Credit Card Fraud Detection Using Feature Fusion-based Machine Learning Model

Qing Meng*

Hunan University, Juzizhou street, Yuelu District, Changsha City, Hunan Province, China 410082

*Corresponding author: 100784@yzpc.edu.cn

Abstract. Credit card fraud is a very common means of financial fraud at present. In order to reduce social and personal economic property losses, this study aims to propose a model to predict credit card fraud, so as to detect possible fraudulent transactions in the transaction process. In this paper, the data set to simulate fraudulent transactions be used, and construct the model after processing the characteristic data and selecting the features. In the model construction part, four algorithms are used as the trained classifiers, which are K Nearest Neighbor algorithm, Bgging algorithm, Logistic Regression algorithm and Gaussian Bayesian algorithm. First, the four algorithms are used to train the data set respectively, and then the parameters of each classifier are adjusted respectively, so that each individual classifier achieves the optimal performance. On this basis, a new model is constructed by using the method of model fusion. The classifier constructed by K Nearest Neighbor algorithm, Bagging algorithm and Gaussian Bayesian algorithm is used as the primary learner, and the classifier constructed by logistic regression algorithm is used as the secondary classifier to build a fused model as a whole. The final experimental results will give the classification performance of the five models, including individual models. The results show that the performance of the model after integrating the four individual models is better than that of the single model, which can provide accurate feedback for credit card holders on whether there is fraud.

Keywords: Credit card fraud detection, financial fraud, model fusion.

1. Introduction

With the increasing development of Internet digital economy, people's life has been convenient. In such an environment, virtual credit card technology is growing in popularity, and some malicious technologies have also surfaced. Among them, credit card fraud is a very common means of fraud that leads to people's economic losses. These malicious technologies play a role in credit cards used for financial transactions on online platforms. Credit card fraud will not only cause economic losses to cardholders, but also affect their reputation because they fail to repay in time. At the same time, it will also result in financial losses for banks, affect the effect of the whole society in controlling the excess liquidity of money, and disrupt the normal financial order. Therefore, it is very urgent and necessary to build a prediction model of credit card fraud. In order to detect such fraudulent transactions and reduce the losses of individuals, banks and society, this study uses machine learning related algorithms to analyze the historical transaction data in the data set, and constructs a machine learning model for credit card fraud detection to detect potential fraudulent transactions and reduce the adverse effects caused by fraudulent transactions.

In the existing research, mainly using one or more algorithms to build the detection model, and select one having the best performance as the most appropriate classifier by comparison. Chen et al. use the classification model of support vector machine to classify fraud and not fraud [1]. In addition, a variety of algorithms are used to compare to find a suitable algorithm for fraud detection [2]. And five classifiers are used to build models to detect fraudulent transactions, and the classification model most suitable for this situation is found after comparison [3]. While in study, the scheme adopted is unsupervised learning technology based on neural network to detect credit card fraud [4]. The model constructed in this way, whether using one or more algorithms, can only obtain the classification performance of the individual model. When the hyper parameters of the model are tuned to a certain extent, the classification performance of the model will not be improved. In fact, this does not make good use of each individual model to get the best classification effect after tuning parameters.

Therefore, based on the adjusted algorithm model, a new model can be constructed by model fusion, which enhances the classification performance of this model compared with other individual models. In this way, the classification performance of each model is fully utilized, and the classification performance is further improved.

In this regard, this study employed four algorithms namely Logistic Regression, Bagging, K-Nearest Neighbor, Gaussian Naive Bayes. First of all, considering that there are 23 features in the data set used, it is necessary to process the data and select the features. Then the hyper parameters of each classifier are adjusted to reduce the occurrence of abnormal conditions such as overfitting, so that the classification effect of each algorithm is the best. Finally, this study uses the method of model fusion to fuse the classification effects of the above four models, then get a model with better classification performance compared to four models alone. The fusion here uses two ways. The first is the most basic way, using the feature output generated by the previous classifier as the final total meta classifier input data; The second method is to set the relevant parameters in the fusion classifier, so as to use the category probability value generated by the first level basic classifier as the input of meta-classifier. Among them, the Area under Curve (AUC) score of logistic regression algorithm is 83.48%, that of bagging algorithm is 87.49%, about K-Nearest Neighbor is 78.44%, and of naive Bayesian algorithm is 72.20%, while the AUC score of the first fused model is 90.02% and the second got 92.14%. It can be seen that the classification performance of the model is improved after using the method of model fusion, and the effect of the second fusion model is better than that of the first.

2. Methods

2.1. Dataset description and preprocessing

Table 1. Sample data in the collected dataset

	Item 1	Item 2	Item 3	Item 4	Item 5
trans_date_trans_time	2020/6/21 12:14:00	2020/6/21 12:14:00	2020/6/21 12:15:00	2020/6/21 12:15:00	2020/6/21 12:16:00
cc_num	2291160000000000	3598220000000000	3591920000000000	30407700000000	3596360000000000
merchant	fraud_Kirlin and Sons	fraud_Swaniawski,	fraud_Haley Group	fraud_Daugherty LLC	fraud_Goyette, Howell
category	personal_care	health_fitness	misc_pos	kids_pets	shopping_pos
amt	2.86	41.28	60.05	19.55	4.37
first	Jeff	Ashley	Brian	Danielle	David
last	Elliott	Lopez	Williams	Evans	Everett
gender	M	F	M	F	M
street	351 Darlene Green	9333 Valentine Point	32941 Krystal Mill Apt.	76752 David Lodge Apt. 064	4138 David Fall
city	Columbia	Bellmore	Titusville	Breesport	Morrisdale
state	SC	NY	FL	NY	PA
zip	29209	11710	32780	14816	16858
lat	33.9659	40.6729	28.5697	42.1939	41.0001
long	-80.9355	-73.5365	-80.8191	-76.7361	-78.2357
city_pop	333497	34496	54767	520	3688
job	Mechanical engineer	Librarian, public	Set designer	Psychotherapist	Advice worker
dob	1968/3/19	1970/10/21	1987/7/25	1991/10/13	1973/5/27
trans_num	2da90c7d74bd46a0caf3777415b3ebd3	c81755dbbbea9d5c77f094348a7579be	2159175b9efe66dc301f149d3d5abf8c	798db04aaceb4febd084f1a7c404da93	71a1da150d1ce510193d7622e08e784e
unix_time	1371816865	1371816893	1371816915	1371816937	1371816970
merch_lat	33.986391	40.49581	28.812398	41.747157	41.546067
Merch_lon	-81.200714	-74.196111	-80.883061	-77.584197	-78.120238
is_fraud	0	0	0	0	0

The study used a simulated credit card transaction data set containing legal and fraudulent transactions from January 1, 2019 to December 31, 2020. The data set covers the credit cards of 1000 customers who transact with 800 merchants. This data set contains 22 features. The overall data is unbalanced, with most of the transactions are legitimate transactions and a few are fraud transactions. The source of the dataset is come from [5]. There are 1.5 million items in the dataset, to be more specific, 500 thousand and 1million items in fraudTest.csv and fraudTrain.csv files respectively. The study merged fraudTest.csv and fraudTrain.csv datasets, select 1 million data as the dataset, and then used the function of randomly dividing training set and test set to divide them in proportion. The training set has 800, 000 rows of data, while the validation set, and test set have 10, 000 rows of data respectively. A part of data samples is shown in Table 1.

2.2. Employed machine learning algorithms

2.2.1. Logistic Regression

Logistic Regression is derived from linear regression [6, 7]. Linear regression is generally used to predict continuous data, and rarely or even not used to classify discrete data. To realize that the regression model can classify discrete data, there is a regression model called logistic regression. In order to achieve the performance of classification, sigmoid function is usually used. Sigmoid function shown in Figure 1 can map any real number to (0, 1), so that it can be used to convert any value function into a function more suitable for binary classification, which is also applicable to the classification of fraud and not fraud required by the study.

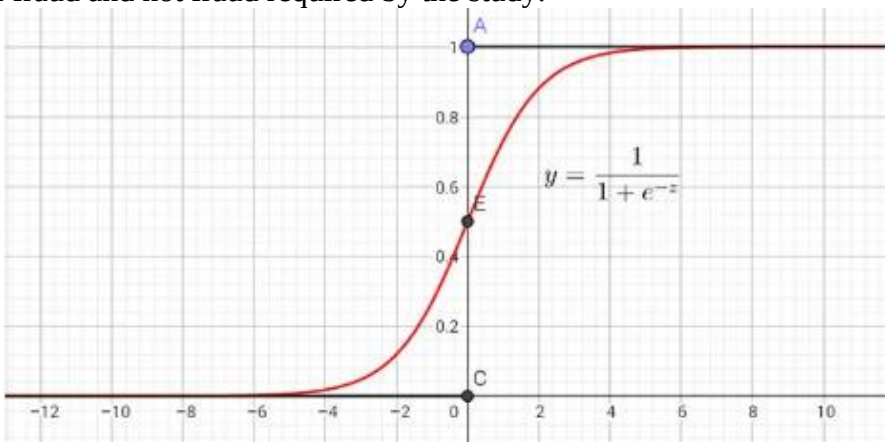

Figure 1. The schematic diagram of Sigmoid function.

The most used method in Logistic Regression is the gradient descent method to solve the minimum value of the loss function. The hyperparameter max_iter represents the maximum number of steps that can be taken for gradient descent. And hyperparameter C is a super parameter used to control the regularization. The smaller C is, the loss function will be larger, the heavier the penalty of the model on the loss function will be, the regularization effect will be stronger, and the parameters will be gradually compressed smaller and smaller. The last important hyperparameter is class_weight, which is used to represent various types of weights.

2.2.2. Bagging

The bagging algorithm assumes that there is a data set D, and K data subsets are obtained by using the method of Bootstrap sample (with put back random sampling) (the number of subset samples is equal): D1, D2…DK, as a new training set, use this k subset to train a classification model respectively (using classification, regression and other algorithms), and finally get k classification models [8]. Input the test data into these K classifiers and will get k classification results. For example, the classification results are 0 and 1. Which of these K results accounts for the most, then the prediction result is which. Figure 2 shows the process of bagging algorithm.

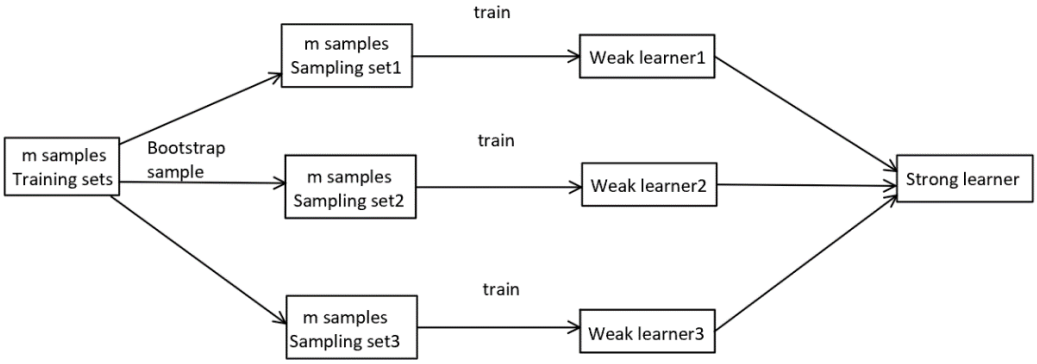

Figure 2. Bagging algorithm flow chart

There are three main hyperparameters in Bagging algorithm. First is n_estimators, representing the basic estimator in the set. Second is max_samples, indicating the number of samples taken from X to train each basic estimator. And the last is max_features, which represents the number of elements drawn from X to train each basic estimator.

2.2.3. K-Nearest Neighbor

The core idea of KNN algorithm is that birds of a feather flock together and people flock together. Assuming that most of the k most similar (i.e. the nearest) samples of a sample in the feature space belong to a certain category, the sample also belongs to this category, which is widely in various tasks [9, 10]. In other words, this method only determines the category of the samples to which it belongs according to the category of the nearest one or several samples in the decision-making of determining the category.

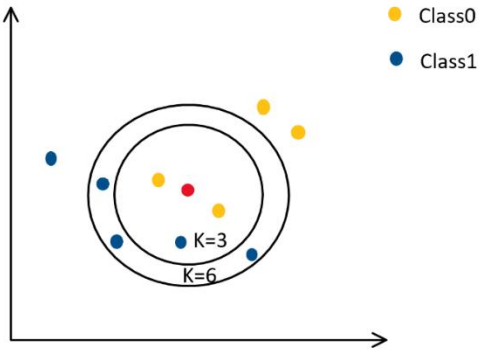

Figure 3. Hyperparameter K

There are two main parameters in KNN algorithm. One of them is K (as Figure 3 shows), indicating the nearest K points; The other is distance measurement, including Euclidean distance, Manhattan distance and Cosine distance.

2.2.4. GaussianNB

Gaussian Bayesian algorithm assumes that the conditional probability of each feature dimension of the sample obeys the Gaussian distribution, and then calculates the posterior probability of the new sample belonging to each category under a certain feature distribution according to the Bayesian formula. Finally, the category of the sample is determined by maximizing the posterior probability.

2.2.5. Proposed method based on model fusion

Here, the stacking method is used for model fusion. Stacking is a set learning technology, which merges multiple classification models through using meta classifiers. Training each classification model based on the complete training set; Then, the meta classifier is fitted based on the overall output meta features of each classification model. The meta classifier can be trained according to the probability in the predicted class label or set. Figure 4 shows the process.

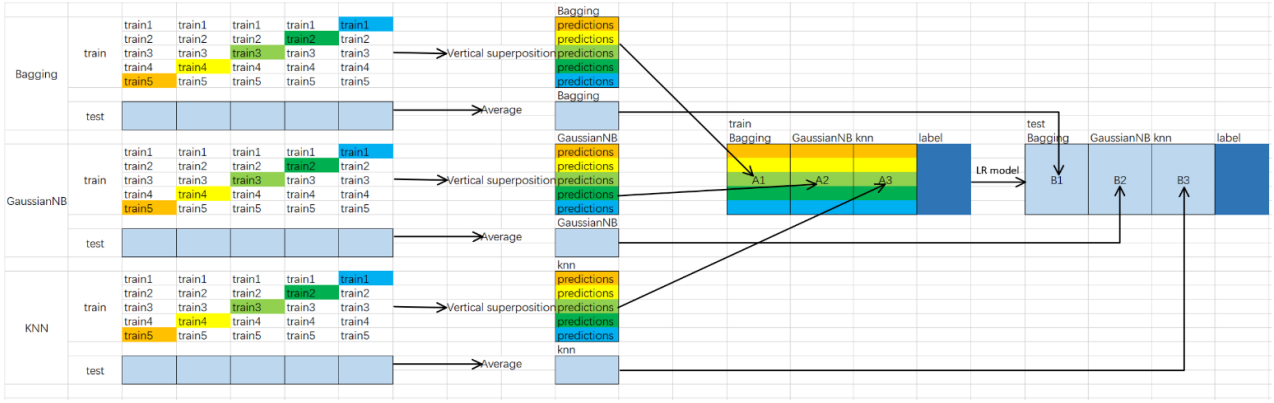

Figure 4. The process of combining

a. First, the training set will be divided into five parts: train1, train2, train3, train4, train5.

b. Three base models are used: bagging, Gaussian Nb and KNN. For example, in the bagging model part, one of the five parts of the training set are used as the validation set, and other four parts are used as the training set to conduct 5-fold cross validation for model training; Then using the test set for prediction. Using this approach, five predictions trained by Bagging algorithm on the training set and one prediction B1 on the test set will be obtained. Combine these five vertical overlaps to get A1. GaussianNB and KNN models are partially the same.

c. After the training of the three base models, take the predicted values of the three models on the training set as the three "characteristics" A1, A2, A3 respectively, and using the logistic regression model for training to establish the logistic regression model.

d. Using the trained logistic regression model, predict on the values (B1, B2, B3) of the three "features" constructed by the predicted values on the test set before the three base models to obtain the final prediction category or probability.

The specific parameters are provided by Table 2.

Table 2. The specific parameters of the final moel

Bagging		KNN		LogisticRegression		
n_estimator	max_samples	n_neighbors	metric	C	max_iter	class_weight
25	0.3	5	manhattan	0.1	30	balanced

3. Result and discussions

Table 3. The performance of different model

Model	AUC Score(Training set)	AUC Score(Test set)
Logistic Regression	84.95%	83.48%
Bagging	91.50%	87.49%
GaussianNB	73.27%	72.20%
KNN	80.32%	78.44%
Stacking	92.76%	90.02%

By tuning the hyperparameters of each basic model, the study set up five experiments to test the data set, solving the problem of overfitting. The results of each group of experiments shown in Table 3 are based on the same preprocessed data. First, build the optimal basic model, and then use the hyper parameter values of the basic model after adjusting the parameters to build the fusion model. The best neutral performance of the basic model is Bagging algorithm, which is 87.49%, followed by Logistic Regression, which is 83.48%. Gaussian Bayesian algorithm and KNN method both perform poorly, with AUC score of 72.20 percent and 78.44 percent, respectively. The performance of the final model is better than that of the basic model, and the AUC value is 90.02%.

The performance improvement of individual models is mainly through parameter adjustment, so using a single model for further performance improvement will have certain limitations. Through

model fusion, individual models can be fused at the model level, and the resulting model has further performance improvement. In the process of model fusion, with the increase of the number of base classifiers in the combination, the error rate of the combination will decline exponentially and finally tend to zero. At the same time, in order to improve the performance of the fusion model, the performance and diversity of base classifiers should be improved. The performance is mainly reflected in the tuning of each individual learner. The diversity is reflected in the fact that the correlation between individual models should be as small as possible, which is also the reason why the above four individual models are selected in this study. The model fusion method used in the study is stacking. Stacking first trains primary learners through the initial data set and then generates a data set for training secondary learners, which is regarded as the input data of the secondary learner, while the label of the initial sample is still regarded as the example label. By improving the performance of individual models and the selection of individual models, the performance of the final fusion model has been improved.

4. Conclusion

In this study, four individual models are established and four basic models are obtained after parameter adjustment and optimization. Then a fused model is proposed according to the stacking method. Four individual models are established by using four algorithms: Logistic Regression, Bagging, Gaussian Bayes and K nearest neighbor, and then the performance of the model is improved by model fusion. After many adjustments and a large number of experiments, the optimal models of the four selected algorithms are established. On this basis, Bagging, Gaussian Bayes and KNN models are used as the most out of base learners, and the logistic regression model is used as the secondary learner to build the fusion model, which increases the performance of the final model to 90.02%. In the future, this research plans to combine other model fusion methods, such as the fusion methods at the result level and the feature level, so that models with better performance can be trained.

References

[1] Chen, R. C., et al. "Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines," Intelligent Data Engineering & Automated Learning-ideal, International Conference, Exeter, Uk, August. DBLP, 2004.

[2] Mjm, A., et al. "Exploratory Analysis of Credit Card Fraud Detection using Machine Learning Techniques," Global Transitions Proceedings, 2022.

[3] Sahu, A., et al. "A Dual Approach for Credit Card Fraud Detection using Neural Network and Data Mining Techniques," 2020 IEEE 17th India Council International Conference (INDICON). IEEE, 2021.

[4] Rai, A. K., et al. "Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020.

[5] Kaggle, "Credit card fraud detection", https://www.kaggle.com/code/gopibollineni/credit-card-fraud-detection, 2022.

[6] Levy, A., et al. "Credit risk assessment: a comparison of the performances of the linear discriminant analysis and the logistic regression," Post-Print, 2021.

[7] LaValley, M. P., "Logistic regression." Circulation 117.18 (2008): 2395-2399.

[8] Quinlan, J. R., "Bagging, boosting, and C4. 5." In Aaai/Iaai, vol. 1, pp. 725-730.

[9] Yu, Q., et al. "Clustering Analysis for Silent Telecom Customers Based on K-means++." 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 1. IEEE, 2020.

[10] Limam, H., "A New Hybrid Multiclass Approach Based on KNN and SVM." Journal of Information & Knowledge Management (2022): 2250061.