Communications of the Korean Statistical Society
2009, Vol. 16, No. 6, 997–1004

An Algorithm for Support Vector Machines with a Reject Option Using Bundle Method

Hosik Choi[1,a], Yongdai Kim[b], Sang-Tae Han[a], Hyuncheol Kang[a]

[a]Department of Informational Statistics and Institute of Basic Science, Hoseo University
[b]Department of Statistics, Seoul National University

Abstract

A standard approach is to classify all of future observations. In some cases, however, it would be desirable to defer a decision in particular for observations which are hard to classify. That is, it would be better to take more advanced tests rather than to make a decision right away. This motivates a classifier with a reject option that reports a warning for those observations that are hard to classify. In this paper, we present the method which gives efficient computation with a reject option. Some numerical results show strong potential of the proposed method.

Keywords: Classification, reject option, support vector machines, bundle method.

1. Introduction

A standard approach for classifying one group from the other group is to classify all of future observations. In some cases, however, it would be desirable to defer a decision in particular for observations which are hard to classify. For example, an observation whose conditional probability of being cancer is around 1/2, it would be better to take more advanced tests rather than to make a decision right away. This motivates a classifier with a reject option that reports a warning for those observations that are hard to classify. Many empirical studies in the engineering community support that a reject option effectively reduces misclassification error rates. See, for example, Lendgrebe *et al.* (2006). Recently, Bartlett and Wegkamp (2008) proposed a learning algorithm with a reject option based on the support vector machines(SVM, Cortes and Vapnik, 1995) called the SVM with a reject option and studied its theoretical properties.

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be input-output pairs of given data where $\mathbf{x}_i \in \mathbb{R}^p$ is an input random vector and $y_i \in \{-1, 1\}$ denotes a class label. We assume that the data are n independent copies of a random vector (X, Y). Traditional learning algorithms try to find the optimal hyperplane which minimizes the misclassification error rate $\mathrm{E}\,(I(Yf(X) < 0))$ among all linear hyperplanes $f(\mathbf{x}) = \beta_0 + \mathbf{x}^T\beta$, $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$ (we use $\boldsymbol{\beta} = (\beta_0, \beta^T)^T$). A method of estimating the optimal hyperplane is to minimize a regularized empirical risk given as $\sum_{i=1}^n \phi(-y_i(\beta_0 + \mathbf{x}_i^T\beta)) + P_\lambda(\beta)$, where $\phi(z)$ is a convex surrogate loss function of the 0–1 loss $I(z < 0)$ and P_λ is a penalty function controlling misclassification error and classifier's complexity. Various surrogate loss functions $\phi(z)$ correspond to various learning algorithms including logistic regression, boosting and SVM (Hastie *et al.*, 2001).

A learning algorithm with a reject option is to construct a classifier $t : \mathbb{R}^p \to \{-1, 1, \text{®}\}$, where the interpretation of the output ® is of being in doubt and making no decision. Chow (1970) introduced a misclassification error rate of a classifier with a reject option as $L_d(t) = d \cdot \Pr(t(X) = \text{®}) + \Pr(t(X) \neq$

[1] Corresponding author: Full Time Lecturer, Department of Informational Statistics and Institute of Basic Science, Hoseo University, Asan Campus, Sechul-ri, Baebang-myun, Asan, Chungnam 336-795, Korea. E-mail: choi.hosik@gmail.com.

Table 1: The misclassification's cost matrix with a reject option

	$\hat{y}$		
y	$+1$	Ⓡ	-1
$+1$	0	d	1
-1	1	d	0

$Y, t(X) \neq$ Ⓡ) where $d \in [0, 1/2)$ is a cost of a reject option. Table 1 shows that the misclassification's cost matrix for classification problem with a reject option.

Proposition 1. (Chow' rule; Chow, 1970) *For 0–1 loss function with a reject option $L_d(t)$ and a given classifier t, let t^* be the Bayes classifier with respect to $L_d(t)$. Then the classifier t^* with a reject option is given as*

$$t^*(\mathbf{x}) = \begin{cases} -1, & \text{if } \eta(\mathbf{x}) < d, \\ \text{Ⓡ}, & \text{if } d \le \eta(\mathbf{x}) \le 1-d, \\ 1, & \text{if } \eta(\mathbf{x}) > 1-d, \end{cases}$$

where $\eta(\mathbf{x}) = \Pr(Y = 1|X = \mathbf{x})$ and $0 \le d < 1/2$.

For given real valued function $f(\mathbf{x})$ and $\delta > 0$, Bartlett and Wegkamp (2008) considered a method constructing a classifier with a reject option $t_f^\delta(\mathbf{x})$ by

$$t_f^\delta(\mathbf{x}) = \text{sign} f(\mathbf{x}) I(|f(\mathbf{x})| > \delta) + \text{Ⓡ} I(|f(\mathbf{x})| \le \delta).$$

Then, the misclassification error rate of t_f^δ becomes $E(l_{d,\delta}(Yf(X)))$ where

$$l_{d,\delta}(z) = \begin{cases} 1, & \text{if } z < -\delta, \\ d, & \text{if } |z| \le \delta, \\ 0, & \text{otherwise.} \end{cases}$$

They proposed the SVM with a reject option by replacing $l_{d,\delta}$ with a convex surrogate loss and applying the l_2 norm of the coefficients as a penalty function. That is, the SVM with a reject option estimates a classifier by minimizing the following regularized empirical risk

$$\sum_{i=1}^{n} \phi_d \left(-y_i \left(\beta_0 + \mathbf{x}_i^T \beta \right) \right) + \frac{\lambda}{2} \|\beta\|_2^2, \tag{1.1}$$

where $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ and the surrogated loss function ϕ_d, the hinge loss with a reject option, is given as

$$\phi_d(z) = \begin{cases} 1 - \dfrac{1-d}{d} z, & \text{if } z < 0, \\ 1 - z, & \text{if } 0 \le z < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.2}$$

Bartlett and Wegkamp (2008) showed that ϕ_d is a reasonable surrogate loss for $l_{d,\delta}$ by proving the Fisher consistency. See Figure 1 for comparison of the $l_{d,\delta}$ and ϕ_d as well as the hinge loss $\phi_H(z) = (1-z)_+$ that is a surrogate loss for the SVM, where $(z)_+ = \max\{z, 0\}$. Note that ϕ_d is piecewise

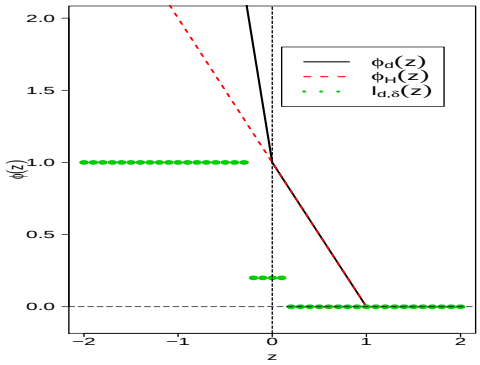


Figure 1: *The hinge loss ϕ_d with a reject option when $d = 0.2$ and other loss functions*

linear as the hinge loss. However, there are two nondifferentiable points - one at $z = 1$ and the other at $z = 0$ in ϕ_d while there is only one nondifferentiable point at $z = 1$ in the hinge loss.

In this paper, we consider a machine learning approach for variable selection with a reject option. For this purpose, the l_1 penalty is applied to the SVM with a reject option. The l_1 penalty is widely used for variable selection in many contexts including Tibshirani (1996) and Zhu *et al.* (2004) since it gives a sparse solution (*i.e.* some estimated coefficients are exactly zero). We call the proposed method the l_1 SVM with a reject option (L1SVM-R). For the purpose of comparison, we denote the standard l_1 SVM (Zhu *et al.*, 2004) as L1SVM.

The paper is organized as follows. In Section 2, we review the bundle method for minimizing non-smooth objective function and apply this to the l_1 SVM with a reject option. Numerical results on simulated data as well as publicly available gene expression data set are presented in Section 3. Concluding remarks follow in Section 4.

2. Methodology

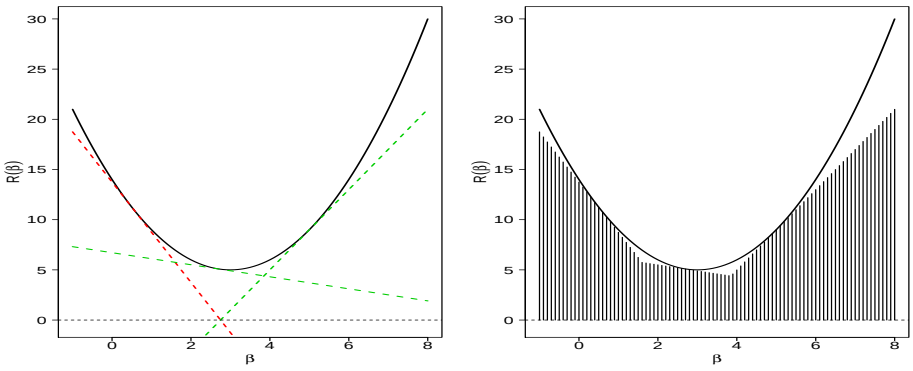
In this section, we present an efficient algorithm to solve

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \phi_d \left(y_i f(\mathbf{x}_i)\right) + \lambda \|\beta\|_1, \tag{2.1}$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. The Equation (2.1) is convex and continuous function but not differentiable. That is, a main difficulty is oriented from non-smooth surrogate function. Recently, the bundle method is developed where it efficiently treats a minimization problem with non-smooth convex objective.

2.1. Bundle method

Conceptually, bundle method can be characterized as cutting plane method. Here, cutting plane is a lower bound of convex objective function. See Figure 2. Figure 2(a) depicts that a convex objective function (not necessary smooth) and its Taylor approximations of first order and the shaded area of Figure 2(b) describes the maximum of Taylor approximations of first order. This maximum function is the lower bound of the objective function. Thus, the bundle method or cutting plane method minimizes the lower bound instead of the objective function. Since the lower bound is linear function, it can be implemented easily with a non-smooth function.



(a) (b)

Figure 2: *A convex function $R(\boldsymbol{\beta})$ is bounded below by Taylor approximation of first order*

Definition 1. *For a convex function F, μ is called a subgradient of F at w if and only if*

$$F(\tilde{w}) \geq F(w) + < \tilde{w} - w, \quad \mu > \text{ for all } \tilde{w}.$$

The set of subgradients at a point is called the subdifferential, and is denoted by $\partial_w F(w)$.

The problem of the regularized empirical is $\min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \triangleq R(\boldsymbol{\beta}) + P_\lambda(\beta)$. Then, Taylor first order approximation for $R(\boldsymbol{\beta})$ at a given solution $\boldsymbol{\beta}^{(t)}$ is

$$R(\boldsymbol{\beta}) \geq R\left(\boldsymbol{\beta}^{(t)}\right) + \partial_{\boldsymbol{\beta}} R\left(\boldsymbol{\beta}^{(t)}\right)^T \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right), \quad \text{for all } \boldsymbol{\beta}.$$

And hence, the lower bound $R_t^{\mathrm{CP}}(\boldsymbol{\beta})$ of the objective given previous set of solutions $\boldsymbol{\beta}^{(l-1)}$, $l < t$, is less than $R(\boldsymbol{\beta})$. That is,

$$R(\boldsymbol{\beta}) \geq R_t^{\mathrm{CP}}(\boldsymbol{\beta}) \triangleq \max_{l<t} \left[R\left(\boldsymbol{\beta}^{(l-1)}\right) + \partial_{\boldsymbol{\beta}} R\left(\boldsymbol{\beta}^{(l-1)}\right)^T \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(l-1)}\right) \right] \tag{2.2}$$

holds.

Note that $R_t^{\mathrm{CP}}(\boldsymbol{\beta})$ is a piecewise linear lower bound of the R and also, this is the tightest among lower bounds in the sense of first order approximation.

Let $J_t(\boldsymbol{\beta}) = R_t^{\mathrm{CP}}(\boldsymbol{\beta}) + P_\lambda(\beta)$ be the regularized objective function at t-iteration step. Along with definitions of $J_t(\boldsymbol{\beta})$ and $R_t^{\mathrm{CP}}(\boldsymbol{\beta})$, the following simple lemma holds.

Lemma 1. (Teo *et al.*, 2007, 2009) *Let $\boldsymbol{\beta}^*$ be the minimizer of $J(\boldsymbol{\beta})$ and let J^* be its minimum value $J(\boldsymbol{\beta}^*)$. Then, the following holds for the upper bound $J_t^+(\boldsymbol{\beta})$ of J^* and the lower bound $J_t^-(\boldsymbol{\beta})$ of J^*:*

$$J_t^+(\boldsymbol{\beta}) \triangleq \min_{l \leq t} J_{l+1}\left(\boldsymbol{\beta}^{(l)}\right) \geq J^* \text{ and } J_t^-(\boldsymbol{\beta}) \triangleq J_t\left(\boldsymbol{\beta}^{(l)}\right) \leq J^*.$$

The sequence J_t^+ is monotonically decreasing and J_t^- is monotonically increasing. Moreover, if we let $\epsilon_t = \min_{l \leq t} J_{l+1}(\boldsymbol{\beta}^{(l)}) - J_t(\boldsymbol{\beta}^{(t)})$ then ϵ_t is monotonically decreasing and $\epsilon_t \geq \min_{l \leq t} J_{l+1}(\boldsymbol{\beta}^{(l)}) - J^* \geq 0$.

Lemma 1 says that the algorithm converges to the optimum for sufficiently small ϵ_t. Figure 3 summarizes the pseudo code of the bundle method for minimizing a non-smooth objective function for a given stopping criterion ϵ.

1. Initialize $t = 0$, $\boldsymbol{\beta}^{(0)} = \left(\beta_0, \beta^T\right)^T = \mathbf{0}$ and calculate $J_0\left(\boldsymbol{\beta}^{(0)}\right)$
2. Repeat until $\epsilon_t \le \epsilon$.
   (a) $t \leftarrow t + 1$,
   (b) Compute (sub)gradient $\mathbf{a}_t$ and offset b_t
   where $\mathbf{a}_t = \partial_{\boldsymbol{\beta}} R\left(\boldsymbol{\beta}^{(t-1)}\right)$, $b_t = R\left(\boldsymbol{\beta}^{(t-1)}\right) - \partial_{\boldsymbol{\beta}} R\left(\boldsymbol{\beta}^{(t-1)}\right)^T \boldsymbol{\beta}^{(t-1)}$
   (c) Update model : $R_t^{\mathrm{CP}}(\boldsymbol{\beta}) = \max_{1 \le l \le t} \mathbf{a}_l^T \boldsymbol{\beta}^{(l-1)} + b_l$.
   (d) Find minimizer $\boldsymbol{\beta}^{(t)} = \arg\min J_t(\boldsymbol{\beta}) := R_t^{\mathrm{CP}}(\boldsymbol{\beta}) + \lambda \|\beta\|_1$.
   (e) $\epsilon_t \leftarrow \min_{0 \le l \le t} J\left(\boldsymbol{\beta}^{(l)}\right) - J_t\left(\boldsymbol{\beta}^{(t)}\right)$
3. Return $\boldsymbol{\beta}^{(t)}$.

Figure 3: *Bundle Method for minimizing a non-smooth objective function for a given stopping criterion* ϵ

2.2. l_1 SVM with a reject option using the bundle method

Let $R(\boldsymbol{\beta}) = \sum_{i=1}^n \phi_d(y_i f(\mathbf{x}_i))$, $P_\lambda(\beta) = \lambda \sum_{j=1}^p |\beta_j|$ and $J(\boldsymbol{\beta}) = R(\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j|$. Then, the problem $\min_{\boldsymbol{\beta}} J_t(\boldsymbol{\beta})$ of t-iteration step of the bundle method is as follows:

$$\min_{\boldsymbol{\beta}} R_t^{\mathrm{CP}}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j| = \min_{\boldsymbol{\beta}} \left[\max_{l \le t} \left[R\left(\boldsymbol{\beta}^{(l-1)}\right) + \partial_{\boldsymbol{\beta}} R\left(\boldsymbol{\beta}^{(l-1)}\right)^T \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(l-1)}\right) \right] + \lambda \sum_{j=1}^p |\beta_j| \right].$$

This problem is equivalent to

$$\min_{\xi, \boldsymbol{\beta}} \left(\xi + \lambda \sum_{j=1}^p |\beta_j| \right) \text{ subject to } \xi \ge R\left(\boldsymbol{\beta}^{(l-1)}\right) + \partial_{\boldsymbol{\beta}} R\left(\boldsymbol{\beta}^{(l-1)}\right)^T \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(l-1)}\right), \quad \text{for all } \; l \le t.$$

This can be solved by usual linear programming(LP). The exact form is

$$\begin{aligned} &\min_{\xi, \beta_0^+, \beta_0^-, \beta^+, \beta^-} \xi + \lambda \sum_{j=1}^p \left(\beta_j^+ + \beta_j^-\right), \\ &\text{subject to } \; \xi \ge \mathbf{a}_l^T \boldsymbol{\beta}^+ - \mathbf{a}_l^T \boldsymbol{\beta}^- + b_l, \quad l = 1, \ldots, t, \\ &\qquad \boldsymbol{\beta}^+ = \left(\beta_0^+, \beta^{+T}\right)^T, \; \boldsymbol{\beta}^- = \left(\beta_0^-, \beta^{-T}\right)^T \ge \mathbf{0}, \; \boldsymbol{\beta}^+ \in \mathbb{R}_+^{p+1} \text{ and } \boldsymbol{\beta}^- \in \mathbb{R}_+^{p+1}. \end{aligned}$$

Since the penalty as well as the surrogate loss are piecewise linear, we can use a LP to solve the optimization problem of the l_1 SVM with a reject option. Also, the optimization problem $\min_{\boldsymbol{\beta}} J_t(\boldsymbol{\beta})$ of inner loops of the bundle method can be solved using LP. However, in the former problem, the number of constraints equals to the number of observations in the LP. Meanwhile, in the latter problem, the number of constraints equals the number of (sub)gradients which are computed previously. Since the number of iterations required for convergence is typically in the order 10s to 100s, the computational cost of the resulting LP of the bundle method is not so expensive. Thus, the computational efficiency via the standard LP becomes worse as the number of observations grows.

3. Numerical Studies

In this section, we compare l_1 SVM with a reject option with the standard l_1 SVM (Zhu *et al.*, 2004) that does not have a reject option by analyzing simulated as well as real data sets in terms of prediction accuracy and variable selectivity.

Table 2: Comparison of prediction accuracy of the L1SVM and L1SVM-R: average misclassification errors (standard errors)

r	Method	Total MIS	Accept MIS	Reject MIS	Reject	p-value
0	L1SVM	.151 (.003)	.134 (.003)	.505 (.017)		
	L1SVM-R	**.147** (.003)	**.131** (.003)	**.483** (.013)	.048 (.007)	.003
0.3	L1SVM	.215 (.003)	.193 (.003)	.480 (.013)		
	L1SVM-R	**.211** (.003)	**.191** (.004)	**.470** (.024)	.078 (.010)	.002
0.6	L1SVM	.258 (.004)	.236 (.004)	**.493** (.013)		
	L1SVM-R	**.251** (.004)	**.230** (.004)	.515 (.024)	.087 (.013)	.001

Table 3: Comparison of variable selectivity of the L1SVM and L1SVM-R: average numbers of selected coefficients (standard errors)

r	Method	Czeros	Cnzeros	Count	Others
0	L1SVM	90.90 (0.976)	5 (0)	20 20 20 20 20	0.863
	L1SVM-R	90.35 (1.164)	5 (0)	20 20 20 20 20	0.979
0.3	L1SVM	92.45 (0.634)	4.300 (.147)	20 17 17 17 15	0.537
	L1SVM-R	90.95 (0.752)	4.500 (.154)	20 18 18 17 17	0.853
0.6	L1SVM	89.15 (1.186)	3.550 (.198)	19 10 11 12 19	1.232
	L1SVM-R	89.25 (1.015)	3.750 (.176)	18 14 13 12 18	1.211

3.1. Simulation

For simulation data, we generate a sample of size n as follows. Let μ^+ be a p-dimensional vector whose first q entries are D and the other $p-q$ entries are zero and let $\mu^- = -\mu^+$. Then, we generate $\mathbf{x}$ from $N_p(\mu^+, \Sigma)$ and assign $y = 1$ for the first $n/2$ observations and generate $\mathbf{x}$ from $N_p(\mu^-, \Sigma)$ and assign $y = -1$ for the last $n/2$ observations. We let the (k, l) entry of Σ be $r^{|k-l|}$ for some $r \in [0, 1)$. Note that all input variables except the first q are noisy.

Table 2 compares the prediction accuracy. The training sample size is 100, the regularization parameters (d, λ) are selected based on an independent validation sample of size 100 and the misclassification errors are calculated based on another independent test sample of size 2000. We repeat the simulation 20 times and report the averages with their standard errors in the parenthesis. In the table, "Total MIS" denotes the misclassification error rates obtained based on all observations in a test sample, "Accept MIS" based only on accepted observations by a L1SVM-R classifier and "Reject MIS" based on rejected observations. "Reject" is the portion of rejected observations. The p-values are obtained by the paired t-test with 20 paired error rates of L1SVM-R and L1SVM.

Table 3 shows variable selectivities of two methods. In the table, "Czeros" and "Cnzeros" are the average numbers of correct 0 (true is 0 and estimated as 0) and correct nonzero coefficients (true is nonzero and estimated as nonzero) in the estimated models, "Count" represents the frequencies of the first q coefficients being estimated as nonzero among 20 simulations and "Others" is the frequencies of selected noisy variables. We set $D = 0.5$, $p = 100$ and $q = 5$. In the Table, we highlight the smallest misclassification error rates by bold face.

The L1SVM-R always has lower misclassification errors significantly (except one case - Reject MIS with $r = 0.6$) than the L1SVM, and the improvements are statistically significant in most cases. Also, note that the misclassification error rates on rejected observations are around 0.5, which indicates that the L1SVM-R successively selects observations near a decision boundary. The superior prediction performance of the L1SVM-R is partly because using only high quality samples (samples from far away a decision boundary) makes a corresponding classifier robust to less informative samples that usually locate near a decision boundary and hence yields a better prediction accuracy. For variable selectivity, we can see that the L1SVM-R and L1SVM are competitive. Based on the results

Table 4: Prediction accuracies, reject portions, p-values and mean model sizes (Vsize, $|\mathcal{V}|$) with the corresponding standard errors in the parentheses of the L1SVM and L1SVM-R on thyroid cancer data

Method	Total MIS	Accept MIS	Reject MIS	Reject	p-value	Vsize
L1SVM	.254 (.018)	.235 (.018)	.512 (.071)			3.450 (.872)
L1SVM-R	**.235 (.009)**	**.218 (.011)**	**.481** (.068)	.059 (.011)	.097	2.850 (.386)

of the simulation, we can conclude that the L1SVM-R improves prediction accuracy significantly without hampering variable selectivity.

3.2. Analysis of a gene expression data

In this section, we analyze gene expression data set for thyroid cancer (Yukinawa *et al.*, 2006). Thyroid cancer is a relatively common cancer accounting for roughly 1% of total cancer incidence. There are two main types of thyroid cancer, papillary carcinoma(PC) and follicular carcinoma(FC). In addition to these malignant types, a benign tumor, follicular adenoma(FA), is also prevalent. It consists of 168 samples and 2000 genes. To calculate misclassification errors, we divide each data set into two parts, training and test data sets, by randomly selecting 2/3 observations and 1/3 observations, respectively. We construct a classifier on training data and select the regularization parameters (d, λ) by minimizing the BIC-type criterion (Schwarz, 1978) calculated on the training data :

$$\frac{1}{n}\sum_{i=1}^{n}\phi_d(y_i f(\mathbf{x}_i)) + \frac{\log n}{2n}|\mathcal{V}|,$$

where $|\mathcal{V}|$ is the number of nonzero coefficients in $\boldsymbol{\beta}$. Then, we measure a misclassification error on test data. We repeat this procedure 20 times and summarize the results in Table 4.

The results also confirm superiority of the L1SVM-R over the L1SVM. That is, the L1SVM-R performs better than the L1SVM in prediction. As shown in Table 4, the L1SVM-R performs better in terms of Total MIS and Accept MIS than the L1SVM. Also, the L1SVM-R has lower error rates even for rejected samples in most cases. The numbers of selected genes (Vsize in the table) are similar.

4. Conclusion

In this paper, we proposed a learning method which can simultaneously select signal variables and produce a highly accurate predictive model by incorporating a reject option. Also, we developed an efficient computational algorithm for minimizing non-smooth objective function without much difficulty. Analysis of simulated and real data set suggested the strong potential of the proposed method.

There would be various applications of the reject option. For example, we can remeasure rejected samples. The remeasurement would improve prediction accuracy further in particular when there are measurement errors.

References

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

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

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning,* **20**, 273–297.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning,* First Edition, Springer-verlag, New York.

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

Lendgrebe, C. W., Tax, M. J. and Duin, P. W. (2006). The interaction between classification and reject performance for distance-based reject-option classifiers, *Pattern Recognition Letters,* **27**, 908–917.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* **6**, 461–464.

Teo, C. H., Le, Q., Smola, A. and Vishwanathan, S. V. N. (2007). A scalable modular convex solver for regularized risk minimization, *International Conference on Knowledge Discovery and Data Mining archive Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining,* 727–736.

Teo, C. H., Vishwanathan, S. V. N., Smola, A. and Le, Q. (2009). Bundle methods for regularized risk minimization, *Journal of Machine Learning Research,* To appear.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B,* **58**, 267–288.

Yukinawa, N., Oba, S., Kato, K., Taniguchi, K., Iwao-Koizumi, K., Tamaki, Y., Noguchi, S. and Ishii, S. (2006). A multi-class predictor based on a probabilistic model: Application to gene expression profiling-based diagnosis of thyroid tumors, *BMC Bioinformatics,* **7**, 1471–2164.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2004). 1-norm support vector machines, *In Thrun,S. et al. (eds). Advances in Neural Information Processing Systems,* **16**, MIT Press, Cambridge, MA.

Received September 2009; Accepted September 2009