Application du Modèle des Croyances Transférables en Reconnaissance de Formes[1]

Application of the Transferable Belief Model to Pattern recognition

Thierry Denœux

U.M.R CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France
email: `Thierry.Denoeux@hds.utc.fr`

1. A Paraître dans *Traitement du Signal*, 1997.

Résumé

Nous présentons dans cet article une nouvelle approche de la Reconnaissance de Formes basée sur le Modèle des Croyances Transférables, une interprétation non probabiliste de la théorie des fonctions de croyance de Dempster et Shafer. Le principe de cette méthode consiste à caractériser sous la forme d'une fonction de croyance l'information apportée par un ensemble d'apprentissage, relativement à la classe d'un nouveau vecteur. Différentes stratégies de décision avec coûts arbitraires, généralisant l'approche bayésienne, sont présentées et illustrées à l'aide d'un exemple.

We present in this paper a new approach to Pattern Recognition based on the Transferable Belief Model, a non probabilistic interpretation of Dempster and Shafer's theory of belief functions. This method uses the formalism of belief functions to represent the information brought by the training set concerning the class of a new pattern. Various decision strategies generalizing Bayes decision theory and presented and demonstrated using a real example.

Mots-clés : fonctions de croyance, théorie de Dempster et Shafer, reconnaissance de formes, discrimination, décision, fusion de données.

Key-words: *belief functions, Dempster-Shafer theory, pattern recognition, classification, decision, data fusion.*

Catégorie : Recherche

1 Introduction

Le problème de la reconnaissance de formes en mode supervisé, ou discrimination, est généralement posé de la manière suivante [11]. On considère une population $\mathcal{P}$ composée d'individus répartis en M groupes ou classes $\omega_1, \ldots, \omega_M$. Chaque individu est caractérisé par un *vecteur forme* $\mathbf{x}$ composé de d variables quantitatives ou qualitatives, et par une variable qualitative z à valeurs dans $\Omega = \{\omega_1, \ldots, \omega_M\}$ indiquant son appartenance à l'un des groupes. On dispose d'un ensemble d'apprentissage $\mathcal{X} = \{(\mathbf{x}^{(1)}, z^{(1)}), \ldots, (\mathbf{x}^{(N)}, z^{(N)})$ relatif à N individus extraits de $\mathcal{P}$. Il s'agit d'en déduire une règle de classement d'un individu quelconque dont seul le vecteur forme est connu.

Ce problème est habituellement abordé dans une perspective probabiliste. Les individus sont supposés prélevés aléatoirement parmi la population d'origine, ce qui permet de considérer chaque observation $(\mathbf{x}, z)$ comme une réalisation d'un couple de variables aléatoires $(\mathbf{X}, Z)$ caractérisé par une certaine loi de probabilité [1]. Lorsque cette loi est connue, le problème du choix d'une règle de classement est résolu dans le cadre de la théorie bayésienne de la décision, la règle optimale consistant à choisir, parmi un ensemble fini d'actions (affectation à l'une des classes ou rejet), celle dont le risque conditionnel est le plus faible [8].

En pratique, cependant, la loi du couple $(\mathbf{X}, Z)$ n'est pas connue, et les probabilités conditionnelles $P(z = \omega_i|\mathbf{x})$ doivent être estimées par une méthode paramétrique ou non paramétrique [8, 11]. On justifie généralement le choix d'un estimateur par ses propriétés asymptotiques, c'est-à-dire par son comportement lorsque N tend vers l'infini.

Ce modèle général présente l'avantage de reposer sur une base théorique bien établie et a donné lieu au développement de méthodes très performantes. Néanmoins, il possède un domaine d'application limité de par la nature des hypothèses assez restrictives sur lesquelles il repose. En particulier, son application est difficile dans les situations où la connaissance disponible est relativement pauvre. Par exemple, lorsque les exemples d'apprentissage sont produits dans des conditions variables, mal connues ou non reproductibles, il est permis de mettre en doute la pertinence de la notion de loi de probabilité comme modèle de génération des exemples. Par ailleurs, la connaissance des classes des éléments de l'ensemble d'apprentissage peut être entachée de différentes formes d'incertitude mal prises en compte par la théorie classique de l'estimation. Enfin, les estimations des probabilités conditionnelles produites en sortie du classifieur ne reflètent qu'imparfaitement l'incertitude qui subsiste quant à la classe du vecteur forme considéré, particulièrement dans le cas où, l'échantillon étant de petite taille, on est très loin des conditions asymptotiques.

Ces remarques nous ont conduit à proposer une approche différente reposant sur la théorie des fonctions de croyance [4]. Cette approche vise à répondre à la question suivante: Etant donné un ensemble d'apprentissage de taille finie, comment caractériser l'incertitude quant au groupe d'origine d'un nouveau vecteur? La méthode proposée consiste à construire une structure de croyance m prenant en compte à la fois l'incertitude relative aux exemples d'apprentissage, et la plus ou moins grande proximité de ces exemples à l'individu à classer.

Cet article commence par un bref rappel des principaux concepts de la théorie des fonctions de croyance et de leur interprétation dans le cadre du Modèle des Croyances Transférables proposé par Smets [14, 15]. Nous présentons ensuite l'application de ces concepts en reconnaissance des formes, en abordant successivement les problèmes de construction d'une fonction de croyance, puis de prise de décision. La méthode proposée est ensuite illustrée à l'aide de données réelles issues d'une

1. On peut également postuler directement l'existence d'une telle loi de probabilité comme modélisation d'un processus physique de génération des couples $(\mathbf{x}, z)$, ce processus ne consistant pas nécessairement en un échantillonnage parmi une certaine population.

application de surveillance de qualité de l'eau de Seine en amont d'une usine de traitement d'eau potable.

2 Théorie des fonctions de croyance

La théorie des fonctions de croyance est issue des travaux de Dempster sur les bornes inférieure et supérieure d'une famille de distributions de probabilités [2]. Elle a été développée par Shafer [13] qui a montré l'intérêt des fonctions de croyance comme formalisme de représentation de l'incertitude. Smets a proposé une justification axiomatique cohérente des principaux concepts sous la forme du Modèle des Croyances Transférables (*Transferable Belief Model*) [14, 15], et a clarifié les liens entre représentation des croyances et prise de décision [15]. Nous adopterons ce point de vue dans la suite de cet article.

2.1 Représentation des croyances

Le problème considéré est le suivant. On s'intéresse à la valeur prise par un paramètre ω dont les valeurs possibles constituent un ensemble fini Ω. On suppose que ω ne peut prendre qu'une valeur et une seule dans Ω. On se place donc à la fois dans l'hypothèse du monde fermé [2] et dans le cadre de la logique classique, une proposition (encore appelée hypothèse) étant soit vraie, soit fausse. Les propositions considérées sont toutes de la forme "la vraie valeur de ω est dans A", où A est une partie de Ω. Une telle proposition peut sans difficulté être assimilée au sous-ensemble A lui-même, et l'ensemble des propositions de ce type à l'ensemble 2^Ω des parties de Ω. Il y a ainsi correspondance entre les concepts logiques de conjonction, disjonction, implication et négation, et les concepts ensemblistes d'intersection, union, inclusion et complémentation [13].

Un agent rationnel est supposé détenir à un instant t, sur la base d'un corps de connaissances, une opinion caractérisée par un degré de croyance en chaque hypothèse. On postule que ces degrés de croyance peuvent être décrits par une *structure de croyance* m définie comme une fonction de 2^Ω dans $[0,1]$ vérifiant :

$$m(\emptyset) = 0 \tag{1}$$

$$\sum_{A\subseteq\Omega} m(A) = 1 \tag{2}$$

La quantité $m(A)$ s'interprète comme la quantité de croyance placée strictement en A, et qui ne peut être allouée à aucune autre hypothèse plus restrictive, faute d'information suffisante. Toute partie A de Ω telle que $m(A) > 0$ est appelée *élément focal* de A. On définit à partir de la fonction m des fonctions de *crédibilité* et de *plausibilité* par :

$$\mathrm{bel}(A) = \sum_{B\subseteq A} m(B) \tag{3}$$

$$\mathrm{pl}(A) = \sum_{B\cap A\neq\emptyset} m(B) = 1 - \mathrm{bel}(\bar{A}) \tag{4}$$

bel(A) représente le degré de croyance en A, compte tenu de tous les éléments qui accréditent, directement ou non, cette hypothèse. pl(A) s'interprète comme la part de croyance qui pourrait potentiellement être allouée à A, compte tenu des éléments

2. Il est possible, dans le cadre du Modèle des Croyances Transférables, de s'affranchir de cette condition en admettant que l'ensemble Ω puisse ne pas être exhaustif (hypothèse du monde ouvert). Nous préférons conserver l'hypothèse d'exhaustivité de Ω, quitte à introduire explicitement dans Ω un élément supplémentaire représentant l'ensemble des valeurs "inconnues" de ω.

qui ne discréditent pas cette hypothèse. Les trois fonctions m, bel et pl constituent trois représentations d'une même information, la donnée de l'une d'elles permettant de retrouver les deux autres. On peut également leur associer une famille $\mathcal{C}$ de distributions de probabilité P dites *compatibles*[3] vérifiant $\mathrm{bel}(A) \leq P(A) \leq \mathrm{pl}(A)$ pour toute partie A de Ω.

La partie *dynamique* du Modèle des Croyances Transférables concerne la révision des croyances suite à la prise de connaissance de nouvelles informations. Le mécanisme de base est la règle de conditionnement de Dempster [14]. Ayant défini une structure de croyance m, supposons que l'on vienne à apprendre que la proposition $\omega \in B \subseteq \Omega$ est vraie. On en déduit une nouvelle structure de croyance $m(\cdot|B)$ en *transférant* la part de croyance initialement allouée à A vers $A \cap B$, pour tout $A \subseteq \Omega$ t.q. $A \cap B \neq \emptyset$. On a donc :

$$m(A|B) = \begin{cases} c^{-1} \sum_{X \subseteq \bar{B}} m(A \cup X) & \text{si } A \subseteq B \text{ et } A \neq \emptyset, \\ 0 & \text{si } A \not\subseteq B \text{ ou } A = \emptyset \end{cases} \tag{5}$$

avec $c = 1 - \sum_{X \subseteq \bar{B}} m(X)$.

Supposons maintenant que l'on dispose d'une structure de croyance m_1 et que l'on prenne connaissance de nouvelles informations qui, considérées isolément, induisent une structure de croyance m_2. On définit la *somme orthogonale* de m_1 et m_2, notée $m = m_1 \oplus m_2$ par :

$$m(\emptyset) = 0 \tag{6}$$

$$m(A) = K^{-1} \sum_{B \cap C = A} m_1(B) m_2(C) \quad \forall A \subseteq \Omega, A \neq \emptyset \tag{7}$$

avec $K = 1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$. La somme m n'est définie que si les fonctions m_1 et m_2 sont combinables, c'est-à-dire si, pour au moins deux sous-ensembles B et C non disjoints de Ω, on a à la fois $m_1(B) > 0$ et $m_2(C) > 0$. Smets [14] a montré que cette règle, appelée *règle de combinaison de Dempster*, se déduit de manière unique de la règle de conditionnement introduite ci-dessus et d'un petit nombre d'axiomes.

2.2 Principes de décision

En situation d'incertitude, la théorie Bayésienne de la décision préconise l'action pour laquelle l'espérance du coût est la plus faible. Lorsque l'incertitude est caractérisée non plus par une distribution de probabilité mais par une fonction de croyance, plusieurs généralisations sont envisageables. Une première approche consiste à étendre la notion d'espérance mathématique dans le cadre des fonctions de croyance, ce qui peut être fait de la manière suivante.

Soit une fonction $f : \Omega \mapsto \mathbb{R}$ et une distribution de probabilité P sur Ω. L'espérance de f relativement à P est :

$$E(f, P) = \sum_{\omega \in \Omega} f(\omega) P(\omega) \tag{8}$$

On définit les espérances inférieure (*lower expectation*) et supérieure (*upper expectation*) de f relativement à une fonction de croyance bel respectivement comme les bornes inférieure et supérieure des espérances de f pour toutes les distributions de probabilité compatibles avec bel [2] :

$$E_*(f) = \min_{P \in \mathcal{C}} E(f, P) \tag{9}$$

$$E^*(f) = \max_{P \in \mathcal{C}} E(f, P) \tag{10}$$

3. Remarquons que cette définition ne présuppose nullement l'existence de probabilités *objectives* dont découlerait la notion de fonction de croyance.

On montre que $E_*(f)$ et $E^*(f)$ peuvent être obtenues par les équations suivantes :

$$E_*(f) = \sum_{A \subseteq \Omega} m(A) \min_{\omega \in A} f(\omega) \quad (11)$$

$$E^*(f) = \sum_{A \subseteq \Omega} m(A) \max_{\omega \in A} f(\omega) \quad (12)$$

Une généralisation de la théorie bayésienne de la décision consiste à s'appuyer sur ces définitions, en proposant comme principes de décision la minimisation de l'espérance inférieure ou de l'espérance supérieure du coût. La première approche correspond plutôt à une attitude optimiste, la seconde à une attitude pessimiste.

Une approche différente, proposée par Smets [15], consiste à choisir dans $\mathcal{C}$ une distribution de probabilité particulière BetP, dite "pignistique", obtenue en répartissant la masse de croyance $m(A)$ à parts égales[4] entre les éléments de A, pour tout $A \subseteq \Omega$. On a donc :

$$\mathrm{BetP}(\omega) = \sum_{A \ni \omega} \frac{m(A)}{|A|} \quad (13)$$

pour tout $\omega \in \Omega$, $|A|$ étant le cardinal de $A \subseteq \Omega$. On définit ensuite l'espérance pignistique de f comme son espérance mathématique relativement à BetP :

$$E_{bet}(f) = \sum_{\omega \in \Omega} f(\omega)\mathrm{BetP}(\omega) \quad (14)$$

Notons qu'il n'y a aucune contradiction dans le fait de revenir à un modèle probabiliste au moment de la prise de décision, si l'on opère une distinction entre deux niveaux : un niveau "cognitif" relatif à la représentation d'un système de croyances, et un niveau décisionnel pour lequel l'utilisation des probabilités permet la prise de décision de manière cohérente et non ambiguë [14]. Remarquons enfin que l'on retrouve l'approche précédente basée sur les espérances inférieure et supérieure du coût, si l'on considère l'ensemble des distributions pignistiques induites par toutes les spécialisations (*refinements*) de Ω [16].

3 Application en discrimination

La plupart des travaux visant à appliquer la théorie des fonctions de croyance en reconnaissance des formes ont consisté jusqu'à présent à convertir les sorties de classifieurs classiques en structures de croyance, en utilisant par exemple une estimation du taux d'erreur de classification [17], la distance à la moyenne des vecteurs de sortie dans chaque classe [12], ou une interprétation des sorties en termes de vraisemblance [1, 10]. La méthode décrite ci-dessous diffère des approches précédentes en ce qu'elle repose exclusivement sur le principe de représentation des croyances par une fonction de crédibilité, et ne fait appel à aucune notion probabiliste [4]. Nous présentons ci-dessous les deux aspects principaux de la méthode : la construction d'une structure de croyance, et les principes de décision.

4. Cette opération peut être vue comme l'application du Principe de Raison Insuffisante, qui consiste à choisir une loi de probabilité uniforme en l'absence d'information. Étendu à chaque élément focal d'une fonction de croyance, ce principe est appelé Principe de Raison Insuffisante Généralisé par Smets [15].

3.1 Construction d'une structure de croyance

3.1.1 Principes généraux

Soit une population $\mathcal{P}$ partitionnée en M classes $\omega_1, \ldots, \omega_M$. Soit un ensemble d'apprentissage $\mathcal{X} = \{(\mathbf{x}^{(1)}, m^{(1)}), \ldots, (\mathbf{x}^{(N)}, m^{(N)})$ où $\mathbf{x}^{(i)}$ est le vecteur forme [5] d'un individu i et $m^{(i)}$ est une structure de croyance sur $\Omega = \{\omega_1, \ldots, \omega_M\}$ caractérisant notre état de connaissance relativement à la classe $z^{(i)}$ du même individu [6]. Dans le cas particulier où cette connaissance est certaine, on a $m^{(i)}(\{\omega_j\}) = 1$ pour un élément ω_j de Ω. On suppose connue une mesure de dissimilarité δ caractérisant de manière pertinente le degré de dissemblance entre deux vecteurs.

Considérons maintenant une nouvelle entité de vecteur forme $\mathbf{x}$ connu et de classe z inconnue. La prise en compte d'un élément $e^{(i)} = (\mathbf{x}^{(i)}, m^{(i)})$ de $\mathcal{X}$ induit pour toute partie stricte A de Ω un degré de croyance $m(A|e^{(i)})$ dans la proposition $z \in A$ au plus égal au degré de croyance $m^{(i)}(A)$ dans la proposition $z^{(i)} \in A$. Plus précisément, on pose :

$$m(A|e^{(i)}) = \varphi(\delta(\mathbf{x}, \mathbf{x}^{(i)}))m^{(i)}(A) \quad \forall A \subset \Omega \tag{15}$$

$$m(\Omega|e^{(i)}) = 1 - \sum_{A \subset \Omega} m(A|e^{(i)}) \tag{16}$$

où φ est une fonction décroissante vérifiant $\varphi(0) \leq 1$, et $\lim_{d \to \infty} \varphi(d) = 0$. La structure de croyance $m(\cdot|e^{(i)})$ s'obtient donc par *affaiblissement* [13] de $m^{(i)}$, le facteur d'affaiblissement étant fonction croissante de la dissimilarité entre les vecteurs $\mathbf{x}$ et $\mathbf{x}^{(i)}$. La condition $\lim_{d \to \infty} \varphi(d) = 0$ traduit le fait que la structure de croyance $m(\cdot|e^{(i)})$ tend vers la structure vide ($m(\Omega|e^{(i)}) = 1$) lorsque la dissimilarité entre les vecteurs $\mathbf{x}$ et $\mathbf{x}^{(i)}$ tend vers l'infini. La structure vide étant l'élément neutre de l'opération de somme orthogonale, une observation $\mathbf{x}^{(i)}$ très éloignée de $\mathbf{x}$ n'aura qu'une influence négligeable sur la croyance en la classe de l'entité décrite par $\mathbf{x}$.

Le choix de la mesure de dissimilarité δ et de la fonction φ traduit en principe une connaissance a priori sur le domaine d'application et doit être laissé à l'utilisateur. Dans le cas où δ est la distance euclidienne, on peut justifier une forme particulière pour φ de la manière suivante. Supposons tout d'abord que le vecteur $\mathbf{x}$ contienne toute l'information utile pour le classement de l'entité correspondante. Il existe donc une fonction déterministe $\Psi : \mathbb{R}^d \mapsto \Omega$ qui à tout vecteur associe une classe. Il est logique dans ce cas de caractériser par une même structure de croyance notre connaissance relativement à la classe de deux entités décrites par le même vecteur forme, d'où $\varphi(0) = 1$. Par ailleurs, supposons que le vecteur $\mathbf{x}$ se décompose en deux parties $\mathbf{x} = [\mathbf{x}' \ \mathbf{x}'']^t$ avec $\mathbf{x}' \in \mathbb{R}^{d'}$, $\mathbf{x}'' \in \mathbb{R}^{d''}$ et $d' + d'' = d$. Chaque élément $e^{(i)}$ de l'ensemble d'apprentissage se décompose de même en deux parties $e'^{(i)} = (\mathbf{x}'^{(i)}, m^{(i)})$ et $e''^{(i)} = (\mathbf{x}''^{(i)}, m^{(i)})$. Si l'on connaît seulement la composante $\mathbf{x}'$ de $\mathbf{x}$, on déduira de l'élément $e'^{(i)}$ une structure de croyance $m(\cdot|e'^{(i)})$ définie par $m(A|e'^{(i)}) = \varphi(\|\mathbf{x}' - \mathbf{x}'^{(i)}\|)m^{(i)}(A)$ pour tout $A \subset \Omega$. Si par la suite la composante $\mathbf{x}''$ de $\mathbf{x}$ vient à être connue, il faudra pour tenir compte de la différence entre $\mathbf{x}''$

5. Nous supposons ici, comme il est d'usage en reconnaissance de formes statistique, que chaque individu est décrit par un vecteur de caractéristiques. Toutefois, notre approche pourrait facilement être étendue à d'autres formalismes de représentation, de nature symbolique par exemple.

6. Le problème de l'obtention en pratique d'une telle information auprès d'un expert relève de la problématique générale de l'acquisition de connaissances et dépasse le cadre de cet article. Dans de nombreux cas, on pourra se restreindre à des structures de croyance possédant une forme relativement simple. Par exemple, le choix d'un seul élément focal : $m^{(i)}(A) = 1$ pour $A \subseteq \Omega$ traduit la connaissance du fait que la classe de l'entité décrite par le vecteur $\mathbf{x}^{(i)}$ appartient certainement à A. On peut supposer qu'une telle information puisse facilement être obtenue auprès d'un expert, particulièrement lorsque les classes sont organisées selon une structure hiérarchique, comme il arrive fréquemment dans les problèmes de diagnostic.

et $\mathbf{x}''^{(i)}$ procéder à un affaiblissement supplémentaire en posant pour tout $A \subset \Omega$:

$$\begin{aligned} m(A|e'^{(i)}, e''^{(i)}) &= \varphi(\|\mathbf{x}'' - \mathbf{x}''^{(i)}\|) m^{(i)}(A|e'^{(i)}) && (17) \\ &= \varphi(\|\mathbf{x}'' - \mathbf{x}''^{(i)}\|)\varphi(\|\mathbf{x}' - \mathbf{x}'^{(i)}\|) m^{(i)}(A) && (18) \end{aligned}$$

Si l'on impose que $m(\cdot|e^{(i)}) = m(\cdot|e'^{(i)}, e''^{(i)})$, il faut donc que :

$$\varphi(\|\mathbf{x} - \mathbf{x}^{(i)}\|) = \varphi(\|\mathbf{x}'' - \mathbf{x}''^{(i)}\|)\varphi(\|\mathbf{x}' - \mathbf{x}'^{(i)}\|) \qquad (19)$$

Sachant que $\|\mathbf{x} - \mathbf{x}^{(i)}\|^2 = \|\mathbf{x}'' - \mathbf{x}''^{(i)}\|^2 + \|\mathbf{x}' - \mathbf{x}'^{(i)}\|^2$, la solution de cette équation fonctionnelle est $\varphi(y) = \exp(-\gamma y^2)$ avec $\gamma > 0$. En pratique, le vecteur forme ne contient que rarement toute l'information discriminante : deux vecteurs identiques peuvent donc représenter des entités appartenant à des classes distinctes, ce qui conduit à poser $\varphi(0) = \alpha < 1$. On est donc conduit, dans le cas où la dissemblance entre deux vecteurs formes est correctement mesurée par la distance euclidienne, à choisir φ de la forme $\varphi(y) = \alpha \exp(-\gamma y^2)$.

Jusqu'à présent, nous n'avons considéré pour le classement de $\mathbf{x}$ qu'un seul élément de $\mathcal{X}$. Si nous répétons la même opération pour les N exemples d'apprentissage, nous obtenons N structures de croyance $m(\cdot|e^{(i)}), i = 1, \ldots, N$. Ces N structures sont issues de sources d'information distinctes et peuvent donc être combinées par la règle de Dempster [7]. On pose donc finalement :

$$m = m(\cdot|e^{(1)}) \oplus \ldots \oplus m(\cdot|e^{(N)}) \qquad (20)$$

En pratique, les éléments caractérisés par des vecteurs formes très éloignés de $\mathbf{x}$ influent peu sur le résultat et peuvent être négligés. On pourra donc, par exemple, se contenter de ne prendre en compte dans la somme orthogonale que les k plus proches voisins de $\mathbf{x}$ [4].

3.1.2 Méthodes pratiques

Dans la forme générale indiquée ci-dessus, la fonction φ définissant l'influence de la distance sur le facteur d'affaiblissement dépend de deux paramètres α et γ. D'une manière générale, on peut choisir φ parmi une famille paramétrée de fonctions. Soit $\mathbf{w}$ le vecteur de paramètres. La structure de croyance calculée pour $\mathbf{x}$ est donc fonction de $\mathbf{x}$, de $\mathcal{X}$ et de $\mathbf{w}$.

Dans [18], nous avons suggéré la méthode suivante pour optimiser le paramètre $\mathbf{w}$, dans le cas où la classe des exemples d'apprentissage est connue avec certitude. Soit $\mathbf{t}^{(i)} = (t_1^{(i)}, \ldots, t_N^{(i)})^t$ le vecteur d'appartenance pour l'exemple i, défini par $t_j^{(i)} = 1$ si cet exemple appartient à la classe ω_j, et $t_j^{(i)} = 0$ sinon. On calcule pour cet exemple i une structure de croyance $m^{(i)}$ en utilisant les autres exemples de l'ensemble d'apprentissage, et on note $\mathbf{P}^{(i)}$ le vecteur des probabilités pignistiques associées. On postule que les vecteurs $\mathbf{t}^{(i)}$ et $\mathbf{P}^{(i)}$ doivent être aussi proches que possible, en prenant par exemple comme critère d'erreur $E^{(i)} = \|\mathbf{t}^{(i)} - \mathbf{P}^{(i)}\|^2$. On cherche ensuite la valeur $\mathbf{w}^*$ du paramètre qui minimise l'erreur moyenne sur l'ensemble d'apprentissage. Cette méthode s'avère très performante, comparée à d'autres méthodes basées sur le principes des k plus proches voisins [18].

Dans [3], nous avons généralisé cette approche au cas où l'ensemble d'apprentissage est résumé sous la forme de n vecteurs de référence ou prototypes. Le même

7. Une remarque s'impose ici sur la complexité calculatoire de l'opération de somme orthogonale, qui pourrait poser problème dans certaines applications (peu nombreuses) où le nombre de classes est très grand. Il faudrait alors soit se limiter à des structures de croyance de forme particulière (à support simple par exemple), soit avoir recours à des algorithmes efficaces tels que la transformée de Möbius rapide [9].

principe de minimisation d'un critère d'erreur permet alors d'optimiser simultanément la position des prototypes, les degrés d'appartenance des prototypes aux différentes classes, et les paramètres de la fonction φ. Les grandes lignes de cette méthode sont résumées en Annexe.

3.2 Principes de décision

Nous supposons maintenant que l'incertitude relative à la classe d'un individu décrit par un vecteur forme $\mathbf{x}$ est représentée par une structure de croyance m définie sur l'ensemble des classes Ω, et nous considérons le problème du choix d'une action parmi un ensemble fini $\mathcal{A}$. La mise en œuvre d'une action α_i lorsque l'individu appartient en réalité à la classe ω_j est supposée entraîner un coût noté $\lambda(\alpha_i|\omega_j)$. On peut donc définir pour chaque action α_i trois risques conditionnels :

$$
\begin{aligned}
R_*(\alpha_i|\mathbf{x}) &= \sum_{A\subseteq\Omega} m(A) \min_{\omega_j\in A} \lambda(\alpha_i|\omega_j) && (21)\\
R^*(\alpha_i|\mathbf{x}) &= \sum_{A\subseteq\Omega} m(A) \max_{\omega_j\in A} \lambda(\alpha_i|\omega_j) && (22)\\
R_{bet}(\alpha_i|\mathbf{x}) &= \sum_{\omega_j\in\Omega} \lambda(\alpha_i|\omega_j) \sum_{A\ni\omega_j} \frac{m(A)}{|A|} && (23)\\
&= \sum_{A\subseteq\Omega} m(A) \frac{1}{|A|} \sum_{\omega_j\in A} \lambda(\alpha_i|\omega_j) && (24)
\end{aligned}
$$

Ces trois risques correspondent respectivement à l'espérance inférieure, à l'espérance supérieure, et à l'espérance relativement à la distribution pignistique du coût associé au choix de l'action α_i. Ces définitions suggèrent trois stratégies différentes de décision : une stratégie "optimiste" de minimisation du risque inférieur, une stratégie "pessimiste" de minimisation du risque supérieur, et une stratégie "neutre" de minimisation du risque pignistique [5]. Le choix d'une stratégie particulière peut être dicté par une attitude générale de l'utilisateur face au risque en présence d'informations incomplètes.

Afin de mettre en évidence l'intérêt de ces règles de décision dans un contexte de reconnaissance de formes, nous proposons de considérer les deux cas particuliers suivants.

Cas 1 : Les classes sont toutes connues. On note $\Omega = \{\omega_1, \ldots, \omega_M\}$ l'ensemble des classes, et $\mathcal{A} = \{\alpha_0, \alpha_1, \ldots, \alpha_M\}$ l'ensemble des actions, α_i désignant l'affectation à la classe ω_i pour $i \in \{1, \ldots, M\}$, et α_0 l'option de rejet. Les coûts sont supposés égaux à 0 pour une bonne classification, à 1 pour une mauvaise classification, et à une constante λ_0 pour le rejet. On a donc $\lambda(\alpha_i|\omega_j) = 1-\delta_{i,j}$ pour $i, j \in \{1, \ldots, M\}$ et $\lambda(\alpha_0|\omega_j) = \lambda_0$ pour $j \in \{1, \ldots, M\}$. Dans ces conditions, on obtient pour chaque action les risques suivants :

$$
\begin{aligned}
R_*(\alpha_i|\mathbf{x}) &= 1 - \sum_{A\ni\omega_i} m(A) = 1 - \mathrm{pl}(\{\omega_i\}) && (25)\\
R^*(\alpha_i|\mathbf{x}) &= 1 - m(\{\omega_i\}) = 1 - \mathrm{bel}(\{\omega_i\}) && (26)\\
R_{bet}(\alpha_i|\mathbf{x}) &= = \sum_{j\neq i} \mathrm{BetP}(\{\omega_j\}) = 1 - \mathrm{BetP}(\{\omega_i\}) && (27)
\end{aligned}
$$

pour $i \in \{1, \ldots, M\}$, et $R_*(\alpha_0|\mathbf{x}) = R^*(\alpha_0|\mathbf{x}) = R_{bet}(\alpha_0|\mathbf{x}) = \lambda_0$.

Selon la stratégie employée, on choisit donc la classe de plus grande plausibilité, crédibilité ou probabilité pignistique, à condition que la valeur correspondante soit supérieure au seuil $1 - \lambda_0$.

Cas 2 : On admet l'existence de classes inconnues ou non représentées dans l'ensemble d'apprentissage. Cette situation se rencontre par exemple fréquemment dans les application de diagnostic de systèmes complexes [7, 6]. On introduit alors explicitement dans l'ensemble Ω un élément supplémentaire ω_u représentant l'ensemble des classes inconnues : $\Omega = \{\omega_1, \ldots, \omega_M, \omega_u\}$. Aux actions définies précédemment vient par conséquent s'ajouter une action α_u d'affectation à la classe ω_u [8]. Les coûts $\lambda(\alpha_i|\omega_j)$ sont supposés définis comme ci-dessus pour $i \in \{0, \ldots, M\}$ et $j \in \{1, \ldots, M\}$. On pose en plus $\lambda(\alpha_0|\omega_u) = \lambda_0$, $\lambda(\alpha_i|\omega_u) = 1$ pour $i \in \{1, \ldots, M\}$, $\lambda(\alpha_u|\omega_u) = 0$ et $\lambda(\alpha_u|\omega_j) = \lambda_1$ pour $i \in \{1, \ldots, M\}$. On obtient alors les mêmes expressions que précédemment pour le risques associés à l'affectation à une classe connue et à l'action de rejet ; de plus, on a :

$$
\begin{aligned}
R_*(\alpha_u|\mathbf{x}) &= \lambda_1(1 - \mathrm{pl}(\{\omega_u\})) && (28)\\
R^*(\alpha_u|\mathbf{x}) &= \lambda_1(1 - \mathrm{bel}(\{\omega_u\})) && (29)\\
R_{bet}(\alpha_u|\mathbf{x}) &= = \lambda_1(1 - \mathrm{BetP}(\{\omega_u\})) && (30)
\end{aligned}
$$

La classe ω_u étant inconnue, on peut raisonnablement s'attendre à ce que le seul élément focal de m contenant ω_u soit Ω (ceci est notamment le cas lorsque la classe des exemples d'apprentissage est connue avec certitude). On a alors $\mathrm{pl}(\{\omega_u\}) = m(\Omega)$, $\mathrm{bel}(\{\omega_u\}) = 0$ et $\mathrm{BetP}(\{\omega_u\}) = \frac{m(\Omega)}{M+1}$ et on obtient finalement :

$$
\begin{aligned}
R_*(\alpha_u|\mathbf{x}) &= \lambda_1(1 - m(\Omega)) && (31)\\
R^*(\alpha_u|\mathbf{x}) &= \lambda_1 && (32)\\
R_{bet}(\alpha_u|\mathbf{x}) &= = \lambda_1(1 - \frac{m(\Omega)}{M+1}) && (33)
\end{aligned}
$$

L'action α_u présente donc un risque supérieur toujours plus grand ou toujours plus petit que l'action α_0, selon que $\lambda_1 > \lambda_0$ ou $\lambda_1 < \lambda_0$. En revanche, les risques inférieur et pignistique de α_u sont fonctions décroissantes de la masse $m(\Omega)$, qui tend vers 1 lorsque la distance à l'ensemble d'apprentissage tend vers l'infini : la propension à choisir l'affectation à la classe inconnue est donc d'autant plus grande que l'entité à classer diffère davantage des exemples d'apprentissage.

4 Exemple

L'approche décrite ci-dessus a été appliquée à des mesures physico-chimiques de qualité de l'eau de la Seine [6]. Il s'agit d'analyses effectuées quotidiennement à Suresnes pendant une période de deux ans, l'objectif étant de surveiller la qualité de l'eau brute à l'entrée de l'usine de traitement. Les données ont été partitionnées en quatre classes de qualité par un algorithme de classification automatique. Afin de permettre une meilleure visualisation des régions de décision, nous n'avons considéré comme variables d'entrée que les deux premiers axes extraits par analyse en composantes principale. La méthode retenue pour le calcul des structures de croyance est l'algorithme connexionniste décrit dans [3] et résumé en Annexe.

Nous donnons ci-dessous quelques exemples de régions de décision obtenues dans le cas où l'ensemble d'apprentissage est supposé exhaustif (Figure 1), et dans le cas où l'on admet l'existence de classes inconnues (Figure 2).

8. Les actions α_0 et α_u peuvent s'interpréter comme deux formes différentes de rejet appelées respectivement rejet d'*ambiguïté* et rejet *de distance* selon la terminologie introduite en diagnostic [7].

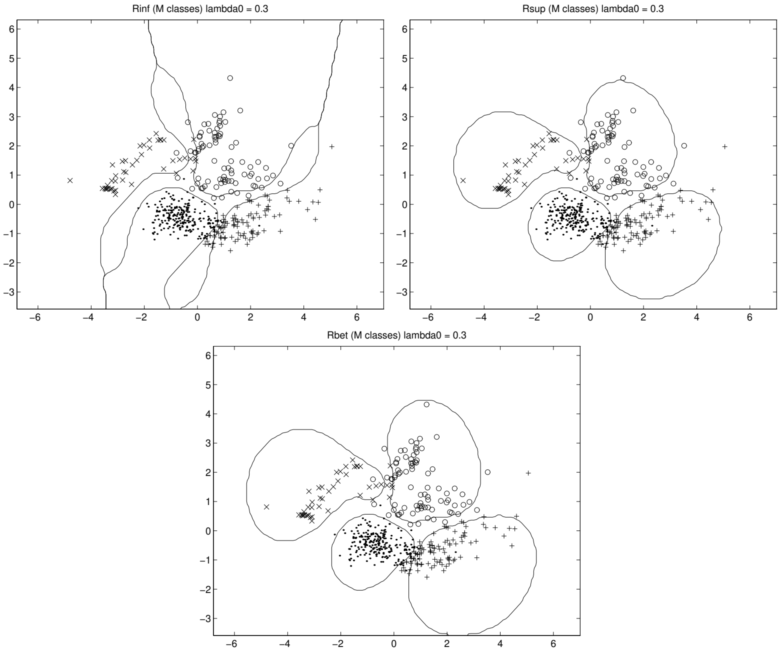

FIG. 1 - *Régions de décision correspondant à la minimisation du risque inférieur (en haut à gauche), supérieur (en haut à droite) et pignistique (en bas), l'ensemble d'apprentissage étant supposé exhaustif ($\lambda_0 = 0.3$).*

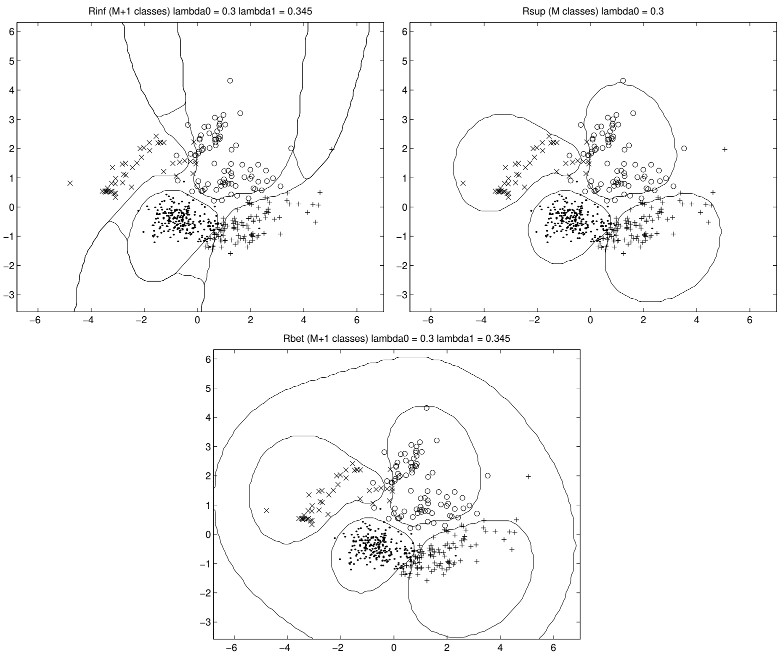

FIG. 2 - *Régions de décision correspondant à la minimisation du risque inférieur (en haut à gauche), supérieur (en haut à droite) et pignistique (en bas), l'ensemble d'apprentissage étant supposé non exhaustif ($\lambda_0 = 0.3$ et $\lambda_1 = 0.345$).*

Dans le premier cas, on remarque que la stratégie de décision optimiste (basée sur R_*) conduit à ne rejeter que les observations situées à la frontière entre deux classes, tandis que la stratégie pessimiste (basée sur R^*) rejette également les observations éloignées de l'ensemble d'apprentissage (Figure 1). Les stratégies pessimiste et neutre donnent des résultats sensiblement identiques dans ce cas.

Sous l'hypothèse d'existence d'une classe inconnue, la stratégie optimiste rejette les individus situés entre les classes, et préconise l'affectation à la classe inconnue lorsqu'il y a à la fois ambiguïté et éloignement par rapport à l'ensemble d'apprentissage. La stratégie pessimiste conduit aux mêmes frontières de décision que précédemment : affectation aux classes connues dans les régions où la densité des observations est élevée, et rejet ou affectation à la classe inconnue (selon le signe de $\lambda_0 - \lambda_1$) hors de ces régions. La stratégie neutre semble donner lieu au comportement le plus conforme à l'intuition : affectation à une classe connue dans les régions "denses" de l'espace de représentation, affectation à la classe inconnue loin de l'ensemble d'apprentissage, et rejet d'ambiguïté dans les zones intermédiaires (Figure 2).

Pour une même fonction de coût, les trois stratégies étudiées conduisent donc en général à des règles de décision différentes. Le choix d'une stratégie particulière nous semble devoir être fait en tenant compte des caractéristiques particulières de l'application envisagée. La nécessité d'effectuer un tel choix peut fournir l'occasion d'intégrer dans la conception du système de décision des éléments difficilement formalisables autrement, tels que la plus ou moins grande prudence de l'utilisateur en l'absence d'information statistiquement fiable.

5 Conclusion

Une nouvelle approche de la reconnaissance des formes en mode supervisé a été proposée. Cette approche repose sur une modélisation du sentiment d'incertitude d'un agent rationnel confronté au problème du classement d'une entité décrite par un vecteur forme $\mathbf{x}$, compte tenu d'une connaissance complète ou partielle de la classe associée à N vecteurs d'apprentissage. Le formalisme adopté est celui du Modèle des Croyances Transférables qui constitue une interprétation non probabiliste de la théorie des fonctions de crédibilité initialement introduite par Shafer. Les notions d'espérances inférieure et supérieure ainsi que de probabilité pignistique conduisent à proposer différentes stratégies de décision avec coûts arbitraires, certaines pouvant donner lieu à des résultats très différents de ceux obtenus par les méthodes classiques. Nous pensons que cette méthodologie pourra se révéler particulièrement utile dans les situations où l'information disponible est très incomplète et entachée d'incertitude.

Références

[1] A. Appriou. Probabilités et incertitude en fusion de données multi-senseurs. *Revue Scientifique et Technique de la Défense*, (11):27–40, 1991.

[2] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, AMS-38:325–339, 1967.

[3] T. Denœux. An evidence-theoretic neural network classifier. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 712–717, Vancouver, October 1995.

[4] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[5] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition (In press)*, 1997.

[6] T. Denœux and G. Govaert. Combined supervised and unsupervised learning for system diagnosis using Dempster-Shafer theory. In P. Borne et al., editor, *CESA'96 IMACS Multiconference. Symposium on Control, Optimization and Supervision*, volume 1, pages 104–109, Lille, July 1996.

[7] B. Dubuisson. *Diagnostic et Reconnaissance des Formes.* Hermès, Paris, 1990.

[8] R. O. Duda and P. E Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New-York, 1973.

[9] R. Kennes. Computational aspects of the Möbius transform of a graph. Technical Report TR/IRIDIA/90-13, Université Libre, Bruxelles, 1990.

[10] H. Kim and P. H. Swain. Evidential reasoning approach to multisource-data classification in remote sensing. *IEEE Transactions on Systems, Man and Cybernetics*, 25(8):1257–1265, 1995.

[11] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley and Sons, New-York, 1992.

[12] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.

[13] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, N.J., 1976.

[14] P. Smets. The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

[15] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[16] N. Wilson. Decision making with belief functions and pignistic probabilities. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 364–371, Granada, November 1993. Springer Verlag.

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

[18] L. M. Zouhal and T. Denœux. Une méthode de discrimination non paramétrique basée sur la théorie de Dempster et Shafer. In *Actes du Quinzième Colloque GRETSI*, pages 689–692, Juan les Pins, Septembre 1995.

Annexe : Calcul des structures de croyance à l'aide de prototypes

L'approche générale pour le calcul des structures de croyance décrite au paragraphe 3.1 nécessite en pratique l'utilisation de la totalité de l'ensemble d'apprentissage pour le classement d'un nouvel individu, ce qui limite son domaine d'application à des ensembles d'apprentissage de taille modeste (quelques milliers d'individus). Une façon de contourner cette difficulté consiste à résumer l'ensemble d'apprentissage sous la forme de $n \ll N$ vecteurs de référence ou prototypes, choisis initialement de manière aléatoire ou déterminés par une procédure de classification automatique. Le calcul d'une structure de croyance pour un nouveau vecteur forme peut ensuite être effectué de la manière suivante [3].

Supposons chaque prototype i décrit par un vecteur forme $\mathbf{p}^{(i)}$, des degrés d'appartenance $u_1^{(i)}, \ldots, u_M^{(i)}$ à chaque classe avec $\sum_{j=1}^{M} u_j^{(i)} = 1$, et un paramètre positif $\gamma^{(i)}$ caractérisant le rayon de la "zone d'influence", ou "champ réceptif" du prototype i. Notons $e^{(i)}$ l'ensemble des informations relatives au prototype i. La classification s'effectue comme décrit au paragraphe 3.1 en trois étapes :

1. Calcul des distances $||\mathbf{x} - \mathbf{p}^{(i)}||$ entre $\mathbf{x}$ et chaque vecteur de référence $\mathbf{p}^{(i)}$.

2. Définition pour chaque prototype i d'une structure de croyance $m(\cdot|e^{(i)})$:

$$\forall j \in \{1, \ldots, M\} \quad m(\{\omega_j\}|e^{(i)}) = \alpha^{(i)} u_j^{(i)} \varphi^{(i)}(||\mathbf{x} - \mathbf{p}^{(i)}||) \tag{34}$$

$$m^{(i)}(\Omega|e^{(i)}) = 1 - \alpha^{(i)} \varphi^{(i)}(||\mathbf{x} - \mathbf{p}^{(i)}||) \tag{35}$$

   avec

$$\varphi^{(i)}(||\mathbf{x} - \mathbf{p}^{(i)}||) = \exp(-\gamma^{(i)} ||\mathbf{x} - \mathbf{p}^{(i)}||^2).$$

3. Combinaison des n structures de croyance $m(\cdot|e^{(i)})$, $i = 1, \ldots, n$ par la règle de Dempster.

L'étape 1 ci-dessus est analogue au calcul effectué dans la première couche d'un réseau de neurones à fonctions de base radiales. Cette analogie suggère une implémentation connexionniste de l'algorithme au sein d'une architecture possédant une première couche cachée L_1 de n neurones, une seconde couche cachée L_2 de n modules de $M+1$ neurones, et une couche de sortie L_3 composée d'unités sigma-pi réalisant la combinaison (Figure 3). Les différents paramètres du modèle (vecteurs de poids, degrés d'appartenance et rayon de la zone d'influence de chaque prototype) peuvent être optimisés par descente de gradient d'un critère d'erreur. Différentes simulations ont montré les bonnes performances de cette méthode par rapport aux réseaux à fonctions de base radiales et LVQ (Learning Vector Quantization) [3].

Biographie

T. Denœux est ingénieur civil et docteur de l'Ecole Nationale des Ponts et Chaussées. Il a obtenu en 1996 l'Habilitation à diriger des recherches à l'Institut National Polytechnique de Lorraine. Il est depuis 1992 enseignant-chercheur à l'Université de Technologie de Compiègne, et membre du laboratoire Heudiasyc (UMR CNRS 6599). Ses recherches portent sur la reconnaissance des formes, les réseaux de neurones, la représentation des connaissances incertaines et la fusion de données.

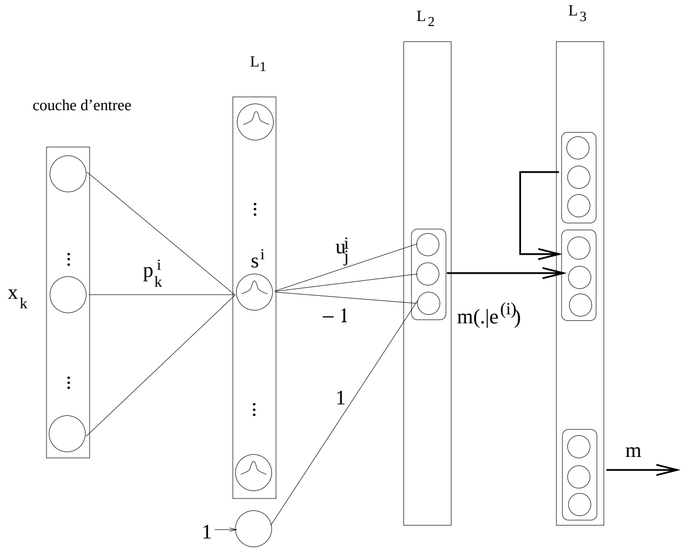

FIG. 3 - *Modélisation connexionniste de l'algorithme de discrimination par calcul de distances à des prototypes et combinaison des distributions de masses de croyance (d'après [3]). Les structures de croyances $m(\cdot|e^{(i)})$ sont combinées selon la règle de Dempster non normalisée dans la couche L_3. La sortie du réseau est une structure de croyance m non normalisée.*