2018 3rd IEEE International Conference on Image, Vision and Computing

Research on Bank Anti-fraud Model Based on K-means and Hidden Markov Model

Xiaoguo Wang, Hao Wu, Zhichao Yi
College of Electronics & Information Engineering
Tongji University
Shanghai, China
e-mail: haowu@tongji.edu.cn

Abstract—**Internet finance is developing rapidly. As online payments such as Alipay and WeChat Pay become more and more popular, cases of fraud associated with are also rising. In this paper, we describe the entire process of fraud detection using Hidden Markov model (HMM). We use the k-means algorithm to symbolize the transaction amount and frequency sequence of a bank account. This sequence is used to build and test the model. An HMM is initially trained with the normal behavior of an account. If an incoming credit card transaction is not accepted by the trained HMM with sufficiently high probability, it is considered to be fraudulent. We illustrate the feasibility of the model through simulation experiments and verify the validity of the model with real-world bank transaction data. Especially, in the case of enough historical transactions, this method performs well for low, medium frequency and amount of user groups.**

Keywords-fraud detection; k-means; hidden Markov model

I. INTRODUCTION

With the rapid development of Internet finance, online transactions using various payment tools have been closely connected with people's daily activities. While the quick payment method brings convenience to people's lives, it is accompanied by the occurrence of internet trading fraud. The concealed and diverse fraud means make consumers hard to detect and bring losses to both consumers and banks. In this situation, how to build the intelligent anti-fraud models and identify suspicious transactions by using the technology has become an urgent demand for the banking industry [1]. This paper combines the K-means clustering algorithm with the Hidden Markov Model (HMM) to study the bank anti-fraud model.

II. HMM AND K-MEANS IN ANTI-FRAUD MODEL

A. Use of HMM for Fraud Detection

The Hidden Markov Model (HMM) is a statistical model developed on the basis of the Markov chain to describe a Markov process with implicit unknown parameters. It is a double stochastic process in which a random process is visible and the other process is invisible [2]. For example, in the user's transaction process, the amount of each transaction is visible from the bank's view, while the types of behaviors such as the user's transfer transaction and purchase of goods are invisible. The transaction amount is ever-changing, but the type of transaction behavior is relatively fixed. Its purpose is to use the parameters in the visible random process to determine the implicit parameters of the other process. And then use these parameters for further analysis. It is widely used in speech recognition, text recognition, biological information science, fault diagnosis, anomaly detection and other fields [3].

Most people will have a relatively stable trading behavior for a period of time, either buying different types of goods, or performing transfer transactions. So a sequence of transactions is formed along with the order of time. There are two forms of this kind of sequence, one is the transaction amount sequence that can be directly observed in the bank database, and the other is a transaction behavior sequence that implies the user's trading habits, as Fig. 1

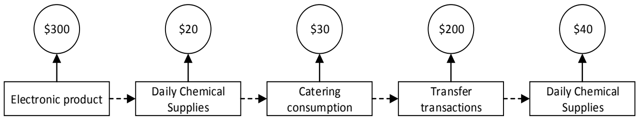

Figure 1. HMM in transactions

In [4], the author regards the consumption amount as the observed state of the HMM and the product type as the hidden state of the HMM. Since the cost of each purchase is various, it is very inconvenient to directly use the amount of consumption to train the model. So we quantify the amount into three price ranges, which are low, medium, and high, respectively, and the symbol is denoted by $\{l,m,h\}$. In this paper, in addition to using the quantified amount as the observation status, we have also added a frequency statistic that counts the number of transactions within a fixed period of time before and after each transaction. Similarly, the frequency is also divided into three intervals corresponding to the low frequency, intermediate frequency, and high frequency, and the symbol is denoted by $\{l,m,h\}$. We combine the transaction amount with the transaction frequency to form low-frequency low-value, low-frequency medium-value, low-frequency high-value, medium-frequency low-value, medium-frequency medium-value, medium-frequency high-value, high-frequency low-value, high-frequency medium-value, high-frequency high-value total nine states as the observed state of HMM. The status symbol is represented by $\{ll,lm,lh,ml,mm,mh,hl,hm,hh\}$.

The entire Hidden Markov model is described by $\lambda=(a,b,\pi)$:

$Q=\{q_1,q_2,...,q_N\}$ is a set of all possible states corresponding to the above-mentioned transactions such as transfer transactions and catering consumption. $V=\{v_1,v_2,...,v_M\}$ is a set of all possible observation states corresponding to the nine observation states consisting of the aforementioned amount and frequency. N is the number of states in the model. M is the number of distinct observation symbols per state. In this paper, $M=9$.

$I=\{i_1,i_2...,i_T\}$ is a state sequence of length T, and $O=\{o_1,o_2...,o_T\}$ is the corresponding observation sequence.

$A=\left[a_{ij}\right]_{N\times N}$ is the state transition probability matrix, where

$$a_{ij}=P(i_{t+1}=q_j\mid i_t=q_i), i=1,2,...,N;\ j=1,2,...N \tag{1}$$

It is the probability of transitioning to state q_j at time t_{i+1} under the condition that state q_i at time t.

$B=\left[b_j(k)\right]_{N\times M}$ is the observation symbol probability matrix, where

$$b_j(k)=P(o_t=v_k\mid i_t=q_j), k=1,2,...M;\ j=1,2,...,N \tag{2}$$

It is the probability that the observation v_k is generated under the condition that the time t in the state q_j.

$\pi=(\pi_i)$ is the initial state probability vector, where

$$\pi_i=P(i_1=q_i), i=1,2,...,N \tag{3}$$

It is the probability that time $t=1$ in state q_i.

This paper deals with learning problems and probability calculation problems in the three basic problems of HMM. First, we use a user's processed amount and frequency training sequence samples, i.e., given the observation sequence, by Baum-Welch algorithm [5] estimates the user HMM model parameters. Then, given a sample of the user's test sequence, The Forward-Backward [6] algorithm is used to calculate the probability. Finally, whether the transaction is abnormal is determined based on the calculated probability.

B. *Dynamically Generate Symbols Using K-means*

K-means is a cluster analysis algorithm, which is faster and easier to implement when given a large dataset and is very popular in data mining [7]. The purpose of the K-means clustering algorithm is to divide the n samples into k clusters such that each sample belongs to a category that corresponding to the closest cluster center. That is, find the cluster S_i that satisfies:

$$\underset{S}{\arg\min}\sum_{i=1}^{k}\sum_{x\in S_i}\left\|x-\mu_i\right\|^2 \tag{4}$$

K-means clustering algorithm is calculated with Euclidean distance. From Section 2.1 we can see that the data is divided into nine clusters. Let $c_1,c_2,...,c_9$ be the center point of the generated cluster. When a new transaction occurs, the distances between the new data point and each of the nine cluster centers are calculated. Then it will be classified to the cluster which is the closet cluster based on the calculated distance metrics. Then, the transaction could be represented by the symbol of the assigned cluster. Finally, the user's transaction data will be symbolically represented to form a set of symbol sequences, such as $ml,mm,ml,lm,ll,ml,ml,ml,...,mm$.

III. Build an Anti-fraud Model

The anti-fraud model proposed in this paper could generate user-specific consumption parameters by analyzing the user's spending habits in the profile. And this model is built on the historical transaction data. Then the generated parameters could be used as a reference to select the matched HMM models in order to finally determine the possibility of fraudulent transactions. Based on this idea, each user will have an independent and user-specific HMM model in the system. So, the risk of fraudulent transaction is decreased in this way. The principle is shown in Fig. 2.

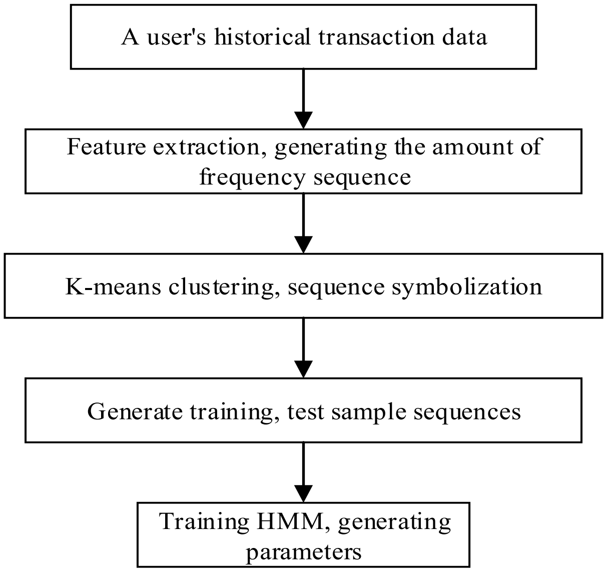
A user's historical transaction data
Feature extraction, generating the amount of frequency sequence
K-means clustering, sequence symbolization
Generate training, test sample sequences
Training HMM, generating parameters

Figure 2. Model training flowchart

A. *User's Trading Behavior Characteristics*

The characteristics of the user's transaction behavior reflect the user's consumption habits. As mentioned above, we roughly divided the users into nine groups: low-frequency low-consumption group, low-frequency middle-consumption group, low-frequency high-consumption group, medium-frequency low-consumption group, medium-frequency medium-consumption group, medium-frequency high-consumption group, high-frequency low-consumption group, high-frequency middle-consumption group, high-frequency high-consumption group. After symbolizing the user transaction data using K-means clustering, the percentage of transactions belonging to the cluster in the user's transaction to the total transaction of the user can be calculated. For example, if a user has a total of 100 transactions, after symbolization, *ml* accounts for 75%, *mm* accounts for 12%, *ll* accounts for 8%, and *lm* accounts for 5%. It can be seen that *ml* accounts for the highest percentage and the user is assigned into the medium-frequency low-consumption group, which indicates that most of the users' transactions belong to the medium-frequency low-consumption cluster. The characteristics of the user's transaction behavior play an important role in the initialization of the following HMM parameters. Different initialization schemes are customized for users with different transaction characteristics, so that the anti-fraud model is more targeted.

B. Parameter Estimation and Training

After the steps of symbolizing the transaction data and grouping the users, the training data of the HMM can be obtained, that is, several sets of observation sequences $\{O_1, O_2, \ldots, O_S\}$. Under the condition of a given observation sequence, the parameters $\lambda = (A, B, \pi)$ of the Hidden Markov model are learned using the Baum-Welch algorithm. It should be noted that when using the Baum-Welch algorithm to learn parameters, it is necessary to provide an initial value $\lambda = (A, B, \pi)$. Normally, it is set to be uniform distribution. However, user grouping can be combined here. Different groupings are set with different initial values so that the learned model can reflect the user's transaction behaviors more accurately.

C. Fraud Detection

The user-specific HMM model $\lambda = (A, B, \pi)$ could be obtained after the parameter training. There is a fraud detection method specialized for credit card proposed in [8]. The user's observation sequence with the length R is extracted in the time period t , denoted as $o_1, o_2, \ldots, o_R$. It is used as the input for the HMM model, and the forward-backward algorithm is used to calculate the probability of being accepted by the HMM, which is denoted as α_1 , as follows:

$$\alpha_1 = P(o_1, o_2, ..., o_R \mid \lambda) \tag{5}$$

Suppose o_{R+1} is the new state symbol generated at time $t+1$ and o_1 is removed from the original sequence $o_1, o_2, \ldots, o_R$ .Then state o_{R+1} is appended to the end of the sequence, so the new sequence would be $o_2, o_3, \ldots, o_R, o_{R+1}$. The probability of being accepted by the HMM model should be calculated under the same method, which is denoted as α_2 :

$$\alpha_2 = P(o_2, o_3, ..., o_{R+1} \mid \lambda) \tag{6}$$

If $\triangle\alpha = \alpha_1 - \alpha_2$, $\triangle\alpha > 0$, it means the new sequence $o_2, o_3, \ldots, o_R, o_{R+1}$ is accepted by the HMM model with a low probability. Then the newly added transaction status symbol o_{R+1} is likely to be a fraudulent transaction. In this case, a threshold could be used to make a decision: if $\triangle\alpha / \alpha_1 > Threshold$, then the new transaction at the tail of the sequence could be determined as a fraud transaction.

IV. Experiments and Results

TP, FP, Precision and Recall are used as the criteria to evaluate the mentioned model in the experiment. TP indicates the actual number of fraud transactions that are determined as fraud transactions, and FP indicates the number of cases when the normal transactions are wrongly detected as the fraud transactions. It is desired that the TP value of the model is as high as possible and the FP value is as low as possible. In the actual application of the bank, the wrong detection of a normal transaction as a fraud one will increase the burden on the banking business, so ensuring a lower FP value can make the anti-fraud model more practical. It is very difficult to directly apply the real-world transaction data into the experiment due to the confidentiality and privacy feature of the data in the bank. Therefore, the simulation data is used at the beginning stage of the experiments in order to verify the feasibility of the model as well as the utility of the parameters in the model. Finally, the second verification is applied by using the real-world transaction data.

A. Simulation Experiment and Result Analysis

In order to compare with the HMM model directly using the amount sequence training, we use the same parameters in the experiment. Number of the hidden states is also set to 10. The sequence length is set to 15. The threshold is set to 50%. Both normal data and fraudulent data is included in the simulated test. The fraudulent data is normally distributed. The setting $\mu = 1$, $\sigma = 0.5$ indicates that there is a fraudulent transaction within the sequence no matter what the input for the model is. For the HMM model with the amount and frequency as the observation sequence, respectively for the low-frequency low-consumption group, low-frequency middle-consumption group, low-frequency high-consumption group, medium-frequency low-consumption group, medium-frequency medium-consumption group, medium-frequency high-consumption group, high-frequency low-consumption group, high-frequency middle-consumption group, high-frequency high-consumption group were simulated and 100 experiments were performed in each group. Finally calculate the average TP and FP for each group. The results are shown in Table I.

TABLE I. Simulation Result

	Amount sequence			Amount and frequency sequence								
Group	l	m	h	ll	ml	hl	lm	mm	hm	lh	mh	hh
TP	0.64	0.62	0.49	0.81	0.72	0.52	0.71	0.69	0.61	0.63	0.62	0.53
FP	0.06	0.07	0.13	0.08	0.04	0.02	0.05	0.03	0.03	0.08	0.04	0.01

As can be seen from Figure 3, in the case of a large TP value, the FP value is also relatively large, and this model performs well in the low, medium amount, and frequency groups.

The HMM model trained by the amount and frequency of the observed sequence is better than the HMM model trained by the amount of observation sequence in the low, medium frequency, and amount groups, while keeping the TP value high, FP value is relatively low, indicating that this method can be more effective in the simulation experiment to complete the task of fraud detection.

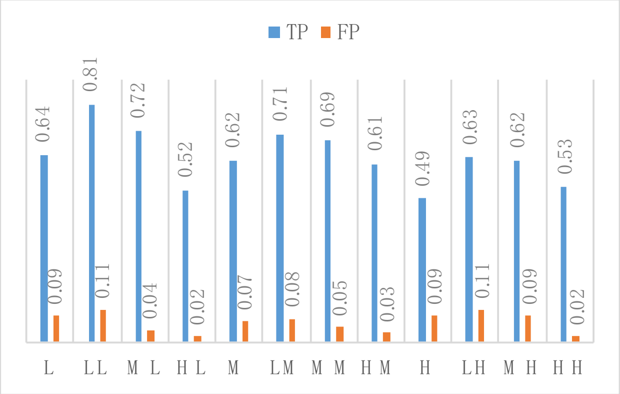

Figure 3. Simulation result

B. *Real Bank Data Verification Experiment and Results*

In order to test the effectiveness of this anti-fraud model under real-world bank transaction data, the third-party transaction data provided by the third party is used in this experiment, and 20000 accounts among over 2 million normal accounts are selected for the experiment. For these test accounts, 21 are identified to have fraud history, totally 61 fraud transactions. The goal is to check whether the method still works under the case where the amount of fraud cases is far away from the amount of normal cases. The thirty most recent transactions in each account are used as test data. And all fraudulent transactions are existed in the last thirty transactions of the corresponding account. All other normal historical transaction data are used as training data to train the HMM model of the account. The experimental results are shown in Table II:

TABLE II. EXPERIMENTAL RESULT

Threshold	TP	FP	precision	recall
0.9	17	142	0.11	0.28
0.7	22	305	0.07	0.36
0.5	24	522	0.04	0.39
0.3	24	1933	0.01	0.39

It can be seen in Table II that the number of false alarms is very large and the model is not ideal. Through the study of transaction data, it is found that some users have less historical transaction data, and the HMM model trained through its historical transaction data cannot fully reflect the user's transaction behaviors. The model would generate false alarm even if the user has a new normal transaction behavior after this kind of training. The model will fail the correct detection if some fraud transactions are only distributed among the first several transactions when the account opened. Besides, it also fails when an account only contains fraud transactions. For these situations, the model could not play a role.

In response to the above situations, we have adjusted the data collection. In the more than 2 million accounts of third-party transaction data provided by the cooperative bank, a total of 438,392 accounts with historical transaction data greater than 100 were selected. Then ran out of 438,392 accounts, randomly selected 20,000 accounts and 11 bank accounts confirmed fraudulent transactions, of which 34 fraudulent transactions. Similarly, the 30 most recent transactions in each account are used as test data. All fraudulent transactions are existed in the last 30 transactions of the corresponding account. All other normal historical transaction data are used as training data to train the HMM model of the account. The final experimental results are shown in Table III:

TABLE III. IMPROVED EXPERIMENTAL RESULT

Threshold	TP	FP	precision	recall
0.9	11	13	0.46	0.32
0.7	16	34	0.32	0.47
0.5	23	74	0.31	0.68
0.3	24	136	0.15	0.71

As can be seen in Table III, the results of the experiment are ideal after removing less than 100 accounts with historical transaction records. When the threshold parameter is taken as 0.5, the comprehensive effect of precision and recall is better.

V. CONCLUSION

This paper focuses on the needs of banking transaction anti-fraud modelling and transaction detection, and does the research on the modeling and application of bank anti-fraud based on K-means clustering and Hidden Markov model. Simulation experiments and bank real data verification experiments have proved that this method can detect bank transaction data to a certain extent, and it could perform well for low, medium frequency and amount of user groups which can provide an effective solution for the bank fraud problems. In the future work, we will focus on the selection of the parameters of the HMM model. We have known that initial parameters that are consistent with consumer habits play an important role in models and the determination of the number of hidden states deserves our consideration at the same time. With the development of electronic commerce and the development of information technology in China, the

HMM model will show its unique superiority in the field of fraudulent transaction risk detection.

REFERENCES

[1] Alfian N, Tarjo T, Haryadi B. THE EFFECT OF ANTI FRAUD STRATEGY ON FRAUD PREVENTION IN BANKING INDUSTRY[J]. 2017, 2(1):61.

[2] Kim J H, An T K, Kim Y N, et al. Abnormal Events Recognition Framework based on HMM[C]// ITC-CSCC :International Technical Conference on Circuits Systems, Computers and Communications. 2015.

[3] Alsharif O, Pineau J. End-to-End Text Recognition with Hybrid HMM Maxout Models[J]. Computer Science, 2013.

[4] Bhingarde A, Bangar A, Gupta K, Karambe S. Credit Card Fraud Detection using Hidden Markov Model[J]. International Journal of Computer Science & Information Technolo, 2015, vol. 76, pp. 169-170.

[5] Lindberg D V, Omre H. Inference of the Transition Matrix in Convolved Hidden Markov Models and the Generalized Baum–Welch Algorithm[J]. IEEE Transactions on Geoscience & Remote Sensing, 2015, vol. 53, pp. 6443-6456.

[6] Merialdo B. On the locality of the forward-backward algorithm [speech recognition][J]. IEEE Transactions on Speech & Audio Processing, 2015, vol 1, pp. 255-257

[7] Cai X, Nie F, Huang H. Multi-view K-means clustering on big data[C]. International Joint Conference on Artificial Intelligence. AAAI Press, 2013, pp. 2598-2604.

[8] Raparty L V P, Nammi S R. Credit Card Fraud Detection Using Hidden Markov Model[J]. International Journal of Soft Computing & Engineering, 2012, vol 2