The Application of Weighted Kernel Fisher Discriminant Analysis in Student Loans Default

Tang Qin[1], Zeng Jianyou[2], Li Xing[1], Zhang Hongyang[1]
1 School of Mathematics and Physics, China University of Geosciences, Wuhan, P. R. China, 430074
2 School of Arts and Communication, China University of Geosciences, Wuhan, P. R. China, 430074
(E-mail: tangqin@cug.edu.cn, jianyou@cug.edu.cn, lixing@cug.edu.cn, zhydzh@yahoo.cn)

Abstract: This paper took 2782 data of non-defaulted and defaulted state-subsidized student loan in a university as samples. Firstly, by using Factor Analysis, 7 factors were picked up from original 12 attributes of every sample. Then 70% data were served as training samples and 30% data were served as test samples. Furthermore, Fisher Discriminated, Bayesian Discriminate and Weighted Kernel Fisher Discriminated were respectively used to classify these data. The result indicated that the accuracy rate of Fisher Discriminated was 54.08%, while the accuracy rate of Bayesian was 67.99% and Weighted Kernel Fisher Discriminated reached 74.0%. To decision and management, this research has guiding significance for banks, and the principle and the method can also be applied into other similar problems.

Key words: Factor analysis; Weighted kernel fisher discriminated analysis; Loan defaults

1 Introduction

National Student Loan has played an important role in helping college students from poor families to receive higher education and training talents for the country. At the end of June 2003, Commercial Bank of China has a total payment of the national student loan ¥4.1 billion, It helps 710 000 poor student achieved the dream of college[1]. But in recent years, as a few students lost their honesty, default in bank loans has been on a rise. A survey shows that the default loan rate of many colleges are more than 20%, and a few ones have even been up to 60%[2]. This phenomenon has led to the gradual loss of momentum for state-owned commercial banks to continue loan business, and even in some districts, commercial banks are suspended for this business. In order to identify the impact factor of default from a large number of loan data, Aiming at this problem and applying factor molecules, Bayesian Discriminant, and weighted kernel Fisher discriminate as well as other mathematical methods, we has established a corresponding mathematical model according to 2782 data samples of actual repayment and default in some colleges and universities from Wuhan. The application of the mathematical model enables us to carry out a pre-discriminant analysis of the borrowers, and enhance tracking and management on the "possible defaulter of contract" in accordance with the corresponding analytical results, which can be provided for relevant staffs to make their decisions. It can help banks and universities to strengthen the management of student loan default. It will promote the sustainable development of the college student loans.

2 The Basic Principles of Factor Analysis and the Weighted Kernel Fisher Discriminant Analysis

In order to apply factor analysis and Fisher discriminant method for discriminant analysis of the data, firstly, we introduce the factor analysis and Fisher discriminant, in particular the basic principles of the weighted kernel Fisher discriminant.

2.1 The basic principles of Factor analysis

Factor analysis was first proposed by British psychologist CE Spearman[3]. Based on the correlation between the observed variables $x_1, x_2,\ldots,x_p$, they are translated into several groups (also known as factor variables) $F_1, F_2,\ldots,F_m$, $(m \le p)$, and the correlation between the observed variables within the group is close, while that among groups is loose. Each group represents an unobservable abstract variable, called as the common factor. The corresponding mathematical model is as follows:

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + \varepsilon_1 \\ x_2 = a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + \varepsilon_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_p = a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pm}F_m + \varepsilon_p \end{cases} \tag{1}$$

In the formula, a_{ij} (i=1,2,...,p; j=1,2,...,m) is called as the factor loading, which can be obtained according to the eigenvalues and eigenvectors of matrix R of sample correlation; $\square_i$ (i=1,2,...,p) is called as the special factor.

Factor variable F_j (j=1,2,...,m) can also be represented by the linear combination of observed variable x_i (i=1,2,...,p), that is

$$\begin{cases} F_1 = \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1p}x_p \\ F_2 = \beta_{21}x_1 + \beta_{22}x_2 + \ldots + \beta_{2p}x_p \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ F_m = \beta_{m1}x_1 + \beta_{m2}x_2 + \ldots + \beta_{mp}x_p \end{cases} \tag{2}$$

Because $m<p$, so β_{ji} (i=1,2,...,p; j=1,2,...,m) in the above formula can only be obtained in the sense of least square, and then the factor scores for each factor variable F_j (j=1,2,...,m) can be calculated. Using factor scores, we can do further analysis of discrimination.

2.2 The basic principles of weighted kernel Fisher discriminant

The basic idea of Fisher discriminant is to project k sets of m metadata to a direction, making the variance of the projection data is as large as possible[4]. Although the method can solve many practical problems of discriminant analysis, but there are still some limitations. And kernel Fisher discriminant method is generated to address the limitations of Fisher discriminant method.

Set Φ as the mapping from the input space to higher dimensional space, x as the input vector, and then the mean vector of the two samples in the higher dimensional space is

$$m_i^{\Phi} = \frac{1}{N_i}\sum_{x\in\omega_i}\Phi(x) \qquad (i=1,2) \tag{3}$$

In the formula, ω_i represents the i^{th} class, and N_t is the number of samples in i^{th} class (i=1,2).

Thus the scatter matrix between classes is as follows:

$$S_b^{\Phi} = (m_1^{\Phi} - m_2^{\Phi})(m_1^{\Phi} - m_2^{\Phi})^T \tag{4}$$

The scatter matrix within classes is

$$S_{w_i}^{\Phi} = \sum_{x\in\omega_i}(\Phi(x) - m_i^{\phi})(\Phi(x) - m_i^{\Phi})^T \qquad (i=1,2)$$

The total scatter matrix within classes is

$$S_w^{\Phi} = \sum_{i=1,2} w_i S_{w_i}^{\Phi} = \sum_{i=1,2} w_i \sum_{x\in\omega_i}(\Phi(x) - m_i^{\Phi})(\Phi(x) - m_i^{\Phi})^T, \tag{5}$$

in which w_i (i=1,2) is the weight according to the imbalance of data. The purpose of Kernel Fisher discriminant is to find a proper vector in the direction of projection, which makes

$$J(Z) = \frac{Z^T S_b^{\Phi} Z}{Z^T S_w^{\Phi} Z} \tag{6}$$

reach its maximum value. According to reproducing kernel theory, Z can be represented by linear mapping of all the samples in higher dimensional space[5], that is

$$Z = \sum_{i=1}^{N}\alpha_i\Phi(x_i), \tag{7}$$

in which N=N_1+N_2. By derivation, it can be obtained that

$$J(Z) = \frac{\alpha^T M\alpha}{\alpha^T H\alpha} \tag{8}$$

in the above formula $M_i = (M_1^{(i)}, M_2^{(i)}, \ldots, M_N^{(i)})^T$, while

$$M_j^{(i)} = \frac{1}{N_i}\sum_{k=1}^{N_i}\exp\left(-\frac{\| x_j - x_k^{\omega_i}\|^2}{2\sigma^2}\right) \quad (i=1,2); \tag{9}$$

$$H = \sum_{i=1,2} w_i K_i (I - L_i) K_i^T,$$

$$K_i = \left(k_{pq}^{(i)}\right)_{N\times N_i},$$

$$k_{pq}^{(i)} = \exp\left(-\frac{\| x_p - x_q^{\omega_i}\|^2}{2\sigma^2}\right),$$

I is the unit matrix of order N_i, while L_i is the matrix of order N_i for those whose elements are all $1/N$; In the formula, $\alpha = (\alpha_1, \alpha_2, ... \alpha_N) = H^{-1}(M_1 - M_2)$, which is the vector that makes $J(Z)$ reach its maximum value.

Furthermore, as the projection of $\Phi(x)$ of higher dimensional feature space in Z direction is

$$y = Z^T \Phi(x) = \sum_{j=1}^{N} \alpha_j \exp(-\frac{\| x_j - x \|^2}{2\sigma^2}) \tag{10}$$

The mean value of the two classes after projection

$$\overline{m}_i^{\phi} = \frac{1}{N_i} \sum_{y \in \omega_i} y \qquad (i=1,2) \tag{11}$$

Therefore, the threshold point of classification

$$y_0 = \frac{N_1 \overline{m}_1^{\phi} + N_2 \overline{m}_2^{\phi}}{N_1 + N_2} \tag{12}$$

can be constructed. When $\overline{m}_1^{\phi} \geq \overline{m}_2^{\phi}$, if $y \geq y_0$,then the sample will be awarded to Class 1, otherwise to Class 2; when $\overline{m}_1^{\phi} \leq \overline{m}_2^{\phi}$, if $y \geq y_0$,then the sample will be awarded to Class 2, otherwise to Class 1.

3 Forecast and Analysis of Results

Based on the above principles, methods, and MATLAB7.0 platform, we compiled software, and then analyzed 2782 loan repayment and defaults in a university. In these samples, there are 2575 non-default samples (accounting for 92.56%), with 207 samples of default (accounting for 7.44%). 70% (1947) samples are extracted as training data with the other 30% (835) data as test data, and a mathematical model of the bank loan is also established.

3.1 Extract the main factor by applying factor analysis

In each data (samples), the following 12 indicators or attributes are included: gender, education, age, employment rate of majors, grade point, the economic category of native place, the average income in native place, tuition fee loans, living expenses loans, accommodation loans, study and work situation, and the total amount of loans. After conducting varimax rotation, extract 7 principal factors from the original data, whose variance contributes has been up to 84.519% and they can be interpreted as follows:

The factor load of the first principal factor (F_1) is 17.1%, whose associated variables include tuition fee loan and its total amount. As the total amount of the loan is mainly composed by the tuition, thus the first principal factor could be interpreted as the situation of tuition loans.

The factor load of the second principal factor (F_2) is 15.2%, whose associated variables include the economic groups and average income of native place. Therefore, it could be interpreted as the family's economic condition.

The factor load of the third principal factor (F_3) is 13.6%, and whose associated variables include the academy degree and age. General speaking, academy degree is proportional with the age, thus it could be interpreted as the educational background.

The factor load of fourth principal factor (F_4) is 10.3%, whose associated variables only include the accommodation loan, so it could be interpreted as the situation of accommodation difficulty.

The factor load of the fifth principal factor (F_5) is 10.1%, whose associated variables include the GPA in school, college admission and the employment status. This main factor mainly reflects students' achievement after graduation, thus it could be interpreted as the personal achievement.

The factor load of sixth principal factor (F_6) is 9.5%, whose associated variables include the gender and the employment status of the specified major. As some majors have been largely affected by gender, this principal factor could be interpreted as the situation of the major.

The factor load of seventh principal factor (F_7) is 8.7%, whose associated variables include the loan of the living expenses, so it could be interpreted as the difficulty in living conditions.

According to the factor analysis method, we can get the scores of each principal factor, and then conduct the discriminated classification. In this paper, the weighted kernel Fisher discriminant is adopted to conduct further analysis and processing.

3.2 Analysis on the discriminant results of weighted kernel Fisher and other discriminant methods

Randomly select 1947 data (70%) from the existing 2782 data as a training sample and the remaining 835 data (30%) is regarded as the testing data. Then apply Fisher discriminant and the linear discriminant function to discriminate the testing data and the obtained results are shown as follows: the accuracy rate of non-default is 53.24%, while the accuracy rate of default is 64.52%, and the total accuracy rate is 54.08%. Meanwhile, the Bayes discriminant method is also adopted to conduct discriminated analysis for comparison[6], and the results are as shown in Table 1, which shows that the accuracy rate of the above two methods are relatively low.

Table 1 The Correct Rate of Bayes Discrimination Method Under Different Loss Ratio

Loss ratio	Non-default class	Default class	Total accuracy rate
100	45.60%	82.26%	48.32%
50	53.11%	74.19%	54.68%
20	68.52%	61.29%	67.99%
15	73.32%	48.39%	71.46%

To improve this situation, in this paper, against the ratio of the number of two samples, respectively set the weights as w_1=0.9256, w_2=0.0744 and then conduct discriminant and classification on the mentioned data using the weighted kernel Fisher discriminant method. Under the premise of ensuring the total accuracy rate reaches the maximum, the accuracy rate of default and non-default are made to be balanced as far as possible. Take different value on the parameter σ in kernel Fisher discriminant to conduct analysis. The results show that when σ=7.5 and the accuracy rates of default and non-default are in an equilibrium state, the accuracy rate reaches its maximum. By training samples, we obtain the vector α and the threshold point y_0=-0.0237, thus conducting discriminant and classification on the testing sample. Compared with the actual classes of testing sample, its total accuracy rate is 74.0%, among which the accuracy rate of non-default is 74.7%, while the accuracy rate of default is 64.5%.

To sum up, the weighted kernel Fisher discriminant method is better than the linear Fisher discriminant method and Bayes method (loss ratio of 20) (as shown in Table 2), which fully reflects its superiority. According to the training parameters and classification criteria, we can predict the new data and ensure high reliability.

Table 2 Comparison of the Accuracy Rate of Two Methods

Discriminated method	default	Non-default	Total accuracy rate
Weighted Fisher	64.50%	74.70%	74.0%
Fisher	64.52%	53.24%	54.08%
Bayes	68.52%	61.29%	67.99%

4 Conclusions

In this paper, we establish a comprehensive mathematical model to conduct discriminant analysis, which has successfully applied the factor analysis and the weighted kernel Fisher discriminant method to forecast the loan and loan repayment against the actual data of student loans, with the accuracy rate of testing samples up to 74.0%. Compared with Bayes discriminant method and the linear Fisher discriminant method, the discriminant kernel Fisher discriminant method has the decisive advantage. The model can be used not only in the forecasts of bank loan, but also for other similar situations, such as, the principal factor identification of nutritional status of different lakes. In this model, the selection method of parameter σ has been studied, but its optimum value still needs further study.

References

[1] Lu Wang, Dajian Li. Sustainable Development of National Student Loan [J]. Peking University Education Review, 2004, (1):13-15 (In Chinese)

[2] Miaofeng Xie. Half of the First Group of Guangdong College Student Loan Default Rates Touches the Red Line [N]. 2010-01-12, NanFang Daily (In Chinese)

[3] Dongjin Xiang, Xiaoya Liu, Hongwei Li. Practical Multivariate Statistical Analysis [M]. Wuhan: China University of Geosciences Press, 2005, 157-171 (In Chinese)

[4] Xuemei Yang, Shipeng Li. Prediction of O-glycosylation Sites in Protein Sequence by Kernel Fisher Discriminant Analysis [J]. Journal of Computer Applications, 2010, 30(11):2959-2961

[5] Yugang Fan, Ping Li, Zhihuan Song. Fisher Discriminant Analysis Based on Nonlinear Mapping [J]. Control and Decision, 2007, 22(4):384-388

[6] Lianwen Zhang, Haipeng Guo. Introduction of Bayesian Network [M]. Science Press, 2006, 20 (In Chinese)