Classification par régression floue et crédibiliste à base de machines à vecteurs de support

Classification by fuzzy and belief regression with support vector machines

H. Laanaya[1 2] A. Martin[1] A. Khenchaf[1] D. Aboutajdine[2]

[1] ENSIETA - E^{3}I^2 EA3876

[2] Université Mohammed V-Agdal

2, rue François Verny 29806 Brest cedex 9, laanayhi, Arnaud.Martin, Ali.Khenchaf@ensieta.fr

Université Mohammed V-Agdal, Faculté des sciences de Rabat, Maroc, aboutaj@fsr.ac.ma

Résumé :

De nombreux classifieurs flous ou crédibilistes ont été développés ces dernières années en vue de répondre au traitement de données de plus en plus incertaines et imprécises. Les approches de classification à base de machines à vecteurs de support (SVM) sont de plus en plus appréciées pour leur simplicité et leur puissance. Le principe des SVM peut également être employé pour réaliser une régression linéaire simple ou multiple. Nous proposons dans cet article une approche de classification à partir de la régression linéaire des SVM sur des fonctions d'appartenance ou des fonctions de croyance. En effet, ces fonctions possèdent les mêmes propriétés, que nous intégrons comme contraintes dans la résolution de notre problème convexe. Nous comparons notre approche à un k-plus proches voisins flou et à un k-plus proches voisins crédibiliste sur des données générées.

Mots-clés :

SVM, régression, fonctions de croyance, fonctions d'appartenance.

Abstract:

In the last few years, many fuzzy or belief classifiers have been proposed in order to take into account imprecise and uncertain data. The support vector machines (SVM) are more and more employed for classification because of their easiness and performance. The SVM principle can also be used for linear regression that can be simple or multiple. We propose in this article a classification approach based on the SVM linear regression on membership functions or on belief functions. Indeed, these functions have the same properties ; we take as constraints in the resolution of our convex problem. We compare our approach with a fuzzy k-nearest neighbors and with a belief k-nearest neighbors

Keywords:

SVM, regression, belief functions, membership functions.

1 Introduction

L'étude d'environnements de plus en plus complexes étudiés à partir de systèmes également complexes est aujourd'hui nécessaire dans de nombreuses applications. Nous devons alors évaluer ces environnements à partir de données incertaines et imprécises. Plusieurs choix s'offrent alors à nous : soit nous tentons de supprimer ces imperfections, ce qui nécessite une compréhension, souvent difficile, de la physique qui a conduit à ces imperfections ; soit nous cherchons à développer des processus de traitement robustes à ces imperfections ; soit nous cherchons à les modéliser.

Une modélisation fine de données incertaines et imprécises peut être réalisée à l'aide des théories de l'incertain telles que la théorie des sous-ensembles flous [22], la théorie des possibilités [6] ou encore la théorie des fonctions de croyance [5, 16]. Ainsi de nombreuses méthodes de classification ont été développées dans le cadre de ces théories à partir de méthodes existantes telles que les réseaux de neurones, les k-plus proches voisins ou les arbres de décision, donnant naissance à des classifieurs flous [1, 10, 11] ou crédibilistes [3, 2, 4, 20].

Les machines à vecteurs de support (*Sup-*

port Vector Machines (SVM)) récemment introduites par Vapnik [21] sont des méthodes d'optimisation linéaire définies pour une classification linéaire à deux classes. La simplicité de cette approche et sa capacité d'extension à une classification non linéaire à un développement et une utilisation importants des SVM, notamment en régression linéaire [17, 7].

Différentes tentatives ont été proposées pour intégrer le flou dans les SVM. Ainsi l'apprentissage peut être réalisé à partir d'une pondération issue de fonctions d'appartenance [8]. Dans [19], une fonction d'appartenance est introduite pour lever l'ambiguïté dans les zones d'ombre où la classification multi classes est problématique. Enfin [9] propose une régression sur des nombres flous triangulaires qui présente des différences avec notre approche sur lesquelles nous reviendrons plus loin.

Nous proposons dans cet article une méthode de classification à partir de données floues ou crédibilistes fondée sur le principe de régression des SVM. En effet les fonctions d'appartenance et les fonctions de croyance possèdent des propriétés similaires qui sont introduites en tant que contraintes dans l'approche d'optimisation. Nous comparons cette approche à un k-plus proche voisin flou [11] et à un k-plus proche voisin crédibiliste [2]. Les résultats montrent l'intérêt de cette nouvelle approche de classification.

Ainsi, nous rappelons le principe des SVM dans la section suivante, puis nous fixons les notations des fonctions d'appartenance et des fonctions de croyance nécessaires à la compréhension de la suite. Nous décrivons ensuite l'approche proposée de régression à partir des vecteurs de support en vue de la classification. Cette approche est comparée et discutée dans une dernière partie à partir de données générées.

2 Principe du classifieur SVM

L'approche des machines à vecteurs de supports, initiées par Vapnik [21], est avant tout une méthode de classification linéaire à deux classes. Elle tente de séparer des individus issus de deux classes (+1 et -1) en cherchant un l'hyperplan optimal qui sépare les deux ensembles, en garantissant que la marge entre le plus proche des exemples positifs et négatifs soit maximale. Intuitivement, cela garantit un bon niveau de généralisation car de nouveaux individus pourront ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être tout de même situés d'un coté ou de l'autre de la frontière. Un autre intérêt est la sélection de vecteurs de support grâce auxquels est déterminé l'hyperplan optimal. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs de support sont utilisés pour classer un nouveau cas. Cela en fait une méthode très rapide. La force des SVM tient à leur simplicité de mise en œuvre face à des problèmes complexes et à des fondements théoriques solides. Dans le cas où les exemples sont linéairement séparables, on cherche l'hyperplan $y = w.x + b$ qui maximise la marge entre les deux ensembles. Ainsi w est la solution du problème d'optimisation convexe :

$$\text{Min} \quad \|w\|^2/2 \tag{1}$$

sous les contraintes :

$$y_t(w.x_t + b) - 1 \geq 0 \quad \forall t = 1, \ldots, l. \tag{2}$$

où les $x_t \in \mathbb{R}^d$ représentent les l données d'apprentissage , et $y_t \in \{-1, +1\}$ la classe. La solution de ce problème d'optimisation est donnée par le point selle du lagrangien associé :

$$L = \frac{\|w\|^2}{2} - \sum_{t=1}^{l} \alpha_t(y_t(w.x_t - b) - 1), \tag{3}$$

où les $\alpha_t \geq 0$ sont les multiplicateurs de Lagrange, vérifiant $\sum_{t=1}^{l} \alpha_t y_t = 0$.

Dans le cas où les données ne sont pas linéairement séparables, on relâche les

contraintes (2) par le bais de termes positifs ξ_t. Dans ce cas, nous cherchons à minimiser :

$$\frac{1}{2} \parallel w \parallel^2 + C \sum_{t=1}^{l} \xi_t, \tag{4}$$

sous les contraintes données pour tout t :

$$\begin{cases} y_t(w.x_t + b) \geq 1 - \xi_t \\ \xi_t \geq 0 \end{cases} \tag{5}$$

où C est une constante choisie par l'utilisateur. Une grande valeur de C correspond à une grande pénalité aux erreurs. Ce problème se résout de la même façon que le premier, mais avec des multiplicateurs de Lagrange vérifiant $0 \leq \alpha_t \leq C$.

Afin de classer un nouvel élément x il suffit d'étudier la fonction de décision donnée par :

$$f(x) = \text{sign}(\sum_{t \in SV} y_t \alpha_t^0 x_t.x - b_0), \tag{6}$$

où $SV = \{t\ ; \alpha_t^0 > 0\}$ pour le cas séparable et $SV = \{t\ ; 0 < \alpha_t^0 < C\}$ pour le cas non séparable, est l'ensemble des vecteurs de support.

Dans les cas non linéaire, le principe des SVM est de projeter, par une fonction noyau, les données de départ dans un espace de grande dimension (éventuellement infinie) dans lequel les données sont séparables par un hyperplan. Ainsi la classification d'un nouvel élément x est donnée par la fonction de décision :

$$f(x) = \text{sign}(\sum_{t \in SV} y_t \alpha_t^0 K(x, x_t) - b_0) \tag{7}$$

où K est la fonction noyau, dont les plus utilisées sont le noyau polynomial $K(x, x_t) = (x.x_t + 1)^d, \quad d \in \mathbb{N}$, et le noyau gaussien $K(x, x_t) = \exp(-\gamma \|x - x_t\|^2), \quad \gamma \in \mathbb{R}^+$.

3 Théories de l'incertain

Dans de nombreuses situations il est important de pouvoir modéliser les environnements complexes mesurés à partir de systèmes d'information eux-mêmes complexes. Les systèmes d'acquisition et la difficulté liée à la scène fait que nous manipulons rapidement des données incertaines et imprécises. Ainsi, pour une tâche d'identification, nous devons employer des classifieurs pouvant gérer de telles données. Les théories de l'incertain telles que la théorie des sous-ensembles flous [22], la théorie des possibilités [6] ou encore la théorie des fonctions de croyance [5, 16] permettent la modélisation de données incertaines et imprécises.

Dans le cadre de ces théories de nombreux classifieurs ont été développés tels que des réseaux de neurones flous [1] ou crédibilistes [3], des k-plus proches voisins flous [10, 11] ou crédibilistes [2, 23] ou encore des arbres de décision crédibilistes [4, 20].

3.1 Les fonctions d'appartenance

Les fonctions d'appartenance permettent de décrire une appartenance floue à une classe. Ainsi l'appartenance d'une observation x à une classe C_i parmi N_c classes, est donnée par une fonction $\mu_i(x)$ telle que :

$$\begin{cases} \mu_i(x) \in [0, 1] \\ \sum_{i=1}^{n} \mu_i(x) = 1. \end{cases} \tag{8}$$

Dans ce cas, nous considérons les classes floues. Dans le cas de classes nettes, il est possible de considérer les distributions de possibilité. Ces fonctions à valeurs dans $[0, 1]$ ont une contrainte de normalisation différente de la somme précédente, nous ne les étudions pas ici.

Exemple de fonction d'appartenance dans le cas d'un k-plus proches voisins flou

Nous considérons ici l'approche de [11]. Dans un premier temps, nous calculons la fonction d'appartenance d'un vecteur d'apprentissage x_t donnée par :

$$\mu_i(x_t) = \frac{k_i(x_t)}{k_f}, \tag{9}$$

où k_f le nombre de plus proches voisins choisi pour le voisinage flou V_{K_f} et $k_i(x_t) = |C_i \cap$

$V_{K_f}(x_t)|$. Dans un second temps, nous calculons la fonction d'appartenance pour un vecteur x à classifier :

$$\mu_i(x) = \frac{\sum_{t=1}^{l} \frac{\mu_i(x_t)}{\|x - x_t\|^2}}{\sum_{t=1}^{l} \frac{1}{\|x - x_t\|^2}}. \quad (10)$$

La norme employée est la norme euclidienne.

La classe d'appartenance de x est ensuite décidée de manière classique comme la classe donnant le maximum des fonctions d'appartenance.

3.2 Les fonctions de croyance

La théorie des fonctions de croyance est fondée sur la manipulation des fonctions de masse. Les fonctions de masse sont définies sur l'ensemble de toutes les disjonctions du cadre de discernement $\Theta = \{C_1, \ldots, C_{N_c}\}$ et à valeurs dans $[0, 1]$, où C_i représente l'hypothèse "l'observation appartient à la classe i". Généralement, il est ajouté une condition de normalité, donnée par :

$$\sum_{A \in 2^\Theta} m_j(A) = 1, \quad (11)$$

où $m(.)$ représente la fonction de masse. La première difficulté est donc de définir ces fonctions de masse selon le problème. A partir de ces fonctions de masse, d'autres fonctions de croyance peuvent être définies, telles que les fonctions de crédibilité, représentant l'intensité que toutes les sources croient en un élément, et telles que les fonctions de plausibilité représentant l'intensité avec laquelle on ne doute pas en un élément.

Afin de conserver un maximum d'informations, il est préférable de rester à un niveau crédal (*i.e.* de manipuler des fonctions de croyance) pendant l'étape de combinaison des informations pour prendre la décision sur les fonctions de croyance issues de la combinaison. Si la décision prise par le maximum de crédibilité peut être trop pessimiste, la décision issue du maximum de plausibilité est bien souvent trop optimiste. Le maximum de la probabilité pignistique, introduite par [18], reste le compromis le plus employé. La probabilité pignistique est donnée pour tout $X \in 2^\Theta$, avec $X \neq \emptyset$ par :

$$\mathrm{betP}(X) = \sum_{Y \in 2^\Theta, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m(Y)}{1 - m(\emptyset)}. \quad (12)$$

Exemple de fonction de croyance dans le cas d'un k-plus proches voisins crédibiliste

Denœux [2] propose une estimation des fonctions de masses à partir d'un modèle de distance :

$$\begin{aligned} m_k(C_i|x^{(t,k)})(x) &= \alpha_i \exp(\gamma_i d^2(x, x^{(t,k)})) \\ m_k(\Theta|x^{(t,k)})(x) &= 1 - \alpha_i \exp(\gamma_i d^2(x, x^{(t,k)})) \end{aligned} \quad (13)$$

où C_i est la classe associée à $x^{(t,k)}$, $x^{(t,k)}$ sont les k vecteurs d'apprentissage les plus proches de la valeur x et la distance employée est la distance euclidienne. α_i et γ_i sont des coefficients d'affaiblissement, et de normalisation qui peuvent être optimisés [23]. Les k fonctions de masse ainsi calculées pour chaque x sont combinées par la règle orthogonale normalisée de Dempster-Shafer donnée pour tout $A \in 2^\Theta$, $A \neq \emptyset$ par :

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}, \quad (14)$$

et $m(\emptyset) = 0$. La décision est ensuite prise par le maximum sur les fonctions de masse qui dans ce cas est équivalent au maximum de probabilité pignistique car les seuls éléments focaux sont les singletons et l'ignorance.

4 Régression floue et crédibiliste à partir des vecteurs de supports

Les machines à vecteurs de support offrent la possibilité de procéder à une régression linéaire pour non plus prédire une classe, mais une

fonction quelconque [21]. Dans le cas de fonctions à valeurs dans $\mathbb{R}^N$ différentes solutions ont été apportées, par exemple dans [17, 7] une régression simple est effectuée sur chacune des dimensions. Dans le cas des fonctions d'appartenance ou des fonctions de croyance la condition de normalisation impose de considérer chacune des dimensions de la fonction à prédire conjointement et de manière indépendante [13, 15]. Les contraintes identiques des fonctions d'appartenance et des fonctions de croyances (à valeurs dans $[0,1]$ et dont la somme est 1), nous permettent de réécrire la régression linéaire multiple tout en généralisant les travaux de [13, 15], comme nous le verrons plus loin.

Hong [9] propose une régression sur des nombres flous triangulaires, il se place ainsi dans le cas d'une régression multiple en dimension 3, uniquement, mais la différence fondamentale avec notre approche vient des contraintes sur les fonctions d'appartenance et les fonctions de croyance.

Considérons les vecteurs d'apprentissage $x_t \in \mathbb{R}^d$ et les fonctions associées $y_t \in \mathbb{R}^N$, où $N = N_c$ le nombre de classes dans le cas des fonctions d'appartenance et $N = 2^{N_c}$ dans le cas des fonctions de masse. Ainsi par la régression multiple linéaire, nous cherchons une fonctionnelle $f = (f_1, \ldots, f_N)$ où les f_n sont linéaires, de forme $f_n(x) = w_n.x + b_n$. Nous cherchons à déterminer cette fonctionnelle telle que pour les (x_t, y_t) de la base d'apprentissage $|y_{tn} - w_n.x_t + b^n|$ ne dépasse pas un certain ϵ fixé pour tout n. Il est possible de considérer un ϵ_j différent selon la dimension lorsque l'application le justifie. Dans cette formulation nous avons supposé que tous les points sont à l'intérieur du cylindre défini par ϵ. Dans le cas général, nous associons un facteur C pour les points qui sont à l'extérieur du cylindre défini par ϵ. Si les composantes de chaque y_t sont indépendantes, l'approche proposée dans [17] peut être envisagée. Dans notre cas, nous n'avons pas l'indépendance puisque $y_{tn} \in [0,1]$ et vérifient $\sum_{n=1}^{N} y_{tn} = 1$ pour tout t. Ainsi, le problème d'optimisation convexe est similaire à celui exposé dans la section 2, et nous cherchons donc à minimiser :

$$\frac{1}{2}\sum_{n=1}^{N}\|w_n\|^2 + C\sum_{n=1}^{N}\sum_{t=1}^{l}(\xi_{tn}+\xi^*_{tn}), \quad (15)$$

sous les contraintes données pour tout t et tout n :

$$\begin{cases} y_{tn} - w_n.x_t - b_n \leq \epsilon + \xi_{tn}, \\ w_n.x_t + b_n - y_{tn} \leq \epsilon + \xi^*_{tn}, \\ \sum_{n=1}^{N}(w_n.x_t + b_n) = 1, \\ w_n.x_t + b_n \geq 0, \\ w_n.x_t + b_n \leq 1, \\ \xi_{tn}, \xi^*_{tn} \geq 0. \end{cases} \quad (16)$$

L'équation (15) diffère de celle proposée dans [13, 15], car les poids donnés pour borner l'erreur ξ_{tn}, ξ^*_{tn} peut être différents pour chaque direction.

Le lagrangien est donc donné par :

$$\begin{aligned} L = & \frac{1}{2}\sum_{n=1}^{N}\|w_n\|^2 + C\sum_{n=1}^{N}\sum_{t=1}^{l}(\xi_{tn}+\xi^*_{tn}) \\ & - \sum_{n=1}^{N}\sum_{t=1}^{l}(\eta_{tn}\xi_{tn} + \eta^*_{tn}\xi^*_{tn}) \quad (17) \\ & - \sum_{n=1}^{N}\sum_{t=1}^{l}\alpha_{tn}(\epsilon + \xi_{tn} - y_{tn} + w_n.x_t + b_n) \\ & - \sum_{n=1}^{N}\sum_{t=1}^{l}\alpha^*_{tn}(\epsilon + \xi^*_{tn} + y_{tn} - w_n.x_t - b_n) \\ & - \sum_{n=1}^{N}\sum_{t=1}^{l}\beta_{tn}(w_n.x_t + b_n) \\ & - \sum_{n=1}^{N}\sum_{t=1}^{l}\beta^*_{tn}(1 - w_n.x_t - b_n) \\ & - \sum_{t=1}^{l}\gamma_t(1 - \sum_{n=1}^{N}(w_n.x_t + b_n)) \end{aligned}$$

où les η, α, β et γ sont les multiplicateurs de Lagrange et sont positifs.

Au point selle du lagrangien L, on a pour tout t et tout n, $\partial L/\partial b_n = 0$, $\partial L/\partial w_n =$

0, $\partial L/\partial \xi_{tn} = 0$ et $\partial L/\partial \xi^*_{tn} = 0$. Ainsi :

$$\begin{cases} \sum_{t=1}^{l} \sigma_{tn} = 0, \\ w_n = \sum_{t=1}^{l} \sigma_{tn} x_t, \\ \eta_{tn} = C - \alpha_{tn}, \\ \eta^*_{tn} = C - \alpha^*_{tn}, \end{cases} \tag{18}$$

avec $\sigma_{tn} = \alpha_{tn} - \alpha^*_{tn} + \beta_{tn} - \beta^*_{tn} - \gamma_t$.

En intégrant ces équations (18) dans le lagrangien (équation (17)), le problème revient à maximiser :

$$-\frac{1}{2} \sum_{n=1}^{N} \sum_{t,t'=1}^{l} \sigma_{tn}\sigma_{t'n} x_t.x_{t'} - \sum_{n=1}^{N} \sum_{t=1}^{l} \beta^*_{tn} + \frac{\gamma_t}{N} - \sum_{n=1}^{N} \sum_{t=1}^{l} \alpha^*_{tn}(\epsilon + y_{tn}) - \sum_{n=1}^{N} \sum_{t=1}^{l} \alpha_{tn}(\epsilon - y_{tn})$$

sous les contraintes :

$$\begin{cases} \sum_{t=1}^{l} \sigma_{tn} = 0, \\ \alpha_{tn} \in [0, C], \\ \alpha^*_{tn} \in [0, C], \\ \beta_{tn} \geq 0, \\ \beta^*_{tn} \geq 0, \\ \gamma_t \geq 0. \end{cases}$$

Enfin, pour prédire la $n^{\text{ème}}$ sortie, $\tilde{y}_n$, d'un nouvel élément x, on calcule :

$$\tilde{y}_n = \sum_{t=1}^{l} \sigma_{tn} x_t.x + b_n,$$

où b_n est déduite des conditions de Kuhn, Karush et Tucker :

$$\begin{cases} \alpha_{tn}(\epsilon + \xi_{tn} - y_{tn} + w_n.x_t + b_n) = 0, \\ \alpha^*{}_{tn}(\epsilon + \xi^*_{tn} + y_{tn} - w_n.x_t - b_n) = 0, \\ (C - \alpha_{tn})\xi_{tn} = 0, \\ (C - \alpha^*_{tn})\xi^*_{tn} = 0, \\ \beta_{tn}(w_n.x_t + b_n) = 0, \\ \beta^*_{tn}(1 - w_n.x_t - b_n) = 0. \end{cases}$$

Si pour un t_0, $\alpha_{t_0 n} \in]0, C[$ alors, $\xi_{t_0 n} = 0$, ce qui implique $b_n = y_{t_0 n} - w_n.x_{t_0} - \epsilon$, le même raisonnement sur α^* donne $b_n = y_{t_0 n} - w_n.x_{t_0} + \epsilon$.

Jusqu'à présent, nous avons supposé la relation entre les x_t et les sorties linéaires. Dans le cas contraire, d'une manière similaire aux SVM pour la classification, nous pouvons représenter les données de départ en utilisant un noyau. Ainsi le produit scalaire entre les données de la base d'apprentissage peut être remplacé par un noyau : $x.x'$ devient $K(x, x')$. Il suffit ensuite d'appliquer une régression linéaire dans l'espace de représentation. Pour prédire la sortie, $\tilde{y}$, d'un élément x, nous considérons donc :

$$\tilde{y}_n = \sum_{t=1}^{l} \sigma_{tn} K(x, x_t) + b_n.$$

A partir de cette approche de régression sur les fonctions d'appartenance ou les fonctions de croyance, nous obtenons un classifieur en prenant la décision via le maximum des fonctions d'appartenance ou le maximum de la probabilité pignistique (12).

5 Expérimentations

Les expérimentations de notre approche de classification floue et crédibiliste ont été conduites sur deux jeux de données gaussiennes générés. Nous avons défini sur ces données une fonction d'appartenance à partir de l'équation (10) et une fonction de masse à partir de la fonction de masse issue de la combinaison orthogonale normalisée des fonctions de masse données par l'équation (13). Les deux classifieurs ainsi obtenus sont comparés aux k-plus proches voisins flous (avec $k = 5$ et $k_f = 7$) et aux k-plus proches voisins crédibilistes (avec k=5) sur les même données générées et avec les mêmes fonctions d'appartenance et de croyance. Les valeurs de k et k_f sont choisis arbitrairement avec le même k pour les deux approches des k-plus proches voisins et un k_f supérieur à k.

Le premier jeu de données est généré à partir de deux distributions gaussiennes dans $\mathbb{R}^2$, respectivement de moyenne $\mu_1 = (1 \quad 0)^T$ et

de covariance $\Sigma_1 = 0.25\mathbf{Id}$ et de moyenne $\mu_2 = (-1 \quad 0)^T$ et une matrice de covariance $\Sigma_2 = \mathbf{Id}$ pour la seconde classe, où $\mathbf{Id}$ est la matrice identité. Chaque classe contient 140 éléments (40 pour l'apprentissage et 100 pour le test).

Voici les matrices de confusion obtenues pour les k-plus proches voisins flous, les k-plus proches voisins crédibilistes, par la régression linéaire floue avec $C = 10^6$ et $\epsilon = 10^{-7}$ et crédibiliste avec $C = 10^5$ et $\epsilon = 0.1$ (les valeurs de C et de ϵ sont choisis en utilisant une recherche linéaire sur les deux intervalles $\log_{10}(C) \in [1, 10]$ et $\log_{10}(\epsilon) \in [-10, -1]$:

k-ppv flou régression floue

$$\begin{array}{|cc|} \hline 97 & 3 \\ 21 & 79 \\ \hline \end{array} \qquad \begin{array}{|cc|} \hline 100 & 0 \\ 18 & 82 \\ \hline \end{array}$$

k-ppv créd. régression créd.

$$\begin{array}{|cc|} \hline 87 & 13 \\ 9 & 91 \\ \hline \end{array} \qquad \begin{array}{|cc|} \hline 100 & 0 \\ 18 & 82 \\ \hline \end{array}$$

Nous avons ainsi obtenu un taux de 91% pour la classification par régression crédibiliste et floue, qui est supérieur à ceux obtenus par les k-plus proches voisins flous (89%) et les k-plus proches voisins crédibilistes (88%). Notons que l'approche par la régression est plus coûteuse pour la phase d'apprentissage, mais beaucoup moins pour celle de la classification.

Le second jeu de données est généré à partir de trois distributions gaussiennes de dimension 3 utilisant, respectivement, une moyenne de $(1\ 1\ 1)^T$, $(-1\ 1\ 0)^T$ et $(0\ -1\ 1)^T$ et des matrices de covariances $0.25\mathbf{Id}$, $0.75\mathbf{Id}$ et $0.5\mathbf{Id}$. Chaque classe contient 140 vecteurs (40 pour l'apprentissage et 100 pour le test).

Voici les résultats obtenus pour les k-plus proches voisins flous, les k-plus proches voisins crédibilistes, par la régression linéaire crédibiliste avec $C = 10^2$ et $\epsilon = 10^{-1}$ et floue avec $C = 10$ et $\epsilon = 10^{-8}$:

k-ppv flou régression floue

$$\begin{array}{|ccc|} \hline 95 & 1 & 4 \\ 19 & 75 & 6 \\ 3 & 0 & 97 \\ \hline \end{array} \qquad \begin{array}{|ccc|} \hline 82 & 11 & 7 \\ 7 & 84 & 9 \\ 1 & 0 & 99 \\ \hline \end{array}$$

k-ppv créd. régression créd.

$$\begin{array}{|ccc|} \hline 87 & 6 & 7 \\ 19 & 76 & 5 \\ 5 & 3 & 92 \\ \hline \end{array} \qquad \begin{array}{|ccc|} \hline 92 & 6 & 2 \\ 9 & 85 & 6 \\ 8 & 9 & 83 \\ \hline \end{array}$$

Le taux de classification par les k-plus proches voisins flou est de 89% et celui par la régression floue est de 88.33%. Nous avons obtenu un taux de 86.67% pour la classification par régression crédibiliste qui est supérieur à celui obtenu par les k-plus proches voisins crédibilistes (85%) et qui est inférieur aux taux obtenus par les deux approches floues.

Comme noté auparavant, l'apprentissage par la régression est coûteux vu la dimension du problème d'optimisation à résoudre. Une solution est d'utiliser des méthodes de décomposition où, à chaque itération, nous résolvons un sous-problème de dimension petite [12]. Dans le cas extrême où la dimension du sous-problème est égale à 2, nous utilisons la décomposition par *Sequential Minimal Optimization* [14].

6 Conclusions

Nous avons proposé dans ce papier une nouvelle approche de régression floue et crédibiliste à partir de machines à vecteurs de support pour la classification. Les résultats comparés à une méthode de k-plus proches voisins flous, et une méthode de k-plus proches voisins crédibilistes ont montré l'intérêt de cette approche. Nous n'avons donné ici que des résultats dans le cas

de la régression linéaire. Selon les données il est préférable d'employer un noyau par exemple polynomial ou gaussien. L'approche proposée reste alors la même. L'inconvénient majeur de l'approche est le choix des constantes C et ϵ, ainsi que les réglages éventuels selon le noyau. Afin de résoudre ce problème, il est possible d'envisager l'optimisation de ces constantes soit en les intégrant dans l'approche soit par une autre approche telle que les algorithmes génétiques.

Références

[1] J.J. Buckley, Y. Hayashi, Fuzzy neural networks : a survey, *Fuzzy Sets and Systems*, 66(1) : 1-13, 1994.

[2] T. Denœux, A k-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory,*IEEE Transactions on Systems, Man, and Cybernetics - Part A : Systems and Humans*, 25(5) : 804-813, 1995.

[3] T. Denœux, A Neural Network Classifier Based on Dempster-Shafer Theory, *IEEE Transactions on Systems, Man, and Cybernetics - Part A : Systems and Humans*, 30(2) : 131-150, 2000.

[4] T. Denœux, M. Skarstein Bjanger, Induction of decision trees from partially classified data using belief function, *Proceedings of SMC'2000, Nashville, USA*, pp 2923-2928, 2000.

[5] A.P. Dempster, Upper and Lower probabilities induced by a multivalued mapping, *Annals of Mathematical Statistics*, 83 : 325-339, 1967.

[6] D. Dubois, H. Prade, Théorie des possibilités, *Masson*, 1987.

[7] S. Gunn, Support Vector Machines for Classification and Regression, *ISIS Tech. Report, University of Southampton*, 1998.

[8] H. Han-Pang, L. Yi-Hung, Fuzzy support vector machines for pattern recognition and data mining, *International Journal of Fuzzy Systems*, 14(3) :25-28, 2002.

[9] D.H. Hong, C. Hwang, Support vector fuzzy regression machines, *Fuzzy Sets and Systems*, 138 :271-281, 2003.

[10] A. Jozwik, A learning scheme for a fuzzy k-NN rule, *Pattern Recognition Letters*, 1 : 287-289, 1983.

[11] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy k-NN neighbor algorithm, *IEEE Transactions on Systems, Man, and Cybernetics*, 15 : 580-585, 1985.

[12] E. Osuna and R. Freund and F. Girosi, Improved training algorithm for support vector machines, *NNSP'97*, 1997.

[13] F. Pérez-Cruz, G. Camps, E. Soria, J Pérez, A.R. Figueiras-Vidal, A. Artés-Rodriguez, Multidimensional function approximation and regression estimation, *International Conference on Artificial Neural Networks*, Madrid, Espagne, Août 2002.

[14] J. Platt, Sequential Minimal Optimization : A fast algorithm for training Support Vector Machines, *Microsoft Research Technical Report MSR-TR-98-14*, 1998.

[15] M. P. Sanchez-Fernandez, M. de-Prado-Cumplido, J. Arenas-Garcia and F. Pérez-Cruz, SVM Multiregression for Non-Linear Channel Estimation in Multiple-Input Multiple-Output Systems, *IEEE Transactions on Signal Processing*, 58(8) : 2298 - 2307, Août 2004.

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

[17] A.J. Smola, B. Schoelkopf, A tutorial on support vector regression, *NeuroCOLT2 Technical Report NC2-TR-1998-030*, 1998.

[18] Ph. Smets, Constructing the pignistic probability function in a context of uncertainty, *Uncertainty in Artificial Intelligence*, 5 : 29-39, 1990.

[19] D. Tsujinishi, S. Abe, Fuzzy least squares support vector machines for multiclass problems, *Neural Networks*, 16 :785-792, 2003.

[20] P. Vannoorenberghe, T. Denœux, Handling uncer-