Online Transaction Fraud Detection System Based on Machine Learning

Online Transaction Fraud Detection System Based on Machine Learning

Bocheng Liu, Xiang Chen* and Kaizhi Yu

School of Software, Nanchang University, Nanchang 330047, China

*Corresponding author e-mail: bcliu@ncu.edu.cn

Abstract. With the rapid development of Internet technology, the scale of online transactions is constantly expanding. At the same time, the related network transaction fraud problem has become more significant. Compared with the credit card transaction, the network transaction has the characteristics of low cost, wide coverage and high frequency, which makes the detection of fraud more complex. Aiming at the problem of difficult fraud detection in network transactions, this paper designed two fraud detection algorithms based on Fully Connected Neural Network and XGBoost, whose AUC values can achieve 0.912 and 0.969 respectively. Meanwhile, we designed an interactive online transaction fraud detection system based on XGBoost model, which can automatically analyze the transaction data uploaded and return the fraud detection results to users.

Keywords: Fraud Detection, Fully Connected Neural Network, Xgboost

1. Introduction

So far, mobile payment has become one of the mainstream payment methods. Thousands of transactions are carried out on the online trading platform all the time. The popularity of network transactions provides some criminals with the opportunity to commit crimes. Personal property in the complex network environment has the risk of theft, which not only damages the interests of consumers, but also seriously affects the healthy development of the network economy. Therefore, the transaction fraud detection is one of the key tools to solve the problem of network transaction fraud [1].

Traditional fraud detection mostly adopts statistical and multi-dimensional analysis techniques. Since they are verification techniques, it is difficult to obtain the laws hidden behind the transaction data. The big data technology and machine learning algorithm provide efficient detection methods for transaction fraud detection [2]. Compared to the traditional statistical methods, machine learning can represent important features through a large amount of data, which cannot be described by the former. By using the corresponding machine learning method, we can establish a model based on the existing transaction data to realize the detection of network transaction fraud, so as to reduce the loss caused by fraud. In 2018, Zhaohui Zhang proposed a reconstructed feature convolutional neural network prediction model applied to transaction fraud detection, which has better stability and availability in

classification effect compared with other convolutional neural network models [3]. However, there is also a problem that the detection accuracy is not high enough due to the imbalance of sample labels.

Combined with requirements, this paper proposed two fraud detection algorithms based on Fully Connected Neural Network and XGBoost. The former algorithm integrated two neural network models with different cross entropy loss functions, and the design process of the integrated model is quick and convenient. The latter algorithm used Hyperopt to optimize the XGBoost classifier, so that the model can be constructed with the best parameters and have a high performance of fraud detection. The two algorithms have different application scenarios. In order to ensure the good detection performance, we decided to use XGBoost model to build an online transaction fraud detection system. This system has obvious advantages in running time and applicability, and can accurately predict the fraud probability of network transaction behaviors.

2. System Design

The system first acquires the raw transaction data uploaded by users, and gets the specified data by data preprocessing. Then use the fraud detection algorithm based on XGBoost to predict the probability that the transaction is a fraud transaction. If the probability detected exceeds the set threshold (This system is set to 0.85), a warning will be issued immediately. Finally, the detection results will be saved in the database and returned to the user. The System Flow Chart is shown in Figure 1.

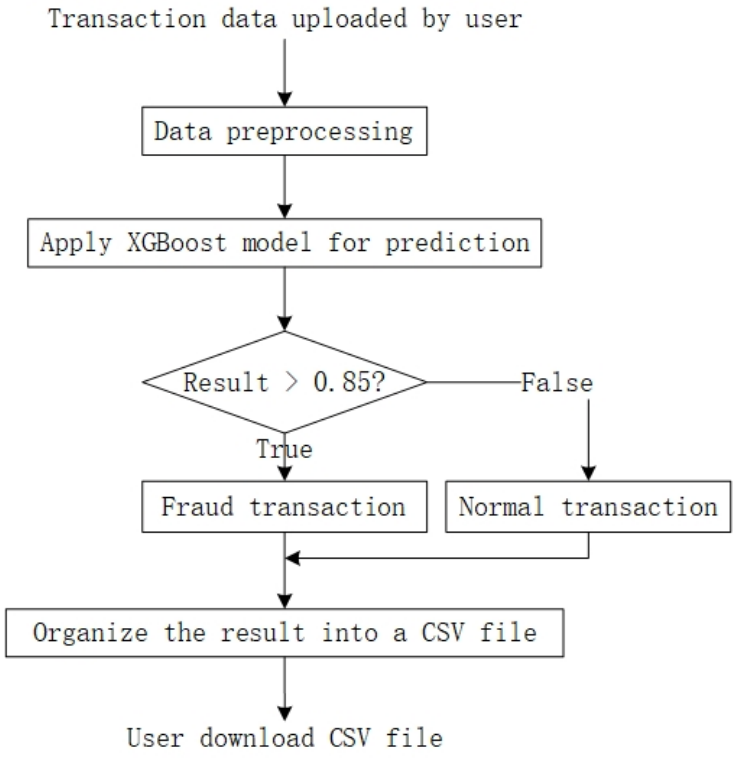

Figure 1. System Flow Chart

3. System Realization

3.1. Fraud Detection Algorithm Based On Fully Connected Neural Network

The fraud detection algorithm based on Fully Connected Neural Network consists of three steps: feature selection, data preprocessing and neural network construction [4].

(1) Feature selection: The experiment contains two types of data sets: Transaction and Identity, which comes from the payment service company Vesta. Since selecting all the features did not

significantly improve the model, we decided to train the neural network model only using the features of the Transaction data set.

(2) Data preprocessing: Features in the Transaction data set can be divided into continuous features and categorical features. For continuous features, we use the logarithmic transformation first to make the processed data conform to the standard normal distribution. Secondly, Z-score standardization is carried out on the data, so that each dimension is dimensionless to avoid the huge influence of different dimensional selection on the distance calculation. For categorical features, the One-Hot Encoding is applied to generate feature vectors, but only for the top 50 most common values of each feature to reduce its sparsity.

(3) Neural network construction: In this step, we use Keras to build the Fully Connected Neural Network. The network structure is shown in Figure 2.

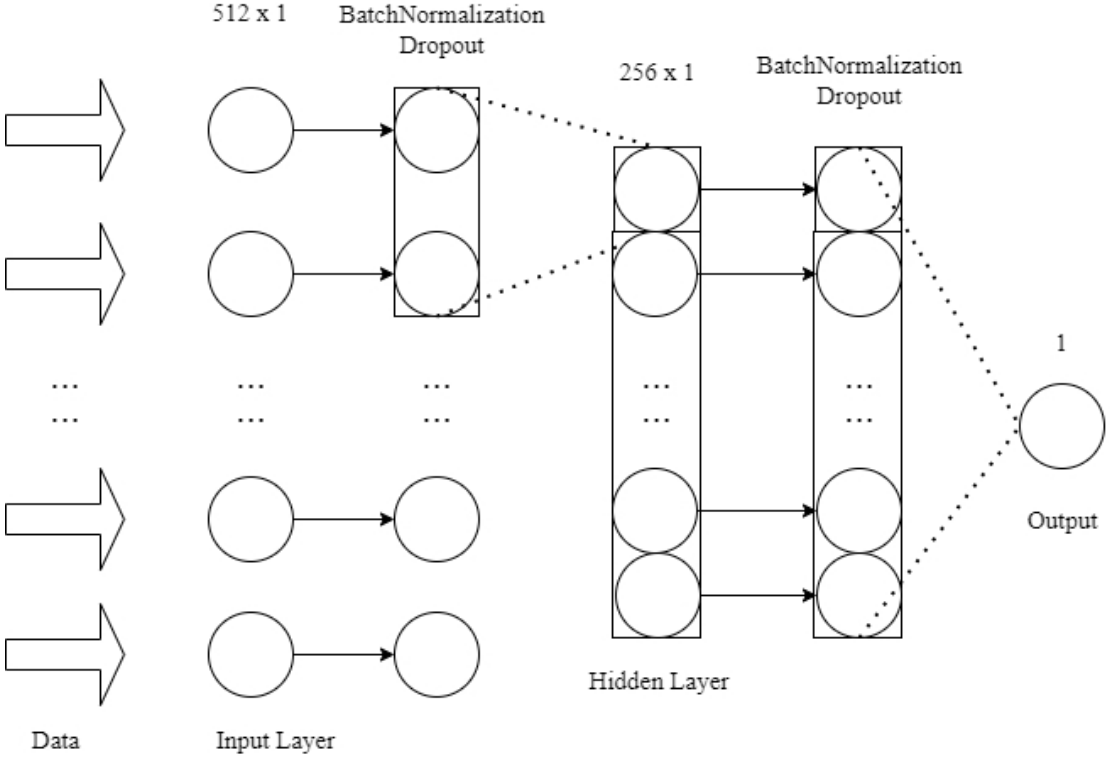

Figure 2. Overall network structure

The activation function used in the input layer and hidden layer of the neural network is GELU. GELU adds statistical characteristic on the basis of ReLU, which has faster convergence speed and higher learning efficiency, and can effectively avoid the problem of gradient disappearance [5, 6].

For the selection of the cross entropy loss function, we can choose either the categorical cross entropy (CE) or the binary cross entropy (BCE). Since we are discussing the binary classification problem, BCE is chosen as the cross entropy loss function. The formula of BCE is as follows.

$$BCE(x)_i = -[y_i \log f_i(x) + (1 - y_i) \log(1 - f_i(x))] \tag{1}$$

In addition, in order to solve the problem of the serious imbalance of the positive and negative sample ratio in the data set, we also adopt the Focal loss (FL) as cross entropy loss function to conduct experiment [7]. The formula of Focal loss is as follows.

$$FL = \begin{cases} -\alpha(1-p)^{\gamma} \log(p), & if\ y = 1 \\ -(1-\alpha)p^{\gamma} \log(1-p), & if\ y = 0 \end{cases} \tag{2}$$

Different cross entropy functions are used to construct the neural network, and we divide 80% of the data set as the training set to train the model. In the case of unbalanced positive and negative samples, taking Accuracy as an evaluation standard has great defects, so we use AUC value, Loss value of training set and Loss value of test set as standards to evaluate the model. The comparison between two models in different training epochs is shown in Table 1.

Table 1. Performance of models in different training epochs

Training epochs	Binary cross entropy-based			Focal loss-based		
	AUC	Loss	Val_loss	AUC	Loss	Val_loss
2	0.8759	0.1006	0.1028	0.8500	0.0132	0.0109
4	0.8876	0.0845	0.0995	0.8739	0.0108	0.0106
6	0.8934	0.0735	0.0981	0.8850	0.0096	0.0104
8	0.9012	0.0636	0.0961	0.8898	0.0080	0.0099

By comparison, we can find that the AUC value of model BCE is higher, while the loss value of model FOCAL is smaller. By using Spearman correlation coefficient, we can calculate that the prediction correlation of the two models is about 80.3%, which has good integration performance. Therefore, we tried to integrate the two models [8]. The comparison between the integrated model and the two basic models is shown in the Table 2.

Table 2. The Performance of the two base models and integrated model

Model	Binary cross entropy-based	Focal loss-based	Integrated model
AUC	0.9012	0.8898	0.9117

3.2. Fraud Detection Algorithm Based On XGBoost

The fraud detection algorithm based on XGBoost also includes three steps: data preprocessing, parameter tuning and model construction [9].

(1) Data preprocessing: Unlike the algorithm based on Fully Connected Neural Network, this algorithm selects all features of both Transaction and Identity data sets as the basis for further processing.

For categorical features, we use LabelEncoder to convert discrete data to numbers between 0 and N −1. In addition, since there are 339 features of the V series, we choose PCA to reduce the dimension of 339 features and reconstruct 30 new features.

(2) Parameter tuning: After data preprocessing, we use Hyperopt to tune the parameters of the model [10]. Hyperopt is an implementation of Bayesian optimization that can quickly and efficiently implement tuning parameters of the XGBoost classifier. Here are the specific steps for using Hyperopt.

Step1: Define an objective function. For the function fmin in the Hyperopt module, it accepts an objective function and a configuration space, and it tries to minimize the objective function, which means finding the input value in the configuration space that generates the minimum output. In the objective function, in order to overcome the problem of unbalanced sample labels, we use StratifiedKFold to divide the training set into 7 folds, and train the model constructed with passed-in parameters 7 times. Each time AUC value is used as standard to score the output of the classifier. After training, the average of the seven scores is calculated as the output of objective function.

Step2: Define a configuration space. A configuration space object describes the domain over which Hyperopt is allowed to search. Here we specify the range for important parameters such as gamma, learning_rate, max_depth, etc.

Step3: Training model. In the last step, we set the parameter max_evals of the function fmin to 25, which represents 25 rounds of parameter tuning. Then get the best parameters through the function space_eval. Table 3 lists the values of the best parameters for the XGBoost classifier.

Table 3. Best parameters for the XGBoost classifier

Parameter	Colsample_bytree	Gamma	Learning_rate	Max_depth	Subsample
Value	0.856	0.224	0.060	14	0.90

The 7 fold cross-validation performance of the XGBoost classifier constructed with the best parameters is shown in the Table 4.

Table 4. Performance of the XGBoost classifier

CV Round	1	2	3	4	5	6	7	Average
AUC	0.968	0.971	0.966	0.968	0.970	0.969	0.971	0.969

(3) Model construction: Construct the XGBoost classifier with the best parameters, and use all of the data to train the classifier, so that the final model is obtained.

3.3. Algorithm Evaluation and Comparison

The comparison of the two fraud detection algorithms is shown in Table 5.

Table 5. Comparison of detection algorithms

Evaluation criteria \ Detection algorithm	Fully Connected Neural Network	XGBoost classifier
AUC	0.912	0.969
Model training speed	quick	slow
Model Predicting time (5 million pieces, unit: s)	74.65	52.83

From the perspective of accuracy, XGBoost performs better and uses shorter time for prediction. Moreover, the input dimension of Fully Connected Neural Network depends on the specific data set. Once the data set changes, the structure of the neural network needs to be adjusted immediately. However, from the perspective of training speed, the Neural Network can achieve a fairly good level of performance, while the XGBoost classifier takes plenty of time to tune the parameters. Therefore, the fraud detection algorithm based on Fully Connected Neural Network is suitable for scenes with high time requirements, and the fraud detection algorithm based on XGBoost is suitable for scenes with high precision requirements.

3.4. Web Development Technology

The transaction fraud detection system finally provides services in the form of Web. In order to ensure the detection accuracy, we built an online detection platform based on XGBoost model with Django framework. Once user uploads the CSV file containing transaction information and identity information, the system behind the platform will preprocess the data and use the trained XGBoost model to predict the probability of fraud transactions. The detection results will be sorted into the CSV file and returned to the user for download. At the same time, the detection history will be saved in the database.

4. Conclusions

With the diversification of online transactions, machine learning is applied to more and more anti-fraud processing tasks. This paper proposed two fraud detection algorithms based on Fully Connected Neural Network and XGBoost, and designed an online transaction fraud detection system base on XGBoost classifier.

The specific work is as follows:

(1)Generate new features through feature combination, feature decomposition and algebraic operation, and add model input of effective features.

(2)Propose the detection algorithm based on Fully Connected Neural Network. This algorithm integrates neural networks using different cross entropy loss functions, which can effectively mine information of various features in a short time.

(3)Propose the detection algorithm based on XGBoost. This algorithm constructs the XGBoost classifier with best parameters by using Hyperopt. The AUC of this classifier can reach 0.969, which highly improves the fraud detection performance of network transactions.

(4)Design the online fraud detection system based on XGBoost classifier. The system can accurately predict the fraud probability of network transaction behavior and return the result to users. In the real situation, the detection system can be directly embedded into the interface of online transaction, and predict the transaction before the user pays, so as to take the initiative to intercept the fraudulent behavior.

Acknowledgements

This work was supported by Student innovation and entrepreneurship training program (No.2020CX166).

References

[1] E. Kurshan, H. Shen: Graph Computing for Financial Crime and Fraud Detection: Trends, Challenges and Outlook. 2020 Second International Conference on Transdisciplinary AI (TransAI), 2020.

[2] Dheepa V, Dhanapal R: Analysis of credit card fraud detection methods. International journal of recent trends in engineering, 2009, 2(3):126.

[3] Zhang Z, Zhou X, Zhang X, et al: A model Based on Convolutional Neural Network for Online Transaction Fraud Detection. Security and Communication Networks, 2018, 10(2):1-9.

[4] Jain V: Perspective analysis of telecommunication fraud detection using data stream analytics and neural network classification based data mining. International Journal of Information Technology, 2017, 9(3):303-310.

[5] Hendrycks D, Gimpel K: Gaussian Error Linear Units (GELUs), 2016.

[6] Glorot X, Bordes A, Bengio Y: Deep Sparse Rectifier Neural Networks. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), 2011, 15:315-323.

[7] Lin T Y, Goyal P, Girshick R, et al: Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 42:2999-3007.

[8] Prusti D, Rath S K: Fraudulent Transaction Detection in Credit Card by Applying Ensemble Machine Learning techniques. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2019.

[9] Lei, Shimin, et al: An Xgboost based system for financial fraud detection. E3S Web of Conferences, 2020, 214(2).

[10] Bergstra J, et al: Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. Computational Science & Discovery, 2015, 8(1).