i}{2}}(1-P_{F1})^{\frac{1-i}{2}}}\log\frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}}}{\sum_{i,j\in\{-1,1\}} \frac{i\,\beta_{i,j}}{P_{D1}^{\frac{1+i}{2}}(1-P_{D1})^{\frac{1-i}{2}}}\log\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}} \tag{3.1.1.16}$$

de même :

$$\lambda_2 = - \frac{\sum_{i,j\in\{-1,1\}} \frac{j\alpha_{i,j}}{P_{F2}^{\frac{1+j}{2}}(1-P_{F2})^{\frac{1-j}{2}}} \log\frac{\alpha_{i,j}}{\alpha_{i,j}+\beta_{i,j}}}{\sum_{i,j\in\{-1,1\}} \frac{j\,\beta_{i,j}}{P_{D2}^{\frac{1+j}{2}}(1-P_{D2})^{\frac{1-j}{2}}} \log\frac{\beta_{i,j}}{\alpha_{i,j}+\beta_{i,j}}} \qquad (3.1.1.17)$$

Nous pouvons remarquer que les équations (3.1.1.16) et (3.1.1.17) sont couplées. Une solution simultanée à ces deux équations permet de calculer les seuils λ_1 et λ_2 qui minimisent $H(H/u_1,u_2)$.

• Exemple

Reprenons l'un des exemples du premier chapitre : Considérons un système formé de deux capteurs dont les observations y_1 et y_2 sont indépendantes. Les fonctions de densité de probabilité sous chaque hypothèse sont des gaussiennes. Sous l'hypothèse H_0, ces fonctions de densité de probabilité sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1, ces fonctions sont supposées être de moyenne m_1 (resp. m_2) et de variance 1.

La règle de décision au niveau de chaque détecteur est : $\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i \quad i=1,2$

Deux systèmes sont considérés, le premier pour lequel $m_1=m_2=1$ et le second pour lequel $m_1=1$ et $m_2=1,5$. Les valeurs des seuils λ_1 et λ_2 qui minimisent $H(H/u_1,u_2)$ sont indiqués dans les figures 14 et 16. Sur les figures 15 et 17, les variations de $H(H/u_1,u_2)$ en fonction de P_0 sont visualisées.

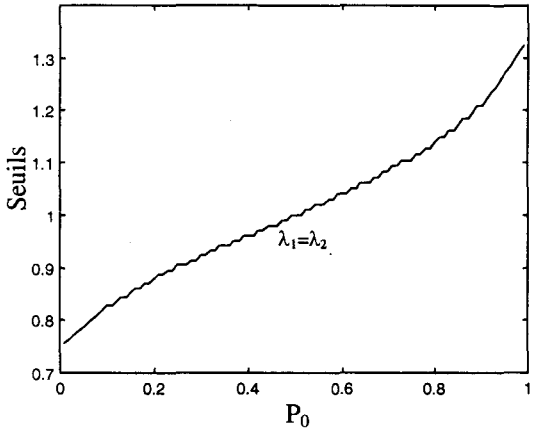


Figure 14 : Seuils en fonction de P_0, $m_1=m_2=1$.

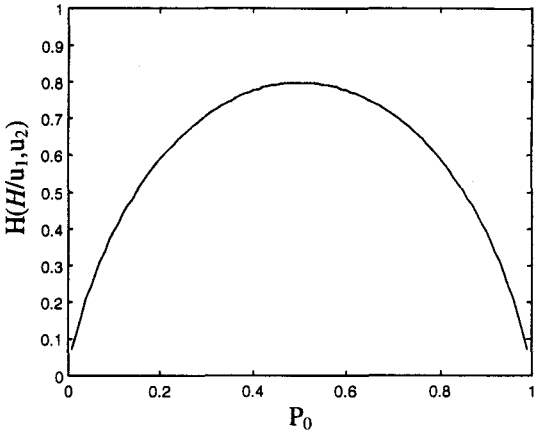


Figure 15 : $H(H/u_1,u_2)$ en fonction de P_0, $m_1=m_2=1$.

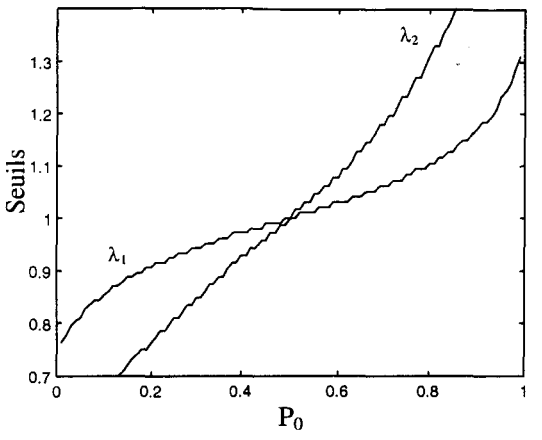


Figure 16 : Seuils en fonction de P_0, m_1=1, m_2=1,5.

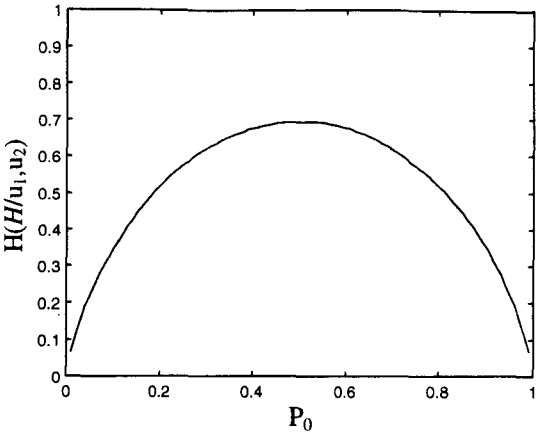


Figure 17 : $H(H/u_1,u_2)$ en fonction de P_0, m_1=1, m_2=1,5.

3.1.2. Le cas de N détecteurs en parallèle

Les résultats valables dans le cas de deux détecteurs en parallèle peuvent être étendus au cas de N détecteurs en parallèle. Dans ce cas, le système est composé de N détecteurs associés à N capteurs en parallèle. Les observations associées aux N capteurs sont toujours notées y_1, y_2, ...,y_N. Les décisions prises par chaque détecteur sont toujours notées u_i, i=1, 2, ..., N :

$$u_i = \begin{cases} 0, \text{ on décide que } H_0 \text{ est vraie} \\ 1, \text{ on décide que } H_1 \text{ est vraie} \end{cases}$$

Chaque décision binaire est caractérisée par ses probabilités de fausse alarme P_{Fi} et de détection P_{Di} telles que :

$$P_{Fi}=P(u_i=1/H_0) \qquad \text{et} \qquad P_{Di}=P(u_i=1/H_1)$$

Les décisions prises par chaque détecteur sont supposées indépendantes.

on note :

$$\alpha_{s_1,s_2,\ldots,s_N} = P_0 \prod_{i\in\{1,2,\ldots,N\}} (P_{Fi})^{\frac{1+s_i}{2}} (1-P_{Fi})^{\frac{1-s_i}{2}} \qquad s_1, s_2, \ldots, s_N \in \{-1,1\} \qquad (3.1.2.1)$$

$$\beta_{s_1,s_2,\ldots,s_N} = (1-P_0) \prod_{i\in\{1,2,\ldots,N\}} (P_{Di})^{\frac{1+s_i}{2}} (1-P_{Di})^{\frac{1-s_i}{2}} \qquad s_1, s_2, \ldots, s_N \in \{-1,1\} \qquad (3.1.2.2)$$

L'entropie conditionnelle $H(H/ u_1, u_2, \ldots, u_N)$ peut s'écrire :

$H(H/ u_1, u_2, \ldots, u_N)$

$$= - \sum_{s_1,s_2,\ldots,s_N\in\{-1,1\}} \left[\alpha_{s_1,s_2,\ldots,s_N} \log \frac{\alpha_{s_1,s_2,\ldots,s_N}}{\alpha_{s_1,s_2,\ldots,s_N}+\beta_{s_1,s_2,\ldots,s_N}} + \beta_{s_1,s_2,\ldots,s_N} \log \frac{\beta_{s_1,s_2,\ldots,s_N}}{\alpha_{s_1,s_2,\ldots,s_N}+\beta_{s_1,s_2,\ldots,s_N}} \right] \qquad (3.1.2.3)$$

Les détecteurs qui minimisent $H(H/ u_1, u_2, \ldots, u_N)$ peuvent être mis sous la forme d'un seuillage du rapport de vraisemblance. Les seuils λ_i, $i\in\{1, 2, .., N\}$, associés à ces détecteurs peuvent être déterminés à partir de la résolution d'un système de N équations non linéaires couplées de la forme :

$$\lambda_i = - \frac{\sum_{s_1,s_2,\ldots,s_N\in\{-1,1\}} \frac{s_i \alpha_{s_1,s_2,\ldots,s_N}}{P_{Fi}^{\frac{1+s_i}{2}}(1-P_{Fi})^{\frac{1-s_i}{2}}} \log \frac{\alpha_{s_1,s_2,\ldots,s_N}}{\alpha_{s_1,s_2,\ldots,s_N}+\beta_{s_1,s_2,\ldots,s_N}}}{\sum_{s_1,s_2,\ldots,s_N\in\{-1,1\}} \frac{s_i \beta_{s_1,s_2,\ldots,s_N}}{P_{Di}^{\frac{1+s_i}{2}}(1-P_{Di})^{\frac{1-s_i}{2}}} \log \frac{\beta_{s_1,s_2,\ldots,s_N}}{\alpha_{s_1,s_2,\ldots,s_N}+\beta_{s_1,s_2,\ldots,s_N}}} \qquad i\in\{1, 2, .., N\} \qquad (3.1.2.4)$$

3.2. Optimisation de l'opérateur de fusion

Nous considérons dans ce paragraphe l'optimisation de l'opérateur de fusion comme dans le paragraphe 5.1.2.2. du chapitre 1. Le système est composé de N détecteurs locaux q_i i=1,...,N associés à N capteurs y_i i=1,...,N tels que :

$$u_i = \begin{cases} 0, \text{ le détecteur i décide que } H_0 \text{ est vraie} \\ 1, \text{ le détecteur i décide que } H_1 \text{ est vraie} \end{cases}$$

Les décisions locales sont ensuite transmises à un opérateur de fusion qui les combine de façon à obtenir la décision finale u_0 (Figure 18).

$$u_0 = \begin{cases} 0, H_0 \text{ est vraie} \\ 1, H_1 \text{ est vraie} \end{cases}$$

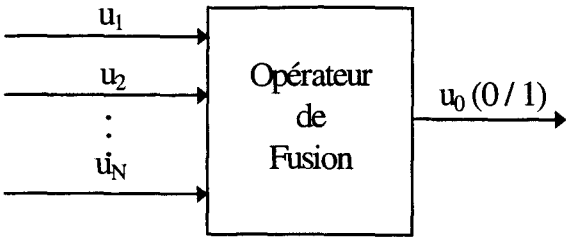


Figure 18 : L'opérateur de fusion

Chaque décision binaire transmise à l'opérateur de fusion est caractérisée par ses probabilités de fausse alarme P_{Fi} et de détection P_{Di} telles que :

$$P_{Fi}=P(u_i=1/H_0) \qquad \text{et} \qquad P_{Di}=P(u_i=1/H_1)$$

Les probabilités de fausse alarme et de détection du système complet sont notées P_F et P_D telles que :

$$P_F=P(u_0=1/H_0) \qquad \text{et} \qquad P_D=P(u_0=1/H_1)$$

L'opérateur de fusion sera une fonction logique des N décisions binaires qui lui sont transmises. Dans le cas général, il y a 2^{2^N} opérateurs de fusion possibles. Considérons l'optimisation à l'aide d'un critère entropique de l'opérateur de fusion. On a démontré au paragraphe 2 que $H(H/u_0)$ peut se mettre sous la forme :

$$\begin{aligned} H(H/u_0)= & - P_0 P_F \log \frac{P_0 P_F}{P_0 P_F + (1-P_0) P_D} \\ & - P_0 (1-P_F) \log \frac{P_0 (1-P_F)}{P_0 (1-P_F) + (1-P_0)(1-P_D)} \\ & - (1-P_0) P_D \log \frac{(1-P_0) P_D}{P_0 P_F + (1-P_0) P_D} \\ & - (1-P_0)(1-P_D) \log \frac{(1-P_0)(1-P_D)}{P_0 (1-P_F) + (1-P_0)(1-P_D)} \end{aligned} \qquad (3.2.1)$$

L'objectif est de déterminer la règle de fusion qui minimise l'entropie conditionnelle $H(H/u_0)$. Dans le cas de deux capteurs on ne considérera, comme au paragraphe 5.1.2.2 du chapitre 1, que les 6 fonctions de fusion monotones. Parmi toutes ces fonctions, il ne faudra retenir que celle(s) qui minimise(nt) $H(H/u_0)$.

Entrées		Décision finale					
u_1	u_2	f_1	f_2	f_4	f_6	f_8	f_{16}
0	0	0	0	0	0	0	1
0	1	0	0	0	1	1	1
1	0	0	0	1	0	1	1
1	1	0	1	1	1	1	1

Figure 19 : Les fonctions de fusion admissibles

Dans le cas où les décisions locales sont indépendantes, les probabilités de fausse alarme P_F et de détection P_D du système complet peuvent s'exprimer en fonction des probabilités de fausse alarme P_{Fi} et de détection P_{Di} (i=1,2) des différents détecteurs (Figure 20).

	f_1	f_2	f_4	f_6	f_8	f_{16}
P_F	0	$P_{D1}P_{D2}$	P_{D1}	$P_{D1}+P_{D2}-P_{D1}P_{D2}$	P_{D2}	1
P_D	0	$P_{F1}P_{F2}$	P_{F1}	$P_{F1}+P_{F2}-P_{F1}P_{F2}$	P_{F2}	1

Figure 20 : Valeurs de P_F et P_D suivant la fonction de fusion

La règle de fusion optimale sera celle qui minimise $H(H/u_0)$. Pour déterminer la valeur de cette fonction, il suffit de remplacer dans l'équation (3.2.1), P_F et P_D par leurs valeurs en fonction de P_{Fi} et P_{Di} pour toutes les règles de fusion possibles.

• Exemple

Considérons un système formé de deux capteurs dont les observations y_1 et y_2 sont indépendantes. Les fonctions de densité de probabilités sous chaque hypothèse sont des gaussiennes. Sous l'hypothèse H_0 ces fonctions de densité de probabilités sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1 ces fonctions sont supposées être de variance 1, et respectivement de moyenne m_1 pour Y_1 et m_2 pour Y_2.

La règle de décision au niveau de chaque détecteur est la suivante : $\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i \quad i=1,2$

Deux systèmes ont été étudiés, le premier pour lequel $m_1=m_2=1$ et le second pour lequel $m_1=1$ et $m_2=1,5$. Les valeurs des seuils λ_1 et λ_2 sont celles trouvées dans le paragraphe 3.1.1. Le but est alors de déterminer la règle de fusion qui minimise $H(H/u_0)$. Sur les figures 21 et 24, la variation de $H(H/u_0)$ en fonction de P_0 est visualisée. Les figures 23 et 26 représentent les courbes C.O.R. pour chaque système en fonction de la fonction de fusion utilisée. Sur les figures 22 et 25, la variation de la probabilité d'erreur en fonction de P_0 et en fonction de la fonction de fusion utilisée est représentée. Au vu des figures 21 et 24, la règle de fusion qui minimise l'entropie conditionnelle $H(H/u_0)$ dépend de P_0. Dans le cas ou $m_1=m_2=1$, c'est le « ou logique » qui est optimal pour $0<P_0<0.5$, puis c'est le « et logique » pour $0.5<P_0<1$. Dans le cas ou $m_1=1$ et $m_2=1.5$, c'est la décision prise par le capteur 2 (f_6) qui est optimale quelque soit la valeur prise par P_0. Pour $P_0=0.5$, les résultats trouvés en terme de seuil, et de probabilité d'erreur sont les mêmes que ceux trouvés dans le cas Bayésien du paragraphe 5.1.2.2. du chapitre 1. Par contre, lorsque P_0 varie, le critère entropique ne permet pas de minimiser la probabilité d'erreur, en revanche les probabilités de détection et de fausse alarme trouvées restent dans des limites acceptables, ce qui n'est pas le cas en appliquant le critère de Bayes.

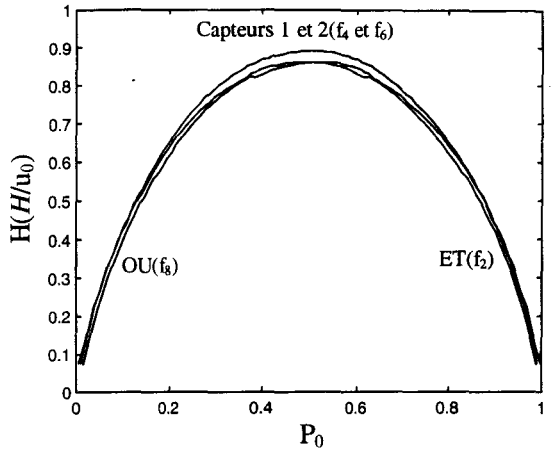


Figure 21 : H(H/u_0) en fonction de P_0, $m_1=m_2=1$.

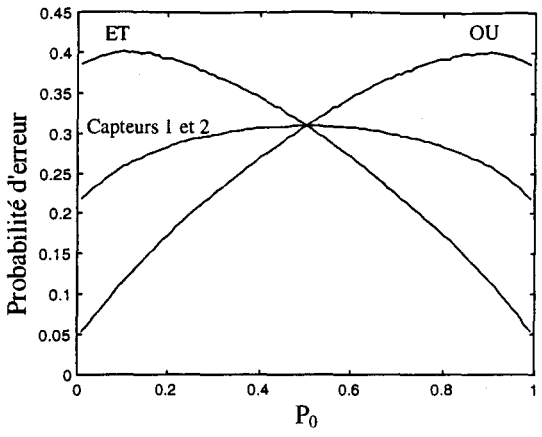


Figure 22 : Probabilités d'erreur en fonction de P_0, $m_1=m_2=1$.

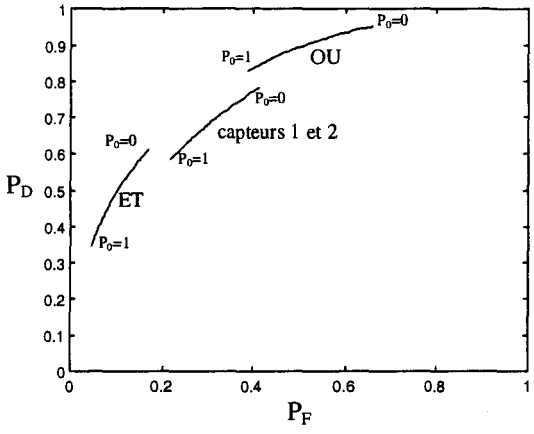


Figure 23 : Courbes C.O.R. $P_D=f(P_F)$, $m_1=m_2=1$.

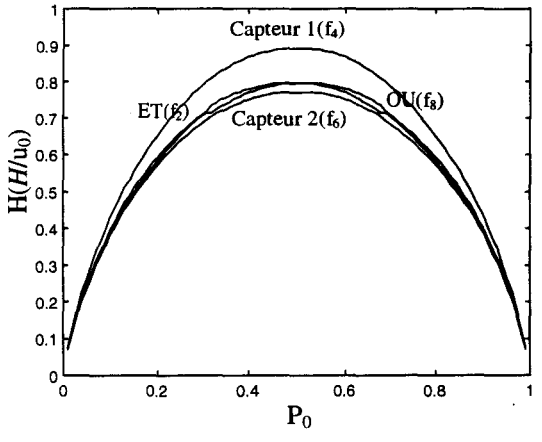


Figure 24 : H(H/u_0) en fonction de P_0

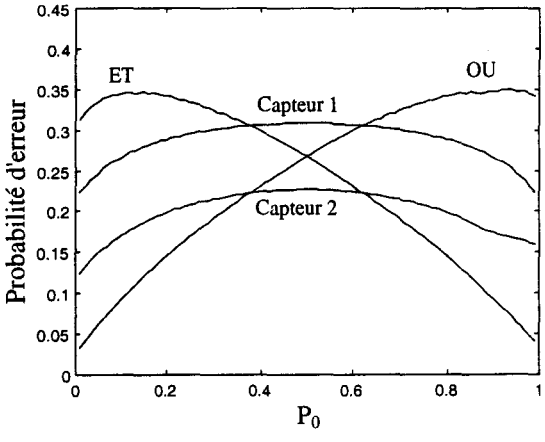


Figure 25 : Probabilités d'erreur en fonction de P_0, $m_1=1$, $m_2=1,5$.

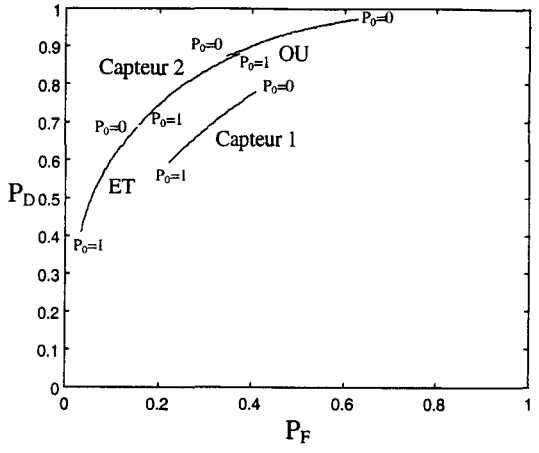


Figure 26 : Courbes C.O.R. $P_D=f(P_F)$, $m_1=1$, $m_2=1,5$.

3.3. Optimisation simultanée des détecteurs locaux et de l'opérateur de fusion

Considérons l'optimisation simultanée des détecteurs locaux et de l'opérateur de fusion. Le système considéré est composé de N détecteurs locaux q_i i=1, ..., N associés à N capteurs Y_i i=1, ..., N qui observent le même phénomène. Les décisions locales sont ensuite transmises à un opérateur de fusion qui les combine de façon à obtenir la décision finale u_0 (Figure 27).

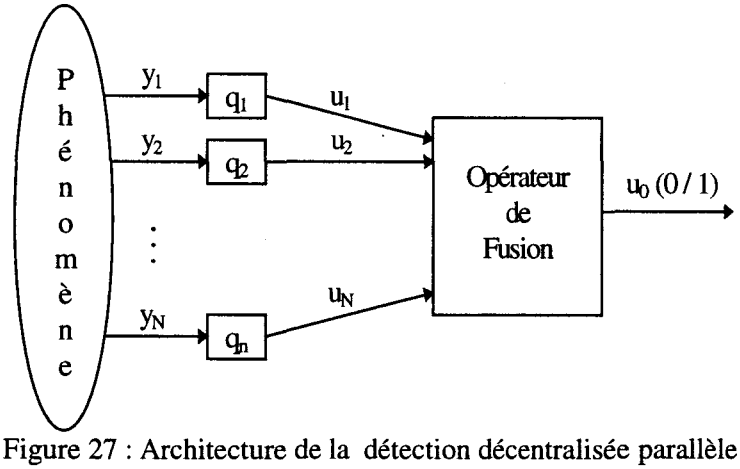


Figure 27 : Architecture de la détection décentralisée parallèle

Chaque détecteur local est caractérisé par ses probabilités de fausse alarme P_{Fi} et de détection P_{Di}. Les probabilités de fausse alarme et de détection du système complet sont, quant à elles, notées P_F et P_D.

• Expression de l'entropie $H(H/u_0)$

Nous cherchons à optimiser l'architecture de ce système en utilisant un critère informationnel. Celui-ci consiste en la minimisation de l'entropie conditionnelle $H(H/u_0)$. Or, on sait que $H(H/u_0)=E\{\log(1/P(H/u_0)\}$. On peut poser :

$$C_{ij}=\log \frac{1}{p(H_j / u_0 = i)} \qquad (i, j=0,1) \tag{3.3.1}$$

$H(H/u_0)$ peut s'écrire :

$$\begin{aligned} H(H/u_0)= \quad & - P(u_0=0,H_0) \log P(H_0/u_0=0) \\ & - P(u_0=1,H_0) \log P(H_0/ u_0=1) \\ & - P(u_0=0,H_1) \log P(H_1/ u_0=0) \\ & - P(u_0=1,H_1) \log P(H_1/ u_0=1) \end{aligned} \tag{3.3.2}$$

$$H(H/u_0)= \quad P_0 (1-P_F) C_{00} + P_0 P_F C_{10} + (1-P_0) (1-P_D) C_{01} + (1-P_0) P_D C_{11}$$

$$H(H/u_0)= \quad P_0 P_F (C_{10} - C_{00}) - (1-P_0) P_D (C_{01}-C_{11}) + P_0 C_{00} + (1-P_0) C_{01} \tag{3.3.3}$$

Minimiser $H(H/u_0)$ revient alors à minimiser une fonction « risque moyen » pour laquelle les fonctions de coût ne sont pas constantes mais dépendent des probabilités *a posteriori*. L'optimisation du système se fait élément par élément. Dans un premier temps, nous chercherons à optimiser l'opérateur de fusion, les détecteurs locaux étant alors supposés fixés.

• **Optimisation de l'opérateur de fusion** [HoV89b]

Soit ν une variable telle que $H(H/u_0)$ soit une fonction de ν et $\nu \in [0,1]$. Si l'on dérive $H(H/u_0)$ par rapport à ν on obtient :

$$\frac{\partial H(H/u_0)}{\partial \upsilon} = F(\partial \nu) + P_0 (C_{10}-C_{00}) \frac{\partial P_F}{\partial \upsilon} - (1-P_0) (C_{01}-C_{11}) \frac{\partial P_D}{\partial \upsilon} \tag{3.3.4}$$

$$\text{où } F(\partial \nu)= \frac{\partial C_{10}}{\partial \upsilon} P_0 P_F + \frac{\partial C_{00}}{\partial \upsilon} P_0(1-P_F) + \frac{\partial C_{11}}{\partial \upsilon} (1-P_0)P_D + \frac{\partial C_{01}}{\partial \upsilon} (1-P_0)(1-P_D) \tag{3.3.5}$$

et

$$C_{10}= \log \frac{1}{p(H_0/ u_0 = 1)} = \log \frac{P(u_0 = 1/ H_0) P_0 + P(u_0 = 1/ H_1)(1-P_0)}{P(u_0 = 1/ H_0) P_0} = \log \frac{P_0 P_F + (1-P_0) P_D}{P_0 P_F}$$

$$C_{00}=\log \frac{P_0 (1-P_F) + (1-P_0)(1-P_D)}{P_0 (1-P_F)}$$

$$C_{11}=\log \frac{P_0 P_F + (1-P_0) P_D}{(1-P_0) P_D} \tag{3.3.6}$$

$$C_{01}=\log \frac{P_0 (1-P_F) + (1-P_0) (1-P_D)}{(1-P_0)(1-P_D)}$$

En introduisant les équations (3.3.6) dans l'équation (3.3.5), on montre que la quantité $F(\partial \nu)$ est égale à zéro. Par conséquent (3.3.4) peut s'écrire :

$$\frac{\partial H(H/u_0)}{\partial \upsilon} = P_0 (C_{10}-C_{00}) \frac{\partial P_F}{\partial \upsilon} - (1-P_0) (C_{01}-C_{11}) \frac{\partial P_D}{\partial \upsilon} \tag{3.3.7}$$

Lorsque cette expression est négative, cela signifie que $H(H/u_0)$ est une fonction décroissante de v. Lorsqu'elle est positive, cela signifie que $H(H/u_0)$ est une fonction croissante de v. Lorsqu'elle est égale à zéro, à la valeur de v trouvée correspond un extremum de la fonction $H(H/u_0)$.

D'autre part, on a :

$$P_D = P(u_0=1/H_1) = \sum_u P(u_0=1/u).P(u/H_1)$$
$$P_F = P(u_0=1/H_0) = \sum_u P(u_0=1/u).P(u/H_0) \qquad (3.3.8)$$

Et si l'on pose $v=P(u_0=1/u^*)$, la probabilité de décider $u_0=1$ sachant que le vecteur u est égal à u^*, l'optimisation élément par élément donne :

$$\frac{\partial H(H/u_0)}{\partial P(u_0=1/u^*)} = P_0 (C_{10}-C_{00}) \frac{\partial P_F}{\partial P(u_0=1/u^*)} - (1-P_0)(C_{01}-C_{11}) \frac{\partial P_D}{\partial P(u_0=1/u^*)}$$

$$\frac{\partial H(H/u_0)}{\partial P(u_0=1/u^*)} = P_0 (C_{10}-C_{00}) P(u^*/H_0) - (1-P_0)(C_{01}-C_{11}) P(u^*/H_1) \qquad (3.3.9)$$

L'opérateur de fusion est obtenu en observant que, lorsque la quantité (3.3.9) est négative (resp. positive) l'entropie $H(H/u_0)$ est un fonction décroissante (resp. croissante) de la variable $P(u_0=1/u^*)$. Par conséquent pour minimiser $H(H/u_0)$ il faut fixer $P(u_0=1/u^*)=1$ lorsque $H(H/u_0)$ est une fonction décroissante de $P(u_0=1/u^*)$, et $P(u_0=1/u^*)=0$ lorsque $H(H/u_0)$ est une fonction croissante de $P(u_0=1/u^*)$. Ce qui équivaut à appliquer la règle suivante :

$$P_0 (C_{10}-C_{00}) P(u^*/H_0) - (1-P_0)(C_{01}-C_{11}) P(u^*/H_1) \underset{P(u_0=1/u^*)=1}{\overset{P(u_0=1/u^*)=0}{\gtrless}} 0 \qquad (3.3.10)$$

En faisant l'hypothèse raisonnable que le coût d'une mauvaise décision est supérieur au coût d'une bonne décision, c'est-à-dire que $C_{10} > C_{00}$ et $C_{01} > C_{11}$, l'équation (3.3.10) peut se mettre sous la forme :

$$\frac{P(u^*/H_1)}{P(u^*/H_0)} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \frac{P_0}{1-P_0} \frac{(C_{10}-C_{00})}{(C_{01}-C_{11})} \qquad (3.3.11)$$

L'équation (3.3.11) est similaire à l'équation (3.1.2.3.11) du chapitre 1. Les fonctions de coût ne sont pas des constantes mais dépendent des probabilités *a posteriori*.

• Optimisation des détecteurs locaux

L'optimisation élément par élément du détecteur k (k=1, ..., N) peut être obtenue en dérivant $H(H/u_0)$ par rapport à P_{Fk}. On pose $v= P_{Fk}$ dans l'équation (3.3.7), et on obtient :

$$\frac{\partial H(H/u_0)}{\partial P_{Fk}} = P_0 (C_{10}-C_{00}) \frac{\partial P_F}{\partial P_{Fk}} - (1-P_0)(C_{01}-C_{11}) \frac{\partial P_D}{\partial P_{Fk}} \qquad (3.3.12)$$

sachant que le seuil λ_k associé au détecteur k est donné par $\lambda_k = \frac{\partial P_{Dk}}{\partial P_{Fk}}$

et que l'on a $\frac{\partial P_D}{\partial P_{Fk}} = \frac{\partial P_D}{\partial P_{Dk}} \lambda_k \qquad (3.3.13)$

en substituant (3.3.13) dans (3.3.12), on obtient :

$$\frac{\partial H(H/u_0)}{\partial P_{Fk}} = P_0 (C_{10}-C_{00}) \frac{\partial P_F}{\partial P_{Fk}} - (1-P_0)(C_{01}-C_{11}) \frac{\partial P_D}{\partial P_{Dk}} \lambda_k \qquad (3.3.14)$$

Minimiser $H(H/u_0)$ revient à dire que (3.3.14) est nulle, ce qui revient à choisir :

$$\lambda_k = \frac{P_0}{1-P_0} \frac{(C_{10}-C_{00}) \frac{\partial P_F}{\partial P_{Fk}}}{(C_{01}-C_{11}) \frac{\partial P_D}{\partial P_{Dk}}} \qquad (k=1, \ldots, N) \qquad (3.3.15)$$

Une solution commune aux équations (3.3.11) et (3.3.15) est une solution de l'optimisation élément par élément utilisant le critère entropique du problème de détection décentralisée de la Figure 27.

• Application

En considérant deux capteurs dont les observations sont indépendantes, une expression explicite des seuils peut être obtenue en fonction de l'opérateur de fusion qui a été retenu. Ainsi l'équation (3.3.15) permet d'écrire :

$$\lambda_1 = \frac{P_0}{1-P_0} \frac{(C_{10}-C_{00}) \frac{\partial P_F}{\partial P_{F1}}}{(C_{01}-C_{11}) \frac{\partial P_D}{\partial P_{D1}}}$$

$$\lambda_2 = \frac{P_0}{1-P_0} \frac{(C_{10}-C_{00}) \frac{\partial P_F}{\partial P_{F2}}}{(C_{01}-C_{11}) \frac{\partial P_D}{\partial P_{D2}}}$$

Dans le cas où l'opérateur de fusion est un « et logique », les performances en terme de probabilités de fausse alarme P_F et de détection P_D du système complet peuvent être calculées et sont données par :

$P_D = P_{D1} \cdot P_{D2}$
$P_F = P_{F1} \cdot P_{F2}$

On en déduit que :

$$\lambda_1 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})P_{F2}}{(C_{01}-C_{11})P_{D2}} \qquad (3.3.16)$$

$$\lambda_2 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})P_{F1}}{(C_{01}-C_{11})P_{D1}} \qquad (3.3.17)$$

où :

$$C_{10} = \log \frac{P_0 P_{F1}P_{F2} + (1-P_0)P_{D1}P_{D2}}{P_0 P_{F1}P_{F2}}$$

$$C_{00} = \log \frac{P_0 (1-P_{F1}P_{F2}) + (1-P_0)(1-P_{D1}P_{D2})}{P_0 (1-P_{F1}P_{F2})}$$

$$C_{11} = \log \frac{P_0 P_{F1}P_{F2} + (1-P_0)P_{D1}P_{D2}}{(1-P_0)P_{D1}P_{D2}}$$

$$C_{01} = \log \frac{P_0 (1-P_{F1}P_{F2}) + (1-P_0)(1-P_{D1}P_{D2})}{(1-P_0)(1-P_{D1}P_{D2})}$$

Les équations (3.3.16) et (3.3.17) sont couplées. Leur résolution permet de trouver les valeurs des seuils λ_1 et λ_2 qui minimisent $H(H/u_0)$ dans le cas où l'opérateur de fusion est un « et logique ».

Dans le cas où l'opérateur de fusion est un « ou logique », les performances en terme de probabilités de fausse alarme P_F et de détection P_D du système complet sont données par :

$$P_D = P_{D1} + P_{D2} - P_{D1} . P_{D2}$$
$$P_F = P_{F1} + P_{F2} - P_{F1} . P_{F2}$$

On en déduit que :

$$\lambda_1 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})(1-P_{F2})}{(C_{01}-C_{11})(1-P_{D2})} \qquad (3.3.18)$$

$$\lambda_2 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})(1-P_{F1})}{(C_{01}-C_{11})(1-P_{D1})} \qquad (3.3.19)$$

où :

$$C_{10} = \log \frac{P_0 (P_{F1}+P_{F2}-P_{F1}P_{F2}) + (1-P_0)(P_{D1}+P_{D2}-P_{D1}P_{D2})}{P_0 (P_{F1}+P_{F2}-P_{F1}P_{F2})}$$

$$C_{00} = \log \frac{P_0 (1-(P_{F1}+P_{F2}-P_{F1}P_{F2})) + (1-P_0)(1-(P_{D1}+P_{D2}-P_{D1}P_{D2}))}{P_0 (1-(P_{F1}+P_{F2}-P_{F1}P_{F2}))}$$

$$C_{11} = \log \frac{P_0 (P_{F1}+P_{F2}-P_{F1}P_{F2}) + (1-P_0)(P_{D1}+P_{D2}-P_{D1}P_{D2})}{(1-P_0)(P_{D1}+P_{D2}-P_{D1}P_{D2})}$$

$$C_{01} = \log \frac{P_0\,(1-(P_{F1}+P_{F2}-P_{F1}P_{F2}))+(1-P_0)\;(1-(P_{D1}+P_{D2}-P_{D1}P_{D2}))}{(1-P_0)(1-(P_{D1}+P_{D2}-P_{D1}P_{D2}))}$$

Les équations (3.3.18) et (3.3.19) sont couplées. Leur résolution permet la détermination des valeurs des seuils λ_1 et λ_2 qui minimisent $H(H/u_0)$ dans le cas où l'opérateur de fusion est un « ou logique ».

• Exemple

Reprenons un exemple du premier chapitre : Considérons un système formé de deux capteurs dont les observations y_1 et y_2 sont indépendantes. Les fonctions de densité de probabilité sous chaque hypothèse sont des gaussiennes. Sous l'hypothèse H_0, ces fonctions de densité de probabilités sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1, ces fonctions sont supposées être de moyenne respectives m_1 et m_2, et de variance 1.

La règle de décision au niveau de chaque détecteur est la suivante : $\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i \quad i=1,2$

Les deux mêmes systèmes ont été étudiés : $m_1=m_2=1$ pour le premier ; et $m_1=1$ et $m_2=1,5$ pour le second. Les seuils optimaux au niveau des deux détecteurs dépendent de l'opérateur de fusion utilisé et sont indiqués dans les figures 28 et 32. Sur les figures 29 et 33, la variation de $H(H/u_0)$ en fonction de P_0 est représentée. Sur les figures 30 et 34, la variation de la probabilité d'erreur en fonction de P_0 est représentée. Comme on peut le remarquer sur les figures 29 et 33, la règle de fusion optimale dépend de P_0. C'est le « ou logique » qui est optimal pour $0<P_0<0.5$, puis c'est le « et logique » pour $0.5<P_0<1$. Les figures 31 et 35 représentent les courbes C.O.R. pour chaque système. Pour $P_0=0.5$, les résultats trouvés en terme de probabilité d'erreur sont les mêmes que ceux trouvés dans le cas Bayésien du paragraphe 5.1.2.3. du chapitre 1. Par contre, lorsque P_0 varie, le critère entropique ne permet pas de minimiser la probabilité d'erreur, en revanche les probabilités de détection et de fausse alarme trouvées restent dans des limites acceptables, ce qui n'est pas le cas en appliquant le critère de Bayes.

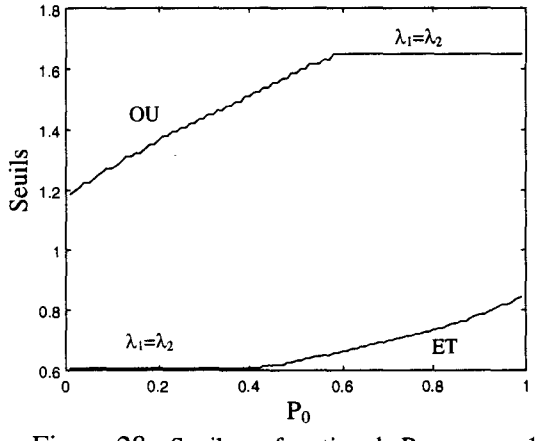


Figure 28 : Seuils en fonction de P_0, $m_1=m_2=1$

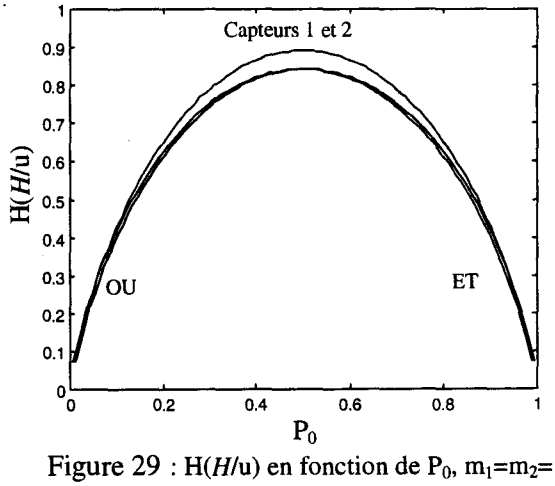


Figure 29 : $H(H/u)$ en fonction de P_0, $m_1=m_2=1$

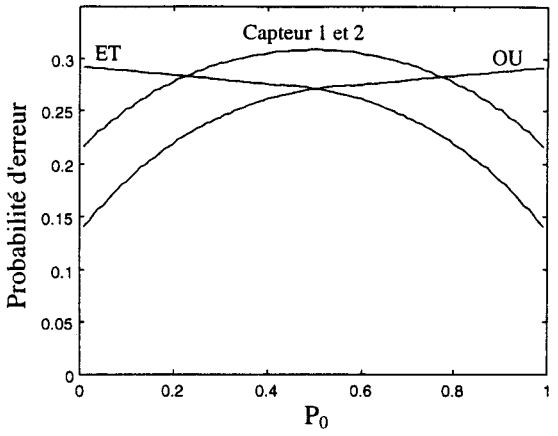


Figure 30 : Probabilités d'erreur en fonction de P_0, $m_1=m_2=1$

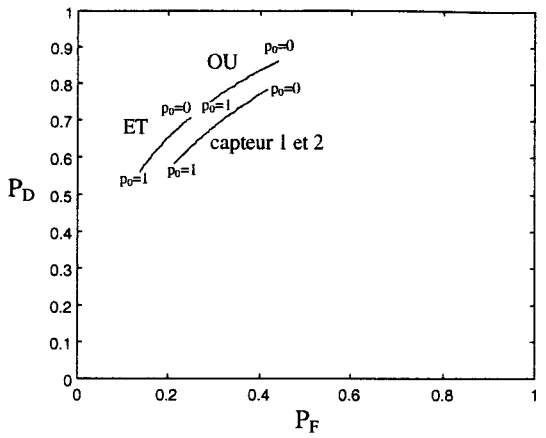


Figure 31 : Courbes C.O.R. $P_D=f(P_F)$, $m_1=m_2=1$

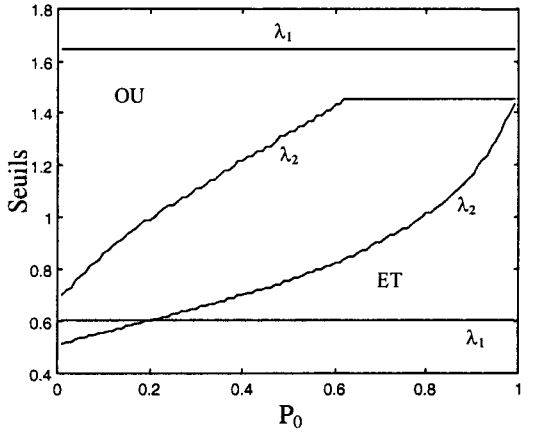


Figure 32 : Seuils en fonction de P_0, $m_1=1$, $m_2=1,5$

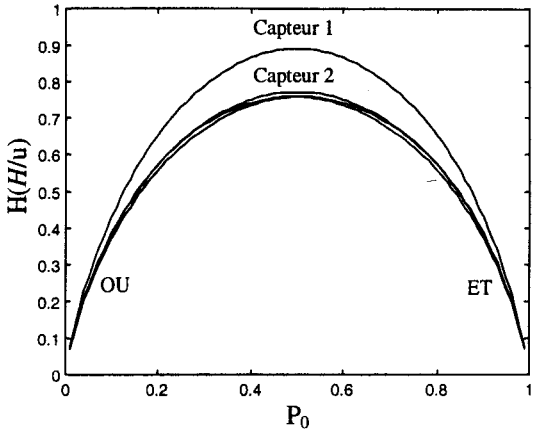


Figure 33 : H(H/u) en fonction de P_0, $m_1=1$, $m_2=1,5$

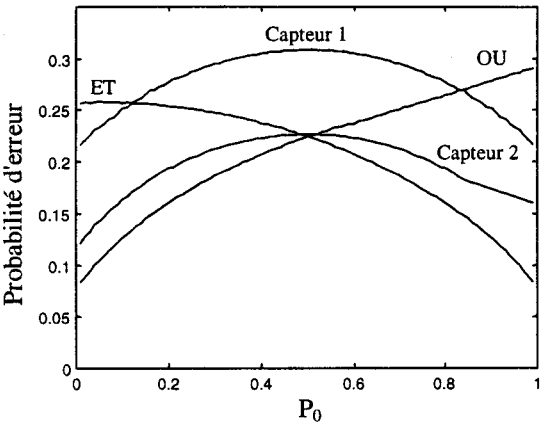


Figure 34 : Probabilités d'erreur en fonction de P_0, $m_1=1$, $m_2=1,5$

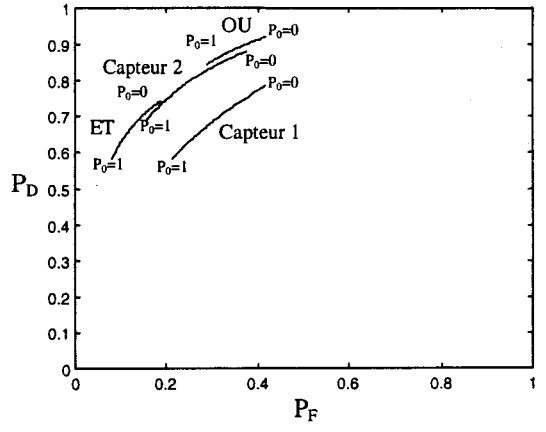


Figure 35 : Courbes C.O.R. $P_D=f(P_F)$, $m_1=1$, $m_2=1,5$

4. Application du critère entropique au cas de la détection décentralisée série

Nous considérons, dans ce paragraphe, le problème de la détection décentralisée série (Figure 36). Chaque détecteur local reçoit une information issue d'un capteur et transmet un message binaire à son successeur. La décision du premier détecteur est basée sur les informations issues d'un seul capteur et c'est le dernier détecteur qui élabore la décision finale.

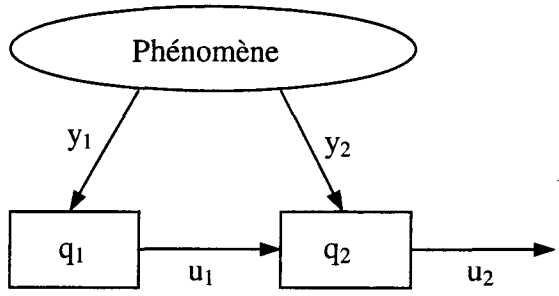


Figure 36 : Architecture de détection décentralisée série à deux détecteurs

Nous étudierons un système en tandem formé de deux détecteurs q_1 et q_2. Les hypothèses possibles sont toujours H_0 et H_1 auxquelles sont associées les probabilités *a priori* P_0 et P_1. Les observations issues de chaque capteur sont notées respectivement y_1 et y_2. Les fonctions de densité de probabilités conditionnelles sous chaque hypothèse sont notées $p(y_1,y_2/H_k)$, k=0,1. Le détecteur q_1 élabore sa décision u_1 à partir de y_1. Cette décision est ensuite transmise au détecteur q_2 qui élabore la décision finale u_2 à partir de u_1 et de y_2. Chaque décision binaire est telle que :

$$u_i = \begin{cases} 0, \text{ le détecteur i décide que } H_0 \text{ est vraie} \\ 1, \text{ le détecteur i décide que } H_1 \text{ est vraie} \end{cases}$$

Chaque détecteur local est caractérisé par ses probabilités de fausse alarme P_{Fi} et de détection P_{Di}. Les décisions prises par chaque détecteur sont supposées indépendantes.

• Expression de l'entropie $H(H/u_0)$

Notre approche consiste à optimiser l'architecture de ce système en utilisant un critère informationnel qui permet de minimiser l'entropie conditionnelle $H(H/u_2)$. Or, on sait que $H(H/u_2)=E\{\log(1/P(H/u_2)\}$. On peut poser :

$$C_{ij}=\log \frac{1}{P(H_j/u_2=i)} \quad (i,j=0,1) \tag{4.1}$$

$H(H/u_2)$ peut s'écrire :

$$\begin{aligned} H(H/u_2) = & - P(u_2=0,H_0) \log P(H_0/u_2=0) \\ & - P(u_2=1,H_0) \log P(H_0/ u_2=1) \end{aligned}$$

$$- P(u_2=0,H_1) \log P(H_1/ u_2=0)$$
$$- P(u_2=1,H_1) \log P(H_1/ u_2=1) \quad (4.2)$$

$$H(H/u_2) = P_0 (1-P_{F2}) C_{00} + P_0 P_{F2} C_{10} + (1-P_0) (1-P_{D2}) C_{01} + (1-P_0) P_{D2} C_{11}$$
$$H(H/u_2) = P_0 P_{F2} (C_{10} - C_{00}) - (1-P_0) P_{D2} (C_{01}-C_{11}) + P_0 C_{00} + (1-P_0) C_{01} \quad (4.3)$$

Minimiser $H(H/u_2)$ revient alors à minimiser une fonction risque moyen pour laquelle les fonctions de coût ne sont pas constantes mais dépendent des probabilités *a posteriori.* L'optimisation de ce système se fait élément par élément. Dans un premier temps, c'est le deuxième détecteur qui est optimisé, le premier détecteur est alors supposé fixé. De même, lors de l'optimisation du premier détecteur, c'est le second détecteur qui sera supposé fixé.

• Optimisation du second détecteur

Soit ν une variable telle que $H(H/u_2)$ soit une fonction de ν et $\nu \in [0,1]$. Si l'on dérive $H(H/u_2)$ par rapport à ν, on obtient :

$$\frac{\partial H(H/u_2)}{\partial \upsilon} = F(\partial \nu) + P_0 (C_{10}-C_{00}) \frac{\partial P_{F2}}{\partial \upsilon} - (1-P_0) (C_{01}-C_{11}) \frac{\partial P_{D2}}{\partial \upsilon} \quad (4.4)$$

$$\text{où } F(\partial \nu) = \frac{\partial C_{10}}{\partial \upsilon} P_0 P_{F2} + \frac{\partial C_{00}}{\partial \upsilon} P_0(1-P_{F2}) + \frac{\partial C_{11}}{\partial \upsilon} (1-P_0)P_{D2} + \frac{\partial C_{01}}{\partial \upsilon} (1-P_0)(1-P_{D2}) \quad (4.5)$$

et

$$C_{10} = \log \frac{1}{P(H_0/ u_2 = 1)} = \log \frac{P(u_2 = 1/ H_0) P_0 + P(u_2 = 1/ H_1)(1-P_0)}{P(u_2 = 1/ H_0) P_0} = \log \frac{P_0 P_{F2} + (1-P_0) P_{D2}}{P_0 P_{F2}}$$

$$C_{00} = \log \frac{P_0 (1-P_{F2}) + (1-P_0)(1-P_{D2})}{P_0 (1-P_{F2})}$$

$$C_{11} = \log \frac{P_0 P_{F2} + (1-P_0) P_{D2}}{(1-P_0) P_{D2}} \quad (4.6)$$

$$C_{01} = \log \frac{P_0 (1-P_{F2}) + (1-P_0) (1-P_{D2})}{(1-P_0)(1-P_{D2})}$$

En introduisant les équations (4.6) dans l'équation (4.5), on montre que la quantité $F(\partial \nu)$ est égale à zéro. Par conséquent (4.4) peut s'écrire :

$$\frac{\partial H(H/u_2)}{\partial \upsilon} = P_0 (C_{10}-C_{00}) \frac{\partial P_{F2}}{\partial \upsilon} - (1-P_0) (C_{01}-C_{11}) \frac{\partial P_{D2}}{\partial \upsilon} \quad (4.7)$$

Lorsque cette expression est négative, cela signifie que $H(H/u_2)$ est une fonction décroissante de ν. Lorsqu'elle est positive, cela signifie que $H(H/u_2)$ est une fonction croissante de ν. Lorsqu'elle est égale à zéro, à la valeur de ν trouvée correspond à un extremum de la fonction $H(H/u_2)$.

D'autre part, on a :

$$P_{D2} = P(u_2=1/H_1) = \sum_{u_1,y_2} P(u_2=1/u_1,y_2).P(u_1,y_2/H_1)$$

$$P_{F2} = P(u_2=1/H_0) = \sum_{u_1,y_2} P(u_2=1/u_1,y_2).P(u_1,y_2/H_0) \qquad (4.8)$$

Et si l'on pose $v=P(u_2=1/u_1{}^*,y_2{}^*)$, la probabilité de décider $u_2=1$ sachant que u_1 et y_2 prennent les valeurs $u_1{}^*$ et $y_2{}^*$, l'optimisation élément par élément donne :

$$\frac{\partial H(H/u_2)}{\partial P(u_2=1/u_1{}^*,y_2{}^*)} = P_0\,(C_{10}-C_{00})\,\frac{\partial P_{F2}}{\partial P(u_2=1/u_1{}^*,y_2{}^*)} - (1-P_0)\,(C_{01}-C_{11})\frac{\partial P_{D2}}{\partial P(u_2=1/u_1{}^*,y_2{}^*)}$$

$$\frac{\partial H(H/u_2)}{\partial P(u_2=1/u_1{}^*,y_2{}^*)} = P_0\,(C_{10}-C_{00})\,P(u_1{}^*,y_2{}^*/H_0) - (1-P_0)\,(C_{01}-C_{11})\,P(u_1{}^*,y_2{}^*/H_1) \qquad (4.9)$$

L'opérateur de fusion est obtenu en observant que, lorsque la quantité (4.9) est négative (resp. positive) l'entropie $H(H/u_2)$ est un fonction décroissante (resp. croissante) de la variable $P(u_2=1/u_1{}^*,y_2{}^*)$. Par conséquent, minimiser $H(H/u_2)$ revient à fixer $P(u_2=1/u_1{}^*,y_2{}^*)=1$ lorsque $H(H/u_2)$ est une fonction décroissante de $P(u_2=1/u_1{}^*,y_2{}^*)$, et $P(u_2=1/u_1{}^*,y_2{}^*)=0$ lorsque $H(H/u_2)$ est une fonction croissante de $P(u_2=1/u_1{}^*,y_2{}^*)$. Ce qui équivaut à utiliser la règle suivante :

$$P_0\,(C_{10}-C_{00})\,P(u_1{}^*,y_2{}^*/H_0) - (1-P_0)\,(C_{01}-C_{11})\,P(u_1{}^*,y_2{}^*/H_1) \underset{P(u_2=1/u_1{}^*,y_2{}^*)=1}{\overset{P(u_2=1/u_1{}^*,y_2{}^*)=0}{\gtrless}} 0 \qquad (4.10)$$

En faisant l'hypothèse raisonnable que le coût d'une mauvaise décision est supérieure au coût d'une bonne décision, c'est-à-dire que $C_{10} > C_{00}$ et $C_{01}>C_{11}$, l'équation (4.10) peut se mettre sous la forme :

$$\frac{P(u_1{}^*,y_2{}^*/H_1)}{P(u_1{}^*,y_2{}^*/H_0)} \underset{u_2=0}{\overset{u_2=1}{\gtrless}} \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})}{(C_{01}-C_{11})} \qquad (4.11)$$

sachant que u_1 et y_2 sont indépendants, on a :

$$\frac{P(u_1{}^*,y_2{}^*/H_1)}{P(u_1{}^*,y_2{}^*/H_0)} = \frac{P(u_1{}^*/H_1)}{P(u_1{}^*/H_0)}\,\frac{P(y_2{}^*/H_1)}{P(y_2{}^*/H_0)}$$

l'équation (4.11) peut s'écrire :

$$\frac{P(y_2{}^*/H_1)}{P(y_2{}^*/H_0)} \underset{u_2=0}{\overset{u_2=1}{\gtrless}} \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})}{(C_{01}-C_{11})}\,\frac{P(u_1{}^*/H_0)}{P(u_1{}^*/H_1)} \qquad (4.12)$$

L'équation (4.12) est similaire à l'équation (5.2.4) du chapitre 1, avec la différence que les fonctions de coût ne sont pas des constantes mais dépendent des probabilités *a posteriori*. De plus, il faut remarquer que le seuil de droite correspond en fait à deux seuils différents, l'un

lorsque $u_1^*=0$, et l'autre lorsque $u_1^*=1$. Ces deux seuils sont notés λ_2^1 et λ_2^0 et dépendent des probabilités de fausse alarme P_{F1} et de détection P_{D1} du détecteur q_1 tels que :

$$\lambda_2^1 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})}{(C_{01}-C_{11})}\frac{P(u_1=1/H_0)}{P(u_1=1/H_1)} = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})}{(C_{01}-C_{11})}\frac{P_{F1}}{P_{D1}} \tag{4.13}$$

$$\lambda_2^0 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})}{(C_{01}-C_{11})}\frac{P(u_1=0/H_0)}{P(u_1=0/H_1)} = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})}{(C_{01}-C_{11})}\frac{(1-P_{F1})}{(1-P_{D1})} \tag{4.14}$$

• Optimisation du premier détecteur

L'optimisation du premier détecteur peut être obtenue en dérivant $H(H/u_2)$ par rapport à P_{F1}. On pose $v=P_{F1}$ dans l'équation (4.7), et on obtient :

$$\frac{\partial H(H/u_2)}{\partial P_{F1}} = P_0\,(C_{10}-C_{00})\,\frac{\partial P_{F2}}{\partial P_{F1}} - (1-P_0)\,(C_{01}-C_{11})\frac{\partial P_{D2}}{\partial P_{F1}} \tag{4.15}$$

sachant que le seuil λ_1 au niveau du premier détecteur est donné par :

$$\lambda_1 = \frac{\partial P_{D1}}{\partial P_{F1}}$$

et que l'on a :

$$\frac{\partial P_{D2}}{\partial P_{F1}} = \frac{\partial P_{D2}}{\partial P_{D1}}\lambda_1 \tag{4.16}$$

En substituant (4.16) dans (4.15), on obtient :

$$\frac{\partial H(H/u_2)}{\partial P_{F1}} = P_0\,(C_{10}-C_{00})\,\frac{\partial P_{F2}}{\partial P_{F1}} - (1-P_0)\,(C_{01}-C_{11})\,\frac{\partial P_{D2}}{\partial P_{D1}}\lambda_1 \tag{4.17}$$

Pour minimiser $H(H/u_2)$ il faut que (4.17) soit nulle, ce qui revient à choisir :

$$\lambda_1 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})\dfrac{\partial P_{F2}}{\partial P_{F1}}}{(C_{01}-C_{11})\dfrac{\partial P_{D2}}{\partial P_{D1}}} \tag{4.18}$$

De plus :

$P_{F2} = P(u_2=1/H_0)$
$P_{F2} = P(u_2=1/u_1=0,H_0).P(u_1=0/H_0) + P(u_2=1/u_1=1,H_0).P(u_1=1/H_0)$
$P_{F2} = P_{F2}(\lambda_2^0).(1-P_{F1}) + P_{F2}(\lambda_2^1).P_{F1}$

De même :

$$P_{D2} = P_{D2}(\lambda_2^0).(1-P_{D1}) + P_{D2}(\lambda_2^1).P_{D1}$$

Où $P_{F2}(\lambda_2^j)$ et $P_{D2}(\lambda_2^j)$ (j=0,1) représentent les probabilités de fausse alarme et de détection du détecteur 2 calculées en utilisant λ_2^j comme seuil.

On en déduit, sachant que le second détecteur est supposé fixé, que :

$$\frac{\partial P_{F2}}{\partial P_{F1}} = P_{F2}(\lambda_2^1) - P_{F2}(\lambda_2^0)$$

$$\frac{\partial P_{D2}}{\partial P_{D1}} = P_{D2}(\lambda_2^1) - P_{D2}(\lambda_2^0)$$

Par conséquent, le seuil λ_1 peut s'écrire :

$$\lambda_1 = \frac{P_0}{1-P_0}\frac{(C_{10}-C_{00})\,(P_{F2}(\lambda_2^1)-P_{F2}(\lambda_2^0))}{(C_{01}-C_{11})\,(P_{D2}(\lambda_2^1)-P_{D2}(\lambda_2^0))} \qquad (4.19)$$

L'équation (4.19) est similaire à l'équation (5.2.9) du chapitre 1, avec la différence que les fonctions de coût ne sont pas des constantes mais dépendent des probabilités *a posteriori.*

Optimiser le système revient donc à résoudre 3 équations non linéaires couplées de la forme (4.13) (4.14) et (4.19). Une solution commune à ces 3 équations est une solution de l'optimisation élément par élément utilisant le critère entropique du problème de détection décentralisée série de la Figure 36.

• Exemple

Reprenons ici l'exemple développé au chapitre 1. Considérons un système formé de deux capteurs dont les observations y_1 et y_2 sont indépendantes. Les fonctions de densité de probabilité sous chaque hypothèse sont des gaussiennes Sous l'hypothèse H_0 ces fonctions de densité de probabilité sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1 ces fonctions sont supposées être de moyenne 1 et de variance 1.

La règle de décision au niveau du premier détecteur est la suivante : $\Lambda(y_1) \underset{u_1=0}{\overset{u_1=1}{\gtrless}} \lambda_1$ avec λ_1 qui vérifie (4.19)

La règle de décision au niveau du deuxième détecteur est :

$$\Lambda(y_2) \underset{u_2=0}{\overset{u_2=1}{\gtrless}} \lambda_2^0 \text{ si } u_1=0 \text{ avec } \lambda_2^0 \text{ qui vérifie (4.14)}$$

$$\Lambda(y_2) \underset{u_2=0}{\overset{u_2=1}{\gtrless}} \lambda_2^1 \text{ si } u_1=1 \text{ avec } \lambda_2^1 \text{ qui vérifie (4.13)}$$

Une expression explicite des seuils λ_1, λ_2^0 et λ_2^1 peut être obtenue en résolvant les équations (4.13) (4.14) et (4.16).

Les seuils optimaux au niveau des deux détecteurs sont indiqués dans la figure 37. Sur la figure 38, on a représenté la variation de $H(H/u_2)$ en fonction de P_0. Sur la figure 39, la variation de la probabilité d'erreur en fonction de P_0 est représentée. La figure 40 représente la courbe C.O.R. associé au système de détection. Pour P_0=0.5, les résultats trouvés en terme de probabilité d'erreur sont les mêmes que ceux trouvés dans le cas

Bayésien du paragraphe 5.2. du chapitre 1. Par contre, lorsque P_0 varie, le critère entropique ne permet pas de minimiser la probabilité d'erreur, en revanche les probabilités de détection et de fausse alarme trouvées restent dans des limites acceptables, ce qui n'est pas le cas en appliquant le critère de Bayes.

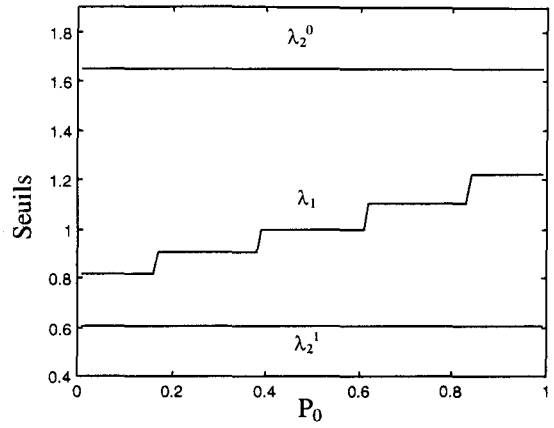


Figure 37 : Seuils en fonction de P_0, $m_1=m_2=1$.

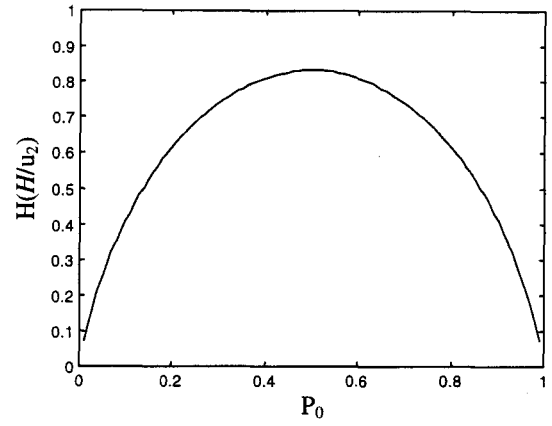


Figure 38 : $H(H/u_2)$ en fonction de P_0, $m_1=m_2=1$

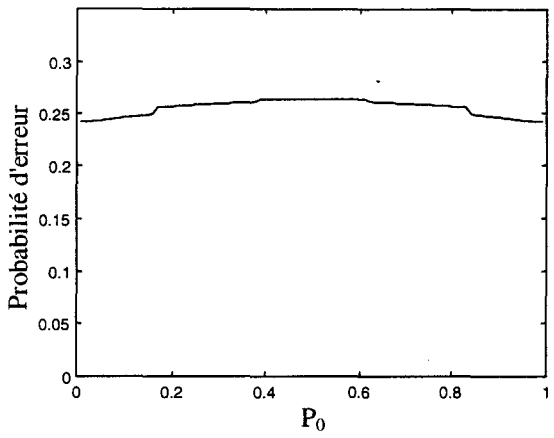


Figure 39 : Probabilité d'erreur en fonction de P_0, $m_1=m_2=1$.

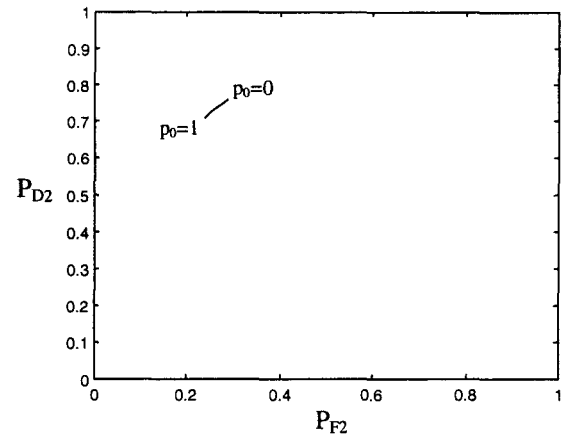


Figure 40 : Courbe C.O.R. $P_D=f(P_F)$, $m_1=m_2=1$.

5. Conclusion

L'utilisation de l'entropie comme critère d'optimisation des systèmes de détection peut être envisagée dans le cas de la détection centralisée, décentralisée parallèle, et série.

Dans le cas de la détection centralisée, les probabilités d'erreur pour $P_0=0,5$ en utilisant le critère de Bayes et le critère entropique sont identiques. Par contre, elle est supérieure dans les autres cas, ce qui est tout à fait normal car le critère de Bayes est basé sur la minimisation de l'erreur ; on ne pourra par conséquent pas obtenir de meilleurs résultats.
En revanche, lorsqu'on utilise le critère de Bayes dans les cas où P_0 est proche de 0 ou de 1, les points de la courbe C.O.R. s'éloignent notablement du point (0,1), ce qui n'est pas le cas lorsqu'on utilise le critère entropique. Cela signifie que lorsqu'un événement n'arrive que très rarement, avec le critère de Bayes, on aura tendance à ne pas le considérer ; tandis qu'avec le

critère entropique, cet événement est pris en compte de façon non négligeable (au détriment bien entendu de l'erreur). Cette propriété caractéristique de l'entropie est dûe au fait que : $H(H/u_0) = -\sum_{i,j} p_{ij}.\log p_{j/i}$

Nous pouvons enfin remarquer que dans l'exemple traité, la courbe C.O.R. en utilisant le critère entropique « colle » tout à fait à la courbe C.O.R. en utilisant le critère de Bayes.

Dans le cas de la détection décentralisée parallèle ou série, dans les deux exemples traités, pour $P_0 \# 0.5$, la probabilité d'erreur est identique à celle trouvée en utilisant le critère de Bayes. Lorsque P_0 varie, les probabilités de fausse alarme et de détection restent respectivement très proche de 0 et de 1.

2ième partie :

APPRENTISSAGE

CHAPITRE 3

DES DONNEES D'APPRENTISSAGE AUX ARBRES DE DECISION

Dans un premier temps, nous présentons le cadre général de la classification automatique afin d'introduire l'apprentissage supervisé (à base d'exemples) qui nous intéresse tout particulièrement.

Dans ce cadre, nous présentons la structure des données recueillies sur un système plus ou moins complexe. Ces données proviennent en général de capteurs qui, par définition, quantifient un signal (continu ou pas). Nous pourrons dès lors introduire la notion de finesse des variables associées aux capteurs correspondant.
Nous introduisons également la notion d'incohérence des données d'apprentissage provenant d'un bruit se superposant aux données, ou provenant d'un manque d'explication dû au fait que l'utilisateur n'aurait pas connaissance de certaines variables pertinentes pour l'étude du système.

Enfin, pour terminer ce chapitre, nous introduisons les méthodes de construction d'arbres de décision qui permettent de traiter des variables qualitatives, mais également numériques. Ces méthodes visent à l'optimisation d'un critère global afin de discriminer les différentes classes présentes. Dans ce cadre, nous présentons l'algorithme C4.5, qui est certainement l'algorithme de construction d'arbres de décision le plus usité à l'heure actuelle. Nous discuterons alors de ses points forts, comme de ses points faibles pour lesquels nous proposons des alternatives.

1. Introduction aux méthodes de classification automatique

Les méthodes de classification...

Les méthodes de classification utilisées sont très nombreuses et sont issues de domaines scientifiques variés. Nous distinguons trois groupes de méthodes :

- Les méthodes statistiques (ou paramétriques) nécessitent la connaissance d'un modèle de distribution des probabilités de chaque classe. A titre d'exemple, citons la méthode des Noyaux de Rosenblatt-Parzen (méthode de détection des zones à forte densité dans l'espace des observations) [Par60] [Pos87]. Les méthodes de détection décrites au chapitre 1 en font partie.
- Les méthodes métriques telles que l'algorithme de « regroupement autour des centres mobiles » [For65], ou l'algorithme des « nuées dynamiques » [Did72].
- Les méthodes issues de l'intelligence artificielle sont des méthodes non paramétriques. On distingue les méthodes symboliques [Ler70] [JaL78] [CDG89] (la procédure de classification produite peut être écrite sous forme de règles) des méthodes non symboliques ou adaptatives (la procédure de classification produite est de type « boîte noire »). A titre d'exemple, les méthodes basées sur la construction d'arbres de décision sont issues des méthodes symboliques. Pour les méthodes adaptatives, on distingue deux grandes classes : les réseaux de neurones [GaG96] [Hin92] et les algorithmes génétiques [Qui88].

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets (ou individus) à partir de certains traits descriptifs. Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier au problème de la prise de décision automatisée. Il s'agira, par exemple, d'établir un diagnostic médical à partir de la description

clinique d'un patient, de donner une réponse à la demande de prêt bancaire de la part d'un client sur la base de sa situation personnelle, de déclencher un processus d'alerte en fonction de signaux reçus par des capteurs.

Une première approche possible pour résoudre ce type de problème est l'approche « systèmes experts ». Dans ce cadre, la connaissance d'un expert (ou d'un groupe d'experts) est décrite sous forme de règles. Cet ensemble de règles forme un système expert qui est utilisé pour classifier de nouveaux cas. Cette approche, largement utilisée dans les années 80, dépend fortement de la capacité à extraire et à formaliser les connaissances de l'expert.

Nous considérons ici une autre approche pour laquelle l'appartenance à une classe sera déterminée au moyen d'algorithmes formalisés, et non pas par des méthodes subjectives ou visuelles faisant appel à l'initiative d'un expert (d'où le terme « automatique »). La procédure de classification sera extraite automatiquement à partir d'un ensemble d'exemples. Un exemple consiste en la description d'un cas avec la classification correspondante.
Nous nous plaçons dans un cadre d'apprentissage supervisé car les exemples fournis sont supposés correctement classifiés (observés ou fournis par un expert). L'ensemble des paramètres observés est ainsi constitué, d'une part, de l'ensemble des paramètres descriptifs (noté X), et d'autre part, d'une variable de classification (notée Y). Les valeurs (ou modalités) prises par Y constituent les classes qu'il nous faudra dès lors discriminer (dans un cadre non supervisé, le système devra déterminer ses propres classes).

La classification supervisée...

Introduisons deux exemples afin de présenter la classification supervisée ; le premier dans le domaine médical, le second dans le domaine bancaire :

Etablir un diagnostic dans le domaine médical signifie être capable d'associer le nom d'une maladie à un certain nombre de symptômes présentés par les malades.
Si on analyse l'exemple précédent, on repère trois objets essentiels : les malades, les maladies et les symptômes. Les malades représentent la population de travail, les symptômes représentent les descriptions qui permettent d'appréhender la population tandis que les maladies représentent l'ensemble qui doit permettre de classer la population. On suppose qu'il existe un classement correct (c'est-à-dire une application de l'ensemble des malades vers l'ensemble des maladies). Apprendre à établir un diagnostic, c'est associer une maladie à une liste de symptômes de telle manière que cette association corresponde au classement défini ci-dessus.
On peut remarquer que l'on différencie la population d'apprentissage de l'ensemble des descriptions, ce qui correspond à la situation réelle. On décrit par exemple un patient par une liste de symptômes. Par conséquent, il est possible que deux individus appartenant à des classes différentes aient les mêmes descriptions (pour deux patients présentant les mêmes symptômes, l'un est malade, l'autre pas). Ce problème sera traité dans le paragraphe concernant l'incohérence des données d'apprentissage.

Dans le domaine bancaire, on peut imaginer que l'on dispose d'un historique des prêts accordés avec, pour chaque prêt, la situation personnelle du demandeur et le résultat du prêt (problèmes de recouvrement ou non). Un système d'apprentissage doit alors, à partir de cet ensemble d'exemples, extraire une procédure de classification qui, au vu de la situation personnelle d'un client, devra décider de l'attribution du prêt (nous avons ici deux conclusions

possibles). Il s'agit donc d'induire une procédure de classification générale à partir d'exemples. La procédure générée devra classifier correctement les exemples de l'échantillon mais surtout avoir un bon pouvoir prédictif pour classifier correctement de nouvelles descriptions.

Le principe de la classification supervisée...

Soit Ω la population d'apprentissage, M_x est l'ensemble des descriptions, et l'ensemble des classes est $M_Y = \{1,\ldots,c_j,\ldots,c_m\}$.
- $X : \Omega \rightarrow M_x$ est une fonction qui associe une description à tout élément de la population d'apprentissage.
- $Y : \Omega \rightarrow \{1,\ldots,c_j,\ldots,c_m\}$ est la fonction de classification qui associe une classe à tout élément de la population d'apprentissage. Ces classes peuvent être, par exemple, des modes de fonctionnement (en diagnostic : panne ou fonctionnement normal), ou des fonctions d'appartenance (en Reconnaissance des Formes).

Le but de la classification est alors de trouver une fonction de décision $D : M_x \rightarrow C$ telle que $DoX = Y$ ou, de manière plus réaliste, telle que DoX soit une bonne approximation de Y. Une telle fonction sera appelée fonction de classement, ou procédure de classification.

Dans la pratique, on dispose souvent d'un ensemble de variables $X_1, X_2, \ldots, X_N$ symboliques ou numériques dont la concaténation constitue la fonction de description X. L'espace des descriptions M_x est alors égal au produit cartésien $M_{X1} \times M_{X2} \times \ldots \times M_{XN}$.

2. Structure des données

2.1. Le tableau initial des données

On souhaite apprendre à associer à une observation, constituée des valeurs prises par les variables de description du système, la classe qui lui correspond. Les seules informations dont on dispose sont un ensemble d'exemples sur une population de taille L. Ces exemples associent à chaque observation la classe qui lui correspond. On note :
- $Y : \Omega \rightarrow \{1,\ldots,c_j,\ldots,c_m\}$ la fonction de classification qui associe une classe à tout élément de la population d'apprentissage Ω. Certains auteurs l'appellent variable endogène, ou variable difficilement observable (par exemple, d'acquisition coûteuse, ou souvent bruitée, ou non accessible à la mesure, ...).
- $X = (X_1, X_2, \ldots, X_N)$ représente le vecteur des N variables de description du système. Elles sont également appelées variables explicatives, exogènes, ou facilement observables (elles peuvent être d'acquisition non coûteuse, fiables, ...).

Nous pouvons dès lors représenter le tableau des données (voir figure 1) :

Ω \ Σ	X				Y
	$[X_1$	X_2	X_i	..... $X_N]$	
	.....				
ω_1			.		
ω_2			.		
...			.		
ω_j	.....		$X_i(\omega_j)$	.....	$Y(\omega_j)$
...			.		
ω_L			.		

Figure 1 : Le tableau initial des données

où :

- $\Omega = \{\omega_1, \omega_2, \ldots, \omega_L\}$ est la population d'apprentissage.
- ω_j est la $j^{ème}$ observation.
- $\Sigma = (X,Y) = (X_1, X_2, \ldots, X_N, Y)$ est le vecteur des variables de description et de classification du système. Ces variables sont définies en général par l'expert du système en question, et sont ainsi considérées comme pertinentes pour son étude. Ce problème sera soulevé dans le paragraphe traitant de l'incohérence des données d'apprentissage.
- $X_i(\omega_j)$ représente la modalité de la variable X_i pour l'observation ω_j (j=1, ... , L).

Dans ce tableau, $X_i(\omega_j)$ (resp. $Y(\omega_j)$) désigne la modalité (ou valeur) prise par la variable X_i (resp. Y) pour le $j^{ième}$ échantillon. Chaque variable de description X_i peut donc être considérée comme une application de l'ensemble d'apprentissage Ω vers son ensemble de modalités M_{Xi}.

$$X_i : \Omega \longrightarrow M_{Xi} = \{\alpha_k^i, k = 1, \ldots, m_i\} \forall i \in \{1..N\} \qquad \text{avec } m_i = \text{card}(M_{Xi})$$
$$\omega \mapsto X_i(\omega)$$

M_{Xi} peut être caractérisé par plusieurs structures, suivant la nature de la variable en question.

Celle-ci peut être :

- quantitative (ou numérique : $M_{Xi} \subset \Re$) : par exemple la température, la taille, ... sont des variables quantitatives. Il existe une évolution continue des valeurs prises par ces variables.
- qualitative ordinale (ou ordonnée) : la variable "taille" composée des modalités (petit, moyen, grand), le "poids" composé des modalités (léger, lourd), ... sont des variables qualitatives ordinales.
- (qualitative) nominale : aucune relation n'est définie sur M_{Xi}. Par exemple, la couleur des yeux, la couleur des cheveux, les variables booléennes, ... sont des variables nominales (sous-entendues « qualitatives »).
- structurée : il existe une ou plusieurs relations plus ou moins complexes sur M_{Xi}.

En général, Y possède (ou est considérée comme ayant) une structure nominale avec m=card(M_Y) petit.

2.2. Finesse des variables de descriptions

Etant donnée une variable de description X_i, il est toujours possible d'en créer d'autres, en partitionnant l'ensemble des modalités M_{Xi}. Soit $M_{\tilde{X}_i}$ une partition de M_{Xi}. On notera $\tilde{X}_i$ la variable déduite de X_i, dont l'ensemble des modalités est $M_{\tilde{X}_i}$. L'ensemble $P(M_{Xi})$ des partitions de M_{Xi}, ordonné par la relation de finesse ≤, constitue un treillis* [Sza62], [Bir67], [BaM70], [KaB78] dont M_{Xi} est le plus petit élément (minorant universel). On dira que la variable $\tilde{X}_i$ est plus grosse que X_i ($\tilde{X}_i \geq X_i$), puisque chacune des classes de M_{Xi} est entièrement incluse (par définition) dans une classe de $M_{\tilde{X}i}$. On peut noter que, pour tout élément de la population d'apprentissage, la connaissance de la modalité prise par X_i implique la connaissance de celle prise par $\tilde{X}_i$.

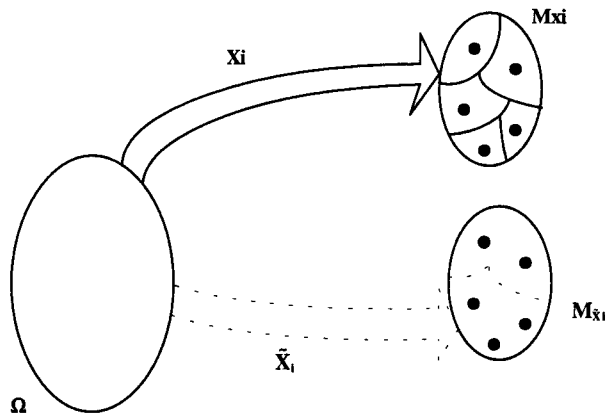


Figure 2 : Finesse d'une variable $Xi \leq \tilde{X}i$

A titre d'exemple, à partir de la variable de description « âge » (constituée de 4 modalités) de la Figure 3, on peut créer les 6 variables plus grosses données ci-dessous :

âge1 : {bébé ou enfant, adolescent, adulte}
âge2 : {bébé ou enfant ou adolescent, adulte}={jeune, adulte}
âge3 : {bébé, enfant ou adolescent, adulte}
âge4 : {bébé, enfant, adolescent ou adulte}
âge5 : {bébé, enfant ou adolescent ou adulte}
âge6 : {bébé ou enfant ou adolescent ou adulte}

* Une partition p d'un ensemble E est plus fine qu'une partition p' de E si et seulement si chaque classe de p est incluse dans une classe de p'. On note p≤p'.
L'ordre (P(E),≤) est muni d'une structure de treillis : ∀p,p'∈P(E) ∃p∧p' (infimum) et ∃p∨p' (supremum).
Le treillis peut alors être représenté par un diagramme de Hasse.

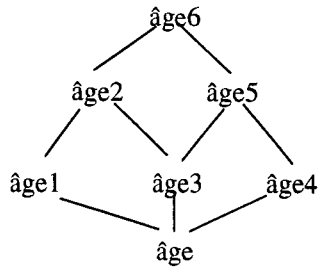


Figure 3 : Le treillis des partitions engendrées à partir des modalités de la variable « âge »

Remarque :

L'ensemble de ces variables constitue « une variable structurée ».

2.3. Partitions de la population d'apprentissage

• Définition

A chaque variable de description X_i de notre système (X_i est une composante du vecteur X), nous pouvons associer une partition de la population d'apprentissage Ω de la façon suivante :

$P_{Xi}(\Omega) = \{X_i^{-1}(u), u \in M_{Xi}\}$ où $X_i^{-1}(u) = \{\omega / X_i(\omega) = u\}$.

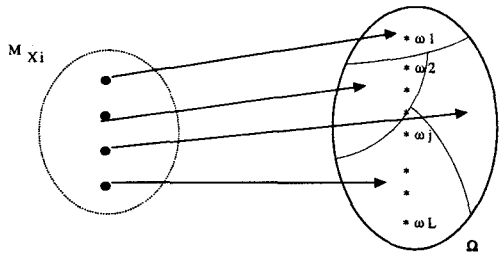


Figure 4 : Partition de la population d'apprentissage Ω engendrée par X_i (i=1,...,N).

L'ensemble $P(\Omega)$ des partitions de Ω, muni de la relation de finesse $\leq$, possède une structure de treillis. Ainsi, il est possible de construire une relation d'ordre et une relation d'équivalence entre les variables.

Considérons X_i et X_j ($i \neq j$) deux variables de description de notre système. Nous définissons :

$X_i \equiv X_j \Leftrightarrow P_{Xi}(\Omega) = P_{Xj}(\Omega)$

On dit que X_i et X_j sont des variables équivalentes.

A titre d'exemple, si on considère le tableau suivant :

Ω	*forme*	*température*	*précision*	*S*
ω_1	hexagonale	200°C	1.5×10^{-3}	0
ω_2	carrée	500°C	10^{-4}	1
ω_3	hexagonale	200°C	1.5×10^{-3}	0
ω_4	hexagonale	200°C	1.5×10^{-3}	0
ω_5	carrée	500°C	10^{-4}	1

Figure 5 : Un exemple de tableau de données.

toutes les variables en question apportent la même quantité d'information sur le système ; il suffit pour s'en convaincre de recoder les modalités (ex. : hexagonale → 200°C ; carrée → 500°C ; ...).

- **Propriété**

$X_i \leq X_j \Leftrightarrow P_{Xi}(\Omega) \leq P_{Xj}(\Omega)$

Autrement dit, si X_j est une variable plus grosse que X_i, la partition de Ω associée à (induite par) X_j est également plus grosse que celle induite par X_i.

Par contre, en considérant l'exemple précédent, la variable *vitesse* ci-dessous apporte, elle, plus d'information que les autres sur le système étudié :

Ω	**vitesse**
ω_1	15 rad/s
ω_2	20 rad/s
ω_3	15 rad/s
ω_4	10 rad/s
ω_5	20 rad/s

Figure 6 : La variable *vitesse* apporte plus d'information que les autres variables.

2.4. Variables vectorielles (ou multidimensionnelles)

$X = (X_1, X_2, \ldots, X_N)$ étant le vecteur des N variables de description du système, nous représentons par $P^v(X)$ l'ensemble de tous les vecteurs possibles dont les composantes sont des composantes élémentaires de X. Chaque élément S de $P^v(X)$ représente une application de Ω dans son ensemble de modalités M_S :

$$S : \Omega \longrightarrow M_S \qquad \text{où } S \in P^v(X)$$
$$\omega \mapsto S(\omega)$$

La variable multidimensionnelle S induit une partition de Ω, qui est l'intersection des partitions de Ω induites par chaque variable de description composant le vecteur S :

$$P_S(\Omega) = \bigcap_{Xi \in S} P_{Xi}(\Omega)$$

Nous pouvons alors en déduire la propriété de finesse suivante :

• **Propriété**

Soient S_1 et S_2 appartenant à $P^v(X)$, telles que $S_1 \subset S_2$.

On a alors les relations suivantes :

$M_{S1} \geq M_{S2}$ et $P_{S1}(\Omega) \geq P_{S2}(\Omega)$.

En d'autres termes, la relation d'inclusion entre sous-ensembles de variables induit une relation de finesse entre les variables vectorielles et les partitions associées de la population d'apprentissage.

A titre d'exemple, dans l'exemple suivant (extrait de [Min89], p.320, et représentant les modalités de deux attributs à partir de données sur le cancer du sein) :

Ω	*radiation*	*menopause*	*class*
ω_1	no	<60	recur
ω_2	no	≥60	recur
ω_3	no	<60	recur
ω_4	no	not	recur
ω_5	yes	≥60	not recur
ω_6	yes	<60	not recur
ω_7	yes	≥60	not recur
ω_8	no	not	not recur
ω_9	no	<60	not recur
ω_{10}	no	<60	recur

Figure 7 : Exemple de tableau de données.

on pourra remarquer que le vecteur de variables *(radiation,menopause)* induit une partition de Ω plus fine que la variable *radiation* seule :

$$\{(\omega_1, \omega_3, \omega_9, \omega_{10}), (\omega_2), (\omega_4, \omega_8), (\omega_5, \omega_7), (\omega_6)\} \leq \{(\omega_1, \omega_2, \omega_3, \omega_4, \omega_8, \omega_9, \omega_{10}), (\omega_5, \omega_6, \omega_7)\}$$

2.5. L'incohérence des données d'apprentissage

Parmi toutes les variables définies par l'expert du système en question, certaines d'entre elles pourront ne pas apporter d'information sur l'appartenance d'un individu à une classe donnée. Ce problème sera traité et résolu en utilisant des indices issus de la théorie de l'information.

A l'inverse, lors d'une étude d'un système plus ou moins complexe, l'expert pourra trouver qu'à deux individus ayant les mêmes valeurs des descripteurs correspondent deux classes différentes : $\exists \omega_i, \omega_j \ (i \neq j) \ / \ X(\omega_i)=X(\omega_j)$ et $Y(\omega_i) \neq Y(\omega_j)$

Nous dirons dans ce cas que la population d'apprentissage considérée est incohérente.

Nous distinguons alors deux cas :

(1) *l'absence d'explication à tort* : le modèle déterministe Y=f(X) est représentatif du système, mais les données relevées sont biaisées. Ceci peut être dû à la présence d'un bruit important sur les mesures effectuées ou/et au mauvais fonctionnement d'un sous-système.

(2) *l'absence d'explication à raison* : certaines variables pertinentes pour l'étude de notre système ont été omises. La solution consistera alors à reprendre l'étude du système en considérant ces autres variables apportant une quantité d'information non négligeable sur la procédure de classification.

2.6. Le tableau de contingence

On note :
- $M_X = \{\alpha_1,\alpha_2,...,\alpha_n\}$, l'ensemble des modalités de la variable vectorielle de descriptions X. Ces modalités sont les conjonctions des modalités des différentes variables X_i qui composent X.
- $M_Y = \{c_1,c_2,...,c_m\}$ est l'ensemble des classes à discriminer.

Afin d'obtenir les probabilités conjointes p_{ij} d'obtention des modalités α_i et c_j (i=1,...,n ; j=1,...,m), nous partons du principe que ces probabilités peuvent être estimées par les fréquences relatives. En effet, lorsque le nombre d'exemples contenus dans le tableau initial de données n'est pas important, la fréquence d'un événement peut notablement changer d'un apprentissage à l'autre. Cependant, lorsque le nombre d'expériences augmente, la fréquence de l'événement perd son caractère aléatoire ; les conditions aléatoires intervenant dans chaque expérience isolée se trouvent mutuellement compensées et la fréquence tend à se stabiliser, s'approchant d'une certaine grandeur constante. On dit que lorsque le nombre d'expériences augmente, la fréquence d'un événement converge en probabilité. Cette propriété de convergence en probabilité est l'objet du théorème de Bernouilli [Vent73]. Partant de ce principe, nous pouvons construire, pour chaque variable de description X_i, et plus généralement pour chaque variable vectorielle $S \in P^v(X)$, un tableau de contingence. Ce tableau de contingence est basé sur la connaissance du tableau initial de données construit sur une population de taille L suffisamment grande.

$M_S \setminus M_Y$	c_1 c_2	c_j	..... c_m	
α_1		.		
α_2		.		
...		.		
α_i	.....	p_{ij}	.....	$p_{i.}$
...		.		
α_{n_s}		.		
		$p_{.j}$		

Figure 8 : Tableau de contingence P(S,Y).

où :
- n_s = card M_S ; m = card M_Y
- p_{ij} est la probabilité d'occurrence conjointe des modalités α_i et c_j.
- $p_{i.} = \sum_j p_{ij}$; $p_{.j} = \sum_i p_{ij}$

On peut également définir un tableau à partir des probabilités conditionnelles $p_{j/i}$, en divisant chaque terme de la $i^{ème}$ ligne du tableau de contingence par la quantité $p_{i.}$ comme le montre la figure 6.

$M_S \setminus M_Y$	c_1 c_2	c_j	..... c_m
α_1		.	
α_2		.	
...		.	
α_i	.....	$p_{j/i}$	.....
...		.	
α_{n_s}		.	

Figure 9 : Tableau de probabilité conditionnelles P(Y/S).

où $p_{j/i} = p_{ij} / p_{i.}$ est la probabilité d'obtenir $Y = c_j$ sachant que $S = \alpha_i$.

3. Les méthodes d'induction par arbre de décision

3.1. Position du problème

Afin d'obtenir des procédures de classification, il nous faut définir des arbres de décision, des règles de décision, ou encore des formalismes logiques plus ou moins complexes. Nous nous limiterons ici aux arbres de décision.

Les systèmes d'induction par arbre de décision visent à l'optimisation d'un critère global afin de *spécialiser les hypothèses* [Ren86]. La démarche est la suivante :

On cherche la variable apportant le plus d'information sur les classes à expliquer, puis une deuxième prise parmi les variables restantes, puis une troisième, ...
L'introduction de variables supplémentaires permet de disposer de partitions de plus en plus fines de la population d'apprentissage et donc d'approcher de mieux en mieux la partition associée à la variable à expliquer. Cette introduction trouve cependant ses limites dans les deux points suivants [Ayg86] :

- l'augmentation de la complexité du modèle d'explication
- la diminution de l'effectif des classes de la partition de la population d'apprentissage.

3.2. Présentation générale des arbres de décision

Formellement, un arbre de décision est un arbre [Pic72] tel que :

- Une feuille (ou nœud réponse) contient un nom de classe.
- Un nœud (qui n'est pas une feuille, ou nœud décision) contient un test sur un attribut (variable de description) avec une branche (donnant naissance à un autre arbre de décision) pour chaque valeur possible de l'attribut en question.

Le principe de construction de l'arbre est de diviser récursivement les exemples de l'ensemble d'apprentissage par des tests définis sur les attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant tous à une même classe.

La démarche est la suivante : On recherche la variable de description apportant le plus d'information (suivant un critère donné) sur les classes à expliquer, puis une deuxième prise parmi les variables restantes, puis une troisième,... L'introduction de variables supplémentaires permet de disposer de partitions de plus en plus fines de la population d'apprentissage et donc de répartir de mieux en mieux les exemples dans les différentes classes.

Dans toutes les méthodes, on trouve les trois opérateurs suivants :

1. Décider si un noeud est terminal. Par exemple : tous les exemples sont dans la même classe, il y a moins d'un certain nombre d'erreurs, ...
2. Sélectionner un test à associer à un noeud. Par exemple : aléatoirement, utiliser des critères statistiques, ...
3. Affecter une classe à une feuille. On attribue la classe majoritaire sauf dans le cas où on utilise des fonctions « coût » ou « risque ».

On peut alors trouver un arbre de décision dont l'erreur apparente (l'erreur calculée sur la population d'apprentissage) est faible, voire nulle dans l'hypothèse d'une population d'apprentissage cohérente. Dans ce cas, un arbre de décision parfait est un arbre de décision tel que tous les exemples de l'ensemble d'apprentissage soient correctement classifiés. Un tel arbre n'existe pas toujours (s'il existe deux exemples tels que à deux descriptions identiques correspondent deux classes différentes). Le meilleur arbre de décision est le plus petit arbre de décision parfait, une mesure de complexité étant choisie.

Les méthodes procèdent toujours en deux phases. Dans une première phase, on calcule récursivement, après avoir définis les trois opérateurs ci-dessus, un « bon » arbre de décision (erreur apparente faible). Dans une seconde phase, on élague l'arbre obtenu pour essayer de faire diminuer l'erreur réelle. On procède en deux phases car il n'existe aucune heuristique satisfaisante permettant d'arrêter au « bon » moment la croissance de l'arbre de décision. En effet, il se peut que l'on choisisse un test qui fasse augmenter temporairement l'erreur mais qui permettra, par la suite, de faire diminuer nettement cette erreur. De plus, le risque d'arrêter trop tôt la croissance de l'arbre est plus important que de l'arrêter trop tard.

Soit X une variable-test. Nous pouvons construire un *tableau de comptage* (ou un tableau de contingence) croisant les modalités de X et les différentes classes à expliquer (*i.e.* les modalités observées de Y). Notons k_{ij} le nombre d'occurrences de la conjonction « $[X=\alpha_i]et[Y=\beta_j]$ ».

Plusieurs critères peuvent être utilisés pour définir, à chaque étape, la variable supplémentaire à prendre en compte [Min89]. Parmi les critères les plus usités, en voici quelques uns :

i) La mesure du Khi^2 :

La mesure du Khi^2 est connue depuis longtemps, mais elle a été utilisée dans des algorithmes d'induction pour la première fois par [Har84].

$$Khi^2 = \sum_i \sum_j \frac{(k_{ij} - E_{ij})^2}{E_{ij}} \quad \text{où} \quad E_{ij} = \frac{k_{i.} . k_{.j}}{N} \quad \text{et} \quad k_{i.} = \sum_j k_{ij} \; ; \; k_{.j} = \sum_i k_{ij}$$

La variable « X » apportant le plus d'information sera alors celle qui maximisera la mesure du Khi^2.

ii) L'indice GINI (ou mesure de l'impureté) :

Cet indice a été proposé par [BFO84].

$$i = \frac{1}{N}\left(\sum_i \sum_j \frac{k_{ij}^2}{k_{i.}} - \sum_j \frac{k_{.j}^2}{N}\right)$$

La variable « X » apportant le plus d'information sera alors celle qui maximisera cet indice.

iii) La mesure d'information de J.R. Quinlan :

Cette mesure est issue de la Théorie de l'Information. Elle est utilisée dans l'algorithme C4.5, et dans de nombreuses variantes de celui-ci. Il est souvent considéré (et à juste titre) comme l'algorithme type de l'apprentissage numérique. Nous le présentons donc très sommairement.

3.3. Présentation sommaire de C4.5

Le système C4.5 a été développé par J.R. Quinlan [Qui83,86,87]. C'est un système *d'inférence inductive* à partir d'*exemples*. Et plus particulièrement, c'est une méthode d'*apprentissage* par *arbre de décision* (cf [HMS66]). Il fait partie de la famille des systèmes d'apprentissage TDIDT (*Top-Down Induction of Decision Trees*), c'est-à-dire qu'il commencera à construire l'arbre par le haut (par la racine). Cette démarche possède de multiples facettes comme par exemples, le traitement des données manquantes, l'incohérence des données d'apprentissage, les variables quantitatives (numériques) ... Notre but est simplement de décrire le fonctionnement de celui-ci sans entrer dans les détails.

Description de la méthode :

Pour simplifier le problème, on ne considère que deux classes notées P et N. P et N représentent respectivement l'ensemble des *instances positives* (p = card P) et l'ensemble des *instances négatives* (n = card N).

L'information moyenne apportée par P et N est :

$$I(p,n) = -\frac{p}{p+n}\mathrm{Log}\frac{p}{p+n} - \frac{n}{p+n}\mathrm{Log}\frac{n}{p+n}$$

Cette quantité représente en fait l'entropie d'une variable à expliquer Y dont l'ensemble des modalités est séparé en deux classes notées N et P.

L'information moyenne apportée par l'arbre ayant X (un attribut ayant les modalités $\{\alpha_1,\alpha_2,\ldots,\alpha_r\}$) comme racine est :

$$M(X) = \sum_{i=1}^{r} \frac{p_i + n_i}{p+n} . I(p_i, n_i)$$

(cette quantité représente en fait l'entropie conditionnelle H(Y/X) présentée dans le chapitre suivant).

J.R. Quinlan définit alors le gain d'information qu'apporte l'attribut X par :

$$\mathrm{gain}(X) = I(p,n) - M(X).$$

(qui représente la transinformation externe H(Y)-H(Y/X)).

Le système examine alors tous les candidats et choisit la variable X qui maximise le gain(X), forme l'arbre à partir de celui-ci, et utilise le même processus récursivement (pour former l'arbre de décision).

Remarque :

Comme le dit J.R. Quinlan [Qui86], p.90 (dans une note), maximiser le gain(X) revient à minimiser M(X), car I(p,n) est constant, quel que soit l'attribut X.

Description de l'algorithme :

; *Début*
1. Sélectionner un noeud impur n
2. Pour chaque attribut X, calculer la valeur de gain(X)
3. Pour l'attribut X maximisant le gain : étendre l'arbre à partir du noeud n (à chaque modalité de X correspond une branche)
4. Retourner en *1* si il existe encore des noeuds impurs et que toutes les variables non encore utilisées ont un pouvoir discriminant.

; *Fin.*

où un noeud *impur* est un noeud ne reconnaissant pas une seule et unique classe.

3.4. Les différents points de vue

C4.5 introduit les variables-test une à une avec une approche de construction descendante de l'arbre. Cependant, V.M. Toro [Tor82] montre que :

$$\text{gain}(X_1,X_2,...,X_p) \neq \sum_{i=1}^{N} \text{gain}(X_i)$$

Dans ce sens, il peut arriver que le pouvoir explicatif de chaque variable d'un vecteur S soit faible (elles ne seront alors pas retenues comme variables discriminantes) alors que le pouvoir explicatif du vecteur S considéré globalement est important. La démarche définie par C4.5 sera alors incapable de mettre en évidence la relation entre la variable à expliquer et le vecteur de variables explicatives S.

A titre d'exemple, considérons le tableau suivant :

X_1	X_2	X_3	Y
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

Figure 10 : Un exemple

Nous cherchons à expliquer Y.
$H(Y/X_1) = H(Y/X_2) = 1$
$H(Y/X_3) = 0.69$
Notre choix va se porter sur X_3, alors que manifestement, Y est le bit de parité de X_1 et X_2.
La solution serait de chercher le sous-ensemble de variables expliquant au mieux Y :
Nous avons $H(Y/(X_1,X_2,X_3)) = 0$, ce qui signifie qu'il existe une relation entre Y et un sous-ensemble (au sens large) de $\{X_1,X_2,X_3\}$.
Ainsi, $H(Y/(X_1,X_2)) = 0$
alors que $H(Y/(X_1,X_3)) = H(Y/(X_2,X_3)) = 0.5$
Notre choix devrait se porter sur le couple (X_1,X_2).

Dans ce sens, on cherchera plutôt à écarter d'emblée les variables n'apportant aucune information, et ne garder que les variables véritablement informatives [PeP97], ce qui revient

à commencer par le bas de l'arbre. On trouve ainsi la dernière variable à tester (la moins informative) que l'on place dès lors en bas de l'arbre. Puis, on recommence l'opération sur l'ensemble des variables restant. La variable trouvée est alors l'avant dernière variable à tester que l'on place sur l'avant dernier niveau de l'arbre, et ainsi de suite... jusqu'à trouver la première variable à tester, qui représente la racine de l'arbre.

L'arbre étant construit, il est important de remarquer qu'il peut arriver, lors d'un parcours de celui-ci (lors d'une phase de décision), de s'arrêter à un niveau pour lequel toutes les variables n'ont pas été testées, simplement parce qu'à ce niveau :

- la population d'apprentissage est complètement discriminée, c'est-à-dire qu'on peut proposer une décision unique.
- ou/et l'introduction de variable-test supplémentaire n'apporterait aucune information supplémentaire sur les conclusions. Dans ce cas, la population d'apprentissage est dite incohérente.

3.5. Conclusion sur l'utilisation des arbres de décision

Un grand avantage des méthodes d'apprentissage par arbres de décision consiste en la considération de paramètres nominaux, ordonnés, mais aussi numériques. En effet, lorsque des variables numériques sont considérées, la recherche d'un ou de plusieurs seuils de discrimination sera effectuée, au lieu de tester chaque modalité comme pour les variables nominales.

L'information pertinente conduisant le plus rapidement possible à une feuille (une conclusion) est traitée d'emblée.

Enfin, l'arbre est directement compréhensible par une personne non experte du domaine considéré. Il apporte de surcroît une explication (via le parcours de l'arbre) à la conclusion trouvée, contrairement, par exemple, aux réseaux neuronaux.

4. Conclusion

Après avoir présenté la structure des données d'apprentissage permettant de définir l'incohérence des données ainsi que la notion de finesse, nous avons introduit les arbres de décision.
Ces méthodes visent à l'optimisation d'un critère global afin de discriminer les différentes classes en présence ; et permettent de traiter des variables qualitatives, mais également numériques.

Le cadre de l'apprentissage étant posé, nous définissons les outils de la théorie de l'information, qui nous permettrons dès lors de déterminer des critères informationnels utilisables dans les algorithmes à base d'arbres de décision.

CHAPITRE 4

LES OUTILS DE LA THEORIE DE L'INFORMATION APPLIQUES A L'ANALYSE STRUCTURALE DES SYSTEMES

Après un bref historique de la théorie de l'information, l'entropie de Shannon sera présentée en mettant en évidence ses propriétés les plus intéressantes. Des indices issus de la théorie de l'information seront construits, et pourront dès lors être utilisables dans des algorithmes de recherche d'informations, ou dans des algorithmes de classification.
Par définition, les outils de la théorie de l'information s'appliquent à des systèmes discrets, nous nous proposons d'étendre ceux-ci à des systèmes continus.

1. Bref historique de la théorie de l'information – Utilisation dans le cadre de l'analyse structurale des systèmes complexes

Bref historique...

La Théorie de l'Information (T.I.) fût amorcée dans les travaux de R.V.L. Hartley [Har28] et de H. Nyquist [Nyq24] et connût son véritable essor vers la fin des années 1940 avec les travaux de C.E. Shannon et W. Weaver [Sha48,49] [ShW49]. Cette théorie répondait à des problèmes pratiques de Télécommunication (« Théorie mathématique de la communication »), comme celui de la transmission des messages par une ligne téléphonique (ou un canal hertzien) brouillée. Le domaine militaire, avec les travaux du cybernéticien N. Wiener*, a également contribué au développement de cette théorie.

La généralisation à N variables fût entreprise avec les travaux de W.R. Garner et N.J. Mc Gill [GarG56].
Les lois de transmission et d'évolution de l'information dans les systèmes complexes fût développées grâce aux travaux de W.R. Ashby [Ashb65].
La T.I. a également été rapprochée des problèmes de régulation et de modélisation d'un système avec les travaux de Conant [Con69] [CoA70] et H.L. Weidmann [Wei69].

W.R. Ashby, R.C. Conant, N.J. Mc. Gill sont parmi les initiateurs de la T.I. appliquée à l'Analyse Structurale des Systèmes. Cette étude s'est poursuivie au L.A.I.L. avec [Tor82], [Bar87], [Pom91], [Sba93].

Enfin, la T.I. a été utilisée dans les problèmes liés à l'apprentissage [Qui83] [Man91] et a donné lieu à de nombreuses communications. Il suffit, pour s'en convaincre, de consulter la revue « Machine Learning ».

* Le mot "cybernétique" dérive du mot grec kubernetes, pilote. Il a été créé par Wiener lui-même. La cybernétique est par définition la science de la communication et de la régulation des messages.

Utilisation de la T.I. dans le cadre de l'analyse structurale des systèmes complexes...

La T.I. se prête bien à l'analyse de systèmes de tous types :

- surtout si on possède peu de connaissances *a priori*. Les seules hypothèses requises sont :
  . un nombre d'échantillons suffisant pour une estimation correcte des probabilités. Dans ce cas, il est nécessaire d'observer suffisamment longtemps les systèmes dynamiques.
  . pour les systèmes dynamiques : stationnarité et ergodicité, ainsi qu'une période d'échantillonnage appropriée [LaT75].

- La T.I. s'applique à tout type de variables (numériques, semi-numériques, qualitatives, logiques, ...) puisqu'elle est indépendante des valeurs mêmes des variables, et prend en compte seulement la partition de l'espace d'échantillonnage Ω à laquelle chaque variable donne lieu.

- Les indices de la T.I. sont sensibles à n'importe quel type de relation entre les variables (linéaire, non linéaire, logique, ...) ce qui n'est pas le cas pour tous les indices issus de la statistique [Bla68].

- Enfin, les équations de la T.I. ont toutes une interprétation intuitive très naturelle. Dans la plupart des cas, il est possible de les interpréter graphiquement.

2. Introduction à la théorie de l'information

Supposons un système physique pouvant se trouver dans un état quelconque. Ce système est alors caractérisé par un certain degré d'incertitude. Les renseignements à obtenir sur ce système sont d'autant plus importants que son incertitude *a priori* est grande. Il paraît donc important de disposer d'une mesure de son degré d'incertitude.

Afin de mieux comprendre cette notion, considérons deux systèmes : une pièce de monnaie bien équilibrée et un dé à six faces non pipé.
La pièce de monnaie, jetée en l'air, peut présenter deux côtés différents : un côté *pile* et un côté *face* équiprobables.
Le dé possède six états potentiels différents, équiprobables entre eux.

On peut alors se poser la question suivante : quel est le système possédant l'incertitude la plus grande ?
Il apparaît logique de dire que le système ayant l'incertitude la plus grande est le dé car son nombre d'états possibles (potentiels) est plus grand.
Nous pourrions donc penser que le degré d'incertitude d'un système physique est caractérisé par le nombre d'états potentiels de ce système. Montrons que la seule connaissance du nombre des états potentiels n'est pas suffisant :
Prenons par exemple un système quelconque ayant les caractéristiques suivantes :

- il peut se trouver dans un état (1) (par exemple, un état de bon fonctionnement), qui a une probabilité *a priori* de 0.99
- il peut également se trouver dans un second état, noté (2) (état de panne ou d'arrêt), qui est réalisé avec une probabilité *a priori* de 0.01.

Le degré d'incertitude sur l'état du système est très faible car il a toutes les chances de se trouver dans l'état (1) beaucoup plus souvent que dans l'état (2).

Nous voyons donc que le degré d'incertitude d'un système physique est déterminé par :
- le nombre d'états potentiels
- la probabilité d'obtention (d'occurrence) de ces états.

Appelons *Entropie* la mesure du degré d'incertitude d'un système physique.

3. Définition de l'entropie

Soient Ω un univers d'événements et A un événement particulier. La réalisation de A apporte une quantité d'information notée $I_\Phi(A)$ [Agg76].

$I_\Phi(A) := \Phi_0(p(A)) - \Phi_0(p(\Omega))$ où :

$\rightarrow \Omega$ est l'ensemble des événements possibles

$\rightarrow p(A)$ est la probabilité d'occurrence de l'événement A

$\rightarrow \Phi$ est une application telle que $\Phi : [0,1] \longrightarrow \Re^+$ et $\Phi(1) \geq 0$.

D'autre part, la réalisation d'un événement peu probable apporte beaucoup d'information, tandis qu'un événement fort probable, en apporte peu. $I_\Phi(A)$ sera donc une fonction décroissante des probabilités.

Considérons une partition $\{A_1, A_2, ..., A_i, ..., A_n\}$ de l'ensemble des états potentiels. On définit $H_\Phi(X)$ l'information fournie par (ou l'*entropie* de) la partition :

$$H_\Phi(X) = \sum_{i=1}^{n} p(A_i).I_\Phi(p(A_i))$$
$$= \sum_i p(A_i).\Phi(p(A_i)) - \Phi(1)$$
$$= \sum_i p_i.\Phi_i - \Phi(1)$$

en notant p_i la probabilité de trouver le système X dans l'état *(i)* (i=1,...,n).

De nombreuses entropies ont été définies (cf [Bar87], pp.I.27-I.28). Citons simplement :

$\rightarrow$ l'entropie du MAX : $\Phi(t) := \left(\frac{1}{t} - 1\right).\delta(1 - p_M)$

où : - $p_M = \underset{i}{\text{Max}}\ p_i$

- $\delta(t-p_M)$ est l'impulsion de Dirac au point p_M.

L'entropie du MAX est donc définie par : $H_M(X) = 1 - \underset{i}{\text{Max}}\ p_i$

→ l'entropie de Shannon : $\Phi(t) := -\text{Log}(t)$

$$H_S(X) = -\sum_{i=1}^{n} p_i . \text{Log}\, p_i = E[-\text{Log}\, p(X)]$$

où E[f(x)] est la fonction *espérance mathématique* (encore appelée *moyenne pondérée*) de f(x).

Compte tenu de ces définitions, nous pouvons remarquer que l'entropie du MAX correspond à une approche *locale*, contrairement à l'entropie de Shannon, mieux adaptée à un problème du type *global* (dû au signe *somme* Σ qui permet de faire une moyenne sur un domaine donné).

D'autre part, l'entropie de Shannon vérifie certaines propriétés intéressantes que nous examinerons par la suite, justifiant son utilisation comme mesure du degré d'incertitude d'un système.

Remarque :

La base du logarithme peut être quelconque (strictement supérieur à 1). Le changement de base est équivalent à une simple multiplication de l'entropie par un nombre constant.
En pratique, on utilise des logarithmes de base 2. On mesure donc l'entropie en unités binaires (bits). Ceci s'accorde bien avec le système binaire de représentation des informations dans les calculateurs. Ultérieurement, le symbole *Log* représentera des logarithmes binaires.

4. L'entropie de Shannon

4.1. Implications directes de la définition

Reprenons la définition de l'entropie de Shannon :

$$H(X) = -\sum_{i=1}^{n} p_i . \text{Log}\, p_i$$

où :
- $\text{Log} \equiv \text{Log}_2$
- $\sum_{i=1}^{n} p_i = 1$
- n est le nombre d'états possibles du système (de la variable X).

1ère implication :

$$\text{Log}\, n = \max_{p_i} \{ H(X) \} \Leftrightarrow \text{système (variable) à n états équiprobables}$$

L'entropie d'un système équiprobable est égale au logarithme du nombre d'états possibles.

démonstration :

$\forall p_i \quad p_i = 1/n \qquad (i=1,...,n)$

$$H(X) = -\sum_{i=1}^{n} p_i . \text{Log}\, p_i = -n.(\frac{1}{n}.\text{Log}\frac{1}{n}) = -\text{Log}(\frac{1}{n}) = \text{Log}\, n$$

cqfd.

On peut facilement montrer que l'entropie est maximale dans ce cas.

2ème implication :

$H(X) = 0 \Leftrightarrow$ Système à état unique (variable constante)

L'entropie d'un évènement certain est nulle :

- pour un seul α donné, on a $p_\alpha = 1$: $p_\alpha . \text{Log}\, p_\alpha = 0$
- $p_i = 0 \ \forall i \neq \alpha$, et par continuité : $\lim_{x \to 0} x.\text{Log}(x) = 0.$

4.2. Entropie d'un système composé

Considérons deux variables, notées X et Y, du système étudié. Chacune de ces variables possède son propre ensemble de modalités observées :

$M_X = \{\alpha_1, \alpha_2, ..., \alpha_n\}$ et $M_Y = \{\beta_1, \beta_2, ..., \beta_m\}$.

Supposons également que les probabilités conjointes d'obtention des modalités α_i et β_j soient correctement estimées par les fréquences relatives.

Nous pouvons construire le tableau de contingence suivant :

X \ Y	β_1	β_2	...	β_j	...	β_m	
α_1				.			
α_2				.			
...				.			
α_i	.	.	.	p_{ij}	.	.	$p_{i.}$
...				.			
α_n				.			
				$p_{.j}$			

Figure 1 : Tableau de contingence $[P_{ij}]$ de X et de Y.

où :

→ X et Y sont deux variables (vectorielles ou non) du système.

→ $\{\alpha_1, \alpha_2, ..., \alpha_n\}$ et $\{\beta_1, \beta_2, ..., \beta_m\}$ sont respectivement les ensembles de modalités potentielles de X et de Y.

→ Les p_{ij} (i=1,...,n ; j=1,...,m) sont les probabilités conjointes de X et de Y.

$$\rightarrow \quad p_{i.} = \sum_{j=1}^{m} p_{ij} \quad \text{pour } i=1,...,n.$$

$$\rightarrow \quad p_{.j} = \sum_{i=1}^{n} p_{ij} \quad \text{pour } j=1,...,m.$$

Par définition de l'entropie de Shannon, nous avons :

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}.\text{Log } p_{ij} = E[-\text{Log } p(X,Y)]$$

Remarques :

→ Nous pouvons montrer l'inégalité suivante :
$H(X,Y) \leq H(X) + H(Y)$

→ En vertu du théorème de multiplication des probabilités des événements indépendants : $P(X,Y)=P(X).P(Y)$
on a $\text{Log } P(X,Y) = \text{Log } P(X) + \text{Log } P(Y)$
et donc $H(X,Y) = H(X) + H(Y)$

4.3. Entropie conditionnelle

Considérons maintenant deux systèmes X et Y dépendants l'un de l'autre. $p(\beta_j/\alpha_i)$ représente la probabilité conditionnelle pour que le système Y se trouve dans l'état β_j lorsque le système X se trouve dans l'état α_i.
Or, l'entropie du système Y lorsque le système X se trouve dans l'état α_i est :

$$H(Y/\alpha_i) = -\sum_{j=1}^{m} p(\beta_j/\alpha_i).\text{Log } p(\beta_j/\alpha_i)$$

L'entropie totale du système Y est donc :

$$H(Y/X) = \sum_{i=1}^{n} p_{i.}.H(Y/\alpha_i) = -\sum_{i=1}^{n} p_{i.}.\sum_{j=1}^{m} p(\beta_j/\alpha_i).\text{Log } p(\beta_j/\alpha_i)$$

$$H(Y/X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{i.}.p(\beta_j/\alpha_i).\text{Log } p(\beta_j/\alpha_i)$$

Remarque :
En vertu de la définition de la probabilité conditionnelle ($p(X,Y)=p(X).p(Y/X)$), on démontre que : $H(Y/X) = H(X,Y) - H(X)$

4.4. Entropie et information

4.4.1. Position du problème

Nous avons vu que l'entropie d'une variable est la mesure du degré d'incertitude de celle-ci. Il semble *naturel* de dire que l'obtention d'informations sur cette variable diminue son incertitude. La quantité d'informations obtenue sur cette variable peut donc être mesurée par la diminution de son entropie.
Prenons le cas particulier d'une variable X_i ayant une entropie *a priori* $H(X_i)$. Si après obtention d'informations, cette variable est totalement connue, alors son entropie s'annule. Notons $\vec{I}(X_i:X_i)$ l'information obtenue avec la détermination de X_i :

$$\vec{I}(X_i:X_i) = H(X_i) - 0 = H(X_i).$$

Ce que nous venons d'énoncer pour la variable X_i peut être généralisé pour tout $S \in P^v(\Sigma)$, où :

- $\Sigma=(X,Y)=(X_1,\ldots,X_N,Y)$ représente le vecteur de toutes les variables considérées comme pertinentes pour l'étude du système.
- $P^v(\Sigma)$ représente tous les vecteurs (toutes les variables multidimensionnelles) possibles dont les composantes sont des composantes élémentaires de Σ.

Ainsi, H(S) représente une mesure de l'information nécessaire pour connaître S.

Si l'on considère maintenant deux variables X et Y, l'inégalité $H(X,Y) \leq H(X)+H(Y)$ peut être interprétée par le fait que, dans l'expression H(X)+H(Y), on compte deux fois une partie de l'information qui se trouve aussi bien dans X que dans Y.
Afin de mesurer cette quantité d'information, nous définissons la *transinformation interne* [Ash65] et de la *transinformation externe*.

4.4.2. Transinformation interne

$$\overset{\circ}{I} : P^v(\Sigma) \rightarrow R$$

$$S \mapsto \overset{\circ}{I} := \sum_{X_i \in S} H(X_i) - H(S)$$

$\overset{\circ}{I}$ mesure la quantité d'information qui est transmise *dans* un groupe de variables.

4.4.3. Transinformation externe

Par définition :

$$\vec{I}: \quad P(P^v(\Sigma)) \quad \rightarrow \quad R$$

$$R = \{R_1, R_2, ..., R_r\} \mapsto \vec{I}(R) := \sum_{i=1}^{r} H(R_i) \; - \; H(\bigcup_{i=1}^{r} R_i)$$

$\vec{I}$ mesure la quantité d'information qui est transmise *entre* plusieurs groupes de variables, où P(E) représente une partie de l'ensemble E.

4.4.4. Propriétés de l'entropie et des transinformations

De nombreuses propriétés intéressantes ont été démontrées [Sha49] [Mil63] [Ash65] [Tor82]. M. Sbaï [Sba83] fait une synthèse de ces propriétés :

$\forall S_1, S_2, S_3 \in P^v(\Sigma)$; $\forall R, Q \in P(P^v(\Sigma))$

Propriété 1 : $H(S_1) \geq 0$; $\mathring{I}(S_1) \geq 0$; $H(S_1/S_2) \geq 0$; $\vec{I}(R) \geq 0$

Propriété 2 : $H(S_1 \cap S_2) + H(S_1 \cup S_2) \leq H(S_1) + H(S_2)$
avec égalité si et seulement si $S_1 \cap S_2 = \varnothing$

Propriété 3 : $S_1 \subset S_2 \quad \Rightarrow \quad H(S_1) \leq H(S_2)$
$\mathring{I}(S_1) \leq \mathring{I}(S_2)$
$H(S_3/S_1) \geq H(S_3/S_2)$

Propriété 4 : $R \leq Q \quad \Rightarrow \quad \vec{I}(R) \geq \vec{I}(Q)$
où : - le 1er signe "$\leq$" représente la relation "... être plus fine que ..."
- le 2ème signe "$\leq$", la relation "... être plus petit que ..."

Propriété 5 : $\forall R = \{R_1, ..., R_r\} \in P(P^v(\Sigma))$
$\vec{I}(R) = 0 \quad \Leftrightarrow \quad R_1, R_2, ..., R_r$ sont des sous-ensembles de variables indépendants entre eux.

Propriété 6 : $\vec{I}(R) + \sum_{i=1}^{r} \mathring{I}(R_i) = \text{Constante} = \vec{I}(\{X_i\}; i = 1, ..., N)$

Propriété 7 : $\forall R = \{R_1, R_2, ..., R_i, ..., R_j, ..., R_r\} \in P(P^v(\Sigma))$ et $R' = \{R_1, R_2, ..., R_i \cup R_j, ..., R_r\}$,
alors : $\vec{I}(R') = \vec{I}(R) - \vec{I}(R_i, R_j)$.

Propriété 8 : Soit $R' = \{R_1, R_2, ..., R_i \backslash S, S, ..., R_r\}$ avec : $S \subset R_i$; $S \neq \varnothing$; $S \neq R_i$; $|R_i| \geq 2$
alors : $\vec{I}(R') = \vec{I}(R) + \vec{I}(R_i \backslash S, S)$.

Propriété 9 : Principe de conditionnement uniforme [GaG56] :

"Si l'on conditionne, avec le même ensemble de variables, tous les termes d'une identité valable pour H=entropie qui contient H(.), H(/.), $\tilde{I}$, $\tilde{I}_\alpha$, $\mathring{I}$, $\mathring{I}_\beta$, on obtient à nouveau une identité valable pour l'entropie (le double conditionnement étant équivalent au conditionnement avec l'union)".

où :

- $\tilde{I}_\alpha(A:B) := H(A/\alpha) + H(B/\alpha) - H(A \cup B/\alpha)$
- $\mathring{I}_\beta(R) := \sum_{X_i \in V} H(X_i/\beta) - H(R/\beta)$

$\forall A,B,\alpha,\beta \in P^v(\Sigma)$ et $\forall R \in P(P^v(\Sigma))$

exemple : $\forall A,B \in P^v(\Sigma)$: $H(A \cup B) = H(A) + H(B) - \tilde{I}(A:B)$.

On a d'après ce principe :

$\forall C \in P^v(\Sigma)$: $H(A \cup B/C) = H(A/C) + H(B/C) - \tilde{I}_C(A:B)$.

4.4.5. Détermination de quelques indices [Tor82]

L'incohérence des données recueillies sur un système complexe implique un manque d'explication de Y par X (où X et Y représentent des variables vectorielles). En effet, nous ne pourrons pas trouver une application f telle que Y=f(X). En pratique, nous chercherons à l'approcher ($Y \approx f(X)$).

Nous déterminerons alors quelques indices mesurant l'*explicabilité* (ou la *modélisabilité*) de Y par X :

Soit $S \in P^v(X)$:

- $\tilde{I}(Y:S) := H(Y) + H(S) - H(Y,S) = H(Y) - H(Y/S)$
  est l'entropie de Y expliquée par S.

- $H(Y/S) := H(Y,S) - H(S)$ est l'entropie de Y non expliquée par S.

- $\%H(Y/S) := H(Y/S) / H(Y)$ est le pourcentage de l'entropie de Y non expliquée par S.

Si l'on remplace S par X dans les trois expressions ci-dessus, nous obtenons :

- $H(Y/X) := H(Y,X) - H(X)$ est la partie non explicable de Y.

- $\tilde{I}(Y:X) := H(Y) - H(Y/X)$ est la partie explicable de l'entropie de Y.

- $\%H(Y/X) := H(Y/X) / H(Y)$ est le pourcentage explicable de l'entropie de Y.

- $1 - \%H(Y/X) = \tilde{I}(Y:X) / H(Y)$ est le pourcentage explicable de l'entropie de Y.

D'autres indices peuvent également être définis :

- %ENEPE(Y/S) := * $\frac{H(Y/S) - H(Y/X)}{\tilde{I}(Y:X)}$ si $\tilde{I}(Y:X)>0$

* 0 si $\tilde{I}(Y:X)=0$

représente le pourcentage encore non expliqué par S de la partie explicable de Y.

4.5. Entropie conditionnelle et classification

L'entropie conditionnelle est particulièrement intéressante dans le cas d'un problème de classification ou d'explication. En effet, H(Y/S) peut être interprétée comme la quantité d'information restant à connaître (l'incertitude) sur Y, une fois que l'on possède déjà l'information fournie par S. H(Y/S) représente l'entropie de Y non expliquée par S. Trois cas distincts sont alors susceptibles d'apparaître :

1er cas : H(Y/S) = H(Y) : Y et S sont indépendants statistiquement. Aucune relation entre ces grandeurs ne pourra être mise en évidence.

2ème cas : H(Y/S) = 0 : Y et S sont totalement liées et il existe une variable $\tilde{S} \geq S$ (au sens large) telle que $Y=f(\tilde{S})$.

3ème cas : 0 < H(Y/S) < H(S) : S n'explique que partiellement Y.

Ces propriétés ont inspiré la définition d'un coefficient de couplage mesurant la modélisabilité de Y par S [Con76]. Ce coefficient de couplage est défini par :

$$m(Y/S)=1-\frac{H(Y/S)}{H(Y)}=\frac{H(Y)-H(Y,S)}{H(Y)}$$

Les trois cas précédent peuvent alors s'écrire en utilisant le coefficient de couplage :

1er cas : m(Y/S) = 1 $\Leftrightarrow$ Y et S sont indépendants statistiquement.
2ème cas : m(Y/S) = 0 $\Leftrightarrow$ Y et S sont totalement liées
3ème cas : 0 < m(Y/S) < 1 $\Leftrightarrow$ S n'explique que partiellement Y.

Remarque : Dans le 3ème cas, S n'explique que partiellement Y. Nous distinguons alors deux causes possibles :

(1) *l'absence d'explication à tort* : le modèle déterministe Y=f(S) est représentatif du système, mais les données relevées sont biaisées. Ceci peut être dû à la présence d'un bruit important sur les mesures effectuées ou/et au mauvais fonctionnement d'un sous-système.

(2) *l'absence d'explication à raison* : certaines variables explicatives, pertinentes pour l'étude de notre système, ont été omises.

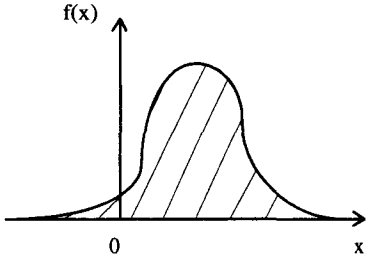
5. Entropie et information des systèmes continus

Jusqu'à présent, nous avons envisagé des systèmes dont les états étaient dénombrables. Les variables considérées étaient qualitatives, ordinales ou non. La condition d'ordre se traduit alors par le fait que l'on travaille sur des modalités consécutives de la variable considérée, ou sur des ensembles consécutifs de modalités. Ces variables qualitatives X_i sont des variables aléatoires discrètes (ou discontinues) prenant les valeurs $\{\alpha_{i1},\alpha_{i2},\ldots,\alpha_{in_i}\}=M_{Xi}$ et dont les probabilités "empiriques" sont respectivement $\{p_1, p_2, \ldots, p_{n_i}\}$.

Dans la pratique, nous rencontrons souvent des variables quantitatives, *i.e.* des variables continues. Leurs états (valeurs) varient de façon continue. La répartition des probabilités est alors caractérisée par une certaine densité. Nous noterons f(x) la densité de probabilité de la variable continue x. Ainsi, une variable vectorielle (multidimensionnelle) $S=(x_1,x_2,\ldots,x_n)$ continue sera caractérisée par sa fonction de densité de probabilité $f(x_1,x_2,\ldots,x_n)$. Cette variable vectorielle est la réunion des variables primaires $x_1, x_2, \ldots, x_n$.

5.1. Entropie d'une variable continue [Ven73]

Considérons une variable continue x caractérisée par une fonction de densité de probabilités f(x) :



avec $\int_{-\infty}^{+\infty} f(x) \, . \, dx = 1$

Figure 2 : Fonction de densité de probabilités d'une variable continue

Il est alors important de remarquer que, même si on considère des variables continues, celles-ci ne nous apparaissent que sous une forme discrète. En effet, ce que nous observons, ce sont les « mesures » réalisées par un capteur plus ou moins précis, et qui, de par sa nature, discrétise le signal obtenu (la variable continue considérée).

Par exemple, la taille d'un être humain pourra être approchée au cm près, voire au mm près. De même, son poids sera vu au gramme près, son âge, au jour près, ...

De plus, le traitement « informatique » des données nécessite, de par la nature de celui-ci, une discrétisation préalable des variables continues.

Considérant ces deux observations, on peut décrire une variable continue par une variable discrète, après avoir établi la précision des mesures Δx : A l'intérieur d'un segment Δx, les états (les valeurs) de la variable x sont indiscernables. La courbe précédente est alors remplacée par la courbe en escalier suivante :

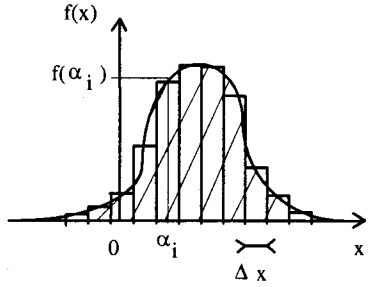


Figure 3 : Discrétisation de la fonction de densité de probabilités

Δx est la précision de lecture (ou « zone d'insensibilité ») de la variable continue x.
Chaque segment Δx est alors remplacé par une unique valeur (un unique point) représentative.
La grandeur $f(\alpha_i).\Delta x$ (aire d'un rectangle) représente alors la probabilité de tomber dans le segment correspondant.
Dans un même segment, les états de la variable x considérée étant indiscernables, on peut approximer l'entropie de x (à Δx près) de la façon suivante :

$$H_{\Delta x}(x) = -\sum_i f(\alpha_i).\Delta x.\log[f(\alpha_i).\Delta x]$$

$$= -\sum_i f(\alpha_i).\Delta x.[\log f(\alpha_i) + \log \Delta x]$$

$$= -\sum_i [f(\alpha_i).\log f(\alpha_i)]\Delta x - \log \Delta x \sum_i f(\alpha_i).\Delta x$$

Pour Δx suffisamment petit :

$$\sum_i f(\alpha_i).\log f(\alpha_i).\Delta x \ \# \ \int_{-\infty}^{+\infty} f(x).\log f(x)\,.\,dx$$

$$\text{et} \quad \sum_i f(\alpha_i).\Delta x \ \# \ \int_{-\infty}^{+\infty} f(x)\,.\,dx \ = 1$$

$$\Rightarrow \quad H_{\Delta x}(x) = -\int_{-\infty}^{+\infty} f(x).\log f(x)\,.\,dx \ - \log \Delta x$$

L'entropie est donc la somme de deux termes, le premier ne dépendant que de la densité de probabilité de la variable considérée, l'autre ne dépendant que de la précision avec laquelle on mesure cette variable.

Remarque :

Si $\Delta x \to 0$, alors $-\log \Delta x \to +\infty$. Plus la mesure de l'état de x est précise, plus l'incertitude croît, et plus l'entropie est importante.

Entre $H_{\Delta X}(x)$ et ce terme croissant indéfiniment avec la précision de mesure, il y a une différence :

$$H^*(x) = -\int_{-\infty}^{+\infty} f(x).\log f(x)\,.\,dx$$

que l'on appellera "entropie réduite" de la variable continue x.

Et donc :

$$H_{\Delta X}(x) = H^*(x) - \log \Delta x$$

La précision de la mesure va donc en quelque sorte fixer l'origine du calcul de l'entropie.

Exemples : Calcul de l'entropie de Shannon d'une fonction de densité paramétrique

- Cas où une variable x suit une densité de probabilités gaussienne :

$$f(x) = \frac{1}{\sqrt{2.\pi}\,.\sigma}\,.\,e^{-\frac{x^2}{2.\sigma^2}}$$

$$H^*(x) = E[-\log f(x)] = E\left[-\log\left(\frac{1}{\sqrt{2.\pi}\,.\sigma}\,.\,e^{-\frac{x^2}{2.\sigma^2}}\right)\right]$$

$$= E\left[\log(\sqrt{2.\pi}\,.\sigma) + \frac{x^2}{2.\sigma^2}\,.\log e\right]$$

$$= \log(\sqrt{2.\pi}\,.\sigma) + \frac{\log e}{2.\sigma^2}\,E[x^2]$$

De plus :

$$E[x^2] = Var[x] = \sigma^2$$

$$H^*(x) = \log(\sqrt{2.\pi}\,.\sigma) + \frac{1}{2}\,.\log e$$

$$= \log(\sigma.\sqrt{2.\pi}\,.e)$$

$$H(x) = \log(\sigma.\sqrt{2.\pi}\,.e) - \log \Delta x$$

$$\Rightarrow \quad \boxed{H(x) = \log\left(\frac{\sigma.\sqrt{2.\pi}\,.e}{\Delta x}\right)}$$

- Cas où x suit une densité de probabilité équiprobable sur un intervalle [a,b] :

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{pour } a \le x \le b \\ 0 & \text{pour } x<a \text{ et } x>b \end{cases}$$

$$H^*(x) = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a}\, dx$$

$$= \log(b-a)$$

Et donc : $H(x) = \log(b-a) - \log \Delta x$

$$\Rightarrow \boxed{H(x) = \log \frac{b-a}{\Delta x}}$$

Remarques :

- Nous omettrons l'indice Δx : $H_{\Delta X}(x)$ sera noté H(x), ceci afin d'alléger les notations.

- Nous avons les relations suivantes :

$$H(x) = -\int_{-\infty}^{+\infty} f(x).\log f(x)\,.\,dx - \log \Delta x$$

$$= -\int_{-\infty}^{+\infty} f(x).[\log f(x) + \log \Delta x]\,.\,dx$$

$$= -\int_{-\infty}^{+\infty} f(x).\log [f(x).\Delta x]\,.\,dx$$

$$= E[-\log[f(x).\Delta x]]$$

- De même : $H^*(x) = E[-\log f(x)]$

- Δx dépend de l'unité employée pour mesurer la variable continue x. Lorsque l'on considère deux variables continues x et y, n'ayant pas les mêmes unités de mesures, peut-on comparer H(x) et H(y) ? Quel sens faut-il donner à cette comparaison ? C'est ce que nous allons étudier dans le paragraphe suivant concernant l'entropie conditionnelle.

5.2. Entropie et information

Nous avons noté la présence du terme Δx (précision) dans le calcul de H(x).
Or, si on apprend une certaine information sur x, l'incertitude sur son état diminue. Donc, plus l'information dont on dispose sur x est importante, plus l'incertitude sur son état va diminuer.
Il est alors naturel de mesurer l'information apportée sur x par la différence d'entropie entre ces deux états (état final et état initial). Le terme en « Δx » va alors disparaître, et par conséquent l'information apportée sur une variable x ne dépend pas de la précision avec laquelle on mesure celle-ci.

Ainsi, l'information apportée par la connaissance de f(x) est :

$$I(x) = H^*(x) = -\int_{-\infty}^{+\infty} f(x).\log f(x)\,.\,dx$$

C'est-à-dire que la quantité d'information acquise lorsque l'état de x devient complètement connu est égale à l'entropie réduite de x.

5.3. Entropie conditionnelle

Soient deux variables continues x et y.
Appellons :

- . f(x,y) la densité de probabilité de la variable (x,y).
- . $f_1(x)$ la densité de probabilité de x.
- . et $f_2(y)$ la densité de probabilité de y.

f(y/x) et f(x/y) sont les densités de probabilités conditionnelles, avec :

$$f(y/x) = \frac{f(x,y)}{f_1(x)} \qquad \text{et} \qquad f(x/y) = \frac{f(x,y)}{f_2(y)}$$

Notons $H(y/x=\alpha_i)$ l'entropie conditionnelle partielle qui représente l'entropie de y sachant que x se trouve dans l'état α_i :

$$H(y/x=\alpha_i) = - \sum_j f(\beta_j/\alpha_i)\,.\,\Delta y\,.\,\log\,[\,f(\beta_j/\alpha_i)\,.\,\Delta y\,]$$

et lorsque Δy est suffisamment petit :

$$H\,(y/x=\alpha_i\,) = -\int_{-\infty}^{+\infty} f(y/\alpha_i).\,\log f(y/\alpha_i)\,.\,dy \qquad - \log\Delta y$$

H(y/x) est alors la moyenne des entropies conditionnelles partielles pour tous les états α_i de x (en prenant en compte leur densité de probabilité $f_1(x)$) :

$$H(y/x) = -\int_{-\infty}^{+\infty}\int \underbrace{f_1(x) \cdot f(y/x)}_{=f(x,y)} \cdot \log f(y/x) \cdot dy.dx \quad - \underbrace{\int_{-\infty}^{+\infty} f_1(x) \cdot dx}_{=1} \cdot \log\Delta y$$

$$\boxed{H(y/x) = -\int_{-\infty}^{+\infty}\int f(x,y) \cdot \log f(y/x) \cdot dy.dx \quad - \log\Delta y}$$

Remarque : $H(y/x) = E[- \log f(y/x)] - \log \Delta y$
$= E[- \log [f(y/x) \cdot \Delta y]]$

5.4. Entropie d'un système composé

La précision des variables continues x et y étant respectivement Δx et Δy, la précision de la variable (x,y) sera le rectangle $\Delta x.\Delta y$.

$H(x,y) = E[- \log [f(x,y) \cdot \Delta x \cdot \Delta y]]$
et $f(x,y) = f_1(x) \cdot f(y/x)$

$H(x,y) = E[- \log f_1(x) - \log f(y/x) - \log \Delta x - \log \Delta y]$
$= E[- \log [f_1(x) \cdot \Delta x]] + E[-\log [f(y/x) \cdot \Delta y]]$

$\Rightarrow H(x,y) = H(x) + H(y/x)$

Le théorème de l'entropie d'un système composé (avec des variables discrètes) est applicable dans le cas de variables continues.

De même : $H(x,y) = H(x) + H(y)$ si x et y sont deux variables continues indépendantes statistiquement.

Remarque : *Calcul de la transinformation interne (information mutuelle contenue dans x et y)*
Par définition, la transinformation interne mesure la quantité d'information qui est transmise dans un groupe de variables :

$$I := H(x) + H(y) - H(x,y)$$

$$= -\int_{-\infty}^{+\infty} f_1(x).\log f_1(x)\,.\,dx \quad -\cancel{\log\Delta x}$$

$$-\int_{-\infty}^{+\infty} f_2(y).\log f_2(y)\,.\,dy \quad -\cancel{\log\Delta y}$$

$$+\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y).\log f(x,y)\,.\,dx\,.\,dy \quad +\cancel{\log\Delta x} + \cancel{\log\Delta y}$$

hors, $f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\,.\,dy$ et $f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\,.\,dx$

ce qui entraîne : $$I = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y).\log\frac{f(x,y)}{f_1(x)\,.\,f_2(y)}\,.\,dx\,.\,dy$$

La généralisation à plus de deux variables est immédiate.

6. Conclusion

Dans ce chapitre, nous avons montré tout l'intérêt de l'utilisation de l'entropie dans le cadre d'une analyse structurale d'un système complexe. Elle permet, entre autres :

- l'utilisation de variables numériques (continues), mais également qualitatives,
- de mettre en évidence des relations autres que linéaires,
- d'avoir une interprétation intuitive naturelle en termes de quantité d'information apportée.

Nous nous proposons donc tout naturellement de l'utiliser dans le cadre de la théorie de la détection.

3ième partie :

APPRENDRE

AFIN DE MIEUX DETECTER

CHAPITRE 5

INTRODUCTION

D'UNE PHASE D'APPRENTISSAGE

DANS LES SYSTEMES

DE DETECTION DECENTRALISEE PARALLELE

1. Introduction

Comme nous l'avons rappelé dans le premier chapitre, le problème de la détection décentralisée reste encore aujourd'hui un problème complexe et multiforme. Le but recherché est la mise en commun de données collectées par N capteurs suivant une architecture donnée. Dans ce but, deux architectures ont été étudiées : la détection décentralisée parallèle, et la détection décentralisée série. L'optimisation de ces différentes architectures a permis de déterminer les opérateurs de traitement locaux permettant d'obtenir les meilleures performances de détection suivant un critère donné. Ces optimisations aboutissent à un difficile problème de résolution de systèmes d'équations. Dans le cas de la détection décentralisée parallèle, que l'on se place dans le cas Bayésien ou de Neyman-Pearson, l'optimisation d'un système comprenant N capteurs aboutit à un système de 2^N+N équations non linéaires couplées à résoudre. Dans le cas de la détection décentralisée série, on aboutit à un système de 2N-1 équations non linéaires couplées à résoudre. Ces systèmes n'ont pour l'instant pu être résolus que pour des cas particuliers en supposant par exemple l'indépendance des observations locales et pour des systèmes comportant peu de capteurs. De façon générale, on s'aperçoit que le nombre d'équations à résoudre simultanément croît très rapidement avec le nombre de capteurs, les calculs nécessaires à la résolution de ces équations deviennent alors très vite inextricables.

Afin de simplifier le problème d'optimisation de ces systèmes, nous proposons de limiter le nombre de capteurs à prendre en compte lors de l'optimisation du système de détection. Considérons par exemple le contexte de la surveillance d'installations industrielles complexes, où un grand nombre de capteurs observant des grandeurs physiques différentes est disponible. L'optimisation d'un système de détection décentralisée utilisant tous les capteurs s'avère très vite impossible à réaliser. Cependant, il est possible que l'on puisse implémenter une structure de détection en ne considérant qu'un sous-ensemble des capteurs, plutôt que l'ensemble des informations disponibles. Dans certains cas, ceux-ci peuvent en effet présenter des redondances, sans améliorer les performances de l'ensemble, ou noyer le système sous un flot d'informations trop coûteux à gérer. C'est pour cette raison que nous proposons d'introduire dans les systèmes de détection décentralisée, avant toute optimisation, une étape de sélection de capteurs [CDS95] [DPS97] [PeP97]. Parmi tous les capteurs disponibles, nous proposons de ne faire intervenir que ceux apportant beaucoup d'information au processus de décision. Dans ce but, nous utilisons une phase d'apprentissage inspirée des problèmes de classification développés dans le chapitre 3. Nous proposons différents algorithmes de sélection basés sur le critère entropique introduit dans les chapitres 2 et 4 et qui est tout à fait adapté à ce problème de sélection.

La sélection de capteurs étant faite, on pourra optimiser le système en utilisant les résultats classiques de la théorie de la détection exposés dans le premier chapitre, ou en utilisant un critère entropique comme nous l'avons montré dans le second chapitre. Dans le premier cas, l'optimisation du système risque d'être encore inextricable, même si la complexité du système a été préalablement diminuée par notre phase de sélection de capteurs. Dans le second cas, nous montrerons que les propriétés particulières de l'entropie peuvent être mises à profit pour limiter la complexité des calculs à mettre en œuvre lors de l'optimisation du système de détection. Nous nous baserons sur les méthodes de construction d'arbres de décision développées en classification pour proposer des algorithmes basés sur le critère entropique présenté au chapitre 2. Ces algorithmes nous permettront de nous approcher du système de

détection décentralisé parallèle optimal en limitant le plus possible la complexité des calculs à mettre en œuvre.

Enfin, il paraît évident que les performances des systèmes de détection centralisée sont meilleures que les performances obtenues via des systèmes de détection décentralisée. Dans ce sens, les techniques d'optimisation précédentes seront étendues au problème de la quantification répartie afin d'obtenir un compromis entre la quantité d'information à envoyer à l'opérateur de fusion et les performances du système de détection.

2. Le problème de la détection vu comme un problème de classification

Les systèmes que nous étudions ont pour but de résoudre un problème de détection, qui se traduit par le choix entre deux hypothèses H_0 et H_1 à partir d'observations collectées sur un ensemble de N capteurs. Chaque source délivre un vecteur y_i i=1,...,N. Les mesures y_i relevées par chaque capteur Y_i sont supposées prendre leurs modalités dans un ensemble M_{yi} et on note $y=(y_1,y_2,...,y_N)$. L'ensemble des observations y fournies par les capteurs constitue l'espace des observations, noté D. Décider que l'on se trouve dans la situation H_0 ou H_1 revient à diviser l'espace des observations D en deux domaines disjoints D_0 et D_1 tels que si l'observation tombe dans D_0 (respectivement dans D_1) la décision prise est H_0 (respectivement H_1). Une partition de l'espace des observations D en deux classes est alors obtenue.

En théorie de la détection, on suppose toujours que l'on dispose d'un modèle statistique du système étudié. Les fonctions de densité de probabilités conditionnelles sous chaque hypothèse $p(y/H_i)$, i=0,1 sont donc connues.

Cependant, plus le nombre de capteurs sera important, plus il sera difficile de disposer d'un modèle statistique du système. Dans ce cas, comme c'est la pratique dans les problèmes de classification, on pourra déterminer ces probabilités en les estimant par leurs fréquences relatives déterminées à partir d'un ensemble d'exemples de taille suffisamment grande.
De plus, on pourra considérer que les variables aléatoires y_i (i=1,...,N) sont des variables discrètes. En effet, il y a tout lieu de noter que dans le contexte de la détection, la notion de variable aléatoire continue représente une certaine « idéalisation » de la réalité. Considérons, par exemple, les mesures relevées par un capteur de température. En considérant ces mesures comme des mesures continues, nous faisons abstraction du fait qu'en réalité, le capteur ne dispose que d'une certaine précision, et que dans le cas de capteurs numériques, l'information délivrée est *a fortiori* discrétisée (dû au quantum élémentaire du Convertisseur Analogique-Numérique, qui est par définition la plus petite variation de l'entrée du CAN permettant de faire varier la valeur du bit de poids faible). Il est impossible de distinguer, par exemple, entre deux températures qui diffèrent entre elles de 0.01°. En pratique, ayant le quantum élémentaire Δy_i associé à la variable aléatoire y_i, c'est-à-dire ayant choisi un certain segment Δy_i à l'intérieur duquel les mesures y_i sont indiscernables, on peut considérer que les mesures délivrées par les capteurs sont discrètes (donc qualitatives). Cette réflexion est renforcée par le fait que le traitement de l'information délivrée par un capteur sur une machine informatique nécessite des grandeurs discrètes.
Le cas où l'on considère des variables dont on connaît un modèle probabiliste continu (ou discret) peut également être traité dans le cadre de ce travail.

Par analogie avec les problèmes de classification, les mesures délivrées par les capteurs peuvent être assimilées aux variables de description du système, la décision à la fonction de classification qui à toute observation associe une hypothèse H_0 ou H_1, et l'espace des observations D à $M_{y_1} \times M_{y_2} \times \ldots \times M_{y_N}$. C'est dans ce contexte que nous introduisons la notion de finesse entre les variables d'un système de détection (mesures, rapports de vraisemblance, décisions locales et finale). A chacune de ces variables, on pourra associer une partition plus ou moins fine de l'espace des observations D, qui induit une partition plus ou moins fine de la population d'apprentissage Ω.

3. Les relations de finesse entre les variables d'un système de détection

3.1. Rapport de vraisemblance et finesse des variables vectorielles

Nous nous proposons d'appliquer la notion de finesse introduite au chapitre 3 au problème de la détection centralisée et décentralisée.

Les fonctions de densité de probabilités transforment une variable discrète multidimensionnelle S (quelle que soit sa dimension) en une variable dont la partition suivant la population d'apprentisage sera plus grosse que la partition suivant la variable S. Nous obtenons ainsi $P_{p(S)}(\Omega) \geq P_S(\Omega)$.
Il en est de même pour les fonctions de densité de probabilités conditionnelles. Nous obtenons ainsi $P_{p(S/Hi)}(\Omega) \geq P_S(\Omega)$. Nous pouvons également extrapoler cette remarque aux variables multidimensionnelles continues.
De même, le rapport de vraisemblance, qui constitue en fait un rapport de densités de probabilités conditionnelles, peut être considéré comme une variable unidimensionnelle dont la partition sur la population d'apprentissage est plus grosse que la partition de la population d'apprentissage engendrée par la variable multidimensionnelle considérée.
Enfin, ce rapport de vraisemblance est comparé à un seuil qui permet dès lors de prendre une décision (u_i=0 ou 1). Nous nous retrouvons par conséquent avec une partition en deux classes de la population d'apprentissage plus grosse que la partition engendrée par le rapport de vraisemblance, elle-même plus grosse que la partition engendrée par la variable en question.
Dans le cadre général de la détection, nous pouvons établir le schéma ci-dessous :

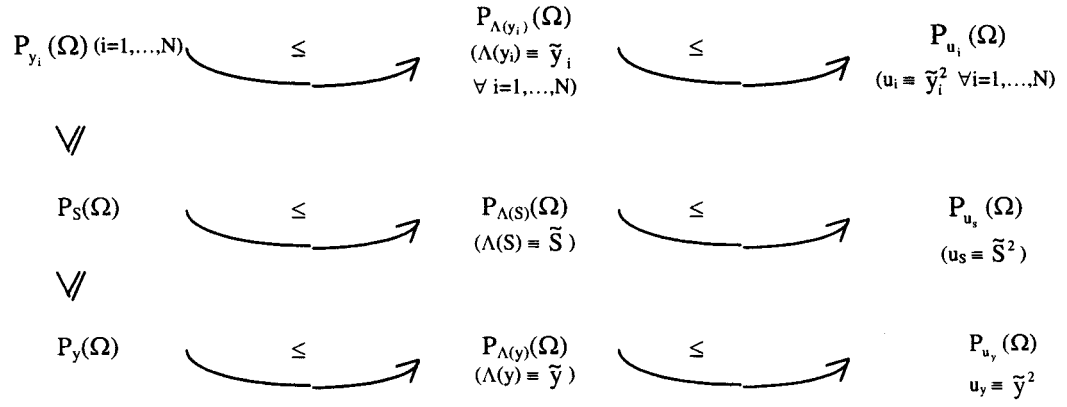


Figure 1 : Relations entre les partitions de l'espace des observations engendrées par les variables multidimensionnelles et le rapport de vraisemblance.

où :
- le symbole « ≤ » représente la notion de finesse entre deux variables multidimensionnelles.
- S représente une variable multidimensionnelle quelconque.
- $\tilde{S}$ représente une variable plus grosse que S et $\Lambda(S)$ représente le rapport de vraisemblance de S.
- $\tilde{S}^2$ représente une variable plus grosse que $\tilde{S}$ en 2 classes.
- y_i (resp.y) n'est qu'un cas particulier de S en ce sens qu'il ne représente qu'une variable unidimensionnelle (resp. le vecteur constitué de toutes les variables considérées) (i=1,...,N).

3.2. Utilisation d'un critère entropique dans le cadre de la détection

Parmi tous les critères utilisables, le critère entropique est certainement l'un des critères les plus intéressants car il possède de nombreuses propriétés et permet de mettre en évidence des relations non linéaires entre les données. Ce critère est, par exemple, à la base du noyau ID3 de l'algorithme C4.5, qui est une référence dans le domaine de l'apprentissage automatique.

Nous pouvons ainsi compléter le schéma précédent de la façon suivante :

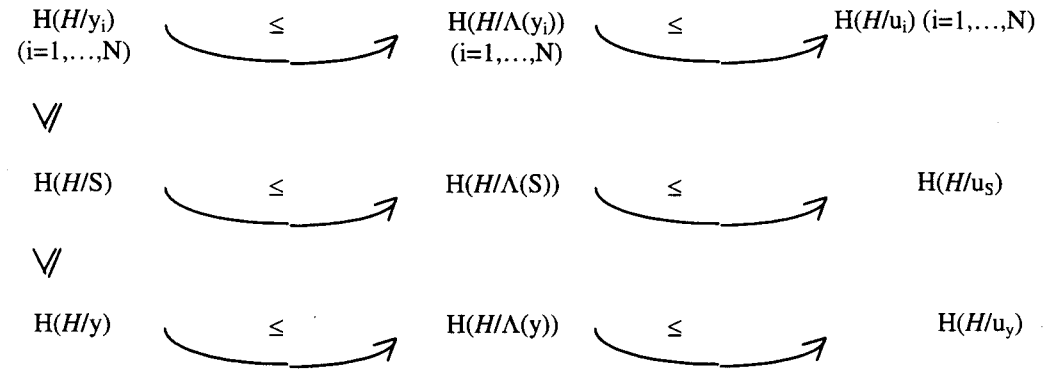


Figure 2 : Relations entre les entropies conditionnelles de l'hypothèse vraie connaissant la variable, son rapport de vraisemblance, ou la décision prise.

L'utilisation d'un critère entropique associé à la notion de finesse nous permet de définir une mesure numérique de la quantité d'information apportée par un ensemble de variables.

Nous nous proposons dès lors de développer cette démarche dans le cadre de la détection centralisée, décentralisée parallèle et série.

3.3. Cas de la détection centralisée

Dans le cadre de la détection centralisée, que le critère soit le critère de Bayes, de Neyman-Pearson ou que l'on utilise un critère entropique, la théorie aboutit à la même structure optimale du détecteur ; l'ensemble des informations délivrées par les N capteurs Y_i (i=1,..,N) est transmis à un opérateur de décision qui prend la décision finale u_0 qui consiste à comparer, pour chaque observation $y=(y_1,\ldots y_N)$, le rapport de vraisemblance à un seuil λ, soit :

$$\Lambda(y) \underset{u_0=0}{\overset{u_0=1}{\gtrless}} \lambda \quad \text{où} \quad \Lambda(y) = \frac{p(y/H_1)}{p(y/H_0)}$$

Dans le cadre d'une phase d'apprentissage, la décision finale u_0 peut être considérée comme une variable plus grosse que $\Lambda(y)$, dû au fait qu'elle engendre une partition en deux classes de la population d'apprentissage plus grosse que la partition engendrée par $\Lambda(y)$. Cette dernière représente de même une variable plus grosse que y. Nous obtenons ainsi la relation suivante : $u_0 \geq \Lambda(y) \geq y$ (Figure 3).

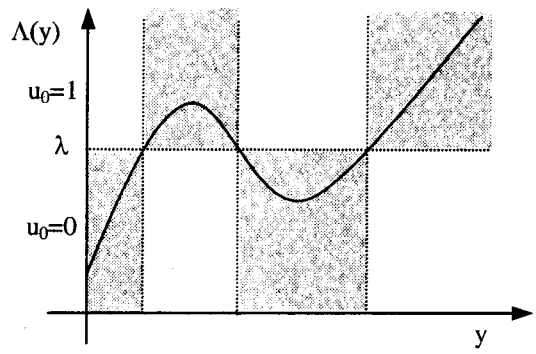


Figure 3 : u_0 est une variable plus grosse que $\Lambda(y)$, elle-même plus grosse que y.

3.4. Cas de la détection décentralisée parallèle

Dans le cadre de la détection décentralisée parallèle, à partir de ses propres observations y_i, chaque détecteur prend une décision locale u_i. La règle de décision au niveau de chaque détecteur est alors la suivante :

$$\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i \quad i=1,2$$

Les décisions locales sont ensuite transmises à un opérateur de fusion qui les combine de façon à obtenir la décision finale u_0. L'opérateur de fusion f qui optimise, suivant un critère donné, la mise en commun des décisions prises par les différents détecteurs locaux est une fonction logique des N décisions binaires qui lui sont transmises.

Dans le cadre de la phase d'apprentissage, chaque décision u_i peut être considérée comme une variable plus grosse que $\Lambda(y_i)$ (i=1,...,N), dû au fait qu'elle engendre une partition en deux

classes de la population d'apprentissage plus grosse que la partition engendrée par $\Lambda(y_i)$. De même, $\Lambda(y_i)$ représente une variable plus grosse que y_i : nous obtenons ainsi la relation suivante : $u_i \geq \Lambda(y_i) \geq y_i$. La décision finale u_0 est, quant à elle, une variable résultant des décisions locales $(u_1,\ldots,u_N)$. Nous avons donc la relation suivante : $u_0 \geq (u_1,\ldots,u_N)$.
En combinant ces relations, nous obtenons finalement :

$$u_0 \geq (u_1,\ldots,u_N) \geq (\Lambda(y_1),\ldots,\Lambda(y_N)) \geq (y_1,\ldots,y_N)$$

3.5. Cas de la détection décentralisée série avec N=2 (2 capteurs)

Dans le cadre de la détection décentralisée série, chaque détecteur local reçoit une information issue d'un capteur et transmet un message binaire à son successeur. La décision du premier détecteur est basée sur les informations issues d'un seul capteur et c'est le dernier détecteur qui élabore la décision finale. La règle de décision au niveau du premier détecteur est :

$$\Lambda(y_1) \underset{u_1=0}{\overset{u_1=1}{\gtrless}} \lambda_1$$

La règle de décision au niveau du deuxième détecteur est :

$$\Lambda(y_2) \underset{u_2=0}{\overset{u_2=1}{\gtrless}} \lambda_2^0 \text{ si } u_1=0$$

$$\Lambda(y_2) \underset{u_2=0}{\overset{u_2=1}{\gtrless}} \lambda_2^1 \text{ si } u_1=1$$

Dans ce cas u_1 peut être considérée comme une variable plus grosse que la variable $\Lambda(y_1)$ en deux classes, elle-même plus grosse que y_1 : $u_1 \geq \Lambda(y_1) \geq y_1$. La décision finale u_2 peut être considérée comme une variable plus grosse que la variable vectorielle $(u_1,\Lambda(y_2))$ en deux classes. De même, $\Lambda(y_2)$ sera considérée comme une variable plus grosse que y_2. En combinant ces résultats, nous obtenons finalement :

$$u_2 \geq (u_1,\Lambda(y_2)) \geq (\Lambda(y_1),\Lambda(y_2)) \geq (y_1,y_2)$$

4. La sélection de capteurs

4.1. Fondements de notre approche

Dans le deuxième chapitre, nous avons montré que l'optimisation d'un système de détection pouvait être effectuée en utilisant un critère entropique. Que l'on se place dans le cadre de la détection centralisée ou dans celui de la détection décentralisée, la quantité qu'il convient de minimiser est l'entropie conditionnelle $H(H/u_0)$ où u_0 est la décision finale et H est l'hypothèse vraie. Comme nous l'avons vu dans le chapitre 4, cette grandeur peut être interprétée comme l'incertitude relative à H, étant connue l'information fournie par u_0.

En détection centralisée, nous avons vu que la décision finale u_0 représentait une variable plus grosse que $\Lambda(y_1,\ldots,y_N)$ et que $\Lambda(y_1,\ldots,y_N)$ était elle-même plus grosse que $(y_1,\ldots,y_N)$ ($u_0 \geq \Lambda(y_1,\ldots,y_N) \geq (y_1,\ldots,y_N)$). Par conséquent, en vertu des propriétés de l'entropie conditionnelle développées dans le chapitre 4, nous obtenons les relations suivantes :
$H(H/u_0) \leq H(H) \leq \log_2 2 = 1$ car l'entropie d'une variable ayant n modalités est inférieure ou égale à $\log_2(n)$.
Nous obtenons ainsi les relations suivantes :

$$1 \geq H(H/u_0) \geq H(H/\Lambda(y_1,\ldots,y_N)) \geq H(H/y_1,\ldots,y_N)$$

En détection décentralisée parallèle, nous avons vu que :
$u_0 \geq (u_1,\ldots,u_N) \geq (\Lambda(y_1),\ldots,\Lambda(y_N)) \geq (y_1,\ldots,y_N)$.
On en déduit de même que :

$$1 \geq H(H/u_0) \geq H(H/u_1,\ldots,u_N) \geq H(H/\Lambda(y_1),\ldots,\Lambda(y_N)) \geq H(H/y_1,\ldots,y_N)$$

En détection décentralisée série avec N=2, nous avons vu que :
$u_2 \geq (u_1,\Lambda(y_2)) \geq (\Lambda(y_1),\Lambda(y_2)) \geq (y_1,y_2)$
On en déduit donc que :

$$1 \geq H(H/u_2) \geq H(H/\, u_1,\Lambda(y_2)) \geq H(H/\Lambda(y_1),\Lambda(y_2)) \geq H(H/y_1,y_2)$$

Quelle que soit l'architecture étudiée, nous pouvons remarquer que l'on a toujours :

$$1 \geq H(H/u_0) \geq H(H/y_1,\ldots,y_N)$$

(voir le chapitre 4)

D'autre part, on sait d'après les propriétés de l'entropie conditionnelle que :
S_1 et S_2 étant des variables vectorielles, si $S_2 \subseteq S_1$ alors $H(H/S_1) \leq H(H/S_2)$.

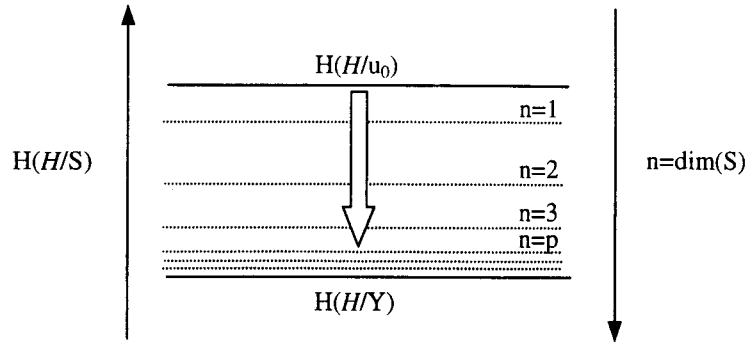


Figure 4 : Convergence de H(H/S) vers H(H/Y) lorsque la dimension de S augmente

Notre démarche de sélection de capteurs est issue de ces remarques.
Parmi les N capteurs disponibles, certains peuvent apporter de l'information au processus de détection ; les autres n'apportant qu'une information redondante (ou pas d'information du tout). Il n'est par conséquent pas utile de les prendre en compte dans le processus de détection. Dans le cadre de la théorie de l'information, cette remarque se traduit par le fait que la quantité $H(H/y_1,\ldots y_p)$ peut être très proche de $H(H/y_1,\ldots y_N)$ qui minore $H(H/u_0)$ (Figure 4). La prise en compte de p capteurs (p fixé par l'utilisateur du système) représente ainsi une condition de sortie des algorithmes.
Cette démarche se justifie par le fait que certains capteurs peuvent être redondants pour le problème de la détection ou peuvent noyer le système sous un flot d'informations trop coûteux à gérer.
L'optimisation du système de détection en minimisant l'entropie conditionnelle $H(H/u_0)$ aura alors toutes les chances d'aboutir à un système pratiquement aussi performant qu'en utilisant l'ensemble des capteurs. Etant donné que la complexité de l'optimisation des systèmes de détection croît très rapidement avec le nombre de capteurs, on aura de cette façon un bon compromis entre les performances et la complexité du système à optimiser.

L'utilisateur du système doit ainsi connaître le nombre de capteurs à sélectionner (le nombre « p »), ce qui n'est pas toujours évident *a priori*. Une autre démarche consisterait alors à demander à l'utilisateur une entropie minimale « seuil » en dessous de laquelle la quantité d'information apportée n'aurait pas une grande signification. Cette entropie conditionnelle minimale correspondrait en fait à l'erreur que s'autorise l'utilisateur en ne prenant en compte que p capteurs au lieu de N.

Le rôle de notre phase de sélection est de déterminer la configuration qui permettra de tirer le meilleur parti de toutes les informations $y=(y_1,\ldots,y_N)$ dont dispose le système. Il faudra alors déterminer le sous-ensemble $S\in P^v(y)$ (où $P^v(y)$ représente tous les vecteurs (toutes les variables multidimensionnelles) possibles dont les composantes sont des composantes élémentaires de y) de capteurs tel que l'incertitude relative à H connaissant l'information fournie par S, c'est-à-dire $H(H/S)$, se rapproche le plus de la quantité $H(H/y_1,\ldots,y_N)$. Dans la pratique, le calcul exhaustif de $H(H/S)$ pour tout $S\in P^v(y)$ est impossible car on se heurte à une explosion combinatoire lorsque le nombre de capteurs devient grand. On construit donc S en utilisant des démarches heuristiques, démarches qui sont ici inspirées des algorithmes de classification par arbres de décision, que nous allons maintenant développer.

4.2. Les méthodes de sélection

Notre but est de sélectionner l'ensemble S des capteurs les plus pertinents pour la détection. Les méthodes développées doivent être simples de façon à pouvoir les appliquer à des systèmes comportant un nombre élevé de capteurs.
Le critère étant défini (on cherchera à minimiser H(H/S)), nous utilisons ici deux approches : une approche agrégative des variables, ainsi qu'une approche désagrégative.

4.2.1. Approche agrégative

• Présentation de la méthode

Dans un premier temps, l'algorithme recherchera le capteur y_i qui minimise la quantité H(H/y_i). Puis il recherchera le capteur y_j qui, associé au capteur y_i initialement trouvé, minimise la quantité H(H/y_i,y_j) ; puis un troisième, ..., ainsi de suite jusqu'à ce que l'on ait sélectionné un ensemble de p capteurs (Figure 5). On voit ici que l'on agrège petit à petit les capteurs apportant de l'information sur l'hypothèse vraie, de façon à déterminer S (d'où le nom de la méthode).

La méthode consiste donc à :

$$\min_{y_i \in (y \setminus S_{k-1})} (H(H/(S_{k-1}, y_i)))$$

avec $S_0 = \varnothing$ et $S_k = (S_{k-1}, y_i)$

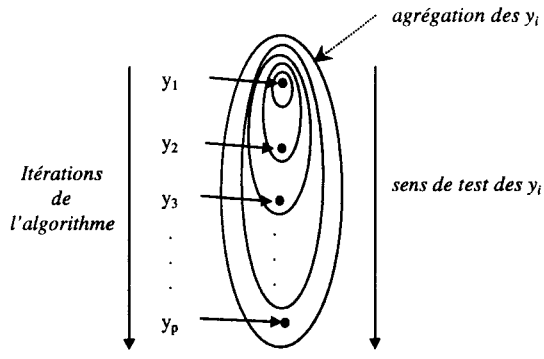


Figure 5 : L'approche agrégative moyennant une renumérotation des variables testées. La première variable trouvée par l'algorithme est appelée y_1, la deuxième, y_2, ...

• **Algorithme proposé**

; APPROCHE AGREGATIVE DE SELECTION DE CAPTEURS
; phase d'initialisation
0. « p » et/ou « H_{min} » sont des paramètres fixés par l'utilisateur du système
1. pour i := 1 à N (N = nombre de capteurs)
2. calculer les $H(H/y_i)$
3. fin pour
4. retenir y_i qui minimise $H(H/y_i)$
5. $S \leftarrow (y_i)$

; boucle
6. répéter
7. calculer les $H(H/(S,y_i))$ avec $y_i \notin S$
8. retenir y_i qui minimise $H(H/(S,y_i))$
9. $S \leftarrow (S,y_i)$
10. jusqu'à [card S = p] ou $[H(H/(S,y_i))<H_{min}]$

; FIN de l'algorithme.

où $S \in P^v(y)$.

• **Remarques**

- Nous avons toujours la relation suivante $H(H/y_1) \geq H(H/y_1,y_2) \geq \ldots \geq H(H/y_1,\ldots,y_N)$; et par conséquent, l'entropie conditionnelle est décroissante, minorée par 0. L'algorithme va donc nécessairement converger (Figure 6). Une autre condition d'arrêt lors de la construction de S pourrait donc être :
$H(H/S_k)-H(H/S_{k-1})<s$ où s est un seuil fixé par l'utilisateur.
Ce seuil représente la quantité d'information à prendre en compte pour passer de k-1 capteurs à k capteurs. Lorsque cette quantité est négligeable (le kième capteur apporte peu d'information sur l'hypothèse vraie), il n'est pas forcément nécessaire de prendre en compte ce capteur dans le processus de décision, sous peine de compliquer considérablement les calculs d'optimisation du chapitre 1 ou 2.

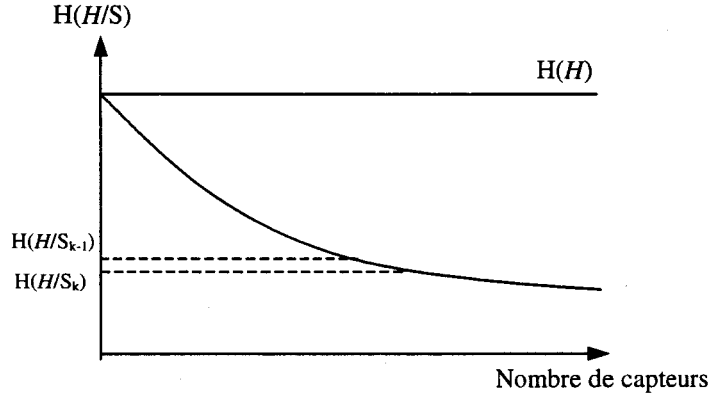


Figure 6 : Convergence du critère de sélection en fonction du nombre de capteurs

- Dans certains cas, la quantité $H(H/(y_i,y_j))$ peut être petite alors que les quantités $H(H/y_i)$ et $H(H/y_j)$ prises séparément peuvent être grandes. Le couple de capteurs (y_i,y_j) apporte alors beaucoup d'informations sur l'hypothèse vraie, alors que les capteurs y_i et y_j pris séparément n'apportent aucune information sur celle-ci. Dans ce cas, l'algorithme agrégatif sélectionnera à tort des capteurs qui ne permettront pas de minimiser l'entropie conditionnelle. Pour remédier à ce problème, nous proposons de construire S en utilisant une démarche désagrégative.

4.2.2. Approche désagrégative

• Présentation de la méthode

Dans un premier temps, au lieu de chercher directement le capteur y_i qui minimise $H(H/y_i)$ pour construire S, on va écarter de y le capteur y_i qui minimise $H(H/y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_N)$; ce qui signifie que ce capteur n'apporte aucune information sur l'hypothèse vraie par rapport à $(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_N)$. L'ensemble résiduel des capteurs les plus pertinents est ainsi constitué des capteurs composant y auquel on a enlevé la composante y_i.
Dans un deuxième temps, on réitère le même processus sur l'ensemble résiduel. On écarte alors y_j, une variable n'apportant que peu (ou pas) d'information sur l'hypothèse vraie ; ainsi de suite jusqu'à ce que l'on ait un ensemble résiduel composé de p capteurs qui représentent les p capteurs les plus pertinents pour le problème de la détection.

Il s'agit donc d'une méthode désagrégative car on évalue le critère avec N-1 capteurs (Figure 7) puis N-2, etc... La méthode consiste alors à :

$$\min_{y_i \in S_{k-1}} (H(H/(S_{k-1} \setminus y_i)))$$

$$S_0 = y$$

$$S_k = S_{k-1} \setminus y_i$$

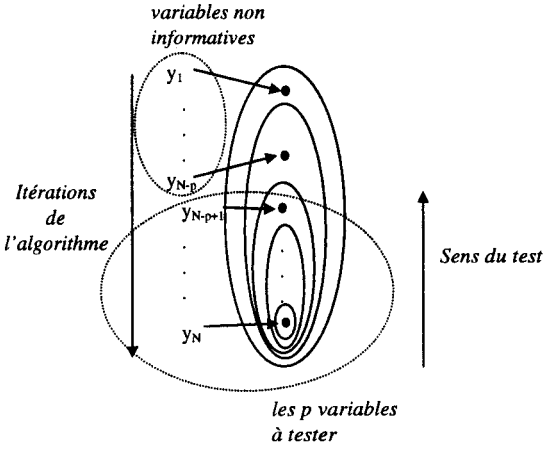


Figure 7 : L'approche désagrégative moyennant une renumérotation des variables testées. La première variable trouvée par l'algorithme est appelée y_1, la deuxième, y_2, ...

La même réflexion peut être menée en considérant une entropie conditionnelle, cette fois-ci, maximale, qui représenterait en fait un seuil de tolérance qu'il ne faudrait pas dépasser lors de la sélection de capteur. Ce seuil pourrait être fixé par l'utilisateur.

• **Algorithme proposé**

; APPROCHE DESAGREGATIVE DE SELECTION DE CAPTEURS
; initialisation
0. « p » et/ou « H_{max} » sont des paramètres fixés par l'utilisateur du système
1. pour i := 1 à N (N = nombre de capteurs)
2. calculer les $H(H/(y \backslash y_i))$
3. fin pour
4. déterminer y_i qui minimise $H(H/(y \backslash y_i))$
5. $S \leftarrow y \backslash y_i$
; boucle
6. répéter
7. calculer les $H(H/(S \backslash y_i))$ avec $y_i \in S$
8. déterminer y_i qui minimise $H(H/(S \backslash y_i))$
9. $S \leftarrow S \backslash y_i$
10. jusqu'à [card S = p] ou $[H(H/(S \backslash y_i)) > H_{max}]$
; FIN de l'algorithme.

avec $y_i \in y$ et $S \in P^v(y)$.

Cette démarche se justifie sur la précédente par le fait qu'il peut par exemple arriver qu'un capteur apporte plus d'information sur H à lui seul que tous les autres capteurs pris un à un, alors que ces mêmes capteurs pris dans leur ensemble annulent quasiment l'incertitude sur H. Pour illustrer ce phénomène, on peut considérer l'exemple suivant.

• Exemple 1

On dispose d'un système comprenant trois capteurs y_1,y_2 et y_3. Il s'agit de détecter l'hypothèse qui représente en réalité le bit de parité associé aux moyennes de y_1 et y_2. Les fonctions de densité de probabilités sous chaque hypothèse associées à ces trois capteurs sont des gaussiennes. Sous chaque hypothèse H_0 et H_1, ces fonctions de densité de probabilités sont supposées être de moyenne 0 ou 1 et de variance 1 telles que :

Moyenne de y_1	Moyenne de y_2	Moyenne de y_3	H
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

Après calcul, on remarque que le capteur y_3 va permettre de minimiser $H(H/y_i)$ (i=1,2,3). Mais c'est le couple de capteurs (y_1,y_2) qui minimise $H(H/y_i,y_j)$ (i,j=1,2,3). Si l'on désire sélectionner deux capteurs parmi les trois disponibles dans le but de bien détecter, l'algorithme agrégatif conduira à choisir y_3 comme étant le capteur qui minimise $H(H/y_i)$, puis y_1 ou y_2 qui minimiserons $H(H/y_3,y_j)$ (j=1,2). Par contre l'algorithme désagrégatif va écarter le capteur y_3 et va permettre de sélectionner les capteurs y_1 et y_2 car ce sont eux qui minimisent $H(H/y_i,y_j)$ (i,j=1,2,3).

Les résultats donnent :

Algorithme	Capteurs sélectionnés
Agrégatif	X_3 X_1
Désagrégatif	X_1 X_2

Dans ce cas, l'algorithme désagrégatif donne de meilleurs résultats que l'algorithme agrégatif.

• Exemple 2

Considérons maintenant un système comprenant trois capteurs. Les fonctions de densité de probabilités conditionnelles au niveau de chaque détecteur sont supposées être des gaussiennes. Sous l'hypothèse H_0 ces fonctions sont de moyenne 0 et de variance 1. Sous l'hypothèse H_1 ces fonctions sont supposées être de moyenne m_1,m_2 et m_3 et de variance 1. Le but est de sélectionner le couple de capteurs qui permettra de faire la meilleure détection sans utiliser les trois capteurs disponibles. A titre d'exemple, nous avons posé $m_1=1$ $m_2=1.5$ $m_3=2$. Sur la Figure 8, on peut visualiser les variations de $H(H)$,$H(H/y_1)$, $H(H/y_2)$, $H(H/y_3)$, $H(H/y_1,y_2)$, $H(H/y_1,y_3)$, et $H(H/y_2,y_3)$ en fonction de P_0. On constate que c'est le capteur y_3 qui permet de minimiser $H(H/y_i)$ (i=1,2,3) et que c'est le couple (y_2,y_3) qui minimise $H(H/y_i,y_j)$ (i,j=1,2,3) et ceci, quelque soit la valeur prise par P_0. Ces résultats sont tout à fait conformes à ce que l'on pouvait attendre par la seule connaissance des fonctions de probabilités associées aux différents capteurs.

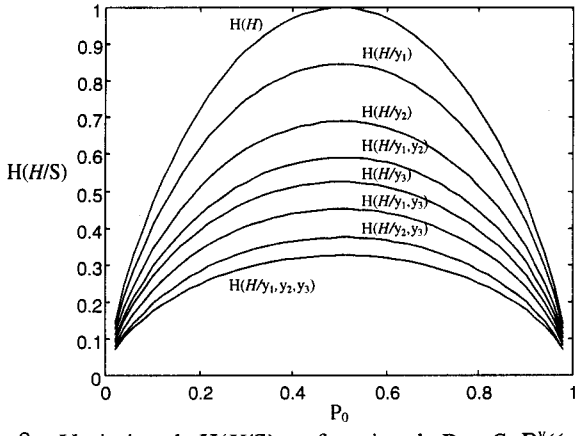


Figure 8 : Variation de H(*H*/S) en fonction de P_0 - $S \in P^v((y_1,y_2,y_3))$

Sur cet exemple, que l'on utilise la méthode agrégative ou désagrégative, ce sera le couple de capteurs (y_2,y_3) qui sera sélectionné en vu de la détection.

5. Détection décentralisée parallèle par arbre de décision

Dans le paragraphe 3 du chapitre 2, nous avons montré que les systèmes de détection décentralisée parallèle pouvaient être optimisés en utilisant un critère entropique. Deux approches ont été étudiées :

- une approche dans laquelle nous considérons le problème de l'optimisation des différents détecteurs locaux sans prendre en compte l'opérateur de fusion. Les détecteurs locaux étant fixés, nous optimisons ensuite l'opérateur de fusion.
- une approche dans laquelle nous considérons l'optimisation simultanée des détecteurs locaux et de l'opérateur de fusion.

Dans le premier cas, l'optimisation des détecteurs locaux consiste à déterminer les seuils associés à chaque détecteur qui permettent de minimiser l'entropie conditionnelle $H(H/u_1,\ldots,u_N)$. Ici, nous proposons d'utiliser les propriétés particulières de l'entropie afin de limiter la complexité des calculs à mettre en œuvre lors de la détermination de ces seuils. Ces seuils étant déterminés, nous proposons de construire un arbre de décision qui minimise la probabilité d'erreur du système.

5.1. Détermination des seuils au niveau des détecteurs locaux

Nous proposons ici un algorithme de type agrégatif qui nous permettra de déterminer les seuils au niveau des détecteurs locaux en minimisant l'entropie conditionnelle $H(H/u_1,\ldots,u_N)$. Cet algorithme nous permettra de nous approcher du système de détection optimal, son principal intérêt étant d'être très peu coûteux en terme de calculs. Etant donné sa simplicité de mise en œuvre, cet algorithme pourra être appliqué à des systèmes comportant de nombreux capteurs.

• Présentation

On cherche ici à trouver le capteur y_i et le seuil λ_i qui minimisent $H(H/u_i)$ tels que $\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i$, puis le capteur y_j (et le seuil λ_j) qui, associé à u_i, minimise $H(H/u_i,u_j)$ et ainsi de suite, ..., jusqu'à ce que l'on ait déterminé chaque seuil. Il s'agit d'une méthode agrégative car on évalue le critère avec la décision prise par un détecteur, puis deux détecteurs, etc... La méthode est alors la suivante :

$$\min_{y_i \in (y \setminus S_{k-1}), \lambda_i \in M_{\Lambda(y_i)}} (H(H/(S_{k-1}, u_i)))$$

$$\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i$$

$$S_0 = \varnothing$$

$$S_k = (S_{k-1}, u_i)$$

où $M_{\Lambda(yi)}$ est l'ensemble des valeurs prises par $\Lambda(y_i)$.

• **Algorithme proposé**

; *APPROCHE AGREGATIVE DE DETERMINATION DES SEUILS LOCAUX*
; *initialisation*
1. pour i := 1 à N (N = nombre de capteurs)
2. calculer les $H(H/u_i)$ pour tous les $\lambda_i \in M_{\Lambda(yi)}$; *voir remarque 2.*
3. retenir u_i et λ_i qui minimisent $H(H/u_i)$
4. fin pour
5. $S \leftarrow (u_i)$

; *boucle*
6. répéter
7. calculer les $H(H/(S,u_i))$ pour chaque $\lambda_i \in M_{\Lambda(yi)}$ avec $u_i \notin S$; *voir remarque 2.*
8. retenir u_i et λ_i qui minimisent $H(H/(S,u_i))$
9. $S \leftarrow (S,u_i)$
10. jusqu'à (card S = N)

; *FIN de l'algorithme.*

Remarque 1 :

Par cette approche, les seuils obtenus pour chaque détecteur dépendent bien entendu de l'ordre de traitement des variables en question. On pourra de même utiliser une démarche de sélection de variables via un algorithme de sélection désagrégatif, qui nous donnera l'ordre des variables à tester. Il nous suffira alors de déterminer les seuils optimaux pour chaque détecteur. Nous obtenons l'algorithme suivant :

; APPROCHE DESAGREGATIVE DE DETERMINATION DES SEUILS LOCAUX
1. Utilisation de l'algorithme désagrégatif de sélection des variables
   $\Rightarrow$ y_N est la 1ère variable à tester, y_{N-1}, la deuxième, ..., y_1, la dernière
2. Calculer les $H(H/u_N)$ pour tous les $\lambda_N \in M_{\Lambda(yN)}$; *voir remarque 2.*
3. Retenir λ_N qui minimise $H(H/u_N)$
4. $S \leftarrow (u_N)$

; *boucle*
5. pour i :=N-1 à 1 faire
6. calculer les $H(H/(S,u_i))$ pour chaque $\lambda_i \in M_{\Lambda(yi)}$; *voir remarque 2.*
7. retenir λ_i qui minimise $H(H/(S,u_i))$
8. $S \leftarrow (S,u_i)$
9. fin pour

; *FIN de l'algorithme.*

Remarque 2 :

Si le nombre de seuils possibles pour $\Lambda(y_i)$ est important, une procédure heuristique (comme par exemple la méthode du gradient) pourra être utilisée afin de limiter le temps de calcul.

Remarque 3 :

Les seuils λ_i étant définis pour l'algorithme proposé, on pourrait proposer une solution permettant un réajustement de ceux-ci par l'intermédiaire d'une méthode du gradient qui peut être facilement implémentée *via* une méthode « axe par axe » de la façon suivante :

On fait varier le seuil λ_1 (les autres λ_i sont constants) de façon à obtenir un minimum local de l'entropie conditionnelle ; puis on réitère l'opération sur $\lambda_2 \ldots \lambda_N$; ... et on réitère le processus jusqu'à ce que les seuils n'évoluent plus.

Les seuils ayant été déterminés, on pourra alors construire l'arbre de décision qui permet de minimiser la probabilité d'erreur du système de détection.

5.2. Construction de l'arbre de décision

Les seuils ayant été déterminés au moyen de l'algorithme précédent, nous proposons de construire un arbre de décision binaire (Figure 9). Chaque nœud de l'arbre contiendra un test sur une variable y_i (i=1,...N). Ce test consistera en un seuillage du rapport de vraisemblance $\Lambda(y_i) \underset{u_i=0}{\overset{u_i=1}{\gtrless}} \lambda_i$ et donnera naissance à deux branches, une pour chaque valeur prise par le détecteur local (u_i=0 ou 1). Chaque feuille contiendra la valeur de la décision finale u_0 (H_0 ou H_1). Cette décision pourra être déterminée en minimisant la probabilité d'erreur ce qui reviendra à affecter à chaque feuille une décision telle que :

$$p(\text{chemin},H_1) \underset{u_0=H_0}{\overset{u_0=H_1}{\gtrless}} p(\text{chemin},H_0)$$

le chemin étant la concaténation des différents valeurs prises par les décisions locales pour arriver à la feuille considérée (par exemple u_2=0 et u_1=0 pour la feuille de gauche de la Figure 9).

• Exemple

Reprenons un des exemples développés dans les premier et deuxième chapitres : Considérons un système formé de deux capteurs dont les observations y_1 et y_2 sont indépendantes. Les fonctions de densité de probabilités sous chaque hypothèse sont des gaussiennes Sous l'hypothèse H_0, ces fonctions de densité de probabilités sont supposées être de moyenne 0 et de variance 1. Sous l'hypothèse H_1, ces fonctions sont supposées être de moyenne m_1 (resp. m_2) et de variance 1. Deux systèmes sont considérés, le premier pour lequel $m_1=m_2=1$ et le second pour lequel $m_1=1$ et $m_2=1,5$.

Les valeurs des seuils λ_1 et λ_2 optimaux trouvées dans le deuxième chapitre et qui minimisent $H(H/u_1,u_2)$ sont indiqués sur les figures 14 et 20. Sur les figures 15 et 21, les variations de $H(H/u_1,u_2)$ correspondant à ces seuils en fonction de P_0 sont visualisées. Sur les figures 13 et 19, la variation de la probabilité d'erreur en fonction de P_0 et en fonction de la fonction de fusion utilisée est représentée.

Pour P_0=0.5 et dans le cas ou m_1=1 et m_2=1.5, à la première itération l'algorithme agrégatif nous indique que c'est le capteur y_2 avec le seuil λ_2=1 qui minimise l'entropie conditionnelle $H(H/u_2)$, à la deuxième itération il nous indique que c'est le capteur y_1 avec le seuil λ_1=1 qui minimise l'entropie conditionnelle $H(H/u_2,u_1)$. L'arbre de décision qui minimise la probabilité d'erreur du système est représenté sur la figure 9. Sur cet arbre, on remarque que seule la mesure de y_2 conditionne la décision finale. On retrouve le fait que la fonction de fusion se limite à la décision prise au niveau du capteur 2.

Les valeurs des seuils λ_1 et λ_2 trouvés avec l'algorithme agrégatif sont indiquées sur les figures 10 et 16. Sur les figures 11 et 17, les variations de $H(H/u_1,u_2)$ correspondant à ces seuils en fonction de P_0 sont visualisées. Sur les figures 12 et 18, la variation de la probabilité d'erreur en fonction de P_0 correspondant à l'arbre de décision trouvé avec l'algorithme est représentée.

Sur ces deux exemples, on peut remarquer que les résultats trouvés avec l'algorithme agrégatif présenté dans ce paragraphe sont très proche des résultats optimaux trouvés dans le chapitre 2. Ces exemples simples montrent le bien fondé de notre démarche.

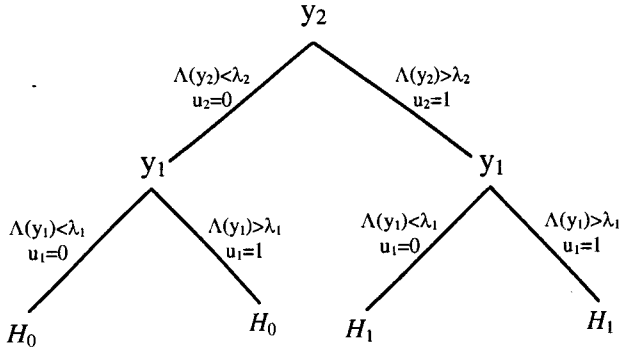


Figure 9 : Arbre de décision trouvé par l'algorithme agrégatif pour P_0=0.5 pour m_1=1, m_2=1.5

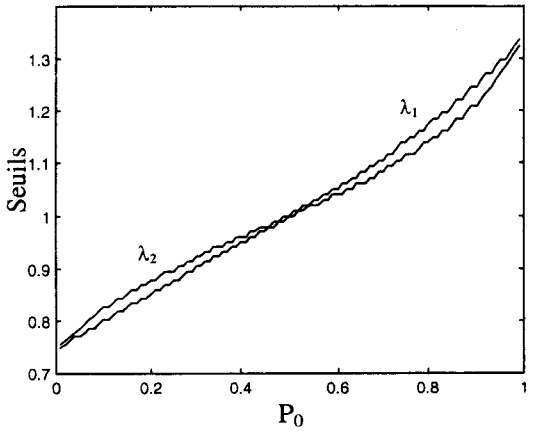


Figure 10 : Seuils trouvés avec l'algorithme agrégatif en fonction de P_0, m_1=m_2=1.

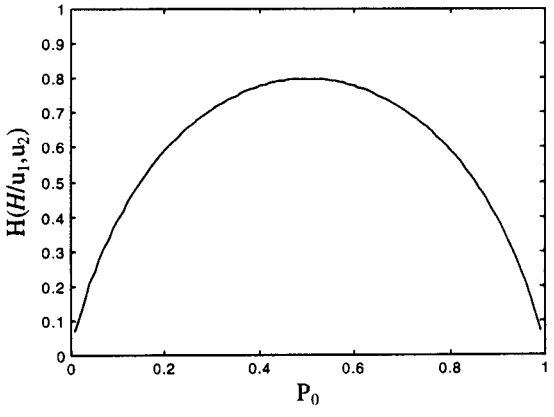


Figure 11 : $H(H/u_1,u_2)$ correspondant aux seuils trouvés avec l'algorithme agrégatif en fonction de P_0, m_1=m_2=1.

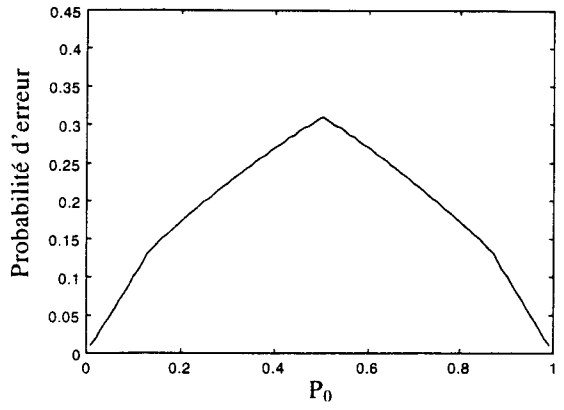


Figure 12 : Probabilité d'erreur en fonction de P_0 correspondant à l'arbre de décision trouvé avec l'algorithme agrégatif, $m_1=m_2=1$.

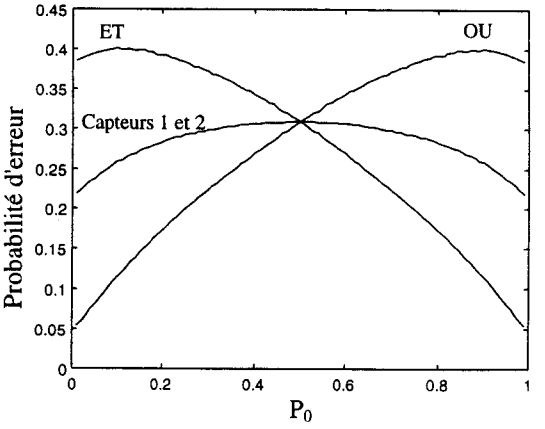


Figure 13 : Probabilité d'erreur en fonction de P_0 correspondant aux seuils optimaux, $m_1=m_2=1$.

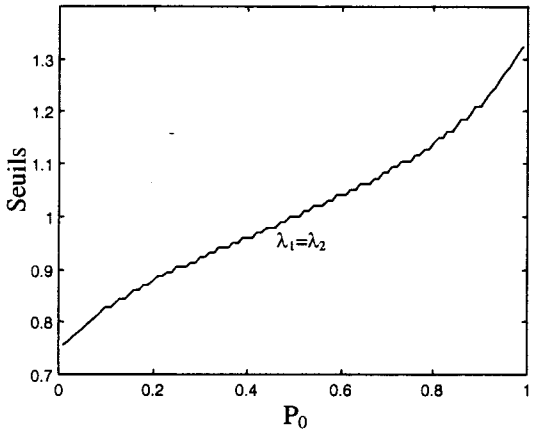


Figure 14 : Seuils optimaux en fonction de P_0, $m_1=m_2=1$.

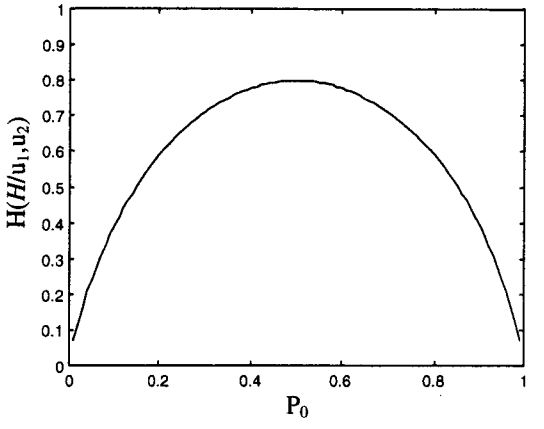


Figure 15 : $H(H/u_1,u_2)$ correspondant aux seuils optimaux en fonction de P_0, $m_1=m_2=1$.

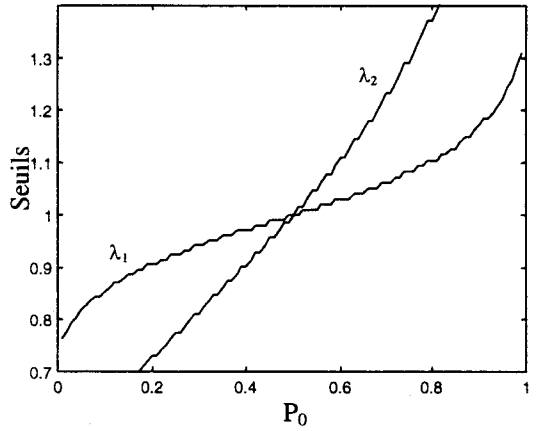


Figure 16 : Seuils trouvés avec l'algorithme agrégatif en fonction de P_0, $m_1=1$, $m_2=1.5$.

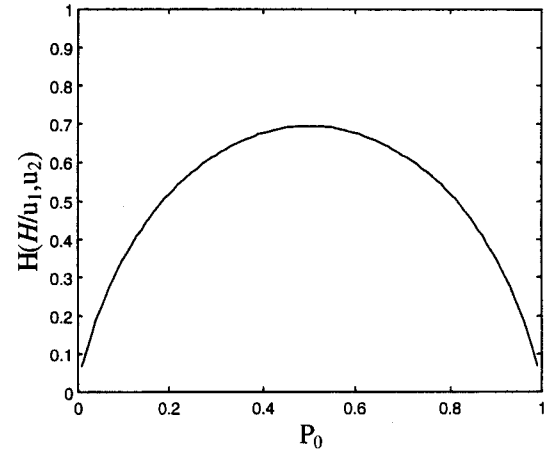


Figure 17 : $H(H/u_1,u_2)$ correspondant aux seuils trouvés avec l'algorithme agrégatif en fonction de P_0, $m_1=1$, $m_2=1.5$.

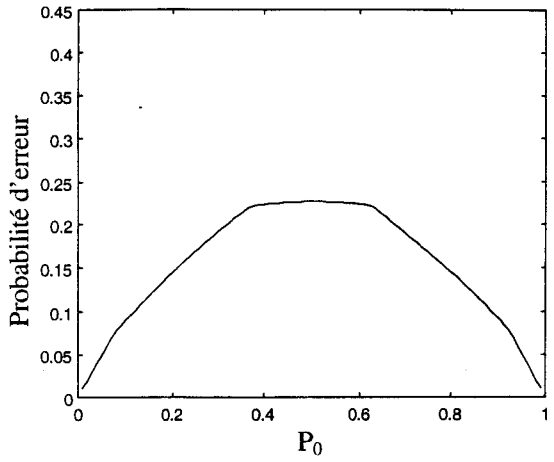


Figure 18 : Probabilité d'erreur en fonction de P_0 correspondant à l'arbre de décision trouvé avec l'algorithme agrégatif, m_1=1, m_2=1,5.

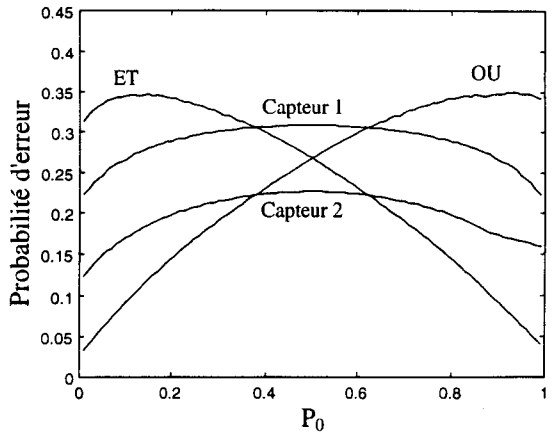


Figure 19 : Probabilité d'erreur en fonction de P_0 correspondant aux seuils optimaux, m_1=1, m_2=1,5.

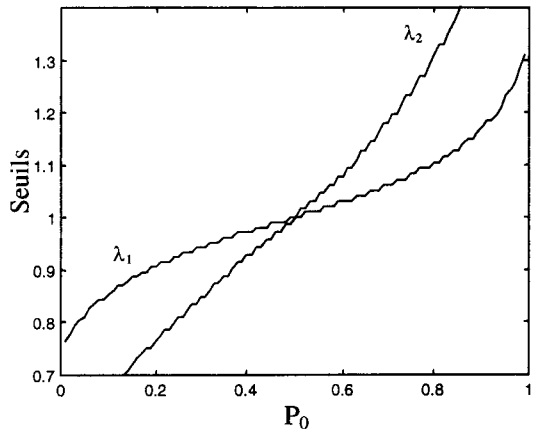


Figure 20 : Seuils optimaux en fonction de P_0, m_1=1, m_2=1.5.

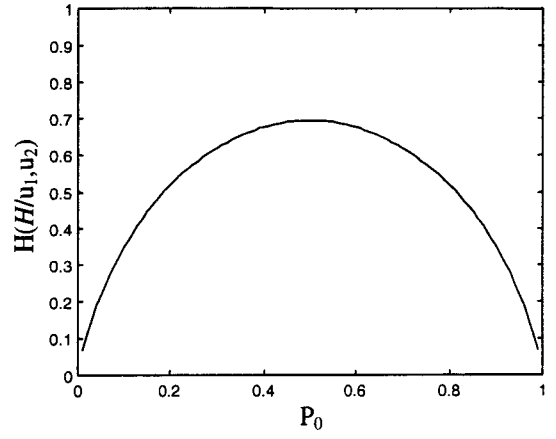


Figure 21 : $H(H/u_1,u_2)$ correspondant aux seuils optimaux en fonction de P_0, m_1=1, m_2=1.5.

6. Quantification répartie et détection par arbre de décision

Dans un système de détection, l'ensemble des observations y fournies par les capteurs constitue l'espace des observations, noté D. Le problème est alors de diviser l'espace des observations en deux domaines disjoints D_0 et D_1 tels que si l'observation tombe dans D_0 (respectivement dans D_1) la décision prise est H_0 (respectivement H_1). Une partition de l'espace des observations D en deux classes est alors obtenue. La détection centralisée permet de diviser ce domaine de façon optimale suivant un critère donné (Figure 22). Les méthodes décentralisées représentent un compromis entre les performances optimales mais coûteuses d'un système centralisé, et celles dégradées mais plus économiques des solutions décentralisées. Le but de ces méthodes est de s'approcher au mieux des domaines de décision optimaux trouvés dans le cas centralisé (Figure 23, Figure 24).
Cependant, en détection décentralisé on se limite à comparer le rapport de vraisemblance au niveau de chaque détecteur à un (détection décentralisée parallèle) voire deux seuils (deuxième détecteur en détection série). La partition de l'espace des observations qui en résulte ne peut donc être que relativement grossière, cette partition ne s'approchant que de façon éloignée de la partition optimale « centralisée ».

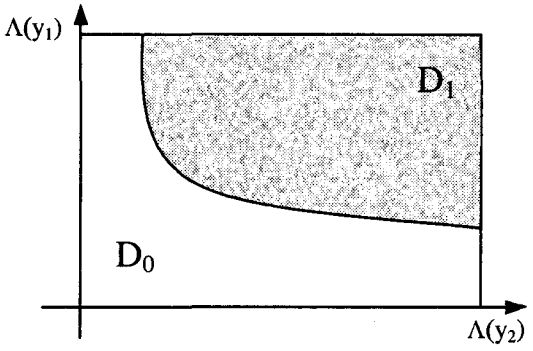


Figure 22 : Les domaines de décision en détection centralisée

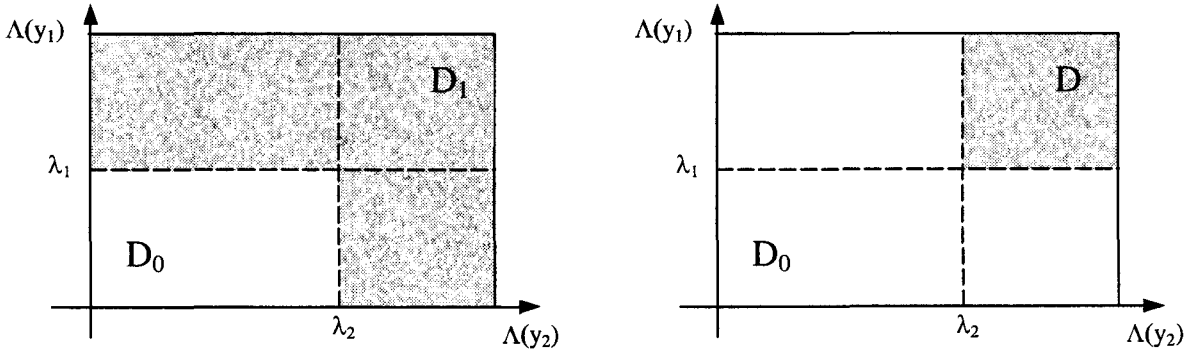


Figure 23 : Les domaines de décision en détection décentralisée parallèle suivant la fonction de fusion (OU ou ET)

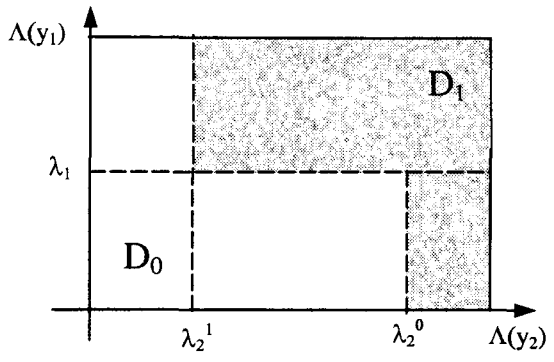


Figure 24 : Les domaines de décision en détection décentralisée série

Nous proposons d'augmenter le nombre de seuils au niveau de chaque détecteur, de façon à obtenir une partition plus fine de l'espace des observations (Figure 25). Pour déterminer les seuils qui nous permettrons de nous approcher de la solution optimale (détection centralisée), nous cherchons toujours à minimiser $H(H/u_1,...u_N)$, où $u_1,...,u_N$ ne représentent plus des variables plus grosses que $\Lambda(y_1),...,\Lambda(y_N)$ en deux classes, mais en un nombre de classes fixé à l'avance par l'utilisateur. Afin de trouver ce compromis, nous utilisons un algorithme agrégatif de construction d'arbre de décision du même type que celui développé précédemment, mais nous lui ajoutons la possibilité de partitionner une variable en utilisant plusieurs seuils, le critère ne changeant évidemment pas.

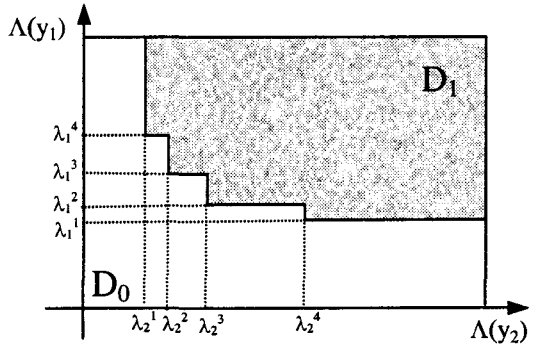


Figure 25 : Les domaines de décision en augmentant le nombre de seuils

6.1. Détermination des seuils au niveau des détecteurs locaux

• Présentation de la démarche

Nous commençons par chercher le capteur y_i et le seuil λ_i^1 qui minimisent $H(H/u_i^1)$, u_i^1 étant une variable plus grosse que $\Lambda(y_i)$ en deux classes telle que $\Lambda(y_i) \underset{<}{\overset{u_i^1=1}{>}} \lambda_i^1$. Puis on détermine le capteur y_j et le seuil λ_j^1 qui, associé à u_i^1, minimisent $H(H/u_i^1,u_j^1)$ ou, si il existe, le seuil λ_i^2 qui, en tenant compte du seuil λ_i^1 trouvé précédemment, minimisent $H(H/u_i^2)$, u_i^2 étant une variable plus grosse que $\Lambda(y_i)$ en trois classes, et ainsi de suite, ..., jusqu'à ce que l'on ait

déterminé le nombre de seuils « nbseuilmax(i) » désiré au niveau de chaque détecteur local. En seuillant au fur et à mesure les valeurs prises par les rapports de vraisemblance associés aux différents capteurs, on détermine une partition de plus en plus fine de l'espace des observations.

Les seuils ayant été déterminés, on pourra alors construire l'arbre de décision permettant de minimiser la probabilité d'erreur du système de détection.

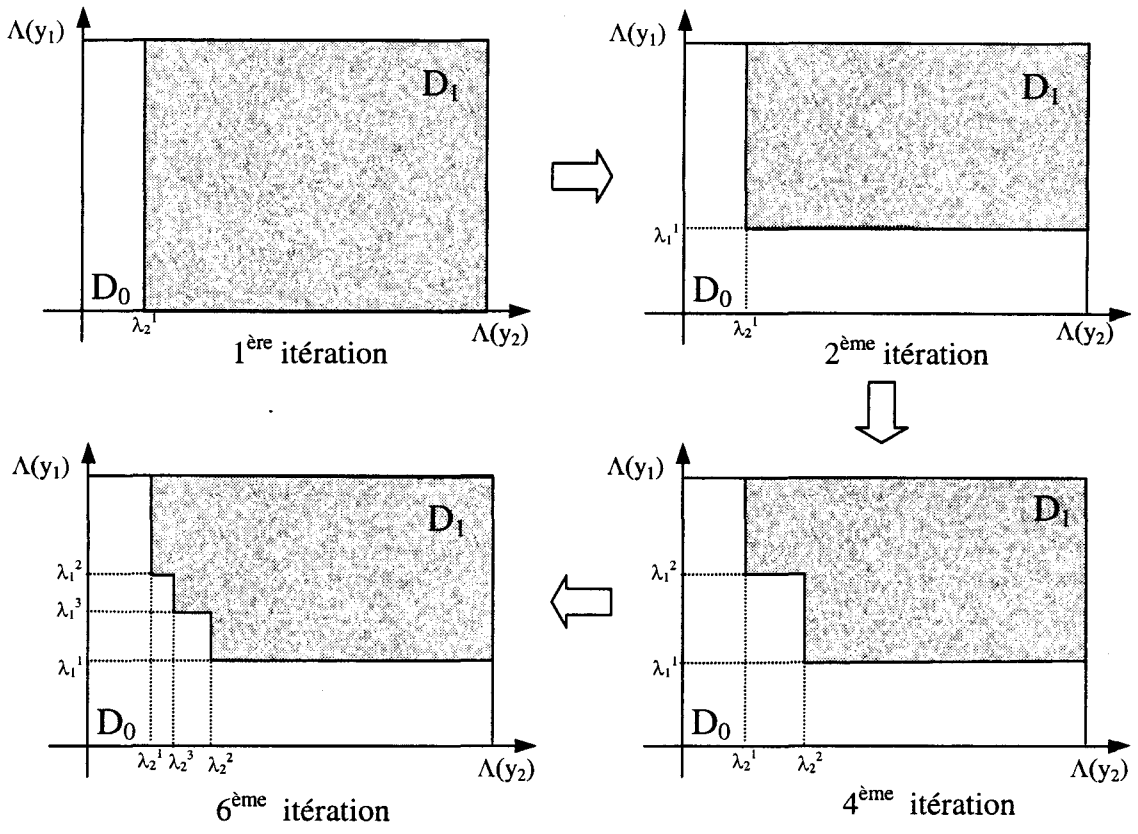


Figure 26 : A chaque itération de l'algorithme, on crée une partition de plus en plus fine de l'espace des observations

La même réflexion menée dans les paragraphes précédents peut être faite ici :
L'utilisateur du système doit connaître le nombre de seuils de chaque capteur ; c'est-à-dire qu'il doit connaître le nombre de bits à envoyer à l'opérateur de fusion (ex. : 3 seuils sur $y_1 \Rightarrow$ 2 bits ; 1 seuil sur $y_2 \Rightarrow$ 1 bit ; ...).
Une autre démarche consisterait à demander à l'utilisateur une entropie minimale « seuil » en dessous de laquelle, il considère que la quantité d'information apportée sur le système est suffisante.

• **Algorithme proposé**

; *APPROCHE AGREGATIVE DE DETERMINATION DES PARTITIONS LOCALES*
; *initialisation*
1. $\forall$ i=1,...,N « nbseuilmax(i) » connu (fixé par l'expert) et/ou « H_{min} » fixé par l'expert
; *boucle*
2. pour i :=1 à N
nbseuil(i) $\leftarrow$ 0
$\lambda_i^{nbseuil(i)} \leftarrow -\infty$
fin pour
; *boucle*
3. répéter
4. pour i := 1 à N (N = nombre de capteurs)
5. créer toutes les dichotomies de $\Lambda(y_i)$ possibles et calculer
H(H/ $\mathcal{P}$(nbseuil(1),nbseuil(2),...,nbseuil(i)+1,...,nbseuil(N)))
fin pour
6. retenir « y_i » et mémoriser $\lambda_i^{nbseuil(i)+1}$ qui minimisent
H(H/ $\mathcal{P}$(nbseuil(1),nbseuil(2),...,nbseuil(i)+1,...,nbseuil(N)))
7. si nbseuil(i)+1>nbseuilmax(i)
alors rejeter la dichotomie
sinon nbseuil(i) $\leftarrow$ nbseuil(i)+1
fin si
jusqu'à ($\forall$ i=1,...,N nbseuil(i)=nbseuilmax(i))
ou (H(H/$\mathcal{P}$(nbseuil(1),nbseuil(2),...,nbseuil(i)+1,...,nbseuil(N)))<H_{min})
; *FIN de l'algorithme.*

où $\mathcal{P}$(nbseuil(1),nbseuil(2),...,nbseuil(i),...,nbseuil(N)) = $\bigcap_{i=1}^{N}$ $\mathcal{P}$(nbseuil(i))

avec $\mathcal{P}$ (nbseuil(i)) représentant la partition du domaine d'observations engendrée par la partition en (i+1) classes de $\Lambda(y_i)$.

6.2. Construction de l'arbre de décision

Les seuils ayant été déterminés au moyen de l'algorithme précédent, nous proposons de construire un arbre de décision (Figure 27). Chaque nœud de l'arbre contiendra un test sur une variable y_i (i=1,...N). Ce test consistera en une partition du rapport de vraisemblance $\Lambda(y_i)$ en (nbseuilmax(i)+1) classes et donnera naissance à (nbseuilmax(i)+1) branches, une pour chaque valeur prise par le détecteur local (u_i=0 ou 1 ... ou (nbseuilmax(i)+1)). Chaque feuille contiendra la valeur de la décision finale u_0 (H_0 ou H_1). Cette décision pourra être déterminée en minimisant la probabilité d'erreur, ce qui reviendra à affecter à chaque feuille une décision telle que :

$$p(\text{chemin},H_1) \underset{u_0=H_0}{\overset{u_0=H_1}{\gtrless}} p(\text{chemin},H_0)$$

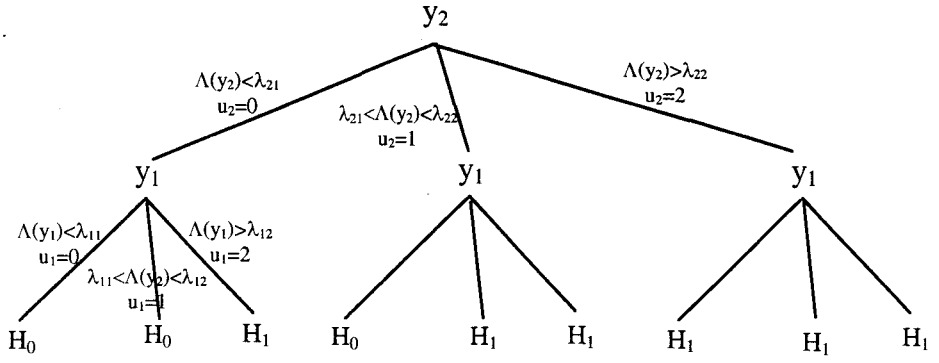


Figure 27 : Exemple d'arbre de décision construit en se limitant à 2 seuils par variable

• Exemple

Nous avons repris ici l'exemple du paragraphe précédent pour lequel $m_1=1, m_2=1,5$. Pour chaque capteur, nous avons limité le nombre de seuils à 3, de façon à pouvoir coder les informations transmises par les quantificateurs locaux à l'opérateur de fusion sur 2 bits. L'algorithme agrégatif permet de trouver les six seuils $\lambda_{11},\lambda_{12},\lambda_{13},\lambda_{21},\lambda_{22}$, et λ_{23} qui vont nous permettre de minimiser $H(H/u_1,u_2)$.

Avec $u_1=0$ si $\Lambda(y_1) < \lambda_{11}$ et $u_2=0$ si $\Lambda(y_2)<\lambda_{21}$
$u_1=1$ si $\lambda_{11} < \Lambda(y_1) < \lambda_{12}$ $u_2=1$ si $\lambda_{21} < \Lambda(y_2) < \lambda_{22}$
$u_1=2$ si $\lambda_{12} < \Lambda(y_1) < \lambda_{13}$ $u_2=2$ si $\lambda_{22} < \Lambda(y_2) < \lambda_{23}$
$u_1=3$ si $\lambda_{13} < \Lambda(y_1)$ $u_2=3$ si $\lambda_{23} < \Lambda(y_2)$

Les valeurs des seuils $\lambda_{11},\lambda_{12},\lambda_{13}$ associées au capteur y_1 trouvées avec l'algorithme agrégatif sont indiquées sur la figure 28. Les valeurs des seuils $\lambda_{21},\lambda_{22},\lambda_{23}$ associées au capteur y_2 trouvés avec l'algorithme agrégatif sont indiquées sur la figure 29. Sur la figure 30, la variation de la probabilité d'erreur en fonction de P_0 correspondant à l'arbre de décision construit en considérant les 6 seuils est représentée. De plus, sur cette figure, nous avons fait figurer les résultats obtenus dans le cas de la détection centralisée avec le capteur 1 seul, le capteur 2 seul et en considérant les deux capteurs, et les résultats obtenus avec l'algorithme précédent pour lequel nous n'avions déterminé que 2 seuils.

On s'aperçoit que la probabilité d'erreur trouvée en utilisant cet algorithme est en dessous de celle trouvée en ne considérant que deux seuils. De plus , elle est très proche de la probabilité d'erreur trouvée en considérant la solution optimale centralisée.

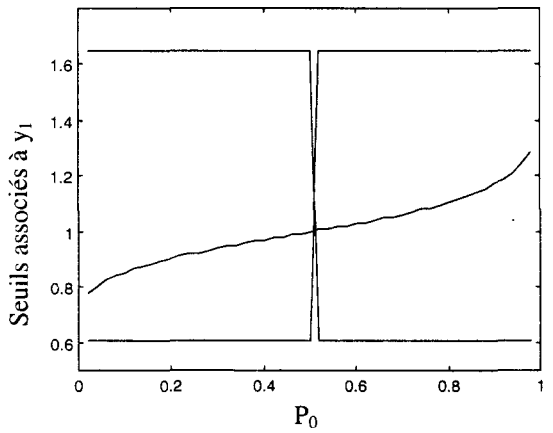


Figure 28 : Seuils associés à y_1 trouvés avec l'algorithme agrégatif en fonction de P_0, $m_1=1$, $m_2=1.5$.

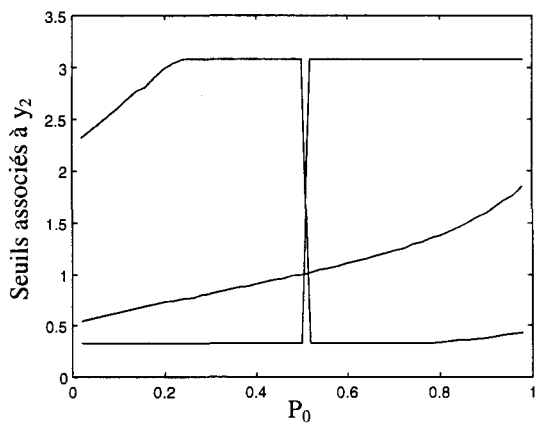


Figure 29 : Seuils associés à y_2 trouvés avec l'algorithme agrégatif en fonction de P_0, $m_1=1$, $m_2=1.5$.

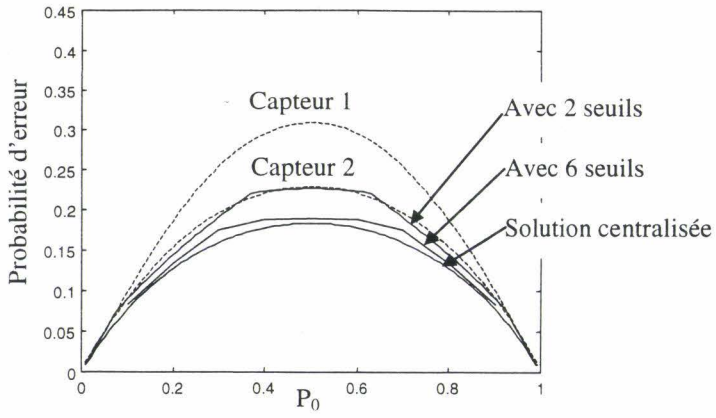


Figure 30 : Probabilité d'erreur en fonction de P_0 correspondant à l'arbre de décision trouvé avec l'algorithme agrégatif, $m_1=1$, $m_2=1,5$.

7. Conclusion

Afin de simplifier le problème d'optimisation des systèmes de détection, nous avons proposé de limiter le nombre de capteurs à prendre en compte.
Il nous est dès lors possible de ne considérer qu'un sous-ensemble de capteurs, plutôt que l'ensemble des informations disponibles. Cette sélection est basée sur l'utilisation d'un critère issu de la théorie de l'information.

Dans un deuxième temps, nous nous sommes proposé de déterminer une architecture décentralisée parallèle basée sur l'utilisation de ce même critère. Nous proposons ainsi la détermination des seuils pour chaque détecteur. Un exemple a été traité, montrant tout l'intérêt de cette approche.

Enfin, les techniques d'optimisation précédentes ont été étendues au problème de la quantification répartie afin d'obtenir un compromis entre la quantité d'information à envoyer à l'opérateur de fusion et les performances du système de détection. Cette approche se justifie pleinement dans le sens où les performances du processus de décision pourront être grandement améliorées si on consent à envoyer un peu plus d'information à l'opérateur de fusion.

CONCLUSION GENERALE

La théorie classique de la détection repose sur le postulat de centralisation de l'information qui suppose que l'information ainsi que le traitement qui lui est appliqué soient regroupés en un même lieu. Une alternative à la structure centralisée a été développée sous la forme d'une architecture imposée où le traitement est décomposé en plusieurs étapes, par exemple où chaque source élabore un résumé de son observation, que l'on transmet ensuite à un opérateur central de décision. L'intérêt de la détection décentralisée est de réduire à sa plus simple expression la transmission des informations locales au niveau central. Ce type d'architecture a été pour la première fois étudiée en 1981 [TeN81], depuis, le sujet de la détection décentralisée a fait l'objet d'une abondante littérature et continue à susciter de nombreux travaux. Aujourd'hui, c'est un sujet qui occupe une place de choix dans les techniques avancées de fusion de données.

Dans ce travail, nous avons rappelé les résultats importants de la théorie de la détection en distinguant les différentes architectures rencontrées dans la littérature : la détection centralisée, décentralisée parallèle et série. Lors de l'optimisation de ces systèmes, deux critères ont été employés : le critère de Bayes et celui de Neyman-Pearson.
L'optimisation de ces différentes architectures a permis de déterminer les opérateurs de traitement locaux permettant d'obtenir les meilleures performances de détection suivant le critère utilisé. En règle générale, ces optimisations aboutissent à un difficile problème de résolution de systèmes d'équations non linéaires couplées. Ces équations n'ont pour l'instant pu être résolues que pour des systèmes comportant peu de capteurs et en supposant, par exemple, l'indépendance des observations.

Partant de l'analogie entre les systèmes de communication numériques et les systèmes de détection, nous nous sommes posés la question de l'utilisation de l'entropie dans les systèmes de détection. Nous avons démontré que les différentes architectures de détection pouvaient être optimisées en utilisant un critère basé sur l'entropie conditionnelle de Shannon. Utiliser ce critère est équivalent à minimiser une fonction « risque moyen » où les fonctions de coût ne sont pas fixées *a priori* mais dépendent des probabilités *a posteriori* associées aux différentes situations. Le critère de Bayes est un critère optimal en terme de probabilité d'erreur, alors que le critère entropique que nous introduisons consiste à trouver un compromis entre une probabilité de fausse alarme faible et une probabilité de détection élevée.

L'utilisation d'un critère entropique ayant été justifié mathématiquement, nous suggérons de l'utiliser dans le cadre d'un apprentissage. Aussi nous avons présenté le problème de l'apprentissage supervisé en général et plus particulièrement la notion d'arbres de décision. Les méthodes de construction d'arbres de décision visent à l'optimisation d'un critère global afin de discriminer les différentes classes en présence ; et permettent de traiter des variables qualitatives, mais également numériques.
Nous avons ensuite défini les outils de la théorie de l'information appliquée à l'analyse structurale des systèmes. Les propriétés les plus intéressantes de l'entropie de Shannon ont été mises en évidence. Des indices ont dès lors pu être construits dans le but de les utiliser dans des algorithmes de classification, et plus particulièrement dans des algorithmes permettant de construire des arbres de décisions ayant la particularité de traiter aussi bien des variables qualitatives que des variables numériques.

Afin de simplifier le problème d'optimisation des systèmes de détection décentralisée, nous avons introduit une phase d'apprentissage inspirée des problèmes de classification. Des algorithmes tirant parti des propriétés de l'entropie ont été développés.
Ainsi, nous avons suggéré de limiter le nombre de capteurs à prendre en compte lors de l'optimisation du système de détection. Parmi tous les capteurs disponibles, nous ne faisons intervenir que ceux apportant de l'information au processus de décision.
D'autre part, nous avons proposé une méthode d'optimisation rapide des systèmes de détection décentralisée parallèle.
Enfin, la technique d'optimisation précédente a été étendue au problème de la quantification répartie afin d'obtenir un compromis entre la quantité d'information à envoyer à l'opérateur de fusion et les performances du système de détection.
Toutes ces méthodes sont relativement simples, faciles à programmer, et peuvent être une alternative très intéressante aux méthodes traditionnelles d'optimisation des systèmes de détection.

PERSPECTIVES...

Jusqu'à présent, l'architecture centralisée, décentralisée parallèle, décentralisée série avec N=2 et la quantification parallèle ont été considérées. Il serait à notre sens intéressant de définir une « architecture mixte » optimale suivant un certain critère (entropique ?). Cette architecture serait une sorte de quantification mixte (parallèle-série) dans laquelle interviendraient des combinaisons d'observations, mais également des combinaisons de décisions locales.
Cette optique de recherche est basée sur l'idée d'une généralisation des processus de détections centralisée et décentralisée. Dans le cas d'un processus de détection centralisée, les observations sont envoyées à un opérateur central qui compare le rapport de vraisemblance à un seuil fixé afin de fournir la décision finale u_0. Dans le cas d'un processus de détection décentralisée, cette même opération se situe localement. Les décisions locales sont ensuite acheminées vers un opérateur central de fusion qui prend la décision finale u_0.
La généralisation semble dès lors évidente (Figure 1), et les systèmes de détection centralisée et décentralisée peuvent alors être vus comme des cas particuliers.

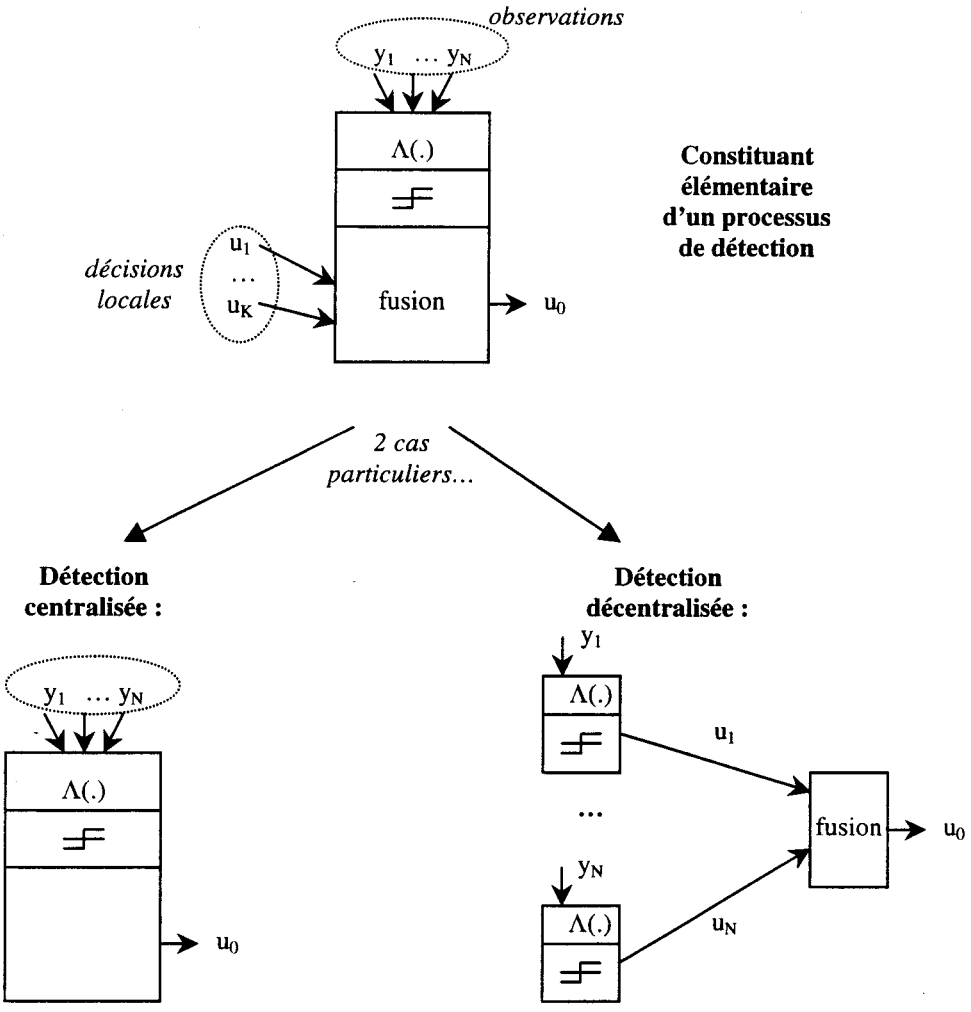


Figure 1 : Généralisation des systèmes de détection.

Il serait dès lors intéressant de rechercher une structure optimale (suivant un critère qui reste à déterminer) de détection constituée de plusieurs « boîtes » interconnectées. Il semblerait qu'une structure arborescente puisse être construite sur cette base.
Des travaux ont déjà été menés dans ce cadre de travail [Bal94]*.

De plus, dans ce travail, nous avons illustré chaque résultat théorique par des exemples simples couramment étudiés dans la littérature, de façon à pouvoir comparer les méthodes que nous proposons avec les méthodes classiques d'optimisation des systèmes de détection. La validité de ces méthodes ayant été démontrée, il serait maintenant intéressant de les appliquer à des systèmes plus complexes dans le cadre de la surveillance de processus industriels.

* [Bal94] F. Baldit « *Détection décentralisée : Théorie et pratique des architectures arborescentes* », thèse de l'Université de Rennes 1, 11 juillet 1994.

Enfin, il serait intéressant d'étendre les résultats obtenus à des problèmes de décisions multi-hypothèses dans lesquels on ne cherchera plus à discriminer deux hypothèses H_0 et H_1 mais m hypothèses $H_0, \ldots, H_{m-1}$ (m>2).
Dans cette optique et dans le cas d'une population d'apprentissage très incohérente, il serait certainement souhaitable de ne prendre aucune décision si la probabilité de fausse alarme est forte (ou que la probabilité de vraie détection est faible). Il est alors conseillé de recueillir quelques observations supplémentaires afin d'augmenter les performances de notre système de décision.

REFERENCES BIBLIOGRAPHIQUES

[AaV89] V. Aalo and R. Viswanathan, « *On Distributed Detection with Correlated Sensors : Two Examples* », *IEEE Trans. on Aerospace and Electronic Systems*, Vol.AES-25, No.3, May 1989, pp.414-421.

[Agg76] N.L. Aggarwal, « *Mesure d'information : caractéristiques et propriétés* », Ecole de l'INRIA, Théorie de l'Information, 1976.

[Ash65] R. Ash, « *Information Theory* », John Wiley and Sons, 1965.

[Ashb65] W.R. Ashby, « *Measuring the Internal Informational Exchange in a System* », Cybernetica Namur, Belgique, vol.8, 1965.

[Ayg86] P. Aygalinc, « *Application de la reconnaissance des formes au diagnostic médical, sélection multi-critère des variables explicatives* », Thèse de 3ème cycle, Université de Lille, 1986.

[BaM70] M. Barbut, B. Monjardet, « *Ordre et classification - algèbre et combinatoire* », tome 1&2, Hachette Université, 1970.

[Bar87] M. Barboucha, « *Modélisation structurale des systèmes complexes, Extraction et validation des règles d'un système expert* », Thèse d'état, Université Lille 1, 26 juin 1987.

[Bar91] J.S. Baras, « *Signal Detection and Estimation* », *Artech House*, Boston, 1991.

[BFO84] L. Breiman, L. Friedman, R. Olshen, C. Stone, « *Classification and Regression Trees* », Belmont, CA : Wadsworth International Group, 1984.

[Bir67] G. Birkhoff, « *Lattice Theory* », Amer. Math. Soc., Providence, 3e éd., 1967.

[Bla68] N.M. Blachman, « *The Amount of Information that Y gives about X* », IEEE Trans. on Information Theory, vol.IT.14, n°1, pp.27-31, 1968.

[CDG89] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, H. Ralambondrainy, « *Classification automatique des données* », Dunod, 1989.

[CDS95] M-P. Carton, D. Pomorski, M. Staroswiecki, « *Sélection de capteurs pour système de détection décentralisée par un algorithme d'apprentissage basé sur l'entropie* », *Gretsi* 1995,18-21 septembre 1995, Juan Les Pins, France.

[ChK92] M . Cherikh, and P.B. Kantor, « *Counterexamples in Distributed Detection* », *IEEE Trans. on Information Theory*, Vol.38, No.1, pp.162-165, Jan.1992.

[ChV86] Z .Chair and P.K. Varshney, « *Optimal Data Fusion in Multiple Sensor Detection Systems* », *IEEE Trans. on Aerospace and Electronic Systems*, Vol.AES-22, No.1, Jan.1986, pp.98-101.

[CoA70] R.C. Conant, W.R. Ashby, « *Every Good Regulator of a System must be a Model of that System* », International Journal of Systems Sciences, vol.1, n°2, 1970.

[Con69] R.C. Conant, « *The Information Tranfer Required in Regulatory Processes* », IEEE Trans. on Systems, Sciences and Cybernetics, vol.SSC-5, n°4, 1969.

[Con76] R.C. Conant, « *Lows of Information wich Govern Systems* », IEEE Trans. on Systems, Man and Cybernetics, vol.SMC-6, n°4, 1976.

[DeK91] H. Delic and D. Kazakos, « *Fusion of Likelihood Ratios in Distributed Bayesian Detection* », *Proc. IEEE SMC Intern. Conf. on Systems, Man, and Cybernetics,* Vol.2, IEEE Service Center, Piscataway, NJ, Oct.1991, pp.755-760.

[Did72] E. Diday « *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes* », Thèse de doctorat, Univ. Paris 6, 4 déc. 1972.

[DPS97] C. Desrousseaux, D. Pomorski, M. Staroswiecki, « *Apprentissage d'un modèle qualitatif dynamique minimal pour la détection décentralisée* », AGIS'97, 9-11 décembre 1997, Angers, France.

[DrL91] E. Drakopoulos and C.C. Lee, « *Optimum Multisensor Fusion of Correlated Local Decisions* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-27, No.4, July 1991, pp.593-605.

[DuH73] P.O. Duda and P.E. Hart, « *Pattern Classification and Scene Analysis* », *John Wiley & Sons*, New York, 1973.

[FiJ68] H.M. Finn and R.S. Johnson, « *Adaptative Detection Mode with Threshold Control as a Foncion of Spatially Sample, Clutter-level Estimates* », *RCA review*, Vol.29, pp.414-464, Sept.1968.

[For65] E.W. Forgy « *Cluster Analysis of Multivariate Data* », Biometrics, vol.21, n°3, sept. 1965.

[Gab66] T.L. Gabrielle, « *Information Criteria for Threshold Determination* », *IEEE Trans. on Information Theory*, Vol.6, pp.484-486, Oct.1966.

[GaG56] W.R. Garner, N.J. Mc Gill, « *The Relation between Information and Variance Analysis* », Psychometrika, vol.21, pp.219-228, 1956.

[GaG96] P. Gallinari, O. Gascuel, « *Statistique, apprentissage et généralisation : Application aux réseaux de neurone* », Revue d'intelligence artificielle, n°10, p.285-343,1996.

[GaL95] P. Gaillard, R. Lengellé, « *Eléments de théorie de la décision* », *cours de DEA*, Université de Technologie de Compiègne, 1995.

[Har28] R.V.L. Hartley, « *Transmission of Information* », Bell System Technical Journal, n°7, p.535, 1928.

[Har84] A. Hart, « *Experience in the Use of an Inductive System in Knowledge Engineering* », M. Bramer ed., Research and developments in expert systems, Cambridge : Cambridge University Press.

[HaR87] H.R. Hashemi and I.B. Rhodes, « *Decentralized Dynamic Decision Making* », *Proc. 26th IEEE Conf. on Decision and Control,* IEEE Service Center, Piscataway, NJ, Dec.1987, pp.1836-1841.

[Hel95a] C.W. Helstrom, « *Element of Signal Detection and Estimation* », *Prentice-Hall*, Englewood Cliffs, N.J., 1995.

[Hel95b] C.W. Helstrom, « *Gradient Algorithm for Quantization Levels in Distributed Detection Systems* », *IEEE Trans. on Aerospace and Elect. Syst.*, Vol.31, pp.390-399, Jan.1995.

[Hin92] G. Hinton, « *Apprentissage et réseaux de neurones* », Pour la science, n°181, p.124-132, nov. 1992.

[HMS66] E. Hunt, J. Marin, P. Stone, « *Experiments in Induction* », New-York, Academic Press, 1966.

[Hob86] I.Y. Hoballah, « *On the Design and Optimisation of Distributed Signal Detection and Parameter Estimation Systems* », Ph.D Dissertation, Syracuse University, Nov.1986.

[HoV89a] I.Y. Hoballah and P.K. Varshney, « *Distributed Bayesian Signal Detection* », *IEEE Trans. on Information Theory,* Vol.IT-35, No.5, Sept.1989, pp.995-1000.

[HoV89b] I.Y. Hoballah and P.K. Varshney, « *An Information Theoretic Approach to the Distributed Detection Problem* », *IEEE Trans. on Information Theory,* Vol.IT-35, No.5, Sept.1989, pp.988-994.

[JaL78] M. Jambu, M.O. Lebeaux, « *Classification automatique pour l'analyse des données* », tome 1 : Méthodes et algorithmes, Dunod Décision, 1978.

[KaB78] A. Kaufmann, G. Boulaye, « *Théorie des treillis en vue des applications* », Masson, Paris, 1978.

[KZG 92] M. Kam, Q. Zhu, and W.S. Gray, « *Optimal Data Fusion of Correlated Local Decisions in multiple Sensor Detection Systems* », *IEEE Trans. on Aerospace and Elec. Syst.*, Vol.28, pp.916-920, July 1992.

[LaT75] P. Larminat, Y. Thomas, « *Automatique des systèmes linéaires* », tome 1 : Signaux et Systèmes, Flammarion Sciences, 1975.

[Ler70] I.C. Lerman, « *Les bases de la classification automatique* », Gauthier-Villars, Coll. Programmation, Paris, 1970.

[LKM91] S.S. Lyengar, R.L. Kashyap, and R.N. Madan, « *Distributed Sensor Networks* », *IEEE Trans. on Systems, Man, and Cybernetics,* Vol.SMC-21, No.5, Sept.-Oct.1991, pp.1027-1031.

[LuK89] R.C. Luo and M.G. Kay, « *Multisensor Integration and Fusion in Intelligent Systems* », *IEEE Trans. on Systems, Man, and Cybernetics,* Vol.SMC-19, No.5, Sept.-Oct. 1989, pp.901-931.

[Man91] R.L. De Mantaras, « *A Distance-Based Attribute Selection Measure for Decision Tree Induction* », Machine Learning, 6, pp.81-92, 1991.

[MeC60] J.L. Melsa, and D.L. Cohn, « *Decision and Estimation Theory* », *Mcgraw-Hill*, New York, 1960.

[Mid60] D. Middleton « *Statistical communication theory* », McGraw-Hill, New York, 1960.

[Mil63] G.A. Miller, « *What is Information Measurement* », American Psychologist, vol.8, n°2, pp.50-51, 1963.

[Min89] J. Mingers, « *An Empirical Comparaison of Selection Measures for Decision-Tree Induction* », Machine Learning, 4, 1989, pp.319-342.

[Nyq24] H. Nyquist, « *Certain Factors affecting Telegraph Speed* », Bell System Technical Journal, vol.3, p.324, 1924.

[PaA90] J.D. Papastavrou and M. Athans, « *Distributed Detection by a Large Team of Sensors in Tandem* », *Proc. 29th IEEE Conf. Decision and Control*, Vol.1, IEEE Service Center, Piscataway, NJ, Dec.1990, pp.246-251.

[Pap90] J.D. Papastavrou, « *Decentralized Decision Making in a Hypothesis Testing Environment* », *Ph.D. Dissertation*, M.I.T., May 1990.

[Par60] E. Parzen, « *Modern Probability Theory and its Application* », John Wiley and Sons, New-York, 1960.

[PeP97] P.B. Perche, D. Pomorski, « *Decision Tree Induction Methods using Entropy Criterion – I. Global Approaches – II. Local Approaches* », Second International ICSC Symposium on Soft Computing (SOCO'97), Nîmes, France, September 17-19, 1997, pp.286-299.

[Pic72] C.F. Picard, « *Graphes et questionnaires* », tome 1&2, Coll. Programmation, Gauthier-Villars, 1972.

[Pom91] D. Pomorski, « *Apprentissage automatique symbolique/numérique – Construction et évaluation d'un ensemble de règles à partir des données* », Thèse de l'Université Lille 1, 6 décembre 1991.

[Poo88] H.V. Poor, « *An Introduction to Signal Detection and Estimation* », *Springer-Verlag*, New York, 1988.

[Pos87] J.G. Postaire, « *De l'image à la décision* », Dunod Informatique, Paris, 1987.

[PoT90] G. Polychronopoulos and N. Tsitsiklis, « *Explicit Solutions for some Simple Decentralized Detection Problems* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-26, No.2, Mar.1990, pp.282-292.

[Qui83] J.R. Quinlan, « *Learning Efficient Classification Procedures and their Application to Chess and Games* », Machine Learning : an Artificial Intelligence Approach, R.S. Michalski, J.G. Carbonell, T.M. Mitchell eds, Tioga Publishing Company, 1983, pp.463-482.

[Qui86] J.R. Quinlan, « *Induction of Decision Trees* », Machine Learning Journal, 1, 1986, pp.81-106.

[Qui87] J.R. Quinlan, « *Simplifying Decision Trees* », International Journal of Man-Machine Studies, 27, 1987, pp.221-234.

[Qui88] J.R. Quinlan, « *An Empirical Comparison of Genetic and Decision-Tree Classifiers* », Proc. Fifth International on Machine Learning Conference, pp.135-141, Morgan Kaufmann, San Mateo, CA, 1988.

[Ren86] L. Rendell, « *A General Framework for Induction and a Study of Selective Induction* », Machine Learning 1, pp.177-226, 1986.

[ReN87a] A.R. Reibman and L.W. Nolte, « *Optimal Detection and Performance of Distributed Sensor System* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-23, No.1, Jan.1987, pp.24-30.

[ReN87b] A.R. Reibman and L.W. Nolte, « *Design and Performance Comparison of Distributed Detection Networks* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-23, No.6, Nov.1987, pp.789-797.

[Sad86] F.A. Sadjadi, « *Hypotheses Testing in a Distributed Environment* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-22, No.2, Mar.1986, pp.134-137.

[Sba83] M. Sbaï, « *Analyse structurale des systèmes complexes : Méthodes d'explication et de partition* », Thèse de 3ème cycle, Université Lille 1, 29 septembre 1983.

[Sba93] M. Sbaï, « *Modélisation structurale – Apprentissage automatique et détection de rupture dans les systèmes complexes* », Thèse d'Etat, Université Mohamed 1er d'Oujda, Maroc, 1993.

[Sha48] C.E. Shannon, « *A Mathematical Theory of Communication* », Bell System Technical Journal, vol.27, 1948.

[Sha49] C.E. Shannon, « *Communication in the Presence of Noise* », IRE, vol.37, p.10, 1949.

[ShW49] C.E. Shannon, W. Weaver, « *The Mathematical Theory of Communications* », Urbana, IL : The University of Illinois Press, 1949.

[Sri86a] R. Srinivan, « *A Theory of Distributed Detection* », *Signal Processing*, 11, pp.319-327, 1986.

[Sri86b] R. Srinivasan, « *Distributed Radar Detection Theory* », *IEEE Proc.*, Vol.133, Part F, Inst. of Electrical Engineers, England, UK, Feb.1986, pp.55-60.

[SRV95] M.D. Srinath, P.K. Rajasekaran and R. Viswanathan, « *An Introduction to Statistical Signal Processing with Applications* », *John Wiley & Sons*, New York, 1995.

[Swa93] P.F. Swaszek, « *On the Performance of Serial Networks in Distributed Detection* », *IEEE Trans. on Aerospace and Elec. Syst.*, Vol.29, No.1, pp.254-260, Jan.1993.

[Sza62] G. Szasz « *Introduction to Lattice Theory* », Ed. Academic Press, New-York and London, 1962.

[Tan90] Z.B. Tang, « *Optimisation of Detection Networks* », *Ph.D. Dissertation*, University of Connecticut, 1990.

[TeS81] R.R. Tenney and N.R. Sandell, « *Detection with Distributed Sensors* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-17, No.4, July 1981, pp.98-101.

[ThV89] S.C.A. Thomopoulos, R. Viswanathan, and D.K. Bougoulias, « *Optimal Distributed Decision Fusion* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-25, No.5, Sept.1989, pp.761-765.

[Tor82] V.M. Toro Cordoba, « *Contribution à l'analyse structurale des systèmes complexes à l'aide de l'entropie et ses généralisations* », Thèse de 3ème cycle, Université Lille 1, 8 mars 1982.

[TPK91a] Z.B. Tang, K.R. Pattipati, and D.L. Kleinman, « *An Algorithm for Determining the Decision Thresholds in a Distributed Detection Problem* », *IEEE Trans. on Sytems, Man, and Cybernetics*, Vol.SMC-21, pp.231-237, Jan-Fev.1991.

[TPK91b] Z.B. Tang, K.R. Pattipati, and D.L. Kleinman, « *Optimisation of Detection Networks : Part I – Tandem Structures* », *IEEE Trans. on Sytems, Man, and Cybernetics*, Vol.SMC-21, No.5, pp.1044-1059, Sept-Oct.1991.

[Tsi86] J.N. Tsitsiklis, « *On threshold Rules in Decentralized Detection* », *Proc.25th IEEE Conf. on Decision and Control*, pp.232-236, Vol.1, Athens, Greece, 1986.

[Tsi93] J.N. Tsitsiklis, « *Decentralized Detection* », *in Advances in Statistical Signal Processing*, Vol.2, pp.297-344, 1993.

[TVB87] S.C.A. Thomopoulos, R. Viswanathan, and D.K. Bougoulias, « *Optimal Decision Fusion in Multiple Sensor Systems* », *IEEE Trans. on Aerospace and Electronic Systems,* Vol.AES-23, No.5, Sept.1987, pp.644-653.

[TVB89] S.C.A. Thomopoulos, R. Viswanathan, and D.K. Bougoulias, « *Optimal Distributed Decision Fusion* », *IEEE Trans. Aerospace and Electronic Systems,* Vol. AES-25, No. 5, Sept. 1989, pp. 761-765.

[Var97] P.K. Varshney « *Distributed Detection and Data Fusion* », Springer Verlag, New York, 1997.

[Ven73] H. Ventsel, « *Théorie des probabilités* », Ed. Mir Moscou, 1973.

[Wal47a] A. Wald, « *Sequential Analysis* », *John Wiley & Sons*, New-York, 1947.

[Wal47b] A. Wald, « *Fondations of General Theory of Sequential Decision Function* », *Econometrica*, Vol. 15, pp.279-313, 1947.

[Wal48] A. Wald, J .Wolfourtz, « *Optimum Character of Sequential Probability Ratio Test* », *Ann. Math. Stat.*, Vol.19, pp.326-339, 1948.

[Wei69] H.L. Weidmann, « *Entropy Analysis of Feedback Control System* », Advances in Control Systems. Theory and Application, pp.225-255, vol.7, edited by C.T. Leondes, Academic Press, New-York, 1969.

[WiW91] P. Willet, and D. Warren, « *Decentralized Detection : When are Identical Sensors Identical* », *Proc. Conf. on Info. Sciences and Systems*, 1991, pp.287-292.

[WWR89] D.Warren, P. Willet, and R. Rampertab, « *Shannon's Information in Decentralized Signal Detection* », *Proc.23rd Conf. on Info. Sciences and Systems*, Baltimore, March 1989.