LE DISCRIMINATEUR A PLUS PROCHE STRUCTURE PROPRE (PPSP)

Jean-Pierre Asselin de Beauville
Eric Darmigny
Nicole Vincent

Laboratoire d'Informatique
Equipe « Reconnaissance des Formes et Analyse d'Images »
Ecole d'Ingénieurs en Informatique pour l'Industrie (E3i)
Université de Tours
64 Avenue Jean Portalis
37200 TOURS (France)
Email :{vincent,asselin}@e3i.univ-tours.fr

1. INTRODUCTION

Le problème de la discrimination est celui de l'apprentissage à partir d'exemples d'une fonction de classement. Il s'agit d'une problématique très vaste, aux applications nombreuses, ayant été déja largement explorée en statistique et en reconnaissance des formes ([Ulmo1973], [Caraux et al.1994] ...).

Une classe importante de méthodes de discrimination repose sur le calcul d'une distance (ou d'une similarité), dans l'espace de représentation, entre l'observation à classer $\underline{x}$ et un prototype représentatif de chaque classe d'apprentissage ([Dubuisson1990]...). Par exemple, un discriminateur à distance minimum pourra se baser sur une distance entre $\underline{x}$ et le centre de gravité (prototype) de chaque classe. Le prototype d'une classe peut être défini de différentes façons. Il peut, par exemple, être constitué des observations de cette classe appartenant à un voisinage ou à un ensemble structuré de voisinages de $\underline{x}$. Dans ce cas, est associé à chaque classe, non pas un unique prototype mais bien un ensemble de prototypes, chacun d'entre eux étant lié à une observation de l'espace de représentation ([Amghar et al.1994], [Sebban1997]).

Dans cet article on présente une nouvelle méthode de discrimination pour laquelle le prototype d'une classe est défini à partir de sa structure propre (valeurs et vecteurs propres). Il suffira ensuite de savoir calculer la similarité entre $\underline{x}$ et chacune des structures propres des classes d'apprentissage pour être en mesure d'effectuer le classement de $\underline{x}$. Le principe de cette approche avait été introduit dans [Asselin de Beauville1995].

Dans le paragraphe 2 on décrira l'algorithme proposé. Le paragraphe 3 sera relatif aux performances de cette méthode pour résoudre le problème-test des formes d'ondes de Breiman ([Breiman et al.1984]). Les résultats obtenus seront comparés à ceux récoltés, dans les mêmes conditions,

par le projet inter-PRC « Méthodes symboliques-numériques de discrimination » (1993-1994). Enfin le paragraphe 4 concluera et ouvrira quelques perspectives à ce travail.

2. L'ALGORITHME PPSP

Comme la plupart des discriminateurs PPSP fonctionne en deux phases successives : Apprentissage et Décision.

2.1 Phase d'apprentissage

Soit $E=\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ un échantillon d'apprentissage de n éléments On suppose que les $\underline{x}_i$ appartiennent à R^p (p>=1). On notera le transposé du vecteur $\underline{x}_i$ par : $\underline{x}'_i = [\underline{x}_{i1}, \ldots, \underline{x}_{ip}]$. Chaque $\underline{x}_i$ est supposé appartenir à une des c classes : $w_1, w_2, \ldots, w_c$ avec c>=2. Soit $n^{(k)}$ l'effectif de la classe w_k. On a donc $n^{(k)}$= Card(w_k). On suppose dans la suite que, pour tout k, $n^{(k)}$ est supérieur ou égal à 2. Le centre de gravité $\underline{g}^{(k)}$ de w_k est défini par : $\underline{g}^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_p^{(k)}]'$ avec $x_j^{(k)} = (\sum_{i \in w_k} x_{ij}) / n^{(k)}, j = 1, \ldots, p$.

Le centre de gravité constitue en lui-même la première structure propre de la classe, celle que l'on peut appeler d'ordre 0. Elle permet de représenter la classe par un point. Evidemment cette structure très pauvre est insuffisante pour représenter des classes dont la forme est éloignée d'une symétrie sphérique ou pour discriminer des classes dont les frontières de séparation ne sont pas linéaires.

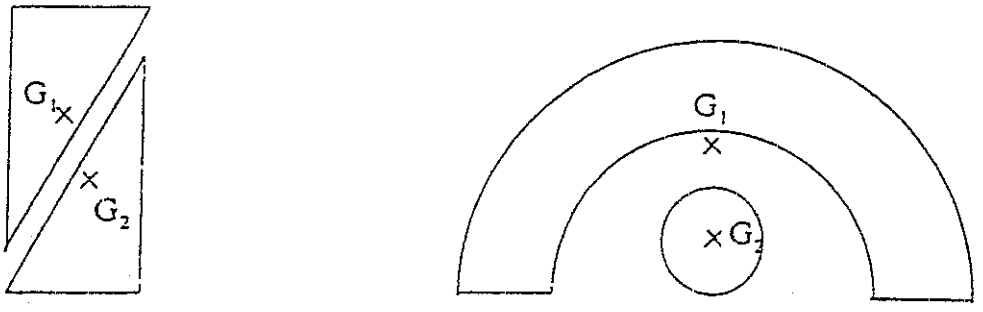

Figure 1 : Exemples de formes de classes

Nous sommes alors amenés à choisir, pour la classe, une structure propre qui soit plus précise. Pour cela nous considérons la matrice $\underline{X}^{(k)}$ des observations centrées pour chaque classe w_k. Le terme général de cette matrice est donc : $\left(x_{ij} - x_j^{(k)}\right)$, i=1,2,....,n ; j=1,2,....,p ; k=1,2,....,c.

La structure propre de la classe peut être plus ou moins complexe, elle doit s'adapter non seulement à la complexité de l'ensemble des représentants de la classe en présence de laquelle on se trouve, mais aussi à la complexité de l'agencement des différentes classes, comme on a pu le voir sur la figure 1.

La structure propre de la classe w_k est constituée par les vecteurs propres associés aux plus grandes valeurs propres. De manière à réaliser un compromis entre l'étude locale de la classe et l'étude globale de la répartition des classes nous avons choisi dans la suite de l'article de considérer systématiquement des structures propres qui soient des plans. C'est-à-dire que nous ne considérons que les deux plus grandes valeurs propres $\lambda_1^{(k)}$ et $\lambda_2^{(k)}$ de la matrice $\underline{X}^{(k)}$, avec $\lambda_1^{(k)} \geq \lambda_2^{(k)}$, soient $\underline{u}_1^{(k)}$ et $\underline{u}_2^{(k)}$ les deux vecteurs propres associés. Les deux axes principaux définissant l'orientation de la classe ont pour direction respectivement ces deux vecteurs.

Soit $\underline{S}^{(k)}$ la matrice de dispersion de la classe w_k : $\underline{S}^{(k)} = \underline{X}^{(k)'} \underline{X}^{(k)}$, k=1,....,c. Les structures propres des c classes sont définies par : $\underline{S}^{(k)} \underline{u}_j^{(k)} = \lambda_j^{(k)} \underline{u}_j^{(k)}$, j=1,2 et $\left\|\underline{u}_j^{(k)}\right\| = 1$.

Evidemment on peut facilement envisager d'améliorer la représentativité locale d'une classe en augmentant le nombre des directions propres intervenant dans la définition de la structure propre de la classe. Ce nombre α peut être choisi en fonction du rapport : $\dfrac{\sum_{i=1}^{\alpha} \lambda_i^{(k)}}{Tr(\underline{X}^{(k)})}$.

- Une fois la structure propre de chacune des classes calculée, la phase d'apprentissage est terminée

2.2 Phase de décision

Maintenant on désire affecter une observation $\underline{x}$ à l'une des c classes w_k, k=1,....,c. Pour chaque w_k on centre $\underline{x}$ par rapport à $\underline{g}^{(k)}$. On obtient ainsi le vecteur $\underline{x} - \underline{g}^{(k)}$. La règle de décision consiste alors à affecter $\underline{x}$ à la classe w_k si une distance $d^{(k)}(\underline{x})$ entre la classe w_k et $\underline{x} - \underline{g}^{(k)}$ est la plus petite parmi toutes les distances $d^{(l)}(\underline{x})$, l=1,....,c.

Le critère d'affectation est donc :

$$w_k(\underline{x}) = w_m \quad m \in \left[1; c\right] \quad et \quad d^{(m)}(\underline{x}) = \underset{l=1,\ldots,c}{Min}(d^{(l)}(\underline{x}))$$

La distance entre w_k et $\underline{x}$ peut être définie de différentes façons. Ici nous proposons d'utiliser une « distance adaptative » du même type que celle employée pour la classification en classes de frontières linéaires ([Bezdek1981],[Davé1989]) :

$$d^{(k)}(\underline{x}) = \alpha_k d_1^{(k)}(\underline{x}) + (1-\alpha_k) d_2^{(k)}(\underline{x}) \text{ où}$$

$$\alpha_k = 1 - \lambda_2^{(k)} / \lambda_1^{(k)} \text{ avec}$$

$$(d_j^{(k)}(\underline{x}))^2 = ((\underline{x} - \underline{g}^{(k)})'(\underline{x} - \underline{g}^{(k)})) - \left[(\underline{x} - \underline{g}^{(k)})' \underline{u}_j^{(k)}\right]^2, \ j=1,2.$$

Cette distance a la propriété de « s'adapter » à la forme des classes selon qu'elles sont « arrondies » ou « aplaties » à travers l'utilisation du coefficient α_k. Ainsi lorsque la classe est de forme très allongée selon l'axe principal $\underline{u}_1$ par exemple, cela entraîne que $\lambda_1 >> \lambda_2$ et donc que α_k sera proche de 1. Dans ces conditions c'est donc la distance d_1, distance à l'axe $\underline{u}_1$, qui sera prépondérante pour le classement. Lorsque la classe est de forme plus sphérique alors α_k est voisin de 0 et c'est alors la distance de $\underline{x}$ au centre de gravité de la classe qui prend le dessus.

3. TESTS ET RESULTATS

Le discriminateur a été testé principalement sur le problème des « ondes de Breiman ». Ces données ont été introduites par [Breiman et al. 1984] pour l'étude des arbres de décision. Nous les avons retenues ici car elles avaient été choisies pour l'étude comparative entre discriminateurs menée par le projet inter-PRC « Méthodes symboliques-numériques de discrimination » (1993-94). Ce problème servira donc de support pour comparer PPSP aux autres discriminateurs.

Il est constitué par la discrimination entre trois classes de formes d'ondes. Chaque forme d'onde simule un phénomène chronologique quantitatif étudié en 21 instants régulièrement espacés. Chaque individu est donc plongé dans R^{21}. Les trois classes sont engendrées par combinaison deux à deux de trois formes d'onde de base. Ces dernières notées h_1, h_2 et h_3 sont représentées sur la figure 2 ci-dessous.

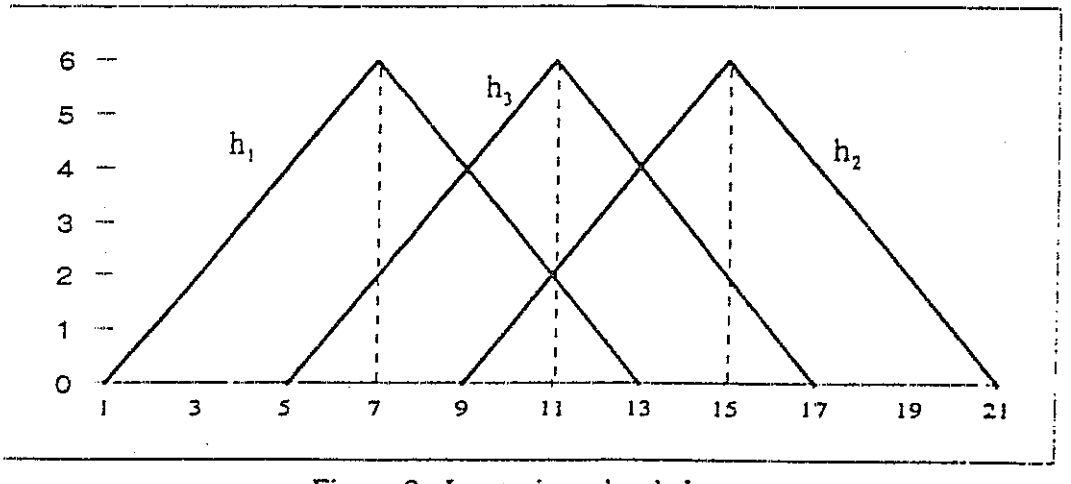

Figure 2 : Les trois ondes de base

Ces associations se font par combinaisons aléatoires convexes avant d'être perturbées par un bruit aléatoire gaussien.

De façon analytique les individus des classes w_1, w_2, w_3 sont engendrés respectivement par les expressions :

$$\underline{x} = r\underline{h}_1 + (1-r)\underline{h}_2 + \varepsilon \quad \text{pour } w_1$$
$$\underline{x} = r\underline{h}_1 + (1-r)\underline{h}_3 + \varepsilon \quad \text{pour } w_2$$
$$\underline{x} = r\underline{h}_2 + (1-r)\underline{h}_3 + \varepsilon \quad \text{pour } w_3 \quad \text{où}$$

r est une variable aléatoire uniforme sur [0,1] et ε un vecteur aléatoire gaussien centré de matrice de variances-covariances identité. Les vecteurs $\underline{h}_i$, i=1,2,3 sont définis par :

$$\underline{h}_1 = [0\ 1\ 2\ 3\ 4\ 5\ 6\ 5\ 4\ 3\ 2\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]'$$
$$\underline{h}_2 = [0\ 0\ 0\ 0\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 5\ 4\ 3\ 2\ 1\ 0\ 0\ 0\ 0\ 0]'$$
$$\underline{h}_3 = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 5\ 4\ 3\ 2\ 1\ 0]'$$

On considère que les trois classes sont équiprobables. Selon [Breiman et al.1984] le taux d'erreur de classement incompressible (erreur de Bayes) est de l'ordre de 14% pour ce problème. D'autres tests, réalisés sur des échantillons gaussiens ne seront pas rapportés ici mais les conclusions que l'on peut en déduire ne remettent pas en cause celles que l'on pourra faire à partir des ondes. La méthodologie appliquée pour ces tests est décrite ci-dessous. Elle a été calquée sur la procédure expérimentale du projet inter-PRC afin d'obtenir des résultats comparables.

On dispose de 10 ensembles d'apprentissage de 300 éléments tirés aléatoirement selon le modèle des ondes de Breiman. Chaque ensemble est constitué de 3 classes équiprobables. Le fait de disposer de plusieurs ensembles d'apprentissage issus de la même population permet de fournir des résultats en moyenne plus précis que des résultats ponctuels et permet aussi de mesurer la variabilité de ces résultats. Indépendamment des ensembles d'apprentissage on génère un ensemble de test de 4000 individus (5000 pour les PRC) à partir de la même population que celle dont sont issus les ensembles d'apprentissage.

Les critères retenus pour évaluer les résultats des simulations sont :

- Le taux d'échec (ou taux d'erreur de classement) évalué sur l'échantillon-test.
- Le taux d'échec évalué sur l'échantillon d'apprentisssage.

En réalité on effectue les calculs sur 10 cycles (un par échantillon d'apprentissage). Pour chaque cycle on effectue un apprentissage (avec l'un des 10 ensembles d'apprentissage) puis un test avec l'ensemble de test. Les valeurs affichées dans les tableaux qui suivent sont donc des moyennes calculées sur 10 cycles. On fournit en outre les écarts-types de ces simulations.

	Taux d'échec moyen	Ecart-type moyen
APPRENTISSAGE	0,11	2,59
TEST	0,17	0,49

Tableau 1 : Résultats sur les ondes de Breiman

Comparé à 0,14, taux d'échec optimum en test pour les ondes de Breiman, on peut apprécier très positivement les performances de l'algorithme PPSP, ce d'autant plus que les temps d'apprentissage sont très courts.

Pour évaluer la robustesse de PPSP nous avons éffectué un certain nombre de tests rassemblés dans les tableaux 2 et 3.

*** Influence de la taille de l'ensemble d'apprentissage (classes équiprobables)**

Nombre d'individus par classe dans l'ensemble d'apprentissage	Taux d'échec moyen en phase de test	Ecart-type moyen en phase de test
10	0,21	2,23
50	0,18	0,83
100	0,17	0,49
250	0,16	0,56
500	0,16	0,58

Tableau 2 : Influence de la taille de l'ensemble d'apprentissage

On constate sur le tableau 2 que PPSP converge rapidement lorsque la taille des classes d'apprentissage croît puisque le taux d'échec de 0,17 est déja atteint entre 50 et 100 individus par classe.

***Influence des probabilités a priori des classes**

On fixe le cardinal de l'ensemble d'apprentissage à 300 et on fait varier les probabilités a priori Ainsi si $P(w_1)=1/2$ alors la classe d'apprentissage correspondante contiendra 150 éléments. Il est clair que l'ensemble de test sera lui aussi réparti selon ces probabilités.

Probabilités a priori des 3 classes			Taux d'échec moyen en test	Ecart-type moyen en test
1/3	1/3	1/3	0,17	0,49
1/2	1/4	1/4	0,16	0,45
2/20	9/20	9/20	0,16	0,58

Tableau 3 : Influence des probabilités a priori.

Comme on peut le constater PPSP s'avère donc une méthode très stable par rapport aux probabilités a priori des classes d'apprentissage.

Pour finir, on présente sur la figure 3 des résultats comparatifs entre PPSP et quelques uns des principaux discriminateurs testés, dans des conditions quasi-identiques, par le projet inter-PRC. On rappelle que dans cette dernière étude, on a utilisé un échantillon test de 5000 éléments alors que notre étude est faite avec 4000 individus ce qui permet d'affirmer que nos résultats sont sous évalués par rapport à ceux des PRC. On pourra néanmoins en tirer des enseignements quant aux performances générales de PPSP. On indique ci-dessous la liste des discriminateurs comparés.

a) Discriminateur linéaire Bayesien (densités conditionnelles aux classes gaussiennes) ([Duda et Hart1973]) (**DLB**) :

La règle de décision est dans ce cas : Affecter $\underline{x}$ à la classe w_k ($k \in \{1,2,3\}$ pour laquelle $d^2(\underline{x},\underline{\mu}_j)$ est-minimum avec :

$$d^2(\underline{x},\underline{\mu}_j) = (\underline{x}-\underline{\mu}_j)' \underline{\Sigma}(\underline{x}-\underline{\mu}_j) \text{ où}$$

$\underline{\mu}_j$ est le centre de gravité estimé de la classe $w_j, j = 1,2,3$.

$\underline{\Sigma}$ est la matrice de variances-covariances estimée pour l'ensemble de la population à partir de l'échantillon d'apprentissage.

On notera que les PRC ont aussi testé le discriminateur Bayesien quadratique ($\underline{\Sigma}_1 \neq \underline{\Sigma}_2 \neq \underline{\Sigma}_3$) mais, les résultats obtenus étant moins bons que dans le cas linéaire, nous n'avons pas jugé utile de retenir ce discriminateur.

b) Discriminateur non paramétrique de Parzen (noyaux gaussiens) (**DNP**) :

Il s'agit encore d'un discriminateur Bayesien pour lequel les densités conditionnelles sont estimées par la méthode du noyau de Parzen ([Parzen1962]). La densité conditionnelle à la classe w_j a pour expression :

$$\hat{f}_n(\underline{x}|w_j) = 1/(h^p Card(w_j)) \sum_{\underline{x}_i \in w_j} K(\underline{x} - \underline{x}_i)/h \quad \text{où p=21 et h=1,}$$

$$K(\underline{z}) = 1/(2\pi)^{p/2} \exp((-1/2)\underline{z}'\underline{z}) \quad , \quad \underline{z} \in R^p .$$

c) Perceptron multi-couches (**PMC**) :

Le perceptron utilisé possède une rétine de 22 neurones (21 variables et une cellule de biais), une couche cachée de 5 neurones et une couche de sortie de 3 neurones (un par classe). Les neurones des couches cachée et de sortie ont des fonctions de transition sigmoïdales. La rétine possède des fonctions de transfert identité. Les connexions entre couches sont totales et l'algorithme d'apprentissage est la rétro-propagation du gradient de l'erreur ([Rumelhart et al.1986]). Le pas du gradient a été fixé à 0,01. En sortie on attribue un exemple à la classe identifiée par la sortie calculée la plus grande.

d) Arbres de décision flous (**ADF**) :

Il s'agit ici d'une variante de l'algorithme ID3 ([Quinlan1986]) destinée à la construction d'arbres de décision en présence d'attributs continus. Ceux-ci sont discrétisés à l'aide de l'entropie de Shannon en vue de pouvoir prendre une décision binaire en chaque noeud de l'arbre. Afin de pallier le problème de l'imprécision des classements pour des observations proches des seuils de discrétisation, les PRC font appel à des « seuils flous ». On trouvera un exposé sur cette approche dans [Marsala1994].

e) Discriminateur des k plus proches voisins (**k-PPV**) :

C'est une méthode très connue et souvent prise comme référence à cause de sa simplicité de mise en oeuvre. La règle d'affectation consiste à attribuer $\underline{x}$ à la classe majoritaire parmi les k plus proches voisins de $\underline{x}$. Ici k=30 (valeur optimale des PRC).

f) Discriminateur des « plus proches voisins adaptatif » ([Mraghni1997]) (**PPVA**) :

Nous avons défini une variante du discriminateur précédent. Cette méthode étant originale nous la décrivons brièvement.

Soit une métrique d(.,.) définie sur R^p. On adoptera ici la métrique euclidienne usuelle (comme pour le classifieur des k-ppv). On commence par calculer le « Moins Proche Voisin », dans l'ensemble d'apprentissage, de la forme à classer $\underline{x}$ (noté MPV($\underline{x}$)). L'ensemble PPV($\underline{x}$) des plus proches voisins de $\underline{x}$ est défini par :

$$\text{PPV}(\underline{x}) = \left\{ \underline{x}_i | d(\underline{x}, \underline{x}_i) < d(MPV(\underline{x}), \underline{x}_i) \right\}$$

L'avantage de cette approche par rapport a celle des k-PPV classique est son adaptativité c'est à dire son indépendance par rapport au choix du seuil k. Par contre elle s'avère (au moins dans sa version non optimisée) très lente à cause du grand nombre de distances à calculer.

*** Résultats comparatifs**

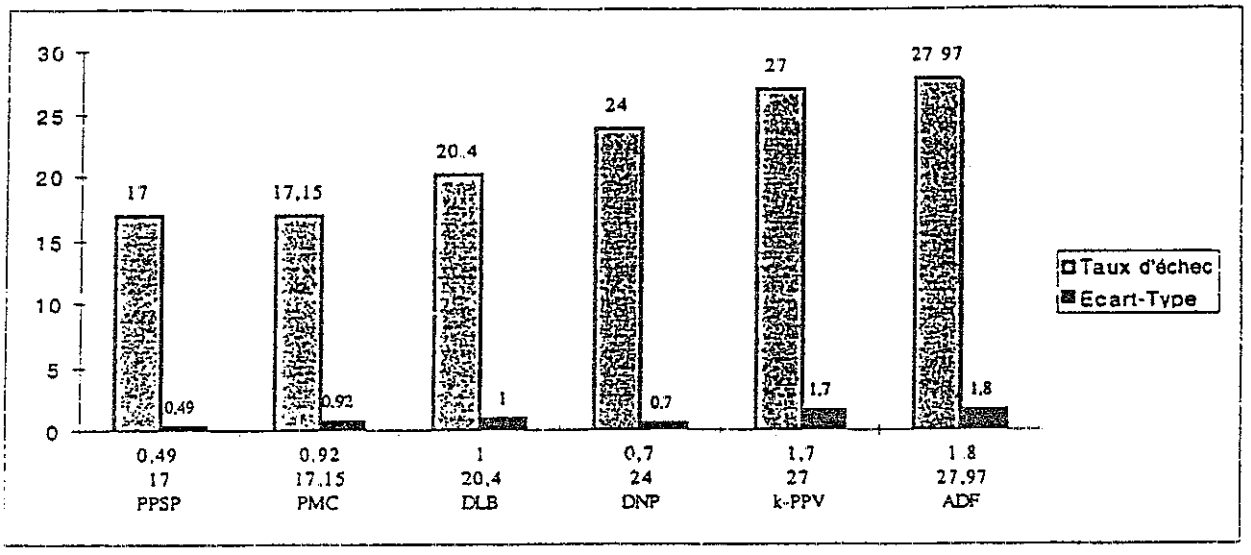

Figure 3 : Résultats comparatifs en test

On constate sur la figure 3 que PPSP est très efficace sur le problème des ondes de Breiman. Les performances obtenues sont meilleures que celles données par le perceptron multi-couches qui était le meilleur discriminateur des PRC. En effet non seulement le taux d'échec moyen est plus faible avec PPSP mais de plus l'écart-type de ces taux est presque deux fois plus faible avec notre méthode. Il est utile de se rappeller ici que les valeurs données en figure 3 sont calculées pour des échantillons de test de taille 4000 alors que les PRC avaient choisis 5000 éléments. Ces résultats prouvent aussi la bonne robustesse de la méthode proposée tant au regard du bruit que par rapport aux principaux paramètres de la population (probabilités a priori, taille de l'ensemble d'apprentissage...). La méthode des plus proches voisins adaptative ne se montre malheureusement pas compétitive sur ce problème, ni sur le plan des résultats, ni sur celui des temps de calcul.

4. CONCLUSION

Nous avons présenté dans cet article une nouvelle méthode de discrimination non paramétrique. Elle se caractérise par une phase d'apprentissage très rapide puisqu'elle se limite au calcul des structures propres des classes de l'ensemble d'apprentissage. Les tests réalisés sur le problème des ondes de Breiman montrent les excellentes performances de cette méthode comparées à celles des classifieurs classiques. La quasi-linéarité des frontières optimales pour les ondes de Breiman peut expliquer en partie la qualité de ces résultats. A l'avenir on envisage de tester ce classifieur sur des données réelles et d'en étudier certaines propriétés théoriques.

REFERENCES BIBLIOGRAPHIQUES

[Ulmo1973] **Ulmo J. (1973).** Différents aspects de l'analyse discriminante, Revue de Statistique Appliquée, 6, 2, pp 17-55.

[Caraux et al.1994] **Caraux G., Lechevallier Y. (1994).** Les méthodes statistiques de classement, Revue d'Intelligence Artificielle.

[Dubuisson1990] **Dubuisson B. (1990).** Diagnostique et reconnaissance des formes, Ed. Hermes.

[Amghar et al.1994] **Amghar S. , Zighed D. (1994).** The universal vote : a new criterion of classification, in 10th International Conference on Systems Engineering.

[Sebban1996] **Sebban M. (1996).** Modèles théoriques en reconnaissance de formes et architecture hybride pour machine perceptive, Thèse de doctorat, université de Lyon 1.

[Asselin de Beauville1995] **Asselin de Beauville J.-P. (1995).** Non parametric discrimination by the Nearest Principal Axis method (NPA)-Preliminary study, in Data science and its applications, Ed. Academic Press Inc, pp 145-154.

[Breiman et al.1984] **Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984).** Classification and regression trees, Ed. Wadsworth Inc.

[Projet PRC 1993] **Projet Inter-PRC (1993-1994).** Rapport final d'activité : Méthodes symboliques-numériques de discrimination, LAFORIA, université de Paris 6

[Bezdek1981] **Bezdek J.C. (1981).** Pattern recognition with fuzzy objective function algorithms, Ed. Plenum Press.

[Davé1989] **Davé R.N. (1989).** Use of the adaptative fuzzy clustering algorithm to detect lines in digital images, SPIE, vol. 1192, Intelligent Robots and Computer Vision VIII : Algorithms and Techniques.

[Duda et al.1973] **Duda R.O., Hart P.E. (1973).** Pattern classification and scene analysis, Ed. John Wiley.

[Parzen1962] **Parzen E. (1962).** On estimation of a probability density function and mode, Ann. Math. Stat. , 33, pp 1065-1076.

[Rumelhart1986] **Rumelhart D.E., Hinton G.E., Williams R.J. (1986).** Learning internal representations by error propagation, in Parallel Distributed Processing, vol. 1, chap 8, Ed. Cambridge MIT Press.

[Quinlan1986] **Quinlan J.R. (1986).** Induction of decision trees, in Machine Learning 1, pp 86-106.

[Marsala1994] **Marsala C. (1994).** Arbres de décision et sous-ensembles flous, rapport du LAFORIA n° 94/21.

[Mraghni1997] **Mraghni M.C. (1997).** Détection de chaines de contours dans une image numérique par approche symbolique et par grammaire de formes, Thèse de doctorat, université de Tours.