International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 1, January- 2013

A Predictive Approach for Fraud Detection Using Hidden Markov Model

S. Esakkiraj
PG scholar, Department of IT,
National Engineering College, Kovilpatti,
Tamilnadu, India

S. Chidambaram
Asst.Professor, Department of IT,
National Engineering College, Kovilpatti,
Tamilnadu, India

Abstract

Today world online banking service is the most popular and provides a fast and easy way to make transactions. As increasing online transaction, the number of fraud transaction is also increased by various thefts. In this paper, we design a model with sequence of operations in online transaction by using hidden markov model (HMM) and decides whether the user act as a normal user or fraud user. The HMM is initially trained with customer's last few transactions. In the trained system, the new transaction is evaluated with transition and observation probability. Depends upon the observation probability, system finds the acceptance probability and decides the transaction will be declined or not. Normally existing fraud detection system for online banking will detect the fraudulent transaction after completion of the transaction. It is additional burden to find fraudulent after transactions. This causes the economic loss and makes the bank name as unsecured. Our model predicts the fraudulent during the transaction time and prevents the money transfer. The main objective is to ensure that the genuine transactions should not be rejected.

Index Terms

Online banking –clustering-Hidden Markov Model-Fraud detection

1. Introduction

Data mining is a powerful tool to help organizations to extract hidden predictive information from large databases. Now days, various data mining tools allowing business to make proactive, knowledge driven decisions. There are different stages of data mining techniques to complete the process of data mining and the organization gains its full benefits. The clustering is one of the data mining techniques for dividing a dataset into groups. The member in each group has similarities and other groups are different as possible. Related records are grouped together on basis of having similar values for attributes. There are various algorithms used for clustering the records, calculating the measure and reassigning records until calculated measures don't change, that is stable segments.

Net banking is a process over the internet to make the banking process effectively. The bank automatically updates the customer accounts and records automatically. E-commerce applications are now widely used

by people and also various companies offer their services through these applications for improving their business.

2. Various Fraud Techniques

There are various ways that fraudsters execute an online fraud. By using various technologies they can do fraudulent activities.

2.1 Identity Theft

The most widely defined online banking fraud is the identity theft; gets the most attention from the customers. Identity theft can be difficult to find victims. To predict the theft, take months or even years to correct the fault. It is one of the easiest methods to get the card holder's information.

2.2 Phishing

It is one of the threats involves using bogus emails or websites. The word "phishing" means from combining the words "password" and "fishing". Fraudulent send emails that appear to be from the customer's bank that direct customers to a fake website. This website impersonates the bank's website and prompts customers for their account access data.

2.3 Trojan horse

It is one of the computer virus type software program stored on the customer's PC. Trojans records the keyboard driver and keystrokes. Once a Trojan detects that the customer opens an online banking website, it captures login name and password, and sends it to the criminal.

2.4 Internal Fraud

Banking sector allows their employees to access customer data. The data is the same information needed to access online banking to customer accounts. So that an employee can easily commit fraud. Instead of this, financial

institutions should require a password or PIN for net banking, and the password or PIN should be stored in the format of encrypted.

3. Literature review

Abhinav Srivastava et al describe the "Credit card fraud detection method by using Hidden Morkov Model (HMM)", [7]. In this paper, they model the sequence of operations in credit card transaction processing using a Hidden Markov Model (HMM) and show how it can be used for the detection of frauds. An HMM is initially trained with the normal behavior of a cardholder.

Sushmito Ghosh and Douglas L.Reilly et al describes the "Credit card fraud detection With Neural Network", [11].In this paper they using data from a credit card issuer, a neural network based fraud detection system was trained on a large sample of labeled credit card account transactions and tested on a holdout data set that consisted of all account activity over a subsequent two-month period of time. The neural network was trained on examples of fraud due to lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud and NRI (non-received issue) fraud. The network detected significantly more fraud accounts (an order of magnitude more) with significantly fewer false positives (reduced by a factor of 20) over rule based fraud detection procedures.

Sunil S Mhamane et al describes the "Use of Hidden Markov Model as Internet Banking Fraud Detection", [1]. In this paper they explained about how Fraud is detected using Hidden Markov Model also care has been taken to prevent genuine Transaction should not be rejected by making use of one time password which is generated by server and sent to Personal Mobile of Customer.

Raghavendra Patidar and Lokesh Sharma describe the "Credit Card Fraud Detection Using Neural Network" [5]. In this paper they will try to detect fraudulent transaction through the neural network along with the genetic algorithm. As they explains artificial neural network when trained properly can work as a human brain, though it is impossible for the artificial neural network to imitate the human brain to the extent at which brain work, yet neural network and brain, depend for there working on the neurons, which is the small functional unit in brain as well as ANN.

Peter Burns, Anne Stanley describes the"Fraud Management in the Credit Card Industry", [10]. In this paper they explain the Internet and the anonymity associated with card not present transactions present unique fraud management challenges. Authentication of the cardholder is a fundamental requirement in managing fraud on the Internet and there are no universally accepted solutions. As a result, credit card fraud on the Internet is substantially greater than in the physical, or even, phone environments.

Sung-Bae Cho and Hyuk-Jang Park describes the "Efficient anomaly detection by modeling privilege flows using hidden Markov model", [13].In this paper Anomaly detection techniques have been devised to address the limitations of misuse detection approaches for intrusion detection with the model of normal behaviors. A hidden Markov model (HMM) is a useful tool to model sequence information, an optimal modeling technique to minimize false-positive error while maximizing detection rate.

Pankaj Richhariya describes "A Survey on Financial Fraud Detection Methodologies", [2].The paper details as follows. Owing to levitate and rapid escalation of E-Commerce, cases of financial fraud allied with it are also intensifying and which results in trouncing of billions of dollars worldwide each year.

4. Clustering Procedure and Its Types

Clustering is a process of data into groups of similar objects. Various clustering methods available to group the dataset and each of them may give different grouping results. The choice of a particular method will depend on the type of output desired.

- Hierarchical Agglomerative methods
- Partitioning Methods
- The Single Link Method (SLINK)
- The Complete Link Method (CLINK)
- The Group Average Method.

4.1 Hierarchical Agglomerative methods

The hierarchical agglomerative clustering methods are most commonly used. The steps of an hierarchical agglomerative clustering by the following steps,

1. Find the two closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains , return to step 2

4.2 Partitioning Methods

The partitioning methods has a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster.

The partition method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

4.3 The Single Link Method (SLINK)

The single link method is the well known of the hierarchical methods and operates by combining, at each step, the two most similar objects, which are not yet in the same cluster. The name *single link* thus refers to the joining of pairs of clusters by the single shortest distance between them.

4.4 The Complete Link Method (CLINK)

The complete link method is most likely to the single link method except that it uses the least similar pair between two clusters to define the similarity between objects (so that every cluster object is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

5. K-Means Clustering Technique

K-Means clustering is a simple and efficient method to cluster the data. Initially, we determine the number of cluster K, and centroid values. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the three steps below [16].

This technique is a nonhierarchical method initially takes the number of objects equal to the final required number of clusters. In this step itself the final ‘k’ number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each object in the population and assigns it to one of the clusters depending on the Euclidean distance. The centroid's position is recalculated every time an object is added to the cluster and this continues until all the objects are grouped into the final required number of clusters.

Iterate until *stable* (= no object move group):

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

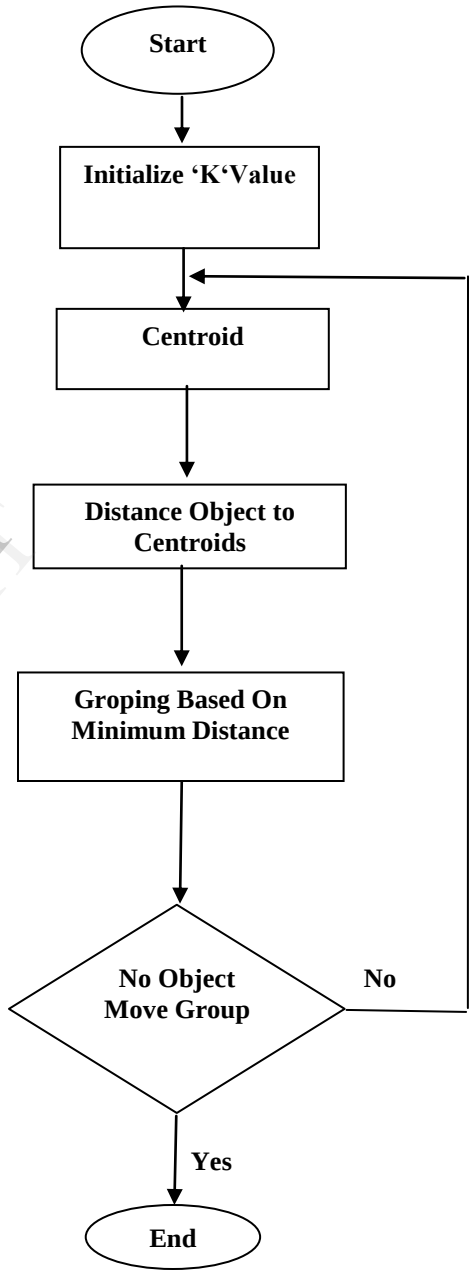

Fig 1: Block diagram of K-Means clustering Process

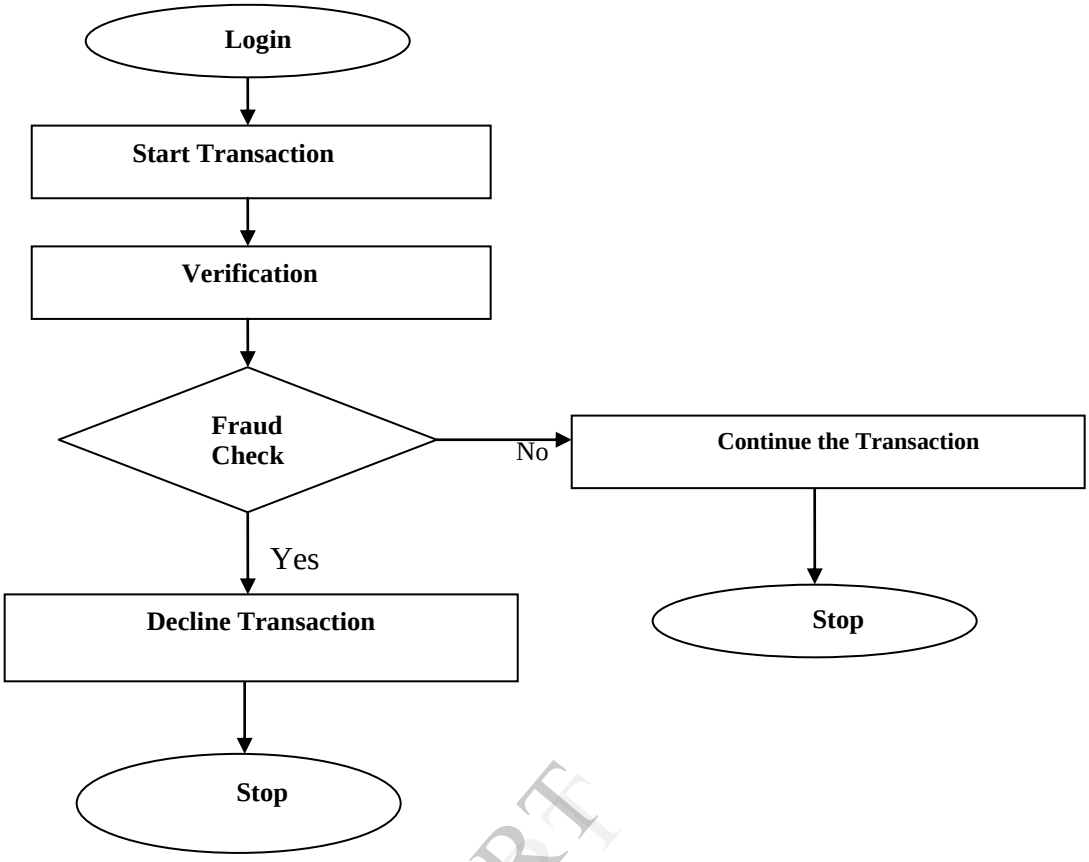

Fig 2. Block diagram of the predictive model

5. Hidden Markov Model Approach

A hidden Markov model (HMM) is a statistical model with set of states and each state is linked with a probability distribution. Transitions among these states are governed by a set of probabilities called transition probabilities. In a Markov process, the state is directly visible to the observer. In a hidden Markov model, the state is not directly visible.

In a particular state a possible outcome can be generated which is associated symbol of observation of probability distribution. It is only the outcome; so that the model named as Hidden Markov Model. Hence, Hidden Markov Model is a correct solution for detecting fraud transaction through online process.

The main feature of the HMM-based model is reducing in False Positive (FP) transactions predict as fraud by a fraud detection system even though they are really genuine customer. [7]In this prediction process, HMM consider mainly three price value ranges such as 1) Low (l), 2) Medium (m), 3) High (h).

6. Proposed System

6.1 Purpose

In this paper we focuses on

- To Design the predictive model for the fraud detection using HMM.
- Analyze the performance of the model.

6.2 Module Description

This application consists of the following modules:

Phase 1: Database development

Phase 2: Cluster Formation

Phase 3: Fraud detection

6.2.1 Database development

Create an application for online ticket reservation. The application that contains mainly home page and reservation page. Home page provides the user to know about the train schedule, Trains between the stations and special train details by separate hyper links. It also validates the user by verifying user name and password and allows registering a new user by registration form. All these information displayed by different web pages. After entering username and password details in the home page the server validates the information and redirect to reservation page. The reservation page that contains the reservation form with details such as train number, class, amount, and passenger details (name, age, sex, mobile number). After filling the reservation form the user selects the bank to pay the amount through online (The user must have the internet banking facility and have the valid username and password provided by the bank). Finally select the pay option which directs into bank website. In bank home page validates the username and password by checking in the bank database. If the information provided by the user is correct then it directs final validation page for pin verification. After the pin verification it moves to the fraud check module.

6.2.2 Cluster formation

The second module of the project to cluster the customer accounts into low, high, middle depending upon the spending profile of the customer. The clustering process based on the K-means clustering

Fig 3. Block diagram for clustering process

Euclidean Distance is the most common use of distance. In most cases will refer to Euclidean distance. Euclidean distance or simply 'distance' examines the root of square differences between coordinates of a pair of objects.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik}-x_{jk})^2}$$

d_{ij} = Distance between coordinates

6.2.3 Fraud detection

To implement the fraud detection by using the Hidden Morkov Model (HMM). we uses the states and transition probability between them. The outcome of the HMM will be the observation symbols (Oi).The observation symbols are the low, high, medium depends on the transaction amount. It can be determined by clustering process.

After the HMM parameters are learned, we take the symbols from a cardholder's training data and form an initial sequence of symbols. Let $O_1,O_2, \ldots O_R$ be one such sequence of length R. This recorded sequence is formed from the cardholder's transactions up to time t. We input this sequence to the HMM and compute the probability of acceptance by the HMM. Let the probability be α1, which can be written as follows $\alpha 1 = P(O_1,O_2,O_3, \ldots O_R|\lambda)$

Let O_{R+1} be the symbol generated by a new transaction at time t+1. To form another sequence of length R, we drop O_1 and append O_{R+1} in that sequence, generating $O_1,O_2,O_3 \ldots O_R,O_{R+1}$ as the new sequence. We input this new sequence to the HMM and calculate the probability of acceptance by the HMM. Let the new probability be $\alpha 2 = P(O_2,O_3,O_4 \ldots O_{R+1}|\lambda)$ Let $\Delta\alpha=\alpha 1-\alpha 2$ If $\Delta\alpha$ value is greater than threshold value then the transaction could be a fraud with low probability. If the transaction be a fraud then the model asks the user to answer the security questions (after ten transactions).The customer answers were checked by the FDS (Fraud Detection System) and decides whether the user be the genuine user or the fraud user. If the customer be a genuine the ticket will reserved otherwise the transaction will be declined.

7. Experimental Results

By using the net beans IDE we can create the JSP web pages and for clustering customer profile we use the weka data mining tool. Create the web pages for the bank customer entry and validate the user by the pin verification page. After the system has to be trained, every transaction sequence compared with the previous transaction sequence and the model predicts whether the user be the genuine or fraud..

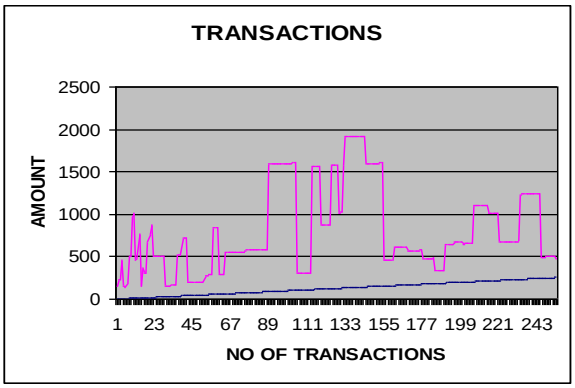

Fig 4.Customer transactions

Initially train the system with 10 transactions of each customer.Fig 4 displays the transactions

Cluster the customer profiles into low, high, middle as cluster0, cluster1 and cluster2 by using weka.

In weka tool load the dataset and cluster the datas as follows in fig 5.

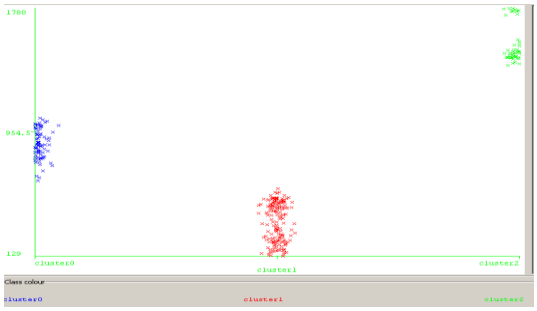

Fig 5.Cluster assignments

Performance of the model depicted as number of true positive (represent as 1) rate and number of false positive rate (represent as 0) as in Fig 6

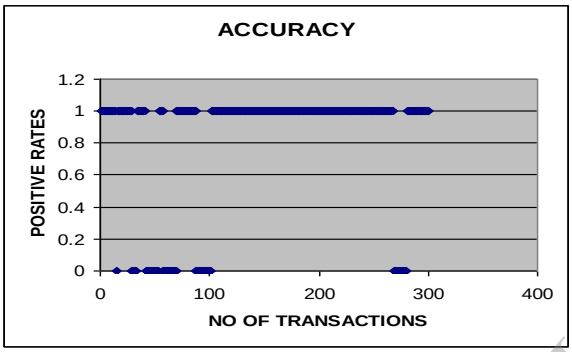

Fig 6.Performance chart

8. Conclusion

In this paper different steps of online transaction processing are represented as the underlying stochastic process of an HMM. Here for clustering process transaction amount as single attribute and in HMM it also the observation symbols, whereas the types of item have been considered to be states of the HMM. This project focuses a method for predicting the fraud users while the transaction taking place. It has also been explained how the HMM can detect whether an incoming transaction is fraudulent or not. As future work, some effective classification algorithms instead of using clustering which can perform well for the prediction.

References

[1]. Sunil S Mhamane and L.M.R.J Lobo "Use of Hidden Markov Model as Internet Banking Fraud Detection" *International Journal of Computer Applications (0975 – 8887)* Volume 45– No.21, May 2012

[2]. Pankaj Richhariya et al "A Survey on Financial Fraud Detection Methodologies" BITS,Bhopal," *International Journal of Computer Applications (0975 – 8887)* Volume 45 No.22, May 2012

[3]. Anshul Singh and Devesh Narayan "*A Survey on Hidden Markov Model for CreditCard Fraud Detection" International Journal of Engineering and Advanced Technology (IJEAT)* ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012

[4]. SHAILESH S. DHOK "Credit Card Fraud Detection Using Hidden Markov Model" *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[5]. Raghavendra Patidar and Lokesh Sharma, "Credit Card Fraud Detection Using Neural Network" *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011

[6]. V.Dheepa and Dr. R.Dhanapal, *"Analysis of Credit Card Fraud Detection Methods", "International Journal of Recent Trends in Engineering",* Vol 2, No. 3, November 2009

[7]. Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model", *IEEE transactions on dependable and secure computing,* vol. 5, no. 1, January-march 2008.

[8]. Clifton Phua1, Vincent Lee, Kate Smith& Ross Gayler *School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia,* "A Comprehensive Survey of Data Mining-based Fraud Detection Research "(2007).

[9]. Peter Burns and Anne Stanley," Fraud Management in the Credit Card Industry*" Federal Reserve Bank of Philadelphia April 2002*

[10].S. Ghosh and D.L. Reilly, "Credit Card Fraud Detection with a Neural-Network," Proc. *27th Hawaii Int'l Conf. System Sciences: Information Systems: Decision Support and Knowledge-Based Systems,* vol. 3, pp. 621-630, 1994.

[11].Sung-Bae Cho and and Hyuk-Jang Park ,"Efficient anomaly detection by modeling privilege flows using hidden Markov model" *Department of Computer Science, Yonsei University,134 Shinchon-*dong,Sudaemoon-ku,Seoul 120-749, Korea.

[12].Falaki Sand Alese B. K. "An Update Research On Credit Card On-Line Transactions" *Department of Computer Science,Federal University of Technology, Akure, Ondo State, Nigeria.*

[13].WEKA Software, The University of Waikato.*http://www.cs.waikato.ac.nz/ml/weka/*

[14].Sapna Jain and et al , "k-means clustering using weka

interface", *Proceedings of the 4th National Conference; Computing For Nation Development*, February 25 – 26, 2010