Classification and Dimension Reduction in Bank Credit Scoring System

Bohan Liu, Bo Yuan, and Wenhuang Liu

Graduate School at Shenzhen, Tsinghua University,
Shenzhen 518055, P.R. China
bohn22@126.com, {yuanb,liuwh}@sz.tsinghua.edu.cn

Abstract. Customer credit is an important concept in the banking industry, which reflects a customer's non-monetary value. Using credit scoring methods, customers can be assigned to different credit levels. Many classification tools, such as Support Vector Machines (SVMs), Decision Trees, Genetic Algorithms can deal with high-dimensional data. However, from the point of view of a customer manager, the classification results from the above tools are often too complex and difficult to comprehend. As a result, it is necessary to perform dimension reduction on the original customer data. In this paper, a SVM model is employed as the classifier and a "Clustering + LDA" method is proposed to perform dimension reduction. Comparison with some widely used techniques is also made, which shows that our method works reasonably well.

Keywords: Dimension Reduction, LDA, SVM, Clustering.

1 Introduction

Customer credit is an important concept in the banking industry, which reflects a customer's non-monetary value. The better a customer's credit, the higher his/her value that commercial banks perceive. Credit scoring refers to the process of customer credit assessment using statistical and related techniques. Generally speaking, banks usually assign customers into good and bad categories based on their credit values. As a result, the problem of credit assessment becomes a typical classification problem in pattern recognition and machine learning.

As far as classification is concerned, some representative features need to be extracted from the customer data, which are to be later used by classifiers. Many classification tools, such as Support Vector Machines (SVMs), Decision Trees, and Genetic Algorithms can deal with high-dimensional data. However, the classification results from the above tools based on the original data are often too complex to be understood by customer managers. As a result, it is necessary to perform dimension reduction on the original data by removing those irrelevant features. Once the dimension of the data is reduced, the results from the classification tools may turn to be simpler and more explicable, which may be easier for bank staff to comprehend. On the other hand, it should be noted that the classification accuracy still needs to remain at an acceptable level after dimension reduction.

F. Sun et al. (Eds.): ISNN 2008, Part I, LNCS 5263, pp. 531–538, 2008.

2 Credit Data and Classification Models

The experimental data set (Australian Credit Approval Data Set) was taken from the UCI repository [2], which has 690 samples, each with 8 symbolic features and 6 numerical features. There are 2 classes (majority rate is about 55.5%) without missing feature values. The data set was randomly divided into training set (490 samples) and test set (200 samples). All numerical features were linearly scaled to be within [0, 1]. In this paper, the SVM model was employed as the classifier, which has been widely used in various classification tasks and credit assessment applications [3, 4, 5].

2.1 Preliminary Results

In order to use the SVM model, all symbolic features need to be transformed into numerical features. A simple and commonly used scheme is shown in Table 1. In this example, a symbolic feature S taking 3 possible values a, b, and c is transformed into 3 binary features (S1, S2, and S3).

Table 1. A simple way to transform symbolic features into numerical features

	S1	S2	S3
S=a	1	0	0
S=b	0	1	0
S=c	0	0	1

In the experimental studies, K-fold cross-validation was adopted [6] where the parameter K was set to 5. In the SVM model, the RBF kernel was used and its parameters were chosen based on a series of trials. The accuracies of the SVM were 86.7347% and 87.5% on the training set and the test set respectively. The implementation of the SVM was based on “libsvm-2.85” [7].

2.2 An Alternative Way to Handle Symbolic Features

There is an alternative way to transform symbolic features, which is based on the idea of probabilities [10]. Let t represent a symbolic feature and its possible values are defined as: $t_1, t_2, \ldots, t_k$. Let ω_i $(i=1,2,\ldots,M)$ denote the i^{th} class label.

For example, the case of $t=t_k$ is represented by:

$$(P(\omega_1 \mid t=t_k), P(\omega_2 \mid t=t_k), \ldots, P(\omega_M \mid t=t_k))$$

Since the sum of probabilities should always equal to 1, each symbolic feature can be represented by M-1 numerical features. As a result, for two-class problems, each symbolic feature can be represented by a single numerical feature. Compared to the scheme in Table 1, this new scheme is favorable when the number of classes is small (two classes in this paper) while the cardinality of each symbolic feature is high. With this type of transformation of symbolic features in the credit data, the accuracies of the SVM were 86.939% and 88.0% on the training set and the test set respectively.

3 Dimension Reduction Techniques

The main objective is to project the original data into a 2D space, which is intuitive to analyze. For this purpose, LDA (Linear Discriminant Analysis) was used to reduce the dimension of the data. Although there are many other dimension reduction tools such as PCA (Principal Components Analysis), LDA is usually preferred in terms of the classification accuracy after dimension reduction. Since LDA can only deal with numerical features, all symbolic features in the original data set were transformed into numerical features by the method in Section 2.2. An improved LDA was also proposed to address some of the weaknesses of the standard LDA technique.

3.1 LDA (Linear Discriminant Analysis)

The purpose of LDA is to perform dimension reduction while preserving as much class discriminatory information as possible [8]. In two-class problems, LDA is often refereed to as FLD (Fisher Linear Discriminant). In this method, the between-class scatter matrix S_B and the within-class scatter matrix S_W are defined as:

$$S_B = \sum_{i,j(i<j)} N_i N_j (\mu_i - \mu_j)(\mu_i - \mu_j)^T \tag{1}$$

$$S_W = \sum_i \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T \tag{2}$$

In Eq.1 and Eq.2, N_i is the number of samples in class ω_i while μ_i is the mean of data in class ω_i. Note that for M-class problems, there are at most M-1 projection directions [9] and consequently it is only possible to project the original data to a line for two-class problems.

The optimal projection is defined as W_{opt} that maximizes the following function:

$$J(W_{opt}) = \left|W_{opt}^T S_B W_{opt}\right| / \left|W_{opt}^T S_W W_{opt}\right| \tag{3}$$

3.2 Clustering Based LDA

Although the objective is to transform the original data into 2D data, for two-class problems, it is only possible to get a single projection vector from the standard LDA. In the following, a new LDA method based on clustering is proposed.

The key idea is to partition the data in each class into subclasses through clustering. The number of subclasses is a tunable parameter of the new LDA method. For example, for a two-class problem, two clusters (subclasses) can be created in each original class and by doing so the number of classes increases from 2 to 4. As a result, it is now possible to get three nonzero eigenvalues (instead of one). The projection directions are determined by finding the nonzero eigenvalues of $S_W^{-1}S_B$. Since the rank of S_B is more than 2, it is now possible to select two projection directions.

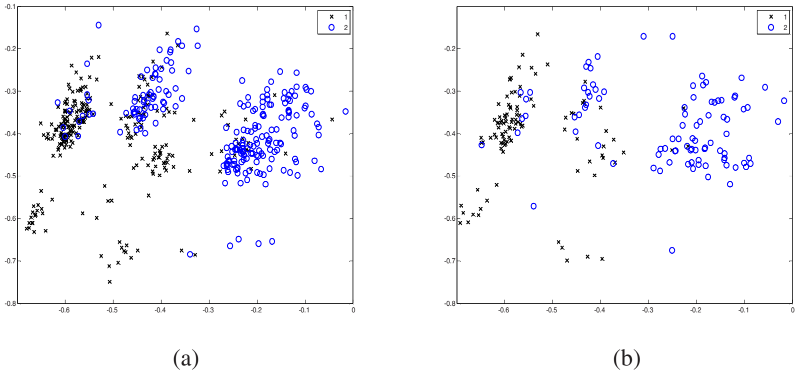
4 Experiments

In order to empirically investigate the performance of the proposed LDA method, experimental studies were conducted to demonstrate its effectiveness. Comparison with two existing LDA extensions capable of producing multiple projection directions for two-class problems was also performed.

4.1 The Effectiveness of Clustering Based LDA

The widely used k-means clustering algorithm with k=2 (divide each original class into two subclasses) was employed. This parameter value was selected based on a few preliminary trials.

Three nonzero eigenvalues were found based on the training set: λ_1=1166, λ_2=571 and λ_3=163. The first two eigenvalues were selected and their corresponding eigenvectors were used as the projection directions. Fig.1 shows the transformed 2D data from the training set and the test set.



(a) (b)

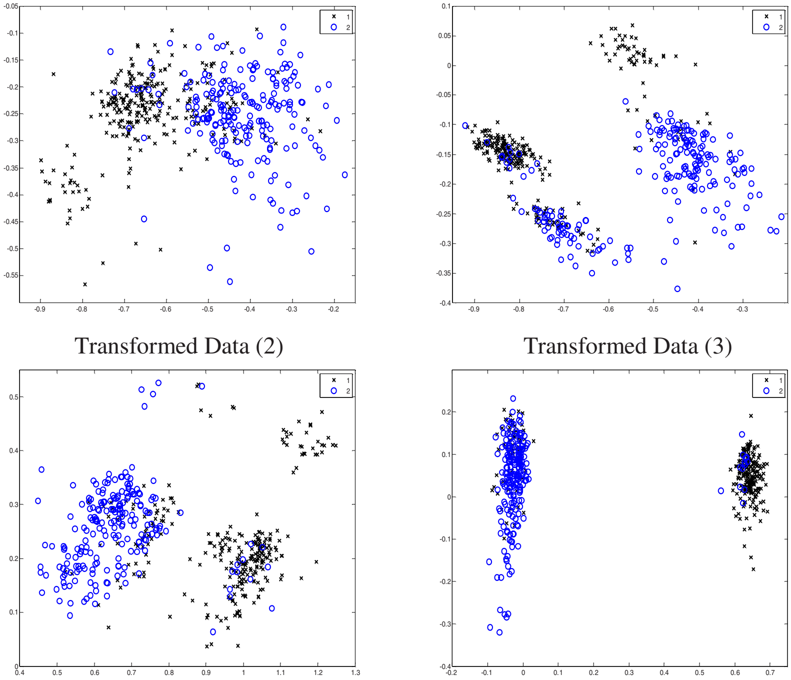
Fig. 1. (a) Training Set; (b) Test Set

Table 2. Classification accuracies of the SVM with the new clustering based LDA

	Training Set	Test Set
Original Data	86.939%	88%
Transformed Data (1)	84.694%	88%
Transformed Data (2)	86.939%	88%
Transformed Data (3)	86.327%	87.5%
Transformed Data (4)	88.163%	87%
Transformed Data (5)	84.082%	89%

As can be seen immediately from Fig.1, the 2D projections make it much easier for people to understand the distribution of the two classes. The accuracies of the SVM on the original data and transformed data, referred to as "Transformed Data (1)", are shown in Table 2. It is clear that the accuracies of the SVM remained almost unchanged while the dimension of the data was reduced from 14 to 2. This result also indicates that the original data set contains significant amount of redundancy as far as classification is concerned.

Since the initial cluster centers are randomly selected in the k-means algorithm, different original cluster centers may result in different final clusters and projection directions. To demonstrate this point, some examples of other 2D projections (training set only) that can be obtained from the same data set are shown in Fig.2.



Transformed Data (2)

Transformed Data (3)

Transformed Data (4)

Transformed Data (5)

Fig. 2. Four different dimension reduction results on the same training set

4.2 Comparison with Other LDA Techniques

There are several variations of the original LDA framework in the literature, which can find multiple nonzero eigenvalues for two-class problems. Two representative examples are briefly described below:

1. Nonparametric Discriminant Analysis (NPLDA) [11] employs the K Nearest Neighbor (KNN) method when calculating the between-class scatter matrix S_B in order to make S_B full of rank. Consequently, it is possible to get more than one nonzero eigenvalues (multiple projection directions).

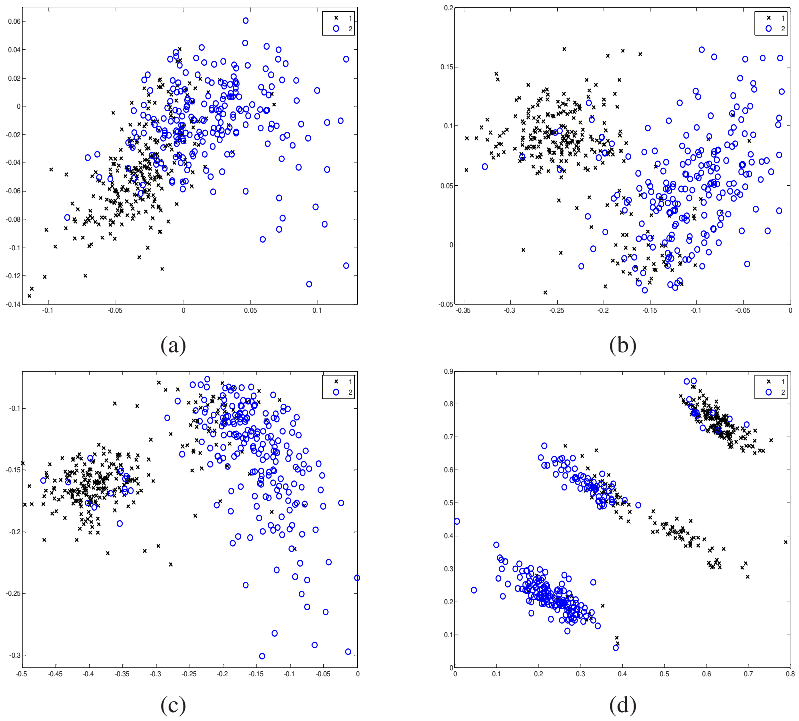


Fig. 3. (a) NPLDA where the parameter K of KNN equals 25; (b) NPLDA when the parameter K of KNN equals 50; (c) NPLDA when the parameter K of KNN equals 100; (d) W2 method

2. The second method (referred to as W2 in this paper) uses the original S_B and S_W. The first projection W_{opt} is the same as in the original LDA. The second projection W_2 (orthogonal to W_{opt}) is defined as the eigenvector corresponding to the nonzero eigenvalue of:

$$\left[S_W^{-1} - \frac{S_B^T \left(S_W^{-1}\right)^2 S_B}{S_B^T \left(S_W^{-1}\right)^3 S_B} \left(S_W^{-1}\right)^2 \right] S_B$$

Table 3. Classification accuracies of the SVM with different LDA methods

	Training Set	**Test Set**
Clustering Based LDA	88.163%	87%
NPLDA, K=25	81.429%	81.5%
NPLDA, K=50	87.551%	85.5%
NPLDA, K=100	88.367%	86%
W2 method	88.571%	87%

As shown in Table 3, in the experiments using NPLDA, when the value of K increased, the accuracy was improved gradually. When K was set to100, the accuracy reached a satisfactory level, although the process of searching for the 100 nearest neighbors for each sample may require extra computational cost. By contrast, the W2 method showed good performance in terms of time complexity and classification accuracy. Note that it can only find a fixed projection map without the flexibility of choosing the number of projection directions as well as selecting the "best" projection maps. In summary, the proposed clustering based LDA method worked reasonably well compared to other representative LDA methods.

5 Conclusion and Future Work

The major focus of this paper is on improving the clarity of the customer data. Generally speaking, dimensionality is a major challenge for data interpretation and understanding by domain experts. For this purpose, various LDA related techniques for dimension reduction were tested, including a new clustering based LDA method. Experimental results showed that all these techniques were effective at reducing the dimension of the customer data set of interest while the classification accuracies of the SVM model remained almost unaffected after dimension reduction.

In addition to the preliminary work reported in this paper, there are a few directions for future work. Firstly, the proposed dimension reduction techniques need to be further tested on large scale customer data sets from commercial banks. Secondly, as shown in this paper, the projection directions as well as the classification accuracies may vary with different cluster patterns from the same data set due to the randomness of the clustering algorithm and different parameter values. As a result, a thorough analysis is required to better understand the relationship between clustering and LDA in order to investigate what kind of cluster patterns are preferred for the purpose of dimension reduction.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC, Project no.: 70671059).

References

[1] Quan, M., Qi, J., Shu, H.: An Evaluation Index System to Assess Customer Value. Nankai Business Review 7(3), 17–23 (2004)

[2] Mertz, C.J., Murphy, P.M.: UCI repository of machine learning databases, http://www.ics.uci.edu/pub/machine-learning-databases

[3] Yang, Y.: Adaptive credit scoring with kernel learning methods. European Journal of Operational Research 183, 1521–1536 (2007)

[4] Martens, D., Baesens, B., Van Gestel, T., Vanthienen, J.: Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research 183, 1466–1476 (2007)

[5] Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, London (2006)

[6] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1143 (1995)

[7] Chang, C., Lin, C.: Libsvm: a library for Support Vector Machine, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[8] Fisher, R.A.: The Use of Multiple Measures in Taxonomic Problems. Ann. Eugenics 7, 179–188 (1936)

[9] Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley, New York (1973)

[10] Duch, W., Grudziński, K., Stawski, G.: Symbolic Features in Neural Networks. In: 5th Conference on Neural Networks and Soft Computing, pp. 180–185 (2000)

[11] Fukunaga, K., Mantock, J.: Nonparametric Discriminant Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 5, 671–678 (1983)