A comparative study of discrimination methods for credit scoring

(Invited Paper)

Hsiang-chun Chen
Department of Statistics
Texas A & M University
College Station, Texas, 77843-3143, USA
Email: ahcchen@neo.tamu.edu

Yi-chin Chen
Department of Industrial and Systems Engineering
Texas A & M University
College Station, Texas, 77843-3131, USA
Email: yichin@neo.tamu.edu

***Abstract*—Credit scoring has become an important management science issue as the credit industry has been experiencing enormous growth during the past few decades. Numerous popular classification methods (e.g. linear discriminant analysis, quadratic discriminant analysis, and logistic regression) have been applied in credit scoring for years. Recently researchers proposed several sophisticated and highly effective data mining techniques, such as Skew-normal discriminant analysis (SNDA), Skew-t discriminant analysis (STDA), Stepwise discriminant analysis (SDA), Sparse discriminant analysis (Sparse DA), Flexible discriminant analysis (FDA), and Mixture discriminant analysis (MDA). The objective of this study is to examine these recently proposed discrimination methods for screening credit applicants. The performance of various credit scoring models is evaluated by one real-world credit scoring dataset. The predictive ability of each credit scoring model is accessed by the total percentage of correctly classified cases (total PCC) and the bad rate among accepts (BRA). The results show that SNDA, STDA, and SDA are outperforming techniques for implementing credit scoring models.**

I. INTRODUCTION

Credit scoring is an evaluation for creditworthiness of a person based on this person's past credit files. Credit scores can not only be used for assessing potential risks but also helps financial institutions discriminate between those applicants whom will repay a loan or card and decide whether to grant or reject credit to applicants. The goal of constructing credit scoring models is to classify credit applicants as good or bad payers. Financial institutions could determine who qualifies for a loan, at what interest rate, and what credit limits in terms of credit scores. The use of credit scoring prior to authorizing access or granting credit is an implementation of a trusted system.

Many researchers have developed sophisticated credit scoring models for credit applications on the basis of the appilcants data. The data of applicants consist of the individual behaviors that have been are observed over a period of time, such as credit history, purpose, gender or ages. A number of credit scoring models have been published and focused on different emphasizes. Researchers treated lending decisions as binary classification problems. Discriminantion methods or regression methods have been applied to credit scoring problems for years. Durand(1941) used quadratic discriminant analysis to classify credit applicants as good or bad payers. Many non-statistical methods are also developed, e.g. decision trees and Neural networks (NN). Data mining approaches are used in establishing credit scoring models as well. A lot of new classificatory algorithms have been researched nowadays. The most of banks use these models to score their applicants.

In this study, some recently proposed discrimination methods are used as tools for constructing classification rules for credit scoring. The remainder of the study is organized as follows. We present a theoretical introduction of various discriminant analyses in section 2. A comparison of these discriminant analyses introduced above is discussed. In section 3, several evaluation methods of the discriminant analysis are provided. One real world dataset: German credit dataset from the UC Irvine Machine Learning Repository (Asuncion and Newman, 2007) is introduced in section 4. Then, all discriminant analyses discussed in section 2 are applied on the German Credit dataset and their predictive performances are evaluated by the total percentage of correctly classified cases (total PCC) and the bad rate among accepts (BRA). Plots of the receiver operating characteristics (ROC) curves are also provided to assess the predictive accuracy of these methods. Section 5 includes with some conclusion and comparison.

II. THEORETICAL BACKGROUND

The discriminant analysis is a popular technique for classifying objects into one of two or more groups based on a set of features that describe the objects. It has been widely used in the fields of economics, education, biology, engineering, etc. Linear discriminant analysis (LDA) is proposed in 1936 by R. A. Fisher. Numerous extended discrimination analyses have been developed since then. In this section, we would like to discuss some of these discrimination methods.

A. REVIEW OF LINEAR DISCRIMINANT ANALYSIS(LDA)

The linear discriminant analysis (LDA) is a useful technique for constructing classification rules. Given a data of multivariate observation, suppose each observation comes from one of K groups which are denoted by $\Pi_1, \Pi_2, \cdots, \Pi_K$. Each group consists of observations with the same or similar characteristics. The discriminant analysis is a procedure for constructing a classification rule that maximizes the separation

among all groups. The measurements on the observation are called the feature variables. The simplest case in discriminant analysis is binary classification problem, that is, there are only two groups ($K = 2$). Credit scoring problem in this study is a binary classification problem.

The Bayes' rule classification is the most commonly used classification strategy. The Bayess rule classifier assigns the observation $X = x$ to Π_1 if $\frac{f_1(x)}{f_2(x)} > \frac{\pi_1}{\pi_2}$, where π_i is the prior probability of Π_i and $f_i(\cdot)$ is the conditional multivariate probability density of X for the ith group, $i = 1, 2$, and assigns $X = x$ to Π_2 otherwise.

Linear discriminant analysis (Fisher, 1936) is a maximum likelihood classification assuming each class is Gaussian distributed. Fisher created a linear discriminant function to find maximum separation among the groups based upon all known measurements. In LDA, assume all measurements are independent normally distributed with arbitrary mean vectors and the common covariance matrices. That is, the K groups have a common covariance matrix $\Sigma_k = \Sigma$ for each K. Then based on the Bayes' rule classification and derivation of logarithm of the ratio of the two posterior probabilities, Fishers linear discriminant function (LDF) is derived from

$$L(x) = \log(\frac{f_1(x)\pi_1}{f_2(x)\pi_2}). \tag{1}$$

LDF is a linear combination of all input variables. The measurements of an observation are used to help collect necessary information for prediction. Then we estimate the coefficients of LDF. It has been proven that the estimated coefficients for LDF are proportional to those for least square regression. For each subject, the discriminant score is the posterior probability that the subject belongs to the second group given all measurements of the subject. If the discriminant score of the function is less than or equal to the cutoff point, the case is classified into the first group, or if above it is classified into the second group.

The aim of the discrimination analysis is forming a classification rule that maximizes the variance between groups but minimizes the variance within groups. Thus solving the discriminant problems is equivalent to estimate the discriminant direction a that

$$\max_a aB^T a \tag{2}$$
$$\text{subject to } aW^T a = 1$$

where B is the between-class covariance and W is the within-class variance. This is called the Fishers criterion.

However, LDA is not appropriate in some situations. For example, when linear boundaries are insufficient in separating the classes or a single prototype per class is insufficient, LDA fails. Various generalizations of LDA are developed to solve those difficulties.

While the covariance matrices are unequal, the quadratic classification rule developed. Separate covariance matrices for each class are estimated. This technique is called the Quadratic discriminant analysis (QDA). The method for creating the discriminant function is similar to what we do in LDA. However, QDA would fail when the covariance matrix is singular. Logistic regression is also widely used in classification problems. Logistic discriminant analysis works well and it works without assumption of distribution. But logistic discriminant analysis is less efficient than LDA.

In order to enhance the efficiency of the classification rules, several modified discrimination methods are proposed, e.g. Skew-normal discriminant analysis (SNDA), Skew-t discriminant analysis (STDA), Stepwise discriminant analysis (SDA), Sparse discriminant analysis (Sparse DA), Flexible discriminant analysis (FDA), and Mixture discriminant analysis (MDA).

B. SKEW-NORMAL AND SKEW-T DISCRIMINANT ANALYSES

Skew-normal and Skew-t discriminant analyses are modified classification techniques. In Skew-normal discriminant analysis (SNDA), instead of assuming normal distribution in LDA, we take the conditional multivariate probability density $f_1(\cdot)$ of X for the first group to be Skew-normal(μ_1, Σ, α) and the conditional multivariate probability density $f_2(\cdot)$ of X for the second group to be Skew-normal(μ_2, Σ, α) . Among sub-populations, the location parameters could be unequal while the scale matrix and the skewness vector parameter keep constant. Similar to LDA, the discriminant function, $L(x) = \log(\frac{f_1(x)\pi_1}{f_2(x)\pi_2})$, is a linear combination of all measurements, and the coefficients can be estimated by maximum likelihood estimation.

Another good method for relaxing the restriction of equal scale matrix and the skewness vector parameter without inflating the number of unknown parameter is assuming that the conditional multivariate probability densities of X for the ith group,$f_i(\cdot)$, are Skew-t distributed. Similarly, the coefficients of the discriminant function could be estimated by maximum likelihood estimation. The classification rule is generally referred to as Skew-t discriminant analysis (STDA).

Due to the nesting of the parametric classes, the use of SN or ST distribution improves the classification a lot. Though there is one more parameter that needs to be estimated in STDA than in SNDA, STDA leads to a visible improvement.

C. STEPWISE DISCRIMINANT ANALYSIS(SDA)

Stepwise discriminant analysis (SDA) is also a widely used technique for building multivariate classifiers. SDA tries to find out the variables in the model that contribute most to the discrimination between groups and builds the discrimination step-by-step. Similar to forwards stepwise variable selection, all variables are reviewed at every step and checking the F-values which indicates the statistical significance in the discrimination between groups. Then the variable that contributes most to the discrimination between groups is included. Keep adding variables until the p-value of the minimum F-value exceeds the predetermined significance level. Also, similar to backward elimination variable selection, all variables are included in the model at each step and then the variable that contributes least to the prediction of group membership

is eliminated. Keep deleting variables until the p-value of the maximum F-value does not exceed the predetermined significance level.
Costanza and Afifi(1979) concluded that using a moderate significance level(0.1 to 0.25) as the predetermined significance level often comes with better performance.
A measure from this group is sometimes used in stepwise discriminant analysis to determine whether adding an independent variable to the model will significantly improve the discrimination between the dependent variables.

D. SPARSE DISCRIMINANT ANALYSIS(SPARSE DA)

Sparse discriminant analysis is a method that performs linear discriminant analysis with a sparseness criterion. By adding an L1 penalty terms constraint on the weights, Sparse DA regularizes the LDA loss function.
Similarly to the Fishers criterion (1), the sparse discriminant criterion is

$$\max_a aB^T a - \lambda_1 \sum_{i=1}^{p} |a_i| \tag{3}$$

$$\text{subject to } aW_p^T a = 1$$

where B is the between-class covariance and W_p is the penalized within- class variance.
The classification, feature selection, and dimension reduction are conducted simultaneously when conducting the Sparse DA. Thus the sparse discriminant analysis is faster in computation and has better classification results.

E. FLEXIBLE DISCRIMINANT ANALYSIS(FDA)

Flexible discriminant analysis (Hastie, Tibshirani, and Buja, 1994) is developed to improve LDA. Since LDA can be viewed as a linear regression problem, FDA replaces the linear regression with a nonparametric regression. A nonparametric regression is fitted to the response on all input variables. And the optimal scores are obtained by computing the eigenvectors matrix. We assign scores to the classes such that the transformed class labels are optimally predicted by regression on the measurement. Then update the model by replacing the optimal scores. The regression method used in optimal scaling in this study is the polynomial regression. FDA permits to use non-linear decision boundaries. FDA can be viewed as the application of LDA on the matrix obtained with the nonparametric regression and on the transformed class matrix.

F. MIXTURE DISCRIMINANT ANALYSIS(MDA)

Mixture discriminant analysis (Hastie and Tibshirani, 1996) is a generalization of LDA. It is developed based on the mixture Gaussian models. MDA keeps the assumption of equal covariance matrix in LDA, but extends the assumption of single normal model in LDA to a mixture of normal distribution. The mixture of normal distributions is used to obtain the density estimation for each class.

$$f_i(x) = \sum_k p_{ik}\phi(x|\mu_{ik}, \Sigma) \qquad i = 1, 2 \tag{4}$$

where $\phi(\cdot)$ is the density of normal distribution with mean μ_{ik} and covariance matrix Σ, and $\sum_k p_{ik} = 1$.
EM algorithm is used to estimate the parameters $(p_{ik}, \mu_{ik}, \Sigma)$ in a mixture of normal distribution for each individual class. Due to the flexibility, MDA outperforms LDA.

III. EVALUATION

The performances of crediting scoring methods are evaluated by the capability of distinguishing the good credit population from the bad credit populations based on the applicants past credit files. In these discrimination analyses, each applicant will be assigned a discriminant score which is the posterior probability that the applicant is a good payer given the behavior measurements of this individual. Applicants with higher score usually have good credits while applicants with lower score usually have bad credits. The selection of cut-off point affect making the granting decisions, hence the performance for each discrimination method. Proficiency of each classification procedure becomes quite important especially when misclassifications are costly.

A. THE TOTAL PERCENTAGE OF CORRECTLY CLASSIFIED CASES(TOTAL PCC)

The actual error rate (AER) is the probability that a classification rule conducted on a given sample will misclassify a future observation (Type I error or Type II error), while the total percentage of correctly classified cases (total PCC) is the probability of correctly classifying a future observation. The total PCC measures how accurately a predictive model will perform in practice, thus it is an evaluation of the precision of a classification rule.
There are several nonparametric methods to estimate the total PCC. The simplest method is the substitution method which one applies the classification rule to the sample, then generates a classification rule and summarizes results in the confusion table. The apparent error rate (APER) is the observed error rate which is defined as the ratio of the total number of misclassified observations to the total number of observations. However, the substitution method usually overestimates the actual correctly classified rates.
If the sample is large, one can divide the sample into the training set and validation set. The classification rule is created using the training set, while the apparent error rate is determined using the validation set. The precision of the classification rule will depend on the split. However, this method is evidently waste of the precious data available because of the omission of the validation set in the training stage.
The leave-one-out cross-validation method (Lachenbruch and Mickey, 1968) is also used to estimate the actual error rate. In leave-one-out cross-validation, one take one observation from the data set as the validation data, use the remaining observations to form the classification rule, and then classify the omitted observation. The procedure is repeated until each observation in the data set is used once as the validation data. However, because of large number of repeats, leave-one-out

cross-validation is usually computationally expensive.
In this study, five-fold cross-validation is used. The original dataset is randomly partitioned into five partitions. Of the five partitions, one partition is retained as the validation set for testing the model, and the remaining four partitions are used as training set. The cross-validation process is then repeated five times with each of the five partitions used exactly once as the validation data. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

B. THE BAD RATE AMONG ACCEPTS(BRA)

The bad rate among accepts (Hand, 2005) is the proportion of the applicants who is with high credit score eventually turn out to be non-creditworthy. Hand (2005) suggests that BRA is an appropriate measure of effectiveness in credit scoring. The BRA also measures the predictive ability over a particular range of cut-off points. The higher predictive ability, the lower the BRA values. In this study, the BRA will be examined over the accept rate from .75 to .90.

C. THE RECEIVER OPERATING CHARACTERISTICS (ROC) curve

A receiver operating characteristic (ROC) curve is a plot of the fraction of true positive rates (TPR) vs. the fraction of false positive rates (FPR) for all possible cutoff points. The ROC curve is more informative than a classification table because it summarizes predictive power for all possible cutoff points. The ROC curves can be used to compare the diagnostic performance of two or more laboratory or diagnostic tests. An idea ROC curve would go through point $(0, 0)$ to point $(0, 1)$ along $Y - axis$ and then through point $(0, 1)$ to point $(1, 1)$ parallel $X-axis$, while a ROC curve of a random performance usually has a concave shape connecting point $(0, 0)$ and point $(1, 1)$. Thus the outperformed discrimination method usually gives a steeper concave curve. The area under the ROC curve is identical to the concordance index which is a measure of predictive power and it is independent of the cutoff points.

IV. EMPIRICAL ANALYSIS

To compare these discrimination methods, we apply them on one real-world credit scoring dataset in this study. This data is from the UCI Machine Learning Repository (Asuncion and Newman, 2007).

A. DATA SAMPLES

The German credit dataset consists of 1000 past credit applicants, of whom 700 applicants were creditworthy applicants and 300 applicants were non-creditworthy applicants. For each applicant, twenty attributes are used as predictors, which include the applicants age, sex, credit history, credit amount, number of existing credits at this bank, etc. Thirteen of these predictors are categorical and seven are quantitative.

TABLE I
TOTAL PCC OF THE AUSTRALIAN CREDIT DATASET FOR THE SNDA, STDA, SDA, SPARSE DA, FDA AND MDA.

Method	Total	Good	Bad
SNDA	.8425	.8869	.7060
STDA	.8375	.8860	.7344
SDA	.7330	.8173	.6736
SPARSE DA	.7450	.7461	.7142
FDA	.7750	.8054	.6863
MDA	.7984	.8107	.5859

TABLE II
BRA OF FOR THE GERMAN CREDIT DATASET FOR THE LDA, SNDA, STDA, SDA, SPARSE DA, FDA AND MDA.

	Accept Rate			
Method	75	80	85	90
SNDA	.1267	.1456	.1796	.2123
STDA	.1151	.1402	.1667	.2022
SDA	.1773	.1988	.2259	.2478
SPARSE DA	.2633	.2688	.2706	.2833
FDA	.2026	.2113	.2341	.2522
MDA	.2013	.2175	.2353	.2556

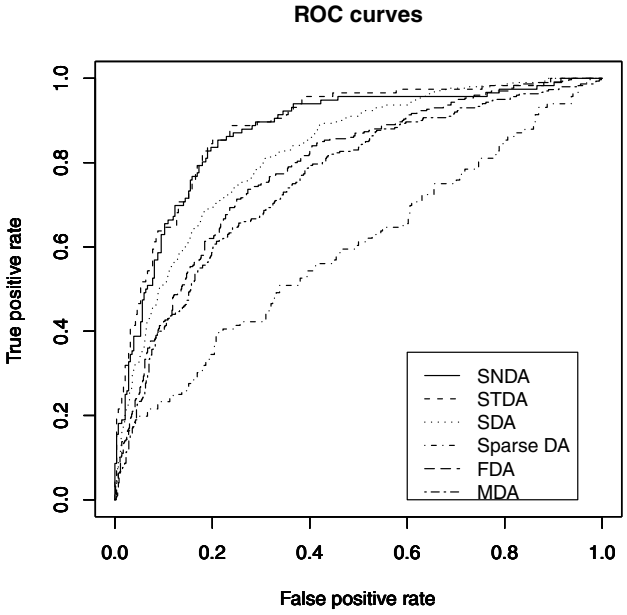
B. RESULTS

The discrimination methods discussed above are applied on the German Credit dataset and five-fold cross validation is used to evaluate their performances. These methods are assessed by their total PCC , BRA and ROC curves.
For the German credit dataset, the total PCC for each method are listed in Table 1. The percentage of correctly classified good risks and percentage of correctly classified bad risks are also provided in Table 1. The results indicate that on basis of the total PCC, the SDA outperform Sparse DA, FDA and MDA while the SNDA and STDA outperform all other discrimination methods.
Table 2 gives the BRA results for the German credit dataset. The results indicate that SNDA and STDA have higher predictive abilities than the other methods in the German credit dataset because of the lower BRA values.
The ROC curves for the German credit dataset are plotted in Figure 1. The ROC curves for SNDA and STDA dominates the other curves over most of the cut-off points. This suggests that SNDA and STDA have higher predictive ability than Sparse DA, FDA and MDA in the German credit dataset.

V. CONCLUSION

In this study, several recently developed discrimination methods are applied for establishing credit scoring models. The performances of these techniques are evaluated by using one real world data. From the results, the SNDA, STDA and SDA outperformed the other discrimination methods on German credit dataset. However, each discrimination method discussed in this study may be robust in some specific models. It is impossible to conclude which method is the most effective technique in general. The main contribution of this paper is to show the comparison of these methods. On the basis of these results, it may conclude that SNDA, STDA and SDA may be

Fig. 1. Receiver operating characteristic (ROC) curves of the German credit dataset for the LDA, SNDA, STDA, SDA, Sparse DA, FDA and MDA.



competitive methods for credit scoring.

REFERENCES

[1] A.C. Antonakis and M.E. Sfakianakis, *Assessing naive Bayes as method for screening credit applicants*, Journal of Applied Statistics, Vol. 36, No. 5, pp537-545, 2009.

[2] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed. John Wiley and Sons Inc., 2007.

[3] A. Asuncion and D.J. Newman, UCI Machine Learning Repository, 2007. Available:http://www.ics.uci.edu/mlearn/MLRepository.html

[4] A. Azzalini and M. G. Genton, *Robust likelihood methods based on the skew-t and related distributions*, International Statistical Review, Vol. 76,Issue 1, pp106-129, 2008.

[5] G.P. Cachon and P.H. Zipkin, *Credit scoring with boosteddecision trees*, MPRA Paper,University Library of Munich, Germany, No. 8034, 2008.

[6] L. Clemmensen, T. Hastie and B. Ersboll, *Sparse discriminant analysis*, Technical report, IMM, Danish Technical University, 2008.

[7] J.N. Crook, D.B. Edelman and L.C. Thomas, *Recent Developements in Consumer Credit Risk Assessment*, European Journal of Operation Research, Vol. 183, pp. 1447-1465 2007.

[8] D. Durand, *Risk Elements in Consumer Instalment Financing*, the Natioal Bureau of Economic Research, 1941.

[9] S. Finlay, *Credit Scoring for profitability objectives*, European Journal of Operation Research, Vol. 202, pp. 528-537, 2010.

[10] D.J. Hand and W.E. Henley *Statistical Classification Methods in Consumer Credit Scoring: A Review*, J. Roy. Stat. Soc. A, Vol. 160, pp. 523-541, 1997.

[11] T. Hastie and R. Tibshirani, *Discriminant Analysis by Gaussian Mixtures*, Journal of the Royal Statistical Society, Series B, Vol. 58, No. 1, pp. 155-176, 1996.

[12] T. J. Hastie, R. Tibshirani and A. Buja, *Flexible Discriminant Analysis by Optimal Scoring*, Journal of the American Statistical Association, Vol. 89, pp. 1255-1270, 1993.

[13] A.J. Izenman, *Modern Multivariate Statistical Techniques*, Springer, 2008.

[14] C. Reynes, R. Sabatier and N. Molinari, *Choice of B-splines with free parameters in the flexible discriminant analysis context*, Computational Statistics and Data Analysis, Vol. 51, Issue 3, pp.1765-1778, 2006.

[15] L.C. Thomas, J.N. Crook and D.B. Edelman, *Credit Scoring and Its Applications*, Society for Industrial and Appplied Mathematics, Philadelphia, 2002.