An AdaBoost-DT Model for Credit Scoring

Jiali Zhao
School of Economics & Management, China Jiliang University, China, zjl13396717601@163.com

Zengyuan Wu
School of Economics & Management, China Jiliang University, China, wuzengyuan@cjlu.edu.cn

Bei Wu
School of Management and E-business, Zhejiang Gongshang University, China

Short Research Paper

An AdaBoost-DT Model for Credit Scoring

Jiali Zhao[1], *Zengyuan Wu*[1], *Bei Wu*[2*]

[1]School of Economics & Management, China Jiliang University, China

[2]School of Management and E-business, Zhejiang Gongshang University, China

Abstract: Credit scoring for loan applicants is an essential measure to reduce the risk of personal credit loan. Due to low percentage of non-performing loans, credit scoring is typically considered as an imbalanced classification problem. It is difficult to adress this kind problem using a single classifier. In order to settle the problem of imbalanced samples in credit scoring system, an ensemble learning classification model named AdaBoost-DT is proposed. In this model, we employ adaptive boosting (AdaBoost) to cascade multiple decision trees (DT). The weights of the base classifier can be adjusted automatically by enhancing the learning of misclassified samples. In order to verify the effectiveness empirically, we use data from Kaggle platform. Ten-fold cross-validation is carried out to evaluate and compare the performance among AdaBoost-DT model, DT, and Random Forest. The empirical results show that AdaBoost-DT model has higher accuracy. This model is valuable for banks and other financial institutions to evaluate customers' credit efficiently.

Keywords: credit scoring, ensemble learning, imbalanced classification

1. INTRODUCTION

The rapid development of the economy and the change in consumer attitudes have driven the development of personal credit loans. According to data from the People's Bank of China, China's overall personal credit consumption balance rose from $18.95 trillion in 2015 to $43.97 trillion in 2019. However, with the increasing amount of credit transactions, the non-performing loan rate has also increased year by year. According to the data released by the China Banking Regulatory Commission, the non-performing loan rate of commercial banks in China reached 1.86% at the end of the fourth quarter of 2019, and the non-performing loan balance reached 2.41 trillion yuan[1]. Non-performing loans not only affect the normal operation of banks, but also induce social moral risks and cause a series of adverse reactions. Credit scoring for loan applicants is an important tool to reduce credit risk and non-performing loan rates.

Credit scoring refers to classifying customers into "good credit" and "bad credit" customers according to their default risk. The idea of credit scoring appeared in the United States, where David Durand[2] proposed firstly the application of statistical methods in this field in 1941 to determine the goodness of loan customers. In the late 1960s, the emergence and development of credit cards made banks and companies with credit operations aware of the importance of credit scoring, and more and more experts began to study it. Altman[3], Meyer[4], Tam[5], Lundy[6] and other scholars used multivariate discriminant analysis, regression analysis, k-nearest neighbor discriminant analysis, and cluster analysis to evaluate individual credit. Leong[7] used a Bayesian network model to solve the truncated sample, real-time implementation problem in credit risk scoring. Comparing to logistic regression and neural networks, this model performs better in several dimensions such as accuracy and sensitivity. Fang & Chen[8] propose a credit scoring model based on semi-supervised generalized additive (SSGA) logistic regression to use both labeled and unlabeled sample information.

In recent years, with the development of big data and Internet finance, some artificial intelligence methods have been widely applied in credit scoring, including ANN[9], decision trees[10], and SVM[11]. Tony & Jonathan[12] conducted a comparison study between SVM and traditional methods such as logistic regression. They found that

[*] Corresponding author. Email: wuzengyuan@cjlu.edu.cn(Zengyuan Wu) , zjl13396717601@163.com(Jiali Zhao)

SVM can be used as a feature selection method to discriminate the most important features that determine the magnitude of default risk. Hussain[13] used an artificial neural network approach to provide technical support for commercial banks' lending decisions. Kambal[14] used decision trees (DT) and artificial neural networks (ANN) to build credit scoring models and performed comparative analysis. He found that the ANN approach mostly outperformed DT, but the results of DT were more explanatory. Artificial neural networks would improve the efficiency of credit decisions and help financial institutions save analysis time and cost. With the deepening of theoretical and practical research, the imbalance of samples was noticed, i.e., the number of customers with good credit is not the same as the number of customers with bad credit. And the current single classifier cannot obtain good classification results when processing unbalanced data. Therefore, Corchado [15] used ensemble learning algorithm by cascading several weak classifiers to overcome the limitations of a single classifier. Wang [16] proposed an ensemble algorithm with decision trees to reduce the effects of data noise and redundancy of features, and confirmed that relatively high classification accuracy could be obtained. Finlay[17] used a variety of Boosting and Bagging for modeling in the credit scoring, and the results showed that the ensemble learning method achieved better classification results than single classifier.

According to the shortcoming of a single classifier to process imbalanced data, we employ the adaptive boosting algorithm for credit scoring. As a typical ensemble algorithm, the AdaBoost algorithm can automatically adjust the weights of the base classifier and improve the classification accuracy by enhancing the learning of misclassified samples. Decision tree is employed as the base classifier. An ensemble learning classification model is proposed for credit scoring, where we employ adaptive boosting algorithm by cascading multiple decision trees (DT), named AdaBoost-DT model. Area Under Curve (AUC) and G-mean are selected as performance evaluation metrics. Furthermore, we empirically test our model using the data from Kaggle platform. In order to verify the effectiveness, we compare the performance of our proposed model with Decision Tree and Random Forest.

2. ADABOOST-DT MODEL

2.1 Decision tree classifier

A decision tree is a tree-like structure that divides a set of input samples into several smaller sets based on certain features of their attributes, and it is a fundamental classification method in machine learning. Unlike traditional statistical classifiers, decision trees use a multi-stage or sequential approach to the label assignment problem. The labelling process is considered as a simple decision chain based on successive test results rather than a single complex decision. In general, decision tree structures include tree nodes, bifurcation paths, and leaf nodes. The root node represents the object, while each branch fork path represents the value of an attribute of the object, and the leaf node represents the value of the object as represented by the path experienced from the root node to that leaf node.

Decision trees were chosen as the base classifier for three main reasons. Firstly, the resulting classification model is easier to explain and illustrate due to its intuitive presentation[18]. Secondly, unlike statistical models, decision trees require fewer assumptions in terms of data distribution[17]. Finally, they are relatively fast to construct compared to other techniques.

2.2 AdaBoost algorithm

The AdaBoost [19] algorithm is a classical ensemble algorithm proposed by Yoav Freund and Robert Schapire in 1995 to achieve better prediction by cascading several weak classifiers. The basic idea is that at the beginning, if there are N samples, each training sample is given the same weight 1/N. If a sample fails in training during the training process, a larger weight is given, which can make the classifier in the next iteration will focus on learning those failed samples. However, for accurately classified samples, their weights are reduced to obtain a new sample distribution. The AdaBoost algorithm training process is as follows.

Input: training sample set

$S=\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_i,Y_i)\}, i=1,2,\ldots,n, Y_i\in\{0,1\}$, M is the number of iterations, and H is the base classifier.

（1）Initialize the weight distribution of each sample in the training data

$$D_1=(u_{11},u_{12},..u_{1i},\ldots,u_{1N}), u_{1i}=\frac{1}{N}, i=1,2,\ldots,N \tag{1}$$

（2）Perform M iterations

（a）The training sample set with the weight distribution Hm is learned to obtain the weak classifier.

$$H_m(x):\chi\to\{-1,+1\} \tag{2}$$

（b）Calculate the classification error rate e_m , and discard the weak classifier if e_m it is greater than 50%.

$$e_m=\sum_{i=1}^{N}P(H_m(x_i)\neq y_i)=\sum_{i=1}^{N}u_{mi}I(H_m(x_i)\neq y_i) \tag{3}$$

（c）Calculate the importance of the weak classifier in the final classifier

$$\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m} \tag{4}$$

（d）Update the weight distribution of the training sample set for the next round of iterations.

$$D_{m+1}(i)=\frac{D_m(i)}{Z_m}\exp(-\alpha_m y_i H_m(x_i)), i=1,2,\ldots,N \tag{5}$$

where Z_m is the normalization factor and is the sum of all samples corresponding to weights of 1.

$$Z_m=\sum_{i=1}^{N}u_{mi}\exp(-\alpha_m y_i H_m(x_i)) \tag{6}$$

（3）Combining weak classifiers to output strong classifiers

$$H(x)=sign(f(x))=sign\left[\sum_{m=1}^{M}\alpha_m H_m(x)\right] \tag{7}$$

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental data set

To verify the effectiveness of the model in this paper, the experiment uses the customer credit dataset from the public dataset provided by Kaggle. We define the customers with two months or more overdue loan repayments in Status as the default sample (positive sample) and the rest as the compliance sample (negative sample), where the number of positive samples is 422 and the number of negative samples is 24,712, with an imbalance ratio of 1:58. In addition to the repayment status, each record also contains 17 attributes, as shown in Table 1.

Table 1Sample properties

Property Name	Meaning	Property Type
ID	Customer Number	Continuous type
CODE_GENDER	Gender	Discrete type
FLAG_OWN_CAR	Is there a car	Discrete type
FLAG_OWN_REALTY	Whether there is a property	Discrete type
CNT_CHILDREN	Number of children	Continuous type
AMT_INCOME_TOTAL	Annual income	Continuous type
NAME_INCOME_TYPE	Income Category	Discrete type
NAME_EDUCATION_TYPE	Education level	Discrete type
NAME_FAMILY_STATUS	Marital Status	Discrete type
NAME_HOUSING_TYPE	Living Style	Discrete type
DAYS_BIRTH	Birthday	Continuous type
DAYS_EMPLOYED	Start date	Continuous type
FLAG_MOBIL	Availability of cell phones	Discrete type
FLAG_WORK_PHONE	Availability of working telephone	Discrete type
FLAG_PHONE	Availability of telephone	Discrete type
FLAG_EMAIL	Is there email	Discrete type
OCCUPATION_TYPE	Career	Text type
CNT_FAM_MEMBERS	Family size	Continuous type
MONTHS_BALANCE	Month of recording	Continuous type
STSTUS	Repayment Status	Discrete type

3.2 Data pre-processing

Customer credit data is characterized by large volume, missing data and anomalies, which are not conducive to finding the required information quickly. Therefore, the above-mentioned characteristics of credit datasets need to be pre-processed before data mining to provide clean and more targeted high-quality data for data mining algorithms, thus improving data mining efficiency.

Firstly, features with more than half of the missing values are removed and the rest are filled with the missing values using the mean value. Secondly, the continuous values of income, age and years of work are discretized to increase the robustness to abnormal data. Finally, the income categories of customers are aggregated, and all job categories are divided into "lab work", "office work", "high-tech work". This is used to analyze the relationship between the customer's income category and the default or non-default.

3.3 Evaluation indicators

The current evaluation metrics for binary classification problems usually use the classification correctness [20] (Accuracy, Acc), but Acc ignores the performance of recognition of a few classes. For example, the prediction of a certain disease, even if the accuracy reaches 99%, but does not identify the people who are really sick, such a high accuracy is meaningless. The effective identification of a few classes in unbalanced data classification is more practically meaningful, so Acc is not sufficient as a performance metric for the model. In order to better evaluate the accuracy of the model, the geometric mean criterion (G-means metric) and AUC (area under the ROC curve) are used as evaluation metrics in this paper. Most of the above evaluation methods are represented by confusion matrix (Table 2). Among them, TP indicates that the positive class sample prediction is still positive, TN indicates that the negative class sample prediction is still negative, FP indicates that the negative class sample misclassification is positive, and FN indicates that the positive class sample misclassification is negative.

Table 2Confusion matrix

Actual category	Predicted results	
	Positive Class Sample	Negative Class Sample
Positive Class Sample	TP	FN
Negative Class Sample	FP	TN

(1) G-mean value

$$G-mean=\sqrt{\frac{TP}{TP+FP}\times\frac{TP}{TP+FN}} \tag{15}$$

The G-mean is based on the positive class classification accuracy and the negative class classification accuracy, so it can better measure the comprehensive performance of the classification method on the unbalanced data set. And its value is in the range of [0,1], the larger the G-mean value is, the better the comprehensive performance of the model is. The G-mean value is large only when both positive and negative classes achieve high accuracy.

(2) AUC

The ROC curve is a curve plotted with the false positive rate as the horizontal coordinate and the true positive rate as the vertical coordinate, depicting the changes of the false positive rate and true positive rate under different parameter variations. The classifier's performance will be better if the curve is close to the upper left corner. The performance of the classifier is generally evaluated by the area AUC between the ROC curve and the coordinate axis. The higher the value of AUC, the better the performance of the classifier.

3.4 Analysis of experimental results

The experiments were conducted using a ten-fold cross-validation method to divide the data set into 10 parts equally, and the ratio of the training set to the test set was 1:9. The average value was finally used as an estimate of the accuracy of the classification algorithm. In this paper, AUC and G-mean values are used as model evaluation indexes, and they are compared and analyzed with decision tree(DT) and random forest. The classification performance results of the three models are shown in Table 3.

Table 3Performance comparison of various algorithms

Models	AUC mean value	G-mean average value
DT	0.5218	0.2072
Random Forest	0.5500	0.5125
AdaBoost-DT	0.6854	0.7495

From the experimental results, it is seen that the AdaBoost-DT model has larger values on both AUC and G-mean than the other two models. Therefore, the proposed model in this paper has a higher accuracy rate.

According to Table 3, it can be seen that the average AUC value of the AdaBoost-DT algorithm proposed in this paper is greater than the other two algorithms, so the classification prediction effect of the model in this paper is better than the other two models. Generally, the closer the AUC value is to 1, the better the classification performance of the model is. And the AUC value of the algorithm proposed in this paper reaches 0.69, which has a better performance.

G-mean value is used to measure the accuracy of the positive and negative classes of the sample, and high G-mean value indicates that the discrimination between the two classes is accurate. The AdaBoost-DT algorithm proposed in this paper achieves 0.75 in the G-mean value. Thus, the model in this paper can effectively discriminate not only the good credit customers but also the bad credit customers, and can fully maintain the classification balance between the two.

In summary, the AdaBoost-DT model proposed in this paper has higher accuracy for customer credit evaluation in unbalanced data classification.

4. CONCLUSIONS

In the era of big data where the financial industry is moving towards information technology, it has become an inevitable trend to apply machine learning to credit scoring models. In this context, the AdaBoost ensemble algorithm is proposed to build a credit scoring model in order to improve customer credit prediction. According to the shortcoming of a single classifier to process imbalanced data, we use the AdaBoost-DT algorithm to build a credit scoring model. This model uses the AdaBoost ensemble method with DT as the base classifier, and solve the propensity problem for most samples by using the AdaBoost algorithm to increase the weight of the sample automatically. The empirical results show that the AdaBoost-DT model has higher accuracy than DT and random forest. The study provides a new credit scoring model for banks and credit companies, which can contribute to predicting the credit level of customers and reducing default loan generation.

Although the AdaBoost-DT model constructed in this paper has better classification effect than other models, the parameter optimization methods and the selection of base classifiers are not comprehensive enough. Future research should try more parameter optimization methods and try to implement more traditional classifiers as base classifiers. In addition, the credit scoring model proposed in this paper only targets the imbalance of data without considering the coexistence of data multidimensionality. Although the traditional classification method can obtain better classification results in low-dimensional data, it is more difficult to handle in high-dimensional data. Therefore, future research will consider classification models for high-dimensional, unbalanced data to improve the generalizability and application value of the methods in real life.

ACKNOWLEDGEMENT

This research was supported by Zhejiang Philosophy and Social Science Project under Grant 19NDJC203YB.

REFERENCES

[1] Li Runfa. (2020). CBRC: China's commercial banks' non-performing loan ratio 1.86% at the end of 2019. http://www.gov.cn/shuju/2020-02/17/content 5480190.html (in Chinese).

[2] Durand D. Risk Elements in Consumer Installment Financing[M]. New York: National Bureau of Economic Research, 1941: 78-129.

[3] Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. Journal of Finance,1968, 23: 589-609.

[4] Meyer P A, Pifer H. Prediction of bank failures[J]. Journal of Finance, 1970, 25: 853-868.

[5] Kar Y T, Melody Y K. Managerial Applications of Neural Networks: The Case of Bank Failure Predictions[J]. Management Science,1992, 38(7): 927-947.

[6] Lundy M. Credit Scoring and Credit Control[M]. New York: Oxford University Press, 1993: 25-36.

[7] Leong C K. Credit Risk Scoring with Bayesian Network Models[J]. Computational Economics, 2016, 47(3):423-446.

[8] Fang K N, Chen Z L. A credit scoring method based on semi-supervised generalizable additive logistic regression[J]. Systems Engineering Theory and Practice, 2020,40(02):124-134. (in Chinese).

[9] Sustersic M, Mramor D, Zupan J. Consumer Credit Scoring Models with Limited Data[J]. Expert Systems with Applications,2009,36(3):4736-4744.

[10] Pathak A N, Sehgal M, Francolin C. A study on fraud detection based on data mining using decision tree[J]. International Journal of Computer Science Issues, 2011, 8(3): 258.

[11] Pawiak P, Abdar M, Acharya U R. Application of New Deep Genetic Cascade Ensemble of SVM Classifiers to Predict

the Australian Credit Scoring[J]. Applied Soft Computing, 2019, 84: 105740.

[12] Bellotti T, Crook J. Support vector machines for credit scoring and discovery of significant features[J]. Expert Systems with Applications, 2009, 36(2):3302-3308.

[13] Bekhet H A, Eletter S F K. Credit risk assessment model for Jordanian commercial banks: Neural scoring approach[J]. Review of Development Finance, 2014, 4(1):20-28.

[14] Kambal E, Osman I, Taha M, et al. Credit scoring using data mining techniques with particular reference to Sudanese banks[C]// International Conference on Computing. IEEE, 2013.

[15] Michał W, Manuel G, Emilio C. A survey of multiple classifier systems as hybrid systems[J]. Information Fusion, 2014,16:3-17.

[16] Wang G, Ma J, Huang L, et al. Two credit scoring models based on dual strategy ensemble trees[J]. Knowledge Based Systems, 2012, 26:61-68.

[17] Finlay S. Multiple classifier architectures and their application to credit risk assessment[J]. European Journal of Operational Research, 2011, 210(2): 368-378.

[18] Xia Y, Zhao J, He L, et al. A novel tree-based dynamic heterogeneous ensemble method for credit scoring[J]. Expert Systems with Applications, 2020, 159:113615.

[19] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to Boosting[J]. Journal of Computer and System Sciences, 1999, 55(01): 119-139.

[20] Razia S, Rao A S. Machine learning techniques for thyroid disease diagnosis- A review[J]. Indian Journal of Science & Technology, 2016, 9(28): 1-9.