Hidden Markov Model based Credit Card Fraud Detection System with Time Stamp and IP Address

Aayushi Gupta
Scholar, Department of Information Technology, Oriental Institute of Science & Technology, Bhopal

Dhananjay Kumar
Department of Information Technology, Oriental Institute of Science & Technology, Bhopal

Atul Barve
Associate Professor, Department of Information Technology, Oriental Institute of Science & Technology, Bhopal

ABSTRACT
The evolution of the new technology supports the online transactions to be held with the assistance of different payment cards. Credit card frauds have become increasingly rampant in living years and critical for banks to enhance fraud detection so as to protect their cardholders from financial loss. The simple way to detect such kind of fraud is to decipher the spending pattern on each card and to highlight any irregularity with respect to the "standard" spending pattern. In this paper we try to review Hidden Markov model which works on such technique. The HMM, trained with the normal behavior of a cardholder needs an enough number of normal transactions and fraud transactions for learning fraud patterns. To make it more effective we have enclosed the provision of determining the IP address of intruder machine along with its time stamp. The simulation analysis include different real dataset to identify the fraud and discover the intruder. Form our model it is proven that it works with more efficiency than existing models.

Keywords
Hidden Markov model, spending pattern, fraud transaction, credit card, time stamp, financial loss.

1. INTRODUCTION

Online activities are well acknowledged to every citizen of the society with the eminent growth of e-commerce. Online activities mainly involve regular purchase of goods, electronic devices and other such things. The online transactions made for such activities are secure payment methods that authorize the transfer of funds. These transactions are supported by different bank cards which makes the operation easy. A huge population use credit card for its undemanding accessibility. The bank has accumulated a vast count of credit card transactions.

Apart from its magnificent advantages they do face some of their pitfalls regarding the security. The illicit use of these credit cards is a major issue to ponder on. The credit card fraud can be done for various reasons, mainly to get unaccredited funds from the account. It is thus the responsibility of the bank to safeguard the amount transferred online on the internet of the card holder. The bank organization can adopt various existing methodologies such as case based reasoning, decision tree, and neural network for fraud detection in order to reduce the financial loss [1].

From among various detection techniques our approach focus on Hidden Markov model which detects the fraud transaction and concurrently report the timestamp and IP address of the intruder's machine. The HMM prior processing include maintaining the record of the card holders transactions to evaluate its normal behavior [2]. Every time a new transaction made is recorded in the system. The Hidden Markov model then automatically generate the spending profile of the user. Now if any intruder tries to make transaction with any registered credit card, then its spending habit will be different from that of authenticate user and can be easily captured. Through this system we make sure that no genuine transaction is rejected. The system is capable of recording the timestamp and IP address of the attacking machine so that the geographic location of intruder can be traced.

2. LITERATURE SURVEY

Credit card fraud detection has been a current evoking issue of major concern. In affect to this various detection techniques such as genetic algorithms, data mining, neural networks, clustering techniques and decision tree are used.

Ghosh and Reily [3] implemented the neural network system which involved cases dealing with lost cards, stolen cards, stolen card details, application fraud etc. Aleskerov, Freisleben and Rao [4] also developed a system on neural network called Card watch. The system focus towards commercial implementation.

Dorronsoro and others [5] developed a neural network based detection system called Minerva. This system proposes the facility to ingrain itself deep in credit card transaction servers to detect fraud in real-time. Kokkinaki [6] suggested to create a user profile for each credit card account and to test incoming transactions against the corresponding user's profile. Chan and Stolfo [7] studied the class distribution of a training set and its effects on the performance of multi classifiers on the credit card fraud domain. Brause and others [8] looked specifically at credit card payment fraud and identified fraud cases by combining a rule-based classification approach with a neural network algorithm. Kim [9] proposed a fraud density map technique to improve the learning efficiency of a neural network. Chiu and Tsai [10] identified the problem of credit card transaction data having a natural skewness towards legitimate transaction. Foster and Stine [11] attempted to predict personal bankruptcy using a fully automated stepwise regression model.

3. PROPOSED MODEL

The Hidden Markov model is undemanding and easily manageable sequential model which is use to model the spending convention of the card holder (user). It is a doubly embedded random process comprising of two disparate levels. One of them remains hidden and other is noticeable to observer. The Hidden Markov model has greater potential in managing complex process than the traditional Markov model. The considerable advantage seen in the model is the diminution in number of FP (False Positives). FP is the transition identified as

fraudulent by the fraud detection system but although they are genuine.

The new model consists of finite set of states which are associated with probability distribution. Transitions among different states are supervised by set of probability called as transition probability [13]. Every state in model originates some outcome called as observation calculated according to corresponding probability distribution. HMM can be successfully applied to various applications in temporal pattern recognitions such as speech, handwriting gesture reorganization part of speech tagging and bioinformatics [12].

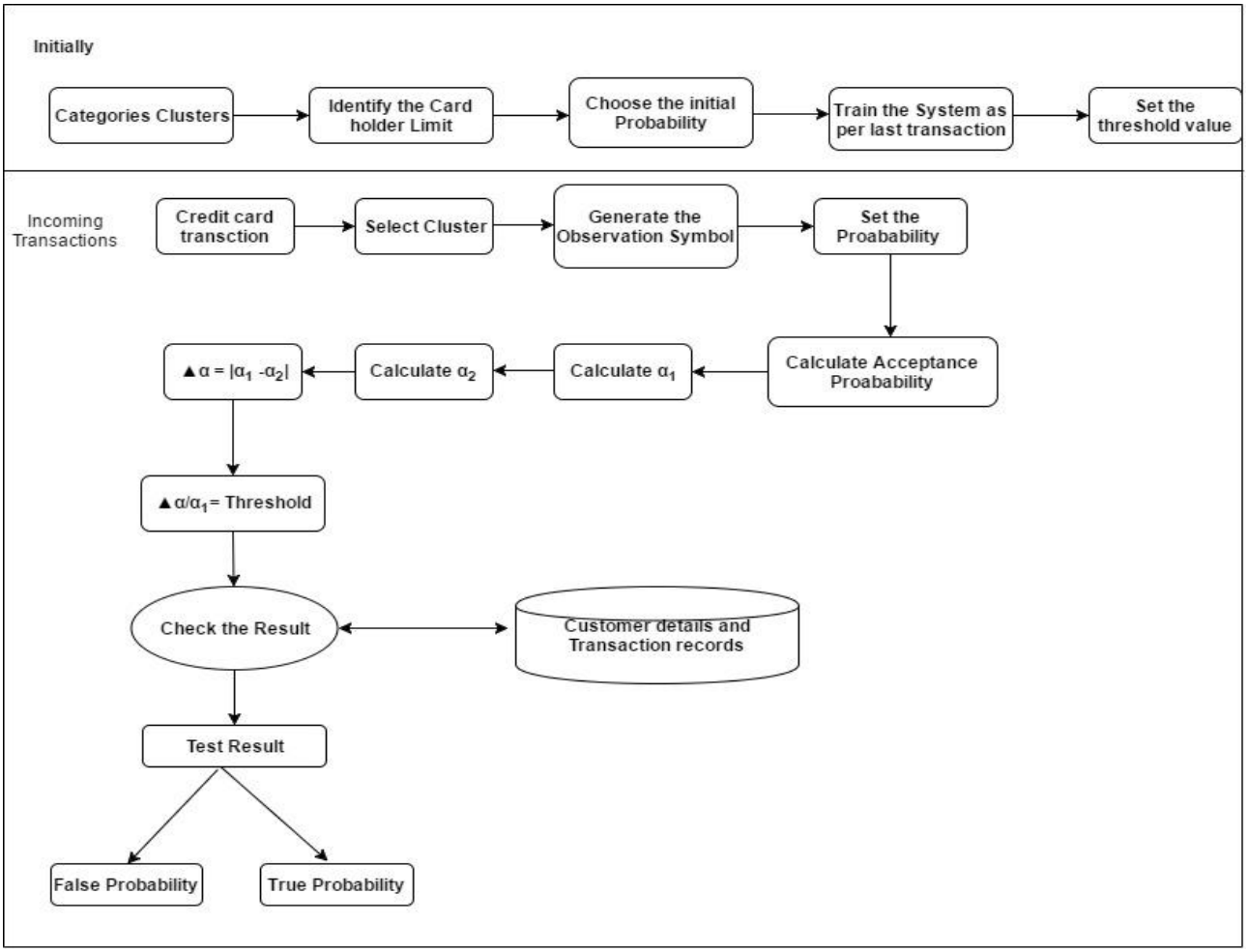

Fig 1: Flow of proposed fraud detection model

The HMM can be well defined with the following elements-

- ✓ N number of states that are hidden denoted by a set S = $\{S_1,S_2,S_3,\ldots. S_N\}$, where i =1, 2,……, N, is count of state and Si, is an individual state.
- ✓ M denotes the total number of observation symbols. When observations are continuous then M is infinite. We denote the set of symbols V = $\{_{V1},_{V2},\ldots._{VM}\}$ where $_{Vi}$, is an individual symbol.
- ✓ A set containing probability of moving from one state to another, defined as transition probability.

  $a_{ij}= P\{q_{t+1} = S_j \mid q_t = S_i\}$, $1\leq i,\ j\leq N$

  where q_t denotes the present state. Transition probabilities should satisfy two constraints

  $a_{ij}\geq 0$, $1\leq i,\ j \leq N$

  and

  $\sum_{i=1}^{N} aij= 1,\ 1\leq i \leq N$
- ✓ Matrix B, indicating observation symbol probability

  $B = \{b_j(k)\}$

  A probability distribution in each of the states is given as,

  $b_j(k) = P\{a_t =V_k| q_t = S_j\}, 1 \leq j \leq N, 1 \leq k \leq M$ where, V_k denotes the k^{th} observation symbol and a_t the current parameter vector. The given equation should satisfy some constraints

  $$b_j(k) \geq 0,\ 1 \leq j \leq N,\ 1 \leq k \leq M$$

  and

  $$\sum_{k=1}^{M} b_j(k) =1,\ 1 \leq j \leq N$$
- ✓ The initial state probability given by

  $$\Pi = \{\Pi_i\}$$

  where,

  $$\Pi_i = P\ \{q_i = S_i\},\ 1 \leq i \leq N$$

  $$\sum_{i=1}^{N} \pi_i= 1$$

The compact notation for the above defined probabilities is given by

$$\Pi = (A, B, \lambda)$$

The observation sequence $O=O_1,O_2,\ldots\ldots O_R$ where observation O_t is one of the symbols from V, and R denotes the number of observation sequence. An observation sequence O can be generated by different possible state sequences. One such particular sequence is

$$Q = q_1, q_2, q_3,\ldots q_R,$$

where q1 is the beginning state. The probability that O is generated from this state sequence is given by

$$P(O|Q, \lambda) = \prod_{t=1}^{R} P(O_t|q_t, \lambda)$$

where statistical independence of observation is assumed. The above equation can be extended as,

$$P(O|Q, \lambda) = b_{q1}(O_1).\, b_{q2}(O_2)\ldots\ldots\ldots b_{qR}(O_R)$$

The probability of state sequence Q given as

$$P(Q|\lambda) = \pi_{q1}\, a_{q1q2}\, a_{q2q3}\ldots\ldots\ldots a_{qR\text{-}1qR}$$

Thus, the generation of probability of the observation sequence O by the HMM identified by λ can be written as,

$$P(O|\lambda) = \sum_{all\ Q} P(O|Q, \lambda) P(Q|\lambda)$$

After knowing all these elements, the HMM is ready to work. We consider the initial sequence of transaction of card holder. The transactions made by the credit card holder is categorized into three clusters l, m, h for low medium and high category transaction respectively. The volume of each cluster is resolved considering the limit of the credit card. The amount up to 35% of card limit belongs to l cluster, upto 65% belongs to m and above that comes under h cluster.

Let $O_1,O_2,\ldots\ldots O_R$ be consisting of R symbols to form a sequence. The HMM now works in the following manner

- ✓ To calculate the probability of acceptance (alpha1) we consider a series of last R transaction of the card holder. The value is given by
  $$\alpha_1 = P(O_1, O_2, O_3,\ldots\ldots O_R|\lambda)$$
- ✓ Let OR+1 is the new generated transaction at time t+1. The total no of sequences R+1. Now for calculating the second probability of acceptance we will drop O1 observation and the sequence will be from O2 to OR+1
  $$\alpha_2 = P(O_2, O_3, O_4,\ldots\ldots\ldots O_{R+1}|\lambda)$$
- ✓ Next we find the standard deviation
  $$\blacktriangle\alpha = \alpha_1 - \alpha_2$$
- ✓ The standard deviation calculated is now compared with threshold value. If percentage change in probability is found more than predefine threshold value then the transaction will be consider as fraud transaction.
  $$\text{If } \blacktriangle\alpha/\alpha_1 \leq \text{threshold value } (\Theta)$$
- ✓ Ideally the threshold value is taken as 0.5 and further the value modifies every time the algorithm runs using the below formula

$$\text{Threshold} = (\blacktriangle\alpha/\alpha_1 + \text{threshold})/2$$

Table 1. Notations and Acronyms

Notation	**Meaning**
N	Number of hidden states
M	Number of observation symbols
R	Sequence length
V_k k=1 to n	Observation symbols
l,m,h	Low medium high cluster
$H_{x\text{-}y}$	Probability of transition from Hidden Markov model state x to y.
α	Probability of acceptance of sequence
Θ	Threshold value calculated
▲α	Represents standard deviation between two acceptance probabilities
Acronym	**Expanded Form**
FDS	Fraud Detection System
HMM	Hidden Markov Model
TP, FP	True Positive, False Positive
hs, ms, ls	High spending, medium spending, low spending

4. RESULTS

Due to the obvious security reasons it is very difficult to fetch the dataset from any bank. So in order to get our results simulated analysis is performed by considering a random dataset of transactions for any credit card holder. Firstly, all the transaction sequence need to be categorized into three clusters namely low medium and high according to the user credit card limit. Assuming the credit card limit to be ₹. 10000 in our case, the range of clusters thus produced will be low {₹. 0, 3000}, medium {₹. 3000, 6000} and high {₹. 6000, 10000}. After deciding the categories the fraud detection of incoming transaction will be verified by last 10 transactions.

Table 2. List of all the transaction of a cad holder

Transaction No	**Amount**	**Category**
1st	1086	1
2nd	2836	1
3rd	5590	2
4th	6920	3
5th	8640	3
6th	102	1
7th	937	1
8th	763	1
9th	1649	1
10th	2980	1
11th	1672	1
12th	3760	2
13th	1578	1
14th	1920	1
15th	940	1
16th	1327	1
17th	2670	1
18th	9260	3

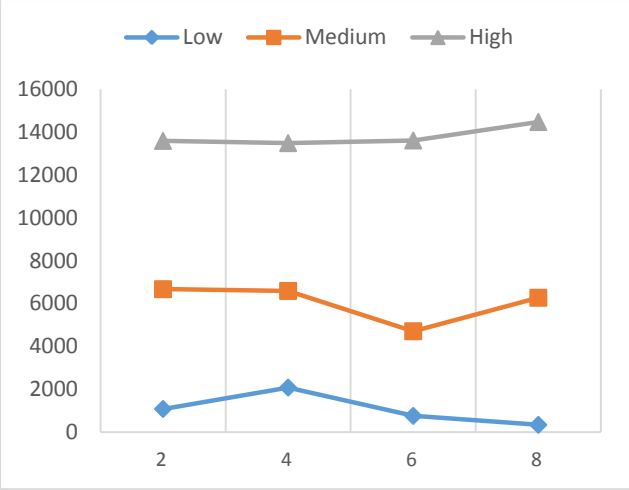

Fig 2: Different categories of the transaction

In the table above the transaction sequence with its amount dataset is represented. From the given datasets we calculate the first acceptance probability to check the spending habit of the user taking 7^{th} to 17^{th} transaction.

$$\text{Acceptance Probability} = \sum_{r=1}^{R} Proability\ of\ clusters(Or)$$

$$\alpha_1 = 0.5923$$

Since the 18^{th} dataset refers to current transaction, second acceptance probability is calculated taking this transaction into consideration. So,

$$\alpha_2 = 0.65413$$

With the values of two acceptance probability we can determine the standard deviation as,

$$\blacktriangle\alpha = |\alpha_1 - \alpha_1|$$

$$\blacktriangle\alpha = |0.5923 - 0.65413|$$

$$\blacktriangle\alpha = 0.06183$$

The percentage change of this standard deviation is compared with the threshold value Θ. Ideally the Θ is taken as 0.1. Further it is continually calculated every time the algorithm runs. So the value of threshold is empirically calculated for every transaction. Following this criteria the current threshold value turns out to be 0.0842.

$$\blacktriangle\alpha/\alpha_1 = 0.06183/0.5923$$

$$\blacktriangle\alpha/\alpha_1 = 0.10438$$

Since,

$$0.10438 > 0.0842(\Theta)$$

the above transaction is detected as fraudulent and the user has to answer the security questions. Simultaneously, the IP and time stamp of the user is also recorded.

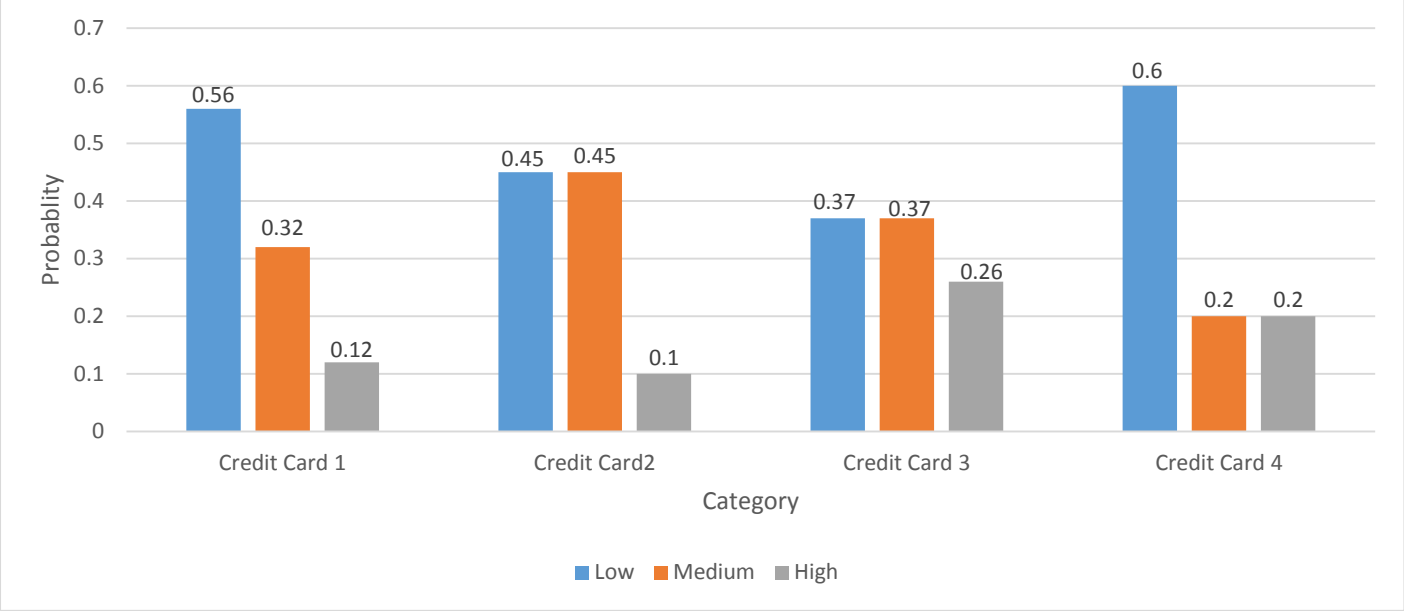

Fig 3: Spending profile of different card holder

5. CONCLUSION

The technology supporting online transaction has provoked the use of payment cards. Though online transactions have paved a smooth path for the customers, at the same time has provoked various threats to security for these transactions. In our system we have proposed the application of HMM in detecting the credit card frauds thereby recording the IP of the fraud system along with the timestamp when malignant attempted to attack. It thus helps in tracing the geographic location of the attacker.

In this paper it is shown that the processing of the HMM starts with grouping of various transaction amount sequence into three categories which forms different hidden states of the model. The range of such clusters is reliant on limit of the credit card which varies with the user. Now with the help of such clusters, model

suggests to find spending profile of the user for given sequence. The percentage change in the probabilities of previous and new transaction sequence is compared with the threshold value which decides whether the upcoming transaction is fraudulent or not.

Comparative studies revealed an accuracy of the system to be about 80% for a wide range of input dataset. Thus the produced system is genuine to a great extent. It has also reduced the complexity when compared with the existing system. In our simulation analysis we have considered a small set of data, but our proposed system is capable of handling larger range of transactions which is quite certain in real life scenarios.

6. FUTURE WORK

The system has the flexibility for the future enhancement at the same time shows its advantage of dynamic nature. There will always be a method to enhance the probability which we use for the fraud detection based on practical datasets and values. Also the algorithm used is applied at one layer. For stronger protection multiple layer algorithm can be implemented. Further in future we can design an application which can be add sophisticated modules like capturing the photo of the attacker.

7. REFERENCES

[1] Khyati Chaudhary, Bhawna Mallick, "Credit Card Fraud: The study of its impact and detectionTechniques", International Journal of Computer Science and Network (IJCSN), pp: 31-35, 2012.

[2] Bilonikar Priya, "Survey on Credit Card Fraud Detection Using Hidden Markov Model", International Journal of Advanced Research in Computer and Communication Engineering 2014.

[3] Ghosh S., Reilly D.L., "Credit Card Fraud Detection with a Neural- Network" Proceedings of the International Conference onSystem Science, pp.621-630, 1994.

[4] Aleskerov E., Freisleben B., and Rao B., "CARDWATCH: A Neural Network Based Database Mining System for Credit CardFraud Detection", Proc. IEEE/IAFE: Computational Intelligence for Financial Eng., pp.:220-226, 1997.

[5] Dorronsoro J.R., Francisco G., Carmen S., and Carlos S.C., "Neural Fraud Detection in Credit Card Operation" IEEETransaction on Neural Network, vol.-08, no.-04, pp.: 827-834, 1997.

[6] Kokkinaki, A., "On Atypical Database Transactions: Identification of Probable Frauds using Machine Learning for UserProfiling." Knowledge and Data Engineering Exchange Workshop. IEEE, pp.:107-113, 1997.

[7] Stolfo S.J., Fan D.W., Lee W., Prodromidis A.L., and Chan P.K., "Credit Card FraudDetection Using Meta-Learning: Issues andInitial Results", Proc. AAAI Workshop AI Methods in Fraud and Risk Management, pp.:83-90, 1997.

[8] Brause R., Langsdorf T., and Hepp M., "Neural Data Mining for Credit Card Fraud Detection", Proc. IEEE Int'l Conf. Toolswith Artificial Intelligence, pp.:103-106, 1999.

[9] Kim, M, and Kim T., "A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection", Proceedings of IDEAL. pp.:378-383, 2002.

[10] Chiu A., Tsai C., "A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection", Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, pp.:177-181, 2004.

[11] Foster and Stine R., "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy", Journal of American Statistical Association, pp.: 303-313, 2004.

[12] V. Bhusari, S. Patil "Application of Hidden Markov Model in CreditCard Fraud Detection", International Journal of Distributed and Parallel Systems (IJDPS), pp: 203-211, 2011.

[13] Arun K Majumdar "Credit Card Fraud Detection Using HiddenMarkov Model"IEEE transactions on dependable and secure computing, pp; 37-47, 2008.