Computing and Informatics, Vol. 39, 2020, 1082–1098, doi: 10.31577/cai_2020_5_1082

IMPROVED DEEP FOREST MODE FOR DETECTION OF FRAUDULENT ONLINE TRANSACTION

Mian Huang

The Key Laboratory of Embedded System and Service Computing of Ministry of Education
Tongji University
Shanghai, China
e-mail: net-cn@163.com

Lizhi Wang, Zhaohui Zhang*

School of Computer Science and Technology
Donghua University
Shanghai, China
e-mail: lizhi_wang2769433@163.com, zhzhang@dhu.edu.cn

Abstract. As the rapid development of online transactions, transaction frauds have also emerged seriously. The fraud strategies are characterized by specialization, industrialization, concealment and scenes. Anti-fraud technologies face many challenges under the trend of new situations. In this paper, aiming at sample imbalance and strong concealment of online transactions, we enhance the original deep forest framework to propose a deep forest-based online transaction fraud detection model. Based on the BaggingBalance method we propose, we establish a global sample imbalance processing mechanism to deal with the problem of sample imbalance. In addition, the autoencoder model is introduced into the detection model to enhance the representation learning ability. Via the three-month real online transactions data of a China's bank, the experimental results show that, evaluating by the metric of precision and recall rate, the proposed model has a beyond 10 % improvement compared to the random forest model, and a beyond 5 % improvement compared to the original deep forest model.

Keywords: Deep forest, online transaction, fraud detection, autoencoder

* Corresponding author

1 INTRODUCTION

Under the general trend of Internet finance, digital technologies such as big data and artificial intelligence (AI) are widely used in the financial field, and the volume and potential development of financial markets are gradually enlarged. At the same time, the risk of exposure is also increasing, and frauds are endless [1]. According to statistics [2], China's fraudulent employees exceed 1.5 million, and the raised annual output value reaches 100 billion in 2017. The financial institutions that use Internet financial technology to carry out financial business are one of the main targets of the attack. The risk control of digital finance faces enormous challenges.

At the background, detecting fraudulent transaction patterns precisely is a highly important research direction in the field of online transaction fraud detection. The traditional expert rule-driven fraud detection technologies require a lot of manual operations, have a high application cost and low efficiency, while the traditional anti-fraud technologies consider simple transaction dimensions, thus they are difficult to form a multi-dimensional user portrait for the user. The online transactions have strong real-time performance, large amount of data, and fraud is characterized by small amount and high frequency. It is challenging for traditional anti-fraud methods to precisely detect fraudulent online transactions.

At present, a large number of machine learning (ML) – based research are widely used in the field of fraud detection, including decision trees [3], support vector machines (SVM) [3], naive bayes [4], random forest (RF) [4,5] and other ML algorithms. ML technology learns existing fraud strategies and explores potential fraud strategies by learning historical transaction information for online transactions, then precisely detects online transactions with fraudulent possibilities. In addition, some research about deep learning (DL) techniques are gradually being used in fraud detection tasks.

DL techniques [6] such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have achieved excellent performance in many popular tasks, such as image recognition [7], natural language processing [8], and so on. DL techniques excel in processing high-dimensional data and nonlinear feature space inputs, which are common in fraud detection tasks. On this basis, some studies begin to introduce DL for fraud detection, and use its powerful representation learning ability to solve the problem of online transaction fraud detection. A research from McKinsey concluded that it is a promising solution to apply DL techniques for the problem of financial fraud detection [9]. However, using only ML techniques or DL techniques does not completely solve the problem of fraud detection [10]. Therefore, integrating the advantages of ML and DL for fraud detection tasks has also become one of the research directions.

Deep forest (multi-Grained Cascade Forest, gcForest) is a novel decision tree ensemble method, which may open the door towards an alternative to deep neural networks for many tasks [11]. By creating a cascade forest structure, the method could enable its representation learning. At the same time, its multi-scanning structure could enhance its representational learning ability. From another perspective,

gcForest is a learning framework that integrates ML and DL techniques. The multi-scanning structure uses the idea of 1D and 2D convolution similar to CNN to establish representation learning. Based on the idea of stacking, cascade forest structure ensembles the RF model [12] and the completely random forest model [13] as base classifiers.

In this paper, we propose an improved gcForest-based online transaction fraud detection model. In view of the problems in online transaction fraud detection, on the one hand, we add the autoencoder DL model [7] to the multi-scanning structure to enhance its representation learning ability, because autoencoder could produce more concise unsupervised representations, which is proved to be a robust algorithm [14]. At the same time, we use XGBoost (eXtreme Gradient Boosting) model [15] to replace the completely random forest base classifiers in cascade forest structure. By combining with the proposed BaggingBalance method, a global sample imbalance processing mechanism is established. XGBoost is a scalable end-to-end tree boosting system [15], which is used widely to achieve state-of-art results on many ML competitions [16]. By combining the above methods, we enhance the original gcForest framework, then establish a detection model for online transaction fraud. The main contributions of this paper are summarized as follows:

- Apply gcForest model and improve the model for fraud detection in online transactions. Based on the accumulated experience of ML in fraud detection tasks in recent years, and with the excellent representation learning ability demonstrated by DL, the structure of the original gcForest is improved for the online transaction fraud detection task, and the experiment result shows the proposed model is superior to RF model and the original gcForest model.
- Aiming at the data characteristics of online transactions, the multi-scanning structure of the original gcForest is enhanced. Autoencoder model with excellent representation learning ability is introduced to enhance the model's feature learning of online transactions.
- The BaggingBalance method is proposed to deal with the sample imbalance problem in online transactions on the data input. At the same time, the XGBoost model is introduced in the cascade forest structure. Combined with the two, a global sample imbalance processing mechanism is established.

In the remainder of the paper, Section 2 describes some related work about status quo of the online transaction fraud detection. Section 3 introduces the methodology proposed in this paper. The data information and experimental results are discussed in Section 4. Finally, conclusion and future work are presented.

2 RELATED WORK

Nowadays, with the continuous development of Internet finance, online transaction fraud detection has become a hot research topic, including credit card fraud de-

tection, mobile payment fraud detection, B2C (Business-to-Customer) transaction fraud detection and so on.

The ML-based fraud detection algorithm is widely used in the field of online transaction fraud detection, including supervised learning model and unsupervised learning model. The supervised learning models establish a fraud detection model based on historical transaction data after manual investigation to determine whether a new transaction is fraudulent. For example, Shiyang Xuan et al. [5] learn the behavior patterns of normal and abnormal transactions via two kinds of RFs, where the two RFs have different base classifiers, and evaluate their performance on credit card transactions. While unsupervised learning models typically treat identified outliers as detected fraudulent transactions using outlier detection or anomaly detection techniques. In 2014, Olszewski [17] uses the self-organizing map (SOM) method to build a user behavior model to look for outliers that deviate from normal user behavior for fraud detection. ML-based detection algorithms have the advantages of learning known fraud patterns and detecting potential new fraud strategies. However, the methods of supervised learning strongly rely on the correctness of the original labels and the need to deal with the existing sample imbalance. Unsupervised learning is very sensitive to the overlapping distribution of normal transactions and fraudulent transactions, which often leads to a serious decline in accuracy [18].

With the excellent performance of DL technology in many classification tasks, DL technology is introduced in the field of online transaction fraud detection. In 2016, Kang Fu et al. [19] propose a CNN-based fraud detection framework, which could learn fraud behavior patterns via transaction data and show its excellent performance compared with some state-of-art methods. In 2017, Jingdong Finance's Shuhao Wang et al. [20] present CLUE framework, a novel DL-based transaction fraud detection system. By using neural network based embedding and RNN, the system achieves over 3 times improvement over the existing fraud detection approaches on real production data for eight months. In 2018, Zhaohui Zhang et al. [21] apply CNN for the task of online transaction fraud detection by constructing an input feature sequencing layer to obtain various input feature patterns, the proposed method outperforms the existing CNN model. At the same year, Abhimanyu Roy et al. [22] deeply study the application of DL technologies in credit card fraud detection tasks, and solve the common problems in fraud by using high-performance distributed cloud computing environment, while providing a parameter adjustment framework for DL topology. However, although DL technology can acquire more sequential information between transactions, it is insufficient for DL to just learn feature information within a single transaction, which can be well learned by ML technologies. But only using ML methods would attenuate the sequential learning ability of detection models [10].

In recent years, the online transaction fraud detection field starts to apply detection techniques that combine the advantages of ML and DL. In 2017, Xurui Li et al. [10] propose a novel "within-between-within" (WBW) sandwich-structured sequence learning architecture by integrating ensemble and DL methods, and intro-

duce attention mechanism to further enhance its performance. In the same year, Zahra Kazemi and Houman Zarrabi [23] use deep autoencoder model and softmax network to learn credit card transaction information and establish a fraud detection model, where results show the advantages of proposed method comparing to state-of-art methods.

In this work, based on the framework of the original gcForest, we improve the model for the online transaction fraud detection task. Introducing the autoencoder model into the multi-scanning structure enables the detection model a stronger representation learning ability on the input of the cascade forest, which could better handle the strong concealment of online transaction fraud patterns. While establishing a global sample imbalance processing mechanism, which could deal with the problem of sample imbalance in online transaction fraud detection. On this basis, we propose an improved gcForest-based method for online transaction fraud detection.

3 METHODOLOGY

3.1 Improved gcForest Framework for Detecting Fraudulent Online Transaction

The improved gcForest-based online transaction fraud detection framework can be seen in Figure 1, including the multi-scanning and the cascade forest. The cascade forest structure uses XGBoost as the base classifier to replace the completely random forest model in the original gcForest. As for the multiscanning structure, it introduces the autoencoder model into the original structure to enhance representational learning, then reconstruct a multi-scanning structure based on autoencoder combined with sample imbalance processing method BaggingBalance.

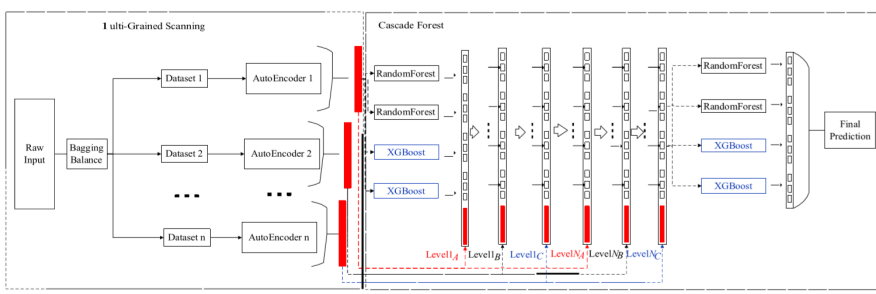


Figure 1. Improved gcForest framework for detecting fraudulent online transaction

As shown in Figure 1, assuming that the features number of initial input is M, there are n autoencoders for multi-scanning structure. For training samples with size N, an autoencoder obtains its hidden layer output representation vector H through the BaggingBalance method. the vector H will be used to train the first level of the first layer of the cascade forest. The same operation is performed on

the other $n-1$ autoencoders, and the obtained hidden layer output vectors are respectively used to train the second level to k^{th} level of the first layer of the cascade forest.

Repeat the same operations for every initial training transaction sample. The expanded feature vector adds the class vectors generated by the previous level, which are used to train the second and third layers of the cascaded forest, respectively, and this process is repeated until the convergence of the model performance. In other words, the final model is actually a ensemble of deep forest, each of which is composed of multiple levels, as shown in Figure 1, with each layer corresponding to a hidden layer vector representation of an autoencoder.

3.2 BaggingBalance: A Method for Processing Sample Imbalance

BaggingBalance is a sample imbalance processing method based on the idea of Bagging [24], by under-sampling operation of raw data at the data input layer, and randomly selecting attribute features, thus obtaining different sampling data sets.

Specifically, the original data set is first divided into a majority class training set D_{major} and a minority class training set D_{minor} based on bootstraping [24]. Sampling the majority class training sets produces a data set D_{sample}: each time randomly pick a sample from the majority class dataset D_{major}, copy it into D_{sample}, and then put the sample back into the initial dataset D_{major}. It is possible to enable the sample sampled at the next sampling via the step. Different from self-sampling, the times of this process is repeatedly executed is the sample size $|\mathrm{D}_{minor}|$ of the minority training set D_{minor}, instead of the size of $|\mathrm{D}_{major}|$ of majority class training set D_{major}.

In addition, unlike Bagging, which only differs by sample perturbation, the BaggingBalance method also introduces the randomness of attribute features, which is similar to the idea of RF, i.e., the attribute feature perturbation is added at the same time, which will improve the generalization performance of final model.

The process of BaggingBalance algorithm is Algorithm 1.

Algorithm 1 BaggingBalance

Input: The majority training set, D_{major}, the minority training set, D_{minor}, the feature space, F, the number of sampling training set, k, and number of features randomly selected, $\mathrm{m}_{feature}$

Output: k sample training sets, $\mathrm{D} = \{D_1, D_2, \ldots, D_k\}$;

$\mathrm{D} = []$;

for $i = 0$ to k **do**

Sampling D_{major} to get sampled data set D_{sample}, where $|\mathrm{D}_{sample}| = |\mathrm{D}_{minor}|$;

Randomly extract feature subset F_{sample} from feature space F, where $|\mathrm{F}_{sample}| = \mathrm{m}_{feature}$;

$D_i = \{\mathrm{D}_{sample}, \mathrm{F}_{sample}\}$

end forreturn D;

3.3 New Multi-Scanning Structure Using Autoencoder

In the original gcForest algorithm, the multi-scanning structure [11] uses the sliding window technique to process the original features. The vectors obtained by each sliding are processed by the RF model and completely random forest model to obtain the class vector, and then all the class vectors are concatenated as a transformation feature vector, which is passed as an input feature vector to the cascade forest for classification.

However, the sliding window-based method has its own limitations. As mentioned in the proposed paper [11], the multi-scanning structure has a good effect on data with sequence relationship or spatial relationship. Because the sliding window method can only slide linearly, thus there is a great demand for the feature space arrangement of the raw data. Specifically, the sliding window is qualified to process the feature vectors with sequence relationship, but there is no strong sequence relationship between the original feature vectors in each online transaction, even in an out-of-order feature space status. In addition, the reconstructed feature vector generated by the original structure is completely composed of the class vectors, which cannot fully map the feature space of the original data.

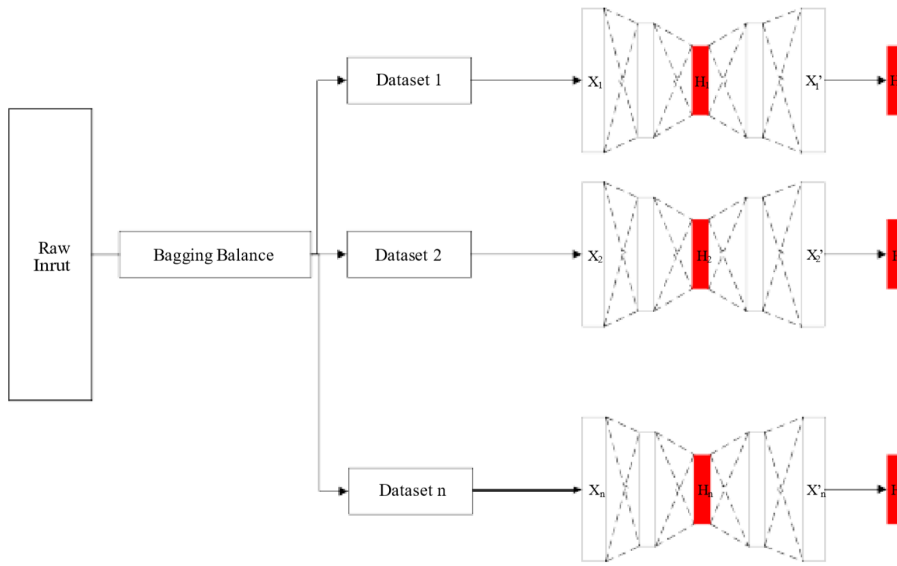


Figure 2. New multi-scanning structure using autoencoder

Therefore we reconstruct the multi-scanning structure by introducing autoencoder algorithm and the BaggingBalance method, and propose a new multi-scanning structure using autoencoder, which is shown in Figure 2.

On the one hand, based on the BaggingBalance method we proposed, the original input data is double-randomly sampled in the sample space and the feature

space, and several different sampling training sets are obtained. Through the introduction of randomness, the disorder of the online transaction feature space is considered while dealing with the problem of sample imbalance. Through the random feature selection, the random combination of different transaction attributes in the original feature space can be realized, and the internal relationship between the fraud patterns and the transaction feature space in online transaction can be deeply explored. This is also reason that why randomness exists in many ML algorithms.

On the other hand, the autoencoder model is introduced in the multi-scanning structure to further enhance the representation learning ability of the fraud detection model. Autoencoder has proven to be a robust algorithm which can be used in several applications and the main advantage is to extract best features for data analysis [23]. The sampling training set obtained by the BaggingBalance method is used as input to train the autoencoder model, and the hidden layer output representation vector of the trained autoencoder is extracted as a new modified feature vector, which is transmitted as input to the cascade forest model for model training. Compared with the class vector generated by RF model and completely random forest model, the hidden layer output representation vector obtained by the autoencoder is a better expression of the original input feature space. What is more, it is more concise and effective, and more fully reflects the distribution of the original feature space.

The overall process flow of the multi-scanning structure using autoencoder is summarized as shown in Algorithm 2.

Algorithm 2 The process flow of the multi-scanning structure using autoencoder

Input: The majority training set, D_{major}, the minority training set, D_{minor}, the feature space, F, the number of initialized autoencoders, k, the number of features randomly selected, $\mathrm{m}_{feature}$, and the number of iterations of the autoencoder, iters

Output: output expression vector of k autoencoders in hidden layer, $\mathrm{H} = \{H_1, H_2, \ldots, H_k\}$;

```
H = [];
for each AutoEncoder_i in k autoencoders do
    Sampling D_major to get sampled data set D_sample, where |D_sample| = |D_minor|;
    Randomly extract feature subset F_sample from feature space F, where |F_sample| = m_feature;
    for t = 0 to iters do
        training AutoEncoder_i by TrainAutoencoder(D_major, D_minor, F_sample);
    end for
    get H_i;
    push H_i to H;
end forreturn H;
```

4 EXPERIMENTS

4.1 Datasets and Indicators

Experimental data comes from real online transaction data of a China's bank, including three-month B2C transaction records (from April 2017 to June 2017). There are original 67 available transaction attributes, and there are more than 70 000 transactions labeled as fraudulent transactions in historical data. In this paper, we use transaction data of the first two months as a training set to train the improved gcForest-based online transaction fraud detection model. The last months transaction data is used as the testing set to evaluate the performance of the detection model. Last but not least, precision rate and recall rate is used to evaluate the performance of the proposed model.

Real / Predicted	True Fraud	True Normal
Predicted Fraud	TP	FP
Predicted Normal	FN	TN

Table 1. Confusion matrix

As shown in Table 1, because it is a fraudulent transaction interception, the focus of the model should be on fraudulent transactions, so the confusion matrix is slightly modified. TP (True Positive) is the number of fraudulent transactions judged as fraudulent transactions by the model. FP (False Positive) is the number of normal transactions that are judged as fraudulent transactions by the model. TN (True Negative) is the number of normal transactions that are judged as normal transactions by the model. FN (False Negative) is the number of fraudulent transactions that are judged as normal transactions by the model. Then, 3 indicators in Table 2 will serve to evaluate the performance.

Indicator Name	Calculation Method
Accuracy	$(TP + TN)/(TP + TN + FP + FN)$
Precision	$TP/(TP + FP)$
Recall	$TP/(TP + FN)$

Table 2. Indicator calculation method

As shown in Table 2, in fraud detection, the accuracy rate refers to the ratio of the number of correct transactions predicted by the detection model to the total electronic transactions. Accuracy is the most commonly used performance metric in classification tasks, which is suitable for both binary classification tasks and multi-classification tasks. However, it cannot meet the needs of fraud detection tasks because this indicator cannot accurately measure the performance of fraud detection models due to the imbalance of samples in fraud detection. On this basis, the concepts of precision and recall are proposed, matching with precision and recall

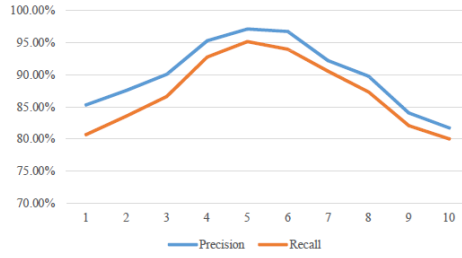
respectively in machine learning. Therefore, the application of precision rate and recall rate in fraud detection task can reflect the performance and effect of the model in electronic transaction fraud detection.

4.2 Model Evaluation

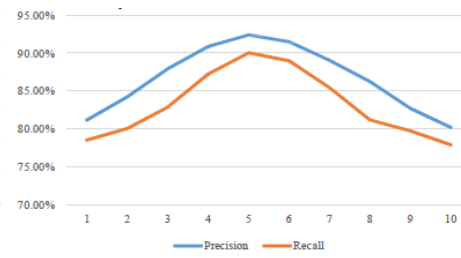
1. Selection of the Number of Autoencoder: In the framework of improved gcForest-based online transaction fraud detection, the selection of the autoencoders' number, i.e., the selection of the number of sample datasets in BaggingBalance, is a problem worth studying. Because the data distribution and feature distribution of sampling datasets generated by the BaggingBalance method tend to be very different due to the randomness of sample selection and feature selection. These datasets will be used as the input of the autoencoder model to train the model, and produce various hidden layer poor model performance. If there are too many autoencoders, overfitting will occur, which leads to the worse generalization ability of the model and the degraded performance.
2. Performance of the Fraud Detection Model: After determining the number of autoencoders, this section conducts an experimental study on the performance of the proposed fraud detection model. The RF model and the original gcForest model are selected as the comparison model. The test was extracted from the online transaction data of June 2017 which are divided into five subsets including the first 10 days, the first 15 days, the first 20 days, the first 25 days and the first 30 days.

Based on the above considerations, this section of the experiment selects the transaction data of the first two months as the training sets and tests it on the transaction data of the first 10 days, the first 20 days and the entire month in June. X-axis is the number of autoencoders which is considered in the ranges from 1 to 10. Figures 3 a), 3 b), 3 c) show the results of fraud detection models via the different autocoders' number, which are tested on the online transaction data for the first 10 days, the first 20 days and the entire month of June. From the above experimental results, the number of autoencoders should not be too large, and should not be too small, generally taking 4 to 6. If the number is too small, the dataset generated by BagingBalance is small in size and cannot fully reflect the original data space, resulting in insufficient learning of the original data and output representation, which has a great influence on the final detection effect of the model.

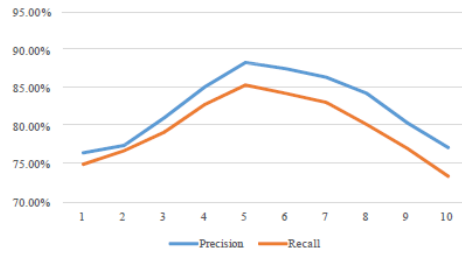
In addition, to verify the effectiveness of the introduced autoencoder in the model, we also evaluate the performance of the original cascaded forest structure with an autoencoder in this section. In this model, we input the raw online transaction data, pass them into the autoencoder and obtain the hidden layer output representation vector, which will be as the input of original cascaded forest structure. From Figures 4 a), 4 b) we can conclude that compared with the original gcForest model, the introduction of the autoencoder has its effectiveness. At the same time,



a) Data set of the first 10 days



b) Data set of the first 20 days

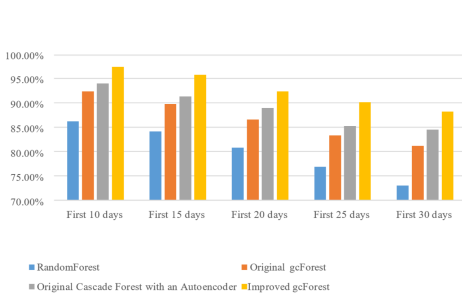


c) Data set of the whole month

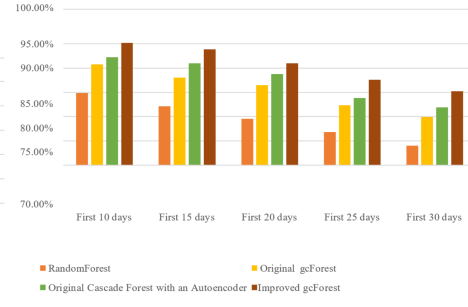
Figure 3. Evaluation on the number of autoencoders

the results also show that the improved gcForest with multi-autoencoders based on BaggingBalance is superior to ones with an autoencoder.

Based on the last experimental result, this section initializes the number of autoencoders to 5. Figure 4 a) shows the precision rate on different models in the five test sets, and Figure 4 b) shows the recall rate on different models in the five test sets. It can be seen that the proposed model has a beyond 10 % improvement compared to RF model, a beyond 5 % improvement compared to the original gcForest model.



a) The precision rate of different models



b) The recall rate of different models

Figure 4. The performances of different models on various sample sets

4.3 System Implementation

In order to verify the comprehensive performance of the model, a fraud detection subsystem was built. Based on the fraud detection model based on deep forest proposed in this paper, the system realized two functions of offline model training and simulated real-time electronic transaction fraud detection to verify the application and effectiveness of the detection model.

For the model off-line training module, the functions of data extraction, data preprocessing, feature differentiation and model training of the deep forest model are completed. The realization of this part of functions in order to make the training model process from data extraction to result analysis can be completed at the system end, convenient for users to control the training process of the model.

For the function of simulating real-time electronic transaction fraud detection module, the detection, interception and release of real-time transaction data are completed. The work of this part is to deploy the trained deep learning model into the system. The system passes the received electronic transaction into the detection model for detection. If the detection is normal, the transaction will be released. If the transaction is identified as fraudulent by the model, it will be intercepted.

After configuring the service based on the deep forest fraud detection model, the model will start running to prevent and monitor the risks of real-time electronic transaction data streams entering the system. The interactive page design for real-time risk control monitoring is shown in Figure 5. This part shows the real-time detection results after real-time transaction data enters the group behavior fraud detection subsystem and the performance analysis and visualization of the running detection model. The detection result display part displays basic information such as the user account of the current transaction, the user's name, the time of the transaction and the interception of the detection model. At the same time, the intercepted electronic transactions are displayed in detail to analyze the characteristics of the intercepted transactions, as shown in Figure 6.

At the same time, the simulated real-time electronic transaction fraud detection function counts the real-time detection performance indicators of the fraud detection model which can display the detection effect of the model in real time and is also beneficial to analyze the specific application of the model. Figure 7 shows the number of intercepted electronic transactions of the detection model. While Figure 8 shows the performance indicators of the detection model running in the system, including hit rate, recall rate, accuracy rate and interference rate.

This chapter designs and implements a B/S-based group behavior fraud detection subsystem. The system mainly includes two functional modules: offline model training function and real-time detection of simulated electronic transactions. On the one hand, the offline model training module is used to access the API interface of the data storage platform of the hierarchical diagnosis and treatment cloud platform to obtain the historical transaction data of electronic transactions as the original training set. At the same time, the model parameters are set through visual interactive operations to realize electronic transactions based on deep forests.

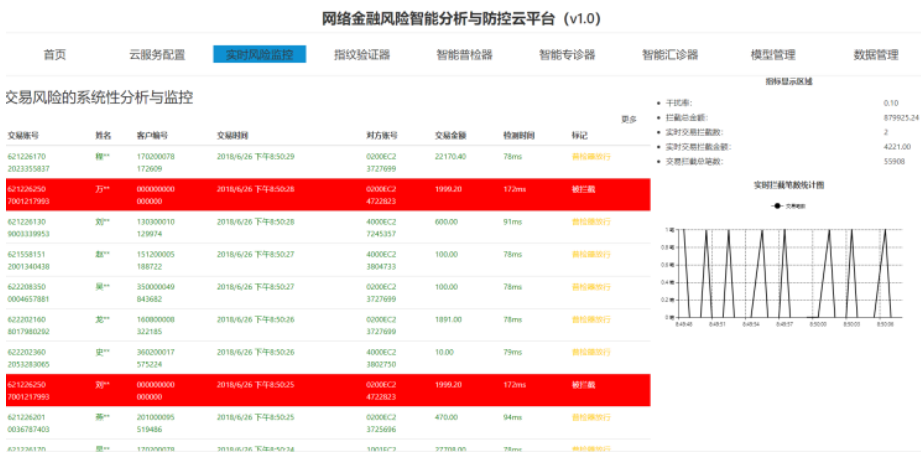


Figure 5. Transaction risk monitoring page

交易账号 Trading account	姓名 User name	客户编号 User ID	交易时间 Trading hours	对方账号 Reciprocal account number	交易金额 Transaction amount	检测时间 Detection time	标记 Test results
622202170 2026602508	万**	000000000 000000	2018/6/26 下午9:11:22	0213EC4 6895132	66.12	172ms	被拦截
622202170 2026602508	徐**	000000000 000000	2018/6/26 下午9:11:09	0213EC4 6895132	8.54	188ms	被拦截
622202170 2026602508	吴**	000000000 000000	2018/6/26 下午9:11:06	0213EC4 6895132	56.95	173ms	被拦截
622202170 2026602508	杨**	000000000 000000	2018/6/26 下午9:11:03	0213EC4 6895132	4.33	171ms	被拦截
622202170 2026602508	李**	000000000 000000	2018/6/26 下午9:10:54	0213EC4 6895132	66.12	172ms	被拦截
622202170 2026602508	毛**	000000000 000000	2018/6/26 下午9:10:51	0213EC4 6895132	8.54	172ms	被拦截
622202170 2026602508	万**	000000000 000000	2018/6/26 下午9:10:40	0213EC4 6895132	66.12	173ms	被拦截
622202170 2026602508	刘**	000000000 000000	2018/6/26 下午9:10:38	0213EC4 6895132	4.33	183ms	被拦截
622202170 2026602508	李**	000000000 000000	2018/6/26 下午9:10:29	0213EC4 6895132	4.33	174ms	被拦截

Figure 6. The set of transactions that the model identifies as fraudulent

Fraud model training and visualization. On the other hand, the real-time detection function of simulating electronic transactions is used to obtain the implementation transaction data stream of the risk control subsystem of the financial risk control platform and the detection model trained by the offline model training function is used to perform the real-time transaction data stream.

Real-time detection of fraudulent transactions and analysis of detection effects. By building a group behavior fraud detection subsystem, the development complexity of developers is reduced and the application value of the detection model proposed in this article is verified.

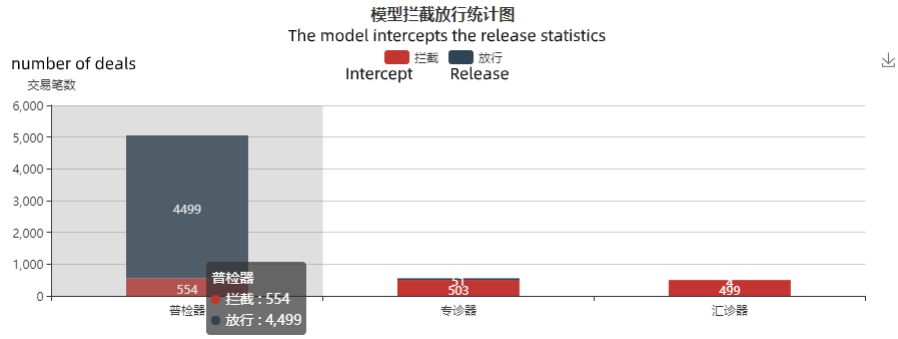


Figure 7. Each module intercepts the number of releases

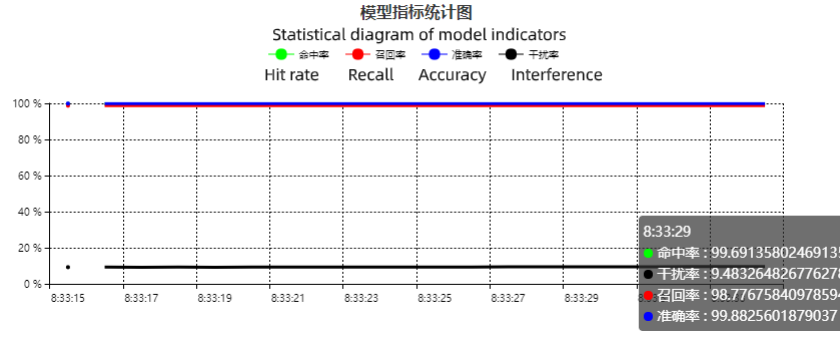


Figure 8. The model detects trade indicators in real time

5 CONCLUSION

This paper establishes an online transaction fraud detection model based on improved gcForest. BaggingBalance, a sample imbalance processing method based on Bagging, is proposed to rebalance the datasets and construct a global sample unbalance processing mechanism with XGBoost used in cascade forest. Based on this, autoencoder algorithm is introduced to the multi-scanning structure, further enhancing the representational learning ability of the model. The experimental results show a superior fraud detection performance of the proposed model on real bank online transaction data. Furthermore, this paper is another exploration about using the advantage of ML techniques and DL techniques. There are more possibilities for combining more ML models and DL models to detect online fraudulent transactions.

Acknowledgement

This work was supported by the Natural Science Foundation of Shanghai (No. 19ZR-1401900) and the Shanghai Science and Technology Innovation Action Plan Project (No. 19511101300).

REFERENCES

[1] Jingdong Financial Research Institute: Digital Finance Anti-Fraud White Paper. Available from: http://finance.qq.com/original/caijingzhiku/yzzk12.html, May 2018.

[2] Security Alliance of E-Commerce Ecosystem: 2017 E-Commerce Ecological Security White Paper. Available from: https://www.saee.org.cn/pc/newsContent/news20170726, July 2017.

[3] Sahin, Y.—Duman, E.: Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. International MultiConference of Engineers and Computer Scientists, Vol. 1, 2011, pp. 442–447, doi: 10.1109/inista.2011.5946108.

[4] Alowais, M. I.—Soon, L. K.: Credit Card Fraud Detection: Personalized or Aggregated Model. Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing, June 2012, pp. 114–119, doi: 10.1109/music.2012.27.

[5] Xuan, S.—Liu, G.—Li, Z.—Zheng, L.—Wang, S.—Jiang, C.: Random Forest for Credit Card Fraud Detection. 15th International Conference on Networking, Sensing and Control, March 2018, pp. 1–6, doi: 10.1109/icnsc.2018.8361343.

[6] LeCun, Y.—Bengio, Y.—Hinton, G. E.: Deep Learning. Nature, Vol. 521, 2015, pp. 436–444, doi: 10.1038/nature14539.

[7] Krizhevsky, A.—Sutskever, I.—Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q. (Eds.): Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.

[8] Sutskever, I.—Vinyals, O.—Le, V. Q.: Sequence to Sequence Learning with Neural Networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. Q. (Eds.): Advances in Neural Information Processing Systems 27 (NIPS 2014), Vol. 2, 2014, pp. 3104–3112.

[9] Corbo, J.—Giovine, C.—Wigley, C.: Applying Analytics in Financial Institutions Fight Against Fraud. McKinsey Analytics, April 2017. Available from: https://www.mckinsey.com/businessfunctions/mckinsey-analytics/our-insights/applying-analytics-infinancial-institutions-fight-against-fraud, retrieved February 2018.

[10] Li, X.—Yu, W.—Luwang, T.—Zheng, J.—Qiu, X.—Zhao, J. et al.: Transaction Fraud Detection Using GRU-Centered Sandwich-Structured Model. 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design, November 2017, pp. 467–472, doi: 10.1109/cscwd.2018.8465147.

[11] Zhou, Z.—Feng, J.: Deep Forest: Towards an Alternative to Deep Neural Networks. Proceedings of the Twenty-Sixth International Joint Conference on Arti-

ficial Intelligence (IJCAI 2017), February 2017, pp. 3553–3559, doi: 10.24963/ijcai.2017/497.

[12] Breiman, L.: Random Forests. Machine Learning, Vol. 45, 2001, No. 1, pp. 5–32, doi: 10.1023/A:1010933404324.

[13] Liu, F.—Ting, K.—Yu, Y.—Zhou, Z.: Spectrum of Variable-Random Trees. Journal of Artificial Intelligence Research, Vol. 32, 2008, pp. 355–384, doi: 10.1613/jair.2470.

[14] Dong, M.—Yao, L.—Wang, X.—Benatallah, B.—Huang, C.—Ning, X.: Opinion Fraud Detection via Neural Autoencoder Decision Forest. Pattern Recognition Letters, Vol. 132, 2020, pp. 21–29, doi: 10.1016/j.patrec.2018.07.013.

[15] Chen, T.—Guestrin, C.: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[16] Nielsen, D.: Tree Boosting with XGBoost – Why Does XGBoost Win "Every" Machine Learning Competition? Master's Thesis, Norvegian University of Science and Technology (NTNU), Trondheim, 2016.

[17] Olszewski, D.: Fraud Detection Using Self-Organizing Map Visualizing the User Profiles. Knowledge-Based Systems, Vol. 70, 2014, pp. 324–334, doi: 10.1016/j.knosys.2014.07.008.

[18] Dal Pozzolo, A.: Adaptive Machine Learning for Credit Card Fraud Detection. Ph.D. Thesis, Université Libre de Bruxelles, December 2015.

[19] Fu, K.—Cheng, D.—Tu, Y.—Zhang, L.: Credit Card Fraud Detection Using Convolutional Neural Networks. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (Eds.): Neural Information Processing (ICONIP 2016). Springer, Cham, Lecture Notes in Computer Science, Vol. 9949, 2016, pp. 483–490, doi: 10.1007/978-3-319-46675-0_53.

[20] Wang, S.—Liu, C.—Gao, X.—Qu, H.—Xu, W.: Session-Based Fraud Detection in Online E-Commerce Transactions Using Recurrent Neural Networks. In: Altun, Y. et al. (Eds.): Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2017). Springer, Cham, Lecture Notes in Computer Science, Vol. 10536, 2017, pp. 241–252, doi: 10.1007/978-3-319-71273-4_20.

[21] Zhang, Z.—Zhou, X.—Zhang, X.—Wang, L.—Wang, P.: A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection. Security and Communication Networks, Vol. 2018, 2018, Art. No. 5680264, 9 pp., doi: 10.1155/2018/5680264.

[22] Roy, A.—Sun, J.—Mahoney, R.—Alonzi, L.—Adams, S.—Beling, P.: Deep Learning Detecting Fraud in Credit Card Transactions. 2018 Systems and Information Engineering Design Symposium (SIEDS), 2018, pp. 129–134, doi: 10.1109/SIEDS.2018.8374722.

[23] Kazemi, Z.—Zarrabi, H.: Using Deep Networks for Fraud Detection in the Credit Card Transactions. 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017, pp. 630–633, doi: 10.1109/kbei.2017.8324876.

[24] Breiman, L.: Bagging Predictors. Machine Learning, Vol. 24, 1996, No. 2, pp. 123–140, doi: 10.1007/bf00058655.

Mian Huang received the M.Sc. degree from Lanzhou University of Technology, Lanzhou, China in 2008, and now he is pursuing the Eng.D. degree from Tongji University, Shanghai, China. His current research interests include network security and identity authentication.

Lizhi Wang is an M.Sc. candidate at the Donghua University. His research area includes machine learning, deep learning, and cloud computing.

Zhaohui Zhang obtained his Bachelor's degree in computer science from Anhui Normal University, Wuhu, China in 1994. He obtained his Ph.D. in computer science from Tongji University, Shanghai, China in 2007. From 1994 to 2015, he worked in Anhui Normal University as Professor. Since 2015 he has been working as Professor in the School of Computer Science and Technology, Donghua University, Shanghai, China. His research interests include big data intelligent processing and behavior analysis.