Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Multiple-class classification: Ordinal and categorical labels

Yuan-chin Ivan Chang

To cite this article: Yuan-chin Ivan Chang (2017) Multiple-class classification: Ordinal and categorical labels, Communications in Statistics - Simulation and Computation, 46:10, 7561-7581, DOI: 10.1080/03610918.2016.1242732

To link to this article: https://doi.org/10.1080/03610918.2016.1242732

Accepted author version posted online: 13 Oct 2016.
Published online: 04 May 2017.

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 46, NO. 10, 7561–7581
http://dx.doi.org/10.1080/03610918.2016.1242732

Taylor & Francis
Taylor & Francis Group

Multiple-class classification: Ordinal and categorical labels

Yuan-chin Ivan Chang

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

ABSTRACT

We study multiple-class classification problems. Both ordinal and categorical labeled cases are discussed. The common approaches for multiple-class classification are built on binary classifiers, in which one-versus-one and one-versus-rest are typical approaches. When the number of classes is large, then these binary-classifier-based methods may suffer from either computational costs or the highly imbalanced sample sizes in their training stage. In order to alleviate the computational burden and the imbalanced training data issue in multiple-class classification problems, we propose a method that has competitive performance and retains the ease of model interpretation, which is essential for a prognostic/predictive model.

ARTICLE HISTORY
Received 10 June 2016
Accepted 23 September 2016

KEYWORDS
classification; imbalanced data; multiple-class; ordinal response

MATHEMATICS SUBJECT CLASSIFICATION
Primary 62H30; Secondary 62J12

1. Introduction

In this study, we focus on multiple-class classification problems. In the machine learning literature, this type of problems is sometimes called "multiclass" or "multinomial" classification, and is different from multi-label classification problems, in which more than one label may be assigned to a subject (see Duan and Keerthi, 2005; Fu, 1968). In this study, we assume that each subject only belongs to one class, and our goal is to construct a classification function based on a given training dataset such that for a given new data point, it can correctly predict the labels of given new subjects.

Suppose that there are K classes. Let p_j, for $j = 1, \ldots, K$ with $\sum_{j=1}^{K} p_j = 1$, denote the proportion of Class j in the whole population, and $f_j = f_j(X)$ denote the distribution of Class j, and $X \in R^p$ be the p-dimensional vector of measures of a subject. It is known that when density f_j and the probability of each class p_j are known for each j, we can predict a subject with variable vector X using the Bayes rule as follows:

$$C(X) = \arg \max_{i=1,\ldots,K} \{p_1 f_1(X), \ldots, p_K f_K(X)\}, \tag{1}$$

where $C \in \{1, \ldots, K\}$. When f_j and p_j are unknown, we can estimate them using a given "training dataset." Then, the above classification function becomes

$$\hat{C}(x) = \arg \max_{i=1,\ldots,K} \left\{\hat{p}_1 \hat{f}_1(X), \ldots, \hat{p}_k \hat{f}_K(X)\right\} \tag{2}$$

CONTACT Yuan-chin Ivan Chang ycchang@stat.sinica.edu.tw Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

Equation (2) is used to approximate the Bayes rule (1), and in Van Ryzin (1966), it has been shown that under some conditions on the densities, as the sample size $n \to \infty$,

$$0 \le L_n - L^* \le \sum_{i=1}^{K} \int \left| p_i f_i(X) - \hat{p}_i \hat{f}_i(X) \right| dX \to 0 \text{ almost surely,} \tag{3}$$

where L^* and L_n denote the probability errors of Bayes rule (1) and the approximate Bayes rule (2), respectively. However, to estimate the density of a case with a high-dimensional vector of the explanatory variables and only limited observations is usually subject to the "curse of dimensionality," and therefore, the estimation cannot provide stable performance. This is especially the case in multiple-class classification problems.

Multiple-class classification rules built on binary classifiers are very common. Two popularly adopted approaches are the one-versus-one (OVO) and one-versus-rest (OVR), and the mixtures of these two approaches are occasionally proposed (Bishop, 2006; Hastie et al., 2001). When an OVO-type method is adopted for a K-class problem, then $K(K-1)/2$ binary classifiers must be built in order to construct a multiple-class classifier. Wu et al. (2004) is a typical example. Therefore, this type of approaches will usually suffer from the computationally intensive when the number of classes K is large. If an OVR-type method is applied, then there are only K binary classifiers to be constructed. However, the construction process of each binary classifier may fall into the imbalanced training data issue. It is reported in the literature that the imbalanced of sample sizes of two classes may largely increase the difficulty in the training stage (Duan and Keerthi, 2005; Sun et al., 2009). For example, even when the sample sizes of all classes in the training set are equal, the ratio of sample sizes for training a binary classifier in an OVR-type method is equal to $1/(K-1)$. It is clear that the difficulty is increased when K is large. Further discussions about imbalanced issues can be found in Sun et al. (2009). Hence, how to alleviate the computational burden and the issue of imbalanced training data in multiple-class classification problems is an important problem, which also motivates us to study this problem. Besides that, the model interpretation is also our concern. The development of modern classifiers mostly emphasizes their performances, instead of the ease of model interpretation. Of course, there is no measure for model interpretation. However, if a classification rule will be used not only for screening/diagnosis, but also prediction/prognosis, then to know which variables play effective roles for the classification rule becomes essential (Knottnerus, 2002; Moons and Grobbee, 2002). Here, we propose a framework, which can be used with most binary classifiers. Thus, it also allows practitioners to select the most suitable binary classification rules as building blocks for a particular application.

In the following sections, we will first introduce the method for cases with categorical labeled subjects; that is, with no relations among classes. Then, we will study the ordinal labeled subject cases, in which the ordinal relations among classes can be defined, such as the status of a disease' progress. The numerical studies based on the synthesized data and real-data examples are reported, and it is followed by a summarization and discussion.

2. Methodology

Besides the computational and imbalanced issues mentioned above, to apply these binary-classifier-based approaches, we need to have a good binary classification scheme as its building blocks. Likelihood-based binary classification methods are most appealing because they can be connected likelihood ratio tests, and therefore have some nice statistical properties such as achieving the maximization of the area under a receiver operating characteristic (ROC) curve (Eguchi and Copas, 2002). To take advantage of likelihood ratio methods, we modify

the method in Fu (1968) as follows. Let us start from an OVR procedure. For each i, let

$$R_i(X) = \log\left(\frac{p_i f_i(X)}{\left[\prod_{j=1}^{K} f_j^{\omega_j}(X)\right]}\right) \tag{4}$$

$$= \log(p_i) + \left\{\log f_i(X) - \sum_{j=1}^{K} \omega_j \log f_j(X)\right\} \tag{5}$$

$$= \log(p_i) + \left\{\sum_{j=1}^{K} \omega_j [\log f_i(X) - \log f_j(X)]\right\} \tag{6}$$

where $\omega_j > 0$ and $\sum_{j=1}^{K} \omega_j = 1$ are weights of classes. Then, we assign the subject with variable X to Class i, if $R_i = \max\{R_1, \ldots, R_K\}$. Note that the last term of (5) is a common term for all i. Thus, Eq. (5) implies that using $\max R_i$ to determine the label of a subject is equivalent to using the maximum value of $\log(p_i) + \log f_i(X)$ $(= \log(p_i f_i(X)))$, which is equivalent to the Bayes rule.

Equation (4) is a logarithm of the ratio of $p_i f_i(X)$ to a product of the "weighted" densities of all classes or a "weighted" version of a geometric mean of the densities of all classes. Thus, Eq. (4) is not a conventional likelihood ratio statistic, because the denominator is no longer a density. It is clear that Eq. (4) basically follows an OVR concept. If $\omega_i = 1/K$ for all i, then this denominator is the original geometric mean of all densities. Moreover, as shown in (5), the log of the denominator of Eq. (4) is a "weighted average" of log-densities of all classes, which can be viewed as the "center" of all classes in terms of their log-densities. In this sense, these weights represent the contribution or "importance," such as proportions or prevalences, of individual classes to this center, and Eq. (5) can be treated as the "distance" of the log-density of Class i to this weighted average. If these proportions/prevalence rates are known, then let $\omega_j = p_j$, $j = 1, \ldots, K$ will be a natural choice. However, when there is no such information, to treat each class equally by letting $\omega_j = 1/K$, $j = 1, \ldots, K$ is a reasonable choice.

Moreover, Eq. (6) says that when p_is are known, $R_i(X)$ is also equivalent to a weighted sum of log-likelihood ratios of Class i versus each other. That is, (4) can also be viewed as a "weighted ensemble" of likelihood ratios of OVO binary classifiers. Because we can start from an OVR formulation, there are only K ratios, R_i's, to be calculated. Eqs. (4)–(6) simply say that these R_i, $i = 1, \ldots, K$ can be calculated as a weighted sum of log-likelihood ratios. Hence, the imbalanced issue, caused by aggregating classes as in the conventional OVR approaches, is rare in our current approach. The remaining issue becomes whether we can have all the log-likelihood ratios without calculating all the pairs. To resolve this computational issue, we will borrow the idea of Begg and Gray (1984), and the categorical and ordinal labeled cases will be treated separately.

As mentioned above, when the class sizes or probabilities are different, we can incorporate this information in our classification rule via the weighting parameters ω_i's. A possible choice of weights is $\omega_i = p_i$'s; that is, the probabilities of individual classes. If these probabilities p_i's and data are collected via a simple random sampling scheme, then they can be estimated using the proportional sizes of all classes in a training set. This scenario is very common in which these weights could be related to the prevalence rates of subgroups. If there is no information about the class probabilities, and the sampling scheme does not allow us to estimate these proportions correctly, then we assume these weights are prefixed and equal, then set

$\omega_j = 1/K$, and the formula above can be re-written as:

$$\begin{aligned} R_i(X) &= -\log K + \log f_i(X) - \frac{1}{K}\sum_{j=1}^{K} \log f_j(X) \\ &= -\log K + \frac{1}{K}\sum_{j=1}^{K} [\log f_i(X) - \log f_j(X)] \\ &= -\log K + \frac{1}{K}\sum_{j=1}^{K} r_j^i. \end{aligned} \tag{7}$$

where

$$r_j^i = \log\left(\frac{f_i(X)}{f_j(X)}\right) = \log f_i(X) - \log f_j(X). \tag{8}$$

The following arguments can be extended to the case with unequal ω_i's with no difficulties. Thus, for simplicity we assume the class weights, ω_is, are equal throughout this article.

Eq. (7) suggests that an OVR (log)-likelihood ratio can be calculated using several OVO (log)-likelihood ratios. There will be no computational advantages if we need to calculate/model all possible combinations of i, j, i, $j = 1, \ldots, K$. However, it is easy to see from (8) that $r_j^i - r_h^i = r_j^h$. Hence, once we calculate r_j^1 for all $j = 2, \ldots, K$, then we can have all r_j^is. That is, there are $\log f_j$, $j = 1, \ldots, K$; that is, K "log densities" to be estimated. In other words, we only need to estimate K one-to-one pairs of log-likelihood ratios (the same number as that in OVR methods). Therefore, the proposed method will not often suffer from issues of imbalanced training data. A similar idea was used in Begg and Gray (1984) in their study of the calculation of polychotomous logistic regression parameters via individual regressions. Moreover, these calculations can be done separately. Hence, the parallel computing methods can be easily applied with no extra programming cost. Moreover, there are many ways to estimate these r_j^1, for $j > 1$. Of course, they can be obtained by direct estimates of densities, or log densities instead (Sugiyama and Kanamori, 2012). Then, we might still more or less suffer from the curse of dimensionality. Here, we adopt some modeling approaches to avoid the curse of dimensionality. There are also some commonly mentioned advantages to modeling-based approaches, such as their ease of interpretation. Moreover, we can also retain the ordinal relations among classes. Of course, there are some disadvantages, including that this type of approaches is likely to introduce model-specific errors and might still need to deal with the variable selection issues. In order to utilize the label information efficiently, we will discuss two general procedures for the cases of categorical and ordinal labeled subjects, separately.

The procedures stated below can be used with different estimating or modeling methods for r_j^is. In this article, the logistic-type binary classification functions are chosen as our building blocks, and in our numerical studies both logistic models and general additive logistic models (Hastie and Tibshirani, 1990) are used. There are always cons and pros for any statistical model. One of the reasons we chose a logistic-type classification function is that we were inspired by the results of Eguchi and Copas (2002), who proved that the logistic-type classification function has certain nice statistical properties, including maximizing the area under the ROC curve (AUC). In addition, the usefulness and importance of the ROC curve and AUC are well-known and intensively studied, which can be easily found in the literature and textbooks, such as Pepe (2004). For the categorical labeled data case, the procedure is similar to the findings in Begg and Gray (1984). The ordinal labeled procedure is new and

has never been reported in the literature as we know. Hence, we will focus on the ordinal one through a brief review of the categorical label case.

2.1. Categorical label case

Categorical Label Procedure: Let $P_i = Pr(Y = i|X)$, $i = 1, \ldots, K$. Then

$$\frac{P_i}{P_1} = \frac{P(Y=i)}{P(Y=1)} \cdot \frac{P(X|Y=i)}{P(X|Y=1)} = \frac{p_i}{p_1} \cdot \frac{f_i(X)}{f_1(X)} \text{ for } i = 2, \ldots, K. \tag{9}$$

Hence, $\log(P_i/P_1) = \log(p_i) - \log(p_1) + r_1^i(X)$. If we use Class 1 as the baseline, then we can estimate f_i, $i = 2, \ldots, K$ using logistic models. There are only $K - 1$ models to be constructed. As mentioned before, we can set $\omega_i = P(Y = i)$; that is, we can use the probabilities of classes as weights. If $P(Y = i)$ is unknown, then equal weights will be suggested. The class probabilities of a subject with measurement vector X are calculated as follows. Suppose that $p_i = \omega_i = 1/K$; then from (9) we have $P_i = \exp(r_1^i) \cdot P_1$, for $i = 2, \ldots, K$. Because $\sum_{i=1}^{K} P_i = 1$, it implies

$$P_1 = \frac{1}{1 + \sum_{i=2}^{K} \exp(r_1^i)}.$$

It then follows that for $j = 2, \ldots, K$, $i > 1$,

$$P_i = \frac{\exp(r_1^i)}{1 + \sum_{j=2}^{K} \exp(r_1^j)}.$$

Therefore, for a subject with measurement vector X, if $P_i = \max_j\{P_1, \ldots, P_K\}$, then we assign this subject to Class *i*. When there is a tie, a randomization scheme among the tied classes can be used to determine the final labels of subjects in order to break the tie.

Note that in the multiple-class problem, a subject only belongs to one class. We can predict the class of a subject based on the r_1^i values directly. However, "rescaling" r_1^i-values to a probability scale provides us useful information, which can be useful in a multiple-label problem. For example, we might assign multiple labels to a subject, depending on applications, based on the top-ranked class probabilities P_i's, or a threshold of probabilities can be used for this purpose.

2.2. Ordinal label case

The ordinal labeled response datasets are very common in many applications. When the label of a dataset is ordinal, we would like to incorporate this information in the classification rule in order to have better performance, since in ordinal label data problems, the "order" of classes can be viewed as a "measure of distance" between classes. That is, the subjects in classes with adjacent labels are usually more similar to each other than to subjects with unconnected labels. Therefore, when we consider the performance in this kind of multiple-class classification problems, the "distance" between the predicted label and the true label is important. For example, in a medical application subjects will usually be diagnosed and "labeled" according to their disease status. The misclassification between normal and suspect cases is very different from the misclassification between normal and malignant cases. This is unlike the situation in the binary classification or the categorical label classification cases, in which the accuracy is a

major concern. The proposed algorithm below allows us to take the ordinal label information among classes into consideration.

Ordinal label procedure: Given a measurement vector X, let $P_i = Pr(Y = i|X)$, $i = 1, \ldots,$ and K, be the class probability as defined before. Let $G_1 = 0$, and for $h = 1, \ldots, K$, define

$$G_h = \log\left(\frac{P_h}{\left(\prod_{j=1}^{h} P_j\right)^{\frac{1}{h}}}\right) = \log(P_h) - \frac{1}{h}\sum_{j=1}^{h}\log(P_j). \tag{10}$$

Note that the denominator in G_h is now the hth root of the product of P_j, $j = 1, \ldots, h, h \leq K$. Hence, for the Class h, we only calculate the log-density ratio of the Class h to the product of densities of Classes 1 to h. Thus, the ordinal relation among K classes is retained in this procedure. It follows that for $h \geq 2$,

$$\begin{aligned} h \cdot G_h &= h \cdot \log(P_h) - \sum_{j=1}^{h} \log P_j \\ &= (h-1)[\log P_h - \log P_{h-1}] + (h-2)[\log P_{h-1} - \log P_{h-2}] \\ &\quad + \cdots + 2[\log P_3 - \log P_2] + [\log P_2 - \log P_1] \\ &= \sum_{j=1}^{h-1} j \cdot [\log P_{j+1} - \log P_j] \end{aligned} \tag{11}$$

Note that

$$\begin{aligned} \log P_h - \log P_{h-1} &= [\log p_h - \log p_{h-1}] + [\log f_h - \log f_{h-1}] \\ &= [\log p_h - \log p_{h-1}] + r_{h-1}^h \end{aligned} \tag{12}$$

As discussed before, the prediction probability can be adjusted if p_i, $i = 1, \ldots, K$, are known or can be estimated. If information is not available, then we can assume they are equal. Therefore, as in the categorical label case, let Class 1 be the baseline, and we can calculate r_1^h, $h = 1, \ldots, K$. Because $r_a^b = r_1^b - r_1^a$, where a and b are integers satisfying that $1 < a, b \leq K$. Hence, it follows from (11) and (12) that we are able to calculate other log ratios. This implies that we then have the estimates of $\log P_h - \log P_{h-1}$ and G_h, for all $h = 2, \ldots, K$. Similarly, this procedure allows us to estimate $r_1^h, h = 2, \ldots, K$, with a suitable modeling strategy, which can depend on the practical needs. However, the probability P_i's should be calculated with an iterated procedure. In the following arguments, we will show that how P_i, $i = 1, \ldots, K$ can be calculated based on the estimates of log ratios; that is, how to calculate P_is via G_hs.

Following Eq. (11), we have that

$$\log P_2 - \log P_1 = 2G_2 - G_1 \tag{13}$$

$$\log P_3 - \log P_2 = \frac{3}{2}G_3 - G_2 \tag{14}$$

$$\cdots$$

$$\log P_K - \log P_{K-1} = \frac{K}{K-1}G_K - G_{K-1}. \tag{15}$$

Let $d_1 \equiv G_1 = 0$ and let $d_h \equiv [h/(h-1)]G_h - G_{h-1}$, for $h = 2, \ldots, K$. It implies that for $h = 2, \ldots, K$,

$$P_h = \exp\left(\sum_{j=2}^{h} d_j\right) \cdot P_1 = e^{\sum_{j=2}^{h} d_j} P_1. \tag{16}$$

Because $\sum_{h=1}^{K} P_h = 1$, it implies that

$$1 = \sum_{h=1}^{K} P_h = \sum_{h=1}^{K} e^{\sum_{j=1}^{h} d_j} P_1.$$

Hence,

$$P_1 = \frac{1}{[1 + \sum_{h=2}^{K} \exp(\sum_{j=2}^{h} d_j)]},$$

and for each $h = 2, \ldots, K$,

$$P_h = \frac{\exp(\sum_{j=2}^{h} d_j)}{1 + \sum_{h=2}^{K} \exp(\sum_{j=2}^{h} d_j)}.$$

Although there are many methods can be used to approximate the log-density ratio (8), the dimensionality of explanatory variables is still an obstacle for applying such kind of methods. If a parametric or a general additive form is taken, then the statistical properties, such as the consistency and asymptotic normality, can be easily proved as the sample sizes of all classes go to infinity. Moreover, the method of log-density estimation (Cule et al., 2010) may also be useful here if the sample sizes of individual classes are large enough and the number of dimensions is not too large. Note that the proposed approach requires only $K - 1$ binary classifiers, as that in the categorical label procedure. Moreover, the binary classifiers used here will not suffer from any extra imbalanced issue from aggregating the classes as the conventional OVR-type algorithms would.

3. Numerical study

Both synthesized data and real examples are used in our numerical demonstration. Here, the logistic models and the general additive logistic models (GAM) of Hastie and Tibshirani (1990) are used. The results obtained for these two methods are conducted using the R software. (Note that in Hastie and Tibshirani (1990), they studied general additive models, and the general additive logistic model is a special case in it. Here, for simplicity, we use GAM to denote the general additive logistic model.) The linear logistic model is easy to interpret and commonly used in many areas, so it is used as a baseline method. Although the highly nonlinear models usually have good classification performance, we know that for many applications, the model interpretation is essential. Moreover, as stated in Hastie and Tibshirani (1990), the advantages of GAM include that "the additive model retains some of the interpretability of the linear model by assuming additivity of effects," and "lets the data show us the appropriate functional form." In other words, the GAM sits in the middle of these two methods; it has some nonlinear advantages while still retaining some ease of interpretation. That is also one of the reasons why we use GAM in our numerical demonstrations.

For the numerical studies with the synthesized data, we will compare their results to the results of the Bayes rule, since we know the distributions used for data generation. We use

R-package *GAMBoost* for the general additive logistic regression model, which is denoted as GAM in this article, and use package *kernlab* for SVM. The idea of GAM was raised in Hastie and Tibshirani (1986), and a more comprehensive discussion can be found in their book (Hastie and Tibshirani, 1990). The book by Wood (2006) is a good source for an introduction to GAM, and it contains some useful programming information of it using the R language. The support vector machine is now a very popular classification tool in classification and machine learning. In order to save the computational time and to avoid the complicated and time-consuming parameter tuning procedures, we use the default parameter settings in their corresponding packages. When the GAMBoost is used for fitting a general additive model, we will use the default B-spline with degree 2, number of boosting steps equal to 500, penalty equal to 100 and the other default options as described in that package. Note that there are other packages for fitting GAM with different kinds of options. Here, we use the GAMBoost package in R, which is based on the work of Tutz and Binder (2006). There is a cross-validation option in this package that can be used for selecting variables. Because this cross-validation procedure takes too much computational time, we did not use it in our illustration. For the detailed information can be found from their original papers and corresponding R documents. For the studies with real examples, the Bayes rule is not available because there is a lack of the information about the true distributions of data. Hence, we will used the results of the support vector machines (SVM) with Gaussian kernels (Cristianini and Shawe-Taylor, 2000; Vapnik, 1995) as a high standard because their remarkable performances are often reported in the literature. Note that here we just apply the SVM method with Gaussian kernels directly and only use the results of SVM as a high standard reference. We do not use the SVM for modeling r_j^i in (8). The original idea of SVM can be dated back to Vapnik (1995). Nowadays, there are a lot of extensions and modifications of it proposed in the literature. Cristianini and Shawe-Taylor (2000) provided readers a good introduction. Our purpose here is to demonstrate the usefulness of the proposed methods. Please also note that the original SVM was designed for binary classification problems; we simply apply it to our dataset with the default setting in *kernlab*. There are several methods proposed to extend it to the multiple-class classification problems. Some empirical studies can be found in Duan and Keerthi (2005). In this package, a modification of the OVO-typed method, called the pairwise coupling method (Wu et al., 2004), is used for the multiple-class classification. The method of OVO, such as the pairwise coupling method in Wu et al. (2004), constructs a rule for discriminating between every pair of classes and then selecting the class with the most winning two-class decisions. Then, one can either use a voting scheme or calculate a class probability for each of the $k(k-1)/2$ models created in the pairwise classification. For the details of its options, please refer to its R document and related papers (Hsu and Lin, 2002; Wu et al., 2004). Further improvements for both methods may be obtained with some data-dependent parameter tuning. Both synthesized data and real-data examples are used for illustration. We did not synthesize the ordinal labeled data here; however, there are ordinal labeled datasets in the real example study and we will compare the results of methods with and without using the ordinal information.

3.1. Synthesized data

In simulation studies, we consider two three-class classification problems: one three-dimensional case and one five-dimensional case. The purpose of these studies is to compare the results of procedures using linear logistic models and GAM to that of the Bayes rule.

3.1.1. Simulation study for three-dimensional case

In the three-dimensional case, we generated 500 samples for each class for each run. Classes 1–3 are generated from multivariate normal distributions with different mean vectors (1, 0, 0), (0, 1, 0), and (0, 0, 1), respectively, with the same variance–covariance matrix

$$\text{Cov} = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}.$$

We then randomly select 90% of the samples from each class to form a training dataset and use the remaining 10% of each class as testing samples. The numerical results are summarized below based on 1000 replications of this simulation procedure. Since the distributions used for generating data are known, we can compute the empirical results of the Bayes rule. The results of the linear logistic models and the general additive logistic models are conducted using R packages as described before.

Figure 1 shows the box-plots of the empirical accuracies, based on 1000 replications, of three methods—the logistic-model-based procedure, general additive logistic regression (GAM)-based procedure, and true density (Bayes rule). The mean and standard deviation for each individual method are also shown in this figure. No significant differences are found among these three methods in terms of their accuracies in both training and testing results. Because the data are generated from the multivariate normal distributions, both the logistic-model-based and GAM-based procedures are as good as the Bayes rule in this situation. Tables 1–4 are confusion tables for the logistic general additive logistic model and Bayes rule (true densities), respectively. Each row of the tables shows the proportions of predicted labels for a given class. For example, row 1 in Table 1 shows that 74.17% of the subjects in Class 1 have been correctly assigned to Class 1, and 12.94% and 12.88% of them are misclassified as

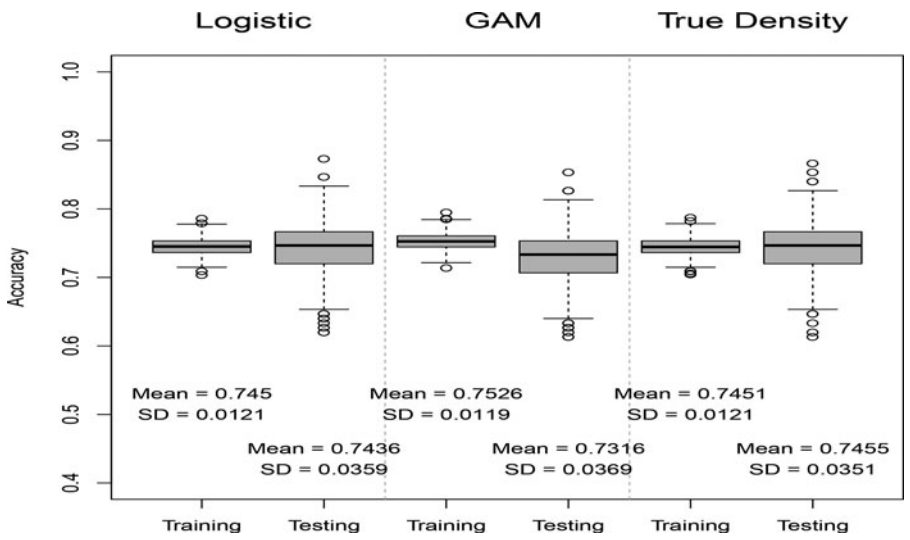

Figure 1. The box-plots of the accuracies of three methods. Both training and testing accuracies are plotted. Notation logistic denotes the logistic-model-based procedure, the GAM denotes the results of general additive logistic models, and true density denotes the results of the Bayes rule when the density function of each individual class is known. Notation SD denotes the corresponding standard deviation.

Table 1. Confusion table of the testing accuracy of the linear logistic model.

		Predict			
		1	2	3	Class
True	1	74.17%	12.94%	12.88%	33.33%
	2	13.05%	74.45%	12.50%	33.33%
	3	12.75%	12.79%	74.46%	33.33%
Pred.		33.32%	33.40%	33.28%	100.00%

Table 2. Confusion table of the testing accuracy of the general additive logistic regression model.

		Predict			
		1	2	3	Class
True	1	73.00%	13.67%	13.33%	33.33%
	2	13.44%	73.28%	13.28%	33.33%
	3	13.23%	13.58%	73.20%	33.33%
Pred.		33.22%	33.51%	33.27%	100.00%

Table 3. Confusion table of the testing accuracy using an OVO procedure.

		Predict			
		1	2	3	Class
True	1	74.9%	11.4%	13.7%	33.3%
	2	12.3%	76.5%	11.1%	33.3%
	3	14. 1%	13.0%	73.0%	33.3%
Pred.		33.8%	33.6%	32.6%	100.00%

Table 4. Confusion table of the testing accuracy of the Bayes rule, i.e., when the true distributions of three classes are known.

		Predict			
		1	2	3	Class
True	1	74.29%	12.86%	12.85%	33.33%
	2	12.96%	74.67%	12.37%	33.33%
	3	12.66%	12.65%	74.69%	33.33%
Pred.		33.30%	33.39%	33.31%	100.00%

Classes 2 and 3, respectively. The last row, named *Pred.*, in each table is a summarization of prediction proportions. The last column of each table, named Class, is the proportions of the true class labels. From these tables, we also found that both linear and general additive logistic models have similar performances to the Bayes rule.

3.1.2. Simulation study for five-dimensional case

For the study of the five-dimensional case, the data for Classes 1–3 are also generated from multivariate normal distributions with mean vectors equal to (1, 0, 0, 0, 0), (0, 1, 0, 0, 0), and

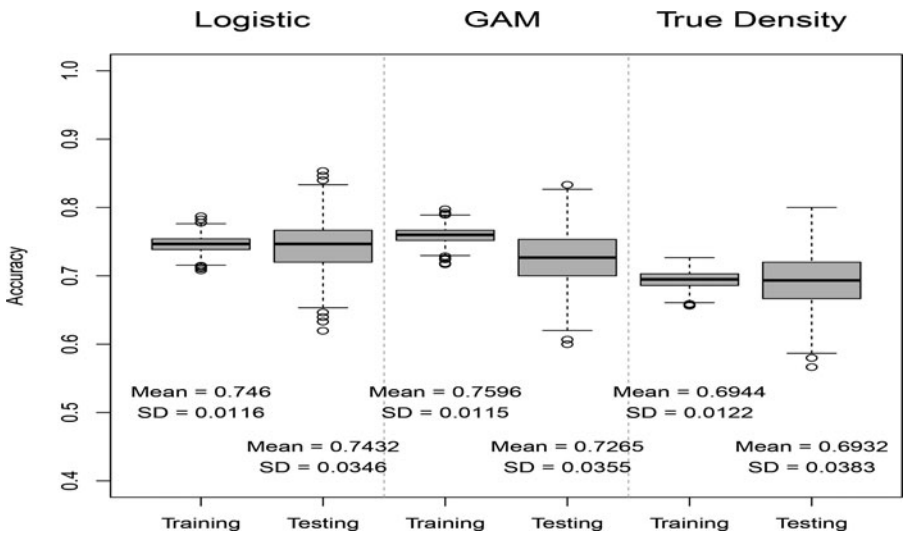

Figure 2. The box-plots of both training and testing accuracies of three methods in the five-dimensional data case. The GAM denotes the results of general additive logistic models and true density denotes the results of the Bayes rule, that is, when the density function of each individual class is known. Notation SD's denote their corresponding standard deviations.

$(0, 0, 1, 0, 0)$, respectively, and a common covariance matrix

$$\text{Cov} = \begin{bmatrix} 1 & .5 & .5 & 0 & 0 \\ .5 & 1 & .5 & 0 & 0 \\ .5 & .5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Again, we generated 500 samples for each class, and randomly selected 90% of the samples from each class for training, and used the remaining 10% for testing. The results summarized below are based on 1000 replications. Figure 2 shows box-plots of both training and testing accuracies of the three different methods. In this five-dimensional case, both the logistic regression model and GAM performed slightly better than the Bayes rule on average. This was because we compared our results with the empirical results of the Bayes rule. However, the differences in the testing accuracies of the three methods are not statistically significant. Tables 5–8 are confusion tables of the logistic-model-based procedure, GAM-based procedure and Bayes rule. The results of the logistic-model-based and GAM-based procedures are similar.

Table 5. Confusion table of the testing accuracy for logistic regression model in five-dimensional case.

		Predict			
		1	2	3	Class
True	1	74.29%	12.66%	13.05%	33.33%
	2	12.77%	74.30%	12.93%	33.33%
	3	12.68%	12.94%	74.37%	33.33%
Pred.		33.25%	33.30%	33.45%	100.00%

Table 6. Confusion table of the testing accuracy for GAM in five-dimensional case.

		Predict			
		1	2	3	Class
True	1	72.93%	13.64%	13.43%	33.33%
	2	13.52%	72.52%	13.95%	33.33%
	3	13.32%	14.18%	72.50%	33.33%
Pred.		33.26%	33.45%	33.29%	100.00%

Table 7. Confusion table of the testing accuracy using an OVO procedure in five-dimensional case.

		Predict			
		1	2	3	Class
True	1	70.3%	14.7%	15.0%	33.3%
	2	11.0%	74.2%	14.8%	33.3%
	3	14. 4%	12.7%	72.9%	33.3%
Pred.		31.9%	33.9%	34.2%	100.00%

Table 8. Confusion table of the testing accuracy Bayes rule in five-dimensional case.

		Predict			
		1	2	3	Class
True	1	65.24%	15.05%	19.71%	33.33%
	2	15.01%	65.34%	19.64%	33.33%
	3	11.23%	11.39%	77.38%	33.33%
Pred.		30.50%	30.59%	38.91%	100.00%

3.2. Real examples

We have applied our two algorithms to three real datasets: *Forest-type mapping*, *Vertebral Column,* and *Cardiotocography* datasets, which are available from the UCI Machine Learning Repository (Lichman, 2013). Among them, the Cardiotocography (CTG) dataset can be treated as either a three-class or a 10-class classification problem. Table 9 summarizes some basic information of the datasets used here. We will briefly introduce these datasets in each subsection. For each case, we randomly sample 90% of each individual class to form a training set and use the remaining 10% for testing, then repeat this process 1000 times. The numerical results for each case are summarized based on these 1000 replications.

3.2.1. Forest type mapping Dataset

The forest type mapping dataset was provided by Johnson et al. (2012), and according to them this dataset was collected from a remote sensing study that mapped different forest types based

Table 9. Summarization of the datasets used in Numerical Studies.

Data	Classes	Label Type	Variables	Size	Proportions
Forest	4	Categorical	27	533	(195:86:159:83)
Vertebral	3	Categorical	6	310	(100:60:150)
CTG*	3	Ordinal	21	2216	(1655:295:176)
CTG*	10	Ordinal	21	2216	(see Table 18)

Forest: Forest type mapping; Vertebral: Vertebral Column; CTG*: Cardiotocography. Variables = number of variables without labels; Size = total sample size; Proportions: sample sizes of individual classes. The proportions of CTG 10-class, please refer to Table 18.

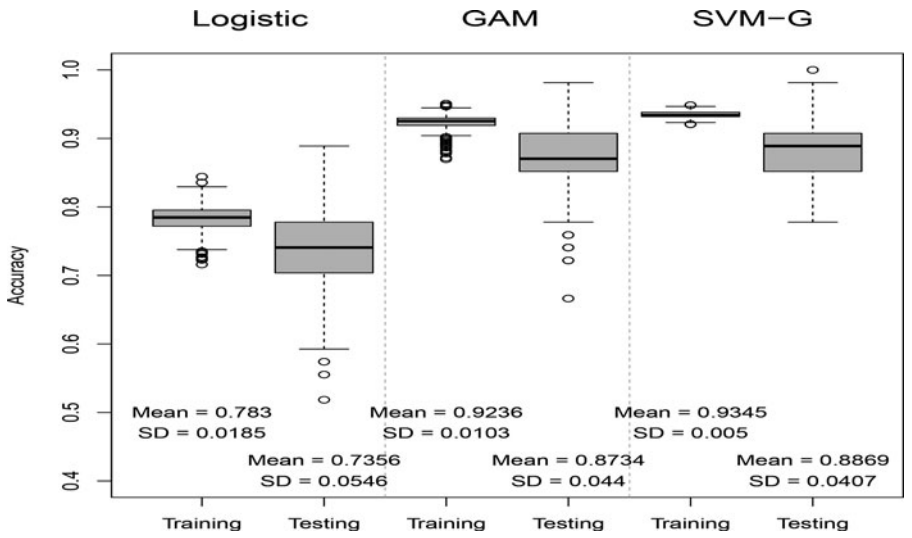

Figure 3. The box-plots of three methods when they are applied to the forest type mapping data. In this dataset, there are four classes in this dataset and there is not clear ordinal relations among labels. The results of the logistic and GAM are obtained by applying the algorithm for the categorical label case, and the results of SVM-G are obtained using gaussian kernels with default setting as that in *kernlab* package of R software. The corresponding mean and standard deviation (SD), based on 1000 runs, are reported.

on their spectral characteristics at visible-to-near-infrared wavelengths, using ASTER satellite imagery. The output (forest type map) or Class labels, 's' ('Sugi' forest), 'h' ('Hinoki' forest), 'd' ('Mixed deciduous' forest), 'o' ('Other' nonforest land), can be used to identify and/or quantify the ecosystem services (e.g., carbon storage, erosion protection) provided by the forest. The other attributes, variables b1–b9, are ASTER image bands containing spectral information in the green, red, and near-infrared wavelengths for three dates (Sept. 26, 2010; March 19, 2011; May 8, 2011.) For further detailed information, please refer to Johnson et al. (2012).

Figure 3 shows box-plots of three methods—the categorical label procedure with logistic models (logistic) and general additive logistic models (GAM)-based, and SVM with Gaussian kernels (SVM-G). Both the training and testing accuracies of each method are plotted, and their means and standard deviations are also shown in this figure. From Fig. 3, we found that the performance of the proposed categorical label procedure with general additive logistic models is statistically significantly better than the one using logistic models (at 0.05 level), and it is a similar to that of SVM-G. It is known that the results of SVM are usually hard to be interpreted, hence if the ability of model interpretation, as stated in Hastie and Tibshirani (1990), is considered here, then in this example, the advantage of GAM-based procedure is obvious. Tables 10 and 11 are the confusion tables of GAM and SVM-G. (The table of the procedure with logistic models is omitted.) In these tables, Classes 'Sugi', 'Hinoki' and 'Mixed deciduous'

Table 10. Confusion table of the testing accuracy for Forest-9-4C dataset based on GAM.

		Predict				
		1	2	3	4	Class
True	1	90.47%	0.60%	4.23%	4.71%	29.63%
	2	0.18%	71.09%	0.00%	28.73%	16.67%
	3	16.13%	0.00%	82.61%	1.26%	16.67%
	4	2.36%	2.80%	0.57%	94.28%	37.04%
Pred.		30.40%	13.06%	15.23%	41.31%	100.00%

Table 11. Confusion table of the testing accuracy for Forest-9-4C dataset based on SVM-G.

		Predict				
		1	2	3	4	Class
True	1	88.18%	1.19%	4.39%	6.24%	29.63%
	2	0.00%	84.63%	0.07%	15.37%	16.67%
	3	12.42%	0.00%	85.48%	2.10%	16.67%
	4	2.41%	4.65%	0.57%	92.37%	37.04%
Pred.		29.09%	16.18%	15.76%	38.97%	100.00%

forests and 'Other non-forest land' are coded as 1, 2, 3, and 4, respectively. Although Fig. 3 shows that the overall accuracies of GAM and SVM-G are similar, from these tables we can still see that the SVM has more balanced accuracies among the four classes. From these tables, we also found that for both of them, a large portion of subjects in Class 2 are misclassified as Class 4, and another portion of subjects in Class 3 are misclassified as Class 1.

3.2.2. Vertebral column data

This Vertebral Column (VC) dataset was provided by Dr. Henrique da Mota during a medical residence period in the Group of Applied Research in Orthopaedics (GARO) of the Centre MÃ©dico-Chirurgical de RÃ©adaptation des Massues, Lyon, France. According to the data description, this dataset can be used in both two-class and three-class classification problems. The original three classes are normal (100 patients), disk hernia (60 patients), or spondylolisthesis (150 patients). For the two-class problem, the classes "Disk Hernia" and "Spondylolisthesis" were merged into a single class labeled as "abnormal." We only consider it as a three-class classification problem in our study, and code three categories normal, disk hernia, and spondylolisthesis, as Classes 1, 2, and 3, respectively. Both categorical and ordinal procedures are applied to this dataset for comparison purposes.

Figure 4 shows the box-plots of both training and testing accuracies for all methods; while Logistic and GAM denote the corresponding procedures assuming the labels are categorical,

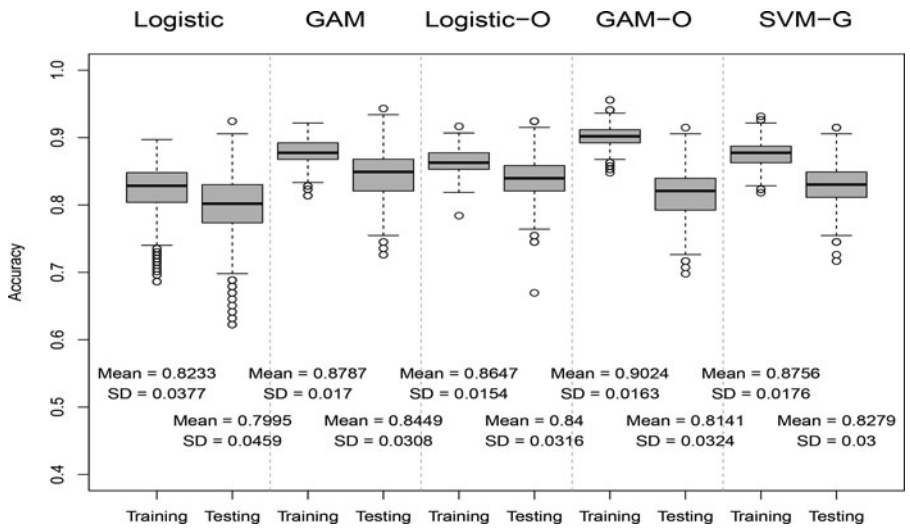

Figure 4. The box-plots of both training and testing accuracies for the Vertebral Column dataset using different methods: the procedure based on logistic models with categorical labels (logistic) and ordinal labels (logistic-O), the procedures based on the general added logistic models using categorical labels (GAM) and ordinal labels (GAM-O), and SVM with Gaussian kernels (SVM-G).

Logistic-O and GAM-O denote the procedures assuming the labels are ordinal. The results of SVM-G are using the *kernlab* package in R with Gaussian kernels and the default setups in it. Among all methods, the GAM has the highest testing accuracy in this figure, although the differences are not statistically significant. The accuracy improvement of Logistic-O on Logistic is minor; that is, there is not much advantage when the ordinal procedure is used. A similar situation is found between GAM and GAM-O. In addition, we found that even the SVM-G is not significantly better than the other methods. In fact, both GAM methods have slightly higher accuracies than SVM-G on average. This situation may be due to the fact that there is no clear ordinal relation among labels, and the boundaries between classes are rather smooth, such that the highly nonlinear models have no advantage in this example. In fact, if we compare the training and testing accuracies of each method, then we find that GAM-O has the highest training accuracy with a relatively lower testing accuracy than the other methods. This suggests that using nonlinear models for this dataset may tend to "overfit" the training data, and using the categorical label procedure with linear logistic models may suffice for this dataset.

Tables 12–14 show the confusion tables of GAM, GAM-O, and SVM-G. The tables of Logistic and Logistic-O are similar to those of GAM and GAM-O, so they are omitted. From these tables, we can see the distributions of the predicted labels of the three methods. The prediction label distribution of GAM-O seems closer to that of the true class labels. When the true categories are 1 and 2, the accuracies of all three methods cannot predict labels correctly.

Table 12. Testing accuracy for the vertebral column dataset based on GAM.

		Predict			
		1	2	3	Class
True	1	82.28%	14.04%	3.68%	32.08%
	2	37.47%	59.59%	2.94%	19.81%
	3	3.02%	0.77%	96.21%	48.11%
Pred.		48.05%	16.68%	35.27%	100.00%

Table 13. Testing accuracy for the vertebral column dataset based on GAM-O.

		Predict			
		1	2	3	Class
True	1	77.99%	17.95%	4.06%	32.08%
	2	40.38%	56.70%	2.91%	19.81%
	3	5.39%	0.75%	93.87%	48.11%
Pred.		35.61%	17.35%	47.04%	100.00%

Table 14. Testing accuracy for the vertebral column dataset based on SVM-G.

		Predict			
		1	2	3	Class
True	1	77.51%	15.04%	7.45%	32.08%
	2	31.24%	64.09%	4.68%	19.81%
	3	4.28%	1.71%	94.02%	48.11%
Pred.		48.55%	18.34%	33.11%	100.00%

Both GAM and GAM-O have less than 60% accuracy. All methods tend to misclassify subjects in Class 2 as Class 1. However, all three methods can identify Class 3 very well; in fact, all of them have more than 93% accuracy when subjects are from Class 3.

3.2.3. Cardiotocography data as a three-class problem

The Cardiotocography (CTG) data-set was originally used in Ayres-de Campos et al. (2000), and the following information about this dataset is quoted from Lichman (2013). This dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms classified by expert obstetricians. The dataset includes 2126 fetal cardiotocograms (CTGs) that were automatically processed and the respective diagnostic features that were measured. These CTGs were also classified by three expert obstetricians and a consensus classification label was assigned to each of them. Classification was both with respect to a morphologic pattern with FHR pattern class code (1–10) and the fetal state code (Normal, Suspect, or Pathologic; which are coded as 1, 2, and 3 in our confusion tables below). Hence, we used this dataset for both the 10-class and three-class studies and applied both categorical and ordinal label procedures to it. Based on the information above, the class labels, normal, suspect, and pathologic have a clearer ordinal relation among these three labels; however, the 10-class labels based on the morphologic pattern of CTGs might not have a clear ordinal relation among them.

Figure 5 shows the box-plots for both training and testing accuracies of five different methods, including both categorical and ordinal procedures with logistic models and GAMs, and SVM with Gaussian kernels. As before, the results of the SVM with Gaussian kernels is used as our benchmark for comparison. It is clearly shown in this figure that the GAM-O is the best among all of the methods in terms of both training and testing accuracy. It even outperforms the SVM-G in this case. This amelioration of the performance of GAM-O is due to both the nonlinear property of GAM and the ordinal label information used in this procedure. We know that the SVM method with Gaussian kernels is known as a highly nonlinear method

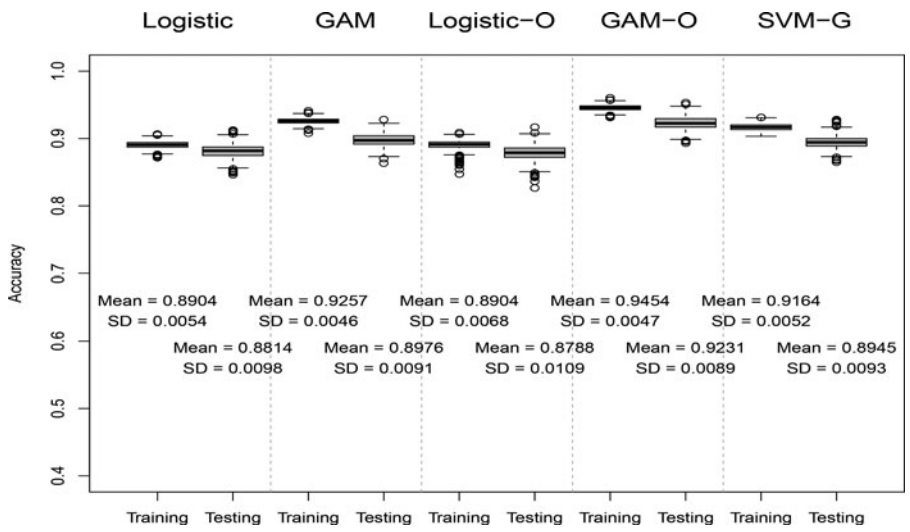

Figure 5. The box-plots of both training and testing accuracies of the five methods: the procedure based on logistic models (Logistc), the procedure based on GAM (GAM), the procedure based on logistic model with ordinal labels (Logisitc-O), the procedure using GAM with ordinal labels (GAM-O), and the SVM with Gaussian kernels (SVM-G).

Table 15. Testing accuracy for the cardiotocography data as a three-class problem using the categorical label procedure with general additive logistic models.

		Predict			
		1	2	3	Class
True	1	95.53%	3.94%	0.54%	77.75%
	2	32.86%	65.09%	2.05%	13.94%
	3	9.55%	13.28%	77.16%	8.31%
Pred.		79.64%	13.24%	7.12%	100.00%

Table 16. Testing accuracy for the cardiotocography data as a three-class case using on the ordinal label procedure with general additive logistic models.

		Predict			
		1	2	3	Class
True	1	96.09%	2.68%	1.23%	77.75%
	2	23.16%	75.26%	1.58%	13.94%
	3	9.81%	4.71%	85.49%	8.31%
Pred.		78.75%	12.97%	8.28%	100.00%

Table 17. Testing accuracy for the cardiotocography data as a three-class problem using the SVM with Gaussian kernels.

		Predict			
		1	2	3	
True	1	96.76%	2.95%	0.29%	77.75%
	2	40.05%	58.26%	1.69%	13.94%
	3	11.75%	14.81%	73.44%	8.31%
	4	81.79%	11.65%	6.56%	100.00%

and that its final classification model is usually very difficult to interpret. Under the additive assumption of GAM, we can have at least some information about the relation between the response and each variable. That is, besides the performance, using GAM-O here also provides us the advantage of its ability to interpret the model. We did not go the details about the model interpretation here; please refer to Hastie and Tibshirani (1986, 1990) for the further discussion. Tables 15–17 are confusion tables of GAM, GAM-O, and SVM-G. (The confusion tables for logistic-model-based procedures are omitted here.) We can see from these tables that all of them do well for Class 1, and the major gain in accuracy for GAM-O are from Classes 2 and 3, which are the suspect and pathologic categories, respectively. In addition, there are fewer subjects in Class 2 that are misclassified as Class 1 by the GAM-O. Since the same models are used in both GAM and GAM-O, the improvement of the GAM-O on GAM is basically from the use of the ordinal label information.

3.2.4. Cardiotocography data as a 10-class problem

We now consider the cardiotocography data as a 10-Class problem; that is, the classes, coded as 1–10, based on a morphologic pattern with the FHR patterns that were used as labels. These patterns have clear no ordinal relation, hence we only apply the logistic, GAM, and SVM with Gaussian kernels procedures to this 10-class problem. The box-plots of both training and testing accuracies of these three methods are shown in Fig. 6. In this example, we can clearly see that the advantage of nonlinear models—that is, the SVM—is clearly the best among the

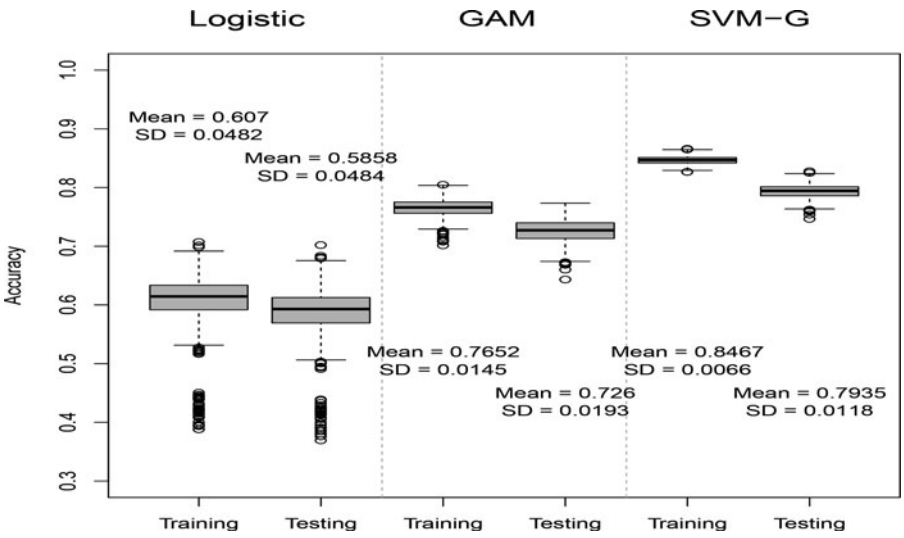

Figure 6. The box-plots of the logistic model, general additive logistic model, and SVM with Gaussian kernels using the cardiotocography dataset as a 10-class case with the categorical labels of subjects.

three methods, and the GAM is also significantly better than Logistic in terms of accuracy. From logistic, GAM to SVM-G, the differences between the training and testing accuracies of each method increase. This phenomenon suggests that the nonlinear properties of GAM and SVM-G may tend to overfit the datasets.

From the confusion tables of GAM and SVM, Tables 18 and 19, we can see that some classes are easily confounded. For example, from these two tables, we can see that there is a big portion of Class 3 subjects misclassified as Class 1. In addition, 38.54% of subjects in Class 5 are misclassified as Class 10 when the procedure with GAM is used, while the SVM-G method misclassified 46% of Class 5 subjects as Class 1, and another 20.8% of subjects as Class 10. (The confusion table for logistic is omitted.) Hence, it can be helpful to use different models are used for some particular pairs. However, this type of method may be tedious.

In this dataset, the proportions of the true labels (the Class column) range from a minimum of 2.52% to a maximum of 27.13%. The ratio of the largest class is more than 10 times the size of the smallest one. In the proposed categorical procedure, the baseline class is prefixed, and when the sizes of all classes are similar it may not be an issue. In this situation, the choice

Table 18. Testing accuracy of applying the proposed method for the ordinal label data to CTG as a 10-class classification with ordinal label responses and using general additive logistic models (GAM).

		Predict										
	Class	1	2	3	4	5	6	7	8	9	10	Class
True	1	73.11%	5.54%	1.14%	0.00%	4.07%	1.40%	0.61%	0.00%	0.25%	13.88%	17.90%
	2	3.57%	91.55%	0.21%	1.06%	0.57%	2.59%	0.22%	0.00%	0.00%	0.23%	27.13%
	3	66.41%	6.51%	24.44%	0.00%	0.69%	0.00%	0.16%	0.00%	0.12%	1.67%	2.52%
	4	0.00%	18.26%	0.00%	80.93%	0.00%	0.82%	0.00%	0.00%	0.00%	0.00%	3.92%
	5	11.76%	8.70%	0.54%	0.14%	49.94%	0.24%	0.07%	0.00%	0.06%	28.54%	3.50%
	6	1.43%	5.52%	0.02%	0.95%	0.00%	91.50%	0.50%	0.08%	0.00%	0.00%	15.52%
	7	5.36%	0.00%	0.04%	0.00%	1.58%	34.95%	55.51%	1.72%	0.01%	0.82%	11.89%
	8	1.22%	0.00%	0.04%	0.09%	0.00%	21.13%	29.48%	48.04%	0.00%	0.00%	5.03%
	9	31.33%	0.00%	0.73%	0.00%	2.22%	0.00%	0.05%	0.02%	6.59%	59.06%	3.36%
	10	33.05%	0.99%	0.02%	0.00%	3.61%	0.00%	0.00%	0.00%	0.56%	61.76%	9.23%
Pred.		21.16%	27.96%	0.93%	3.61%	3.24%	20.42%	8.34%	2.64%	0.32%	11.37%	100.00%

Table 19. Testing accuracy of applying the SVM with the default Gaussian kernel to CTG, as a 10-class classification problem.

		Predict										
	Class	1	2	3	4	5	6	7	8	9	10	Class
True	1	83.48%	6.36%	0.55%	0.00%	0.33%	0.74%	0.53%	0.00%	0.01%	8.01%	17.90%
	2	5.53%	90.89%	0.12%	0.15%	0.28%	2.10%	0.36%	0.00%	0.00%	0.57%	27.13%
	3	66.12%	8.34%	25.38%	0.00%	0.00%	0.01%	0.15%	0.00%	0.00%	0.00%	2.52%
	4	0.00%	32.67%	0.00%	66.53%	0.00%	0.80%	0.00%	0.00%	0.00%	0.00%	3.92%
	5	46.07%	7.40%	0.01%	0.00%	25.52%	0.00%	0.14%	0.00%	0.00%	20.84%	3.50%
	6	1.69%	7.34%	0.03%	0.09%	0.00%	85.50%	4.93%	0.42%	0.00%	0.00%	15.52%
	7	4.05%	0.30%	0.51%	0.00%	0.01%	6.38%	85.81%	2.55%	0.00%	0.40%	11.89%
	8	0.05%	0.00%	0.00%	0.00%	0.00%	1.38%	8.15%	90.41%	0.00%	0.00%	5.03%
	9	13.29%	0.32%	0.01%	0.00%	0.00%	1.20%	0.18%	0.14%	45.92%	38.95%	3.36%
	10	31.73%	1.05%	0.00%	0.00%	0.03%	0.01%	0.40%	0.00%	1.32%	65.47%	9.23%
Pred.		23.84%	28.83%	0.84%	2.66%	1.03%	14.88%	11.62%	4.93%	1.66%	9.72%	100.00%

of the baseline class may affect the overall performance when the sizes of classes are highly imbalanced.

A possible way to improve the performance of the categorical procedure further is the choice of the baseline class. In general, to use the class with highest prevalence in the population as the baseline class will provide us the most stable results as suggested in the literature (see, e.g., Begg and Gray, 1984; Albert and Anderson, 1984; Santner and Duffy, 1986). However, to do this we will need the prevalence information of each class. If the training data are collected via, for example, a stratified sampling method, then the proportions of the sizes of classes can be obtained with a pilot study such as in a two-stage method or a sequential sampling method. In this case, we can select the baseline based on the estimated prevalence of each class and the results of it will be reported elsewhere.

4. Summary

There are always many modern classification methods proposed, and they mostly focus on the classification performance. The model interpretation is not well discussed in these methods. For example, the kernel-based methods, such as SVM, can usually produce a highly nonlinear separation boundary for classification; nevertheless, the relation between this boundary and variables in term of "model interpretation" can be very complicated. However, if a classification rule is used as a predictive/prognostic model, and not just for screening/diagnosis, then the model interpretation becomes essential. In this article, we propose two procedures for multiple-classification problems for the categorical or ordinal labeled data, for which we can use binary classifiers as building blocks with flexible modeling options. The formulation under this framework allows us to restore the Bayes rule through its binary classifiers. We use both parametric and additive logistic regression models for demonstration purposes and compare their results to those of the Bayes rule, when the synthesized data are used, or to those of SVM with Gaussian kernels when real examples are used. It is clear that this framework allows us to take advantage of novel binary classification methods via integrating them as its building blocks. If the conventional logistic regression models are used, then the modern variable selection schemes, such as lasso can also be used. In this article, we want to emphasize that it is possible to retain the ease of model interpretation without diminishing the performance. That is why we choose the general additive logistic model for our illustration and compare its results with both that of linear logistic regression models and SVM. The numerical results show that the proposed methods are very promising and as competitive as

SVM when the additive models are used. For the categorical label case, the number of binary classifiers required is much less than the number needed for conventional OVO-type procedures. For the ordinal case using the proposed procedure, the proposed methods can avoid the issue of imbalanced training data issue due to the class aggregation, which is very common in the OVR-type procedures. In our numerical study, we show that for the ordinal label data case, the proposed ordinal label procedure can successfully take this information into account and actually improve the classification performance on that of the procedures ignoring this information. The model interpretation character is essential in many classification applications, and the proposed procedures provide such a modeling option to practitioners without diminishing the performance much; that is, the proposed methods can retain the model interpretation ability depending on the selected models, and the practical needs. The proposed method is launched from binary classifiers, so other advanced binary classifiers can be integrated into this framework; for example, one can consider using a variable selection in the parametric logistic regression models in order to deal with high-dimensional problems. In addition to the selection of the baseline class as mentioned before, this part will also be an important extension for our future research.

ORCID

Yuan-chin Ivan Chang http://orcid.org/0000-0002-4977-7721

References

Albert, A., Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71(1):1–10.

Ayres-de Campos, D., Bernardes, J., Garrido, A., Marques-de Sá, J., Pereira-Leite, L. (2000). Sisporto 2.0: A program for automated analysis of cardiotocograms. *J Matern Fetal Med.* 9(5):311–318.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Begg, C., Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* 71:11–18.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer-Verlag.

Cristianini, N., Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

Cule, M., Samworth, R., Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72:545–607.

Duan, K. B., Keerthi, S. S. (2005). *Which Is the Best Multiclass SVM Method? An Empirical Study*. Chapter Multiple Classifier Systems. Lecture Notes in Computer Science Vol. 3541. Berlin: Springer-Verlag, pp. 278–285.

Eguchi, S., Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika* 89(1):1–22.

Fu, K. (1968). *Sequential Methods in Pattern Recognition and Machine Learning*. New York: Academic Press.

Hastie, T., Tibshirani, R. (1986). Generalized additive models. *Statistical Science* 1(3):297–318.

Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. Vol. 43. Boca Raton, FL: CRC Press.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*. Berlin: Springer.

Hsu, C.-W., Lin, C.-J. (2002). A comparison on methods for multi-class support vector machines. *IEEE Transactions on Neural Networks* 13:415–425.

Johnson, B., Tateishi, R., Xie, Z. (2012). Using geographically-weighted variables for image classification. *Remote Sensing Letters* 3(6):491–499.

Knottnerus, J. A. (2002). Challenges in dia-Prognostic Research. *Journal of Epidemiology and Community Health* 56:340–341.

Moons, K., Grobbee, D. E. (2002). Diagnostic studies as multivariable, prediction research. *Journal of Epidemiology and Community Health* 56:337–338.
Pepe, M. S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. USA: Oxford University Press.
Santner, T. J., Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 73(3):755–758.
Sugiyama, M. S. T., Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge, New York: Cambridge University Press.
Sun, Y., Wong, A., Kamel, M. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence* 23(4):687–719.
Tutz, G., Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics* 62(4):961–971.
Van Ryzin, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhya, Series A* 28(2/3):261–270.
Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.
Wood, S. (2006). *Generalized Additive Models: An Introduction with R*, London: Chapman & Hall/CRC, Taylor and Francis Group.
Wu, T.-F., Lin, C.-J., Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5:975–1005.