Scientific Research Publishing

Open Journal of Business and Management, 2017, 5, 244-252
http://www.scirp.org/journal/ojbm
ISSN Online: 2329-3292
ISSN Print: 2329-3284

Research on Personal Credit Assessment Based on Neural Network-Logistic Regression Combination Model

Yajuan Huo, Huazhou Chen*, Jiechao Chen

College of Science, Guilin University of Technology, Guilin, China
Email: *hzchengut@foxmail.com

How to cite this paper: Huo, Y.J., Chen, H.Z. and Chen, J.C. (2017) Research on Personal Credit Assessment Based on Neural Network-Logistic Regression Combination Model. *Open Journal of Business and Management*, **5**, 244-252.
https://doi.org/10.4236/ojbm.2017.52022

Received: March 19, 2017
Accepted: April 9, 2017
Published: April 14, 2017

Abstract

With the development of economic globalization and financial liberalization, credit assessment plays an important role in maintaining the normal relationship of social economy. Personal credit assessment requires establishing calibration models with statistic methods. The mono-method-based models are not capable to simultaneously hold the robustness, interpretation and prediction accuracy of the models. In this paper, back-propagation neural network (BPNN) was used to generate a new comprehensive variable for logistic regression (LR) by tuning the number of hidden nodes. The optimal back-propagation neural network-logistic regression combination model (BPNN-LR) was established with 5 input nodes, 7 hidden nodes and 1 output node. The model performance was slightly improved. The prediction accuracy was raised up to 86.33% and 87.96% for the training samples and the test samples, respectively. Results showed that the BPNN-LR model had higher classification accuracy than the LR model. It is concluded that the outcome performance provides technical reference for the corporation's decision making.

Keywords

Personal Credit Assessment, Back-Propagation Neural Network, Logistic Regression, Information Value, Weight of Evidence

1. Introduction

Credit maintains the normal relationship of social economy. Personal credit is the basis of the social credit. As a result of the imperfect personal credit system, the confusion of personal credit relationship, fraud, repudiation and other dishonesty behaviors prevail universally. This situation has not only become a huge obstacle to the development of market economy. But also has seriously affected

the progress and development of society [1] [2]. Credit assessment is used to evaluate credibility for individuals or corporations. It plays a key role in the development of economic globalization and financial liberalization [3]. It is necessary to establish statistic model for predicting personal credibility. Thus the customers can be discriminated as having good credit or having bad credit [4].

Statistic methods are widely used for the personal credit assessment. Linear discriminant and logistic regression (LR) are the basic robust and interpretative linear methods. They generate a linear scorecard to evaluate customer's credit [5] [6] [7]. But the linear methods meet the difficulties of extracting the nonlinear relationship from data. Therefore, nonlinear machine learning methods are introduced, such as classification decision tree, support vector machine (SVM), neural network (NN), etc. They are feasible to effectively improve the predictive performance of the model [8] [9] [10]. However, the mono-method-based models cannot take into consideration all demands of model robustness, prediction accuracy and variables' interpretation. Therefore, combination optimization of multiple methods has become an important developing for credit assessment [11]. Logistic regression-support vector machine combination (LR-SVM) model could extract significant variables, simplify the model structure, reduce the computational complexity and improve the model efficiency [12]. Neural network combined with multivariable linear regression model could optimize the initial solution of the neural network, thus to shorten the calculation time consumption and to improve the classification accuracy [13]. In this paper, the personal credit assessment models were established by combining back-propagation neural network (BPNN) and logistic regression, to estimate the customers' behaviors for a communication corporation.

The records of customers' behaviors (January-June, 2015) were collected from the transaction data of China Unicom. The numerical variables were transformed using normalization and classification variables are transformed to the binary classification form. The contribution of each variable was evaluated by IV and some of the most contributive valuables were selected for model calibration and prediction. The selected transformed variables were the basic data for model optimization. First, the LR models were established to calculate the probability of being bad credit for each customer, in order to discriminate that this customer has good credit or bad credit. Then, a BPNN model with one hidden layer was used to generate a new comprehensive variable for model optimization by tuning and selecting the number of hidden nodes. Thus, an optimal back-propagation neural network-logistic regression combination model (BPNN-LR) was determined for the evaluation of customer's personal credit, so that the LR classification accuracy was improved. In our study, Python and Matlab are combined used for sample classification, data transformation and model optimization.

2. Data and Methods

2.1. Experimental Data

A total of 4518 customer behavior records were collected from transaction data

of China Unicom. These records from January to June 2015 were scrambled as samples for establishing statistic models. And each sample included 24 variables (as shown in Table 1). One of these valuables was used to evaluate customer's personal credit, by which the customers could be discriminated as having good credit or having bad credit. The customers who had good credit were labeled as "good" customers and the customers who had bad credit were labeled as "bad" customers, so that this valuable was defined as "customer credibility". The remaining 23 variables described customers' daily communication behaviors. They were regarded as the characteristic variables in the modeling processes. As the customer credibility recording, there were 2574 actual "good" customers(valued as 0), and 1944 actual "bad" customers (valued as 1). To achieve the modeling and predicting tasks, we randomly divided all of the 4518 samples into two parts at around the ratio of 3:1. One part was used to train the model, containing 3388 samples as the training set, and the other part containing 1130 samples as the test set.

2.2. Data Transformation

Establishing personal credit assessment model was used to predict the value of customer credibility (Y) with 23 characteristic variables, so that the customers could be discriminated as "good" or "bad" customers. The 23 characteristic variables included 16 numerical variables (T) and 7 classification variables (C). It was necessary to normalize the numerical variables, because they were acquired

Table 1. The variables.

Classification variable		Numerical variable			
Variable	Symbol	Variable	Symbol	Variable	Symbol
Customer credibility	Y	Times of payment arrearage	T1	Totalnumber of communicational friends	T9
IS converged with other service	C1	Amount of payment arrearage	T2	Number of intranet communicational friends	T10
IS real-name registered	C2	Online duration	T3	Number of outer communicational friends	T11
IS bound with bankcards	C3	Accumulative days of communication	T4	Accumulation of used data	T12
IS an attractive number	C4	Times of accumulative calls	T5	Monthly average of the on-bill amounts	T13
IS registered to APP's	C5	Monthly average of arrearages	T6	Monthly average of payments	T14
IS a grouped number	C6	Accumulative callduration	T7	Accumulation of overdue payment arrearages	T15
On/off line	C7	Times of being suspended	T8	Times of international roaming	T16

using different measurement scales and valued in different algebraic ranges. While the seven classification variables had different number of classification target. For example, the values of "IS converged with other service" (C1) included 0, 1, 2 and 3. This valuing diversity was probable to result in high dimensional calculation. In order to reduce model's complexity, they were transformed to binary classification variable, equaling to either 0 or 1.

2.3. Variable Selection

The contributions of 23 characteristic variables were different. The calibration model could be much complex and the prediction accuracy be affected if all the 23 variables were used for personal credit assessment. Therefore, it was necessary to select characteristic variables according to their contribution. The contributions of variables are quantitatively measured by calculating the Weight of Evidence (WOE) for divided sample groups. WOE is used to measure the difference between variables and between different sample groups [14]. The value of WOE_i for the i-th group was calculated as follows,

$$\mathrm{WOE}_i = \ln\left(\frac{G_i/G}{B_i/B}\right)$$

where G represented the "good" customers. B represented the "bad" customers. The Information Value (IV) of each variable is successively defined by weighted sum of the grouping WOE's,

$$IV = \sum_{i=1}^{n} \mathrm{WOE}_i \times \left(\frac{G_i}{G} - \frac{B_i}{B}\right)$$

A larger IV represent the variable contributes greater to the models [15].

2.4. The Logistic Regression Model

Logistic regression is a common classification model in machine learning field. The personal credit was estimated by establishing logistic regression models with the selected transformed variables. To achieve the calibrating and predicting tasks, the customer credibility (Y) was defined as the dependent variable. And the selected variables were defined as the independent variables (labeled as X_1, X_2, ···, X_n). The logistic regression formula was as follows,

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where, P was the probability of a customer being "bad", *i.e.* $P = P(\mathrm{Y} = 1)$. The P-value resulting from logistic regression model was used to distinguish "good" customers and "bad" customers.

2.5. The BPNN-LR Model

The BPNN-LR model was to utilize back-propagation neural network to train a new comprehensive valuable for logistic regression. The neural network was formed as having one hidden layer with k hidden neurons. The parameters (the

weights) of the whole network should be trained by iteration.

Step 1, the weights of any neuron between the input and the hidden layer (V_{ij}) were preliminarily initialized. The selected characteristic variables X_1, X_2, ···, X_n were input and the weighted sum of them was passed to the hidden layer by logistic function. The output of the hidden layer was defined as follows,

$$H_j = f\left(\sum_{i=0}^{n} V_{ij} X_i\right), j = 1, 2, \cdots, k$$

Step 2, the weights of any neuron between the hidden and the output layer (W_{jp}) were initialized. The output of the hidden layer were weighted and summed, then transferred to the output layer. The final output of the neural network was as follows,

$$Y_p = f\left(\sum_{j=0}^{k} W_{jp} H_j\right), p = 1, 2, \cdots, m$$

where, the logistic function was used as the transfer function *f*, V_{ij} was defined as the weight of the *i*-th input node and the *j*-th hidden node and W_{jp} was defined as weight of the *j*-th hidden node and the *p*-th output. For each weighted sum, the *i* and *j* equaling to 0 meant that V_{0j} and W_{0p} were the alternatives of the thresholds of the neurons when X_0 equaled to 1 and H_0 equaled to 1.

The output of BP neural network was regarded as a new comprehensive variable added to logistic regression, so that the predictions of customer credibility of all samples were accomplished. According to the predictive results, the optimal number of hidden neurons was selected for the improvement of the BPNN-LR model and the "good/bad" customers identified.

3. Results and Discussions

A total of 4518 samples were randomly divided into training set and test set at around the ratio of 3:1, containing 3388 training samples and 1130 test samples. Each sample included 24 variables. One of these variables was customer credibility (Y), by which the customers could be discriminated as "good" (valued as 0) or "bad" (valued as 1). We defined the customer credibility as the dependent variable and the 23 characteristic variables as the independent variables to establish LR and BPNN-LR models for personal credit assessment, so that the customers could be distinguished as "good" or "bad".

In order to reduce the complexity of the model, it was necessary to select variables according to their contribution (*i.e.* the IV values). A larger IV represent the variable contributes greater to the models. The variables were sorted in descending order of IV (shown in **Figure 1**). Several variables at the top of the IV chart were expected to produce most contribution. For example, "On/off line" (C7) was at the top of the figure, with its IV equaling to 2.77. The first runners-up were "Times of payment arrearage" (T1), "Amount of payment arrearage" (T2), "Online duration" (T3) and "Accumulative days of communication" (T4). Each of them contributed much less than C7. As for dimension reduction, we established the self-adaptive models by tuning the different numbers of variables (from 4 to 10). The prediction accuracy of different number of selected

variables were shown in Figure 2. The prediction accuracy of using 5 variables was slightly better than using four variables, but almost the same as using more-than-five variables. Thus the top five variables with the largest IV values were selected to establish an optimal model. The 5 selected variables were: On/off line (C7), Times of payment arrearage (T1), Amount of payment arrearage (T2), online duration (T3) and Accumulative days of communication (T4).

The logistic regression method was used to establish the personal credit assessment model to distinguish "good" customers and "bad" customers by the P-value resulting from the model. The training samples with the most contributive characteristic valuables were input to the model to predict the P-values of the test samples. If P was greater than 0.5, the sample was classified as "bad" customers (Y = 1), otherwise it was classified as "good" customers (Y = 0).

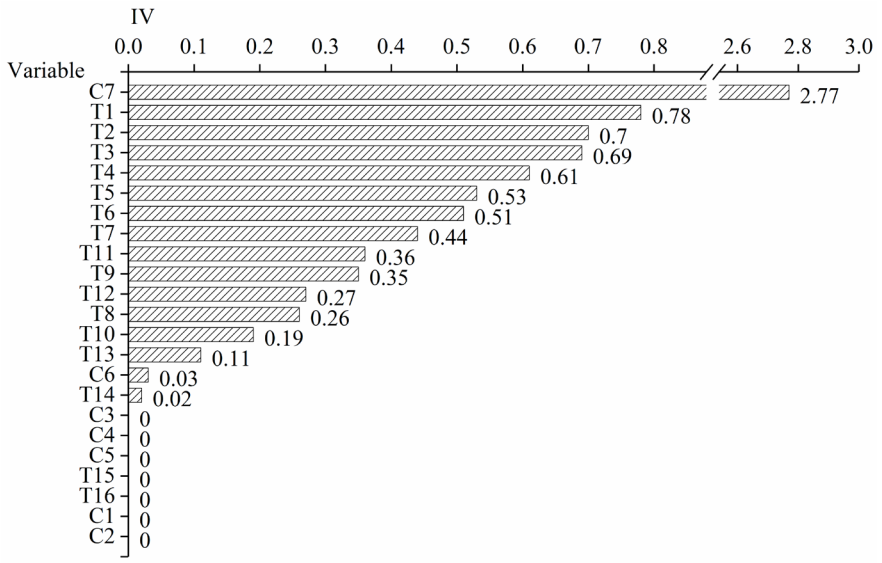

Figure 1. The IV values of the 23 characteristic variables in descending order.

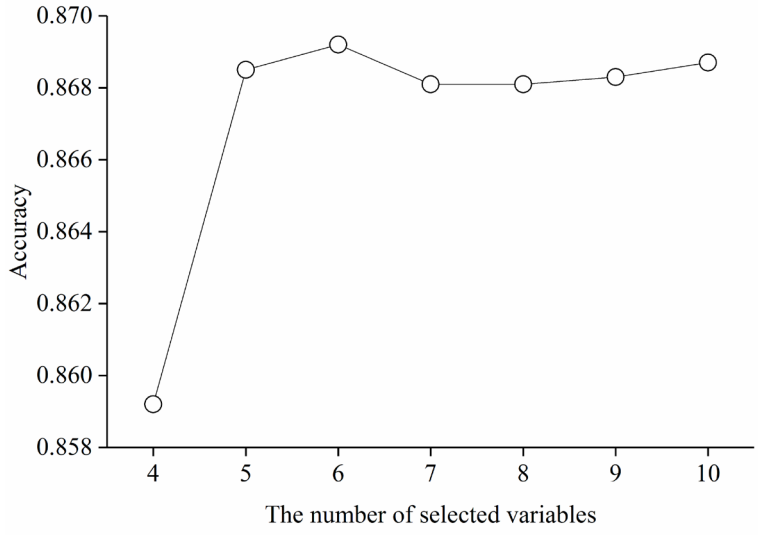

Figure 2. Self-adaptive prediction accuracy of different number of selected variables.

Results of the logistic regression model were shown in Table 2. The classification accuracy of the training samples was 86.25% and the classification accuracy of the test samples was 84.87%.

In order to improve the predictive performance of the logistic regression model, the BPNN method was used to generate a new comprehensive valuable for logistic regression model. And a BPNN-LR model was established. To select the number of BPNN's hidden neurons (k), we established the models with tuning the number of hidden neurons (from 4 to 10). Initializing the weights between the input and the hidden layer (V_{ij}), and between the hidden and the output layer (W_{j1}), the logistic function was used as the transfer function. Iterative training the parameters (the weights) of the whole network, the classification accuracy of the test samples corresponding to different k values were shown in Figure 3. As shown in Figure 3, the classification accuracy of the BPNN-LR model was superior to the logistic regression model. And the optimal number of hidden neurons was 7. Therefore, the BP neural network included 7 hidden nodes, 5 input nodes (*i.e.* The selected 5 most contributive characteristic variables) and 1 output node (the customer credibility). The optimal BPNN-LR model was optimized with the output neuron having the weights of $W_{01} = -4.14$, $W_{11} = -3.48$, $W_{21} = -0.71$, $W_{31} = 0.34$, $W_{41} = 6.81$, $W_{51} = -3.04$, $W_{61} = -3.37$ and $W_{71} = 5.33$. Results of the BPNN-LR model were shown in Table 3. The classification accuracy of the training samples was 86.33%. And the classification accuracy of the test samples was 87.96%.

4. Conclusion

In this paper, the customer transaction data from China Unicom were used to establish personal credit assessment model. These data were from January to

Table 2. The confusion matrix for the binary classification model based on logistic regression.

Training set		Prediction		Accuracy (%)	Test set		Prediction		Accuracy (%)
		0	1				0	1	
Reality	0	1784	129	93.25	Reality	0	572	89	86.53
	1	337	1138	77.15		1	82	387	84.00
Total accuracy (%)			86.25		Total accuracy (%)			84.87	

Table 3. The confusion matrix for the binary classification model based on BPNN-LR model.

Training set		Prediction		Accuracy (%)	Test set		Prediction		Accuracy (%)
		0	1				0	1	
Reality	0	1789	124	93.52	Reality	0	630	31	95.31
	1	339	1136	77.02		1	105	364	77.61
Total accuracy (%)			86.33		Total accuracy (%)			87.96	

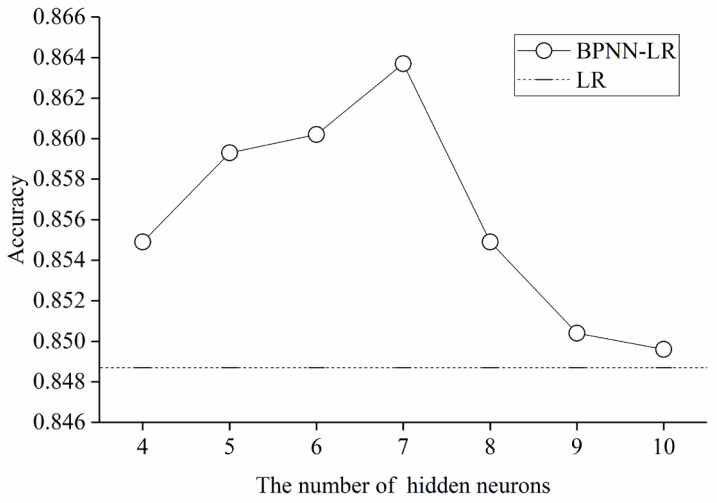

Figure 3. The prediction accuracy of different number of hidden neurons.

June 2015. The LR model and the BPNN-LR model were respectively established to discriminate between "good" customers and "bad" customers. First of all, the numerical variables were normalized and classification variables were transformed to binary classification variables. Then, variables were selected according to their contribution measured by calculating the IV. The selected five variables with the largest IV were used to establish the model and predict. The P-value resulting from the logistic regression model was used to distinguish "good" customers and "bad" customers. The accuracy of the training samples was 86.25%. And the accuracy of the test samples was 84.87%. In order to optimize the model, BPNN method was used to generate a new comprehensive valuable for logistic regression model. And a BPNN-LR model was established. The optimal BPNN included 5 input nodes, 7 hidden nodes and 1 output node. The classification accuracy of the training samples was 86.33%. And the classification accuracy of the test samples was 87.96%. The results showed that the classification accuracy of the BPNN-LR model was higher than the LR model. This indicated that the new comprehensive valuable generated by BPNN was feasible to highly interpret the characteristic valuables. The method of BPNN combined with LR modeling provided technical references to distinguish "good" customers and "bad" customers and provide decision-making basis for a corporation.

Acknowledgements

This work was supported by National Natural Scientific Foundation of China (61505037) and Natural Scientific Foundations of Guangxi (2016GXNSFBA-380077, 2015GXNSFBA139259).

References

[1] Fang, K.N., Wu, J. and Zhu, J. (2010) Forecasting of Credit Card Credit Risk Under Asymmetric Information Based on Nonparametric Random Forests. *Economic Research Journal*, **39**, 97-107.

[2] Liu, L. (2007) Retail Exposures Credit Scoring Models for Chinese Commercial Banks: Model Specification and its Application. *Journal of Finance and Economics*, **33**, 26-36.

[3] Wang, Y., Nie, G. and Shi, Y. (2012) A Research on Customers Default Rate of Commercial Banks in China Based on Credit Scoring Models. *Management Review*, **24**, 80-89.

[4] Guo, Y., Zhou, W. and Luo, C. (2015) Instance-Based Credit Risk Assessment for Investment Decisions in P2P Lending. *European Journal of Operational Research*, **24**, 417-426.

[5] Shi, Q. and Jin, Y. (2004) A Comparative Study on The Application of Various Personal Credit Scoring Models in China. *Statistical Research*, **21**, 43-47.

[6] Miyamoto, M. (2014) Credit Risk Assessment for a Small Bank by Using a Multinomial Logistic Regression Model. *International Journal of Finance & Accounting*, **3**, 327-334.

[7] Menard, S. (2002) Applied Logistic Regression Analysis. *Technometrics*, **38**, 184-186.

[8] Yang, S., Zhu, Q. and Cheng, C. (2013) The Building of the Combined Model for Personal Credit Rating: A Study Based on the Decision Tree-Neural Network. *Financial Forum*, **27**, 57-61.

[9] Xiao, W. and Fei, Q. (2006) A Study of Personal Credit Scoring Models on Support Vector Machine with Optimal Choice of Kernel Function Parameters. *Systems Engineering Theory & Practice*, **26**, 73-79.

[10] Galindo, J. and Tamayo, P. (2006) Evaluation of Neural Networks and Data Mining Methods on a Credit Assessment Task for Class Imbalance Problem. *Nonlinear Analysis*: *Real World Applications*, **7**, 720-747.

[11] Salehi, M. and Mansoury, A. (2011) An Evaluation of Iranian Banking System Credit Risk: Neural Network And Logistic Regression Approach. *International Journal of Physical Sciences*, **6**, 6082-6090.

[12] Lu, H., Li, Y. and Hong, W. (2013) Credit Scoring Model Hybridizing Artificial Intelligence With Logistic Regression. *Journal of Networks*, **8**, 253-261.

[13] Khashei, M. and Bijari, M.S. (2012) A Novel Hybrid Classification Model of Artificial Neural Networks and Multiple Linear Regression Models. *Expert Systems with Applications*, **39**, 2606-2620.

[14] Zheng, L. (2002) Commercial Bank Credit Risk Management. China Renmin University Press, Beijing.

[15] Galindo, J. and Tamayo, P. (2000) Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*, **15**, 107-143. https://doi.org/10.1023/A:1008699112516