Classification avec contraintes : problématique et apprentissage d'une règle de décision par SVM

Edith Grall-Maës, Pierre Beauseroy, Abdenour Bounsiar
Institut des Sciences et Technologies de l'Information de Troyes (CNRS FRE 2732)
Équipe Modélisation et Sûreté des Systèmes
Université de Technologie de Troyes
12, rue Marie Curie - BP 2060 -10010 Troyes cedex - FRANCE
{Edith.Grall,Pierre.Beauseroy,Abdenour.Bounsiar}@utt.fr

Résumé – Le travail présenté porte sur la détermination d'une règle de décision pour un problème avec deux classes et deux contraintes qui fixent des bornes supérieures pour les probabilités d'erreur conditionnelles aux classes. Dans le cas où il existe des règles de décision satisfaisant conjointement les contraintes, la règle choisie sera celle qui minimise un coût combinant les probabilités de décision conditionnelles. Dans le cas contraire, il est nécessaire de définir une règle qui ajoute une classe de rejet. La règle optimale recherchée est alors est celle qui minimise la probabilité de rejet. Dans un premier temps, la règle de décision est définie lorsque les densités de probabilités de chacune des classes sont connues : elle consiste à comparer le rapport de vraisemblance à un ou deux seuils selon que du rejet est nécessaire ou pas. Dans un second temps, une méthode basée sur les SVM est proposée pour élaborer une règle de décision lorsque le processus est uniquement décrit par un ensemble d'apprentissage. La règle consiste à comparer la sortie du SVM avec un ou deux seuils. Ceux-ci sont déterminés à partir des sorties du SVM pour un ensemble d'apprentissage, soit directement avec les valeurs de sortie, soit à partir d'estimation des densités de probabilités. Le biais et la variance des probabilités de rejet et d'erreurs conditionnelles inhérents à la taille de l'ensemble d'apprentissage sont étudiés.

Abstract – The paper deals with the definition of a decision rule for a two classes and two constraints problem. The considered constraints bound the conditional error probabilities. When decision rules verifying both constraints exist, the rule that minimizes a cost combining the conditional decision probabilities is chosen. If such rule does not exist, it is necessary to introduce a rejection class and the optimal rule is the one that minimizes the reject probability. Firstly the rule is defined assuming that the conditional density functions are known. It consists in comparing the likelihood ratio with one or two thresholds depending on the use of rejection or not. Secondly, a method based on SVM is proposed to design decision rules for processes described by sample sets. The rule consists in comparing the function of the SVM output with thresholds. They are determined using outputs of a training set, either directly, either by estimating the output densities. The bias and the standard deviation of the reject and of the error probabilities arising from the size of the training set are studied.

1 Introduction

Les approches théoriques de la classification entre deux classes ont permis d'établir différentes règles de décision optimales au sens de certains critères, par exemple la règle de Bayes, le test de Neyman-Pearson, le minimax [1]. Elles peuvent être utilisées pour autant que le problème posé vérifie les hypothèses d'établissement de ces règles.
L'introduction de contraintes dans ces problèmes peut conduire à l'absence de solutions dans un cadre classique. Pour traiter ces situations, la notion de rejet a été développée notamment dans le cas de contraintes portant sur la probabilité d'erreur totale [2, 3, 4]. Toutefois un certain nombre de problèmes réels peuvent impliquer un ensemble de contraintes plus complexe. Les contraintes peuvent s'exprimer selon un ou plusieurs coûts qui combinent différentes probabilités d'erreur (totale ou conditionnelle). Elles peuvent impliquer des conditions d'inégalité ou d'ordre. Pour tous ces scenarii, deux cas se présentent : soit les contraintes peuvent être satisfaites en appliquant une règle de décision classique, alors cette solution est adoptée ; soit pour satisfaire les contraintes il est nécessaire d'introduire du rejet, c'est-à-dire une non décision pour certaines observations. Dans ce cas la règle vise à minimiser la probabilité de rejet. Pour qualifier ce type de règle il convient de mesurer les performances mais aussi de s'assurer du respect des contraintes. Dans les cas d'apprentissage à partir d'exemples étiquetés la qualification pose un problème particulier compte tenu de la variabilité inhérente à l'ensemble d'échantillons considéré.
Le travail proposé est une illustration de ce type de problème au cas particulier où les probabilités d'erreur conditionnelles à chacune des classes sont bornées. Le paragraphe 2 décrit précisément le problème de classification posé. Le paragraphe 3 est consacré au développement de la règle de décision lorsque les densités de probabilités sont connues. Le paragraphe 4 traite des problèmes d'estimation lorsque le processus est seulement connu à l'aide d'un ensemble de données étiquetées. Dans le paragraphe 5, une méthode est proposée pour construire un classifieur dans le cas d'un processus décrit par un ensemble d'apprentissage ; elle s'appuie sur les machines à vecteurs de support (SVM). Son application à un exemple simulé est présentée dans le paragraphe 6, qui précède la conclusion.

2 Classification avec contraintes

Soient deux classes C_1 et C_2. On considère le problème de classification binaire avec les deux contraintes suivantes :

$$\begin{cases} P(D_1/C_2) \leq e_{12} \\ P(D_2/C_1) \leq e_{21} \end{cases} \quad \text{où } e_{12} \text{ et } e_{21} \in [0,1] \tag{1}$$

avec $P(D_i/C_j)$ la probabilité de décider la classe i conditionnellement à la classe j et e_{ij} la borne supérieure pour la probabilité $P(D_i/C_j)$.

Le problème à résoudre diffère selon qu'il existe ou non au moins une partition (Z_1, Z_2) de $\mathbb{R}^n$ telle que les contraintes

soient vérifiées. S'il existe au moins un couple (Z_1, Z_2), le problème consiste à rechercher celui qui minimise un critère, qui de façon générale est un coût $\overline{c}$ qui s'exprime selon :

$$\begin{aligned}\overline{c} = \quad & c_{11}P(D_1/C_1)P_1 + c_{12}P(D_1/C_2)P_2 \\ & + c_{21}P(D_2/C_1)P_1 + c_{22}P(D_2/C_2)P_2.\end{aligned}$$

où P_i est la probabilité *a priori* de la classe i. En prenant $c_{ii} = 0$, $c_{12} = c_{21} = 1$, le critère $\overline{c}$ est égal à la probabilité d'erreur P_e.

S'il n'existe pas de partition (Z_1, Z_2) telle que les contraintes soient vérifiées, il est nécessaire, pour satisfaire les contraintes, d'introduire une zone de rejet, dans laquelle aucune décision sur les réalisations n'est prise. Le problème consiste alors à rechercher la partition (Z_1, Z_2, Z_R) qui minimise la probabilité de rejet P_R.

La qualification de la règle de décision nécessite d'estimer des grandeurs caractéristiques (moyenne, écart-type) du critère minimisé ($\overline{c}$ ou P_R selon le cas) et également des probabilités intervenant dans la définition des contraintes.

3 Élaboration d'une règle de décision à partir des densités de probabilités

Lorsque les densités de probabilités conditionnelles à chacune des classes $P(x/C_1)$ et $P(x/C_2)$ et les probabilités *a priori* P_1 et P_2 sont connues, la règle de décision théorique peut être élaborée.

Tout d'abord, il est nécessaire de déterminer si les contraintes peuvent être conjointement satisfaites. À cette fin, il est possible de considérer chaque contrainte indépendamment puis d'analyser la compatibilité des solutions obtenues.

Soit le domaine Z_1^* tel que la contrainte $P(D_1/C_2) = e_{12}$ soit vérifiée et le coût $\overline{c}$ soit minimum. Pour déterminer la solution à ce problème, la démarche est similaire à celle pour le test de Neymann-Pearson. Z_1^* est défini à partir d'un seuil λ_1^* sur le rapport de vraisemblance $l(x)$ défini par :

$$l(x) = \frac{P(x/C_1)}{P(x/C_2)}. \tag{2}$$

Z_1^* est donné par :

$$Z_1^* = \{x | l(x) \geq \lambda_1^*\} \text{ avec } \int_{Z_1^*} P(x/C_2)dx = e_{12}. \tag{3}$$

Par symétrie, on définit le domaine Z_2^* qui satisfait la contrainte $P(D_2/C_1) = e_{21}$ et qui minimise $\overline{c}$. On montre que Z_2^* est défini à partir d'un seuil λ_2^* sur le rapport de vraisemblance $l(x)$ selon :

$$Z_2^* = \{x | l(x) \leq \lambda_2^*\} \text{ avec } \int_{Z_2^*} P(x/C_1)dx = e_{21}. \tag{4}$$

Dans le cas où $\lambda_1^* \leq \lambda_2^*$, l'intersection entre les domaines Z_1^* et Z_2^* est non nulle, et il est donc possible de trouver une loi qui satisfasse les deux contraintes sans introduire de rejet. Toute règle de décision qui engendre une partition à partir d'un seuil $\lambda \in [\lambda_1^*, \lambda_2^*]$ sur le rapport de vraisemblance vérifie les contraintes. La règle de décision qui minimise le coût $\overline{c}$ et vérifie les contraintes est :

$$l(x) \underset{D_2}{\overset{D_1}{\gtrless}} \lambda \quad \text{avec } \lambda = \begin{cases} \lambda_0 \text{ si } \lambda_1^* \leq \lambda_0 \leq \lambda_2^* \\ \lambda_1^* \text{ si } \lambda_0 \leq \lambda_1^* \leq \lambda_2^* \\ \lambda_2^* \text{ si } \lambda_1^* \leq \lambda_2^* \leq \lambda_0. \end{cases} \tag{5}$$

où λ_0 est le seuil sur le rapport de vraisemblance qui minimise $\overline{c}$ sans contraintes :

$$\lambda_0 = \frac{P_2(c_{12} - c_{22})}{P_1(c_{21} - c_{11})}. \tag{6}$$

Dans le cas où $\lambda_1^* > \lambda_2^*$, il est nécessaire de faire appel à du rejet pour satisfaire les deux contraintes. La probabilité de rejet minimum est obtenue lorsque les contraintes sont saturées. La règle de décision est donc :

$$l(x) \quad \begin{cases} \geq \lambda_1^* \text{ alors } D_1 \\ \leq \lambda_2^* \text{ alors } D_2 \\ \notin [\lambda_2^*; \lambda_1^*] \text{ alors } D_R. \end{cases} \tag{7}$$

4 Performances d'une règle de décision élaborée à partir d'échantillons

Généralement les probabilités théoriques ne sont pas connues et le processus est uniquement décrit par un ensemble d'apprentissage. Le classifieur doit alors être élaboré à partir de cet ensemble d'échantillons étiquetés. Une famille d'approches pour construire la règle de décision consiste à déterminer une fonction représentative du rapport de vraisemblance et à déterminer les seuils à appliquer sur la fonction. Les performances du classifieur alors sont liées aux erreurs d'estimation de ces deux grandeurs.

Il est assez aisé de mesurer l'effet de la taille de l'ensemble d'apprentissage sur l'erreur d'estimation des seuils. Considérons que le rapport de vraisemblance est parfaitement connu mais pas les densités conditionnelles. la détermination des seuils λ_1^* et λ_2^* se fait sans biais mais avec une certaine variance dépendant du nombre d'échantillons. En effet, considérons une loi donnée et un domaine D, avec p la probabilité théorique qu'un échantillon appartienne au domaine D et N le nombre d'échantillons tirés de cette loi. Soit la variable aléatoire P_N mesurant la proportion d'échantillons dans le domaine D ; sa loi de probabilité est donnée par :

$$P\left[P_N = \frac{k}{N}\right] = C_N^k p^k (1-p)^{N-k}. \tag{8}$$

On peut montrer que la moyenne et l'écart-type de la la variable aléatoire P_N sont :

$$E[P_N] = p \quad \text{et} \quad \sigma[P_N] = \sqrt{\frac{p(1-p)}{N}}. \tag{9}$$

Par conséquent, même si l'on est capable de disposer une estimation parfaite du rapport de vraisemblance, l'élaboration d'un classifieur à partir d'un ensemble de N échantillons permettra d'obtenir un classifieur sans biais mais avec une variance inversement proportionnelle au nombre d'échantillons.

5 Règle de décision basée sur les SVM

Lorsque le processus est uniquement décrit par un ensemble d'apprentissage, le rapport de vraisemblance n'est pas connu et difficilement estimable avec précision. Il est alors nécessaire de construire une règle de décision à l'aide d'une autre méthode. Les SVM ont été utilisés en raison de leurs aptitudes à trouver des frontières non linéaires et leur robustesse vis à vis de l'ensemble d'apprentissage.

Considérons un ensemble de vecteurs d'apprentissage appartenant à deux classes :

$$\mathcal{A} = \{(x_1, y_1), ...(x_l, y_l)\} \quad x_i \in \mathbb{R}^n, y_i \in \{-1, 1\} \quad (10)$$

Le principe des vecteurs à support machine (SVM) [5, 6] repose sur la transformation d'un vecteur x dans un espace de grande dimension et la construction d'un hyperplan optimal dans cet espace. Le produit scalaire dans l'espace transformé est effectué à l'aide d'un noyau K défini par un vecteur de paramètres θ. Le noyau gaussien est couramment utilisé ; il est défini par :

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\rho^2}\right). \quad (11)$$

La fonction de décision du SVM est la solution de la minimisation de :

$$\frac{1}{2}\|w\|^2 + C\sum_i \xi_i \quad (12)$$

où C est une constante de régularisation qui détermine le compromis entre l'erreur empirique et le terme de complexité, et w est le vecteur normal à l'hyperplan séparateur dans l'espace transformé.

Pour le problème avec contraintes, par analogie avec la règle obtenue sur le rapport de vraisemblance, il est possible d'élaborer une règle de décision en exploitant la dynamique de la sortie du SVM :

$$\widetilde{f}_\theta(x) = \sum_{i=1}^{l} a_i y_i K_\theta(x_i, x) + b \quad (13)$$

où les a_i sont les multiplicateurs de Lagrange du problème dual de (12).

La règle est alors définie comme suit :

$$\widetilde{f}_\theta(x) \quad \begin{cases} \geq \xi_1^* \text{ alors on décide } D_1 \\ \leq \xi_2^* \text{ alors on décide } D_2 \\ \notin [\xi_2^*; \xi_1^*] \text{ alors on décide } D_R \end{cases} \quad (14)$$

L'estimation des seuils ξ_1^* et ξ_2^* doit être faite en utilisant un ensemble d'apprentissage $\mathcal{A}'$ indépendant de l'ensemble $\mathcal{A}$ utilisé pour construire la fonction de décision du SVM.

Notons $z_j^{(i)}$ (où (i) indique la classe et j indique l'échantillon) les valeurs des $\widetilde{f}_\theta(x)$ de l'ensemble d'apprentissage $\mathcal{A}'$. Pour déterminer les seuils, le procédé le plus simple est d'estimer l'erreur empirique pour différents seuils et de sélectionner les seuils permettant de respecter les contraintes :

$$\frac{1}{N_2}\sum_{i \in C_2} I_1(i) = e_{12} \qquad \tfrac{1}{N_1}\textstyle\sum_{i \in C_1} I_1(i) = e_{21}$$

$$I_1(i) = \begin{cases} 1 \text{ si } \widetilde{f}_\theta(z_j^{(2)}) > \xi_1^* \\ 0 \text{ sinon} \end{cases} \qquad I_2(i) = \begin{cases} 1 \text{ si } \widetilde{f}_\theta(z_j^{(1)}) < \xi_2^* \\ 0 \text{ sinon.} \end{cases}$$

Une autre approche consiste à utiliser les valeurs $z_j^{(i)}$ pour estimer les densités de probabilité conditionnelles aux classes. L'estimation pour chacune des classes peut être obtenue à l'aide de l'estimateur de Parzen :

$$\hat{p}(z^{(i)}) = N_i^{\alpha-1}\frac{(2\pi)^{1/2}}{\widehat{\sigma}} \sum_{j=1}^{N_i} \exp\left(-\frac{1}{2}N_i^{2\alpha}\frac{(z^{(i)} - z_j^{(i)})^2}{\widehat{\sigma}}\right)$$

où $\widehat{\sigma}$ est une estimation de l'écart-type des valeurs, N_i est le nombre d'échantillons dans la classe considérée, et α est un coefficient de lissage.

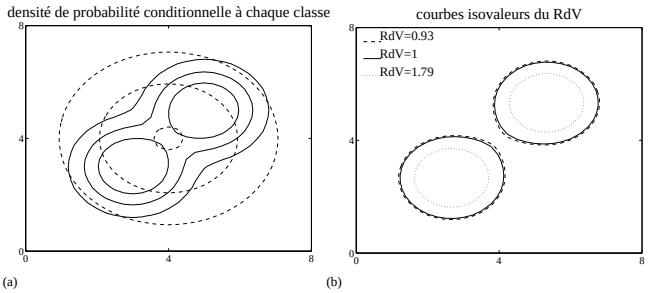

FIG. 1 – courbes isovaleurs de (a) la densité de probabilité conditionnelle à chacune des classes (b)du rapport de vraisemblance

6 Application à un problème simulé

6.1 Description du problème simulé

On recherche une règle de décision qui satisfasse les deux contraintes présentées ci-dessus avec $e_{12} = e_{21} = 0.2$ pour le processus observé dans $\mathbb{R}^2$ décrit ci-après. Les classes 1 et 2 sont équiprobables et respectivement composées du mélange de deux distributions normales, et d'une distribution normale seule. Elles sont représentées par des courbes isovaleurs sur la figure 1a.

Différentes probabilités théoriques ont été calculés numériquement. Pour un rapport de vraisemblance égal à 1, la probabilité d'erreur totale est minimum et égale à 0.312, $P(D_2/C_1) = 0.228$ et $P(D_1/C_2) = 0.396$. Pour les valeurs de e_{12} et e_{21} choisies les seuils obtenus sont $\lambda_2^* = 0.93$ et $\lambda_1^* = 1.79$. Les contraintes ne peuvent donc être satisfaites qu'en introduisant du rejet. Les courbes isovaleurs pour ces valeurs de rapport de vraisemblance (RdV) sont représentées sur la figure 1b.

L'expérimentation a été effectuée avec 50 ensembles d'apprentissage comptant chacun par classe 100 échantillons pour la construction du classifieur et 100 échantillons pour la détermination du seuil. Pour la règle de décision optimale donnée par (7), les moments théoriques et estimés à partir des 50 ensembles sont :

$$\begin{array}{ll} E[P(D_1/C_2)] = 0.2, & \widehat{E}[\widehat{P}(D_1/C_2)] = 0.205 \\ \sigma[P(D_1/C_2)] = 0.04, & \widehat{\sigma}[\widehat{P}(D_1/C_2)] = 0.036, \\ E[P(D_2/C_1)] = 0.2, & \widehat{E}[\widehat{P}(D_2/C_1)] = 0.204 \\ \sigma[P(D_2/C_1)] = 0.04, & \widehat{\sigma}[\widehat{P}(D_2/C_1)] = 0.038, \\ E[P_R] = 0.248, & \widehat{E}[\widehat{P}_R] = 0.247 \\ \sigma[P_R] = 0.030, & \widehat{\sigma}[\widehat{P}_R] = 0.031 \end{array}$$

Ces valeurs mettent en évidence qu'avec 100 échantillons dans chaque classe, le problème peut être résolu au mieux avec un écart-type de l'ordre de 20% ($\sigma = 0.04$) de l'erreur de consigne ($e_{12} = 0.2$).

6.2 Performances du classifieur

Le classifieur a été construit en prenant un noyau gaussien défini par (11). Les paramètres ont tout d'abord été optimisés en utilisant les méthodes proposées dans [7] et [8]. La figure 2 montre les courbes isovaleurs de la sortie du SVM obtenu à partir d'un ensemble d'apprentissage pour les deux seuils correspondant aux contraintes d'erreur e_{12} et e_{21}.

La figure 3 montre les densités de probabilité conditionnelles des sorties du SVM pour un ensemble d'échantillons donné,

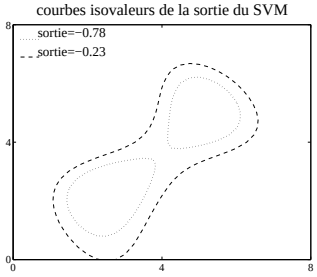

FIG. 2 – courbes isovaleurs de la sortie du SVM

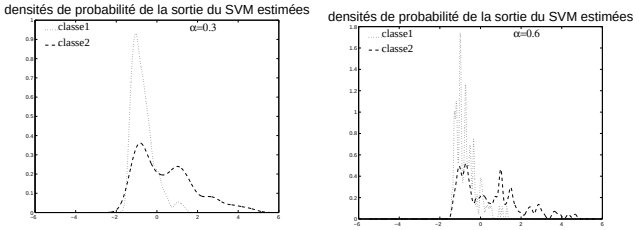

FIG. 3 – densités de probabilité conditionnelles des sorties du SVM pour $\alpha = 0,3$ et $\alpha = 0,6$

dans le cas de $\alpha = 0.3$ et $\alpha = 0.6$.

Pour chaque ensemble d'apprentissage, le SVM est construit et une règle de décision est définie selon (14). Les performances de chaque règle de décision sont estimées en mesurant les grandeurs $\widehat{P}(D_1/C_2)$, $\widehat{P}(D_1/C_2)$ et $\widehat{P_R}$. Pour cette estimation, plutôt que d'utiliser un ensemble test, nous avons utilisé les densités de probabilité théoriques de chacune des deux classes : le plan a été fragmenté en régions égales suivant une grille, la probabilité conditionnelle associée à chaque région a été calculée à partir des densités théoriques, puis la décision pour chaque région a été déterminée et pondérée par les probabilités conditionnelles pour obtenir les grandeurs estimées. L'opération répétée pour les 50 ensembles d'apprentissage a permis d'estimer la moyenne et l'écart-type des grandeurs mesurées. Les résultats sont représentés sur la figure 4.

Cette figure montre que la valeur de α doit être judicieusement choisie : moins le lissage est important, plus l'écart-type est grand, mais un lissage trop important déforme la densité de probabilité. L'estimateur de Parzen introduit un produit de convolution qui a pour effet d'étaler la densité de probabilité. Pour un seuil donné, il existe donc un biais entre la probabilité théorique et la probabilité obtenue par estimateur de Parzen.

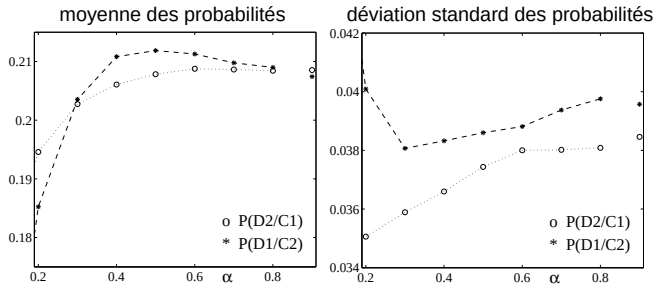

FIG. 4 – Moyenne et écart-type des probabilités $P(D_1/C_2)$ et $P(D_2/C_1)$. L'axe des abscisses indique la valeur de utilisée pour estimer les densités, sauf pour la dernière valeur, qui correspond à α très grand.

En prenant $\alpha = 0.4$ la valeur de l'écart-type est proche de la valeur théorique.

7 Conclusion

Le problème porte sur la classification avec contraintes. Dans le cas particulier de deux classes et où deux contraintes sont à satisfaire (probabilités d'erreur conditionnelles à chacune des classes bornées), la règle de décision théorique consiste à comparer le rapport de vraisemblance à un ou deux seuils selon que les contraintes peuvent être satisfaites sans ou avec rejet.

Une solution pour l'élaboration d'une règle de décision pour ce problème lorsque le processus est uniquement décrit à l'aide d'une base d'exemples est proposée. Elle est basée sur l'utilisation des SVM. Les densités de probabilité conditionnelles des sorties du SVM sont estimées à partir de la base d'apprentissage, à l'aide d'un estimateur de Parzen. Les résultats expérimentaux montrent qu'un lissage adéquat des densités estimées permet de limiter la variance des probabilités d'erreur conditionnelles intervenant dans les contraintes. Ce principe permet d'améliorer la robustesse de la règle de décision apprise.

Le problème de classification traité dans ce papier constitue un cas particulier. Les méthodes peuvent être généralisées à d'autres types de contraintes et à plus de deux classes.

Références

[1] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Boston, 1990.

[2] C.K. Chow, *On Optimum Reject Error and Reject Tradeoff*, IEEE Transactions on Information Theory, Vol. IT-16, N°.1, pp. 41-46, January 1970.

[3] B. Dubuisson and M. Masson, *A statistical decision rule with incomplete knowledge about classes*, Pattern recognition, Vol. 26, N°.1, pp. 155-165, 1993.

[4] G. Fukmera, F. Roli and G. Giacinto, *Reject option with multiple thresholds*, Pattern recognition, Vol. 33, N°.12, pp. 2099-2101, 2000.

[5] C. Cortes and V. Vapnik. *Support Vector Networks*. Machine Learning, 20 :273-279, 1995.

[6] C. J. C. Burges. *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 2(2) :121-167, 1998.

[7] T. Hastie, S. Rosset, R. Tibshirani and J. Zhu *The Entire Regularization Path for the Support Vector Machine.*, Journal of Machine Learning Research, 5 :131-159, 2004.

[8] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukhrjee. *Choosing kernel parameters for support vector machines.* Machine Learning, 46 :131 :159, 2002.