World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:9, No:5, 2015

The Classification Performance in Parametric and Nonparametric Discriminant Analysis for a Class-Unbalanced Data of Diabetes Risk Groups

Lily Ingsrisawang, Tasanee Nacharoen

***Abstract*—**The problems arising from unbalanced data sets generally appear in real world applications. Due to unequal class distribution, many researchers have found that the performance of existing classifiers tends to be biased towards the majority class. The k-nearest neighbors' nonparametric discriminant analysis is a method that was proposed for classifying unbalanced classes with good performance. In this study, the methods of discriminant analysis are of interest in investigating misclassification error rates for class-imbalanced data of three diabetes risk groups**.** The purpose of this study was to compare the classification performance between parametric discriminant analysis and nonparametric discriminant analysis in a three-class classification of class-imbalanced data of diabetes risk groups**.** Data from a project maintaining healthy conditions for 599 employees of a government hospital in Bangkok were obtained for the classification problem. The employees were divided into three diabetes risk groups: non-risk (90%), risk (5%), and diabetic (5%). The original data including the variables of diabetes risk group, age, gender, blood glucose, and BMI were analyzed and bootstrapped for 50 and 100 samples, 599 observations per sample, for additional estimation of the misclassification error rate. Each data set was explored for the departure of multivariate normality and the equality of covariance matrices of the three risk groups. Both the original data and the bootstrap samples showed non-normality and unequal covariance matrices. The parametric linear discriminant function, quadratic discriminant function, and the nonparametric k-nearest neighbors' discriminant function were performed over 50 and 100 bootstrap samples and applied to the original data. Searching the optimal classification rule, the choices of prior probabilities were set up for both equal proportions (0.33: 0.33: 0.33) and unequal proportions of (0.90:0.05:0.05), (0.80: 0.10: 0.10) and (0.70, 0.15, 0.15). The results from 50 and 100 bootstrap samples indicated that the k-nearest neighbors approach when k=3 or k=4 and the defined prior probabilities of non-risk: risk: diabetic as 0.90: 0.05:0.05 or 0.80:0.10:0.10 gave the smallest error rate of misclassification. The k-nearest neighbors approach would be suggested for classifying a three-class-imbalanced data of diabetes risk groups.

***Keywords*—**Bootstrap, diabetes risk groups, error rate, k-nearest neighbors.

I. Introduction

DISCRIMINANT analysis (DA) is often used for classification to classify subjects or cases into one of the pre-defined groups when we have a categorical response and one or more measurement variables as predictors. Sometimes, a qualitative or categorical variable may be a useful classifier in nature and is generally preferred as a measured predictor in the classification procedure. For example, gender may be a good classifier and can be treated as a measured variable by creating numerical value 1 if gender is female and 0 if gender is male [1], [9]. Normally, DA, as well as other traditional classification models such as logistic regression, classification trees, or neural networks is based on the implicit assumption of the well balanced distribution of the responses over the sample [6]. In many applications such as medical diagnosis of rare disease, fraud detection in credit card operation, identifying bird species in forest, or etc., there are problems arising from unbalanced data in which one class has higher number of sample observations than others. Classification of imbalance data becomes the interesting issue in statistics and machine learning approach. Most of the existing classifiers tend to be biased towards the majority class of data and give low classification rate to the minority class [11]**.** The class imbalance problems can affect the model estimation and the performance of classification accuracy [4], [8]. In this study, both parametric and nonparametric discriminant analyses are of interest in evaluating the classification performance in a three-class imbalanced data of diabetes risk groups; non-risk (90%), risk (5%), and diabetic (5%). The methods of DA for investigation the classification accuracy consist of parametric linear discriminant function, quadratic discriminant function, and the nonparametric k-nearest neighbors' discriminant function. The parametric DA requires the assumptions of multivariate normality distribution of independent variables and the equality of covariance matrices across groups or classes. The linear discriminant function is applied when both assumptions are met. If the covariance matrices across classes are unequal, the parametric quadratic discriminant function is suggested. In case of the assumption of multivariate normality is not satisfied, the nonparametric k-nearest neighbors' discriminant function is the choice. In evaluation the classification accuracy the lower misclassification error rate indicates the better classification DA technique. Therefore, this study aimed to compare the classification performance between parametric and nonparametric discriminant analyses in a three-class classification with application of class-imbalanced data of diabetes risk groups.

Lily Ingsrisawang and Tasanee Nacharoen are with the Department of Statistics, Kasetsart University, Bangkok 10900, Thailand (phone: +66 (02)-5625555 ext 3873, fax: +66 (02)-9428384;phone: +66 (02)-5625555 ext 3873, fax: +66 (02)-9428384; e-mail: fscilli@ ku.ac.th, tas_tadnee@hotmail.com).

II. Methods

A. Data

Data used in this study were obtained from a project maintaining healthy conditions for 599 employees of a government hospital in Bangkok during a period of August, 2008 and April, 2009. The employees were diagnosed for type II diabetes and classified into one of three diabetes risk groups; non-risk (539 persons), risk (29 persons), and diabetic (31persons). The observed data contained the following variables; diabetes risk group (1 = non-risk, 2 = risk, and 3 = diabetic), age (years), gender (0=male, 1=female), blood glucose (mg/dl), and BMI (kg/m^2). The diabetes risk group was treated as a dependent variable for classification purpose while the other four variables were used as predictors or classifiers for classification procedure.

B. Study Design

A two factor design [2] was generated for analyzing the misclassification error rates in the sample of three-class imbalanced data. One factor was the DA methods which consisted of five different discriminant functions, two functions for parametric DA (linear discriminant function and quadratic discriminant function) and three functions for nonparametric DA (k-nearest neighbors' discriminant function when k = 3, 4, and 5).

The other factor was the prior probability for classifying each subject or sample unit into the t^{th} class (or group), where t = 1, 2 and 3. Four different levels of prior probability setting were specified for those three risk groups of diabetes (non-risk: risk: diabetic) as the followings: 1) equal probability (0.33:0.33:0.33) and 2) unequal proportions with three choices of (0.90:0.05:0.05), (0.80: 0.10: 0.10) and (0.70, 0.15, 0.15), respectively.

A total of 20 combinations of two factor levels (5×4) were investigated for misclassification error rate in a three-class imbalanced data.

Another important aspect of classification is cost. For simplicity, the cost of misclassification for each risk group was assumed to be equal.

C. Discriminant Analysis for Classification

Checking the Assumptions of DA

To choose the appropriate method between parametric and non-parametric DA for classification purpose, it requires checking the common assumptions of multivariate normality and equality of covariance matrices of the three risk groups. Firstly, the multivariate normality in each risk group can be evaluated by estimating its multivariate skewness and kurtosis, and testing for significance levels using Mardia's multivariate skewness and kurtosis. If the estimated values of skewness and kurtosis and the testing results strongly support that the four classifiers do not distributed as the multivariate normal in some risk group, the DA based on non-parametric distributions should be applied for classification the three risk groups [7], [9].

Secondly, if the data on each risk group distributes as multivariate normal with $N_p(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, $t = 1, 2, 3$; $p = 4$, the test of equality of covariance matrices for these three risk groups based on information on the four classifiers: age, gender, blood glucose, and BMI, will be conducted to choose the appropriate parametric discriminant function. The null hypothesis, $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3$, where $\boldsymbol{\Sigma}_t$ is a 4×4 variance-covariance matrix for the t^{th} risk group, can be tested via the likelihood ratio test statistics with Bartlett's correction which is distributed approximately as Chi-squares [7]. If the testing result supports the null hypothesis of equality of covariance matrices, the linear discriminant function (LDF) is needed; otherwise the quadratic discriminant function (QDF) is suggested.

Thirdly, after having tested the equality of covariance matrices for the three populations of diabetes risk group, the difference in their population means should be examined to see whether it is indeed different to each other. Then, the null hypothesis of equality of population means, $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$, where $\boldsymbol{\mu}_t$ is a 4×1 vector of the true mean of age, gender, blood glucose, and BMI, for the t^{th} risk group, can be tested by using Hotelling's T^2 via the quantity $\frac{[n_t - p(t-1)]}{[(n_t - 1)p(t-1)]} T^2$. This amount follows the F distribution with $p(t-1)$ and $n_t - p(t-1)$ degrees of freedom, where n_t is the sample size of the t^{th} risk group [7]. If the null hypothesis of equality of means is rejected, it is useful to perform DA on the observed sample.

Finally, DA was performed based on the selected choice of discriminant function in parametric or nonparametric approach. For parametric DA, the observation $\mathbf{x}$ will be classified into the t^{th} population if the square distance from $\mathbf{x}$ to the t^{th} population, with prior probability of π_t, is minimum based on selected LDF defined in (1) or selected QDF defined in (2), as:

$$d_t^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_t)' S_{pl}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_t) - 2\ln(\pi_t) \quad (1)$$

$$d_t^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_t)' S_t^{-1} (\mathbf{x} - \bar{\mathbf{x}}_t) + \ln|S_t| - 2\ln(\pi_t) \quad (2)$$

In the nonparametric approach for DA, the k-nearest-neighbor (kNN) method was chosen to measure the nearest of an observation $\mathbf{x}_i$ to all other points $\mathbf{x}_j$ by the square distance function defined in (3):

$$d_t^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)' S_t^{-1} (\mathbf{x}_i - \mathbf{x}_j), \quad (3)$$

where $\mathbf{x}_i, \mathbf{x}_j$ are observation vectors for i^{th} and j^{th} objects in group t, $\mathbf{S}_t$ is the variance-covariance matrix within group t, $\mathbf{S}_{pl}$ is the pooled variance-covariance matrix, and π_t is the prior probability of the population in group t. To classify the observation $\mathbf{x}$ into the t^{th} population, the k neighbor observations that are closest to $\mathbf{x}$ will be searched. The choice

for the best k is not clear suggested. From the defined integer k, there are k_t observations come from the t^{th} population with its prior probability π_t. Then, the object with observation $\mathbf{x}$ will be classified into the t^{th} population if its estimated posterior probability for belonging in the t^{th} population is highest [1], [7].

Cross-Validation and Misclassification Error Rate

For evaluation the performance of discriminant function in classification model, cross-validation should be employed for protection the possibility that the model works well only the observed sample, but may not work in the population of the unbalanced three risk groups. Then, splitting the observed data randomly into two data sets, the first set contained two third of the original data, called the training data set, was used to construct the discriminant function. The second set with the rest one third of the original data, called the test data set, was used for evaluation the classification ability of the selected discriminant function [5]. The misclassification error rate (MER) can be used as an indicator for classification performance. The smaller value of MER indicates the better classification function. In this study, the MER was estimated by using the apparent error rate (APER) because it does not depend on the form of the parent population and can be calculated for any classification procedure [9]. The overall apparent error rate for the given discriminant function was estimated by $\sum_{t=1}^{3}\pi_t \sum_{s=1,s\neq t}^{3}\hat{P}(s\mid t)$, where π_t is the assumed prior probability for classifying each subject into the t^{th} population, and $\sum_{s=1,s\neq t}^{3}\hat{P}(s\mid t)$ is the estimated proportion of misclassifications from the t^{th} population into the s^{th} population [7].

Bootstrapping Misclassification Error Rate

In order to judge whether there is enough evidence to conclude that the true parameter of misclassification error rate is less than or equal to some constant value, the idea of bootstrapping was applied [10]. The bootstrap is one in resampling techniques that Efron [3] viewed it as an analysis tool based solely on the data. To generate the misclassification error rates, one bootstrap sample could be obtained by sampling with replacement from the original sample of 599 observations with the same size as the original sample. Based on the one bootstrap sample, the estimated MER was calculated by applying the cross-validation procedure for providing a training set (2/3 of the bootstrap observations) and a test set (1/3 of the bootstrap observations). The training set was used to perform DA using appropriate discriminant function which depends upon the assumptions of multivariate normality and equality of covariance matrices across groups, and the four different patterns of prior probability for the three risk groups. The test set was used to evaluate the classification accuracy by computing the misclassification error rates via the APER for the given discriminant function and prior probability level. Next, repeating this bootstrap procedure over and over again about 50 and 100 times, obtained the bootstrap distribution of estimated MER. The overall MERs from the 50 and 100 bootstrap samples for the given discriminant function and prior probability level were presented by their average values of APER and their corresponding standard errors. All analyses were conducted by using PROC DISCRIM in SAS software.

III. Results

Two parts of results would be presented. The first part was the results obtained from the original sample of imbalanced three risk groups of diabetes, and the second part was the results received from the 50 and 100 bootstrap samples.

A. Results from the Original Sample

A total of 599 employees in the original sample were diagnosed into one of three diabetes risk groups: 539 non-risk persons (90%), 29 risk persons (5%), and 31 diabetic persons (5%).The non-risk group has average age 42.20 years (standard deviation: SD = 5.43 years), average blood glucose 79.76 mg/dl (SD = 8.02mg/dl), average BMI 23.67kg/m^2 (SD = 3.56kg/m^2), and 80% female. For the risk group, they have average age 43.52 years (SD = 4.99 years), average blood glucose 108.33mg/dl (SD = 6.52 mg/dl), average BMI 25.81 kg/m^2 (SD = 4.68kg/m^2), and 76% female. And the diabetic group, they have average age 44.42 years (SD = 6.92 years), average blood glucose 155.70 mg/dl (SD= 6.67 mg/dl), average BMI 26.91 kg/m^2 (SD = 4.46kg/m^2), and 76% female. The mean age for these three risk groups are close together, but the diabetic group has more variability than other groups. There is much difference in mean values of blood glucose across groups; especially the non-risk group has lowest mean and largest variability. For BMI, the diabetic has the highest mean and its value is not far from the mean of the risk group. Moreover, these four predictors were checked for the problem of multicollinearity and found that the pairwise Pearson's correlation coefficients have values ranged from 0.04 to 0.2 (P-values < 0.001).

The skewness and kurtosis values for each predictor in each risk group are presented in Table I. The Madia's test for multivariate normality for each risk group shows that the non-risk group does not have a multivariate normal distribution (P -values for skewness = 0.000 and for kurtosis < 0.001). The risk and the diabetic groups are each distributed as a multivariate normality (P-values for skewness = 0.469 and 0.096 and for kurtosis = 0.203 and 0.852, respectively). Therefore, the three risk groups of diabetes could not be assumed individually distributed as the multivariate normal. The nonparametric DA should be the appropriate approach. Furthermore, the test for homogeneity of variance-covariance matrices for the three risk group populations was rejected by the observed value of chi-square of 522.098 with P-value < 0.0001. The equality of means for the three risk populations was also tested and the exact value of F-statistics revealed that there were statistically different in their population means with P-value < 0.0001. Hence, it might be meaningful to conduct a DA for this observed data. Although, the nonparametric DA

was suggested by theory to be the appropriate method for classifying three risk groups in the observed sample, the parametric DA, linear and quadratic discriminant functions, were also parallel conducted to achieve the objective of comparison the classification performance. The computed MERs from the test sets using the kNN method of nonparametric DA for the unbalanced three risk groups of diabetes were summarized in Table II. As seen, whatever k is equal to 3, 4, or 5 and if the prior probabilities were defined as 0.90: 0.05: 0.05, the same proportions of occurrence for non-risk: risk: diabetic in the original sample, their MERs were about 0.04 which is the smallest value when compared with those in other possibilities.

TABLE I
SKEWNESS AND (KURTOSIS) FOR FOUR PREDICTORS IN EACH RISK GROUP OF DIABETES

Risk groups of Diabetes	Predictors			
	Age	Blood glucose	BMI	Gender
Non-risk	0.009	-0.006	0.713	-1.670
	(0.416)	(0.091)	(0.481)	(0.792)
Risk	0.366	1.173	0.318	-1.327
	(0.784)	(0.656)	(-0.498)	(-0.276)
Diabetic	-0.065	1.626	0.479	-1.372
	(0.729)	(2.036)	(-0.747)	(-0.149)

TABLE II
MISCLASSIFICATION ERROR RATES FOR THREE RISK GROUPS OF DIABETES FROM TEST DATA IN ORIGINAL SAMPLE

k-Nearest Neighbor	Prior Probability			
	0.33 : 0.33: 0.33	0.90: 0.05 : 0.05	0.80: 0.10: 0.10	0.70: 0.15: 0.15
k = 3	0.172	0.041	0.062	0.086
k = 4	0.231	0.038	0.088	0.118
k = 5	0.280	0.040	0.089	0.143

B. Results from the Bootstrap Samples

To be confident that the true value of MER less than or equal to 0.040, the evidence from 50 and 100 bootstrap samples, 599 observations per sample, were investigated. For bootstrap samples, the number of observations appeared in the non-risk, risk, and diabetic groups were in a range of 70 - 90 (%), 5 - 15 (%), and 5 - 15 (%), respectively. The assumptions of multivariate normality and the equality of variance-covariance matrices across three risk groups were tested from the associated training set of each bootstrap sample. Using a 0.05 significance level, the Madia's test indicated that the non-risk group did not have a multivariate normal distribution but the other two risk groups distributed as multivariate normal. Also, the equality of variance-covariance matrices among these three risk groups was statistically significant. Then, DA was performed using both parametric and nonparametric discriminant functions. Applying each discriminant function for each level of prior probability with the test set of the given bootstrap sample; the estimated MER was calculated using APER for measuring the classification performance. The overall MERs for a total of 50 and 100 bootstrap samples were calculated by taking the average values of APERs as presented in Tables III and IV respectively.

The results of DA from either 50 or 100 bootstrap samples showed that the average MER and its standard error are lowest under the kNN method of nonparametric discrimination when k=3 or k=4, and the prior probability setting at a level of 0.90: 0.05: 0.05 or 0.80: 0.10: 0.10 for non-risk: risk: diabetic groups. When the bootstrap samples are increased to 100, the choice of kNN method with k=4 and the prior probability 0.90: 0.05: 0.05 gave the smallest value of 0.01 for MER.

TABLE III
AVERAGE MISCLASSIFICATION ERROR RATES (STANDARD ERROR) FOR THREE RISK GROUPS OF DIABETES FROM TEST DATA IN 50 BOOTSTRAP SAMPLES

Discriminant Method	Prior Probability			
	0.33 : 0.33: 0.33	0.90: 0.05 : 0.05	0.80: 0.10: 0.10	0.70: 0.15: 0.15
Parametric Discriminant Analysis:				
Linear Discriminant Function	0.15 (0.15)	0.13 (0.11)	0.14 (0.12)	0.14 (0.16)
Quadratic Discriminant Function	0.06 (0.08)	0.03 (0.10)	0.04 (0.10)	0.05 (0.11)
Nonparametric Discriminant Analysis: k-Nearest Neighbor (kNN)				
k = 3	0.05 (0.07)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)
k = 4	0.06 (0.08)	0.02 (0.02)	0.02 (0.03)	0.03 (0.04)
k = 5	0.07 (0.08)	0.03 (0.06)	0.05 (0.04)	0.05 (0.04)

TABLE IV
AVERAGE MISCLASSIFICATION ERROR RATES (STANDARD ERROR) FOR THREE RISK GROUPS OF DIABETES FROM TEST DATA IN 100 BOOTSTRAP SAMPLES

Discriminant Method	Prior Probability			
	0.33 : 0.33: 0.33	0.90: 0.05 : 0.05	0.80: 0.10: 0.10	0.70: 0.15: 0.15
Parametric Discriminant Analysis:				
Linear Discriminant Function	0.15 (0.15)	0.13 (0.11)	0.14 (0.12)	0.34 (0.16)
Quadratic Discriminant Function	0.06 (0.08)	0.03 (0.10)	0.04 (0.09)	0.05 (0.10)
Nonparametric Discriminant Analysis: k-Nearest Neighbor (kNN)				
k = 3	0.05 (0.07)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)
k = 4	0.06 (0.07)	0.01 (0.02)	0.02 (0.02)	0.03 (0.03)
k = 5	0.07 (0.08)	0.03 (0.05)	0.05 (0.03)	0.05 (0.03)

IV. CONCLUSION

Although the results from bootstrap samples confirmed that the misclassification error rate is less than or equal to 0.04 when using the nonparametric discrimination approach of kNN method for classifying a three-class-imbalanced data of diabetes risk groups, our concern on the proportions of unbalanced data in different class and the total sample size may affect the classification accuracy. Some classes of the available data set may have small samples and the estimated apparent error rate may be biased downward. To reduce the bias, the method of leaving-one-out cross-validation will be more reliable than the simple splitting the sample into two parts, a training set and a test set.

For future work, simulation study should be the good choice for generating all possible total sample sizes and proportions

of unbalanced data in different class. The leaving-one-outcross-validation is the preferred method for developing discriminant function and evaluating the classification performance. Besides parametric and nonparametric discriminant analyses procedures, the machine learning techniques such as decision tree, support vector machine, or other classification methods should be applied for choosing the best classifier for the case of a three-class-imbalanced data.

REFERENCES

[1] A.C. Rencher, *Methods of multivariate analysis*, New York: Join Wiley & Sons, 1995, ch 9.
[2] A.J.A. Ferrer, and W. Lin, "Comparing the classification accuracy among nonparametric, parametric discriminant analysis and logistic regression methods," *Meeting Papers. Annu. Meeting of the American Educational Research Association*, Montreal, April 13-17, 1999, pp. 1–23.
[3] B. Efron, and R.J. Tibshirani, *An introduction to the bootstrap*, New York: Chapman & Hall, 1993, ch 6.
[4] G. Menardi, "Statistical issues emerging in modeling unbalanced data sets (Abstract)," full text not available for download.
[5] M.H. Kutner, C.J. Nachtsheim, and J. Neter, *Applied linear regression models*, Singapore: McGraw Hill, 2008, pp. 372-375.
[6] N. Japkowicz, and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, pp. 429-450, January 2002.
[7] R. Khattree, and D.N. Naik, *Multivariate data reduction and discrimination*, Cary, NC: SAS Institute Inc., 2000, ch 5.
[8] R. Longadge, S.S. Dongre, and L. Malik, "Class imbalance problem in data mining: review," *IJCSN*, vol. 2, no. 1, pp. 83-87, February 2013.
[9] R.A. Johnson, and D.W. Wichen, *Applied multivariate statistical analysis*, 4th ed., New Jersy: Printice Hall, 1998, ch.11.
[10] R.J. Rossi, *Applied biostatistics for the health sci*ences, Montana: Join Wiley & Sons, 2010, pp.215-218.
[11] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *IJETAE*, vol. 2, no. 4, pp. 42-47, April 2012.