BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 1. 13–26
Advance Access published on 9 March 2012
doi:10.1093/bib/bbs006

Class-imbalanced classifiers for high-dimensional data

Wei-Jiun Lin and James J. Chen

Submitted: 26th October 2011; Received (in revised form): 2nd February 2011

Abstract

A class-imbalanced classifier is a decision rule to predict the class membership of new samples from an available data set where the class sizes differ considerably. When the class sizes are very different, most standard classification algorithms may favor the larger (majority) class resulting in poor accuracy in the minority class prediction. A class-imbalanced classifier typically modifies a standard classifier by a correction strategy or by incorporating a new strategy in the training phase to account for differential class sizes. This article reviews and evaluates some most important methods for class prediction of high-dimensional imbalanced data. The evaluation addresses the fundamental issues of the class-imbalanced classification problem: imbalance ratio, small disjuncts and overlap complexity, lack of data and feature selection. Four class-imbalanced classifiers are considered. The four classifiers include three standard classification algorithms each coupled with an ensemble correction strategy and one support vector machines (SVM)-based correction classifier. The three algorithms are (i) diagonal linear discriminant analysis (DLDA), (ii) random forests (RFs) and (ii) SVMs. The SVM-based correction classifier is SVM threshold adjustment (SVM-THR). A Monte–Carlo simulation and five genomic data sets were used to illustrate the analysis and address the issues. The SVM-ensemble classifier appears to perform the best when the class imbalance is not too severe. The SVM-THR performs well if the imbalance is severe and predictors are highly correlated. The DLDA with a feature selection can perform well without using the ensemble correction.

Keywords: *class-imbalanced prediction; feature selection; lack of data; performance metrics; threshold adjustment; under-sampling ensemble*

INTRODUCTION

Recent advancements in high-throughput technology have accelerated interest in the development of class prediction model (classifiers) for safety assessment, disease diagnostics and prognostics and prediction of response for patient assignment in clinical studies [1–5]. Although many classification algorithms and their applications have been published, classification of imbalanced class size data, where one class is under-represented relative to another, remains among the leading challenges in the development of prediction models.

Classification of the imbalanced data sets arises in many practical biomedical applications. For example, in clinical diagnostic tests of rare diseases or pre-clinical drug-induced adverse toxicity, positive outcomes are rare compared to negative outcomes. Other examples include using gene-expression signatures to distinguish primary from rare metastatic adenocarcinomas [6], prediction of early intrahepatic recurrence of patients with hepatocellular carcinoma [7] and identification of different subtypes of cancer [8]. For these applications, the interest is to correctly identify the samples with outcomes of interest or

Corresponding author. James J. Chen, HFT-20, 3900 NCTR Rd., Jefferson, AR 72079, USA. Tel: +870-543-7007; Fax: +870-543-7662; E-mail: jamesj.chen@fda.hhs.gov

Wei-Jiun Lin is Assistant Professor in the Department of Applied Mathematics, Feng Chia University, Taiwan. He was a postdoctoral fellow at the National Center for Toxicological Research, U.S. Food and Drug Administration. His current research interests include statistical classification and statistical testing for genetic association.

James J. Chen is Mathematical Statistician, National Center for Toxicological Research (NCTR), U.S. Food and Drug Administration. He heads the biometry Branch at NCTR. He is an Adjunct Professor in the Graduate Institute of Biostatistics and Biostatistics Center, China Medical University, Taiwan. His current research interests are statistical methods for biomarker identification for personalized medicine and statistical modeling for quantitative risk assessment.

Published by Oxford University Press, 2012. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

classify the patients into appropriate subgroups as accurately as possible for better intervention.

Most of the current standard classification algorithms are designed to maximize the overall number of correct predictions. This criterion is based on an assumption of an equal cost of misclassifications in each class. When the class sizes differ considerably, most standard classifiers would favor the larger class. In general, the majority class will have a high accuracy in prediction (sensitivity if the positive class is the majority and specificity if the negative class is the majority) and the minority class will have a low accuracy. The procedures are not useful for the above applications. A main challenge in the class-imbalanced classification is to develop a classifier that can provide good accuracy for the minority class prediction [9–17].

Class-imbalanced prediction of high-dimensional data presents an additional challenge. High-throughput genomic, proteomic and metabolomic data are characterized by a large number of predictors (variables) with a relatively small number of samples. In most studies, the majority of predictors are irrelevant to the class membership. Selection of a subset of relevant predictors (feature selection) to enhance predictive performance has become an integral part in the development of classifiers [18].

The poor performance of standard classifiers in minority class prediction can be attributed to these factors: (i) the imbalance ratio, the ratio of the minority class size to the majority class size, (ii) the level of data complexity, the separableness of minority and majority class distributions and (ii) the lack of training data. The first factor reflects the extent of the imbalance between the majority class and minority class sizes. A small imbalance ratio implies more difficulty of the classification problem. The second factor is the characteristics of imbalanced data, including small disjuncts (subclusters in the minority class), ambiguous boundary between classes, overlapping of two classes in feature spaces, dimensionality of data and noisy data. These issues will be addressed in terms of the underlying minority class and majority class distributions. The third factor of the lack of training data influences the first two factors. When the training data size is small, there is a lack of minority class data. The boundary space of the minority class is likely to be underestimated and results in poor performance on the minority class prediction.

Figure 1 demonstrates how the data complexity and lack of data affect the performance of a standard linear classifier. The true boundaries are displayed with dashed lines, and the learned boundaries are displayed with solid lines. The solid points represent the data points in the majority class, and the open points represent the data points in the minority class. In the left figure, two class distributions are well separated. The standard classifier would produce good discrimination, regardless of the imbalance ratio and lack of data. In the middle figure, the two classes are overlapped; the boundary between two classes is ambiguous. The two circles in the overlapped region contain small numbers of the minority class data surrounded by a large number of majority class data. The boundary of the minority class is likely to be underestimated resulting in poor performance of the prediction. In the right figure, when there is a large number of data, both majority class and minority class data in the overlapped region can be misclassified.

Recently, Blagus and Lusa [19] investigated the joint effect of high dimensionality and class imbalance focusing on the effect of variable selection and effectiveness of some correction strategies. They evaluated performance of six classification algorithms with three data-based correction approaches (over- and under-sampling and under-sampling ensemble) and an algorithm-based threshold approach for logistic regression modeling and random forests (RFs). Their analyses provided some useful insights, such as matching the prevalence of the classes in training and test set did not guarantee good performance of classifiers; the problems were exacerbated when dealing with high-dimensional data. However, their

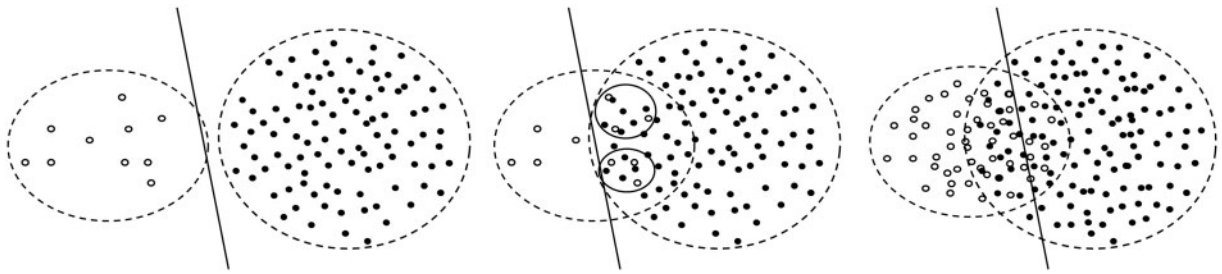

Figure 1: Effects of data complexity and lack of data on classification problem.

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

analyses and interpretations are incomplete. For example, they did not take the correlation structure among genes into consideration in their simulation studies and completely omitted investigating the effect of lack of data, which is a main factor that contributes to the poor performance of standard classification algorithms. Furthermore, their presentation did not put sufficient emphasis on increasing the accuracy of minority class prediction.

In this study, we review and discuss some most important useful methods and fundamental issues for classifying high-dimensional imbalanced class data. First, we review two general strategies for correction of imbalanced data classification and use them with the three classification algorithms—the diagonal linear discriminant analysis (DLDA) [20], RF [21], support vector machines (SVMs) [22, 23]—to develop class-imbalanced classifiers. Four class-imbalanced classifiers are considered, including three standard classification algorithms each coupled with an ensemble correction strategy and one SVM-based correction classifier. Second, we describe some commonly proposed metrics for performance evaluation of class-imbalanced classifiers. Third, we present a simulation study to address the three issues in the class-imbalanced classification problem: imbalance ratio, level of data complexity and lack of data. Fourth, we present an analysis of class-imbalanced classification for five public data sets. Finally, we summarize the important issues in class-imbalanced classification and recommend algorithms and correction strategies for dealing with classification of high-dimensional imbalanced data.

CLASS-IMBALANCED CLASSIFIER

A class-imbalanced classifier is a decision rule on the basis of a training data set where the class sizes differ considerably. The performance of a class-imbalanced classifier depends on the classification algorithm and the strategy for correction of class imbalance as well as the measures of performance (given below). A class-imbalanced classifier typically either modifies a standard classifier by a correction strategy or incorporates a strategy in the training phase to account for differential class sizes. Standard classifiers, such as the decision tree classification C4.5 or C5.0 [14–16], neural networks [24], *k*-nearest neighbor [25, 26], and support vector machines [27–33], were evaluated and modified for imbalanced classification. Many correction strategies have been proposed to improve standard classifiers; these strategies generally can be categorized into the data-based and the algorithm-based approaches described below.

Data-based approach

The data-based approach, also known as the sampling approach, uses a sampling technique to account for class imbalance without modifying a classification algorithm. This is the most common practice in dealing with class-imbalanced data by either under-sampling the majority class or over-sampling the minority class [12]. Chawla *et al.* [34, 35] proposed a technique to generate new synthetic minority class members by interpolating between several positive examples that lied close together; this method is known as SMOTE (synthetic minority over-sampling technique). Chen *et al.* [36] generalized a single sampling approach to two ensemble classifiers, multiple over-sampling and multiple under-sampling ensembles, by generating different bootstrap samples of equal class size in the training set to build ensemble classifiers. Their analysis showed that the under-sampling ensemble method performed more consistent than the over-sampling ensemble method. Recently, using simulation studies, Blagus and Lusa [19] showed that the multiple under-sampling (down-sizing) method was effective if class imbalance was not too severe. The data-based approach can be applied to any classification algorithm.

Algorithm-based approach

This approach modifies the standard classification algorithm to account for class imbalance. Standard classification algorithms generally use a default decision threshold to assign class membership (for example, the probability of 0.5 is used to assign positive and negative class in the logistic regression prediction) for maximizing the classification accuracy, based on an assumption of an equal cost of misclassifications. An algorithm-based approach such as an adjusting decision threshold is a simple modification of the standard algorithm by changing the decision threshold (boundary) in assigning class memberships to account for differential misclassification costs and/or prior probabilities [10, 11, 13]. This approach is also referred to as a cost-sensitive learning. Chen *et al.* [37] considered the decision threshold adjustment approach for four classification algorithms, including classification tree, logistic regression

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

model, Fisher's linear discriminant and a modified nearest neighbor. Their simulation analysis showed that a change of decision threshold can increase the sensitivity and decrease specificity, or vice versa, but the accuracy remains more or less constant. Alternatively, the one-class learning approach, where the classifier learns only on one class to determine the decision boundary, can also be regarded as an algorithm-based approach [38, 39]. Raskutti and Kowalczk [27] demonstrated that a SVM learned only from one class performed well for extremely imbalanced data sets. Another algorithm-based approach is the 'meta imbalanced classification ensemble (MICE)' algorithm, which partitions the majority group and integrates the subclassifiers trained with the partitions and the minority group to deal with the class imbalance issue [40]. One drawback of the algorithm-based approach is that it requires algorithm-specific modification.

Although much research has been conducted, few works provided systematic evaluation on the effects of high dimensionality on the performance of class-imbalanced classifiers. We consider the three algorithms with their associated feature selection methods, DLDA with between–within variance ratio [20], RFs with mean decrease in accuracy [21], SVMs with recursive feature elimination [22, 23], coupled with an under-sampling ensemble correction strategy. A brief description of the three classifiers and feature selections are given in Supplementary Data. These three algorithms have been widely used and shown to perform well in class prediction of the high-dimensional data [18]. Each algorithm also represents unique characteristics for classifying high-dimensional imbalanced data. Only the under-sampling ensemble approach is considered because the other sampling techniques do not perform as well [19, 36]. The ensemble approach in this study generated 101 base classifiers using different bootstrap samples of equal class size; each bootstrap sample set consisted of the entire minority group samples and a set of random samples of equal size generated from the majority group. The majority voting was then used for assignment of new samples.

There have been many recent works proposed to improve the performance of SVM algorithms for class-imbalanced classification [27–33]. However, most of the SVM-based correction strategies still rely on sampling techniques. Some approaches were also used with an ensemble approach. In this study, a simple SVM-based correction classifier is presented: SVM threshold adjustment (SVM-THR).

SVM-THR

Let $f(x)$ be the decision function of SVM for a given sample x. A new sample x will be assigned to the positive class (minority class) if $f(x) \geq 0$ and the negative class (majority class) if $f(x) < 0$. However, for classifying imbalanced data, using the default decision threshold will lead to high accuracy in predicting the majority class and low accuracy in predicting the minority class [28–33]. Chen *et al.* [37] showed that a change of decision threshold can increase the accuracy in predicting the minority class. Based on this concept, we propose a simple approach by adjusting the decision threshold as

$$\theta = -1 + 2\frac{n_+ + a}{n_+ + n_- + 2a} = \frac{n_+ - n_-}{n_+ + n_- + 2a}$$

where n_+ and n_- are the class sizes of minority class and majority class, respectively, and a is a constant to modify the magnitude of the adjustment. In addition to the misclassification costs and/or prior probabilities, the constant a can be specified based on prior analyses such as the performance of the standard SVM, cross-validation, or ROC analysis. In this study, the constant a is simply set to be 1. Since the class size of minority class is smaller than the size of majority class, the adjusted threshold θ is negative. The new threshold will force the classification rule to pay more attention on the minority class. Note that when the class sizes of two classes are equal, the adjusted threshold θ is equal to the default 0.

METRICS FOR CLASS-IMBALANCED CLASSIFIERS

Many metrics have been used for assessment of the performance of classifiers. All of them are based on the four simple measures: the number of true positives (TP), the number of false positives (FP), the number of true negatives (TN) and the number of false negatives (FN). Four commonly used performance metrics are sensitivity, specificity, precision (positive predictive value) and overall accuracy. The sensitivity is $SN = TP/(TP + FN)$, specificity is $SP = TN/(TN + FP)$, precision is $PV = TP/(TP + FP)$ and the accuracy is $ACC = (TP + TN)/(TP + FP + TN + FN)$. The accuracy can be expressed as $\rho SN + (1-\rho)SP$, where ρ is the

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

proportion of positive samples. Accuracy is the most commonly used single metric for performance evaluation; however, it is not an appropriate metric when the class sizes differ considerably. Two alternative metrics, G-mean (geometric mean) and *F*-measure, have been used for performance assessment of class-imbalanced classifiers, where G-mean $= (\mathrm{SN} \times \mathrm{SP})^{1/2}$ and *F*-measure $= (1+\beta^2) \times \mathrm{SN} \times \mathrm{PV}/(\beta^2 \times \mathrm{PV} + \mathrm{SN})$. G-mean is a measure of the ability of a classifier to balance sensitivity and specificity. For a fixed total (SN + SP), G-mean has the maximum when sensitivity and specificity are equal. F-measure is the weighted harmonic mean of the sensitivity and precision. The coefficient β represents the relative preference of sensitivity against precision. The sensitivity is preferred if $\beta > 1$, while the precision is preferred if $\beta < 1$. If $\beta = 1$, the sensitivity and precision are equally important. The β is set as the ratio of the majority class size to the minority class size to emphasize the sensitivity. In evaluation of classifiers, the high values of SN and SP are desirable, but typically there is a tradeoff between the two. These metrics are computed to evaluate performance of the standard and imbalanced data classifiers for imbalanced class data. We only report the four metrics, SN, SP, ACC and G-mean, in the text, and results for the *F*-measure are reported in the Supplementary Tables.

SIMULATION EXPERIMENTS

Four simulation experiments were conducted to investigate the effects of class imbalance on the performance of four class-imbalanced classifiers. The three factors that affect the performance of classifiers were investigated: imbalance ratio, distributions of minority and majority class data, and sample size. The imbalance ratio is the ratio of the minority class size to the majority class size. The performance of class-imbalanced classifiers typically decreases as the ratio decreases. The factors for the distributions of minority and majority class data considered involve mean differences between two classes, the variance–covariance structure and the number of variables (dimensions). Intuitively, large mean differences would likely lead to a well separation of the two classes; thus, the standard classifiers can work well. In contrary, a large standard deviation (variance) would likely have a (large) overlapping area between two classes, resulting in poor performance. A correlation matrix of a real data set was used to generate the correlated model versus independent model in the investigation. The effects of dimensionality were investigated, and subsets of variables were selected to compare with all variables. Finally, the sample size of the training data set is expanded to explore its effect on the performance of class-imbalanced classification.

The simulation design was based on the public colon data set, which contained 2000 genes [41]. One hundred genes were randomly selected as marker genes. The means for the majority class were 0 for all 2000 genes, and the means for the minority class were 0 for non-marker genes and 1 for the marker genes. The data were generated from a 2000-dimensional multivariate normal distribution with the correlation matrix based on the correlation matrix $\mathbf{r}$ of the colon data set. Let V(S,$\mathbf{R}$) denote the structure for the correlation matrix of the simulated data, where S and $\mathbf{R}$ represent, respectively, the standard deviation and (off-diagonal) correlation matrix among the 2000 genes. The standard deviation S is 1 or 2 and the correlation matrix $\mathbf{R}$ is either $\mathbf{0}$ or the correlation matrix of colon data $\mathbf{r}$. For example, V(2,$\mathbf{0}$) denote the standard deviation 2 and correlation 0 for all 2000 genes. The simulation was repeated 1000 times. The results were averaged over these 1000 repetitions.

The first experiment was to investigate the effects of the imbalance ratio on the performance of classifiers. We compared the independent model and correlated model with standard deviation of 1. The total number of training samples was 80 with the imbalance ratios of 1/15, 1/7, 1/3 and 1/1, and additional 80 samples were generated as test data. The effect of the feature selection was evaluated by selecting 50 top-ranked genes (see Supplementary Data) and comparing with 2000 genes without a feature selection.

Table 1 shows empirical estimates of the sensitivity, specificity and accuracy as well as G-mean for the three standard classifiers, DLDA, RF and SVM, and their ensemble classifiers. The results with the *F*-measure are given in Supplementary Table S1. As previously discussed, when the positive class size is much smaller than the negative class size, the sensitivity is low and the specificity is high, which results in high overall predictive accuracy. The estimates of the G-mean are high only when both sensitivity and specificity are high. Thus, the G-mean is a more appropriate measure of performance than the accuracy in imbalanced classification. Comparing the

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

Table 1: Performance of the standard and ensemble classifiers for DLDA, RF and SVM classification algorithms based on 1000 repetitions

			Independent model: V(1,0)								Correlated model[a]: V(1,r)							
			Standard				Ensemble				Standard				Ensemble			
Classifier	Ratio[b]	m_g[c]	SN	SP	ACC	G-mean	SN	SP	ACC	G-mean	SN	SP	ACC	G-mean	SN	SP	ACC	G-mean
DLDA	1/15	All	0.00	1.00	0.94	0.00	0.97	0.97	0.97	0.97	0.38	0.75	0.73	0.49	0.57	0.57	0.57	0.54
		50	0.44	1.00	0.97	0.62	0.97	0.97	0.97	0.96	0.56	0.85	0.83	0.66	0.67	0.69	0.69	0.66
	1/7	All	0.03	1.00	0.88	0.08	1.00	1.00	1.00	1.00	0.47	0.71	0.68	0.57	0.59	0.60	0.60	0.58
		50	0.98	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.69	0.81	0.80	0.74	0.76	0.77	0.77	0.76
	1/3	All	0.86	1.00	0.96	0.92	1.00	1.00	1.00	1.00	0.57	0.68	0.66	0.62	0.63	0.63	0.63	0.62
		50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.74	0.78	0.77	0.76	0.77	0.77	0.77	0.77
	1/1	All	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.65	0.65	0.65	0.64	0.65	0.65	0.65	0.64
		50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.77	0.77	0.77	0.76	0.77	0.77	0.77	0.76
RF	1/15	All	0.00	1.00	0.94	0.00	0.97	0.96	0.96	0.96	0.01	1.00	0.94	0.02	0.61	0.62	0.62	0.59
		50	0.00	1.00	0.94	0.00	0.96	0.96	0.96	0.96	0.12	0.99	0.94	0.22	0.67	0.69	0.69	0.66
	1/7	All	0.00	1.00	0.88	0.00	1.00	1.00	1.00	1.00	0.05	0.99	0.88	0.13	0.72	0.74	0.74	0.72
		50	0.14	1.00	0.89	0.32	0.99	0.99	0.99	0.99	0.29	0.97	0.89	0.49	0.77	0.77	0.77	0.77
	1/3	All	0.02	1.00	0.76	0.09	1.00	1.00	1.00	1.00	0.32	0.98	0.82	0.54	0.83	0.85	0.85	0.84
		50	0.78	1.00	0.95	0.88	1.00	1.00	1.00	1.00	0.53	0.92	0.82	0.69	0.78	0.78	0.78	0.78
	1/1	All	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
		50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
SVM	1/15	All	0.00	1.00	0.94	0.00	0.96	0.95	0.95	0.95	0.56	1.00	0.97	0.71	0.66	0.68	0.68	0.65
		50	0.26	1.00	0.95	0.42	0.97	0.97	0.97	0.97	0.61	1.00	0.97	0.76	0.78	0.82	0.82	0.79
	1/7	All	0.02	1.00	0.88	0.05	0.99	0.99	0.99	0.99	0.96	1.00	0.99	0.98	0.87	0.89	0.89	0.88
		50	0.91	1.00	0.99	0.95	1.00	1.00	1.00	1.00	0.90	1.00	0.99	0.95	0.96	0.97	0.97	0.97
	1/3	All	0.75	1.00	0.94	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		50	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1/1	All	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

[a]The correlated model was based on the correlation matrix **r** of a public colon data set and the standard deviation 1. [b]The class ratio of positive-to-negative samples. [c]The classifiers are performed based on all genes (m_g = all) or 50 selected genes (m_g = 50).

estimates with and without a feature selection, feature selection did improve the performance of DLDA substantially. SVM and RF only showed little improvement since the SVM and RF algorithms implicitly performed variable selection in the development of the prediction model; however, the many selected variables were not useful for improvement because of imbalanced class sizes. With regard to the standard versus ensemble classifiers, all three algorithms showed considerable improvement by using the ensemble methods when the genes were independent. When the genes were correlated, the effect of class imbalance on performance was less pronounced, and the sensitivity and specificity were more balanced, especially for DLDA and SVM. In other words, the ensemble approach for the three classifiers improved less when the genes were correlated. Besides, both DLDA and RF performed less well in the presence of correlation. The minimum estimate of G-mean was 0.96 for both DLDA- and RF-ensemble classifiers when the genes were independent, but the maximum estimates were 0.77 and 0.89, respectively, when the genes were correlated.

Table 2 shows the performance of the SVM algorithm-based correction classifier, SVM-THR. The results with *F*-measure are given in Supplementary Table S2. SVM-THR performed reasonably well. An adjustment of decision threshold to favor minority class prediction made a tradeoff between the specificity and sensitivity which led to slightly lower specificities and much higher sensitivities (Table 1). SVM-THR performed very well under the correlated model with the imbalance ratio of 1/15. The estimate of G-mean for SVM-THR was 0.91 which was higher than the best estimate 0.79 among all ensemble classifiers (Table 1). The feature selection generally improved the performance of SVM-THR except that the imbalance ratio was 1/7 and 1/3 under the correlated model.

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

Table 2: Performance of the SVM-based classifier, SVM-THR, based on 1000 repetitions

Correlation structure	Ratio[b]	m_g[c]	SN	SP	ACC	G-mean
Independent—model: V(1,**0**)	1/15	All	1.00	0.71	0.73	0.84
		50	1.00	0.80	0.81	0.89
	1/7	All	1.00	0.79	0.81	0.89
		50	1.00	0.88	0.90	0.94
	1/3	All	1.00	0.94	0.95	0.97
		50	1.00	0.97	0.98	0.99
	1/1	All	1.00	1.00	1.00	1.00
		50	1.00	1.00	1.00	1.00
Correlated[a] model: V(1,**r**)	1/15	All	0.96	0.84	0.84	0.89
		50	0.98	0.85	0.86	0.91
	1/7	All	1.00	0.92	0.93	0.96
		50	1.00	0.89	0.91	0.94
	1/3	All	1.00	1.00	1.00	1.00
		50	1.00	0.98	0.99	0.99
	1/1	All	1.00	1.00	1.00	1.00
		50	1.00	1.00	1.00	1.00

[a]The correlated model was based on the correlation matrix **r** of a public colon data set and the standard deviation 1. [b]The class ratio of positive-to-negative samples. [c]The classifiers are performed based on all genes (m_g = all) or 50 selected genes ($m_g = 50$).

The second experiment investigated how the standard deviation of the class distributions affected the performance of classifiers. The same parameters for the sample size and imbalance ratio as the first experiment were used with the standard deviation 2 for the two covariance models V(2,0) and V(2,**r**). The results are as expected: the sensitivity, specificity and G-mean estimates were lower under the models V(2,0) and V(2,**r**) than the corresponding estimates under the models V(1,0) and V(1,**r**), respectively. The detailed simulation results are given in Supplementary Tables S3 and S4.

The third experiment explored the effect of dimensionality (the number of genes) on classification of class-imbalanced data. We compared 500, 1000, and 2000 genes with the imbalance ratio 1/15 for models V(1,**0**) and V(1,**r**) (Table 3 and Supplementary Table S5). Note that 500 and 1000 genes were randomly generated from the 2000 genes. The class-imbalanced classifiers performed better in classifying low-dimensional data than high-dimensional data, especially for the standard classifiers under the independent model.

The last experiment investigated the effect of the sample size on the performance of standard classifiers. We considered the sample sizes of 80, 240 and 400 with the imbalance ratio 1/15 for models V(1,**0**) and V(1,**r**) (Table 4 and Supplementary Table S6). When the number of training data was 240 or 400, the standard DLDA and SVM classifiers improved considerably. However, RF showed only small improvement.

EXAMPLES

Five publicly available data sets, colon cancer data, gene-imprint data, breast cancer data, lung cancer data and lymphoma data, were analyzed for further evaluation. Five-fold cross-validation was used to evaluate the performance of each class-imbalanced classifier. Each cross-validation took the class ratio into account and was repeated 50 times to obtain different partitions. The estimates of SN, SP, ACC and G-mean were the averages of the estimates over the 50 repetitions, and the standard deviations were calculated.

Colon cancer data

The colon cancer data set consisted of 40 colon tumor and 22 normal colon tissue samples from an Affymetrix oligonucleotide array with more than 6500 genes. A clustering algorithm revealed broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues [41]. The current data set contained the expression of the 2000 genes with highest minimal intensity across the 62 tissues. The ratio of positive-to-negative was about 1.8:1. The data set is available on the web at (http://genomics-pubs.princeton.edu/oncology/).

Gene-imprint data

The gene-imprint data set was collected to study imprinted genes from the UCSC Genome Browser (http://genome.ucsc.edu/). Imprinted genes tend to affect growth in the womb and behavior after birth. Aberrant imprinting is the cause of various diseases [42]. Greally [43] described that a lack of short interspersed transposable elements (SINEs) is a genomic characteristic of regions undergoing genomic imprinting, which can help predict the presence and extent of imprinted regions. The current data set contained 131 samples and 1446 predictors, where 43 were imprinted and 88 were non-imprinted genes. The ratio of positive-to-negative was about 1:2.1. The data set was obtained from GreallyLab web at http://greallylab.aecom.yu.edu/.

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

Table 3: Effects of numbers of genes on the performance of the class-imbalanced classifiers based on 1000 repetitions and the ratio of positive-to-negative is 1/15

			Independent model: V(1,0)								Correlated model[a]: V(1,r)							
			Standard				Ensemble				Standard				Ensemble			
Classifier	m[b]	m_g[c]	SN	SP	ACC	G-mean	SN	SP	ACC	G-mean	SN	SP	ACC	G-mean	SN	SP	ACC	G-mean
DLDA	500	All	0.59	1.00	0.97	0.75	1.00	1.00	1.00	1.00	0.56	0.83	0.81	0.65	0.69	0.73	0.73	0.70
		50	0.89	1.00	0.99	0.94	1.00	1.00	1.00	1.00	0.61	0.85	0.84	0.70	0.73	0.77	0.77	0.74
	1000	All	0.01	1.00	0.94	0.02	0.99	0.99	0.99	0.99	0.45	0.79	0.77	0.56	0.61	0.64	0.63	0.60
		50	0.71	1.00	0.98	0.83	0.99	0.99	0.99	0.99	0.59	0.85	0.83	0.67	0.71	0.73	0.73	0.70
	2000	All	0.00	1.00	0.94	0.00	0.97	0.97	0.97	0.97	0.38	0.75	0.73	0.49	0.57	0.57	0.57	0.54
		50	0.44	1.00	0.97	0.62	0.97	0.97	0.97	0.96	0.56	0.85	0.83	0.66	0.67	0.69	0.69	0.66
RF	500	All	0.00	1.00	0.94	0.00	1.00	1.00	1.00	1.00	0.04	1.00	0.94	0.08	0.73	0.76	0.76	0.73
		50	0.00	1.00	0.94	0.00	0.99	0.99	0.99	0.99	0.10	0.99	0.94	0.20	0.74	0.77	0.77	0.74
	1000	All	0.00	1.00	0.94	0.00	0.99	0.99	0.99	0.99	0.02	1.00	0.94	0.04	0.67	0.69	0.69	0.66
		50	0.00	1.00	0.94	0.00	0.98	0.98	0.98	0.98	0.08	0.99	0.94	0.16	0.70	0.73	0.72	0.69
	2000	All	0.00	1.00	0.94	0.00	0.97	0.96	0.96	0.96	0.01	1.00	0.94	0.02	0.61	0.62	0.62	0.59
		50	0.00	1.00	0.94	0.00	0.96	0.96	0.96	0.96	0.12	0.99	0.94	0.22	0.67	0.69	0.69	0.66
SVM	500	All	0.50	1.00	0.97	0.67	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.99	0.90	0.94	0.94	0.92
		50	0.75	1.00	0.98	0.86	1.00	1.00	1.00	1.00	0.70	1.00	0.98	0.82	0.87	0.92	0.92	0.89
	1000	All	0.01	1.00	0.94	0.01	0.99	0.99	0.99	0.99	0.88	1.00	0.99	0.93	0.79	0.82	0.82	0.79
		50	0.52	1.00	0.97	0.69	0.99	0.99	0.99	0.99	0.65	1.00	0.98	0.78	0.84	0.88	0.88	0.85
	2000	All	0.00	1.00	0.94	0.00	0.96	0.95	0.95	0.95	0.56	1.00	0.97	0.71	0.66	0.68	0.68	0.65
		50	0.26	1.00	0.95	0.42	0.97	0.97	0.97	0.97	0.61	1.00	0.97	0.76	0.78	0.82	0.82	0.79

[a]The correlated model was based on the correlation matrix **r** of a public colon data set and the standard deviation 1. [b]The number of genes is 500, 1000 or 2000. [c]The classifiers are performed based on all genes (m_g = all) or 50 selected genes ($m_g = 50$).

Table 4: Effects of increasing the training data on the performance of the standard classifiers based on 1000 repetitions and the ratio of positive-to-negative is 1/15

			Independent model: V(1,0)				Correlated model[a]: V(1,r)			
Classifier	n[b]	m_g[c]	SN	SP	ACC	G-mean	SN	SP	ACC	G-mean
DLDA	80	All	0.00	1.00	0.94	0.00	0.38	0.75	0.73	0.49
		50	0.44	1.00	0.97	0.62	0.56	0.85	0.83	0.66
	240	All	0.22	1.00	0.95	0.44	0.50	0.72	0.71	0.59
		50	1.00	1.00	1.00	1.00	0.71	0.80	0.79	0.75
	400	All	0.83	1.00	0.99	0.91	0.57	0.72	0.71	0.64
		50	1.00	1.00	1.00	1.00	0.74	0.79	0.78	0.76
RF	80	All	0.00	1.00	0.94	0.00	0.01	1.00	0.94	0.02
		50	0.00	1.00	0.94	0.00	0.12	0.99	0.94	0.22
	240	All	0.00	1.00	0.94	0.00	0.01	1.00	0.94	0.05
		50	0.07	1.00	0.94	0.22	0.17	0.99	0.94	0.37
	400	All	0.00	1.00	0.94	0.00	0.01	1.00	0.94	0.05
		50	0.14	1.00	0.95	0.36	0.18	0.99	0.94	0.40
SVM	80	All	0.00	1.00	0.94	0.00	0.56	1.00	0.97	0.71
		50	0.26	1.00	0.95	0.42	0.61	1.00	0.97	0.76
	240	All	0.15	1.00	0.95	0.36	1.00	1.00	1.00	1.00
		50	0.97	1.00	1.00	0.98	1.00	1.00	1.00	1.00
	400	All	0.71	1.00	0.98	0.84	1.00	1.00	1.00	1.00
		50	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00

[a]The correlated model was based on the correlation matrix **r** of a public colon data set and the standard deviation 1. [b]The total number of training data is 80, 240 or 400. [c]The classifiers are performed based on all genes (m_g = all) or 50 selected genes ($m_g = 50$).

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

Breast cancer data

The breast cancer data set consists of 99 tumor specimens from breast cancer patients with 7650 genes [44]. Breast cancer is the most commonly diagnosed cancer in women and the second leading fatal cancer among women in the United States [45]. Breast cancer tumors could be divided into two subgroups based on the estrogen receptor (ER) status. Patients with ER-positive tumors have a better survival than those with ER-negative tumors [46]. The ER-positive patients can benefit from anti-estrogens, such as tamoxifen. Sotiriou *et al.* [44] found that gene expression patterns were strongly associated with ER status. Of the 99 patients in this data set, 34 were ER negative and 65 were ER positive. The ratio of positive-to-negative was about 1.9:1. The data set is publicly available at http://www.pnas.org/.

Lung cancer data

Lung cancer is the leading cause of cancer death in the United States [45]. It has been suggested that gene expression profiling could serve as a diagnostic tool in lung cancer [47]. The lung cancer data set contained a total of 203 examples with 12 600 genes. Of the 203 samples, 17 were normal (negative) lung samples and the remaining were 186 tumor (positive) samples. The ratio of positive-to-negative was about 10.9:1. The genes with standard deviations <50 expression units were removed, and the new data set consists of 3312 genes. The data are available at http://www.pnas.org/ and www.genome.wi.mit.edu/MPR/lung.

Lymphoma data

There are two common non-Hodgkin's lymphoma types: diffuse large B-cell lymphomas (DLBCL) and follicular lymphoma (FL) [48]. The lymphoma data set contained 58 patients with DLBCL and 19 patients with FL with 6817 genes [8]. We set the DLBCL samples as positives and FL samples as negatives, and then the ratio of positive-to-negative is about 3.1:1. The gene-expression data set is available at www.genome.wi.mit.edu/MPR/lymphoma.

Table 5 shows the sensitivity, specificity, accuracy and G-mean for the standard DLDA, RF and SVM classifiers. Table 6 shows the results for the RF classifier with the ensemble correction strategy, and the detailed results for the DLDA, RF, SVM and SVM-THR classifiers are shown in the Supplementary Tables S7–S9. The results generally agree with the results from the simulation study. The standard classifiers have high accuracy in predicting the majority class where the performance of RF and SVM is highly sensitive to imbalanced class sizes and DLDA is less affected. The feature selection substantially improves the performance of DLDA, and only little improvement for SVM and RF. Both SVM and RF ensemble correction methods improved the performance considerable.

DISCUSSION

Many biomedical applications have suffered the problem of class-imbalanced classifications where the class imbalances may hinder the performance of standard classifiers. The issues of the class-imbalanced problem have been known for some time and much research has been conducted for addressing it. Only a few studies [19] have partly investigated the effect of class imbalances on classification of high-dimensional data. Many key problems of classification of high-dimensional imbalanced data still remain to be addressed.

This article investigates the five major factors that affect the performance of classifiers for high-dimensional data classification: (i) imbalance ratio, (ii) the minor and majority class distributions, (iii) sample size, (iv) feature selection and (v) the class-imbalanced classifier (classification algorithm and the strategy for correction of class imbalance). The first three factors characterize the underlying issues in the class-imbalanced problem. Any of the three factors or combinations can affect the performance of a classifier. The last two factors address the methods to improve the minority class prediction.

The imbalance ratio measures the degree of difficulty for a standard classifier to address the imbalance problem. The performance of a classifier generally depends on the magnitude of the imbalance ratio (Tables 1 and 2; Supplementary Tables S1 and S2). The performance is also affected by the class distributions, dimensionality and sample size. The performance decreases as the standard deviation or number of genes increases or sample size decreases (Tables 3 and 4; Supplementary Tables S3–S6).

The performance of a classifier highly depends on the underlying distributions of the data of each class. Regardless of the imbalance ratio between the class sizes, a large variance will hinder the performance, whereas a large mean difference will enhance the performance. When the variance is large, there will be between class overlap; the minority data will

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

Table 5: Examples of predictive performance of three standard classification algorithms for five imbalanced genetic and gene expression data

				DLDA			RF			SVM		
Data	**No. of predictors[a]**	**#P/#N[b]**	**m_g[c]**	**SN[d]**	**SP**	**ACC**	**SN**	**SP**	**ACC**	**SN**	**SP**	**ACC**
Colon Cancer	2000	40/22	All	0.70	0.61	0.67	0.88	0.69	0.81	0.88	0.78	0.84
			50	0.85	0.84	0.85	0.89	0.76	0.84	0.89	0.78	0.85
Gene imprint	1446	43/88	All	0.57	0.98	0.84	0.64	0.99	0.87	0.68	0.90	0.83
			50	0.83	0.70	0.75	0.68	0.92	0.84	0.70	0.84	0.79
Breast cancer	7650	65/34	All	0.86	0.81	0.84	0.93	0.71	0.85	0.91	0.74	0.86
			50	0.89	0.85	0.87	0.92	0.83	0.89	0.90	0.76	0.85
Lung cancer	3312	186/17	All	0.98	0.97	0.98	0.99	0.83	0.98	0.99	0.90	0.99
			50	0.99	0.94	0.99	0.99	0.88	0.99	0.99	0.88	0.98
Lymphoma	6817	58/19	All	0.81	0.76	0.80	0.98	0.62	0.89	0.98	0.98	0.98
			50	0.89	0.94	0.90	0.97	0.76	0.92	0.97	0.90	0.95

[a]The number of predictors (genes). [b]The ratio of the positive-to-negative samples. [c]The classifiers are performed based on all genes (m_g = all) or 50 selected genes (m_g = 50). [d]Three measures of performance for evaluation of the three algorithms, sensitivity (SN), specificity (SP) and accuracy (ACC). [e]The performance was evaluated based on 5-fold cross-validation with 50 repetitions.

Table 6: Performance[a] of RF standard and ensemble classifiers on the five imbalanced data sets

		RF-Standard				RF-Ensemble			
Data	**m_g[b]**	**SN[c]**	**SP**	**ACC**	**G-mean**	**SN**	**SP**	**ACC**	**G-mean**
Colon cancer	All	0.88	0.69	0.81	0.78	0.83	0.83	0.83	0.83
		(0.02)	(0.07)	(0.03)	(0.04)	(0.03)	(0.05)	(0.03)	(0.03)
	50	0.89	0.76	0.84	0.82	0.84	0.83	0.84	0.84
		(0.02)	(0.06)	(0.03)	(0.04)	(0.03)	(0.05)	(0.03)	(0.03)
Gene imprint	All	0.64	0.99	0.87	0.79	0.79	0.86	0.84	0.83
		(0.04)	(0.01)	(0.02)	(0.03)	(0.03)	(0.02)	(0.01)	(0.02)
	50	0.68	0.92	0.84	0.79	0.80	0.83	0.82	0.81
		(0.04)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Breast cancer	All	0.93	0.71	0.85	0.81	0.87	0.84	0.86	0.85
		(0.02)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
	50	0.92	0.83	0.89	0.87	0.89	0.85	0.87	0.87
		(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.00)	(0.01)	(0.01)
Lung cancer	All	0.99	0.83	0.98	0.91	0.98	0.94	0.98	0.96
		(0.00)	(0.02)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
	50	0.99	0.88	0.99	0.94	0.99	0.94	0.99	0.97
		(0.00)	(0.04)	(0.00)	(0.02)	(0.00)	(0.01)	(0.00)	(0.01)
Lymphoma	All	0.98	0.62	0.89	0.78	0.87	0.97	0.89	0.92
		(0.00)	(0.07)	(0.02)	(0.05)	(0.01)	(0.03)	(0.01)	(0.02)
	50	0.97	0.76	0.92	0.86	0.89	0.96	0.91	0.92
		(0.01)	(0.07)	(0.02)	(0.04)	(0.02)	(0.03)	(0.01)	(0.02)

[a]The performance was evaluated based on 5-fold cross-validation with 50 repetitions. [b]The classifiers are performed based on all genes (m_g = all) or 50 selected genes (m_g = 50). [c]The estimates (standard deviations) of the four measures, sensitivity (SN), specificity (SP), accuracy (ACC) and G-mean.

likely be classified to the majority class since there are more majority data in the overlapping area. On the other hand, the class data will be well separated or less overlapped when the difference between the two class means is large. The standard classifiers can perform well (Figure 1). An additional simulation study was conducted with a mean difference of 2 between two classes with the total sample size 80 and the imbalance ratio 1/15. DLDA, for example, had sensitivities of 0.68 and 1 without and with variable selection, respectively. Furthermore, genomic variables are correlated; the performance of a classifier is also affected by the underlying correlation structures. In general, the standard classifiers produce more balanced sensitivity and specificity for the correlated model than the independent model, resulting

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

in that the correction strategies are more effective for the independent model (Tables 1 and 2; Supplementary Tables S1–S4).

A lack of data is the primary contributor to the poor accuracy in the minority class prediction. The data complexity characteristics, such as small disjuncts, ambiguous boundary and overlapping between classes, can be attributed to the lack of minority data. When there is a lack of minority class data, the training data set is unlikely to include sufficient instances of the minority class in the boundary area; therefore, the estimated decision boundary can be far less than the true boundary and results in poor minority prediction. On the other hand, when there are sufficient data, the estimated decision will approximate well to the true boundary; the classifier may not be affected by the imbalance between classes (Figure 1). Thus, a standard classifier could perform well when the sample size is sufficiently large or the difference between classes is not too small (Tables 4 and Supplementary Table S6).

All classifiers suffer from lack of data; each factor or combination affects particular classifiers differently. It is useful to identify which algorithm and correction strategy is more robust to a particular imbalance factor or combination. For the algorithms and correction strategies investigated in this study, DLDA appeared to be less affected by the imbalanced class sizes; this result is consistent with a conclusion by Blagus and Lusa [19]. The reason for the performance of DLDA might be that the decision boundary for DLDA was based on the sample means and variances of the two classes which are independent of the ratio of class sizes. For DLDA, the feature selection is essential for classification of high-dimensional data, even when the class size is balanced. All three ensemble classifiers improve the balance between sensitivity and specificity when the class imbalance was modest or severe. It appears that SVM generally performs better than RF. The SVM-THR was shown to perform better than the three ensemble classifiers when the feature variables were correlated and the class imbalance was severe.

The ensemble voting classification is a bagging method [36] based on the idea that a combination of the results of several classifiers will have more accurate prediction than an individual component classifier. However, the main use of the ensemble approach in class-imbalanced problem is to obtain more balanced estimates of sensitivity and specificity. It does not necessarily improve the overall predictive accuracy.

The decision threshold adjustment was developed to estimate the optimal decision threshold for specified misclassification costs and/or prior probabilities of the prevalence [10, 11, 13]. When the class sizes are unequal, a shift in a decision threshold to favor the minority class can increase minority class prediction. There are two challenges in the proposed SVM-THR approach: the choices of shifted distance and the function to adjust the threshold. The linear kernel distance function and the adjusted threshold θ, a function composed of sizes of positive and negative samples with a constant a, were used in this study based on empirical comparisons. Specifically, the numerator of the new threshold represents the difference between the two class sizes and denominator represents the total sample sizes plus $2a$. When $a = 0$, the new threshold only depends on the two class sizes. This new threshold may lead to the over adjustment when the imbalance ratio is too small or the size of training data is large. The constant a was added to alleviate the magnitude of the adjustment. Based on our empirical analysis, the constant a was determined as 1. This adjustment is simple and seems to perform reasonably well.

In the microarray experiments, the collected data typically contains tens of thousands genes; however, many gene are unexpressed, expressed at a relative small level or in only a few samples. Filtering out these genes before data analysis is generally essential to increase the power for identifying the differentially expressed genes (predictive features) [49]. The example colon data set took this issue into consideration in the simulation experiments. This data set contained the expressions of 2000 genes after filtering out the genes with low intensity from more than 6500 genes. The feature selection methods were further identified as 50 most discriminating genes from the underlying 100 differentially expressed genes. The number of genes was varied to investigate the relations between class-imbalanced and dimensionality. A simulation study with 500 and 1000 genes, in addition to 2000 genes, was conducted (Tables 3 and Supplementary Table S5). The results show that dimensionality has impact on classifiers' performance. The performances of class-imbalanced classifiers decrease as the dimensions increase, regardless of the correction strategies and feature selection.

Blagus and Lusa [19] concluded that matching the prevalence of the classes in training and test set does

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

not guarantee good performance of classifiers. A classification model is developed to predict class category of future samples based on the fundamental assumption of that the training samples are representative of the future samples. That is, each future sample is either from the major class distribution or from the minor class distribution, regardless of population size. The performance of a classifier is independent of prevalence of classes. Matching the prevalence should not have an effect on the performance of a classifier.

Three classification algorithms, DLDA [20], RF [21] and SVM [22, 23], were considered to investigate the class-imbalanced problem in the simulation study and real examples, because these classification algorithms have been commonly used and work well in classification of high-dimensional data [18]. The performance of other algorithms on classifying high-dimensional class-imbalanced data, such as *k*-nearest neighbor (k-NN) and prediction analysis of microarrays (PAM), can be found in Blagus and Lusa [19]. In addition, the effect of the minority and majority class distributions (the complexity of data) was studied in terms of the standard deviations 1 and 2, and mean difference of 1 and 2 between two classes. Other aspects, such as subgroups with different means and/or standard deviation or small/no differences can also be studied. For the effect of no difference or small differences, Blagus and Lusa [19] showed that the class-imbalanced problem is more severe.

One goal in oncology studies is to classify the tumor tissues and normal tissues or divide cancers into biologic subtypes. Both the sensitivity and specificity are required to be high for better treatment decisions and avoidance of unnecessary side effects. In this study, five publicly available imbalanced data sets were used for evaluation of the four class-imbalanced classifiers. It seems that the difficulty in classifying the imbalanced data sets is not as serious as that in the simulation study. The predictive performance (sensitivity, specificity and accuracy) depends on the three factors, imbalance ratio, class distributions and sample size. For the examined data sets, it could be due to the large mean difference or large sample size.

In summary, the effect of class imbalance depends on the imbalance ratio, total sample size in the training phase, distributions of the data in each class and selection of the relevant variables as well as the classification algorithm and correction strategy. The poor prediction accuracy of the minority class is primarily caused by the lack of data. In the examined situations, the SVM-ensemble classifier generally outperforms others except when the imbalance is severe and the variables are correlated, where the SVM-THR performs the best. Building a good classifier requires sufficient numbers of minority and majority class samples in the training data. For predicting high-dimensional data, we suggest that one should collect the samples for each class as balanced as possible, a prior cost estimation of the collecting minority class data is useful. Afterward, standard classifiers are identified and the effects of class imbalance are evaluated. Finally, an appropriate class-imbalanced classifier is selected based on the effect of class imbalance and the misclassification costs.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- The performance of classifying high-dimensional imbalanced data is affected by imbalance ratio, distributions of minority and majority class data, sample size and feature selection as well as the classification algorithm and correction strategy.
- A standard classifier could perform well in classification of imbalanced data when the sample size is sufficiently large.
- DLDA appears to be less affected by the class imbalanced, and the feature selection is essential for classification of high-dimensional data, even when the class size is balanced.
- The SVM-ensemble classifier generally outperforms others except when the imbalance is severe and the variables are correlated, where the SVM-THR performs the best.
- For predicting high-dimensional data, we suggest that one should follow the steps: (i) to collect the samples for each class as balanced as possible, (ii) to evaluate the effects of class imbalance using standard classifiers and (iii) an appropriate class-imbalanced classifier is selected based on the effect of class imbalance and the pre-specified misclassification costs.

Acknowledgements

The views presented in this article are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration. The authors would like to thank three reviewers for providing valuable comments and suggestions that improved this article considerably.

FUNDING

This research was supported in part by the National Science Council, Taiwan, under contract NSC 100-2119-M-035-002-.

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

References

1. Helma C, Kramer S. A survey of the Predictive Toxicology Challenge 2000-2001. *Bioinformatics* 2003;**19**:1179–82.
2. Young J, Tong W, Fang H, *et al*. Building an organ-specific carcinogenic database for SAR analyses. *J Toxicol Environ Health A* 2004;**67**:1363–89.
3. Tong W, Xie Q, Hong H, *et al*. Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity. *Environ Health Perspect* 2004;**112**: 1249–54.
4. Rosenkranz HS. SAR modeling of genotoxic phenomena: the consequence on predictive performance of deviation from a unity ratio of genotoxicants/non-genotoxicants. *Mutat Res* 2004;**559**:67–71.
5. Zou W, Lin WJ, Foley SL, *et al*. Evaluation of pulsed-field gel electrophoresis profiles for identification of Salmonella serotypes. *J Clin Microbiol* 2010;**48**:3122–6.
6. Ramaswamy S, Ross KN, Lander ES, *et al*. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003;**33**:49–54.
7. Iizuka N, Oka M, Yamada-Okabe H, *et al*. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet* 2003;**361**:923–9.
8. Shipp MA, Ross KN, Tamayo P, *et al*. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002;**8**: 68–74.
9. Fawcett T, Provost F. Adaptive fraud detection. *Data Min Knowl Discov* 1997;**1**:291–316.
10. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn* 1997;**30**:1145–59.
11. Provost F, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning (ICML98), 1998*. Morgan Kaufmann;445–53.
12. Ling CX, Li C. Data mining for direct marketing: problems and solutions. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), 1998*. AAAI Press;73–79.
13. Provost F, Fawcett T. Robust classification for imprecise environments. *Mach Learn* 2001;**42**:203–31.
14. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. *Intell Data Anal* 2002;**6**:203–31.
15. Weiss GM, Provost F. Learning when training data are costly: the effect of class distribution on tree induction. *J Artif Intell Res* 2003;**19**:315–54.
16. Weiss GM. Mining with rarity: a unifying framework. *SIGKDD Explorations* 2004;**6**:7–19.
17. Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations* 2004;**6**:1–6.
18. Baek S, Tsai CA, Chen JJ. Development of biomarker classifiers from high-dimensional data. *Brief Bioinform* 2009;**10**: 537–46.
19. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2010;**11**:523.
20. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002; **97**:77–87.
21. Breiman L. Random forest. *Mach Learn* 2001;**45**:5–32.
22. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
23. Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002;**46**:389–422.
24. Zhou ZH, Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans Knowl Data Eng* 2006;**18**:63–77.
25. Barandela R, Sánchez JS, García V, *et al*. Strategies for learning in class imbalance problems. *Pattern Recogn* 2003;**36**: 849–51.
26. Zhang J, Mani I. kNN approach to unbalanced data distributions: a case study involving information extraction. *Proceedings of the Workshop on Learning from Imbalanced Datasets II* 2003.
27. Raskutti B, Kowalczyk A. Extreme re-balancing for SVMs: a case study. *SIGKDD Explorations* 2004;**6**:60–9.
28. Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. *Proceedings of the Fifteenth European Conference on Machine Learning (ECML), 2004*; 39–50.
29. Wu G, Chang EY. KBA: kernel boundary alignment considering imbalanced data distribution. *IEEE Trans Knowl Data Eng* 2005;**17**:786–95.
30. Tang Y, Zhang YQ. Granular SVM with repetitive undersampling for highly imbalanced protein homology prediction. *Proceedings of IEEE International Conference on Granular Computing, 2006*;457–60.
31. Lessmann S. Solving imbalanced classification problems with support vector machines. *Proceedings of the International Conference on Artificial Intelligence (IC-AI), 2004*; 214–20.
32. Tang Y, Zhang YQ, Chawla NV, *et al*. SVMs modeling for highly imbalanced classification. *IEEE Trans Syst Man Cybern B Cybern* 2009;**39**:281–8.
33. Byon E, Shrivastava AK, Ding Y. A classification procedure for highly imbalanced class sizes. *IEEE Trans* 2010;**42**: 288–303.
34. Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321–57.
35. Chawla NV, Lazarevic A, Hall LO, *et al*. SMOTEBoost: Improving prediction of the minority class in boosting. *Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003*;107–19.
36. Chen JJ, Tsai CA, Young JF, *et al*. Classification ensembles for unbalanced class sizes in predictive toxicology. *SAR QSAR Environ Res* 2005;**16**:517–29.
37. Chen JJ, Tsai CA, Moon H, *et al*. Decision threshold adjustment in class prediction. *SAR QSAR Environ Res* 2006; **17**:337–52.
38. Juszczak P, Duin RPW. Uncertainty sampling methods for one-class classifiers. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets, 2003*.
39. Schölkopf B, Smola AJ, Williamson RC, *et al*. New support vector algorithms. *Neural Comput* 2000;**12**:1207–45.

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020

40. Lin SC, Chang YCI, Yang WN. Meta-learning for imbalanced data and classification ensemble in binary classification. *Neurocomputing* 2009;**73**:484–94.

41. Alon U, Barkai N, Notterman DA, *et al*. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999;**96**:6745–50.

42. Reik W, Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* 2001;**2**:21–32.

43. Greally JM. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci USA* 2002;**99**:327–32.

44. Sotiriou C, Neo SY, McShane LM, *et al*. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA* 2003; **100**:10393–8.

45. Jemal A, Siegel R, Xu J, *et al*. Cancer statistics, 2010. *Ca Cancer J Clin* 2010;**60**:277–300.

46. Bishop HM, Blamey RW, Elston CW, *et al*. Relationship of oestrogen-receptor status to survival in breast cancer. *Lancet* 1979;**2**:283–4.

47. Bhattacharjee A, Richards WG, Staunton J, *et al*. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001;**98**:13790–5.

48. The Non-Hodgkin's Lymphoma Classification Project A clinical evaluation of the international lymphoma study group classification of non-Hodgkin's lymphoma. *Blood* 1997;**89**:3909–18.

49. Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. *BMC Bioinformatics* 2009;**10**:11.

Downloaded from https://academic.oup.com/bib/article/14/1/13/304457 by Kyungpook National/Bin 164 user on 18 August 2020