Personal Credit Evaluation Under the Big Data and Internet Background Based on Group Character

Cheng Liu, Dan Wang, Wenxin Wang and Zhenyi Ji*
Sichuan Agricultural University, Dujiangyan, China, 611830
*Corresponding author

Keywords: SVM, Logistic, Personal credit, Combination model.

Abstract. Personal credit evaluation is one of the important means of financial risk prediction.Ttraditional method of personal credit evaluation is Single model analysis. In order to accurately evaluate personal credit and reduce the default loss caused by credit economy to internet finance, combinatorial thinking is needed. In this paper, SVM model and Logistic regression model are analyzed by single analysis, and we set up SVM-Logistic combination model. The results show that the SVM-Logistic model has higher robustness andaccuracy.

Introduction

In recent years, with the rapid development of information technology, "Internet +" has become more and more popular. More and more measurement methods have been applied to the field of credit evaluation. The rapid development of big data provides more basic conditions for credit evaluation.A single model has its own limitations. Based on a single model, some scholars make use of the advantages of multiple models and synthesize the advantages of various aspects to combine two or more models to evaluate personal credit. This reduces the probability of errors to a certain extent, and improves the accuracy and stability of the evaluation model.Jiang Based on the advantages of different single models in personal credit evaluation, Jiang Minghui[1] established the Logistic-RBF combination model, and verified the advantages of the combination model. Xianghui and Yang Shenggang[2] propose a method of building composite model based on multi-classifier. Based on Support Vector Machine (SVM) and Logistic regression model, this paper establishes a SVM-Logistic combination model based on group feature characterization, evaluates and analyses personal credit, and cross-validates the results in order to improve the accuracy and robustness of the model.

Analysis Modeling

Support Vector Machine

Let the training sample set be $\{x_i, y_i\}, i = 1,2,\mathrm{L}\ ,N,$ is introduced as model input, $Y_i \in R$ is the corresponding output variable. The corresponding linear regression function can be defined as:

$$f(X) = w \cdot \phi(X) + b \tag{1}$$

Where, w is the weight vector, b is bias. According to statistical theory, the fitting problem of equation (1) can be described as the following optimization problem:

$$\min_{w,b,e}(w,e) = \frac{1}{2}\|w\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{N} e_i^2 \tag{2}$$
$$s.t. \qquad y_i = w^T\phi(x_i) + b + e_i, i = 1,2,\cdots,N$$

Where, $w \in R^{n^h}$ is the weight vector, e_i is error, $\gamma_i > 0$ is the penalty coefficient. By transforming equation (2) into dual space, and introducing Lagrange multiplier a_i, the following Lagrange functions are obtained.

$$L(w,b,e,a)=\frac{1}{2}\|w\|^2+\frac{1}{2}\gamma\sum_{i=1}^{N}e_i^2-\sum_{i=1}^{N}a_i(w^T\phi(x_i)+b+e_i-y_i) \tag{3}$$

According to KKT (Karush-Kuhn-Tucker) condition, a system of linear equations is obtained, which is

$$\begin{bmatrix}0 & I^T\\ I & \Omega+\gamma^{-1}I\end{bmatrix}\begin{bmatrix}b\\ a\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{4}$$

Where, $y=\left[y_1,,y_2,...,y_N\right]^T$, $a=\left[a_1,,a_2,...,a_N\right]^T$, $I=[1,1,...,1]^T$; Ω is a core matrix, Satisfying Mercer condition, and the elements in the i row and the j column are $\Omega_{ij}=K\left(x_i,x_j\right)=\phi\left(x_i\right)^T\phi\left(x_j\right)^T,i,j=1,2,...,N$, $K\left(x_i,x_j\right)$ are Gauss kernels in this paper.

The LSSVM fitting model can be obtained by solving linear equations (4).

$$y(X)=\sum_{i=1}^{N}a_iK\left(x_i,X\right)+b\text{ .} \tag{5}$$

Logistic Regression

Logistic regression is a further distortion of linear regression, and its equation can be expresse d as:

$$\ln\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1x_1+\cdots\beta_nx_n \tag{6}$$

Where, p_i is the probability of event occurrence, β_i is the parameter to be estimated, x_i is the explanatory variable. Equation (6) establishes the relationship between probability of occurrence of events and explanatory variables.

This paper calculates the prediction probability according to the different levels of the predictive variable X. If the predicted probability is fairly large, then we predict that it will happen. Conversely, if the predicted probability is fairly small, we don't expect it to happen. How to be "fairly large" or " fairly small" requires the determination of the point of separation between the two. There are three commonly used methods to determine the segmentation point, one is to select the "best" segmentation point, the other is to use 0.5 as the segmentation point, and the third is to determine the segmentation point according to the prior probability and misjudgement loss. In order to calculate conveniently, the second method is adopted in this paper.

SVM-Logistic Composite Model

The classification accuracy of SVM model is high, but it lacks robustness and explanability. Logistic regression model is robust and explanatory, but its classification accuracy is lower than that of artificial intelligence model. A natural solution is to combine models with different characteristics to form complementary advantages, and finally build a combination model with good classification accuracy and robustness. According to this idea, this paper establishes SVM-Logistic model for analysis. The structure diagram is shown in Figure 1. The detailed calculation steps are as follows:

(1)SVM model is used to select the characteristic variables that have a significant impact on distinguishing "good" and "bad" customers, and build a scoring model. The higher the score, the more likely the sample belongs to this category.

(2)These salient characteristic variables are used as input units of Logistic model.

(3) The score of each sample obtained by SVM model is also used as input unit.

(4) SVM-Logistic model is established.

When building a composite model, the output Y of the SVM model is essentially a measure of whether the sample belongs to a "good" customer or a "bad" customer. Similarly, the output Y'of the Logistic regression model can be regarded as a measure that the sample belongs to "good" or "bad" customers. The greater the probability value of the combination variable input into the SVM-Logistic combination model, the greater the probability that the sample belongs to the "good"

customer or the "bad" customer.

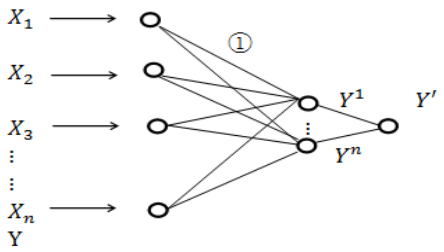

Figure 1. Structural Diagram of SVM-Logistic Composite Model (①:Discriminant analysis is needed when input variables in logistic model.)

Simulation

Data Source and Preprocessing

This paper analyses and studies German credit data, which is provided by the School of Computer Science and Information, University of California, Irvine. German Credit data set has 1000 credit data records, and defines two types of people. The first type (Good Credit) sample is 700, and the second type (Bad Credit) sample is 300. Each sample is described by 21 variables. The first 20 variables are attributes of the sample and the twenty-first variables are categories of the sample.

Before using sample data to model, data need to be preprocessed. There are vacancy data in the data set, so we need to fill the vacancy data or remove the sample corresponding to the vacancy data. In the code, we use the method of filling, filling each vacancy data to twice the maximum value of the corresponding parameters, in order to ensure a sufficient distance from the normal parameters. In addition, in order to eliminate the influence of dimension between the features of samples, the mapminmax () function is used to normalize the features of samples to [-1,1]. In order to ensure the accuracy of modeling, the original data are randomly divided into three parts. Two of them were used as training data and the last as test data. Using cross validation method, each test set and training set are different, and the average of correctness rate calculated ten times is used as the final correctness rate.

SVM Analysis

The data are analyzed by SVM using the software of Matlab. Because the data are linear inseparable, we need to transform the non-linear separable problem in the low-dimensional input space into the linear separable problem in the high-dimensional feature space by using the kernel function. Here we use RBF kernels.

Table 1. Test sample results (SVM)

Test Set	"good" Correctness Rate (%)	"bad" Correctness Rate (%)	Total Correctness Rate (%)
1	0.917	0.352	0.740
2	0.917	0.359	0.763
3	0.952	0.330	0.754
4	0.902	0.330	0.731
5	0.899	0.368	0.731
6	0.882	0.409	0.757
7	0.920	0.367	0.740
8	0.921	0.359	0.766
9	0.891	0.352	0.722
10	0.935	0.369	0.760
Avg-Correctness Rate(%)	0.914	0.359	0.746
Avg-error rate(%)	0.086	0.641	0.254

On the other hand, the bank is a profit-making organization, which wants to maximize profits and

minimize risks. The ideal credit evaluation model should try to make the "good" correct rate as high as possible, that is, to predict good customers into good customers. The lower the "bad" error rate, the better, that is, to predict bad customers as good customers. The test results are shown in Table 1. From Table 2, we know that the average correct rate of "good" is 0.914, but the average wrong rate of "bad" is 0.641. The average misclassification rate is very high, which is obviously not conducive to the risk control of banks.

Logistic Regression Analysis

Logistic regression analysis of data was carried out by using MATLAB software. According to the regression equation, the probability is 0.5 as the dividing point, and greater than 0.5 means "good" customers; conversely, it means "bad" customers. The analysis results are shown in Table 2. On the whole, the correct rate of logistic regression is obviously higher than that of SVM model. The average error rate of "bad" is 0.463, which is much lower than that of SVM model, but the correct rate of "good" is 0.853 at the same time. The result may be related to the low accuracy, robustness and strong interpretability of Logistic classification.

Table 2. Test sample results (Logistic)

Test Set	"good" Correctness Rate(%)	"bad" Correctness Rate(%)	Total Correctness Rate(%)
1	0.838	0.613	0.775
2	0.828	0.604	0.760
3	0.902	0.318	0.710
4	0.885	0.485	0.766
5	0.852	0.549	0.769
6	0.864	0.535	0.766
7	0.850	0.521	0.757
8	0.860	0.558	0.757
9	0.819	0.608	0.754
10	0.835	0.577	0.754
Avg-correctness Rate(%)	0.853	0.537	0.757
Avg-error rate(%)	0.147	0.463	0.243

SVM-Logistic Combination Analysis

Firstly, the SVM model is used to select the characteristic variables which have significant influence on distinguishing "good" and "bad" customers, and the scoring model is established. These salient characteristic variables are used as input units of Logistic, and the scores of each sample obtained by SVM model are also used as input units. Then, the SVM-Logistic model is established.

The analysis results are shown in Fig. 2 and Table 3. The combination model can significantly reduce the high "good" error rate of Logistic model and the high "bad" error rate of SVM model. And the total misjudgment rate is the lowest in the training set and the test set. Compared with the robustness of the three models, the combination model has the highest accuracy and the strongest robustness. It shows the rationality of the combination model and the applicability of the combination model in personal credit scoring model is higher than that of the single model.

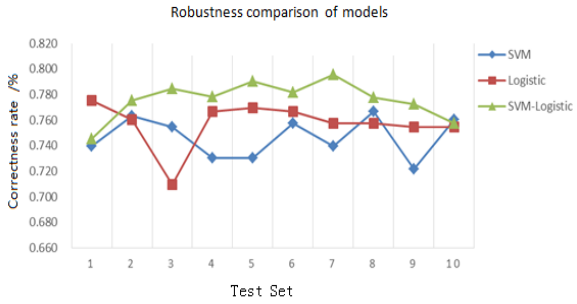

Figure 2. Robustness comparison

Table 3. Comparison of error rates of three models

CLASSIFIER	"good" Correctness RATE(%)		"bad" Correctness RATE(%)		Total Correctness Rate(%)	
	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set
SVM	0.033	0.086	0.531	0.641	0.287	0.254
Logistic	0.087	0.147	0.426	0.463	0.223	0.243
SVM-Logistic	0.024	0.059	0.514	0.613	0.183	0.235

Conclusion

This paper lists two single personal credit scoring models, and on the basis of these two single models, establishes a combination scoring model. Through the application analysis, we can see that the single credit scoring model has its own advantages and disadvantages, but none of them can achieve the perfect unity of stability, prediction accuracy and interpretability. Although different single models have their own advantages and disadvantages, they are not mutually exclusive, but complementary and interrelated. The prediction results of each method contain useful information from different perspectives. Therefore, we establish a combination model based on SVM and Logistic regression. The simulation results show that the SVM-Logistic combination model can significantly reduce the high "good" error rate of the Logistic model and the high "bad" error rate of the SVM model, which shows the rationality of the combination model and that in the personal credit scoring model, the combination model is indeed more accurate than the single model, and has better adaptability to credit risk control. Usefulness.

Acknowledgment

This paper is funded by the National Natural Science Foundation Youth Fund Project (41701324), the Spark Program Project of the Ministry of Science and Technology (2012 GA810002), and the Key Project of Sichuan Science and Technology Department (2018NZ0057).

Reference

[1] Jiang Minghui,Chen Yufang.. Combining forecasts of personal credit scoring based on RBF neural network [J]. Journal of Harbin Engineering University. 2006.27(7): 138-141

[2] XIANG Hui,YANG Sheng-gang. An Ensemble Credit Scoring Model Based on Multiple Classifiers [J]. Journal of Hunan University(Social Sciences). 2011.25(3): 30-33

[3] REN Xiao, JIANG Minghui, CHE Kai, WANG Shang. The research on methods of personal credit scoring combined model selection based on optimized index system [J]. Journal of Harbin Institute of Technology.2016.48(5):67-71

[4] Ding Juanjuan. Cui Yuanyuan. Application of Combination Forecasting in Personal Credit Evaluation [J]. Academic Exchange. 2006.9:117-120

[5] Liao Guomin Tu Wenhua Ning Jing. Measuring the Risk of Individual Consumption Loan: Toward a Logistic Model [J]. Journal of Guangdong University of Foreign Studies. 2013(5):27-33.

[6] WANG GuoFu, ZHANG QingFeng. A New Application of Logistic Regression to Discriminant Analysis [J]. Journal of Anyang Institute of Technology. 2009(6): 89-91.

[7] GAO-Li. The Application of Support Vector Machine in Personal Credit Scoring [J]. Journal of Xinxiang Teachers College. 2008(2):12-15.

[8] Xiang Hui. Research and Application of Personal Credit Scoring Combination Model [D] Hunan: Hunan University. 2011:1-125

[9] Huang Wei , Zhang Liang, Tang You. Perfection of Personal Credit Evaluation Method Based on SVM Algorithm [J]. Journal of Heilongjiang August First Land Reclamation University. 2016. 28(2): 105-110.

[10] Dai Tingting, Han Yan, Hu Xiaofei. Personal Credit Evaluation Based on PLS-SVM [J]. Journal of Dehong Normal College 2017.1(25):105-107