TECHNICAL TERM PAPER

CREDIT CARD FRAUD DETECTION

PHISHING Under the guidance of Dr.R.V Ramana Chary,

Department of Information Technology

B V Raju Institute of Technology, Narsapur.

By:

SAMUDRALA SAI SANTHOSH

18211A12A2

Abstract

Online shopping and banking has increased by the growth of the internet and by use of credit cards. Along with this, credit card fraud has also increased. Many modern techniques based on Artificial Intelligence, Data warehousing have evolved in detecting various credit card fraudulent transactions. We proposed a system which detects fraud in credit card transaction processing using a decision tree with a combination of Luhn's algorithm and Hunt's algorithm. Luhn’s algorithm is used to validate the card number. Address matching rule checks whether the Billing Address and Shipping Address match or not. This check does not guarantee whether a transaction is fraud or genuine. But if the two addresses match, the transaction can be classified as genuine with a high probability. Else, the transaction is labelled as suspect. A customer usually carries out similar types of transactions in terms of amount, which can be visualized as part of a cluster. Since a fraudster is likely to differ from the customer’s account, his transactions can be detected as exceptions to the cluster – a process known as outlier detection.

General Terms : Credit card fraud, online Transaction, Electronic Commerce

Key words: Electronic Commerce, Credit card fraud, address matching, spending pattern, Luhn's Algorithm, Outlier Detection

1. INTRODUCTION

With rapid advancement of e-commerce, use of credit cards for purchases has exponentially increased. Unfortunately, fraudulent use of credit cards has also become a source of crime.

Credit card fraud is a most popular term for theft and fraud committed using or involving a payment card, such as a credit card or debit card, as a fraudulent source of funds in a transaction. Credit card fraud is also an appendage to identity theft.

According to the United States Federal Trade Commission, while identity theft had been holding steady for the last few years, it saw a 21 percent increase in 2008. However, credit card fraud, the crime which most people’s privilege with ID theft, decreased as a percentage of all ID theft complaints for the sixth year in a row.

Financial institutions employ various fraud prevention models for tackling this problem. But fraudsters are adaptive, and given time, they devise several ways to intrude such protective models. Despite the best efforts of the financial institutions, law enforcement agencies and the government, credit card fraud continues to rise. Fraudsters nowadays may constitute a very inventive, intellect and fast moving fraternity. Several techniques for the detection of credit card fraud have been proposed in the last few years .Today technology is a basic mandatory need of humans. Just look around and you will know why. Literally, at every instant of time, you are surrounded by technology.

Today there is no such place where technology is not present. Due to technology communication is easy and quick, travel is fast, and movements are also fast. There are lots of advantages to technology, but with that it causes Fraud also. Fraud is behavior of humans which is out of rule and causes crime.

One of the biggest facilities provided by technology is that we can shop using various facilities provided by banks e.g Credit Card, Debit Card, Internet Banking [1] etc. Here is a major chance for fraud. Credit card becomes the most popular mode of payment for both online as well as regular purchase so mostly frauds happen in Credit Card System. A Credit Card Fraud is a transaction that is complete with your credit card by someone else.

Credit card fraud happens when someone steals your credit card, credit card information, or Personal Identification Number (PIN), and uses it without your permission to make purchases in stores, online or by telephone, or to withdraw money from an automated bank machine (ABM).

Many modern techniques[1] based on Artificial Intelligence, Data mining[3][4], Neural Network[2], Bayesian Network[6], Fuzzy logic[5], Artificial Immune System, Knearest neighbor algorithm, Support Vector Machine[7][8], Decision Tree, Fuzzy Logic Based System, Machine learning, Sequence Alignment, Genetic Programming etc., has evolved in detecting various credit card fraudulent transactions. Each method has its own pros & cons.

The Nelson Report, 2016

Credit card frauds make a greater impact on the merchants when compared to the consumers; merchants are considered to face more risks in the cr dit card transactions. While consumers may face trouble trying to get a fraudulent charge reversed, merchants lose the cost of the product s ld, pay charg back fees, and fear from the risk of having th ir merchan account closed. Increasingly, the card not present scenario, such as hopping on the internet poses a greater threat as the merchant (the web ite) is no longer protected with advantages of physical verifi ation such as gnatu ch ck, photo ident fication, etc.

In the last d ca e, credit card fraud has started to pose a great threat to businesses all over th world and it seems to have an impact on the e onom . It has become very important for busi ss organisat ons to count r these c edit card frauds effectively, for which understanding the redi ard is consider d to be qu lly important.

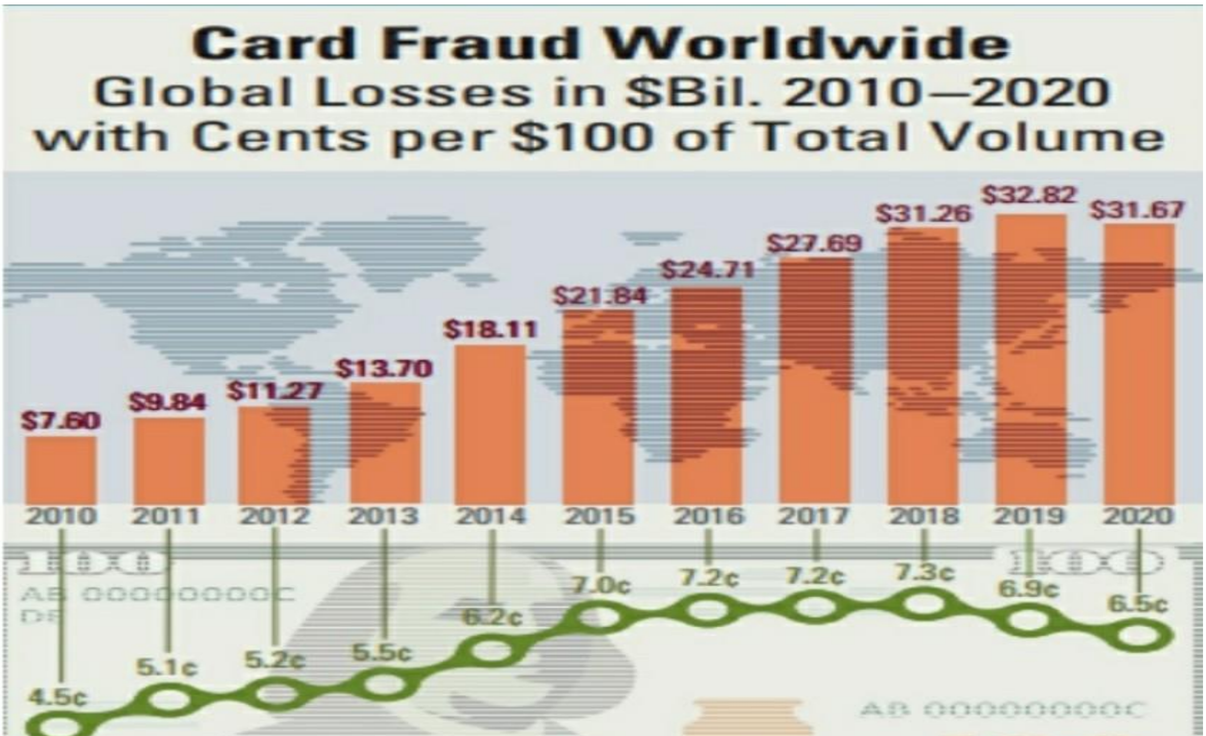


Fig.1 Credit Card Frauds Worldwide (The Nelson Report, 2016)

2. LITERATURE SURVEY

Chen et al. [3] propose a method in which an online questionnaire is used to collect questionnaire-responded transaction (QRT) data of users. Further it uses a support vector machine (SVM) trained with this data and the QRT models are used to predict new transactions. Chen et al. [4] have recently presented a personalized approach for credit card fraud detection that employs both SVM and ANN. It tries to prevent fraud for users even without any transaction data. However, these systems are not fully automated and depend on the user‟s expertise level.

Chan et al. [5] divide a large set of transactions into smaller subsets and then apply distributed data mining for building models of user behavior. The resultant base models are then combined to generate a meta-classifier for improving detection accuracy. Brause et al. [6] have explored the possibility of combining advanced data mining techniques and neural networks to obtain high fraud coverage along with a low false alarm rate. Use of data mining is also developed by Chiu and Tsai [7] .

Since the last two decades, research on the data mining techniques for credit card fraud detection has been started; Chan, et.al. (1999) addressed the growing credit card transactions in the US payment system that is considered to be leading to greater stolen credit card accounts. In the early years of credit card usage, banks faced a huge problem in analysing massive amounts of transaction data that efficiently computed fraud detectors in a timely manner.

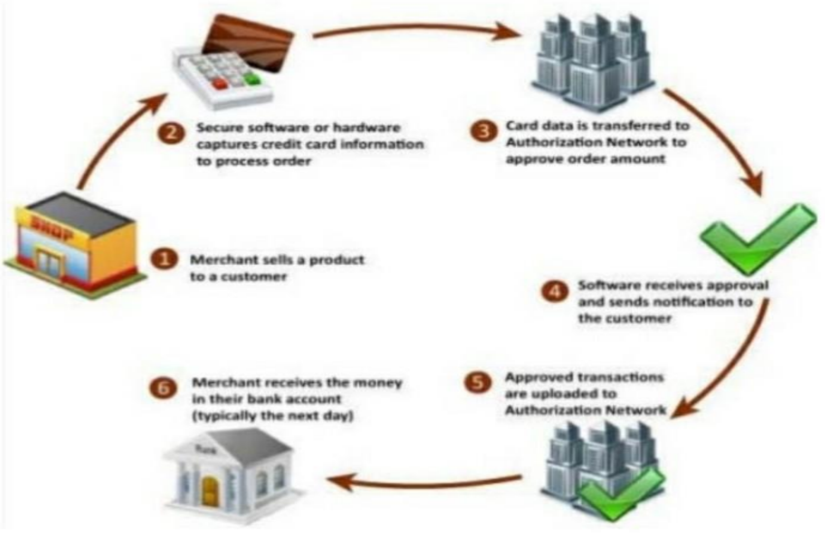
There are also several problems associated with the skewed distributions of training data and non-uniform cost per error. Chan, et.al (1999) conducted a study to address the three most important problems associated with the credit card transactions especially in e-commerce such as scalability, efficiency and technical issues.

2.1 Merchant Related Frauds

Merchant related frauds are initiated either by owners of the merchant establishment or their employees. The types of frauds initiated by merchants are described below:

i. Merchant Collusion: This type of fraud occurs when merchant owners or their employees conspire to commit fraud using the cardholder accounts or by using the personal information. They pass on the information about cardholders to fraudsters.

ii. Triangulation: Triangulation is a type of fraud which is done and operates from a web site. The products or goods are offered at heavily discounted rates and are also shipped before payment. The customer browses the site and if he likes the product he places the online information such as name, address and valid credit card details to the site. When the fraudsters receive these details, they order goods from a legitimate site using stolen credit card details. The fraudsters then by using the credit card information purchase the products.



1 Merchant sells a product to a customer
2 Secure software or hardware captures credit card information to process order
3 Card data is transferred to Authorization Network to approve order amount
4 Software receives approval and sends notification to the customer
5 Approved transactions are uploaded to Authorization Network
6 Merchant receives the money in their bank account (typically the next day)
SHOP
Bank

Fig.2 Credit Card Transaction Processing Steps

INTERNET RELATED FRAUDS

The internet is the base for the fraudsters to make the frauds in the simplest and the easiest way. Fraudsters have recently begun to operate on a truly transnational level.

With the expansion of trans-border, economic and political spaces, the internet has become a new world market, capturing consumers from most countries around the world.

The below described are the most commonly used techniques in Internet fraud:

i. Site cloning: Site cloning is where fraudsters close an entire site or just the pages from which the customer made a purchase. Customers have no reason to believe they are not dealing with the company that they wished to purchase goods or services from because the pages that they are viewing are identical to those of the real site.

The cloned site will receive these details and send the customer a receipt of the transaction through the email just as the real company would do. The consumer suspects nothing, while the fraudsters have all the details they need to commit credit card fraud.

Some sites often offer a cheap service for the customers. That site requests the customer to fill in his complete details such as name and address to access the webpage where the customer gets his required products.

CREDIT CARD GENERATORS

These are the computer programs that generate valid credit card numbers and expiry dates. These generators work by generating lists of credit card account numbers from a single account number. The software works by using the mathematical Luhn algorithm that card issuers use to generate other valid card number combinations.

ERASING THE MAGNETIC STRIP

This is the type of the fraud where the fraudsters erase the magnetic stripe by using the powerful electro-magnet. The fraudster then tampers with the details on the card so that they match the details of a valid card, which they may have attained, for example, when the fraudster begins to use the card, the cashier will swipe the card through the terminal several times, before realizing that the metallic strip does not work.The cashier will then proceed to manually input the card details into the terminal.

CREATING A FAKE CARD

Today we have sophisticated machines where one can create a fake card from scratch. This is common fraud though fake cards require a lot of effort and skill to produce it. Modern cards have many security features, all designed to make it difficult for fraudsters to make good quality fraudulent cards. After introducing the Holograms in the credit cards it makes it very difficult to forge them effectively.

FALSE MERCHANT SITES

Some sites often offer a cheap service for the customers. That site requests the customer to fill in his complete details such as name and address to access the webpage where the customer gets his required products. Many of these sites claim to be free, but require a valid credit card number to verify an individual's age. These kinds of sites collect as many as credit card details. The sites themselves never charge individuals for the services they provide.

3. METHODOLOGY

FRAUD DETECTION METHODS On doing the literature survey of various methods for fraud detection we come to the conclusion that to detect credit card fraud there are multiple approaches like

Gass algorithm

Bayesian networks

Hidden markov model

Genetic algorithm

A fusion approach using dempster shafer theory and bayesian learning.

Decision tree

Neural network

Logistic Regression

Gass algorithm

This algorithm is a combination of genetic algorithm and scatter search (Benson, Raj, & Portia, 2011). The basic operating principles of genetic algorithms and scatter search and then explain the steps of the suggested GASS algorithm. Genetic algorithms are inspired from natural evolution.

Normally the new generations will be produced by the crossover of two parent members. However, sometimes some random mutations can also occur on individuals which in turn increase the diversity in the population.The less fit members of this generation are eliminated and the fitter members are selected as the parents for the next generation.

The SS is another evolutionary algorithm which shares some common characteristics with the GA. It operates on a set of solutions, the reference set, by combining these solutions to create new ones.

Bayesian networks

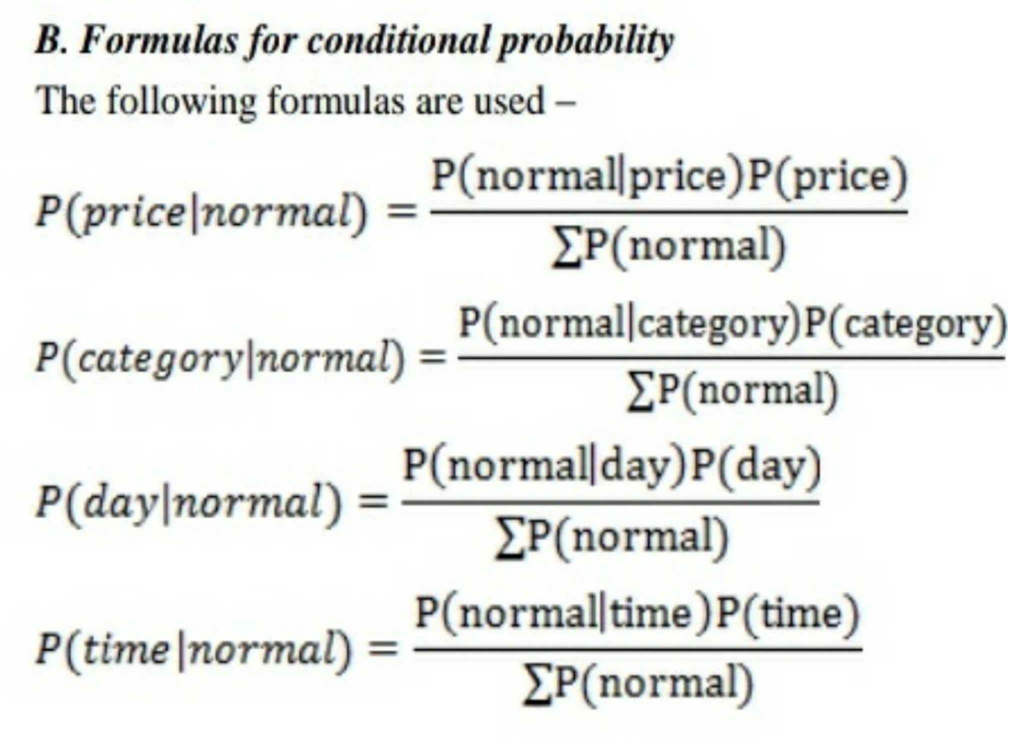
For the purpose of fraud detection, two Bayesian networks to describe the behavior of users are constructed. First, a Bayesian network is constructed to model behavior und r the assumption that the user is fraudulent (F) and another model under the assumption the user i l gitimate (NF).

The ‘user net’ is set up by u ing data from non fraudulent users. During operation the user net is adapted to a specific user based on emerging data. By inserting evid nce in these networks and propagating i th ough the network, the probability of the measurement x less than two above mentioned hypotheses is obtained.Thi means, it gives judgments to what degree observed us   behavior meets typical frau ulen or non fraudulent behavior. These quant ti s w call p(X NF) and p (X| F).

A. Cluster

is we are considering fo r clusters like “Price”, “Category”, “Day” & “Time” as hown in figur II. Our work is carr ed ou by considering the last en tran action and get the probabi ity of each cluster over “No mal and Susp cious”.

Formu as f r cond tional probability



B. Formulas for conditional probability

The following formulas are used –

$$P(price|normal) = \frac{\mathrm{P(normal|price)P(price)}}{\sum \mathrm{P(normal)}}$$

$$P(category|normal) = \frac{\mathrm{P(normal|category)P(category)}}{\sum \mathrm{P(normal)}}$$

$$P(day|normal) = \frac{\mathrm{P(normal|day)P(day)}}{\sum \mathrm{P(normal)}}$$

$$P(time|normal) = \frac{\mathrm{P(normal|time)P(time)}}{\sum \mathrm{P(normal)}}$$

TRANSACTION HISTORY DATABASE

A. Probability normal given category

For this we take all transactions from the database that match the current category with transactions.

B. Probability normal given day

For this we take all transactions from the database and match current day with transactions.

C.Probability normal given time

For this we take all transactions from the database to match current time with transactions. We match the current time between five hour before and five hour after.

Hidden markov model

A Hidden Markov Model is a double embedded stochastic process used to model much more complicated stochastic processes as compared to a traditional Markov model. If an incoming credit card transaction is not accepted by the trained Hidden Markov Model with sufficiently high probability, it is considered to be fraudulent transactions. HMM[5], Baum Welch algorithm is used for training purposes and K-means algorithm for clustering.HMM stores data in the form of clusters depending on three price value ranges low, medium and high (Bhusari & Patil, 2011).

Genetic algorithm

Genetic algorithms, inspired from natural evolution, were first introduced by Holland (1975). Genetic algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses.Fraud detection has been usually in the domain of Ecommerce, data mining.

Neural network

Fraud detection methods based on neural networks are the most popular ones. An artificial neural network (Chang, et.al., 2006) consists of an interconnected group of artificial neurons .The principle of neural network is motivated by the functions of the brain, especially3 pattern recognition and associative memory (Patidar & Sharma, 2011). The neural network recognizes similar patterns, predicts future values or events based upon the associative memory of the patterns it was learned.

Logistic regression

Two advanced data mining approaches, support vector machines and random forests, together with the well known logistic regression (Bhattacharyya, et.al., 2011), as part of an attempt to better detect (and thus control and prosecute) credit card fraud.

The study is based on real-life data of transactions from an international credit card operation. It is well understood, easy to use, and remains one of the most commonly used for data-mining in practice.

It thus provides a useful baseline for comparing performance of newer methods. Supervised learning methods for fraud detection face two challenges.

Support vector machine

The basic idea of SVM classification algorithm is to construct a hyper plane as the decision plane which makes the distance between the positive and negative mode maximum (Pun, 2011).

The strength of SVMs comes from two important properties they possess - kernel representation and margin optimization. Kernels, such as the radial basis function (RBF) kernel, can be used to learn complex regions. A kernel function represents the dot product of projections of two data points in a high dimensional feature space.

Random forests

The popularity of decision tree models in data mining arises from their ease of use, flexibility in terms of handling various data attribute types.

Single tree models can be unstable and overly sensitive to specific training data. Ensemble methods seek to address this problem by developing a set of models and aggregating their predictions in determining the class label for a data point. A random forest model is an ensemble of classification trees.

DEMPSTER–SHAFER ADDER

The role of the DSA is to combine evidence from the rules observation of Bayesian Network and by conditional probability and compute an overall belief value for each transaction.

Suppose from the current transaction we get – [Normal, Suspicious, Suspicious, Normal] By using DSA we can make the conclusion that the overall result is "Normal".

We demonstrate the effectiveness and usefulness of our FDS by testing it with large scale data. Due to unavailability of real life credit card data or benchmark data set for testing, we used dummy data that represent the behaviour of genuine cardholders as well as that of fraudsters.

Consider a particular example, we demonstrate results by observation and by using mathematical analysis. Mathematical. Analysis is cross checked with threshold. These thresholds are calculated by using above threshold formulas.

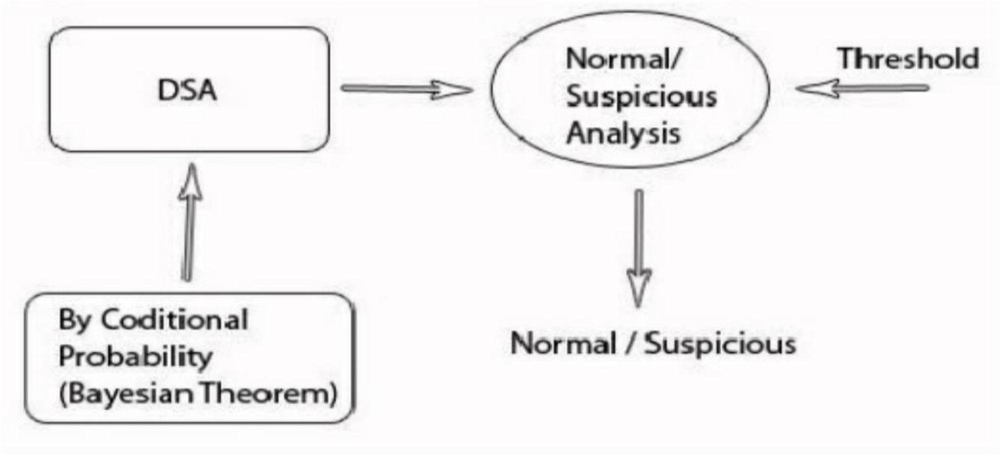
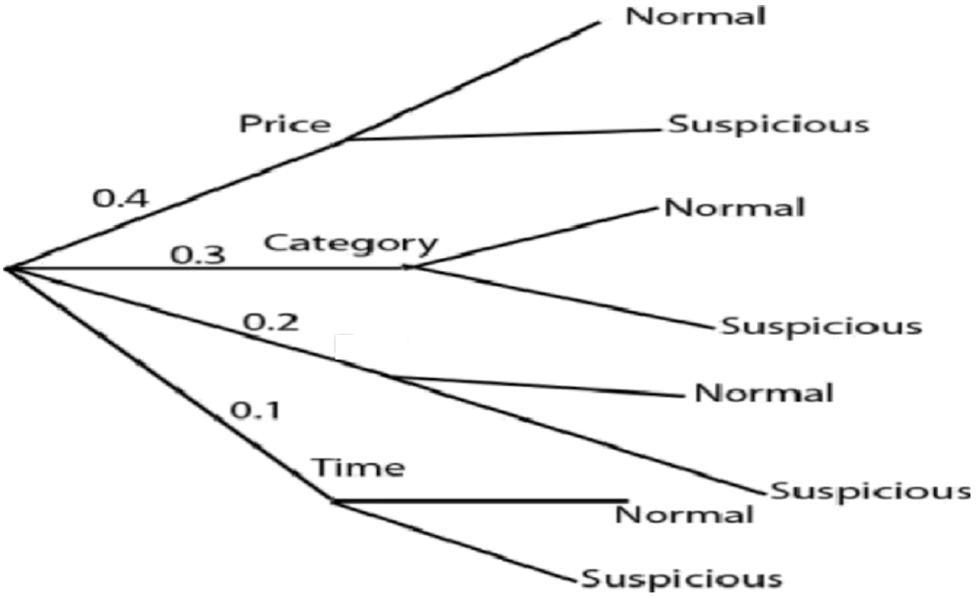

DSA
Normal/ Suspicious Analysis
Threshold
By Coditional Probability (Bayesian Theorem)
Normal / Suspicious

Figure III. DSA



We can represent result by bar chart.

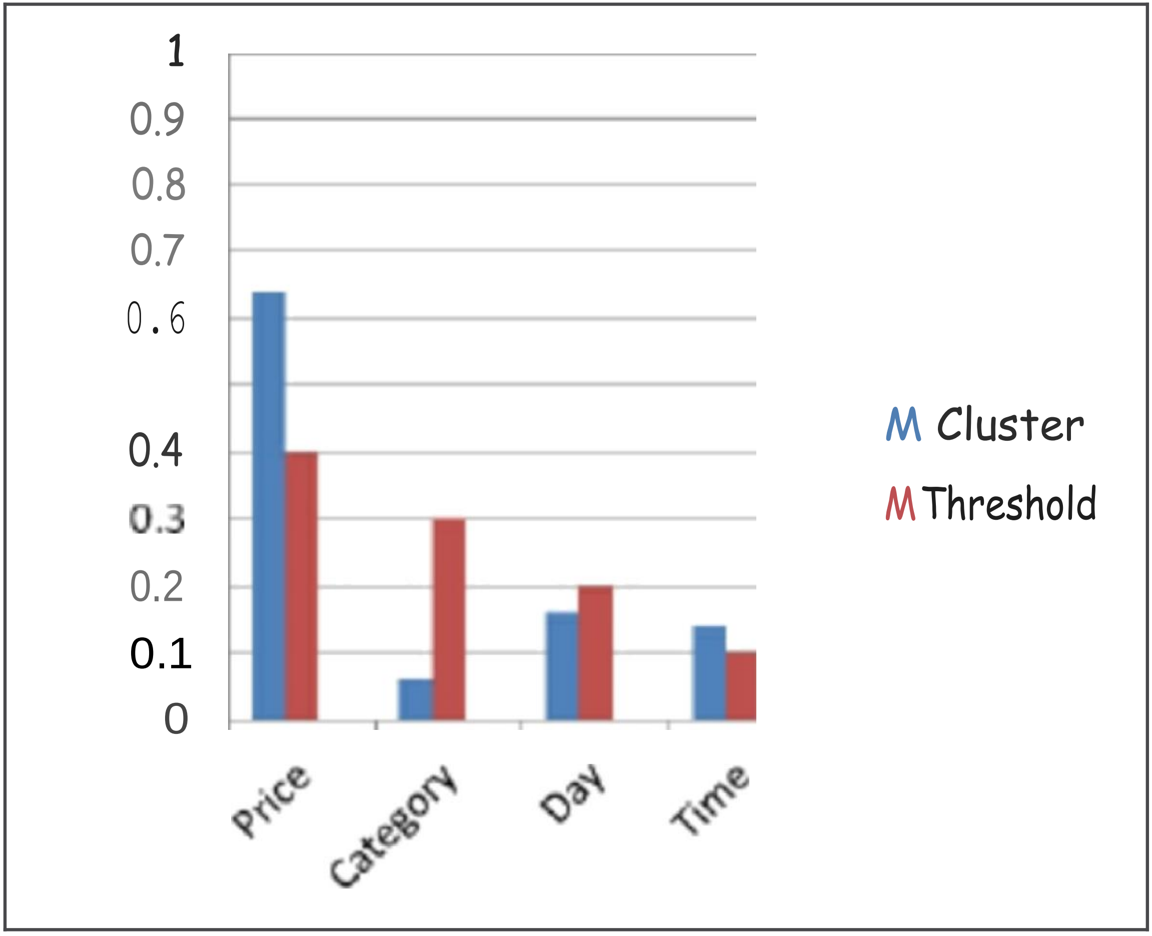


Figure V. Bar Cha8 Representation Of Reult

PROPOSED SYSTEM

The proposed model consisted of six steps.

Firstly, Luhn's Test is used to validate card numbers. Then, two rules ie. Address Mismatch and Degree of Outlierness are used to analyze the deviation of each incoming transaction from the normal profile of the cardholder. These two steps compute initial beliefs. The initial belief values are combined to obtain an overall belief by applying Advanced Combination Heuristic in step four. Step five looks into the spending history to extract characteristic information about genuine and fraud transactions. The overall belief is further strengthened or weakened in the final step using Bayes‟ Theorem, followed by recombination of the calculated probability with initial belief of fraud using advanced combination heuristic.

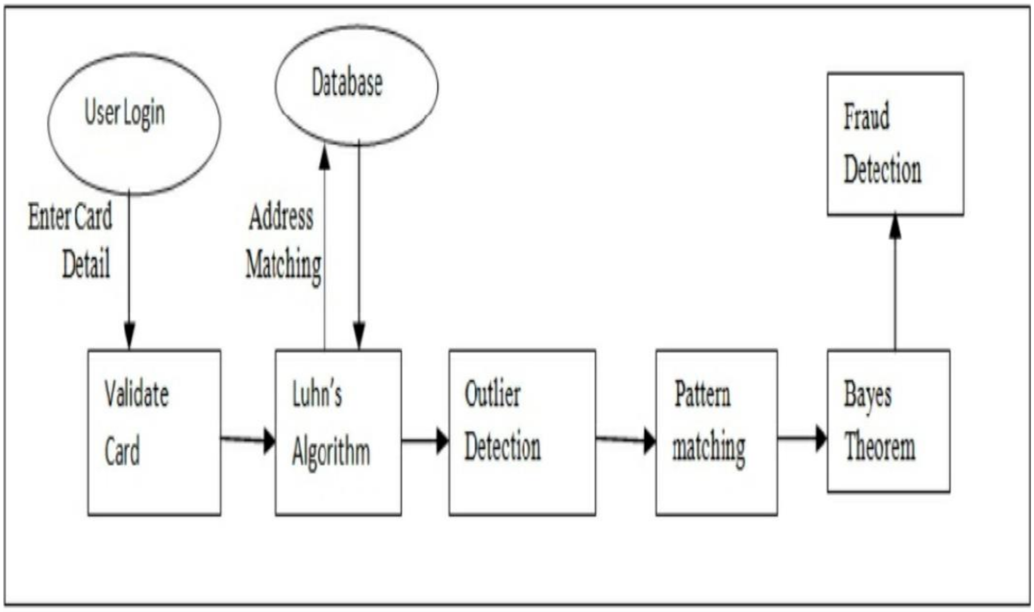


Fig 1: Architecture Diagram

Card Number Validation

Luhn's Algorithm is used to validate card numbers that distinguish valid numbers from mistyped or otherwise incorrect numbers.

Following standard algorithm is used to validate credit card numbers, [14]

1. Reverse the order of the digits in the number.

2. Take the first, third, ... and every other odd digit in the reversed digits and sum them to form the partial sum S1.

3. Taking the second, fourth and every other even digit in the reversed digits. Multiply each digit by two and sum the digits if the answer is greater than nine to form partial sums for the even digits.

4. Sum the partial sums of the even digits to form S2. If S1+S2 ends in zero,then the original number is in the form of a valid credit card number as verified by the Luhn test.

For example, if the trial number is 49927398716,

1. Reverse the digits: 61789372994

2. Sum the odd digits: 6 + 7 + 9 + 7 + 9 + 4 = 42 = s1

3. The even digits: 1, 8, 3, 2, 9 Two times each even digit: 2, 16, 6, 4, 18 Sum the digits of each multiplication: 2, 7, 6, 4, 9

4. Sum the last: 2 + 7 + 6 + 4 + 9 = 28 = s2

5. S1+S2 = 70 which ends in zero which means that 49927398716 passes the Luhn's test.

Algorithm I

1. First rcmo vc spaccs / hyphcns.
2. Find the length of card number (Input).
3. Find parity / Checksum / check digit

   Parity = Length % 2
4. Define total = 0 (Input)

1. Then we move as —

```
For (I = 0; I < lenp•th; I ++)

Digit — number | I]
If (I %« 2 == parity)

      Digit *- 2
      If (digit > 9)
      Digit - = 9
   )
Total + = digit
)
((total One 10) == O)? TRUE: FACS E
```

Example:

Card Number = 41 S1 3835 0000 0140

Length = 1 6

Tratal - ()

Parity = 16 % 2 = 0

If total % 10 - 0 thcn ctird is valid according to Luhn Algorithm else invalid card.

Table II. VALUE FOR CARD UIIfBERVALIDATION

i	Card number	Oigit
0	4	8
2	8	7
H	1	1
6	3	6
7	9	9
s	0	0
9	0	0
10	0	0
11	0	D
i z	0	0
i 3	1	1
14	4	8
is	0	0

Here total = 50 so in current example card i.s “valid”.

Address Verification

This step is used for comparing Billing Address with Shipping Address and the check is whether it matches or not. This check does not guarantee whether a transaction is fraud or genuine. But if the two addresses match,t he transaction can be classified as genuine with a high probability. Else, the transaction is labeled as suspect.

Outlier Detection

We have used DBSCAN (Density Based Spatial Clustering of Application with Noise) to generate clusters, using transaction amount as attribute. Any incoming transaction amount that does not belong to any cluster is detected as fraudulent. These two steps compute initial belief.

Advanced Combination Heuristic Function

The initial belief values are combined to obtain an overall belief.

Spending History Databases

It comprises genuine Transaction Record(for individual customers from their past behaviour) and Fraud Transaction Record(from different types of past fraud data).We represent each history transaction by a set of attributes containing information like card number, transaction amount and time since last purchase. to extract characteristic information about genuine and fraud transactions.

Bayes Theorem

The idea of belief revision is that, whenever new information becomes available, it may require updating of prior beliefs. Bayes Theorem theorem expresses how a subjective degree of belief should rationally change to account for availability of related evidence.

4. RESULTS/PERFORMANCE ANALYSIS

ONE TIME PASSWORD

One time password [10] is one of those changes every time while using.

One time password is very important for authentication because an intercept static password is useless because it cannot be reused.

One time generated password SMS is sent to users register mobile number. User puts that password to our system and authenticates.

By studying this paper we get to know many things about this domain, its importance and among other things, the impact that a safe and sound community makes a difference.

However, it is important to note that the customers are not necessarily active during the whole period. In fact, some of them perform transactions only in the first part of the considered time frame, others only at the end, and others in the middle. Our approach based on the ARIMA model requires sufficient legitimate transactions in the training set in order to learn the legitimate behaviour of the customers. In addition, our approach requires at least one fraud in the testing set to evaluate the performance of the model. In this context, initially, we propose to split the dataset into the training and testing set with a 70–30 ratio.

With this setting, there is at least one fraud in the testing set and no fraudulent transactions in the training set, but, unfortunately, this reduces the number of customers’ time series from 24 to 9. The composition of the final 9 time series that are used in the next section. The last column indicates the number of frauds over the total number of transactions occurring on the same day; as can be seen, only in one of the time series (number 10) do frauds occur on two different days.

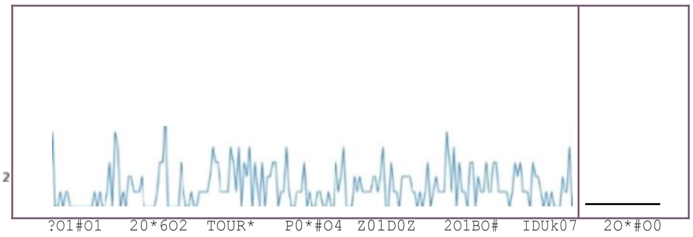


Figure 1. Plot of daily number of transactions for a customer in the dataset. Legitimate transactions are represented by the blue dot, whereas fraudulent transactions are represented by the red dot.

Table 1. Frequency of fraud and legitimate tnnsacaons in the whole dataset.

	Legitimate	£’raué	**Totnl**
Number	11,384	87	11,471
Percentage	99.24%	0.76%	100%

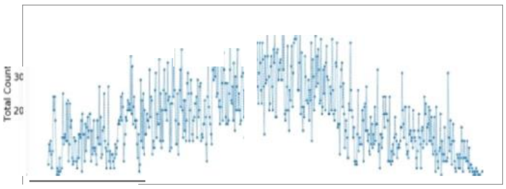
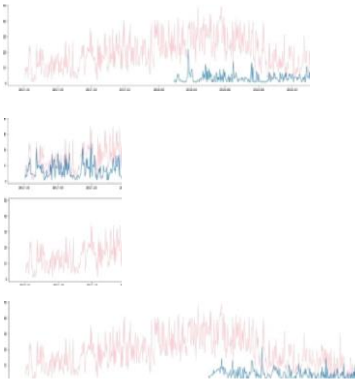


Figure 2. Number of daily transactions summing up all customers.

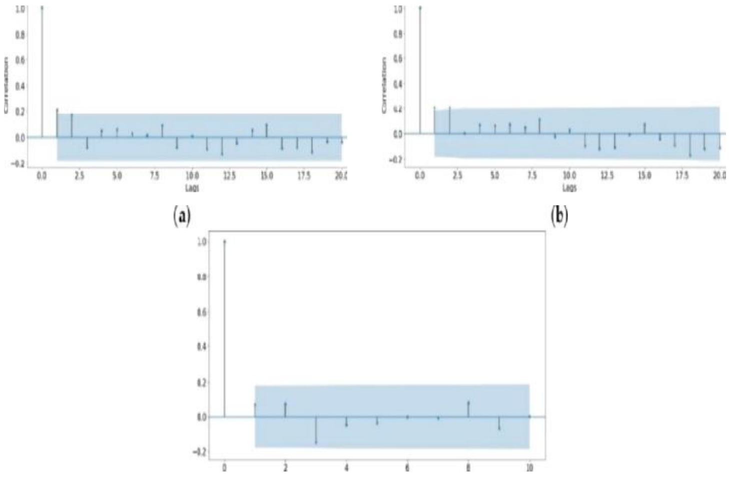
T*bl‹ 35AbésdAD5mWuBamairfg4tnme.

p-value	3.18O17Df5x10-'



fi$«e 3. Number of daily barisacbon6' Jfle blue éot tep sents a 8pgrtfic rusto»ia uid tie red dot

flmeSeñe€D	0D*y£inWim	#Ihy inTee0	FravdPepeWbn
	192	83	1/14
	193	84	1/3
7	186	80	1/11
9	164	71	8/2t
10	193	84	I/17
t5	191	82	t/t\ and I/2



figatt(a)7tttialtvboazk‘tonyl tbruopk ‘kiexée;(¥)tvBaeldonyl Ibiuzylr *kzretiu;(c)oiel za

Data analysis

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make a valid prediction (Edelstien, 1999).

The six basic steps of the data mining process are defining the problem, preparing data, exploring data, building models, exploring and validating models, deploying and updating models.

Fraud detection solutions are used in the fields like credit card, e-commerce, telecommunication, insurance etc in order to protect personal information of customers.

The two main disadvantages involved in the research process of fraud detection in data mining are lack of personal information for conducting experiments and lack of well developed techniques and research methods.

In order to overcome these issues proper categorization and comparison has been done using similar literature in this paper.

SECURITY QUESTION

A security question [11] is used as an authenticator by banks, many other companies as an extra security layer. They are a form of shared secret. Financial institutions have used questions to authenticate customers since at least the early 20th century.

In a 1906 speech at a meeting of a section of the American Bankers Association, Baltimore banker William M. Hayden [12] described his institution's use of security questions as a supplement to customer signature records.

In the 2000s, security questions came into widespread use on the Internet. As a form of self-service password reset, security questions have reduced information technology help desk costs. By allowing the use of security questions online, they are rendered vulnerable to keystroke logging attacks. In addition, whereas a human customer service representative may be able to cope with inexact security answers appropriately, computers are less adept.

As such, users must remember the exact spelling and sometimes even case of the answers they provide, which poses the threat that more answers will be written down, exposing them to physical theft.

5. CONCLUSION

In this paper we have a brief discussion on credit card fraud detection. Here we have shown how the system detects whether an incoming transaction is fraud or genuine.

In our proposed model, we have found out validations of cards are genuine and very low false alarm.The relative studies and our results are sure that the correctness and effectiveness of the proposed system is secure.

The day by day credit card use is increasing online and offline. So according to that, credit card fraud also increases. Every bank, finance company and other finance related institutes require this system.There are a number of techniques present to implement this system.

The objective of this survey was to review the major research in the area of intrusion detection using the Dempster-Shaffer theory of evidence. Most of the researchers have discussed the resolution of various issues and intended future work in this area. It is a very fast and effective method using minimum effort so widely preferred.

Credit card fraud detection has drawn quite a lot of interest from the research community and a number of techniques have been proposed to count credit fraud. Bayesian learning takes place so that the FDS dynamically adapts to the changing behavior of genuine customers as well as fraudsters over time.

Dempster–Shafer theory gives good performance, especially in terms of true positives, Bayesian learning helps to further improve the system accuracy.

Finally Fraud detection system gives more performance in terms of

[1]