iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VII **Month of publication:** July 2024

DOI: https://doi.org/10.22214/ijraset.2024.63525

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VII July 2024- Available at www.ijraset.com

Network Based Feature Extraction Method for Fraud Detection Using Label Propagation

Ravula Muralidhar Reddy[1], N. Naveen Kumar[2]
[1]*M. tech (Computer Science), Student,* [2]*Associate Professor of CSE, Department of Information Technology, JNTUHUCESTH, Hyderabad, Telangana – 500085*

Abstract: ***Nowadays, judging the current transaction based on user history transactions is an important detection method. However, different users have different transaction behaviors, when all users use the same limit to judge whether the transaction is abnormal, it will result in higher misjudgment for some users. Aiming at the above problems, this paper proposes an individual behavior transaction detection method based on hypersphere model. In this model, considering multiple dimensions of normal historical transaction records, the characteristics of user's transaction behavior is generated with the trend of transaction. Then, the user optimal risk threshold algorithm is proposed to determine the optimal risk threshold for each user. Finally combining the transaction behavior and the optimal risk threshold, the user behavior benchmark is formed, which is used to construct the multidimensional hypersphere model. On this basis, a mapping method for transforming transaction detection into midpoint in multidimensional space is proposed. The experiment proves that the proposed method is superior to other models, and it is found that the characterization effect of user behavior is related to the frequency of users' transactions.***
Applied computing → Secure online transactions; Digital cash;
Computing methodologies → Instance-based learning; Rule learning.
Keywords: Online transaction detection, Individual behavior, Transaction behavior, Optimal risk threshold, Behavior benchmark, Hypersphere model

I. INTRODUCTION

The fraud detection in online lending establishes whether the user is a fraudster—someone who is unwilling to repay the loan. Fraud has become a major concern for consumer finance organizations worldwide due to quick development of consumer finance and the rise of fraudulent occurrences. Finding the unusual traits of fraudsters is the basic concept behind fraud detection, which has evolved from old techniques like blacklists and rules-based models to machine learning based on big data. Therefore, the crucial issue that ultimately determines the effectiveness of fraud detection is how to extract the traits from the available data that most accurately represent the fraudster. Currently, the most common approach to feature engineering is to extract certain intrinsic attributes from applicants as features. Examples of this include determining whether the applicant's credit information contains any past-due records and using the RFM technique to take details from the applicant's most recent transaction records, such as the quantity of consumption in the previous month or the number of transactions in the previous week. Moreover. Finding out if the registration name matches a specific pattern, the IP block being used at the time of application, whether it's a temporary IP is required. A person's likelihood of being a fraudster increases if they related to the other fraudsters.

With the rapid development of e-commerce, online payment has become more and more popular. The data shows that in 2017, the B2C market transaction volume accounted for 60.0% of the online shopping market in China, and the transaction scale reached 3.6 trillion yuan[6], and in the 2018 Double 11 Shopping Carnival, the final transaction volume of 24 hours reached an astonishing 213.5 billion yuan[18], however, the booming electronic trading market also provides opportunities for fraudsters, causing huge economic losses to users and institutions, disrupting the normal financial order and restricting the long-term healthy development of electronic transactions. According to the investigation and analysis of payment fraud cases by the payment control department, the main means of fraud crimes include hacking, stolen cards, credit card cashing, phishing websites, Trojan horses, etc. [14] How to effectively prevent the risk of online transaction fraud has become a problem to be solved.

Banks generally adopt a rule-based expert system as a method of fraud detection. Through anti-fraud experts analyze the behavior patterns of fraudsters in the case, find out the effective features, and write expert rules to identify fraudulent behaviors [17]. However, the recognition effect of this method is highly dependent on the artificial rules written by anti-fraud experts, at the same time, due to the excessive number of rules, there will be a certain degree of rule redundancy.

Most scholars generally use data mining methods to compensate for the lack of rule systems, such as neural networks, Markov Chains, Bayesian Networks, Decision Trees, Support Vector Machines, and Logistic Regression [5][2][15][16][13][1] and unsupervised methods such as Clustering algorithm and Self- organizing Mapping[12]. Some scholars start from the perspective of individual users, obtaining the transaction pattern of the user according to the historical data of a single user, and then matching the current record with the user transaction mode, thereby performing fraud detection, such as literature [4][11][23][24][9]. Despite the current efforts to resolve the transaction fraud problem, it still faces many difficulties:

1) Using machine learning and other related models, the training data needs to be marked. However, in reality, there is a case where the sample is extremely unbalanced, and it is difficult for the model to fully learn the characteristics of fraudulent transactions[7].

2) Occasionally, there may be an abnormal amount of money or an abnormal time in the normal transaction of the user, and the model can easily intercept such transactions, resulting in a higher false positive rate[23].

3) How to extract user behavior characteristics to construct user behavior benchmark, user transaction abnormality judgment standards still face many challenges.

4) The same limit is applied to all users to detect transactions, and the differences between users are not considered, resulting in higher misjudgment for some users.

Aiming at the above problems, this paper proposes a new transaction detection model based on individual behavior. Compared with other models, the proposed model can alleviate the above problems and has the following advantages:

- This paper from the perspective of individual users, using the user's normal transactions to establish behavior benchmark, can well avoid the problem of sample imbalance.
- In the establishment stage of the user behavior benchmark, this paper considers the user's various dimension information and considers the trend of user transactions, which can better describe the user behavior.
- This paper considers the difference between users, and proposes the optimal risk threshold division algorithm for users, and determines the optimal risk threshold for each user according to the transaction behavior of users.
- This paper builds a behavioral benchmark for each user, and proposes a hypersphere model based on behavioral benchmarks, transforming transaction detection into a mapping of points in multidimensional space.

As mentioned above, the main contributions of this paper are as follows. First, a more accurate individual behavior model is proposed, which considers user transaction behavior from multiple dimensions such as transaction amount, transaction time, transaction frequency, transaction IP and amount change trend. Secondly, considering the difference between different users, the user optimal risk threshold algorithm is proposed based on the user transaction behavior, and the user transaction behavior and the optimal risk threshold are combined into a user behavior benchmark. Based on the user behavior benchmark, a multidimensional hypersphere model is proposed. Third, it is found through experiments that the characterization of user behavior is related to the frequency of user transactions.

The rest of the paper is organized as follows, the second section details the related work, the third section discusses the model approach presented in this paper, the fourth section introduces the data source and the experimental results of this paper, and the fifth section summarizes the research results of the paper and the planning of future research work.

II. RELATED WORK

Fraud detection is a classification problem, so based on group users, using machine learning and other technologies to achieve fraud detection by learning pre-marked transaction data has been widely studied in recent years. Kolalikhormuji et al. [10] propose cascade artificial neural networks based on existing neural networks. AC Bahnsen et al.[3] propose a cost-sensitive method based on Bayesian minimum risk. Zareapoor et al.[20] use integrated learning techniques to build classifiers based on existing machine learning, and introduce decision-making mechanisms for classifier integration evaluation. Xuan et al.[19] use random forests to train the characteristics of normal and abnormal behaviors and perform well in credit fraud detection. Zhang Z et al.[21] can use the convolutional neural network to derive the characteristics of the feature, by inputting the original features and adding the feature arrangement layer to combine the inputs, the transaction is detected and a good result is obtained. J.Cui et al.[22] propose an agile sensing method, which includes an agile perception model of system anomaly and a Petri net model for repeated behavior detection, this method can effectively perceive impending system anomalies and locate them before they occur. However, in reality, high-frequency users have a large number of transactions, which makes it difficult for the model to learn the transaction characteristics of other users. Therefore, the misjudgment of some users is more serious.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

In recent years, based on a single user, anomaly detection based on user history data has gradually gained attention. User behavior is more mature in the user portrait field, including intelligent marketing and personalized recommendation. By analyzing the basic information, social characteristics and transaction characteristics of the group users, the user is tagged and classified according to the tags to which the user belongs, and to achieve advertising recommendations, precision marketing, etc.[8] But the user portrait is actually a "typical" user obtained by refining the attribute characteristics, a virtual representation of the user's real data, a collection of common features of a user group with similar behaviors, and a conceptual model of a user group with some distinctive features. However, the financial fraud detection based on individual users pay more attention to the users themselves. From the perspective of users, the user behavior model is constructed by analyzing the user's transaction behavior patterns, and then the model is used to detect the transactions of the users. Ji Bingshuai et al.[4] propose a method for e-commerce user abnormal behavior detection research, collecting user historical behavior data, using data mining algorithm to establish the user's normal behavior pattern, and judging whether the user's transaction behavior is abnormal. Yigit Kultur et al.[11] propose a new cardholder behavior model for credit card fraud detection, focusing on the cardholder's transaction behavior and detecting abnormal transactions through user's historical transaction behavior. Zheng et al.[23] propose a new behavior certificate-based credit card fraud detection system that use user behavior certificates to identify user transactions. J.Zhong et al.[24] propose a method based on browsing behavior authentication, which constructs a user browsing behavior model from the Web usage log, the model identifies the true identity of the user in the visited web page. C.Jiang et al.[9] propose a new method using aggregation strategy and feedback mechanism. Firstly, all cardholders are divided into different groups through aggregation strategy, and a series of specific behavior patterns are extracted for each group of cardholders. Finally, a classifier set is used to detect fraud online.

In the above work, the literature[11] only starts from the perspective of the user transaction amount, if the amount is very different from usual, it will be regarded as an abnormality, but the user behavior cannot be fully characterized only by the amount. Although the literature [4] and [23] portray user behavior from multiple angles, it does not consider the independence between users when judging the normal transaction and abnormal transaction of the user, so the misjudgment of some users is more serious. Although the literature [24] establishes a browsing behavior model for each user, it pays more attention to the distinction between users rather than the judgment of user behavior. Although the literature [9] pays attention to the judgment of the transaction, it is to group similar users and establish a model for each group of users to judge, and does not fully consider each user. At the same time, due to the lack of real transaction data, some of the work is carried out on the simulated data, which is deviated from the actual situation, and the applicability needs to be evaluated.

III. MODEL METHOD

This section introduces a new model for transaction detection based on individual behavior. As shown in Figure 1, the first part is the user transaction behavior generation, the second part is the determination of the user's optimal risk threshold, and the third part is the transaction detection. The user transaction behavior generating part generates a transaction behavior for the user from multiple dimensions according to each user normal transaction. The optimal risk threshold determination section determines an optimal risk threshold for the user based on the transaction behavior and the transaction record of the user. The fraud detection part constructs a multidimensional hypersphere model based on the user behavior benchmark, and according to the model, a detection algorithm is proposed to judge the user's transaction record.

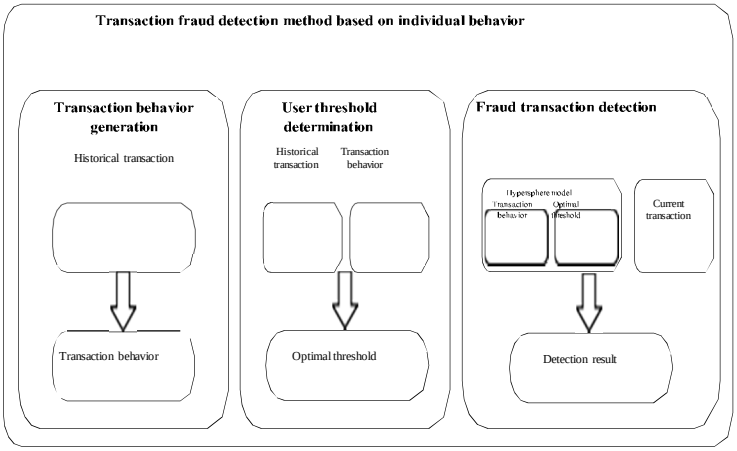


Figure 1: Method architecture

A. Transaction Behavior Generation

This section describes the generation process of user behavior benchmark, which are generated from user history transaction data.

1) *Definition 1(Transaction Record):* A record r contains m attributes, recorded as $r = \{a_1, \cdots, a_m | a_1 \in A_1, \cdots, a_m \in A_m\}$, the transaction attributes in this paper include transaction user number, transaction time, transaction amount and transaction *IP*, that is $r = \{a_{no}, a_{time}, a_{amount}, a_{ip}\}$. User u's transaction log is the user's historical transaction set, write as $L_u = \{r^u, r^u, \cdots, r^u\}$, and $n = |L_u|$, as the number of transaction records of the user, define each attribute set in the user transaction log as $A^u = \{a, a^i, \cdots, a^i u\}$,where $A^u \subseteq A_i$, $n^u = |A^u|$. The normal transaction $T_u = \{t \in L_u | label = T\}$ is extracted in the user transaction log L, where $n_{tu} = |T_u|$.

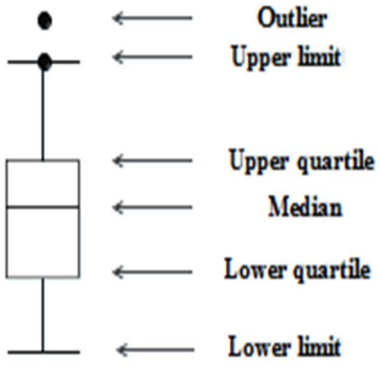


Figure 2: Method architecture

2) *Definition 2(Transaction Amount Attribute):* User u's transaction amount attribute is defined as $TAR^u = (range_1, range_2, \cdots, range_n)$. Representing the probability that the user u's transaction amount is within each interval, reflecting the user u's consumption habits in the transaction amount dimension. The calculation rule of $range_i$ is as follows. A^u is taken in the normal transaction log T_u of the user u, where $n^u_{amount} = |A^u_{amount}|$ as the number of transactions for this user.

For set A^u_{amount}, this paper uses the box plot method to find the value of $range_i$. As shown in the Figure 2, The biggest advantage of the box plot is that it is not affected by the outliers and can accurately depict the distribution of the data. For the set A^u, the aggregate transaction amount is arranged from small to large, and then the lower quartile *Q1*, the median *Q2*, the upper quartile *Q3*, the upper limit *Qmax*, the lower limit *Qmin* of the set are obtained, and the following five subsets are obtained.

$$TAR_1 = \{a \in A^u_{amount} | Q_{min} \le a < Q_1\}$$
$$TAR_2 = \{a \in A^u | Q \le a < Q\}$$
$$TAR_3 = \{a \in A^u_{amount} | Q \le a < Q\}$$
$$TAR_4 = \{a \in A^u_{amount} | Q_3 \le a \le Q_{max}\}$$
$$TAR_5 = \{a \in A^u_{amount} | a \le Q_{min}, a \ge Q_{max}\}$$

Therefore, $range_i$ can be gotten by (1), and then the user transaction amount attribute $TAR^u = (rang_1, \cdots, range_n)$ is obtained.

$$range_i = |TAR_i| / n^u_{amount} \quad (1)$$

3) *Definition 3(Amount Change Attribute):* The transaction amount change attribute of user u is defined as $TAC^u = (trend_1, \cdots, trend_n)$, which represents the probability that the user transaction amount increases in each interval, and reflects the trend of the user transaction amount over time. Calculating the set A^u_{trend} of the amount of change between each two transactions according to the set A^u_{amount}, and $n^u_{trend} = |A^u_{trend}|$, the calculation rule for each element a^{trend} is as shown in (2).

$$a_i^{trend} = \begin{cases} 0 & i = 0 \\ \dfrac{a_i^{amount} - a_{i-1}^{amount}}{a_{i-1}^{amount}} & i > 1 \end{cases} \quad (2)$$

After obtaining the lower quartile *Q1*, the median *Q2*, the upper quartile *Q3*, the upper limit *Qmax*, and the lower limit *Qmin* of the set A_{trend}^{u}, the following five subsets are obtained.

$$TAC_1 = \{a \in A_{trend}^{u} \mid Q_{min} \le a < Q_1\}$$
$$TAC_2 = \{a \in A_{trend}^{u} \mid Q_1 \le a < Q_2\}$$
$$TAC_3 = \{a \in A_{trend}^{u} \mid Q_2 \le a < Q_3\}$$
$$TAC_4 = \{a \in A_{trend}^{u} \mid Q_3 \le a \le Q_{max}\}$$
$$TAC_5 = \{a \in A_{trend}^{u} \mid a \le Q_{min}, a \ge Q_{max}\}$$

Therefore, $trend_i$ can be gotten by formula similar to (1), and then the user transaction change attribute $TAC^u = (trend_1, \cdots, trend_5)$ is obtained.

4) *Definition 4(Transaction Workday Attribute):* The user transaction workday attribute is defined as $TIW^u = (isworkday, noworkday)$, which represents the probability that the user transaction occurs on workdays and non-workdays, non- workdays include weekends and holidays, reflecting the user *u*'s spending habits in the transaction date dimension. A_{time}^{u} is taken from the normal transaction log T_u of the user *u*, where $n^u = |A_{time}^{u}|$ as a set of transaction time attributes of the user. For each element, we determine whether it belongs to the working day, label each element with 0 and 1 respectively, and then get the following two subsets.

$$TIW_1 = \{a \in A_{time}^{u} \mid label=0\}$$
$$TIW_2 = \{a \in A_{time}^{u} \mid label=1\}$$

Therefore, *isworkday* and *noworkday* can be gotten by formula similar to (1), and then the user transaction workday attribute $TIW^u = (isworkday, noworkday)$ is obtained.

5) *Definition 5(Transaction Time Attribute):* The user transaction time attribute is defined as $TTR = (time_1, time_2, \cdots, time_n)$, where $\sum time_i = 24$, which represents the probability of the user *u*'s transaction time in each interval, reflecting the user *u*'s transaction habits in the transaction time dimension. According to the set A^u, the time period is divided into working time and non- working time, thereby obtaining the following two subsets.

$$TTR_1 = \{a \in A_{time}^{u} \mid a \ in \ worktime\}$$
$$TTR_2 = \{a \in A_{time}^{u} \mid a \ not \ in \ worktime\}$$

Therefore, $time_i$ can be gotten by formula similar to (1), and then the user transaction time attribute $TTR^u = (time_1, time_2)$ is obtained.

6) *Definition 6(Transaction Frequency Attribute):* The user transaction frequency attribute is defined as $TFA^u = (intarval_1, \cdots, intarval_n)$, and the transaction frequency is defined as the user transaction volume per unit time, this paper uses the transaction time interval to reflect the user *u*'s transaction frequency. The smaller the transaction interval, the more transaction volume the user has in a unit of time, and the higher the frequency of user transactions. Therefore, this attribute represents the probability of the user transaction time interval in each interval According to the set A^u, the transaction time interval set $A_{interval}$ is obtained, where $n_{interval} = |A_{interval}|$, and the calculation rule of each element $a^{interval}$ is as shown in the formula (3).

$$interval_i = \begin{cases} t_1^{time} & i = 1 \\ t_i^{time} - t_{i-1}^{time} & i > 1 \end{cases} \quad (3)$$

After obtaining the lower quartile *Q1*, the median *Q2*, the upper quartile *Q3*, the upper limit *Qmax*, and the lower limit *Qmin* of the set interval

$$TFA_1 = \{a \in A^u_{interval} \mid Q_{min} \le a < Q_1\}$$
$$TFA_2 = \{a \in A^u_{interval} \mid Q_1 \le a < Q_2\}$$
$$TFA_3 = \{a \in A^u_{interval} \mid Q_2 \le a < Q_3\}$$
$$TFA_4 = \{a \in A^u_{interval} \mid Q_3 \le a \le Q_{max}\}$$
$$TFA_5 = \{a \in A^u_{interval} \mid a \le Q_{min}, a \ge Q_{max}\}$$

Therefore, $interval_i$ can be gotten by formula similar to (1), and then the user transaction frequency attribute $TFA^u = (interval_1, \cdots, interval_5)$ is obtained.

7) *Definition 7(Transaction IP Attribute):* The user transaction address represented by the user transaction *IP* which reflects the user *u*'s transaction habits in the transaction location dimension, defined as $TIP^u = (iscommonip, nocommonip)$, which represents whether the user transaction is a common *IP*. A^u is taken from the normal transaction log T_u of the user *u*, where $n^u = |A^u|$, as the set of *IP* attributes of the user transaction, *IPNum* represents the number of *IPs* in the historical transaction of the user *u*. In the set A, the probability P of the transaction of each *IP* of the user is obtained, if $P_{ipi} > 1/IPNum^{IP}$, the *IP* is regarded as the common *IP* of the user, and then the user *u*'s common *IP* address library is established. According to the user *u*'s common *IP* address library, the elements in A^u are marked with 0, 1 label, and 0 represents non-common *IP*, 1 represents the common *IP*, according to which the following two subsets will be obtained.

$$TIP_1 = \{a \in A^u_{IP} \mid a = 1\}$$
$$TIP_2 = \{a \in A^u_{IP} \mid a = 0\}$$

Therefore, *iscomonip* and *nocommonip* can be gotten by formula similar to (1), and then the user transaction *IP* attribute TIP_u is obtained.

8) *Definition 8(Previous Transaction Status):* Considering fraudulent methods, such as hacking, stolen cards, phishing websites, Trojan viruses, etc. After the fraudsters steal the user identity information, they transfer the user funds to different places, most of the fraudulent transactions often occur continuously, therefore, considering the user *u*'s previous transaction status. The user *u*'s previous transaction status is defined as $PTS^u = (T, F)$,which represents whether the transaction status of the user before the current transaction is normal, so that the following two subsets can be obtained.

$$PTS_1 = \{R \in T_u \mid prestate = T\}$$
$$PTS_2 = \{R \in T_u \mid prestate = F\}$$

Therefore, T and F can be gotten by formula similar to (1), and then the user previous transaction status PTS^u is obtained.

9) *Definition 9(User Behavior Benchmark):* According to the user *u*'s normal transaction T_u, the user behavior benchmark $UBB_u = [TB, Threshold^u]$ is obtained, TB is the transaction behavior, and $Threshold^u$ is the optimal risk threshold,

$TB_u = (TAR^u, TAC^u, TIW^u, TTR^u, TFA^u, TIP^u, PTS^u)$.

TAR^u represents the user transaction amount attribute.

TAC^u represents the user transaction change attribute.

TIW^urepresents the user transaction workdays attribute.

TTR^u represents the user transaction time attribute.

TFA^u represents the user transaction frequency attribute.

TIP^u represents the user transaction *IP* attribute.

PTS^u represents the user previous transaction status.

B. User Optimal Risk Threshold Algorithm

Each user's transaction behavior is different from others. Considering the differences between users, each user has its own abnormal threshold, instead of all users sharing the same risk threshold.

$$TT = \frac{count(rellable = 0)}{count(prelable = 0)} \quad (4)$$

$$TF = \frac{count(rellable = 1)}{count(prelable = 1)} \quad (5)$$

$$SE = \alpha * TT + (1-\alpha) * TF \quad (6)$$

In the above formula, *TT* represents the proportion of true normal transactions that are judged to be normal transactions. *TF* represents the proportion of real abnormal transactions that are judged to be abnormal transactions. SE is the weighted sum of TT and TF , which represents the judgment effect, the larger the SE value is, the better the judgment effect is. The α represents the weight, if you pay more attention to the interception of abnormal transactions and ignore the misjudgment of normal transactions, the α is smaller, if you pay more attention to the misjudgment of normal transactions and ignore the interception of abnormal transactions, the α is larger.

$$d(\overrightarrow{TB^u}, \overrightarrow{UTV^u}) = \sqrt{\sum_{i=1}^{n} (tb_i - utv_i)^2} \quad (7)$$

The user transaction behavior TB_u is regarded as a n- dimensional vector $\vec{T}\vec{B}\vec{u}$. For each transaction, it is matched with the user transaction behavior to obtain a n-dimensional user transaction vector $\vec{U}\vec{T}\vec{V}\vec{u}$, then uses (7) to calculate the distance between the user *u*'s historical transaction and the transaction behavior, and gets a distance set $D^u = \{dist^u, dist^u_1, \cdots^2, dist^u\}^{n^u}$, and the maximum value $max_D u$ and the minimum value $min_D u$ are gotten in the set D^u, and then the optimal risk threshold of the user is obtained by the Algorithm 1

Algorithm 1: User Optimal Risk Threshold Algorithm

Input: User *u*'s historical behavior benchmark TB_u and user *u*'s history transaction record $L_u = \{r^u, r^u, \cdots, r^u\}$.1 2 r

Output: The best threshold for this user.

1. calculate D_u , max_D , min_D
2. bestSE := 0;
3. bestThreshold := min_D
4. for ($j = min_D; j \le max_D; j = j + 0.01$) do
5. for ($i = 1 ; i \le n^u ; i$++) do prelable = 1 if $dist_i \ge j$ else 0
6. end for
7. calculate *TT* , *TF* and *SE*
8. bestSE = *SE* if *SE* > bestSE else bestSE
9. bestThreshold = j if SE > bestThreshold else bestThreshold

10. end for
11. $Threshold^u$= bestThreshold
12. return $Threshold^u$;

C. Online Fraud Detection

Through the above method, the transaction behavior TB_u of the user and the user *u*'s optimal risk threshold $Threshold^u$ is obtained, then user *u*'s transaction behavior benchmark UBB_u = $[TB_u, Threshold^u]$ can be constructed. Considering the *u*'s transaction behavior TB_u as a point in multidimensional space and the *u*'s optimal risk threshold $Threshold^u$ as a radius, using the sphere expression in three-dimensional space, a hypersphere model of multidimensional space based on user behavior benchmark UBB_u is construct, the model expression is as follow.

$$f(x) = \sum_{i=1}^{n} (x_i - tb_i) - (Threshold) \quad (8)$$

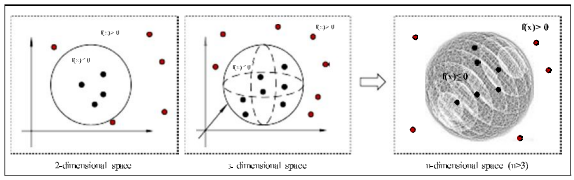


Figure 3: Schematic diagram of the Hypersphere Model

According to the user *u*'s hypersphere model *f(x)*, the multidimensional space is divided into two parts, as shown in Figure 3. The space of the $f(x) \leq 0$ is regarded as the user *u*'s normal transaction behavior space, and the $f(x) > 0$ is regarded as the user abnormal behavior space, and each transaction of the user is regarded as a point in the multidimensional space, therefore, the user transaction detection can be transformed into the mapping problem of the midpoint in the multidimensional space. Each transaction is used as the input of the user behavior benchmark hypersphere model, and the value of *f(x)* is calculated, if $f(x) > 0$, the transaction is intercepted. The algorithm details are show in Algorithm 2.

Algorithm 2: Online Transaction Fraud Detection

Input: User *u*'s historical behavior benchmark UBB_u = $[TB_u, Threshold^u]$ and user *u*'s current transaction r_u = $\{a_{no}, a_{time}, a_{amount}, a_{ip}\}$.

Output: The outcome of the transaction.

1. UBB_u = $[TB_u, Threshold^u]$;
2. UTV_u = $(tar^u, tac^u, tiw^u, ttr^u, tfa^u, tip^u, pts^u)$ = 0;
3. for every element in r_u do
4. tar^u : $range_i$ = 1 if a_{amount} in range *i* else 0;
5. tac^u : $trend_i$ = 1 if amount trend in range *i* else 0;
6. tiw^u : *isworkday* = 1 if workday else *noworkday* = 1;
7. ttr^u : $time_i$ = 1 if a_{time} in range i else 0;
8. tfa^u : $interval_i$ = 1 if interval in range *i* else 0;
9. tip^u :*iscommonip*=1 if *commonip* else *nocommonip* =1;
10. pts^u :*T* = 1 if trans pre_state is True else *F* = 1;

11. end for
12. use (8) calculate $f(x)$;
13. return Fraud if $f(x) > 0$ else return Normal;

IV. PERFORMANCE EVALUATION

In this section, we verify the effect of the model by experiment. First, we introduce the training set and test set used in this experiment, and then explain the effect of the experiment and the comparison with other models.

A. Data Set

The current research data set comes from a domestic bank data, which covers 3 months (April to June) transactions; contains 92,133 users, 35020,48 transaction records, each transaction record is marked by the bank for the label, the white sample data accounted for 98.14% of the transaction data set, the black sample data accounted for 1.86% of the transaction data set.

In Figure 4, all transaction data sets are referred to as data set *A*, and all fraudulent transactions are in data set *F*, which contains 147,829 transaction data for 14,751 users. The data set shows that the main types of fraud include phishing sites, Trojan viruses, etc. Transactions involving 14,751 users are in data set *B*, all transactions for fraud-free trading users are in data set *C*, the transaction data of any number of users is randomly extracted from the data set *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: All transaction data
- B: All transactions with fraudulent trading customers
- C: All transactions without fraudulent trading customers
- F: All fraudulent transactions

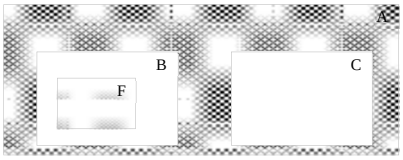


Figure 4: Data set distribution

B. Comparative Performance

In this section we compare our model (shown below with *OM*) with another model based on user behavior certificate [23] (shown below with *UBC*). We randomly select 8 sets of data in data set *B* according to the historical transaction volume of users in the same time period, and 8 sets of experiments are performed. The data information is shown in Table 1.

Table 1: Experimental data details

NO	Count_user	C_F	C_T	C_train	C_test
Ⅰ	13	1949	9818	8828	2940
Ⅱ	14	971	3338	3354	825
Ⅲ	11	921	1798	2346	373
Ⅳ	12	725	979	1557	147
Ⅴ	15	848	530	945	433
Ⅵ	12	341	313	485	169
Ⅶ	12	75	261	301	65
Ⅷ	12	46	169	168	47

The average transaction volume per user in group Ⅰ is more than 500, the average transaction volume of each user in group Ⅱ is between 300 and 500, the average transaction volume per user in group Ⅲ is between 200 and 300, the average transaction volume of each user in group Ⅳ is between 100 and 200, the average transaction volume of each user in group Ⅴ is between 50 and 100, the average transaction volume per user in group Ⅵ is between 30 and 50, the average transaction volume per user in group Ⅶ is between 20 and 30, the average transaction volume per user in group Ⅷ is less than 20. For each user in the data set, we train separately to obtain the user's transaction behavior and the optimal risk threshold, which constitutes the behavior benchmark of the user, then use the hypersphere model to detect the transaction in June and verify the effect of the model.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Table 2: Confusion matrix

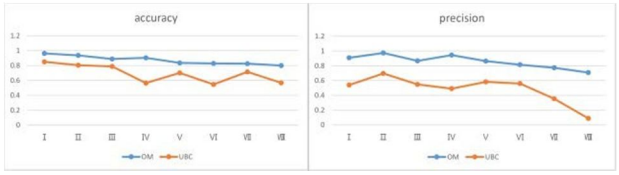
	True Fraud	True Normal
Pre Fraud	TP	FP
Pre Normal	FN	TN

As shown in Table 2, 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.

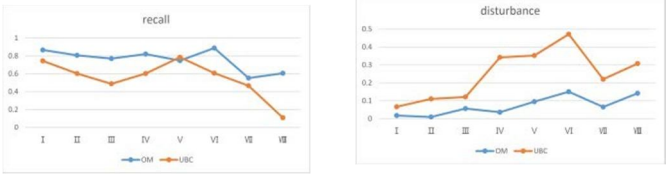
Table 3: Indicator calculation method

Indicator name	calculation method
Accuracy	*(TP+TN)/(TP+TN+FP+FN)*
Precision	*TP/(TP +FP)*
Recall	*TP/(TP +FN)*
Disturbance	*FP/(TN+FP)*
F1 value	*2*Precision*Recall/(Precision+Recall)*

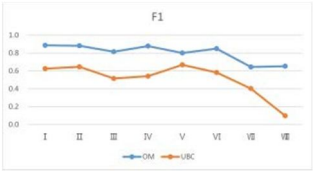
In order to make the comparison results more convincing, we use several indicators commonly used in fraud detection as the evaluation indicators, including accuracy, precision, recall, disturbance and F1 value, as shown in Table 3. Accuracy indicates that the model determines the correct number of transactions as a percentage of the total number of transactions. The precision rate is the ratio of the model's judgment to true fraudulent transactions in fraudulent transactions. The recall rate is the percentage of the model that detected the number of real fraudulent transactions as a percentage of all fraudulent transactions. The disturbance is the ratio of the model misjudgment a normal transaction as an abnormal transaction to all normal transactions. The F1 value is the harmonic mean of the accuracy rate and the recall rate.



(a) accuracy (b) precision



(c) recall (d) disturbance



(e) F1

Figure 5: Comparative Results

The comparison results are shown in Figure 5, the results can be seen under the same data set, the indicators of *OM* are higher than *UBC*. The accuracy rate, which is shown in Figure 5(a), is on average 10% higher than *UBC*, indicating that *OM* can more accurately determine the user's normal transactions and fraudulent transactions. The precision rate, which is shown in Figure 5(b), is on average 10% higher than *UBC*, indicating that *OM* is better than *UBC* in intercepting fraudulent transactions. The recall rate and disturbance rate, which are shown in Figure 5(c) and Figure 5(d), although there are fluctuations, the *OM* recall rate is still about 15% higher than *UBC*, and the disturbance rate is much lower than *UBC*, indicating that the model *OM* is accurately intercepting fraud, at the same time, the misjudgment of normal transactions is very rare. The F1 value, which is shown in Figure 5(e), which represents the overall performance of the model, it can be seen from the figure that *OM* is more than 20% higher than *UBC* under all data sets, and the overall performance of the model *OM* is better than *UBC*.

At the same time, it can be seen from the 8 sets of experiments that the overall performance of the model fluctuates with the gradual decrease of the user's historical transaction volume in the same time period, but the model *OM* can maintain better performance than *UBC*. When the user data volume is between 200 and 300, the overall performance of the model *UBC* shows a downward trend, the accuracy and precision rate shows a sharp decline, and the disturbance rate also increases rapidly, however, the model *OM* performs even better. When the number of user historical transaction records is above 100, the overall performance of model *OM* is excellent, and the accuracy and precision are above 90%. The recall rate and F1 value are also higher than other data sets, and the disturbance rate is less than 5%. When the user history transaction records between 30 and 100, the indicators are relatively good, and the accuracy, precision, recall and F1 values are all above 80%. When the number of user history transactions is less than 30, the performance of the model *OM* shows a downward trend, and all indicators have declined. At the same time, it can be found that the characterization effect of the user behavior is related to the frequency of the user's transaction. The more the user's transaction volume in the same time period, the better the model effect, that is, the better the characterization effect of the user behavior.

Through the analysis of the experimental results, it can be seen that *OM* has better overall performance than *UBC*, the main reasons are as follows. First, the data set used in this paper is real data, and the data set used by *UBC* is simulation data, and the data simulation is idealized, which is not completely consistent with the actual situation. Second, the characteristics derived from the establishment of the user behavior benchmark are more comprehensive, including transaction time, transaction frequency, amount, amount change, transaction location, whether it is a working day and the last transaction status, which more fully represent a person's transaction behavior. The third is to use the box plot method when dealing with transaction frequency and amounts, considering the case of outliers, it is more able to describe the data distribution characteristics. The fourth is to propose a user's optimal risk threshold algorithm based on the difference among users, and construct a hypersphere model based on user transaction behavior and user optimal risk threshold. Therefore *OM* has better overall performance than *UBC*.

V. CONCLUSION

In this paper, a fraud detection model for online transaction based on individual is proposed. Compared with other works, this paper considers the amount, time, location of transaction, as well as more detailed information, such as transaction frequency, transaction trend, states of previous transaction, and whether the transaction occurs during the workday, which can describe a user's transaction behavior more comprehensively. Furthermore, for considering the difference between user's transaction, we design a user optimal risk threshold algorithm that avoids misjudgment on users. Combined with above methods, user's behavior benchmark is modeled as a hypersphere model, which transforms fraud detection action into the mapping relationship of points in multidimensional space. Experiments have shown that our method is more accurate than other models and maintains a very low interference rate. In the future, we will focus on the relationship between the accuracy of user behavior and the area of transaction data (transaction frequency and transaction dimension). When we find this relationship, we will pay more attention to the behavior of low- frequency users.

REFERENCES

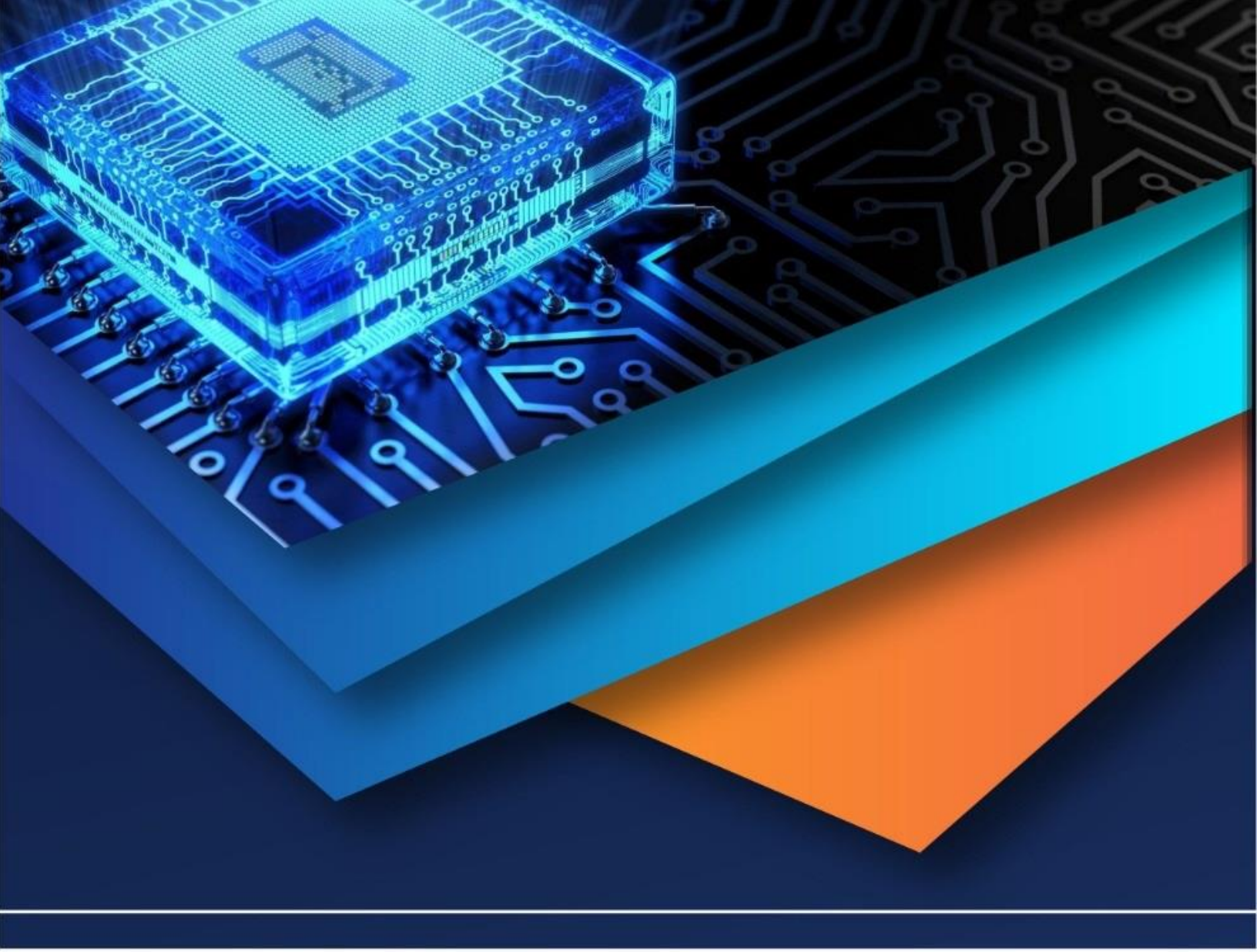
[1] SamanehSorournejad , Zahra Zojaji, Reza Ebrahimi Atani, and Amir Hassan Monadjemi. 2016. A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective. (11 2016).

[2] Rong Chang Chen, Shu Ting Luo, Liang Xun, and V. C. S. Lee. 2005. Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud. In 2005 International Conference on Neural Networks and Brain, Vol. 2. 810–815.

[3] A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten. 2013. Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk. In 2013 12th International Conference on Machine Learning and Applications, Vol. 1. 333– 338.

[4] J. I. Bing-Shuai, L. I. Hu, Wei Hong Han, and Yan Jia. 2014. Research on E- commerce-oriented User Abnormal Behaviour Detection. Netinfo Security (2014).

[5] R. Brause, T. Langsdorf, and M. Hepp. 1999. Neural data mining for credit card fraud detection. In Proceedings 11th International Conference on Tools with Artificial Intelligence. 103–106.
[6] Chyxx 2018. Forecast of the market size of China's online shopping industry in 2018. Chyxx. http://www.chyxx.com/industry/201803/614936.html.
[7] Andrea Dal Pozzolo, Olivier Caelen, Yann Aël Le Borgne, Serge Waterschoot, and Gianluca Bontempi. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications 41 (08 2014), 4915–4928.
[8] Liu Haiou. 2018. Literature Review of Persona at Home and Abroad. Information Studies:Theory Application (2018).
[9] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan. 2018. Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism. IEEE Internet of Things Journal 5, 5 (Oct 2018), 3637–3647.
[10] Morteza KolaliKhormuji, Mehrnoosh Bazrafkan, Maryam Sharifian, Seyed Mirabedini, and Ali Harounabadi. 2014. Credit Card Fraud Detection with a Cascade Artificial Neural Network and Imperialist Competitive Algorithm. International Journal of Computer Applications 96 (06 2014), 1–9.
[11] Yigit Kultur and Mehmet Ufuk Caglayan. 2015. A novel cardholder behavior model for detecting credit card fraud. In 2015 9th International Conference on Application of Information and Communication Technologies (AICT). 148–152.
[12] Dominik Olszewski. 2014. Fraud detection using self-organizing map visualizing the user profiles. Knowledge-Based Systems 70 (11 2014), 324–334.
[13] A. Shen, R. Tong, and Y. Deng. 2007. Application of Classification Models on Credit Card Fraud Detection. In 2007 International Conference on Service Systems and Service Management. 1–4.
[14] Souhu 2018. Research report on the trend of Network Fraud in 2017. Souhu. https://www.sohu.com/a/222391501_100017648.
[15] Dheepa V and Dhanapal R. 2012. Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft Computing 02 (07 2012), 391–397.
[16] S. Wang. 2010. A Comprehensive Survey of Data Mining-Based Accounting Fraud Detection Research. In 2010 International Conference on Intelligent Computation Technology and Automation, Vol. 1. 50–53.
[17] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams. 2009. Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery 18, 1 (01 Feb 2009), 30–55.
[18] Ws 2018. big data observation report on double 11 in 2018. Ws. http://www.100ec. cn/detail--6481169.html.
[19] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang. 2018. Random forest for credit card fraud detection. In 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). 1–6.
[20] Masoumeh Zareapoor and Pourya Shamsolmoali. 2015. Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science 48 (12 2015), 679–686.
[21] Zhaohui Zhang, Xinxin Zhou, Xiaobo Zhang, Lizhi Wang, and Pengwei Wang. 2018. A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection. Security and Communication Networks 2018 (08 2018), 1–9.
[22] Z.-H Zhang and J Cui. 2017. An Agile Perception Method for Behavior Abnormality in Large-Scale Network Service Systems. Jisuanji Xuebao/Chinese Journal of Computers 40 (02 2017), 505–519.
[23] L. Zheng, G. Liu, W. Luan, Z. Li, Y. Zhang, C. Yan, and C. Jiang. 2018. A new credit card fraud detecting method based on behavior certificate. In 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). 1– 6.
[24] J. Zhong, C. Yan, W. Yu, P. Zhao, and M. Wang. 2014. A Kind of Identity Au- thentication Method Based on Browsing Behaviors. In 2014 Seventh International Symposium on Computational Intelligence and Design, Vol. 2. 279–284.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)