World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:12, 2007

A Kernel Based Rejection Method for Supervised Classification

Abdenour Bounsiar, Edith Grall, and Pierre Beauseroy

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

Abstract**—** In this paper we are interested in classification problems with a performance constraint on error probability. In such problems if the constraint cannot be satisfied, then a rejection option is introduced. For binary labelled classification, a number of SVM based methods with rejection option have been proposed over the past few years. All of these methods use two thresholds on the SVM output. However, in previous works, we have shown on synthetic data that using thresholds on the output of the optimal SVM may lead to poor results for classification tasks with performance constraint. In this paper a new method for supervised classification with rejection option is proposed. It consists in two different classifiers jointly optimized to minimize the rejection probability subject to a given constraint on error rate. This method uses a new kernel based linear learning machine that we have recently presented. This learning machine is characterized by its simplicity and high training speed which makes the simultaneous optimization of the two classifiers computationally reasonable. The proposed classification method with rejection option is compared to a SVM based rejection method proposed in recent literature. Experiments show the superiority of the proposed method.

Keywords**—** rejection, Chow's rule, error-reject tradeoff, Support Vector Machine.

I. Introduction

THE aim of a classifier in pattern recognition is to optimize a performance criterion which can be error probability or any classification cost for example. In real world applications, there is usually a performance constraint inherent to the application in hand that must be considered in addition to the performance criterion to optimize. Performance constraints refer to constraints on error probabilities. In the pattern recognition community, performance constraints usually refer to a constraint on the false positive rate (false alarm rate) in Neyman-Pearson tests. However performance constraints can be more complex, they can combine different error probabilities, consider order constraints between different error probabilities or even constraints on the ratio between two different error probabilities, see [1][2]. Moreover, performance constraints have been considered in multi-class [3][4] and multi-label classification problems [5].

Manuscript received April 26, 2006.

Abdenour Bounsiar is with Institut Charles Delaunay, Équipe Modélisation et Sûreté des Systèmes, 12 rue marie curie, 10010, Troyes Cedex, France (phone: + (33) 325718450, fax: + (33) 325715699, email: abdenour.bounsiar@utt.fr)

Edith Grall is with Institut Charles Delaunay, Équipe Modélisation et Sûreté des Systèmes, 12 rue marie curie, 10010, Troyes Cedex, France (email: edith.grall@utt.fr)

Pierre Beauseroy is with Institut Charles Delaunay, Équipe Modélisation et Sûreté des Systèmes, 12 rue marie curie, 10010, Troyes Cedex, France (email: pierre.beauseroy@utt.fr)

When considering performance constraints on error probabilities in classification problems, if the constraints cannot be altogether satisfied, it is necessary to introduce a reject option in order to satisfy the classification constraints. The reject option consists of withholding decision of samples whose decision is less confident in order to reduce error probabilities so as to verifyi the constraints. In this paper we will consider the binary classification problem of designing a classifier with an error probability constraint lower than the error rate of the best classifier, i.e., the best one separating two classes. In this case, rejection must be introduced.

Allowing for the reject option is of great importance in practice, as for example, in the case of medical diagnoses where misclassifying a sick patient as healthy may have serious consequences. Nevertheless, since the publications of Chow on the error-reject tradeoff [6][7], this option has not received a great deal of attention up until now and is often ignored in statistical literature. Notable exceptions in the engineering literature are Fumera and Roli [8], Fumera, Roli and Giacinto [9], Golfarelli, Maio and Maltoni [10] and Hansen, Liisberg and Salomon [11].

Other works have considered rejection with Support Vector Machines (SVMs) [12][13][14][15][16]. All of the proposed rejection techniques use two thresholds on the output of the SVM classifier and produce a reject region delimited by two parallel hyperplanes in the feature space. The SVM classifier is generally the one giving minimum error.

However, it has been experimentally shown in [1] and [17] that using the output dynamic of the optimal SVM (giving minimum error) as decision thresholds for any other classification task (classification problems with constraints on error probabilities) instead of minimizing total error, gives poor results. The authors propose not only to optimize the bias of the SVM but also all the other parameters (kernel parameters and training costs) in order to obtain high performances. Hence in order to design a good classification method with rejection option, one may use two different SVMs jointly optimized to obtain the best rejection. However tuning the parameters of the two SVMs (two training costs, kernel parameters and the bias for each SVM) together is computationally highly expensive.

In this paper we propose using a kernel based linear learning machine recently presented in literature [18]. This machine is fast training and hence easier to optimize. The proposed rejection method based on this learning machine is compared to a SVM based rejection method proposed by Tortorella in [16]. The comparison is based on the error-reject curve. Results of experiment on synthetic data comparing the proposed method to the SVM based method show encouraging

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

results.

In the next section, a fast overview on the error reject tradeoff is given. Assuming that the conditional density probabilities, and the a priori probabilities are known, or can be sufficiently well estimated, the results of Chow's theory can be directly applied. However in many real world problems where data are represented by sample sets, probability density functions cannot be well estimated. For such problems, discriminant function based classifiers such as SVMs are used. Section (III) is devoted to SVM based rejection methods; firstly, SVMs are introduced, thereafter, an overview on existing SVM based rejection methods is given and finally, the method to which the proposed method will be compared is reviewed in more details. In section (IV-B) the proposed rejection method is presented after the used learning machine has been developed. Experimental results and discussions are presented in section (V). Conclusions are given in section (VI).

II. Chow's theory on the error-reject tradeoff

Assuming symmetrical *0-1 loss function* for decision costs, a binary decision rule with reject option is optimum if for a given error rate (error probability) it minimizes the reject rate (reject probability). Chow [7] demonstrates that the optimum rule is to reject the pattern if the maximum of the a posteriori probabilities is less than a certain threshold. More explicitly, the optimum decision rule is to accept a pattern x for decision and to identify it as of the kth $(k = 1, 2)$ pattern class whenever

$$P_k p(x|\omega_k) \geq P_i p(x|\omega_i), \qquad \text{for all} \quad i = 1, 2,$$

or

$$P_k p(x|\omega_k) \geq (1-t) \sum_{i=1}^{2} P_i p(x|\omega_i),$$

and to reject the pattern whenever

$$\max_i [P_i p(x|\omega_i)] < (1-t) \sum_{i=1}^{2} P_i p(x|\omega_i),$$

where (P_1, P_2) are the a priori probabilities of the two classes ω_1 and ω_2 respectively, $p(x|\omega_i)$ is the conditional probability density function for x given the ith class and t is a constant between 0 and 1/2 $(0 < t \leq 1/2)$.

Now let $m(x)$ denote the maximum of the a posteriori probabilities of the classes given the pattern x

$$m(x) = \max_i [p(\omega_i|x)] = \frac{\max_i [P_i p(x|\omega_i)]}{P(x)}$$

The optimal decision rule can then be reformulated so as to reject the pattern x whenever

$$m(x) < 1 - t, \tag{1}$$

or accept the pattern otherwise.

When $P_1 p(x|\omega_1) > P_2 p(x|\omega_2)$ then (1) implies

$$\begin{aligned} \frac{P_1 p(x|\omega_1)}{P(x)} &< 1 - t \Leftrightarrow \\ P_1 p(x|\omega_1) &< (1-t)\left[P_1 p(x|\omega_1) + P_2 p(x|\omega_2)\right] \Leftrightarrow \\ \frac{p(x|\omega_1)}{p(x|\omega_2)} &< \frac{P_2}{P_1}\frac{1-t}{t} \end{aligned}$$

Similarly, if $P_2 p(x|\omega_2) > P_1 p(x|\omega_1)$ then (1) implies

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P_2}{P_1}\frac{t}{1-t}$$

This decision rule can be formulated so as to decide

$$\begin{cases} \omega_1 \text{ if } \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P_2}{P_1}\lambda \\ \omega_2 \text{ if } \frac{p(x|\omega_1)}{p(x|\omega_2)} < \frac{P_2}{P_1}\lambda^{-1} \\ \text{reject pattern } x \text{ otherwise,} \end{cases}$$

with $\lambda = (1-t)/t > 1$.

So the optimal decision rule with reject option, consists of a likelihood ratio based decision rule with two inversely symmetrical thresholds with respect to Bayes threshold P_2/P_1.

In classification problems described by sample sets, probability density functions of data are usually not known or difficult to estimate and the results of Chow's theory cannot be applied. To deal with such problems, a number of rejection methods based on classifiers with discriminant functions such as SVMs have been developed. Such rejection methods involve a labelled training data $\{x_i, y_i\}$, $i = 1, ..., l, x_i \in \mathbb{R}^d$ and y_i is the x_i pattern label taking value -1 or +1. In such considerations we define positive class , say ω_1 containing positive labelled patterns $\{x_i, y_i = +1\}$ and negative class , say ω_2 containing negative labelled patterns $\{x_i, y_i = -1\}$.

III. SVM based rejection methods for supervised binary classification problems

Because almost all the rejection solutions that have been recently proposed in literature are SVM based, and the fact that SVMs have gained considerable popularity in the machine learning community in recent years, this section will be devoted to SVM based rejection methods. Starting with a short review on SVMs, a brief presentation of different existing SVM based rejection methods is given. Special attention is given to a method proposed by Tortorella that has been experimentally shown in [16] to be successful. This method will be the subject of more details in section III-B and will be compared to the method proposed in section IV-B.

A. Support Vector Machines

A practical application of the principle of *Structural Risk Minimization* (SRM) [19] to the problem of pattern recognition leads to the definition of *Support Vector Machines* (SVM). Support Vector Machines map a pattern $x \in \mathbb{R}^d$ into a high (possibly infinite) dimensional space and construct an optimal separating hyperplane in this space [20]. The mapping $\phi(.)$ is implicitly introduced in the decision function via a dot product of the data in that space, that is performed by a kernel function $K(.,.)$ so that $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. The kernels that have these properties satisfy the Mercer conditions [19], i.e. for any $g(x)$ with finite L_2 norm (2), equation (3) must hold. Any positif definite kernel satisfies this condition [21].

$$\int_{-\infty}^{+\infty} g^2(x)dx < \infty \tag{2}$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K(u,v)g(u)g(v)dudv > 0. \quad (3)$$

Here we consider a kernel K_θ depending on a set of parameters θ. The decision function given by a SVM is thus:

$$\begin{aligned} f_\theta(x) &= \text{sign}\Big(\langle w_\theta^T, \phi_\theta(x)\rangle + b\Big) \\ &= \text{sign}\Big(\sum_{i=1}^{l} \alpha_i^0 y_i K_\theta(x_i, x) + b\Big), \end{aligned}$$

where w_θ and b are referred to as *weight vector* and *bias* respectively, and

$$w_\theta = \sum_{i=1}^{l} \alpha_i^0 y_i \phi(x_i)$$

The coefficients α_i^0 are obtained by maximizing the following functional [20] [22]:

$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i\alpha_j y_i y_j K_\theta(x_i, x_j)$,
subject to
$\sum_{i=1}^{l} \alpha_i y_i = 0$ and $\alpha_i \geq 0, \quad i = 1, ..., l.$

The coefficients α_i^0 define the optimal hyperplane with the maximal distance (in the high dimensional space) to the closer image $\phi_\theta(x_i)$ from the training data, called the *maximal margin*. For the non-separable case, one need to allow for training errors which results in the so called *soft margin SVM* [23], in which the coefficients α_i^0 are obtained by maximizing the same functional [22]:

$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i\alpha_j y_i y_j K_\theta(x_i, x_j)$,
subject to
$\sum_{i=1}^{l} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C, \quad i = 1, ..., l,$

where C is the training cost penalizing the training errors, and will be considered as just another parameter of the SVM:

$$f(x) = \text{sign}\Big(\sum_{i=1}^{l} \alpha_i^0 y_i K_\theta(x_i, x) + b\Big), 0 \leq \alpha_i^0 \leq C. \quad (4)$$

All of the SVM based rejection techniques that have been proposed in literature produce a reject region delimited by two parallel hyperplanes in the feature space. In [12], Fumera and Roli developed a maximum margin classifier with reject option, i.e. a SVM whose rejection region is determined during the training phase. As a result, the rejection region provided by their algorithm is delimited by a pair of parallel hyperplanes whose positions and orientations depend on the rejection cost. A different classification method with reject option is presented in [13], where the reject decision is made for the samples near the optimal hyperplane for which the classifier may have not confidence in the class label. The authors introduced confidence levels on the SVM output dynamic, d, which provides the signed distance of the sample from the optimal hyperplane. This allows rejection of samples below a certain value of $|d|$. Another possible way to establish a rejection rule for the SVM is to use the results of Chow's theory on the error-reject tradeoff [7] and apply them to the posterior class probabilities estimated using the SVM outputs [14][15]. Another rejection method with SVMs was proposed recently by Tortorella. This method is compared to the proposed method which is presented in section IV-B and will be detailed in the next section.

B. SVM based reject method with two independent thresholds

In [16], Tortorella introduces a cost-sensitive reject rule for SVM classifiers which is able to minimize the expected cost of classification, defined on the basis of correct classification and on the reject and error costs particular to the application. The proposed approach is based on the *Receiver Operating Characteristic* (ROC) curve, and defines two different reject thresholds for the two classes. This way the reject region is defined by two independent hyperplanes parallel to the optimal separating hyperplane. The thresholds depend on the costs defined for the application, but it is not necessary to retrain the SVM when the costs change. The author showed experimentally using toy and real data that with two independent hyperplanes he obtained better rejection than with symmetrical hyperplanes. Note that because the class of the classifiers engendered by this method includes the class of classifiers with symmetrical hyperplanes, the supremacy of this method is not surprising.

If we consider that the optimal separating hyperplane between two different classes is

$$f_{(\theta,b)}(x) = \langle w_\theta, \phi_\theta(x)\rangle + b,$$

then the decision rule with rejection option provided by this method, can be formulated so as to decide

$$\begin{cases} \omega_1 \text{ if } f_{(\theta,b_1)}(x) = \langle w_\theta, \phi_\theta(x)\rangle + b_1 > 0 \\ \omega_2 \text{ if } f_{(\theta,b_2)}(x) = \langle w_\theta, \phi_\theta(x)\rangle + b_2 < 0 \\ \text{reject pattern } x \text{ otherwise,} \end{cases} \quad (5)$$

with $b_2 > b_1$. In this paper, this method is compared to the proposed rejection method which is presented in the next section. Note that as explained above, the method as presented in [16] uses ROC curves in order to obtain a decision rule with rejection option that minimizes the cost of classification. However, in this paper, the aim of the introduction of a reject option is to reduce error probability in order to respect a constraint on the latter using a minimum of rejection. Hence ROC curves are not used to determine the biases b_1 and b_2, but an optimization method is.

IV. Proposed classification method with rejection option

A. Description of the used learning method

To construct a decision rule with reject option, a kernel based linear learning machine proposed in [18] is used. Performances of this machine on standard data bases are compared to those of other state of the art paradigms, such as Support Vector Machines and Kernel Fisher Discriminant. This machine has several interesting characteristics such as simplicity, high training speed and good performance. The principle of this learning machine will now be presented.

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

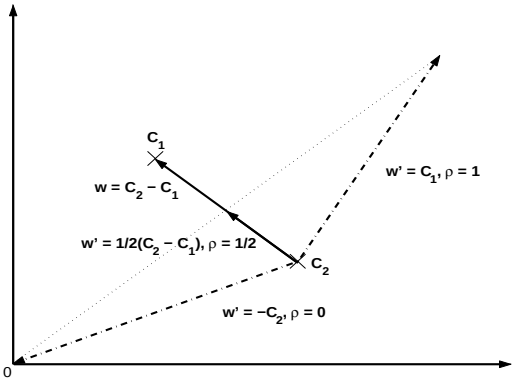


Fig. 1. Varying the variable ρ from 0 to 1 in (10) makes the weight vector varying from $-c_2$ to c_1.

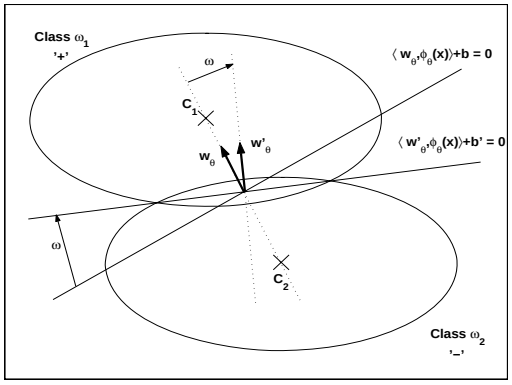


Fig. 2. The best separating hyperplane may be obtained by rotation of the one of (6).

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

In [24], authors have exposed a simple linear classifier. The basic idea is to assign a new pattern to the class with the closest mean. The means of the two classes are estimated from training samples, they are denoted $c_1 = \sum_{\{x_i \in \omega_1\}} \frac{x_i}{m_1}$ and $c_2 = \sum_{\{x_i \in \omega_2\}} \frac{x_i}{m_2}$ for classes with positive and negative labelled samples respectively, where m_1 and m_2 are the number of positive and negative labelled training patterns. Half way between c_1 and c_2 lies the point $c = (c_1+c_2)/2$. The class of an input x is determined by comparing the absolute angle between the vector $x - c$ and the vector $c_1 - c_2$ to $\pi/2$. This leads to the following analytical formulation of the decision function

$$\begin{aligned} f(x) &= \text{sign}\left(\langle x - c, c_1 - c_2\rangle\right) \\ &= \text{sign}\Big(\sum_{\{x_i \in \omega_1\}} \frac{\langle x, x_i\rangle}{m_1} - \sum_{\{x_i \in \omega_2\}} \frac{\langle x, x_i\rangle}{m_2} + b\Big), \end{aligned}$$

with $b = \frac{1}{2}(\|c_2\|^2 - \|c_1\|^2)$. Note that other values of the bias may lead to better performances.

In general, real world applications require discriminant functions that are more complex than linear ones. Kernel representations such as defined in section (III-A) offer a solution by projecting the data from X into a high dimensional *feature space* $F = \{\phi(x)|x \in X\}$. Using such kernels, the decision rule of the previous classifier can be expressed as

$$\sum_{\{x_i \in \omega_1\}} \frac{K_\theta(x, x_i)}{m_1} - \sum_{\{x_i \in \omega_2\}} \frac{K_\theta(x, x_i)}{m_2} + b \underset{D_2}{\overset{D_1}{\gtrless}} 0, \quad (6)$$

where D_1 and D_2 are the decisions to assign a pattern to the classes ω_1 and ω_2, respectively.

Assuming that K_θ is a probability density i.e., it is positive and has a unit integral: $\int_X K_\theta(x, y)dx = 1$ for all $y \in X$. Assuming also that the conditional probability density of each of the two classes is estimated by the *Parzen windows* estimator:

$$\hat{p}_1(x) \sim \sum_{\{x_i \in \omega_1\}} \frac{K_\theta(x, x_i)}{m_1}, \hat{p}_2(x) \sim \sum_{\{x_i \in \omega_2\}} \frac{K_\theta(x, x_i)}{m_2}. \quad (7)$$

In such conditions, if $b = 0$, (6) becomes an estimated likelihood ratio based classifier

$$\frac{\hat{p}_2(x)}{\hat{p}_1(x)} \underset{D_1}{\overset{D_2}{\gtrless}} \lambda, \quad (8)$$

with decision threshold $\lambda = 1$. By varying λ in $[0, +\infty[$, classifier (8) may cover a large scope of likelihood ratio based decision rules (Bayes rule, Neyman-Pearson test, Mini-Max test).

Using a parameter $\rho \in [0, 1[$, decision rule (8) with $\lambda \in [0, +\infty[$ can be reformulated as

$$\frac{\hat{p}_2(x)}{\hat{p}_1(x)} \underset{D_1}{\overset{D_2}{\gtrless}} \frac{\rho}{1-\rho}, \quad (9)$$

which is equivalent to

$$\rho\hat{p}_1(x) - (1-\rho)\hat{p}_2(x) \underset{D_2}{\overset{D_1}{\gtrless}} 0.$$

This corresponds to a linear classifier in the feature space without bias ($b = 0$) and the following weight vector

$$\begin{aligned} w_\theta &= \rho \sum_{\{x_i \in \omega_1\}} \frac{\phi_\theta(x_i)}{m_1} - (1-\rho) \sum_{\{x_i \in \omega_2\}} \frac{\phi_\theta(x_i)}{m_2} \\ &= \rho C_1 - (1-\rho)C_2, \quad (10) \end{aligned}$$

where C_1 and C_2 are the class means in the feature space.

Generally for linear classifiers, the bias is one of the classifier parameters that must be optimized jointly with the other classifier parameters in order to get a reliable classifier. Considering a bias in (10) leads to

$$\rho\hat{p}_1(x) - (1-\rho)\hat{p}_2(x) + b \underset{D_2}{\overset{D_1}{\gtrless}} 0 \Leftrightarrow \quad (11)$$

$$\frac{\hat{p}_2(x)}{\hat{p}_1(x)} \underset{D_1}{\overset{D_2}{\lessgtr}} \frac{\rho}{1-\rho} + \frac{b}{(1-\rho)\hat{p}_1(x)} \Leftrightarrow$$

$$\frac{\hat{p}_2(x)}{\hat{p}_1(x)} \underset{D_1}{\overset{D_2}{\lessgtr}} \frac{\rho}{1-\rho} + \delta(x). \quad (12)$$

Equation (12) is a likelihood ratio based decision rule, where the probability densities are estimated by *Parzen windows* estimators. The decision threshold consists of a constant term that is defined by the parameter ρ, and a variable term $\delta(x)$ that depends on ρ, b and the pattern under consideration .

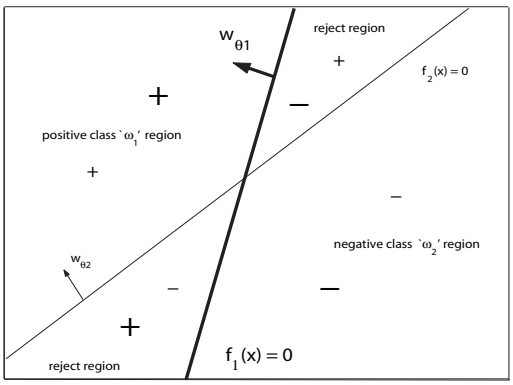


Fig. 3. An illustration of decision regions in feature space for the proposed method.

The obtained result show that $\delta(x)$ can be considered as a correction term to the decision rule (8). Once the probability density functions are estimated by choosing the convenient kernel function K_θ, the decision threshold and the bias b are jointly chosen to give the best average correction to the decision threshold in order to improve performances.

Note that because the estimation error of the likelihood ratio $\hat{p}_2(x)/\hat{p}_1(x)$ is a function of the pattern x, correction of the decision threshold is also a function of the pattern x.

Decision rule (11) can also be reformulated as

$$\begin{array}{ll} \dfrac{\hat{p}(\mathbf{x}|\omega_2)}{\hat{p}(\mathbf{x}|\omega_1)} + \delta(\mathbf{x}) & \underset{D_1}{\overset{D_2}{\gtrless}} \quad \dfrac{\rho}{1-\rho} \Leftrightarrow \\ \dfrac{\hat{p}(\mathbf{x}|\omega_2) - b/(1-\rho)}{\hat{p}(\mathbf{x}|\omega_1)} & \underset{D_1}{\overset{D_2}{\gtrless}} \quad \dfrac{\rho}{1-\rho}. \end{array}$$

With such a formulation, the bias appears to modify the discriminant function rather than the decision threshold. It can be interpreted as an offset of the estimated probability density function $\hat{p}(\mathbf{x}|\omega_2)$, which gives a correction to the estimated likelihood ratio. For different patterns x, this correction is not the same. Considering this formulation or that of (12), the bias b can be considered as a correction term of the decision rule (9) estimation errors.

The family of classifiers defined by (6) is included in the proposed more general family of classifiers defined by (11). Disregarding the problem of parameter estimation, this family of classifiers achieves better performances.

A geometric interpretation of this classifier is given as the following: the estimated hyperplane separating the two classes is perpendicular to the weight vector w_θ whose direction varies from $-C_2$ to C_1 when varying the value of the parameter ρ from 0 to 1, as illustrated by figure 1. The position of this hyperplane in w_θ's direction is set by the value of the bias b. Figure 2 shows on a two dimensional case, the influence of parameter ρ on the determination of the separating hyperplane. The hyperplane $\langle w_\theta, \phi_\theta(x)\rangle + b = 0$ given by (6) gives a clearly poor separation. In this case, the best separating hyperplane $\langle w'_\theta, \phi_\theta(x)\rangle + b' = 0$ in the sense of minimum error is a member of the family of hyperplanes given by (11), which can be obtained by rotation of the first hyperplane (rotation angle ω in figure 2) and a suitable value of b. Note that for bi-dimensional classification problems such as the one of figure 2, the weight vector of the best separating hyperplane is obviously in the plane defined by (C_1, C_2) except if these two vectors are collinear. This is not necessarily true for higher dimensional problems, in such cases the best separating hyperplane may not be a member of the family of hyperplanes defined by (11).

A special attention to the choice of the kernel function K_θ is necessary to verify that the two means C_1 and C_2 are not collinear. If they are, the rotation of the weight vector w_θ in (10) would not be possible.

In the case of a RBF kernel

$$K_\sigma(x, y) = \exp\left(-\sigma\|x - y\|^2\right), \tag{13}$$

since $K_\sigma(x, x)$ = 1 for all patterns x, all the vectors $\phi_\sigma(x_i)$ in the feature space are located on a unit radius hyper-sphere: $\|\phi_\sigma(x_i)\| = 1, \forall i$. Furthermore, for all patterns x and y: $0 < \widehat{\phi_\sigma(x)\phi_\sigma(y)} < +\pi/2$ because $0 < K_\sigma(x, y) < 1$. Thus all data in the feature space are located on a surface bounded by a solid angle of $\pi/2$. So, the two means can be collinear only in the case where $C_1 = C_2$, a situation that is extremely improbable.

B. Proposed decision rule with rejection option

The proposed rejection rule using the proposed learning machine, consists of two different classifiers belonging to the family of classifiers (11), say $f_1(x) = \langle w_{(\theta_1,\rho_1)}, \phi_{\theta_1}(x)\rangle + b_1$ and $f_2(x) = \langle w_{(\theta_2,\rho_2)}, \phi_{\theta_2}(x)\rangle + b_2$, combined to obtain four regions of decision in the feature space corresponding to only three regions of decision in the input space: one for each class and a region of rejection. The corresponding hyperplanes are not necessarily parallel. Each classifier defines two regions on the feature space; one region corresponds to positive evaluated patterns ($f(x) \geq 0$) and the other to negative evaluated patterns ($f(x) < 0$). Pattern x is rejected if the decision of the two classifiers on it are incompatible. Thus the decision rule for a pattern x is defined as (see figure 3):

$$\begin{cases} \text{decide } \omega_1 \text{ if } f_1(x) \geq 0 \text{ and } f_2(x) \geq 0 \\ \text{decide } \omega_2 \text{ if } f_1(x) < 0 \text{ and } f_2(x) < 0 \\ \text{reject pattern } x \text{ if } f_1(x) \geq 0 \text{ and } f_2(x) < 0 \\ \text{reject pattern } x \text{ if } f_1(x) < 0 \text{ and } f_2(x) \geq 0 \end{cases} \tag{14}$$

or

$$\begin{cases} \text{decide } \omega_1 \text{ if } f_1(x) \geq 0 \text{ and } f_2(x) \geq 0 \\ \text{decide } \omega_2 \text{ if } f_1(x) < 0 \text{ and } f_2(x) < 0 \\ \text{reject pattern } x \text{ otherwise,} \end{cases} \tag{15}$$

Note that since the parameters of the kernel function are not necessarily the same in $f_1(x)$ and $f_2(x)$, the two separating hyperplanes presented in figure (3) may not be in the same feature space. Hence the four regions as defined in figure (3) may not exist in any feature space; figure (3) is just an illustration of the decision rule (14). However, returning to the input space $\mathcal{X}$ these four regions can be defined and are easily identified.

The optimal rejection rule as defined in (15) is obtained by optimizing the parameters $\rho_1, \rho_2, \theta_1, \theta_2, b_1$ and b_2 all together

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

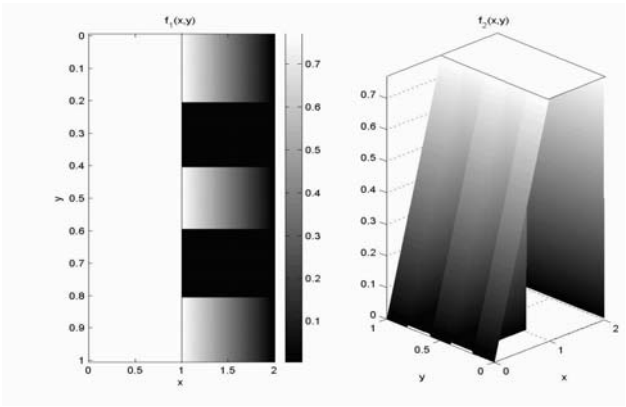


Fig. 4. Probability Density Functions of the synthetic data.

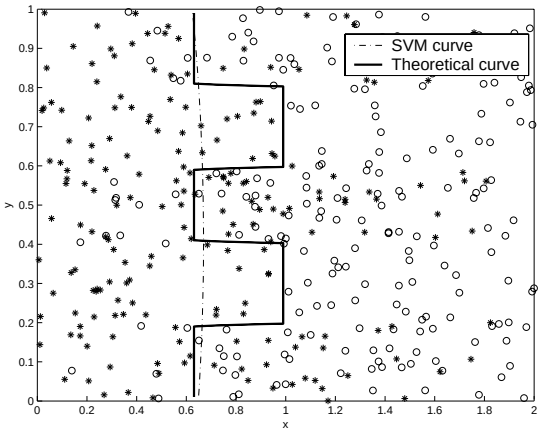


Fig. 6. Partitions corresponding to FPR= 0.1.

in order to minimize reject probability subject to an error probability constraint. The simultaneous optimization of six parameters may be computationally very heavy for other learning machines such as SVM. In the case of SVMs, for each combination of SVM parameters, the optimization of the l Lagrange coefficients α_0^i (4) is necessary. In the case of the proposed learning machine, the weight vector is immediately determined without any complex calculus or optimization step (10). This makes the optimization of the six parameters a computationally reasonable task.

V. Experimental results

A good method to compare two classifiers with a rejection option, is to perform the comparison for different values of the constraint on error rate, and to plot the reject-error curve for each of them. In order to compare the proposed classification method with a rejection option to the one with SVM, two series of tests on synthetic data were performed; the first one using 50 training data sets of 50 samples and the second one using larger 50 training data sets of 200 samples. The data were drawn from the distribution used in [1] and [17]. It consists in two equiprobable and mirrored probability density functions (pdfs) with respect to the axe $x = 1$ and took the form of letter 'E' (see figure 4). The trunk of the letter is uniformly distributed and the distribution of the three branches are uniform according to variable y and linearly decreasing according to x. The left shape in figure 4 represents a top view of the pdf of one class (consider the class ω_1) and the right shape represents a lateral view of the pdf of the second class (consider the class ω_2). This specific distribution was introduced in [1] and [17] due to the difficulty to estimate properly the corresponding decision function using drawn data. Because the two pdfs are symmetrical, the theoretical optimal separation (Bayes classifier) between the two classes is obtained for the plane $x = 1$. An optimal separating SVM may approach the optimal theoretical solution. However the optimal separation for any Neyman-Pearson test is not linear but fits the shape of the three legs of the probability density function. Hence any solution of a Neyman-Pearson test using the dynamic of the output of an optimal separating SVM on this data will lead to poor solution [1][17]. The ER= f(FPR) (error rate (ER) in y-axis versus false positive rate (FPR) in x-axis) curves obtained on the data with theoretical results (solid line) and with optimal SVM (bold solid line) by varying only the bias, are represented on figure 5 [1][17]. This figure shows clearly that an optimal SVM can separate optimally the two classes, but performs badly in the case of Neyman-Pearson tests (regions on the ER= f(FPR) curve corresponding to non minimum error probability). For example the boundaries corresponding to FPR $= 0.1$ obtained with theoretical results and optimal SVM are depicted both on figure 6. It is clear that the solution obtained with the optimal SVM is very bad.

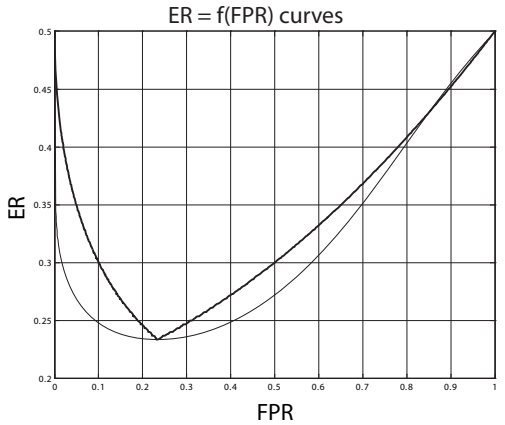


Fig. 5. ER= f(FPR) curves obtained with optimal SVM (bold solid line) and theoretical results (solid line).

In figure 5 the performances of classifiers are presented in ER= f(FPR) curves instead of the standard ROC (Receiver Operating Characteristic) curves which are defined as plots of true positive rate (TPR) as the ordinate versus false positive rate (FPR) as the abscissa. In fact, ROC analysis does not directly commit to any particular measure of performance. This is sometimes considered as an advantageous feature of ROC curves. For example, Van Rijsbergen [25] quotes Swets [26] who argues that this is useful as it measures "discrimination

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

power independent of any 'acceptable criterion' employed". But, being independent of any particular performance measure can be a disadvantage when one has a particular performance measure in mind. ROC curves do not visually depict the quantitative performance of a classifier or the difference in performance between two classifiers [27]. This is why the ER versus FPR curves have been adopted in [1] and [17]. Such curves enable to appreciate the quantitative difference between classifiers in term of Error Rate subject to a given False Positive Rate.

A. Tests on classes of 50 samples

The first series of tests was achieved using 50 training sets of 50 samples per class. On each data set we have trained the two rejection methods. For 10 predefined values of error probability, the parameters of the two methods where optimized in order to give the minimum reject probability subject to each of these 10 values of error probability. The optimization of the parameters was done using validation error and reject probabilities obtained using real probability density functions on the decision areas. This gives unbiased estimates of validation error and reject probabilities. Then, the error-reject curve for each set is constructed using the 10 obtained couples of reject and error probabilities plus the trivial point (reject = 1, error = 0): by using interpolation, 10 values of reject probability corresponding to the 10 prefixed values of error rates were picked up. The obtained mean Error-Reject curves with standard-deviation bars on the rejection values for the two classification methods are depicted on figure 7.

Note that because, the validation error and reject probabilities are estimated using the true probability density functions on the decision areas, the Error-Reject curves of the two classification methods that are represented in figure 7, are the best ones we can obtain.

B. Tests on classes of 200 samples

The second series of tests was achieved using 50 training sets of 200 samples. Similarly to the first series of test, the mean Error-Reject curves with standard-deviation bars on the rejection values obtained by using the same experimental procedure than with the fist series for the two classification methods, are depicted on figure 8.

C. Discussion of experimental results

The comparison between the proposed rejection method and the SVM based one that was presented in this paper, was performed on the error-reject curve. Any decision rule minimizing reject probability subject to a given error rate, can be formulated as to decide

$$\begin{cases} \omega_1 \text{ if } f(x) > t_1 \\ \omega_2 \text{ if } f(x) < t_2 \\ \text{reject pattern } x \text{ otherwise,} \end{cases} \quad (16)$$

where $f(x)$ is any discriminant function and thresholds t_1 and t_2 are real quantities that may be dependent or not. If we suppose that for a given problem we obtain the validation

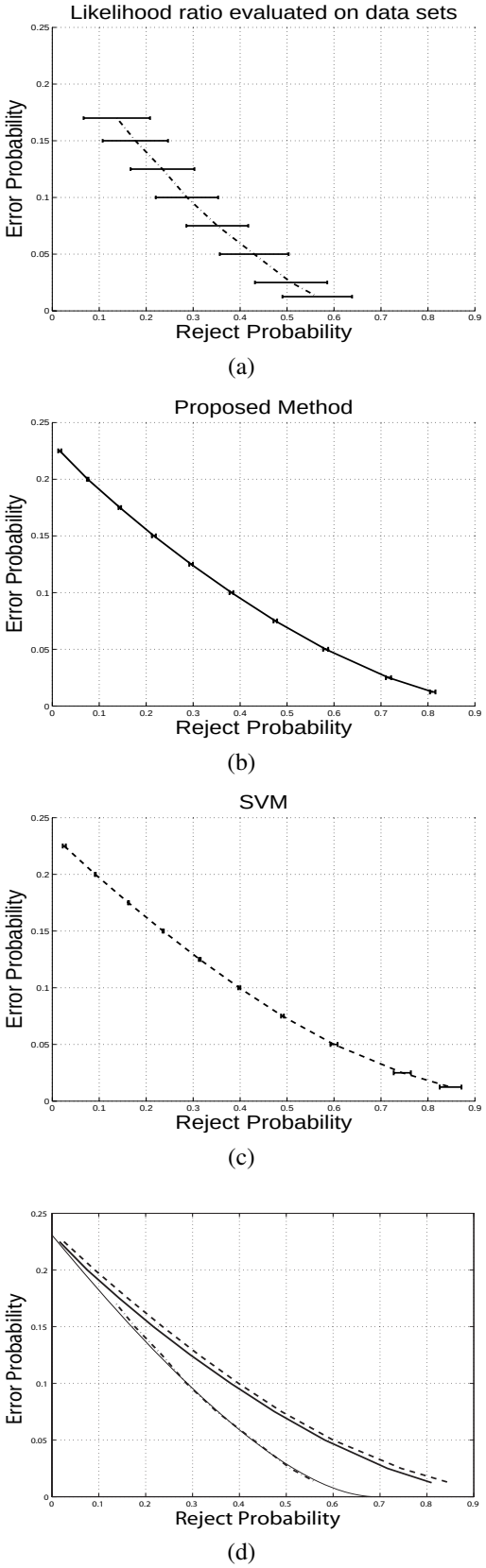


Fig. 7. Error-Reject curves (mean and standard-deviation bars) evaluated on 50 sets of 50 samples (a) using the true likelihood ratio evaluated o data sets, (b)using the proposed method, (c) using the SVM based method. In (d) the previous curves and the theoretical error-reject curve is also represented (solid), are represented all together for comparison

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

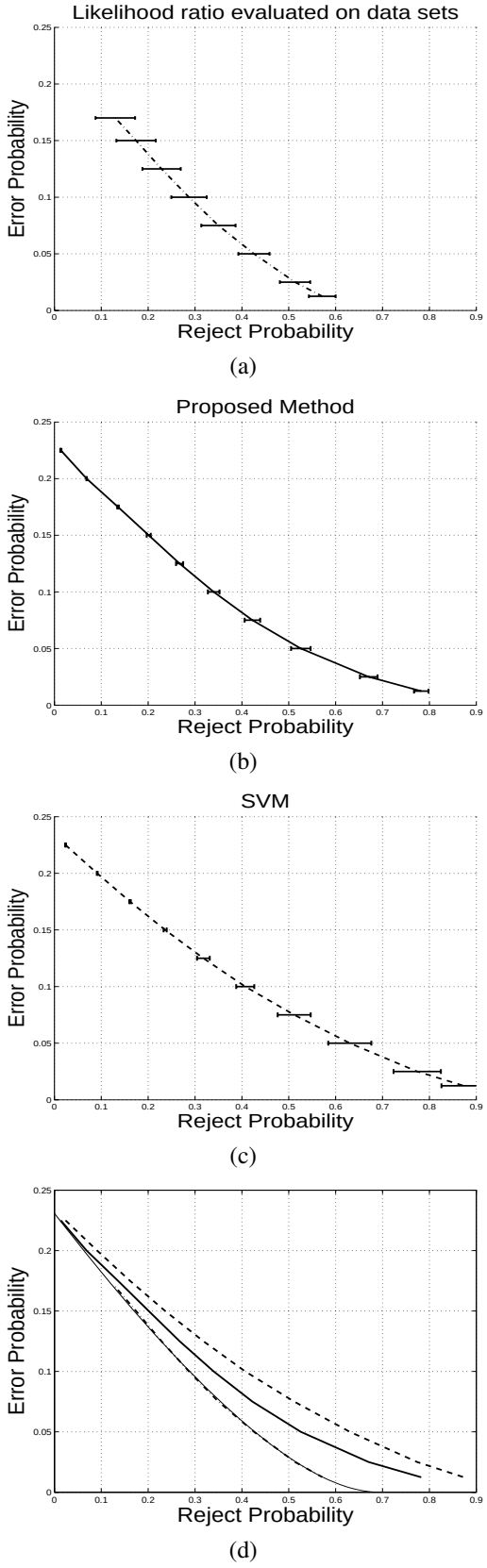


Fig. 8. Error-Reject curves (mean and standard-deviation bars) evaluated on 50 sets of 200 samples (a) using the true likelihood ratio evaluated o data sets, (b)using the proposed method, (c) using the SVM based method. In (d) the previous curves and the theoretical error-reject curve is also represented (solid), are represented all together for comparison

errors $P(D_1|\omega_2) = \text{FPR}^*$ and $P(D_2|\omega_1) = \text{FNR}^*$ (False Negative Rate), with $(\text{FPR}^*, \text{FNR}^*) \in [0,1]^2$ and where D_1 is to decide ω_1 if $f(x) \geq t_1$ and D_2 is to decide ω_2 if $f(x) < t_2$, then in such conditions, the initial problem can be seen as the problem of designing a binary classifier verifying

$$P(D_1|\omega_2) = \text{FPR}^* \quad \text{and} \quad P(D_2|\omega_1) = \text{FNR}^*, \tag{17}$$

for which the solution is the intersection of the following two Neyman-Pearson tests [1]:

$$\begin{aligned} f(x) \underset{D_2}{\overset{D_1}{\gtrless}} t_1 \text{ with } \int_{Z_1} P(x|\omega_2)dx = \text{FPR}^*, \\ f(x) \underset{D_2}{\overset{D_1}{\gtrless}} t_2 \text{ with } \int_{Z_2} P(x|\omega_1)dx = \text{FNR}^*. \end{aligned} \tag{18}$$

The decision areas Z_1 and Z_2 are determined by the first and the second tests (18) respectively [28]. It has been shown in [1] that in the case where $t_1 > t_2$ then the decision rule of the classifier verifying (17) is the one of (16). So the use of the output dynamic of the optimal SVM must be not successful for problems consisting of minimizing reject probability subject to a given error rate, since such practice has been shown to be not successful for Neyman-Pearson tests [1][17].

On previous works [1][17] we have proposed a SVM based solution for problems of Neyman-Pearson type. It consists of tuning all SVM parameters to find a set of parameters giving the best solution. However for a rejection technique we may need two SVMs and tuning the parameters of the two SVMs together is computationally highly expensive. This is why the training machine presented in section (IV-A) have been used. This machine is fast training and hence convenient with optimization algorithms.

On figures (10), (11) and (9) are represented decision areas obtained respectively by SVM based method, the proposed method and the true likelihood ratio. These decision areas correspond to minimal reject probability given a constraint on error rate ER $= 9.4\%$. The decision areas obtained with the proposed method (figure 11) are closer to the optimal ones (figure 9) than those obtained with SVM based method (figure 10). For comparison, the optimal reject probability corresponding to the constraint ER $= 5\%$ is 42.5%, the average one obtained by SVM based method is 63.5% and the average one obtained using our method is 52.5%, this corresponds to an average amelioration of performance of 47.6% which is an encouraging result.

The inadequacy of the SVM based rejection method can be seen on figures (7) and (8) through the SVM curves especially in figure (8) with training sets of 200 samples/class. In figure (10) this inadequacy is clearly seen, the SVM rejection method with two thresholds is incapable to detect the structure of the training data, contrary to the proposed method which produces decision areas that follow this structure (figure 11).

On figures (7) and (8) the first plots represent the error-reject curves obtained with Chow's rule, where error and reject rates were evaluated on the training sets of 50 samples/class and 200 samples/class, respectively. The aim of these two plots is to show the variability or the dispersion existing between different training sets. Taking into account the standard-deviations of different values of the reject rate in these two plots on the

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf

one hand and the standard-deviations on the plots obtained with the SVM based method and the proposed one on the other hand, it seems that the two methods present good robustness. However, the fact that the size of the training sets increases don't bring any amelioration in term of mean and standard-deviation to the performance of the SVM based method, for the considered data. On the contrary, more the size of the training sets is larger, better is the average performance of the proposed method.

It seems that the average performance of the proposed method is asymptotically approaching the performance of the optimal Chow's rule based on likelihood ratio. Such result is not surprising if we take into consideration the fact that the used learning machine may have asymptotically the same performance as any likelihood based classifier (12). Indeed, the used learning machine uses *Parzen windows estimators* for the probability density functions of the two classes (7), such estimators are known to be asymptotically convergent to the true probability density functions.

VI. Conclusion

In this paper the problem of classification with a constraint on error probability has been considered. When the constraint is lower than the error probability obtained with the best classifier, an option of reject needs to be introduced. For supervised classification, SVM based rejection methods proposed in literature use one SVM with two thresholds on its output, the area between these two thresholds is the rejection area. Some use two symmetrical thresholds to the optimal separating hyperplane and others use non dependent thresholds. However having shown in previous works on synthetic data that using the output dynamic of optimal SVM for classification tasks with performance constraint may lead to poor results (especially for Neyman-Pearson tests), we have proposed a SVM based solution consisting in tuning all SVM parameters to find a set of parameters giving the best performance with respect to the constraint, a method that we have shown to be efficient. However for a rejection technique we need two SVMs and tuning the parameters of the two SVMs together is computationally very heavy because of the additional optimization of Lagrange coefficients α_0^i (4) for each set of SVM parameters.

In order to construct an efficient classification method with reject option, the learning machine presented in section (IV-A) has been used. This machine is fast training and hence convenient with optimization algorithms. The results on synthetic data obtained using the proposed classification method with rejection option are better than those obtained with the SVM based method, especially for large training sets. Indeed, the average performance of the proposed method improves with the increase of the number of training samples, contrary to the SVM based method for which there is no amelioration of performance due to the complexity of the data structure.

Any classifier minimizing classification errors is optimized to provide a separation that approximates the optimal separation between classes. However, there is no reason to use the output of such classifiers for classification tasks with performance constraints since these classifiers were not specifically

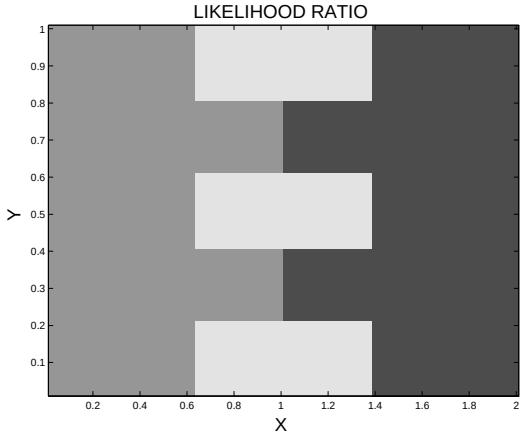


Fig. 9. Optimal decision areas corresponding to ER = 9.4% obtained with true likelihood ratio.

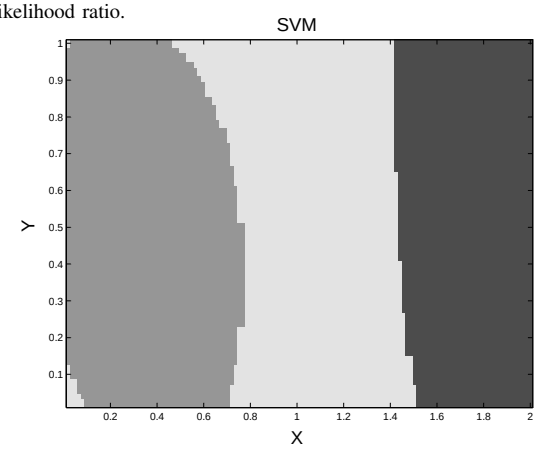


Fig. 10. Best decision areas over the 50 sets of 200 samples/class, corresponding to ER = 9.4% obtained with SVM based method.

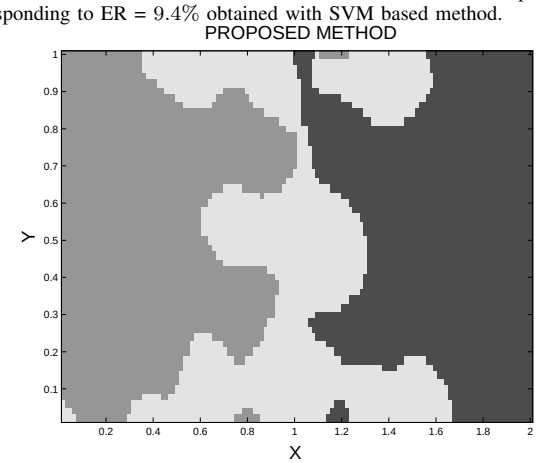


Fig. 11. Best decision areas over the 50 sets of 200 samples/class, corresponding to ER = 9.4% obtained with the proposed method.

optimized for. So, for any classification problem with or without performance constraints, the best solution is to optimize classifiers in order to provide separations that approximates the optimal ones for and only for the considered problem. The proposed method is based on this. Thus, the use of the output of optimal SVM for classification with rejection option is not justified and may lead to poor performances.

Note that because the aim of this article was to propose an alternative classification method for the bad one using the output dynamic of optimal SVM for classification problems with a rejection option, all results presented in this paper were based on validation error and reject probabilities evaluated on decision areas with real probability density functions. This resulted in classifiers with high robustness to the variations between training sets. However, testing the performance of the proposed method on real world problems, needs to use validation techniques on test sets. This will result in negative effects on the estimation of the proposed method parameters.

Future works will consider this problem and will try to find efficient methods for parameter selection that allow for good performance and robustness. Then, using the obtained results, this work will be extended to multi-category classification problems with a rejection option.

REFERENCES

[1] A. Bounsiar, P. Beauseroy, and E. Grall, "A straightforward svm approach for classification with constraints," in *proceedings of EUSIPCO'05*, Antalya, Turkey, September 2005.

[2] E. Grall, P. Beauseroy, and A. Bounsiar, "Classification avec contraintes : problématique et apprentissage d'une règle de décision par svm," in *Proceedings of GRETSI'05*. 2005, Louvain-la-Naeuve, Belgium.

[3] T. Ha, "The optimum class-selective rejection rule," *Transactions on Pattern Analysis ans Machine Intelligence*, vol. 19, no. 6, pp. 608–615, 1997.

[4] T. Horiuchi, "Class-selective rejection rule to minimize the maximum distance between selected classes," *Pattern recognition*, vol. 31, no. 10, pp. 579–588, 1998.

[5] E. Grall, P. Beauseroy, and A. Bounsiar, "Multilabel classification rule with performance constraints," To appear in proceedings of ICASSP'06. Toulouse, France, May 14-19 2006.

[6] C. K. Chow, "An optimum character recognition system using decision functions," *IEEE Trans. Electronic computers*, vol. EC-6, pp. 247–254, December 1957.

[7] C. K. Chow, "On optimum error and reject trade-off," *IEEE Transactions on Information Theory*, vol. 16, pp. 41–46, 1970.

[8] G. Fumera, F. Roli, and G. Giacinto, "Analysis of error-reject trade-off in linearly combined multiple classifiers," *Pattern Recognition*, vol. 33(12), pp. 2099–2101, 2000.

[9] G. Fumera and F. Roli, "Analysis of error-reject trade-off in linearly combined multiple classifiers," *Pattern Recognition*, vol. 37(6), pp. 1245–1265, 2004.

[10] M. Golfarelli, D. Maio, and D. Maltoni, "On the error-reject trade-off in biometric verification systems," *IEEE Transactions on Pattern analysis and Machine Intelligence*, vol. 19(7), pp. 789–796, 1997.

[11] L. K. Hansen, C. Lissberg, and P. Salomon, "The error-reject tradeoff," *Open systems and Information Dynamics*, vol. 4, pp. 159–185, 1997.

[12] G. Fumera and F. Roli, "Support vector machines with embedded rejection option," *In: Lee S, Verri A (eds) Pattern recognition with support vector machines. Lecture notes in computer science*, vol. 2388, pp. 68–82, Springer, Berlin Heidelberg New York, 2002.

[13] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J.P. Mesirov, and T. Poggio, "Support vector machine classification of microarray data," *AI Memo 1677*, Massachusetts Institute of Technology, 1999.

[14] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *In: Smola AJ, Bartlet PL, Schölkopf B, Schurmans D (eds) Advances in large margin classifiers*, pp. 61–74, MIT Press, 2004.

[15] J.T. Kwok, "Moderating the outputs of support vector machine classifiers," *IEEE Transactions on Neural Networks*, vol. 10, pp. 1018–1031, 1999.

[16] F. Tortorella, "Reducing the classifcation cost of support vector classifiers through an roc-based reject rule," *Pattern Anal. Applic.*, vol. 7, pp. 128–143, 2004.

[17] A. Bounsiar, E. Grall, and P. Beauseroy, "Using svm for binary classification with first type error constraint," in *proceedings of ICSIT'05*, Algiers, Algeria, july 2005, pp. 494–499.

[18] A. Bounsiar, P. Beauseroy, and E. Grall, "Fast training and efficient linear learning machine," To appear in proceedings of ICASSP'06. Toulouse, France, May 14-19 2006.

[19] V. Vapnik, *The Nature of Statistical Learning Theory*, Spring Verlag, 1995.

[20] V. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[21] N. Cristianini and J. Shawe-Taylor, *Support vector machines and other kernel-based learning methods*, Combridge University Press, 2000.

[22] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[23] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–279, 1995.

[24] B. Schölkopf and A. J. Smola, *Learning with kernels*, MIT Press, MA, 2002.

[25] C. J Van Rijsbergen, *Information retrieval*, Butterworths, London, 1979.

[26] J. A. Swets, *Information Retrieval Systems*, Bolt, Beranek and Newman. Cambridge, Massachusetts, 1967.

[27] C. Drummond and R. Holte, "What roc curves can't do (and cost curves can)," in *Proceedings of the ROC Analysis in Artificial Intelligence, 1st International Workshop*. 22 août 2004, pp. 19–26, Valencia, Espagne.

[28] K. Fukunaga, *Introduction to Statistical Patern Rcognition*, Academic Press, New York, 2^{nd} edition, 1990.

Open Science Index, Computer and Information Engineering Vol:1, No:12, 2007 publications.waset.org/14477/pdf