Modélisation de l'imprécis et de l'incertain en apprentissage supervisé par la théorie des fonctions de croyance

Modeling imprecision and uncertainty in supervised learning using the theory of belief functions

T. Denœux
Heudiasyc – UMR CNRS 6599
Université de Technologie de Compiègne
BP 529 - F-60205 Compiègne cedex - France
tdenoeux@hds.utc.fr

Résumé

Nous présentons dans cet article une théorie générale de l'apprentissage supervisé basée sur le modèle des croyances transférables, une interprétation subjectiviste et non probabiliste de la théorie de Dempster et Shafer. Grâce à la grande richesse d'expression des fonctions de croyance, cette approche permet en particulier la prise en compte de données imprécises et incertaines, ainsi que la modélisation des incertitudes de différentes natures (non-spécificité, conflit, flou) rencontrées dans les problèmes de discrimination et de régression.

Mots Clef

Apprentissage supervisé, discrimination, régression, fonctions de croyance, théorie de Dempster-Shafer, ensembles flous.

Abstract

This paper presents a general theory of supervised learning based on the Transferable Belief Model, a subjectivist, non probabilistic interpretation of the Dempster-Shafer theory of evidence. Thanks to the expressive power of belief functions, this approach allows in particular the processing of imprecise and uncertain data, as well as the modelling of verious kinds of uncertainty (non-specificity, conflict, fuzziness) arising in classification and regression problems.

Keywords

Supervised learning, classification, regression, belief functions, Dempster-Shafer theory, fuzzy sets.

1 Introduction

1.1 L'apprentissage supervisé

D'une manière très générale, l'apprentissage statistique supervisé peut être défini comme la recherche de dépendances entre variables à partir d'un nombre limité d'observations [30].

Plus précisément, on considère une population $\mathcal{P}$ d'entités décrites par deux variables $\mathbf{x} \in \Omega_x$ et $y \in \Omega_y$. Le plus souvent, $\mathbf{x}$ est une variable vectorielle composée de p descripteurs ou attributs $x_1, \ldots, x_p$ qualitatifs ou quantitatifs, tandis que la variable y est soit une variable nominale à c modalités (problème de discrimination), soit une variable quantitative (problème de régression). Typiquement, le problème posé est le suivant :

Etant donné un ensemble d'apprentissage $\mathcal{E}$ composé de n observations précises

$$(\mathbf{x}^i, y^i), i = 1, n$$

des variables $\mathbf{x}$ et y relatives à un échantillon de n individus issus de $\mathcal{P}$, construire une fonction de décision

$$f : \Omega_x \mapsto \Omega_y \tag{1}$$

permettant d'effectuer une « prédiction » de la variable y pour chaque nouvel individu issu de $\mathcal{P}$ sur lequel on a observé le vecteur $\mathbf{x}$.

L'approche classique de ce problème consiste à considérer $\mathbf{x}$ et y comme un couple de variables aléatoires muni d'une loi de probabilité $F(\mathbf{x}, y)$, et les n observations $(\mathbf{x}^i, y^i), i = 1, n$ comme une réalisation d'un échantillon i.i.d. issu de F. La qualité d'une prédiction $f(x)$ étant mesurée par une fonction de perte

$L(y, f(\mathbf{x}))$ à valeurs dans $\mathbb{R}_+$, on cherche la fonction f qui minimise le risque

$$R(f) = \int L(y, f(\mathbf{x}))dF(\mathbf{x}, y).$$

De nombreuses techniques d'apprentissage statistique supervisé (dont les réseaux de neurones et les arbres de décision) visent à construire des estimateurs de la fonction optimale f^* ayant asymptotiquement de « bonnes propriétés » [30, 13].

1.2 Généralisation

Malgré son apparente généralité, le formalisme précédent s'applique difficilement à certaines situations réelles, dans lesquelles l'imprécision des observations ou l'incertitude des prédictions ne peuvent être négligées.
Afin de montrer de prendre en compte dans certaines application l'imprécision des observations, considérons les exemples suivants :

1. Dans un problème de diagnostic médical, $\mathbf{x}$ désigne typiquement les symptômes et y le diagnostic. L'évolution des connaissances conduit fréquemment à affiner l'ensemble Ω_y des diagnostics. Suposons que, jusqu'à une date t_0, Ω_y comporte deux éléments ω_1 et ω_2. A t_0, on opère une distinction entre 2 variantes de la pathologie correspondant à ω_2 et on pose $\omega_2 = \{\omega_{21}, \omega_{22}\}$. Parmi les cas collectés avant t_0, certains seront donc de la forme $(\mathbf{x}^i, Y^i)$, Y^i désigant un ensemble de valeurs possibles pour la variable y^i. Il s'agit là d'un exemple d'*étiquetage imprécis* de l'ensemble d'apprentissage.

2. Dans certaines applications, la variable y ne peut être déterminée avec certitude. Par exemple, Richard [20] décrit un problème de détection de complexes K dans des électroencéphalogramme, dans lequel plusieurs experts sont sollicités pour inspecter visuellement des signaux de sommeil; les désaccords entre experts se traduisent dans ce cas par un *étiquetage incertain* de l'ensemble d'apprentissage.

3. Dans d'autres situations, la variable y est déterminée par un ou plusieurs capteurs plus ou moins fiables, les mesures étant entâchées d'incertitudes. Supposons que k capteurs fournissent k mesures de $y^i \in \mathbb{R}$ sous formes d'intervalles $Y^i_j = [\underline{y}^i_j, \overline{y}^i_j]$ (respectivement, de nombres flous $\widetilde{Y}^i_j$), $j = 1, k$. D'autre part, notons $m^i_j \in [0, 1]$ la probabilité que le capteur j fonctionne correctement. L'information disponible relativement à la variable y^i se traduit dans ce cas par k ensembles Y^i_j (respectivement, k ensembles flous $\widetilde{Y}^i_j$) de valeurs possibles, chaque ensemble étant affecté d'un « degré de confiance » m^i_j. L'étiquetage est dans ce cas *à la fois imprécis et incertain.*

Par ailleurs, il importe dans de nombreuses applications de prendre en compte l'incertitude non seulement sur les entrées du système d'apprentissage, mais également sur les sorties. Dans certains problèmes d'aide à la décision tels que le diagnostic médical ou la gestion de portefeuilles boursiers, les prédictions fournies par une méthode statistique ne sont souvent que des éléments parmi d'autres permettant à une expert de prendre une décision. Il est souhaitable dans ce cas que l'incertitude sur chacune des prédictions soit décrite de manière suffisamment fidèle. De même, dans les problèmes de fusion de décisions, la stratégie de combinaison des informations doit pouvoir prendre en compte la plus ou grande fiabilité des experts ou capteurs fusionnés.
Les considérations précédentes nous conduisent à proposer une formulation plus générale du problème de l'apprentissage statistique supervisé, dans laquelle les entrées et les sorties du système sont représentées à l'aide du formalisme des fonctions de croyance [21, 28]. Nous adoptons pour cela le modèle des croyances transférables (MCT) de Smets, dont rappelons brièvement ci-dessous les principaux éléments. Différentes extensions permettant la représentation d'informations floues sont également présentées.

2 Le modèle des croyances transférables

2.1 Concepts fondamentaux

Nous nous contenterons ici de rappeler les principaux éléments du MCT introduit par Smets [28, 26], afin notamment de préciser les notations et la terminologie.
Le problème posé consiste à modéliser la croyance d'un agent rationel relativement à la valeur prise par une variable y, sur la base d'un corpus de connaissance. On suppose que cette croyance peut être représentée par l'allocation d'une « masse de croyance » unité à certains sous-ensembles du domaine Ω_y de y, appelés *éléments focaux*[1]. On appelle *structure de croyance* (SC) l'application qui, à chaque sous-ensemble A de Ω_y, associe la fraction de masse $m(A)$ affectée à A. On a

1. Dans ce paragraphe, l'ensemble Ω_y est supposé fini. L'extension à un ensemble infini, même non dénombrable, est cependant immédiate pourvu que l'on impose à chaque SC d'avoir un nombre fini d'éléments focaux.

donc nécessairement :

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

Remarquons que l'on n'impose pas systématiquement la condition $m(\emptyset) = 0$, la masse allouée à $\emptyset$ pouvant recevoir une interprétation dans le cadre de l'hypothèse du monde ouvert [23].

A la SC m est associée une *fonction de croyance* bel_m, qui associe à toute partie A de Ω_y la somme des masses allouées aux sous-ensembles de A:

$$\mathrm{bel}_m(A) = \sum_{\emptyset\neq B\subseteq A} m(B). \quad (3)$$

Chaque nombre $\mathrm{bel}_m(A) \in [0, 1]$ s'interprète comme un *degré de croyance* en l'hypothèse « $y \in A$ ». On montre que la fonction bel_m est une capacité de Choquet d'ordre infini, et qu'il existe une bijection entre l'ensemble des SC et celui des fonctions de croyance [21]. Smets [25] a montré que l'usage de telles fonctions pour la représentation des croyances pouvait être déduit d'une série axiomes.

Supposons maintenant que deux sources d'informations distinctes induisent séparément deux SC m_1 et m_2. La considération simultanée de ces deux sources induit une nouvelle SC $m_1 \cap m_2$, appelée somme conjonctive de m_1 et m_2, et définie par :

$$(m_1 \cap m_2)(A) = \sum_{B\cap C=A} m_1(B)m_2(C) \quad \forall A \subseteq \Omega. \quad (4)$$

On montre facilement l'associativité et la commutativité de l'opérateur $\cap$. Notons que l'on peut définir, à partir d'autres opérations ensemblistes, d'autres opérateurs sur les SC, comme par exemple la somme disjonctive [24], qui correspond à un mode de combinaison beaucoup plus « prudent » (on suppose que l'une *ou* l'autre des deux sources doit être prise en considération).

Enfin, le MCT opère (contrairement au modèle bayésien) une distinction très nette entre la représentation des croyances et la prise de décision, cette dernière devant être basée sur le formalisme des probabilités afin de guarantir la cohérence des décisions [28]. On appelle *distribution de probabilité pignistique* associée à une SC m la fonction $\mathrm{BetP}_m : \Omega_y \mapsto [0, 1]$ définie par :

$$\mathrm{BetP}(\omega) = \sum_{B\ni\omega} \frac{m^*(B)}{|B|}, \quad (5)$$

m^* étant la SC normalisée associée à m. Face à un ensemble d'action possibles, le MCT préconise l'action de risque minimal au sens de la probabilité pignistique.

2.2 Extensions

Le formalisme précédent peut être étendu afin de représenter (1) des croyances en des propositions floues (par exemple, « y est grand »), et (2) des degrés de croyance flous (« $m(A)$ = environ 0.3 »). Nous présentons brièvement ces deux extensions.

Croyances en des propositions floues. L'idée de généraliser aux ensembles flous les notions relatives aux fonctions de croyance est due à Zadeh [36], et a été explorée par de nombreux auteurs [31, 32, 22, 34]. On définit une structure de floue (SCF) comme une application de l'ensemble $[0, 1]^{\Omega_y}$ des sous-ensembles flous de Ω_y dans $[0, 1]$, telle que

- l'ensemble $\mathcal{F}(m) = \{F \in [0, 1]^{\Omega_y} | m(F) > 0\}$ est fini ;
- $\displaystyle\sum_{F\in\mathcal{F}(m)} m(F) = 1.$

La SCF m est dite normalisée si ses éléments focaux sont des sous-ensembles flous normalisés (de hauteur égale à 1).

La notion de degré de croyance $\mathrm{bel}_m(A)$ en une proposition floue « y est A », pour $A \in [0, 1]^{\Omega_y}$ peut être définie par

$$\mathrm{bel}_m(A) = \sum_{j=1}^{k} m(F_j)N(A|F_j). \quad (6)$$

où les F_j, $j = 1, k$ sont les éléments focaux de m, et $N(A|F_j)$ la nécessité conditionnelle de A sachant F_j (cf. [35], et [15] pour le cas doù les ensembles A et F_j ne sont pas normalisés). Remarquons que $N(A|F_j)$ peut être interprété comme un degré d'intersection de F_j dans A, ce qui met peut-être mieux en évidence le lien entre les Eqs (3) et (6).

La somme conjonctive peut facilement être généralisée en remplaçant dans l'Eq. (4) l'intersection classique par un opérateur d'intersection entre ensembles flous.

Enfin, la probabilité pignistique associée à une SCF m peut également être définie [8, 18, 19] par une formule analogue à l'Eq. (5), en utilisant la définition de la cardinalité pour les ensembles flous : $|B| = \sum_{\omega\in B} B(\omega)$, et en étendant aux SCF l'opération de normalisation [33].

Degrés de croyance flous. Une étape supplémentaire dans la « fuzzification » du MCT peut être franchie en autorisant l'affectation aux différents éléments focaux (éventuellement flous) de *masses de croyance imprécises*, définies comme des intervalles ou des nombres flous [10, 6, 8]. Cette généralisation est plus délicate que la précédente, car les différentes masses de croyance

$m(F_j), j = 1, k$ sont des variables interactives liées par la relation $\sum_{j=1}^{k} m(F_j) = 1$; on ne peut donc appliquer aux Eqs (4), (5) et (6) les principes simples de l'arithmétique d'intervalles ou de l'arithmétique floue. Dans [10], nous avons abordé ce problème en définissant une structure de croyance à valeurs par intervalles comme un ensemble de SC dont les masses sont comprises entre certaines bornes ; nous avons montré que les notions de croyance et de plausibilité, de probabilité pignistique, d'opérateurs de combinaisons pouvaient être étendues rigoureusement dans ce contexte comme la solution de programmes linéaires ou quadratiques. Une généralisation supplémentaire a été effectuée dans [6, 8], où nous définissons une structure de croyance à valeurs floues comme un ensemble flou de SC, dont les masses sont contraintes par des distributions de possibilité, le principe d'extension de Zadeh permettant de retrouver l'ensemble des notions du MCT (degrés de croyance, probabilité pignistique, etc.).

3 Application en apprentissage

3.1 Principe

Le MCT et ses extensions floues permettent de reformuler le problème de l'apprentissage supervisé, énoncé au paragraphe 1.1, dans le cadre plus général suivant : On considère comme précédemment n individus extraits d'une population $\mathcal{P}$, et décrits par deux variables $\mathbf{x}$ et y, à valeurs respectivement dans Ω_x et Ω_y. On note $\mathbf{x}^i$ et y^i les valeurs prises par les variables $\mathbf{x}$ et y pour l'individu i. Pour chaque i, la variable $\mathbf{x}^i$ est supposée connue, mais la variable y^i est partiellement inconnue, et décrite par une SC m^i sur Ω_y. On a donc un ensemble d'apprentissage de la forme :

$$\mathcal{E} = \{e^i = (\mathbf{x}^i, m^i) | i = 1, \ldots, n\}. \qquad (7)$$

Il convient ici de remarquer que cette formulation très générale englobe un grand nombre de situations distinctes, dont certaines ont déjà été étudiées dans un cadre plus restrictif :

- dans le cas où chaque SC m^i a un seul élément focal singleton, il s'agit d'un problème classique d'apprentissage supervisé (discrimination ou régression) avec étiquetage précis et certain ;
- lorsque chaque chaque SC m^i a un seul élément focal $F^i \subseteq \Omega_y$, l'étiquetage est imprécis : on ignore la valeur exacte de y^i, mais on sait avec certitude que $y^i \in F^i$. Dans le cas où $\Omega_y = \mathbb{R}$ et où les F^i sont des intervalles, on retrouve le problème de la régression sur variable de type intervalle, étudiée en particulier par Tanaka et Diamond [14, 29].
- lorsque chaque SC m^i a un unique élément focal flou $F^i \in [0,1]^{\Omega_y}$, l'étiquetage peut-être qualifié de flou ; on retrouve alors comme cas particuliers la régression floue [29, 14] et la discrimination floue, abordée dans un cadre évidentiel dans [37, 40, 39, 12].
- le cas le plus général où m^i est une SC quelconque possédant un nombre arbitraire d'éléments focaux (l'étiquetage étant alors à la fois imprécis et incertain) n'a pas encore, à notre connaissance, été étudié, bien qu'il soit adapté à la modélisation de nombreux problèmes réels, comme nous l'avons montré au paragraphe 1.2.

Etant donné un ensemble d'apprentissage de la forme décrite par l'Eq. (7), on considère maintenant un nouvel individu sur lequel on observe la variable $\mathbf{x}$, et pour lequel on cherche à prédire la variable y. Une telle prédiction étant nécessairement incertaine, on cherche à décrire l'information relative à y sous la forme d'une SC m_y sur Ω_y (la notation m_y étant introduite pour mettre nettement en évidence la variable à laquelle se rapporte la SC). On généralise ainsi la fonction de décision définie par l'Eq. (1) sous la forme d'une fonction f de Ω_x dans l'ensemble $\mathcal{S}(\Omega_y)$ des SC sur Ω_y.

La méthode que nous avons initialement proposée dans [1, 3] pour construire une telle fonction repose sur deux principes fondamentaux : l'affaiblissement des SC m^i en fonction d'une mesure de dissimilarité δ entre éléments de Ω_x, et la combinaison des SC résultantes par l'opérateur de somme conjonctive.

Considérons tout d'abord l'impact d'un exemple d'apprentissage $e^i = (\mathbf{x}^i, m^i)$ sur notre croyance relativement à la valeur de y. Afin d'introduire la méthode dans un cas simple, supposons tout d'abord que m^i a un seul élément focal F^i, et que $\mathbf{x} = \mathbf{x}^i$. Le fait $y^i \in F^i$ accrédite l'hypothèse selon laquelle $y \in F^i$, et aucune autre hypothèse. S'il n'existe pas *a priori* de relation fonctionnelle entre $\mathbf{x}$ et y, on n'aura cependant aucune certitude relativement à la valeur de y, ce qui conduit à poser :

$$\begin{aligned} m_y(F^i|e^i) &= \alpha && (8)\\ m_y(\Omega_y|e^i) &= 1-\alpha, && (9) \end{aligned}$$

avec $0 < \alpha \leq 1$. Si maintenant $\mathbf{x} \neq \mathbf{x}^i$, il est naturel de faire dépendre la masse de croyance $m_y(F^i|e^i)$ de la *dissimilarité* entre $\mathbf{x}$ et $\mathbf{x}^i$ en posant :

$$\begin{aligned} m_y(F^i|e^i) &= \phi[\delta(\mathbf{x}, \mathbf{x}^i)] && (10)\\ m_y(\Omega_y|e^i) &= 1-\phi[\delta(\mathbf{x}, \mathbf{x}^i)], && (11) \end{aligned}$$

où ϕ est une fonction décroissante vérifiant $\phi(0) \leq 1$ et

$$\lim_{d \to \infty} \phi(d) = 0.$$

La généralisation au cas où la SC m^i est de forme quelconque est immédiate si l'on remarque que m_y définie par les Eq (10) et (11) s'obtient à partir de m^i par l'opération d'*affaiblissement* [21], le facteur d'affaiblissement étant égal à $1-\phi[\delta(\mathbf{x},\mathbf{x}^i)]$. En étendant ce principe, on obtient donc dans le cas général :

$$\begin{aligned} m_y(F|e^i) &= \phi[\delta(\mathbf{x},\mathbf{x}^i)]m^i(F) \quad \text{si } F \neq \Omega_y \quad (12)\\ m_y(\Omega_y|e^i) &= 1-\phi[\delta(\mathbf{x},\mathbf{x}^i)]\sum_{\{G\neq\Omega_y\}} m^i(G). \quad (13)\end{aligned}$$

Ayant quantifié sous la forme d'une SC $m_y(\cdot|e^i)$ l'impact de chaque exemple d'apprentissage e^i sur notre croyance en la valeur de y, il nous reste maintenant à combiner les n SC induites par les n exemples e^i, $i=1,n$. Plusieurs opérateurs de combinaison peuvent *a priori* être envisagés, mais un argument décisif milite selon nous en faveur de la somme conjonctive. Supposons que, pour un exemple e^i, on n'ait aucune information sur y^i : m^i est donc la SC vide et l'on a $m^i(\Omega_y)=1$, d'où par conséquent également $m_y(\Omega_y|e^i)=1$. Or, un tel exemple n'apportant aucune information sur y, la SC vide $m_y(\cdot|e^i)$ doit être un élément neutre de l'opérateur de combinaison, condition vérifiée pour la somme conjonctive et non pour la somme disjonctive ou la moyenne. Nous définirons donc la SC m induite par la totalité de l'ensemble d'apprentissage $\mathcal{E}$ comme la somme conjonctive des SC induites par chacun des éléments de $\mathcal{E}$:

$$m_y(\cdot|\mathcal{E}) = \bigcap_{i=1}^{n} m_y(\cdot|e^i). \quad (14)$$

La fonction $m_y(\cdot|\mathcal{E})$ modélise donc l'information apportée par l'ensemble d'apprentissage, relativement à la variable à expliquer y. Plusieurs stratégies de décision en discrimination basée sur cette information (affectation à une ou plusieurs classes, rejet, etc.) ont été étudiées dans [4, 7]. En régression, une prédiction ponctuelle de y peut être obtenue en calculant l'espérance de y relativement à la mesure de probabilité pignistique [18, 19, 17].

3.2 Optimisation des performances

Nous avons jusqu'à présent passé sous silence le problème du choix de la mesure de similarité δ et de la fonction ϕ dans les Eq. (10-13). Or, il s'agit bien entendu d'un problème fondamental puisque ces choix conditionneront les performances de la méthode. Une approche pour résoudre ce problème consiste à choisir δ et ϕ parmi des familles paramétrées de fonction. Notons

$$\Phi_\theta(\mathbf{x},\mathbf{x}^i) = \phi_{\theta_2}[\delta_{\theta_1}(\mathbf{x},\mathbf{x}^i)],$$

où θ_1 est un paramètre définissant la dissimilarité, θ_2 un paramètre définissant la fonction ϕ, et $\theta=(\theta_1,\theta_2)^t$. Un choix naturel lorsque $\Omega_x=\mathbb{R}^p$ consiste à chosir δ dans la famille des distances euclidiennes généralisées :

$$\delta_{\theta_1}(\mathbf{x},\mathbf{x}^i) = \left[(\mathbf{x}-\mathbf{x}^i)^t\Sigma(\mathbf{x}-\mathbf{x}^i)\right]^{1/2},$$

Σ étant une matrice symétrique définie positive. On peut dans ce cas justifier [7] le choix suivant pour ϕ :

$$\phi_{\theta_2}(\delta) = \alpha\exp(-\gamma\delta^2)$$

avec $\alpha\in]0,1[$, $\gamma\in]0,+\infty[$, et $\theta_2=(\alpha,\gamma)^t$.
Une fois le choix de la famille $\{\Phi_\theta|\theta\in\Theta\}$ effectué, il est possible d'optimiser les performances du système d'apprentissage par une technique de validation croisée [38, 41]. Soit $\mathcal{E}^{-i}$ l'ensemble d'apprentissage privé de l'exemple i, et $m^{-i}=m_{y^i}(\cdot|\mathcal{E}^{-i})$ la SC représentant l'information apportée par $\mathcal{E}^{-i}$ sur y^i, obtenue en appliquant les Eq. (12), (13) et (14) à $\mathbf{x}^i$. Etant donné un critère d'erreur $\Delta(m^{-i},m^i)$, il s'agit de trouver la valeur du paramètre θ minimisant l'erreur moyenne sur l'ensemble d'apprentissage :

$$\widehat{\theta} = \arg\min_\theta E(\theta)$$

$$E(\theta) = \sum_{i=1}^{n}\Delta(m^{-i},m^i).$$

Plusieurs choix de critères d'erreur ont été proposés pour la discrimination [41], et la régression [18, 19, 17].

3.3 Lien avec les méthodes de noyaux

Dans le cas où l'ensemble d'apprentissage a la forme classique $\{(\mathbf{x}^i,y^i)|i=1,n\}$ (étiquetage précis et certain), la méthode d'apprentissage décrite dans cet article présente d'intéressantes similitudes avec certaines techniques statistiques non paramétriques appelées méthodes de noyaux [13, 16]. Considérons à titre d'exemple le cas de la discrimination (cf. [18, 19, 17] pour la régression).
Soit donc un problème à c classes $\Omega_y=\{\omega_1,\ldots,\omega_c\}$ et $\mathcal{E}_j$ le sous-ensemble de $\mathcal{E}$ composé des exemples étiquetés dans la classe j.
Soit $e^i=(\mathbf{x}^i,y^i)\in\mathcal{E}_j$. D'après les Eq. (10) et (11), on peut écrire :

$$\begin{aligned} m_y(\{\omega_j\}|e^i) &= 1-\exp(-w^i)\\ m_y(\Omega_y|e^i) &= \exp(-w^i),\end{aligned}$$

avec

$$w^i = -\ln(1-\phi[\delta(\mathbf{x},\mathbf{x}^i)]).$$

La quantité w^i est le *weight of evidence* introduit par Shafer [21]. Lorsque l'on combine plusieurs SC simples de mêmes éléments focaux, les w^i s'ajoutent, et l'on a donc, en notant

$$m_y(\cdot|\mathcal{E}_j) = \bigcap_{\{i|e^i \in \mathcal{E}_j\}} m_y(\cdot|e^i),$$

$$\begin{aligned} m_y(\{\omega_j\}|\mathcal{E}_j) &= 1 - \exp\left(-\sum_{\{i|e^i \in \mathcal{E}_j\}} w^i\right) \\ m_y(\Omega_y|\mathcal{E}_j) &= 1 - m_y(\{\omega_j\}|\mathcal{E}_j). \end{aligned}$$

En remarquant que

$$m_y(\cdot|\mathcal{E}) = \bigcap_{j=1}^{c} m_y(\cdot|\mathcal{E}_j),$$

il s'ensuit que

$$m(\{\omega_j\}|\mathcal{E}) \geq m(\{\omega_\ell\}|\mathcal{E}) \quad \forall \ell \neq j$$

si et seulement si :

$$\sum_{\{i|e^i \in \mathcal{E}_j\}} w^i \geq \sum_{\{i|e^i \in \mathcal{E}_\ell\}} w^i \quad \forall \ell \neq j.$$

Il s'agit donc bien d'une méthode de noyaux (*kernel rule* [13, p. 147]), avec le noyau :

$$K(\mathbf{x}, \mathbf{x}^i) = -\ln(1 - \phi[\delta(\mathbf{x}, \mathbf{x}^i)]).$$

Sous réserve de choisir ϕ de telle sorte que ce noyau soit régulier, ce fait permet de transposer à notre approche les résultats théoriques relatifs à la consistence des méthodes de noyaux [13]. Insistons néanmoins sur le fait que les méthodes statistiques non paramétriques habituelles ne permettent pas de prendre en compte des données imprécises au sens où nous les avons définies dans cet article.

3.4 Réduction du temps de calcul

Dans le cas où les SC m^i ont une forme générale et où n est grand, la combinaison des n SC $m_y(\cdot|e^i)$ peut devenir pénalisante par suite de la complexité exponentielle de la somme conjonctive. Plusieurs approches peuvent être envisagées pour remédier à ce problème :

- une première approche, initialement proposée dans [3], consiste à n'utiliser que les k plus proches voisin de $\mathbf{x}$ dans $\mathcal{E}$ (ou, alternativement, les voisins situés dans une boule de rayon r centrée sur $\mathbf{x}$) ; néanmoins, cette approche nécessite encore le calcul de toutes les distances entre $\mathbf{x}$ et les $\mathbf{x}^i$;
- une autre méthode consiste à résumer l'ensemble d'apprentissage sous la forme d'un nombre restreint de prototypes [2, 5, 9, 19], éventuellement optimisés par minimisation d'un critère d'erreur ;
- enfin, dans le cas où Ω_y est très grand ou continu, il est possible d'envisager l'utilisation de techniques d'approximation de fonctions de croyance ; une telle approche, basée sur une classification hiérarchique des éléments focaux, est décrite dans [18].

3.5 Généralisation à des entrées imprécises

Nous avons jusqu'à présent supposé que seule la variable à expliquer pouvait être entachée d'imprécision ou d'incertitude. Or il arrive également, par suite d'incertitudes de mesures par exemples, que les attributs x_j ne soient que partiellement connus. Le cas où les composantes du vecteur d'entrée $\mathbf{x}$ sont restreintes par des intervalles a été étudié dans [10], puis étendu aux nombres flous dans [6, 8]. L'approche retenue pour résoudre ce problème s'appuie sur les notions de SC imprécise (par intervalle ou floue), et consiste schématiquement à appliquer aux Eq. (12-14) le principe d'extension.

4 Conclusion

Nous avons résumé dans cet article les grandes lignes d'une approche très générale de l'apprentissage supervisé basée sur le MCT. Nous nous sommes notamment attachés à dégager les principes fondamentaux de cette approche, et à en présenter de manière unifiée l'application en discrimination et en régression. De nombreux résultats expérimentaux ont montré que cette méthode permet non seulement de traiter efficacement des problèmes nouveaux tels que l'apprentissage à partir de données incertaines et imprécise, mais offre par ailleurs des performances comparables aux meilleures méthodes existantes sur des problèmes « classiques » (cf. par exemple [3, 41, 9, 27]). Cette approche se révèle également particulièrement efficace dans les problèmes de fusion de décisions [9]. L'extension à l'apprentissage non supervisé et les applications en diagnostic de systèmes complexes sont abordées dans [11].

Références

[1] T. Denœux. Application of evidence theory to k-NN pattern classification. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice IV*, pages 13–24. Elsevier, Amsterdam, 1994.

[2] 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.

[3] 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.

[4] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[5] T. Denœux. Function approximation in the framework of evidence theory: A connectionist approach. In *Proceedings of the 1997 International Conference on Neural Networks (ICNN'97)*, volume 1, pages 199–203, Houston, June 1997. IEEE.

[6] T. Denœux. Allowing imprecision in belief representation using fuzzy-valued belief structures. In *Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'98)*, volume 1, pages 48–55, Paris, July 1998. Editions EDK.

[7] T. Denœux. Application du modèle des croyances transférables en reconnaissance de formes. *Traitement du Signal*, 14(5):443–451, 1998.

[8] T. Denœux. Modeling vague beliefs using fuzzy-valued belief structures. *Fuzzy Sets and Systems (In Press)*, 1999.

[9] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE transactions on Systems, Man and Cybernetics - Part A (accepted for publication)*, 1999.

[10] T. Denœux. Reasoning with imprecise belief structures. *International Journal of Approximate Reasoning*, 20:79–111, 1999.

[11] 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.

[12] T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. Technical Report 98/160, Université de Technologie de Compiègne, Heudiasyc Laboratory, November 1998.

[13] L. Devroye, L. Gyorfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New-York, 1996.

[14] P. Diamond and H. Tanaka. Fuzzy regression analysis. In R. Slowinski, editor, *Fuzzy sets in decision analysis, operations research and statistics*, pages 349–387. Kluwer Academic Publishers, Boston, 1998.

[15] D. Dubois and H. Prade. An alternative approach to the handling of subnormal possibility distributions. *Fuzzy Sets and Systems*, 24:123–126, 1987.

[16] W. Härdle. *Applied nonparametric regression.* Cambridge University Press, Cambridge, 1990.

[17] S. Petit-Renaud and T. Denœux. Application de la théorie des fonctions de croyance en régression. In *LFA'99*, Valenciennes, October 1999.

[18] S. Petit-Renaud and T. Denœux. Handling different forms of uncertainty in regression analysis: a fuzzy belief structure approach. In A. Hunter and S. Pearsons, editors, *Symbolic and quantitative approaches to reasoning and uncertainty (ECSQARU'99)*, pages 340–351, London, June 1999. Springer Verlag.

[19] S. Petit-Renaud and T. Denœux. Regression analysis using fuzzy evidence theory. In *Proceedings of FUZZ-IEEE'99*, Seoul, August 1999.

[20] C. Richard. *Une méthodologie pour la détection à structure imposée. Applications au plan temps-fréquence.* PhD thesis, Université de Technologie de Compiègne, 1998.

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

[22] P. Smets. The degree of belief in a fuzzy event. *Information Sciences*, 25:1–19, 1981.

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

[24] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[25] P. Smets. The normative representation of quantified beliefs by belief functions. *Artificial Intelligence*, 92(1–2):229–242, 1997.

[26] P. Smets. The Transferable Belief Model for quantified belief representation. In D. M. Gabbay and P. Smets, editors, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 1, pages 267–301. Kluwer Academic Publishers, Dordrecht, 1998.

[27] P. Smets. Practical uses of belief functions. In *UAI'99*, Stockholm, Sweden, August 1999.

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

[29] H. Tanaka and P. Guo. *Possibilistic data analysis for operations research.* Physica-Verlag, Heidelberg, 1999.

[30] V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, New-York, 1995.

[31] R. R. Yager. Generalized probabilities of fuzzy events from fuzzy belief structures. *Information Sciences*, 28:45–62, 1982.

[32] R. R. Yager. Arithmetic and other operations on Dempster-Shafer structures. *Int. J. Man-Machines Studies*, 25:357–366, 1986.

[33] R. R. Yager. On the normalization of fuzzy belief structure. *International Journal of Approximate Reasoning*, 14:127–153, 1996.

[34] J. Yen. Generalizing the Dempster-Shafer theory to fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics*, 20(3):559–569, 1990.

[35] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[36] L. A. Zadeh. Fuzzy sets and information granularity. In R. K. Ragade M. M. Gupta and R. R. Yager, editors, *Advances in Fuzzy Sets Theory and Applications*, pages 3–18. North-Holland, Amsterdam, 1979.

[37] L. M. Zouhal and T. Denœux. A comparison between fuzzy and evidence-theoretic k-NN rules for pattern recognition. In *Proceedings of EUFIT'95*, volume 3, pages 1319–1325, Aachen, August 1995.

[38] 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.

[39] L. M. Zouhal and T. Denœux. Reconnaissance de formes floues par la théorie de Dempster et Shafer. In *Rencontres Francophones sur la Logique Floue et ses Applications*, pages 3–8, Nancy, December 1996. Cépaduès.

[40] L. M. Zouhal and T. Denœux. Generalizing the evidence-theoretic k-NN rule to fuzzy pattern recognition. In *Proceedings of the Second International Symposium on Fuzzy Logic and Applications ISFL'97*, pages 294–300, Zurich, February 1997. ICSC Academic Press.

[41] L. M. Zouhal and T. Denœux. An evidence-theoretic k-NN rule with parameter optimization. *IEEE Transactions on Systems, Man and Cybernetics – Part C*, 28(2):263–271, 1998.