Une méthode de discrimination non paramétrique basée sur la théorie de Dempster et Shafer*

Lalla Merieme Zouhal[1,2] et Thierry Denœux[1]

[1] Université de Technologie de Compiègne - U.R.A CNRS 817 Heudiasyc
BP 649 - F-60206 Compiègne cedex - France
Tel. (33) 44 23 44 23. Fax. (33) 44 23 44 77
email: lzouhal@hds.univ-compiegne.fr

[2] Lyonnaise des Eaux
Laboratoire d'Informatique Avancée de Compiègne (LIAC)

Résumé

La méthode de discrimination introduite dans [3] consiste à considérer chaque voisin d'un vecteur à classer comme une source d'information renforçant certaines hypothèses concernant la classe de ce vecteur. Cet source d'information est représentée par une distribution de masse de croyance (DMC), et les différentes DMCs sont combinées par la règle de Dempster. Dans cet article, nous proposons de déterminer par apprentissage les paramètres de la méthode en optimisant un critère de performance. Plusieurs jeux de données artificielles et réelles sont utilisés pour effectuer une comparaison avec les règles des k-plus proches voisins classique et pondérée.

Abstract

Recently, a new classifier using neighborhood information in the context of the Dempster-Shafer theory of evidence has been introduced [3]. This approach consists in considering each neighbor of a pattern to be classified as an item of evidence supporting certain hypotheses concerning the class membership of that pattern. The evidence of the k nearest neighbors is pooled by means of Demspter's rule of combination. In this paper, we propose to determine the optimal values of the parameters used in the method by gradient descent of an error function. Several sets of artificial and real-world data are used for comparison with the voting and distance-weighted k-NN classifiers.

*In *Actes du Quinzième Colloque GRETSI*, pages 689–692, Juan les Pins, Septembre 1995.

1 Introduction

La théorie des fonctions de croyance, encore appelée théorie de Dempster et Shafer (D-S) [8] est largement utilisée en Intelligence Artificielle pour le traitement de l'incertitude dans les bases de connaissances [7]. Or, la gestion de l'incertitude et la combinaison d'informations issues de différentes sources est également l'un des problèmes fondamentaux en Reconnaissance de Formes (RdF), notamment dans des applications telles que le diagnostic [4]. Une nouvelle méthode des k plus proches voisins (k-ppv) basée sur la théorie de Dempster et Shafer a récemment été proposée [2, 3]. La classification d'un nouveau vecteur est basée sur la combinaison des informations apportées par chacun de ses plus proches voisins dans l'ensemble d'apprentissage. Chaque voisin est considéré comme renforçant certaines hypothèses concernant la classe d'appartenance du vecteur forme à classer. A cette source d'information est associée une distribution de masse de croyance (*basic belief assignment*) définie en fonction de la distance entre les deux vecteurs. Les différentes distributions de masse sont ensuite combinées par la règle de Dempster [8]. Le vecteur forme est finalement affecté à la classe de plus grande crédibilité (ou, de manière équivalente, de plus grande plausibilité). Dans cet article, nous proposons une méthode adaptative pour déterminer les paramètres de la distribution de masse associée à chaque voisin. Cette méthode est basée sur la minimisation d'un critère d'erreur. Des simulations ont été réalisées sur plusieurs jeux de données artificielles et réelles. Les performances des règles des k-ppv basées sur la théorie de D-S (avec et sans apprentissage des paramètres) ont été comparées à celles obtenues avec les règles des k-ppv classique et pondérée [1].

2 Théorie de D-S en RdF

Soient $\mathcal{X}$ un ensemble d'apprentissage composé de vecteurs répartis en M classes $\Omega = \{\omega_1, \dots, \omega_M\}$, et $\mathbf{x}^s$ un nouveau vecteur forme à classer. Soit ϕ_k^s l'ensemble des k ppv de $\mathbf{x}^s$ dans l'ensemble d'apprentissage. Chaque voisin $\mathbf{x}^i$ appartenant à ϕ_k^s est supposé avoir une étiquette $L^i = q \in \{1, \dots, M\}$ définissant son appartenance à la classe ω_q. La connaissance de la classe de $\mathbf{x}^i$ apporte une certaine information sur la classe d'appartenance du vecteur $\mathbf{x}^s$. Cette information peut être représentée par une masse de croyance [8] affectée à la classe ω_q. La classification de $\mathbf{x}^s$ peut être basée sur la combinaison des informations apportées par chacun de ses plus proche voisin. A chaque voisin $\mathbf{x}^i$ est associée une distribution de masse de croyance $m^i : 2^\Omega \longmapsto [0,1]$ définie par :

$$\begin{aligned} m^i(\{\omega_q\}) &= \alpha_q^i && (1)\\ m^i(\Omega) &= 1 - \alpha_q^i && (2) \end{aligned}$$

et $m^i(A) = 0$ pout tout $A \in 2^\Omega \setminus \{\Omega, \{\omega_q\}\}$. Le terme α_q^i est une fonction décroissante de la distance d^i entre les vecteurs $\mathbf{x}^s$ et $\mathbf{x}^i$:

$$\alpha_q^i = \alpha_0 \exp(-\gamma_q^2 \ (d^i)^2) \qquad (3)$$

où γ_q est un paramètre caractérisant la classe ω_q et α_0 un paramètre fixé. La distribution de masse m^i possède ω_q et Ω comme éléments focaux (sous-ensembles de Ω de masse non nulle). La règle de combinaison de Dempster [8] permet de construire une nouvelle distribution de masse m en combinant les distributions $m^1, m^2, \dots, m^k$ associées aux k voisins :

$$m = m^1 \oplus \ldots \oplus m^k \qquad (4)$$

Les éléments focaux de M sont les classes représentées dans ϕ_k^s, et Ω. La crédibilité et la plausibilité de ω_q peuvent être calculées en fonction de m :

$$\begin{aligned} Bel(\{\omega_q\}) &= m(\{\omega_q\}) && (5)\\ Pl(\{\omega_q\}) &= m(\{\omega_q\}) + m(\Omega) && (6) \end{aligned}$$

Toute distribution de probabilité P telle que $Bel(A) \leq P(A) \leq Pl(A)$ pour toute partie A de Ω est dite compatible avec m [8]. Un cas particulier est la distribution de probabilité "pignistique" [9] $BetP$ obtenue en répartissant uniformément la masse $m(\Omega)$ entre toutes les classes :

$$BetP(\{\omega_q\}) = m(\{\omega_q\}) + \frac{m(\Omega)}{M}. \qquad (7)$$

Le vecteur forme $\mathbf{x}^s$ est finalement affecté à la classe ω_{qmax} de plus grande crédibilité (qui est également celle de plus grandes plausibilité et probabilité pignistique) :

$$qmax = \arg\max_q m(\{\omega_q\}) \qquad (8)$$

3 Méthode proposée

En ce qui concerne la détermination des valeurs de γ_q pour $q = 1, \dots, M$, l'heuristique suivante a été suggérée [3] :

$$\gamma_q = 1/d_q \qquad (9)$$

où d_q est la distance moyenne entre les éléments de la classe ω_q. Bien que ces valeurs donnent de bon résultat en moyenne, la performance de classification peut varier considérablement en fonction des valeurs de γ_q ($q = 1, \dots, M$). Dans cet article, nous proposons une nouvelle méthode pour déterminer automatiquement les valeurs de ces paramètres en utilisant l'information contenue dans l'ensemble d'apprentissage.

A chaque vecteur $\mathbf{x}$ appartenant à la classe ω_q est associé un vecteur étiquette $\mathbf{t} = (t_1, \dots, t_M)^t$ avec $t_i = \delta_{i,q}$ si $\mathbf{x} \in \omega_q$. Le résultat de classification par la méthode de k-ppv basée sur la théorie de D-S consiste en une distribution de masse de croyance m à partir de laquelle peut être défini un vecteur de probabilités pignistiques $\mathbf{P} = (P_1, \dots, P_M)^t$ défini par :

$$P_q = m(\{\omega_q\}) + \frac{m(\Omega)}{M} \qquad (10)$$

Le but recherché est de déterminer les valeurs de γ_q ($q = 1, \dots, M$), pour lesquelles le vecteur $\mathbf{P}$ sera aussi proche que possible du vecteur étiquette $\mathbf{t}$. Ceci est réalisé en minimisant la fonction erreur $E(\mathbf{x})$ égale à :

$$E(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{M}(P_i - t_i)^2 \qquad (11)$$

L'erreur moyenne sur l'ensemble d'apprentissage $\mathcal{X}$ de taille N est définie par :

$$E = \frac{1}{N}\sum_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \qquad (12)$$

La dérivée de $E(\mathbf{x})$ par rapport à chaque paramètre γ_q a l'expression suivante [10] :

$$\begin{aligned} \frac{\partial E(\mathbf{x})}{\partial \gamma_q} &= -2\ \gamma_q \sum_{\mathbf{x}^j \in \omega_q} \alpha_q^j \ (d^j)^2 \ \frac{\partial E(\mathbf{x})}{\partial \alpha_q^j} && (13)\\ \frac{\partial E(\mathbf{x})}{\partial \alpha_q^j} &= \frac{1}{K^2}\sum_{i=1}^{M}(P_i - t_i)[K((\bar{m}^j(\{\omega_i\}) + \\ &\quad \bar{m}^j(\Omega))\delta_{iq} - \bar{m}^j(\{\omega_i\})) - (m'(\{\omega_i\}) + \\ &\quad \frac{m'(\Omega)}{M})\frac{\partial K}{\partial \alpha_q^j} - \frac{K}{M}\bar{m}^j(\Omega)] && (14)\\ \frac{\partial K}{\partial \alpha_q^j} &= \sum_{i=1}^{M}((\bar{m}^j(\{\omega_i\}) + \bar{m}^j(\Omega))\delta_{iq} - \\ &\quad \bar{m}^j(\{\omega_i\})) - \bar{m}^j(\Omega) && (15) \end{aligned}$$

avec

- m' la somme orthogonale non normalisée de $m^1, \dots, m^k$;

– $\bar{m}^j$ la somme orthogonale non normalisée de m^ℓ, $\ell \in \{1, \ldots, k\} \setminus \{j\}$;

– $K = \sum_{q=1}^{M} m'(\{\omega_q\}) + m'(\Omega)$.

Le calcul du gradient de E permet de déterminer par une méthode itérative les valeurs de γ_q ($q = 1, \ldots, M$) minimisant E. Ces valeurs peuvent ensuite être utilisées pour la classification de nouveaux vecteurs.

4 Simulations

Des simulations ont été réalisées sur plusieurs jeux de données artificielles et réelles (Tableau 1). Deux ensembles de données B_1 et B_2 ont été générés suivant la méthode proposée dans [5]. Ces données sont issues de 3 classes Gaussiennes en dimension 10. Les matrices de covariance D_1, D_2 et D_3 des trois distributions sont diagonales. Les probabilités d'erreurs de classification ont été estimées sur l'ensemble test par le taux de vecteurs mal classés, et représentent des moyennes obtenues sur 6 ensembles d'apprentissage indépendants.

Les jeux de données réels utilisés dans cette étude proviennent de la base de données d'apprentissage développée par Murphy et Aha [6]. Les données "ionosphère" sont issues de mesures de réflectivité radar, les cibles étant constituées par des électrons libres dans la ionosphère [6]. Les observations ont été réparties en deux classes en fonction de la structure de la ionosphère. Le jeu de données "véhicules" a été construit à partir de caractéristiques extraites de silhouettes de quatre types de véhicules (bus, Chevrolet van, Saab 9000 et Opel Manta 400).

Les performances des règles de k-ppv basées sur la théorie de D-S (avec et sans apprentissage des paramètres) ont été comparées à celles obtenues par les règles des k-ppv classique et pondérée [1] (Figure 1). Les résultats obtenus montrent une nette supériorité de la méthode proposée, dont les performances en termes de taux de bonne classification sont supérieures à celles des autres méthodes pour toutes les valeurs de k (sauf sur les données "véhicules" pour lesquelles la méthode des k-ppv pondérés donne des résultats équivalents pour $k = 5$). La méthode s'avère d'autre part extrêmement robuste au choix du paramètre k.

5 Conclusion

Une méthode adaptative de discrimination non paramétrique basée sur la théorie de Dempster et Shafer et une approche de type "k plus proches voisins" a été présentée. La méthode consiste à déterminer, par optimisation d'un critère de performance, les paramètres de la distribution de masse de croyance associée à chaque voisin. Des simulations effectuées sur quatre ensembles d'apprentissage ont démontré l'intérêt de la méthode par rapport aux règles des k-ppv avec et sans pondération. Ce résultat semble dû à une meilleure utilisation de l'information contenue dans l'ensemble d'apprentissage, qui est utilisée à la fois *localement* pour la recherche des voisins, et *globalement* pour optimiser les paramètres. La méthode apparaît également peu sensible au choix du nombre de voisins.

Références

[1] S. A. Dudani, "The distance-weighted k-nearest neighbor rule", *IEEE Trans. Syst. Man. Cybern.*, vol. SCM-6(4), pp. 325-327, 1976.

[2] T. Denoeux, "Application of evidence theory to k-NN pattern classification. In E. S. Gelsema and L. N. Kanal (Eds)", *Pattern Recognition in Practice IV*, Elsevier, Amsterdam, pp. 13-24, 1994.

[3] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory", *IEEE Trans. Syst. Man. Cybern.*, vol. SCM-25(5):804-813, 1995.

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

[5] J. H. Friedman, "Regularized discriminant analysis", *J. Am. Statist. Ass.*, vol. 84, pp. 165-175, 1989.

[6] P. M. Murphy and D. W. Aha, *UCI Reposition of machine learning databases [Machine-readable data repository]*, Irvine, CA, 1994.

[7] H. Prade, "A computational approach to approximate and plausible reasoning with applications to expert systems", *IEEE Trans. Pattern Anal. Machine. Intell.*, PAMI-7(3):260-283, 1985.

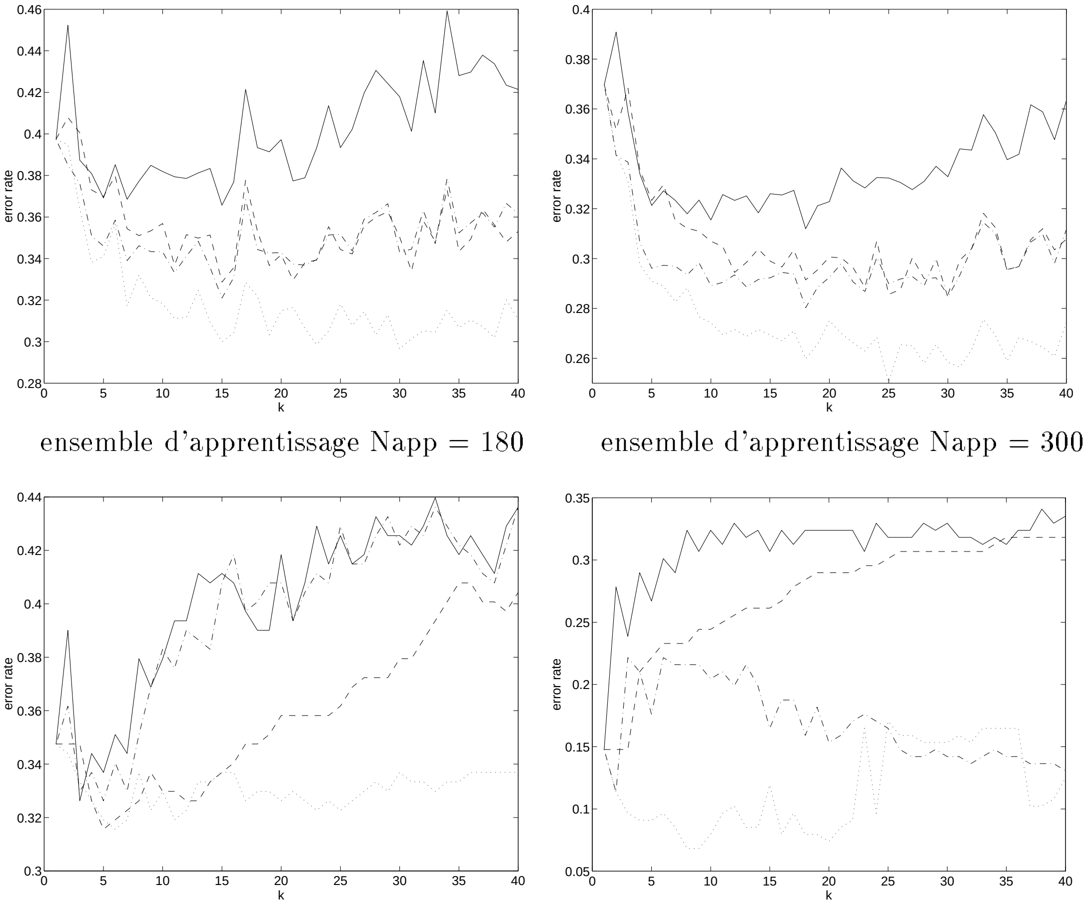
[8] G. Shafer, *A mathematical theory of evidence*, Princeton N.J, Princeton University Press, 1976.

[9] P. Smets and R. Kennes, "The combination of evidence in the transferable belief model", *Artificial Intelligence*, vol. 66, 1994.

[10] L. M. Zouhal and T. Denœux. An adaptive k-NN rule based on Dempster-Shafer theory. In *Proc. of CAIP'95*, Prague, Sept. 1995.

Tab. 1 - *Description des données.*

données	M	apprentissage	test	dimension
B1	3	120	1000	10
B2	3	300	1000	10
Ionosphère	2	175	176	34
véhicules	4	564	282	18



ensemble d'apprentissage Napp = 180

ensemble d'apprentissage Napp = 300

données véhicules

données ionosphère

Fig. 1 - *Taux d'erreur de classification des règles : k-ppv classique (-), k-ppv basée sur la théorie de D-S (sans apprentissage (-.), avec apprentissage (:)) et k-ppv pondérés (- -), pour différentes valeurs de k.*