CSCanada
Management Science and Engineering
Vol. 7, No. 3, 2013, pp. 81-85
DOI:10.3968/j.mse.1913035X20130703.2416

ISSN 1913-0341 [Print]
ISSN 1913-035X [Online]
www.cscanada.net
www.cscanada.org

Applied Research of Genetic Algorithm in Personal Credit Risk Combined Assessment

SHUAI Li[a]; LI Tingting[a]; XU Chao[a]; ZHOU Zongfang[a],*

[a]School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China.
*Corresponding author.

Supported by the Natural Science Foundation of China (71271043), the Special Research Foundation of Ph.D Program of China (20110185110021), and Sichuan Province Science and Technology Support Project (2012SZ0001).

Received 17 June 2013; accepted 16 August 2013

Abstract

With the increasing scale of individual credit consumption, the individual credit risk assessment has become more and more important. This paper selected 640 samples from the Germany personal credit database as study object. First, this preliminary screened the primitive indexes, and then used sample classification accuracy as fitness function, making combined assessment model based on linear regression and logistic regression through genetic algorithm. The results showed that the combined assessment model based on genetic algorithm had higher accuracy compared to a single model, and combined assessment model based on the least sum of square error had an advantage in the individual credit risk assessment over the others.

Key words: Personal credit risk assessment; Genetic algorithm; Logistic regression; Combined assessment

SHUAI Li, LI Tingting, XU Chao, ZHOU Zongfang (2013). Applied Research of Genetic Algorithm in Personal Credit Risk Combined Assessment. *Management Science and Engineering*, *7*(3), 81-85. Available from: URL: http://www.cscanada.net/index.php/mse/article/view/j.mse.1913035X20130703.2416
DOI: http://dx.doi.org/10.3968/j.mse.1913035X20130703.2416

INTRODUCTION

Credit risk, also known as default risk, refers to the potential losses that the bank might suffer from the unwillingness or incapability of the borrower to fulfill the contract because of various reasons (Zhou et al., 2010). Causes for individual credit risk can be concluded into the following aspects: asymmetric information, the borrower's moral, repayment will and ability. Personal credit assessment methods including using rigorous scientific analysis methods, taking into account the inner and outer, subjective and objective environment that influence the individual and the family, making comprehensive judgment and assessment to their ability to fulfill the economic commitments, and using certain symbols to indicate their credit state (Huang, Zhou, & Yu, 2010).

The existing personal credit assessment methods are mainly divided into two classes, one is the method based on statistic model, such as multiple discriminant analysis, linear regression, logistic regression; the other is non-parametric estimation and artificial intelligence method, including neural network, classification tree, and genetic algorithm etc.

Research on personal credit assessment is continuing developing and maturing at home and abroad. For example, Liang, Guo, Li, & Fang (2002) analyzed the main cause of the rapid development of modern credit risk model, comparatively analyzed the theory, advantage and disadvantage of each model; Shi & Jin (2004) comparatively analyzed the application of various methods in domestic small sample data; Jiang & Chen (2006) provided the LS estimation and hypothesis test of the weight of combined forecast model, from statistic aspect; Ma & Tang (1998) described and analyzed the accuracy feature of the linear combined forecast model, based on the absolute error series of the linear combined model; Jiang (2006) took the idea of combined forecast, proposed the forecasting method of combining the multiple linear regression and logistic regression by using RBF neural network, and applied it in personal credit assessment; Chen & Xu (2000) applied and compared three combined forecast model: linear combination, variable weights combination, and non-linear combination; Fogarty & Ireson (1993) were two pioneer scholars in applying genetic algorithm in scoring; Zhu & Feng (2003) proposed the method to determine the weight factor of the combined forecast method based on genetic algorithm.

This paper select 640 German personal credit sample data as research object, firstly, this preliminary screens the primitive indexes, then establish two single statistic model: linear regression and Logistic regression, and build the personal credit combined assessment model, finally, analyze the empirical results.

1. THE SELECTION AND QUANTIFICATION OF PERSONAL CREDIT ASSESSMENT INDEXES

The personal credit assessment indexes that this paper selects conclude the following three aspects. (1) The naturel condition of the individuals, such as age, sex, marriage, education, housing and vehicle condition. (2) The financial condition of the individuals, such as profession and occupation, job and title, years of working, monthly income, annual family revenue, monthly payment to monthly disposable income rate. (3) Social condition of individuals, on one hand, it is about the status of the business among the individual and the bank and other financial institutions, mainly includes: Bad credit record, sum of business with the bank, saving account in the bank, default record, year of credit record, on the other hand, it is about the status of the business among the individual and the non-bank financial institutions, main includes: insurance, tax evasion, tax loophole record, record of malicious default on utilities, and judicial, public records.

Due to the difficulty of collecting the original personal credit assessment data and files in the domestic, this paper selected 1000 personal credit sample data of German as the study objects, which have 20 attributes indexes, according to the Chinese context, the indexes are finally determined as following table shows:

Table 1
Selection and Quantification of Personal Credit Assessment Indexes

Indexes	Variable	Definition
Age	X_1	Actual value
Marriage	X_2	1 = Single; 2 = Married
Supporting family members	X_3	Actual value
Working condition	X_4	1 = Unemployed/manual workers, non-resident; 2 = Non-proficient worker, resident; 3 = Proficient worker/officer; 4 = Manager/independent entrepreneurs
Year of working	X_5	1 = Unemployed; 2 = Less than 1 year; 3 = 1-4 years; 4 = 4-7 years; 5 = More than 7 years
Housing condition	X_6	1 = Rent; 2 = Owned; 3 = Free housing
Year of living 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 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 dollars or salary contract has been signed for at least a year; 4 = No payment account
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

2. IDEA OF BUILDING THE COMBINED ASSESSMENT MODEL BASED ON GENETIC ALGORITHM

Combined assessment is the weighted combination of different assessment model, and it has unique effect in improving the accuracy and robustness of the assessment, which leads to its wide application and massive study to it. The theory of linear combination assessment is as follows:

We presume there are *m* assess method to one problem, f_{it} is value of the *i* th kind of assess method for the *t* th person (i=1,2...,m;t=1,2,...,n), w_i represents the weight of *i* th assess method, f_t represents the value of combined assessment for the *t* th person, then the mathematical model of the linear combination assessment is as follow:

$$f_t = \sum_{i=1}^{m} w_i f_{it}$$

$$st, \sum_{i=1}^{m} w_i = 1, t = 1, 2, ..., n \qquad (1)$$

The key to combined assessment model is to solve the weight of it, this paper use genetic algorithm to get the weight w_i.

Genetic Algorithm (GA) initially proposed by Holland (1975), which is a heuristic optimization algorithm to a simulated biological population genetic and evolutionary mechanism, the search is guided by the fitness value of the individual (Lei, Zhang & Li, 2004) . The selection of the fitness function and the genetic operator (selection operator, cross operator, mutation operator) is the core of the genetic algorithm. GA determines the search direction by choosing the operator to weed out the poor individual of the group, and the selection of the operator is determined by the fitness value, then carry out the operation of cross and mutation, finally get the optimum solution, some key steep are as follow:

(1) The coding of the genetic algorithm and the setting of the initial group

Coding is the first problem that needs to be solved before applying the genetic algorithm, usually there two coding method of coding: binary coding (0,1 coding) and floating number coding. Binary coding is the main coding way in genetic algorithm, because the parameters that need to be determined in constructing the combined assessment model are few, and the value range of it is small, this paper use binary coding. When the number of the group is small it can improve the running speed in the cost of reducing the diversity of the group, and leading to premature convergence of the genetic algorithm; When the number of the group is large, it will reduce the efficiency. Generally, proposed group scale is 20-100. This paper select two of them to construct a combined predict model, with 2 parameters to be determined, thus, the initial group scale is set 20, and the generation of the group evolution is set 100.

(2) Selection of fitness function

Fitness value measures the adaptation ability of the individual in evolution. Individuals with high fitness value is more likely to survive, and produce offspring with high adaptation ability through genetic mechanism, otherwise, individuals with low fitness value will gradually be weed out. This paper use classification accuracy to calculate the fitness function: $fit = 1 - \frac{m}{n}$, *n* is the number of the sample, *m* is misclassified number.

(3) Selection of the three operators

The selection of operator is built based on the valuation of the adaptation ability of individuals. Its main purpose is to avoid missing gene, and to improve the overall convergence, and operating speed. This paper use proportional selection method, first calculate the fitness function value of each individual, the sort them, and select individual according the following probit:

$$p_i = \frac{fit_i}{\sum_{i=1}^{N} fit_i} \qquad (2)$$

Among them, fit_i is the fitness value of individual i , p_i is the probability of each individual to be selected, N is the scale of the group, this paper set 20.

Crossover operation is to exchange part of the gene of two pared chromosome according to some method, and form two new individual. Crossover operation includes single point crossover, multi-point crossover, balanced crossover, and then exchange the gene after that point, the probability of crossover is generally set 0.4-0.99, this paper choose 0.45.

The mechanism of mutation operator is similar to gene mutation, which is one of the methods to produce new individual. This paper adopts the basic mutation operator, to make mutation operation of the value of one or some gene, randomly appointed by the mutation probability of individual coding. Mutation probability is normally 0.0001-0.1, this paper chooses 0.01.

3. BUILDING AND APPLICATION OF THE MODEL

3.1 Selection and Standardization of the Sample Data

This paper selected 640 sample data of Germany personal credit as the study object, in which 340 samples were classified into class 1 (good credit), and 300 samples into class 2 (bad credit). Every sample included 17 attributes indexes. Then we normalized the indexes, and the personal data was classified into continues data and random data, for each of them needed different methods. As to discrete data, min-max normalization method is used to precede the original data, and to make them into the district [0, 1]:

$$x'_{ij} = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}} \qquad (3)$$

Among them, x'_{ij} is the new attribute value of indicator i of sample j, x_{ij} is the origin attribute value of the indicator i of sample j, min x_{ij} is the minimum of all the sample attributes indexes value of indicator i , max x_{ij} is the maximum of all the sample attributes indexes value of indicator.

As to the sample date that this paper used, age, loan amount, and loan period, which were continuous data, by observing their distribution map, they were found to be similar to Gaussian distribution, $x \sim N(\mu,\sigma^2)$, converting them into (0,1) by method of standardization:

$$x' = \Phi(\frac{x-\mu}{\sigma}) \tag{4}$$

Among them, $\Phi(x)$ is the Gaussian distribution. Through this converting, we got to probability according to the indicator.

This paper uses stratified sampling, in order to reduce influence to the model classification brought by the uneven of the data, the number of each kind of sample were equal. Among then, the training sample and the testing sample all included 170 "good" samples and 150 "bad" samples.

3.2 Single Statistical Model

3.2.1 Multiple linear regression

Multiple linear regression has a strict requirement for the distribution of the data, however, when the sample is large, valuation model eliminated the multicollinearity between explanatory variables is efficiently explicable, the variables in the model are of economic importance. To satisfy the requirement that there are no multicollinearity among variables, this paper used gradually stepwise to eliminate effect brought by the multicollinearity between variables. The final explanatory variables were x_5,x_9,x_{10},x_{12},x_{13},x_{14},x_{15}, and the result of linear regression was:

$$\begin{aligned} f_{1t} &= 17.6151 + 0.2499x_{5t} - 0.0273x_{9t} - 77.5211x_{12t} \\ &+ 0.1289x_{13t} - 0.1737x_{14t} + 0.3958x_{15t} \end{aligned} \tag{5}$$

Using the static software to exam the significance of the parameter and the whole function, the result proved their significance.

3.2.2 Logistic regression model

Logistic regression has comparatively advantage in proceeding data, and is robust. The method of variable screen was to backward conditional screen, according to proposed parameter make the likelihood ratio probability test, the obtained Logistic regression function was:

$$\begin{aligned} &1\mathrm{n}\frac{f_{2t}}{1-f_{2t}} = 103.6398 + 0.3667x_4 + 0.4915x_8 + 1.5301x_{10} \\ &- 463.6845x_{12} - 0.7878x_{14} + 1.9918x_{15} \end{aligned} \tag{6}$$

Among them, f_{2t} was the probability of person t of "good" credit.

3.3 Linear Combined Predict Model Based on Minimum Error Square

To compare the valuation effect of combined valuation model based on genic algorithm, first, we built the linear combined model:

$$\begin{aligned} &f_t = w_1 f_{1t} + w_2 f_{2t} \\ &st, w_1 + w_2 = 1 \end{aligned} \tag{7}$$

The weight w_1,w_2 may be negative. The combined predict model based on minimum error square was:

$$f_t = -0.6737 f_{1t} + 1.6737 f_{2t} \tag{8}$$

3.4 Combined Valuation Model Based on Genic Algorithm

Genic algorithm evolved 100 generation, finally we get the weight were 0.0693 and 0.9307, the fitness value now was 0.8977, which meant the accuracy rate of this model in classification was 89.77%, combined valuation model was:

$$f_t = 0.0693 f_{1t} + 0.9307 f_{2t} \tag{9}$$

Among them, f_{1t},f_{2t} were the valuation result of person t in linear regression model and logistic regression model.

Apply the above models in sample test, using 0.5 as Classification boundaries, which meant the value result was higher than 0.5, then the credit was classified "good", the value result was lower than 0.5, the credit was classified "bad"(Ma & Tang, 1998), the classification result was as the following table:

Table 2
Comparison of the Classification Results of Single Model and Combined Valuation Model

Model	Actual value	Assessed value		Sample number	Accuracy	Error rate
		1 (good)	0 (bad)			
Multiple linear regression model	1 (good)	153	17	170	90.00%	10.00%
	0 (bad)	26	124	150	82.67%	17.33%
		Total classification percentage		320	86.57%	13.43%
Logistic regression model	1 (good)	155	15	170	91.18%	8.82%
	0 (bad)	20	131	150	86.67%	13.33%
		Total classification percentage		320	89.07%	10.93%
Combined valuation model based on minimum error square	1 (good)	155	15	170	91.18%	8.82%
	0 (bad)	17	133	150	88.69%	11.33%
		Total classification percentage		320	90.00%	10.00%
Combined valuation model based on genic algorithm	1 (good)	157	13	170	92.36%	7.64%
	0 (bad)	16	134	150	89.34%	10.66%
		Total classification percentage		320	90.94%	9.06%

CONCLUSION

Compare the classification result of combined model and single model. Table 2 shows that, both the two of the combined model were better than single model in general classification accuracy, and they had improvement in reducing false accept rate, which meant that the combined valuation model were superior than single model in classification accuracy.

Compare the combined model based on minimum error square and the combined model based on genic algorithm. Table 2 showed that the latter were better than the former, in both accuracy, and the Error Type I (reject the truth) and Error type II (accept the false), especially in the Error type II.

Generally speaking, regarding to the classification, in personal credit classification, the combined model was better than the single model, especially in avoiding the Error type II (accept the false). The combined valuation model based on genic algorithm was better than the combined valuation model based on minimum error square.

REFERENCES

Chen, H. Y., & Xu, Y. S. (2000). The estimation of weight of combined prediction and its significance test. *Operations Research and Management Science*, (2), 75-78.

Fogarty, T. C., & Ireson, N. I. (1993). Evolving Bayesian classifiers for credit control: A comparison with other machine learning methods. *Proceedings of the 3rd IMA Conference on Credit Scoring and Credit Control* (*Vol.5*, pp.63-76).

Huang, H. Z., Zhou, Z. F., & Yu, J. K. (2010). ILMBP neural network model and its application in personal credit valuation. *Managerialist*, (10).

Jiang, M. H. (2006). *Research of the combination forecast method of individual credit evaluation for commercial bank* (Doctoral dissertation). Retrieved from CNKI (http://cdmd.cnki.com.cn/Article/CDMD-10213-2007040225.htm). (In Chinese).

Jiang, M. H., & Chen, Y. F. (2006). Combining forecasts of personal credit scoring based on RBF neural network. *Journal of Harbin Engineering University*, *27*(z1). (In Chinese).

Lei, Y. J., Zhang, S. W., & Li, X. W. (2004). *MATLAB genetic algorithm toolbox and its application*. Xi'an: Xidian University Press. (In Chinese).

Liang, S. D., Guo, B., Li, Y., & Fang, Z. B. (2002). Comparative analysis of credit risk model. *China Management Science*, *10*(1), 17-22. (In Chinese).

Ma, Y. K., & Tang, X. W. (1998). Research of the optimization of linear combined prediction model. *Systems Engineering-Theory & Practice*, (9), 110-115. (In Chinese).

Shi, Q. Y., & Jin, Y. H. (2004). Comparative research of the application of multiple personal credit rating model in China. *Statistic Research*, (6), 43-47. (In Chinese).

Zhou, Z. F., et al. (2010). *The study of the evolution mechanism and valuation method of emerging technology enterprise*. Beijing: Science Press.

Zhu, X. D., & Feng, T. J. (2003). Personal credit valuation based on GA neural network. *Systems Engineering-Theory & Practice*, (12), 34-38. (In Chinese).