International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011
1793-8201

Cost-sensitive SVM with Error Cost and Class-dependent Reject Cost

En-hui Zheng, Chao Zou, Jian Sun, Le Chen

Abstract—**In such real data mining applications as medical diagnosis, fraud detection and fault classification, and so on, the two problems that the error cost is expensive and the reject cost is class-dependent are often encountered. In order to overcome those problems, firstly, the general mathematical description of the Binary Classification Problem with Error Cost and Class-dependent Reject Cost (BCP-EC2RC) is proposed. Secondly, as one of implementation methods of BCP-EC2RC, the new algorithm, named as Cost-sensitive Support Vector Machines with the Error Cost and the Class-dependent Reject Cost (CSVM-EC2RC), is presented. The CSVM-EC2RC algorithm involves two stages: estimating the classification reliability based on trained SVM classifier, and determining the optimal reject rate of positive class and negative class by minimizing the average cost based on the given error cost and class-dependent reject cost. The experiment studies based on a benchmark data set illustrate that the proposed algorithm is effective.**

Index Terms—**SVM, cost-sensitive, error cost, class-dependent reject cost.**

I. Introduction

In pattern classification community, the main concern of conventional algorithms of designing classifier is to minimize the error rate based on the assumption of both 0-1 loss and all classification results being reliable. However, in such fields as fraud detection, medical diagnosis, and so on, the high classification reliability is necessary in order to avoid running the high risk of misclassification [1, 2]. It is reasonable to reject the example whose classification reliability is less than a specific threshold that is we do not believe in the results of automatic classification with low reliability. In current literatures, the reject option is adopted to achieve the best tradeoff between error rate and reject rate [3-5]. On the basis of multi-expert system, Foggia (1999) presents a method of determining the best tradeoff between the error rate and reject rate based on the estimation of the reliability of each classification act and on the evaluation of the convenience of rejecting the input example when the reliability is under a domain-dependent threshold [3]. Thomas (2006) proposed a two-stage classifier to explore the interaction between classification performance and reject performance for distance-based reject-option classifiers [4]. Claudio (2000) extended Foggia's results and adapted the behavior of the reject option to the requirements of the considered application domain [5]. The reject option involves two stages: 1) evaluating the classification reliability [3]; 2) rejecting the unreliable classified samples with a fixed threshold [5, 6] or an optimum one [7].

In present paper, we define the reject cost as overall cost (or loss) resulted from the introduction of the reject act, and consider that the cost of rejecting a positive example is unequal to that of rejecting negative one. For example, in medical diagnosis, the cost of rejecting a patient example is bigger than that of rejecting healthy one. The former involves both deteriorating the state of an illness (even the loss of life) and further diagnosis, while the latter only involves further diagnosis. Obviously, the cost of the former is higher than that of the latter, which is the reject cost is class-dependent, which can not be found in current researches. In those conditions, the currently-available algorithms without taking the assumption of reject cost being class-dependent into account do not perform well [8]. In order to solve the above problem, firstly, the general mathematic description of Binary Classification Problem with Error Cost and Class-dependent Reject Cost (BCP-EC^{2}RC) is proposed. Secondly, as one of implementation methods of solving the BCP-EC^{2}RC problem, the novel algorithm, named as Cost-sensitive Support Vector Machines with Error Cost and Class-dependent Reject Cost (CSVM-EC^{2}RC), is presented and is derived by integrating the above costs into SVM. The researches in this paper belong to the context of Cost-sensitive Learning and Cost-sensitive Data Mining [8-11].

The rest of this paper is organized as follows. In Section II, the standard SVM is reviewed. In Section III, we formulate the binary classification problem with error cost and class-dependent reject cost. The algorithm procedure is realized in Section IV. In Section V, the numerical experiments on UCI data sets are presented. Finally, conclusions and future works are eventually reported in Section VI.

Manuscript received December 18, 2009. This work was supported by the Natural Science Foundation of Zhejiang Province P. R. China under Grant Y1080950, the National Natural Science Foundation of P. R. China under Grant 60905034 and the National Department Public Benefit Research Foundation of P. R. China under Grant 2007GYJ016.

En-hui Zheng, corresponding author of this paper, is with the China Jiliang University, Hangzhou, CO 310018 P. R. China (phone: +86-571-86914549 5; e-mail: ehzheng@cjlu.edu.cn).

Chao Zou is with the China Jiliang University, Hangzhou, CO 310018 P. R. China (e-mail: zouc@cjlu.edu.cn).

Jian Sun is with the China Jiliang University, Hangzhou, CO 310018 P. R. China (e-mail: sunjian@cjlu.edu.cn).

Le Chen is with the China Jiliang University, Hangzhou, CO 310018 P. R. China (e-mail: cl7788@126.com).

Ping Li is with the Zhejiang University, Hangzhou, CO 310027 P. R. China (e-mail: pli@iipc.zju.edu.cn).

II. Support Vector Machines

Support Vector Machines (SVM), motivated by results of statistical learning theory, are a new class of machine learning algorithms, which capture the main insight of statistical learning theory: in order to obtain a small risk function, one needs to control both training error and model complexity. Instead of the local minima of neural networks, the optimization for constructing SVM has a unique solution; therefore SVM has the well generalization ability and solid theoretical foundation.

In SVM, we are given the following training data

$$\begin{aligned} &(x_1, y_1), ..., (x_i, y_i), ..., (x_k, y_k), \\ &x_i \in R^l, \; y_i \in \{+1, -1\}, \; i = 1, ..., k, \end{aligned} \tag{1}$$

where k denotes the number of examples, and l denotes the dimension number of predictive attributes x_i. Suppose the hyper-plane $(\boldsymbol{x} \cdot \boldsymbol{w}) - b = 0$ is used to classify the training data (1), the task of SVM algorithm is to minimize the expected cost

$$\begin{aligned} &R(\boldsymbol{w}, \xi) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C(\sum_{i=1}^{k} \xi_i), \\ &s.t. \;\; y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \geq 1 - \xi_i \;, \;\; \xi_i \geq 0 \;, \;\; i = 1, ..., k, \end{aligned} \tag{2}$$

where $\|\boldsymbol{w}\|^2$ and $\sum_{i=1}^{n} \xi_i$ approximate the model complexity and the empirical risk, respectively, and the regularization constant $C > 0$ determines the tradeoff between the model complexity and the empirical risk [6].

III. General Mathematical Description of the Binary Classification Problem with Error Cost and Class-dependent Reject Cost

For the binary classification problem, suppose we are given a set of labeled training data as following

$$\begin{aligned} &(x_1, y_1, r_1, m_1), ..., (x_i, y_i, r_i, m_i), ..., (x_n, y_n, r_n, m_n), \\ &x_i \in R^l, \;\; y_i \in \{n, p\}, \;\; i = 1, 2, ..., k, \end{aligned} \tag{3}$$

where n and p denote the label of negative class (i.e., -1 class) and that of positive class (i.e., +1 class), respectively. m_i, $m_i \in R$ and r_i, $r_i \in R$, $i = 1, 2, ..., k$ are the error cost and reject cost of the i th example, respectively.

Let r_n and r_p be the cost of rejecting an example of negative class and that of positive class, respectively, and σ_n and σ_p are the reject threshold of negative class and that of positive class, respectively. Suppose that the constrains $m_i = m_0$ and $r_i \in \{r_n, r_p\}$, $i = 1, 2, ..., k$, are satisfied. Let P_r^n and P_r^p denote the reject rate of negative class and that of positive class, and P_e the error rate of all examples. Based on the above settings, suppose that the reject thresholds (σ_n, σ_p) be given, the learning problem BCP-EC^{2}RC minimizes the average cost

$$P_r^n \cdot r_n + P_r^p \cdot r_p + P_e \cdot m_0, \tag{4}$$

where P_r^n, P_r^p and P_e in (4) are estimated according to

$$\begin{aligned} &P_r^n \approx \frac{1}{k}\sum_{i=1}^{k} a_i, \; P_r^p \approx \frac{1}{k}\sum_{i=1}^{k} b_i, \\ &P_e \approx \frac{1}{k}\sum_{i=1}^{k} c_i \cdot sign(|\, f(x_i) - y_i \,|), \end{aligned} \tag{5}$$

where $f(x_i)$, $f(x_i) \in \{n, p\}$, denotes the label of classification based on the trained model. a_i, b_i and c_i are determined according to

$$\begin{cases} a_i = 1, & \text{if } \varphi(x_i) \leq \sigma_n \text{ and } y_i = n \\ a_i = 0, & \text{if } \varphi(x_i) > \sigma_n \text{ and } y_i = n \end{cases}, \tag{6.1}$$

$$\begin{cases} b_i = 1, & \text{if } \varphi(x_i) \leq \sigma_p \text{ and } y_i = p \\ b_i = 0, & \text{if } \varphi(x_i) > \sigma_p \text{ and } y_i = p \end{cases}, \tag{6.2}$$

$$\begin{cases} c_i = 1, & \text{if } \begin{cases} (\varphi(x_i) > \sigma_p \text{ and } y_i = p) \\ \text{or } (\varphi(x_i) > \sigma_n \text{ and } y_i = n), \end{cases} \\ c_i = 0, & \text{otherwise} \end{cases} \tag{6.3}$$

where $\varphi(x_i)$, $\varphi(x_i) \in R$, denotes the estimation function of classification reliability of the i th example based on the trained model. It is worth noting that the reject thresholds are determined in the training stage based on the known data set (3), while the known label y_i in (6.1), (6.2) and (6.3), in application stage, are substituted by the label $f(x_i)$ of winning class based on the trained classifier.

Suppose that the constrains $\sigma_n = \sigma_p$, $m_i = 1$, $i = 1, 2, ..., k$, satisfied, the BCP-EC^{2}RC problem is similar to the problem studied in current literatures [3-5]. Further, if the constrain $\sigma_n = \sigma_p = 0$, are satisfied, the BCP-EC^{2}RC problem is equivalent to the conventional classification problem characterized by both minimizing 0-1 loss and zero reject rate. That is the former is the generalization of the latter, and the latter is the special case of the former.

IV. The CSVM-EC2RC Algorithm

As one of the implementation methods of solving the BCP-EC^{2}RC problem presented in section 3, based on SVM, the algorithm CSVM-EC^{2}RC is proposed and is derived in the following sections.

A. Estimation of Post-probability based on SVM

Based on the known training data (1) and the SVM algorithm, the optimal decision hyper-plane is derived as $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$. The distance between example x_i, $i = 1, ..., k$, and the above hyper-plane is

$$d(x_i) = (\boldsymbol{w} \cdot x_i + b) / \|\boldsymbol{w}\|. \quad (7)$$

The post-probability of example x_i belonging to the positive class is estimated by

$$P_p^i = \frac{1}{1+\exp(-a \cdot d(x_i))}, \quad (8)$$

where a is the parameter of Sigmoid function. The post-probability of x_i belonging to the negative class is

$$P_n^i = 1 - P_p^i. \quad (9)$$

B. Evaluation of Classification Reliability

In order to avoid running a high risk of misclassification, the input example is rejected when its classification reliability is lower than a specific threshold. The low classification reliability is generally due to the situation that two classes are estimated to have a comparable likeness [3, 5]. For the i th example, let π_1^i denote the value of the post-probability of the winning class and π_2^i denote the value of the post-probability of the other class, which is formulated as following

$$\pi_1^i = \max(P_p^i, P_n^i), \ \pi_2^i = \min(P_p^i, P_n^i). \quad (10)$$

Then the classification reliability parameter of example x_i is defined by

$$\psi^i = 1 - (\pi_2^i / \pi_1^i). \quad (11)$$

The low parameter ψ^i, $0 \le \psi^i \le 1$ means that the example x_i located at the overlapping region of two classes in the feature space [3, 5].

C. Optimal Reject Rate

Based on the classification reliability parameter φ^i, $i = 1, 2, ..., k$, and the decision function $f(\cdot)$ of SVM, the optimal reject threshold is determined when the error cost m_i and the class-dependent reject cost r_i are taken into account. Given the reject threshold (σ_n, σ_p), $\sigma_n, \sigma_p \ge 0$, the example x_i, $i = 1, 2, ..., k$, is rejected if its classification reliability is lower than its reject threshold, and the corresponding average cost can be evaluated according to (4). The rejected samples involve not only the ones that could be misclassified but also the ones that could be classified correctly, that is the introduction of reject option has a side effect. The optimal reject thresholds can be solved by

$$(\sigma_n^*, \sigma_p^*) = \arg\min_{0 \le \sigma_n, \sigma_p \le 1} (P_r^n \cdot r_n + P_r^p \cdot r_p + P_e \cdot m_0), \quad (12)$$

where the probability P_r, P_e^p and P_e^p are estimated by (5) and (6).

D. Algorithm Procedure

Algorithm: The CSVM-EC^{2}RC algorithm

1: {Input: positive class P and negative class N }

2: $i \leftarrow 0$

3: **repeat**

4: $i \leftarrow i + 1$

5: Randomly sample a training subset P_i form P and a training subset N_i form N, $(\mathrm{P} - \mathrm{P}_i) + (\mathrm{N} - \mathrm{N}_i)$ is the testing subset

6: Train SVM classifier with P_i and N_i : P_i & $\mathrm{N}_i \rightarrow d_i$

7: Estimate the post-probability: $d_i \rightarrow P_i$

8: Evaluate the classification reliability: $P_i \rightarrow \psi_i$

9: Optimize the reject rate: $\psi_i \rightarrow \sigma_i^n * \ \& \ \sigma_i^p *$

10: **until** $i = T$ (T is the repeat times)

11: Output: $\frac{1}{T}\sum_{i=1}^{T} \sigma_i^n * \ \& \ \sigma_i^p * \rightarrow \sigma^n * \ \& \ \sigma^p *$

V. Numerical Experiment

A. Data Sets and Settings

Two data sets from the UCI Machine Learning Repository [12] are used to empirically evaluate the performance of CSVM-EC^{2}RC algorithm. Information about these data sets is summarized as follow:

1) Data Sets:
   - *German Credit data:* There are 1000 instances in this data set, 300 (30%) of which belong to positive class (i.e., fraudulent class), and 700 (70%) negative class (i.e., genuine class). Each instance contains 24 predictive attributes and 1 class attributes.
   - *Australian Credit Approval:* There are 690 instances in this data set, 307 (44.5%) of which belong to positive class (i.e., fraudulent class), and 383 (55.5%) negative class (i.e., genuine class). Each instance contains 14 predictive attributes and 1 class attributes.

2) Settings:

In each experiment, 150 positive examples and 150 negative examples are randomly selected for training, while 150 positive examples and 150 negative examples randomly selected from the remaining positive examples and negative examples constitute the test data. All the symbolic variables are converted to numeric values and then are normalized. The experiment is repeated 100 times with different training-test partition, and the parameters setting is as: $m_0 = 1$, $r_p = 0.4$ and $r_n = 0.3$.

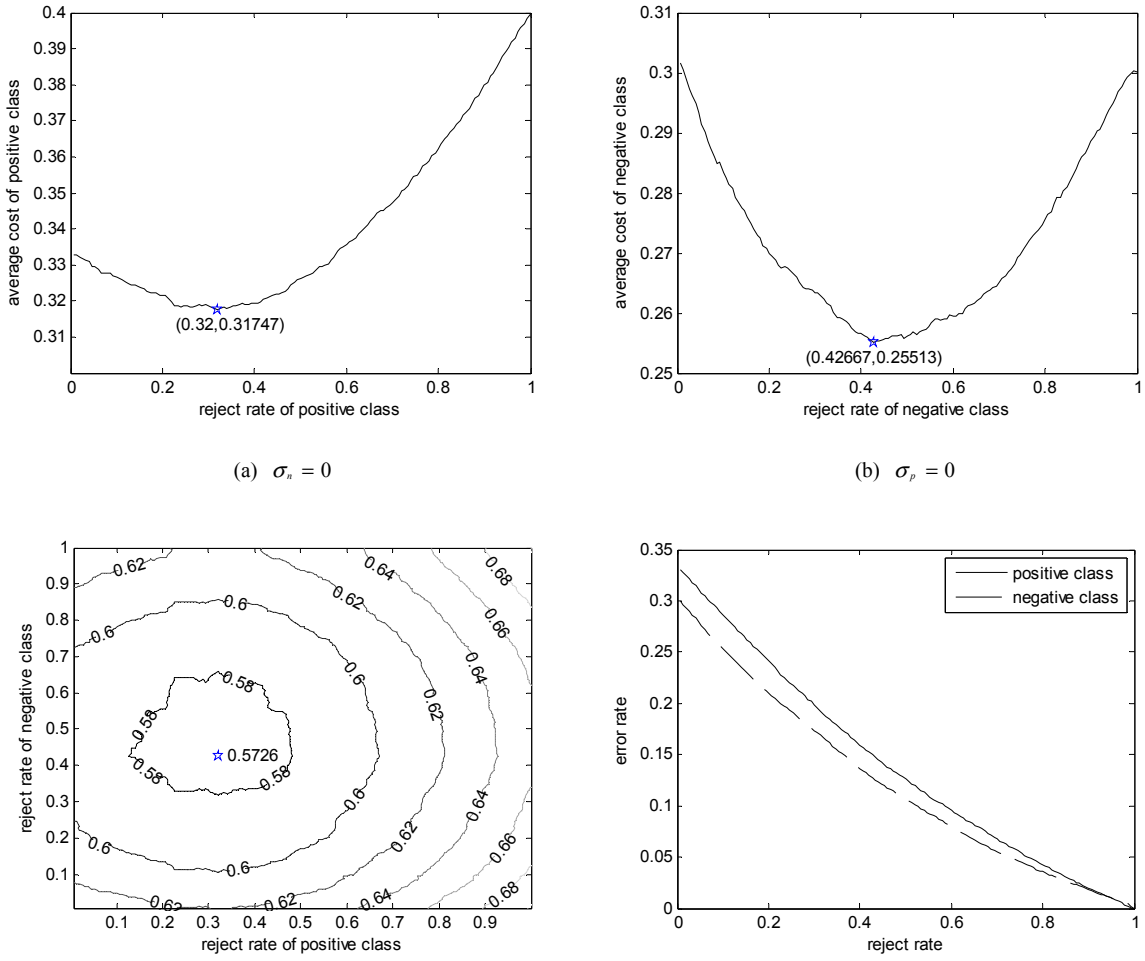
B. Experimental Results

The experimental results on *German Credit* data set and *Australian Credit Approval* data set are shown in Fig. 1 and Fig. 2.

When the reject rate of negative class is zero, it is shown that with the reject rate of positive class increases from 0 to 0.32 (in Fig. 1 (a)) and from 0 to 0.087 (in Fig. 2 (a)), the average cost associated with both rejecting and misclassifying the examples of positive class decreases from

0.333 to the minimum 0.317 (in Fig. 1 (a)) and from 0.263 to the minimum 0.257 (in Fig. 2 (a)), while with the reject rate of positive class increases further, the average cost related with positive class increases also and reaches 1 (at this point, all examples of positive class are rejected) finally. Those results mean that the introduction of rejecting examples of positive class could reduce the average cost of positive class. It is worth noting that, in Fig. 1 (d) and Fig. 2 (d), the increase of reject rate of positive class leads to the decrease of error rate of positive class, and the former degrades the performance of classifier, while the latter improves the classification performance. The optimal reject rate of positive class when the reject threshold of negative class is equal to 0, therefore, means that the summation of error cost and reject cost of positive is minimized.

The similar results are presented in Fig. 1 (b) and Fig. 2 (b) when the reject rate of positive class is zero. By comparing Fig. 1 (a) and Fig. 1 (b), Fig. 2 (a) and Fig. 2 (b), it can be seen that the optimal reject rate of positive class when $\sigma_n = 0$ is less than that of negative class when $\sigma_p = 0$, and we argue that the introduction of domain knowledge $r_p > r_n$ causes the above results. In Fig. 1 (c) and Fig. 2 (c), the effect of reject rate of positive class and negative class on global average cost is illustrated based on the contour of average cost. It is depicted that the average cost declines by integrating the error cost and class-dependent reject cost into classical SVM.



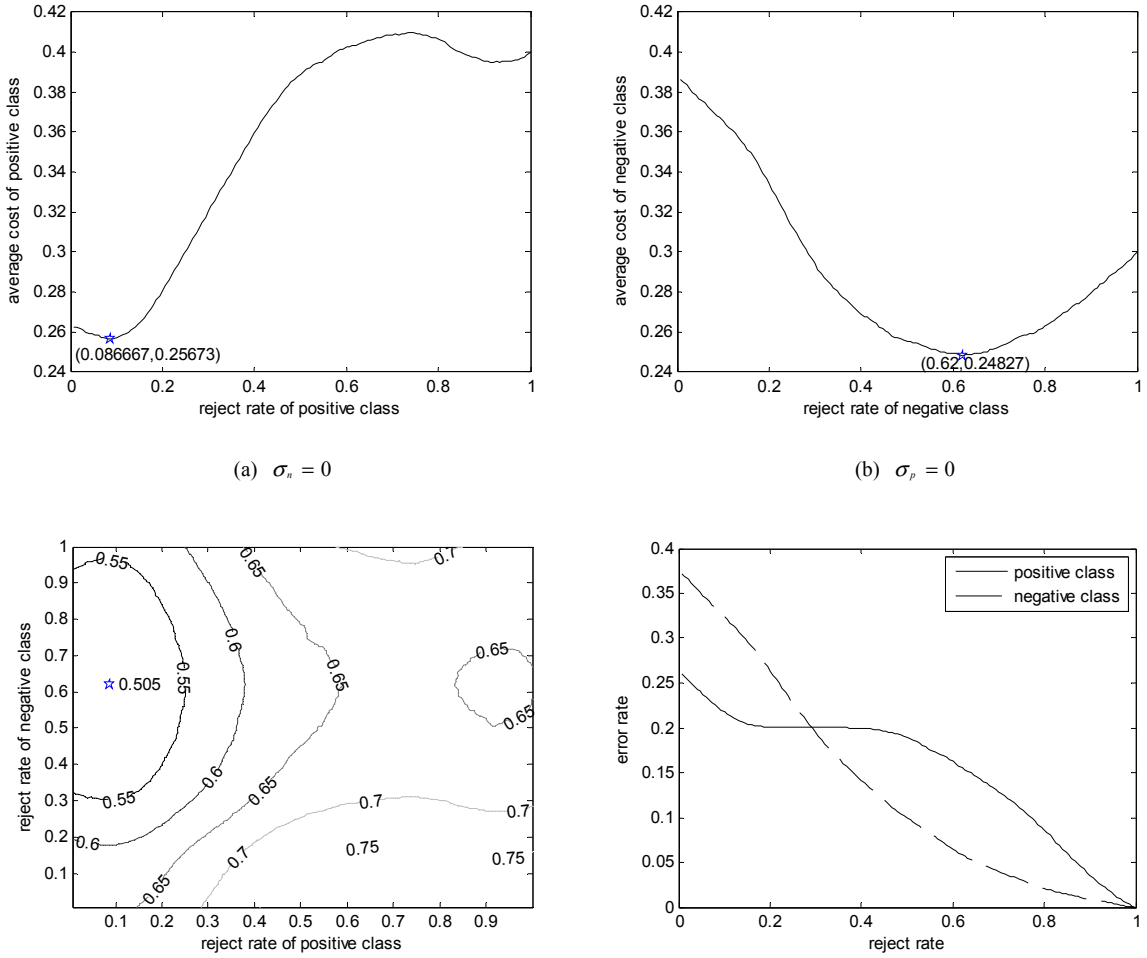
(a) $\sigma_n = 0$

(b) $\sigma_p = 0$

(c) The effect of reject rate on global average cost

(d)

Figure 1. Experimental results on German Credit data set



(a) $\sigma_n = 0$

(b) $\sigma_p = 0$

(c) The effect of reject rate on global average cost

(d)

Figure 2. Experimental results on Australian data set

TABLEI PERFORMANCE OF SVM AND CSVM-EC^{2}RC CLASSIFIERS ON TESTING DATA

Data Sets		SVM			CSVM-EC^{2}RC		
		Positive Class	*Negative Class*	*Global*	*Positive Class*	*Negative Class*	*Global*
German Credit	*Reject Rate*	0	0	—	0.3200	0.4267	—
	Error Rate	**0.3303**	**0.2996**	—	**0.1895**	**0.1271**	—
	Average Cost	0.3329	0.3016	**0.6345**	0.3175	0.2551	**0.5726**
Australian	*Reject Rate*	0	0	—	0.0867	0.6200	—
	Error Rate	**0.2601**	**0.0605**	—	**0.2221**	**0.0583**	—
	Average Cost	0.2628	0.3864	**0.6492**	0.2567	0.2483	**0.5050**

TABLE I lists the performance of SVM and CSVM-EC^{2}RC classifiers on testing data. Overall, comparing with SVM classifier, CSVM-EC^{2}RC classifier achieves a better performance: 1) global average cost decrease from 0.6345 to 0.5726 on *German Credit* data set and from 0.6492 to 0.5050 on *Australian Credit Approval* data set; 2) error rate decrease significantly both in positive class and negative class on these two data sets. The results in TABLE I reinforce the conclusion from previous experiment: the performances of our CSVM-EC^{2}RC algorithm are very encouraging.

VI. CONCLUSIONS AND FUTURE WORKS

It is considered in this paper that one of problems encountered in real data mining applications is that the reject cost is class-dependent. The new algorithm, CSVM-EC^{2}RC, by embedding the error cost and class-dependent reject cost into conventional SVM, is proposed to solve the above problem. The conclusions based on the experiment studies on a benchmark data set are as followings. Firstly, the effectiveness of classifier designed by introducing the class-dependent reject cost is improved, which leads to the decrease of average cost. Secondly, given the error cost and class-dependent reject cost, the optimal reject rate of negative class is bigger than that of positive class based on the

application of proposed algorithm into fraud detection.

The CSVM-EC^{2}RC algorithm formulated in present paper is one of the implementation methods of the proposed BCP-EC^{2}RC problem related to many real data mining applications. The future works could involve developing another method to solve the problem and applying the proposed algorithm to medical diagnosis or fault diagnosis domains.

REFERENCES

[1] D. Sánchez, M. A. Vila, L. Cerda and J. M. Serrano, "Association rules applied to credit card fraud detection," Expert Systems with Applications, vol. 36, Issue 2, 2009, pp. 3630-3640.
[2] A. Srivastava, A. Kundu, S. Sural and A.K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model," IEEE Transactions on Dependable and Secure Computing, vol. 5, Issue 1, 2008, pp. 37-48.
[3] P. Foggia, C. Sansone, F. Tortorella and M. Vento, "Multi-classification: Reject criteria for the Bayesian combiner," Pattern Recognition, vol. 32, issue 8, 1999, pp. 1436-1447.
[4] T. C. W. Landgrebe, D. M. J. Tax, P. Paclík and R. P. W. Duin, "The interaction between classification and reject performance for distance-based reject-option classifiers," Pattern Recognition Letters, vol. 27, issue 8, 2006, pp. 908-917.
[5] C. D. Stefano, C. Sansone and M. Vento, "To reject or not to reject, that is the question-an answer in case of neural classifiers," IEEE Trans. on SMC, vol. 30, issue 1, 2000, pp. 84-94.
[6] Z. P. Hu and Y. Zhang, "Micro-calcification Detection Algorithm Based on Fast Double-layer Support Vector Classifier with Reject Performance," Chinese Journal of Scientific Instrument, vol. 28, issue 3, 2007, pp. 446-450.
[7] E. H. Zheng, C. Zou, J. Sun and L. Chen, "SVM-based Credit Card Fraud Detection with Reject Cost and Class-dependent Error Cost," Proceedings of the PAKDD' 09 Workshop: Data Mining When Classes are Imbalanced and Errors Have Cost, 2009, pp. 50-58.
[8] E. H. Zheng, "Cost Sensitive Data Mining Based on Support Vector Machines: Theories and Applications," Dissertation of Zhejiang University, 2006.
[9] Z. H. Zhou and X. Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," IEEE Transactions on Knowledge and Data Engineering, vol.18, issue 1, 2006, pp. 63-77.
[10] E. H. Zheng, P. Li and Z. H. Song, "Cost Sensitive Support Vector Machines," Control and Decision, vol. 21, issue 4, 2006, pp. 473-476.
[11] D. J. Michie and C. C. Taylor, "Machine Learning, Neural and Statistical Classification," Prentice Hall, Englewood Cliffs, N.J. , 1994. Data available: http://www.liacc.up.pt/, 2004.
[12] UCI Machine Learning Repository, available: http://archive.ics.uci. edu/ml/, 2009.

En-hui Zheng was born in Xinmin, China in 1975. He received his B.S. and M.S. degree from Northeast Dianli University, China in 1999 and 2002, and Ph.D. degree from Zhejiang University, China in 2006.
He is working as an associate professor in College of Mechatronics Engineering, China Jiliang University. His research interests include cost-sensitive data mining and modeling and control of complicated system.

Chao Zou was born in Luoyang, China in 1986. He received his B.S. degree from The First Aeronautic Institute of Air Force, China in 2008.
He is currently pursuing his M.S. degree in China Jiliang University. His research interests include cost-sensitive data mining and machine learning.

Jian Sun was born in Hangzhou, China in 1964. He received his B.S. degree from Zhejiang University, China in 1985.
He is working as a research professor and Vice President of College of Mechatronics Engineering of China Jiliang University. His research interests include dynamic measurement, online measurement and automation control instrumentation.

Le Chen was born in Hangzhou, China in 1959. She received her B.S., M.S. and Ph.D. degree from Zhejiang University, China in 1982, 1994, and 2007.
She is working as a research professor and director of Science & Technology Department of China Jiliang University. Her research interests include dynamic measurement, online measurement and temperature metrology.