SCJ SOFT COMPUTING JOURNAL
ISSN: 2322-3707
University of Kashan

An ensemble method based on bagging SVM for credit rating problem

Sayyed Mohammad Hoseini[1,*] Majid Ebtia[2] and Ramin Khochiani[3†]

[1,2]Department of Mathematics, Ayatollah Boroujerdi University, Boroujerd, Iran.
[3]Deparment of Economics, Ayatollah Boroujerdi University, Boroujerd, Iran.

Abstract. In this paper, a classification model based on an ensemble approach is used for credit scoring of the bank costumers. The proposed method is based on the bagging scheme for the support vector machines classifier. Using the bootstrap method, the dataset is first divided into some subsets and then the classifier is implemented on each subsets. The support vector machines method is used as classifier. Then the final model is made by voting among all of the classifiers. The proposed method has many advantages for implementation, including the reduction of computational costs. Two credit datasets are used to show the efficiency and applicability of the present method.

Keywords: Credit rating, SVM, ensemble method, bootstrap, bagging

1. Introduction

Banks play a significant role in influencing the economic performance of countries through activities such as resource allocation, liquidity provision, issuing payment instruments, and extending credit facilities. Engaging in these activities exposes banks to various risks, with credit risk being the most significant. Credit risk is the potential for loss arising from a borrower's inability to repay a loan or fulfill contractual obligations [1]. Typically, it encompasses the risk of a lender not receiving the principal and interest owed. Credit risk gives rise to numerous challenges, including disruptions to cash flows and increased costs associated with collection and the profit margins of banks. Banks also face a sudden reduction in resources and the risk of bankruptcy. Thus the correct and optimal provision of financial facilities is one of the important activities of banks. To do this, the characteristics of the customers should be correctly identified and this will be achieved by properly validating the customers based on their ability and willingness to fully repay their received facilities.

Many researchers have used new and modern data analysis techniques on the subject of customer credentials. Methods such as logistic regression, neural network, genetic algorithm, decision tree, support vector machine (SVM), and many other methods that are covered by data mining, machine learning, and artificial intelligence have been used to validate customers and estimate credit risk. In this article, the issue of credit rating of real applicants for bank facilities has been studied. To perform the above credit rating, group methods based on SVM have been used. Also, a comparison was made between the

*Corresponding author. Email: sm.hoseini@abru.ac.ir
†Email(s): smhoseiny@gmail.com, majid.ebtia@gmail.com, khochiany@abru.ac.ir

four common cores in the group SVM method. Finally, a comparison is made between this method and random forest, which is one of the most common and widely used group methods.

2. Literature review

The first research and model for measuring credit risk on bonds was done by John Murray in 1909 [2]. Various methods such as data envelopment analysis [1] and logit regression [3] have been proposed to design a model for measuring credit risk. The use of data-driven methods has received much attention in recent years. Mehrara et al. [4] compared the two methods of logistic regression and neural networks for credit rating of bank customers, the results of which show 80% accuracy for regression and 87% for neural networks and show superior performance and better prediction of networks. Thomas [5] used the logistic regression model to predict the credit risk of borrowers and achieved an accuracy of nearly 72%. Keshavarz Haddad and Aytigazar [3] compared decision tree and logistic regression methods in the process of credit rating bank applicants for facilities, each of which presented an accuracy of nearly 96% and 82%, respectively.

The neural network method has also been used to classify loan applicants, which presented an accuracy of 69-84% [6]. Tang et al. [7] used the random forest method to assess credit risk by credit cards in the energy industry in China. The purpose of the study is to scientifically measure the credit risk of credit cards used in the energy industry. The results show that credit card features such as credit added ratio and credit card costs over a month have a significant impact on credit risk. This valuable information will help banks to improve their risk management. Also, the SVM method, which is one of the most widely used methods in classification problems, in addition to the credit rating problem, has been implemented on various field issues such as medicine [8, 9] and on the issue of educational status [10].

In order to obtain a better model, the ensemble methodology combines several models, each of which treatments the same task. The philosophy of this method is base on that an aggregating voting from several simple models is more accurate and reliable estimates or decisions [11, 12]. The motivation of this study is to present the reliability and accuracy of the credit scoring model base on ensemble SVM method.

3. Main results

In this section, we introduce the ensemble bagging SVM method and then we apply the method on the credit rating problem. Let X is the given two labeled dataset as follows

$$X = \{(x_i, y_i) \,|\, x_i \in \mathbb{R}^n,\ y_i = \pm 1,\ i = 1, 2, \ldots, N\} \tag{1}$$

where x_i represents the vector of variables related to the i-th loan applicant and labels indicate whether customers are well-accounted for or not, i.e. the label 1 is used for good customers and -1 is used for bad ones. The goal of solving this problem is to find a classification model based on the training data that can best distinguish between good and bad customers among them and also, this model should be able to predict the labels of the testing data and, especially the new customers as well as possible. Recently, ensemble classification methods have received much attention from researchers. In the ensemble learning approach (as shown in Fig. 1), several different classifiers are created on the data set, then the final model is generated based on voting between them.

We first divide the dataset, X, into two subsets, say training data and testing data. Here we use 80% of the dataset for training, and the remaining 20% of the dataset for

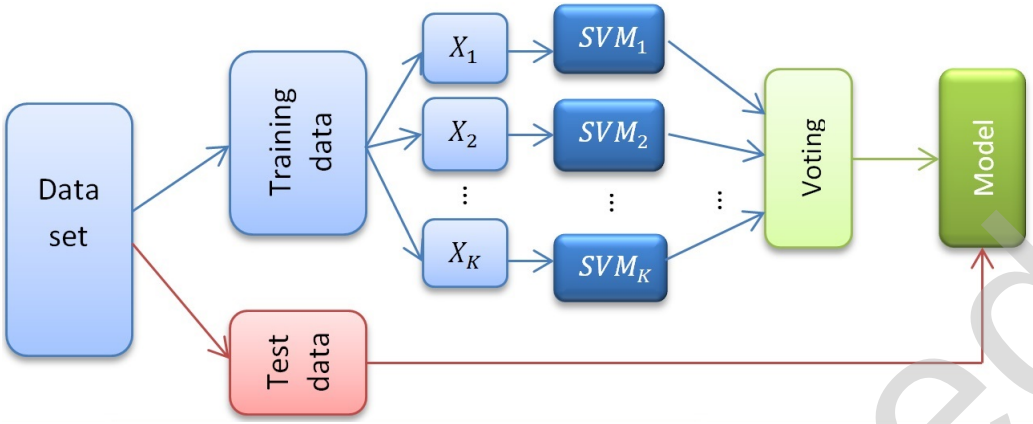

FIGURE 1. Flowchart of the ensemble SVM classification method.

testing. The training data then is divided into K subsets X_k, for $k = 1, 2, \ldots, K$. Here we use the bootstrap approach to obtain these subsets. Now, we apply the SVM method on each subset X_k, for $k = 1, 2, \ldots, K$ to get K local classifiers and the final and global classifier can be obtained by voting on these local classifiers. We used the bagging method to construct the ensemble classification model.

Let TP denotes the number of good costumers that are predicted as $+1$; TN refers the number of bad costumers that are predicted as -1; FP denotes the number of bad costumers that are predicted as $+1$; FN refers the number of good costumers that are predicted as -1. Then the following criteria can be used to evaluate the final model:

$$E_I = \frac{FP}{FP+TN}, \qquad E_{II} = \frac{FN}{FN+TP},$$
$$ACC = \frac{TP+TN}{TP+TN+FP+FN},$$
$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(FP+TP)(FN+TP)(FP+TN)(FN+TN}}.$$

4. Numerical results

We apply the SVM and ensemble SVM methods to two benchmarking datasets including the Australian and German credit datasets, from the UCI Machine Learning Repository, published by the University of California Irvine. The Australian dataset has 690 data points composed of 307 good costumers and 383 bad costumers. The data points of this dataset have 14 features including 6 continuous and 8 categorical.

The German credit dataset has 1000 data points composed of 300 bad costumers and 700 good costumers. Each data point has 20 features, which include 3 continuous and 17 discrete features: age, household, credit history, account balances, job, employment status, savings account, status of existing checking account, housing, income, other debtors, present residence, property, number of existing credits, other installment plans, telephone, foreign worker, credit amount, duration, and purpose of the loan.

The models of Australian and German credit datasets are built using the individual and ensemble methods, and in Table 1, the results in terms of ACC, MCC, E_I and E_{II} are reported, which are expressed as a percentage. In the proposed model, the improvement of ensemble SVM evaluation criteria is observed compared to individual SVM. For example,

TABLE 1. Results of credit ranking problem using the SVM and ensemble SVM together with the desicion tree and random forest methods.

	German Dataset									
	SVM				ensemble SVM				DT	RF
	Rbf	Lin.	Poly.	Sig.	Rbf	Lin.	Poly.	Sig.		
ACC	76.50	77.50	79.00	70.50	78.50	78.50	**80.50**	70.50	76.50	72.50
MCC	39.32	42.01	46.89	50.19	46.20	45.04	52.03	**57.14**	39.75	23.34
E_I	27.99	26.48	24.79	64.75	25.67	25.21	**23.28**	64.75	28.22	33.62
E_{II}	32.44	31.24	28.21	50.00	28.07	29.55	**24.68**	50.00	31.95	41.69
	Australian Dataset									
	SVM				ensemble SVM				DT	RF
	Rbf	Lin.	Poly.	Sig.	Rbf	Lin.	Poly.	Sig.		
ACC	84.05	86.23	84.05	84.05	84.78	86.23	86.95	84.05	86.95	**87.68**
MCC	66.09	71.24	65.78	67.88	67.01	71.67	71.63	67.88	71.68	**73.24**
E_I	17.21	14.99	17.11	17.02	16.05	14.96	13.12	17.02	13.50	**12.50**
E_{II}	16.71	13.76	17.11	15.08	16.94	**13.36**	15.22	15.08	14.81	14.24

criteria E_{II}, which indicates the misclassification rate of bad customers, in the Bagging SVM with polynomial kernel has lowest value and is fewer than the values compared to other methods summarized in table 1. A comparison is made between the proposed method and decision tree (DT) and random forest (RF) methods in table 1. It should be mentioned that RF algorithm is an ensemble approach based on the DT method. The results show that for German dataset, the best values of the defined criteria are achieved by the proposed ensemble method.

It is worth noting that the customer credit factors in the Australian and German datasets have undergone significant changes due to rapid and substantial shifts in customer behavior. Consequently, there is an urgent and critical need for research based on current real-world datasets.

5. Conclusion

As the credit industry continues to expand, credit scoring has emerged as a crucial tool for distinguishing between good and bad applicants. In this study, we introduce an ensemble strategy utilizing SVM to develop credit scoring models. We apply the proposed method to two different credit datasets to illustrate the efficiency of the ensemble method. A comparison between individual SVM and the proposed ensemble SVM method shows that the evaluation criteria are improved for the latter. In the German credit dataset, the best results are obtained via the proposed method.

References

[1] S. Isazade and B. Oryani, "Credit Risk Rating of the Bank's Customers by Data Envelopment Analysis: Case Study the Branches of Keshavarzi Bank," *Quarterly Journal of Economic Research and Policies,* vol. 18, no. 55, pp. 59–86, 2010. Available: http://qjerp.ir/article-1-230-en.html. In Persian.

[2] M. F. Fallah Shams and H. Mahdavi Rad, "Validating Model and Risk Forecasting for Leasing Customers (Case Study: Iran Khodro Leasing Company)," *Economics Research,* vol. 12, no. 44, pp. 213–234, 2012. Available: https://joer.atu.ac.ir/article_967.html. In Persian.

[3] G. R. Keshavarz Haddad and H. Ayati Gazar, "A Comparison between Logit Model and Classification Regression Trees (CART) in Customer Credit Scoring Systems," *The Economic Research (Sustainable Growth and Development)*, vol. 7, no. 4, pp. 71–97, 2008. Available: http://ecor.modares.ac.ir/article-18-5897-en.html. In Persian.

[4] M. Mehr Ara, M. Mousaii, M. Tasavori, A. Hassan Zadeh, "Credit ranking of Parsian bank legal customers," *Journal of Economic Modelling* vol. 3, no. 10, pp. 121-150, 2009. In Persian.

[5] L. C. Thomas, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers," *International Journal of Forecasting*, vol. 16, no. 2, pp. 149–172, 2000, doi: https://doi.org/10.1016/S0169-2070(00)00034-0.

[6] Y. S. Kim and S. Y. Sohn, "Managing loan customers using misclassification patterns of credit scoring model," *Expert Systems with Applications*, vol. 26, no. 4, pp. 567–573, 2004, doi: https://doi.org/10.1016/j.eswa.2003.10.013.

[7] L. Tang, F. Cai, and Y. Ouyang, "Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China," *Technological Forecasting and Social Change*, vol. 144, pp. 563–572, 2019, doi: https://doi.org/10.1016/j.techfore.2018.03.007.

[8] A. Vasighi Zaker and S. Jalili, "Candidate disease gene prediction using One-Class classification," *Soft Computing Journal*, vol. 4, no. 1, pp. 74–83, 2021. Available: https://scj.kashanu.ac.ir/article_111393.html. In Persian.

[9] H. Veisi, H. R. Ghaedsharaf, and M. Ebrahimi, "Improving the performance of machine learning algorithms for heart disease diagnosis by optimizing data and features," *Soft Computing Journal*, vol. 8, no. 1, pp. 70–85, 2021, doi: 10.22052/8.1.70, In Persian.

[10] A. Khosravi, H. Abdulmaleki, and M. Fayazi, "Predicting the academic status of admitted applicants based on educational and admission data using data mining techniques," *Soft Computing Journal*, vol. 9, no. 2, pp. 94–113, 2021, doi: 10.22052/scj.2021.242837.0, In Persian.

[11] R. Polikar, "Ensemble based systems in decision making," in *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, Third Quarter 2006, doi: 10.1109/MCAS.2006.1688199.

[12] M. Ebtia, S. M. Hoseini, and R. Khochiani, "Credit rating of bank customers using a new ensemble method based on support vector machine: a case study of Pasargad bank," *Soft Computing Journal*, vol. 10, no. 2, pp. 2–15, 2022, doi: 10.22052/scj.2022.243227.1016, In Persian.