Application de la théorie des croyances floues en régression

Application of Evidence Theory in Regression Analysis

S. Petit-Renaud T. Denœux

Université de Technologie de Compiègne,

HeuDiaSyc-UMR C.N.R.S. 6599, BP 20529, 60205 Compiègne, France

Résumé

Nous proposons une nouvelle méthode de régression basée sur la théorie de l'évidence floue. La sortie est estimée sous la forme d'une structure de croyance floue à l'aide d'un ensemble d'apprentissage, ce qui permet de mettre en valeur différents types d'incertitude, tels que l'imprécision, le conflit ou l'ignorance. Cette méthode est capable de prendre en compte des données très diverses (réels, intervalles, nombres ou quantités floues, « structures de croyance floues »). Afin de limiter le temps de calcul et d'améliorer l'interprétabilité de la sortie, nous proposons une procédure de simplification de structures de croyance, basée sur la similarité entre ensembles flous. Notre modèle fournit des estimations sous des formes très variées : ponctuelle, probabiliste, floue ou sous forme de structure de croyance. Des simulations permettent de le comparer à quelques méthodes de régression classiques ou floues.

Mots Clef

Dempster-Shafer, Données floues, Données imprécises, Incertitude, Régression, Classification hiérarchique.

Abstract

We propose a new approach to functional regression based on fuzzy evidence theory. This method uses a training set for computing a fuzzy belief structure which quantifies different types of uncertainties, such as nonspecificity, conflict, or low density of input data. The method can cope with a very large class of training data, such as numbers, intervals, fuzzy numbers, and, more generally, fuzzy belief structures. In order to limit calculations and improve output readability, we propose a belief structure simplification method, based on similarity between fuzzy sets. The proposed model can provide predictions in several different forms, such as numerical, probabilistic, fuzzy or as a fuzzy belief structure. To validate the model, we propose some simulations and compare the results with classical or fuzzy regression methods.

Keywords

Dempster-Shafer, Fuzzy data, Imprecise data, Uncertainty, Regression analysis, Hierarchical classification.

1 Introduction

La modélisation de phénomènes à partir de données observées $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^n$ relatives à des variables d'entrée $\boldsymbol{x}$ et de sortie $\boldsymbol{y}$ est un problème qui couvre des disciplines diverses telles que l'estimation fonctionnelle, la théorie de l'apprentissage ou l'identification de systèmes. En général, le principe consiste à sélectionner parmi une classe $\mathcal{M}$ de fonctions de l'entrée $\boldsymbol{x}$ du système considéré, celle qui « s'adapte le mieux » à la sortie y selon un critère C déterminé. Les méthodes statistiques classiques supposent une modélisation probabiliste des entrées et des sorties. Si le modèle est connu *a priori* à un ensemble de paramètres près, le choix de la fonction « idéale » f revient à identifier les paramètres. Dans le cas le plus général où l'on ne dispose d'aucune information particulière, la modélisation est non-paramétrique. Le choix du critère C et de la classe $\mathcal{M}$ conduisent à différentes méthodes d'approximation fonctionnelle, le critère le plus fréquemment utilisé étant la minimisation du coût empirique quadratique. On peut répartir également les différentes approches statistiques en deux catégories, selon que l'approximation de la fonction est réalisée globalement, c'est-à-dire sur l'ensemble du domaine de variation possible de l'entrée, ou au contraire localement. Parmi les estimateurs fonctionnels globaux, on peut citer les modèles paramétriques comme les modèles linéaires généralisés, les splines [8], les perceptrons multi-couches et les méthodes de projection, comme les ondelettes ou les séries polynomiales. Dans les méthodes locales, comme les fonctions de base radiale généralisées, les régresseurs à noyaux [8], les arbres de régression, la méthode des k plus proches voisins ou les méthodes de mémorisation [2], la description du modèle est divisée en différentes parties de l'espace d'entrée.

Bien que les méthodes de régression statistiques soient les plus répandues pour modéliser les relations entre variables, elles ne sont pas toujours bien adaptées aux différents types d'incertitude que l'on peut rencontrer dans des applications réelles. La théorie des croyances [3, 13], la théorie des ensembles flous, la théorie des possibilités, permettent la représentation de plusieurs types d'incertitude, comme le conflit, l'imprécision ou l'ignorance [11]. La théorie des croyances offre un cadre plus général que la théorie des probabilités pour l'analyse des données. Plusieurs approches, comme la régression linéaire par intervalle ou floue [6], et les systèmes flous [19, 16] ou neuro-flous [10] intégrent des éléments flous, entrées ou paramètres, dans l'estimation fonctionnelle.

Dans cet article, nous proposons une nouvelle méthode de régression basée sur la théorie des croyances floues, que l'on peut appliquer non seulement à des données définies par des nombres réels, mais aussi à des données imprécises représentées par des intervalles ou des nombres flous. Le cas des sorties conflictuelles est également pris en compte : pour une même entrée, on dispose de plusieurs sorties qui peuvent être contradictoires. L'information fournie sur la sortie est délivrée sous la forme très générale d'une structure de croyance. Cette structure est déterminée à partir de l'ensemble d'apprentissage, selon un principe récemment développé par Denœux [5, 4] dans un contexte de discrimination. La généralisation de la définition de l'espérance en théorie de Dempster - Shafer permet de définir une sortie ponctuelle à partir d'une distribution de probabilité compatible avec la structure de croyance obtenue. En particulier, le concept de probabilité pignistique défini par Smets [15] permet de transformer simplement la structure de croyance en une distribution de probabilité sur les valeurs possibles.

Afin d'éviter une augmentation excessive et coûteuse du nombre d'éléments focaux pendant la phase de combinaison des structures, une méthode de simplification est également proposée. Celle-ci tient compte de deux critères, la similarité entre les éléments focaux, qui peuvent être des ensembles flous, et la masse affectée à ces éléments. Les ensembles flous semblables sont regroupés et les ensembles flous de faible masse sont éliminés ou absorbés par des ensembles flous plus importants.

Cet article est organisé de la façon suivante. Dans la section 2, nous rappelons brièvement les notions de la théorie de lévidence et son extension à la théorie des ensembles flous. En section 3, nous présentons notre approche et ses variantes. L'algorithme de simplification est présenté dans la section 4. Finalement, nous présentons quelques simulations dans la section 5 et nous concluons dans la section 6.

2 Structures de croyance floues

Dans cette section, nous rappelons des éléments de la théorie des croyances [3, 13, 15] que nous utiliserons ultérieurement. Pour les définitions de base, nous adoptons ici essentiellement le point de vue que Smets a développé dans son *Modèle de Croyances Transférables* [15]. Soit Ω un ensemble fini, et $S(\Omega)$ l'ensemble de ses parties. Le concept fondamental de la théorie de l'évidence est celui de la *structure de croyance*, définie comme une fonction m de $S(\Omega)$ dans $[0,1]$ et vérifiant: $\sum_{A \subseteq \Omega} m(A) = 1$. Une structure de croyance m telle que $m(\emptyset) = 0$ est dite *normale*. Un sous-ensemble de Ω tel que $m(A) > 0$ est appelé élément focal de m. Nous noterons $F(m)$, l'ensemble des éléments focaux de m. L'information fournie par une structure de croyance peut être représentée par une *fonction de crédibilité* ou une *fonction de plausibilité* definies, respectivement, par : $\text{bel}(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$ et $\text{pl}(A) = \sum_{B \cap A = \emptyset} m(B) = \text{bel}(\Omega) - \text{bel}(\bar{A})$. Ces fonctions peuvent être considérées respectivement comme les bornes inférieure et supérieure de la famille des distributions compatibles avec m. L'une d'elles, la probabilité *pignistique*, est particulièrement utile pour la prise de décision [15]; elle est définie par :

$$\text{BetP}(\omega) = \sum_{A \subset \Omega} \frac{m(A)}{|A|} \delta_A(\omega) \quad \forall \omega \in \Omega$$

où δ est le symbole de Kronecker, et $|A|$ la cardinalité de A. La théorie des croyances peut facilement se généraliser au cas d'un référentiel Ω *continu*, si le nombre d'éléments focaux $|F(m)|$ est *fini* [7]. Toutes les notions vues précédemment se généralisent. La seule modification majeure consiste à remplacer la cardinalité d'un ensemble A par sa mesure de Lebesgue. La distribution pignistique devient alors une loi de probabilité continue ou mixte. C'est cette approche que nous utiliserons dans la suite de cet article.

Une des opérations les plus importantes en théorie de l'évidence est la procédure d'agrégation de structures définies sur le même espace Ω. Soit un opérateur binaire ensembliste ∇. La fusion de deux structures de croyances compatibles m_1 et m_2, notée $m = m_1 \nabla m_2$ est définie par : $m(A) = \sum_{B \nabla C = A} m_1(B) m_2(C)$, pour tout $A \subset \Omega$. On peut alors obtenir une structure non normalisée, c'est-à-dire telle que $m(\emptyset) > 0$, en particulier dans le cas de l'opérateur conjonctif defini initiale-

ment par Dempster, où $\nabla = \cap$. La procédure de normalisation de Dempster convertit une structure non normalisée m en une structure normalisée m^* définie comme suit : $m^*(A) = \frac{m(A)}{1-m(\emptyset)}$ si $A \neq \emptyset$ et $m^*(\emptyset) = 0$.
Une extension de la théorie de l'évidence aux ensembles flous a été proposée par plusieurs auteurs [18, 14, 20]. L'idée de base est que les éléments focaux peuvent être des ensembles flous. Si A est un ensemble flou de Ω, nous noterons $A(.)$ sa fonction d'appartenance et $h(A)$ sa hauteur. Si $\mathcal{F}(\Omega)$ est l'ensemble des sous-ensembles flous de Ω, une structure de croyance floue (SCF) est une fonction m de $\mathcal{F}(\Omega)$ dans $[0, 1]$ telle que:

$$\begin{cases} m(A) = 0 \quad \text{si} \quad A \notin F(m) \\ \sum_{A \in F(m)} m(A) = 1 \end{cases} \tag{1}$$

pour un sous-ensemble *fini* $F(m)$ of $\mathcal{F}(\Omega)$. Là encore, A est un élément focal de m si $m(A) \neq 0$. Si au moins un des éléments focaux n'est pas normalisé, m est dite non normalisée. Yager [19] a proposé une procédure de normalisation des structure de croyance floues appelée « smooth normalization procedure » (SNP), qui généralise à la fois la normalisation d'un ensemble flou et la procédure de normalisation de Dempster définie pour les structures de croyances classiques. La SNP convertit une structure non normalisée m d'éléments focaux $(F_i)_{i=1}^n$ en une structure normalisée m^* d'éléments focaux $(E_i)_{i=1}^n$ de la façon suivante:

$$\begin{cases} E_i(\omega) = F_i(\omega)/h(F_i) \quad \forall \omega \in \Omega \\ m^*(E_i) = \frac{m(F_i)h(F_i)}{\sum_{i=1}^n m(F_i)h(F_i)}. \end{cases} \tag{2}$$

La combinaison de structures de croyance peut également se généraliser aux ensembles flous. Quelque soit l'opérateur de combinaison employé ∇, on utilise sa version floue. Par exemple, si m_1 et m_2 sont deux structures de croyance d'éléments focaux flous $(A_i)_{i=1}^n$ and $(B_i)_{i=1}^p$ respectivement, et $m = m_1 \cap m_2$, les éléments focaux de m sont définis par : $C_k = A_i \cup B_j$, où $C_k(\omega) = T(A_i(\omega), B_j(\omega))$ et T est une t-norme. Les concepts d'espérance, de crédibilité, de plausibilité ou de probabilité pignistique peuvent également se généraliser au flou. Ainsi, $\mathrm{BetP}(\omega) = \sum_{A \in F(m)} \frac{m(A)}{|A|} A(\omega)$, où $|A|$ représente la cardinalité de l'ensemble flou A, définie par $|A| = \sum_{\omega \in \Omega} A(\omega)$ si Ω est discret.

3 Application en régression

3.1 Modèle proposé

Dans cet article, nous proposons d'utiliser la théorie des croyances dans le cadre de la régression. Le principe de notre approche a été introduit par Denœux [4, 5] dans le contexte de la discrimination.

Soient $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \ldots \mathcal{X}_r$ et $\mathcal{Y}$ des espaces de référence réels et continus, domaines de variations respectifs de l'entrée de dimension r et de la sortie monodimensionnelle d'un système. On suppose dans cette section que les entrées $\boldsymbol{x}$ sont réelles mais que les sorties y peuvent être des réels, des intervalles, des nombres flous ou de façon générale, des quantités floues. Soit $(\boldsymbol{x}_i, \tilde{y}_i)_{i=1,\ldots,n}$ un ensemble de vecteurs d'apprentissage appartenant à $\mathcal{X} \times \mathcal{F}(\mathcal{Y})$.
Soit $\boldsymbol{x}$ un vecteur d'entrée arbitraire, et y la sortie correspondante, supposée inconnue. Chaque élément $(\boldsymbol{x}_i, \tilde{y}_i)$ de l'ensemble d'apprentissage donne une information sur la valeur possible de y qui peut se représenter par une fonction de croyance. Si $\boldsymbol{x}$ est « proche » de $\boldsymbol{x}_i$ selon une certaine métrique d, nous pouvons raisonnablement supposer que y sera également « proche » de $\tilde{y}_i$. Ceci peut se traduire par l'affectation d'une certaine masse de croyance à $\tilde{y}_i$, dont la valeur dépend de la distance entre $\boldsymbol{x}$ et $\boldsymbol{x}_i$. En l'absence d'information supplémentaire, le complément à 1 de la masse peut être alloué à l'ensemble de référence $\mathcal{Y}$. Nous obtenons alors une SCF $m_i(.|\boldsymbol{x})$ sur $\mathcal{Y}$, d'éléments focaux $F(m_i(.|\boldsymbol{x})) = \{\tilde{y}_i, \mathcal{Y}\}$, définie par :

$$\begin{cases} m_i(\tilde{y}_i|\boldsymbol{x}) &= \phi[d(\boldsymbol{x}, \boldsymbol{x}_i)] \\ m_i(\mathcal{Y}|\boldsymbol{x}) &= 1 - \phi[d(\boldsymbol{x}, \boldsymbol{x}_i)] \\ m_i(A|\boldsymbol{x}) &= 0 \quad \forall A \in \mathcal{F}(\mathcal{Y}) \backslash F(m_i(.|\boldsymbol{x})), \end{cases} \tag{3}$$

où ϕ est une fonction décroissante telle que $\phi(0) \in]0, 1[$ et $\lim_{d \to \infty} \phi(d) = 0$. Dans la suite, nous noterons $\phi_i(\boldsymbol{x}) = \phi[d(\boldsymbol{x}, \boldsymbol{x}_i)]$. Par exemple, ϕ est la fonction exponentielle et d la distance euclidienne induite par une matrice symétrique définie positive Σ: $\phi_i(\boldsymbol{x}) = \exp[-(\boldsymbol{x} - \boldsymbol{x}_i)^T \Sigma (\boldsymbol{x} - \boldsymbol{x}_i)]$.
Ainsi, plusieurs types d'incertitude peuvent être clairement identifiés: le flou, l'imprécision ou la nonspécificité (relative à la cardinalité de $\tilde{y}_i$) et l'ignorance, representée par le poids affecté à l'ensemble $\mathcal{Y}$.
Afin de combiner l'information fournie par chaque élément de l'ensemble d'apprentissage, nous utilisons la généralisation de la règle de combinaison de Dempster aux SCF. La structure de croyance finale est alors : $m(.|\boldsymbol{x}) = \cap_{i=1}^n m_i(.|\boldsymbol{x})$. Afin de diminuer le coût de calcul, la combinaison peut éventuellement se restreindre aux k plus proches voisins de $\boldsymbol{x}$ dans l'ensemble d'apprentissage : $\{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(k)}\}$. La SCF finale devient alors : $m(.|\boldsymbol{x}) = \cap_{i=1}^k m_{(i)}(.|\boldsymbol{x})$.
Parmi les diverses caractéristiques déduites de $m(.|\boldsymbol{x})$, la probabilité pignistique est définie par :

$$\mathrm{BetP}(y|\boldsymbol{x}) = \sum_{A \in F(m^*)} \frac{m^*(A|\boldsymbol{x})}{|A|} A(y),$$

où $|A| = \int_{\mathcal{Y}} A(y)dy$, puisque $\mathcal{Y}$ est un espace continu et m^* est la structure normalisée correspondant à m.

Une estimation ponctuelle $\hat{y}$ peut alors être obtenue en prenant l'espérance de y selon cette loi de probabilité particulière :

$$\hat{y} = \mathbb{E}_{\text{BetP}}(Y|X = \boldsymbol{x}) = \sum_{A \in F(m)} \frac{m(A|\boldsymbol{x})}{|A|} \int_{\mathcal{Y}} uA(u)du.$$

Soit y_A^* le centre de gravité de A : $y_A^* \triangleq \frac{\int_{\mathcal{Y}} yA(y)dy}{\int_{\mathcal{Y}} A(y)dy}$. Alors $\hat{y} = \sum_{A \in F(m(.|\boldsymbol{x}))} m(A|\boldsymbol{x}) y_A^*$.

Dans la suite, le modèle defini par Eq. (3)sera noté SCF1.

Comme cela a été souligné par Ph. Smets (communication personnelle), si les $\tilde{y}_i$ sont des nombres réels, $m_i(.|\boldsymbol{x})$ et $m(.|\boldsymbol{x})$ deviennent des structures de croyance classiques avec seulement $n+1$ éléments focaux et la structure finale peut s'exprimer facilement (modèle SCF2) :

$$\left\{ \begin{array}{lcl} m(\{y_i\}|\boldsymbol{x}) & = & \frac{1}{K}\phi_i(\boldsymbol{x}) \prod_{j \neq i}[1 - \phi_j(\boldsymbol{x})] \\ m(\mathcal{Y}|\boldsymbol{x}) & = & \frac{1}{K} \prod_{i=1}^{n}[1 - \phi_i(\boldsymbol{x})] \\ m(A|\boldsymbol{x}) & = & 0 \quad \forall A \in \mathcal{F}(\mathcal{Y}) \setminus F(m(.|\boldsymbol{x})), \end{array} \right. \quad (4)$$

où $K = \prod_{i=1}^{n}[1 - \phi_i(\boldsymbol{x})] + \sum_{i=1}^{n} \phi_i(\boldsymbol{x}) \prod_{j \neq i}[1 - \phi_j(\boldsymbol{x})]$ est le facteur de normalisation.

3.2 Sorties conflictuelles

Nous proposons de généraliser l'équation (3) au cas où il existe une ambiguïté entre différentes sorties possibles pour une même entrée $\boldsymbol{x}_i$. On suppose ici que la connaissance relative à y_i peut se représenter de manière très générale par une SCF que l'on notera m_{y_i}, telle que $F(m_{y_i}) = \{\tilde{y}_{i1}, \ldots, \tilde{y}_{ij}, \ldots, \tilde{y}_{i,J(i)}\}$, chaque $\tilde{y}_i$ étant une quantité floue et $m(\tilde{y}_i) = p_{ij}$.

Pour un nouveau vecteur $\boldsymbol{x}$, l'information fournie par $(\boldsymbol{x}_i, m_{y_i})$ concernant la sortie y correspondant à $\boldsymbol{x}$, peut encore être modélisée par une SCF $m_i(.|\boldsymbol{x})$ sur $\mathcal{Y}$, dont les éléments focaux sont $F(m_i(.|\boldsymbol{x})) = F(m_{y_i}) \cup \mathcal{Y}$. La SCF $m_i(.|\boldsymbol{x})$ est définie par (SCF 3):

$$\left\{ \begin{array}{lcl} m_i(\tilde{y}_{ij}|\boldsymbol{x}) & = & p_{ij}\phi_i(\boldsymbol{x}) \quad \forall j \in \{1, \ldots, J(i)\} \\ m_i(\mathcal{Y}|\boldsymbol{x}) & = & 1 - \phi_i(\boldsymbol{x}) \\ m_i(A|\boldsymbol{x}) & = & 0 \quad \forall A \in \mathcal{F}(\mathcal{Y}) \setminus F(m_i(.|\boldsymbol{x})). \end{array} \right. \quad (5)$$

La structure finale $m(.|\boldsymbol{x})$ est obtenue en combinant comme précédemment les structures individuelles m_i. Là encore, le nombre d'éléments focaux $|F(m(.|\boldsymbol{x}))|$, qui peut atteindre dans le pire des cas $\prod_{i=1}^{n} J(i)$, peut se réduire nettement si l'on considère uniquement les k plus proches voisins de $\boldsymbol{x}$.

3.3 Utilisation de prototypes

Nous supposerons dans cette section que les y_i sont réels. Si l'ensemble d'apprentissage est de « grande » taille, l'information qu'il contient peut être résumée par un ensemble de prototypes. Soit $C = \{\boldsymbol{c}_i \in \mathcal{X}, i \in \{1, \ldots, I\}\}$, et $E = \{e_j \in \mathcal{Y}, j \in \{1, \ldots, J\}\}$ deux ensembles respectivement de I et J prototypes des vecteurs d'entrée et de sortie. On suppose que ces prototypes ont été obtenus à l'aide d'une méthode de classification automatique quelconque, comme par exemple l'algorithme des centres mobiles (flous ou non), un de ses dérivés, ou les cartes auto-organisatrices de Kohonen. Si nous associons à chaque $\boldsymbol{x} \in \mathcal{X}$ son plus proche prototype dans C selon une certaine distance, l'espace $\mathcal{X}$ peut être partitionné en I classes $(C_i)_{i=1}^{I}$, caratérisés par les centres $\boldsymbol{c}_i$ et les matrices de dispersion estimées $\Sigma_i^x = \frac{1}{N}\sum_{k=1}^{N} u_{ik}(\boldsymbol{x}_k - \boldsymbol{c}_i)^T(\boldsymbol{x}_k - \boldsymbol{c}_i)$, où u_{ik} est le degré d'appartenance de $\boldsymbol{x}_k$ à la classe C_i. De même, $\mathcal{Y}$ est partitionné en J classes $(E_j)_{j=1}^{J}$.

On suppose que les $(\boldsymbol{x}_k, y_k)$ sont des réalisations de N vecteurs aléatoires indépendents $(\boldsymbol{X}_k, Y_k)$ de l'espace probabilisé sous-jacent $((\mathcal{X}, \mathcal{Y}), \mathcal{A}, \mathcal{P}_{\mathcal{X}\mathcal{Y}})$, et on note p_{ij} la probabilité conditionnelle $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}(Y \in E_j | X \in C_i)$. Afin de tenir compte de la dispersion autour des prototypes e_j, ceux-ci peuvent être « fuzzifiés » en faisant intervenir l'écart-type estimé σ_j^y de E_j. On notera l'ensemble flou ainsi créé : $\tilde{e}_j \in \mathcal{F}(\mathcal{Y})$. Par exemple, puisque les variables sont réelles et continues, les $\tilde{e}_j$ peuvent être des nombres flous gaussiens, notés (e_j, σ_j^y) et définis par : $\tilde{e}_j(u) = \exp\left\{-\frac{1}{2}(\frac{u - e_j}{\sigma_j^y})^2\right\}, \forall u \in \mathcal{Y} \subset \mathbb{R}$. D'autres types de fuzzification tenant compte de manière équivalente de l'imprécision relative au degré d'appartenance à la classe considérée peuvent être bien entendu envisagés.

Pour un vecteur $\boldsymbol{x}$ quelconque, chaque $\boldsymbol{c}_i$ fournit une information concernant la sortie inconnue y, représentée par une SCF $m_i(.|\boldsymbol{x})$ de façon analogue aux équations (3-5):

$$\left\{ \begin{array}{lcl} m_i(\tilde{e}_j|\boldsymbol{x}) & = & p_{ij}\phi[d(\boldsymbol{x}, \boldsymbol{c}_i)] \quad \forall j \in \{1, \ldots, J\} \\ m_i(\mathcal{Y}|\boldsymbol{x}) & = & 1 - \phi[d(\boldsymbol{x}, \boldsymbol{c}_i)] \\ m_i(A|\boldsymbol{x}) & = & 0 \quad \forall A \in \mathcal{F}(\mathcal{Y}) \setminus \{\tilde{e}_1, \ldots \tilde{e}_J, \mathcal{Y}\}. \end{array} \right. (6)$$

Ici, $|F(m(.|\boldsymbol{x}))| \leq 2^J$. Le nombre d'éléments focaux est généralement beaucoup plus faible que dans le modèle complet SCF1. Dans la suite, on notera ce modèle SCF4.

Si nous «fuzzifions » également les coordonnées c_{ij} de chaque vecteur $\boldsymbol{c}_i$ à l'aide de nombres flous gaussiens $\tilde{c}_{ij} = (c_{ij}, \sigma_{ij}^x) \in \mathcal{X}_j$, en notant $\tilde{\boldsymbol{c}}_i$ le produit cartésien des $(\tilde{c}_{ij})_{j=1}^{r}$, on obtient l'égalité suivante quel que soit $x \in \mathcal{X}$: $\tilde{\boldsymbol{c}}_i(\boldsymbol{x}) = \phi[d(\boldsymbol{x}, \boldsymbol{c}_i)]$ si ϕ est la fonction exponentielle.

La matrice de probabilités conditionnelles d'appartenance aux différentes classes $P = (p_{ij})$ n'est en général pas connue. Nous proposons d'utiliser une approche fréquentiste pour estimer ces probabilités conditionnelles. Soit π_{ij} la probabilité qu'un couple $(\boldsymbol{x}, y)$ appartienne simultanément aux classes C_i et E_j. Par définition, $\pi_{ij} = P_{\mathcal{X}\mathcal{Y}}((X,Y) \in C_i \cap E_j)$. Si N_{ij} est la variable aléatoire dénombrant les couples (x_k, y_k) appartenant à $C_i \times E_j$, par définition, N_{ij} suit une loi binomiale: $N_{ij} \sim \mathcal{B}(n, \pi_{ij})$. La Loi Forte des Grands Nombres justifie théoriquement l'estimation proposée :

$$\widehat{p}_{ij} = \widehat{P_{Y|X}}(E_j|C_i) = \frac{\widehat{P_{\mathcal{X}\mathcal{Y}}}(E_j \cap C_i)}{\widehat{P_{\mathcal{X}}}(C_i)} = \frac{n_{ij}}{\sum_j n_{ij}}$$

En particulier, $\widehat{p}_{ij} \overset{n\to+\infty}{\longrightarrow} \mathcal{P}_{\mathcal{Y}|\mathcal{X}}(Y \in E_j | X \in C_i)$ presque sûrement.

Dans cette approche, se pose également le problème de la détermination du nombre optimal de prototypes. Cet aspect dépasse le cadre spécifique de notre modèle et relève du problème général de l'identification de systèmes. Les techniques classiques de rééchantillonnage (Bootstrap, Jackknife, Validation Croisée) basées sur la minimisation d'un critère C global comme le coût quadratique (cf. section 1) sont envisageables ici, dans le cas où les données sont réelles et où l'ensemble d'apprentissage couvre le domaine d'étude, mais sont assez coûteuses en temps de calcul. D'autres techniques, utilisées par exemple dans le contexte de l'identification de systèmes flous [9], qui offre des similitudes avec notre modèle, sont plutôt basées sur des critères de classification optimale. Parmi les nombreux critères proposés, certains sont basés sur les matrices de covariance empiriques floues, d'autres sur des mesures de dispersion intra-classes :

$$SSW = \sum_{i=1}^{m} \sum_{k=1}^{n} u_{ik} d^2(\boldsymbol{x}_k, \boldsymbol{c}_i).$$

En particulier, une approche simple mais efficace est de considérer le pourcentage d' « inertie expliquée » par le modèle SSW/SST, où $SST = \sum_{k=1}^{n} d^2(\boldsymbol{x}_k, \bar{\boldsymbol{x}})$ et $\bar{\boldsymbol{x}}$ est le vecteur - moyenne. Ce rapport I est une fonction décroissante du nombre de prototypes. Une valeur convenable I^* peut être choisie en imposant un seuil minimal acceptable.

Nous pouvons de même déterminer le nombre optimal de prototypes de sortie e_i, et ce, de manière indépendante aux $\boldsymbol{c}_i$. Cependant, il paraît également légitime de construire e_i à partir de $\boldsymbol{c}_i$ en utilisant une fonction de régression g simple, elle-même facile à identifier. On a alors $e_i = g(\boldsymbol{c}_i)$.

Dans notre méthode, nous avons utilisé le régresseur de Nadaraya-Watson à noyau gaussien uniforme. Ainsi :

$$e_i = \frac{\sum_{k=1}^{n} y_k \exp[-(\boldsymbol{c}_i - \boldsymbol{x}_k)^T \Lambda (\boldsymbol{c}_i - \boldsymbol{x}_k)]}{\sum_{k=1}^{n} \exp[-(\boldsymbol{c}_i - \boldsymbol{x}_k)^T \Lambda (\boldsymbol{c}_i - \boldsymbol{x}_k)]},$$

où Λ est une matrice diagonale: $\Lambda = (\lambda_j)_{j=1}^r$. Les valeurs optimales des λ_j peuvent être obtenues en minimisant le critère classique de validation croisée.

4 Procédure de simplification

L'inconvénient majeur des méthodes basées sur la théorie de l'évidence est qu'elles sont généralement coûteuses en temps de calcul pendant la phase de combinaison des structures, à cause de la multiplication des éléments focaux. Par exemple, dans le modèle SCF4, on peut obtenir jusqu'à 2^J éléments focaux. De plus, outre la lenteur des calculs, le résultat de la structure finale m pose également un problème de lisibilité. En effet, la plupart des éléments focaux de m sont difficiles à interpréter et certains sont quasiment indiscernables. Pour ces deux raisons, la rapidité des calculs et l'interprétabilité de la structure, il convient donc de contrôler sa complexité au cours de la phase de combinaison. Dans cet esprit, nous proposons une procédure de simplification d'une structure de croyance quelconque (classique ou floue). Plusieurs méthodes ont été proposées dans le cas des structures de croyances classiques [1]. La méthode que nous proposons est basée sur l'*agrégation d'ensembles flous semblables*. De façon générale, toute méthode de regroupement, de classification automatique, étendue aux ensembles flous peut convenir. Dans cet article, nous avons envisagé une classification de type hiérarchique. La procédure est itérative : à chaque étape t, on remplace deux ensembles similaires par un seul. Ceci nécessite de définir, d'une part, un *critère de proximité I* entre ensembles flous, d'autre part, un *mode d'agrégation* d'ensembles flous.

Soit $m^{(t)}$ la SCF à l'étape t, dont les éléments focaux sont $\{F_1, \ldots, F_n\}$. Une nouvelle SCF $m^{(t+1)}$ est obtenue en remplaçant les deux plus proches éléments focaux F_i et F_j de m par un ensemble F' et en conservant les autres éléments focaux. Ainsi, $F(m^{(t+1)}) = (F(m^{(t)}) \cup F') \backslash (F_i \cup F_j)$, avec $m^{(t+1)}(F') = m^{(t)}(F_i) + m^{(t)}(F_j)$.

Parmi les nombreuses mesures de similarité entre ensembles flous définies dans la littérature [12, 21], nous avons choisi la mesure ensembliste suivante : $I(A,B) = S(A,B) = \frac{|A\cap B|}{|A \cup B|}$. On choisit les ensembles F_i et F_j tels que $S(F_i, F_j) = \max_{k,l} S(F_k, F_l)$. D'autres critères, tenant également compte de la masse des éléments focaux, sont envisageables. Ainsi, par analogie avec la méthode de Ward [17] en classification hiérarchique classique, on peut définir le crière suivant :

$$I(A,B) = \frac{m^{(t)}(A) + m^{(t)}(B)}{m^{(t)}(A) m^{(t)}(B)} \frac{S(A,B)}{1 - S(A,B)}$$

qui permet d'absorber les éléments focaux de faible importance. On peut aussi de manière plus systématique définir une distance entre les structures $m^{(t)}$ et $m^{(t+1)}$, basée sur des critères d'information mutuelle, mais ceci dépasse le cadre de cet article.

De nombreux opérateurs binaires d'agrégation ∇ sont *a priori* envisageables, produisant un nouvel ensemble F'. L'utilisation de l'opérateur d'union, ie $F' = F_i \cup F_j$ présente la propriété d'augmenter systématiquement un critère d'information bien connu, la nonspécificité N, définie par :

$$N(m) = \sum_{F \in F(m)} m(F) \log_2 |F|. \tag{7}$$

On vérifie en effet facilement que $N(m^{(t)}) \leq N(m^{(t+1)})$. Cependant, il semble plus judicieux de faire intervenir là encore les masses des éléments focaux. En outre, l'objectif n'est pas ici de chercher des structures « emboîtées » mais des structures semblables. Nous proposons alors l'opérateur plus neutre de la *moyenne pondérée*, telle que:

$$\forall \omega \quad F'(\omega) = \frac{m^{(t)}(F_i)F_i(\omega) + m^{(t)}(F_j)F_j(\omega)}{m^{(t)}(F_i) + m^{(t)}(F_j)}$$

toujours par analogie avec la méthode de Ward.

Cette procédure d'agrégation s'arrête dès que, pour un seuil δ donné, $\max_{k,l} S(F_k, F_l) < \delta$.

5 Résultats

Afin de montrer la validité et les capacités de prédiction de notre méthode, nous présentons les résultats de quelques simulations. Dans le cas de données réelles, la précision de l'estimation est mesurée en terme d'erreur quadratique moyenne (EQM) : $C = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2$. Afin de décrire l'information que contiennent les sorties, deux mesures d'incertitude classiques [11] ont été utilisées : la nonspécificité (cf. équation 7), qui représente l'imprécision des ensembles, et la discordance (*strife*):

$$S(m) = -\sum_{A \in F(m)} m(A) \log_2 \sum_{B \in F(m)} m(B) \frac{|A \cap B|}{|A|}.$$

Exemple 1: Dans ce problème de régression, $x \in \mathbb{R}$ est tiré d'un mélange de 2 distributions gaussiennes de mêmes proportions : $X \sim 0.5\mathcal{N}(-1.5, 0.5) + 0.5\mathcal{N}(1.75, 0.5)$ et $y = \sin(3x) + x + \epsilon(x)$, où $\epsilon(x) \sim \mathcal{N}(0, 0.01)$ si $x \leq 0$ and $\epsilon(x) \sim \mathcal{N}(0, 1)$ si $x > 0$. Ici, $N = 200$. Nous avons utilisé la version SCF 4 de notre méthode avec $I = J = 25$ prototypes et $k = 3$ voisins pour la combinaison. La figure 1 montre l'intervalle interquartile de la distribution pignistique BetP calculée pour chaque x. Nous avons comparé les résultats à ceux d'un système

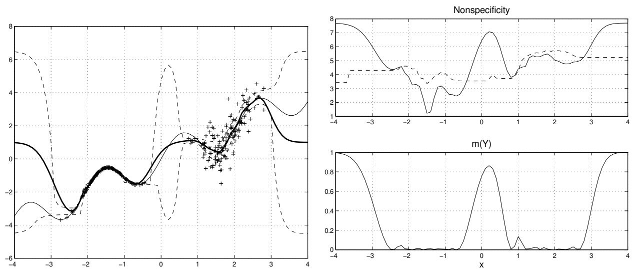

FIG. 1 - **Exemple 1. A gauche**: *données réelles (-); estimations (-); 1er et 3eme quartiles de BetP (- -); ensemble d'apprentissage (+)*; **A droite**: *nonspécificité avec SCF 4 (-) et le modèle de Yager (- -), et masse affectée au domaine entier $m(\mathcal{Y}|\boldsymbol{x})$ avec SCF 4.*

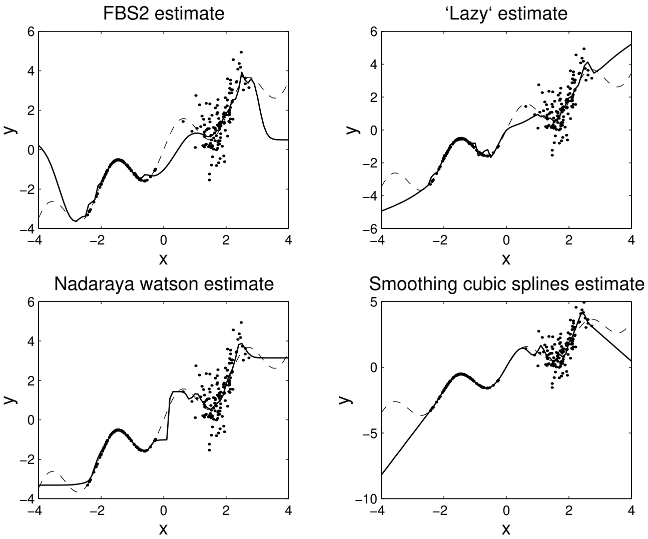

FIG. 2 - **Exemple 1.** *ensemble d'apprentissage (.); données réelles (–); estimation ponctuelle pour différents modèles (-)*

flou proposé par Yager [19]. Les régions de faible densité sont caractérisées par des valeurs élevées pour la nonspécificité et l'ignorance $m(\mathcal{Y})$ et par un intervalle interquartile très étendu. Par contre, dans le modèle de Yager, la nonspécificité n'augmente pas dans ces régions. Dans notre modèle, les résultats confirment également une plus grande nonspécificité et un intervalle interquartile moins précis pour les régions ou la variance est plus grande.

Afin de vérifier l'efficacité de notre méthode en termes de précision, nous l'avons comparé à plusieurs méthodes de régression classiques (cf. figure 2). Le tableau 1 indique l'EQM obtenue pour deux méthodes de régression non paramétriques classiques, les splines cubiques (SC) et le régresseur de Nadaraya-Watson

	SC	NW	"Lazy"	Yager	SCF 2
EQM 2	0.035	0.016	0.019	0.010	0.012
EQM 1	10^{-5}	0.003	0.013	0.008	0.011

TAB. 1 - *EQM pour différents modèles de régression: EQM 1 pour* $x \in [-2.5, -0.7]$*, EQM 2 pour* $x \in [-2.5, -0.7] \cup [0.8, 2.6]$.

(NW), une méthode de mémorisation ("Lazy")[2], le modèle de Yager et une version de notre méthode (SCF 2). Les EQM ont été calculées uniquement dans les régions de fortes densités, dans deux zones : l'une contient les données bien spécifiées, ($x \in [-2.5, -0.7]$). L'autre inclut également la région de plus grande variance : ($x \in [-2.5, -0.7] \cup [0.8, 2.6]$). Comme le montre le tableau 1, les performances de notre méthode sont à peu près équivalentes aux autres, même si les splines semblent être plus adaptées au cas des données homogènes, dans cet exemple. Cependant, cette méthode globale s'adapte moins aux données hétérogènes, ce qui ne semble pas le cas pour notre méthode.

Exemple 2 : Soit un ensemble de données issues de l'équation de Monod $y = \frac{ax}{b+x} + \epsilon(x), x \in [0, 2]$, où $\epsilon(x) \sim \mathcal{N}(0, 0.001)$ si $x \leq 1$ et $\epsilon(x) \sim \mathcal{N}(0, 0.01)$ si $x > 1$. Si $x \in [0.5, 1.5[$, $(a, b) = (.35, .5)$, $(a, b) = (.35, .5)$ ou $(.45, .45)$ avec la même probabilité $p_1 = p_2 = .5$ et $(a, b) = (.35, .5)$ si $x \in [0, 2] \backslash [0.5, 1.5[$ (cf. figure 3). Nous avons utilisé la version SCF 4, avec $I = 20, J = 30$ et $k = 3$ voisins, et les prototypes sont créés comme dans l'exemple 1, mais en utilisant deux fonctions - cibles, dépendant des valeurs des paramètres a et b. Nous avons utilisé l'opérateur « moyenne pondérée » dans la procédure de simplication, avec différents seuils d'agrégation δ pendant la combinaison. S'il n'y a pas de simplification, $\delta = 1$: même dans ce cas extrême, le nombre déléments focaux $|F(m)|$ est très inférieur à 2^{30} car la plupart des probabilités conditionnelles $\widehat{p_{ij}}$ sont nulles. La figure 4 montre que $|F(m)|$ augmente dans les régions de sorties incertaines et décroît très vite en fonction de δ. Les figures (3-4), présentent des résultats pour le cas où $\delta = 0.5$. Les 4 régions peuvent être facilement caractérisées par les mesures d'incertitude choisies. Les régions (1-2) ($x \in [0, 1]$) ont des valeurs de nonspécificité élevées et le conflit est bien représenté dans les régions (2-3) ($x \in [0.5, 1.5]$). La probabilité pignistique calculée pour 4 valeurs typiques ($x = 0.2; 0.6; 1.3; 1.6$) résume bien les quatre situations.

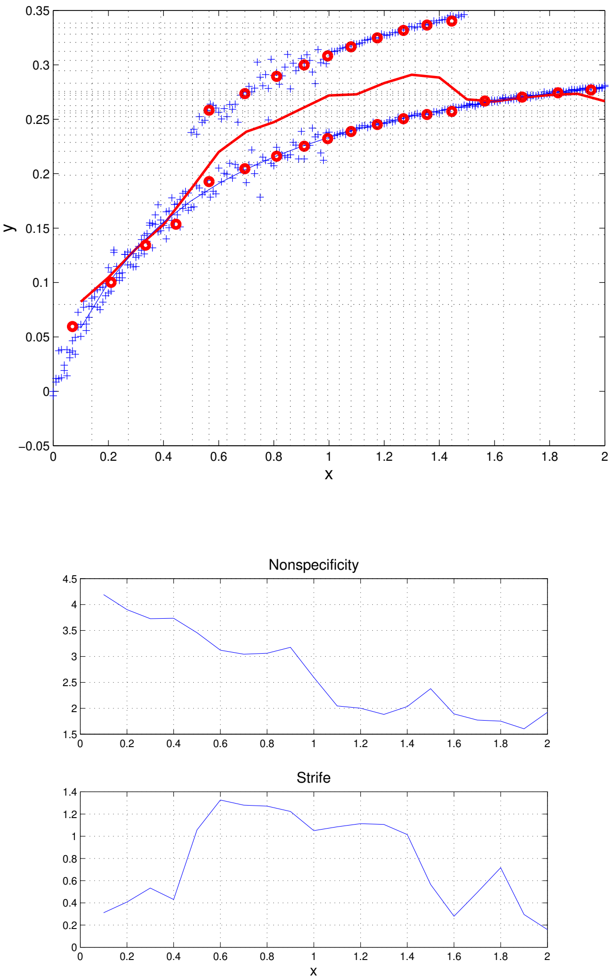

FIG. 3 - **Exemple 2. En haut**: *ensemble d'apprentissage (+); estimation ponctuelle (-), prototypes (o) et classes (:)* **En bas**: *nonspécificité et discordance (SCF 4)(-).*

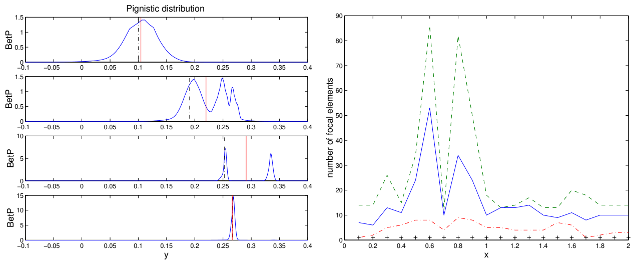

FIG. 4 - **Exemple 2. A gauche**: *distribution pignistique pour 4 entrées:* $x = 0.2 - 0.6 - 1.3 - 1.6$ *(-); donnée réelle (–); estimation ponctuelle (-)* **A droite**: *Nombre d'éléments focaux pour* δ *= 1 (–), 0.5(-), 0.2 (-.-) and 0 (+).*

6 Conclusion

Dans cet article, nous avons proposé une nouvelle méthode de régression basée sur la théorie des croyances floues. Cette approche permet de prendre en compte différents types de données, comme les intervalles, les nombres flous ou plus généralement des structures de croyance floues. En termes de précision, cette méthode obtient des résultats comparables aux techniques statistiques dans le cas de données réelles. Ainsi, elle peut être considérée comme une généralisation de l'estimation fonctionnelle à des données incertaines. Deux types d'information sont proposés : une représentation de l'incertitude (imprécision, indétermination, ignorance) et une estimation ponctuelle. Une méthode de simplification de structures de croyance est proposée, afin de pallier l'augmentation indésirable du nombre d'éléments focaux pendant la phase de combinaison. La méthode repose sur l'agrégation d'éléments focaux similaires. La motivation de cette simplification est double : *pendant* la combinaison, le temps de calcul est plus faible, *après* la combinaison, les résultats sont plus facilement interprétables.

Références

[1] M. Bauer. Approximations for decision making in the Dempster-Shafer theory. In *Uncertainty in Artificial Intelligence*, pages 339–344, Saarbrücken, 1996.

[2] H. Bersini, G. Bontempi, and M. Birattari. Is readability compatible with accuracy? from neuro-fuzzy to lazy learning. In *Fuzzy-Neuro Systems '98*, pages 10–25, Munich, 1998.

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

[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*, 30(7):1095–1107, 1997.

[6] 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, Norwell, 1998.

[7] J.W. Guan and D.A. Bell. A generalization of the Dempster-Shafer theory. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 10–25, Chambéry, France, 1993.

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

[9] H. Hellendoorn and D. Driankov. *Fuzzy model identification: selected approaches.* Springer, 1997.

[10] J.S.R Jang, C.T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing.* Prentice-Hall, 1997.

[11] G. J. Klir and B. Yuan. *Fuzzy sets and fuzzy logic: Theory and applications.* Prentice Hall, Upper Sadddle River, N.J., 1995.

[12] M. Setnes, R. Babuška, U. Kaymak, and H. R. van Nauta Lemke. Similarity measures in fuzzy rule base simplification. *IEEE Transactions on Systems, Man and Cybernetics - part B: Cybernetics*, 28:376–386, 1998.

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

[14] P. Smets. The degree of belief in a fuzzy event. *Information sciences*, 28:45–62, 1982.

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

[16] T. Takagi and M. Sugeno. Fuzzy identification of systems an its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15:116–132, 1985.

[17] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, 58:236–244, 1963.

[18] R. R. Yager and D. P. Filev. Generalized probabilities of fuzzy events from fuzzy belief structures. *Information sciences*, 28:45–62, 1982.

[19] R. R. Yager and D. P. Filev. Including probabilistic uncertainty in fuzzy logic controller modelling using Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25:1221–1230, 1995.

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

[21] R. Zwick, E. Carlstein, and D. V. Budescu. Measures of similarity among fuzzy concepts: a comparative analysis. *International Journal of Approximate Reasoning*, 1:221–242, 1987.