2013 IEEE 13th International Conference on Data Mining Workshops

# Cost-Free Learning for Support Vector Machines with a Reject Option

Guibiao Xu and Bao-Gang Hu
*NLPR*
*Institute of Automation, Chinese Academy of Sciences*
*Beijing, P.R. China*
*Email: {guibiao.xu, hubg}@nlpr.ia.ac.cn*

***Abstract*—In this work, we investigate into the abstaining classification of binary support vector machines (SVMs) based on mutual information (MI). We obtain the reject rule by maximizing the MI between the true labels and the predicted labels, which is a post-processing method. The gradient and Hessian matrix of MI are derived explicitly so that Newton method is used for the optimization which converges very fast. Different from the existing reject rules of SVM, the present MI-based reject rule does not require any explicit cost information and is under the framework of cost-free learning. As a matter of fact, the cost information embedded in MI can also be derived from the method, which provides an objective or initial reference to users if they want to apply cost-sensitive learning. Numerical results confirm the benefits of the proposed MI-based reject rule in comparison with other reject rules of SVM.**

***Keywords*-abstaining classification; mutual information; support vector machines.**

## I. Introduction

Researchers may want to build abstaining classifiers that can refrain from classifying ambiguous examples in applications such as medical diagnosis and identity verification where errors cause severe losses. We can derive abstaining classifiers from nonabstaining probabilistic or margin-based classifiers by setting appropriate lower and upper reject thresholds [1]. However, we must concern that instances which are refrained from classification need a more powerful classification system which involves nonnegligible costs, and as a consequence, we need to find the optimal error-reject trade-off. Chow [2] proposed the optimal error-reject trade-off based on the Bayesian classifier. Pietraszek [3] built abstaining classifiers by three optimization criteria of cost-based, bounded-abstention and bounded-improvement using ROC analysis. Friedel [4] presented abstaining classifiers based on abstention cost curves.

Support vector machine (SVM) [5] is widely used in classification because of its remarkable generalization performance. The output margin of SVM can be used as estimation of confidence of a prediction [6]. Naturally, when errors cause severe losses, we had better refrain from classifying samples around the decision boundary of SVM since these samples are more prone to be misclassified. Researchers have investigated kinds of reject rules for SVM [7], [8], [9], [10]. Although they succeed in building up kinds of abstaining SVM classifiers, most of them need cost information which is given subjectively. Letting the data speak for themselves is a popular rule in data mining. Thus, we let the data determine the optimal reject rule of SVM in the mutual information (MI) sense. The new reject rule is obtained by maximizing the MI between the true labels and the predicted labels which needs no explicit cost information. Our reject rule of SVM is a kind of post-processing method and because it needs no explicit cost information, it is under the framework of cost-free learning. Cost-free learning was first proposed in [15] which means getting the optimal classification results without using explicit cost information. In all, our reject rule gives MI-optimal error-reject trade-off without any explicit cost information. The main contribution of this paper is:

- We develop a MI-based reject rule for SVM which belongs to cost-free learning. The gradient and Hessian matrix of MI between the true labels and the predicted labels are derived. With the help of Parzen Window method, we apply Newton method to optimize the MI objective which converges in several iterations.
- We establish the relationship between our MI-based reject rule and cost-sensitive learning [17] and derive the cost information embedded in MI. Based on the performance of our MI-based reject rule and practical requirements, the embedded cost information can serve as the reference of selecting reasonable cost information for cost-sensitive learning.

## II. Related Work

Table I briefly shows the characteristics of reject rules of SVM. Mukherjee et al. [7] introduced confidence levels based on the SVM outputs and rejected the patterns whose signed distances from the decision boundary were below a certain value. Fumera et al. [8] developed a maximum margin SVM classifier whose reject region was determined during the training phase. However, their approach leads to considerable computational overheads. A conventional way to introduce a reject option is to map the outputs of SVM into posterior probabilities [12] so that Chow's reject rule could be used. But the estimated posterior probabilities are seldom sufficiently faithful to produce good results. In [9], a novel reject rule of SVM was proposed by minimizing

978-0-7695-5109-8/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDMW.2013.45

Table I
THE CHARACTERISTICS OF REJECT RULES OF SVM

| Source | Category of method | Key point | Needed information |
|---|---|---|---|
| Mukherjee et al. [7] | post-processing | confidence level | confidence level |
| Fumera et al. [8] | pre-processing | embedded reject option | cost matrix |
| Chow's reject rule | post-processing | mapping outputs of an SVM into posterior probabilities | cost matrix |
| Tortorella [9] | post-processing | ROC analysis | cost matrix |
| Grandvalet et al. [10] | pre-processing | double hinge loss | cost matrix |
| Our MI-based reject rule | post-processing | mutual information | no cost matrix |

the expected classification cost based on ROC analysis. From the simple desiderata of consistency and sparsity of the classifier, Grandvalet et al. [10] introduced a piecewise linear and convex training criterion dedicated to abstaining classification which, in fact, extended the loss suggested by Bartlett et al. [13] to arbitrary asymmetric misclassification and rejection costs. Among the reject rules in Table I, there are only two reject rules introducing the reject option using pre-processing method, which is embedding rejection during the construction of SVMs, for the reason that it may bring some new problems of optimization that are hard to deal with. Post-processing methods are constructing the reject regions after SVMs have been trained. Compared with pre-processing methods, post-processing methods are simpler but have comparable performance.

In this paper, we propose a novel MI-based reject rule for SVM based on information theoretic learning [11]. It can be clearly verified from Table I that the main distinction of our MI-based reject rule is that we do not apply explicit cost matrix, but there is cost information embedded in our reject rule. Thus we let the data determine the optimal reject rule in the MI sense without any subjective cost information. Besides, the embedded cost information is also derived by us which can serve as the reference of selecting reasonable cost information of cost-sensitive learning.

The paper is organized as follows. Section III gives a brief introduction of SVM and MI measure for classification. The proposed MI-based reject rule is derived in Section IV and the embedded cost information is shown in Section V. Finally, Section VI gives the experimental results and Section VII concludes the whole paper.

## III. BACKGROUND

### A. SVM Classifiers

Supposing that an SVM classifier has been trained on a training set $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $t_i \in \{\pm 1\}$. We use positive to represent class $+1$ and negative to represent class $-1$. The decision function of SVMs is [5]:

$$y = \text{sgn}\Big(\sum_{\boldsymbol{x}_i \in SV_s} t_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b\Big), \tag{1}$$

where $k(\boldsymbol{x}_i, \boldsymbol{x})$ is a kernel function and $SV_s$ represents the set of support vectors. If we introduce the reject option into SVMs and the lower and upper reject thresholds respectively are $f_-$ and $f_+$ ($f_- \le f_+$), then the decision function is [9]:

$$y = \begin{cases} +1 & \text{if} \quad f(\boldsymbol{x}) \ge f_+, \\ 0\,(reject) & \text{if} \quad f_- \le f(\boldsymbol{x}) < f_+, \\ -1 & \text{if} \quad f(\boldsymbol{x}) < f_-, \end{cases} \tag{2}$$

where $f(\boldsymbol{x}) = \sum_{\boldsymbol{x}_i \in SV_s} t_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b$. Fig. 1 shows the effect of introducing two reject thresholds. Many of the errors caused by the class overlap are rejected. Truly, we can eliminate more errors by enlarging the reject region, but this has a disadvantage of eliminating more correct classification. Thus, we need to find an effective error-reject trade-off. In this paper, we use MI to find the error-reject trade-off.

### B. MI Measure for Classification

Hu et al. [14] presented a systematic study of information-theoretic measures for classification. They derived the Shannon-based MI measures based on an augmented confusion matrix $\boldsymbol{C}$:

$$\boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & c_{1(n+1)} \\ c_{21} & c_{22} & \cdots & c_{2n} & c_{2(n+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & c_{n(n+1)} \end{bmatrix},$$

where rows represent the true labels and columns represent the predicted labels; $n$ is the number of classes; $c_{ij}$ represents the number of samples of class $i$ classified as class $j$; and $n+1$ represents the reject option. Three features are proposed by them to evaluate a measure for abstaining classification: 1) monotonicity with respect to the diagonal terms of $\boldsymbol{C}$; 2) variation with the reject rate; 3) intuitively

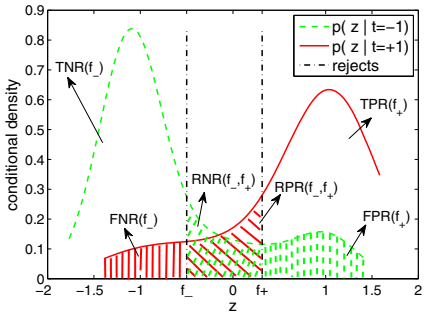


Figure 1. The positive and negative class-conditional densities obtained in an artificial dataset.

consistent costs among error types and reject types. Based on the above three features, Hu et al. [14] suggested the Shannon-based modified MI $I_m(T;Y)$ as the best MI measure for abstaining classification, where $T$ denotes the true label; $Y$ denotes the predicted label and the subscript "m" means modified:

$$I_m(T;Y) = \sum_t \sum_y p(t,y) \log \frac{p(t,y)}{p(t)p(y)} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{c_{ij}}{N} \log \frac{N c_{ij}}{N_i \hat{N}_j}, \quad (3)$$

where $N_i = \sum_{k=1}^{n+1} c_{ik}$, $\hat{N}_j = \sum_{k=1}^{n} c_{kj}, i, j = 1, \cdots, n$; and $N$ is the total sample number. Modified MI satisfies the above feature 2) and feature 3) which have been proved in [14]. Although it does not satisfy feature 1), it is also proper to evaluate the abstaining classification results because local minima only occurs in some extreme conditions. Based on their conclusions, Zhang and Hu [15] proposed cost-free learning which means getting the optimal classification results without any explicit cost information, and they succeeded to apply it to KNN and Bayesian classifiers. Inspired by their work, we also use modified MI (3) to get the modified-MI-optimal error-reject trade-off for SVMs. The difference between our work and their work is that we apply Newton method for the optimization which is more efficient than Powell's Algorithm [15]. We have a discussion about cost-free learning and the work of Zhang and Hu [15] in the experimental section. Without ambiguity, we also call "modified MI" MI for simplicity in this paper.

## IV. MI-Optimal Reject Rule for SVM

We suppose that we have get a trained binary SVM classifier and the augmented confusion matrix is shown in Table II with reject thresholds $f_-$ and $f_+$ ($f_- \le f_+$). Supposing that $\varphi_P(z) = p\big(f(\boldsymbol{x}) = z | t = +1\big)$ and $\varphi_N(z) = p\big(f(\boldsymbol{x}) = z | t = -1\big)$ respectively are the positive and negative class-conditional densities of SVM outputs, $TPR(f_+)$ (*true positive rate*), $FNR(f_-)$ (*false negative rate*), $FPR(f_+)$ (*false positive rate*), $TNR(f_-)$ (*true negative rate*), $RPR(f_-, f_+)$ (*reject positive rate*) and $RNR(f_-, f_+)$ (*reject negative rate*) respectively are (see Fig. 1) [9]:

$$\begin{aligned}
TPR(f_+) &= \frac{TP(f_+)}{N_+} = \int_{f_+}^{+\infty} \varphi_P(z)\mathrm{d}z; \\
FNR(f_-) &= \frac{FN(f_-)}{N_+} = \int_{-\infty}^{f_-} \varphi_P(z)\mathrm{d}z; \\
FPR(f_+) &= \frac{FP(f_+)}{N_-} = \int_{f_+}^{+\infty} \varphi_N(z)\mathrm{d}z; \\
TNR(f_-) &= \frac{TN(f_-)}{N_-} = \int_{-\infty}^{f_-} \varphi_N(z)\mathrm{d}z; \\
RPR(f_-, f_+) &= \frac{RP(f_-, f_+)}{N_+} = \int_{f_-}^{f_+} \varphi_P(z)\mathrm{d}z; \\
RNR(f_-, f_+) &= \frac{RN(f_-, f_+)}{N_-} = \int_{f_-}^{f_+} \varphi_N(z)\mathrm{d}z.
\end{aligned} \quad (4)$$

When the reject thresholds are $f_-$ and $f_+$, we use

Table II
THE AUGMENTED CONFUSION MATRIX OF A BINARY SVM CLASSIFIER WITH REJECT THRESHOLDS $f_-$ AND $f_+$. $RN(f_-, f_+)$ IS THE NUMBER OF REJECT NEGATIVES AND $RP(f_-, f_+)$ HAS A SIMILAR MEANING.

| | | Predict ($Y$) | | | $\sum$ |
|---|---|---|---|---|---|
| | | $+1$ | $-1$ | $0(Reject)$ | |
| True ($T$) | $+1$ | $TP(f_+)$ | $FN(f_-)$ | $RP(f_-, f_+)$ | $N_+$ |
| | $-1$ | $FP(f_+)$ | $TN(f_-)$ | $RN(f_-, f_+)$ | $N_-$ |
| $\sum$ | | $\hat{N}_+$ | $\hat{N}_-$ | $N_R$ | $N$ |

$I_m(f_-, f_+)$ to denote $I_m(T;Y)$ in order to simplify the expression. According to (3), our objective function is:

$$\begin{aligned} \max \quad & I_m(f_-, f_+) \\ s.t. \quad & f_- \le f_+, \end{aligned} \quad (5)$$

where $I_m(f_-, f_+)$ is given by:

$$\begin{aligned}
& I_m(f_-, f_+) \\
&= \tfrac{TP(f_+)}{N} \log \tfrac{TP(f_+)}{P(+1)\big(TP(f_+)+FP(f_+)\big)} \\
&+ \tfrac{FN(f_-)}{N} \log \tfrac{FN(f_-)}{P(+1)\big(FN(f_-)+TN(f_-)\big)} \\
&+ \tfrac{FP(f_+)}{N} \log \tfrac{FP(f_+)}{P(-1)\big(TP(f_+)+FP(f_+)\big)} \\
&+ \tfrac{TN(f_-)}{N} \log \tfrac{TN(f_-)}{P(-1)\big(FN(f_-)+TN(f_-)\big)},
\end{aligned} \quad (6)$$

where $P(+1) = \frac{N_+}{N}$ and $P(-1) = \frac{N_-}{N}$ are the prior probabilities of classes, where $N_+$ and $N_-$ are the numbers of positives and negatives respectively (see Table II). In (6), we use the convention that $0 \log 0 = 0$ and $0 \log \frac{0}{0} = 0$. In the next section, we introduce Newton method to optimize (6) in order to get the MI-optimal reject thresholds.

### A. Newton Method

According to the definitions of $TPR(f_+)$, $FNR(f_-)$, $FPR(f_+)$, and $TNR(f_-)$ (4), their derivatives with respect to $f_-$ and $f_+$ respectively are:

$$\begin{aligned}
\frac{dTPR(f_+)}{df_+} &= -\varphi_P(f_+); \quad \frac{dFPR(f_+)}{df_+} = -\varphi_N(f_+); \\
\frac{dTNR(f_-)}{df_-} &= \varphi_N(f_-); \quad \frac{dFNR(f_-)}{df_-} = \varphi_P(f_-).
\end{aligned} \quad (7)$$

Taking (4) into (6) and making use of (7), we can get the gradient $\boldsymbol{g}$ of $I_m(f_-, f_+)$:

$$\begin{aligned}
\boldsymbol{g}(1) =& \frac{\partial I_m(f_-, f_+)}{\partial f_-} \\
=& P(+1)\varphi_P(f_-) \log \frac{FN(f_-)}{P(+1)\big(FN(f_-)+TN(f_-)\big)} \\
&+ P(-1)\varphi_N(f_-) \log \frac{TN(f_-)}{P(-1)\big(FN(f_-)+TN(f_-)\big)}; \\
\boldsymbol{g}(2) =& \frac{\partial I_m(f_-, f_+)}{\partial f_+} \\
=& - P(+1)\varphi_P(f_+) \log \frac{TP(f_+)}{P(+1)\big(TP(f_+)+FP(f_+)\big)} \\
&- P(-1)\varphi_N(f_+) \log \frac{FP(f_+)}{P(-1)\big(TP(f_+)+FP(f_+)\big)}.
\end{aligned} \quad (8)$$

We further assume that $\psi_P(z)$ and $\psi_N(z)$ are the derivatives of $\varphi_P(z)$ and $\varphi_N(z)$ respectively. Taking the derivatives of (8) with respect to $f_-$ and $f_+$ again under the help of (4) and (7), we can get the Hessian matrix $H$ of $I_m(f_-, f_+)$:

$$\begin{aligned}
&H(1,1)\\
=&P(+1)\psi_P(f_-)\log\frac{FN(f_-)}{P(+1)\big(FN(f_-)+TN(f_-)\big)}\\
+&P(-1)\psi_N(f_-)\log\frac{TN(f_-)}{P(-1)\big(FN(f_-)+TN(f_-)\big)}\\
+&\frac{\big(\varphi_P(f_-)N_+TN(f_-)-\varphi_N(f_-)N_-FN(f_-)\big)^2}{N\cdot FN(f_-)TN(f_-)\big(FN(f_-)+TN(f_-)\big)};\\
&H(2,2)\\
=&-P(+1)\psi_P(f_+)\log\frac{TP(f_+)}{P(+1)\big(FP(f_+)+TP(f_+)\big)}\\
-&P(-1)\psi_N(f_+)\log\frac{FP(f_+)}{P(-1)\big(FP(f_+)+TP(f_+)\big)}\\
+&\frac{\big(\varphi_P(f_+)N_+FP(f_+)-\varphi_N(f_+)N_-TP(f_+)\big)^2}{N\cdot FP(f_+)TP(f_+)\big(FP(f_+)+TP(f_+)\big)};\\
&H(1,2)=H(2,1)=0;
\end{aligned}\tag{9}$$

where we use the convention that $\frac{0}{0}=0$. We want to use Newton method to optimize (5), but we have no idea of the values of $\varphi_P(z)$, $\varphi_N(z)$, $\psi_P(z)$ and $\psi_N(z)$. Fortunately, we can estimate them from the training set using Parzen Window method [16]. With the help of Parzen Window method, we can apply Newton method to optimize problem (5):

$$(f_-,f_+)_k^T=(f_-,f_+)_{k-1}^T+\alpha H_{k-1}^{-1}\boldsymbol{g}_{k-1},\tag{10}$$

where $\alpha$ is the learning rate. Newton method has the property of quadratic convergence and in our experiments, the optimization usually converges around six iterations.

## V. The Embedded Cost Information

Cost-sensitive learning is very popular in abstaining classification and the expected classification cost is usually minimized through ROC analysis [3], [9]. The main problem with cost-sensitive learning is unavailability of proper cost information about misclassification and rejection. We think that there is cost information embedded in maximizing $I_m(f_-, f_+)$. As the cost matrix has the properties of scaling invariance and shifting invariance [17], it can be normalized to the one shown in Table III. The elements in the cost matrix should satisfy the following constraints [3] so that the reject option is applicable:

$$\begin{cases}0<CRP<CFN;\\ 0<CRN<1;\\ CRP+CRN\cdot CFN<CFN.\end{cases}\tag{11}$$

Then the expected classification cost is:

$$\begin{aligned}
&EC(f_-,f_+)\\
=&P(+1)\cdot FNR(f_-)\cdot CFN+P(+1)\cdot RPR(f_-,f_+)\cdot CRP\\
+&P(-1)\cdot FPR(f_+)+P(-1)\cdot RNR(f_-,f_+)\cdot CRN.
\end{aligned}\tag{12}$$

Table III
Cost matrix for binary abstaining classification

| | | Predict ($Y$) | | |
|---|---|---|---|---|
| | | $+1$ | $-1$ | $0(Reject)$ |
| True ($T$) | $+1$ | 0 | $CFN$ | $CRP$ |
| | $-1$ | 1 | 0 | $CRN$ |

Similar to the derivation of the gradient and Hessian matrix of $I_m(f_-, f_+)$, the gradient $\boldsymbol{g}_{ec}$ of $EC(f_-, f_+)$ is:

$$\begin{aligned}
\boldsymbol{g}_{ec}(1)=&\frac{\partial EC(f_-,f_+)}{\partial f_-}\\
=&P(+1)\varphi_P(f_-)(CFN-CRP)-P(-1)\varphi_N(f_-)CRN;\\
\boldsymbol{g}_{ec}(2)=&\frac{\partial EC(f_-,f_+)}{\partial f_+}\\
=&P(+1)\varphi_P(f_+)CRP+P(-1)\varphi_N(f_+)(CRN-1);
\end{aligned}\tag{13}$$

the Hessian matrix $H_{ec}$ of $EC(f_-, f_+)$ is:

$$\begin{aligned}
&H_{ec}(1,1)=P(+1)\psi_P(f_-)(CFN-CRP)-P(-1)\psi_N(f_-)CRN;\\
&H_{ec}(2,2)=P(+1)\psi_P(f_+)CRP+P(-1)\psi_N(f_+)(CRN-1);\\
&H_{ec}(1,2)=H_{ec}(2,1)=0.
\end{aligned}\tag{14}$$

In fact, given a cost matrix, we can also use Newton method to minimize the expected classification cost (12). However, this approach has a drawback of re-optimization if the cost matrix is changed while in such condition, the optimization method based on ROC analysis [9] can get the optimal reject thresholds directly from the estimated ROC curve. We use $\boldsymbol{g}_{ec}$ to calculate the cost information embedded in maximizing $I_m(f_-, f_+)$. We assume that the solution of problem (5) is $(f_-^*, f_+^*)^T$. Because there are three variables in the cost matrix, we have to provide a constraint among $CFN$, $CRP$ and $CRN$ in order to obtain a unique solution. We consider two kinds of constraints here, one is $CRN=CRP$ and the other is $CRN=\eta$, $\eta$ is a given constant $(0<\eta<1)$. Then we can obtain the solutions of $CFN$, $CRP$ and $CRN$ by solving $\boldsymbol{g}_{ec}\big|_{(f_-,f_+)^T=(f_-^*,f_+^*)^T}=0$ with one of the above constraints. If the constraint is $CRN=CRP$, then the solutions are:

$$\begin{cases}CRN=CRP=\frac{P(-1)\varphi_N(f_+^*)}{P(+1)\varphi_P(f_+^*)+P(-1)\varphi_N(f_+^*)};\\ CFN=CRP+CRP\cdot\frac{P(-1)\varphi_N(f_-^*)}{P(+1)\varphi_P(f_-^*)}.\end{cases}\tag{15}$$

If the constraint is $CRN=\eta, 0<\eta<1$, then the solutions are:

$$\begin{cases}CRP=(1-\eta)\cdot\frac{P(-1)\varphi_N(f_+^*)}{P(+1)\varphi_P(f_+^*)};\\ CFN=CRP+\eta\cdot\frac{P(-1)\varphi_N(f_-^*)}{P(+1)\varphi_P(f_-^*)}.\end{cases}\tag{16}$$

The above embedded cost information is important to us. Firstly, it establishes the relationship between our MI-based reject rule and cost-sensitive learning. Secondly, if we are unsatisfied with the abstaining classification results of our MI-based reject rule, we can choose the proper cost information for practical requirements based on this embedded

cost information, and then use cost-sensitive learning [3], [9] to re-optimize the reject thresholds. Because we use the abstaining classification results to calculate the cost information which is an inverse process of cost-sensitive learning, the obtained embedded cost information always satisfies the constraints (11).

### A. Relation between Two Reject Rules

In [23], Hu systematically describes the relationship between Bayesian classifiers and MI classifiers which helps us understand classification from MI perspective. Under his framework, we briefly describe the relationship between our MI-based reject rule and the reject rule of Grandvalet et al. [10] through the embedded cost information (Table III). The abstaining decision function in [23] is:

$$y = \begin{cases} +1 & \text{if} \quad p(+1|\boldsymbol{x}) \geq 1 - T_{r2}, \\ 0\,(reject) & \text{if} \quad T_{r1} \leq p(+1|\boldsymbol{x}) < 1 - T_{r2}, \\ -1 & \text{if} \quad p(+1|\boldsymbol{x}) < T_{r1}, \end{cases} \tag{17}$$

where $T_{r1}$ and $T_{r2}$ are two thresholds which reflect Chow's reject rule [2]. $T_{r1}, T_{r2} > 0$ and $0 < T_{r1} + T_{r2} \leq 1$ [23] so that the rejection is applicable. They can be computed using the embedded cost information in Table III:

$$T_{r1} = \frac{CRN}{CRN + CFN - CRP}, T_{r2} = \frac{CRP}{CRP + 1 - CRN}. \tag{18}$$

Grandvalet et al. [10] built up abstaining SVM classifiers based on a probabilistic interpretation of SVM that the hinge loss can be viewed as a relaxed minimization of negative log-likelihood. Thus, instead of estimating $p(+1|\boldsymbol{x})$ in the whole range, they believe that the estimation of $p(+1|\boldsymbol{x})$ can only be accurate in the vicinity of $T_{r1}$ and $T_{r2}$. Fig. 2 intuitively shows the relationship between MI-based reject rule and the reject rule of Grandvalet et al. [10]. The loss functions of negatives and positives, which are used in the reject rule of Grandvalet et al., are different that we redefine as "dual double-hinge losses":

$$\begin{aligned} l_- &= \max\{T_{r1} f(\boldsymbol{x}) + H(T_{r1}), (1 - T_{r2}) f(\boldsymbol{x}) + H(1 - T_{r2}), 0\}, \\ l_+ &= \max\{-(1 - T_{r1}) f(\boldsymbol{x}) + H(T_{r1}), -T_{r2} f(\boldsymbol{x}) + H(1 - T_{r2}), 0\}, \end{aligned} \tag{19}$$

where $H(q) = -q\log(q) - (1-q)\log(1-q)$. And the corresponding reject thresholds $(f_-, f_+)^T$ of SVM is computed as follows:

$$f_- = \log(\frac{T_{r1}}{1 - T_{r1}}), f_+ = \log(\frac{1 - T_{r2}}{T_{r2}}). \tag{20}$$

Along the $f(\boldsymbol{x})$ axis, the two hinge points $f_1$ and $f_2$, and another point $f_3$ respectively are:

$$\begin{aligned} f_1 &= f_- + \frac{\log(1 - T_{r1})}{T_{r1}}, f_2 = f_+ - \frac{\log(1 - T_{r2})}{T_{r2}}, \\ f_3 &= \frac{-T_{r1} f_- - T_{r2} f_+ + \log\frac{1-T_{r2}}{1-T_{r1}}}{1 - T_{r1} - T_{r2}}. \end{aligned} \tag{21}$$

Within each double-hinge loss function, e.g. $l_+$, four regions are formed along the $f(\boldsymbol{x})$ axis with different meanings, namely, *error region* $f(\boldsymbol{x}) < f_-$, *reject region* $f_- \leq f(\boldsymbol{x}) < f_+$, *correct region with penalty* $f_+ \leq f(\boldsymbol{x}) < f_2$ and *correct region without penalty* $f(\boldsymbol{x}) \geq f_2$. The *correct region with penalty* is necessary for a unique classification solution. From the constraint $0 < T_{r1} + T_{r2} \leq 1$, we can easily verify that $abs(k_-^1) \geq abs(k_-^2)$ and $abs(k_+^1) \geq abs(k_+^2)$ which indicate that a misclassification will cost more than a reject in the same class. The above is a typical relationship between our MI-based reject rule and another reject rule through the embedded cost information.

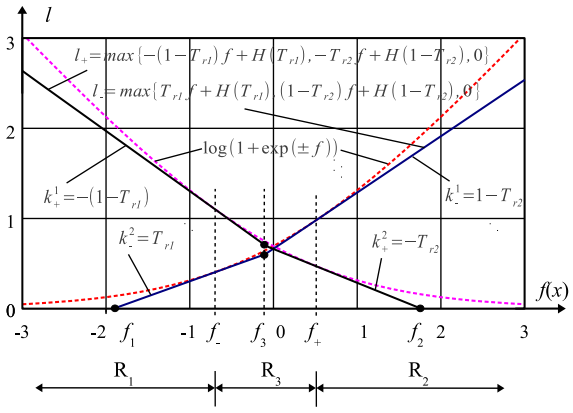


Figure 2. Graphical descriptions of dual double-hinge losses (modified based on [10]).

Table IV
CHARACTERISTICS OF THE REAL-WORLD DATASETS

| Dataset | # of Samples | # of Positive Samples(%) | # of Features |
|---|---|---|---|
| WDBC | 569 | 212 (37.26) | 30 |
| Pima | 768 | 268 (34.90) | 8 |
| German Credit | 1000 | 300 (30.00) | 24 |
| KC1 | 2109 | 326 (15.46) | 21 |
| Mammography | 11180 | 260 (2.33) | 6 |

## VI. EXPERIMENTS

We implement the experiments with the help of libsvm [18]. Table IV presents the characteristics of the real-world datasets. KC1[1] is from the domain of software engineering measurements; Mammography was generously provided by Dr. Nitesh Chawla [19]; and the remaining three datasets are from the UCI data repository. To reduce the bias in comparison, we use 10-fold cross-validation in all the datasets. In each of the runs, we select 60% of the whole dataset as the training set, 20% as the validation set and the remaining 20% as the test set. For imbalanced dataset, positive denotes the minority class and negative denotes the majority class. Although in some imbalanced datasets there are only a few positives in the validation set, experimental results show that Newton method can also converge.

Firstly, the original nonabstaining classification results are presented in order to show the merits of abstaining classification. Secondly, to evaluate the effectiveness of our

[1]http://promise.site.uottawa.ca/SERepository/datasets-page.html

MI-based reject rule, we choose **Chow's** reject rule and MI-based cross-validation reject rule (**MICVR**) as baselines. We apply Chow's reject rule after mapping SVM outputs into posterior probabilities using Platt's method [12] and we set the "rejection threshold" of Chow's reject rule as 0.4 [2]. MICVR is doing cross-validation based on MI to select the optimal reject thresholds. In our MI-based reject rule, during the optimization of (5), we randomly choose the initial points around zero several times and then pick the reject thresholds with the highest MI. There are two reasons that we do not compare our MI-based reject rule with the reject rules based on cost-sensitive learning. One is that we assume that there is no cost information at hand, the other is that our MI-based reject rule is related to cost-sensitive learning which will be experimentally discussed in the following section.

### A. An Illustrative Example

We give a toy example of MI-based reject rule in Fig. 3(a) where the classes are balanced. Because of noise, instances around the decision boundary of SVM are more prone to be misclassified which had better better been rejected for further investigation. After using MI-based reject rule in this artificial dataset, we obtain two reject thresholds $f_+ = 0.6470$ and $f_- = -0.5609$. We can easily find that these two reject thresholds just locate on the places that all the misclassifications are rejected. It intuitively shows the effectiveness of MI-based reject rule in finding the reasonable error-reject tradeoff. The embedded cost information with constraint $CRP = CRN$ is $CFN = 1.0050$ and $CRP = CRN = 0.3883$. And Fig. 3(b) shows the dual double-hinge losses used in the reject rule of Grandvalet et al. [10].

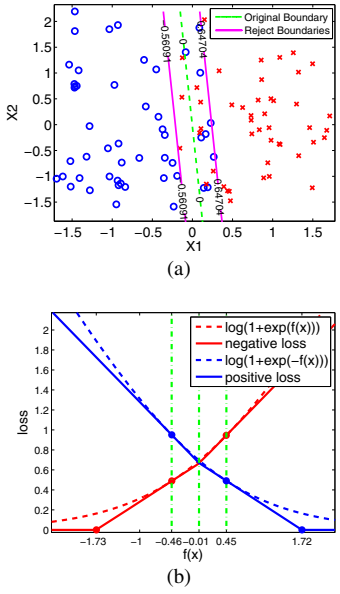


Figure 3. (a) A toy example of reject boundaries of MI-based reject rule; (b) The dual double-hinge loss functions of Grandvalet et al. [10].

### B. Abstaining Classification Results

As far as we know, there is no general measure to evaluate abstaining classification until now. We apply normalized MI (NMI) to evaluate the abstaining classification results which is proposed by Hu et al. [14] and is used in [15]. NMI is the ratio of MI over the entropy of true class distributions which is a constant. Besides NMI, we also use *error rate* and *reject rate* to reflect the error-reject trade-off; *FNR*, *RPR* and *precision* to evaluate the classification of positives. Table V shows the experimental classification results.

Firstly, the *error rates* of abstaining classification are smaller than nonabstaining classification thanks to the reject option. As a result, we can make less wrong predictions in reality and make more careful predictions of those rejected instances by further investigation. Besides, in the imbalanced datasets like KC1 and Mammography where we care more about the positive classification, abstaining classification reduces *FNRs* through proper *RPRs* and improves *precisions* of positive classification.

Secondly, in terms of NMI, Chow's reject rule performs the worst among all the classification methods. In the balanced datasets, WDBC and Pima, compared with our MI-based reject rule, Chow's reject rule has higher *error rates* but lower *reject rates*, and we can choose one of them based on practical requirements. While in the imbalanced datasets, KC1 and Mammography, *error rate* is insufficient to evaluate the classification results, and people concern more with positive classification. Although *RPRs* of our reject rule are higher, *FNRs* of our reject rule are much smaller than Chow's reject rule. Thus our reject rule would rather reject a positive instance than make a wrong positive classification. In fact, in imbalanced datasets, Chow's reject rule even performs worse than nonabstaining classification. This is because the performance of Chow's reject rule depends on the estimation of posterior probability [20] which is seldom sufficiently accuracy. Above all, our MI-based reject rule can give reasonable abstaining classification results in both balanced and imbalanced datasets.

Thirdly, the purpose of making comparison between MICVR and our MI-based reject rule is to evaluate the efficiency of Newton method. We use student test with unequal variance at the 1% significance level [21] to test the difference of NMI between these two reject rules. The result is that there is no statistical difference between them which reflects that Newton method is efficient in optimizing (5). Except in the German Credit and KC1 datasets, the reject thresholds of MICVR and our MI-based reject rule are very close. In our experiments, Newton method usually converges around six iterations which is computationally efficient while MICVR has a heavy time complexity,

Table V
ABSTAINING CLASSIFICATION RESULTS OF THREE REJECT RULES AND NONABSTAINING CLASSIFICATION RESULTS OF SVM. THE NUMBER BELOW THE DATASET IS THE RATIO OF POSITIVES OVER NEGATIVES. MEAN(STD)

| Dataset | Method | NMI | Error Rate(%) | Reject Rate(%) | FNR(%) | RPR(%) | Precision(%) | Reject Thresholds $[f_-, f_+]^T$ |
|---|---|---|---|---|---|---|---|---|
| WDBC (1:1.68) | Chow's | 0.8526(0.0558) | 2.00(0.93) | 1.38(1.24) | 2.55(1.96) | 1.51(1.87) | 97.30(2.28) | $[-0.1056(0.0211), 0.1097(0.0216)]^T$ |
| | MICVR | 0.8776(0.0508) | 1.07(0.97) | 5.59(5.06) | 1.94(2.07) | 3.70(4.06) | 99.16(1.74) | $[-0.3636(0.3640), 0.2271(0.1373)]^T$ |
| | Ours | **0.8971(0.0479)** | 0.88(0.70) | 4.30(2.61) | 1.89(1.71) | 3.45(3.11) | 99.62(0.90) | $[-0.2958(0.1927), 0.2236(0.1039)]^T$ |
| | NC | 0.8293(0.0888) | 2.73(1.60) | —— | 3.30(2.76) | —— | 96.09(2.96) | —— |
| Pima (1:1.87) | Chow's | 0.1796(0.0038) | 20.07(2.16) | 17.14(3.15) | 30.59(4.99) | 22.31(6.10) | 73.80(6.62) | $[0.0890(0.0240), 0.6921(0.0369)]^T$ |
| | MICVR | 0.2414(0.0456) | 16.42(4.98) | 34.28(10.83) | 9.13(6.13) | 36.42(13.45) | 72.46(8.04) | $[-0.6591(0.2343), 0.5182(0.2976)]^T$ |
| | Ours | **0.2491(0.0416)** | 16.41(3.89) | 31.13(6.15) | 9.24(4.36) | 34.59(9.72) | 71.29(5.96) | $[-0.5692(0.1533), 0.4753(0.1342)]^T$ |
| | NC | 0.1808(0.0567) | 25.83(2.73) | —— | 26.45(8.99) | —— | 61.24(2.31) | —— |
| German Credit (1:2.33) | Chow's | 0.1167(0.0493) | 19.62(3.22) | 18.44(3.43) | 42.00(6.26) | 31.15(9.09) | 70.69(12.44) | $[0.2046(0.0272), 0.8373(0.0679)]^T$ |
| | MICVR | 0.1631(0.0413) | 21.78(5.14) | 36.37(11.37) | 13.00(7.24) | 34.63(13.90) | 61.76(7.97) | $[-0.6629(0.2482), 0.3218(0.2258)]^T$ |
| | Ours | **0.1665(0.0441)** | 20.48(2.87) | 21.06(5.62) | 25.67(5.70) | 23.22(6.90) | 64.68(5.80) | $[-0.1854(0.0922), 0.3603(0.0887)]^T$ |
| | NC | 0.1306(0.0437) | 27.78(3.30) | —— | 32.59(6.45) | —— | 53.12(4.52) | —— |
| KC1 (1:5.47) | Chow's | 0.0074(0.0045) | 12.86(0.74) | 8.02(2.27) | 76.08(6.16) | 23.58(5.99) | 10.74(23.88) | $[1.0085(0.0510), 1.6862(0.0827)]^T$ |
| | MICVR | **0.1617(0.0376)** | 26.11(6.28) | 27.20(10.02) | 8.35(3.37) | 27.20(15.55) | 34.83(5.04) | $[-0.8644(0.1277), 0.2636(0.4415)]^T$ |
| | Ours | 0.1589(0.0399) | 22.79(3.69) | 24.12(4.44) | 12.34(4.65) | 32.97(7.47) | 35.75(3.69) | $[-0.5913(0.0890), 0.4496(0.1134)]^T$ |
| | NC | 0.1182(0.0378) | 28.47(2.15) | —— | 29.36(7.03) | —— | 31.36(2.55) | —— |
| Mammography (1:42) | Chow's | 0.1853(0.0456) | 1.67(0.15) | 0.72(0.22) | 58.08(6.03) | 14.66(5.17) | 68.06(8.22) | $[1.3597(0.0631), 1.7746(0.0789)]^T$ |
| | MICVR | 0.4947(0.0403) | 2.33(0.58) | 8.18(2.39) | 9.06(4.41) | 16.37(4.83) | 48.41(6.46) | $[-0.6496(0.2204), 0.6382(0.1794)]^T$ |
| | Ours | **0.5026(0.0429)** | 2.35(0.38) | 6.20(1.15) | 9.96(4.50) | 13.55(3.93) | 47.86(4.22) | $[-0.5091(0.1087), 0.5837(0.0901)]^T$ |
| | NC | 0.4373(0.0342) | 4.67(0.34) | —— | 15.13(3.47) | —— | 31.45(1.79) | —— |

Table VI
THE EMBEDDED COST INFORMATION. THE NUMBER BESIDE THE DATASET IS THE RATIO OF POSITIVES OVER NEGATIVES. MEAN(STD)

| | Constraint: CRP=CRN | | Constraint: CRN=0.4 | |
|---|---|---|---|---|
| Dataset | CFN | CRP=CRN | CFN | CRP |
| WDBC(1:1.68) | 1.701(0.896) | 0.290(0.065) | 2.200(0.883) | 0.253(0.075) |
| Pima(1:1.87) | 2.612(0.706) | 0.480(0.039) | 2.218(0.372) | 0.540(0.079) |
| German Cr.(1:2.33) | 2.573(0.231) | 0.548(0.036) | 2.257(0.230) | 0.737(0.098) |
| KC1(1:5.47) | 8.551(0.924) | 0.758(0.015) | 6.077(0.818) | 1.867(0.159) |
| Mammography(1:42) | 85.77(12.23) | 0.836(0.016) | 44.29(5.193) | 3.160(0.429) |

especially when the search interval becomes larger and the step size becomes smaller.

### C. The Results of Embedded Cost Information

In Section V, we have discussed that there is cost information embedded in maximize $I_m(f_-, f_+)$. Table VI shows the results of embedded cost information with constraints $CRP = CRN$ and $CRN = 0.4$. It is verified that all the cost matrices satisfy constraints (11). Moreover, given an embedded cost matrix, if we use Newton method proposed in Section V to minimize (12), we will get the same reject thresholds. From the results of embedded cost information with constraint $CRN = 0.4$, we find that the negative misclassification cost may be lower than the positive rejection cost which, we think, sometimes is rational. If the abstaining classification results of our MI-based reject rule do not meet the practical requirements, we can estimate the cost matrix based on the embedded cost information and then use cost-sensitive learning to get the optimal reject thresholds. In all, this embedded cost information can help us find the reasonable cost information more easily and quickly.

### D. Discussion

By making comparison with Chow's reject rule and MICVR, we come to the conclusion that our MI-based reject rule can give reasonable abstaining classification results and Newton method is efficient in optimizing (5). Zhang and Hu proposed cost-free learning in [15] which is to learn to get the optimal classification results without using any explicit cost information. It is easy to verify that our MI-based reject rule is a kind of cost-free learning. From another perspective, we can view cost-free learning as cost-sensitive learning with embedded cost information. It is pointed out that the embedded cost information can be derived by analysing the relationship between cost-free learning and cost-sensitive learning which has already been shown in Section V and in [15]. We are not sure that whether abstaining classification results of cost-free learning satisfy practical requirements, but its embedded cost information could help us to find the proper cost information more easily and quickly. Overall, cost-sensitive learning is more flexible and powerful. However, as Hu [23] said, cost-free learning was proposed to enrich the content of cost-sensitive learning and let us understand cost-sensitive learning from a different viewpoint. Furthermore, we can let the data speak for themselves firstly and then embed our prior knowledge into the classifiers.

In this paper, we extend cost-free learning from KNN and Bayesian classifiers [15] to SVM which is more popular in classification. Besides, compared with Powell's algorithm

used by Zhang and Hu [15], Newton method is more efficient for the optimization.

## VII. Conclusion

The reject option is a key option in classification where errors cause severe losses. Then the rejected samples can be classified more carefully. Ferri et al. [22] have used this principle to build delegating classifiers. In this paper, we extend the cost-free learning of Zhang and Hu [15] to SVM and develop a novel MI-based reject rule for SVM which is a method of post-processing. We derive the gradient and Hessian matrix of MI between the true labels and the predicted labels so that Newton method, which is more efficient than Powell's algorithm, can be used to optimize the MI-based objective function. Besides, we also obtain the cost information embedded in MI which reflects the relationship between our MI-based reject rule and cost-sensitive learning, and furthermore, if the abstaining classification results of our MI-based reject rule are not satisfied, the embedded cost information can help us find the proper cost information that meets the practical requirements more easily and quickly. Conventional measures for classification such as *G-mean* and *F-measure* can not be used to evaluate abstaining classification because the reject option is neglected by them. In the future, we will further investigate how to properly evaluate abstaining classification results.

## Acknowledgment

Thanks the anonymous reviewers for their valuable comments to improve the paper. This work is supported by NSFC grants #61075051 and #61273196.

## References

[1] C. Ferri, J. Hernández-Orallo, "Cautious classifiers," In Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence, pp. 27–36, 2004.

[2] C. Chow, "On optimum recognition error and reject tradeoff," IEEE Transactions on Information Theory, Vol. 16, No. 1, pp. 41–46, Jan. 1970.

[3] T. Pietraszek, "Optimizing abstaining classifiers using ROC analysis," In Proceedings of International Conference on Machine Learning, pp. 665–672, 2005.

[4] C. Friedel, U. Rückert, S. Kramer, "Cost curves for abstaining classifiers," In Proceedings of the ICML 2006 Workshop on ROC Analysis in Machine Learning, pp. 33–40, 2006.

[5] V. Vapnik, The nature of statistical learning theory, 1st ed., Springer-Verlag New York, New York, 1995.

[6] J. Shawe-Taylor, "Classification accuracy based on observed margin,", Algorithmica, Vol. 22, No. 1–2, pp. 157–172, Springer, Sept. 1998.

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

[8] G. Fumera, F. Roli, "Support vector machines with embedded reject option," Pattern Recognition with Support Vector Machines, pp. 68–82, 2002.

[9] F. Tortorella, "Reducing the classification cost of support vector classifiers through an ROC-based reject rule," Pattern Analysis & Applications, Vol. 7, No. 2, pp. 128–143, July 2004.

[10] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, S. Canu, "Support vector machines with a reject option," Advances in Neural Information Processing Systems, pp. 537–544, 2009.

[11] J. C. Principe, Information theoretic learning: Renyi's entropy and kernel perspectives, Springer, 2010.

[12] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," In Proceedings of Advances in Large Margin Classifiers, Vol. 10, No. 3, pp. 61–74, MIT Press, 1999.

[13] P. L. Bartlett, M. H. Wegkamp, "Classification with a reject option using a hinge loss," Journal of Machine Learning Research, Vol. 9, pp. 1823–1840, June 2008.

[14] B.G. Hu, R. He, X.T. Yuan, "Information-theoretic measures for objective evaluation of classifications," Acta Automatica Sinica, Vol. 38, No. 7, pp. 1169–1182, July 2012.

[15] X.W. Zhang, B.G. Hu, "Learning in the class imbalance problem when costs are unknown for errors and rejects," 12th International Conference on Data Mining Workshops, pp. 194–201, 2012.

[16] E. Parzen, "On estimation of a probability density function and mode," The Annals of Mathematical Statistics, Vol. 33, No. 3, pp. 1065–1076, Sept. 1962.

[17] C. Elkan, "The foundations of cost-sensitive learning," International Joint Conference on Artificial Intelligence, pp. 973–978, 2001.

[18] C.C. Chang, C.J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 27, pp. 1–27, April 2011.

[19] N. V. Chawla, K. W. Bowyer, T. E. Moore, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal Of Artificial Intelligence Research, Vol. 16, pp. 321–357, June 2002.

[20] G. Fumera, F. Roli, G. Giacinto, "Reject option with multiple thresholds," Pattern Recognition, Vol. 33, No. 12, pp. 2099–2101, March 2000.

[21] E. L. Lehmann, J. P. Romano, Testing statistical hypotheses, Springer, New York, 2005.

[22] C. Ferri, P. Flach, J. Hernández-Orallo, "Delegating classifiers," In Proceedings of the 21st International Conference on Machine Learning, 2004.

[23] B.G. Hu, "What are the differences between Bayesian classifiers and mutual-information classifiers?," IEEE Transactions on Neural Networks and Learning Systems, August 2013, in press.