Biometrics 65, 159–168
March 2009
DOI: 10.1111/j.1541-0420.2008.01017.x

Adaptive Weighted Learning for Unbalanced Multicategory Classification

Xingye Qiao and Yufeng Liu*

Department of Statistics and Operations Research, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
**email:* yfliu@email.unc.edu

Summary. In multicategory classification, standard techniques typically treat all classes equally. This treatment can be problematic when the dataset is unbalanced in the sense that certain classes have very small class proportions compared to others. The minority classes may be ignored or discounted during the classification process due to their small proportions. This can be a serious problem if those minority classes are important. In this article, we study the problem of unbalanced classification and propose new criteria to measure classification accuracy. Moreover, we propose three different weighted learning procedures, two one-step weighted procedures, as well as one adaptive weighted procedure. We demonstrate the advantages of the new procedures, using multicategory support vector machines, through simulated and real datasets. Our results indicate that the proposed methodology can handle unbalanced classification problems effectively.

Key words: Adaptive learning; Mean within group error; Multicategory classification; Unbalanced data; Weighted learning.

1. Introduction

Classification, as a means for information extraction, is a very commonly used statistical tool. It has been successfully applied in many different fields such as engineering, health science, and social science. There are numerous classification methods proposed in the literature. Well-known ones include the k-nearest neighbors (k-NN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), the logistic regression, and the support vector machine (SVM) among others. See Hastie, Tibshirani, and Friedman (2001) and Duda, Hart, and Stork (2001) for a comprehensive review of different classification techniques.

Despite progresses in classification, problems may still occur when the training sample or population is unbalanced, i.e., some classes have relatively small proportions compared to others. For example, suppose we have a three-class problem with class proportions (0.49, 0.49, 0.02). A classifier misclassifying all points in class 3 may still give 90% overall classification accuracy. However, it is not a desirable classifier because it cannot discriminate class 3 at all. This phenomenon is especially problematic if class 3 itself is our main interest. For instance, in a study of a certain rare disease, one hopes to build a classification rule to classify the disease versus normal patients. Compared to the normal group, the disease group sample can be very small. The measure of overall classification accuracy may be misleading if the overall accuracy is high whereas the accuracy for the disease group is very poor. This calls for alternative classification criteria to build and select classifiers.

In the literature, a lot of work has been done to deal with unbalanced classification by adjusting the class proportions. One approach is to undersample the classes with larger proportions and the other one is to oversample the classes with small proportions. See an extensive survey on this topic by Chawla, Japkowicz, and Kolcz (2004). Recently, Owen (2007) studied the asymptotic behavior of binary logistic regression, where one class has a finite sample size and the other class's sample size grows to infinity. The asymptotic results help to compute the logistic regression more efficiently for unbalanced data.

In this research, we first discuss the reason why minority classes can be ignored by a classifier if the overall misclassification rate is used as the evaluation criterion. Based on the fact that one may be more interested in the classification performance on a certain particular class, we propose two new criteria—the mean within group error criterion and the mean square within group error criterion in Section 2. To cope with the difficulty of unbalanced datasets, in Section 3 we propose weighted learning to increase the impact of the minority classes so that they will not be ignored due to their small proportions. Three different weighting schemes are given: The first is the one-step weighting based on class proportions. The second is the one-step weighting using both the information of class proportions and the information of within-group misclassification rates. The last proposal is to use adaptive weighted learning, which modifies the weights for different classes adaptively. The proposed weighted learning methods can be incorporated into many existing classification techniques. The algorithm for the adaptive weighted learning is given in Section 4. In Section 5, we illustrate the proposed weighted learning methods on the SVM. Numerical studies on simulated and real data

are given in Section 6. Some final discussion is included in Section 7.

2. Classification Criteria and Bayes' Rules

2.1 *Multicategory Classification*

In a multicategory classification problem, we are given a training dataset with n observations $\{\mathbf{x}_i,\ y_i\}$, $i = 1, \ldots, n$. Here each input $\mathbf{x}_i$ is a d-dimensional vector and y_i is the class label, indicating the class that $\mathbf{x}_i$ belongs to. Suppose there are k different classes with $y_i \in \{1,\ 2, \ldots, k\}$, and suppose that $\{\mathbf{x}_i,\ y_i\}$'s are drawn from an unknown population, with the underlying $(d + 1)$-dimensional joint probability distribution $P(\mathbf{x},\ y)$.

The goal of multicategory classification problem is to construct a decision rule $\phi(\mathbf{x}) : R^d \to \{1,\ 2, \ldots, k\}$, based on the information of the training data $\{\mathbf{x}_i,\ y_i\}$'s, to predict the class label for future input $\mathbf{x}$. Here $\phi(\mathbf{x})$ needs to not only classify those $\mathbf{x}_i$'s in the training data well, but also have good generalization ability, so that it works well on the entire population for prediction.

Let $(\mathbf{X},\ Y)$ denote a $(d + 1)$-dimensional random vector from a probability distribution $P(\mathbf{X},\ Y)$, with the joint probability density $g(\mathbf{x},\ y)$. Let $p_j(\mathbf{x}) = P(Y = j \,|\, \mathbf{X} = \mathbf{x})$; $j = 1,\ 2, \ldots, k$. Denote the conditional probability density of $\mathbf{X}$ given $Y = j$ as $g_j(\mathbf{x}) = g_{\mathbf{X}|Y=j}(\mathbf{x})$. Let π_j denote the proportion of class j in the population, i.e., $\pi_j = P(Y = j)$, the marginal probability mass function of Y. Then by the Bayes formula, we have $p_j(\mathbf{x}) = P(Y = j \,|\, \mathbf{X} = \mathbf{x}) = g(\mathbf{x}, j)/\sum_{i=1}^{k} g(\mathbf{x}, i) = g_j(\mathbf{x})\pi_j/\{\sum_{i=1}^{k} g_i(\mathbf{x})\pi_i\}$.

Once $p_j(\mathbf{x})$; $j = 1, 2, \ldots, k$ is given, one can form a classification rule according to a given criterion. In Sections 2.2 and 2.3, we discuss different classification criteria and the corresponding Bayes' rules.

2.2 *The Classic Classification Criterion*

In the standard case where the misclassification costs for different classes are all equal, the associated loss function is the $0 - 1$ loss function, i.e., $l(y,\ \phi(\mathbf{x})) = I\{y \neq \phi(\mathbf{x})\}$. The corresponding Bayes decision rule, $\phi_B(\mathbf{x})$, minimizes the posterior expected risk $\rho(\mathbf{x},\ \phi)$, which can be written as: $\rho(\mathbf{x},\ \phi) = E[I\{Y \neq \phi(\mathbf{X})\} \,|\, \mathbf{X} = \mathbf{x}] = 1 - p_{\phi(\mathbf{x})}(\mathbf{x})$. Then we can simplify $\phi_B(\mathbf{x})$ as $\phi_B(\mathbf{x}) = arg\min_{j=1,\ldots,k}\{1 - p_j(\mathbf{x})\} = arg\max_{j=1,\ldots,k}\ p_j(\mathbf{x})$.

Note that the risk function corresponding to the $0 - 1$ loss represents the probability of misclassification given $\mathbf{x}$. Its empirical version can be written as $\sum I\{\phi(\mathbf{x}_i) \neq y_i\}/n$, the ratio between the number of misclassified observations and the total number of observations. This is a commonly used classification criterion, namely "the overall misclassification rate."

It is interesting to note that if there exists $j \in \{1, \ldots, k\}$, such that $\forall \mathbf{x} \in R^d$, $p_j(\mathbf{x}) \neq \max_{i\in\{1,\ldots,k\}} p_i(\mathbf{x})$, then the Bayes rule never classifies any $\mathbf{x}$ to class j. The class j will be ignored with an empty classification set. By the Bayes formula, $p_j(\mathbf{x}) = \{g_j(\mathbf{x})\pi_j\}/\{\sum_{i=1}^{k} g_i(\mathbf{x})\pi_i\}$. Thus as π_j gets small enough, so does $p_j(\mathbf{x})$, even though $g_j(\mathbf{x})$ might be large. For this reason, in an unbalanced scenario where a minority class has very small π_j compared to the other classes, the minority class can be ignored by both the Bayes decision rule and any classification methods based on it, if the overall misclassification rate is used as the classification criterion.

We now consider a simple example to illustrate the problem of minority classes. Consider a three-class classification problem. The proportion $\pi_1 : \pi_2 : \pi_3$ for these three classes is $3 : 3 : 1$. The conditional distributions for these three classes are $N(-0.5,\ 0.4)$, $N(0.5,\ 0.4)$, and $N(0,\ 0.4)$, respectively. The left panel in Figure 1 displays the class-conditional probability density functions $g_j(\mathbf{x})$; $j = 1,\ 2,\ 3$, whereas the right panel shows the posterior probability functions $p_j(\mathbf{x})$; $j = 1,\ 2,\ 3$, given the information that $\pi_1 : \pi_2 : \pi_3 = 3 : 3 : 1$. As indicated in Figure 1, if the overall misclassification rate is used as the classification criterion, the Bayes rule assigns $x \in [0,\ 0.5]$ for class 1 and $x \in [0.5,\ 1]$ for class 2. Thus class 3 is totally ignored simply because of its small proportion $\pi_3 = 1/7$. This calls for alternative criteria that can deal with the unbalanced class proportions.

2.3 *Mean Within Group Error Rate and Mean Square Within Group Error Rate*

In this section, we propose two alternative classification criteria to take minority classes into account.

As stated in Section 2.2, the classic criterion aims to find a decision rule minimizing the overall misclassification rate, $E[I\{\phi(\mathbf{X}) \neq Y\}]$, which can be written as

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

Notice that in (1), if π_j is small, the contribution of class j to the error will be small. The term $E[I\{\phi(\mathbf{X}) \neq j\} \,|\, Y = j]$ represents the within group error for class j. To reduce the effect of π_j, we consider the average of within group errors. Then the corresponding objective function, after removing π_j's in (1) and up to a constant $1/k$, becomes,

$$\sum_{j=1}^{k} E[I\{\phi(\mathbf{X}) \neq j\} \,|\, Y = j] = E\left[\sum_{j=1}^{k} \frac{1}{\pi_j} \sum_{l\neq j} I\{\phi(\mathbf{X}) = l, Y = j\}\right] = E\left[\sum_{j=1}^{k} \frac{1}{\pi_j} \sum_{l=1}^{k} \bar{\delta}_{jl} I(Y = j) I\{\phi(\mathbf{X}) = l\}\right],$$

where $\bar{\delta}_{jl} = 0$ if $j = l$, and 1 otherwise.

The intuition for this new criterion is that instead of aiming for a classification rule with a small overall misclassification rate, we take into account of each within group error. To prevent a minority class from being ignored, we consider the average of within group error rates. Thus the corresponding classification rule will put more emphasis on minority classes compared to the classic one. We call this new criterion "mean within group error rate." If the within group error for a minority class is very large, it will increase the average and consequently help to reduce the unbalanced data effect.

As a remark, we note that the mean within group error rate criterion is closely related to existing classification measures

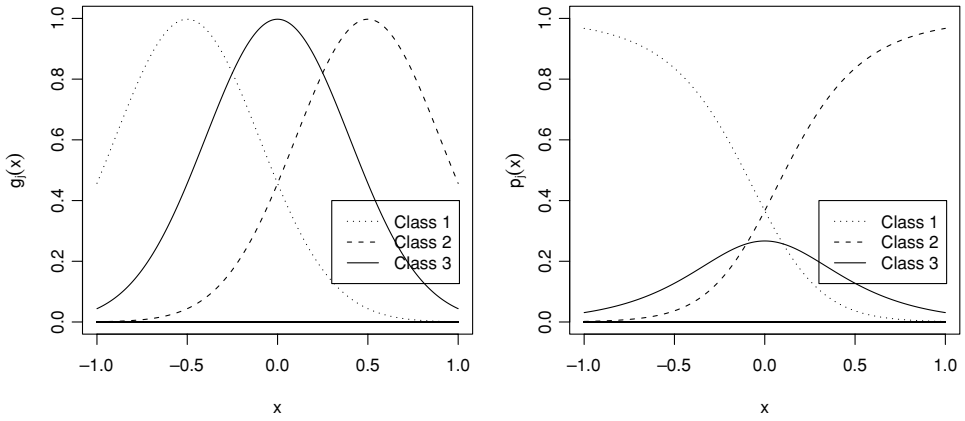

Figure 1. An illustrating plot of a three-class unbalanced problem with classes $1-3$ from $N(-0.5, 0.4)$, $N(0.5, 0.4)$, and $N(0, 0.4)$, respectively, with $\pi_1 : \pi_2 : \pi_3 = 3:3:1$. The left panel displays the class-conditional probability density functions $g_j(\mathbf{x})$ and the right panel displays the posterior class probability functions $p_j(\mathbf{x})$ given $\mathbf{X} = \mathbf{x}$; $j = 1, 2, 3$.

such as sensitivity and specificity for disease diagnosis. In disease diagnosis, sensitivity and specificity refer to the probability that the test is positive given that the patient is sick and the probability that the test is negative given that the patient is not sick, respectively (see, e.g., Yerushalmy, 1947; Leisenring, Alonzo, and Pepe, 2000). It is easy to show that the mean within group error rate can be calculated as $1 - (\text{sensitivity} + \text{specificity})/2$. Thus, the two criteria are equivalent for binary classification. However, one important difference here is that sensitivity and specificity are mainly for binary classification, whereas the new criteria can be used for general classification problems.

Proposition 1 below demonstrates the Bayes rules under the classic and the mean within group error criteria.

PROPOSITION 1: *The Bayes rules under the classic and the mean within group error criteria are* $\phi_B(\mathbf{x}) = arg\min_l \sum_{j=1}^k \bar{\delta}_{jl} p_j(\mathbf{x}) = arg\max_l p_l(\mathbf{x})$ *and* $\phi_B(\mathbf{x}) = arg\min_l \sum_{j=1}^k p_j(\mathbf{x})\bar{\delta}_{jl}/\pi_j = arg\max_l p_l(\mathbf{x})/\pi_l$, *respectively.*

Proposition 1 can be shown using the results of theoretical classification Bayes' rules under the general misclassification cost structure (Johnson and Wichern, 2001). To save space, its detailed proof is omitted.

We now introduce another criterion, which is an extension of the mean within group error. Sometimes we may want not only the average of the within-class misclassification rates to be small, but also the variation to be small as well, if all classes are equally important. For example, when we are given two classification rules with group errors of (20%, 22%) and (1%, 40%), even though the latter one has smaller average of within group error ($20.5\% < 21\%$), we may prefer the first one, because of its smaller within group error variation. Notice that the latter one performs well on one class but does badly on the other, whereas the first one performs reasonably well for both classes. If these two classes are equally important, the first classification rule may be more preferable.

Denote $e_j = E[I\{\phi(\mathbf{X}) \neq j\} \,|\, Y = j]$; $j = 1, \ldots, k$, as the within group error for the jth class, and $\bar{e} = \sum e_j/k$ as the average of within group errors. The sample version of $e_j, \hat{e}_j$, can be calculated as the ratio of the number of misclassifications in class j and total number of observations in class j. Now we try to minimize $\bar{e}^2 + \sum(e_j - \bar{e})^2/k = \sum e_j^2/k$, where the first term on the left-hand side is the square of the average within group error and the second term reflects the variation of within group errors for different classes. Then the new criterion, namely, the mean square within group error, tries to minimize the following

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

$$= \sum_{j=1}^k \frac{1}{\pi_j^2} E^2 \left[\sum_{l=1}^k \bar{\delta}_{jl} I\{\phi(\mathbf{X}) = l\} I(Y = j) \right].$$

Unlike the other two criteria, the Bayes rule for this criterion does not appear to have a close form.

For a given classification procedure, different criteria used can lead to different classifiers. Specifically, many procedures involve selection of multiple parameters. New criteria can be used to select these parameters. For example, the SVM involves several parameters, such as the regularization parameter λ and the parameters used in its kernel function. Selection of these parameters is an important step for choosing the final classifier. The new proposed criteria can be used for this purpose.

2.4 *Unequal Misclassification Costs*

So far we only focus on situations where the costs for different types of misclassification are all equal, that is, the cost of misclassifying an observation in class j to class l is the same for any $j \neq l \in \{1, \ldots, k\}$. However, in many problems, unequal misclassification costs are necessary. For example, misclassifying a cancer patient to be healthy is much more severe than classifying a healthy person to be a cancer patient. The first type of error may cause delay of early treatment of the patient and one should try to avoid that if possible. For the second type of error, it can be corrected using further diagnosis. Similarly, cost of misdetecting a good product to be flawed should

be much smaller than the cost of omitting a defective product into the market.

Denote C_{ij} as the cost of misclassifying an observation in class i to class j. Lin, Lee, and Wahba (2002) discussed the SVM for classification in a nonstandard situation, where the cost of different types of misclassification are not the same or there is sampling bias for the training set. Then the adjusted overall misclassification cost can be expressed as $E[\sum_{j=1}^{k}\sum_{l=1}^{k} C_{jl}\bar{\delta}_{jl}I(Y=j)I\{\phi(\mathbf{X})=l\}]$.

Under the mean within group error criterion, the objective function becomes $E[\sum_{j=1}^{k}\sum_{l=1}^{k} I(Y=j)I\{\phi(\mathbf{X})=l\}C_{jl}\bar{\delta}_{jl}/\pi_j]$. Thus, our proposed criteria in Sections 2.1 and 2.2 can be directly generalized to unequal cost situations. The following corollary gives the Bayes rules using two different criteria with unequal costs.

COROLLARY 1: *Let C_{ij} be the cost of misclassifying observation in class i to class j. Then the Bayes rules under the classic criterion and the mean within group error criterion are $\phi_B(\mathbf{x}) = arg\min_l \sum_{j=1}^{k} C_{jl}\bar{\delta}_{jl}p_j(\mathbf{x})$ and $\phi_B(\mathbf{x}) = arg\min_l \sum_{j=1}^{k} C_{jl}\bar{\delta}_{jl}p_j(\mathbf{x})/\pi_j$, respectively.*

3. Weighted Learning

In Section 2, we discussed different criteria and their associated Bayes' rules. In this section, we propose several weighted learning schemes for classification with unbalanced datasets.

Standard classifiers treat all classes equally. This may cause problems for unbalanced problems because the minority classes can be possibly ignored. Intuitively, one can put a relatively big weight for a minority class so that it cannot be ignored easily. For example, consider a binary problem with the class 1 to be the minority class and class 2 to be the majority class. Suppose we use the weights (4, 1) for the two classes. This implies that one misclassified point of class 1 is treated to be equivalent to four misclassified points of class 2. Using a bigger weight for class 1, one can increase the impact of class 1 for the classification rule so that it will not be ignored due to its small proportion. Denote $(w_1, \ldots, w_k)$ as the weights for k classes. In this section, we consider three different weighting schemes, the first two being one-step fixed weights and the third one being an adaptive procedure.

Without loss of generality, let $C_{ij} = 1$. Then analogous to the $0-1$ loss function, the expected weighted loss can be written as

$$E\left[\sum_{j=1}^{k} w_j I\{Y=j, \phi(\mathbf{X})\neq j\}\right]$$

$$= E\left[\sum_{j=1}^{k}\sum_{l=1}^{k} \bar{\delta}_{jl}w_j I(Y=j)I\{\phi(\mathbf{X})=l\}\right].$$

Clearly, the classic criterion and the mean within group error criterion are special cases with the weights being $(1, \ldots, 1)$ and $(1/\pi_1 : 1/\pi_2 : \ldots : 1/\pi_k)$, respectively. Similar to Proposition 1, we can show that the Bayes rule of the weighted learning is $\phi(\mathbf{x}) = arg\max_l w_l p_l(\mathbf{x})$. Similar results for other loss functions such as the hinge loss of the SVM can be derived as well.

In Sections 3.1 and 3.2, we discuss different ways of selecting the weights to achieve good classification performance under a given criterion.

3.1 *One-Step Fixed Weights*

A natural choice of weights is to make use of the true proportions $\{\pi_j\}$ of different classes, if they are available. Let $(w_1 : w_2 : \ldots : w_k) = (1/\pi_1 : 1/\pi_2 : \ldots : 1/\pi_k)$. Using this choice, we aim to put a big weight for class j if π_j is small. In Section 5, we show that this choice of weights corresponds to the mean within group error rate under the SVM setting. This indicates that this choice of weights can eliminate the unbalanced data effect because it is equivalent to finding a classifier that minimizes the mean within group error rate.

Despite its nice theoretical properties, this choice of weights may have problems for practice. First of all, π_j's are typically not available. Consequently, we need to estimate them. A natural estimator is $\hat{\pi}_j = n_j/n$, where n_j represents the number of observations for class j in the training dataset. Using $\hat{\pi}_j$ has some drawbacks. Clearly, the accuracy of the estimator $\hat{\pi}_j$ affects the performance. The smaller n_j and n are, the less reliable $\hat{\pi}_j$ is as an estimator of π_j. Moreover, this choice of weights only utilizes the class proportions. It may be beneficial to take into account the classification accuracy for different classes as well.

Recall that $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_k$ denote the within group errors on the training dataset with equal weights. Then we can use $(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_k)$ as well as the proportion information $(\pi_1, \pi_2, \ldots, \pi_k)$ to form weights $w_j = W(\hat{e}_j, \pi_j)$ through a function $W(\cdot, \cdot)$. When π_j's are not available, we can replace them by their estimators. Our goal is to put those classes with larger within group error rates (i.e., larger $\hat{e}_j$'s), as well as with smaller proportions, bigger weights so that it can be classified more accurately. As a result, $W(\hat{e}, \pi)$ should be an increasing function in $\hat{e}$ and a decreasing function in π. In this research, we recommend $W(\hat{e}, \pi) = (1/\pi)^{\hat{e}}$. Clearly, this function satisfies our requirement. Moreover, $(1/\pi)^{\hat{e}} \leq 1/\pi$, with the upper bound achieved when $\hat{e} = 1$. Once the new weights are obtained, we can scale them so that the standardized weights add up to one. Besides this choice of $W(\hat{e}, \pi)$, we also examine the performance of other weight functions such as $W(\hat{e}, \pi) = \hat{e}/\pi$. More discussions can be found in Section 6.

One potential drawback of one-step weights is that the weights may not be optimal for unbalanced learning. A natural solution is to update the weights adaptively. In Section 3.2, we introduce the procedure of adaptive weighted learning.

3.2 *Adaptive Weighted Learning*

To choose an optimal weight to deal with unbalanced classification problems, we propose an adaptive weighting procedure, which adaptively updates the weights by utilizing the classification information, including the within group error rates $(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_k)$ and the estimated class proportions $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)$.

Before we discuss the adaptive procedure, we first show an interesting property of the choice of weights. Define the equivalent weight set on R^k to a given weight vector as follows:

DEFINITION 1: *The weight vector* $\{w_j,\ j = 1, \ldots, k\}$ *belongs to the equivalent weight set* $\mathcal{W}_\pi$ *of* $(1/\pi_1, 1/\pi_2, \ldots, 1/\pi_k)$, *if* $\phi(\mathbf{x}) = arg\max_l w_l p_l(\mathbf{x})$ *coincides with* $\psi(\mathbf{x}) = arg\max_l p_l(\mathbf{x})/\pi_l$.

If there are multiple elements in $\mathcal{W}_\pi$, one can find a different weight vector in $\mathcal{W}_\pi$ to get the same Bayes rule as that of $\{1/\pi_j; j = 1, \ldots, k\}$ and it may have different finite sample behaviors. Thus, an adaptive procedure may find better weights than $(1/\hat{\pi}_1, 1/\hat{\pi}_2, \ldots, 1/\hat{\pi}_k)$. The following example further illustrates the equivalent weight set of $\{1/\pi_j\ ; j = 1, \ldots, k\}$.

EXAMPLE*:* Let $d = 1$, $x \in [0,\ 1]$ and $k = 3$, with $\pi_1 = 0.45$, $\pi_2 = 0.45$, and $\pi_3 = 0.1$. Then class 3 is the minority. The class-conditional probability density functions $g_j(x)$ and the conditional class probability functions $p_j(x)$ are given as follows

x	$[0, \frac{1}{3})$	$[\frac{1}{3}, \frac{2}{3})$	$[\frac{2}{3}, 1]$
$g_1(x)$	$\mathbf{\frac{3}{2}}$	1	$\frac{1}{2}$
$g_2(x)$	$\frac{1}{2}$	$\mathbf{\frac{3}{2}}$	1
$g_3(x)$	1	$\frac{1}{2}$	$\mathbf{\frac{3}{2}}$

x	$[0, \frac{1}{3})$	$[\frac{1}{3}, \frac{2}{3})$	$[\frac{2}{3}, 1]$
$p_1(x)$	$\mathbf{\frac{27}{40}}$	$\frac{18}{47}$	$\frac{9}{33}$
$p_2(x)$	$\frac{9}{40}$	$\mathbf{\frac{27}{47}}$	$\mathbf{\frac{18}{33}}$
$p_3(x)$	$\frac{4}{40}$	$\frac{2}{47}$	$\frac{6}{33}$

By some simple calculation, the Bayes decision rule under the classic criteria becomes

$$\phi_B(x) = \begin{cases} 1 & \text{when } 0 \le x < \dfrac{1}{3} \\ 2 & \text{when } \dfrac{1}{3} \le x < \dfrac{2}{3} \\ 2 & \text{when } \dfrac{2}{3} \le x \le 1. \end{cases}$$

Therefore, class 3 is ignored with an empty classification set.

Using the mean within-group error criterion, we have

$$\frac{p_1}{\pi_1} : \frac{p_2}{\pi_2} : \frac{p_3}{\pi_3}(x) = \begin{cases} \mathbf{3} : 1 : 2 & \text{when } 0 \le x < \dfrac{1}{3} \\ 2 : \mathbf{3} : 1 & \text{when } \dfrac{1}{3} \le x < \dfrac{2}{3} \\ 2 : 1 : \mathbf{3} & \text{when } \dfrac{2}{3} \le x \le 1. \end{cases}$$

Thus we classify x into class 3 when $x \in [2/3,\ 1]$. Notice that $1/\pi_1 : 1/\pi_2 : 1/\pi_3 = 1 : 1 : 4.5$. Consequently class 3 has a bigger weight due to its small proportion.

Consider a smaller weight for class 3. Let $w_1 : w_2 : w_3 = 1 : 1 : 4$, then,

$$p_1 w_1 : p_2 w_2 : p_3 w_3(x) = \begin{cases} \mathbf{27} : 9 : 16 & \text{when } 0 \le x < \dfrac{1}{3} \\ 18 : \mathbf{27} : 8 & \text{when } \dfrac{1}{3} \le x < \dfrac{2}{3} \\ 9 : 18 : \mathbf{24} & \text{when } \dfrac{2}{3} \le x \le 1. \end{cases}$$

Notice that $\phi(x) = \text{argmax}_l w_l p_l(x)$ coincides with $\psi(x) = arg\max_l p_l(x)/\pi_l$. Thus the weight vector $\{(w_1,\ w_2,\ w_3)\,|\,w_1 : w_2 : w_3 = 1 : 1 : 4\}$ and the weight vector $\{(w_1,\ w_2,\ w_3)\,|\,w_1 : w_2 : w_3 = 1 : 1 : 4.5\}$ are equivalent. One can show that the equivalent weight set of $(1/\pi_1 : 1/\pi_2 : 1/\pi_3)$ is $\{(w_1,\ w_2,\ w_3)\,|\,w_1 : w_2 : w_3 = 1 : 1 : \alpha\}$, where $3 \le \alpha \le 6.75$.

This example has a set of weight vectors equivalent to the weight $(1/\pi_1,\ 1/\pi_2, \ldots, 1/\pi_k)$ in the sense that these weights correspond to the same Bayes rule. This implies that instead of estimating π_j, it may be better to search optimal weights adaptively. In Section 4, we discuss an adaptive weighting algorithm, which applies a certain classification procedure iteratively on the data sample and updates the weights accordingly. The new weight $\mathbf{w}_{new}$ evolves from the weight $\mathbf{w}_{old}$ obtained from the previous iteration, and the updating rule depends on both the within group error rate $\hat{e}_j$ and the estimated proportion $\hat{\pi}_j$ of each class, i.e., $w_{new,j} = w_{old,j} W(\hat{e}_j, \hat{\pi}_j)$. The iteration stops when the within group error cannot be improved further, i.e., the within group error rates are sufficiently close to the rates in the previous step.

4. Algorithm

We propose the following algorithm to achieve adaptive learning:

Adaptive weighting algorithm:

Step 1: (Initialization) Fit a classifier using a certain classification procedure on the training dataset using the weight vector $\{w_j^{(1)},\ j = 1, \ldots, k\} = \{1,\ 1, \ldots, 1\}$ with the corresponding within group error rates for the training dataset $\{e_j^{(1)},\ j = 1, \ldots, k\}$.

For iteration $s = 2 \ldots$, with given $\{e_j^{(s-1)},\ j = 1, \ldots, k\}$ and $\{w_j^{(s-1)},\ j = 1, \ldots, k\}$ from iteration $s - 1$,

Step 2: (Weight updating) Set $w_j^{(s)} = w_j^{(s-1)} W(e_j^{(s-1)}, \hat{\pi}_j)$; $j = 1, \ldots, k$, and standardize the weights so they sum up to one. Then fit a weighted classifier with corresponding within group error rates for the training dataset $\{e_j^{(s)},\ j = 1, \ldots, k\}$.

Step 3: (Stopping rule) Stop the iteration if $\sum_j (e_j^{(s)} - e_j^{(s-1)})^2 \le \varepsilon$ for some prespecified $\varepsilon > 0$ or s reaches the prespecified maximum number of iterations. Otherwise, let $s = s + 1$ and go to Step 2.

Here the updating rule $W(\hat{e}_j, \hat{\pi}_j)$ is critical. We propose to use $W(\hat{e}_j, \hat{\pi}_j) = (1/\hat{\pi}_j)^{\max(\hat{e}_j, \delta)}$, a modified version of that for the one-step fixed weight in Section 3.1, where $\delta \in (0,\ 1)$ is a filter parameter, which lower bounds the within group error. We use $\delta = 0.1$ in the numerical studies. The use of δ is to prevent potential significance decrease of the weight for class j when $\hat{e}_j$ becomes very small or equals to 0. Otherwise, a small weight on the minority class may lead to bad performance on that class again in the next step and result in a dead loop.

As a remark, we note that for the adaptive procedure, we modify the weights by a small amount depending on the within group errors $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_k\}$. To decide whether we should stop the iteration, we compare the current within group errors to those in the previous iteration. In the case that misclassification costs are unequal as mentioned in Section 2.4, we can replace $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_k\}$

by their generalization using the unequal cost information, $\{\sum_{j=1}^k C_{1j}m_{1j}/n_1, \sum_{j=1}^k C_{2j}m_{2j}/n_2, \ldots, \sum_{j=1}^k C_{kj}m_{kj}/n_k\}$, where m_{ij} denotes the number of misclassifications from class i to class j for $i \neq j$ and $m_{ii} = 0$ for $i = 1, \ldots, k$.

5. Weighted Multicategory Support Vector Machine

The SVM (Boser, Guyon, and Vapnik, 1992; Cortes and Vapnik, 1995) is a large margin based classification method. In a separable scenario with two classes, the SVM aims to find a hyperplane in the input space R^d, which perfectly separates these two classes for the training set, such that the separation distance between the two classes is maximized. In a nonseparable case where a separating hyperplane does not exist, the SVM finds a hyperplane with good separation as well as small violation of the perfect separation constraints. See Burges (1998) for a good tutorial on the SVM.

It is now known that the SVM can be fit into a regularization framework in the form of *Loss + Penalty*. The loss function used in the SVM is the hinge loss function (Wahba, 1998). One can replace the hinge loss function or the penalty term to get different classifiers. For example, ψ-learning (Shen et al., 2003) and robust truncated-hinge-loss SVM (Wu and Liu, 2007) use bounded nonconvex loss functions to achieve robustness of the resulting classifiers. Multicategory extensions of the SVM and ψ-learning can be found in Lee, Lin, and Wahba (2004) and Liu and Shen (2006), respectively. Zhu et al. (2004) and Zhang (2006) proposed the L_1 SVM and the COSSO SVM, respectively, to achieve simultaneous variable selection and classification.

Under the regularization framework, the SVM aims to find a function $f(\mathbf{x}) = h(\mathbf{x}) + b$, where $h \in H_K$, a reproducing kernel Hilbert space (RKHS), and b is an intercept, so that it minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(\mathbf{x}_i)]_+ + \lambda\|h\|^2_{H_K}, \tag{2}$$

where $[\cdot]_+ = \max(\cdot, 0)$, and $\|h\|^2_{H_K}$ denotes the norm of the function h defined in the RKHS with the reproducing kernel function $\mathbf{K}(\cdot,\cdot)$. Here $y_i \in \{1, -1\}$ represents the binary class label and the first component in (2) measures the goodness of fit on the training data using the hinge loss function.

Denote $\mathbf{v}_i = (-1/(k-1), -1/(k-1), \ldots, 1, \ldots, -1/(k-1))$, where the jth entry is 1 when observation i belongs to class j. The multicategory SVM proposed by Lee et al. (2004) finds $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})) \in \Pi_1^k(1 + H_K)$, with the sum-to-zero constraint $\sum_{j=1}^k f_j(\mathbf{x}) = 0$, which minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{1\le l\le k, l\neq y_i}[f_l(\mathbf{x}_i) - v_{il}]_+ + \frac{1}{2}\lambda\sum_{j=1}^{k}\|h_j\|^2_{H_K}, \tag{3}$$

where v_{il} is the lth coordinate of $\mathbf{v_i}$. Once $\mathbf{f}(\mathbf{x})$ is obtained, the corresponding classification rule is $\phi(\mathbf{x}) = arg\ \max_j f_j(\mathbf{x})$. Lee et al. (2004) showed that the minimizer of this multicategory hinge loss, $E[\sum_{l=1}^k I(Y \neq l)[f_l(\mathbf{X}) - V_l]_+]$, under the sum-to-zero constraint is $f_j^*(\mathbf{x}) = 1$ if $j = arg\max_{l=1,\ldots,k} p_l(\mathbf{x})$ and $-1/(k-1)$ otherwise. Thus the corresponding decision rule is same as the Bayes rule under the classic classification criterion.

Naturally, one can modify the multicategory hinge loss function to derive the weighted SVM. In particular, given the weight $(w_1, \ldots, w_k)$, the weighted SVM solves

$$\begin{aligned}&\min_{\mathbf{f}} \frac{1}{n}\sum_{i=1}^{n}\sum_{1\le l\le k, l\neq y_i} w_{y_i}[f_l(\mathbf{x}_i) - v_{il}]_+ + \frac{1}{2}\lambda\sum_{j=1}^{k}\|h_j\|^2_{H_K}\\ &= \min_{\mathbf{f}} \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{k} w_{y_i}\bar{\delta}_{y_i l}[f_l(\mathbf{x}_i) - v_{il}]_+ + \frac{1}{2}\lambda\sum_{j=1}^{k}\|h_j\|^2_{H_K}.\end{aligned} \tag{4}$$

Using a direct generalization of Lemma 3 in Lee et al. (2004), we can show that the minimizer of $E[\sum_{l=1}^k w_Y\bar{\delta}_{Y,l}[f_l(\mathbf{X}) + 1/(k-1)]_+]$ is $f_j^*(\mathbf{x}) = 1$ if $j = arg\min_l \sum_{j=1}^k w_j\bar{\delta}_{jl}p_j(\mathbf{x})$, and $-1/(k-1)$ otherwise. Thus the corresponding decision rule $\phi(\mathbf{x}) = arg\max_j f_j^*(\mathbf{x})$ is the same as the Bayes rule of a weighted learning with weights $(w_1, \ldots, w_k)$. In particular, if we choose $(w_1, \ldots, w_k)$ to be $(1/\pi_1, 1/\pi_2, \ldots, 1/\pi_k)$, then the theoretical decision rule is the same as the Bayes rule under the mean within group error criterion.

Once the weight vector $(w_1, \ldots, w_k)$ is given, we can calculate the weighted SVM (4) as a quadratic programming problem. Consequently, we can implement the adaptive learning algorithm given in Section 4 using the weighted SVM.

6. Numerical Study

In this section, we illustrate the performance of learning with one-step fixed weights and adaptive weighted learning under classic, mean within group error, and mean square within group error criteria. We use the multicategory SVM procedure discussed in Section 5 for demonstration on both simulated and real datasets.

6.1 *Simulation*

Consider a simple three-class two-dimensional example. First we randomly generate y_i with three types of proportions: $\pi_1 : \pi_2 : \pi_3 = 1:2:2$, $1:4:4$, and $1:8:8$. Then for a given y we simulate $\mathbf{x}$ with $\mathbf{X} \sim N_2((1, 0)^T, 0.6I_2)$, $N_2((0, 1)^T, I_2)$, and $N_2((0, -1)^T, I_2)$ for $y = 1, 2, 3$, respectively. Clearly, class 1 is the minority class, whose expected proportions are 1/5, 1/9, and 1/17 for three different cases.

Figure 2 displays a typical training dataset with $\pi_1 : \pi_2 : \pi_3 = 1:8:8$. To demonstrate the effect of different weights, we increase the weight of class 1 and display how the decision boundary changes when the weight for class 1 increases from $1:1:1$ to $6:1:1$. From Figure 2, we can see that the classification set of class 1 gets bigger as we increase its weight. Consequently, the classification accuracy for class 1 becomes better as well. This demonstrates the effect of weighted learning for unbalanced datasets.

To further examine the classification accuracy, we simulate the training, tuning, and test sets of sizes 100, 100, and 10,000. We use the training set to build classifiers and use the tuning set to select tuning parameters. The test set is used to evaluate accuracy of the final classifier. We repeat the procedure 100 times to compare three different criteria (classic criterion, mean within group error criterion, and mean square within group error criterion) as the tuning criterion, and four different weighting procedures (equal weights,

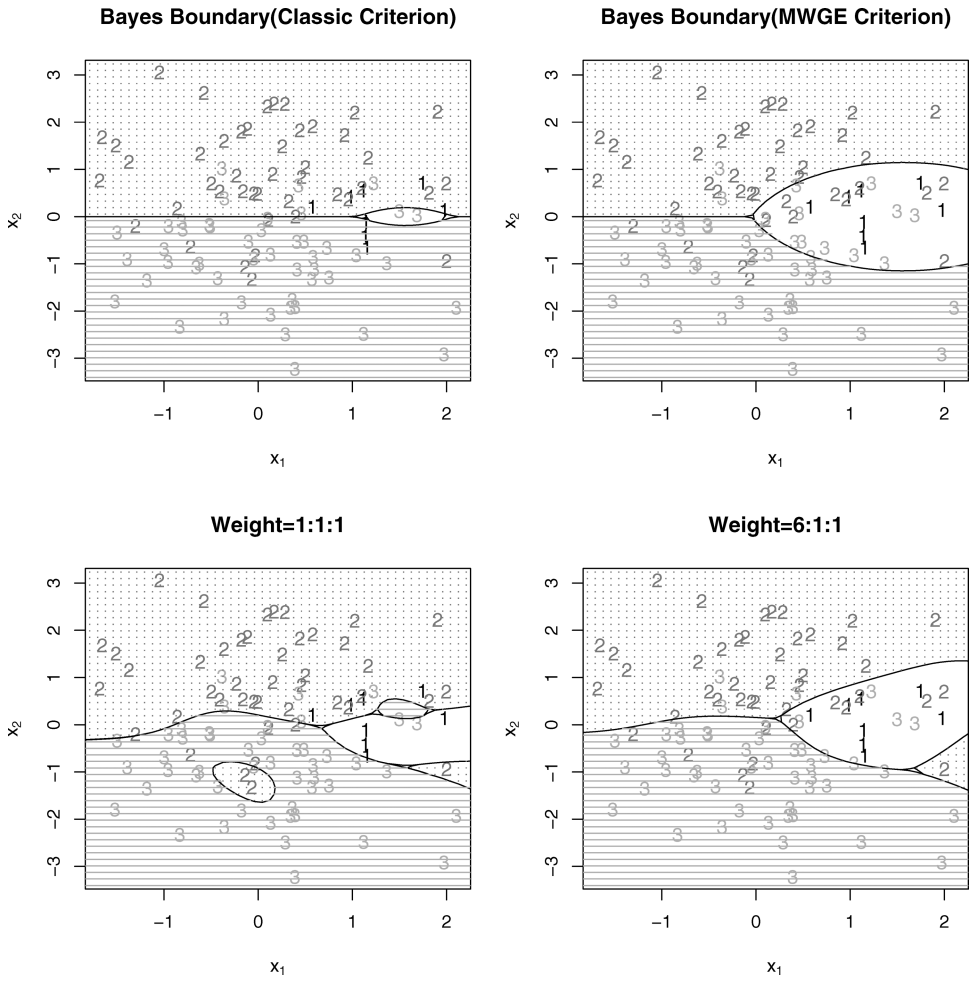

Figure 2. Plots of one typical training dataset with proportions of the three classes being 1 : 8 : 8. The top left and right panels display the Bayes boundaries under the classic and the mean within group error criteria, respectively. The bottom row displays the corresponding classification boundaries using weights 1 : 1 : 1 and 6 : 1 : 1. Note that the classification set for class 1 becomes bigger with the increased weight. This figure appears in color in the electronic version of this article.

one-step fixed weights using $\hat{\pi}_j$, one-step fixed weights using $W(e_j, \pi)$, and adaptive weighting scheme). The within group error rates as well as two types of Bayes errors using $W(\hat{e}_j, \hat{\pi}) = (1/\hat{\pi})^{\max(\hat{e}_j, \delta)}$, associated with the Bayes rules given in Proposition 1), are reported in Table 1 for the case of $\pi_1 : \pi_2 : \pi_3 = 1 : 4 : 4$. The final weights chosen for different methods are reported in Table 2. The cases of $\pi_1 : \pi_2 : \pi_3 = 1 : 2 : 2$ and $1 : 8 : 8$ are not reported to save space.

From the results, we have several observations: First of all, in each case, and under each classification criterion, the adaptive weighted learning method yields better overall classification accuracy than one-step weighted learning procedures. As the data become more and more unbalanced with proportions changing from $1 : 2 : 2$ to $1 : 8 : 8$, the adaptive weighted learning is more powerful compared to other methods. Secondly, the two types of one-step fixed learning procedures appear to perform similarly in terms of classification performance. Thirdly, the misclassification rates for the equal weights are close to the misclassification rate for the classic Bayes' classifier (Bayes' I). This matches our theoretical results in Section 2. Lastly, we observe that the mean within group error criterion and the mean square within group error criterion help to remove some of the unbalanced data effects compared to the classic classification criterion. Overall, we can conclude that our weighted learning procedures as well as the new criteria help to solve the unbalanced data effects.

To further investigate the effect of different weight-updating functions, we also study the two alternative choices: (a) $W(\hat{e}_j^{(s-1)}, \hat{\pi}_j) = (1/\hat{\pi}_j)^{I(s=2)} \max(\hat{e}_j, \delta)$; (b) $W(\hat{e}_j, \hat{\pi}_j) = \max(\hat{e}_j, \delta)/\hat{\pi}_j$. Compared to the suggested updating function $W(\hat{e}_j, \hat{\pi}_j) = (1/\hat{\pi}_j)^{\max(\hat{e}_j, \delta)}$ in Section 4, the updating functions in (a) and (b) are proportional to $\max(\hat{e}_j, \delta)$. In addition, the choice (a) is proportional to $1/\hat{\pi}_j$ for the second step with $s = 2$ whereas choice (b) is proportional to $1/\hat{\pi}_j$ for any $s \geq 2$. Thus, in contrast to $W(\hat{e}_j, \hat{\pi}_j) = (1/\hat{\pi}_j)^{\max(\hat{e}_j, \delta)}$, choices (a) and (b) update the weights more aggressively using $\hat{e}_j$ and $\hat{\pi}_j$. The results are provided in the Web Appendix. Our numerical experience suggests that a good choice of $W(\hat{e}_j, \hat{\pi}_j)$ such as $(1/\hat{\pi}_j)^{\max(\hat{e}_j, \delta)}$ should gradually increase the weights of classes with small $\hat{\pi}_j$'s and big $\hat{e}_j$'s. A weight function that increases

Table 1
Results of the weighted SVMs on the simulated data with proportions $1:4:4$, $W(\hat{e}_j, \hat{\pi}_j) = (1/\hat{\pi}_j)^{\max(\hat{e}_j, \delta)}$, *and* $n = 100$

	CLSC	MWGE	MSWGE
Equal weights	0.61 0.25 0.27	0.57 0.27 0.28	0.56 0.27 0.28
	(0.14 0.06 0.07)	(0.13 0.05 0.07)	(0.14 0.05 0.06)
OSW by $\hat{\pi}_j$	0.49 0.31 0.32	0.38 0.32 0.32	0.41 0.31 0.32
	(0.14 0.06 0.08)	(0.13 0.06 0.08)	(0.12 0.06 0.07)
OSW by $W(\hat{e}_j, \hat{\pi}_j)$	0.49 0.28 0.28	0.46 0.29 0.29	0.45 0.28 0.29
	(0.12 0.06 0.07)	(0.12 0.06 0.07)	(0.12 0.05 0.06)
Adaptive weights	0.43 0.31 0.32	0.39 0.31 0.30	0.40 0.31 0.31
	(0.12 0.06 0.07)	(0.14 0.05 0.07)	(0.12 0.05 0.07)
Bayes' I	0.59 0.19 0.19	0.59 0.19 0.19	0.59 0.19 0.19
	(0.01 0.01 0.01)	(0.01 0.01 0.01)	(0.01 0.01 0.01)
Bayes' II	0.18 0.29 0.29	0.18 0.29 0.29	0.18 0.29 0.29
	(0.01 0.01 0.01)	(0.01 0.01 0.01)	(0.01 0.01 0.01)

The first four rows show the within group error rates of three different classes under three classification criteria, using four different types of weights. The last two rows give the Bayes error rates corresponding to the Bayes rules given in Proposition 1. The numbers in the parentheses are the standard deviations for the corresponding within group errors. CLSC, classic criterion; MWGE, mean within group error criterion; MSWGE, mean square within group error criterion; and OSW, one-step fixed weights.

Table 2
Results of average final chosen weights for weighted SVMs using $W(\hat{e}_j, \hat{\pi}_j) = (1/\hat{\pi}_j)^{\max(\hat{e}_j, \delta)}$, *on the simulated data with proportions* $1:4:4$ *and* $n = 100$

	CLSC	MWGE	MSWGE
Equal weights	0.33 0.33 0.33	0.33 0.33 0.33	0.33 0.33 0.33
OSW by $\hat{\pi}_j$	0.67 0.15 0.18	0.67 0.15 0.18	0.67 0.15 0.18
OSW by $W(\hat{e}_j, \hat{\pi}_j)$	0.48 0.26 0.26	0.44 0.28 0.28	0.44 0.28 0.28
Adaptive weights	0.58 0.21 0.21	0.52 0.24 0.24	0.52 0.24 0.24

CLSC, classic criterion; MWGE, mean within group error criterion; MSWGE, mean square within group error criterion; and OSW, one-step fixed weights.

the weights for minority classes too aggressively may lead to suboptimal classification performance. Our suggested weight function in Sections 3 and 4 indeed gives the best performance in our simulation studies.

As a remark, we note that although the weighted learning helps to improve the classification accuracy of class 1, the corresponding test error of class 1 is still much larger than that of the Bayes rule under the within group error rate criterion (Bayes' II). This is because of the small sample size of class 1 with $n = 100$ and consequently the information of class 1 contained in the training sample is quite limited. As n increases, the performance improves accordingly as shown in the results with $n = 200$ and 300 (reported in the Web Appendix).

6.2 *Application to the Thyroid Data*

In this section, we apply different weighted multicategory SVMs on the thyroid database obtained from the UCI Machine Learning website `http://www.ics.uci.edu/~mlearn/`. The goal here is to determine whether a patient referred to the clinic is hypothyroid. Three classes are subnormal functioning (class 1), hyperfunction (class 2), and normal (not hypothyroid) (class 3). The database includes 7200 observations. This is an unbalanced dataset because around 92% of the patients are normal (class 3). Thus, classes 1 and 2 are minority classes. There are 21 variables, among which 15 are binary and 6 are continuous.

For illustration, we randomly select 200 observations for training and another 200 observations for tuning. Both datasets have the same class proportion, $5:10:185$, as the whole dataset. The remaining observations are used for testing.

The SVM results based on 50 replications are reported in Table 3. We can see that the test errors for classes 1 and 2 using the equal weights are substantially higher than class 3 due to the unbalanced proportion of the data. Our weighted learning procedures perform better in discriminating class 2 and at the same time, the accuracy of class 3 is sacrificed. The adaptive procedure works better than the fixed weight procedures because it gives much better classification accuracy for class 2 with a bigger weight (the final weights are reported in the Web Appendix). The error rate of class 3 has increased due to its low weight.

Interestingly, the weighted learning does not improve the accuracy of class 1 as much as that of class 2. To explore this issue further, we project the data on the first two principal component analysis (PCA) directions (Figure 3). From the projection plot 3, we can easily see that class 1 and class 3 are farther apart with class 2 in the middle. As a result, decreasing the weight for class 3 will help the accuracy of class 2 much more than that of class 1. A further examination on the training errors (in the Web Appendix) shows that the training error for class 1 is already very small without any weight adjustment. Moreover, different weights are used for learning on the training data. Consequently, increasing the weight for class 1 cannot further improve its accuracy much.

Table 3
Results of the weighted SVMs on the thyroid example with $W(\hat{e}_j, \hat{\pi}_j) = (1/\hat{\pi}_j)^{\max(\hat{e}_j, \delta)}$

	CLSC	MWGE	MSWGE
Equal weights	0.630 0.913 0.016	0.623 0.914 0.017	0.644 0.906 0.018
	(0.175 0.056 0.010)	(0.173 0.056 0.010)	(0.169 0.051 0.010)
OSW by $\hat{\pi_j}$	0.705 0.709 0.030	0.643 0.675 0.087	0.623 0.650 0.151
	(0.135 0.083 0.013)	(0.164 0.094 0.093)	(0.150 0.100 0.096)
OSW by $W(\hat{e}_j, \hat{\pi}_j)$	0.844 0.862 0.021	0.803 0.771 0.050	0.795 0.779 0.048
	(0.133 0.105 0.016)	(0.112 0.100 0.037)	(0.112 0.097 0.033)
Adaptive weights	0.618 0.477 0.397	0.662 0.556 0.291	0.667 0.573 0.273
	(0.153 0.187 0.204)	(0.157 0.165 0.175)	(0.156 0.156 0.161)

This table shows the within group error rates of three different classes based on the test dataset, under three classification criteria, using four different types of weights. CLSC, classic criterion; MWGE, mean within group error criterion; MSWGE, mean square within group error criterion; and OSW, one-step fixed weights.

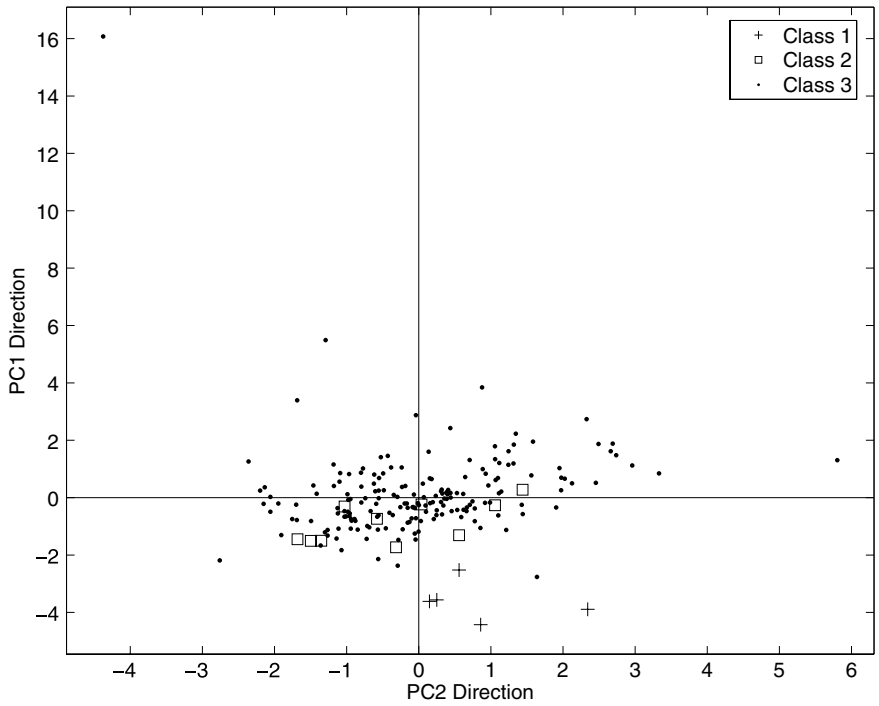

Figure 3. The PCA projection plot of one typical training data of the thyroid example. Note that class 1 and class 3 are farther apart compared to class 2.

7. Discussion

In this article, we have discussed classification of unbalanced datasets. We have shown that when the overall classification error rate is used as the classification criterion, the minority classes with relatively small class proportions tend to be ignored during the classification process. To overcome the problem, we have made two contributions: our first contribution is the proposal of two alternative classification criteria to increase the influence of minority classes so that they cannot be easily ignored; for the second contribution, we introduce several weighted learning procedures to get reasonable classifiers even if the datasets are unbalanced. Our numerical examples demonstrate the effectiveness of the new procedures.

The choice of the weight-updating function $W(\cdot, \cdot)$ is an important factor for the weighted learning procedures. Although the current recommendation works for numerical examples, it will be interesting to explore theoretical properties of various $W(\cdot, \cdot)$'s.

8. Supplementary Materials

The Web Appendix referenced in Section 6 is available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors are indebted to the editor, the associate editor, and two referees, whose helpful comments and suggestions led to a much improved presentation. YL is partially supported by grant DMS-0606577 from the National Science

Foundation and the UNC University Research Council Small Grant Program.

References

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifier. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* **5,** 144–152.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2,** 121–167.

Cortes, C. and Vapnik, V. (1995). Support vector network. *Journal of Machine Learning Research* **20,** 273–279.

Chawla, N. V., Japkowicz, N., and Kolcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter* **6**(1):1C6.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification*, 2nd edition. New York: Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Johnson R. A. and Wichern, D. W. (2001). *Applied Multivariate Statistical Analysis*, 5th edition. New Jersey: Prentice Hall.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* **99,** 67–81.

Leisenring, W., Alonzo, T., and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* **56,** 345–351.

Lin, Y., Lee, Y., and Wahba, G. (2002). Support vector machine for classification in nonstandard situation. *Machine Learning* **46,** 191–202.

Liu, Y. and Shen, X. (2006). Multicategory ψ-learning. *Journal of the American Statistical Association* **101,** 500–509.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research* **8,** 761–773.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003). On ψ-learning. *Journal of the American Statistical Association* **98,** 724–734.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods Support Vector Learning*, B. Schöelkopf, C. Burges, and A. Smola (eds), 69–87. Cambridge, Massachusetts: MIT Press.

Wu, Y. and Liu, Y. (2007). Robust truncated-hinge-loss support vector machines. *Journal of the American Statistical Association* **102,** 974–983.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports* **62,** 1432–1449.

Zhang, H. H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica* **16,** 659–674.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004). 1-norm support vector machines. *Neural Information Processing Systems* **16,** 49–56.

Received July 2007. *Revised January* 2008.
Accepted January 2008.