Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Fraud Detection Method for Low-frequency Transaction

Zhaohui Zhang [1,2,3], **Ligong Chen** [1], **Qiuwen Liu** [1], **Pengwei Wang** [1]
[1] School of Computer Science and Technology, Donghua University, Shanghai, China
[2] The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China
[3] Shanghai Engineering Research Center of Network Information Services, Shanghai, China

Corresponding author: Zhaohui Zhang (e-mail: zhzhang@dhu.edu.cn).

This work was supported by Natural Science Foundation of Shanghai (No. 19ZR1401900), Shanghai Science and Technology Innovation Action Plan Project (No. 19511101802), and National Natural Science Foundation of China (No. 61472004, 61602109).

ABSTRACT The effectiveness of transaction fraud detection methods directly affects the loss of users in online transactions. However, for low-frequency users with small transaction volume, the existing methods cannot accurately describe their transaction behaviors for each user, or lead to a high misjudgment rate. So we propose a new method for individual behavior construction, which can make the behavior of low-frequency users more accurate by migrating the current transaction group behavior and transaction status. Firstly, we consider the user's only historical transactions, combined with the optimal risk threshold determination algorithm, to form the user's own transaction behavior benchmark. Secondly, through the DBSCAN clustering algorithm, the behavior characteristics of all current normal samples and fraud samples are extracted to form the common behavior of the current transaction group. Finally, based on historical transaction records, the current transaction status is extracted using a sliding window mechanism. The combination of the three constitutes a new transaction behavior of the user. On this basis, a multi-behavior detection model based on new transaction behavior is proposed. According to the result of each behavior, Naive Bayes model is used to calculate the probability that current transaction belongs to fraud, and finally determine whether current transaction is fraud. Experiments prove that the method proposed in this paper can have a good effect on low-frequency users, which can accurately identify fraud transactions and has a low misjudgment rate for normal transactions.

INDEX TERMS Transaction detection; Low-frequency users; Individual behavior; Group behavior; DBSCAN; Naive Bayes

I. INTRODUCTION

With the rapid development of e-commerce, online payment has become more and more popular. However, the safety of online transactions has become increasingly serious. According to report by the Beijing Municipal Public Security Bureau Network Security Team and the 360 Company's Hunting Network Platform, the number of cyber fraud criminals exceeded 1.6 million yuan, and online scam market reached 110 billion yuan. According to the relevant departments to analyze existing fraud cases, the main means of fraud crimes are hacking, stolen cards, credit card cashing, phishing websites, Trojan horses, etc. [1]. Criminals steal user information indirectly or directly through these methods, thus stealing user money. Although fraud is serious, for a large number of normal transactions, the number of fraud transactions accounts for only a small proportion, so the sample of financial transactions is unbalanced. If a model is trained directly on unbalanced samples, it often fails to achieve good results. How to effectively prevent the risk of transaction fraud has become a problem to be solved.

In view of above phenomenon, it is very effective to solve such problems from the perspective of user behavior authentication. Because user behavior is unique, the credibility of the current behavior of the system can be verified from the perspective of behavior authentication. And the proportion of normal transactions and abnormal transactions of a single user is better than proportion of all users. At present, major financial institutions mainly adopt the following methods to verify identity and behavior of users. The first is based on account password matching mechanism [2], however existing matching mechanisms are only used to verify password, and can't identify the malicious user's network behavior. The second is to use expert system based on expert rules to achieve fraud detection by anti-fraud experts[3]. However, expert system can't adapt to new fraud methods, and there will be rule redundancy[19]. The third is to analyze group trading users and use machine learning and deep learning methods to mine common behaviors of users, such as neural networks[4], random forests[7][8], relationship networks[13], HMM[14], and so on. However, these methods train models based on transaction data for all users. They can only mine the common characteristics of all users, it is difficult to learn the individual behavior of each user[15][24], and such models cannot effectively identify new types of fraud[19][20][21]. The fourth is to extract user behavior pattern from the perspective of individual user, and then determine whether user's current behavior is fraud, including user's browsing behavior [15], mouse and keystrokes behavior[16], transaction behavior[17][20][21], etc. However, if user does not have enough historical transaction records, it is difficult to ensure that the extracted user behavior pattern is accurate enough[14][22][23][24][25][26]. In addition, in view of imbalanced characteristics of financial transaction samples, there are currently many studies to alleviate this problem, such as the GMM-based undersampling method[27] and the hybrid cross feature extraction method[28].

In summary, the current main problem is that it is difficult for the existing individual behavior models to accurately learn the behavior characteristics of low-frequency users with low transaction volume, which leads to a high misjudgment rate for low-frequency users. Considering that fraudsters often commit fraud against multiple people, this article proposes that it is easier to find traces of fraudster behavior by supplementing the transaction behavior and transaction status of the current transaction group, the specific contributions are as follows. First, it is proved through experiments that the frequency of transactions has a great impact on accuracy of individual behaviors. Existing models can't accurately describe transaction behaviors of low-frequency users. Secondly, a new low-frequency user transaction behavior construction method is proposed. The clustering algorithm is used to extract current group user's behavior to supplement low-frequency user's behavior, and considering current transaction status, the behavior of low-frequency user can be fully characterized. Finally, this paper proposes a multi-behavior detection model based on user behavior, and uses the Naive Bayes formula to determine whether the current transaction is a fraudulent transaction based on the results of each behavior detection.

The rest of this paper is organized as follows. The second section introduces the related work, the third section discusses the construction process of low-frequency user behavior, the fourth section introduces transaction detection method, the fifth section introduces data source and experimental results, and the sixth section summarizes research results and future prospects.

II. RELATED WORK

Fraud detection is similar to classification problems. Therefore, based on user group behavior, using machine learning or deep learning to detect fraud transactions has been widely studied in recent years. Zhang *et al.*[4] used the features of the convolutional neural network to derive features. By inputting the original features and adding a feature arrangement layer to combine the inputs, they achieved good results in fraud detection. Zhou *et al.*[5] proposed a siamese neural network structure based on CNN and LSTM. They used the siamese neural network structure to solve the problem of sample imbalance in online transactions, and used the LSTM structure to make model memory user's transaction information.Wang et al.[6] used neural network-based embedding to capture detailed information about user click actions, and used recurrent neural networks to model such click sequences. Xuan *et al.*[8] used random forests to train the characteristics of normal and abnormal behaviors, compared two random forest models with different base classifiers, and proved that random forests perform well in credit fraud detection. Whitrow *et al.*[9] considered a framework for transaction aggregation and used various classification methods and cost-based indicators to evaluate the effectiveness of various transaction detection methods, and found that random forests outperform other methods. Fu *et al.*[10] proposed the derived characteristics of transaction entropy to characterize the user's transaction behavior, and converted the original one-dimensional transaction data into a two-dimensional transaction matrix and input the convolutional neural network to establish a credit card fraud detection model based on deep learning. Dal Pozzolo *et al.*[11] Considered a concept drift and sample imbalance in fraud detection, proposed a new learning strategy, and verified it on real data. Kim *et al.*[12] Made in-depth comparisons of hybrid-to-hybrid integration and deep learning methods, and developed two detection models. The real data proved that the deep learning model has good performance. Meng *et al.*[13] Proposed a detection method based on relational networks and boosting trees. First, a transaction relationship network was constructed for users, and secondly, information was extracted through the network for transaction identification, which can well identify fraudulent transactions.

From another perspective, most of the existing fraud methods use direct or indirect methods to steal user information and use the identity of normal users to perform fraud. Therefore, there have been some studies in recent years to solve the problem of transaction fraud from the perspective of individual user behavior. Ji *et al.*[18] proposed a method for e-commerce user abnormal behavior detection research, based on user historical behavior data to establish user's normal behavior pattern, and finally used the pattern comparison method to determine user's transactions. Kültür *et al.*[19] proposed a new cardholder behavior model for credit card fraud detection, which used this model to detect fraud transactions in combination with user historical consumption behavior. Zheng *et al.*[20] proposed a new credit card fraud detection system based on behavior certificate. They built user behavior certificates based on user historical transactions and used them for fraud detection. At the same year, Zheng *et al.*[21] proposed a transaction fraud detection method based on total order relationship and behavior diversity. They defined user transaction logic diagrams, attribute transition probabilities, diversity coefficients, and state transition matrices to build a behavior summary for each user. Xie *et al.*[22] proposed a rule-based feature project that took into account individual and group behaviors and depicted individual behaviors as group characteristics, which can more effectively distinguish between normal transactions and fraud transactions. Nami *et al.* [23] proposed a cost-sensitive payment card fraud detection method based on dynamic random forest and k-nearest neighbors, including two stages: cardholder behavior pattern matching and dynamic random forest detection. Jiang *et al.* [24] proposed a new method that utilizes aggregation strategies and feedback mechanisms. First, all cardholders were divided into different groups, and then the behavior patterns of each group were extracted to train classifiers for each group. Finally, a classifier set was used to detect fraud online.

In the above work, the group behavior method is to train a classifier based on all user transaction data. They can only find the common characteristics of users, and the unique characteristics of each user will be overwhelmed during training[15][24]. And compared with individual behavior models, they are not very effective in identifying new types of fraud[19][20][21]. However, the individual behavior method also faces the problem of sparseness of low-frequency user data. The existing individual behavior research is based on the user's possession of sufficient historical data and have not provided effective solutions for low-frequency users. For example, the data amount of each user in the literature [15], [18], [19], [20], [21] is more than 50. The literature [23] considers that the detection method is invalid for users with sparse data, so users with a transaction volume of less than 50 are deleted during the individual behavior modeling stage. The literature [24] considers that a single user has data sparseness, so it considers multiple similar users as similar and trains a classifier for a class of users.

III. LOW-FREQUENCY USER BEHAVIOR BENCHMARK

This section will introduce a behavioral benchmark construction method for low-frequency users. The method mainly consists of three parts, as shown Fig. 1. The first part is user's own transaction behavior. According to user's historical normal transaction records, user's transaction behavior is extracted from the transaction records. The second part is the existing fraud behavior and normal behavior in current trading group, clustering multiple behavior patterns through DBSCAN as the group behavior of current transaction. The third part is through sliding window mechanism, which counts the proportion of fraud transactions in all transactions in current time period as the current transaction status. The combination of the three together builds the behavior benchmark of low-frequency users.

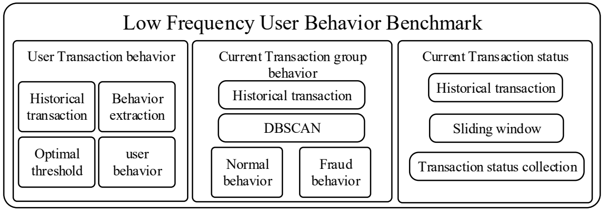


FIGURE 1. user behavior benchmark

A. USER BEHAVIOR EXTRACTION

A transaction r contains m attributes, write as $\mathrm{r} = \{a_1, \cdots, a_m | a_1 \in A_1, \cdots, a_m \in A_m\}$. The transaction attributes in this paper include transaction user number, transaction time, transaction amount and transaction IP address, transaction $r = \{a_{no}, a_{time}, a_{amount}, a_{ip}\}$. User u's transaction log is a collection of historical transactions up to the current date, which is denoted as $L_u = \{r_1^u, r_2^u, \cdots, r_n^u\}$, where $n_u = |L_u|$, as the number of transaction records of the user. Extract the normal transaction $T_u = \{t \in L_u | label = normal\}$ from user transaction log, where $n_{tu} = |T_u|$. For user's normal transaction T_u, we need to do further processing to get user's behavior benchmark.

The specific processing method is introduced in the user behavior extraction part in literature [17]. First, according to the user's historical normal transaction records, the user's transaction amount, transaction time, and transaction IP address are processed separately, taking into account changes in user transactions over time, and performing feature processing and feature derivation to obtain user transaction behavior. Then calculate the optimal risk threshold of each user according to the optimal risk threshold algorithm, and finally we can get the behavior benchmark UBB_u based on user behavior, $UBB_u = [TB_u, Threshold^u]$. Where $Threshold^u$ represents the optimal risk threshold of each user, TB_u represents the user's transaction behavior, $TB_u = (TAR^u, TAC^u, TIW^u, TTR^u, TFA^u, TIP^u, PTS^u)$.

TAR^u represents the user's transaction amount attribute.
TAC^u represents the user's transaction change attribute.
TIW^urepresents the user's transaction workdays attribute.
TTR^u represents the user's transaction time attribute.
TFA^u represents the user's transaction frequency attribute.
TIP^u represents the user's transaction IP address attribute.
PTS^u represents the user's previous transaction status.

B. EXTRACT CURRENT TRADING GROUP BEHAVIOR

The study finds that the accuracy of user behavior description has a great relationship with the user's transaction frequency [17]. However, the low-frequency user transaction volume is scarce. The user behavior generated by the above methods are difficult to guarantee high accuracy, so it is also needed enrich the user's own behavior through other means.

By analyzing the existing fraud cases, it is found that fraud is often dominated by gangs, and fraud transactions initiated by the same gang often have the same behavior pattern, and fraudsters often commit fraud against multiple people at the same time. So, this paper analyzes the fraud transactions and normal transactions that have occurred, and uses algorithms to extract the respective behavior patterns of normal transactions and fraud transactions to compensate the transaction behavior of low-frequency users.

As the distribution of transaction data is uneven, and the data shape is different, the number of transaction categories can't be predicted in advance. Therefore, the DBSCAN method is used to cluster historical transaction records. Compared with other clustering algorithms, DBSCAN has the following advantages[29]. First, it can cluster dense data sets of any shape. Second, it can't predict the number of clusters in advance, and it is not limited by the shape of the sample. Third, the model can find outliers while clustering, and is not sensitive to outliers in the data set. Fourth, there is no bias in the clustering results.

Record all transactions that have occurred as $L = \{r_1^{u_i}, \dots, r_{\mathrm{k}}^{u_j}, \cdots, r_n^{u_k}\}$, $n = |L|$, where $r_{\mathrm{k}}^{u_i}$ represents the *kth* transaction generated by user u_i. The normal transaction $T = \{t \in L | label = normal\}$ and the fraud transaction $F = \{t \in L | label = fraud\}$ are extracted from transactions L, where $n_T = |T|, n_F = |F|$, and $n = n_T + n_F$. Then take the normal transaction and the fraud transaction as input, get the result of each transaction through DBSCAN, and label them with $\{B_1, \dots, B_P\}$and $\{W_1, \dots, W_Q\}$. In the current historical transactions, fraud transactions are classified into P categories, and normal transactions are classified into Q categories. which are represented as follows.

$$\begin{cases} r_1^{u_i}, label = normal, target = W_1 \\ \dots \\ r_{\mathrm{k1}}^{u_j}, label = normal, target = W_Q \\ r_{\mathrm{k2}}^{u_i}, label = fraud, target = B_1 \\ \dots \\ r_{\mathrm{n}}^{u_j}, label = fraud, target = B_P \end{cases}$$

Based on the above results, we deal with transaction records L according to label categories, and obtain $P + Q$ transaction sets, which are recorded as $T_{B1}, \dots, T_{BP}, T_{W1}, \dots, T_{WQ}$, where $T_{B1}, \dots, T_{BP}$ belong to fraud transaction, and $T_{W1}, \dots, T_{WQ}$ belong to normal transaction.

For each transaction set, we use the behavior extraction method in literature [17] to determine the transaction behavior of each type of transaction. As the transaction status attributes of each type of transaction belong to the same category, we do not need to consider the user's previous transaction state when extracting the transaction behavior. Therefore, we will get the behavior of the current trading group G$BB = [TBB, FBB]$, where TBB and FBB are the normal transaction behavior matrix and fraud transaction behavior matrix.

$$TBB = \begin{bmatrix} TAR^{W1}, TAC^{W1}, TIW^{W1}, TTR^{W1}, TFA^{W1}, TIP^{W1} \\ \dots \\ TAR^{WQ}, TAC^{WQ}, TIW^{WQ}, TTR^{WQ}, TFA^{WQ}, TIP^{WQ} \end{bmatrix}$$

$$FBB = \begin{bmatrix} TAR^{B1}, TAC^{B1}, TIW^{B1}, TTR^{B1}, TFA^{B1}, TIP^{B1} \\ \dots \\ TAR^{BP}, TAC^{BP}, TIW^{BP}, TTR^{BP}, TFA^{BP}, TIP^{BP} \end{bmatrix}$$

C. CALCULATE CURRENT TRANSACTION STATUS

Internet transaction fraud is mostly based on phishing websites and Trojan viruses. It has been found that fraudsters tend to transfer the property of normal users in a short period of time, so the time interval of transactions is much shorter and occurs continuously. Based on this, by analyzing the number of fraud transactions in the current time period, it is used to indicate a state of the current time period transactions. If there are a large number of fraud transactions, the suspiciousness of the current transactions will increase. Therefore, this paper proposes a sliding window mechanism to count the proportion of fraud transactions in the current time period, and use this to indicate the status of current transactions.

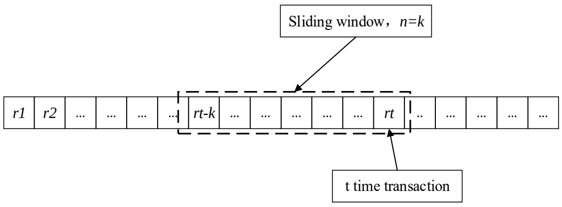


FIGURE 2. sliding window

As shown in Fig. 2, the transaction at time *t* is denoted as r_t, and the transaction sequence occurring before time *t* is denoted as $L_t = \{r_1, \dots, r_t\}$. The sliding window is h, and the sliding window length is n_h, and $n_h = k$. The transaction sequence in the sliding window is recorded as $L_{ht} = \{r_{t-k}, \dots, r_t\}$. And in the sliding window ht at time *t*, the fraud transaction $F_{ht} = \{r \in L_{ht} | label = fraud\}$, $n_{F_{ht}} = |F_{ht}|$. The transaction status in the current time period is recorded as S_t, which represents the proportion of fraud transactions in the first *k* transactions in the current time period. The calculation method is shown in (1).

$$s_t = n_{F_{ht}} / n_h \qquad (1)$$

We use the sliding window *h* to aggregate all the transaction sets L that have occurred, and find the proportion of fraud transactions in each window, will get $n - n_h$ ratio sequence $s = [s_1, \dots, s_{n-n_h}]$. The mean value of the sequence $Threshold^S$ is taken as a critical value of the historical transaction state. If the transaction state s_i is higher than the critical value, $s_i = 1$, otherwise $s_i = 0$.

D. LOW-FREQUENCY USER TRANSACTION BEHAVIOR

Through the above work, the user's own behavior UBB_u, current trading group behavior GBB and current transaction status S_t are obtained. So the supplementary low-frequency user behavior is represented by a triple, which is recorded as $New\_UBB_u = (UBB_u, GBB, S_t)$, and described as follows.

$UBB_u = [TB_u, Threshold^u]$ is the user's own transaction behavior extracted through the user's historical normal transaction records.

$GBB = [TBB, FBB]$ is a group of normal behavior patterns TBB and fraud behavior patterns FBB obtained by analyzing the current trading group.

S_t indicates the status of the transactions in the current time period.

IV. USER BEHAVIOR BASED DETECTION METHOD

In the above work, the user behavior New_UBB_u is divided into three parts, UBB_u, GBB and S_t. Based on the hyper-sphere model of the literature [17], we propose a new detection method. When the user's transaction r^u enters the current system, the system will give the judgment result to the current transaction through the following method, the specific process is shown in Fig. 3.

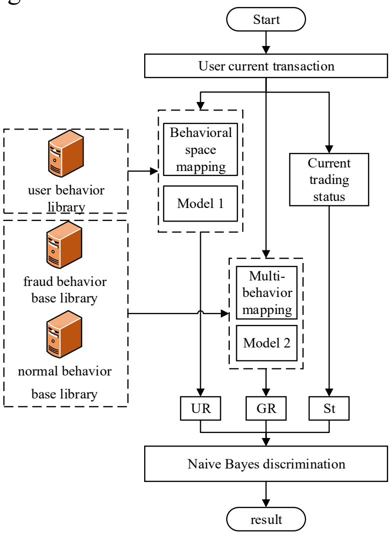


FIGURE 3. detection method flow chart

A. SELF-BEHAVIOR DETECTION

The user's own behavior benchmark $UBB_u = [TB_u, Threshold^u]$ consists of two parts, they are user's own transaction behavior TB_u and the user's optimal risk threshold $Threshold^u$. According to the hyper-sphere model proposed in literature [17], TB_u is regarded as a point in the space, and $Threshold^u$ is regarded as the radius in space. The transaction detection is transformed into a mapping problem of points in multidimensional space, and the (2) is used to judge whether the current transaction is normal.

$$f_1(x) = \sum_{i=1}^{n}(x_i - tb_i)^2 - (Threshold^u)^2 \qquad (2)$$

Where x_i is the value of each dimension of the user's current transaction r^u mapped to the point in the current multidimensional space, tb_i is the value of each dimension of the user's own transaction behavior. The above formula divides the space into two parts, if $f_1(x) > 0$, record as a fraud transaction, otherwise record as a normal transaction, thus obtaining the user's current transaction result UR_t.

$$UR_t = \begin{cases} 0, f_1(x) \le 0 \\ 1, f_1(x) > 0 \end{cases} \qquad (3)$$

B. CURRENT GROUP BEHAVIOR DETECTION

Current trading group behavior $GBB = [TBB, FBB]$, where TBB is a matrix of $Q \times 6$, representing the normal transaction behavior of Q class, and FBB is a matrix of $P \times 6$, representing the fraud transaction behavior of P class. For the transaction r^u currently entering the system, convert it to a matrix $r^T = [x1, x2, x3, x4, x5, x6]$. Then, use (4) to calculate the deviation distance of this transaction from each group behavior benchmark (TBB and FBB), and the following two distance sets D_{TBB} and D_{FBB} will be obtained, where $D_{TBB} = [d_1, \dots, d_q]$, $D_{FBB} = [d_1, \dots, d_P]$. The (5) will be used to determine which behavior the current transaction is biased in. If the transaction is more biased towards the fraud transaction behavior, the transaction is more suspicious.

$$d\left(\overrightarrow{r^T}, \overrightarrow{tbb_{j1}(fbb_{j2})}\right) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (4)$$

Formula (4) represents the calculation of deviation distance between current transaction r^u and each of the behavioral benchmarks in TBB or FBB, where $\overrightarrow{r^T}$ represents the matrix vector after current transaction is converted into a matrix, $\overrightarrow{tbb_{j1}(fbb_{j2})}$ represents the behavior vector of each behavior in TBB or FBB, where $1 \le j1 \le Q$， $1 \le j2 \le P$.

$$f_2(x) = \frac{1}{P}\sum_{i=1}^{P} d_i - \frac{1}{Q}\sum_{j=1}^{Q} d_j \qquad (5)$$

Formula (5) represents the current transaction tendency degree. The first part is the mean deviation of current transaction r^u from fraud behavior. The second part is the mean deviation of current transaction r^u from normal behavior, the difference between the two is $f_2(x)$. If $f_2(x)$ is too small, the distance between current transaction r^u and fraud sample is less than distance from normal sample, and the suspicious degree is increased; otherwise, the suspicious degree is reduced. Therefore, the current r^u transaction result GR_t of user can be obtained, as shown in (6), where $Threshold^G$ is the optimal threshold calculated by optimal risk threshold algorithm.

$$GR_t = \begin{cases} 1, f_1(x) \le Threshold^G \\ 0, f_1(x) > Threshold^G \end{cases} \qquad (6)$$

C. TRANSACTION DETECTION

After the user's own behavior detection and the current group behavior detection, the judgment results $UR_t$and GR_t of the transaction will be obtained, combined with current transaction status S_t, using Naive Bayes model judges the current transaction results.

$$\mathrm{P(Y = k), k = 0,1} \qquad (7)$$

In the historical transaction record L, UR_i, GR_i and S_i are obtained for each transaction, and the category k to which the transaction currently belongs. Based on this, the probability of historical fraud transactions and normal transactions can be calculated.

$$P(X=x|Y=k)=\mathrm{P}(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}|Y=k) \tag{8}$$

Calculate the conditional probability distribution according to the historical transaction record, where X is the random vector in the input space. In this paper, $X=(UR,GR,S)$. Because the user's own detection result, the current trading group detection result and the current transaction status do not affect each other, and the conditions of the independent hypothesis are met. The conditional probability formula is as shown in (9).

$$\begin{aligned}P(X=x|Y=k)&=\mathrm{P}\left(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}|Y=k\right)\\&=\prod_{j=1}^{n}\mathrm{P}\left(X^{(j)}=x^{(j)}|Y=k\right)\end{aligned} \tag{9}$$

For the current transaction r^{u}, according to the user's own behavior and the current group behavior, the judgment results of this transaction are UR_t and GR_t, and the transaction status S_t, $x=(UR_t,GR_t,S_t)$ is the current input. Then, using (10) and (11), the transaction r_u is calculated to belong to the fraud probability $P(Y=1|X=x)$ and the normal probability $P(Y=0|X=x)$.

$$\begin{aligned}P(Y=1|X=x)&=\frac{P(X=x|Y=1)P(Y=1)}{\sum_{k=0}^{1}P(X=x|Y=k)P(Y=k)}\\&=\frac{P(X=x|Y=1)P(Y=1)}{\sum_{k=0}^{1}P(Y=k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}|Y=k\right)}\end{aligned} \tag{10}$$

$$\begin{aligned}P(Y=0|X=x)&=\frac{P(X=x|Y=0)P(Y=0)}{\sum_{k=0}^{1}P(X=x|Y=k)P(Y=k)}\\&=\frac{P(X=x|Y=0)P(Y=0)}{\sum_{k=0}^{1}P(Y=k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}|Y=k\right)}\end{aligned} \tag{11}$$

If $P(Y=1|X=x)>P(Y=0|X=x)$, the transaction r^{u} belongs to a fraud transaction, otherwise it belongs to normal transaction.

V. EXPERIMENTAL RESULTS

This section describes the detection effects of the proposed model. First introduce the source of the experimental data, then introduce the low-frequency user division and clustering results, and finally introduce the comparison of this paper with other models.

A. DATA SET

The data set comes from a domestic bank data, which covers 3 months transactions and contains 92133 users, 3502048 transaction records, each transaction record is marked by the bank for the label. In Fig. 4, all transactions are referred to as data set *A*, and all fraud transactions are in data set *F*, which contains 65138 transaction data for 14751 users. The data set shows that the main types of fraud include phishing sites, Trojan viruses, etc. Transactions involving 14751 users are in data set *B*, all transactions for fraud-free transaction users are in data set *C*. We randomly extract transaction data of any number of users from dataset B as experimental data. The test set and the training set are divided according to time, the data of April and May are used as the training set, and the data of June is used as the test set.

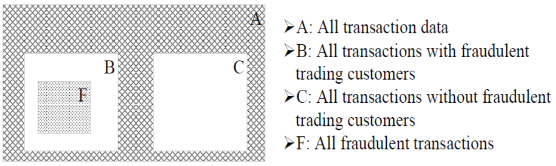

A
B
C
F
➢A: All transaction data
➢B: All transactions with fraudulent trading customers
➢C: All transactions without fraudulent trading customers
➢F: All fraudulent transactions

FIGURE 4. data set

B. DIVISION OF LOW-FREQUEQUNCY USERS

In order to solve the problem of high misjudgment rate of low-frequency users, it is first necessary to determine what is a low-frequency user. Literature [17] proposed that the model detection effect is related to the frequency of user transactions, and the results shown in the Fig. 5 also confirm this. Therefore, this paper uses its proposed model, and uses the model detection precision as an evaluation index to divide high-frequency users and low-frequency users.

TABLE 1. data set D1

data set	D10	D11	D12	D13	D14
Average user volume	9.41	17.58	26.56	36.52	47.37
data set	D15	D16	D17	D18	D19
Average user volume	52.50	65.65	79.48	88.53	99.37

TABLE 2. data set D2

data set	D20	D21	D22	D23	D24
Average user volume	6.50	14.58	26.50	35.50	45.25
data set	D25	D26	D27	D28	D29
Average user volume	55.25	61.75	75.25	88.17	96.63

TABLE 3. data set D3

data set	D30	D31	D32	D33	D34
Average user volume	8.75	19.16	28.33	36.58	44.16
data set	D35	D36	D37	D38	D39
Average user volume	53.92	66.17	73.92	88.50	93.91

In this paper, three sets of data *D1*, *D2* and *D3* are randomly selected in data set *B*, each set of data contains three months of transaction data of 150 users. Among them, each set of data set D_i contains ten sets of data, denoted asD_{ij}, $0\le j\le 9$, the transaction frequency is between $j\times 10\sim(j+1)\times 10$, the details are shown in table 1 to 3. The precision of each group of users is calculated by the hyper-sphere model to determine the range of low-frequency user transaction volume. The experimental results are shown in Fig. 5. When the user transaction volume is more than 30, the model detection precision is relatively stable, and both are higher than 80%; when the user transaction frequency is less than 30, the precision of the model detection will fluctuate greatly. And the precision rate is less than 80%, so this paper considers users with less than 30 transactions as low-frequency users. The next experiment in this paper will focus on these users.

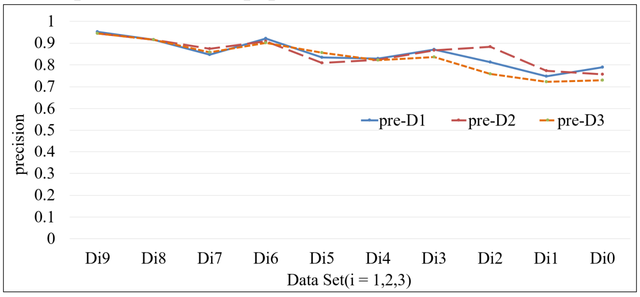
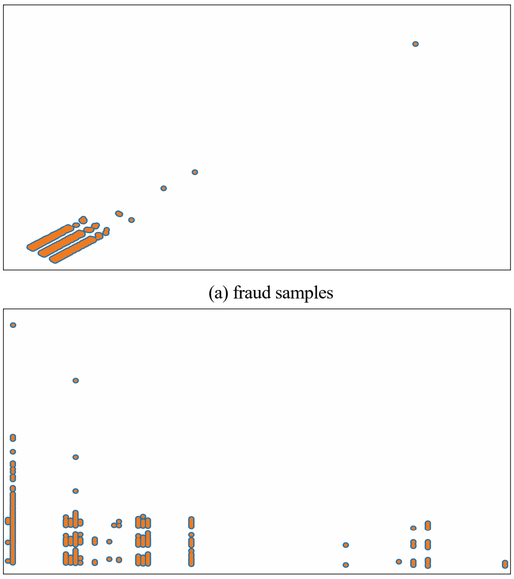


FIGURE 5. Relationship between precision and transaction frequency

C. CLUSTERING RESULTS

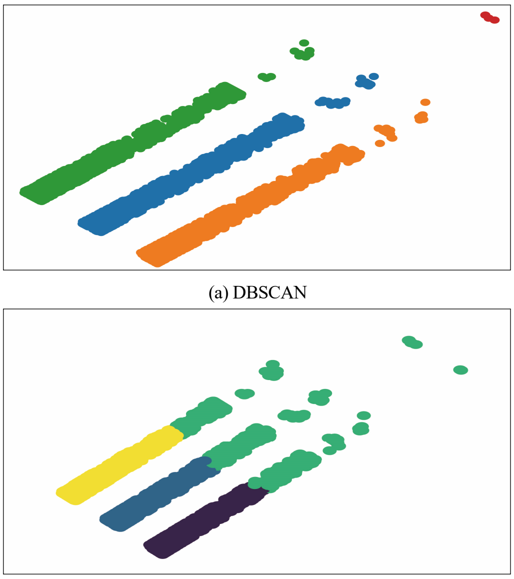


(a) fraud samples

(b) normal samples

FIGURE 6. samples distribution

Fig. 6 shows the distribution of historical transactions, with the distribution of fraud samples is (a) and the normal sample is (b). As can be seen from the figures, the data is not cluster-like and has different shapes. The traditional K-means clustering algorithm is not suitable for such data sets. Therefore, when extracting the behavior of the current trading group, this paper uses the density clustering algorithm DBSCAN to cluster the transactions. The following figures are comparison of the clustering effects of the two methods on fraud transactions.



(a) DBSCAN

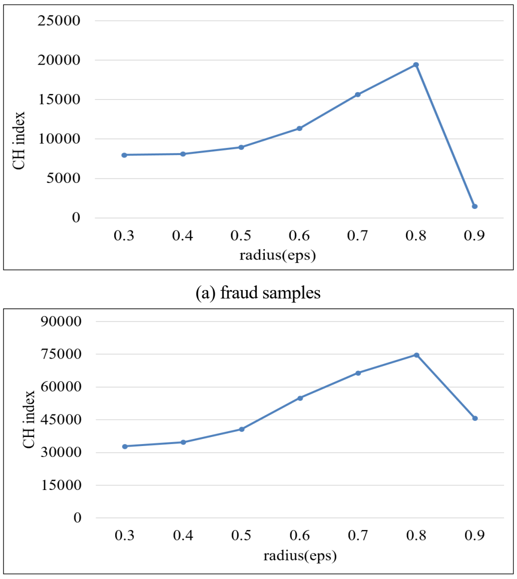
(b) K-means

FIGURE 7. Clustering effect comparison

It can be seen from the Fig. 7. DBSCAN can clearly cluster the samples into different classes, but K-means can't, so DBSCAN has better clustering effect. But the DBSCAN clustering effect is affected by two parameters, radius and density. In this paper, the *CH* index is used to judge the clustering effect of the algorithm, and its calculation method is shown in (12).

$$CH(k) = \frac{trB(k)/(k-1)}{trW(k)/(n-k)} \tag{12}$$

Where n denotes the number of clusters, k denotes the current class, $trB(k)$ denotes the trace of the inter-class dispersion matrix, and $trW(k)$ denotes the trace of the intra-class dispersion matrix. If the larger the *CH*, the closer the class itself is, the more dispersed the classes are, the clustering result is better. As shown in Fig. 8, when the density is 3, the value of the radius is continuously adjusted. When the value of the CH index reaches the maximum, the clustering effect is the best. At this time, the number of fraud sample clusters is 4 and the number of normal sample clusters is 16. Therefore, the fraud samples in this paper are divided into 4 categories, and the normal samples are divided into 16 categories.



(a) fraud samples

(b) normal samples

FIGURE 8. CH index

D. COMPARISON OF RESULTS

In this section, we compare our model (NM) with the following models, which are the hyper-sphere model (UR) proposed in [17], and the detection model based on user user behavior certificate (UBC) proposed in [20], DBSCAN-based group behavior model (GR) and XgBoost (XB) model. Among them, UR and GR are the individual behaviors and group behaviors extracted when constructing low-frequency user behaviors.

TABLE 4. Confusion matrix

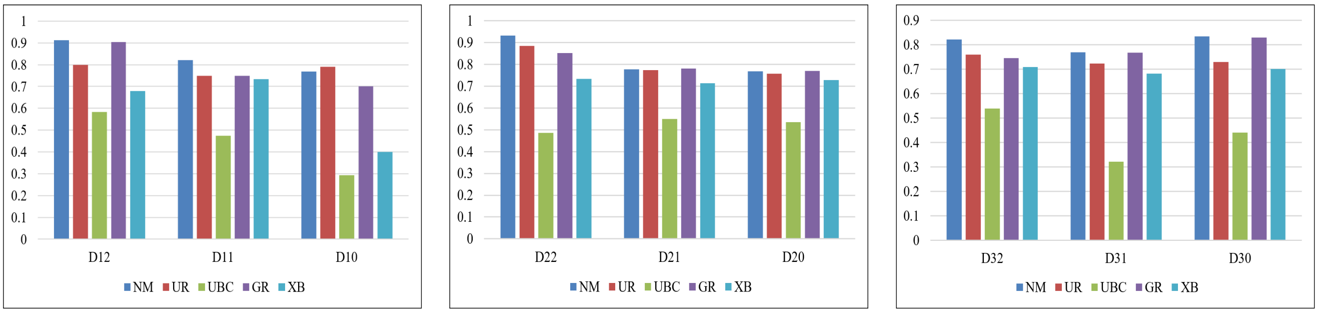
	True Fraud	**True Normal**
Pre Fraud	TP	FP
Pre Normal	FN	TN

As shown in table 4, because it is detecting fraudulent transactions, so the confusion matrix is slightly modified. TP is the number of fraud transactions that are judged as fraud transactions by the model. FP is the number of normal transactions that are judged as fraud transactions by the model. TN is the number of normal transactions that are judged as normal transactions by the model. FN is the number of fraud transactions that are judged as normal transactions by the model.

TABLE 5. Indicator calculation method

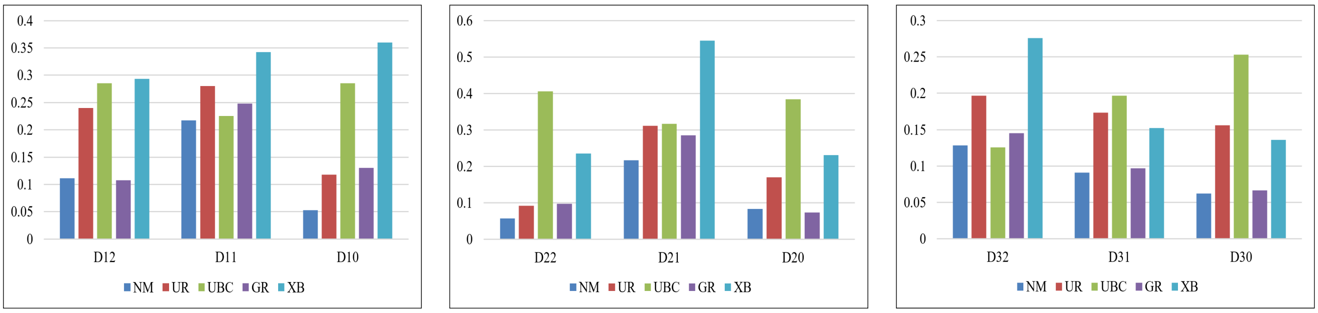
Indicator name	calculation method
Precision	TP/(TP +FP)
Disturbance	FP/(TN+FP)
Recall	TP/(TP+FN)
F1	2 × precision × recall/(precision+ recall)

In order to make the comparison results more convincing, we use several indicators commonly used in fraud detection as the evaluation indicators, including precision, recall, and F1. In order to consider the model's misjudgment of normal transactions, the disturbance also needs to be considered. The calculation method is shown in Table 5. The precision is the ratio of the number of fraudulent transactions judged by the model to the total fraudulent transactions detected by the model. The recall is the ratio of the number of fraudulent transactions judged by the model to the total fraudulent transactions. The F1 is the harmonic average of the precision and recall. The disturbance is the proportion of the model that miscalculated the real normal transactions as abnormal transactions to the total normal transactions.



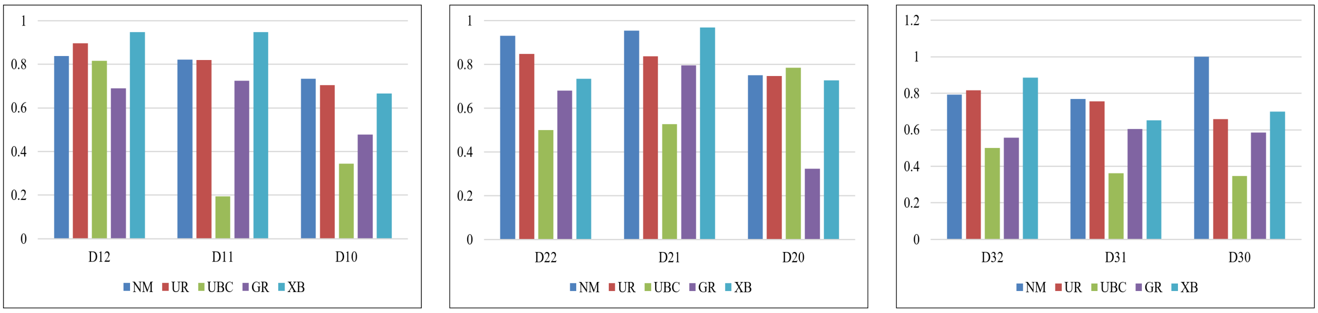
(a) precision(D1) (b) precision(D2) (c) precision(D3)

FIGURE 9. plot of precision comparison results



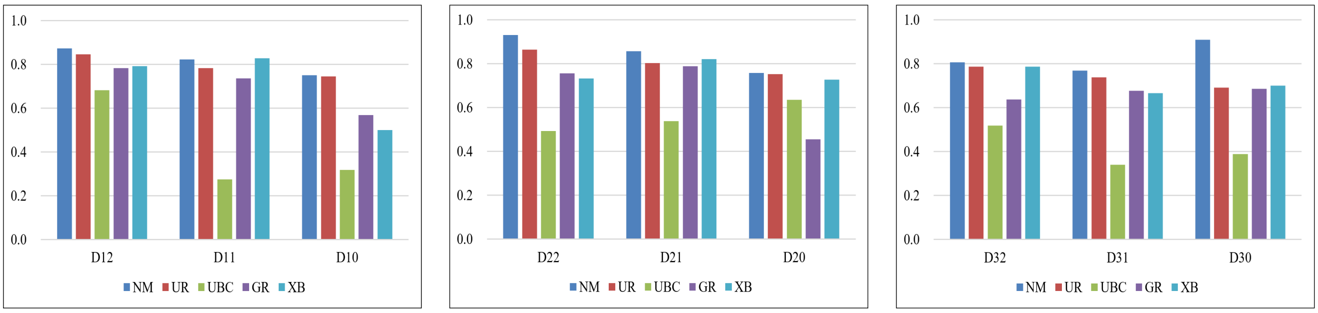
(a) disturbance(D1) (b) disturbance(D2) (c) disturbance(D3)

FIGURE 10. plot of disturbance comparison results



(a) recall (D1) (b) recall (D2) (c) recall (D3)

FIGURE 11. plot of recall comparison results



(a) F1 (D1)

(b) F1 (D2)

(c) F1 (D3)

FIGURE 12. plot of F1 comparison results

As can be seen from Fig. 9 to Fig. 12, compared with the individual behavior models (UR, UBC) and group behavior models (GB, XB), the model NM performs better among the three groups of low-frequency users. The average precision is improved by about 5%, 35%, 4%, and 14%, and the average disturbance is reduced by about 8%, 15%, 3%, and 17%. In the recall, the model NM still performs better than other models. Although occasionally the model XB and UR recall rates are higher than NM, but the precision of the model XB and UR at this time is lower and the disturbance is higher, indicating that the model XB and UR have higher recall rates in the case of misjudgment of a large number of normal transactions. The F1 value is the harmonic average of the recall and precision, and represents the overall performance of the model. At the F1 value, the model NM is higher than the models UR, UBC, GR, and XB, and the average increase is 5%, 35%, 15% and 10%. It has been proved through experiments that the model NM performs better than other models (UR, UBC, GR, XB) in the detection of low-frequency user transactions. It accurately detects fraudulent transactions and reduces the misjudgment of normal transactions. If only the individual behavior model UR and UBC are compared, the model UR performs better regardless of the disturbance rate or F1 value. If only the group behavior model GR and XB are compared, although the F1 values perform similarly, the model GR is better than XB on the disturbance rate, indicating that the individual behavior and group behavior proposed in this paper perform well on low-frequency users.

After analysis, the main reasons are as follows. The group behavior model cannot accurately describe the transaction behavior of each user, resulting in a decrease in model effect. However, the model GR effect is better than XB, because GR is a clustering process for transactions that have already occurred, and the behavior of similar classes is extracted, and the effect is better than the model XB. The individual behavior model UR and UBC only consider the low-frequency users themselves, because the low-frequency user data is scarce, the behavioral characterization is not accurate enough. However, the model NM not only considers the users themselves, but also considers the behavior of the current trading group, and can supplement the low-frequency user's own behavior to a certain extent. At the same time, the model NM proposes a sliding window mechanism to calculate the current transaction status. Moreover, the model NM proposes a comprehensive decision-making method based on the detection results of various behaviors, which is more fully considered than other models. Therefore, the overall performance of the model NM is better than the model UR, UBC , XR and XB.

VI. CONCLUSION

In this paper, a new method for low-frequency user transaction detection is proposed for the problem that low-frequency users cannot accurately describe transaction behavior. Compared with other methods, the proposed method not only considers the low-frequency users' own transaction behavior, but also considers the current trading group behavior and current transaction status, and constitutes a new low-frequency user behavior. Based on this, a method based on user behavior and Naive Bayes detection is proposed to judge the user's current transaction. And in the experiment, the division problem of high-frequency users and low-frequency users is considered. Experiments prove that the precision and F1 of the proposed method is higher than other models, although the recall is not the highest on a few data sets, the disturbance rate has always been lower than other models. It shows that in the detection of low-frequency user transactions, the model proposed in this paper can more accurately identify fraud transactions and has lower misjudgment of normal transactions. In the future we will pay more attention to the problem of online model updating and make the method more effective.

REFERENCES

[1] S. 2018, "Research report on the trend of network fraud in 2017," https://www.sohu.com/a/222391501_100017648, 2018.

[2] R. Soram and E. S. Meitei, "On the performance of rsa in virtual banking," in 2015 International Symposium on Advanced Computing and Commu-nication (ISACC). IEEE, 2015, pp. 352–359.

[3] A. Jarovsky, T. Milo, S. Novgorodov, and W.-C. Tan, "Rule sharing for fraud detection via adaptation," in 2018 IEEE 34th International Confer-ence on Data Engineering (ICDE). IEEE, 2018, pp. 125–136.

[4] Z. Zhang, X. Zhou, X. Zhang, L. Wang, and P. Wang, "A model based on convolutional neural network for online transaction fraud detection," Security and Communication Networks, vol. 2018, 2018.

[5] X. Zhou, Z. Zhang, L. Wang and P. Wang, "A model based on siamese neural network for online transaction fraud detection," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-7.

[6] S. Wang, C. Liu, X. Gao, H. Qu, W. Xu, "Session-based fraud detection in online e-commerce transactions using recurrent neural networks," Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2017: 241-252

[7] C. Liu, Y. Chan, S. H. Alam Kazmi, and H. Fu, "Financial fraud detection model: based on random forest," International journal of economics and finance, vol. 7, no. 7, 2015.

[8] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, C. Jiang, "Random forest for credit card fraud detection", the 15th IEEE ICNSC, Zhuhai, China, March 27-29, 2018,pp. 1–6.

[9] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," Data mining and knowledge discovery, vol. 18, no. 1, pp. 30–55, 2009.

[10] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural network," International Conference on Neural Information Processing. Springer, Cham, 2016: 483-490.

[11] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784-3797, Aug. 2018.

[12] E. Kim, J. Lee, H. Shin, H. Yang, S. Cho, S. Nam, Y. Song, J. Yoon, J. Kim, "Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning," Expert Systems with Applications, 2019, 128: 214-224.

[13] Y. Meng, Z. Zhang, W. Liu, L. Chen, Q. Liu, L. Yang, and P. Wang, "A novel method based on entity relationship for online transaction fraud detection," in Proceedings of the ACM Turing Celebration Conference-China. ACM, 2019, p. 121.

[14] X. Wang, H. Wu and Z. Yi, "Research on Bank Anti-Fraud Model Based on K-Means and Hidden Markov Model," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, 2018, pp. 780-784.

[15] P. Zhao, C. Yan, and C. Jiang, "Authenticating web user's identity through browsing sequences modeling," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). IEEE, 2016, pp. 335–342.

[16] H. Zhang, C. Yan, P. Zhao, and M. Wang, "Model construction and authentication algorithm of virtual keystroke dynamics for smart phone users," in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2016, pp. 000 171–000 175.

[17] L. Chen, Z. Zhang, Q. Liu, L. Yang, Y. Meng, and P. Wang, "A method for online transaction fraud detection based on individual behavior," in Proceedings of the ACM Turing Celebration Conference-China. ACM, 2019, p. 119.

[18] B. Ji, L. Hu, W. Han, and J. Yan, "Research on e-commerce-oriented user abnormal behaviour detection," Netinfo Security, no. 9, p. 19, 2014

[19] Y. Kültür and M. U. Çaˇglayan, "A novel cardholder behavior model for detecting credit card fraud," Intelligent Automation & Soft Computing, pp. 1–11, 2017.

[20] L. Zheng, G. Liu, W. Luan, Z. Li, Y. Zhang, C. Yan, and C. Jiang, "A new credit card fraud detecting method based on behavior certificate," in 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2018, pp. 1–6.

[21] L. Zheng, G. Liu, C. Yan, and C. Jiang, "Transaction fraud detection based on total order relation and behavior diversity," IEEE Transactions on Computational Social Systems, vol. 5, no. 3, pp. 796–806, 2018.

[22] Y. Xie, G. Liu, R. Cao, Z. Li, C. Yan, C. Jiang, "A feature extraction method for credit card fraud detection," the 2nd IEEE International Conference on Intelligent Autonomous Systems (ICoIAS'2019), Singapore, 2019, pp. 70-75.

[23] S. Nami and M. Shajari, "Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors," Expert Systems with Applications, vol. 110, pp. 381–392, 2018.

[24] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit card fraud detection: a novel approach using aggregation strategy and feedback mechanism," IEEE Internet of Things Journal, vol. 5, no. 5, pp. 3637–3647, 2018.

[25] C. Wand, J. Luo, B. Yang, C. Jiang, "On complementary effect of blended behavioral analysis for identity theft detection in mobile social networks," Communications in Computer and Information Science, 2018, 747:32-44.

[26] N. Soltani, M. K. Akbari and M. Sargolzaei Javan, "A new user-based model for credit card fraud detection based on artificial immune system," The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), Shiraz, Fars, 2012, pp. 029-033.

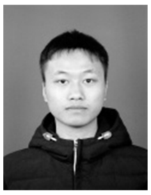
[27] F. Zhang, G. Liu, Z. Li, C. Yan, C. Jiang, "GMM-based undersampling and its application for credit card fraud detection," the 32nd International Joint Conference on Neural Network (IJCNN'2019), Budapest, Hungary, 2019, pp. 1-8

[28] H. Wu, G. Liu, "A hybrid model on learning cross features for transaction fraud detection," the 19th Industrial Conference on Data Mining (ICDM'2019), July 17-19, 2019, New York, USA.

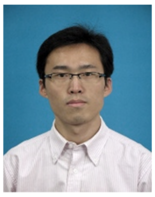
[29] I. Savvas, A. Chernov, M. Butakova, and C. Chaikalis, "Increasing the quality and performance of n-dimensional point anomaly detection in traffic using pca and dbscan," in 2018 26th Telecommunications Forum (TELFOR). IEEE, 2018, pp. 1–4.

Zhaohui Zhang received the B.S. degree in Computer Science from Anhui Normal University, Wuhu, China, in 1994, and became a teacher at the University. He completed his master's courses program of University of Science and Technology of China from 1999 to 2000. Respectively, he received the Ph.D. degree in Computer Science from Tongji University, Shanghai, China, in 2007. He was a professor with Anhui Normal University before 07/2015. Currently, he is a Professor with the School of Computer Science and Technology, Donghua University, Shanghai, China. His research interests include network information services, service computing and cloud computing.

Ligong Chen was born in 1995. He received the B.S degrees in Internet of Things Engineering from Anhui Normal University. He is an M.S. candidate of Donghua University. His research area includes Artificial intelligence, big data and financial transaction risk prevention and control.



Qiuwen Liu was born in 1995. He received the B.S degrees in Software Engineering from Heilongjiang University. He is an M.S. candidate of Donghua University. His research area includes big data management, distributed data stream processing and resource scheduling.



Pengwei Wang received the B.S. and M.S. degrees in Computer Science from Shandong University of Science and Technology, Qingdao, China, in 2005 and 2008, respectively, and the Ph.D. degree in Computer Science from Tongji University, Shanghai, China, in 2013. He finished his postdoctoral research work at the Department of Computer Science, University of Pisa, Italy, in 2015. Currently, he is an Associate Professor with the School of Computer Science and Technology, Donghua University, Shanghai, China. His research interests include cloud computing, service computing, and big data.