Modern Economy, **2013, 4, 461-465**
http://dx.doi.org/10.4236/me.2013.47049 Published Online July 2013 (http://www.scirp.org/journal/me)

The Discrimination Method and Empirical Research of Individual Credit Risk Based on Bilateral Clustering*

Li Shuai, Hui Lai, Chao Xu, Zongfang Zhou
School of Economics and Management, University of Electronic Science and Technology of China, Chengdu, China
Email: xcwdl@126.com

Received April 1, 2013; revised May 6, 2013; accepted June 6, 2013

ABSTRACT

Individual credit risk evaluation has played an extremely important role in the credit risk management of commercial banks. Firstly, through Logistic regression, this paper selects and determines the clustering factors. Then the bilateral clustering structure is proposed. Based on the clustering structure, we cluster to the test samples, and distinguish the individual credit risk as well. Finally, we use the ROC method to test the proposed model and Logistic regression model. The results of comparison show that the discrimination method of individual credit risk based on bilateral clustering can better identify the risk.

Keywords: Individual Credit Risk; Bilateral Clustering; ROC

1. Introduction

The rapid development of Chinese economy led to rapid growth of the credit consumption and continuous growing of the personal credit scale. However, there is neither a scientific and practical system built for personal credit risk valuation, nor robust and reliable valuation models.

Personal credit risk valuation method mainly includes statistics, operation research, artificial intelligence method etc. Among them statistics method includes classification tree, cluster method, linear discriminant analysis etc. David D. (1941) pioneered in applying discriminant analysis in credit risk valuation system [1]; Zhang X., Zhu T. and Yu L. (2011) build credit critical value model based on real sample of some bank through Fisher discriminant analysis model [2]. Classification tree is a nonparametric identification technology, Makowski (1985) and Coffman (1986) applied this method in credit risk valuation area [3]. The application of cluster analysis in credit risk valuation is mainly to classify the sample, Tam *et al.* applied nearest neighbour analysis method in credit risk analysis, using mahalanobis distance to classify the sample, Lundy used cluster analysis to classify and make regression marking for consumer loans applicant according to their application data and age, occupation *etc.* [4] Regression analysis model includes linear regression, Logistic regression, Probit regression etc. Foreign scholars who did research in personal credit risk valuation using regression analysis method include, Fitzpatrick (1976), Lucas (1992) and Henley (1996) [5], etc. There are quite a few research at home in this area, Zheng Y. (2009) did application research in personal credit risk of some bank in Zhejiang Province using traditional probit model [6]; Yang Y. and Shi X. (2009) build bilateral clustering probability model based on artificial immune mechanism, and compared it with Logistic regression model [7]. Operation research method includes integer programming and linear programming; Freed (1981) used linear programming in personal credit risk classification [8]. The most popular artificial intelligence method is a neural network, Security Pacific Bank (SPB) applied neural network intelligence system in the credit valuation of small business loan [9]; Huang H. and Zhou Z. (2010) proposed improved LMBP algorithm to mend the defect of applying BP neural network model in personal credit valuation, and applied ILMBP model in credit risk valuation [10].

Cluster analysis can be applied even when no performance result is available, while Logistic is characterized as simple result, small burden, and propounding classification performance. In order to take advantage of both Logistic regression and cluster analysis, this paper firstly use Logistic regression to regress the element

*This research was be supported by National Natural Science Foundation of China (No. 70971015, 71271043); the special research Foundation of Ph.D program of China (20110185110021).

needed in cluster regression, determine the cluster element, so as to build the Bilateral Clustering Structure, and use the minimum distance to classify the sample data, then we get the default rate, finally, use ROC curve to test the model.

2. To Determine the Cluster Element

2.1. Cluster Element

In cluster analysis, it is very important to determine the cluster elements, which directly influence the accuracy and reliability of the classification result. This paper selects the original data from some Germany commercial bank. As is shown in the **Table 1**, there includes indexes of the original data [11].The left side of **Table 1** are the indexes as follows: X_1 Age, X_2 Marriage, X_3 Supporting family members, X_4 Occupation, X_5 Year of working , X_6 Housing condition, X_7 Year of live in current house, X_8 Installment to deposable disposable income rate, X_9 assets, X_{10} Current payment account status, X_{11} The rest plan for the installment, X_{12} Debt amount, X_{13} Saving account/bonds, X_{14} Loan Period, X_{15} Credit record, X_{16} Existing loan project number in this bank, X_{17} other note debtor/guarantor. While accordingly, the right side in **Table 1** are the definitions of the variables from the original data. Therefore, when we do cluster elements selection, the cluster elements will be picked from 17 indexes form **Table1**.

2.2. Pre-Process of the Sample Data

Table 1 indicates that, the indexes should be standardized: 1) For the discrete data, we use minimum max standardization methods to linear transform the original data, make them into the interval [0,1]; 2) Use scaling transformation to proceed the continuous data [11].

2.3. To Determine the Cluster Element

2.3.1. Collinearity Diagnostics of the Explanatory Variables

In order to make the parameter estimation more accurate, this paper use SPSS16.0 to diagnose the collinearity of the 17 variables, and then use the statistic TOL and VIF to diagnose the existence of collinearity between the explanatory variables. **Table 2** lists the diagnose result of the former 9 variables: Generally, when TOL < 0.1 or VIF > 10, the variables have collinearity problem, **Table 2** shows that the TOL and VIF of variable X_7 and X_8 in

Table 1. The indexes and the definition of the variables from the original data.

Indexes	Variables	Definition
Age	X_1	Actual value
Marriage	X_2	1 = Single; 2 = Married
Supporting family members	X_3	Actual value
Occupation	X_4	1 = Unemployed/Manual workers, Non-Resident; 2 = Non-Proficient worker, resident; 3 = Proficient worker/Officer; 4 = Manager/Independent Eentrepreneurs
Year of working	X_5	1 = Unemployed; 2 = Less than 1 year; 3 = 1 - 4 year; 4 = 4 - 7 year; 5 = More than 7 year
Housing condition	X_6	1 = Rent; 2 = Owned; 3 = Free housing
Year of live in current house	X_7	Actual value
Installment to deposable disposable income rate	X_8	Actual value
Assets	X_9	1 = Real estate; 2 = If not 1: Agreement of public construction savings/Life insurance; 3 = If not 1or 2: Automobile or other; 4 = Vain
Current payment account status	X_{10}	1 = Less than 0 mark; 2 = 0 - 200 dollar; 3 = More than 200 dollar or Salary contract has been signed for at least a year; 4 = No payment account
The rest plan for the installment	X_{11}	1 = Bank; 2 = Stock; 3 = No
Debt amount	X_{12}	Actual value
Saving account/bonds	X_{13}	1 = Less than 100 mark; 2= 100 - 200 dollar; 3 = 500 - 1000 dollar; 4 = More than 1000 dollar; 5 = No saving account/bonds
Loan Period	X_{14}	Actual value
Credit record	X_{15}	0 = No bad credit record; 1 = Has overdue payment record/Other bad credit record; 2 = Overdue payment; 3 = Has late payment record; 4 = No credit record/credit record is no in this bank
Existing loan project number in this bank	X_{16}	Actual value
Other note debtor/guarantor	X_{17}	1 = No; 2 = Joint applicants; 3 = Secured
Sample classification	Y	0 = “Bad” credit; 1 = “Good” credit

Table 2. Collinearity diagnostics.

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
TOL	0.777	0.694	0.934	0.875	0.957	0.669	0.041	0.041	0.586
VIF	1.287	1.441	1.143	1.045	1.431	24.135	24.337	1.761	1.147

dicate collinearity between them.

2.3.2. Logistic Stepwise Regression

To avoid the adverse effect caused by the collinearity of variables, this paper adopts Logistics step wise regression, **Table 3** is the result of the regression. As is shown in the **Table 3**, after 8 steps of screening, the model finally picked variables X_{14}, X_9, X_{10}, X_4, X_3, X_{15}, X_{16} and X_{11}, that is loan period, supporting family members, cur- rent check account, year of working, credit records, loan project number in the current bank and the rest install- ment plan, the coefficient are 0.007, 0.044, −0.098, 0.064, 0.071, −0.066, 0.085, 0.030, The following model can be used to assess the Default status of the individuals:

$$y = 0.925 + 0.007X_{14} + 0.044X_9 - 0.098X_{10} + 0.064X_4 + 0.071X_3 - 0.066X_{15} + 0.085X_{16} + 0.030X_{11}$$

2.3.3. To Determine the Cluster Elements

By using Logistic stepwise regression we selected 8 variables X_{14}, X_9, X_{10}, X_4, X_3, X_{15}, X_{16} and X_{11}, Because Logistic regression is highly descriptive, this paper select the most descriptive variable from the above 8 variables as the cluster element of the sample data, screen according to the condition ROC > 0.5, according to **Table 4** shows, the cluster elements are X_{14}, X_9, X_4 and X_{16}, recorded as cluster element g_1, cluster element g_2, cluster element g_3, and cluster element g_4.

Table 3. Logistic regression MLE.

Variable	Coefficient	Std. Error	t-Statistic	Prob
C	0.925	0.104	8.882	0.000
X_{14}	0.007	0.001	5.750	0.000
X_9	0.044	0.013	3.253	0.000
X_{10}	−0.098	0.011	−9.253	0.001
X_4	0.064	0.018	3.476	0.000
X_3	0.071	0.032	2.195	0.028
X_{15}	−0.066	0.014	−4.688	0.000
X_{16}	0.085	0.025	3.370	0.001
X_{11}	0.030	0.016	1.799	0.072

Table 4. ROC value of the explanatory variables.

Test	X_{14}	X_9	X_{10}	X_4	X_3	X_{15}	X_{16}	X_{11}
ROC value	0.629	0.584	0.292	0.515	0.481	0.373	0.507	0.456

3. To Build the Bilateral Cluster

3.1. The Structure of the Bilateral Cluster

Through normalizing preprocessing, the sample client are randomly divided into 3 groups, the first group is 500 observed samples; the rest of the two group is divided from the remaining 500 data, as test samples. **Figure 1** is the demonstration of the bilateral structure. As is shown in **Figure 1**, the observed samples are divided into default group and non-default group, also called normal client cluster, and default client cluster, so as to form the bilateral cluster. While the remained data, as the test data, also called the newly entered sample w_l, which will be clustered to the default group and non-default group, thus forming a bilateral cluster structure.

Cluster analysis is based on the "distance" and "similarity coefficient", while "distance" is commonly used to measure the similarity of samples. This paper according to the division of observed sample into normal and default client cluster, and based on the similarity between samples, use "distance" as the standard for clustering.

3.2. The Definition of Clustering Distance

Figure 1 not only shows us the structure of bilateral cluster, but also shows us how to definite the "distance", in order to make the newly entered sample w_l clustered to the default and non-default group. As is shown in the figure, supposing $\{u_i, I = 1, 2, 3, \cdots, n_1\}$ is the normal client cluster in the observed sample, among which u_i is the normal client i, U_{ik} $(k = 1, 2, 3, 4)$ is the attribute value of the kth cluster element g_k of client i in the observed sample normal client cluster; Similarly, supposing $\{v_j, j = 1, 2, 3, \cdots, n_2\}$ is the default client cluster in the observed sample, among which v_j is the jth *default client, and* V_{jk} $(k = 1, 2, 3, 4)$ is the attribute value of the kth cluster element g_k of the ith client in the observed sample; $\{w_l, l = 1, 2, 3, \cdots, n_3\}$ is the test sample cluster, among the w_l is the lth *client in the observed sample to be tested*, W_{lk} $(k = 1, 2, 3, 4)$ is the attribute value of the cluster element g_k of the lth client to be tested in the test sample.

This paper use Euclidean distance as the distance between the normal client u_i and the default client v_j in the observed sample:

$$d_{li} = \sqrt{\sum_{k=1}^{4}\left(W_{lk} - U_{ik}\right)^2} \quad (1)$$

$$d_{lj} = \sqrt{\sum_{k=1}^{4}\left(W_{lk} - V_{jk}\right)^2} \quad (2)$$

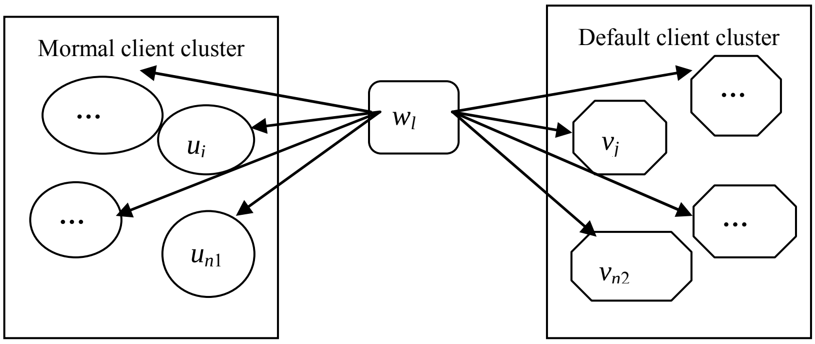

Figure 1. Demonstration of the bilateral cluster structure.

The algorithm of the cluster classification:

1) Select test sample w_l, $l = 1, 2, 3, \cdots, n$, stop while $l > n$;

2) Calculate the bilateral distance of observed sample w_l $D_U = \left(d_{li}\right)_{l=1,2,3\cdots n_1}$; $D_V = \left(d_{lj}\right)_{j=1,2,3\cdots n_2}$;

3) $d_1 = \min(D_U)$, $d_2 = \min(D_V)$, if $d_1 \geq d_2$, then $w_l \in$ default client cluster ,$n_2 = n_2 + 1$; or $d_1 < d_2$, then $w_l \in$ normal client cluster, $n_1 = n_1 + 1$;

4) Repeat the above operation, until the termination conditions occur.

During formation of the normal and default client cluster, n_1 represents the number of the normal clients, n_2 represents the number of defaulted clients, and then we can estimate the default rate of the entire sample client cluster:

$$P_d = \frac{n_2}{n_1 + n_2} \quad (3).$$

4. Model Test

Ususally, ROC value and ROC curve is used to assess the test of the personal credit risk valuation model .

4.1. ROC Curve Test

Figure 2 shows that, after stepwise cluster, the space under the ROC curve grows with adding the test sample. According to the principle that the bigger the space under the ROC curve is indicates the better the discrimination ability the model has. Comparing to the original 500 observed data, after stepwise adding the remaining test data (each of them has 250 data), the discrimination ability gradually increases, which indicates that the cluster model is highly feasible.

4.2. To Compare the Discrimination Ability between Models

Using the same data, the comparison of the bilateral cluster model and the Logistic regression model is shown above (**Table 5**). Logistic regression used 1000 data, get the ROC value of the model is 0.692, while for bilateral cluster model, after adding the test sample a_2 and clustering the ROC value of the model is 0.739 which is higher than the ROC value of the Logistic regression value 0.692, the result indicates that the cluster model can use less data to achieve higher efficiency.

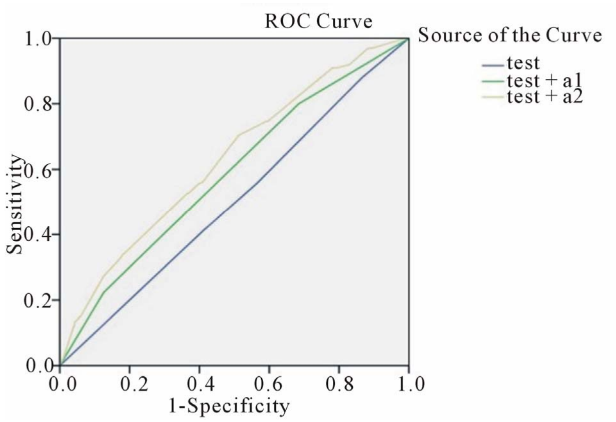

Figure 2. ROC curves of cluster model after twice adding test sample.

Table 5. Influence of the sample scale to the ROC value.

Sample number			ROC value of the cluster model			ROC value of the Logistic regression
Test	a_1	a_2	Test	Test + a_1	Test + a_2	Test
500	250	250	0.617	0.686	0.739	0.692

5. Conclusion

This paper used the data from some Germany commercial bank, built the personal credit risk valuation model based on bilateral clustering, and conducted empirical research. The research result indicates that this method is unusually practicable and effective in the discrimination of personal credit risk, which overcome the defect of traditional personal credit risk valuation and obtains the quality of strong explanatory; The bilateral clustering reduced the complexity of common cluster analysis, and has the advantage of high accuracy and less data oriented, this is the main innovation of this paper. The method which discriminate the clustering result according to "similarity” is very subjective, so, further work can be done in the weight determination, which can be determined by the contribution of the cluster element, and then calculate the distance.

REFERENCES

[1] D. Durand, “Risk Elements in Consumer Installment Financing,” National Bureau of Economy Research, New York, 1941, pp. 189-201.

[2] X. L. Zhang, T. X. Zhu and L. X. Yu, “Research on the Personal Credit Risk Valuation of Commercial Bank Based on Discriminatory Analysis,” *Industrial Technology Economics*, Vol. 10, 2011, pp. 131-137.

[3] J. Y. Coffman, “The Proper Role of Tree Analysis in the Forecasting the Risk Behaviour of Borrowers,” MDS Reports, Management Decision Systems, Atlanta, 1986, pp. 47-59.

[4] E. Mays, “Credit Risk Modeling: Design and Application,” Fitzroy Dearborn Publishers, Chicago, 1998, pp. 190-200.

[5] D. B. Fitzpatrick, “An Analysis of Bank Credit Card Profit,” *Journal of Bank Research*, Vol. 7, 1976, pp. 199-205.

[6] Y. Zheng, “Application Research of Personal Credit Risk Based on Probit Model,” *Shanghai Finance*, Vol. 10, 2009, pp. 85-89.

[7] Y. Yang and X. H. Shi, “Personal Credit Risk Measurement: Bilateral Antibody Artificial Immune Probability Model,” *Systems Engineering-Theory & Practice*, Vol. 29, No. 12, 2009, pp. 89-92.

[8] N. Freed and F. Glover, “A Linear Programming Approach to the Discriminant Problem,” *Decision Sciences*, Vol. 12, No. 1, 1981, pp. 68-74. doi:10.1111/j.1540-5915.1981.tb00061.x

[9] D. West, “Neural Network Credit Scoring Models,” *Computers & Operations Research*, Vol. 27, No. 11-12, 2000, pp. 1131-1152. doi:10.1016/S0305-0548(99)00149-5

[10] H. Z. Huang, Z. F. Zhou and J. K. Yu, “ILMBP Neural Network Model and Its Application in Personal Credit Valuation,” *Managerialist*, Vol. 10, 2010.

[11] “Research of the Combination Forecast Method of Individual Credit Evaluation for Commercial Bank,” Harbin Industrial University Press, Harbin, 2011, pp. 44-45.

