Credit Risk Assessment of P2P Lending Borrowers based on SVM

Wenjing Tao [a, *], Dan Chang [b]
School of Economic and Management, Beijing Jiaotong University, Beijing 100049, China.
[a, *] 16120618@bjtu.edu.cn, [b]18810720705 @163.com

Abstract. With the development of Internet finance, peer to peer online (P2P) lending, which makes a win-win situation between lenders and borrowers, has become one of the most popular means of Internet finance in China. However, problem platforms and borrower default events have also occurred frequently with an explosive-speed growth of P2P online lending. Reducing credit risk of P2P lending borrowers still holds the key to the steady development of P2P online lending platforms. The results show that the SVM model based on cuckoo algorithm to optimize the parameter has a better classification accuracy. This model can be used to judge the potential credit risk of P2P lending borrowers and provides a theoretical basis for the risk management of Internet financial institutions at the same time.

Keywords: P2P Online lending; Credit risk; SVM; Parameter Optimization; Cuckoo Algorithm.

1. Introduction

P2P online lending is an innovative means of Internet finance, which means that borrows are able to get fund directly from investors through P2P lending platforms without collateralizing and investors also can obtain higher returns than other investment means. Compared with traditional private borrowing, P2P lending has the characteristics of disperse risk, high efficiency, low threshold and intuitive transaction, which makes it become a new trend of future financial service development. [1] By the end of 2016, China's P2P online lending turnover has reached up to 3.8 trillion RMB, which has an increase of 138% compared with last year. Since 2018, the online lending industry turnover increases at a speed of 10. 08% monthly, the cumulative turnover has reached 600 billion RMB reach up to now (As shown in Fig. 1). Although China's P2P online lending industry develops rapidly and P2P lending platforms spring up like mushrooms, problem platforms and borrower default events have occurred frequently. Many P2P lending platforms have the difficulty to withdraw cash and some of them even face with liquidate problems. Stepping into the era of big data, P2P lending platforms generate a large number of transaction data which include a variety of borrower information on a daily basis. How to use these data in a reasonable and effective way to obtain useful information and improve risk control ability plays a vital role in steady development of P2P lending platforms. The borrower's credit risk is one of the most critical risks in P2P online lending. Assessing the credit risk of P2P borrowers, which is regarded as an important means to reduce the loss of investors and P2P lending platforms, has been a hotspot topic in academic research.

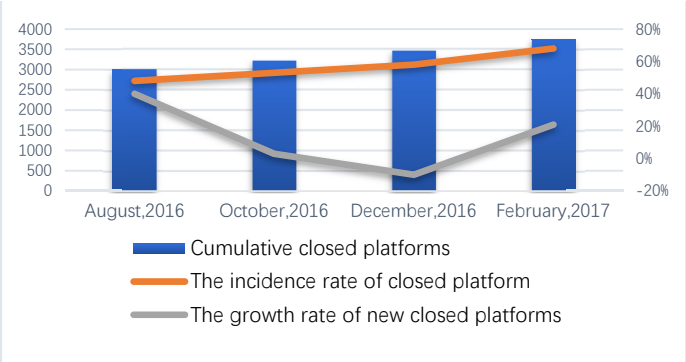

Figure 1. The condition of closed P2P lending platforms.

2. Theoretical Basis

The traditional assessment methods of credit risk tend to rely more on statistical and econometric models. Due to the complex and nonlinear relationship between a series of indicators, it is difficult to establish an accurate classification model to assess the credit risk of P2P lending borrowers by using traditional methods. In recent years, the SVM model based on VC dimension theory and structure-risk minimization principle has been favored in the study of credit risk assessment of P2P borrowers due to the advantage in dealing with high-dimension small samples and non-linear problems. The SVM model is a kind of supervised learning method. The basic idea is to map the sample from input space into the high-dimensional feature space by nonlinear transformation to obtain the optimal classification surface in the feature space, which is linearly separated from the sample. As shown in Fig. 2, the lines H represent the classification line and the lines H_1 and H_2 that separate the samples and are parallel and closest to the classification line. The distance between H_1 and H_2 called classification intervals.

The SVM algorithm converts a complex optimization problem into an inner product operations to the sample. We only need to select appropriate parameters, which is similar to a neural network in form, outputs are linear combinations of S intermediate nodes and each intermediate node is corresponding to a support vector. As shown in Fig. 3.

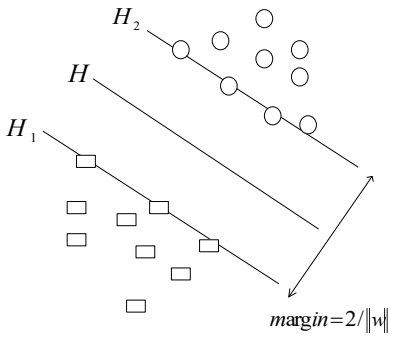

Figure 2. The optimal classification plane of the SVM.

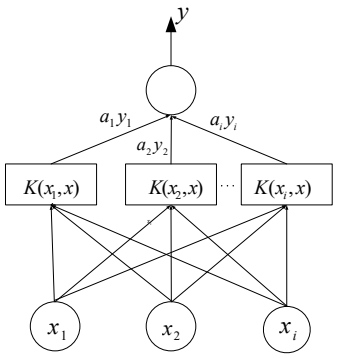

Figure 3. Sketch map of the SVM model.

3. Problem Analysis

P2P lending operation process is shown in Fig. 4, the borrower, the lender and the P2P online lending platform constitute the main body of the online lending. The borrower releases the lending information in the P2P online lending platforms and uses credit rank or property as the guarantee. The Lender through the P2P online lending platform to browse lending information to tender. After trade is successful in both sides, the lender just needs to wait for the borrower to repay the principal and interest on schedule. In the entire transaction process, the P2P online lending platform plays an intermediary role and charges a certain percentage of service fees as a means of making profit.

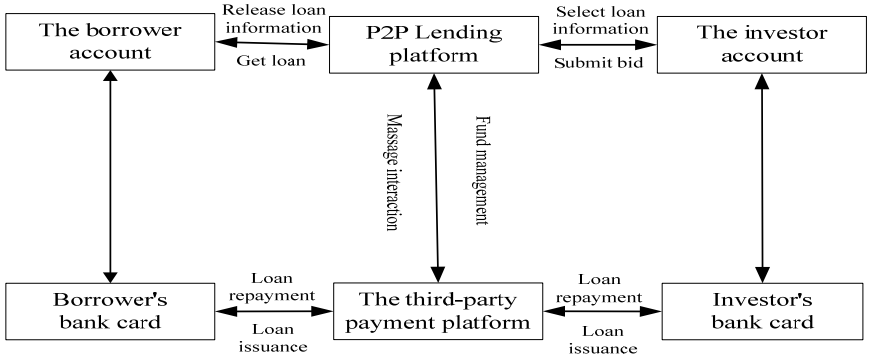

Figure 4. The operation process of P2P online lending

P2P lending industry lacks internal and external supervisions which lead to industry chaos. P2P lending platform as a new means of financial services whose development is in a relatively short time, the relevant laws and regulations are not perfect due to the lack of appropriate legal supervision. This paper mainly analyzes and studies the formation factors of credit risk in the process of lending in the P2P lending platform from the perspective of borrowers. The credit risk assessment of P2P loan borrowers can reduce information asymmetry to some extent to help investors understand the borrower's financial background, the management ability, the credit level, the use of funds and the historical credit record. On the other hand it also play a vital role in improving risk management level of P2P lending platforms, reducing the cost of risk control and strengthening the anti-risk ability of the Internet financial systems.

4. Construct Indicator System

The credit risk assessment of P2P net borrowers is to find the key indicators that can influence their future repayment performance from the borrower's information. This paper constructs a set of credit risk evaluation indicator system of P2P net loan borrowers.

Based on the traditional personal credit evaluation theory FICO, this paper collects some personal basic information of the borrower on the P2P lending platform and takes these as an alternative indicator to evaluate the credit risk of the P2P lending borrower comprehensively. Borrower information is divided into four dimensions, namely, the borrower basic personal information, economic capacity information, loan information, credit history information. As shown in Table 1.

Table 1. The credit evaluation pre selection indicators

Dimension	**Indicator**				
Personal Information	Age	Gender	Marital Status	Educational Level	Location
Economic Status	Industry	Income	Property Status	Vehicle Status	Years of Work
Loan Information	Loan Amount	Loan Interest	Loan Duration		
Credit History Information	Overdue Times				

Since each candidate indicator has different effects on the explanatory variables, it is necessary to select the indicators that have a greater impact on the borrower's credit risk after determining the above-mentioned alternative indicators. This paper will use the way of information gain method to calculate the contribution of each indicator to the explanatory variables, select indicators that has larger impact on the results and delete indicators the has little impact on the results. The information gain values of all the preselected indicators can be calculated, as shown in Table 2.

Table 2. Information gain values of preselected indicators

Indicator	**G (A)**	**Indicator**	**G (A)**
Age	0	Years of Work	0.0081
Gender	0.0263	Property Status	0.0540
Marital Status	0.3521	Vehicle Status	0.0181
Educational Level	0.0306	Loan Amount	0
Location	0.0124	Loan Interest	0.2811
Industry	0.0511	Loan Duration	0.2544
Income Situation	0.0336	Overdue Times	0

4.1 Construction of the Indicator System

The Table 2 shows that information gain values of these indicators, which indicate the value of those indicators (age, loans amount and overdue items) are too small and have little effect on the explanatory variables, that is, these variables have little impact on the credit risk assessment of P2P lending borrowers, so they can be excluded from alternative indicators. In this paper, we further select the evaluation indicator based on the method and then divide the evaluation indicator into qualitative and quantitative indicators according to the Internet financial nature and the usability of the data. The quantitative indicator is represented by its actual value and the qualitative indicator needs to be transformed by the way of converting into the discrete variable, by doing which it's more easily to input data for the SVM model. The final credit evaluation indicator system includes the following 11 indicators. As shown in Table 3.

Table 3. Credit risk assessment indicator system

Indicator	**Indicator Define**				
Gender	Male			Female	
	0			1	
Marital Status	Single		Married		Divorce
	0		1		2
Location	Western Area		Central Area		Eastern Area
	0		1		2
Educational Level	High School or Below		College	Bachelor	Master or Above
	0		1	2	3
Industry	Individual	Private Company	Public Company	Public Institution	Government
	0	1	2	3	4
Income	2-5K	5-10K		10-20K	0-50K
	0	1		2	3
Working	<1 year	1 - 3 years		3 - 5 years	>5 years
	0	1		2	3
Property Status	Homeless	House Mortgage			House-Owner
	0	1			2
Vehicle Status	Carless	Car Mortgage			Car-Owner
	0	1			2
Loan Duration	Actual Value				
Loan Interest	Actual Value				

5. Establish Model

5.1 Determine Optimized Parameters

The different inner kernel functions in the SVM algorithm have different effects on the classification ability of the sample. According to the previous research results, the Redial Bias Kernel can achieve better classification performance than the other three kernel functions, so this paper selects the radial basis function as the kernel function of the SVM model. It's also necessary to select the appropriate parameters for the SVM model to training and test samples. The main parameters are the penalty coefficient C of the SVM model and classification width g of the RBF kernel function. The penalty parameter C is designed to balance the ratio between the empirical risk and the error so as to achieve the optimal generalization capability.

Based on the above-mentioned method, this paper selects training set as initial data samples to obtain the C and g, after which adopts 5-fold cross validation method to verify the classification

accuracy. Finally, the group C and g with the highest classification accuracy of the training set are selected as the best parameter.

5.2 Find the Optimal Parameters

This paper, the Cuckoo search algorithm is used to find the optimal parameters of the SVM model. The steps of the SVM parameter optimization method based on CS are as follows:

Step 1: Input the data set and process the data set. The original data set, the other is used as the test set.

Step 2: Initialize the population M, the initial position of the $nest = (nest_1, nest_2, ..., nest_m)$, in which the position is $nest_i = (C, g)$ for the bird's nest i, the iterations number of the population is N, the probability of being found is $p_a \in [0,1]$.

Step 3: Put the value of each $nest_i = (C, g)$ into the SVM model, which are regarded as the selected parameters, to train the data set and record the fitness value of each nest through the evaluation function. Select the highest fitness as the best bird's nest position (the current optimal position of the population).

Step 4: The position of the nest is updated by $x_{inew} = x_{iold} + a \oplus levy(\lambda)$ and $levy(\lambda) \sim u = t^{-\lambda}, 1 < \lambda < 3$. Calculate the new nest fitness value and compare with the previous generation and retain the bird's nest position that has best fitness value.

Step 5: If $r < p_a$, $r \in [0,1]$, which means that the probability of being found is relatively large, we need to randomly change the location of the nest by Levy flight and calculate the fitness value of the new position. Compared with the previous generation, the nest with high fitness is preserved as the next generation of breeding nest. On the contrary, the location of the nest unchanged.

Step 6: Re-select the current optimal position of the population according to the fitness value.

Step 7: If $j < N$, skip to step 4 to continue the iteration update. The j represents the number of current iterations. If the model has reached the stop condition, output the current optimal position in step 6.

Step 8: The optimal solution output is substituted into the SVM model to classify the test set. Regarding the classification accuracy obtained as the criterion for verifying the validity of the selected parameter. Set Q as the maximum number of iterations, set p as the current iteration number. If $p \leq Q$ then p=p+1 and goes to step 3; if p>Q, the operation will be terminated and the optimal solution will be output.

Step 9: Validation. The output of the optimal solution is substituted into the support vector machine to classify the test set, and the classification accuracy is obtained as the criterion of the validity of the selected parameters.

5.3 Evaluation Criteria

This paper selects three evaluation criteria commonly used in personal credit risk assessment areas as a measure of the quality of the SVM model. The three principles including the overall classification accuracy, the classification accuracy of default samples, the classification accuracy of normal samples, which are defined as follows:

- The overall classification accuracy = correct classification samples / total samples;
- The classification accuracy of the default sample = correct classification samples of the default sample / default samples ;
- The classification accuracy of normal sample = correct classification samples of the normal samples / normal samples.

The overall classification accuracy reflects the performance of the model on the whole test sample. However, considering the imbalance problem of the sample data, it is necessary to further consider the classification accuracy of the default sample and the classification accuracy of the normal sample. A well-behaved model requires a relatively high classification accuracy in both the default sample and the normal sample.

6. Empirical Research

6.1 Data Preparation

"Renrendai" is one of the leading P2P lending platforms. In this paper, through the data crawler software to collect 200 samples during the March, 2018, each sample has 14 attributes and is divided into two categories (normal samples and default samples). Among them, there are 100 samples of normal and default samples respectively. The normal sample category is expressed in 0 and the default sample category is expressed in 1. The intercepted part of the sample data is shown in the following Table 4.

Table 4. Screenshot of fractional data

Gender	0	1	1	0	…	0	0	0	0
Marital Status	0	1	1	1	…	2	0	2	0
Educational Level	3	2	2	2	…	1	0	1	0
Location	2	2	2	2	…	2	2	0	2
Industry	1	2	1	1	…	2	2	2	2
Income	2	2	1	2	…	2	2	0	0
Property Status	1	0	2	1	…	2	0	2	0
Vehicle Status	0	0	2	0	…	0	0	0	0
Years of Work	3	1	3	3	…	0	0	2	3
Loan Interest	10.8	10.8	10.8	10.8	…	13	12	13	12
Loan Duration	36	36	36	48	…	24	12	24	12

6.2 Data Preprocessing

This experiment is mainly to use MATLAB r2012b LIBSVM toolbox to complete. LIBSVM is an effective SVM pattern recognition and regression toolkit that is easy to use.Due to the different range between the lending samples, it is necessary to normalize the data. Assume $\min_A$ and $\max_A$ is the minimum and maximum values in the sample respectively. Using the maximum and minimum method, the normalized formula is as follows:

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

In order to better reflect the learning ability and the generalization ability for small sample of the SVM model, the 200 samples are divided into five groups, each group includes 40 samples. The proportion of default samples and normal samples in each group was 1: 1. The number of normal samples and default samples is 20 respectively. Then select 4 groups as the training set, 1 group as the test set. That is, 1-80 of normal samples, 1-80 of default samples as the training set, 81-100of normal sample, 81-100 of negative samples as the test set. The former is mainly used to establish the credit risk assessment model of P2P loan borrower. The latter is used to verify the classification accuracy of the model and the experimental data are verified according to the two parts. The accuracy of the training set and the classification of the test set were observed by 5-fold cross validation.

6.3 Parameter Optimization Process

According to the above analysis, the sample set (X, Y) is constructed, in which the dimension of X is 11 and Y is the class attribute of the sample. For the "normal customer" Y=0, for the "default" customer Y=1. To the selection parameters of the SVM model is one of the key factors to model performance. The inappropriate choice will lead to over fitting or under fitting. The following three optimization methods are used to optimize the SVM penalty coefficient C and the parameter g of the radial basis function. The parameter optimization process is as follows:

There are two lines in the picture, the best fitness curve with solid dots represents the change of the best fitness value in each generation, the average fitness curve with hollow dots represents the change of the average fitness in each generation. It can be seen from the figure in operation process, when the iteration numbers = 8 generation, SVM penalty coefficient C, the optimal kernel function parameter g and the optimal values of 5 fold cross validation have reached convergence, when C=99.8375, g=0.151744, which can be seen from the figure the highest fitness value is 99.375%. The fitness curve of the cuckoo algorithm optimization SVM parameters is shown in Fig. 5.

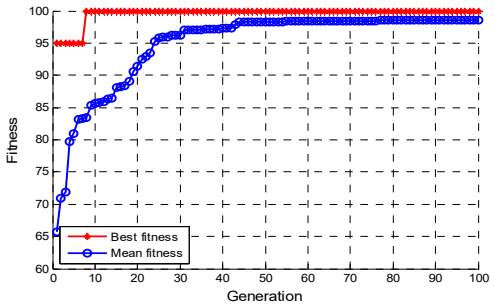

Figure 5. The fitness curve based on cuckoo algorithm

6.4 Results Analysis

This paper uses the SVM model as the classifier to assess credit risk of P2P lending borrowers. Since the classification performance of the SVM model depends on the choice of penalty parameter C and kernel function parameter g, which makes it necessary to choose optimal parameters of the SVM model. In this paper, we use parameter optimization method based on cuckoo algorithm and compared with genetic algorithm and grid search.

Table 5. Classification accuracy of svm model based on parameter optimization algorithms

Method	Sample		Accuracy	Total accuracy
Cuckoo Algorithm	Training Set	Normal Sample	98.75%	99.375%
		Default Sample	100%	
	Test Set	Normal Sample	100%	100%
		Default Sample	100%	
Genetic Algorithm	Training Set	Normal Sample	96.25%	98.125%
		Default Sample	100%	
	Test Set	Normal Sample	95%	97.5%
		Default Sample	100%	
Grid Search	Training Set	Normal Sample	91.25%	95%
		Default Sample	98.75%	
	Test Set	Normal Sample	95%	95%
		Default Sample	100%	

The classification accuracy of the SVM model using the 5-fold cross validation based on three parameter optimization methods are shown in Table 5. The results show that the classification accuracy of the SVM model based on the cuckoo search algorithm is the highest, the classification accuracy of the training set is 99.375% and the classification accuracy of the test set is 100%. The classification accuracy of the training set and the test set of SVM model based on genetic algorithm is 98.125% and 97.5% respectively. The accuracy of the training set and the testing set based on the grid search algorithm is 95%. The classification accuracy of the SVM model using cuckoo search algorithm to optimize parameter is 2.5% higher than using the genetic algorithm and also is 5% higher than using the grid search algorithm.

Considering the classification accuracy of the two kinds of samples, the P2P lending platform is likely to suffer a greater default risk for accepting the default sample. Therefore, in the reality we put

more attention on the default probability of P2P lending borrower. The classification accuracy of the training set and the test set based on cuckoo search algorithm for the default sample both reach up to 100%. The experimental results show that the classification performance of the SVM model based on the proposed method in this paper is better than the other two methods. The cuckoo algorithm can effectively optimize the parameters of SVM model so that the classification ability and generalization ability of the SVM model can achieve the best effect. In general, it is feasible to assess the credit risk of P2P lending borrowers based on the SVM model which uses cuckoo search algorithm to find best parameters. The model has good learning precision and robustness. It's also proves the superiority and practicality of the SVM model in solving the classification problems of small samples.

7. Conclusion

P2P online lending, which is regarded as an innovative lending mode, has made a great deal of contribution to SME financing and personal loans. This paper takes the data of P2P lending platform "renrendai" as an example to conduct deep information mining and comprehensive analysis, in the basic of which to build credit risk assessment indicator system of P2P lending borrower from multiple dimensions and makes an empirical analysis for the classification performance of the SVM model based on cuckoo search algorithm. The research shows that this model has the advantages of simple structure, low computational cost and higher classification accuracy.

Acknowledgments

This paper is funded by Beijing Social Science Fund (No. 16JDGLB001).

References

[1]. S. Angilella, S. Mazzu, "The financing of mit SMEs: A multi criteria credit rating model", European Journal of Operational Research, vol.244, 2015.

[2]. E. Angelini, G. Tollo and Roli A, "A neural network approach for credit risk Evaluation", Econ and Finance, vol.48, 2008.

[3]. G. L. Tang, "Application Research on credit risk assessment of borrowers based on P2P online lending platform", M.S. thesis, Shanghai University of Engineering Science, Shanghai, China, 2016.

[4]. Y. Q. Fan, "Research on personal credit evaluation based on Bias classifier," Xi'an Electronic and Science University, Xi'an, China, 2014.

[5]. S. G. Yang and Q. Zhu, "Construction of a combination model of personal credit assessment based on decision tree and neural network", Financial Forum, vol.18, pp.61-67, 2013.

[6]. Malekipirbazari, Milad and Aksakalli, "Vural. Risk assessment in social ATM via random forests", Expert Systems with Application. voll42, 2015.

[7]. X. D. Liu, F. Li, "The credit risk assessment of online lending under the background of large data", Journal of statistics and information, vol.5, pp.41-48, 2016.

[8]. Y. Liu, "Research on application of decision tree algorithm in credit risk evaluation of P2P online lending", M.S. thesis, Hunan University, Changsha, China, 2016.

[9]. T. Du, "Personal Credit Evaluation Model Based on Rough Set Support Vector Machine", Statistics & Decision, vol.1, pp. 94-96, 2012.

[10]. X. Yao, L Yu, "Fuzzy Approximate Support Vector Machine Model and Its Application in Credit Risk Assessment", Systems Engineering Theory and Practice, vol.32, 2012.

[11]. H. Lai, l. Shuai and Z. F. Zhou, “A New Approach to Credit Evaluation of Personal Loans Customer Technology”, vol33, pp.97-103, 2014.

[12]. G. Z. Zhang, W. Y. Chen and C. H. Liu, “Research on Risk Assessment of Personal Consumption in Commercial Banks Based on Logistic Model”, Journal of Financial Theory and Practice, vol.34, pp.53-57, 2015.

[13]. H. Xia, “Electronic commerce credit risk principal component analysis and support vector regression combination model of classification”, Modern information, vol.35, pp.76-79, 2015.