Research Article

Support Vector Machines for Unbalanced Multicategory Classification

Kang-Mo Jung

Department of Statistics and Computer Science, Kunsan National University, Gunsan 573-701, Republic of Korea

Correspondence should be addressed to Kang-Mo Jung; kmjung@kunsan.ac.kr

Received 16 December 2014; Revised 6 February 2015; Accepted 7 February 2015

Academic Editor: Yaguo Lei

Copyright © 2015 Kang-Mo Jung. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Classification is a very important research topic and its applications are various, because data can be easily obtained in these days. Among many techniques of classification the support vector machine (SVM) is widely applied to bioinformatics or genetic analysis, because it gives sound theoretical background and its performance is superior to other methods. The SVM can be rewritten by a combination of the hinge loss function and the penalty function. The smoothly clipped absolute deviation penalty function satisfies desirably statistical properties. Since standard SVM techniques typically treat all classes equally, it is not well suited to unbalanced proportion data. We propose a robust method to treat unbalanced cases based on the weights of the class. Simulation and a numerical example show that the proposed method is effective to analyze unbalanced proportion data.

1. Introduction

Classification is a very important research topic and is applied to many applications such as health science and bioinformatics. Many classification methods have been proposed in the literature, for example, linear discriminant analysis, logistic regression, the K -nearest neighbors, and support vector machine (SVM) as in [1]. Among them SVM is considered to be popular in engineering, because it gives sound theoretical background. It has been widely applied to bioinformatics or genetic analysis and its performance is superior to other methods.

In these days we can easily obtain high dimensional data thanks to computer technology. Since the number of predictors is very large, variable selection is crucial to get a meaningful model. In fact the model with one thousand nonzero predictors cannot be interpretable and it does not give any information on the data. Furthermore if the true model is sparse, the fitted model should be sparse. Variable selection is an important research topic in linear regression modeling especially when there are tremendous predictors (see [2]). Thus simultaneous variable selection and estimation methods have been suggested. It is called a penalized method. The SVM can be considered as a penalized method consisting of the hinge loss function and the L_2 penalty function.

After the Ridge regression estimator many penalty functions have been proposed. Among them the least absolute shrinkage and selection operator (LASSO) proposed by [3] with the L_1 penalty function is very popular, because it has the sparse property. However, the LASSO estimate can be biased for the large absolute coefficients, which was pointed out by [4]. They proposed a nonconvex penalty function, the smoothly clipped absolute deviation (SCAD) penalty satisfying desirable properties: the unbiasedness, the sparseness, and the continuity. Since the SCAD function is not convex, the standard optimization algorithm cannot be applied. Thus the locally quadratic approximation (LQA) algorithm can be adopted to solve the optimum of the objective function as in [5].

The traditional SVM treats the observations for each class equally. When the observations for each class are unequal, especially the observation for one of classes being relatively small to others, the SVM does not consider the classes corresponding to the smallest numbers of observations. Ignoring the minority classes the traditional SVM can decrease the overall misclassification rate. However, the minority classes have important characteristics for the data. For example, the cancer patients class from the data of hospital patients is much smaller than other patients. However, the cancer class

has significant information on the death rate and so the cancer class should not be ignored.

In this paper we develop SVM algorithms to treat unbalanced cases based on the weights of the loss function and the SCAD penalty function. Since it increases the impact of the minority classes, the result of classification does not ignore the minority classes while the traditional SVM ignores the minority classes. We used the local linear approximation (LLA) of the SCAD function to cover the deficiencies of the LQA algorithm.

This paper is organized as follows. In Section 2 we review the SVM for unbalanced cases and its statistical properties. We consider two classifiers based on the overall misclassification rate and the sum of within group error rate. The latter classifier does not ignore the minor classes and it is more applicable to unbalanced cases. The L_1 and the SCAD penalty functions are briefly reviewed. Section 3 gives an algorithm to implement the proposed method. Reference [6] proposed a LQA for the SVM with the SCAD function. We use a LLA algorithm which can be written by a linear programming problem to minimize the nondifferentiable and nonconvex objective function. Section 4 provides the simulation results and a numerical example. They show that the performance of the proposed method is superior to the traditional SVM in the view of the consideration of the minority classes. Section 5 gives some discussion and concluding remarks.

2. SVM for Unbalanced Cases

For a multiclass classification problem, a training sample $\{(\mathbf{x}_i, y_i)\}$, $i = 1, \dots, n$, is given, where the input vector $\mathbf{x}_i \in R^d$ and the output label $y_i \in \{1, \dots, K\}$, n is the number of observations, and d is the dimension of the input vector. Suppose that the samples are drawn from an unknown joint probability distribution $P(\mathbf{x}, y)$.

The classifier $\phi : R^d \rightarrow \{1, \dots, K\}$ trained by the training sample can predict the class of future input vector $\mathbf{x}$. The standard classification criterion gives the same misclassification costs for different classes. The loss function in this case is the 0-1 loss function, $l(y, \phi(\mathbf{x})) = I(y \neq \phi(\mathbf{x}))$, and the risk function corresponding to the 0-1 loss function can be written as $\sum_{i=1}^n I(y_i \neq \phi(\mathbf{x}_i))/n$ in the empirical version. Here the function $I(x)$ implies the indicator function. However, when there are minority classes in the output label, the classifier based on the overall misclassification rate does not give the information on the minority classes having very important characteristics of the data. For example, the proportion of the cancer patients is a minor class among general patients visiting a hospital. However, the cancer patients can be especially considered for the hospital. Unbalanced proportion samples are often found in real world data.

The classical criterion finds a decision rule minimizing the overall misclassification rate:

$$E[I(\phi(\mathbf{X}) \neq Y)] = P(\phi(\mathbf{X}) \neq Y) \\ = \sum_{k=1}^K E[I(\phi(\mathbf{X}) \neq k) | Y = k] \pi_k, \quad (1)$$

where $(\mathbf{X}, Y)$ is a $(d + 1)$ -dimensional random vector from a probability distribution $P(\mathbf{X}, Y)$ and π_k is the proportion of class k in the population. If π_k is very small, the overall misclassification rate (1) can be very small, because the misclassification rate for the k th class does not influence the quantity. Even if the misclassification rate for the k th class is very high, it could be ignored. Thus if the proportion term π_k in (1) was deleted, then the minor class having very small π_k could influence the classifier. We consider a classifier which minimizes

$$\sum_{k=1}^K E[I(\phi(\mathbf{X}) \neq k) | Y = k], \quad (2)$$

by discarding the term π_k in (1). Denote $e_k = E[I(\phi(\mathbf{X}) \neq k) | Y = k]$, $k = 1, \dots, K$, as the within group error for the k th class. The empirical version of term (2) can be calculated as the sum of the ratio of the number of misclassifications in class k , the sum of the within group error rate (see [7]). Criterion (2) is called the within group error rate criterion.

Research on this topic has focused on the methods at the data and algorithmic levels and it can be categorized such as resampling methods for balancing the dataset, modification of existing learning algorithm. Undersample and oversample are methods of resampling technique. Reference [7] proposed an adaptive weighted learning using the SVM.

The SVM is a classification method based on a large margin. It finds the hyperplane maximizing the separation distance between the two classes in the separable case. The SVM uses the slack variables, the so-called soft margin method to minimize the separation distance with the mislabeled samples. There are several extensions from the binary SVM to the multiclass SVM such as one-versus-one and one-versus-rest (see [8]). However, these methods provided poor performance when the data is dominated by only one class. Reference [9] proposed a simultaneous multiclass SVM. Reference [5] suggested a simultaneous SVM algorithm with the SCAD penalty function.

The multiclass SVM minimizes the objective function (see [9])

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K I(y_i \neq k) (f_k(\mathbf{x}_i) + 1)^+ + \lambda \sum_{k=1}^K \|\mathbf{w}_k\|^2, \quad (3)$$

where $z^+ = \max(z, 0)$, the function $f_k(\mathbf{x})$ is the k th element of the K -dimensional decision function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_K(\mathbf{x}))$ with a sum-to-zero constraint $\sum_{k=1}^K f_k(\mathbf{x}) = 0$, and the parameter λ controls the trade-off between the training error and the model complexity. Equation (3) is limited to the linear SVM, where $f_k(\mathbf{x}) = b_k + \sum_{j=1}^d w_{jk} x_j$. The SVM objective function consists of the loss function and the penalty function, which is the same as the objective function in penalized linear regression (see [10]).

For a sparse model the L_1 SVM as in [11] provides the information on valuable variables by discarding redundant noise input variables like the LASSO as in [3] which is popular in penalized linear regression models. Reference [12] proposed a multiclass SVM, which performs classification

and variable selection simultaneously through an L_1 -norm penalized sparse representation. However, the L_1 solution is biased for large coefficients. Reference [4] proposed a nonconvex penalty function, the SCAD penalty function, and desirable properties for penalty functions such as the unbiasedness of the estimator, the sparseness of the model, and the continuity of the estimator on the tuning parameter λ . Unfortunately, the L_2 penalty does not satisfy the unbiasedness and the sparsity, and the L_1 penalty does not satisfy the unbiasedness.

The SCAD function can be written by

$$p_\lambda(|x|) = \begin{cases} \lambda|x| & \text{if } |x| \leq \lambda \\ \frac{|x|^2 - 2a\lambda|x| + \lambda^2}{2(a-1)} & \text{if } \lambda < |x| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |x| > a\lambda, \end{cases} \quad (4)$$

where $a > 2$ and $\lambda > 0$. Reference [4] recommended that the parameter $a = 3.7$ from the simulation results. Since the derivatives of the SCAD function outside the range $[-a\lambda, a\lambda]$ are zero, the SCAD SVM estimates have the unbiasedness. Because the SCAD function is singular at zero, the SCAD SVM provides a sparse model. Like the standard SVM (3) the SCAD SVM minimizes the objective function

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K I(y_i \neq k) (f_k(\mathbf{x}_i) + 1)^+ + \sum_{k=1}^K p_\lambda(\|f_k\|). \quad (5)$$

In linear SVM case, $f_k(\mathbf{x}_i) = b_k + \mathbf{w}_k^T \mathbf{x}_i$ gives the objective function of the SCAD SVM:

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K I(y_i \neq k) (\mathbf{w}_k^T \mathbf{x}_i + b_k + 1)^+ + \sum_{k=1}^K \sum_{j=1}^d p_\lambda(|w_{jk}|), \quad (6)$$

$$\text{subject to } \sum_{k=1}^K b_k = 0, \quad \sum_{k=1}^K w_{jk} = 0, \quad \text{for } j = 1, \dots, d. \quad (7)$$

The classical classification rule (1) naturally becomes

$$\phi(\mathbf{x}) = \underset{k}{\operatorname{argmax}} f_k(\mathbf{x}). \quad (8)$$

And the classifier based on the within group error rate (2) can be written by

$$\phi(\mathbf{x}) = \underset{k}{\operatorname{argmax}} \frac{f_k(\mathbf{x})}{\pi_k}. \quad (9)$$

If one class dominates all population, in the point of the misclassification rate all of input vector can be classified as the dominating class. However, it does not yield a meaningful classifier. For unbalanced proportion data (6) cannot detect the minority class. The equal hinge loss function in (6) can ignore the minority class and so the unequal hinge loss function with the weight for the class can solve the unbalanced case as in [7]. Now instead of the unweighted SVM objective

function (6) we consider a weighted SVM objective function with the SCAD penalty function in unbalanced cases:

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K v_{ik} I(y_i \neq k) (\mathbf{w}_k^T \mathbf{x}_i + b_k + 1)^+ + \sum_{k=1}^K \sum_{j=1}^d p_\lambda(|w_{jk}|), \quad (10)$$

$$\text{subject to } \sum_{k=1}^K b_k = 0, \quad \sum_{k=1}^K w_{jk} = 0, \quad \text{for } j = 1, \dots, d, \quad (11)$$

where v_{ik} is the weight for the i th observation with the k th class. A weighted SVM is proposed for the robustness of the SVM which is not sensitive to outlier or leverage points (see [13]).

We consider the weight for each class as

$$v_{ik} = \frac{1}{\hat{\pi}_k}, \quad (12)$$

where $\hat{\pi}_k$ is the estimate of the population proportion of the k th class as in [7]. It is called a proportional weight to the number of observations for each class. Weight (12) considers only the unbalanced proportion. When the dataset has outlying observations, the SVM based on weight (12) may not give the exact underlying decision function $f_k(\mathbf{x})$. Thus we propose the weight for the data as

$$v_{ik} = v_k(\mathbf{x}_i) = \left(\frac{1}{\hat{\pi}_k} \right)^{\hat{e}_k} \times \frac{1}{1 + |f_k(\mathbf{x}_i)|}, \quad (13)$$

where $\hat{e}_k$ denotes the within group errors on the training dataset with equal weights. We call this weight an adaptive weight. The within group errors can be calculated as the misclassification rate for the k th class. Weight (13) gives much more weights on the minority class and the well-classified group got the less weight. The larger values of $|f_k(\mathbf{x}_i)|$ in (13) represent well-classified observations. Thus well-classified observations are given by small weights. Also the small values of $|f_k(\mathbf{x}_i)|$ imply the corresponding observations to be misclassified. Therefore, the corresponding weights get larger and the learning machine keeps the observations having important information. The proposed weight (13) considers both robustness and unbalanced proportion of the data, because the terms $(1/\hat{\pi}_k)^{\hat{e}_k}$ and $1/(1 + |f_k(\mathbf{x}_i)|)$ reflect the unbalanced proportion of the data and the resistance to outlying observations, respectively. However, weight (12) only considers the unbalanced proportion for the classes.

3. Algorithm

Since the SCAD function is not convex, the objective function in (10) is not convex. Thus standard optimization techniques cannot be adopted. Usually the nonconvex penalty can give good statistical properties but the implementation is not easy. Reference [4] used the LQA algorithm which has drawbacks like the backward elimination. That is, for numerical stability if the coefficient of the variable is close to zero, the variable

was deleted and it is not included in the final model. Furthermore, the solution of the LQA algorithm can be written in a Ridge-type and it does not guarantee the sparseness of the solution like the property of the Ridge regression estimator. Reference [14] proposed a perturbation of the LQA algorithm and the proposed algorithm renders the objective function differentiable. Then it optimizes this differentiable function using a minorize-maximize algorithm. But it is not easy to select the size of perturbation. Reference [15] proposed a methodology on a one-step sparse estimation procedure in nonconcave penalized likelihood models which is called the LLA algorithm. It has neither the drawbacks of the LQA algorithm nor the numerical instability at zero.

By the Taylor expansion of the SCAD function we obtain the following approximation equality:

$$p_\lambda(|x|) \approx p_\lambda(|x_0|) + p'_\lambda(|x_0|)(|x| - |x_0|) \quad \text{for } x \approx x_0, \quad (14)$$

where the first derivative function of the SCAD function $p'_\lambda(|x|)$ is

$$p'_\lambda(|x|) = \begin{cases} \lambda & \text{if } 0 \leq |x| < \lambda \\ \frac{a\lambda - |x|}{a-1} & \text{if } \lambda \leq |x| < a\lambda \\ 0 & \text{if } |x| \geq a\lambda. \end{cases} \quad (15)$$

By putting (14) into (10), we obtain the objective function up to constants

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K v_{ik} I(y_i \neq k) (\mathbf{w}_k^T \mathbf{x}_i + b_k + 1)^+ \\ & + \sum_{k=1}^K \sum_{j=1}^d p'_\lambda(|w_{jk}^0|) |w_{jk}|, \end{aligned} \quad (16)$$

where w_{jk}^0 is an initial solution near the true value w_{jk} and the restriction of (10) is still effective. We introduce the variable $\xi_l = v_{ik} I(y_i \neq k) (\mathbf{w}_k^T \mathbf{x}_i + b_k + 1)^+$ with adjusted subindex l and use the fact that $|z| = z^+ + z^-$, $z = z^+ - z^-$, and $z^+ \geq z$, where z^- is defined similar to z^+ . Then the weighted SVM objective function (10) becomes

$$\frac{1}{n} \sum_{l=1}^{n(K-1)} \xi_l + \sum_{k=1}^K \sum_{j=1}^d p'_\lambda(|w_{jk}^0|) (w_{jk}^+ + w_{jk}^-), \quad (17)$$

$$\text{subject to } v_{ik} \left(\sum_{j=1}^d x_{ij} (w_{jk}^+ - w_{jk}^-) + b_k^+ - b_k^- + 1 \right) \leq \xi_l,$$

$$\sum_{k=1}^K (b_k^+ - b_k^-) = 0, \quad \sum_{k=1}^K (w_{jk}^+ - w_{jk}^-) = 0,$$

$$\xi_l \geq 0, \quad w_{jk}^+ \geq 0, \quad w_{jk}^- \geq 0,$$

$$b_k^+ \geq 0, \quad b_k^- \geq 0,$$

$$\text{for } i = 1, \dots, n; \quad j = 1, \dots, d;$$

$$k = 1, \dots, K. \quad (18)$$

The first derivative function of the SCAD function (4) was evaluated at the initial value w_{jk}^0 . Equation (17) can be minimized by standard optimization packages. We can obtain the optimum solution $b_k^+, b_k^-, w_{jk}^+, w_{jk}^-$ and then the parameters can be estimated by $b_k = b_k^+ - b_k^-$ and $w_{jk} = w_{jk}^+ - w_{jk}^-$. Equation (17) can be solved by lpSolve in R program.

The linear programming problem can be formulated by

$$\begin{aligned} & \min_{\mathbf{z} \in \mathbb{R}^N} \quad \mathbf{c}' \mathbf{z} \\ & \text{subject to } \mathbf{A} \mathbf{z} \geq \mathbf{f}, \quad \mathbf{z} \geq \mathbf{0}, \end{aligned} \quad (19)$$

where $\mathbf{z}$ is an N -vector of variables with $N = (n+2d+1)K - n$, which is composed of the variables $\xi_l, w_{jk}^+, w_{jk}^-, b_k^+, b_k^-$. Now we set the constraints $\mathbf{A}_{n(K-1) \times N}$, $\mathbf{f}_{n(K-1) \times 1}$ and the coefficient vector of the objective function from the linear programming problem in (17).

Since the SCAD function is approximated by the absolute function, the method is similar to the L_1 penalty SVM. From the point of the variable selection view it is well known that the L_1 SVM gives very useful information on the variable selection for classification with greatly exceeding the number of predictors (see [12]). Thus the proposed method is very effective in variable selection, especially when the number of variables is large.

We summarize the proposed algorithm for the weighted multiclass SVM with the SCAD penalty function by the following steps.

(1) Tuning process is as follows.

- (1.1) Find the values w_{jk}, b_k which minimize (17) for tuning data with initial estimates w_{jk}^0, b_k^0 and $v_{ik} = 1$ from $\lambda = 2^{-10}$ to $\lambda = 2^{10}$.
- (1.2) Obtain the value λ minimizing the misclassification rate which is the overall misclassification rate (CLSC) or the within group error rate (MWGE).

(2) Training process is as follows.

- (2.1) Set the initial solutions b_k^0, w_{jk}^0 for $j = 1, \dots, d$ and $k = 1, \dots, K$.
- (2.2) Solve the linear programming problem (17) for training data with $v_{ik} = 1$ for the given tuning parameter λ .
- (2.3) Set the weights v_{ik} by (12) or (13) with the above solution.
- (2.4) Solve (17) and set the parameters w_{jk} and b_k for $j = 1, \dots, d$ and $k = 1, \dots, K$.

(3) Test process is as follows.

- (3.1) Calculate the misclassification rates for test data based on the tuning parameter λ and the parameters w_{jk} and b_k .

4. Numerical Experiment

In this section, we compare the performance of the proposed algorithm and the classical SVM of the multiclass SVM for unbalanced proportion data. We conducted the experiments on simulation and a real dataset.

4.1. Simulation. We consider a simple three-class dimensional data. The class information on the i th data can be randomly allocated according to the proportions $\pi_1 : \pi_2 : \pi_3 = 1:2:2$ and $1:4:4$, respectively. Here we can see that the first class is the minority class having the expected proportions $1/5$ and $1/9$, respectively. The training data are generated as follows. First, t_1, t_2 are generated from the t distribution with degrees of freedom 3. Second, the responses are allocated randomly and the predictors $x_1 = t_1 + a_1$ and $x_2 = t_2 + a_2$, where $(a_1, a_2) = (\sqrt{3}, 1), (-\sqrt{3}, 1), (0, -2)$ for classes 1–3, respectively (see [16]). The number of observations is 20, 30, and 30 for each class in the data with the proportion $1:2:2$ and 20, 80, and 80 for the other data.

Our simulation data consists of training data, tuning data, and independent test data. The coefficients w_{jk}, b_k are estimated based on the training data and the performance is evaluated on the test data of the sample sizes 10,000 and 18,000 generating from the same distribution described above. The tuning data of sizes 100 and 180 are used to determine the tuning parameter λ . We conducted simulation iterations to evaluate the within group error for each class. All simulations are carried out using *R* program.

Tables 1 and 2 summarize the within group error rate for equal weights, proportion weights (12), and the adaptive weight (13) combining the proportion, misclassification rates, and the robustness to outliers. The values in the tables are evaluated on the test data. We consider two classifiers, the classical classifier (CLSC) of (1) and the minimum within group error (MWGE) of (2). We compared the proposed algorithm with the equal weight SCAD SVM which is a classical SCAD SVM. Tables 1 and 2 show that the equal weight scheme ignores the minor class 1 to minimize the overall misclassification rate. However, the proposed algorithms (proportional weights and adaptive weights) did not ignore the minor class 1 which may be a main interesting class. For the equal weight scheme the misclassification rate for class 1 is close to 1, but the overall misclassification rate is not high, because class 1 has the small proportion. Particularly for the proportion $1:4:4$ case the equal weights scheme did not discriminate class 1 at all. It is not desirable, when class 1 itself is the main interest. As the degree of unbalanced data goes up, the proposed algorithms will get the effectiveness of weights.

We can see that the MWGE method is more powerful compared to the CLSC method as the minority of the data become higher with proportions changing from $1:2:2$ to $1:4:4$. Two weight schemes have similar results in terms of within group error rate. Tables 1 and 2 give us that the proposed weighted SVM helps to decrease the unbalanced proportion of data.

The final weights for classes are reported in Table 3. The adaptive weight scheme gives more large weights on the minor class than the equal weight or the proportional weight.

TABLE 1: The within group error rates of the weighted SCAD SVMs in the unbalanced case with proportion $1:2:2$.

	CLSC			MWGE		
Equal weights	0.891 (0.185)	0.119 (0.092)	0.104 (0.029)	0.836 (0.248)	0.121 (0.092)	0.105 (0.025)
Proportional weights	0.176 (0.085)	0.164 (0.077)	0.199 (0.112)	0.141 (0.096)	0.184 (0.063)	0.238 (0.143)
Adaptive weights	0.092 (0.035)	0.240 (0.167)	0.422 (0.290)	0.070 (0.025)	0.298 (0.188)	0.471 (0.283)

TABLE 2: The within group error rates of the weighted SCAD SVMs in the unbalanced case with proportion $1:4:4$.

	CLSC			MWGE		
Equal weights	1.000 (0.000)	0.107 (0.019)	0.105 (0.031)	1.000 (0.000)	0.106 (0.018)	0.105 (0.031)
Proportional weights	0.165 (0.092)	0.149 (0.042)	0.236 (0.162)	0.106 (0.080)	0.212 (0.094)	0.304 (0.159)
Adaptive weights	0.098 (0.028)	0.229 (0.180)	0.421 (0.314)	0.068 (0.015)	0.302 (0.168)	0.458 (0.271)

It gives that the adaptive weights scheme has small within group error for the minor class 1 in Tables 1 and 2.

4.2. Real Dataset. In this section, we apply the proposed method to the thyroid gland database from the UCI Machine Learning website <http://www.ics.uci.edu/~mllearn/>. The data was used to try to predict whether a patient's thyroid will be classified as the class euthyroidism, hypothyroidism or hyperthyroidism. The diagnosis (the class label) was based on 5 medical records such as anamnesis and scan which are continuous. Three classes are normal (class 1), hyperthyroidism (class 2), and hypothyroidism (class 3). The data has 215 observations which consisted of 150 observations (class 1), 35 observations (class 2), and 30 observations (class 3). It is unbalanced data because near 70 percent of the patients are normal (class 1). Classes 2 and 3 are minority classes.

For evaluation of the performance of the proposed algorithm we divide the data randomly into three categories such as a training set with 72 observations, a tuning set with 72 observations, and a test set with 71 observations. The training and tuning sets have the same class proportion $25:6:5$. We obtained the tuning parameter λ of SVM from the tuning data and the estimates of the model parameters w_{jk}, b_k of SVM from the training data. The within group error rates of the SVMs are evaluated based on the test data.

We conducted 100 replications of such experiment and summarized the results in Tables 4 and 5. We can see that the test errors for classes 2 and 3 (patients class) using the equal weights scheme are higher than class 1, because class 1 (normal class) dominates highly the classes of the data. The proportional weights and the adaptive weights algorithms perform better in discriminating the patients from the normal group, while the correction rate of the normal group becomes low. In Table 5 the weights for class 1 in the proportional and adaptive weights schemes become low and

TABLE 3: The average final chosen weights for the weighted SCAD SVMs on the simulation data.

	1:2:2						1:4:4					
	CLSC			MWGE			CLSC			MWGE		
Equal weights	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Proportional weights	0.50	0.25	0.25	0.50	0.25	0.25	0.67	0.16	0.16	0.67	0.16	0.16
Adaptive weights	0.53	0.23	0.24	0.54	0.23	0.23	0.68	0.16	0.16	0.69	0.15	0.16

TABLE 4: The within group error rates of the weighted SCAD SVMs on the new thyroid data.

	CLSC			MWGE		
Equal weights	0.007	0.374	0.283	0.017	0.245	0.238
	(0.013)	0.172	0.124)	(0.022)	0.160	0.123)
Proportional weights	0.162	0.139	0.242	0.201	0.053	0.189
	(0.214)	0.144	0.155)	(0.185)	0.083	0.124)
Adaptive weights	0.006	0.394	0.281	0.014	0.275	0.243
	(0.012)	0.176	0.124)	(0.023)	0.201	0.120)

TABLE 5: The average final chosen weights for the weighted SCAD SVMs on the new thyroid data.

	CLSC			MWGE		
Equal weights	0.33	0.33	0.33	0.33	0.33	0.33
Proportional weights	0.10	0.41	0.49	0.10	0.41	0.49
Adaptive weights	0.07	0.42	0.51	0.07	0.42	0.51

TABLE 6: The overall misclassification rate of the weighted SCAD SVMs on the new thyroid data.

	CLSC		MWGE	
Equal weights	0.103	(0.031)	0.083	(0.029)
Proportional weights	0.170	(0.136)	0.176	(0.124)
Adaptive weights	0.105	(0.031)	0.087	(0.035)

it makes the within group error rate increase for class 1. The weights for classes 2 and 3 become high. However, the weight for all classes in the classical weight scheme is $1/3$. It does not consider the unbalanced proportion of the data.

We summarize the overall misclassification rate in Table 6. It shows that the classical equal weight scheme with MWGE classifier has the best performance in the view of overall misclassification rate. However, the within group error rates of classes 2 and 3 for this scheme are higher than those of the proportional weight scheme. We should consider the weight SVM for the classification of a thyroid patient.

Suppose that a thyroid patient was classified as a normal person. The patient missed the opportunity to be given medical treatment. It may be a dangerous situation. This example shows that the minor classes should be significantly treated. We recommend that we use both the classical equal weight scheme and the weighted SVM.

5. Conclusion

In this paper we proposed the weighted SCAD SVM for unbalanced multicategory classification problems. When a class has small proportion among all classes, numerical simulation and data analysis show that the proposed method is a considerable method, especially in the view of the within group error rates of the minor classes. We will study the method to consider both the overall misclassification rate and the within group error rate in the future.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Basic Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Engineering (NRF-2012R1A1A4A01004594).

References

- [1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, 2001.
- [2] K.-M. Jung, "Robust statistical methods in variable selection," *Journal of the Korean Data Analysis Society*, vol. 10, pp. 3057–3066, 2008.
- [3] R. J. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.
- [4] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [5] K.-M. Jung, "Multiclass support vector machines with SCAD," *Communications of the Korean statistical society*, vol. 19, no. 5, pp. 655–662, 2012.
- [6] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.
- [7] X. Qiao and Y. Liu, "Adaptive weighted learning for unbalanced multicategory classification," *Biometrics*, vol. 65, no. 1, pp. 159–168, 2009.
- [8] J. Weston and C. Watkins, "Support vector machines for multiclass pattern recognition," in *Proceedings of the 7th European Symposium on Artificial Neural Networks*, pp. 219–224, 1999.
- [9] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: theory and application to the classification of

- microarray data and satellite radiance data,” *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67–81, 2004.
- [10] G. Wahba, “Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV,” in *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. J. C. Burges, and A. J. Smola, Eds., pp. 69–87, MIT Press, 1998.
- [11] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, “1-norm support vector machines,” in *Advances in Neural Information Processing Systems*, vol. 16, pp. 49–56, The MIT Press, Cambridge, Mass, USA, 2004.
- [12] L. Wang and X. Shen, “On L_1 -norm multiclass support vector machines: methodology and theory,” *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 583–594, 2007.
- [13] K.-M. Jung, “Weighted support vector machines with the SCAD penalty,” *Communications for Statistical Applications and Methods*, vol. 20, no. 6, pp. 481–490, 2013.
- [14] D. R. Hunter and R. Li, “Variable selection using MM algorithms,” *The Annals of Statistics*, vol. 33, no. 4, pp. 1617–1642, 2005.
- [15] H. Zou and R. Li, “One-step sparse estimates in nonconcave penalized likelihood models,” *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.
- [16] Y. Liu and X. Shen, “Multicategory ψ -learning,” *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 500–509, 2006.