Unsupervised Learning for Credit Card fraud detection

Professor. Vikrant Agaskar[1] , Megha Babariya[2], Shruthi Chandran[3] , Namrata Giri[4]

[1]*Professor, Department Of Computer Engineering, VCET, Mumbai University, India*
[2]*Student, Department Of Computer Engineering, VCET, Mumbai University, India*
[3]*Student, Department Of Computer Engineering, VCET, Mumbai University, India*
[4]*Student, Department Of Computer Engineering, VCET, Mumbai University, India*

---***---

Abstract - *Today, Internet banking has led to an increase of frauds, resulting in substantial financial losses. Banking frauds increased 93% in 2009-2010, and 30% in 2012-2013. Internet banking frauds are difficult to analyze and detect because the fraudulent behavior is dynamic, spread across different customer's profiles, and dispersed in large and highly imbalanced data sets. Customers do not check their banking history daily to analyze any kind of fraud. We propose in this paper a technique of synthetic model of the data structure for efficient storage of data, and a measure of dissimilarity between these representations for the detection of change in the stream structure , in order to detect different types of fraud during a period of time.*

***Key Words*: Clustering , fraud detection , Transaction , Unsupervised Learning , data mining algorithms.**

1. INTRODUCTION

In this digital age, it is the time of online banking, one of the well-organized and easier modes of transaction. With the evolution of internet in the banking sectors, people have changed the way they used to bank. But this digital transformation is providing new ways for fraudsters to hack people's private accounts. Banking sector frauds have been in existence for centuries, with the earliest known frauds pertaining to insider trading, stock manipulation, accounting irregularity/ inflated assets etc. Fraud is a superior form of white collar crime that persist to extract a significant toll not only on the organizations, but also on investors, financial institutions, and the economy in general.

Most IT entities today in use are transactional. This means that the data transactions are processed in the system and the data of transaction is stored in the system's database. The relationships and patterns in these reserved transactional data is analyzed by the data mining softwares. One of the major issues during this is the fraud which may occur during transactions..

Databases constantly keep on changing and the size of data keeps on increasing . This makes the transformation and mass of data so vital that it is impossible to store them in a database. Here the 'data stream analysis' is done. The upcoming of new data is endless. Thus, effective algorithms must be able to work with a constant memory footprint, regardless of the evolution of the stream, as the whole database cannot be kept in memory.

There are many issues that make effective fraud management a challenging task. These include: large and ever-expanding mass of data, the growing complicatedness of systems, changes in business processes and activities and continuous transformation of new fraud schemes to avoid the existing detection techniques. To detect the fraudulent financial statements is a difficult job when using normal audit procedures due to limitation in understanding the characteristics of financial statements, lack of experience and dynamically changing strategies of fraudsters.

Supervised methods, using samples from the fraudulent/non-fraudulent classes as the basis to construct classification rules to detect future cases of fraud, to prolong from the problem of unbalanced class amount: the legitimate transactions are more in number then the fraudulent ones.

In this paper, we propose an unsupervised method for detecting fraudulent transactions using records of the amount and location details of previous transactions carried out by the customers.

2. RELATED WORK

Construction of a synthetic model at daily intervals over a data repository which vacants itself as new data is stored. This synthetic representation is being derived from the learning of a weighted SOM (Self-Organizing Map) and admits automatic data clustering. During the learning procedure, each pattern is extended with novel information uprooted from the data.

First, there is a shortage of information concerning the characteristics of management scam. Secondly, with the irregularity it has, most of the auditors need the necessary experience to detect it. Finally, managers intentionally try to mislead the auditors. These limitations demonstrate that there is a need for supplementary analytical procedures which can effectively detect and manage fraud. Fraud classification model using neural networks has been developed. Neural, operation based, real-time fraud detection systems are not only technically feasible, but highly interesting from a purely economic point of view. Neural network with statistical methods have been used to detect fraud. Statistical regression analysis and statistical method of logistic regression have also been tried to detect fraud in banks.

3. VARIOUS METHODS TO DETECT CREDIT CARD FRAUD.

There are various emerging technologies that are capable of detecting the credit card fraud. Some of technologies that will work on some parameters and able to detect fraud earlier as well are listed below:

Learning: Learning is generally done with or without the help of teacher.Generally division of learning take place as shown in Figure 1.

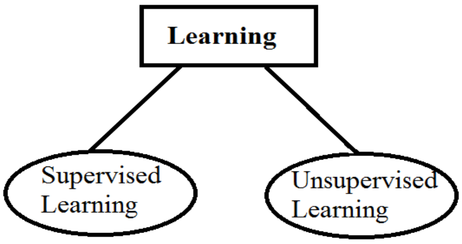

Figure 1: Types of learning.

Supervised Learning : The learning that takes place under the guidance of a teacher is said to be as supervised learning. But in which there is no guidance of teacher is termed as Unsupervised learning. These Learning are explored as:

Decision Tree Induction:

Dipti Thakur & Bhatia defined this as a type of supervised learning in which we make a decision tree to reach at a appropriate solution. As shown in figure-2 they explained that in decision tree we have some interior nodes and each node symbolize a test on a particular attribute and each branch in decision tree represent an outcome of test and each leaf node will represent class label means output. Decision trees are being used in the situation where the classification in which a new transaction is given with the class label as unknown, which means that it is not known whether the transaction is fraudulent or legitimate and the transaction value obtained is tested against the decision tree. A path is traced from root node to output label for that transaction.

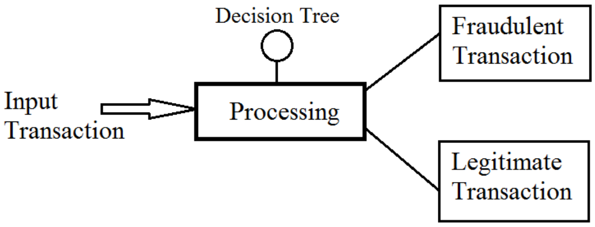

Figure 2: Decision Tree Induction.

Unsupervised Learning:

Peer Group Analysis Approach :

In 2008, Watson and Whitrow et.al worked on Peer Group Analysis in Plastic Card Fraud Detection. They defined it as an unsupervised method that analyzes the behavior over time by monitoring it. This approach can be used to identify credit card fraud detection by analyzing the fraudulent transactions. In this those transactions deviate from their peer group termed as anomalous/fraudulent transactions. They defined that there are generally two type of approach to detect fraud. One is, in which form of the fraud is known, this can be detected by pattern matching. And when the type of fraud is not known then we approach anomaly detection techniques. Peer group analysis is an anomaly detection technique. Suppose we have (a1,a2.......................an-1, an) time series representing the weekly amount spent on a credit card from a particular account and 'an' is the target amount. We wish to determine whether the recently spent transaction 'an' is fraudulent or not at time t=n. In this, in order to detect outlier transaction we can use the Mahalanobis distance of 'an' from the centroid of its peer group. As it is shown in figure 3 that by the help of the Peer group technique anomaly input are separated.

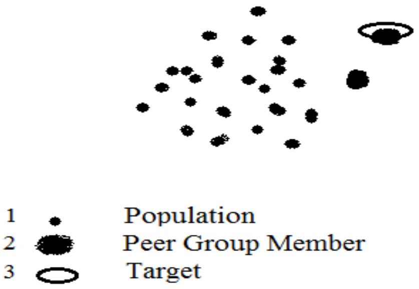

Figure 3 : Peer group Analysis

Unsupervised Learning in Neural Network:

Self Organinsing Map In resaerch of Kuah & Sriganesh Self organizing Map configure its neuron according to topological structure of input data as shown in Figure 4. This process is called Self Organization because of iterative tuning weight of neurons. The result is clustering of input data.In Self Organizing map Zaslavsky and Strizhak[8] defined that we need no external teacher in this mode of learning. So, verification of resultant matrix will be done on behalf of the presented past learning.As our approach is also based on the Unsupervised learning.There are mainly four step of processing take place in Self organizing Map that we applied on our inputs.

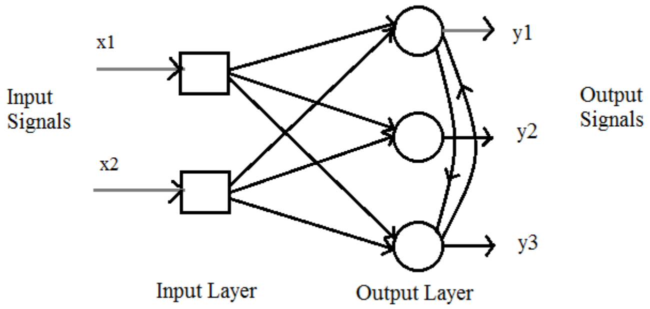

Figure 4: Self organizing Map

1) Initialization : First of all we have to select small random value for synaptic weight in interval[0,1] and have to assign a small positive value of learning parameter 'a'.

2) Activation and the Similarity matching: Here in this step we activate the Kohonen network for input vector X and find the Winning neuron.

3) Learning of Adaptive: Weights are trained by performing various number of steps.

4) 4) Iteration: At that point we performed iterations,till we did not get a stability in our network.

4. OUR APPROACH

A variety of techniques can be applied to answer the presented problem. The most simplest method used in the earliest transaction monitoring systems was control of transaction parameters. Once we have obtained the clusters by using SOM technique we revalidate our clusters by using association rules on each cluster. To apply association rules we need to provide categorical data as input, so we convert numeric data into categorical data based on few criteria as in our case:

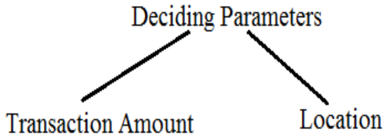

Figure 5: Parameters used in design

Mathematically,

|Lg-Ug| will be associated parameter to the range over expected transaction.

To compute these we will relate each tuple in the Database to a certain amount thus initializing or updating its threshold.

$$\text{Vector\_distance}_{ij} = \sum_{j=i-1}^{j=i-10} transaction\_value[j]$$

Temp=max({parameteric_constant transaction$_i$},Vector_distance$_{ij}$)

Transaction_control_Flag=val(θ,temp)

Where,

θ is threshold value

val(p1,p2) returns binary value after comparing and temp

parametric_constant is mean of transaction attribute for all tuples of the projected customer

Association rules are the if-then statements that support in determining the relationships among attributes in unrelated data of a database. Relationships between objects which frequently occur together are identified by association rules. Support and confidence are two primary criteria used by association rules. They help in identifying the relationships in the data by analyzing on frequently used if/then patterns. Association rules to satisfy minimum support and minimum confidence is provided by user simultaneously. Support gives an idea of the frequency with which the items appear in the data, and confidence describes the proportion in which the if/then statements have been found to be true.

5. CLUSTERING ANALYSIS

Let universe of discourse X={$x_{1,}x_{2,}$.... x_n}be the object to detect. Every object has m indexes, namely x_i ={ x_{i1}, x_{i2}, ...x_{im}} , i= (1,2, ... n) .The following data matrix shows it in the following way:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \qquad (1)$$

Now we are required to figure out the outlier sets of n objects. To judge the diffusion degree of every object in X, We comprises the d_{ij} which denotes the distance between any two objects and composes distance matrix R, described as follows,

$$R = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \qquad (2)$$

It is very important to select any distance function. This paper selects the Euclidean distance.

$$\mathrm{DIS}(U, f) = \sqrt{\sum_{i=1}^{n}(o_i - f_i)^2} \qquad (3)$$

$$\text{Let } P_i = \sum_{j=1}^{n} d_{ij} \qquad (4)$$

Where P_i is the sum of ith row in matrix R. The bigger P_i is, the longer the distance between i object and other object is. Then P_i is said to be the candidate item of outlier set.

$$\lambda_i = \frac{p_i - p_{\min}}{p_{\min}} \times 100\% \qquad (5)$$

Where, λ denotes threshold and the objects with $\lambda_i \geq \lambda$ are taken to be the outlier set.

6. FUTURE SCOPE

With increasing number of bank fraudlency and cyber crime cases need of a secure testing system is on rise. And this is a direct solution to this problem. It can be extended to a duplex verification of not only a customer(debit-ant) but also of the seller(credit-ant). It can be taken and used on a regular basis just like OTP. It can be used to even assess past transaction in database to find whether certain transactions were fraudulent or not and also would be able to produce evidence in such cases.

7. CONCLUSION

This article has proposed a new approach to transaction monitoring and credit card fraud detection using unsupervised learning . It enables automated creation of transaction monitoring rules in a learning process and makes possible their continuous improvement in an environment of dynamically changing information in an automated system.

REFERENCES

[1] G. Cabanes and Y.Bennani, "A local density-based simultaneous two-level algorithm for the topographic clustering, " in Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2008, pp. 1176–1182.

[2] G. Cabanes, Y. Bennani, and D. Fresneau, "Enriched topological learning for the cluster detection and the visualization",neural-networks,no.0,2012.

[3] G. Cabanes and Y. Bennani, "Unsupervised topographic learning for the spatio-temporal data-mining," Advances in Artificial Intelligence, vol. 2010, Article ID 832542, 12 pages, 2010.

[4] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak "Plastic Card Fraud Detection using the Peer Group Analysis" Springer, Issue 2008.

[5] John T.S Quah, M Sriganesh "Real time Credit Card Fraud Detection using the Computational Intelligence" ELSEVIER Science Direct,35 (2008) 1721-1732.

[6] Linda Delamaire ,Hussein Abdou and John Pointon, "Credit Card Fraud and Detection technique", Bank and Bank System,Volume 4, 2009.

[7] D. Sivanandanam, Principles of the Soft Computing

[8] Linda Delamaire ,Hussein Abdou and John Pointon, "Credit Card Fraud and Detection technique", Bank and Bank System,Volume 4, 2009

[9] Philip K Chan,Wei Fan,Andias Prodromidis,J.Stolfo, "Distributed Datan Mining For Credit Card Fraud Detection", IEEE Intelligent System,Special Issue On Data Mining, 1999

[10] Raghavendra Patidar,Lokesh Sharma "Credit Card Fraud Detection using Neural Network" International journal of Soft Computing(IJCSE),Volume 32,38,Issue 2011.

[11] Vladdimir Zasalavsky and Anna Strizhak, "Credit Card Fraud Detection using Self Organising Map",Information and Security An International Journal,Vol 18,2006.