International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

Credit Card Fraud Detection Using HMM and DBSCAN

Bharati H. N.[1], Soumya Bastikar[2], Mita Gavade[3], Sangita Samota[4]

[1]Professor, Head of Department, Department of Computer Engineering, K. J. Somaiya College of Engineering

[2, 3, 4]Student, Department, Department of Computer Engineering, K. J. Somaiya College of Engineering

Abstract: ***With the advent of cashless economy, the demand for credit cards has been rising steadily. With the increase in such transactions, fraud detection systems play a vital role. In this paper, we have modeled the operating phases in a credit card transaction processing. In the prototyped environment, the transaction detail of location is traced from the IP address. We have used two stages to detect the authenticity of a transaction: HMM algorithm, a stochastic model for sequential data, that works on amount as the parameter, and DBSCAN that works on location of the transaction as the parameter. If the transaction doesn't pass through any of these phases, the card holder is alerted via an email. We have backed the efficiency of the approach by presenting an experimental analysis of the same.***

Keywords: Hidden Markov Model, Probability, Fraud Detection System, Credit Card Transaction, DBSCAN

1. Introduction

Credit Cards are favoured for their convenience of use, security and easy tracking capabilities. Use of credit cards is not just limited to online transactions and can be used in retail transactions as well. However, fraudulent transactions can be made if the credit card is lost or stolen.

These cases of credit card frauds have been rising every year. In a country like India, where the government aims at digitising everything, the problem of credit card frauds takes a central role. Therefore means are needed to detect fraudulent transactions and stop their execution.

Neural networks, Genetic algorithms, decision trees have been used for detection of credit card frauds. While some methods lack in offering efficient results others require large computing time. There is a need to develop detection methods that give efficient results within no time as the processing of credit card transaction takes place almost instantly.

The major shortcoming of existing detection methods is that it is a one stage process. And once any transaction falls into the "possibly fraud" transaction category, the credit card is suspended at that instant. While this method almost always prevents fraud transactions to take place, it can be of great annoyance to a customer who is trying to make genuine transactions. Our proposed idea makes the process of credit card fraud detection a multi-stage process. And only after confirmation from the customer, the card is blocked.

Hidden Markov Model is a predictive model algorithm that can be applied on the amount of a transaction. Often the card thief would make transactions of small amount to check if a card is working before actually exploiting it. HMM tracks these abrupt changes in the existing sequence of transactions and detects the fraudulent ones. Also the coordinates where a transaction is taking place can be compared to history of transactions to find suspicious activity.

The originality of this paper lies in the combination of HMM, a doubly stochastic model that defines a cardholder's spending profile and works on the sequence of transactions, and DBSCAN, a density based clustering algorithm that works on locations.

2. Related Work

There are many approaches which have been used over the past few years in Fraud Detection Systems(FDS). Some of them being Decision trees, Genetic algorithms, clustering techniques, and neural networks.

The idea behind Decision tree model is that of a similarity tree created by using decision tree logic. In this case, a similarity tree is defined recursively; the nodes are labeled with the use of attribute names, edges are labeled using values of attributes, and then there are the leaves, which contain an intensity factor that is defined as the ratio of the number of transactions that satisfy the outlined conditions. The main advantage of this method of fraud detection is that it is easy to implement, understand and display. However, there are disadvantages when you are forced to check every transaction one by one. But as this method has been able to offer tangible results , it's still used as an effective method in credit card fraud detection.

Genetic Algorithm[5] is used as a predictive method in fraud detection. It establishes rules that classify credit card transactions into suspicious and non-suspicious classes. So it can be beneficially used in detecting and countering credit card frauds.

Another approach was a parallel granular neural network[8] (GNN) which is developed to speed up the data mining and knowledge discovery process for credit card fraud detection.

The system works on Silicon Graphics. The parallel fuzzy network is trained in parallel using training data sets and the system discovers fuzzy rules for future prediction. Higher the value of fraud detection error, higher is the possibility of the transaction being actually fraudulent.

As for the detection of fraudulency in transactions based on location, one of the approaches is OPTICS. In this approach, instead of fixing MinPts and the Radius just like in DBSCAN, only MinPts are fixed, and the radius at which an object would be considered dense by DBSCAN is plotted. In order to sort the objects on this plot, they are sorted in a priority heap, so that nearby objects are also positioned nearby in the plot. OPTICS comes at a higher cost as compared to DBSCAN. Largely because of the priority heap, but also as the nearest neighbor queries are more complicated than the radius queries of DBSCAN. So it will be slower, the only advantage being that the parameter epsilon doesn't need to be set.

3. Proposed System

The proposed model is a fraud detection system for credit card transactions taking place in real-time.

The system is developed as a web application. The system works in the following stages:

1) Authentication of user:
At this stage an existing user can sign into the system using his/her login credentials.
The new user can sign up for a new account to access the system.

2) Transaction details:
The user enters transaction amount in this stage that gets stored. Also, the location of the user is tracked using his/her IP address and stored in the database.

3) K-means and HMM Algorithm:
Based on the past transactions of user, his/her spending profile is generated by K-means. This spending profile along with current transaction amount is processed by HMM to determine if the current transaction is genuine or fraud. If the algorithm returns false, an email and SMS is sent to the user to confirm the transaction.

4) DBSCAN:
In the previous stage if the outcome turns out to be a genuine transaction, then location of the user is taken into consideration. DBSCAN is applied, and based on the results the transaction is committed or an alert message is sent to user.

5) Verification
If the current transaction is declared as suspicious by the FDS, an alert sms will be sent to the user's phone asking for validation of the transaction.

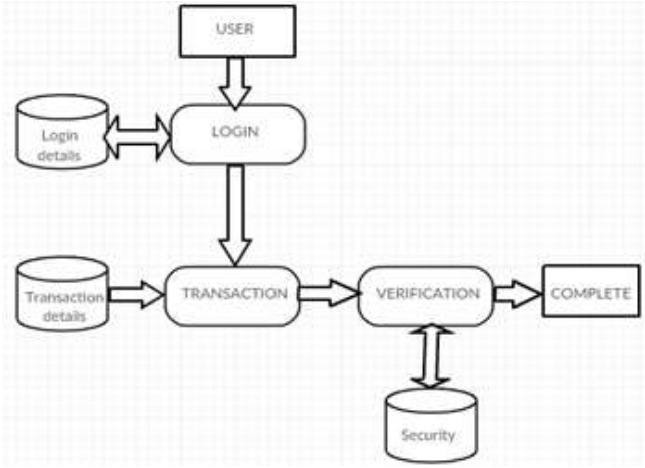


Figure 1: Multi-tier system architecture

4. Implementation Details

a) K-Means Clustering Algorithm
Clustering in statistics refers to how data is gathered. It's a process of partitioning a group of data points into a small number of clusters. It can be quantitative or qualitative kind of clustering. For example, books on Amazon.com are listed both by category (qualitative) and by best seller (quantitative). Now this was just an exemplified version of clustering.

In general, we consider having n data points xi,i=1...n that have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μi,i=1...k of the clusters that minimize the distance from the minimum distance of a data point from a cluster[2].
$\text{argmin}c\sum i=1k\sum x\in \text{cid}(x,\mu i)=\text{argmin}c\sum i=1k\sum x\in ci\| x-\mu_i$ [1]
where ci is the set of points that belong to cluster i.

The K-means clustering uses the square of the Euclidean distance $d(x,\mu i)=\| x-\mu i\|$. The K-means algorithm or Lloyd's algorithm is used to solve the k-means clustering problem. The algorithm first initializes the centre of the clusters which can be any random value at the beginning. It then attributed the closest cluster to each data point. After this, the position of each cluster is then set to the mean of all data points belonging to that cluster. These two steps are repeated until the assignments of the cluster to the data points does not change from one iteration to another.

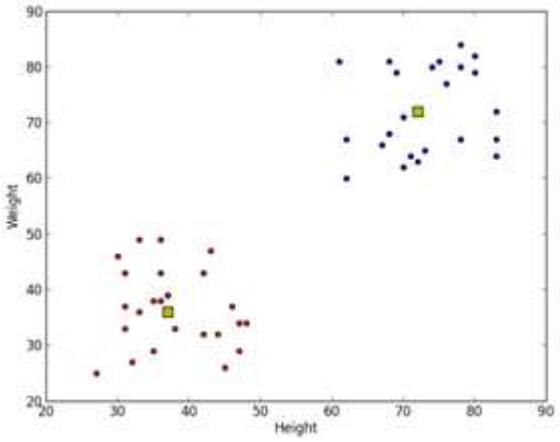


Figure 1

In Figure 1.0, we can see the data points being attributed to two different clusters based on the how close the data point is from the centroid of the cluster.

In our approach, the K-means algorithm serves as a first stage where in the user transaction history is clustered in order to generate the spending profile of the user namely high, medium and low. The Users can always be clustered into groups such as those who regularly does high cost transactions, or the ones who rarely do that. The algorithm of clustering of users based on spending behaviour must be as effective as possible to get better accuracy in output of the system and hence is considered as the backbone of the entire Fraud Detection System.

1) HMM

An HMM is a double embedded stochastic process which is much more complicated than the stochastic processes of a traditional Markov model.

The HMM based of the FDS is set up in two phases:

a) Training
b) Detection

Hmm is trained with the normal behaviour of a cardholder, with the clustered data set consisting of all legit transactions.

HMM model is defined by the specification of following parameters [6]:

i. N: number of states
ii. M: number of observation symbols
iii. Observation symbols
iv. Probability measures

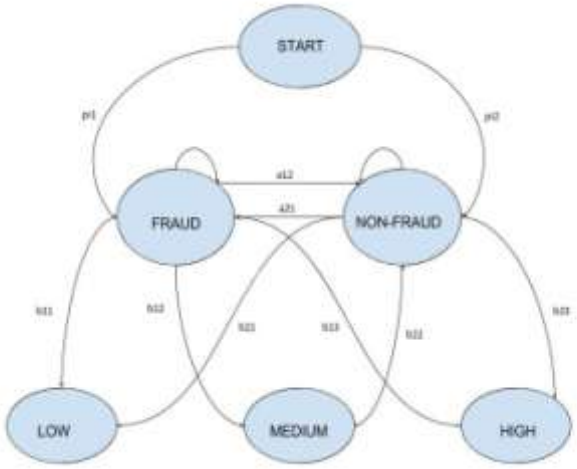
a) State transition probability
b) Emission probability distribution
c) Initial state distribution

The compact notation used to indicate the complete parameter set of the model:

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

In the training phase, it starts off with assumed values for the matrices, and then uses the Baum-Welch and forward backward algorithm to refine the probability matrices.

In Credit Card Fraud Detection systems, the hidden states would be whether the transaction is fraud or legit {F,N}, and observed would be the spending profile {H,M,L}.



We test the data for various sequence lengths of observed states. Initially, a set of say n transactions is considered, the probability alpha is calculated for the length [1].

$$\alpha_1 = P(O_1,O_2,O_3,..,O_n \mid \lambda)$$

The O_{n+1} transaction is the new sequence recorded, at the time the transaction is going to be processed. To maintain the HMM model, the first transaction is dropped and new one is added.

$$\alpha_2 = P(O_2,O_3,O_4,..,O_{n+1} \mid \lambda)$$

The difference is then compared with the threshold value which is calculated empirically.
If the difference is greater than the threshold the transaction is labelled as fraud and the cardholder gets an alerting SMS asking for confirmation of details.

If the FDS identifies the transaction as legit, the amount is added to the dataset and then the control passes to the next filter: DBSCAN.

2) DBSCAN

DBSCAN (Density based spatial clustering of applications with noise) algorithm acts as the next filter after HMM. Once the HMM returns a transaction as non-fraudulent (based on the amount), it is passed onto DBSCAN to check if the transaction is made from a location close to past transactions.

DBSCAN works by forming groups or clusters of points that are close to each other and leaving the remaining ones as noise points or outliers.

The working of DBSCAN is as follows: We have a set of points to be clustered, i.e. the x coordinates and y coordinates of the locations of all transactions. These points are classified as core points (the centres), reachable points

(radius points) and outliers (noise points).
The two important parameters used in DBSCAN are minimum points (minPts) and epsilon distance (eps).

The core point is any point having at least defined minPts at a distance of epsilon from itself. These neighbouring points are directly reachable from the core point. We also have indirectly reachable points, those which can be reached from the neighbours of the core point. Those points which are not reachable from any other point in the graph are termed as outliers.

DBSCAN initialises the clustering process with a random start point that hasn't been visited yet. This start point's neighbours at epsilon distance are calculated and stored, and if it contains sufficient number of points, a cluster starts to form. Else, the point is labeled as outlier or noise for now. Later this point may later be found to become a part of a different cluster as it might be at epsilon distance from any point in the previously formed cluster or a new cluster.

If we find a point in the dense part of a cluster, its epsilon-neighbors are also become part of that cluster. Therefore, all such points found within the epsilon distance are added. The process will continue until the density-connected cluster is formed completely.

A new unvisited point or the coordinates of the current transaction in our case is retrieved and processed, that leads us to classify it as a noise point or not. If it is a noise point then the current transaction is suspicious, else it is a valid transaction.

5. Analysis and Evaluation

To train the Hidden Markov Model the following probability matrices are formed based on the observed sequence i.e. sequence of amounts. The Hidden states are Fraud or Non-fraud.

Initial Probabilities:

Fraud	0.2
Non-Fraud	0.8

Transition Probabilities:

	Fraud	Non-Fraud
Fraud	0.67	0.33
Non-Fraud	0.1	0.9

Emission Probabilities:

	Low	Medium	High
Fraud	0.2	0.1	0.7
Non-Fraud	0.7	0.25	0.05

After Applying Baum-Welch Algorithm:

Initial Probabilities:

Fraud	0.03
Non-Fraud	0.97

Transition Probabilities:

	Fraud	Non-Fraud
Fraud	0.56	0.44
Non-Fraud	0.11	0.89

Emission Probabilities:

	Low	Medium	High
Fraud	0.18	0.02	0.8
Non-Fraud	0.84	0.15	0.01

Since the datasets in credit card transactions are confidential due to security reasons, we are working on a generated dummy dataset of 40 values for a user having maximum transactions in the low amount category. The Hidden Markov Model system is trained on the first 30 amount values and then it is tested on the remaining 10 amounts. Out of these 10 values, system returns the expected output for 9 values. For this small dataset, system provides 90% efficiency.

Amount	HMM Result (Probability Based)	Expected Result
4000	Non-fraud	Non-fraud
1050	Non-fraud	Non-fraud
700	Non-fraud	Non-fraud
2000	Non-fraud	Non-fraud
5	Non-fraud	Fraud
25000	Fraud	Fraud
50000	Fraud	Fraud
1000	Non-fraud	Non-fraud
400	Non-fraud	Non-fraud
600	Non-fraud	Non-fraud

6. Conclusion and Future Scope

In this paper, we have proposed an application of HMM and DBSCAN algorithm for detecting fraud in credit card transactions.

Till now, approaches such as Genetic Algorithm, Neural networks or only HMM had been used for detecting fraud. We propose a system that takes two parameters into consideration namely amount and location which serve as inputs for HMM and DBSCAN algorithms respectively. The user's past transaction amount serve as a data set for training the HMM and is also used for finding the spending profile of cardholders with the help of K-Means Clustering algorithm. The DBSCAN algorithm uses the location of the transaction being performed and detects fraudulency in case of any

anomaly. If fraudulency is detected the users are alerted while the transaction is being processed, thus making the system real-time. The system is scalable for handling large volumes of transactions. Also, care has been taken to reduce the number of false positives and false negatives and generate accurate results as far as possible.

The results generated from our system were promising. However we see our system as an initial framework which has been tested on a dataset of limited records. Future work will concern improving the system by considering more parameters other than amount and location. We hope this work serves as a foundation to much more encompassing and automated systems.

7. Acknowledgement

We would like to thank the faculty and the staff of Computer Department of K. J. Somaiya College of Engineering to support us throughout the project. We also thank Prof. Bharathi H. N. for her guidance right from the beginning of the project, all the creative improvements and healthy discussions.

References

[1] IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, JAN-MARCH 2009, "Credit card Fraud detection using Hidden Markov Model",Abhinav Shrivastava,Amlan Kundu.
[2] Credit Card Fraud Detection: A Case Study, IEEE 2015, Ayushi Agrawal, Shiv kumar.
[3] Implement Credit Card Fraud Detection System Using Observation Probabilistic in HMM, Ashphak Khan; Tejpal Singh; Amit Sinhal, IEEE 2012.
[4] A Novel Approach for Credit Card Fraud Detection Amit kumar Mishra, Ayushi Aggarwal, 2015.
[5] Fraud Detection of Credit Card Payment System by Genetic Algorithm, K. RamaKalyani, D.UmaDevi, 2012.
[6] A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of IEEE 1989, Lawrence R. Rabiner.
[7] Use of Hidden Markov Model as Internet Banking Fraud Detection, International Journal of Computer Applications,May 2012, Sunil S Mhamane, L.M.R.J Lobo.
[8] Parallel granular neural networks for fast credit card fraud detection, Proceedings of IEEE International Conference, 2002.