Optimizing the Classification Cost using SVMs with a Double Hinge Loss

Amirou Ahmed†, Ould-Abdeslam Djaffar‡ and Zidelmal Zahia†
†Mouloud Mammeri University , Tizi-Ouzou, Algeria
E-mail: a-amirou@mail.ummto.dz, djaffar.ould-abdeslam@uha.fr

Aidene Mohamed† and Merckle Jean‡
‡MIPS Laboratoiry, Haute Alsace University, France
E-mail: m-aidene@mail.ummto.dz

Keywords: support vector machines, double hinge loss, rejection, classification cost

Received: March 6, 2014

The objective of this study is to minimize the classification cost using Support Vector Machines (SVMs) Classifier with a double hinge loss. Such binary classifiers have the option to reject observations when the cost of rejection is lower than that of misclassification. To train this classifier, the standard SVM optimization problem was modified by minimizing a double hinge loss function considered as a surrogate convex loss function. The impact of such classifier is illustrated on several discussed results obtained with artificial data and medical data.

Povzetek: Predstavljena je optimizacija cene klasificiranja z metodo strojnega učenja SVM.

1 Introduction

Support Vector Machines (SVMs) are becoming one of the most popular pattern recognition schemes due to their remarkable generalization performance. This is motivated by the application of Structural Risk Minimization principle [1, 2]. Because of their good performance in terms of accuracy and generalization, SVMs are frequently used in very complex two-class classification problems.

Even though the generalization performance of support vector classifiers, misclassifications cannot be completely eliminated and, thus, can produce severe penalties. The expected error of a prediction is a very relevant point in many sensitive applications, such as medical diagnosis or industrial applications.

To improve the reliability of classifiers, new machine learning algorithms have been introduced such us conformal prediction determining levels of confidence [3]. Hence, classifications with less confidence than a given threshold may be rejected. This also motivates the introduction of a reject option in classifiers, by allowing for a third decision ® (Reject) when the conditional probability that an example belongs to each class is close to $1/2$.

Rejecting ambiguous examples has been investigated since the publications of [4, 5] on the error reject tradeoff. A notable attempts to integrate a reject rule in SVMs has been presented in [6]. The authors developed an SVM whose reject region is determined during the training phase. They derived a novel formulation of the SVM training problem and developed a specific algorithm to solve it. Some works have proposed rejection techniques using two thresholds on the output of the SVM classifier and produce a reject region delimited by two parallel hyperplane in the feature space [7, 8]. Other works used mixture of classifiers [9]. This approach is computationally highly expensive.

Recently, some remarkable works have proposed SVM classifier with a reject option using a double hinge loss function. This option was proposed in [10, 11, 12, 13]. The formulation in [10, 11, 12] is restricted to symmetric losses. In [13], the authors have proposed a cost-sensitive reject rule for SVM using an asymmetric double hinge loss. This formulation is based on probabilistic interpretation of SVM published in [14, 15] providing accurate estimation of posterior probabilities. It also generalizes those suggested in [11, 12] to arbitrary asymmetric misclassification and rejection costs. In all these model classifiers, the reject region is defined during the training phase.

In this paper, we develop the training criterion for a generalized SVM with a double hinge loss and then compare the performance of symmetric and asymmetric classification. The optimal classification cost and the error-reject tradeoff have been highlighted through several illustrated tests.

Note that the minimal classification cost must correspond to a good error-reject tradeoff. It is desirable that most of the rejected patterns would have been erroneously classified by the ideal Bayes classifier.

The remainder of this paper is structured as follows. After problem setting in section 2, section 3 recalls Bayes rule with rejection. In section 4, SVM classifier with rejection is developed using the generalized double hinge loss function. After this, the training criterion is detailed. In Section 5, the implementation is tested empirically. Il shows results comparing the considered classifiers. Finally, Section 6 briefly concludes the paper.

2 Problem setting

Let us consider a binary classification problem in which each example belongs to one of two categories. A discriminant $f : \mathcal{X} \mapsto \mathbb{R}$ classifies an observation $x \in \mathcal{X}$ into one of two classes, labeled +1 or -1 . Viewing $f(x)$ as a proxy value of the conditional probability $P = \mathbb{P}(Y = 1|X = x)$, one is less confident for small values of $\mid f(x) \mid$ corresponding to P around $1/2$. The strategy used in this work is to report $sgn(f(x_i)) = +1$ or -1 if $|f(x_i)|$ exceeds a threshold δ_i and no decision otherwise.

In binary problems, the two types of errors are:

- FP: False Positive, where examples labeled -1 are categorized in the positive class, incurring a loss C_n

- FN: False Negative, where examples labeled $+1$ are categorized in the negative class, incurring a loss C_p.

We also assume that the decision ® incurs a loss, R_n and R_p for rejected examples labeled -1 and $+1$, respectively. This formulation corresponds to [13] . For symmetric classification [10, 11, 12], we have $C_p = C_n = 1$ and $R_p = R_n = r$ with $0 \leq r \leq 1/2$. The expected losses pertaining to each possible decision d are displayed in Figure 1, assuming that all costs are strictly positive. The lower risk $\mathcal{R}$ is:

$$\mathcal{R}(d) = min\{C_p P(x), C_n(1 - P(x)), \\ R_p P(x) + R_n(1 - P(x))\} \quad (1)$$

where $P(x)$ denotes $P(Y = 1|X = x)$. According to (1), one can see in Figure 1 that rejecting a pattern is a viable option if and only if the point G is located above the segment AB. In other terms, if and only if $\frac{R_p}{C_p} + \frac{R_n}{C_n} < 1$ corresponding to $0 \leq r \leq 1/2$ in [10, 11, 12].

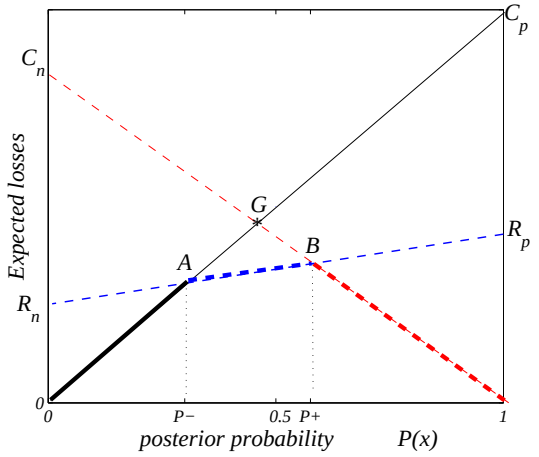

Figure 1: Expected losses against posterior probabilities

3 Bayes rule with rejection

From Figure 1, we deduce that Bayes classifier d^* defined as the minimizer of the risk $\mathcal{R}(d)$ can be expressed simply, using two thresholds:

$$P_+ = \frac{C_n - R_n}{C_n - R_n + R_p} , \quad (2)$$

$$P_- = \frac{R_n}{C_p - R_p + R_n} , \quad (3)$$

corresponding to symmetric thresholds $P_- = r$ and $P_+ = 1 - r$ in [10, 11, 12].

As Bayes decision rule is defined by conditional probabilities, many classifiers first estimate the conditional probability $\widehat{P}(Y = 1|X = x)$, and then plug this estimate in Eq.4 to build the decision rule.

$$f^*(x) = \begin{cases} +1 & \text{if } \widehat{P}(Y = 1|X = x) > P_+ , \\ -1 & \text{if } \widehat{P}(Y = 1|X = x) < P_- , \\ 0 & \text{otherwise .} \end{cases} \quad (4)$$

where $f^*(x)$ corresponds to the decision d^*, minimizer of the risk (1).

4 SVM classifier with Reject option (SVMR)

To minimize the empirical counterpart of the risk (1) computationally not feasible, one could replace it by surrogate loss functions. The most popular are the hinge loss motivated by [1] leading to sparse solutions [13, 12] and the logistic regression model offering ability to estimate the posterior probability $\widehat{P}(Y = 1|X = x) = 1/(1 + exp(-yf(x)))$ and then a good choice of the thresholds δ_i. In this study, $\widehat{P}(Y = 1|X = x)$ have to be accurate only in the neighborhood of P_+ and P_- (see equation 4).

4.1 Double hinge loss

The generalized double hinge loss introduced in [13] is a convex and piecewise linear loss function that is tangent to the negative log-likelihood loss at $\delta_+ = log(P_+/(1-P_+))$ and at $\delta_- = log(P_-/(1 - P_-))$ (see Figure 2). This proposal retains the advantages of both loss functions mentioned above: the sparsity of the hinge loss and the ability of the neg-log-likelihood loss to estimate the posterior probability P_+ and P_-, respectively at the tangency points δ_+ and δ_-. So the decision rule can be expressed as:

$$f(x) = \begin{cases} +1 & \text{if } f(x) > \delta_+ , \\ -1 & \text{if } f(x) < \delta_- , \\ 0 & \text{otherwise .} \end{cases} \quad (5)$$

These thresholds are symmetric in [10, 11, 12], $\delta_+ = -\delta_- = \delta_o$ and the recommended value of δ_o belongs to the interval $[r, 1 - r]$. To express the generalized double

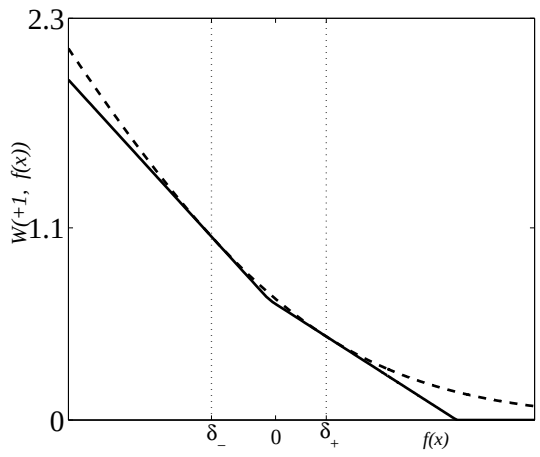

Figure 2: *Double hinge loss function for positive examples, with* $P_- = 0.35$ *and* $P_+ = 0.6$ *(solid: double hinge, dashed: likelihood)*

hinge function [13], we consider firstly the standard logistic regression procedure where φ is the negative log-likelihood loss:

$$\varphi(y, f(x)) = log(1 + exp(-yf(x))) \ . \tag{6}$$

that is $\varphi(+1, f(x)) = log(1 + exp(-f(x)))$ for positive examples and $\varphi(-1, f(x)) = log(1 + exp(f(x)))$ for negative examples. Let us work on Figure 3 corresponding to positive examples ($y_i = +1$).

$W = a_1 f(x) + g_1$ is the first slop (right to left) of $W(+1, f(x))$ where $a_1 = \frac{d[\varphi(+1,f(x))]}{d[f(x)]}|_{\delta_+} = -(1 - P_+)$.

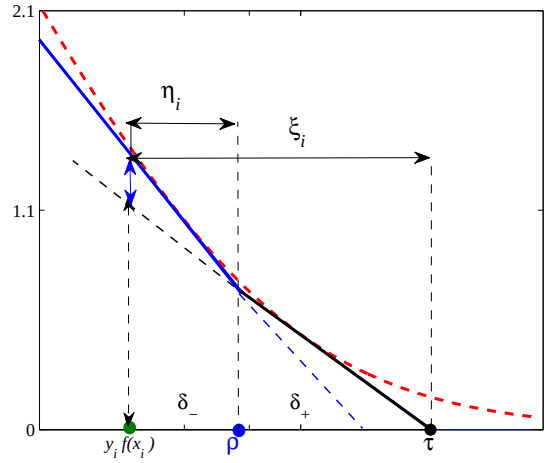

Figure 3: *Double hinge loss function for positive examples, with* $P_- = 0.4$ *and* $P_+ = 0.7$ *(solid: double hinge, red dashed: likelihood)*

At the tangency point δ_+, we have $\varphi(+1, f(x)) = W(+1, f(x))$, hence $g_1 = -p_+ log(P_+) - (1 - P_+) log(1 - P_+) = H(P_+)$.

The second slop of $W(+1, f(x))$ is $W = a_2 f(x) + g_2$ where $a_2 = \frac{d[\varphi(+1,f(x))]}{d[f(x)]}|_{\delta_-} = -(1 - P_-)$ and $g_2 = -P_- log(P_-) - (1 - P_-) log(1 - P_-) = H(P_-)$.

For $a_1 f(x) + g_1 = 0$, we have $f(x) = \tau_+ = \frac{H(P_+)}{1-P_+}$ and for $a_1 f(x) + g_1 = a_2 f(x) + g_2$, we have $f(x) = \rho_+ = \frac{H(P_-) - H(P_+)}{P_+ - P_-}$. The double hinge function for positive examples is then expressed as:
$W(+1, f(x)) =$

$$\begin{cases} -(1 - P_-) f(x) + H(P_-) & \text{if } f(x) < \rho_+ \\ -(1 - P_+) f(x) + H(P_+) & \text{if } \rho_+ \leq f(x) < \tau_+ \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

The same strategy of calculation leads to the double hinge function for negative examples: $W(-1, f(x)) =$

$$\begin{cases} P_+ f(x) + H(P_+) & \text{if } f(x) > \rho_- \ , \\ P_- f(x) + H(P_-) & \text{if } \tau_- \geq f(x) > \rho_- \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

where $\tau_- = \frac{-H(P_-)}{P_-}$ and $\rho_- = \rho_+ = \rho$. The double hinge loss ψ_r introduced in [10, 11, 12] is a scaled version of the loss W. It is given by $\psi_r(yf(x)) =$

$$\begin{cases} 1 - \frac{1-r}{r} yf(x) & \text{if } yf(x) < 0 \\ 1 - yf(x) & \text{if } 0 \leq yf(x) < 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

hence,

$$\psi_r(yf(x)) = \frac{1}{H(r)} W\left(y, \frac{H(r)}{r} f(x)\right) \ .$$

where $H(r) = H(P_-)$ and $H(P_-) = H(P_+)$ in the symmetric case. Note that, although minimizing $\psi_r(yf(x))$ or W will lead to equivalent solutions for f. With minimizing $\psi_r(yf(x))$, the decision rule recommended by [11] classifies an example when $|f(x)| > \delta_o = \frac{1}{2}$, while in [13], an example is classified when $|f(x)| > \frac{r}{H(r)} log \frac{r}{1-r}$. The last decision rule rejects more examples when the loss incurred by rejection is small and fewer examples otherwise. The two rules are identical for $r = 0.24$.

4.2 Training Criterion

As in standard SVMs, we consider the regularized empirical risk on the training sample. Introducing the double hinge loss (7-8) results in an optimization problem that is similar to the standard SVMs problem.

4.2.1 Primal problem

Let C_o a constant to be tuned by cross-validation, we define $D = C_o(P_+ - P_-)$, $B_i = C_o(1 - P_+)$ for positive examples and $B_i = C_o P_-$ for negative examples. The primal optimization problem reads

$$\begin{aligned} min_{f,b} \ & \tfrac{1}{2}\|f\|^2_{\mathcal{H}} + \\ & \textstyle\sum_{i=1}^{n} B_i \left|\tau_i - y_i(f(x_i) + b)\right|_+ + \\ & D \textstyle\sum_{i=1}^{n} \left|\rho - y_i(f(x_i) + b)\right|_+ \end{aligned} \tag{10}$$

where $|\cdot|_+ = max(\cdot, 0)$. The (squared) norm of f is a regularization functional in a suitable Hilbert space. The primal problem (10) is best seen with introduction of slack variables ξ and η shown in Figure(3).

$$\begin{cases} min_{f,b,\xi,\eta} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + \sum_{i=1}^{n} B_i\xi_i + D\sum_{i=1}^{n}\eta_i \ , \\ Sc & y_i(f(x_i)+b) \geq \tau_i - \xi_i,\ i = 1,\ldots,n \\ & y_i(f(x_i)+b) \geq \rho - \eta_i,\ i = 1,\ldots,n \\ & \xi_i \geq 0\ , \quad \eta_i \geq 0, \qquad i = 1,\ldots,n. \end{cases} \tag{11}$$

4.2.2 Dual problem

The Lagrangian of (11) is given by:

$$\begin{cases} L(f,b,\xi,\eta,\alpha,\beta,\upsilon,\omega) = \frac{1}{2}\|f\|^2 + \sum_{i=1}^{n} B_i\xi_i \\ +D\sum_{i=1}^{n}\eta_i - \sum_{i=1}^{n}\alpha_i\left[y_i(f(x_i)+b) - \tau_i + \xi_i\right] \\ -\sum_{i=1}^{n}\beta_i\left[y_i(f(x_i)+b) - \rho + \eta_i\right] \\ -\sum_{i=1}^{n}\upsilon_i\xi_i - \sum_{i=1}^{n}\omega_i\eta_i \end{cases} \tag{12}$$

with:
$\upsilon_i \geq 0,\ \omega_i \geq 0,\ \alpha_i \geq 0,\ \beta_i \geq 0,\ and\ i = 1,\ldots,n.$

The Kuhn-Tucker conditions imply:

$$\begin{cases} \frac{\partial L}{\partial b} = 0 & \Rightarrow & \sum_{i=1}^{n}(\alpha_i + \beta_i)y_i = 0 \\ \frac{\partial L}{\partial f} = 0 & \Rightarrow & \sum_{i=1}^{n} f(\cdot) = (\alpha_i + \beta_i)y_i k(\cdot, x_i) \\ \frac{\partial L}{\partial \xi_i} = 0 & \Rightarrow & B_i - \upsilon_i - \alpha_i = 0 \Rightarrow 0 \leq \alpha_i \leq B_i \\ \frac{\partial L}{\partial \eta_i} = 0 & \Rightarrow & D - \omega_i - \beta_i = 0 \Rightarrow 0 \leq \beta_i \leq D \end{cases} \tag{13}$$

for $i = 1,\ldots,n$. Thanks to these conditions, we can eliminate f, ξ and η from the Lagrangian.

$$\begin{cases} L(\alpha,\beta) = & \frac{1}{2}(\alpha+\beta)^T G(\alpha+\beta) - \tau^T\alpha - \rho^T\beta \\ Sc & y^T(\alpha+\beta) = 0 \\ & 0 \leq \alpha_i \leq B_i, i = 1,\ldots,n \\ & 0 \leq \beta_i \leq D, i = 1,\ldots,n \end{cases} \tag{14}$$

where $\tau = (\tau_1,\ldots,\tau_n)^T$ et $\rho = (\rho_1,\ldots,\rho_n)^T$ are the threshold vectors of $\mathbb{R}^n$, G is the $n \times n$ influence matrix with general term $G_{ij} = y_i y_j k(x_i,\ x_j)$ and $k(.,.)$, is the reproducing kernel of the Hilbert space $\mathcal{H}$. Let $\gamma = \alpha + \beta$, the problem (14) can be rephrased as:

$$\begin{cases} max_{\alpha,\gamma} & -\frac{1}{2}\gamma^T G\gamma + (\tau-\rho)^T\alpha + \rho^T\gamma \ , \\ Sc & y^T\gamma = 0\ , \\ & 0 \leq \alpha_i \leq B_i, \qquad i = 1,\ldots,n\ , \\ & 0 \leq \gamma_i - \alpha_i \leq D, \quad i = 1,\ldots,n \end{cases} \tag{15}$$

The problem (15) is a quadratic problem under box constraints. Compared to the standard SVM dual problem, one has an additional vector to optimize, but we will show that α is easily recovered from γ.

4.2.3 Solving the problem

To solve the dual (15), the strategy used in the active set method [17] is considered. Firstly, the training set is partitioned in support and non support vectors. the training criterion is optimized considering this partition. Then, this optimization results in an updated partition of examples in support and non-support vectors. These two steps are iterated until predefined level of accuracy is reached. Table (1) shows how the training set is partitioned into five subsets defined by the constraints in (15).

The outcomes of the membership of example i to one of the subsets described above has the following consequences on the dual variables (α, γ):

$$\begin{cases} i \in I_0 \ \Rightarrow \alpha_i = 0 & \gamma_i = 0 & ; \\ i \in I_\tau \ \Rightarrow 0 \leq \alpha_i \leq B_i & \gamma_i = \alpha_i & ; \\ i \in I_B \Rightarrow \alpha_i = B_i & \gamma_i = B_i & ; \\ i \in I_\rho \ \Rightarrow \alpha_i = B_i & B_i < \gamma_i < B_i + D; & \\ i \in I_D \Rightarrow \alpha_i = B_i & \gamma_i = B_i + D & . \end{cases} \tag{16}$$

Hence, provided that the partitioning is known, γ_i has to be computed only for $i \in I_\tau \cup I_\rho$. Furthermore, α_i is either constant or equal to γ_i.

We saw that, assuming that the examples are correctly partitioned, problem 15 can be solved by considering a considerably smaller problem, namely the problem of computing γ_i for $i \in I_\tau \cup I_\rho$. Let $I_c = \{I_B, I_D\}$ and $I_h = \{I_\tau, I_\rho\}$. The problem (15) becomes:

$$\begin{cases} L(\gamma) = & \frac{1}{2}\gamma^T G\gamma - (S)^T\gamma \\ Sc: & y^T\gamma = 0 \\ 0 \leq \gamma_i \leq C,\ i = 1,\ldots,n\ and\ C_i = B_i + D \end{cases} \tag{17}$$

The relation between the parameters of the preceding formulation and the initial parameters of the problem (11) can be obtained after formulating the Lagrangian of the dual (17)

$$\begin{cases} L(\gamma,\lambda,\mu,\nu) = \\ \frac{1}{2}\gamma^T G\gamma - S^T\gamma + \lambda\gamma^T y - \nu^T\gamma + \mu^T(\gamma - C\mathbb{I}^n) \end{cases} \tag{18}$$

where the Lagrange multipliers λ, μ, ν must be positive or null and $\mathbb{I}^n$, a vector of 1. This Lagrangian can be compared with the Lagrangian of the primal (11) reformulated as follows:

$$\begin{cases} L = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n}\gamma_i y_i f(x_i) - \sum_{i=1}^{n}\gamma_i y_i b \\ +\sum_{i=1}^{n}\alpha_i(\tau-\rho) + \sum_{i=1}^{n}\gamma_i\rho \\ +\sum_{i=1}^{n}\xi_i(B_i - \alpha_i - \upsilon_i) \\ +\sum_{i=1}^{n}\eta_i(D - \beta_i - \omega_i) \end{cases} \tag{19}$$

by replacing the variable f by γ, the problem (19) becomes:

$$\begin{cases} L(\gamma,b,\xi,\eta) = \\ \frac{1}{2}\gamma^T G\gamma + b\gamma^T y - S^T\gamma \\ +\xi^T(\alpha + \upsilon - B) + \eta^T(\gamma + \omega - D\mathbb{I}^n) \end{cases} \tag{20}$$

I_0	saturated part of the loss	$I_0 = \{i \mid y_i(f(x_i)+b) > \tau\}$
I_τ	first hinge of the loss	$I_\tau = \{i \mid y_i(f(x_i)+b) = \tau\}$
I_B	first slop of the loss	$I_B = \{i \mid \rho < y_i(f(x_i)+b) < \tau\}$
I_ρ	second hinge of the loss	$I_\rho = \{i \mid y_i(f(x_i)+b) = \rho\}$
I_D	second sop of the loss	$I_D = \{i \mid y_i(f(x_i)+b) < \rho\}$

Table 1: Partitioning the training set

To reveal the relations between the primal and dual variables, we will check the KKT conditions stipulating the cancellation of the gradient of the Lagrangian (20) according to the primal variable γ in the different subsets.

Table 2 describes the properties of each set regarding the original variables and the Lagrange multipliers.

4.2.4 Algorithm

Let us assume the repartition in each set (I_0, I_h and I_c) to be known. Only the values of γ belonging to I_h remain unknown, they will then be given by the solution of the following optimization problem whose dimension is lower than initial dimension. After slightly abusing notations, we define: $\gamma_h = \gamma(I_h)$, $y_h = y(I_h)$, $G_{hh} = G(I_h, I_h)$, $c_C = \sum_{(i \in I_B)} B_i y_i + \sum_{(i \in I_D)} C_i y_i$ and
$S_h =$
$(\tau(I_\tau)^T \rho(I_\rho)^T)^T - G(I_h, I_D)(B(I_B)^T D\mathbb{I}(I_D)^T)^T$.
The problem (17) becomes:

$$\begin{cases} L(\gamma_h) = & \frac{1}{2}\gamma_h^T G_{hh}\gamma_h - S_h^T\gamma_h \\ Sc & y_h^T\gamma_h + c_C = 0 \end{cases}, \qquad (21)$$

The Kuhn-Tucker conditions gives us the system to be solved to find the values of γ that are still unknown.

$$\begin{cases} G_{hh}\gamma_h = S_h - y_h^T\lambda \\ y_h^T\gamma_h = -c_C \end{cases}, \qquad (22)$$

After resolving this system, a component of γ violating the primal or dual constraints must be moved to the suitable set. The process is iterated until all box constraints are satisfied.

During the learning process, the time consuming step is the resolution of the linear system (22). For this, we used the incremental strategy outlined in [18] whose complexity is close to $\mathcal{O}(n^2)$ The presented SVMR computational complexity is comparable to that of the standard SVM [18]. The only computational overhead is that the presented SVMR uses 5 categories of examples while SVM uses three.

5 Results and discussions

Data:
To evaluate the performance of the SVMR classifiers, three types of data have been used:

- synthetic data generated with a classical dataset with two gaussianly distributed classes with similar variances but different means chosen to create many ambiguous examples.
- as medical decision making is an application domain for which rejection is of primary importance, data related to medical problems will be considered. ElectroCardioGram (ECG) records from (www.physionet.org /physiobank/database/mitdb) are used. Each tape is accompanied by an annotation file. in which ECG beats have been labeled by expert cardiologists. Since this study is to evaluate the performance of a binary classifier with a reject option, we followed the AAMI recommended practice [19] to form two heartbeat classes: (i) the positive class representing the ventricular ectopic beats (V); (ii) the negative class representing the normal beats (N), including Normal beats, Left Bundle Branch Block beats (LBBB) and Right Bundle Branch Block beats (RBBB). In agreement with [19], records containing paced beats (102, 104, 107, 217) and 23 records with no V beat or less than 40 V beat were excluded leaving 21 records of interest. We have stored each beat by a 7-feature vector. The feature extraction is described in [20]
- For experimenting with large data, the forest CoverType database from UCI repository was also used. (http://kdd.ics.uci.edu/databases/covertype/). We consider the subproblem of discriminating the positive class Cottonwood (2747 examples) against the negative class Douglas-fir (17367 examples).

Tests:
The first series of experiments are done with the ECG data to explain the effectiveness of the classification with rejection. We selected record 214 and 221 containing together 3546 of N beats and 652 of V beats. As no cost matrix is provided with this data, we assume that $R_p = R_n = r$ as in [10, 11, 12] and $P_+ = 1 - \frac{C_p}{C_n}P_- = 1 - \theta P_-$ where $\theta = 1$ in [10, 11, 12] and $\theta \geq 1$ in [13]. Often, in practice, especially in medical applications, FN errors are more costly than FP errors ($\theta > 1$). Figures 4 and 5 show respectively an example of the reject region produced by the SVMR classifier for $\theta = 1$ and for $\theta > 1$. In Figure 5, the SVMR classifier encourages the rejection of more FN examples because they are more costly than FP examples.

All previous classifiers comparatives studies have been based on the error rates obtained, but error rate is not the

Set	Initial constraints	Primal constraints	Dual constraints
I_0	$y_i[f(x_i)+b] > \tau_i$	$\xi_i = \eta_i = 0$	$\mu = 0, \nu = G\gamma + by - \tau \neq 0$
I_τ	$y_i[f(x_i)+b] = \tau_i$	$\xi_i = \eta_i = 0$	$\mu = 0, \nu = G\gamma + by - \tau = 0$
I_B	$\rho < y_i[f(x_i)+b] < \tau_i$	$\xi_i \neq 0, \eta_i = 0$	$\nu = 0, \mu = -G\gamma - by + \tau = \xi$
I_ρ	$y_i[f(x_i)+b] = \rho$	$\xi_i \neq 0, \eta_i = 0$	$\nu = 0, \mu = G\gamma + by - \rho = 0$
I_D	$y_i[f(x_i)+b] < \rho$	$\xi_i \neq 0, \eta_i \neq 0$	$\nu = 0, \mu = -G\gamma - by + \rho = \eta$

Table 2: Situation of the constraints for the five types of examples

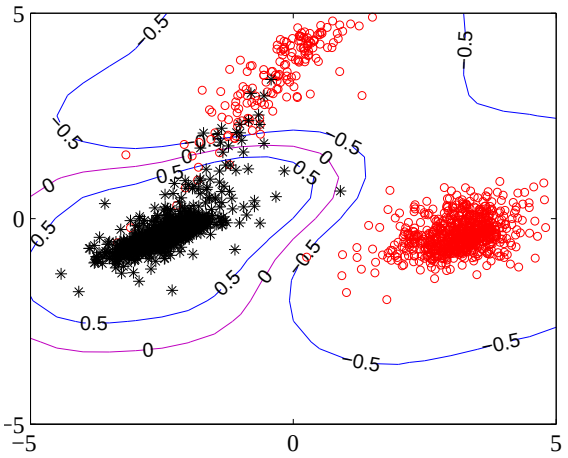

Figure 4: Scatter plot showing the reject region induced by the reject thresholds in correspondence to the costs of misclassifying and rejecting samples. Positive cases are represented by black asts and negative cases by red circles. The lines $+0.5$ and -0.5 correspond respectively to δ_o and $-\delta_o$ and the line 0 corresponds to $f(x) = 0$ or $P(Y = 1 \mid X = x) = 0.5$.

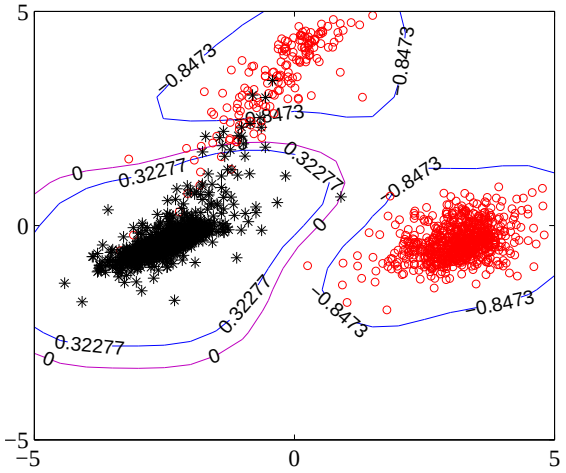

Figure 5: Scatter plot showing the reject region induced by the reject thresholds in correspondence to the costs of misclassifying and rejecting samples. Positive cases are represented by black asts and negative cases by red circles. The lines $+0.32277$ and -0.8473 correspond respectively to δ_+ and δ_- and the line 0 corresponds to $f(x) = 0$ or $P(Y = 1 \mid X = x) = 0.5$.

only measurement that can be used to judge a classifier's performance. In many applications, the classification cost is a parameter witch will be considered since Bayes classifiers with or without rejection aim to minimize the classification cost.

For illustration, we compare the reject rates obtained with the SVMR classifiers proposed in [10, 11, 12] where the reject threshold $\delta_o \in [r, \quad 1-r]$ and the SVMR classifier proposed in [13] where the reject thresholds are $\delta_+ = \log(P_+/(1-P_+))$ and $\delta_- = \log(P_-/(1-P_-))$ respectively for positive and negative examples. For this purpose, we consider the symmetric classification, $P_+ = 1 - P_-$. Figure 6 and 7 obtained with synthetic data and ECG data (record 214 and 221) show that in all cases, the decision rule [13] rejects fewer examples when the loss incurred by rejection is high and more examples otherwise. The rule in [10, 11, 12] considers the reject threshold $\delta_o = 1 - r$ as the largest value of δ_o and then rejects more examples for all reject costs. For $\delta_o = r$, this rule rejects less frequently especially when r close to zero, it becomes almost with no rejection. For the middle value $\delta_o = 0.5$ seen as a compromise among r and $1 - r$, the rule [10, 11, 12] and the one proposed in [13] are identical at $r = 0.24$.

As pointed out in [5], the advantage of classifying with rejection can be judged by the error-reject tradeoff. Since the error rate E and the reject rate R are monotonic functions of r. We compute the tradeoff E versus R from $E(r)$ and $R(r)$ when r varies between 0.5 and 0.12 and the threshold $\delta_o = 0.5$ recommended in [11, 12]. Figure 8 shows the error reject tradeoff for the rule proposed in [13] (black curves) and for the rule proposed in [10, 11, 12] (red curves). The obtained results differ due to the size of the rejection region induced by the rules. From these results, we can conclude another interesting parameter that is the error-reject ratio defined in [5] that is $\frac{\triangle E}{\triangle R}$ (dashed lines). For high reject costs $(0.4 \leq r \leq 0.5)$, the rule [13] indicates an error-reject ratio of -0.58, -0.84 and -0.42 respectively for synthetic data, ECG data and forest data. This means that 58%, 84% and 42% respectively of the rejected patterns would have been erroneously classified. Using the rule proposed in [10, 11, 12] with $\delta_o = 0.5$, the error-reject ratios obtained are -0.15 for synthetic data, -0.23 for ECG

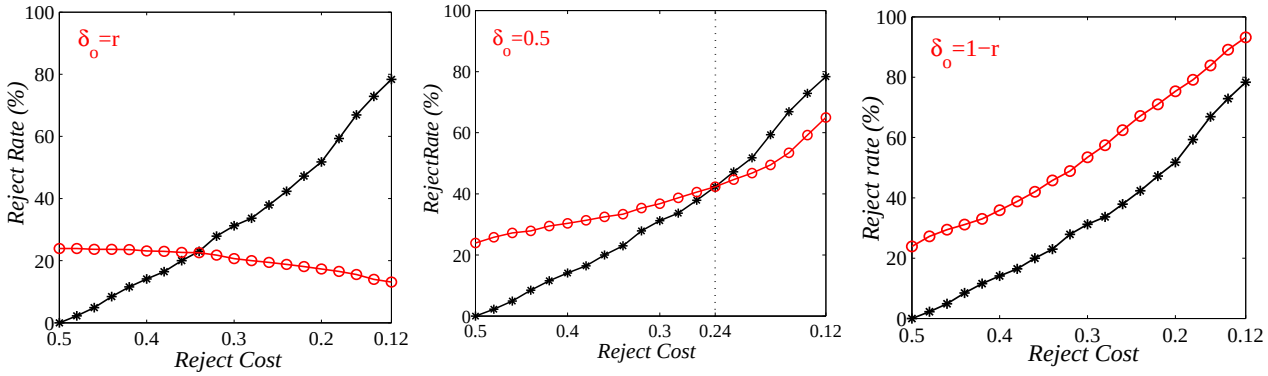

Figure 6: Comparison of the reject rate versus the reject cost r obtained with the SVMR in [13] (black curves) and with the SVMR introduced in [10, 11, 12] (red curves). These results are obtained with synthetic data.

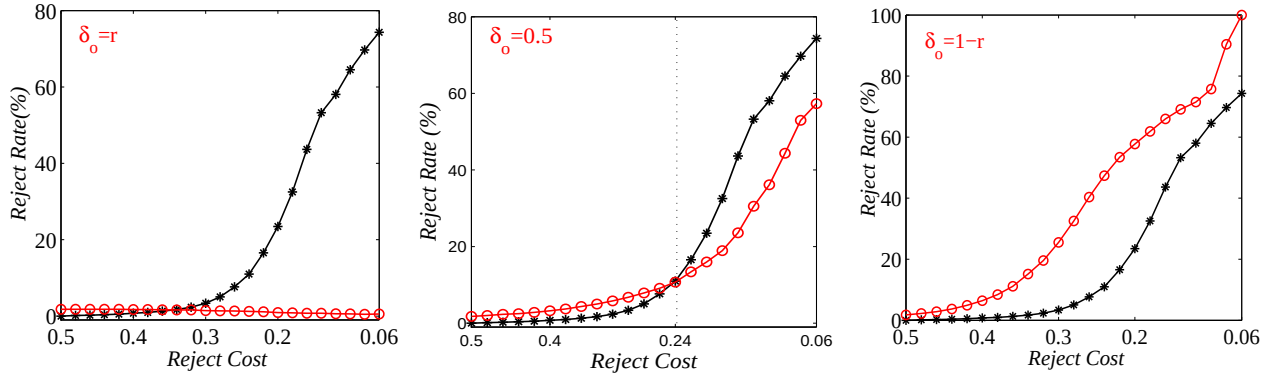

Figure 7: Comparison of the reject rate versus the reject cost r obtained with the SVMR in [13] (black curves) and with the SVMR introduced in [10, 11, 12] (red curves). These results are obtained with medical data.

data and -0.22 for forest data. This means that only 15%, 23% and 22% respectively of the rejected patterns would have been erroneously classified. Hence, it is clear that the rule [13] should lead to a better classification cost.

The last series of tests was carried out using all the selected ECG records. The mean results obtained are reported in Figures 9 and 10. The error against the reject decreases until a quasi constant rate (Figure 9) . Another interesting plot in the same figure represents the error reject ratio. The inflection point in this plot is interesting since it indicates the most important variation of the error against the variation of the reject rate. Two statistical parameters are also used to highlight the performance of the reject rule [13]. The sensitivity and positive predictivity are computed by

$$S_e = \frac{TP}{TP+FN}; \qquad P_p = \frac{TP}{TP+FP}$$

where True Positive (TP) are the samples labeled $+1$ categorized in the positive class. Figure 10 (top) indicates the variation of the classification cost given by

$$C_c = [C_p FN + C_n FP + rR_{rej}]/N_{tot} \qquad (23)$$

where R_{rej} is the number of rejected patterns and N_{tot}, the total number of examples. The same Figure shows that the optimal classification cost C_c corresponds to a good error-reject tradeoff (see Figure 9). Figure 10 (bottom) shows that the positive predictivity is close to 99.8%. In the same figure, it is shown that we obtained more than $98,2\%$ of sensitivity with no rejection and more than 99% of sensitivity for the minimal classification cost with rejection considering $R_p = R_n = r$ and $C_n = 1$ and $\theta = 1.2$. In the same figure, it is clearly shown that the optimal classification cost is not obtained for r=0.5 (simple Bays rule) but for a rejection rate equal to 1.8%. In any application, one must choose the error rate and the rejection rate corresponding to the minimal classification cost. It is the goal of using a cost sensitive classifier.

For a better appreciation of such reject schemes, it should be desirable to perform tests on data accompanied by real cost matrix.

Even though the considered classifier based on sparse probabilistic interpretation of SVM, providing an accurate estimation of posterior probabilities, it should be interesting to assign confidence values to each classification. This can be considered by introducing conformal predic-

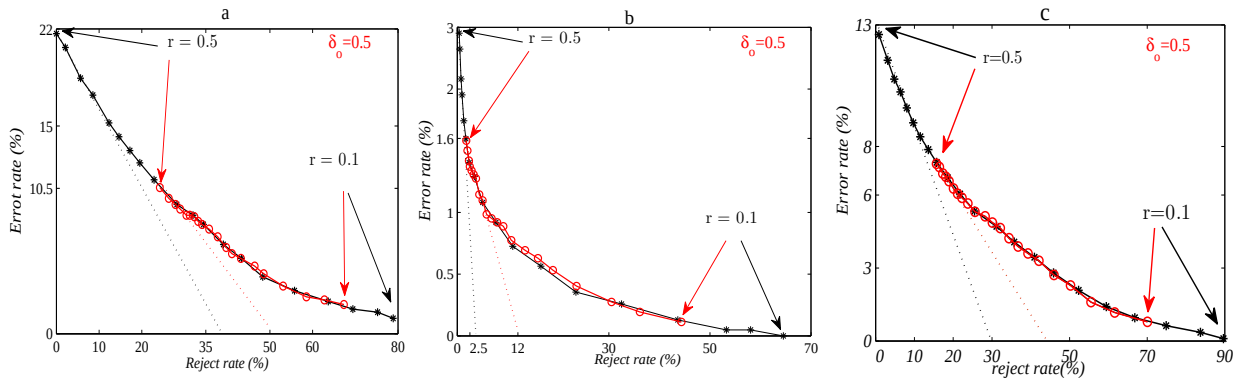

Figure 8: Error versus reject tradeoff obtained using synthetic data (a), ECG data (b) and forest data (c); with [13] (black curves) and with [10, 11, 12] using $\delta_o = 0.5$(red curves).

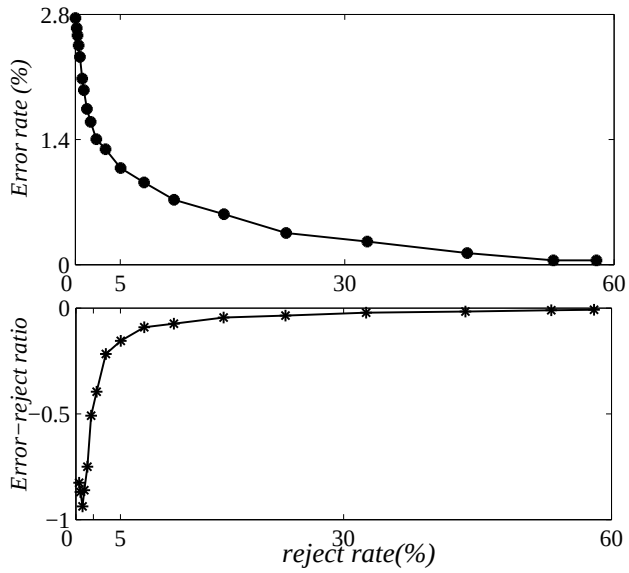

Figure 9: Top: Error rate vs. Reject rate. Bottom: Variation of the error rate against the variation of the reject rate

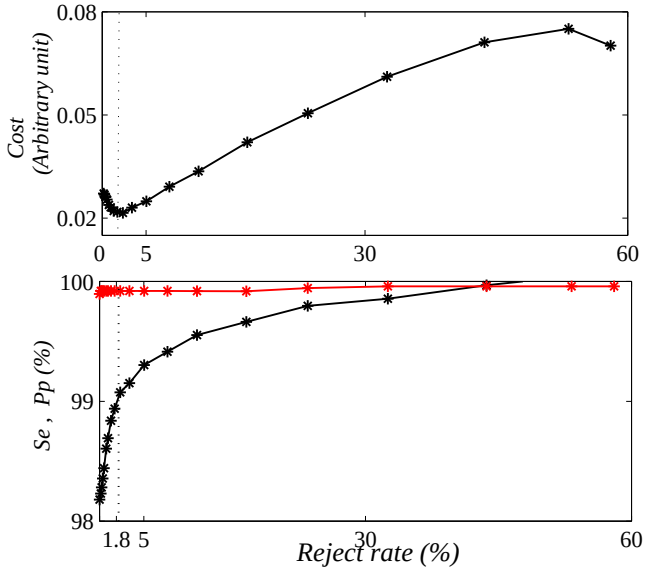

Figure 10: Top: Classification cost against Reject Rate. Bottom: Sensitivity (black curve) and positive predictivity (red curve) against reject rate.

tion whose relationship with rejection is clearly relevant, whether the rejection related to the ambiguity of examples or that related to their atypical characters.

6 Conclusion

This paper presents a cost-sensitive reject rules for SVMs using a double hinge loss. The solution inspired by the probabilistic interpretation of SVM, owns the advantage of the hinge loss function which leads to a consistent solution and the advantage of negative log-likelihood loss which allows a good estimation of posteriori probabilities in the vicinity of the decision thresholds. Note that these dynamic reject thresholds follow the cost of rejecting a sample and the cost of misclassifying a sample. This viewpoint aims to minimize the classification cost.

A possible improvement of this study is to estimate the level of confidence of the classifier by introducing the conformal prediction. This will be a crucial advantage, especially for medical applications, the risk of clinical errors may be controlled by an acceptable level of confidence for a given decision.

References

[1] V. N. Vapnik (1995) The Nature of Statistical Learning Theory *Springer Series in Statistics*

[2] N. Cristianini and J. Shawe-Taylor (2000) An Introduction to Support Vector Machines, *Cambridge University Press*.

[3] Glenn Shafer and Vladimir Vovk (2008) A Tutorial on Conformal Prediction, *Journal of Machine Learning Research*, 9, pp 371-421.

[4] C. K. Chow. (1957) An optimum character recognition system using decision function, *IRE Trans. Electronic Computers*, EC-6(4), pp. 47–254.

[5] C. K. Chow (1970) On optimum recognition error and reject tradeoff, *IEEE Trans. on Information Theory*, 16(41),pp. 41–46.

[6] G. Fumera and F. Roli (2002) Support vector machines with embedded reject option, *In S.-W. Lee and A. Verri, editors, Pattern Recognition with Support Vector Machines: First International Workshop*, volume 2388 of Lecture Notes in Computer Science-Springer, pp. 68–82.

[7] J. T. Kwok (1999) Moderating the outputs of support vector machine classifiers, *IEEE Trans. on Neural Networks*, 10(5), pp 1018–1031.

[8] F. Tortorella. (2004) Reducing the classification cost of support vector classifiers through an ROC-based reject rule, *Pattern Analysis and Applications*, 7(2) pp. 128–143.

[9] A. Bounsiar, E. Grall, P. Beauseroy. (2007) A Kernel Based Rejection Method for Supervised Classification. , *International Journal of Computational Intelligence*, 3(4), pp. 312–321.

[10] R. Herbei and M. H. Wegkamp (2006) Classification with reject option, *The Canadian Journal of Statistics*, 34(4), pp. 709–721.

[11] P. L. Bartlett and M. H. Wegkamp (2008) em Classification with a reject option using a hinge loss, *Journal of Machine Learning Research*, 9, pp. 1823–1840.

[12] M. Wegkamp, M. Yuan (2010) em Classification methods with reject option based on convex risk minimization, *Journal of Machine Learning Research*, (11), pp. 111–130.

[13] Y. Grandvalet, A. Rakotomamonjy, J. Keshet et S. Canu (2009) em Support Vector Machines With a Reject Option, *Advances in Neural Information Processing Systems*, (21), pp. 537–544.

[14] Y. Grandvalet, J. Mariethoz, and S. Bengio (2006) A probabilistic interpretation of SVMs with an application to unbalanced classification. *In Y. Weiss, B. Scholkopf, and J. C. Platt, editors, Advances in Neural Information Processing Systems 18 MIT Press*, pp. 467–474.

[15] P. L. Bartlett and A. Tewari (2007) Sparseness vs estimating conditional probabilities: Some asymptotic results, *Journal of Machine Learning Research*, 8, pp. 775–790.

[16] Z. En-hui and Z. Chao, S. Jian and L. Chen (2011) Cost-sensitive SVM with Error Cost and Class-dependant Reject Cost, *International Journal of Computer Theory and Engineering*, 3(1).

[17] S. V. N. Vishwanathan, A. Smola, and N. Murty (2003) SimpleSVM. *In T. Fawcett and N. Mishra, editors, Proceedings of the Twentieth International Conference on Machine Learning AAAI*, pp. 68–82.

[18] G. Loosli, S. Canu, S. Vishwanathan, and M. Chattopadhay (2005) Boite outils SVM simple et rapide. *Revue d'Intelligence Artificielle RIA*, 19(4), pp.741-767, 2005.

[19] R. Mark and R Wallen, (1987) AAMI-recommended practice: Testing and reporting performance results of ventricular arrhythmia detection olgorithms, *Association for the Advencement of Medical Instrumentation, Tech. Rep. AAMI ECAR*, 1987.

[20] Z. Zidelmal, A. Amirou et A. Belouchrani. (2012) Heartbeat classification using Support Vector Machines (SVMs) with an embedded reject option,*International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, vol,26,no. 1,DOI:10.1142/S0218001412500012.