Classification supervisée avec option de rejet partiel et contraintes de performance basée sur l'estimation des densités de probabilité

Abdenour Bounsiar, Edith Grall-Maës, Pierre Beauseroy

Université de Technologie de Troyes. Institut Charles Delaunay-LM2S
abdenour.bounsiar@utt.fr, edith.grall@utt.fr, pierre.beauseroy@utt.fr

Résumé – Cette communication traite les problèmes de classification avec option de rejet partiel et contraintes de performances. L'objectif est d'étudier la possibilité d'exploiter la solution obtenue dans le cadre des tests d'hypothèses statistiques en y introduisant des densités de probabilité conditionnelles estimées. Deux estimateurs de densités sont utilisés et deux modalités d'optimisation des estimateurs sont comparées et analysées.

Abstract – This work deals with classification problems with class-selective rejection option and performance constraints. The goal is to study the feasibility of applying the solution obtained in the framework of statistical hypothesis testing by using estimated conditional probability densities. Two density estimators are used and two optimization modalities of the estimators parameters are compared and discussed.

1 Introduction

Les problèmes de classification sont souvent formulés à l'aide des différentes classes à discriminer et d'un risque de décision à minimiser. Il n'est toutefois pas très naturel, dans de nombreuses situations, de spécifier la règle de décision recherchée à l'aide du risque et de ces coûts. Ce formalisme peut être généralisé en ajoutant deux éléments. Le première est la possibilité de définir des options de décision regroupant plusieurs classes pour pratiquer le rejet d'ambiguïté [1] ou le rejet sélectif de classes [3]. Le second est l'introduction de contraintes en terme d'objectifs de performance. Ce type de problèmes a été traité du point de vue des tests d'hypothèses statistiques dans [2].

La solution fait intervenir les densités de probabilités des classes. L'objectif de cette communication est d'étudier la possibilité d'exploiter ces résultats théoriques dans le cas d'un apprentissage supervisé. Dans ce cas, le problème peut être résolu en introduisant les estimées des densités de probabilité dans la solution théorique. Pour comparer les performances des règles de classification obtenues, un critère de performance a été proposé dans [6]. Les techniques d'estimation considérées sont l'estimateur de Parzen et une méthode proposée récemment, qui est basée sur les régressions à vecteurs supports (RVS) et la théorie du champ moyen (CM) [4]. Dans cette communication, deux modalités d'optimisation des paramètres des estimateurs de densités sont comparées. L'une consiste à optimiser ces paramètres de façon à minimiser les erreurs quadratiques moyennes d'estimation des densités par rapports aux densités théoriques. L'autre consiste à optimiser ces paramètres de façon à minimiser directement le critère de performance proposé dans [6] sans information a priori sur les densités théoriques. L'objectif est de comparer les performances de solutions obtenues avec ces deux modalités d'optimisation.

Le problème traité est exposé dans la section 2. La section 3 est consacrée aux méthodes utilisées pour l'estimation des densités de probabilité. La section 4 présente les modalités adoptées pour l'évaluation des performances des solutions proposées. Les données synthétiques utilisées et les résultats expérimentaux sont présentés dans la section 5. La section 6 est consacrée à la discussion des résultats et aux conclusions.

2 Le problème de classification

Dans les problèmes de classification à n classes avec option de rejet partiel, un échantillon $\mathbf{x}$ qui appartient à une classe est affecté par la règle de décision à un sous-ensemble de classes ψ_i. L'ensemble des sous-ensembles de classes Ψ représente l'ensemble des options de décision : $\Psi = \{\psi_1, \psi_2, .., \psi_I\}$, où $I \leq 2^n - 1$ est le nombre de sous-ensembles de classes. Chaque sous-ensemble ψ_i est non vide et contient les indices des classes candidates. Les problèmes de classification avec option de rejet partiel et contraintes de performance, sont de la forme [2] :

$$\begin{cases} \underset{\boldsymbol{\mathcal{Z}}}{\text{minimiser}}\ \bar{c} \\ \text{avec les contraintes } e^{(k)} \leq \gamma^{(k)}, \forall k = 1..K, \end{cases} \quad (1)$$

où $\boldsymbol{\mathcal{Z}}$ est la partition de l'espace des observations en régions $\{\mathcal{Z}_i\}_{i=1}^{I}$, $\bar{c} = \sum_{i=1}^{I}\sum_{j=1}^{n} c_{ij} p(\omega_j) p(\mathfrak{D}_i|\omega_j)$ est la fonction coût à minimiser , $e^{(k)} = \sum_{i=1}^{I}\sum_{j=1}^{n} \alpha_{ij}^{(k)} p(\omega_j) p(\mathfrak{D}_i|\omega_j)$ est l'expression de la contrainte de performance, $\mathfrak{D}_i$ est la décision d'affectation au sous-ensemble de classes ψ_i, et $\gamma^{(k)}$ est un seuil de performance à ne pas dépasser. Les c_{ij} et $\alpha_{ij}^{(k)}$ sont des réels donnés. La partition optimale est

définie par les régions [2] :

$$\mathcal{Z}_i^* = \{\mathbf{x} \in \mathcal{X} | \lambda_i(\mathbf{x}, \boldsymbol{\mu}^*) < \lambda_l(\mathbf{x}, \boldsymbol{\mu}^*), \forall l = 1, .., I, l \neq i\},$$

où $\lambda_i(\mathbf{x}, \boldsymbol{\mu}^*) = \sum_{j=1}^{n} p(\omega_j) p(\mathbf{x}|\omega_j) \left(c_{ij} + \sum_{k=1}^{K} \mu_k^* \alpha_{ij}^{(k)} \right)$ et $\boldsymbol{\mu}^*$ est le vecteur des coefficients de Lagrange obtenus par résolution du problème dual de (1) [2].

3 Méthodes d'estimation

Deux estimateurs de densités sont considérés : l'estimateur de Parzen, présentée dans la section 3.1 et une méthode basée sur les régressions à vecteurs supports et la théorie du champ moyen, présentée dans la section 3.2.

3.1 Estimateur de Parzen

L'estimateur de Parzen pour la densité de probabilité $p(\mathbf{x}|\omega_j)$ d'une classe ω_j est défini par [10] :

$$\hat{p}(\mathbf{x}|\omega_j) = \frac{1}{N_j} \sum_{l=1}^{N_j} K_h(\mathbf{x}, \mathbf{x}_l), \quad j = 1, 2, 3, \qquad (2)$$

où K_h est une fonction noyau de paramètre h représentant généralement la largeur du noyau. Cette fonction doit être finie, positive et doit vérifier $\int_{-\infty}^{+\infty} K_h(\mathbf{x})\, d\mathbf{x} = 1$, ce qui veut dire qu'elle représente une fonction de densité de probabilité. Largeur h du noyau dépend de N de telle sorte que [11] : $\lim_{N \mapsto \infty} h = 0$ et $\lim_{N \mapsto \infty} N h^d = \infty$. Il peut être montré que sous ces conditions, la densité estimée $\hat{p}(\mathbf{x}|\omega_j)$ converge en erreur quadratique moyenne vers la vraie densité $p(\mathbf{x}|\omega_j)$ si celle ci est continue [10] :

$$\lim_{N \to \infty} \int \left[\hat{p}(\mathbf{x}|\omega_j) - p(\mathbf{x}|\omega_j) \right]^2 d\mathbf{x} = 0.$$

3.2 Estimation de densités par RVS

Afin d'estimer la fonction de densité de probabilité $p(\mathbf{x}|\omega)$ d'une classe ω à partir d'un ensemble d'échantillons $\{\mathbf{x}_i\}_{i=1}^N$, les auteurs dans [4] proposent d'estimer les valeurs de la fonction de distribution empirique aux points d'apprentissage : $t_i = \hat{P}_\omega(\mathbf{x}_i) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(\mathbf{x}_k \leq \mathbf{x}_i), i = 1..N$, puis d'utiliser un algorithme de régression sur l'ensemble des données $\mathcal{D} = \{(\mathbf{x}_i, t_i), i = 1, .., N\}$ afin d'obtenir une estimation de la fonction de distribution qui soit continue et différentiable. La fonction de densité de probabilité estimée $\hat{p}(\mathbf{x}|\omega)$ peut alors être obtenue par dérivation : $\hat{p}(\mathbf{x}|\omega) = \frac{d}{d\mathbf{x}} \hat{P}_\omega(\mathbf{x})$. En choisissant un noyau adéquat, les régressions à vecteurs supports (RVS) permettent d'obtenir une estimée qui est continue et différentiable.

3.2.1 Régression à vecteurs supports

Les RVSs sont de la forme linéaire $y(\mathbf{x}) = \langle \boldsymbol{w}, \mathbf{x} \rangle + b$. L'objectif est de trouver une fonction $y(\mathbf{x})$ qui soit la plus régulière possible et en même temps qui approxime aux mieux les t_i avec une précision $\epsilon > 0$. Afin de contrôler la régularité de la solution, on peut minimiser la norme du vecteur $\boldsymbol{w}$ [5]. Les erreurs d'approximation qui sont supérieures à ϵ sont représentées par des variables de décalage positives ζ_i et ζ_i^*. Le problème à résoudre est alors formulé par [5] :

$$\begin{array}{ll} \underset{\boldsymbol{w}, b, \boldsymbol{\zeta}, \boldsymbol{\zeta}^*}{\text{minimiser}} & \frac{1}{2}||\boldsymbol{w}||^2 + C \sum_{i=1}^{N} (\zeta_i + \zeta_i^*) \\ \text{sous} & \begin{cases} t_i - (\langle \boldsymbol{w}, \mathbf{x}_i \rangle + b) \leq \epsilon + \zeta_i, \\ (\langle \boldsymbol{w}, \mathbf{x}_i \rangle + b) - t_i \leq \epsilon + \zeta_i^*, \end{cases} \end{array} \qquad (3)$$

où $C > 0$ est un coût d'apprentissage. Afin d'adapter l'algorithme de régression à vecteurs supports aux problèmes non linéaires, des fonctions noyau $K_h(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ de paramètre h sont introduites, où la fonction ϕ est une transformation non linéaire dans un espace de grande dimension et $\langle ., . \rangle$ dénote un produit scalaire. La solution est un point selle du lagrangien associé au problème (3). L'expression de la solution est donnée par $y(\mathbf{x}) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K_h(\mathbf{x}_i, \mathbf{x}) + b$, où les α_i, α_i^* sont les multiplicateurs de Lagrange associés aux deux contraintes du problème (3).

3.2.2 RVS et théorie du champ moyen

Il a été montré que le problème d'optimisation des RVSs peut être interprété comme un problème d'estimation à maximum a posteriori (MAP) sous certaines hypothèses [7]. Pour ce faire, la probabilité $p(t|y(\boldsymbol{x}))$ est supposée régie par un modèle exponentiel dépendant de la fonction perte de Vapnik $\mathcal{L}_\epsilon$ [4], de plus la sortie des RVSs est considérée comme un processus gaussien $N_{\mathbf{y}(\mathbf{X})}(\mathbf{0}, K_N)$ de moyenne nulle et de matrice de variances-covariances K_N. En appliquant le théorème de Bayes, l'estimateur MAP de $p(\mathbf{y}(\mathbf{X})|\mathbf{t})$ est alors donné par :

$$\min_{\mathbf{y}(\mathbf{X})} \quad C \sum_{i=1}^{N} \mathcal{L}_\epsilon(t_i - y(\mathbf{x}_i)) + \frac{1}{2} \mathbf{y}(\mathbf{X}) K_N^{-1} \mathbf{y}(\mathbf{X})^T. \qquad (4)$$

où $\mathbf{t} = [t_1, t_2, ..., t_N]$ et $\mathbf{y}(\mathbf{X}) = [y(\mathbf{x}_1), y(\mathbf{x}_2), ..., y(\mathbf{x}_N)]$. La fonction dans (4) représente la fonction objectif minimisées par les RVSs [9].

Le problème d'optimisation (3) des RVSs est d'une complexité temporelle de $\mathcal{O}(N^3)$. Ceci rend les RVSs peu commodes pour l'estimation de densités qui généralement nécessite un grand nombre d'échantillons. Au lieu de résoudre le problème (3), il est possible d'exploiter l'interprétation probabiliste donnée au RVSs et d'essayer d'approximer l'espérance de la sortie des RVSs [8].

La probabilité a posteriori $p(y(\mathbf{x})|\mathbf{t})$ définit la connaissance sur la sortie $y(\mathbf{x})$ après l'observation de l'ensemble d'apprentissage. Selon [8], la prédiction moyenne suivant cette probabilité est donnée pour une observation $\mathbf{x}$ par :

$$\langle y(\mathbf{x}) \rangle = \sum_{i=1}^{N} K_h(\mathbf{x}, \mathbf{x}_i) w_i, \qquad (5)$$

où $\langle \ldots \rangle$ dénote une moyenne et les w_i's sont des constantes qui sont estimées en utilisant le principe de "Leave-One-Out" : l'échantillon $\mathbf{x}_i$ est écarté de l'ensemble d'apprentissage, et le coefficient w_i est estimé en utilisant les autres échantillons suivant la distribution $p(y(\mathbf{x}_i)|\overline{\mathbf{t}}_i)$, où $\overline{\mathbf{t}}_i$ est obtenu en écartant t_i de $\mathbf{t}$. En effet, $p(y(\mathbf{x}_i)|\overline{\mathbf{t}}_i)$ est la distribution de la prédiction au point "test" $\mathbf{x}_i$ sachant l'ensemble de données $\overline{\mathbf{t}}_i$. En notant la moyenne suivant cette distribution par $\langle \cdots \rangle_i$, les coefficients w_i peuvent s'écrire par [8] :

$$w_i = \frac{\left\langle \frac{\partial}{\partial y(\mathbf{x}_i)} \exp\left\{-C\mathcal{L}_\epsilon(t_i - y(\mathbf{x}_i))\right\} \right\rangle_i}{\left\langle \exp\left\{-C\mathcal{L}_\epsilon(t_i - y(\mathbf{x}_i))\right\} \right\rangle_i}. \qquad (6)$$

Afin d'éviter de telles intégrales numériquement difficiles à calculer, Opper et Winther [8] proposent d'utiliser des résultats de la théorie du champ moyen pour simplifier les intégrales multiples intervenant dans le calcul des w_i en approximant $p(y(\mathbf{x}_i)|\bar{\mathbf{t}}_i)$ par une distribution gaussienne de moyenne $\langle y(\mathbf{x}_i)\rangle_i$ et de variance $\sigma_i^2 = \langle y(\mathbf{x}_i)^2\rangle_i - \langle y(\mathbf{x}_i)\rangle_i^2$. Des équations de champ moyen et un algorithme permettant de calculer ces deux quantités sont proposés dans [8].

4 Modalités d'évaluation

La qualité des solutions est évaluée à l'aide du critère de performance κ proposé dans [6] :

$$\kappa = \sum_{i,j=1}^{I,n} \left(c_{ij} + \sum_{k=1}^{K} \mu_k^* \alpha_{ij}^{(k)} \right) p(\omega_j) p(\tilde{\mathfrak{D}}_i|\omega_j) d\mathbf{x} - \sum_{k=1}^{K} \mu_k^* \gamma^{(k)},$$

où les probabilités de décision $p(\tilde{\mathfrak{D}}_i|\omega_j)$ sont évaluées sur les régions de décision $\tilde{\mathcal{Z}}_i$ déterminées à partir des densités estimés. Les multiplicateurs de Lagrange μ_k devront être estimés à partir des données d'apprentissage. Cependant, pour ne pas introduire en premier temps les problèmes d'estimation de ces multiplicateurs, le critère κ est calculé en utilisant les multiplicateurs de Lagrange théoriques μ^*. Avec ce mode de calcul, le meilleur modèle sera toujours retenu.

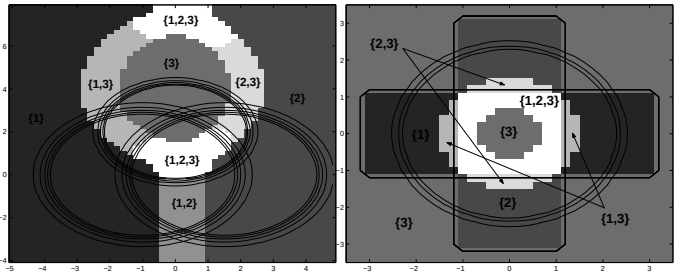
Deux modalités d'optimisation des performances ont été utilisées. La première modalité consiste à optimiser les paramètres de chacun des deux estimateurs afin de minimiser l'erreur d'estimation quadratique moyenne des densités marginales : $E_i = \int \left[\hat{p}(\mathbf{x}|\omega_i) - p(\mathbf{x}|\omega_i)\right]^2 d\mathbf{x}, i = 1, 2, 3$. La seconde modalité consiste à choisir les paramètres de chacun des deux estimateurs qui minimisent directement le critère de performance κ. Rappelons que l'estimateur de Parzen dépend du paramètre h du noyau utilisée (2), la méthode RVS-CM dépend du paramètre h du noyau utilisé (5) et du coût d'apprentissage C (3).

5 Étude expérimentale

Afin de tester expérimentalement la possibilité de l'utilisation d'estimées des densités de probabilité, deux problèmes de classification synthétiques différents ont été considérés. Pour chaque problème, les deux modalités d'optimisation des paramètres des estimateurs, citées ci-dessus, ont été considérées pour l'estimateur de Parzen, et pour l'estimateur basé sur les régressions à vecteurs supports et la théorie du champ moyen (RVS-CM). La définition des données synthétiques utilisées et la présentation du problème de classification sont donnés dans la section 5.1. Les résultats expérimentaux sont donnés dans la section 5.2.

5.1 Présentation du problème simulé

Deux problèmes synthétiques, définis dans $\mathbb{R}^2$ sont considérés. Le premier est défini par trois classes gaussiennes équiprobables. Le deuxième problème est défini par trois classes équiprobables : deux de densités uniformes et une de densité gaussienne. Les fonctions des densités de probabilité des classes sont représentées sur les figures 1.a et 1.b par des courbes de niveau pour les problèmes 1 et 2 respectivement.



(a) (b)

FIG. 1 – Partitions optimales, représentées en niveaux de gris, pour les deux problèmes à : (a) trois densités gaussiennes, (b) deux densités uniformes et une gaussienne.

Pour les deux problèmes, les sous-ensembles de classes sont : $\psi_1 = \{1\}, \psi_2 = \{2\}, \psi_3 = \{3\}, \psi_4 = \{1,2\}, \psi_5 = \{1,3\}, \psi_6 = \{2,3\}, \psi_7 = \{1,2,3\}$, les contraintes sont : $P_E \leq 0.03$ et $P_I \leq 0.12$ pour le problème à trois gaussiennes, $P_E \leq 0.1$ et $P_I \leq 0.1$ pour le problème à deux densités uniformes et une gaussienne. La probabilité d'erreur P_E est définie par : $P_E = \sum_{i=1}^{3} \sum_{j=1, i \notin \psi_j}^{6} p(\mathfrak{D}_j|\omega_i) p(\omega_i)$, et la probabilité d'indétermination P_I est définie par : $P_I = \sum_{i=1}^{3} \sum_{j=4, i \in \psi_j}^{6} p(\mathfrak{D}_j|\omega_i) p(\omega_i)$. Pour les deux problèmes, le coût moyen est défini par $\bar{c} = P_E + 0.5 P_I + p(\mathfrak{D}_7)$.

Les partitions théoriques optimales de ces deux problèmes sont représentées en niveaux de gris sur les figures 1.a et 1.b respectivement.

Pour chacun des deux problèmes, les expérimentations ont été menées sur deux groupes de 50 ensembles d'apprentissage : l'un avec des ensembles de 50 échantillons par classe et l'autre avec des ensembles de 200 échantillons par classe.

5.2 Résultats expérimentaux

Les résultats moyens avec écart-type des erreurs quadratiques d'estimation des densités sont données dans les tables 1.a et 1.b respectivement. Les valeurs moyennes et les écart-type du critère κ sont reportées dans la table 2. Les valeurs théoriques optimales de *kappa* pour ces deux

TAB. 1 – Erreurs quadratiques moyennes, avec écart-types, sur l'estimation des densités de probabilité : (a) trois gaussiennes, (b) deux uniformes et une gaussienne.

(a)

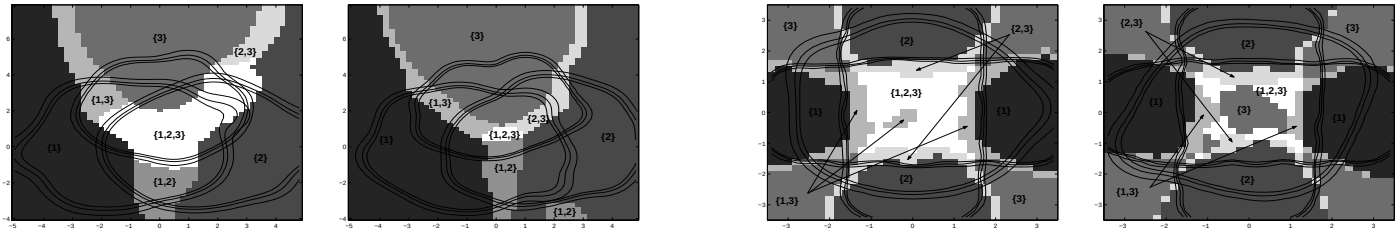
	50 échantillons par classe		200 échantillons par classe	
	Parzen	RVS-CM	Parzen	RVS-CM
FdpG1	0.019 ± 0.009	0.012 ± 0.007	0.008 ± 0.002	0.006 ± 0.002
FdpG2	0.016 ± 0.006	0.010 ± 0.004	0.008 ± 0.003	0.006 ± 0.002
FdpG3	0.034 ± 0.017	0.024 ± 0.012	0.016 ± 0.007	0.012 ± 0.005

(b)

	50 échantillons par classe		200 échantillons par classe	
	Parzen	RVS-CM	Parzen	RVS-CM
FdpU1	0.054 ± 0.008	0.050 ± 0.006	0.038 ± 0.003	0.037 ± 0.003
FdpU2	0.055 ± 0.006	0.051 ± 0.005	0.037 ± 0.016	0.036 ± 0.003
FdpG	0.036 ± 0.016	0.026 ± 0.014	0.016 ± 0.007	0.013 ± 0.006

TAB. 2 – Valeurs du critère de performance κ obtenues sur les deux problèmes synthétiques.

	3 densités gaussiennes		2 densités uniformes + 1 gaussienne	
	Parzen	RVS-CM	Parzen	RVS-CM
50 éch/classe : optimisation des erreurs d'estimation E_i	0.3730 ± 0.0190	0.3636 ± 0.0115	0.5852 ± 0.0132	0.5767 ± 0.0143
50 éch/classe : optimisation du critère de performance κ	0.3613 ± 0.0170	0.3574 ± 0.0125	0.5805 ± 0.0114	0.5654 ± 0.0108
200 éch/classe : optimisation des erreurs d'estimation E_i	0.3508 ± 0.0064	0.3496 ± 0.0063	0.5525 ± 0.0097	0.5483 ± 0.0093
200 éch/classe : optimisation du critère de performance κ	0.3437 ± 0.0060	0.3410 ± 0.0048	0.5453 ± 0.0090	0.5382 ± 0.0064



(a.1) Parzen (a.2) RVS-CM (b.1) Parzen (b.2) RVS-CM

FIG. 2 – Exemple de partition obtenue sur un ensemble de 200 échantillons par classe pour le problème à : (a.1, a.2) trois densités gaussiennes, (b.1, b.2) deux densités uniformes et une gaussienne, en utilisant les deux estimateurs.

problèmes sont 0.333 et 0.499 respectivement.

Un exemple de partition obtenue sur un ensemble de 200 échantillons par classe par l'estimateur de Parzen et la méthode RVS-CM sont représentées par la figure 2 pour les deux problèmes synthétiques.

6 Discussions et conclusions

Les erreurs d'estimation de densités données par les tables 1.a et 1.b montrent que la méthode RVS-CM estime légèrement mieux la densité vraie que l'estimateur de Parzen. Le gain en estimation observé avec la méthode RVS-CM est plus important pour les ensembles de 50 échantillons, ce qui s'explique par le fait que l'estimateur de Parzen tend asymptotiquement vers la vraie densité. Ce gain est plus important pour les densités gaussiennes qui ne posent pas de problèmes d'effets de bord contrairement aux densités uniformes. La table 2 montre que les performances des deux estimateurs sont très proches avec une légère supériorité pour la méthode RVS-CM.

Les résultats obtenus par minimisation directe du critère d'optimisation κ sont meilleurs que ceux obtenus par minimisation des erreurs d'estimation. Ce résultat est logique puisque la solution optimale correspond au minimum de κ. Ces résultats ne sont pas trop éloignés des résultats théoriques (notamment pour les ensembles de 200 échantillons par classe). Ils montrent d'une part la faisabilité de la solution proposée basée sur l'estimation des densités de probabilité, et montrent d'autre part que le choix du critère d'optimisation des estimateurs est plus important que le choix de l'estimateur.

Il reste toutefois à étudier les modalités d'estimation de κ en l'absence de connaissances statistiques a priori. L'utilisation de méthodes d'apprentissage et de validation sera alors nécessaire pour sélectionner les paramètres optimaux des classifieurs. Il sera également nécessaire de vérifier que le critère de performance κ estimé demeure pertinent.

Références

[1] C.K. Chow, On optimum recognition error and reject tradeoff, *IEEE Transactions on Information Theory*, vol. IT-16, no. 1, pp. 41-46, 1970.

[2] E. Grall, P. Beauseroy, and A. Bounsiar. Multilabel classification rule with performance constraints. In proceedings of IEEE conference ICASSP'06. Toulouse, France, 14-19 May 2006.

[3] T. Ha. The optimum class-selective rejection rule. *Transactions on Pattern Analysis and Machine Intelligence*, 19(6) :608-615, 1997.

[4] R. M. Mohamed, A. El-Baz, and A. A. Farag. Probability density estimation using advanced support vector machines and the em algorithm. *International Journal of Signal Processing*, 1(4) :260-264, 2004.

[5] V. Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.

[6] E. Grall, P. Beauseroy, and A. Bounsiar. Quality assessment of a supervised multilabel classification rule with performance constraints. EUSIPCO'06, Florence(Italy), 04-08 September 2006.

[7] P. Sollich. Probabilistic interpretations and bayesian methods for support vector machines. Technical report, King's College London, London, 1998.

[8] M. Opper and O. Winther. Gaussian processes for classification : Mean field algorithms. *Neural Computation*, 12(11) :2655-2684, 2000.

[9] C. Wei and S. S. Keerthi and J. O. Chong. Bayesian support vector regression using a unified loss function. *IEEE Trans. on Neur. Net.*, 15(1) :29-44. 2004.

[10] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statistics*, 33 :1065-1076, Sept. 1962.

[11] G. A. Babich and O. I. Camps. Weighted parzen windows for pattern classification. IEEE Trans. on Pattern analysis and Machine Intelligence, 18(5) :567-570, May 1996.