LOCATION BASED FRAUD DETECTION IN E-COMMERCE TRANSACTIONS

THESIS REPORT

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted by:

SUCHITA JAIN
(2K16/SWE/17)

Under the supervision of

Dr. RAJNI JINDAL
(HOD CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road , Delhi – 110042

JUNE 2018

CANDIDATE'S DECLARATION

I, SUCHITA JAIN, 2K16/SWE/17 a student of M.TECH (Software Engineering) declare that the project Dissertation titled "Location Based Fraud Detection in E-Commerce Transactions" which is submitted by me to Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Fellowship or other similar title or recognition.

Place: DTU, Delhi. SUCHITA JAIN

Date: (2K16/SWE/17)

CERTIFICATE

I, hereby certify that the Project titled "Location Based Fraud Detection in E-Commerce Transactions" submitted By SUCHITA JAIN, Roll number: 2K16/SWE/17, Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

Dr. RAJNI JINDAL
(HOD CSE,DTU)
SUPERVISOR

ACKNOWLEDGEMENT

I am very thankful to Dr. Rajni Jindal (Head of the Department, Computer Science Eng. Dept.) and all the faculty members of the Computer Science Engineering Dept. of DTU. They all provided immense support and guidance for the completion of the project undertaken by me. It is with their supervision that this work came into existence.

I would also like to express my gratitude to the university for providing the laboratories, infrastructure, test facilities and environment which allowed me to work without any obstructions.

I would also like to appreciate the support provided by our lab assistants, seniors and peer group who aided me with all the knowledge they had regarding various topics.

SUCHITA JAIN
M.TECH (SWE)
2K16/SWE/17

ABSTRACT

There was a time when payment was made in the forms of cash, barters, checks and other means. With the advent and rise of Internet, people have shifted from conventional forms of payment to online and digital forms of payment. It has become rather convenient to make payments online so as to avoid tensions to keep money always in pockets.

As more and more payments and digital transactions have appeared online, it has attracted unwanted audience or so called scammers, hackers and threats to eavesdrop on this. Be it MITM attack or skimming or any other form of hacking attempt, these persons keep on developing newer methods for hacking.

Fraud detections and risk prevention has become the foremost step to prevent frauds, robbery etc. Machine learning approaches have been put forth in the research and literature to develop models to stop and cease all this.

Locations based fraud detection is one of those methods so as to identify any vulnerable attempt to steal money and rob the accounts. HMM technique will be used for this purpose as it has advantage of detecting the frauds before the actual transaction is processed with low false alerts. Distance is used as a perspective to detect the frauds. This perspective is helpful in detecting those frauds which were not detected when amount was used as an aspect.

CONTENTS

CANDIDATE'S DECLARATION i

CERTIFICATE ii

ACKNOWLEDGEMENT iii

ABSTRACT iv

CONTENTS v

LIST OF TABLES vii

LIST OF FIGURES viii

LIST OF ABBREVIATIONS ix

CHAPTER 1 INRODUCTION 1

1.1 Literature Survey 3

1.2 Problem Statement 7

1.3 Objective 9

1.4 Organization 10

CHAPTER 2 TECHNIQUES AND TERMINOLOGIES 11

2.1 Outlier 11

2.2 Fraud 11

2.3 K-Means Clustering 12

2.4 Hidden Markov Model 15

2.5 Baum-Welch Detector 18

CHAPTER 3 IMPLEMENTATION DETAILS 21

3.1 Data Set Used 21

3.2 Location 21

3.3 Clustering 22

3.4 Training 23

3.5 Detection 24

CHAPTER 4 RESULTS AND EVALUATION 26

4.1 Basic Terminologies 26

4.2 Results of a customer 28

4.3 Overall Results 32

CHAPTER 5 CONCLUSION AND FUTURE WORK 35

5.1 Conclusion 35

5.2 Future Work 36

REFERENCES 37

LIST OF TABLES

2.1 Data values for K –means algorithm…….....…14

2.2 Different clusters using K-means algorithm.....14

4.1 Overall Results.....34

LIST OF FIGURES

2.1 K-means clustering method algorithm…….....13

2.2 K-means clustering method…….15

2.3 HMM states…….....16

2.4 Training phase of HMM…….....17

2.5 Testing Phase of HMM……18

2.6 States and symbols transitions…….....19

3.1 Fraud Detection System using HMM…….…….25

4.1 Output classes……..27

4.2 Percentage of each distance cluster…….29

4.3 Results when distance values are given as input…….....30

4.4 Percentage of each spending cluster. ……31

4.5 Results when amount values are given as input…….32

4.6 Overall Results when distance values are given as input.....33

4.7 Overall results when amount values are given as input. 33

LIST OF ABBREVIATIONS

ML	Machine Learning
KNN	K-Nearest Neighbour
HMM	Hidden Markov Model
FDS	Fraud Detection System
GA	Genetic Algorithm
OTP	One Time Password
MITM	Man In The Middle
MITB	Man In The Browser
SQL	Structured Query Language
ATM	Automated Teller Machine
RTI	Right To Information

CHAPTER 1 INTRODUCTION

As the world is heading towards the digitization, the problems linked with it are also increasing. The major issue is the frauds. Fraud is a crime in which someone illegally gains cash or goods using tricks. The credit cards utilization is increasing day by day so as the frauds linked with it. When someone is using other's credit card without informing him and with no intention to return the amount back to him, then we refer it as a credit card fraud. This use can be either physical or virtual. In the physical use, one can steal other's card and use it. Such frauds have been decreased these days. In virtual case, one just needs the credit card details of a person like the expiry date, card number. Once you have the details you can easily make transaction.

Credit card frauds have turned out to be rampant as of late. To enhance dealers' risk level in a programmed and powerful way, constructing a precise and simple taking care of card checking management framework is one of the key errands for the shipper banks. One point of this investigation is to form model that best recognizes fraud cases. There are numerous methods for identification of MasterCard fraud. On the off chance that one of these or mix of calculation is connected into bank card fraud identification framework, the likelihood of fraud exchanges can be anticipated after MasterCard exchanges by the banks. Also, a progression of hostile to fraud methodologies can be embraced to keep banks from extraordinary frauds previously and decrease risks.

It has become very important need to detect these frauds. Different techniques and methods are used to detect the frauds like data mining techniques, machine learning techniques etc. Normally we look into the amount of transaction to detect for the frauds and in that normally the transactions with small amount are not taken into consideration .But sometimes these small transaction can also be a fraudulent transaction. Now we will look into the other aspect other than the amount of transaction to detect the fraud which is the location.

Generally the location of these cyber criminals is not close to the person's place. It has been seen that these cyber criminals have their branches in the metropolitan cities from where they conduct their illegal activities.

According to the report by RTI, a total value of 21,860 debit/ATM card and 35,551 frauds related with credit cards have been reported in the India from January 2015 to December 2017. The amount involved in these frauds is Rs14, 869 lacs for ATM/debit card and Rs 14,165 lacs for credit cards.

There exist diverse methods with which fraudsters executes the credit card frauds. As innovation changes, so does the procedures of fraudsters with which they execute frauds. Frauds can be comprehensively arranged into three classifications, i.e. merchant related frauds, traditional card related frauds, and web frauds.

Different types of frauds classification:

- Lost/Stolen Cards: When someone steals the person's card and use it for his benefit, then that leads to the frauds. These frauds have decreased comparatively and are given less importance these days.
- Account Takeover: This type of fraud exists when a person illegally obtains a substantial client's personal data.
- Fake And Counterfeit Cards: Fraudsters are continually putting efforts into innovative and new strategies to make duplicate cards. A portion of the procedures used to make false and fake cards are recorded underneath:

  a) Creating a fake card starting with no outside help utilizing advanced machines.

  b) Skimming which includes electronically replicating information on a card's attractive stripe onto another. This can be done by erasing the attractive strip by messing with a current card that has been obtained unlawfully by deleting the metal strip with an intense electro-attractive gadget.
- Merchant Related Frauds: Merchant related fakes are those which are initiated by

either the owners of the vendor company or their representatives.

- Internet Related Frauds: The Internet gives a perfect stage to fraudsters to effortlessly confer Visa frauds. The most generally utilized procedures in web frauds are depicted beneath:

  a) False merchant sites offering clients to a lot of offers are normally part of a bigger criminal system, and demand a client's card details.

  b) Site Replicating means the fraudsters replicate the sites or simply those web pages where the client puts the order. In this way the hackers get to know the credentials of the owner.

1.1 Literature Survey

It has turned out to be essential to identify these frauds. Diverse procedures have been discussed in previous study to recognize the frauds like data mining systems, machine learning strategies and so.

The research by Sathish and Rajeshwari [2], with the developing use of the web, everything is accessible at our doorstep and accommodation. As the world is heading towards the digitization, the problems linked with it are also increasing. The major issue is the frauds. Frauds can happen due to many reasons like site cloning, lost credentials etc. Streaming Analysis is a time based collecting and pre processing of information which is further utilized for real time decision making by reviewing, connecting and analyzing the information when it is fed into databases and apps from the distinctive sources. We are making use of streaming investigation to recognize and prevent the frauds. Their investigations records the information of exchange to maintain, build up a framework that can distinguish further frauds.

The paper by Pushpa and Malini [6], proposes that credit card fraud tricks has turned out to be considerably broader. To advance security measures of the money exchange frameworks in a constant and viable way, structure an exact and well organized card fraud discovery framework is one of the important capacities for cash exchanges. By performing

oversampling and extracting the key of the information we can utilize our KNN strategy to decide the irregularity of the target data. Subsequently the KNN strategy can suit for recognizing frauds with the confinement of memory. By the mean time outlier detection mechanism helps to distinguish the charge card misrepresentation utilizing less memory and calculation prerequisites. Particularly outlier detection works quick and well on online vast datasets. But in comparison with power method techniques and other known anomaly detection strategies, exploratory outcomes reveal that the KNN strategy is precise and accurate.

The research by Bhusari and Patil [1], examines how Hidden Markov Model will help to stop fraudulent transactions through cards. The Fraud Detection System is additionally versatile for dealing with tremendous volumes of exchanges handling. The HMM based FDS isn't taking long time and having complex procedure to detect fraud like the current framework and it gives preferred and quick outcomes over existing framework. The Hidden Markov Model makes the preparing of location simple. At the underlying state HMM checks whether the upcoming exchange is deceitful or not and it will permit to acknowledge the following exchange on the likelihood result. The diverse scopes of exchange sum like low, medium, and high as the observation images were considered. The kind of the items purchased has been thought to be states of the Hidden Markov Model. It is suggested that a system for finding the spending behavioral of cardholders, likewise the utilization of this learning in choosing the estimation of observation symbols and introductory estimation of the model parameters In their proposed show, they have discovered over 84% exchanges are genuine and low false alert which is around 7 % of aggregate number of exchanges. The relative investigations and our outcomes beyond any doubt shows that the accuracy of the proposed framework is secure to 80 % over a wide deviation in the data.

According to the research paper by Riyanarto Sarno, Chastine Fatichah, Dewi Rahmawati, and Dewi Sunaryono [5], Hidden Markov Model can be utilized for computing likelihood plausibility of the fraud on the events log of the business processes in the bank credit application. The outcomes from their research show that the HMM strategy can recognize fraud fittingly and is very effective. The trial that comes about additionally demonstrates that the accuracy of the outcomes is 94%. In their research, the information access from

each event log data on the event log's of bank credit system can be examined utilizing a calculation Hidden Markov Model (HMM) keeping in mind the end goal to distinguish fraud side effects. The investigation procedure is especially helpful for identifying frauds on business forms that happen within the procedure of the Bank Credit Application.

The paper by Ankit and Chinmay [3] and by Priyanka, Pavan [7], suggested that HMM can be successfully used to detect the frauds in the credit card transactions. The HMM has low false alerts and even it can detect the fraud before the transaction is executed, so no overhead on how to get the amount of the fraudulent transaction back. This system works by considering the behavior of the user by taking into consideration the amount of the transaction as a major factor. The pattern is constructed for the amount and then when any further new transaction comes, it is compared with the pattern and then considered fraud or genuine. Hmm works on different volumes of data successfully. The accuracy of the system comes out to be nearly 80 percent.

The research by Fuzail Misarwala, Kausar Mukadam, and Kiran Bhowmick [4], audits all research methodologies in this field, Their point was to exhibit a survey that can educate the industry experts ,academicians and researchers about the flow condition of research in this field as far as parts focused on and techniques utilized. The authors grouped the writing in light of (i) Fraud type (ii) Year (iii) Data mining class, and (iv) Data mining procedure. Out of the four classifications of misrepresentation that we explored, financial fraud has pulled in the most consideration from specialists. Money related frauds is more likely to be submitted by fraudsters and influences organizations and associations of all sizes, along these lines it involves grave worry for most, so this is the reason most research has been done in this area.

According to K.Rama Kalyani and D.UmaDevi [8], genetic algorithm can be used to detect the frauds. For the evolutionary algorithms like the Genetic algorithms, the point is to get the better and ideal solutions. They examined a sample data set and fraud transactions were generated on the basis of this algorithm. In the method of electronic payment system, fraud exchanges are ascending on the general premise. The point is to build up a technique for creating test information and to recognize fraud exchange with this calculation. This calculation is a optimization method and evolutionary search based in light of the standards

of genetic and natural selection, heuristic used to take care of high complexity computational issues. This paper presents to discover the recognition of FDS and looks at the outcome in light of the standards of this calculation.

The paper by Assis and Pereira [11], proposes a system in light of genetic programming to perform fraud identification in electronic exchanges. The solutions are represented as SQL WHERE-condition, which recognize tests of one of the two classes (frauds or genuine). The appropriateness of the planned calculation has been assessed in light of famous datasets from the writing and in two genuine situations of one of the biggest Latin American electronic installment organizations. The fundamental contribution of the proposed work is the plan and execution of this strategy of genetic programming as a structure. In addition, the work has exhibited great outcomes for the extortion identification issues. Furthermore, the calculation has indicated great outcomes when it was connected to the UCI Machine Learning dataset. This reality shows the robustness and generality of the approach, which can be effectively adjusted to deal with various areas and applications.

The paper by Krishna and Reshma [9], unites different techniques to recognize fraud exchanges and examination of these strategies. One of these or mix of these techniques can be utilized to identify false exchanges. New highlights can be included and different testing techniques can be utilized to prepare the model which is more precise.

The research by Rashi and Shailendra [10], bring together all techniques and methods, in a calm arranged and planned manner for outlier detection. Outlier Detection is the real issue in Data Mining; discovering anomalies from an accumulation of examples is a mainstream topic in the field of mining of information. An outlier is that instance which is divergent from all the rest of the instances in the collection of information. Anomaly location is calm commonplace of research in information mining. It is a critical undertaking in various application areas. Previously anomalies that were considered as noisy data, has now become an extreme problem in different sections of research. The introduction of anomaly is helpful in identification of unpredictable and unidentifiable knowledge, in specific areas like fraud detection, PC interruption, intrusion detection, maintenance and criminal practices etc.

1.2 Problem Statement

Our aim is to develop the system which can detect the frauds before the transaction is processed and with a different perspective other than the amount of the transaction.

It has been seen that there is an impressive increase in the use of credit cards. All thanks to the leading e-commerce companies like Amazon, Flipkart, Walmart, Jabong etc. which had 80% of the e-commerce market share of India in 2016. This digitization has also led to frauds. For cyber criminals this has become gold mine.

According to the report by Nilson, the loses due to card related frauds rose from about US$8 billion in 2010 to US$21 billion in 2015. By 2020, this may reach US$31 billion. When somebody is utilizing other's Mastercard without educating him and with no aim to restore the sum back to him, at that point we consider it as a credit card fraud.

The utilization of charge cards is expanding step by step so as the frauds connected with it. Basically the frauds can be classified into two types: one where card is present and the other where it is not present. In the physical use, one can steal other's card and use it. Such frauds are less common these days. In the other case, the fraudsters illegally use the credit card details of a person like the expiry date, card number to make transaction.

It has become very important requirement to detect these frauds. It has turned out to be imperative to recognize these fakes. Diverse strategies are utilized to identify the cheats like information mining procedures, machine learning methods and so forth. Typically we investigate the measure of exchange to recognize for the cheats and in that regularly the exchanges with little sum are not thought about .But rather in some cases these little exchange can likewise be a false exchange.

No doubt to control frauds, OTPs is used. But they can be hacked also to make the fraudulent transaction. The only advantage of OTPs is that they can't be replayed i.e. they are for a particular time. Basically there are two types of attack linked with the use of OTPs one is MITM and the other is MITB.

A MITM attack is a dynamic eavesdropping stealthily in which the attacker consolidates a few systems to influence the two legitimate entities that they are associating specifically in a protected association when in fact the hacker controls the information being traded. A case of straightforward MITM is a mix of sniffing and spoofing. Sniffing is utilized to 'see' the information bundles that are sent to the end client by the bank. Spoofing is the demonstration of producing the client's IP address in messages sent to the bank's site.

A MITB is the major one amongst the most progressive assaults utilized by programmers. The malware (regularly a worm) does not dwell in the typical areas (stockpiling hard drive, and so on.) however introduces itself inside the program application; this is the reason it is so difficult to identify and remove. The greater part of MITBs was particular to one program and one bank's site. More up to date forms reach out to various browsers and different banks' sites. In basic terms, a browser is generally made out of a GUI (Graphic User Interface) and a engine (among different segments). The MITB comprises in controlling information trades between the GUI and the motor, showing whatever the programmer needs to the client and sending precisely what he needs to the bank.

So there is need to detect the fraudulent transactions even in the presence of OTPs. There are different strategies to detect these frauds. Even there are different attributes of the user profile to detect the frauds. In most of the cases, we just look into the value of transaction to detect for the frauds and in that generally the transactions with small amount are not taken into consideration .But sometimes these small transaction can also be a fraudulent transaction. In this proposal, our objective is to use the Hidden Markov Model to detect the digital frauds but not according to the amount of transaction but by using the location where the transaction is being processed as an aspect. The advantage of using HMM is that it can detect the frauds when the transaction is processing and it doesn’t have to wait for the transaction to get process. It also has low false alarms.

1.3 Objective

There is a need to detect the frauds linked with the credit card transactions as the loss is faced by either the merchants or the owners. The bank also faces a lot of losses due to the frauds. So there is a need to prevent these frauds to happen. For this a system should be made which can detect the frauds much before the transaction is processed. There are different methods which can find out the frauds from the input of the transactions that had actually happened till date. But if a system can find the frauds before the transaction is processed will be of more use.

The main aim of this proposition is to detect the fraud from different perspective. Rather than making amount of the transaction as the aspect to detect the fraud, consider the location of the transaction. Normally a person performs the transaction from the area where he is currently living which is normally his hometown or in some cases the area of their work or education. So the behavior of the person's transactions on the basis of the area where the transaction is processed can be used to detect frauds. This is helpful in detecting those frauds which are not detected on the basis of the amount.

The HMM is used because it is helpful in detecting the fraud before the transaction. So the alerts can be generated before the transaction processing, that is if the system finds out that the transaction being processed shows the abnormal behavior that is it deviates from the usual behavior of the user, it will generate an alarm. It also has low false alarms.

In this way, we can detect the fraud before actual processing and instead of limiting our detection to just the spending behavior of the user, we are actually considering the other important aspect of the user profile, which is location. For this first clustering is done using the k-means method and then using Baum Welch algorithm parameters of HMM are calculated.

1.4 Organization

This thesis has been sorted out into various chapters.

- The first and current part of the thesis represents the goals, related work done in the research and the associations of the work.
- The second part presents the brief review of the research methodology and various techniques which will be used for the fraud detection.
- In chapter 3, whole process used in the implementation of the FDS is described.
- The results after application of the FDS on the dataset have been computed in the fourth part.
- The fifth part concludes the proposition and explains the future work which can be done to improve the system further.
- Finally, all the references used in the research have been mentioned.

CHAPTER 2 TECHNIQUES AND TERMINOLOGIES

In this section, we will take a brief overview of all the techniques and terms which will be used in the Fraud Detection System and also other concepts related with it.

2.1 Outlier:

An outlier is an instance whose appearance is different from the rest of that arrangement of information. Casually, an outlier can be clarified as each data value which shows an extraordinary or disparate behavior with reverence to the rest of the information. Various definitions have been proposed for outlier. It's very important to detect the outliers and it's a very important and trendy topic these days. Fraudulent transactions can also be considered as the outliers because they also show the abnormal behavior. The applications of outlier detection also include fraud detection.

2.2 Fraud:

Fraud can be described as any wrongful or criminal misdirection planned to achieve money related or individual gain. Credit card frauds can be characterized as "Unapproved account movement by an individual for whom the account was not authorized. Actions have to be taken to stop such mishandles in advance and risk administration practices should be used to secure against such activities later on". In fundamental terms, Credit Card Fraud is portrayed as when somebody utilizes someone else's card for personal reasons while the card's proprietor don't have any idea about how the card is being used. Furthermore, the person using the card has no objective of making the repayments for the purchase they have done. Fraud discovery incorporates identifying Fraud as quick as possible once it has been executed.

2.3 K-Means Clustering:

K- Means clustering is an unsupervised method to group the instances into a fixed number of groups say ‘k’. Say we are given a data set of items. The task is to classify those items into groups. To achieve this, we will make use of the k-Means algorithm. In our FDS, we are provided with the dataset which contains the distance of the card holder current city from the location where transaction is processed. We will use the K-mean algorithm to categorize these values into three ranges i.e. Close, Halfway, Far.

In general the algorithm will classify the instances into k groups of similarity. To find out that similarity, the Euclidean distance as measurement is used. The aim of this algorithm is to minimize the objective function, which is squared error function in this case.

The objective function considered is:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is the measure of distance between a \the cluster mean c_j and data point $x_i^{(j)}$, that is the distance of all the data points from their respective mean of the cluster.

The main task is to make clusters in such a way that there is maximum similarity within the clusters and minimum similarity between the clusters.

The algorithm works as follows:

- Initialize k points randomly and call them means.
- Then all the remaining values are categorized to their closest mean and then when all the items are classified, we update the center’s coordinates, which will be the means of the items categorized in that cluster so far.
- We repeat the process iteratively till there will be no change in the means of the cluster and at the end, we have our clusters.

The below figure explains the whole k -means method from initialization of the data set to assigning the cluster centers and then assigning all other data instances to the clusters. This process is repeated iteratively till the cluster's centers don't change anymore.

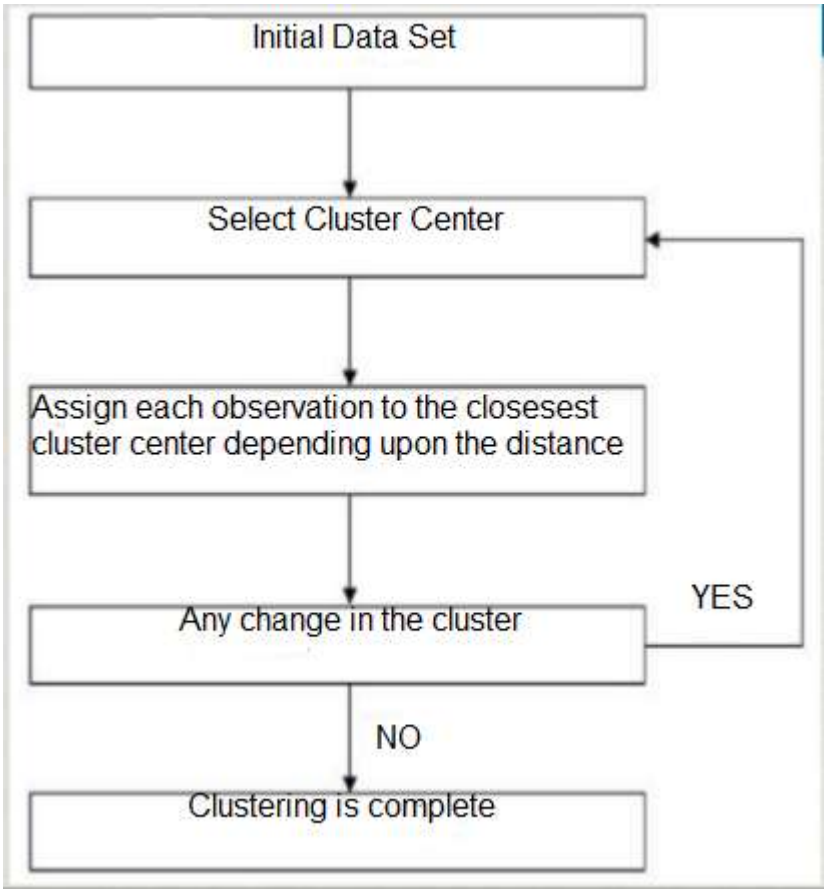

Figure 2.1 K-means clustering method algorithm.

In our FDS, the k means algorithm will take the distance between the location where the transaction is processed and the current city where the cardholder lives as the input. Then it classifies these data values into three clusters where each cluster specify behavior of the user i.e. whether it is near, far or in between the distance. The benefit of k-means is that it is simple and is comparatively fast.

Instance No.	1	2	3	4	5	6	7	8	9	10
Instance Value	25	80	40	15	5	25	10	20	10	15

Table 2.1 Data values for K –means algorithm.

Consider the above example and k=3 .If we apply the K-means algorithm on the data in the table, the outcome is the clusters with the centroids as c1, c2, and c3 which is shown below.

Cluster	C1	C2	C3
Value of the mean of the cluster	8.3	20	60
% of the instances in the cluster	30	50	20

Table 2.2 Different clusters using K-means algorithm.

It can be seen that the instances 5, 10, and 10 are grouped into the cluster with c1 as the mean which is 8.3. The total number of items in this cluster is hence 30 percent. Similarly data instances 25, 25, 20, 15, and 15 have been grouped in the c2 with mean 20, whereas values 80 and 40 have been grouped together in cluster c3. c2 and c3 contains 50 percent and 20 percent respectively of the total number of data instances.

This is how K-means clustering worked for our example. The advantages of k-means are that it is fast and simple.

In the following figure, k mean algorithm is represented. In a) initial clustering is done .In b) the clusters are updated and in c) the new means are calculated.

The b and c part continues iteratively until no further means are updated.

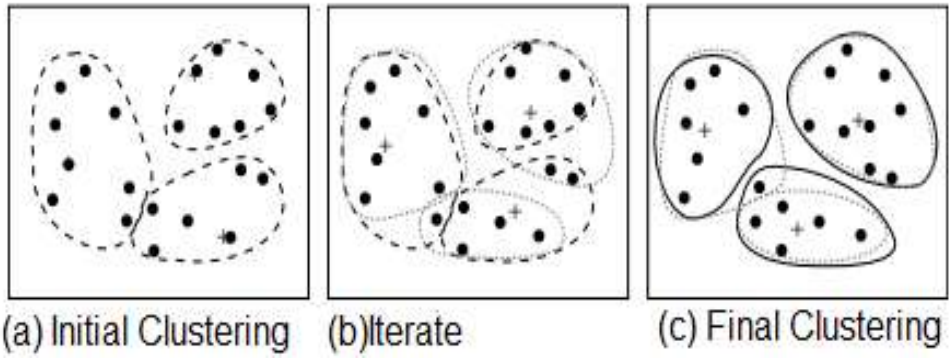

Figure 2.2 K-means clustering method.

2.4 Hidden Markov Model

HMM consists of the states that are actually hidden from the user and each state is linked with the transition probabilities. The states are not evident to the user, the outcomes linked with the states are important to the user. HMM works on the basis of the behavior profile of the user. The area of the transaction will be our states .We will consider three different ranges of the area:

1. Close
2. Halfway
3. Far.

HMM can be explained in the following manner:

- There will be 'N' states in the system and the set of these states is written as S={S_1 ; S_2 ; . . S_N}. The state at any given instance is represented as q_t.
- The no. of the observation symbols is considered to be 'M'. These observation symbols forms the output of the system and these symbols are represented by the set V where V is {V_1; V_2; . . . V_M}.

The below diagram represents a 3 states HMM where a_{ij} represents the probabilities of the transition of the state.

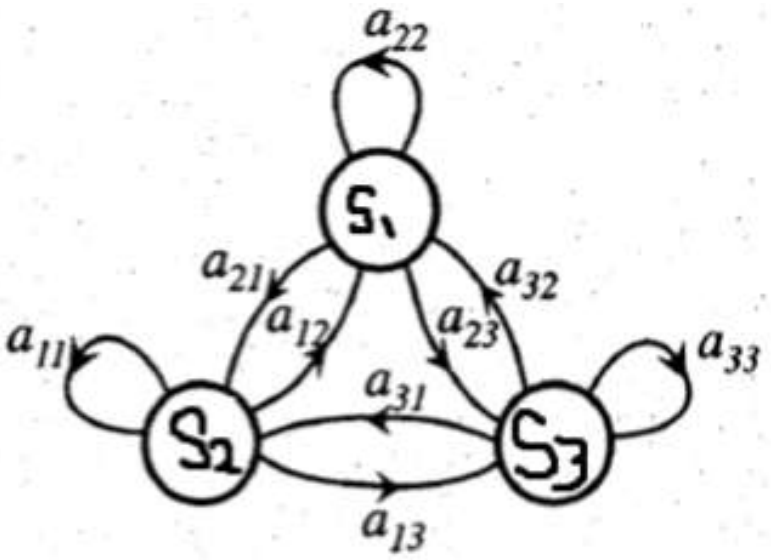

Figure 2.3 HMM states.

- A = [a_{ij}] shows the transition likelihood of the states i.e. the likelihood of state change from 'i' at any time't' to 'j' at the next time i.e. 't+1'. So a_{ij} can be numerically written as P (q_{t+1} =S_j | q_t=S_i) where 1<=i<=N and 1<=j<=N. On the off chance that we think about a typical situation where any state 'i' we can reach to any other state 'j' at that point a_{ij} >0 for all I and j. Additionally the sum of all the probabilities of transition from any state to all other states is 1.
- The observation symbol probability matrix i.e. the observation is k given the state is j can be represented as B where

  B= $b_j(k)$, in which $b_j(k)$= P(V_k|Sj).

- π is the initial state likelihood vector where π = [π_i] and π_i =P(q_1=S_i) i.e. the probability that the start state is S_i.

- $O = O_1, O_2, O_3,...O_R$, is the observation sequence where R is the no. of perception in the sequence and each observation O_t belongs to V.

Doubtlessly a whole working of a HMM needs the calculation of two parameters, M and N, and three prob. allocations A, π, and B. The documentation $\lambda = (A, B, \pi)$ is used to demonstrate the whole course of action of parameters of the model. The perception succession O, as said above, can be made by various state courses of action. Consider one such particular arrangement Q=q1, q2,……,qR where q1 is the hidden state.

The likelihood that O is created from this state succession is given by by $P(O|Q,\lambda) = \pi^{R}_{t=1} P(O_t|q_t, \lambda)$, where genuine opportunity of observations is acknowledged.

The above condition can be stretched out as $P(O|Q,\lambda)= b_{q1}(O_1).b_{q2}(O_2)\ldots .b_{qR}(O_R)$. The likelihood of the state succession Q is given as $P(Q|\lambda)=\pi_{q1},a_{q1q2},a_{q2q3}\ldots \ldots .a_{qR-1,qR}$.

In this manner, the likelihood of construction of the perception design O from the HMM showed by λ can be formed as : $P(O|\lambda) = \sum P(O|Q,\lambda) P(Q|\lambda)$ for all Q . Along these lines, a framework named as Forward-Backward strategy is used to figure $P(O|\lambda)$.

The HMM in proposed FDS is characterized into two sections.

One is the preparation part where the pattern of the sequences are recorded by clustering them into different symbols and the parameters of the HMM are calculated using the forward-backward algorithm.

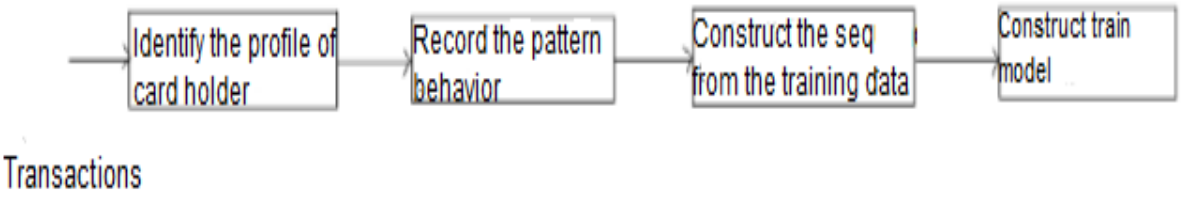

Figure 2.4 Training phase of HMM.

The second part is the testing phase, in which new transaction value is given as input to the system. The system will first find the cluster to which it belongs and then check the

likelihood of the system to accept the transaction. If it can accept the transaction with a major probability then it is termed as genuine, otherwise fraud.

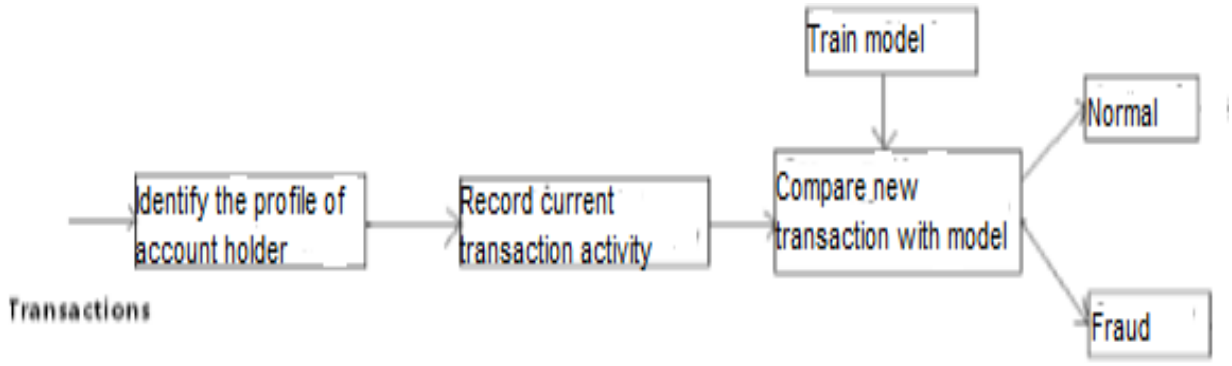

Figure 2.5 Testing Phase of HMM.

The HMM is used because it is helpful in detecting the fraud before the transaction. So the alerts can be generated before the transaction processing, that is if the system finds out that the transaction being processed shows the abnormal behavior that is it deviates from the usual behavior of the user, it will generate an alarm. It also has low false alarms.

2.5 Baum Welch Detector:

The forward and in backward algorithm is an imperative idea of the Hidden Markov Model. This calculation is likewise called as Baum-Welch Calculation or Baum-Welch Detector. It is utilized for inferencing in the Hidden Markov Model. This model is based on the idea of Dynamic Programming as it tries to register the results ideally.

Dynamic Programming was first utilized by Bellman to discover effective and ideal answer for non-trival things. Dynamic programming begins at small issues , figures the ideal answers for these and after that by working from bottom to top , gets the ideal arrangement of the primary large issue.

Consider the figure below:-

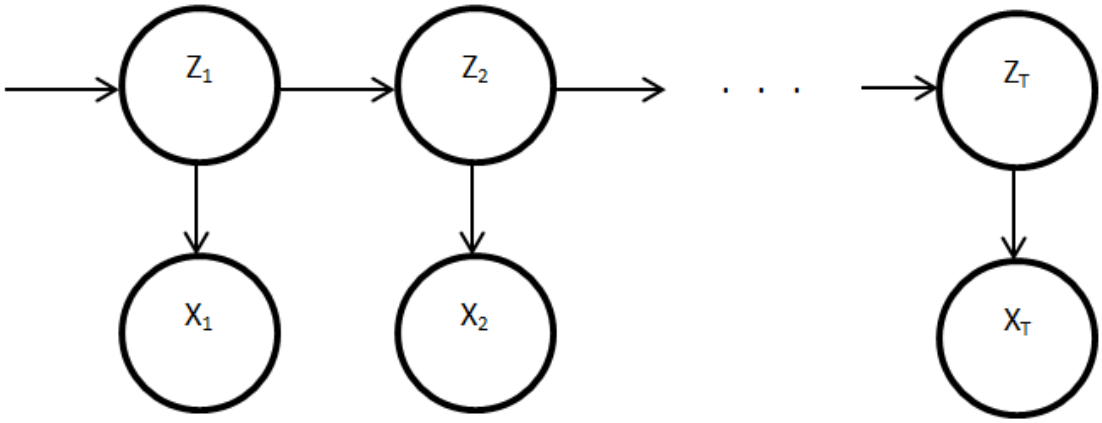

Figure 2.6 States and symbols transitions.

The problem statement of Forward/Backward or Baum-Welch algorithm states that P(x|z), P(zk|zk-1) and P(xi) is known to us. These probabilities are also called as emission, transition and initial probabilities. We are required to compute P(zk|x), where we have n such instances in our Markov model. Effectively the forward algorithm states that we have to compute P(zk,x1:k)(for example calculating P(z3 for x1x2x3). And the backward algorithm states that we have to compute P(xk+1|zk) where k = 1.....n .

Used to consequently evaluate parameters of a HMM or the Forward-Backward calculation an exceptional instance of the Expectation Maximization (EM) calculation

1. Begin with beginning likelihood gauges

2. Figure desires for how frequently each change/outflow is utilized

3. Re-appraise the probabilities in view of those desires

Baum Welch Algorithm can be explained as follows:

- An arrangement of observation patterns, O^1, O^2…..is given as input and arbitrary model parameters are initialized, $\lambda' = a_{ij}, e_i()$.
- The score of each λ' is calculated where score $= \sum_d P(O^d | \lambda')$.
- Repeat till the change of score is below the predefined value for each λ', S' and for each arrangement, O^d, calculate
- "Likely paths" $Q^d = q^d1, q^d2, ...$
- Compute α(t, i) for O^d utilizing Forward calculation.
- Compute β(t, i) for O^d utilizing Backward calculation.
- Ascertain the commitment of O^d.
- Figure the commitment of O^d to E.
- Calculate aij, ei(∂) and score$= \sum_d P(O^d | a_{ij}, e_i())$.

This is needed because the paths of the state are hidden, and we can't solve the equations analytically.

CHAPTER 3 IMPLEMENTATION DETAILS

Normally the frauds are detected on the basis of the amount of transaction. But in this work we are considering another important perspective i.e. location. In this 20 transactions are fed to the system to train the model and then will test the system on the basis of the location and amount individually. To prepare the dataset the distance between the person's permanent location and the place where the transaction is done is considered.

3.1 Data Set Used

The FDS is conducted on the data from the different customer's account. It was conducted on the 14 different customers having different style of work and belonging to different age groups. For each customer, first 20 transactions were used as the training set to build up the model and then further transactions were tested by the system. The no. of transactions for each account is different. The system is used to test the individual effect of the distance and amount separately. The data set is preprocessed as we have nothing to do with the type of the purchase or the exact location. We are concerned with the amount of the transaction and the distance of the current city of the owner with the location where transaction is performed.

3.2 Location

For a normal person, the amount of the transaction may not find all the frauds as now everyone is making almost all the transactions from the credit cards. So a particular pattern for their spending style is difficult. Normally a person performs the transaction from the area where he is currently living which is normally his hometown or in some cases the area of their work or education. So the behavior of the person's transactions on the basis of the area where the transaction is processed can be used to detect frauds as the patterns of the person

according to the area are more primitive. This is helpful in detecting those frauds which are not detected on the basis of the amount.

3.3 Clustering

To begin credit card fraud identification framework utilizing HMM, we start by first picking the perception images in our model. We quantize the distance value into M esteem ranges V1; V2;….VM, i.e. ordering the qualities into observation images in light of grouping. In our work, we consider only three esteem ranges, to be particular, close (c), halfway (h), and far (f). Our game plan of perception images is, in this way, V ={c,h,f} making M =3. For example consider, c ={1, 200}, h ={200, 450} and f ={450,800}. If a cardholder plays out an exchange at a distance of 190, at that instance the observation symbol is c. The actual place of transaction is not considered rather the distance between the places is considered.

For clustering, K means clustering is used. First K –means clustering is used to form the clusters on the training data set and then for each test data, k-mean clustering work by assigning the cluster to it. The strategy adopts a fundamental and straightforward strategy to portray a given data into a particular number of gatherings (expect k groups) settled apriori. The central idea is to describe k focuses, one for each bunch. The better choice is to put them anyway as much far as possible from each other. The accompanying stage is to take each point and connect it to the nearest focus. When no occurrence is remaining, the initial step is finished and an early grouping age is done. Presently we need to re-designate other k new centroids as mean of the gatherings appearing as an outcome of the past advances. Also, when we have these k new means, another progression of coupling should be done between comparable information occasions and the closest new focus. A circle of steps has been created. Due to this circle we may see that the k centers change their region all around requested until the point come where no more changes are done or accordingly centers don't move any more. For our usage we took k=3.

3.4 Training

Subsequent to choosing the state and observation symbols, the next stage is to decide the likelihood parameters A, B, and π with the goal to complete the training of HMM. These three system parameters are resolved in a training stage by utilizing the Baum-Welch calculation. Once these are evaluated then the final detection can be made.

The forward and backward algorithm is an important concept of the Hidden Markov Model. This algorithm is also called as Baum-Welch Learning or Baum-Welch Detector. It is used for inference in the Hidden Markov Model. This model is built on the concept of Dynamic Programming as it tries to compute the results optimally.

Calculation of the "forward-backward" probabilities in view of the current parameters is done as follows:

Forward likelihood is computed as underneath.

a. At any particular time t, the likelihood that system is in state i when the past perception as yet has been o1 … ot.

b. $\alpha(i) = P(i,o1....ot/\lambda)$.

Backward likelihood is computed as:

a. At any particular time t given the system is in state i, the likelihood that the perception that takes after will be ot+1 … oT.

b. $\beta(i)= P(Ot+1....OT/i, \lambda)$

At that point use the forward-backward probabilities to assess the normal frequencies like expected number of changes from state i and approx. number of being in state i. Then we use the ordinary frequencies to assess the parameters.

3.5 Detection

After the parameters are found out, we take the observations from a owner's training information and shape an underlying succession of symbols. We are considering the last 20 transactions for the sequence.

- Let O1; O2; ... Or then again be one such gathering of length R. This recorded gathering is encircled from the owner's trades up to time t. We input this progression to the HMM and figure the likelihood of acknowledgment by the HMM.
- Given the likelihood a chance to be $\alpha 1$, which can be composed as takes after: $\alpha 1 = P(O1, O2, O3,\ldots \ldots OR\ |\lambda)$.
- Let OR+1 be the image created by another exchange at time t +1. To shape another grouping of length R, we affix O_{R+1} and drop O1 in that succession, creating O2 ,O3,... as the new grouping.
- We input this new package to the HMM and figure the likelihood of acknowledgment by the HMM. Give the new likelihood a chance to be $\alpha 2 = P(O2, O3, O4,\ldots \ldots O_{R+1}\ |\lambda)$.
- Let $\Delta\alpha = \alpha 1 - \alpha 2$, now If $\Delta\alpha > 0$ then it implies that the new succession is acknowledged by the HMM with low probability, and it can be a fraud. The recently added exchange is considered to be fraud if the change rate in the likelihood is over a limit that is $\Delta\alpha/\alpha 1 \geq$ Threshold.

In the event that OR+1 turn out to be vindictive, the bank will not certify the trade, and the system discards the image. Else, O_{R+1} is included the succession for all time, and the new grouping is utilized as the base arrangement for deciding the legitimacy of the following exchange. The explanation behind considering new non malicious images in the succession is to catch the changing pattern in the behavior of a owner.

FDS is separated into two sections—first is the training portion and the next is detection. The whole process is explained with the help of the below figure.

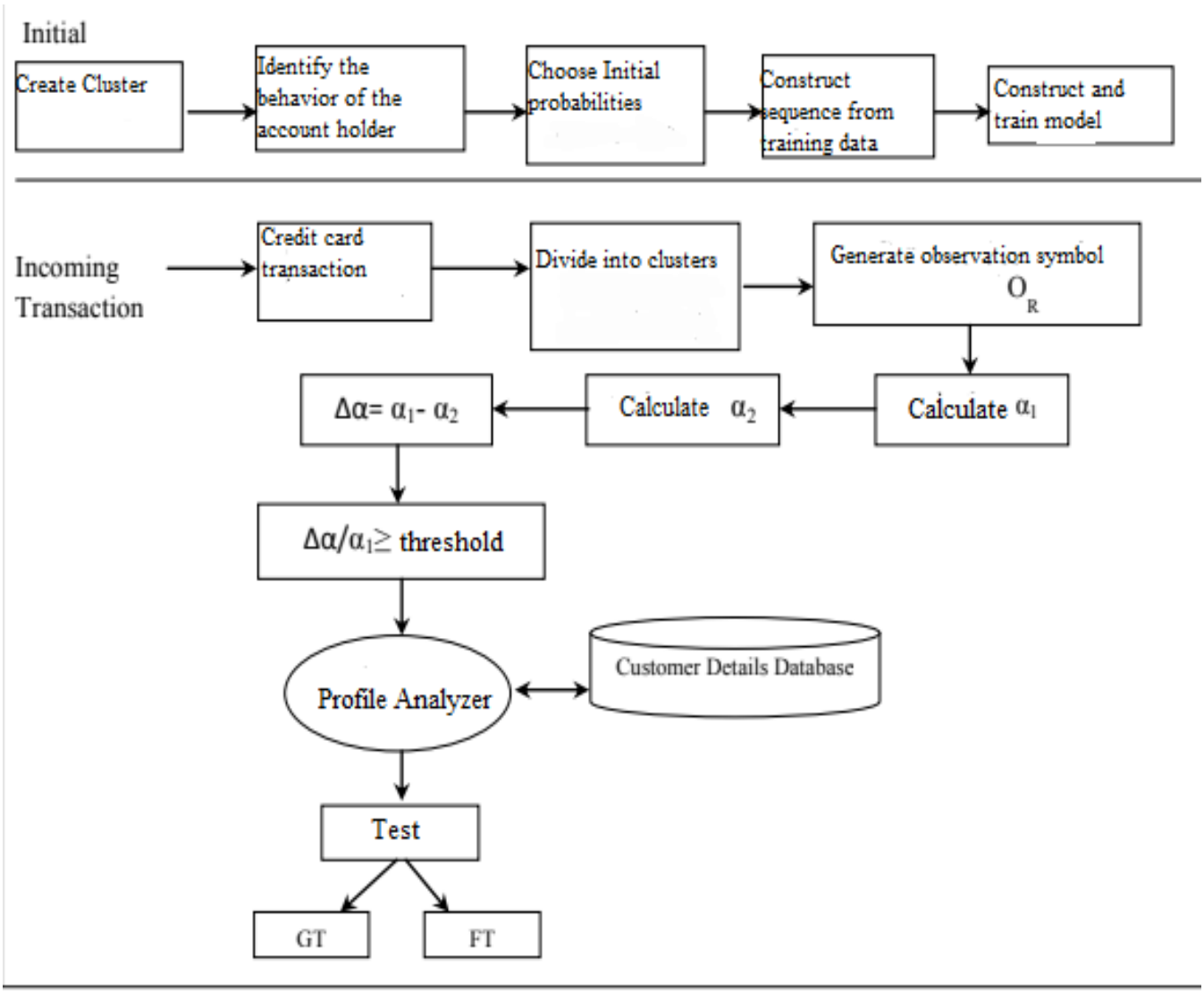

Figure 3.1 Fraud Detection System using HMM.

CHAPTER 4 RESULTS AND EVALUATION

The testing data is tested on the proposed FDS both on the basis of the amount of transaction and on the location independently. The FDS is conducted on the data from the different customer's account. It was conducted on the 14 different customers.

First the result of a single customer is shown in detail and then the overall result of the system on the different profile's is explained. The training data for this customer consists of 20 transactions. On the basis of the distance and amount separately first clustering is done and then the training of HMM is done. Then the detection phase starts where the 100 test cases are fed to the system one by one to check whether it detects fraud or not.

4.1 Basic Terminologies

Before the results, some terminologies are explained that will be used further. To understand the basic terminologies related with the result evaluation, consider that the output is divided into two classes, class 1 and class2 .Now depending upon the actual classification and predicted classification, different terminologies are defined as below:

- **True positives (TP):**

These refer the positive instances that were accurately marked by the classifier. TP is used to represent the true positives. "TP of Class1" is all Class1 cases that are named Class1.

- **True negatives (TN)**:

These are total of the negative instances that were accurately marked by the classifier. TN is used to represent the true negatives. "TN of Class1" is all non-Class1 examples that are not delegated Class1.

- **False positives (FP):**

These are the instances which are actually negative but were erroneously named as positive (e.g., instances of class fraud = no which the system anticipated fraud= yes). FP is used to represent the false positives. "FP of Class1" is all non-Class1 examples that are delegated Class1.

- **False negatives (FN):**

These are the instances which are actually positive but were predicted as negative (e.g., instances of class fraud = yes which the system anticipated fraud= no). FN is used to represent the false negatives. "FN of C1" is all C1 examples that are not delegated C1.

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 4.1 Output classes.

The above figure explains classification of the output classes.

- **Accuracy:**

Accuracy is the most instinctive performance measure and it is just a proportion of accurately predicted instances to the aggregate instances. Accuracy is generally measured when the dataset is symmetric i.e. when the value of false negatives and false positives are nearly same.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:**

To get the estimation of precision we divide the aggregate number of effectively grouped positive cases by the aggregate number of predicted positive cases. High Precision demonstrates an instance marked as positive is indeed positive (small number of FP).

$$\textit{Precision} = \frac{TP}{TP + FP}$$

- **Recall:**

0 Recall can be characterized as the proportion of the aggregate number of accurately arranged positive cases divide to the aggregate number of positive cases. High Recall shows the class is effectively recognized (less number of FN).

$$\textit{Recall} = \frac{TP}{TP + FN}$$

4.2 Results of a customer

Results of a particular customer are described in detail below. The training dataset for this customer consists of 20 transactions. First the clustering is done on the training data set. The testing dataset of 100 transactions are fed into the system one by one. Out of the 100 transactions, there were total 21 fraudulent transactions and 79 genuine transactions. The model is executed two times, One on the basis of the location and one on the basis of the amount. First the results for the distance factor are described.

The clustering result on the basis of the distance of the training dataset shows that most of the transactions belong to the halfway cluster which is 50% of the total and only 10%

belongs to the far and 40% belongs to the close cluster. So according to this, a pattern is formed of 20 transactions.

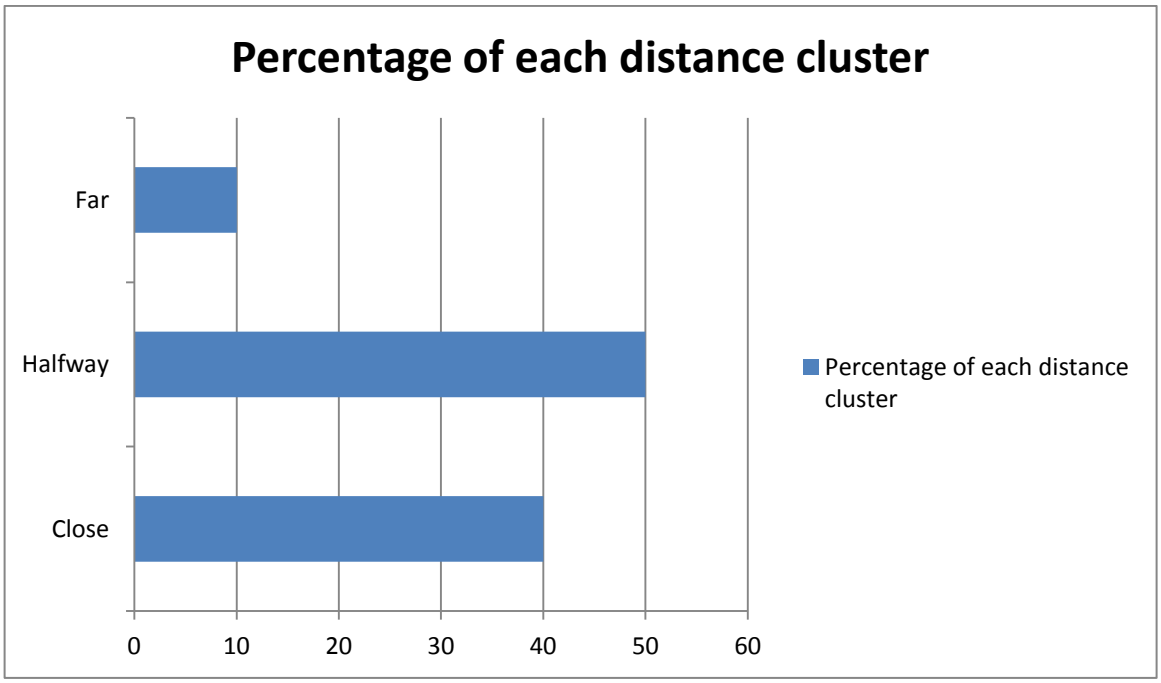

Figure 4.2 Percentage of each distance cluster.

Whenever a new transaction comes, the first transaction in the pattern is discarded and the new transaction cluster symbol is added.

When the model is run for the test dataset with the distance values as the input then the following results come out. Out of the 21 frauds, 15 frauds were detected correctly whereas 6 fraudulent transactions were accepted by the model as genuine transactions. Only 7 transactions were detected as frauds which were actually genuine transactions. Using HMM, the false alarms are decreased.

Actual	Predicted		
		FRAUD	GENUINE
	FRAUD	15	06
	GENUINE	07	72

Figure 4.3 Results when distance values are given as input.

According to the results,

TP=15, FN=6, TN=72, FP=7.

Accuracy = 87/100=87%.

Recall= 15/21= 71.4%.

Precision= 15/22=68.1%.

If we detect the frauds on the basis of the amount of the transaction, then the amount of the transaction is fed to the system.

First clusters are made on the testing data, in this case the type of the products bought are not important rather the amount is considered for clustering. The results shows that most of the transactions belong to the low cluster which is 45% of the total and only 15% belongs to the high and 40% belongs to the medium cluster. So according to this, a pattern is formed of 20 transactions.

The clustering result on the basis of the transaction amount of the training dataset is as follows:

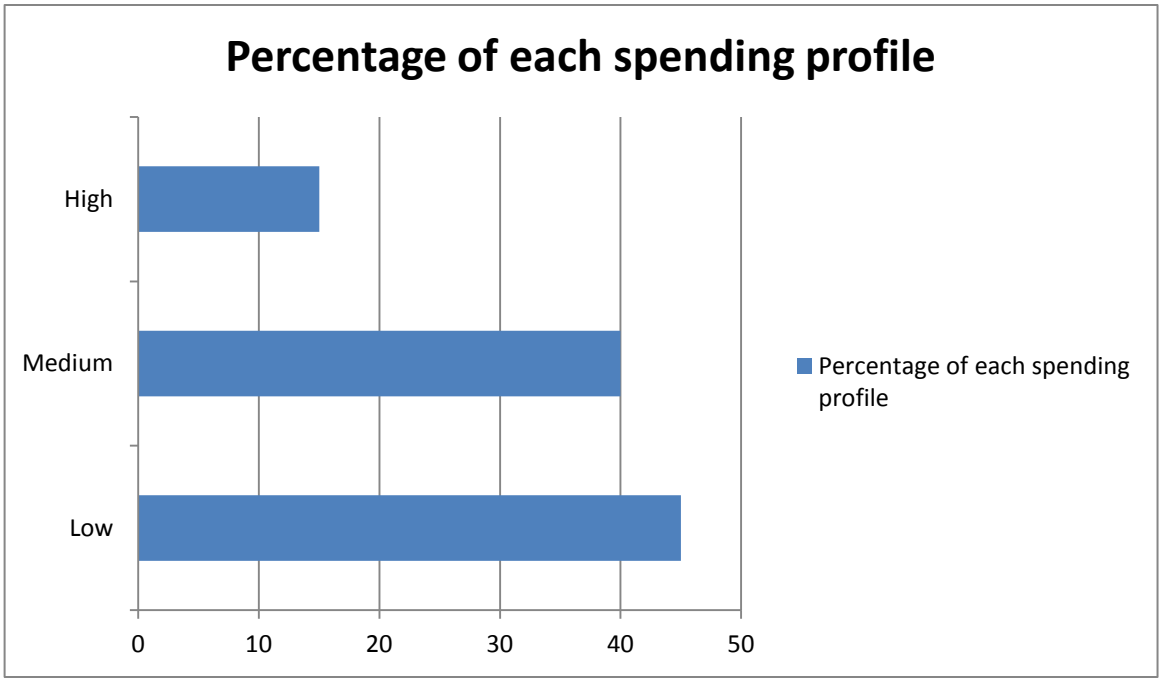

Figure 4.4 Percentage of each spending cluster.

Now after passing the training data for the amount of the transaction, test data is passed. When the model is run for the dataset with the amount values as the input then the following results come out. Out of the 21 frauds, 12 frauds were detected correctly whereas 9 fraudulent transactions were accepted by the model as genuine transactions. 9 transactions were detected as frauds which were actually genuine transactions. So, we have seen that the model behaves better for the distance factor than the amount factor.

Actual	Predicted		
		FRAUD	GENUINE
	FRAUD	12	09
	GENUINE	09	70

Figure 4.5 Results when amount values are given as input.

According to the results,

TP=12, FN=9, TN=70, FP=9.

Accuracy = 82/100=82%.

Recall= 12/21= 57%.

Precision= 12/21=57%.

4.3 Overall Results

If we evaluate the results from all the customers, we conclude that out of 14 customers, for 6 customer's account, amount behaved better but for the rest 8 profiles distance acted as a better aspect. The min. accuracy was 60% in case of the amount and in case of distance it was 65%. The max accuracy for the distance parameter was 87% where as for amount it was 84%. So it can be seen that the distance is better than the amount to detect the frauds. Accuracy in both the cases for all the customer's profile is shown in the below figures.

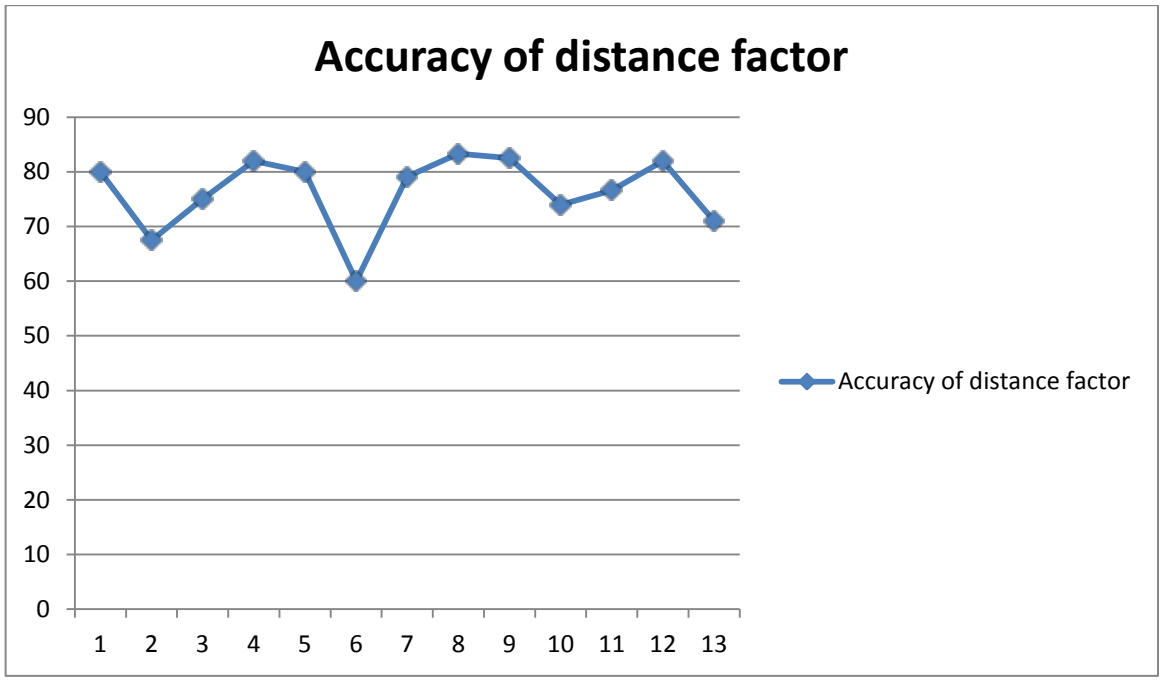

Figure 4.6 Overall Results when distance values are given as input.

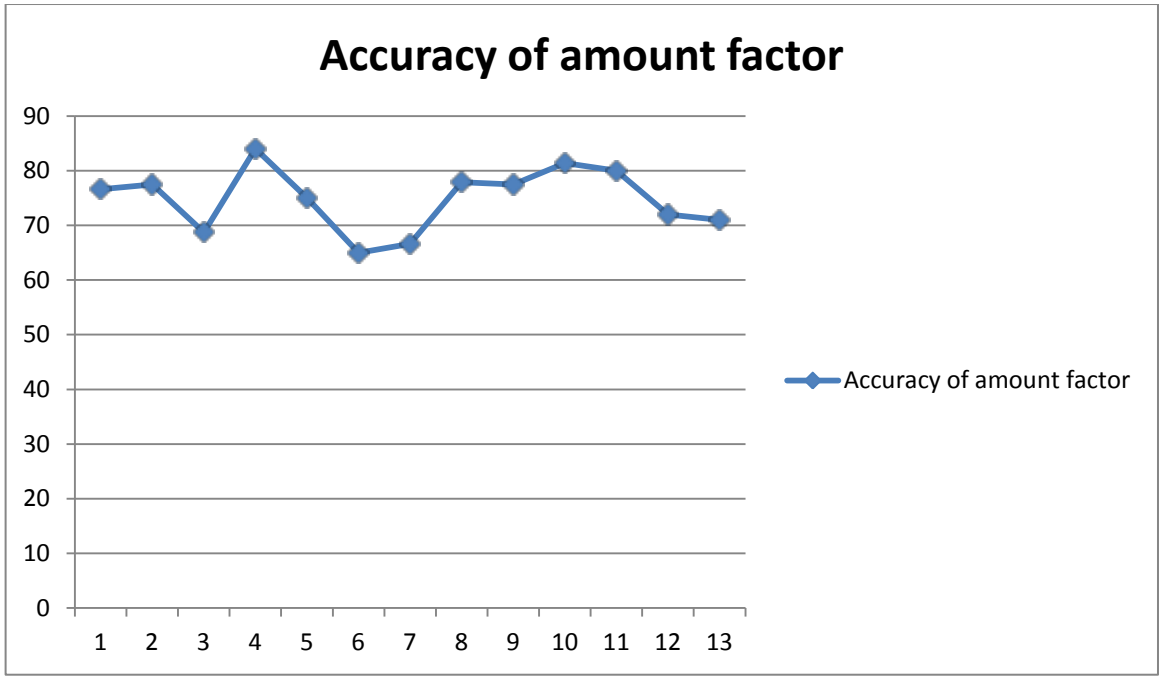

Figure 4.7 Overall results when amount values are given as input.

CUSTOMER NUMBER	ACCURACY FOR DISTANCE FACTOR(%)	ACCURACY FOR AMOUNT FACTOR(%)
Customer 1	87	82
Customer 2	80	76.6
Customer 3	67.5	77.5
Customer 4	75	68.7
Customer 5	82	84
Customer 6	80	75
Customer 7	60	65
Customer 8	79.1	66.6
Customer 9	83.3	77.9
Customer 10	82.5	77.5
Customer 11	74	81.4
Customer 12	76.6	80
Customer 13	82	72
Customer 14	71	71

Table 4.1 Overall Results.

The overall results of all the customers are shown in the table above. Out of 14 customers, for 6 customer's account, amount behaved better but for the rest 8 profiles distance acted as a better aspect. The min. accuracy was 60% in case of the amount and in case of distance it was 65%. The max accuracy for the distance parameter was 87% where as for amount it was 84%. So it can be seen that the distance is better than the amount to detect the frauds.

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

Credit card related frauds are increasing at a very high rate due to the increase in use of the internet. It has turned out to be essential to identify these frauds. Rather than making amount of the transaction as the aspect to detect the fraud, we consider the location of the transaction. The results show that for a regular individual the distance from where the transaction is made is more helpful to detect the frauds as compared with the amount of the transaction. So the behavior of the person's transactions on the basis of the area of the transaction can be used to detect those frauds which are not detected on the basis of the amount.

After examining the results with different transactions with the help of table and bar graph and using different perspectives of amount and distance, it can be inferred that the accuracy in case of the distance as the attribute is more than when the amount of transaction is used. Also the false alarms in case of the amount are more. The no. of the fraudulent transaction that are predicted genuine is also more in the case of the amount, one reason for this can be that as people are using more and more of the credit cards these days, so their transactions comprised of the mix of the small, medium and high patterns without any specific sequence, so it sometimes consider even a fraud transaction as genuine because of the mix. While if we consider the distance, a person normally makes transaction from some fixed no. of the

locations, so in this case patterns can be easily used to detect whether the transactions are genuine or fraudulent.

The HMM is used because it is helpful in detecting the fraud before the transaction. So the alerts can be generated before the transaction processing, that is if the system finds out that the transaction being processed shows the abnormal behavior that is it deviates from the usual behavior of the user, it will generate an alarm. HMM have low false alarms.

5.2 Future Work

In the future we can also combine different aspects together in the same model. The location and amount both can be used to detect the fraud simultaneously. This will make the system better.

REFERENCES

[1] V.Bhusari and S.Patil, “Study of Hidden Markov Model in Credit Card Fraudulent Detection,” World Conference of Futuristic Trends in Research and Innovation for Social Welfare, vol. 20, no.6, pp. 33-36, 2016.

[2] Rajeshwari U and B Sathish Babu,, “Real-time credit card fraud detection using Streaming Analytics,” in 2nd International Conference on Applied and Theoretical Computing and Communication Technology, pp. 439-444, 2016.

[3] Ankit Vartak, Chinmay D Patil and Chinmay K Patil, “Hidden Markov Model for Credit Card Fraud Detection,” International Journal of Computer Science and Information Technologies, vol. 5, pp. 7446-7451, 2014.

[4] Fuzail Misarwala, Kausar Mukadam, and Kiran Bhowmick., “Applications of Data Mining in Fraud Detection ,” International Journal of Computer Sciences and Engineering vol.-3, pp. 45-53, 2015.

[5] Dewi Rahmawati, Riyanarto Sarno, Chastine Fatichah and Dwi Sunaryono, “Fraud Detection on Event Log of Bank Financial Credit Business Process using Hidden Markov Model Algorithm,” IEEE 3rd International Conference on Science in Information Technology, pp. 35-40, 2017.

[6] N.Malini and M.Pushpa, "Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection ", 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEEICB17) ,2017.

[7] Priyanka Yadav, Pavan Wangade, Manish Thakur, Mohammed Fakih, Gayatri Hegde, "Proposed Distributed Data Mining In Credit Card Fraud Detection," International Research Journal of Engineering and Technology, pp. 460-463, 2016.

[8] K.RamaKalyani, D.UmaDevi , "Fraud Detection of Credit Card Payment System by Genetic Algorithm," International Journal of Scientific & Engineering Research Volume 3, Issue 7, pp. 215-220, July-2012.

[9] Krishna Modi and Reshma Dayma, "Review On Fraud Detection Methods in Credit Card Transactions," International Conference on Intelligent Computing and Control, 2017.

[10] Rashi Bansal, Nishant Gaur and Shailendra Narayan, "Outlier Detection: Applications and Techniques in Data Mining," 6th International Conference - Cloud System and Big Data Engineering, pp. 373-377, 2016.

[11] Carlos A. S. Assis, Adriano C. M. Pereira and Marconi A. Pereira, " A Genetic Programming Approach for Fraud Detection in Electronic Transactions", Computational Intelligence in Cyber Security IEEE, pp. 1-8, 2014.

[12]S.Benson Edwin Raj, A. Annie Portia, "Analysis on Credit Card Fraud Detection Methods", IEEE International Conference on Computer, Communication and Electrical Technology, IEEE March 2011.

[13] Haibing Li, Man-Leung Wong" Financial Fraud Detection by using Grammar-based Multi-objective Genetic Programming with ensemble learning", Evolutionary Computation IEEE, pp. 1113-1120, 2015.