Borrower Risk Identification of P2P Platform Based on Support Vector Machine

Xiaozhen Yu

Shanghai University, Shanghai, China

xiaozhen123xz@163.com

Keywords: P2P online lending; Support vector machine; Radial basis function; Default risk identification

Abstract. In order to improve the risk control ability of the P2P online lending platform, the paper applies the support vector machine to the identification of borrower default risk. The paper chooses nonlinear support vector machine to analyze the data of borrowers of a China online lending platform, in which the kernel function selects the radial basis function. The results show that the SVM algorithm can effectively help the P2P network lending platform to improve the ability of identify the default risk of borrowers.

Introduction

With the rapid development of Internet technology, the Internet financial ecosystem is also beginning to form. P2P online lending platform effectively connects the borrowers and the lenders. Not only can the borrowers get the required funds in time, but also the lenders can earn interest rates. However, there are also many problems with the P2P online lending platform. For example, the platforms make off with money or the borrowers defaulted on his debts. Both borrowers and lenders bear a lot of risks.

According to asymmetric information theory, some people have the information that others don't have, so they have an information advantage. When there is an information asymmetry problem, there may be adverse selection and moral hazard problems. The problem of information asymmetry may lead to Adverse selection and Moral hazard. Similarly, in the P2P online lending market, the borrower has more knowledge of his or her situation, including family, work, finance, etc. The platform is unable to obtain complete and accurate information about the borrower for various reasons. This may result in the borrower using fraudulent information to get the loan. This is also the mechanism for the default risk of P2P online lending. The paper studies the credit risk identification problem of borrowers on P2P online lending platform, based on machine learning.

Reviews

Credit risk research is an important topic in the field of risk control research. Effective credit assessment methods are critical to the healthy development of credit business. Emekter, Tu, Jirasakuldech and LuD (2015) analyzed the data of Lending club and pointed out credit rating, debt-to-income ratio, FICO score, and application of cyclic quota have important influence on whether the borrower default[1]. H.Y. Gu and Z. Yao (2015) believes that the borrower's classification information has different effects on the borrower default rate. Their research provides support for building quantitative indicators of borrower default risk[2]. C.H. Zhang and B.H. Wu (2017) draw on the theoretical basis and method of typical personal credit risk measurement of commercial banks, and construct a P2P online lending credit risk rating index system based on soft information of borrowers[3]. Freedman and Jin (2017) studied the impact of social networks on P2P network lending, and found that borrowers who provide social relationships are more likely to borrow successfully and have lower borrowing rates[4].

At present, with the development of various algorithms in the field of machine learning, more and more scholars apply machine learning algorithms to credit risk identification. Malekipirbazari (2015) used the random forest, support vector machine, logistic regression, K-Nearest neighbors

  DOI: 10.25236/icess.2019.203

method to study the data of Lending club[5]. X. Li and Y.C. Dai (2018) selected basic information as the input layer of BP neural network. The results show that the BP neural network model can accurately assess the credit risk of borrowers[6]. P.J. Ma (2018) proposed a risk assessment method based on multi-feature fusion and cost sensitive decision tree to evaluate the network loan risk. The results show that the model has a good evaluation effect[7]. C. L. Huang (2007) selected two credit data sets in the UCI database as experimental data to verify the accuracy of the SVM classifier. Compared with neural networks, genetic programming and decision tree classifiers, support vector machine have the same classification accuracy in the case of fewer input features[8].

Data

The data used in this paper is the borrowers data of a certain online lending platform in 2017. After data cleaning, the available data is 12398 records of borrower. Based on analysis of previous research results, the paper selects characteristics from several aspects: personal information, financial status, credit status, platform information authentication, and borrowing target information. Including age, gender, marriage, education, working hours, work income, real estate, car production, mortgage, car loan, number of applications for borrowing, number of successful loans, number of payments, credit limit, number of overdue, severe overdue, academic qualification , residence certification, marriage certification, job certification, loan certification, income certification, video certification, credit report, real estate certification, vehicle certification, total loan amount, borrowing time, annual interest rate. And whether it is overdue as a response variable or called label.

The analysis shows that the majority of the borrowers are between 30 and 40, and the majority are man. And most of them married. This reflects the economic burden of married men of this age. Most of them have shorter working hours and less wages. Interest rate is between 10 and 12, most people borrowing time is 36 months. This tells us interest rate between 10% and 12% could be accepted by most people, and borrowers tends to borrow money for long time.

Table 1 Explanatory data analysis

index	variable interpretation	variable type	mean	mode
age	real value	qualitative	38.57	32
time of work	1,1-3,3-5,5	quantitative	2.22	1
Income	5000,5000-10000,10000-20000,20000-50000,50000	quantitative	10645.55	7500
Number of applications	real value	quantitative	1.25	1
Number of successful	real value	quantitative	1.14	1
Pay off	real value	quantitative	1.07	1
credits	real value	quantitative	61825.07	47100
Number of overdue	real value	quantitative	0.57	0
Serious overdue	real value	quantitative	0.03	0
Total amount of borrowings	real value	quantitative	68557.15	47100
Borrowing time	real value	quantitative	30.23	36
Annual interest rate	real value	quantitative	11.11	10.2

In order to better adapt to the model, the paper encodes the data. The real values are used for the quantitative variables, and the data is normalized. The two categorical variables in the qualitative variables are integer-encoded. Variables of multiple classifications in qualitative variables are coded by the One-hot encoding.

One-hot encoding is similar to integer encoding. They are used for the categorical variables. The difference is that One-hot coding uses binary vectors to represent different categorical variables. Integer coding uses different integers to represent different categorical variables. Since integer

coding uses integer variables to represent different categorical variables, the integer values have no practical meaning for categorical variables, but different integer values for model algorithms represent different mathematical meanings. Therefore, One-hot coding is often used for multi-category variables, and integer coding is used for two-category variables.

Model

Vapnik proposed a support vector machine algorithm based on VC Dimension theory and structural risk minimization in 1995. Compared to other machine learning algorithms, SVM has a solid theoretical foundation and a clear mathematical model. Some scholars believe that SVM can be used without modification. This means that a lower error rate can be achieved with a basic form of SVM[9].

The support vector machine is a two classification model. Support Vector Machine is a linear classifier with the largest interval defined in the feature space. The support vector machine also includes a kernel technique, which makes it a nonlinear classifier. The learning strategy of support vector machine is interval maximization, which can be formalized to solve convex quadratic programming problem. Its learning algorithm is the optimization algorithm for convex quadratic programming. Simply, the goal of SVM is to find an optimal hyperplane in a multidimensional space. The optimal hyperplane requirements not only make it easy to separate the two types, but also make the distance between the two types of distances closest to each other as large as possible.

For linearly separable samples, this problem can be expressed as:

$$\max_{w,b} \gamma \tag{1}$$

$$s.t \quad y_i \left(\frac{w}{||w||} \cdot x_i + \frac{b}{||w||} \right) \geq \gamma, i = 1,2, \cdots, N$$

The objective function indicates that we want the hyperplane (w, b) to have the largest geometric margin γ for the training data set, and the constraint indicates that the hyperplane (w, b) has a geometric maigin of at least γ for each training sample point.

We replace the geometric margin with function margin, so we can get the following equivalence problem:

$$\min_{w,b} \frac{1}{2} ||w||^2 \tag{2}$$

$$s.t \quad y_i(w \cdot x_i + b) - 1 \geq 0, i = 1,2, \cdots, N$$

Problem (2) that the basic model of SVM is a convex quadratic programming problem, We take it as primitive problem. Applying Lagrange duality, we can get the dual problem. The optimal solution of the primitive problem is obtained by solving the dual problem. The paper uses the SMO (Sequence Minimal Optimization) algorithm which is proposed by John Platt in 1996 to solve the dual problem.

Solve the Eq.2, we could get the maximum interval classification hyperplane $w^* \cdot x + b^* = 0$ and the decision function $f(x) = sign(w^* \cdot x + b^*)$. This is linear separable support vector machine.

For non-linearly separable samples, As shown in Figure 1, we can use the kernel function. Using the kernel function method, the learning method of linear classification can be extended to nonlinear classification. Extending linear support vector machines to nonlinear support vector machines, we need to replace the inner product in the dual form with a kernel function. Convert data from low dimensions to high dimensions by kernel function, we can turn problems into linear separable problems. In the paper, we choose Radial Basis Function (RBF):

$$K(x, y) = \exp\left(\frac{-||x-y||^2}{2\sigma^2}\right) \tag{3}$$

Finally, we get the decision function:

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i^* y_i K(x \cdot x_i) + b^*) \quad (4)$$

Empirical Studies

As mentioned above, the paper selects 12398 records of borrower information as sample data. There are 408 borrower default records. The data show a feature of class imbalance, so the paper judges the classification performance of the model by analyzing the precision rate, recall rate, and ROC curve[10].

Table 2 Confusion matrix of binary classification

		Predicted result	
		+1	-1
Actual result	+1	True positive(TP)	False negative (FN)
	-1	False positive(FP)	True negative(TN)

If a positive case is predicted as a positive case, a true positive is produced; if a negative case is predicted as a negative case, a true negative is produced.

The article uses the python program to get the following results. In the support vector machine (SVM) classification model, we choose the radial basis function (RBF). Data is divided into training set and test set. Training set has 8265 records.

At the beginning, we set the penalty parameter C = 10, Kernel function parameter gamma = 20. Run the program to get the following results:

Table 3 Training set- confusion matrix

	Not overdue(-1)	Overdue(+1)
Not overdue(-1)	7999	0
Overdue(+1)	0	266

Table 4 Test set- confusion matrix

	Not overdue(-1)	Overdue(+1)
Not overdue(-1)	3990	1
Overdue(+1)	127	15

Table5 Test set result evaluation index

	Explanation	Result
Precision	TP/(TP+FP)	0.937
Recall(Sensitive)	TP/(TP+FN)	0.106

Analysis of the above results, we can see that the model performs very well on the training set, but the performance on the test set is very bad. The precision rate is calculated as TP/(TP+FP), so the precision rate reflect the ability that the model can correctly classify the retrieved data. The recall rate is calculated as TP/(TP+FN), so the recall rate reflect the ability that the model correctly retrieve data. So we require high precision rate and high recall rates.

We adjust C and gamma values, increase the value of C, and decrease the value of gamma. Then we get the results as follow:

Table 6 Increase C, decrease gamma

C	Gamma	Precision rate	Recall rate
15	15	0.95652	0.15492
20	10	0.96667	0.20422
25	5	0.96078	0.34507
30	1	0.94845	0.64788
35	0.1	0.95302	1
40	0.01	0.95302	1

We can see when constantly adjusting the C and gamma values, the model is getting better and better on the test set. As the parameters change, there is a trade-off between precision rate and recall rate. Our goal is to find a set of parameters that make both precision rate and recall rate high. When C = 35, gamma=0.1, precision rate is 0.95302 and recall rate is 1.

The following is the ROC curve corresponding to different parameter values:

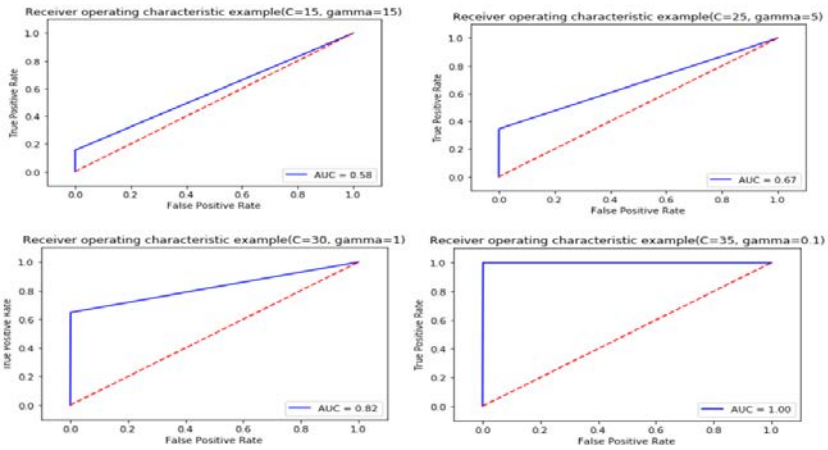

Figure. 1 ROC curve corresponding to different parameter values

Conclusions

Most platforms have their own credit rating system, however, relying solely on the level of credit cannot effectively judge whether the borrower will default. Through the above analysis, the SVM model based on the radial basis function shows good classification performance in the recognition of whether the borrower defaults. According to the borrower's various data information, including personal information, financial information, historical credit records, etc., the SVM based on radial basis function can effectively enhance the accuracy of platform estimation.

The shortcoming of this paper is that the support vector machine is not efficient enough to deal with large sample problems. If the support vector machine can be combined with boosting technology, the model may have higher efficiency.

References

[1] Emekter, Tu, Benjamas Jirasakuldech and Min Lu, Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending, Applied Economics, Vol. 47 (2015) No.1, pp.54-70.

[2] H.Y Gu and Z. Yao, Research on Influencing Factors of Borrower Default Risk in P2P Network Lending Platform, Shanghai Journal of Economics,(2015) No.11, pp.37-46. (In Chinese)

[3] C.H. Zhang and B.H. Wu, Measurement of credit risk of Chinese P2P network lending, Statistics & Information Forum.Vol.32 (2017) No.5, pp.110-115. (In Chinese)

[4] Freedman S, Jin G Z. The information value of online social networks: Lessons from peer-to-peer lending, International Journal of Industrial Organization,(2017) No.51, pp.185-222.

[5] Malekipirbazari M, Aksakalli V, Risk assessment in social lending via random forests, Expert Systems with Applications, Vol. 42 (2015) No.10, pp.4621-4631.
[6] X. Li and Y.C. Dai, Research on Credit Risk Assessment of P2P Network Loan Borrowers Based on BP Neural Network, Wuhan Finance, (2018) No.2, pp.33-37. (In Chinese)
[7] P.J Ma, Y. Wang, L. Yu, C.B. Li and L. Kuang. P2P Network Borrowing Risk Assessment Based on Cost Sensitive Decision Tree, Computer Integrated Manufacturing Systems, Vol. 24 (2018) No.7, pp.1880-1885. (In Chinese)
[8] C. L. Huang, M. C. Chen and C. J. Wang, Credit scoring with a data mining approach based on support vector machines, Expert Systems with Applications, (2007) No.33, pp.847-856.
[9] Peter Harrington, Machine Learning in Action(Posts & Telecom Press, US 2013).
[10] H. Li, Statistical Learning Method(Tsinghua University Press, China 2012). (In Chinese)