ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Using Hidden Markov Model in Detection of Frauds in Credit Cards

Revanth Kumar R[1], Prakruthi R Bharadwaj[2], Kuruva Lavanya[3], Rakshitha H V[4]

Student, Department of Computer Science and Engineering, Sambhram Institute of Technology, Bengaluru, India[1]

Student, Department of Computer Science and Engineering, Sambhram Institute of Technology, Bengaluru, India[2]

Student, Department of Computer Science and Engineering, Sambhram Institute of Technology, Bengaluru, India[3]

Student, Department of Computer Science and Engineering, Sambhram Institute of Technology, Bengaluru, India[4]

ABSTRACT: Due to a fast progression in the electronic trade technology, the use of credit cards has intensely amplified. As credit card becomes the most general mode of sum for both online as well as regular buying, cases of fraud associated with it are also increasing. In this paper, we model the order of methods in credit card transaction processing using a Hidden Markov Model and illustrate how it can be used for the detection of frauds. An HMM is primarily trained with the normal behaviour of a cardholder. If an arriving credit card transaction is not accepted by the trained HMM with necessarily high chance, it is measured to be fraudulent. At the same period, we try to ensure that real transactions are not rejected. We present detailed experimental outcome to show the effectiveness of our method and compare it with other methods available in the works.

KEYWORDS: Internet, E-shopping, Credit Card, E-commerce Security, Fraud Detection, Hidden Markov Model.

I. INTRODUCTION

The popularity of online shopping is increasing gradually. According to Nielsen study conducted in 2005, one-tenth of the world's population is shopping online. Germany and Great Britain have the major number of online shoppers, and credit card is the most general mode of payment (70% percent). About 400 million transactions/year were reportedly carried out by Barclaycard, the largest credit card company in the UK, near the end of the last century. Retailers like Wal-Mart typically handle much larger number of credit card transactions including online and regular purchases. As the number of credit card users increases world-wide, the opportunities for attackers to giveaway credit card details and, then, commit fraud are also growing. The total credit card fraud in the US itself is reported to be $2.7 billion in 2005 and valued to be $3.0 billion in 2006, out of which $1.6 billion and $2billion, respectively, are the estimates of online fraud.

Credit-card based consumptions can be classified into two types: 1) physical card, 2) virtual card. In a physical-card- based buying; the cardholder presents his card physically to a trader for making a payment. To carry out fraudulent transactions in this kind of buying, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In the second kind of purchase, only some important information about a card (card number, expiration date, secure code) is required to make the payment. Such purchases are normally done on the Internet or over the telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyze the spending patterns on every card and to figure out any inconsistency with respect to the "typical" spending patterns. Fraud detection based on the analysis of prevailing purchase data of cardholder is a talented way to lessen the rate of efficacious credit card frauds. Since people tend to exhibit precise behaviouristic profiles, every cardholder can be characterized by a set of patterns comprising information about the distinctive purchase category, the time

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential hazard to the system.

II. RELATED WORK

Credit card fraud detection has drawn a lot of research interest and a number of methods, with special importance on data mining and neural networks, have been suggested. Ghosh and Reilly have proposed credit card fraud detection with a neural network. They have built a detection system, which is trained on a large sample of labelled credit card account transactions. These transactions contain example fraud cases due to lost cards, stolen cards, application fraud, mail-order fraud, and non-received issue (NRI) fraud. Recently, Syeda et al. have used parallel granular neural networks for improving the speed of data mining and knowledge discovery process in credit card fraud detection. A complete system has been implemented for this purpose. Stolfo et al. suggests a credit card fraud detection system (FDS) using meta-learning techniques to learn models of fraudulent credit card transactions. Meta-learning is a general approach that offers a means for combining and integrating a number of distinctly built correlations of the estimates of the base classifiers. The same group has also operated on a cost-based model for fraud and interference detection. They use Java agents for Meta-learning, which is a distributed data mining system for credit card fraud detection. A number of noteworthy routine metrics like True Positive—False Positive (TP-FP) spread and precision have been defined by them.

Aleskerov et al. present CARDWATCH, a database mining system used for credit card fraud detection. The system, based on a neural learning module, provides an interface to a variety of commercial databases. Kim and Kim have identified skewed distribution of data and mix of legitimate and fraudulent transactions as the two main reasons for the complexity of credit card fraud detection. Based on this observation, they use fraud density of real transaction data as a confidence value and generate the weighted score that is fraud score to reduce the number of misdetections.

Fan et al suggest the application of distributed data mining in credit card fraud detection. Brause et al. have developed an approach that involves advanced data mining techniques and neural network algorithms to obtain high fraud coverage. Chiu and Tsai have proposed Web services and data mining techniques to establish a collaborative scheme for fraud detection in the banking industry. With this scheme, participating banks share knowledge about the fraud patterns in a heterogeneous and distributed environment. To establish a smooth channel of data exchange, Web services techniques such as XML, SOAP, and WSDL are used. Phua et al. have done an extensive survey of existing data-mining-based FDSs and published a comprehensive report. It is based on artificial intelligence and combines inductive learning algorithms and meta-learning methods for achieving higher accuracy.

The problem with most of the aforementioned methods is that they need labeled data for both unaffected, as well as untrue transactions, to train the classifiers. Getting real-world fraud data is one of the biggest difficulties related with credit card fraud detection. Also, these methods cannot sense new types of frauds for which labeled data is not obtainable. In contrast, we present a Hidden Markov Model (HMM)-based credit card FDS, which does not require fraud signatures and yet is able to detect frauds by considering a cardholder's spending habit. We model a credit card transaction processing sequence by the stochastic process of an HMM. The details of items purchased in individual transactions are usually not known to an FDS running at the bank that issues credit cards to the cardholders. This can be represented as the underlying finite Markov chain, which is not observable. The transactions can only be observed through the other stochastic process that produces the sequence of the amount of money spent in each transaction. Hence, we feel that HMM is an ideal choice for addressing this problem. Another important advantage of the HMM-based approach is a drastic reduction in the number of False Positives (FPs)—transactions identified as malicious by an FD Sal though they are actually genuine. Since the number of genuine transactions is a few orders of magnitude higher than the number of malicious transactions, an FDS should be designed in such a way that the number of FPs is as low as possible. Otherwise, due to the "base rate fallacy" effect, bank administrators may tend to ignore the alarms. To the best of our knowledge, there is no other published

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

literature on the application of HMM for credit card fraud detection. The rest of the paper is organized as follows: We first briefly explain the working principle of an HMM. We then show how to model credit card transaction processing using HMM further in this page. We also describe the complete process flow of the proposed FDS in this section. Finally, we conclude the paper with some discussions.

III. PROPOSED METHOD (HIDDEN MARKOV MODE)

An HMM is a double embedded stochastic procedure with dual hierarchy stages. It can be used to model much more difficult stochastic procedures as compared to a traditional Markov model. An HMM has a limited set of states overseen by a set of transition probabilities. In a specific state, an outcome can be produced rendering to an associated probability distribution. It is only the result and not the state that is observable to an outside viewer.

HMM-based applications are shared in various areas such as speech recognition, bioinformatics. In recent years, Joshi and Phoba have investigated the capabilities of HMM in anomaly detection. They classify TCP network traffic as an attack or standard using HMM. Cho and Park suggest an HMM-based intrusion detection system that improves the modeling time and performance by considering only the privilege transition flows based on the domain knowledge of attacks. Ourston et al. have proposed the application of HMM in sensing multistage network attacks. Hoang et al. show a new method to process sequences of system calls for irregularity recognition using HMM. The key knowledge is to build a multilayer model of program actions based on both HMMs and counting approaches for anomaly discovery. Once human actions are appropriately modeled, any deviation is a source for concern since an intruder is not expected to have a behavior similar to the unaffected user. Hence, an alarm is raised in case of a detected deviation.

An HMM can be characterized by the following:

1. N is the number of states in the model. We denote the set of states' S = {S1, S2…SN}, where S_i, i = 1, 2...., N is a distinct state. The state at time instant t is denoted by qt.
2. M is the number of discrete observation symbols per state. The observation symbols resemble to the physical result of the system being modeled. We represent the set of symbols V= {V1; V2; ... VM}, where V_i, i = 1, 2....., M is a separate symbol.
3. The state transition probability matrix A= $[a_{ij}]$, where $a_{ij} = P (qt+1=S_j \mid q_t=S)$, $1 \le i \le N$, $1 \le j \le N$; t=1, 2, 3….

   For the overall case where any state j can be reached from any other state i in a sole step, we have $a_{ij} > 0$ for all i, j. Also, $\sum_{j=1}^{N} aij = 1$, $1 \le i \le N$.
4. The observation symbol probability matrix

   B = $[b_j (k)]$, where

   $b_j (k) = P (V_k|S_j)$, $1 \le j \le N$, $1 \le k \le M$ and

   $\sum_{k=1}^{M} bj(k) = 1$, $1 \le j \le N$.
5. The original state probability vector $\pi = [\pi_i]$, where

   $\pi_i = P (q_1 = S_i)$, $1 \le i \le N$, such that $\sum_{i=1}^{N} \pi i = 1$.
6. The observation categorization O = O1, O2, O3...O_R, where individual observation O_t is one of the symbols from V, and R is the number of observations in the sequence.

It is evident that a complete specification of an HMM requires the estimation of two model parameters, N and M, and three probability distributions A, B, and π. We use the notation λ = (A, B, π) to indicate the complete set of parameters of the model, where A, B implicitly include N and M. An observation sequence O, as mentioned above, can be generated by many possible state sequences. Consider one such particular sequence Q = q_1, q_2....., q_R, where q_1 is the initial state. The probability that O is generated from this state sequence is given by

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

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

where statistical independence of observations is presumed. Equation can be long-drawn-out as

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

The probability of the state sequence Q is given as

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

Thus, the probability of generation of the observation sequence O by the HMM stated can be written as follows:

$$P(O|\lambda) = \sum(\text{all } Q) = P(O|Q, \lambda)\, P(Q|\lambda).$$

Deriving the value of $P(O|\lambda)$ by means of the direct definition of the above equation is computationally severe. Hence, a practice called as Forward-Backward procedure is used to figure $P(O|\lambda)$.

Table I: Notations and Acronyms

Notations	**Meaning**
M	Number of observation symbols
N	Number of hidden states
V_k k=1….M	Observation symbols
l, m, h	Price ranges – low, medium, high
$a_{x\text{-}y}$	Probability of transition from the HMM state x to state y
K	Number of clusters
c_i	Centroid of cluster i
R	Sequence length
A	Probability of acceptance of a sequence by HMM
Acronym	**Expanded Form**
FDS	Fraud Detection System
HMM	Hidden Markov Model
SP	Spending Profile
hs, ms, ls	High spending, Medium spending, Low spending
TP, FP	True Positive

IV. USE OF HMM FOR CREDIT CARD FRAUD DETECTION.

An FDS runs at a credit card supplying bank. Each incoming transaction is submitted to the FDS for authentication. FDS obtains the card specifics and the value of purchase to authenticate whether the transaction is unaffected or not. The types of goods that are bought in that transaction are not recognized to the FDS. It tries to find any irregularity in the transaction built on the expenditure profile of the cardholder, and billing address, etc. If the FDS authorizes the transaction to be affected, it raises an alarm, and the supplying bank declines the transaction. The anxious cardholder may then be telephoned and warned about the possibility that the card is imperilled. In this section, we describe how HMM can be used for fraud detection. A set of symbolizations used in the paper is given in Table 1.

HMM Model for Credit Card Transaction Processing

To record the credit card transaction handling operation in terms of an HMM, is started by determining the observation symbols in our model. We quantize the purchase values x into M amount ranges V1, V2...V_M, creating the observation symbols at the supplying bank. The real value range for each symbol is configurable grounded on the expenditure habit of distinct cardholders. These values can be determined by applying a clustering algorithm on the values of respective cardholder's transactions. We use V_k, k = 1, 2, 3 ... M, to represent both the observation symbol, as well as the corresponding value range. Here, we consider only three price ranges, namely, low (l), medium (m), and high (h). Our set of observation symbols is, V = {l, m, h} making M = 3. For example, let l = (0, Rs100], m = (Rs100,

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Rs500], and h = (Rs500, credit card limit]. If a cardholder makes a deal of $190, then the equivalent observation symbol is m. A credit cardholder makes dissimilar kinds of procurements of diverse sums over a period of time. One possibility is to deliberate the classification of transaction amounts and look for aberrations in them. However, the sequence of categories of buying is more constant compared to the sequence of transaction sums. The purpose is that, a cardholder makes purchases depending on his need for acquiring diverse types of items over a period of time.

This, in turn, produces a sequence of transaction sums. Each individual transaction sum typically depends on the matching type of acquisition. Hence, we consider the transition in the type of purchase as state transition in our model. The type of each buying is linked to the line of business of the conforming business. This data about the merchant's line of business is not acknowledged to the supplying bank running the FDS. Thus, the type of buying of the cardholder is secreted from the FDS. The set of all imaginable types of purchase and, the set of all thinkable lines of business of merchants forms the set of hidden states of the HMM. It should be noted at this stage that the line of business of the merchant is known to the acquiring bank, since this information is well-appointed at the time of registration process of a merchant. Also, some merchants may be trading in several types of commodities. Such types of line of business are considered as Miscellaneous, and we do not try to determine the authentic types of items purchased in these transactions. Any guess about availability of this material with the supplying bank and, hence, with the FDS, is not practical and would not have been operative. We show the outcome of the choice of the number of states on the system routine. After deciding the state as well as symbol illustrations, subsequent step is to determine the probability matrices A, B and Gamma so that illustration of the HMM is complete. These three model strictures are determined in a training stage by means of the Baum-Welch algorithm. The preliminary choice of strictures affects the routine of this algorithm so, they should be chosen judiciously.

We study the special case of completely connected HMM in which each state of the model can be reached in a sole step from every other state, as shown in Fig. 1. Gr, El, Mi, etc., are names given to the states to represent buying types like, Electronic items, and miscellaneous purchases. Expenditure profiles of the specific cardholders are used to gain a preliminary approximation for probability matrix B of (2). We describe how to regulate observation symbols dynamically from a cardholder's transactions.

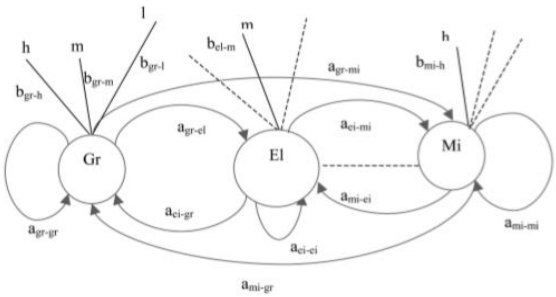

Fig.1: HMM for credit card fraud detection.

Dynamic Generation of Observation Symbols

For each cardholder, we train and preserve an HMM. To find the observation symbols equivalent to specific cardholder's businesses dynamically, we run a clustering algorithm on his history of transactions. Generally, the connections that are kept in the supplying bank's record contain numerous characteristics. We consider only the sum that the cardholder expended in his transactions. Although many clustering techniques could be used, we use K-means clustering algorithm to determine the clusters. K-means is an unsupervised learning algorithm for grouping a given set of data built on the resemblance in their attribute (often called feature) values. Each group formed in the process is called a cluster. The number of clusters K is fixed a priori. The grouping is achieved by lessening the sum of squares of distances between each data point and the centroid of the cluster to which it fits. In our work, K is the number of observation symbols M. Let c_1, c_2 …. c_M, be the centroids of the generated clusters. These centroids or mean values are used to decide the observation symbols when a new transaction comes in. Let x be the amount spent by the cardholder u in transaction T. FDS generates the observation symbol for x (denoted by O_x) as follows:

$$O_x = V_{\arg\min i |x-ci|}.$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

As cited before, the number of symbols is 3 in our system. Considering M = 3, if we perform K-means procedure on the instance transactions in Table 2, we get the clusters, as shown in Table 3, with c_l, c_m, and c_h as the respective centroids. It may be noted that the amounts 5, 10, and 10 have been clustered together as c_l resulting in a centroid of 8.3. The percentage (p) of total number of transactions in this cluster is thus 30 percent.

Table II: Example Transactions with the dollar amount spent in each transaction.

Transact no.	1	2	3	4	5	6	7	8	9	10
Dollar amount	50	35	25	53	15	30	2	30	20	90

Similarly, amounts 15, 15, 20, 25 and 25 have been grouped in the cluster cm with centroid 20, whereas sums 40 and 80 have been gathered together in cluster c_l. c_m and c_h, thus, contain 50 percent and 20 percent of the total number of dealings. When the FDS obtains a transaction T for this cardholder, it measures the distance of the purchase amount x with respect to the means c_l, c_m, and c_h to choose the cluster to which T belongs and, hence, the conforming observation symbol. As an example, if x = Rs10, then in Table 3 using (9), the observation symbol is V_1 =l.

Spending Profile of Cardholders

The expenditure profile of a cardholder advises his usual expenditure behaviour. Cardholders can be generally branded into 3 groups based on their expenditure behaviours, namely, high-spending (hs) group, medium-spending (ms) group, and low-spending (ls) group. Cardholders who belong to the high spending (hs) group, generally use their credit cards for purchasing high-priced things. Likewise, definition applies to the other two categories also. Spending profiles of cardholders are determined at the end of the clustering step. Let pi be the percentage of over-all amount of transactions of the cardholder that fit to cluster with mean ci. Then, the expenditure profile (SP) of the cardholder u is determined as follows:

$$SP(u) = \arg\max_i (p_i).$$

Model Parameter Estimation and Training

We use Baum-Welch algorithm to estimate the HMM parameters for each cardholder. The algorithm starts with an initial estimate of HMM parameters A, B, and gamma and converges to the nearest local maximum of the likelihood function. Initial state probability distribution is considered to be uniform, that is, if there are N states, then the initial probabilityofeachstateis1=N. Initial guess of transition and observation probability distributions can also be considered to be uniform. However, to make the initial guess of observation symbol probabilities more accurate, expenditure profile of the cardholder, as determined and is taken into account. We make three sets of preliminary probability for observation symbol generation for three spending groups - ls, ms, and hs. Based on the cardholder's spending profile, we select the corresponding set of preliminary observation probabilities. The preliminary guess of symbol generation probabilities using this method leads to precise learning of the model. Since there is no a priori information about the state transition probabilities, we consider the initial guesses to be unvarying. In case of a joint work between an obtaining bank and a supplying bank, we can have better initial guess about state transition probabilities as well.

We now start training the HMM. The training algorithm has the subsequent steps: 1) initialize the HMM strictures, 2) advancing procedure, and 3) backward procedure. For training the HMM, we adapt the cardholder's transaction sum into observation symbols and form sequences out of them. At the end of the training phase, we get an HMM equivalent to each cardholder. Since this phase is completed offline, it does not disturb the credit card transaction processing routine, which requires online answer.

Table III: Notations and Acronyms

Cluster mean / Centroid name	c_l	c_m	c_h
Observation symbol	V_1=l	V_2=m	V_3=h
Mean Value	8.3	20	60
%age of total transactions (p)	30	50	20

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Fraud Detection

After the HMM parameters are cultured, we take the symbols from a cardholder's training data and create an early sequence of symbols. Let O1, O2 … O_R be one such sequence of extent R. This recorded sequence is melded from the cardholder's dealings up to period t. We input this arrangement to the HMM and calculate the possibility of acceptance by the HMM. Let the probability be α_1, which can be written as follows:

$$\alpha_1 = P(O_1, O_2, O_3 \ldots. Or|\lambda).$$

Let O_{R+1} be the symbol produced by a new transaction at time t + 1. To form alternative sequence of length R, we drop O_1 and assign O_{R+1} in that sequence, generating O_2, $O_3 \ldots O_{R+1}$ as the new sequence. We input this new sequence to the HMM and compute the probability of acceptance by the HMM. Let the new possibility be α_2,

$$\alpha_2 = P(O_2, O_3, O_4 \ldots. O_{R+1}|\lambda),$$

$$\text{Let } \Delta\alpha = \alpha_1 - \alpha_2.$$

If $\Delta\alpha > 0$, it means that the new sequence is recognized by the HMM with low probability, and it could be a scam. The newly added transaction is determined to be untrue if the percentage disparity in the probability is above the threshold value. The threshold value can be learned empirically. If O_{R+1} is malicious, the supplying bank does not support the transaction, and the FDS discards the symbol. Otherwise, O_{R+1} is added in the sequence permanently, and the new sequence is used as the base sequence for determining the validity of the subsequent transaction. The reason for including new non malicious symbols in the sequence is to capture the changing spending behaviour of a cardholder. Fig. 2 shows the complete process flow of the proposed FDS. FDS is divided into two parts-one is the training module, and the other is discovery. Training stage is achieved offline, whereas detection is an online method.

V. RESULTS

Challenging credit card FDSs using real data set is a tough task. Banks do not, approve to share their information with investigators. There is also no standard dataset obtainable for investigation. We have, made large-scale mock-up studies to test the effectiveness of the system. A simulator is used to produce a mix of frank and deceitful transactions. The number of fraudulent transactions in a given length of mixed transactions is normally distributed with a user specified (mean) and (standard deviation), taking cardholder's expenditure behaviour into account specifies the average number of fraudulent transactions in a given transaction mix. In a typical scenario, a supplying bank and hence, its FDS receives a large number of genuine transactions sparingly mingled with fake dealings. The genuine transactions are produced bestowing to the cardholders' profiles. The cardholders are classified into three classes as cited before the low, medium, and hs collections. We use standard metrics—True Positive (TP) and FP, as well as TP-FP spread and Accuracy metrics, as planned in to quantify the efficiency of the system. TP denotes the portion of fraudulent transactions correctly identified as fraudulent, whereas FP is the fraction of unaffected transactions identified as deceitful. Most of the design choices for FDS that effect as higher values of TP, also cause FP to surge. To eloquently seize the performance of such a system, the variance between TP and FP, often called the TP-FP spread, is used as a metric. Precision denotes the fraction of whole number of transactions (both genuine and fraudulent) that have been sensed appropriately. It can be articulated as follows:

$$\text{Accuracy} = \frac{\text{No.of good trans.detected as good} + \text{No.of bad trans.detected as bad}}{\text{Total No.of trans.}}$$

We first carried out a set of experiments to determine the correct combination of HMM design parameters, namely, the number of states, the sequence length, and the threshold value. Once these parameters were decided, we performed comparative study with another FDS.

Choice of Design Parameters

Since there are 3 parameters in an HMM, we need to fluctuate one at a time keeping the other 2 fixed, thus generating a large number of possible combinations. For selecting the design parameters, we generate transaction sequences using 95 percent low value, 3 percent medium value, and 2 percent high value transactions. The reason for using this mix is that it represents a profile that strongly resembles a customer profile. We also consider the μ and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

sigma values to be 1.0 and 0.5, respectively. This is chosen so that, on the average, there will be 1 fraudulent transaction in any incoming sequence with some scope for variation. After the parameter values are fixed, we will see how the system performs as we vary the profile and the mix of fraudulent transactions.

For parameter selection, the sequence length is varied from 5 to 25 in steps of 5. The threshold values considered are 30 percent, 50 percent, 70 percent, and 90 percent. The number of states is varied from 5 to 10 in steps of 1. We consider both TP and FP for deciding the optimum parameter values. Thus, there are a total of 120 (5 * 4 * 6) possible combinations of parameters. The number of simulation runs required for obtaining results with a given confidence interval (CI) was derived as follows:

An initial set of five simulation runs, each with 100 samples, was carried out to estimate the mean and standard deviation of both TP and FP for a fixed sequence length, number of states, and threshold value. Mean TP was found to be an order of magnitude higher than mean FP. Standard deviation of TP was 0.1 and that for FP was 0.005. We set the target 95 percent CI for TP and FP, respectively, as +/- 2.5 percent and +/- 0.25 percent around their mean values. Using Student's t-distribution, the minimum number of simulation runs required for obtaining desired CI for TP was derived as 83 and that for FP as 23. Based on these observations, we set the number of simulation runs for all the experiments to be 100. The results obtained were within the desired CI.

VI. CONCLUSION

In this paper, we have proposed an application of HMM in credit card fraud detection. The different steps in credit card transaction processing are represented as the underlying stochastic process of an HMM. We have used the ranges of transaction amount as the observation symbols, whereas the types of item have been considered to be states of the HMM. We have suggested a method for finding the spending profile of cardholders, as well as application of this knowledge in deciding the value of observation symbols and initial estimate of the model parameters. It has also been explained how the HMM can detect whether an incoming transaction is fraudulent or not. Experimental results show the performance and effectiveness of our system and demonstrate the usefulness of learning the spending profile of the cardholders. Comparative studies reveal that the Accuracy of the system is close to 80 percent over a wide variation in the input data. The system is also scalable for handling large volumes of transactions.

REFERENCES

[1] "Global Consumer Attitude towards On-Line Shopping,"
[2] D.J. Hand, G. Blunt, M.G. Kelly, and N.M. Adams, "Data Mining for Fun and Profit," Statistical Science, vol. 15.
[3] "Statistics for General and On-Line Card Fraud,"
[4] S. Ghosh and D.L. Reilly, "Credit Card Fraud Detection with a Neural-Network,"
[5] M. Syeda, Y.Q. Zhang, and Y. Pan, "Parallel Granular Networks for Fast Credit Card Fraud Detection,"
[6] S.J. Stolfo, D.W. Fan, W. Lee, A.L. Prodromidis, and P.K. Chan, "Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results"
[7] S.J. Stolfo, D.W. Fan, W. Lee, A. Prodromidis, and P.K. Chan, "Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project,"
[8] E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A Neural Network Based Database Mining System for Credit Card Fraud Detection," Proc. IEEE/IAFE: Computational Intelligence for Financial Eng., pp. 220-226, 1997.
[9] M.J. Kim and T.S. Kim, "A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection," Proc. Int'l Conf. Intelligent Data Eng. and Automated Learning, pp. 378-383, 2002.
[10] W. Fan, A.L. Prodromidis, and S.J. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," IEEE Intelligent Systems, vol. 14, no. 6, pp. 67-74, 1999.
[11] R. Brause, T. Langsdorf, and M. Hepp, "Neural Data Mining for Credit Card Fraud Detection," Artificial Intelligence
[12] C. Chiu and C. Tsai, "A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection,"
[13] C. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-Based Fraud Detection Research,"
[14] S. Stolfo and A.L. Prodromidis, "Agent-Based Distributed Learning Applied to Fraud Detection,"
[15] C. Phua, D. Alahakoon, and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data," ACM SIGKDD Explorations Newsletter, vol. 6.
[16] V. Vatsa, S. Sural, and A.K. Majumdar, "A Game-theoretic Approach to Credit Card Fraud Detection,"