Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 3, 2023

The Credit Card Anti-fraud Detection Model in the Context of Dynamic Integration Selection Algorithm

Jiajian Zheng [1, *], Le Yang [2], Duan Xin [3], Miao Tian [4]

[1] Bachelor of Engineering, Guangdong University of Technology, Shenzhen, China
[2] Computer Information Science, Sam Houston State University, Huntsville, TX, USA
[3] Accounting, Sun Yat-Sen University, Hong Kong, China
[4] Computer Science, San Fransisco Bay University, Fremont CA, USA

[*] **Corresponding author**: Jiajian Zheng (Email: im.JiaJianZheng@gmail.com)

Abstract: Because of its close relationship with information technology, every change in the field of information technology will have an important impact on the means, time and space distribution of the final fraud of the credit card network. The rapid development and popularization of artificial intelligence (AI) software based on Deepfake technology has greatly promoted the intelligent transformation of credit card network fraud. Through the use of advanced hardware and software equipment, criminals have iterated on their own fraud tools, from the use of traditional phone calls and text messages to the use of the latest AI software. Therefore, for the problem of credit card fraud detection with missing data set labels and highly unbalanced category distribution, A credit card approval anomaly detection model DES-HBOS (Dynamic Ensemble Selection based on Histogram-based Outlier Score) is proposed. Firstly, the unsupervised anomaly detection algorithm is used to construct the false label of the training set customer. Then, the customer capability area to be measured is determined, and the classifier performance is evaluated according to Pearson correlation coefficient. Finally, a set of optimal classifiers is selected to integrate the test customers. Experiments on real credit card customer data sets show that DES-HBOS has a higher Recall and can identify more fraudulent customers than other 6 classical anomaly detection models. Comparative experiments were carried out on 4 unbalanced data sets, and the experimental results showed that DES-HBOS had stronger anomaly detection ability than HBOS.

Keywords: Selection Algorithm; Artificial Intelligence; Anti-fraud; Credit Card Anomaly.

1. Introduction

With the rapid development of China's economy, the size of the credit card market has also grown rapidly. By the end of 2021, the number of credit cards issued in China has reached 800 million. As the number of transactions increases, the use of credit cards for fraud is on the rise. Although the proportion of fraud in the whole credit card transaction is very low, once it occurs, it will cause huge economic losses to commercial banks. According to the Aggregate Index for the Payment and Clearing Industry in the Second Quarter of 2023 released by the Payment and Clearing Association of China, the total amount of overdue credit cards is expected to reach 89.646 billion yuan by the end of 2023, up 6.36% year-on-year. It is understood that the total amount of overdue credit cards is still at a high level, a number of banks in order to adjust the credit card overdue situation, reduce the overdue rate of credit cards, take measures such as shrinking the line of credit and issuing card specifications. In view of this, how to quickly and effectively identify credit card fraud and prevent risks [1]. It has become a research topic in the field of bank risk control. However, with the advent of the digital age, the financial industry is also changing. In this new era, innovative technologies such as blockchain, artificial intelligence and big data are profoundly changing the forms and methods of financial services and reshaping the financial market landscape[2-3]. This radical technological change is also accompanied by new risks and challenges, as well as a stricter financial regulatory system. In the face of the high complexity, global cross-border transactions, and data privacy issues that digital finance brings, traditional regulatory models are no longer applicable. Therefore, there are two major problems in the field of credit card fraud detection: first, in real life, it is difficult to obtain data labels for fraud samples, the cost of manually labeling data is high, and the amount of labeled sample data is not enough to reflect the real fraud situation, in most cases, commercial banks are faced with unlabeled data sets; second, there is an extreme imbalance in the category of credit card transaction data The phenomenon of... That is, the fraud sample is much smaller than the normal sample[4]. In view of this, in view of the absence of labels in the data set, this paper mined the potential information in customer characteristics, issued an early warning for applications with high potential risks, identified "abnormal customers", and aimed to reduce the risk of fraud from the aspect of credit approval.

2. RELATED WORK

Since the 1990s, scholars have begun to explore credit card fraud detection methods based on data mining, such as decision tree, neural network, support vector machine, etc. With the development of artificial intelligence technology, Some scholars have applied deep learning technology to the field of credit card fraud detection :Jurgovsky et al. transform fraud detection problems into sequence classification tasks and use long and short term memory neural networks for prediction, thus effectively improving detection accuracy: Fiore et al. train Generative Adversarial networks Networks (GAN) model, which is used to generate fraud samples and combine these samples with original data sets to build an effective fraud detection mechanism[5-7]. At present, the development of "Internet + finance" has made people's

transaction methods more convenient. Among them, credit card transaction has become one of the most popular online and offline payment methods, and the increase in the number of credit card transactions has made credit card fraud often occur. According to the Blue Book on the Development of China's Bank Card Industry (2022), by the end of 2021, the cumulative issuance of bank cards reached 9.25 billion in China, with 270 million new credit cards issued in the year, an increase of 3.0%.The outstanding credit balance of bank cards was 8.62 trillion yuan, an increase of 8.9% over the previous year; The total outstanding credit of overdue credit cards for six months was 86.04 billion yuan, up 2.6% year-on-year; The card fraud rate was 0.32 basis points, down 0.43 basis points from the previous year[8].

In the case that the data sample contains only features without labels, anomaly detection can reveal the internal rule between samples through the analysis of the characteristics of the data sample, so as to find a few samples that are significantly different from the general behavior or pattern. Because it is extremely difficult to obtain labels of fraudulent transactions, some scholars regard fraudulent samples as outliers and separate them from normal samples by anomaly detection technology. Van et al[9-10]. H used unsupervised anomaly detection technology to identify fraud samples of medical insurance claims, and the experimental results showed that potential new fraud patterns could be detected through anomaly detection technology. Porwal et al. adopt the clustering based integration method to obtain the anomaly score of each data sample, which can detect abnormal samples in large data sets. And it can have a strong robustness to changing fraud patterns. It is of practical significance and value to use unsupervised anomaly detection method to identify fraudulent samples, and effectively solve the problem of missing labels. At present, the methods to deal with unbalanced data can be summarized into two categories: one is to adjust the sample category distribution by under sampling or oversampling at the data level; the other is to deal with the algorithm level represented by cost-sensitive learning and ensemble learning. The classification prediction of unbalanced data based on a single machine learning classification algorithm may lead to certain bias, while ensemble learning fuses diverse and complementary multiple base classifiers into a strong classifier, which can effectively improve the accuracy and stability of the model.

Therefore, in the construction of credit evaluation model, we should pay attention to the problems of category imbalance and cost sensitivity. However, in a large number of credit scoring studies, only a few models have considered these two types of problems. Among them, the dynamic integration selection algorithm, the large class is divided into several subsets, the subsets and small classes are combined to form training samples to train the base classifier, so as to ensure that the global does not lose important information, this method is simpler but more efficient than SMOTE. SUBAGGING algorithm for high unbalance rate data sets under samples the large class during the training set sampling process and combines it with all the small class samples to obtain a balanced training data set to train the base classifier. Similarly, multiple differentiated integrated molecular models are trained using different training subsets, feature sampling and parameter perturbation methods. The above methods all use sampling methods to deal with class imbalance[11]. Changing the distribution of training data to make up for the missing representative samples of small and medium-sized classes in model training is an effective way to solve such problems.

3. HBOS Anomaly Detection Algorithm

3.1. HBOS Data Model

The Histogram based Outlier Score (HBOS) was developed by Goldstein an unsupervised anomaly detection algorithm based on nonparametric statistics is proposed by et al., which does not rely on hyperparameters and avoids bias caused by improper selection of hyperparameters: Based on the assumption of independence between features, the method breaks down the processing of high-dimensional data into multiple single-feature calculations. In the context of Internet, customer characteristics are increasing.

Algorithm	Breast-cancer	Pen-global	Pen-local
HBOS	**0.9910**	**0.9214**	**0.7651**
k-NN	0.9826	0.9892	0.9852
LOF	0.9916	0.8864	0.9878
Fast-LOF	0.9882	0.9050	0.9937
COF	0.9888	0.9586	0.9688
INFLO	0.9922	0.8213	0.9875
LoOP	0.9882	0.8492	0.9864
LOCI	0.9678	0.8868	– [3]
CBLOF	0.8389	0.6808	0.7007
u-CBLOF	0.9743	0.9923	0.9767
LDCOF	0.9804	0.9897	0.9617

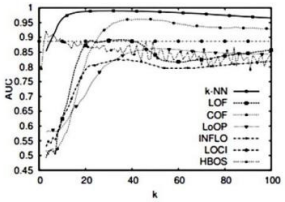

Figure 1. STraining HBOS Model

Put forward higher requirements for data processing. The histogram method has good adaptability to high-dimensional data, and its fast calculation performance makes it very efficient to deal with massive high-dimensional data. Therefore, this experiment adopts HBOS as the method to identify abnormal customers, and its model expression is as follows:

$$\mathrm{HBOS}(p)=\sum_{i=1}^{d}\log\left(\frac{1}{\mathrm{hist}_{i}(p)}\right) \tag{1}$$

Where :P represents customer characteristics; HBOS(P) anomaly scores for customers; hist(P) is the probability density estimate of the customer's I-th feature; d is the number of customer features.

HBOS constructs a univariate histogram for each feature and standardizes it so that the maximum height of the histogram is 1, and the height of each box represents the probability density estimate, which is roughly a "bell curve". The lower the probability density, the customer's characteristic value deviates from most customers, and the higher the anomaly score. The histogram can reflect the distribution of a certain feature of all customers, and the feature value with lower probability density. The more likely it is to be abnormal. Finally, all the characteristics of the customer are integrated to determine the abnormal situation.

Therefore, a set of base classifiers is constructed by means of isomorphic classifier generation. HBOS is used as the learning algorithm, and a series of different base classifiers are obtained by changing the number of parameters, that is, boxes.

3.2. Detection Algorithm

The parameters are random integers between 10 and 50, and the set of base classifiers C= {C1, C2, ..., Cn} X_{train} I R $^{N-m}$, represents the set to be tested, where each customer has m characteristics. All base classifiers are trained under the same training set and the abnormal score matrix $S(X_{train})$ of training set X_{train} is obtained.

$$S(X_{train}) = [C_1(X_{train}), \ldots, C_\eta(X_{train})] = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,\eta} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,\eta} \\ \vdots & \vdots & \cdots & \vdots \\ s_{N,1} & s_{N,2} & \cdots & s_{N,\eta} \end{pmatrix} \quad (2)$$

Where: Cj (Xtrain) (j= 1,2,... n) represents the abnormal score vector of the n classifier on the training set, and has been standardized; Si.j represents the anomaly score of the i th customer in the training set under the j n classifier. In this experiment, customers in the training set are labeled by averaging the output of all classifiers. The false label of the customer in the training set can be represented as

$$\text{target}(X_{train}) = \text{Average}(S(X_{train})) = (s_1, s_2, \cdots s_N)^T \quad (3)$$

$$s_i = \frac{1}{\eta}\sum_{j=1}^{\eta} s_{i,j} \ (i = 1, 2, \ldots, N) \quad (4)$$

False label representing the I-th customer; S represents the exception of the ith guest under the JTH classifier where: s:= score.

3.3. Results and Analysis

In this experiment, Histogram based Outlier Score (HBOS), K-Nearest Neighbor (KNN), and One-Class Support Vector Machine (one-class support vector machine) are selected. OCSVM, Local Outlier Factor (LOF), Principal Component Analysis (PCA), Isolation Forest (Isolation Forest) These six classifiers, which are widely used in the field of anomaly detection and have good results, are compared. The important parameter Settings of the algorithm are shown in Table 1, and the default parameters in sklearn library of Python3.8 are used for other parameters.

Table 1. Model important parameter

Method	Data
HBOS	n-bis=10, contamination= 0.01
KNN	n-neighbors=5, contamination = 0.01
OCSVM	kernel= 'rbf', nu=0.5contamination=0.01
LOF	n-neighbors=20, contamination = 0.01
PCA	contamination = 0.01
IForest	n-neighbors=100, contamination = 0.01

Considering the impact of the number of base classifiers η in DES-HBOS on model performance, too large or too small will affect model performance. Therefore, in this experiment, η is set to 10,20,30,40,50 to explore the impact of η on model performance. Recall and AUC are the average values of DES-HBOS over 15 test sets:

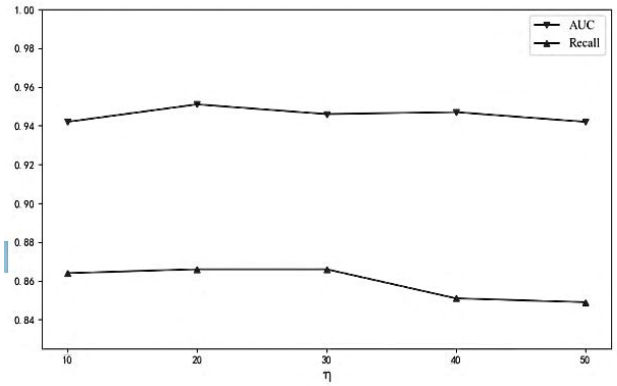

Figure 2. Detect the change map

To solve the problem of credit card fraud detection with missing data set labels and unbalanced category distribution, this paper proposes a credit card approval anomaly detection model based on dynamic integration selection algorithm. In order to solve the problem of missing labels, the unsupervised anomaly detection algorithm was used to construct false labels for the customers in the training set[12]. In order to alleviate the problem of unbalanced category distribution, CR of the customers under test was determined and the performance of the classifier in the classifier set was evaluated on behalf of the customers under test according to Pearson correlation coefficient. A strong classifier is obtained by fusing several classifiers with excellent classification performance.

4. Conclusion

In financial regulation, using histogram algorithm to detect credit card anomalies is a common method. A histogram is a statistical chart used to show the distribution of data. For credit card transaction data, histograms can be used to analyze the distribution of transaction amounts. The steps to detect credit card anomalies using the histogram algorithm are as follows: 1. Collect credit card transaction data: Collect normal credit card transaction data, including transaction amount, transaction time and other information. 2. Construct histogram: According to the collected normal transaction data, the transaction amount is grouped, and then the number of transactions within each amount range is counted. According to the statistical results, the histogram is constructed. 3. Detect abnormal transactions: For new credit card transaction data, compare the transaction amount to the histogram. If the transaction amount falls in the extreme range of the histogram, and the distribution difference from the normal transaction data is large, it can be judged as an abnormal transaction. 4. Additional rules: In addition to histogram algorithms, financial regulators can also develop additional rules to detect credit card anomalies. For example, if the transaction amount exceeds a certain amount threshold, or is contrary to common consumption patterns, it can also be judged as an abnormal transaction.

By using histogram algorithms to detect credit card anomalies, financial regulators can more accurately monitor and prevent credit card fraud and protect consumers' property safety. At the same time, financial institutions can also use this abnormal transaction data to improve their anti-fraud systems and improve the security of their transactions.

Acknowledgments

We thank Bo Liu and JixHuang for their related works within artificial intelligence in the biomedical engineering field. Their articles have provided our research with directions and insights.

References

[1] URCOVSKY J, CRANITZER M, ZIECLER K, et al. Sequence Classification for Credit-Card Fraud DetectionlJl. Expert Systems with Applications, 2018, 100: 234 245.

[2] FIORE U, DE SANTIS A, PERLA F, et al. Using Cenerative Adversarial Netorks for lmproving Classification Effectiveness inCredit Card Fraud Detection J . Information Sciences, 2019, 479: 448-455.

[3] SMITI A. A Critical Overview 0f Outlier Detection Methods Jl. Computer Science Review, 2020, 38: 100306.Van Capelleveen C, Poel M, Mueller R M, et al. Outlier Detection In Healthcare Fraud: A Case Study in the Medicaid

DentalDomainl.Jl. International Journal of Accounting Information Systems, 2016, 21: 18-31.

[4] PORWAL U, MUKUND S. Credit Card Fraud Detection in e-Commerce: An Outlier Detection ApproachlJ . arXiv preprintar Xiv: 1811.02196.2018.

[5] Chang Che, Bo Liu, Shulin Li, Jiaxin Huang, and Hao Hu. Deep learning for precise robot position prediction in logistics. Journal of Theory and Practice of Engineering Science, 3 (10): 36–41, 2023.DOI: 10.1021/acs.jctc.3c00031.

[6] Hao Hu, Shulin Li, Jiaxin Huang, Bo Liu, and Change Che. Casting product image data for quality inspection with xception and data augmentation. Journal of Theory and Practice of Engineering Science, 3(10):42–46, 2023. https://doi. org/10. 53469/jtpes.2023.03(10).06.

[7] Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. 2023. Enhancing Multimodal Understanding with CLIP-Based Image-to-Text Transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies (ICBDT '23). Association for Computing Machinery, New York, NY, USA, 414–418. https://doi. org/10. 1145/3627377.3627442.

[8] Lin, Q., Che, C., Hu, H., Zhao, X., & Li, S. (2023). A Comprehensive Study on Early Alzheimer's Disease Detection through Advanced Machine Learning Techniques on MRI Data. Academic Journal of Science and Technology, 8(1), 281–285.DOI: 10.1111/jgs.18617.

[9] Che, C., Hu, H., Zhao, X., Li, S., & Lin, Q. (2023). Advancing Cancer Document Classification with R andom Forest. Academic Journal of Science and Technology, 8(1), 278–280. https:// doi.org/10.54097/ajst.v8i1.14333.

[10] S. Tianbo, H. Weijun, C. Jiangfeng, L. Weijia, Y. Quan and H. Kun, "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition," 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2023, pp. 834-837, doi: 10.1109/ICCECE58074.2023.10135464.

[11] Y. Wang, K. Yang, W. Wan, Y. Zhang and Q. Liu, "Energy-Efficient Data and Energy Integrated Management Strategy for IoT Devices Based on RF Energy Harvesting," in *IEEE Internet of Things Journal*, vol. 8, no. 17, pp. 13640-13651, 1 Sept.1, 2021, doi: 10.1109/JIOT.2021.3068040.

[12] Wang, Y, Yang, K, Wan, W, Mei, H. Adaptive energy saving algorithms for Internet of Things devices integrating end and edge strategies. Trans Emerging Tel Tech. 2021; 32: e4122. DOI: https://doi.org/10.1002/ett.4122.

[13] Xu, J., Pan, L., Zeng, Q., Sun, W., & Wan, W. Based on TPUGRAPHS Predicting Model Runtimes Using Graph Neural Networks. https://api.semanticscholar.org/Corpus.

[14] Yao, J., Zou, Y., Du, S., Wu, H., & Yuan, B. Progress in the Application of Artificial Intelligence in Ultrasound Diagnosis of Breast Cancer. DOI:https://api.semanticscholar.org/Corpus.

[15] Zhou Y, Chen S, Wu Y, Li L, Lou Q, Chen Y, Xu S. Multi-clinical index classifier combined with AI algorithm model to predict the prognosis of gallbladder cancer. Front Oncol. 2023 May 10;13:1171837. DOI: 10.3389/fonc.2023.1171837. PMID: 37234992; PMCID: PMC10206143.

[16] Li L, Xu C, Wu W, et al. Zero-resource knowledge-grounded dialogue generation[J]. Advances in Neural Information Processing Systems, 2020, 33: 8475-8485. DOI: https:// doi.org/ 10.48550/arXiv.2008.12918.