EXTENDED CENTROID BASED CLUSTERING TECHNIQUE FOR ONLINE SHOPPING FRAUD DETECTION

Priya J Rana[1], Jwalant Baria[2]

[1] *ME IT, Department of IT, Parul institute of engineering & Technology, Gujarat, India*
[2] *ME CSE, Department of Computer science & Engineering, Parul institute of engineering & Technology, Gujarat, India*

ABSTRACT

with the enhancement in technology on-line banking like credit Card and Debit Card, Mobile Banking and most useful Internet Banking is the popular medium to transfer the money from one account to another is call customer to business. online Banking is gaining popularity day by day is very usefully for everyone, which increases the online transaction with the increase in online shopping, other charges, so the fraud cases related to this are also increasing and it puts a great burden on the economy, affecting both customers and company. It not only costs money, but also a great amount of time to restore the harm is all so done. The purpose is to prevent the customer from online transaction by using specific technique i.e. based on Data Mining. Customer Spending Behavior of product parching and payment transfer. The customer than spending behavior about the product that can be identified by searching approach by using K-Mean or k-Medoid clustering algorithm. If risk score is greater 0.75 then transaction is taken as to be fraudulent and then the security mechanism authenticates the user by entering the random number generate digit on the screen and the genuine user enters the code in a correct manner.

Keyword: - *Online Shopping, Data mining techniques, Clustering, K-medoid Algorithm*

---

1. INTRODUCTION

Today technology is basic mandatory need of human in online shopping. Just look around and you will know why. Literally, at every instant of time, you are surrounded by advance 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 of online shopping, but with that it causes Fraud also. Fraud is behavior of human which is out of rule and causes crime. One of the biggest facility provided by technology is that we can able do a shopping using various facility provided by bank e.g Credit Card, Debit Card, Internet Banking [1] etc. Here is 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 and debit card. Now bank 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, debit/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, Nearest neighbour 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 is having its own pros & corns.

2. FRAUD TECHNIQUES

With the old days of timing fraudster were using traditional techniques such as application fraud and Lost and stolen cards. But now a day fraudster becomes too much smarter than the previous time. They are very innovative and fast moving kind people. With the modern technology they are using fake and doctored cards with the skimming techniques.

Skimming: Skimming (fraud) is the theft of money from a business prior to its entry into the accounting system for that company. Skimming is one of the smallest frauds that can occur, but they are also the most difficult to detect[1].

Taking examples of Skimming fraud, Grant owns a sandwich shop named Real Good Wich. Lately, Grant has noticed that the cash account has been dwindling, and decides to hire Mason a CFE to investigate. Mason arrives on the scene and orders a sandwich. After observing the cashier for a while he notices that one of the employees is pocketing the cash when the exact amount for a sandwich is paid. Mason takes note of this and returns to Grant and explains the problem. He explains that when the exact amount is paid the employee can simply pocket the cash because there is no need to open the register for change and no sale needs to be recorded. Mason also provides a quick fix to the problem. If Grant were to provide a free sandwich to his customers if no receipt were given to them on the sale it would significantly reduce the amount of skimming. Grant takes Mason's advice and fires the employee, and implements the free sandwich. After a while Grant sees that his cash has increased as well as overall profits [1].

3. DATA MINING TECHNIQUES

Fraud Detection has been usually seen as a data mining problem where the objective is to correctly classify the transactions as legitimate or fraudulent.

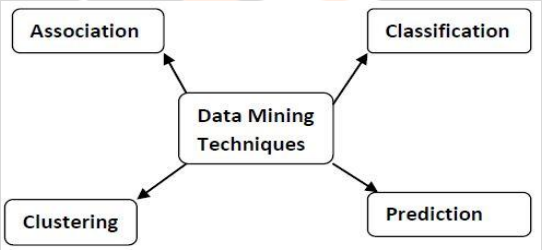

Fig -1: Data Mining Techniques

3.1 Association: In association, patterns are discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together.

3.2 Classification: Classification is a data mining technique i.e. based on machine learning. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. For example, Classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, classification divides the employee's records into two groups that are "leave" and "stay".

3.3 Clustering: A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. For example, in a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. In clustering technique, the books that have some kind of similarities in one cluster or one shelf and the books that is different in other cluster or in another shelf.

3.4 Prediction: The prediction is one of a data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables. In this technique, prediction analysis

technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable.

4. DETECTION TECHNIQUES FOR CREDIT CARD FRAUD

4.1 Decision tree: The idea behind this technique is that of a similarity tree created using decision tree logic. In this case, a similarity tree is defined recursively; the nodes are labeled with the use of attribute names, edges are labeled using values of attributes, and then there are the leaves, which contain an intensity factor that is defined as the ratio of the number of transactions that satisfy the outlined conditions [2]. The main advantage of this method of fraud detection is that it is easy to implement, understand and display. However, there are disadvantages when you are forced to check every transaction one by one. Nevertheless, over the years, similarity trees have been able to offer tangible results. Therefore, it is an effective method when it comes to fraud detection [2].

4.2 Genetic algorithms and a range of additional algorithms: In most instances, algorithms are recommended as predictive methods or means of fraud detection. Most of the algorithms are created to establish logic rules that classify credit card transactions into suspicious and non-suspicious classes. Basically, this method tends to follow suit on the scoring process [2]. This credit card fraud detection method has been proved to deliver results when it comes to giving credible home insurance data. It might be a more efficient method when it comes to detecting and countering credit card fraud. This technique also incorporates a range of methods that are used to predict any suspicious behaviors [2].

4.3 Clustering techniques: There are two main clustering techniques that are used to detect behavioral fraud. Peer group analysis is a system that allows for the identification of accounts that are behaving differently from one another at any moment in time, particularly when they were behaving the same previously. Those accounts are flagged as being suspicious. Fraud analysts can then proceed to investigate such discrepancies [2]. The hypothesis in these clustering techniques is that if accounts behave the same over a certain period of time and then a certain account starts to behave significantly differently, the account holder should be notified. Some signals of suspicious behavior are a sudden high-dollar transaction and a high frequency of usage of a credit card [2].

4.4 Neural networks: Neural networks are also recommended as effective credit card fraud detection methods. The only issue with this method is that all data has to be clustered by the type of account it belongs to. Credit card fraud is a major issue that if not dealt with effectively, it can result in myriad complications. It is vital to try and find ways of detecting the issues and resolving them as soon as they arise [2].

5. RELATED WORK

Clustering is a process of arranging data into groups of similar objects. Different grouping results are obtained from various clustering methods available to group the dataset. The choice of a particular method will depend on the desired output.

The clustering methods are [3]:

- Hierarchical methods
- Partitioning Methods
- Density-based Methods
- Grid-based Methods
- Model-based Method

Clustering technique is unsupervised technique which is useful when there is no prior knowledge about the particular class of observations in a data set. K-Means clustering is a simple and efficient method to cluster the data [7]. As the real data set is not available, here the assumption is made to generate the data set for the transaction randomly as shown in the fig. The data table is generated for card no, Customer Name, Item Name, Order_ID, Amount, Date, Time and City.

Table -1: Dataset

No.	Card No	Customer Name	Item Name	Order_id	Amount	QTY	Total amount	Date	Time	City
1	125887419632	Priya Rana	Black Cotton Kurti	20130101	550	1	550	30/01/2013	17:15:24	Vadodara
2	587964583648	Dipika Raval	AISH SABIA C VIS-2	20130102	1425	1	1425	30/01/2013	17:55:01	Delhi
3	65874932589647	Jagu Patel	RANI RED ROSE	20130103	987	1	987	30/01/2013	19:16:52	Bangalore
4	58961453698745	Hiral Bhavsar	KARINA KAPOOR	20130104	450	1	450	30/01/2013	22:45:01	Hyderabad
5	125489635847	Prerak Parikh	White T-shirt	20130105	698	1	698	31/01/2013	10:15:01	Ahmedabad
6	36578954236987	Shruti Shah	BIPASHA LOVELY BEAUTY	20130106	355	1	355	31/01/2013	12:59:08	Chennai
7	951753258456	Ami Makadia	SONAKSHI RAINBOW	20130107	645	1	645	31/01/2013	17:19:45	Kolkata
8	456852321486	Rashmi Goswami	SRI DEVI PINK BEAUTY	20130108	700	1	700	31/01/2013	19:25:32	Surat
9	85489625863279	Rashmi Bhavsar	SONAM HIGHLIGHT	20130109	713	1	713	31/01/2013	19:52:02	Pune
10	45695275648235	Laksh Maheswari	Fasttrack watch	20130110	800	1	800	31/01/2013	20:35:07	Jaipur
11	45869821536875	Sushant Rajput	Puma shocks	20130111	666	1	666	31/01/2013	21:09:35	Lucknow
12	741369456951	Helly Shah	Priya Maglo Lady	20130112	649	1	649	31/01/2013	23:00:36	Kanpur
13	110919912013	Namish Taneja	lumia 730	20130113	12000	1	12000	10/2/2013	9:16:52	Delhi
14	250119902009	Harvi Sisodiya	DIPIKA RAMLEELA	20130114	1000	1	1000	10/2/2013	10:16:52	Indore
15	85742536159875	Preksha Bhojani	MADHURI RAJGHRANA	20130115	1500	1	1500	10/2/2013	11:16:52	Thane
16	109638075600	Ankita Lokhande	MANDIRA BLACK BEAUTY	20130116	1269	1	1269	10/2/2013	12:16:52	Bhopal
17	900684089614	Priya Tripathi	VIDHYA GREY BEAUTY	20130117	1589	1	1589	11/2/2013	13:16:52	Visakhapatnar
										Pimpri &

- **Order_ID:** is automatically generated and incremented sequentially by the tool.
- **Card_No:** The different credit cards numbers are manually given for transaction. We used five credit card numbers manually.
- **Date:** The date on which the amount is deposited or withdrawn.
- **Amount:** is entered manually which cannot be generated automatically as amount is deposited or withdrawn.

Most of the clustering techniques have implemented K-mean clustering algorithm to detect a fraud. But here we are proposing to use K-medoid algorithm instead of K-mean algorithm. Because K-mean having a disadvantage that it is sensitive to outliers since an object with an extremely large value may distort the distribution of data [8]. So taking the mean value of the objects in a cluster as a reference point, a medoid can be used, which is the most centrally located object in a cluster. Thus, the partitioning method can still be performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point.

6. PROPOSED WORK

Bayes Theorem: this approach using for distributed grid system purpose only Approach a theorem of probability theory originally stated by the Reverend Thomas Bayes. It can be seen as a way of understanding how the probability that a theory is true is affected by a new piece of evidence.

Using this idea of conditional probability to express what we want to use Bayes Theorem to discover, we say that P(Ai|Bn), the probability that A is true given that B is true, is the posterior probability of Ai. The idea is that P(Ai|Bn) represents the probability assigned to A after taking into account the new piece of evidence, Bn. To calculate this we need, in addition to the prior probability P(Ai), two further conditional probabilities indicating how probable our piece of evidence is depending on whether our theory is or is not true.

Algorithm I (checking card)

1. First remove spaces/ history / hyphens.
2. Find the length of card number or bank name (Input).
3. Find the online shopping website and product (Input).
4. Find the Price of product and Qty of Product.
5. Find parity of number of time pursuching product.
6. Find parity / Checksum / check digit
   Parity = Length % 2
7. Define total = 0 (Input)
8. Then we move as –
   For (I = 0; I < length; I ++)

```
 {
   Digit = number [I]
If (I % 2 == parity)
{
Digit *= 2
If (digit > 9)
Digit - = 9
 }
Total + = digit
 }
((total % 10) == 0)? TRUE: FALSE
```

Example:

Card Number = 3456 2456 1235 2784 Length = 16
Total = 0
Parity = 16 % 2 = 0

If total % 10 = 0 then card is valid according to Algorithm else invalid card

For this we are considering four clusters like “Price”, “Category”, “Day” & “Time” as shown in figure II. Our work is carried out by considering last ten transactions and gets the probability of each cluster over “Normal” and “Suspicious”.

K-Medoid: this approach using for searching purpose only Approach this concept forms the basis of the k-Medoid method. The basic strategy of k- Medoid clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the medoid) for each cluster. Each remaining object is clustered with the medoid to which it is the most similar. The k-Medoid method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects [11]. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k number of clusters. This k: the number of clusters required is to be given by user. This algorithm works on the principle of minimizing the sum of dissimilarities between each object and its corresponding reference point. The algorithm randomly chooses the k objects in dataset D as initial representative objects called medoid. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set. Then for all medoid, after every assignment of a data object to particular cluster the new medoid is decided.

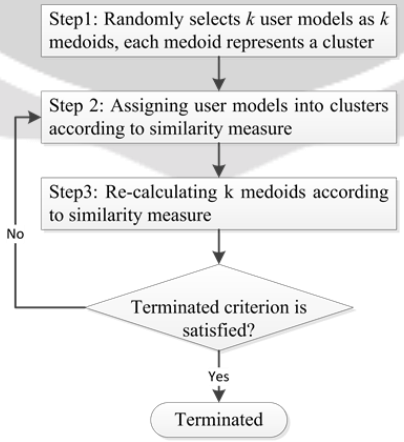

Fig -1: K-medoid architecture

Algorithm

Input:
k: the number of clusters,
D: a data set containing n objects.

Output: A set of k clusters.

Method: Arbitrarily choose k objects in D as the initial representative objects or seeds;

Repeat assigns each remaining object to the cluster with the nearest representative object;

Randomly select a nonrepresentative object, orandom;

Compute the total cost, S, of swapping representative object, oj, with orandom;

If S < 0 then swap oj with orandom to form the new set of k representative objects;

Until no change;

7. CONCLUSIONS

These shopping cart applications typically provide a means of capturing a client's payment information, but in the case of a credit card or Debit card they rely on the software module of the secure gateway such type provider (such as CCAvenew, PayPal, Bill desk or other bank payment gateway), in conjunction with in the secure payment gateway, in order to conduct secure credit card transactions online shopping . Some setup must be done in the HTML code or depend of shopping product company web side of the website, and the shopping cart software must be installed on the server which hosts the site, or on to the secure server (https protocol) which accepts sensitive ordering write information. Later to at the process of finalizing the transactions, the information is accessed and an order is generated against the selected item thus clearing the shopping cart and product on its Qty.

REFERENCES

[1] http://strategiccfo.com/wikicfo/skimming-fraud/

[2] http://www.mydigitalshield.com/credit-card-fraud-detection-techniques/

[3] Celebi, Kingravi and Vela, "A comparative study of efficient initialization methods for the K-Means clustering algorithm", Expert systems with applications, 2013.

[4] Sailesh S. Dhok, "Credit Card Fraud Detection using Hidden Markov Model". 2012 IJSCE.

[5] Neha Sethi, Anju Gera, "A Revived survey on Various Credit Card Fraud Detection Techniques", 2014 IJCSMC.

[6] John Akhilomen, "Data mining Application for Cyber Credit-Card Fraud Detection System", 2013 WCE.

[7] Vaishali, "Fraud detection in Credit Card by Clustering Approach", 2014 IJCA.

[8] Noor Kamal Kaur, Usvir Kaur, Dr. Dheerendra Singh, "K-medoid Clustering Algorithm- A Review", 2014 IJCAT.

[9] Pooja Chougle, A.D. Thakare, Prajkta Kale, Madhura Gole, Priyanka Nanekar, "Genetic K-means Algorithm for Credit Card Fraud Detection", 2015 IJCSIT.

[10] Mahesh Singh, Aashima, Sangeeta Raheja, "Credit crad fraud detection by improving K-means", 2014 IJETR.

[11] MohdAvesh Zubair Khan, Jabir Daud Pathan, Ali Haider Ekbal Ahmed, "Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering", 2014 IJARCCE.