Synthèse des approches de classement en reconnaissance des formes. Nouveaux outils pour l'adaptation d'un système à son environnement

Classification in Pattern Recognition. New Tools to Adapt a System to its Environment

par Pierre LOONIS, Michel MÉNARD

Laboratoire d'Informatique et d'Imagerie Industrielle
Equipe Raisonnement Appliqué à l'Image
Université de La Rochelle
Avenue Marillac
17042 La Rochelle Cedex 1, France
e-mail : pierre.loonis, michel.menard@l3i.univ-lr.fr

résumé et mots clés

Cet article présente un nouvel outil théorique fondé sur la Théorie de l'Information afin de réaliser une évaluation d'un outil de classement plus fine que les mesures classiques.
Nous travaillons dans le cadre de la Reconnaissance d'objets naturels complexes et compliqués. La nature même du problème incite à travailler à l'aide d'une approche multi-points de vue décisionnels, fusionnés de façon adaptative. Nous montrons que les réseaux connexionnistes permettent l'apprentissage d'une fonction de fusion optimisée selon la nature du problème et la structure du Système de Reconnaissance.
Nous montrons aussi que la répartition de l'information sur chaque outil de classement contribue à une meilleure reconnaissance. Une approche de type génétique est alors conçue pour adapter la partition de l'espace des paramètres relativement à l'ensemble des outils disponibles.

Théorie de l'information, fusion d'informations, multi-points de vue, combinaison multi-classifieurs, réseau connexionniste, reconnaissance des formes, algorithme génétique.

abstract and key words

This paper presents a new theoretic tool based on Information Theory, the main interest of which is to acutely evaluate the classification tools.
The particular nature of real-world objects recognition involves us to design systems based on multi-points of view approaches. The fusion stage has to adapt itself to the environment. We show that neural networks allow to learn the fusion function, optimized to the data and the structure of the composite system.
The performance of a composite recognition system is closed to the partition of the available information on each classification tools. A Genetic algorithm is designed to adapt the parameters space partition with the set of classification tools among the quality of the composite system, genetic algorithm.

Information theory, information fusion, multi-points of view, classifiers combination, neural network, pattern recognition, genetic algorithm.

1. introduction

Classiquement, un processus de Reconnaissance des Formes (RdF) extrait des Interprétations à partir des Représentations d'une image.

La qualité d'un processus de RF admet deux caractéristiques induites par le lien entre Représentation et Interprétation : sa complexité et sa complication. Ainsi, dans le cas d'objets manufacturés, les modèles sont clairement établis et la solution au problème de reconnaissance, pour compliquée qu'elle soit, n'en est pas moins déterminée.

A contrario, les objets naturels, *i.e.* non manufacturés, apparaissent complexes, au sens où des événements nombreux et non tous explicites peuvent influer sur la Représentation d'objets possédant la même Interprétation.

De nombreux facteurs viennent apporter indifféremment complexité ou complication au problème de RdF à résoudre. Par exemple, le degré de contrôle de l'éclairage entraîne une complexité de l'image à traiter. La prise en compte des erreurs dues à la physique des capteurs oblige à compliquer le modèle.

Ces considérations sur les Représentations et les Interprétations influent sur le choix d'une stratégie de reconnaissance : Le cas le plus simple, *i.e.* de modèle et d'image non complexes et non compliqués, peut être traité à partir d'un vecteur de caractéristiques, que les techniques de classement peuvent regrouper en classe [26].

Le cas le plus difficile est celui d'un modèle et d'une image complexes et compliqués. Il entraine la nécessité d'une combinaison de stratégies.

Cet article traite de ce dernier type de problème, et plus particulièrement celui d'applications liées à la reconnaissance des objets naturels (reconnaissance de scène naturelle, tri d'objets naturels, ...).

L'essor actuel de l'Intelligence Artificielle Distribuée a permis l'avènement d'approches de type multi-agents ou de type tableau noir [22,45,38]. Ceci à l'avantage d'apporter des informations complémentaires, mais induit un important problème en aval : comment réaliser la fusion de ces différents avis?

A notre connaissance, si les systèmes généraux se développent, la planification des phases critiques (cellules décisionnelles, phase de fusion, mode de combinaison) est encore extrèmement dépendante du savoir-faire d'un expert du domaine d'application. A notre sens, ceci n'est pas homogène avec le but même du Système Général. En effet ces systèmes sont souvent conçus pour traiter des problèmes où les connaissances des experts sont non fiables, ou incomplètes. C'est le cas de la reconnaissance d'objets naturels, de scènes d'extérieur, de scènes de mouvement, par exemple.

A cause de l'incomplétude de nos connaissances sur de telles applications, l'idée, comme le formule Kanal dans [32], est de doter le système d'une capacité d'auto-adaptation. De cette notion ont émergé les techniques génétiques [30], telles que la programmation génétique [11], les systèmes de classeurs [39,23,6], et connexionnistes (voir [35] ou [31] pour une synthèse). Des applications de ces outils d'Intelligence Artificielle voientt le jour dans des domaines variés comme la commande de processus [28,48], le traitement d'images [1,3], la RF [12,29] ou encore l'apprentissage [39,55].

A partir de ces concepts, nous proposons un protocole d'adaptation d'un système de Reconnaissance des Formes mixte envers son environnement. Ici, nous entendons par *environnement*, non seulement le monde extérieur (le signal à identifier), mais encore les outils d'extraction de paramètres, les méthodes de RdF qui sont à la disposition du système mixte.

L'objectif essentiel de nos algorithmes d'adaptation d'un tel système est la performance. Ceci nous amène à travailler sur des techniques de classement, mais surtout de fusion, dites adaptatives.

Dans un premier temps, nous posons les fondement d'une mesure de qualité plus fine que le classique taux de reconnaissance, ce qui nous permet de mieux guider notre construction du système. Dans la section suivante, nous présentons le protocole permettant d'adapter un système mixte général à une application donnée. Après avoir présenté deux formes de fusion adaptatives parmi les plus usitées, nous proposons une nouvelle fusion adaptative, sous la forme d'une mémoire auto-associative de type connexionniste. Puis nous utilisons un Algorithme Génétique pour trouver la meilleure partition de l'espace des paramètres au sens d'un critère de performance.

2. mesure de performance fondée sur la théorie de l'information

2.1. objectif

Nous appellons *classifieur* un outil de RF mis en œuvre pour effectuer un classement.

Afin de définir un modèle faisant abstraction de l'architecture interne du classifieur, ce dernier est vu comme une fonction qui, à une observation x, associe une partie de l'ensemble des classes d'interprétation Ω :

$$\begin{array}{rcl} e^k : X & \rightarrow & \wp(\Omega) \\ x & \mapsto & e^k(x) \end{array}$$

Selon le niveau d'information caractérisant les sorties du classifieur e^k, il est possible d'affiner cette définition et de répartir les

outils classiques de RF selon une taxonomie [43]. Nous utilisons ces outils à la construction de systèmes de reconnaissance mixtes, fondés sur les principes de coopération et/ou de combinaison et/ou de fusion. L'amélioration d'un tel système passe avant tout par la réponse à : quel observateur du système me renseigne sur le niveau performance du système?

A ce jour, la mesure de performance la plus usitée est encore la fréquence des réussites obtenues sur un jeu de test connu. Mais ce taux, dit de reconnaissance, ne rend pas compte de la répartition des erreurs de classement (il faut alors déterminer les fréquences d'échecs, comme les taux de confusion ou rejet d'ambiguité). Toutes ces informations proviennent de la matrice de confusion. Or, dans le cas des outils avec compilation, c'est-à-dire ayant un comportement défini par la phase d'apprentissage, il est trivial de noter que, pour un même classifieur, les taux de reconnaissance pour chaque classe sont différents. De plus, entre classifieurs, les taux de reconnaissance relativement à une même classe sont différents. Ainsi la reconnaissance d'un outil de classement rendant 80% sur trois classes peut être réparti (99% 94% 50%) ou encore (80% 80% 80%). Si l'objectif est de reconnaître très bien une classe, il est préférable d'utiliser le premier système. Au contraire, si le système doit avoir une confusion homogène, la seconde approche est plus pertinente. Il est alors possible d'affiner la reconnaissance en tenant compte des taux de reconnaissance partiels.

Cette diversité de la mesure de la qualité d'information produite par différents classifieurs, sur les différentes classes de l'espace d'Interprétation, dépendante du corpus de validation, nous a conduit à préférer des outils dépendants du contexte[1] [43].

La théorie de l'Information offre un cadre intéressant pour caractériser un système. L'entropie de Shannon [36] permet de mesurer l'état d'organisation d'une structure et est utilisée en RF pour optimiser le système. Ainsi, partant du principe que la minimisation de l'entropie équivaut à une minimisation des erreurs de classement, de nombreuses techniques sont élaborées pour répondre à la question : « Comment construire un classifieur avec un taux d'erreur minimum? »

Les solutions apparaissent naturelles dans le cadre des problèmes de classification par aggrégation des échantillons en classes homogènes. Il s'agit souvent de probabiliser l'espace à travers une distribution de Gibbs. L'optimisation de la partition revient alors à minimiser l'entropie de Shannon. Cette approche se voit non seulement pour des classifieurs statistiques fondés sur la règle de Bayes [47,2,34], mais aussi relativement aux arbres de décision sur ID3 [50], ou [56,53] ou bien encore au sujet du classement structurel [52].

Néanmoins, c'est la communauté connexionniste qui utilise le plus cette information, soit dans le but de concevoir une architecture plus performante [37,44], soit directement intégrée dans l'algorithme d'apprentissage [35,40].

1. Au sens où l'entend Bloch avec les opérateurs flous [5].

Dans un premier temps, nous nous intéressons plutôt aux travaux visant à caractériser les performances d'un outil de classement à travers une mesure informationnelle. Dans la littérature, c'est l'entropie qui est classiquement utilisée. Des études comparatives ont posé les limites d'utilisation des entropies de Gini-Simpson et de Shannon [9,56]. Mais de nombreux travaux se portent sur l'adaptation de la mesure d'entropie des systèmes fondés sur la modélisation de l'incertitude et de l'imprécision. Ainsi Couso dans [10] montre les relations existantes entre les mesures floues et les mesures informationnelles. Fioretto dans [19] montre les avantages de l'approche par maximalisation de l'entropie pour définir une grandeur sur les mesures de confiance de la théorie des croyances.

L'incertitude structurelle, *i.e.* liée à la description d'une forme, peu aussi être modélisée sous une forme d'entropie comme l'exprime De Luca et Termini dans [13].

Aussi, nous sommes-nous intéressés à caractériser la confusion d'un classifieur à travers la théorie de l'Information, afin d'analyser les comportements d'un outil de RF. Cette étude mène à ce que nous avons appellé : l'Outil d'Analyse Informationnelle (OAI). Des propriétés quant aux bornes de fiabilité d'un classifieur s'en dégagent. Puis nous avons appliqué cette mesure à l'évaluation de la construction d'un système mixte de classement, sur des applications concrêtes.

2.2. mesures classiques de Qualité

Face à cette diversité de méthodes de RF, une seule mesure d'évaluation est classiquement adoptée : le taux de reconnaissance. Le fondement de cette mesure est l'élaboration de la matrice de confusion. A partir de ce descripteur de l'état du classifieur, nous pouvons préciser certaines notions couramment employées.

Faisons remarquer que **la matrice de confusion $N = (n_{ij})_{i,j\in[1,M]}$ d'un classifieur e doit être déterminée sur un ensemble d'apprentis de test T, différent de l'ensemble d'apprentis ayant servi à l'apprentissage du classifieur**. Chaque élément n_{ij} correspond au nombre d'individus de la classe i rangés dans la classe j, et ce pour les M classes d'Interprétation.

Posons $T = (T_i)_{i\in[1,M]} = (T'_j)_{j\in[1,M]}$ où :

- T_i représente l'ensemble des individus de test de la classe i)
- T'_j représente l'ensemble des individus de test classé dans la classe j)

Soit τ_g, taux de reconnaissance global, donné par la somme des éléments de la diagonale de la matrice de confusion sur l'effectif total du corpus d'apprentissage :

$$\tau_g = \frac{\sum_{i=1}^{M} n_{ii}}{|T|}$$

Plus finement les taux partiels de reconnaissance relativement à chaque classe se définissent par :

$$\tau_i = \frac{n_{ii}}{|T_i|}$$

La Confusion est la mesure duale du Taux de reconnaissance :

$$cf_g = 1 - \tau_g$$

Cette notion fait référence aux termes extra-diagonaux de la matrice de confusion : *lorsqu'un classifieur commet une erreur de classement, il confond la « bonne » classe d'Interprétation avec une autre.*

Comme précédemment, nous parlerons de confusion partielle relativement à la classe i :

$$cf_i = 1 - \tau_i$$

2.3. une mesure informationnelle de Qualité : l'Outil d'Analyse Informationnelle

Toutes les validations d'outils de RF se fondent sur le taux de reconnaissance, ou sur des mesures dérivées. Toutes ces informations proviennent de la matrice de confusion. Or, pour un même classifieur, les taux de reconnaissance pour chaque classe sont différents. Il en est de même pour les taux de reconnaissance des classifieurs d'un système mixte, entre eux.

Dans [5], Bloch exprime des spécifications d'opérateurs plus particulièrement efficaces sur certaines classes. Mais l'auteur précise que la sélection s'effectue par des heuristiques et que *« nous manquons pour l'instant de méthodes systématiques, voire automatiques, pour déterminer le meilleur opérateur de fusion »*. Dans cette section, nous proposons une étude du point de vue de la théorie de l'Information permettant d'analyser les comportements de tout outil de RF.

Soit un classifieur e et Ω, l'univers d'Interprétation avec $|\Omega| = M$ où $|.|$ représente la cardinalité.

La matrice de confusion définie ci-dessus peut être exprimée par le biais probabiliste. Soient deux événements :

- « $X : x \in \omega_i$ » se rapportant à la connaissance du Superviseur,
- « $Y : e(x) = \omega_j$ » correspondant au choix d'une étiquette par le classifieur.

Par abus de langage, nous écrirons i à la place de $x \in \omega_i$ et j à la place de $e(x) = \omega_j$.

Chacun de ces deux événements prend ses valeurs dans l'ensemble fini et discret des classes d'Interprétation disjointes.

On peut définir une mesure de probabilité sur l'espace joint par la probabilité conjointe :

$$P_{XY}(i,j) \quad \forall i \in [1,M], \quad \forall j \in [1,M] \tag{1}$$

Par définition, la probabilité d'occurrence de l'événement « $X : x \in \omega_i$ » se déduit de la probabilité conjointe :

$$P_X(i) = \sum_{j=1}^{M} P_{XY}(i,j)$$

De même, pour l'événement « $Y : e(x) = \omega_j$ », on obtient :

$$P_Y(j) = \sum_{i=1}^{M} P_{XY}(i,j)$$

Les probabilités conditionnelles s'expriment par :

$$P_{X|Y}(i \mid j) = \frac{P_{XY}(i,j)}{P_Y(i)}$$

$$P_{Y|X}(j \mid i) = \frac{P_{YX}(j,i)}{P_X(j)}$$

Les probabilités conditionnelles définies précédemment s'expriment en fonction des termes de la matrice de confusion par :

- probabilité conditionnelle de l'événement « $X : x \in \omega_i$ » sous l'hypothèse d'occurrence de l'événement « $Y : e(x) = \omega_j$ » :

$$P_{X|Y}(i \mid j) = \frac{n_{ij}}{|T'_j|} \tag{2}$$

- probabilité *a priori* d'occurrence de l'événement « $X : x \in \omega_i$ » :

$$P_X(i) = \frac{|T_i|}{|T|} \tag{3}$$

De façon symétrique, les probabilités $P_{Y|X}(j \mid i)$ et $P_Y(j)$ peuvent être aisément construites.

Il est alors possible d'utiliser le formalisme de la théorie de l'information de Shannon [21] pour caractériser l'état d'un classifieur construit à partir d'un ensemble de test.

L'objectif est **d'exprimer, par l'intermédiaire de mesures d'informations, le lien entre les connaissances de l'expert** (données par l'événement : « $X : x \in \omega_i$ ») **et les connaissances apprises par le classifieur** (données par l'événement : « $Y : e(x) = \omega_j$ »).

Exprimons la notion d'information mutuelle de ces deux événements reliés par la probabilité conditionnelle $P_{X|Y}(i \mid j)$:

$$I_{X;Y}(i;j) = \log\left(\frac{P_{X|Y}(i \mid j)}{P_X(i)}\right) \tag{4}$$

En fonction des termes de la matrice de confusion, cette information mutuelle s'exprime par :

$$I_{X;Y}(i;j) = \log\left(\frac{n_{ij}|T|}{|T'_j||T_i|}\right)$$

A partir de l'équation (4), les notions d'information et d'entropie classiques peuvent être replacées dans le cadre de la RF.

L'information mutuelle moyenne est alors l'espérance de $I_{X;Y}(i;j)$. Elle s'énonce par une double sommation sur l'ensemble des classes :

$$I(X;Y) = \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\frac{|T_i|}{|T|}\log\left(\frac{n_{ij}|T|}{|T'_j||T_i|}\right)\right) \qquad (5)$$

Pour un très bon classifieur (lorsque la matrice de confusion normalisée par $\frac{1}{|T|}$ est égale à la matrice identité), l'entropie conditionnelle est alors nulle, l'information mutuelle est alors maximale. Ainsi la théorie de l'Information offre un cadre intéressant pour exprimer les performances au sens de la réduction d'incertitude entre les avis du Superviseur et les décisions prises par la méthode de RF.

Cette notion de comportement est une approche qualitative du Taux de reconnaissance. On peut suggérer la correspondance de la figure 1. Cette notion fait référence aux termes de variance de la matrice de confusion. L'importance relative des termes de la diagonale de la matrice indique les spécificités du classifieur envers les différentes classes en présence. *Ce sont ces spécificités qui indiquent le comportement du classifieur.*

La Fiabilité quantifie subjectivement le comportement :

1) un comportement idéal signifie que la source apporte de l'information,

2) une source non-informante a un comportement aléatoire ou mauvais.

Il semble alors que nous puissions utiliser l'Information pour caractériser la fiabilité, au sens de la théorie des Croyances, d'une source par rapport à une autre, et ainsi être prise en compte dans l'étape de fusion.

Nous proposons alors deux critères fondés sur la mesure de l'Information, permettant d'évaluer la qualité d'un classifieur

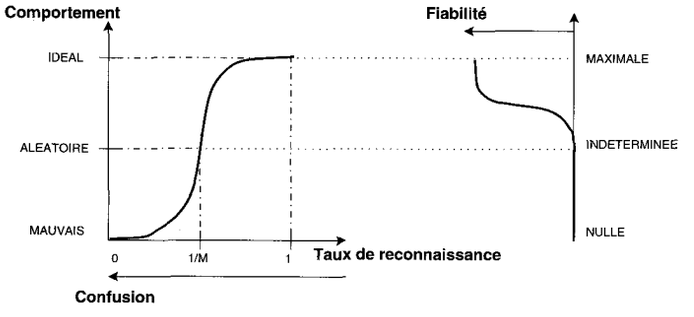


Figure 1. – Définitions - M classes.

2.3.1. *l'information mutuelle*

L'expression (5) permet de mesurer la confusion qui existe dans le classifieur construit. Elle offre donc un moyen de quantifier la discriminance du classifieur considéré, à l'issue de la phase d'apprentissage. En d'autres termes, cette valeur permet de mesurer la qualité d'un classifieur.

Afin de quantifier la part d'information classe par classe pour un classifieur donné, nous proposons de décomposer l'information mutuelle $I_{X;Y}$ relativement à chaque classe, pour donner M informations mutuelles partielles.

De l'expression (5) on obtient :

$$I_{X;Y} = \sum_{i=1}^{M}\sum_{j=1}^{M} P_X(i)P_{Y|X}(j \mid i)\log\frac{P_{Y|X}(j/i)}{P_Y(j)}$$

ou encore :

$$I_{X;Y} = \sum_{i=1}^{M}\sum_{j=1}^{M} P_Y(j)P_{X|Y}(i \mid j)\log\frac{P_{X|Y}(i \mid j)}{P_X(i)}$$

et en forçant l'événement « $X : x \in \omega_i$ » à appartenir à une classe précisément,

$$I_{X;Y}(i;Y) = -\sum_{j=1}^{M}\left(P_{X|Y}(i \mid j)\log\left(\frac{P_Y(j)}{P_{X|Y}(i \mid j)}\right)\right)$$

En fonction des éléments de la matrice de confusion, on obtient l'expression cherchée :

$$I_{X;Y}(i;Y) = -\sum_{j=1}^{M}\left(\frac{n_{ij}}{|T_i|}\log\left(\frac{|T'_j|}{|T|}\frac{|T_i|}{n_{ij}}\right)\right)$$

Cette mesure permet donc de caractériser la pertinence de l'événement « $X : x \in \omega_i$ », relativement au niveau d'apprentissage du classifieur.

2.3.2. *la distorsion moyenne*

De façon complémentaire à la mesure d'information, nous définissons une mesure de non-qualité du classifieur : la distorsion moyenne $\bar{d}$. Cette mesure doit être maximale lorsque le taux de reconnaissance du classifieur est nul, et minimale lorsque ce taux est de 100%.

$$\bar{d}(P_{Y|X}(j \mid i)) = \sum_{i=1}^{M}\sum_{j=1, j\neq i}^{M} P_{YX}(j;i) \qquad (7)$$

soit

$$\bar{d}(P_{Y|X}(j \mid i)) = \sum_{i=1}^{M}\sum_{j=1, j\neq i}^{M}\left(\frac{n_{ij}|T'_j|}{|T||T_i|}\right)$$

3. application à l'évaluation de la partition en M classes d'interprétation

Un classifieur peut alors être représenté par les relations d'informations mutuelles moyennes et de distorsion. Ces mesures tiennent compte aussi bien de l'application (Estimation des probabilités *a priori* sur T) que des performances intrinsèques au classifieur lui-même (Eléments de la matrice de confusion).

L'évaluation d'un outil de RF, qui s'effectue classiquement de façon globale sur la matrice de confusion, peut ici être affinée par la constatation suivante. L'information de confusion que met en évidence la matrice de confusion de taille $M \times M$ peut être exprimée par M matrices de dimension 2×2, en conservant une classe et en regroupant les $M-1$ restantes en une seule classe :

$$N^i = \begin{bmatrix} n_{ii} & |T_i| - n_{ii} \\ |T_i'| - n_{ii} & |T| - |T_i| - |T_i'| + n_{ii} \end{bmatrix}$$

Chaque matrice réduite permet de comparer une classe par rapport à l'ensemble de l'univers d'Interprétation restant. En effet, elle peut être utilisée dans l'estimation des probabilités conditionnelles comme suit, en traitant le cas général de l'appartenance à la classe ω_i :

$$\frac{N^i}{|T|} = \begin{bmatrix} P_{Y|X}(i \mid i) & P_{Y|X}(\bar{i} \mid i) \\ P_{Y|X}(i \mid \bar{i}) & P_{Y|X}(\bar{i} \mid \bar{i}) \end{bmatrix} \qquad (8)$$

où $\omega_{\bar{i}}$ représente le complément de la classe ω_i dans Ω.

Ainsi la détermination des critères informationnels (6) et (7) sur cette matrice réduite permet de caractériser la qualité de la méthode de RF, classe par classe.

3.1. expression de l'information mutuelle

Lorsque l'on étudie le cas d'une classe ω_i et de son complémentaire $\omega_{\bar{i}}$ dans Ω, on peut écrire

$$N^i = \begin{bmatrix} z & N - z \\ c & N' - c \end{bmatrix}.$$

avec

- z, le nombre d'échantillons appartenant à la classe ω_i bien classés,
- c, le nombre d'échantillons n'appartenant pas à la classe ω_i mais classés dans la classe ω_i,
- N, le nombre d'échantillons de T appartenant à la classe ω_i,
- N', le nombre d'échantillons de T n'appartenant pas à la classe ω_i,
- $|T| = |T_i| + |T_{\bar{i}}|$

D'après (6), l'expression de l'information mutuelle $I_{X;Y}(i;Y)$ découle :

$$I_{X;Y}(i;Y) = P_{Y|X}(i \mid i) \log \left(\frac{P_{Y|X}(i \mid i)}{P_Y(i)} \right) + P_{Y|X}(\bar{i} \mid i) \log \left(\frac{P_{Y|X}(\bar{i} \mid i)}{P_Y(\bar{i})} \right)$$

soit,

$$I_{X;Y}(i;Y) = \frac{1}{N} \left[z \log \left(\frac{z}{z+c} \right) + (N-z) \log \left(\frac{N-z}{N-z+N'-c} \right) \right]$$

L'étude [41] montre que cette courbe convexe admet un unique minimum qui survient pour une valeur :

$$z_{01} = \frac{cN}{N'}$$

Cette valeur de z conduit donc au minimum d'information. Physiquement, cela signifie que la confusion est maximale si l'effectif de réussite de la classe concernée est tel que le rapport de la probabilité que le classifieur e vote la classe ω_i sans faire d'erreur, ramenée à la probabilité que ce même classifieur se soit trompé est donné par :

$$\frac{P_{Y|X}(i \mid i)}{P_{Y|X}(i \mid \bar{i})} = \frac{N}{N'}$$

Ce rapport $\frac{N}{N'}$ est constant et représentatif du nombre d'échantillons de la classe considéré dans l'ensemble d'apprentissage, (*i.e.* l'information totale de départ). Le but de l'apprentissage est donc de sortir de ce minimum en augmentant z et diminuant c.

La figure 2 illustre les courbes résultant des expressions de l'information $I_{X;Y}(i;Y)$ et de sa dérivée pour le cas classique où $N < N'$. On retrouve les minima en z_{01}.

De même l'expression de l'information mutuelle $I_{X;Y}(\bar{i};Y)$ peut être déterminée :

$$I_{X;Y}(\bar{i};Y) = \frac{1}{N'} \left[c \log \left(\frac{Ac}{N'(z+c)} \right) + (N'-c) \log \left(\frac{A(N'-c)}{N'(N-z+N'-c)} \right) \right]$$

Une étude similaire à l'étude de $I_{X;Y}(\bar{i};Y)$ donne un minimum unique pour : $z_{02} = \frac{cN}{N'}$

Cela montre de nouveau la symétrie du problème : la confusion maximale se produit au même instant du point de vue de l'événement X ou de l'événement Y.

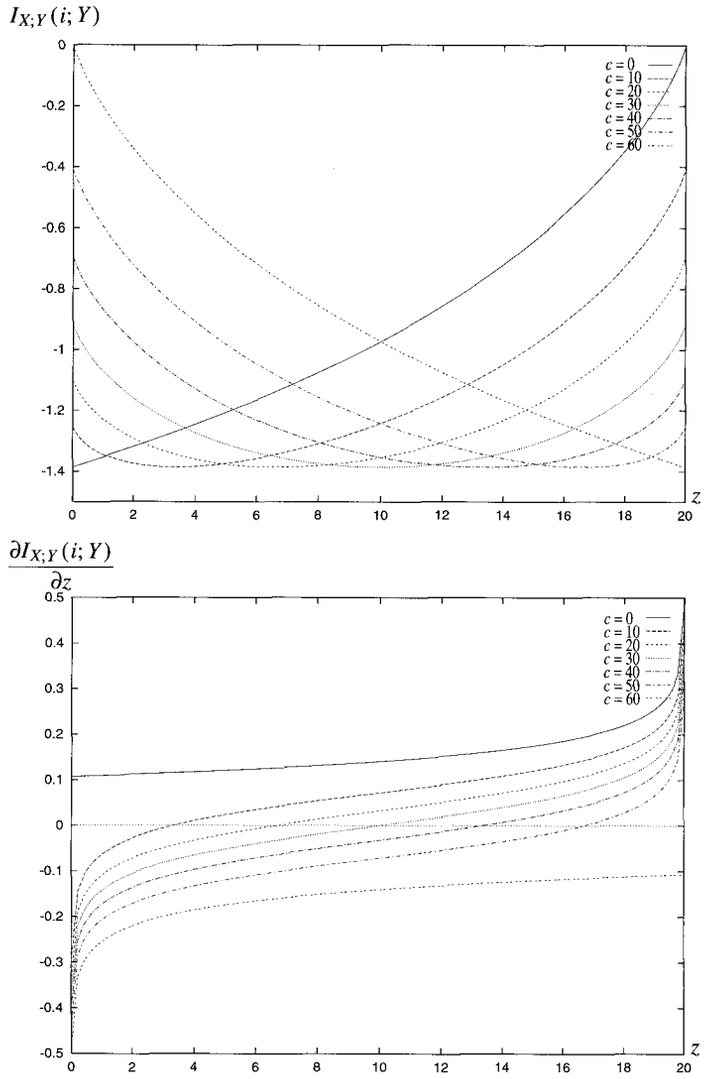


Figure 2. – Illustration de $I_{X;Y}(i;Y)$ et de sa dérivée pour $N = 20$ $N' = 60$.

3.2. comparaison de l'information mutuelle et du taux de reconnaissance

z_{01} correspond au seuil au-delà duquel, pour c constant, tout accroissement de z conduit à une amélioration de la qualité de la méthode de RF considérée :

- pour $z < z_{01}$, la confusion augmente avec z,
- pour $z > z_{01}$, la confusion diminue lorsque z augmente,

Cette valeur minimale est donc obtenue lorsqu'on a l'égalité des rapports :

$$\frac{z}{N} = \frac{c}{N'}$$

ce qui signifie, en terme de probabilité conditionnelle sur la matrice réduite, que l'information mutuelle partielle commence à croître lorsque :

$$P_{Y|X}(i \mid i) \geq P_{Y|X}(i \mid \bar{i})$$

D'après l'expression de c issue de la matrice réduite, on perçoit que la mesure de l'information partielle est beaucoup plus fine que le taux de reconnaissance. En effet, en reprenant la définition du taux de reconnaissance, on peut écrire :

$$z_{01} = c\frac{\frac{N}{N+N'}}{\frac{N'}{N+N'}}$$

soit

$$z_{01} = c\frac{\tau_i}{1-\tau_i}$$

Le seuil de qualité classique est donné par $\tau_i = \frac{N}{N+N'}$. La mesure que nous proposons est plus fine puisqu'elle est ramenée sur le taux du reste de l'univers (le terme $\frac{\tau_i}{1-\tau_i}$), et surtout **parce qu'elle est modulée par les termes de confusion de la classe considérée (le terme c).**

3.3. exemple

De nombreux tests ont été réalisés à partir des données issues du projet européen Statlog. Nous présentons plus avant dans cet article des résultats sur l'application « scène d'extérieur », ou les classes sont les arbres, les murs, la route, ... (voir [46] ou [41] pour plus de détails). Un problème à trois classes montre l'intérêt de la mesure informationelle et son utilisation conjointe avec la mesure z_{01}. L'apprentissage d'une méthode de classement a rendu :

$$\mathcal{M} = \begin{bmatrix} 33 & 0 & 0 \\ 5 & 20 & 8 \\ 0 & 8 & 25 \end{bmatrix}$$

Le nombre d'échantillons est de 99. L'équi-répartition amène un rapport $\frac{N}{N\prime} = \frac{1}{3}$. Par notre OAI, la construction de cet outil de RF est analysée comme suit :

classe i	τ_i	N^i	$I_{X;Y}(i;Y)$	z_{01}
1	1	$\begin{matrix} 33 & 0 \\ 5 & 61 \end{matrix}$	-0.061	$\frac{5}{3}$
2	0.61	$\begin{matrix} 20 & 13 \\ 8 & 58 \end{matrix}$	-0.499	$\frac{8}{3}$
3	0.76	$\begin{matrix} 25 & 8 \\ 8 & 58 \end{matrix}$	-0.314	$\frac{8}{3}$

4. fusion multi-classifieurs par une approche connexionniste supervisée

4.1. fusion multi-classifieurs

Xu, dans [58] décrit trois modes de combinaison de classifieurs :

1) combinaison sous le formalisme du Vote,

2) combinaison sous le formalisme bayésien,

3) combinaison sous le formalisme de Dempster-Shafer.

En introduisant les combinaisons fondées sur les modèles flous, Bloch, dans son étude sur les différentes approches de fusion d'information fondées sur les techniques numériques utilisées en Traitement d'Image [5], généralise la combinaison sous le formalisme du Vote, on parle alors de combinaison floue et possibiliste.

Les opérateurs flous sont les instruments des combinaisons floue et possibiliste. La multitude des possibilités de création de ces opérateurs [33,59] en font le principal avantage de cette approche de la combinaison. Des catégories de comportements (sévère, indulgent ou prudent) précisés par Dubois dans [15] peuvent être rattachées à la position de la fonction de fusion étudiée par rapport aux opérateurs max et min.

Bloch précise sa taxonomie en fonction du comportement de l'opérateur « selon les valeurs des informations à combiner » :

1) « des opérateurs Autonomes à Comportement Constant (ACC) : le résultat ne dépend que des valeurs à combiner et le comportement est le même quelques soient ces valeurs,

2) des opérateurs Autonomes à Comportement Variables (ACV) : le comportement dépend des valeurs numériques des informations à fusionner,

3) des opérateurs Dépendants du Contexte (DC) : dépendant d'une connaissance plus globale comme la fiabilité des capteurs, ou encore le conflit entre les sources. »

Ce dernier type d'opérateurs nous intéresse particulièrement dans notre recherche de la minimisation de l'a priori et de l'adaptativité du système. Le problème est alors de « trouver une bonne mesure de conflit ».

L'avantage de ces opérateurs réside, de par leur grande souplesse d'utilisation, en leur faculté de s'adapter à l'application présentée.

La section suivante présente les approches de fusions adaptatives les plus usitées dans la gestion de sources différentes. Il est montré la forte implication de l'expert dans l'élaboration de ces fusions. Ceci nous amène à développer notre approche qui s'inscrit dans cet objectif d'adaptation au contexte, mais qui nécessite moins de connaissance a priori pour la construction du module de fusion.

4.2. les approches adaptatives de la fusion

De nombreux opérateurs d'aggrégation ou de combinaison d'informations sont disponibles pour les ensembles flous et les possibilités. La plupart des travaux les concernant reposent sur leur utilisation pour la classification supervisée, citons [24], [25],[27] pour ceux concernant les intégrales floues et [18], [14] et [51] ayant pour cadre la théorie des possibilités. Les opérateurs simples les plus couramment utilisés sont les t-normes, les t-conormes, les opérateurs moyennes de différents types, la médiane et le produit. Nous présentons ici, brièvement, deux classes d'opérateurs gérant de façon souple les degrés de conflits et de fiabilités entre sources, propriétés indispensables pour la combinaison de décisions issues d'algorithmes de classement.

4.2.1. les intégrales floues

Dans cette partie, nous présentons les définitions d'une mesure floue, des intégrales floues correspondantes, et des propriétés et interprétation relatives à la combinaison de décisions [24].

Définition 4.1

Soit $X = \{x_1, ..., x_n\}$ un ensemble fini non vide. Une mesure floue est une application g de $P(x)$ (l'ensemble des parties de X) dans $[0, 1]$ et vérifiant les 2 axiomes suivant :

1) $g(\emptyset) = 0, g(X) = 1$

2) $A \subseteq B \Rightarrow g(A) \leq g(B)$

Les fonctions de croyance et de plausibilité incluant les mesures de probabilité sont des exemples de mesures floues.

Définition 4.2

Une g_λ est une mesure floue qui satisfait la propriété supplémentaire : $\forall A, B \subset X \text{et} A \cap B = \emptyset$ alors

$$g(A \cup B) = g(A) + g(B) + \lambda g(A)g(B) \text{ et } \lambda > -1$$

Ceci permet après détermination de λ, de calculer directement la mesure de l'union de deux sous-ensembles disjoints à partir des mesures de chacun des deux sous-ensembles [49]. λ est alors déterminé en résolvant l'équation :

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda g^i)$$

où $g^i = g(\{x_i\})$ et g est une g_λ-mesure floue.

Définition 4.3

Soit g une g_λ mesure floue, alors g est une mesure de croyance, respectivement plausibilité si $\lambda \geq 0$, (respectivement $\lambda \leq 0$) [54]

La notion d'intégrale floue découle de celle de mesure floue [49], [57]. Ce sont des intégrales de fonctions réelles définies par rapport à une mesure floue. Supposons que $0 \leq h(x_1) \leq ... \leq h(x_n) \leq 1$, et posons $A_i = \{x_i, x_{i+1}, ..., x_n\}$. Il existe deux types d'intégrales floues :

1) L'intégrale de Sugeno est définie de la manière suivante :

$$S_g(h(x_1), ..., h(x_n)) = \bigvee_{i=1}^{n} (h(x_i) \wedge g(A_i))$$

2) L'intégrale de Choquet par :

$$C_g(h(x_1), ..., h(x_n)) = \sum_{i=1}^{n} (h(x_i) - h(x_{i-1}) g(A_i))$$

Le cadre de la fusion de décisions est pertinent pour interpréter l'intégrale de Sugeno. Supposons qu'un objet soit classé par un ensemble d'algorithmes de classement X (*i.e.* classifieurs). Soit $h(x_i) \in [0, 1]$ la décision du classifieur x_i concernant cet objet et soit $g(\{x_i\})$ la fiabilité de cette source. Supposons que l'objet soit évalué par un ensemble de sources tel que $A \subseteq X$. La décision ou la combinaison des décisions qui apporte le plus de sécurité est celle qui obéit à $min_{x \in A} h(x)$:

$$min_{x \in A_i} h(x) \wedge g(A_i) = h(x_i) \wedge g(A_i)$$

où $g(A_i)$ exprime le degré de fiabilité du sous ensemble de classifieurs A_i. $g(A_i)$ pondère la décision qui offre le plus de sécurité lorsque les classifieurs $x_i \in A_i$ sont confrontés (intégrale de Sugeno). L'ordre de combinaison des classifieurs privilégie les sources les moins conflictuelles (classement des mesures $h(x_i)$ dans les intégrales de Sugeno et de Choquet).

L'intégrale floue peut donc s'interpréter comme la maximisation (dans le cas où $\vee = max$) des décisions pondérées des sous-ensembles de classifieurs. Nous remarquons aussi que la mesure donnée par l'intégrale floue définie sur une mesure de plausibilité ($\lambda < 0$) est supérieure à celle définie sur une mesure de croyance ($\lambda > 0$). Cependant un comportement de type compromis est atteint puisque :

$$\bigwedge_{j=i}^{n} h(x_i) \leq S_g(h(x_1), ..., h(x_n)) \leq \bigvee_{j=i}^{n} h(x_i).$$

La richesse des intégrales floues pour la fusion d'information, provient de la gestion souple de la fiabilité des classifieurs.

A partir de ces intégrales floues, nous pouvons construire des mesures de fiabilités pour nos algorithmes de classement. Selon le niveau d'intégration souhaité, nous présentons deux mesures que nous avons définies :

- Mesure de fiabilité selon la classe

  Si τ_i^j désigne le taux de reconnaissance partiel relative à la classe j (obtenu sur un ensemble d'apprentissage) pour le classifieurs x_i alors il est possible de définir : $g^j(\{x_i\}) = \tau_i^j$.

- Mesure de fiabilité associée à chaque décision

  Le calcul s'effectue en fonction de la séparabilité (sep) des décisions pour un même classifieur. sep est définie comme la distance qui sépare la classe la plus probable des autres classes : $g(\{x_i\}) = 1 - sep$, où

$$sep = h^j_{max}(x_i) - \frac{1}{N-1} \sum_{j=1}^{N} h^j(x_i)$$

  $j \neq max$ et $\{h^j(x_i)\}_{w_j \in \Omega}$ le vecteur d'interprétation du classifieur x_i.

Lors de la combinaison des décisions, le choix de la classe s'effectue en prenant en considération la valeur de l'intégrale la plus grande. Les intégrales floues offrent une gestion souple des fiabilités, combinent les mesures de décisions dans un ordre privilégiant les classifieurs les moins conflictuels et, selon la mesure floue choisie, offrent un comportement optimiste ou pessimiste. Le processus de fusion n'est cependant pas guidé par un expert, la structure de combinaison étant imposée *a priori*.

4.2.2. combinaison adaptative

Nous présentons ici succinctement comment est abordé le problème de la combinaison d'information incertaines dans le cadre de la théorie des possibilités. Trois comportements de combinaison existent suivant que les informations disponibles sont en conflits ou non : la conjonction, le compromis et la disjonction. Pour passer graduellement d'un mode de combinaison à un autre Dubois et Prade ont proposé une règle de combinaison adaptant automatiquement le mode d'aggrégation à la quantité de conflit présente entre les sources.

La théorie des possibilités repose sur le concept de mesure de possibilité [17], [16], [60]. Soit Ω le référenciel et $P(\Omega)$ l'ensemble des parties de Ω.

Définition 4.4

Π est une mesure de possibilité ssi :

1) $\Pi : P(\Omega) \rightarrow [0\ 1]$

2) $\Pi(\Omega) = 1$ et $\Pi(\emptyset) = 0$

3) $\forall A, B \in P(\Omega) \Pi(A \cup B) = max(\Pi(A), \Pi(B))$

De plus, une distribution de possibilités peut être assimilée à un ensemble flou normalisé :

$$\pi : \Omega \rightarrow [0\ 1]$$

telle que $sup_{\omega \in \Omega} \pi(\omega) = 1$. La mesure de possibilité se détermine alors par : $\Pi(A) = sup_{x \in A} \Pi(\Omega)$.

Dubois et Prade [18] ont proposé une règle de fusion adaptative prenant en compte le niveau de conflit et la fiabilité entre les sources (*i.e.* classifieurs) (Voir [51] et [14] pour deux exemples d'applications). Supposons que parmis k sources, il y a m sources fiables et qu'en aucun cas il n'y en a plus de n ($m \leq n$). Nous partons du principe que les sources fiables sont nécessairement

concordantes. Pour sélectionner ces m sources, nous utilisons un degré de cohérence entre sources représentant le chevauchement entre leur distribution de possibilité :

$$m = sup\{|J|, h(J) = 1\} \text{ et } h(J) = sup_{\omega\in\Omega}(min_{i\in J}(\pi_i(\omega)))$$

où j est un sous-ensemble de sources. Pour sélectionner les n sources, nous calculons une estimation optimiste du nombres de sources fiables : $n = sup\{|J|, h(J) > 0\}$. Dubois et Prade définissent alors la règle de fusion suivante :

$$\pi(\omega) = max(\pi_{(n)}(\omega)/h(n), min(\pi_{(m)}(\omega), 1 - h(n)))$$

où

$$\pi_{(j)}(\omega) = max_{J\subseteq K_{|J|=j}} \bigwedge \pi_{ii\in J}(\omega)$$

et $(h(n) = max\{h(J), |J| = n\}$. La règle est ainsi composée de deux termes généralisant les modes conjonctifs et disjonctifs. L'expression $\pi_{(n)}(\omega)/h(n)$ est une opération où la normalisation est rendue nécessaire par la faible concordance entre les n sources. L'expression $min(\pi_{(m)}(\omega), 1-h(n))$ est une opération de fusion appliquée à un petit nombre de sources hautement concordantes $(h(|J|) = 1)$ limitée à la quantité de conflit $1 - h(n)$.

On peut remarquer dans cette règle que la fiabilité est directement associée au degré de consensus entre sources. La fiabilité n'a donc pas la même signification que celle définie pour les intégrales floues. Pour associer au processus de fusion une fiabilité « exogène » (*i.e.* issues d'observations par exemple), Dubois et Prade définissent une règle de fusion pour des sources plus ou moins fiables.

Il est possible de classer un ensemble de K sources d'informations selon leur fiabilité. Soit $\{K_1, ..., K_n\}$ une partition de K telle que les sources réunies dans K_i sont jugées de fiabilités égales, et plus fiables que celle de K_j dès que $j > i$. La combinaison d'informations issues de sources $K_1, ..., K_n$ peut s'effectuer comme suit : combiner symétriquement (par la règle précédente) les informations relatives à K_1 puis ne raffiner le résultat par les informations issues de K_2 que lorsque celles-ci ne contredisent pas les sources de K_1, etc.

On utilisera donc dans la règle adaptative définie précédemment (dans le cas particulier de deux sources) à la place de la conjonction l'opération de conjonction pondérée suivante :

$$\pi_{\wedge}^{1>2} = min(\pi_1, max(\pi_2, 1 - h(\pi_1, \pi_2)))$$

qui exprime que les sources de K_1 (π_1) fournissent une information fiable tandis que l'information issue de K_2 (π_2) n'est considérée comme fiable qu'avec un degré de certitude égal au degré de consensus entre K_1 et K_2. L'opération disjonctive correspondante est :

$$\pi_{\vee}^{1>2} = max(\pi_1, min(\pi_2, h(\pi_1, \pi_2)))$$

4.3. approche connexionniste

Notre étude des différentes approches des fusions multi-classifieurs [41] met en avant différentes contraintes :

1) mécanisme d'interprétation fixe,

2) coopération de classifieurs de même type,

3) prépondérances des notions *a priori* de cohérence, ignorance, ...

La fusion que nous proposons se fonde sur une approche connexionniste. L'algorithme de rétro-propagation du gradient est utilisé pour apprendre un réseau à une couche cachée.

4.3.1. une fusion par optimisation déterministe

Nous avons choisi de pallier la rigidité liée à la pré-détermination de la fusion (la règle de Dempster-Shafer, de Bayes, ...) par la souplesse de la phase d'apprentissage d'un Réseau connexionniste à rétro-propagation du gradient (MLP). Nous appellerons la fusion qui en découle : la Fusion Connexionniste Dirigée (FCD).

Comme l'exprime à un propos différent, Jodouin dans [31], cette phase d'apprentissage permet de « *comprendre un système complexe, non en s'attardant sur le comportement individuel de ses composants, mais en étudiant le comportement collectif du système dans son ensemble* ».

Kolmogorov a montré que, en un sens, le MLP peut être perçu comme un approximateur de fonctions. En effet, le réseau est de « propagation vers l'avant ». Ainsi le vecteur de sorties est fonction du vecteur d'entrée et des paramètres w correspondant à la matrice de pondération du réseau :

$$y = F(x; w)$$

Dans le cadre d'une fusion par la théorie des Croyances, la fonction F correspond à la règle fixée par Dempster et Shafer. Ici nous cherchons à élaborer F pas à pas selon la technique de l'apprentissage par l'exemple.

Cette phase d'apprentissage supervisé revient à minimiser l'erreur quadratique entre la sortie désirée et la mesure proposée par le réseau. Ainsi force-t'on le MLP à réaliser une hétéro-association entre une Forme $x \in X$ et son Interprétation $C_i \in \Omega$. De façon probabiliste la fonction F peut être perçue comme l'ensemble des frontières de décisions dans l'espace des Interprétations.

Or si cette interprétation du comportement du classifieur MLP s'entend dans le cadre d'un classement, elle est moins satisfaisante dans celui d'une fonction de fusion où les entrées et les sorties appartiennent au même espace.

Le module de fusion connexionniste se fonde sur l'idée de ne pas utiliser le MLP dans sa fonction classique hétéro-associative, mais plutôt de le faire travailler en auto-association, par le fait que la nature des informations de sortie est identique à celle des informations en entrée du réseau.

Ainsi, chaque classifieur de l'étage précédent renvoie sa réponse. Le superviseur présente la solution attendue en sortie de réseau. Le but de la minimisation de l'erreur est alors de renforcer les pondérations de façon à copier les avis des classifieurs lorsqu'ils sont similaires aux sorties désirées, tout en minimisant l'influence des classifieurs trop loin de la solution.

4.3.2. la minimisation de l'a priori

Dans le cadre du MLP à fonction de transfert non-linéaire, l'optimisation est réalisée par une descente du Gradient. L'avantage de cette fonction d'optimisation est dans la phase d'apprentissage. Elle permet d'injecter la connaissance du Superviseur, uniquement sous la forme de la sortie désirée, et non plus sous forme de règles empiriques.

La généralisation due à l'adaptativité du MLP permet de diriger, implicitement et spécifiquement à l'application considérée, la cohérence des informations des classifieurs.

Une autre caractéristique réside en l'opportunité de fusionner différents types de classifieurs. En effet, le fusionneur connexionniste peut accepter toute mesure sur sa couche d'entrée. Les expériences qui suivent vont montrer la fusion de classifieurs de Type 2, retournant une mesure binaire, et de Type 3, retournant une mesure continue. Un ensemble « Classifieur/jeu de paramètres associé » est appellé par la suite Specific Processing Unit (SPU).

Ce réseau de classifieurs, présenté figure 3, fonctionne comme suit :

- Apprentissage : il s'effectue en deux étapes :

  1) apprentissage de l'étage des classifieurs sur un corpus d'apprentissage $E1$, permettant la spécification des outils de classement vis-à-vis de l'application, via un espace de Représentations propre,

  2) apprentissage du module de fusion à l'aide des résultats issus des classifieurs fonctionnant en reconnaissance, sur un corpus $E2 \neq E1$, permettant de caractériser la fusion des informations au contact de l'application considérée.

- Classement. Sur un corpus de test $E3 \neq E1$ et $E3 \neq E2$, les paramètres extraits d'une forme sont répartis sur les SPUs respectifs. Le module de fusion reçoit les avis (décisions, fonction d'appartenance, ...) de chaque classifieur en parallèle, et retourne la décision finale.

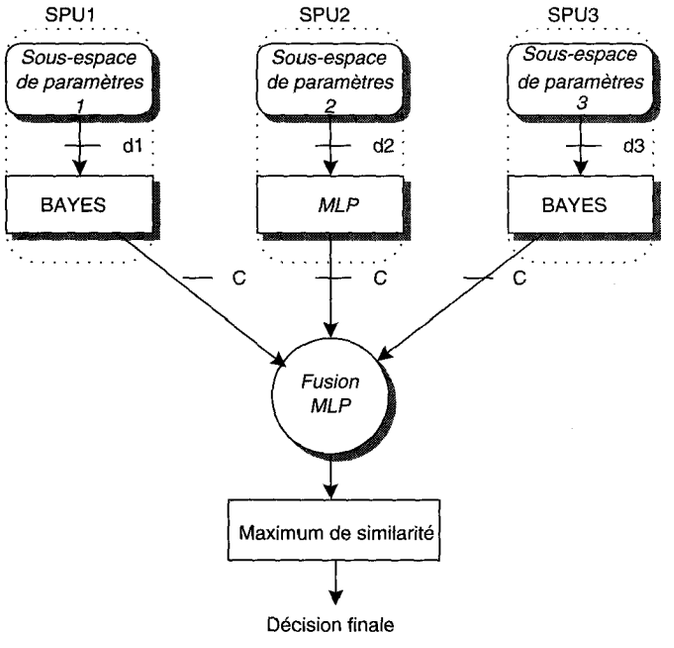


Figure 3. – Fusion par réseau connexionniste.

4.4. application à la reconnaissance de scènes d'extérieur

De nombreux tests ont été réalisés à partir des données issues du projet européen Statlog. Nous présentons ici les résultats obtenus sur le jeu de données *segmentation.data*. Ces données sont extraites de 7 images de vues d'extérieur. Chaque image a été segmentée à la main en 7 classes (brickface, sky, foliage, cement, window, path, grass) et 19 paramètres ont été extraits de ces régions. Après une étude statistique (Analyse en Composante Principale), nous avons conservé 12 des 19 variables continues.

L'expérience présentée ici concerne un classifieur bayésien et un perceptron multi-couche fusionnés par un réseau connexionniste. Le tableau suivant montre le désordre de ce système :

Classifieur	*Reconnaissance*	*distorsion*
BAYES	0.69	0.31
MLP	0.13	0.86
Fusion	0.80	0.19

Puis nous montrons les résultats de l'analyse du système global pour chacune des 7 classes, à travers : le taux de reconnaissance partiel, la matrice réduite, l'information mutuelle partielle et le minimum z_{01}.

classe i	τ_i	N^i	$I_{X;Y}(i;Y)$	z_{01}
1	0.92	72 6 4 498	−0.39	0.62
2	0.99	83 1 0 496	−0.07	0.00
3	0.46	40 47 22 471	−1.50	3.88
4	0.76	64 20 28 468	−1.04	4.74
5	0.85	73 13 58 436	−1.03	10.10
6	0.83	63 13 0 504	−0.63	0.00
7	0.84	71 14 2 493	−0.28	0.34

Remarquons que le taux de reconnaissance des classes 5 et 7 sont quasi-identiques. Or les termes de confusion sont très différents : 58 pour la classe 5 et 2 pour la classe 7. Cela signifie que 58 éléments qui n'appartennaient pas à la classe 5 ont été classés dans cette classe, alors que seulement 2 échantillons ont subi la même erreur pour la classe 7.

Le fait que le taux de reconnaissance ne prenne pas en compte cette confusion vient du mode de calcul de la fréquence de réussite qui ne s'effectue que sur une ligne de la matrice de confusion.

A contrario, l'indice z_{01} est plus faible dans le cas de la classe 7 et l'information mutuelle est plus grande, car les doubles sommations ont pris en compte, d'une façon globale, la répartition des effectifs dans l'ensemble des autres classes d'interprétations. Ce minimum z_{01} est proportionnellement lié au terme de confusion c, issu de la matrice réduite.

Nous pouvons donc conclure que, pour une classe donnée, plus le terme z est éloigné de z_{01} et plus celui-ci est faible, plus l'ordre (ou l'organisation) de cette classe est élevé, donc plus la fiabilité du système de RF associé est importante, relativement à cette classe.

4.5. application à la reconnaissance de poissons de rivière

De nombreux tests ont validé notre approche [42], nous présentons ici les résultats obtenus sur le problème de la reconnaissance de poissons de rivière, qui sont, par nature, déformables, non précisément déterminés et variables dans le temps.

La reconnaissance de poissons de rivière a été traitée récemment par N. Castignolles dans [8]. Les poissons sont vivants, dans leur milieu naturel, ce qui amène des problèmes spécifiques comme les variations de turbidité de l'eau, le retour en arrière d'un individu ayant quitté le champ de la caméra, ... (*cf.* figure 4). De très bons résultats ont été obtenus avec un classement bayésien sur des paramètres morphologiques classiques.

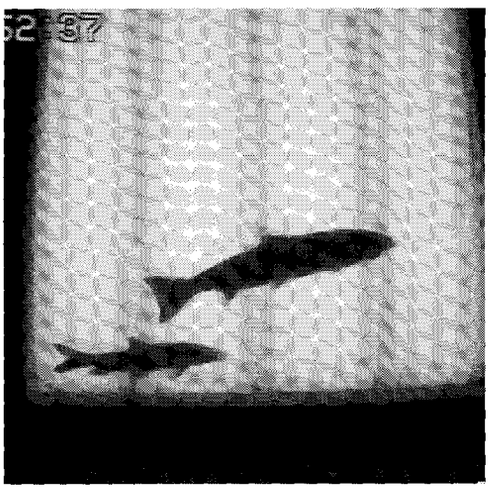


Figure 4. – Type d'image de poissons vivants.

Nous avons utilisé cette base de donnée pour valider notre approche. La cellule de classement est constituée de trois classifieurs bayésiens fusionnés par notre FCD. Les paramètres injectés sont de type morphologique. Après une analyse discriminante, une sélection de 13 mesures extraites permet le tri de 12 espèces. Nous avons réalisé l'apprentissage des SPUs sur des partitions disjointes de l'espace de Représentation. L'architecture du système est présentée sur le schéma 5.

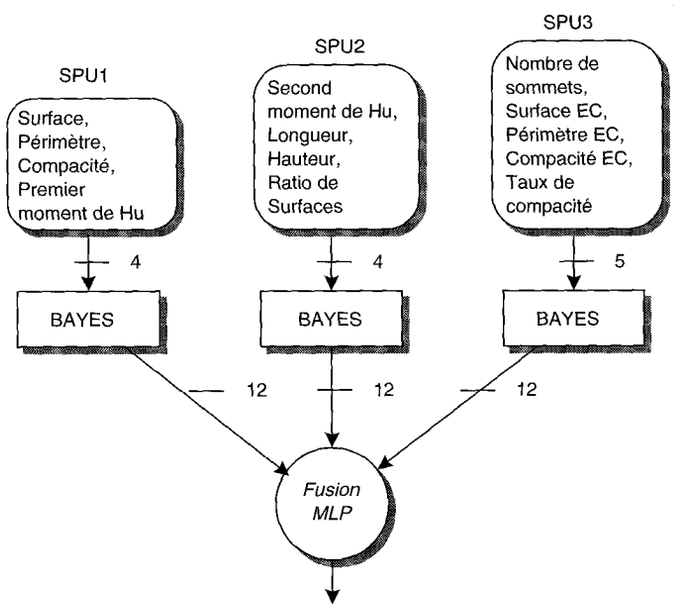


Figure 5. – Systèmes de Classifieurs réalisant la FCD des données poissons de rivière.

L'apprentissage des SPUs a été effectué sur un ensemble de 600 échantillons. L'apprentissage du réseau connexionniste fusionneur a été effectué sur un ensemble distinct de 600 échantillons. Le classement du Système a été effectué sur 1000 échantillons distincts des précédents, dont les occurences sont :

Alose	*Anguille*	*Barbeau*	*Brème*	*Carpe*
119	40	130	116	127
Lamproie	*Muge*	*Perche*	*Sandre*	*Saumon*
18	108	25	35	21
Truite fario	*Truite de mer*			
120	141			

Les résultats obtenus par la Fusion Connexionniste Dirigée sont présentés dans le tableau suivant :

Classifieur	*Reconnaissance*	*Confusion*	*Rejet total*	$I_{X;Y}$
SPU1	0.75	0.14	0.11	−1.06
SPU2	0.72	0.10	0.18	−1.09
SPU3	0.75	0.16	0.09	−1.07
FCD	0.88	0.09	0.03	−0.28

Les taux de reconnaissance et mesures informationnelles classe par classe sont donnés par le tableau suivant. En guise de comparaison, les taux obtenus par N. Castignolles en utilisant un classifieur bayésien unique sont représentés.

	Alose	*Anguille*	*Barbeau*	*Brème*	*Carpe*
$\tau_g(FCD)$	0.93	1	0.91	0.95	0.93
$\tau_g(Bayes)$	0.84	1	0.85	0.96	0.76
$I_{X;Y}$	−0.17	0	−0.15	−0.09	−0.22
z_{01}	0.33	0	0.25	0.03	0.28
	Lamproie	*Muge*	*Perche*	*Sandre*	*Saumon*
$\tau_g(FCD)$	0.72	1	0.68	0.83	0.71
$\tau_g(Bayes)$	1	0.75	0.90	0.90	0.63
$I_{X;Y}$	−1.15	0	−1.22	−1.13	−1.17
z_{01}	4.11	0	11.86	5.04	8.52
	Truite fario	*Truite de mer*			
$\tau_g(FCD)$	0.75	0.80			
$\tau_g(Bayes)$	0.64	0.67			
$I_{X;Y}$	−0.18	−0.21			
z_{01}	0.05	0.16			

Interprétation :

On peut noter l'augmentation générale du taux de reconnaissance de chaque classe, sauf pour les classes *Lamproie, Perche* et *Sandre*[2]. La répartition des différentes espèces dans l'ensemble d'apprentissage est du même ordre de grandeur que celle de l'ensemble de test présenté plus haut. Notons le faible nombre d'échantillons pour les trois espèces dont le taux de reconnaissance diminue avec la FCD : c'est le manque d'apprentis sur ces trois espèces, d'un ordre de grandeur de $\frac{1}{100}$ par rapport aux autres espèces, qui, à travers un mauvais apprentissage du réseau fusionneur, apporte une diminution des taux de reconnaissance.

Par contre l'OAI, en prenant en compte la mesure c de confusion, est relativisé par rapport au déséquilibre de l'ensemble d'apprentissage. Et les mesures montrent une augmentation du taux d'information dans le système organisé autour de la fusion.

Le taux global de reconnaissance des 12 classes, par notre Fusion Connexionniste Dirigée est de 88.2%. Il est de 79.6% par l'utilisation d'un seul classifieur bayésien sur l'intégralité des paramètres [8]. La partition en trois sous-espaces de paramètres sur trois classifieurs bayésien fusionnés par notre fusion adaptative a donc permis un gain de l'ordre de 10% sur l'utilisation d'une seule méthode de RF. De plus l'analyse plus fine apportée par les mesures informationnelles permet de montrer que l'organisation de la classe des Barbeaux est moins désordonnée que celle des Aloses, alors que les taux de reconnaissance ne le laissaient pas prévoir. Il en est de même entre les Truites fario et les Truites de mer.

2. La classe des *Brème* admet un taux de reconnaissance quasiment identique sur les deux expériences.

5. optimisation de la partition de l'espace de représentation

5.1. objectif

L'étude précédente met en avant la forte dépendance Paramètres/Classifieur vis-à-vis du comportement du système global.

Nous proposons ici une méthode d'optimisation dont le but est de trouver la « meilleure » configuration Classifieur/jeu de paramètres, au sens d'une performance optimale du Système Multi-Classifieurs. Notre approche s'appuie sur les mécanismes des Algorithmes Génétiques (AG), que nous avons adaptés à notre problème de RF.

5.2. optimisation génétique

Beightler *et al.*, dans [4], expriment que la théorie de l'optimisation « *étudie comment décrire et atteindre ce qui est meilleur, une fois que l'on connaît comment mesurer et modifier ce qui est bon et mauvais ... La théorie de l'optimisation comprend l'étude quantitative des optima et les méthodes pour les trouver* » [4].

Le problème de l'optimisation d'une fonction f, ou d'un processus p sur un domaine D consiste à chercher un point correspondant à l'optimum global de f, ou au fonctionnement idéal de p sur D. Il existe trois grandes catégories de méthodes d'optimisation :

1) les méthodes fondées sur le calcul utilisent une fonction de recherche qui évolue dans une direction dépendante du gradient de la fonction au point considéré. La dérivée de la fonction doit exister dans le domaine d'étude. Exemple : gradient, Gauss-Newton, ...

2) les méthodes énumératives ne calculent pas directement la solution mais procèdent par énumération des solutions possibles. Deux approches sont à distinguer :

a) les méthodes déterministes sont beaucoup plus générales que les méthodes fondées sur le calcul et nécessitent moins de connaissance *a priori*. En pratique, la fonction f à optimiser doit remplir certaines conditions (continuité, dérivabilité ...).

b) les méthodes aléatoires explorent l'espace des solutions selon une loi de probabilité. La meilleure solution ainsi générée est alors sélectionnée. L'avantage de ces méthodes est qu'elles ne demandent que la connaissance de la fonction f. Cette recherche sans *a priori* ne garantit pas de trouver le point optimal, mais elle permet d'apporter une amélioration de l'état originel. Exemple : Relaxation, ...

L'idée développée dans cette étude est que la performance globale du système peut conduire la sélection des paramètres d'entrée de chaque classifieur. Le mécanisme d'optimisation correspond au moyen de réaliser cette rétro-action de la mesure de performance, afin de corriger l'entrée (cf. figure 6).

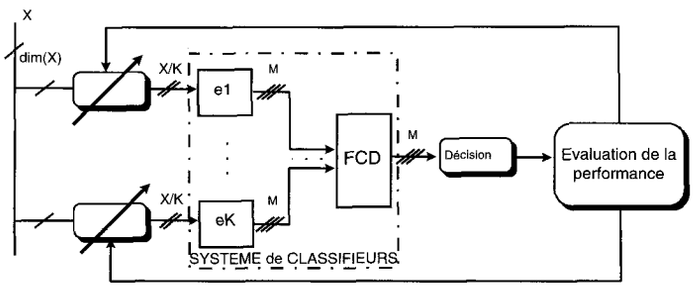


Figure 6. – La performance globale dirige la partition de l'espace d'entrée.

Dans cette section nous employons des notations conformes aux problèmes d'optimisation. Soit K le nombre de classifieurs, posons :

• $\theta = \left(\theta^k\right)_{k\in[1,K]}$ les ensembles de paramètres en entrée du système de K classifieurs,

• $\theta^k = \left(\theta_1^k, .., \theta_{Pk}^k\right)$ ensemble de P_k mesures sur l'ensemble des parties de l'espace de Représentation X, associé au classifieur k

Le problème posé revient à maximiser la fonction coût f :

$$\begin{array}{rcl} f : \wp(X)^K & \rightarrow & [0,1] \\ \theta & \mapsto & \tau_g \end{array}$$

Ceci signifie que le domaine de définition de f est constitué par un K-uplet de parties de l'espace des Représentations X, et produit une mesure, le taux de reconnaissance. La fonction coût f possède les caractéristiques suivantes :

• forte dimensionnalité : chaque classifieur prenant ses paramètres dans X, la taille de l'espace à optimiser est $Card(\wp(X))^K = 2^{Card(X)K}$,

• complexe : en effet, f qui désigne le taux de reconnaissance, ne peut s'exprimer en fonction des éléments de $\wp(X)$, ne serait ce que du fait des diverses procédures d'apprentissage des différents étages (classifieurs, fusionneur).

Cette forte dimensionnalité du problème, à laquelle s'ajoute une fonction de coût implicite, non-linéaire et discrète sur les paramètres, nous conduit à opter pour une technique d'optimisation stochastique.

Les études actuelles présentant ces caractéristiques sont les problèmes d'optimisation de structures de réseaux connexionnistes. Classiquement, les méthodes de relaxation stochastiques, comme le recuit simulé, étaient les plus usitées. Or récemment, certains auteurs (dont Fleurent dans [20], ou Heistermann dans [28]) ont montré que cette approche est un cas particulier d'une autre technique stochastique : les Algorithmes Génétiques (AG).

« *L'algorithme génétique (AG) est un exemple de procédure d'exploration qui utilise un choix aléatoire comme outil pour guider une exploration hautement intelligente dans l'espace des paramètres codés* » (*cf.* [23]).

Cette technique garantit non pas d'atteindre l'optimum global de la fonction, mais de converger vers cet optimum. La littérature est importante dans le domaine de la configuration des réseaux connexionnistes. Comme ce problème ne nous concerne pas expressément, citons seulement Brancke [7] pour une étude complète du sujet, et notamment l'optimisation mixte des poids et de la structure du réseau. Pour des systèmes complexes, il est plus intéressant d'obtenir de bons résultats rapidement que d'atteindre le point optimal au terme d'une longue exploration. Ainsi, des applications industrielles de plus en plus nombreuses voient le jour, dont la plus célèbre est aux Etats Unis, la commande d'un réseau de gazoduc, ou encore au Japon, le contrôle de stockage de fruits [48].

5.3. sélection optimale de paramètres

Nous avons justifié la nécessité d'utiliser une méthode exploratoire robuste comme les AG, pour traiter notre problème de grande dimension.

Dans cette section, nous montrons comment il est possible d'utiliser les mécanismes d'un AG afin de réaliser un choix optimal des paramètres de l'espace de Représentation, relativement au taux de reconnaissance global du système de classifieurs pris en compte.

Précisons que dans notre approche, le but n'est pas de trouver la meilleure configuration du système de classifieurs pour obtenir le taux de reconnaissance le plus élevé, mais plutôt le meilleur partitionnement de l'information en entrée d'un système donné. Le système en lui même est perçu comme une boite noire, de structure fixe.

5.3.1. codage de l'information

La représentation classique consistant à coder un individu par un chromosome n'est pas adaptée à notre problème. En effet, chaque classifieur doit avoir en entrée un vecteur de paramètres prenant ses valeurs dans l'ensemble des paramètres disponibles. **Ainsi les SPUs d'un même individu doivent pouvoir posséder des paramètres en commun**.

De plus, l'évolution du chromosome, c'est-à-dire les modifications des sous-ensembles de paramètres, doit pouvoir être contrôlée. Ceci signifie qu'il y a une certaine homogénéité des SPUs à conserver au sein de la population des systèmes de classifieurs : **les opérateurs génétiques doivent contrôler le mélange des paramètres des différents SPUs**.

Nous avons défini une représentation multi-chromosomique particulière :

- Soit $C = (C^n)_{n\in[1,N]}$ la suite de chromosomes représentant une population composée de N individus.
- Un chromosome C^n représentatif d'un individu n est composé d'une suite de K sous-chromosomes $\left(C^{nk}\right)_{k\in[1,K]}$, où K est le nombre de classifieurs.

Chaque sous-chromosome C^{nk} code l'ensemble de paramètres θ^k pour tout classifieur $k \in [1, K]$. Un des sous-chromosomes de l'individu n, relatif au SPU k, est une suite d'éléments prenant ses valeurs dans $\{0, 1\}$. Il peut se formaliser comme une fonction de l'ensemble des parties de X vers $\{0, 1\}^P$ où $P = dim(X)$:

$$\begin{array}{rcl} \text{Codage} : \wp(X) & \rightarrow & \{0,1\}^P \\ \theta^{n,k} & \mapsto & C^{nk} = \left(C_i^{nk}\right)_{i\in[1,P]} \end{array}$$

où :

- $C_i^{nk} = 1$ lorsque le paramètre i est utilisé par le classifieur k de l'individu n.
- $C_i^{nk} = 0$ sinon

La contrainte de taille sur le vecteur d'entrée de chaque classifieur impose $\sum_{i=1}^{P} C_i^{nk} = P^k = \text{Card}(\theta^{nk})$

Dans notre étude, nous prendrons des longueurs de sous-chromosomes identiques :

$$\text{Card}(\theta^{nk}) = \frac{\text{Card}(X)}{K} \; \forall n \in [2, N] \; \forall k \in [1, K]$$

Ceci signifie qu'il y a le même nombre d'allèles à 1 dans chaque sous-chromosome. *A fortiori*, chaque chromosome prend en compte $dim(X)$ paramètres (distincts ou non).

Ceci permet, entre autre, de diminuer la taille des sous-espaces de Représentation associés à chaque Système de Classifieurs, lorsqu'augmente le nombre de classifieurs disponibles.

5.3.2. la fonction d'évaluation

Les performances du Système de Classifieurs sont clairement données par les taux de reconnaissance globaux. Pour déterminer cette valeur, le protocole expérimental précisé à la sous-section 4.3.2 est lancé sur chaque individu de la population. Un individu est donc évalué par la détermination du taux de reconnaissance représentatif du comportement global du système. Cela revient à juger du pouvoir discriminant du jeu de paramètres associé à l'individu, pour une structure imposée du système.

La population, à un instant donné t, est donc évaluée par un vecteur de N mesures d'évaluation (*i.e.* N taux de reconnaissance) :

$$\mathcal{T}(t) = \left(\tau_g^1(t), .., \tau_g^N(t)\right)$$

Ce vecteur permet de distinguer les individus entre eux, en fonction de leur comportement : les « bons » individus ont un taux élevé, contrairement aux « mauvais » individus.

Ceci permet une évaluation de la pertinence des sous-ensembles de paramètres associés à chaque individu : les sous-ensembles fortement discriminants, sont qualifiés d'une mesure d'évaluation proche de 1. Un comportement aléatoire est obtenu pour les SPUs constitués d'un sous-ensemble de paramètres non-pertinent et conduit à des mesures d'évaluation voisines de $\frac{1}{M}$, avec M le nombre de classes.

Tout le problème de la fonction coût f qui dirige la recherche de l'AG, est de gérer ce vecteur $\mathcal{T}(t)$. En posant $C(t)$, la famille des vecteurs de paramètres de la population à l'itération t. $C(t)$ est une matrice de dimension $K \times N$:

$$C(t) = \begin{bmatrix} C^{11}(t) & \dots & C^{n1}(t) & \dots & C^{N1}(t) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ C^{1k}(t) & \dots & C^{nk}(t) & \dots & C^{Nk}(t) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ C^{1K}(t) & \dots & C^{nK}(t) & \dots & C^{NK}(t) \end{bmatrix}$$

Le problème d'optimisation peut s'exprimer alors par la maximisation du vecteur des taux de reconnaissance :

$$\begin{bmatrix} \tau^1(t) \\ \vdots \\ \tau^N(t) \end{bmatrix} = f\left(C(t-1), \begin{bmatrix} \tau^1(t-1) \\ \vdots \\ \tau^N(t-1) \end{bmatrix}\right)$$

5.3.3. modification des opérateurs génétiques

L'importance des définitions des opérateurs se perçoit en cours de fonctionnement de l'AG par la mauvaise résolution du dilemme Exploitation/Exploration :

- mutation excessive, et l'AG explore plus qu'il n'exploite,
- sélection prépondérante, et l'AG ne fait que de l'exploitation,
- croisement non contrôlé, et l'AG ne conserve pas les schémas intéressants.

Nous proposons les solutions suivantes, afin de pallier ces inconvénients.

La sélection

Nous utilisons le principe, défini par Goldberg, de la « roue de la fortune ». Pour participer à la création de la population à l'instant t, soit $C(t)$, à partir de la population à l'instant $(t-1)$, soit $C(t-1)$, chaque individu a une chance d'être sélectionné, proportionnelle à sa mesure d'évaluation.

Le Croisement

Cet opérateur doit permettre l'échange d'informations entre deux chromosomes « parents » pour former deux chromosomes « fils ». Nous nous sommes fixés la contrainte P^k = Constante, ce qui oblige à redéfinir le croisement. En effet, si le lieu de croisement est totalement aléatoire, le nombre d'allèles à 1, à l'intérieur d'un sous-chromosome peut varier. Ceci signifie que le sous-espace des paramètres associés à un SPU est de taille variable. L'exemple suivant met en avant ce problème : Soit C^1 et C^2 les

deux chromosomes devant être croisés sur le sous-chromosome 1. (*i.e.* C^{11} et C^{21} sont les sous-chromosomes devant être croisés). Un site de croisement est défini au niveau du troisième allèle.

$$\begin{array}{ccc} \text{soit } \theta^{11} = \{1,5,8,9,10\} & \text{et} & \theta^{21} = \{1,2,3,6,8\} \\ \text{avec } dim(\theta^{11}) = 5 & \text{et} & dim(\theta^{21}) = 5 \\ C^{11} = 1\,0\,0\,0\,1\,0\,0\,1\,1\,1 & \oplus & C^{21} = 1\,1\,1\,0\,0\,1\,0\,1\,0\,0 \\ \text{donne } C^{31} = 1\,0\,0\,0\,0\,1\,0\,0\,1\,0 & \text{et} & C^{41} = 1\,1\,1\,0\,1\,0\,0\,1\,1\,1 \\ \text{soit } \theta^{31} = \{1,6,9\} & \text{et} & \theta^{41} = \{1,2,3,5,8,9,10\} \\ \text{avec } dim(\theta^{31}) = 3 & \text{et} & dim(\theta^{41}) = 7 \end{array}$$

Ce croisement a donc violé la contrainte imposée.

L'opérateur que nous avons écrit ici se fonde sur la recherche aléatoire de deux lieux de croisement. Nous montrons que le résultat du croisement vérifie toujours la contrainte $dim(C^{nk}) =$ Constante.

Soit :

- $\times$ le lieu de croisement (i.e. les sous-chromosomes devant être croisés).
- $\theta'^{n\times}$ l'ensemble des paramètres non utilisés dans le classifieur $\times$ de l'individu n ($\theta'^{n\times} = X - \theta^{n\times}$.
- n_1 et n_2 les parents sélectionnés.
- c le nombre de croisement déjà effectué pour former la population $t+1$ à partir de la population t.
- p un élément sélectionné aléatoirement dans $\theta^{n_1\times} \cup \theta'^{n_2\times}$ et q un élément sélectionné aléatoirement dans $\theta^{n_2\times} \cup \theta'^{n_1\times}$ [3].

Les deux fils $\theta^{2c+1}(t+1)$ et $\theta^{2c+2}(t+1)$ seront composés des paramètres suivants :

$$\begin{aligned} \theta^{2c+1\times}(t+1) &= \left(\theta^{n_1\times}(t) - \{p\}\right) \cup \{q\} \\ \theta^{2c+2\times}(t+1) &= \left(\theta^{n_2\times}(t) - \{q\}\right) \cup \{p\} \\ \theta^{2c+1\,k}(t+1) &= \theta^{n_1 k}(t) \;\; k < \times \\ \theta^{2c+2\,k}(t+1) &= \theta^{n_2 k}(t) \;\; k < \times \\ \theta^{2c+1\,k}(t+1) &= \theta^{n_2 k}(t) \;\; k > \times \\ \theta^{2c+2\,k}(t+1) &= \theta^{n_1 k}(t) \;\; k > \times \end{aligned}$$

La mutation

Afin de conserver un nombre d'allèles égal à 1 constant pour chaque chromosome, nous réalisons une double mutation en permutant 2 allèles de valeurs différentes. Dans l'espace de Représentation, ceci revient à prendre en compte un paramètre qui n'appartenait pas au SPU lors de la génération précédente, et à en abandonner un.

3. Note : si $\theta^{n_1\times} \cup \theta'^{n_2\times} = \emptyset$, les deux parents sont identiques, il n'y a pas de croisement possible

5.4. résultats

Nous présentons ici les résultats obtenus sur l'application « Reconnaissance de vues d'extérieur ».

Notre étude concerne l'optimisation d'un système de 2 classifieurs bayésiens fusionnés par FCD. Les paramètres du problème sont les suivants :

- espace de Représentation : $dim(X) = 12$,
- espace d'Interprétation : $dim(\Omega) = 7$,
- nombre de SPUs : $K = 2$,
- nombre d'individus dans la population : $N = 20$,
- nombre de générations : $G = 15$;
- probabilités de croisement et de mutation fixées à $0,8$ et $0,01$.

Le protocole expérimental est le suivant. A $t = 0$, la population d'individus est générée aléatoirement. Ceci revient à sélectionner 20 couples de sous-espaces de paramètres. Chaque couple est associé à deux classifieurs bayésiens pour former 20 Systèmes de Classifieurs. Les phases d'apprentissage des classifieurs puis du module de fusion sont alors lancées. Enfin l'évaluation de la population de systèmes s'effectue sur un corpus de test distinct des corpus d'apprentissage.

A la fin de la génération, l'AG dispose de 20 taux de reconnaissance associés aux 20 systèmes de classifieurs. Ces taux constituent les 20 mesures d'évaluation des individus de la population initiale. Les opérateurs génétiques sont alors exécutés sur cette population.

L'observation de l'évolution de l'algorithme permet de voir que la méthode converge effectivement vers un optimum (voir figure 7). Ainsi, dans le cas relativement simple d'un Système de Classifieurs composé de 2 SPUs sur 12 paramètres, la convergence ne nécessite que très peu d'itérations : le problème d'optimisation de la sélection des paramètres en entrée du Système de Classifieurs ne semble pas être un problème AG-difficile (selon la définition employée classiquement par la communauté s'intéressant aux AG, voir Venturini [55]).

Cette rapidité de convergence nous permet de pouvoir utiliser notre mesure $I_{X;Y}$ associée à z_{01} comme fonction d'évaluation. Les résultats précédents ont montré que la qualité du système est d'autant meilleure que la valeur absolue de l'information mutuelle est faible. Nous prenons en compte les mesures z_{01} issues des information mutuelles partielles comme un coefficient multiplicatif moyen qui doit tendre vers 0. Il s'en suit la fonction suivante :

$$f = M.\frac{I_{X;Y}}{\sum_{i=1}^{M} z_{01}}$$

La figure 8 montre que l'amélioration du système est beaucoup plus lente, mais reste abordable. Les taux de reconnaissance sont du même ordre que pour l'expérience précédente, évaluée par la mesure de réussite, par contre l'homogénéité du système mixte est plus grande, tant au niveau des SPUs qu'au niveau global (*i.e.* après la fusion).

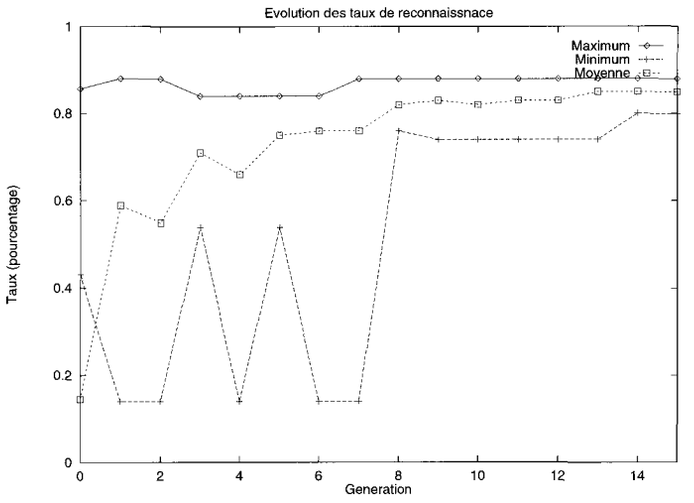


Figure 7. – Evolution des taux maximum, moyen et minimum de la population (expérience 1).

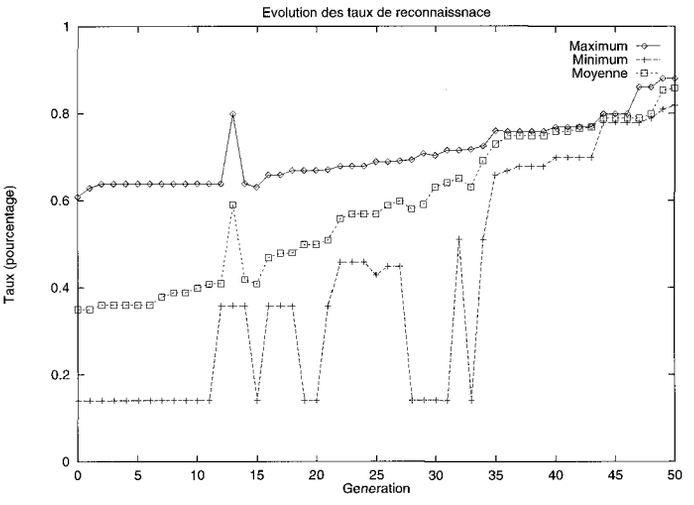


Figure 8. – Evolution des taux maximum, moyen et minimum de la population (expérience 2).

Une Analyse en Composantes Principales montre que les paramètres contenant le plus d'information ont été sélectionnés. Le complément du vecteur de paramètres se fait à l'aide d'attributs contenant peu d'information.

Au sein des approches d'optimisation stochastique, un grand avantage des AG réside en leur particularité à réaliser l'exploration sur un ensemble de solutions possibles, plutôt que sur un unique candidat à évaluer (comme pour le recuit simulé). Cette évaluation simultanée de plusieurs solutions permet aux AG d'explorer, en parallèle, de nouvelles branches de l'arbre de recherche, tout en exploitant des directions de recherche intéressantes au sens de la fonction d'évaluation proposée.

L'algorithme présenté précédement, a donc été « virtuellement » parallélisé. Ainsi une évaluation de population de N individus, en disposant de N machines, ne dure que 20 mn sur un Pentium 90.

6. conclusion

Cette étude porte sur la reconnaissance d'objets naturels complexes et compliqués. Des mécanismes de reconnaissance déconnectés de l'application sont nécessaires et réunis en un système général, par opposition aux Systèmes Dédiés spécifiques d'un contexte déterminé.

Nous avons travaillé à dissocier les mécanismes de reconnaissance, de la connaissance du domaine, selon une approche multi-points de vue, afin de construire un Système Général. Il en résulte un Système de Classifieurs, constitué d'un ensemble de techniques mixtes, reliées par une approche adaptative. L'étape déterminante est alors la *fusion* : il s'agit de gérer la confrontation de différents avis, dans le but de prendre une décision robuste. Dans la plupart des travaux, à notre connaissance, les notions d'imprécision, d'incertitude, ainsi que le contrôle des incohérences inhérentes aux multiples sources du système, sont gérées **explicitement** par la fusion.

Notre approche se situe dans une toute autre direction : **une fusion non explicite dans laquelle l'imperfection, comme l'incohérence, sont des notions gérées globalement et de façon adaptée à l'application concernée, par le système lui-même.**

Nos sources d'informations sont des méthodes de RF. Dans le cadre de nos applications de tri, la dimension nettement plus importante de l'espace de Représentation par rapport à l'espace d'Interprétation nous conduit à utiliser un ensemble restreint de méthodes rendant un maximum d'informations.

Nous employons donc la fusion de méthodes de RF selon deux objectifs :

1) **utiliser de multiples points de vue décisionnels**, afin de profiter des mécanismes de discrimination statistique (Bayes), aussi bien que structurels (Arbre de décision) ou connexionniste (Perceptron Multi-Couches).

2) **réaliser une partition de l'espace de Représentation**, afin d'offrir à chaque méthode de RF un espace d'entrée de taille réduite, dans le but de simplifier l'élaboration des hyper-frontières de décision.

Une étude des techniques de fusions adaptatives nous permet de tirer les conclusions suivantes. Chacun des deux types de fusion, intégrales floues et combinaison adaptative, permet la gestion souple de la fiabilité et de la cohérence entre classifieurs. Dans le cadre des intégrales floues, c'est la cohérence qui pilote la séquentialité des opérations de combinaisons ($0 \leq h(x_1) \leq ... \leq h(x_n) \leq 1$). Dans le cas de la règle adaptative proposée par Dubois et Prade, la fiabilité est directement associée au degré de consensus et, ici aussi, la fusion est organisée selon le degré de consensus entre classifieurs. Dans le cas de sources plus ou moins fiables (la fiabilité n'est plus associée uniquement au degré de consensus mais à une information exogène du même type que celle que nous avons définie pour les intégrales floues), la séquentialité des opérations est organisée autour de cette fiabilité.

Ces deux approches diffèrent de la fusion connexioniste adaptative où aucune séquentialité (*i.e.* règles d'aggrégation) dans la combinaison des classifieurs n'est imposée *a priori*. Cette dernière minimise l'apport d'une connaissance exogène explicite dans le

mécanisme du raisonnement. En effet, la phase d'apprentissage optimise la matrice des poids du réseau, uniquement en fonction des sorties désirées, ce qui rend notre système :

• **adaptatif**, car le raisonnement se construit de façon automatique, au cours d'une phase d'apprentissage, adaptant ainsi l'organisation du module de décision à l'application en cours, en regard des performances globales.

• **général**, car la phase d'apprentissage permet à la fusion de se spécifier à l'application considérée,

• **optimisé**, car nous avons élaboré une procédure stochastique permettant de spécifier un sous-espace de paramètres à chaque classifieur, en regard des performances globales du système. Nous avons utilisé l'exploration aléatoire réalisée par l'Algorithme Génétique afin de minimiser l'*a priori* dans la répartition des paramètres sur les méthodes de RdF du Système de Classifieurs.

La validation de l'outil de classement développé s'est effectué sur des données réelles : reconnaissance d'images de scènes d'extérieur, reconnaissance de données issues d'analyse chimique, reconnaissance d'images de poissons de rivière acquises dans une passe à poissons.

L'étude soulève plusieurs questions en cours d'étude au laboratoire. Premièrement, l'initialisation et la construction du perceptron multi-couches fusionneur introduisent des notions *a priori* (comme le mode d'initialisation des poids du réseau, le nombre de cellules cachées, la nature des fonctions d'activations, ...) que nous désirons améliorer. Or nous avons précisé que le mode de fonctionnement classique du MLP est l'hétéro-associativité. Dans la Fusion Connexionniste Dirigée, nous forçons notre MLP fusionneur à fonctionner en mode auto-associatif. Cela ouvre la voie à des fusions fondées sur des techniques d'optimisation par relaxation, comme les réseaux connexionnistes récurrents. Ces méthodes devraient permettre d'améliorer les deux points précédents car elles sont moins sensibles aux initialisations et leur construction est bien régie dans le cadre de la physique statistique.

Deuxièmement, l'Outil d'Analyse Informationnel, et plus particulièrement de la borne de fiabilité z_{01}, doit permettre d'étudier la dérive des paramètres. Des travaux sont en cours pour spécifier certaines méthodes de RF au sein même du système mixte à certaines classe, en fonction de leur propres indices de performance partiels.

Enfin, ce système de classement est actuellement mis en pratique sur une plate-forme industrielle que nous avons conçue : un prototype de trieur automatique de poissons par Vision Artificielle.

Les algorithmes de segmentation et de reconnaissance, validés au laboratoire, sont en cours de parallélisation sur DSP. La mise en place de ce prototype, dans la Halle à Marée du nouveau Port de Pêche de La Rochelle, doit finaliser notre étude, dans un cadre professionnel.

BIBLIOGRAPHIE

[1] J. Albert, F. Ferri, J. Domingo, and M. Vincens. An approach to natural scene segmentation by means of genetic algorithms with fuzzy data. In *Pattern Recognition and Image Analysis – 10(5)*, pages 97–113, september 1990.

[2] Ch. Amboise and G. Govaert. Spatial clustering and the em algorithm. Report, UTC Compiègne, february 1996.

[3] P. Andrey and P. Tarroux. Unsupervised image segmentation using a distributed genetic algorithm. *Pattern Recognition*, 27(5) :659–673, 1994.

[4] C.S. Beightler, D.T. Phillips, and D.J. Wilde. *Foundations of optimization*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1979.

[5] I. Bloch. Fusion de données, ensemble flous et morphologie mathématique en traitement d'images. application à l'imagerie médicale cérébrale et cardio-vasculaire multi-modalités. Habilitation à Diriger des Recherches 007, Ecole Nationale Supèrieure des Telecommunications – Groupe Image, avril 1995.

[6] P. Bonelli and A. Parodi. An efficient classifier system and its experimental comparison with two representative learning methods on three medical domains. In *4th International Conference on Genetic Algorithm*, pages 288–295, may 4-8 1991.

[7] J. Branke. Evolutionary algorithms for neural network design and training. Technical Report 322, Institute AIFB, University of Karlsruhe, september 1995.

[8] N. Castignolles. *Automatisation du comptage et de la reconnaissance des espèces dans les passes à poissons par l'analyse de séquence d'images*. Thèse de doctorat, Université Paul-Sabatier – Toulouse, janvier 1995.

[9] A. Colubi. Comparative studies of two diversity indices based on entropy measures : Gini-simpson's vs shannon's. In *Information Processing and Management of Uncertainty in Knowledge Based Systems*, pages 675–680, july 1996.

[10] I. Couso, P. Gil, and S. Montes. Measure of fuziness and information theory. In *Information Processing and Management of Uncertainty in Knowledge Based Systems*, pages 501–505, july 1996.

[11] H. de Garis. Modular neural evolution for darwin machines. In *IJCNN*, pages 194–197. Washingon, DC, 1990.

[12] K. De Jong and W.M. Spears. Learning concept classification rules using genetic algorithms. In *11 th ICPR IAPR*, pages 651–656. 12th International Joint Conference on Artificial Intelligence, 1992.

[13] A. De Luca and S. Termini. A definition of a non-probabilistic entropy in the settingg of fuzzy sets theory. *Information and control*, 20 :301–312, 1972.

[14] S. Deveughle. *Etude d'une méthode de combinaison graduelle d'informations incertaines dans un cadre possibiliste*. Thse, Université de Technologie de Compiegne, Compiegne, France, 1993.

[15] D. Dubois and H. Prade. *Théorie des possibilités, applications à la représentation des connaissances en informatique*. Masson, Paris, 1988.

[16] D. Dubois and H. Prade. Combination of fuzzy information in the framework of possibility theory. In *M.A.Abidi and R.C.Gonzales editors, Data fusion in Robotics and Machine Intelligence. Academic Press*, pages 481–505, Boston, 1992.

[17] H. Dubois, D. et Prade. Théorie des possibilités : Applications la représentation des connaissances en informatique. In *Masson*, Paris, May 1985.

[18] H. Dubois, D. et Prade. La fusion d'informations imprécises. In *Revue Traitement du Signal Vol. 11, n°6*, pages 447–458, 1994.

[19] A. Fioretto and A. Sgarro. A second step information measure and the uncertainty of bodies of evidence. In *Information Processing and Management of Uncertainty in Knowledge Based Systems*, pages 687–691, july 1996.

[20] C. Fleurent and J.A. Ferland. Algorithmes génétiques hybrides pour l'optimisation combinatoire. Technical report, Université de Montréal-Département d'Informatique et de Recherche Opérationnelle, 1994.

[21] R.G. Gallager. *Information theory and reliable communication*. John Wiley and sons, inc., 1968.

[22] G. Giraudon, P. Garnesson, and P. Montesinos. Messie : un système multi spécialistes en vision. application à l'interprétation en imagerie aérienne. *Traitement du signal*, 9(5) :403–419, 1992.

[23] D.E. Goldberg. *Algorithmes génétiques*. Addison-Wesley, 1994.

[24] Michel Grabisch. Characterization of fuzzy integrals viewed as aggregation operators. In *IEEE Tr.on Fuzzy Systems*, pages 1927–1932, 1994.

[25] M. Grabish. A survey of application of fuzzy measures and integrals. In *5th IFSA Congress*, pages 294–297, Seoul, July 1993.

[26] F.C. A. Groen, K. Ten, A.W.M. Smeulders, and I.T. Young. Human chromosome classification based on local band descriptors. *Pattern Recognition Letters*, 9(3) :211–222, 1989.

[27] Tahani H. and Keller James M. Information fusion in computer vision using the fuzzy integral. In *IEEE Tr.on Systems, man and Cybernetics, Vol. 20, n° 3*, pages 733–741, May 1990.

[28] J. Heistermann. A mixed genetic approach to the optimization of neural controllers. In *COMPEURO92 : Computer Systems and Software Engineering*, pages 459–464, may 4-8 1992.

[29] A. Hill and C.J. Taylor. Model-based image interpretation using genetic algorithm. In *Image and Vision Computing*, pages 295–301, june 1991.

[30] J.H. Holland. *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, 1975.

[31] J.F. Jodouin. *Les réseaux neuromimétiques. Modèles et applications*. Hermès, 1993.

[32] L. N. Kanal. On pattern, categories and alternative realities. *Pattern RecognitionLetters*, 14(3) :241–255, 1993.

[33] A. Kaufmann. *Introduction à la théorie des sous-ensembles flous - Applications à la linguistique, à la logique et à la sémantique*, volume 2. Masson, Paris, 1975.

[34] G.D. Kleiter and R. Jirousek. A maximum entropy approach for optimal statistical classification. In *IEEE Workshop on Neural Network for Signal Processing*, 1995.

[35] B. Kosko. *Neural networks and fuzzy systems. A dynamical systems approach to machine intelligence*. Prentice Hall International Editions, 1992.

[36] S. Kullback. *Information theory and statistics*. Peter smith edition, 1978.

[37] G.C. Langelaar and J.C.A. van der Lubbe. Adding inputs to existing neural networks. In *Information Processing and Management of Uncertainty in Knowledge Based Systems*, pages 1085–1090, july 1996.

[38] V. Lefèvre, Y. Pollet, S. Philipp, and S. Brunessaux. Un système multi - agents pour la fusion de données en analyse d'images. *Traitement du signal*, 13(1) :99–112, 1996.

[39] K.S. Leung, Y. Leung, and K.F. Yam. Rule learning in expert systems using genetic algorithm : 1, concepts. In *2nd International Conference on fuzzy logic and neural networks*, pages 201–204. Iizuka, Japan, july 1992.

[40] H. Li, Y. Gong, and J.P. Haton. Apprentissage par maximum d'information mutuelle pour des modèles neuronaux probabilistes. *Dizième Congrès Reconnaissance des Formes et Intelligence Artificielle (Rennes)*, 2 :1043–1050, 15-19 janvier 1996.

[41] P. Loonis. *Contribution à la minimisation de l'*a priori *en Reconnaissance des Formes. Conception d'un prototype de trieuse automatique de poissons par Vision Artificielle en milieu industriel*. Thèse de doctorat, Université de La Rochelle, janvier 1996.

[42] P. Loonis, M. Ménard, and J.P. Bonnefoy. Fusion d'information multi-sources : étude comparative entre une approche connexionniste dirigée et la règle orthogonale de dempster-shafer. *Dizième Congrès Reconnaissance des Formes et Intelligence Artificielle (Rennes)*, 2 :606–614, 15-19 janvier 1996.

[43] P. Loonis, M. Ménard, and C. Demko. A new genetic algorithm for the multi-classifiers fusion optimization. *Information Processing and Management of Uncertainty in Knowledge Based Systems*, 2 :957–961, july 1996.

[44] R.J. Machado and A. F. da Rocha. Evolutive fuzzy neural networks. *IEEE NN*, pages 493–500, 1992.

[45] T. Matsuyama and V. Hwang. Sigma : a framework for image understanding. integration of bottom-up and top-down analyses. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3) :908–915, 1992.

[46] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.

[47] D. Miller, A. Rao, K. Rose, and A. Gersho. Learning bayesian networks under the conrol of mutual information. In *Information Processing and Management of Uncertainty in Knowledge Based Systems*, pages 985–990, july 1996.

[48] T. Morimoto, J. De Baerdemaeker, and Y. Hashimoto. Optimization of storage system of fruits usinf neural networks and genetic algorithms. In *IEEE International Conference on Fuzzy Systems*, pages 289–294. Jarmo T. Alander, 1995.

[49] T. Murofushi and M. Sugeno. Fuzzy t-conorm integrals with respect to fuzzy measures : generalization of sugeno integral and choquet integral. In *Fuzzy Sets et Systems. Vol. 42*, pages 57–71, 1991.

[50] M. Ramdani. *Système dinduction formelle base de connaissances imprécises*. PhD thesis, Université Paris 6, février 1994.

[51] M. Roux and J. Desachy. Information fusion for supervised classification in a satellite image. In *Proceedings of the Fourth IEEE International Conference on Fuzzy Systems. Vol. 3*, pages 1119–1124, Yokohama, Japan, 1995.

[52] K. Sengupta and K.L. Boyer. Information theoretic clustering of large structural databases. *IEEE NN*, pages 174–179, 1993.

[53] I. K. Sethi. Entropy nets : From decision trees to neural networks. *Proceedings of the IEEE*, 78(10) :1605–1613, 1990.

[54] G.A. Shafer. A mathematical theory of evidence. In *NJ :Princeton Univ.Press*, 1976.

[55] G. Venturini. *Apprentissage adaptatif et Apprentissage supervisé par Algorithme Génétique*. Thèse de doctorat, Université Paris-sud, janvier 1994.

[56] L. Wehenkel. On uncertainty measures used for decision tree induction. In *Information Processing and Management of Uncertainty in Knowledge Based Systems*, pages 413–418, july 1996.

[57] S.T. Wierzchon. On fuzzy measure and fuzzy integral. In *Fuzzy Information and Decision Processes. Eds New York :North-Holland*, pages 79–86, 1982.

[58] L. Xu, A. Krzyzak, and Suen C.Y. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3) :418–435, 1992.

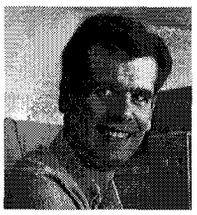
[59] R.R. Yager. Connectives and quantifiers in fuzzy sets. *Fuzzy Sets and Systems*, 40 :39–75, 1991.

[60] L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. In *Fuzzy Sets ans Systems,1*, pages 3–28, 1978.

Manuscrit reçu le 5 décembre 1996.

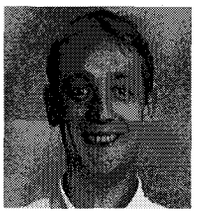
LES AUTEURS

Pierre LOONIS



Pierre Loonis, né en 1968, obtient un Doctorat en Génie Informatique, Signaux et Images à l'Université de La Rochelle en 1996. Ses recherches actuelles portent sur la fusion d'informations multi-sources, principalement dans le domaine de la reconnaissance d'objets naturels, et sur l'évaluation de la qualité des différentes étapes de traitement et d'analyse d'images.

Michel MENARD



Michel Ménard est né en 1966. Docteur en Electronique de l'Université de Poitiers en 1993, il mène des recherches en traitement et en analyse d'images avec une attention particulière pour la reconnaissance d'objets naturels. Les domaines étudiés sont la modélisation floue d'objets et la fusion multi-sources au sein du groupe RAI du L3i.