An Empirical Study on Prediction of the Default Risk on P2P Lending Platform

An Empirical Study on Prediction of the Default Risk on P2P Lending Platform

Meng Qian and Fangqin Hu*

School of computer and imformation, Anqing Normal University, Anqing, China

*Corresponding author e-mail: 1065681918@qq.com

Abstract: P2P lending platform contains many risks of loan default. The top priority of P2P lending platform is to predict the risk of default accurately and to take corresponding measures. The paper selects public lending loan transaction data of users from *Lending Club* to carry on the empirical study on the risks of loan default by respectively using logistic regression model, the bat optimization algorithm of feedforward(BAPA) neural network and least squares support vector machine(LSSVM) to analyze the experimental data, and then to evaluate the applicability of three methods in predicting the risk of loan default on the P2P network platform. The experimental results show that the least squares support vector machine has a better prediction effect.

1. Introduction

Peer-to-peer lending, also abbreviated as P2P Lending, is the practice of directly lending money to individuals or businesses through the online service that match lenders with borrowers. The P2P online lending platform has become a new type of financial management for investors with its high yield, which has met the needs of the borrowers to some extent. On the other hand, the high interest rate of loan, the lack of sound credit system and the imperfect supervision system and information asymmetry pose certain risks to investors and online lending platforms. In recent years, the number and incidence of closed and problematic platforms have been rising, which makes it particularly important to correctly predict the default risks as it not only involves investors' income but also affects the development prospects of the online loan industry to a certain extent.

In recent years, experts and scholars have used the method of machine learning in the risk prediction and already have made some achievements. Jingrui Bai proposed that the combination of neural network and logistic regression can be better applied in P2P credit field [1]. Yu Yuan, based on personal financial indexes, used logistic model to conduct modeling analysis of Prosper credit platform, which proves that this method has a significant accuracy [2]. Qing Zhang made use of the support vector machine model to analyze the personal credit data from the bank and provided references for the risk examination and modeling control for P2P projects [3]. Yanming Fu found that non-linear models, such as support vector machine, have practical potential in credit prediction [4].

However, many related research are mainly from the perspective of policy. The paper uses open date from foreign website for reference, respectively using logistic regression models, bat optimization algorithm of feedforward neural network and least squares support vector machine to predict and

analyze, and to validate the applicability of these three methods in prediction of default risk so as to find out the optimal prediction model among the three methods to provide certain reference about default risk prediction and the construction of personal credit system based on the future perfect lending data for our country.

2. Data processing

In order to ensure the validity and preciseness of the empirical study, the original experimental data is processed and implemented by Python software.

2.1 Processing of missing values

Due to its relatively big original sample data and low proportion of missing value, direct delete method is used in the processing of the missing value to directly delete irrelevant variable indexes and indexes with more null values. After processing, there are 495,717 loan data, 52 indexes.

2.2 Assignment of classified indexes

This article selects loan-status index as the dependent variable, among these indexes, the "Current" field, which represents borrowing period, has not yet expired, this kind of sample data does not belong to both categories, so the data is deleted, ending up with 83641 data. The assignment is shown in Table 1 below, so that the default risk prediction is reduced to a two-category problem.

Table 1. assignment table of the dependent variable index

Field	Assignment	Sample Size	Proportion
Fully Paid	0	59706	71.38%
Charged off		13592	16.25%
Default		12	0.01%
In Grace Period	1	4658	5.57%
Late(16-30 days)		2568	3.07%
Late(31-120 days)		8866	10.60%

In addition to the above dependent variable loan-status index, eight variable indexes in the sample data, which are converted into numerical classification variables after the classification are shown in Table 2 below.

Table 2. index assignment

Index	Type	Assignment
Term	36 months	1
	60 months	2
Grade	A	1
	B	2
	C	3
	D	4
	E	5
	F	6
	G	7
Sub-grade	A1-A5	1-5
	B1-B5	6-10
	C1-C5	11-15
	D1-D5	16-20
	E1-E5	21-25
	F1-F5	26-30
	G1-G5	31-35
Emp-length	<1 year	1
	1 years-10years	2-11
	10+ years	12
Home Ownership	Own	1
	Mortgage	2
	Rent	3
Verification Status	Verified	1
	Source verified	2
	Not verified	3
Purpose		1
	Credit card	2
	Debt consolidation	3
	Home improvement	4
	House	5
	Major purchase	6
	Medical	7
	Moving	8
	Other	9
	Renewable energy	10
	Small business	11
	Vacation	12
Application Type	Individual	1
	Direct-pay	2
	Joint	3

2.3 Data normalization processing

The aim of improving the classification precision of the algorithm can be achieved by date normalization, by which the statistical probability distribution of data is classified into the range of [-1, 1]. In this paper, the maximum minimum method is adopted to standardize the data, and the function is as follows:

$$x_k = \frac{x_k - x_{\min}}{x_{\max} - x_{\min}}$$

$x_{\min}$ refers to minimum value in the data sequence; $x_{\max}$ refers to the maximum in the sequence. The normalized function adopts the function mapminmax in MATLAB

2.4 Dimension reduction

To make the dimension reduction of the sample meet the standards of modeling, this paper adopts the method of Principal Component Analysis to carry out dimension on the independent variables, lowering dimension of the new input variable and lessening the correlation between components. The results are shown in Table 3.

Table 3. variance and principal component contribution rate

Component	Eigenvector		
	eigenvalue	Variance contribution rate(%)	Accumulative contribution rate(%)
1	13.026	25.051	25.051
2	6.061	11.655	36.706
3	5.487	10.553	47.258
4	3.78	7.268	54.527
5	3.065	5.894	60.421
6	2.433	4.678	65.099
7	2.222	4.274	69.373
8	2.022	3.889	73.261
9	1.933	3.717	76.979
10	1.655	3.183	80.161
11	1.402	2.697	82.858
12	1.135	2.183	85.041
13	1.016	1.953	86.994

About principal component selection, firstly, defining the limit of the cumulative percentage of the variance, which can be in the range of 75% to 95%. Then check the scree plot of variance relative to the number of components, and select the point of the graph approximation level. By using the method of cross-validation to select of principal components that can minimize the verification error and with the observation of Table 3 and Figure 1, 2 principal components are chosen finally. By using the corresponding eigenvectors of eigenvalues of these principal components, namely after transformation, the comprehensive indexes are standardized as the input into the neural network prediction model.

3. Experiments about prediction of default risk

3.1 Prediction of default risk based on logistic regression model

As shown in table 4 below, without any independent variables, the accuracy rate of default prediction is only 65.9%, which is obviously not expected and acceptable to the online loan platforms.

Table 4. accuracy of initial classification

Initial classification		Prediction		Accuracy rate(%)
		0	1	
Observation	0	43168	0	100
	1	22367	0	0
Total				65.9

The following table 5 is a comprehensive test of logistic regression model coefficient. From the following table, the significance value of the regression model is less than 0.05, which is statistically significant.

Table 5. comprehensive test of logistic regression model coefficient

		Chi-square	df	Sig.
Step1	Step	54540.14	9	0.00
	Block	54540.14	9	0.00
	Model	54540.14	9	0.00

The estimated results of the regression model can predict whether the borrower defaults. The prediction ability of the model can be evaluated by comparing the predicted results with the actual results, which is shown in table 6. It can be seen that logistic regression model is ideal for the application in predicting default risks of borrowers

Table 6. Accuracy of logistic regression model in prediction

Logistic regression model		prediction		Accuracy rate(%)
		0	1	
Observation	0	40185	2982	93.1
	1	3208	19159	85.7
Total				90.6

3.2 Risk prediction based on the BAPA neural network model

3.2.1 Selection of training samples and test samples

The paper adopts the empirical proportion(7:3). Because of the large amount of sample data in this paper (83641), if all the sample data used in BABP neural network, it will greatly increase the time of network training and predicting. And influenced by factors such as the machine configuration, this paper selects the lending data of 5000 borrowers as the training sample, 1500 data as test samples.

3.2.2 Selection of network structure

For the layer, this paper selects a three-layer BP network with only one hidden layer and the activation function of the hidden layer and output layer is the Sigmoid function and the linear function. The best implicit layer node number selection can be referred to the following formula:

$$l < \sqrt{(m+n)} + a$$

$$l = \log_2 n$$

n stands for the number of input layer nodes; *l* stands for the number of hidden layer nodes; *m* stands for the number of output layer nodes; a is the constant of 0-10.

The choice of the number of hidden layer nodes is to use the formula to determine the approximate range of the number of nodes and then the optimal number of nodes is determined by the method of trial. The number of nodes in this paper is set to 7.

3.2.3 Parameter setting of bat optimization algorithm

BABP neural network model has nine input parameters, an output parameter, so BABP neural network structure is set for the 9-7-1, that is 9 input layer nodes, 7 hidden layer nodes, two output layer nodes, a total of 9 * 7 + 7 * 1 = 70 weights, 1 * 7 + 1 = 8 threshold, so the bat algorithm

individual coding length is 70 + 8 = 78. Other parameter settings are shown in Table 7 below.

Table 7. parameter setting of bat optimization algorithm

Parameter	Maximum iteration	Quantity	Alpha	Gama
Value	400	25	0.98	0.98
Parameter	Maximum Frequency	Speed range	Individual size range	
Value	2	[-9,9]	[-5,5]	

And then, on the basis of the training sample, the network is built to predict the users' credit default rates, setting the maximum training step length 400, training target accuracy 0.1, vector 0.0001, BABP neural network learning curve shown in figure 1. The network reached the target accuracy after 13 iterations, and the network fitting squared error (SSE) reached the minimum value of 0.071185.

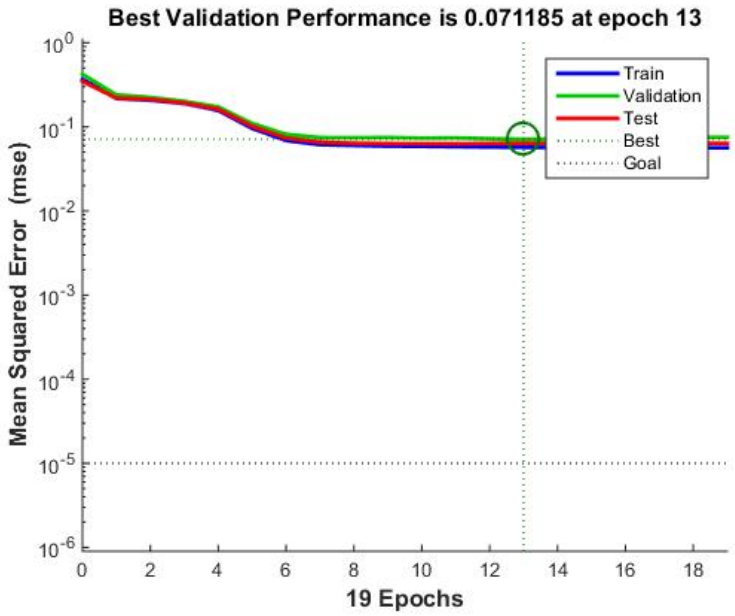

Fig. 1 Training performance of BABP neural network

Figure 2 below is the variation curve of BP network adaptation based on the bat optimization algorithm, and after 230 times iteration, fitness value no longer change and lend to convergence, and the convergence rate is faster. Therefore, the fitting effect of BABP network is preferable.

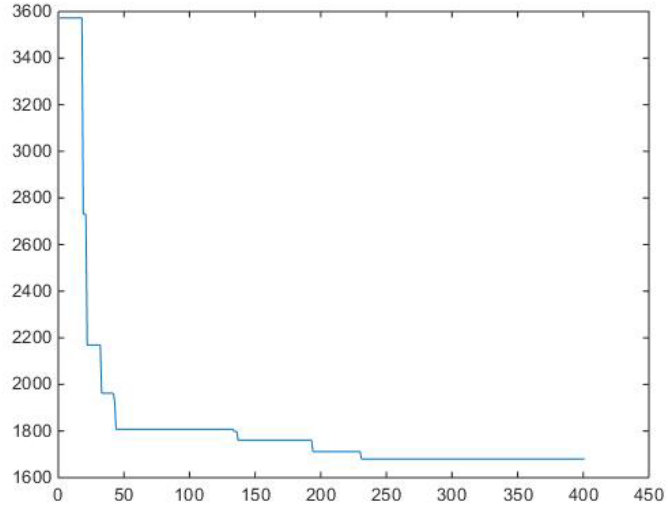

Fig. 2 adaptive evolution of the BABP neural network

The output value of default risk prediction of borrowers based on BABP neural network is a decimal number between 0 and 1, not 0 s and 1 s classification results, in order to calculate prediction accuracy of BABP neural network, this paper assigns the output value [0,0.5) to 0, [0.5, 1] to 1, then classify the output. The result shown in the following table 8, from which can be concluded that the accuracy rate of classification results of prediction of default risk based on BAPA is 91.76%, with only 4.22% error rate of classification of none-default. However, error rate in default category is higher, up to 12.66%. Compared with the logistic regression model, this classification result is satisfactory. Relatively small error rate proves BABP neural network has high feasibility in default risk prediction of borrowers

Table 8. test results of BABP neural network

Type	Outcome	Accuracy number	Error number	Accuracy rate	Error rate	Total accuracy rate
Default	876	830	37	94.75%	4.22%	91.67%
None-default	624	545	79	87.34%	12.66%	

3.3 Prediction of default risk based on LSSVM model

This section uses the same 5000 data of the borrowers as training set and the same 1500 as the test set. Some of the model parameters are also needed to be built. two of the fine-turning parameters γ (gam) can determine the balance between training error minimization and smoothness, square error (sig2). The paper combined the method of Coupled Simulated Annealing(CSA) with standard single-line method to determine two fine tuning parameters, gam = 24.9521, sig2 = 13.0956. In addition, Gaussian RBF kernel is used in the kernel function of LSSVM algorithm

Figure 3 below is the ROC curve of LSSVM classification prediction, which shows that the value under the ROC area is 0.9888 and the standard deviation is 0.0011, proving that accuracy of the default prediction is higher.

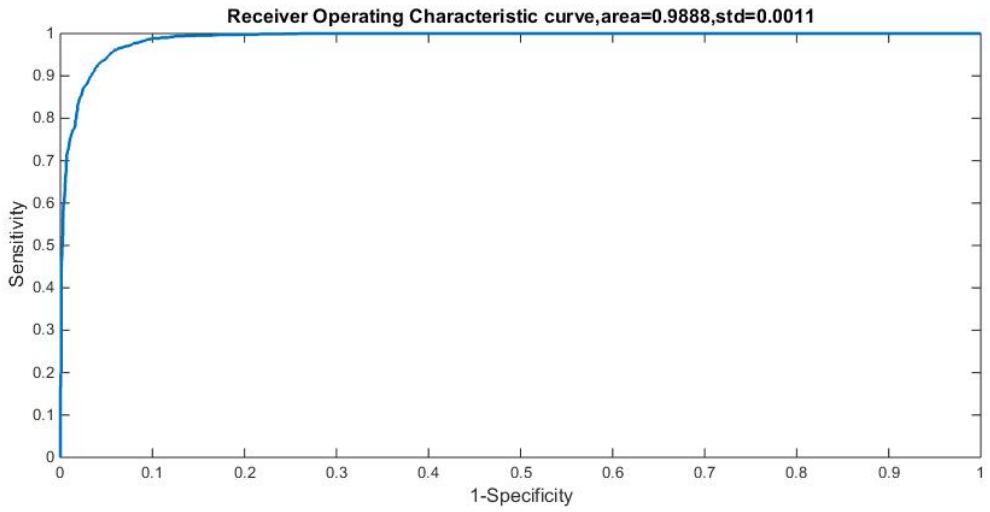

Fig. 3 ROC curve of LSSVM classification prediction

Table below 9 is result of the classification of default prediction based on LSSVM, from which can be seen the accuracy rate of classification prediction reached 92.20%. Compared with logistic regression model and BABP neural network, its accuracy rate is the highest, which proves the high feasibility of LSSVM neural network model in predicting loan default risk of the users. The prediction

error rate of in the non-default classification is only 6.71%, while the error rate in default classification is higher, which is up to 11.24%.

Table 9. summary and evaluation of LSSVM test results

Type	Outcome	Accuracy number	Error number	Accuracy rate	Error rate	Total accuracy
Default	906	849	57	93.71%	6.71%	
None-default	594	534	60	89.90%	11.24%	92.20%

3.4 Analysis of experimental results of three methods

According to the above experiments, the result of one random experiment shows that the classification resultsof LSSVM is relatively more accurate than logistic regression model and BABP neural network. As weights and thresholds are random when BABP neural network is used to predict default rate of borrowers, causing each prediction accuracy is different and different fine-tuning parameters also makes the result of prediction based on LSSVM various. A random experiment is not enough to explain which approach is better. In order to avoid the influence of random error and make the comparison more rigorous, these two methods are randomly run 10 times. Table 10 below is the result of 10 random experiment conducted without random error. The average accuracy of prediction based on LSSVM is 92.17%, slightly higher than that of BABP neural network (91.45%), which shows that the effect least squares support vector machine (LSSVM) is superior to the bat optimization algorithm of BP neural network in prediction of default risk of borrowers on the lending platforms.

Table 10. result of random experiment based on BABP neural network and LSSVM

method	BABP neural network			LSSVM		
	Accuracy rate	Error rate of none-default	Error rate of default	Accuracy rate	Error rate of none-default	Error rate of default
1	91.87	5.03	12.34	92.07	6.21	10.54
2	91.13	5.50	13.54	92.27	6.09	10.22
3	91.93	4.61	12.68	92.27	6.09	10.22
4	91.53	5.26	12.80	92.07	6.21	10.54
5	91.27	5.08	13.72	92.07	6.50	10.14
6	91.53	5.77	12.34	92.40	5.98	10.05
7	91.20	5.90	12.78	92.27	6.19	10.08
8	91.13	5.40	13.51	91.73	6.63	10.76
9	91.27	5.80	12.90	92.40	5.98	10.05
10	91.67	4.63	13.36	92.13	6.30	10.25
Average	91.45	5.30	13.00	92.17	6.22	10.28

4 Conclusion

As P2P lending platforms inevitably have many risks since its birth, it is a long and arduous task to predict the default risk. This paper has carried on the empirical research of default risk of P2P network platform based on a large number of real data and the results show that the method based on LSSVM model is the most effective one among these three methods in predicting the default risk on P2P platforms. However, it is not likely that an absolute optimal approach will be found at this stage to accurately predict the risk of default. Later, these methods can be improved or used in combination in predicting the risk of default to make the effect of prediction closer to the true value and further improve the development of P2P lending platforms in China.

References

[1] Bai Ruijin, “The application of neural network model based on logistic regression in individual credit evaluation”, Inner Mongolia university, 2012.

[2] Yu Yuan, “P2P network loan credit risk prediction ,based on logistic regression,” Shanghai academy of social sciences, 2014.

[3] Qing Zhang and Jiaying Yu, “ A study on the bankruptcy of Hainan Development Bank from the perspective of credit risk, Zhejiang Normal University, XingZhi College, 2014.

[4] Yanming Fu, Dungang Zhang and Yu Qi, “Risk assessment of P2P network loan credit”, Statistics and Decision, 2014, 1002-6487（2014）21-0162-03.

[5] Mengjia Wang, “Credit risk assessment of P2P online loan platform based on Logistic regression model”, Bejing Foreign Studies University, 2015.

[6] Shuo Li, “A study on credit risk measurement of listed companies in China based on Logistic regression model” Tianjin University of finance and economics, 2016.

[7] Qiuyue Zhang,” The Comparative Study of Individual Credit Evaluation Model in P2P Lending”, Statistics and Application, 2017, 6(3), 292-297.