International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 59-64

An Improved Credit Card Fraud Detection Using K-Means Clustering Algorithm

D. Viji[1], S. Kothbul Zeenath Banu[2]
[1, 2](Assistant Professor, Department of Computer Application, Krishnasamy College of Science, Arts and Management for Women, Cuddalore – 607 109, Tamil Nadu, India)

Abstract: *Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behaviourist profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system. In this chapter, we model the sequence of operations in credit card transaction processing using a Hidden Markov Model (HMM) and show how it can be used for the detection of frauds. An HMM is initially trained with the normal behaviour of a cardholder. If an incoming credit card transaction is not accepted by the trained HMM with sufficiently high probability, it is considered to be fraudulent. At the same time, we try to ensure that genuine transactions are not rejected.*

Keywords: *Credit Card, Clustering, Fraud, Fraud Detection, Hidden Markov Model, K-Means*

I. Introduction

Today, Internet has become the essential component of life. As telephone, fridge it becomes the important feature. Now a day's people don't have much time and they wish to shop sitting at home. Credit card is purchase now and pays later. As Credit card has the power to purchase the things, its frauds also increased. The operation performed to validate the Credit card number, which is done as a combination of Luhn algorithm and K-Means algorithm. Luhn Algorithm will be applied if a credit card number is not accepted by K-Means Algorithm. K-Means is then enhanced to addition of epochs. Epochs are the maximum number of iterations and error value is calculated. The main is to increase the security system of credit card and debit card using k-means clustering algorithm. In addition, the proposed model should collect the detailed user profile, security questions and good model in verification and validation of credit card. Collection of User profile (2) Create a new cluster (3) Verify with the existing transactions (4) Validate the records. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization and online updating.

II. Data Mining

2.1 Data Mining invovles six common classes of tasks:

Anomaly detection/Outlier/Change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

2.2 Credit card System and its Architecture:

Credit card fraud is a wide ranging term for theft and fraud committed using or involving a payment card, such as credit card or debit card, as a fraudulent source of frauds in a transaction.

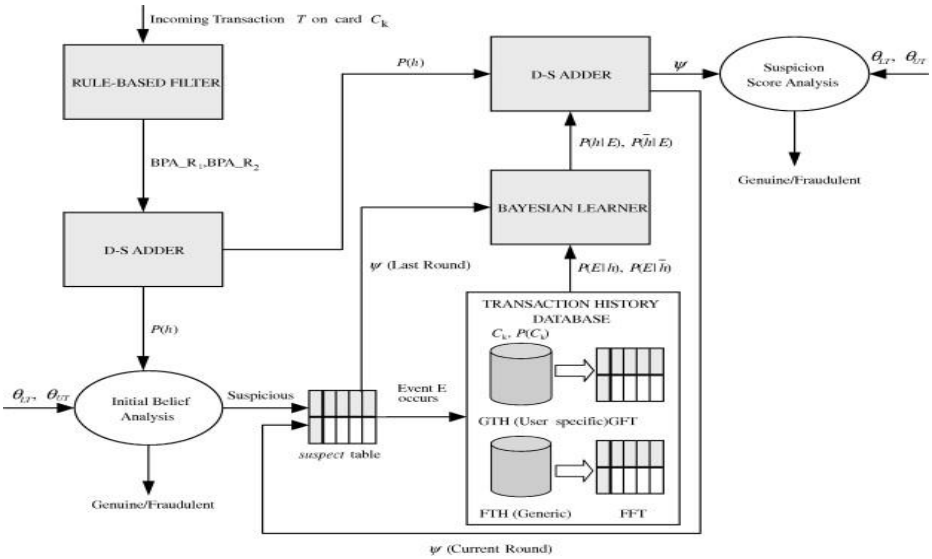

Fig.1. Block Diagram for Creidt Card Fraud Detection System

However, credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complains for the sixth year in a row. A credit card is a payment card issued to users as a system of payment. It allows the cardholder to pay for goods and services based on the holder's promise to pay for them. The issued card creates a revolving account and grants a line of credit to the cardholder, from which the user can borrow money for payment to a merchant or as cash in advance. The size of most credit cards is 3 3/8 X 2 1/8 in (85.60 X 53.98 mm), conforming to the ISO/IEC 7810 ID-1 standard. Credit cards have a printed or embossed bank card number complying with the ISO/IEC 7812 numbering standard.

III. Basic K-Mean Clustering Technique

Suppose that a dataset of n data points $l_1,l_2,\ldots\ldots,l_n$ such that each data point is in R^d, the problem of extracting the minimum variance clustering of the data set into k clusters is that of finding k points $\{m_1\}$ (j=1,2,…..,k) in R^d such that

$$1 \sum_{n\ i=1}^{n} [\min_{j} d^2 (xi, mj)]$$

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j.The points $\{m_j\}$ (j=1,2,….,k) are known as cluster centroids. The thing in above Equation is to find k cluster centroids, so that the average squared Euclidean distance (mean squared error, MSE) among a data point and its nearest cluster centroid is minimized.

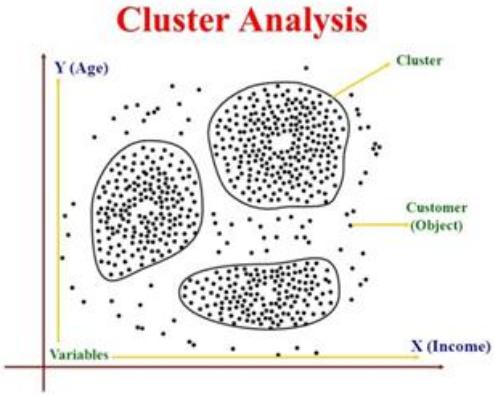

Fig. 2 Cluster Analysis

The k-means algorithm provides an easy method to implement approximate solution to this equation. The cause of the popularity of k-means is ease and simplicity of execution, scalability, speed of convergence and adaptability to sparse data. The k-means algorithm can be thought of as a gradient descent procedure, which initiate at starting cluster centroids, and iteratively modify these centroids to decrease the objective function in

the equation list above. The k-means always congregate to a local minimum. The particular local minimum search depends on the starting cluster centroids. The problem of searching the global minimum is NP-complete. The K-means algorithm modifies cluster centroids till local minimum is found.

3.1 K-Means Clustering Algorithm:

MSE = largenumber;
Select intial cluster centroids {mj}jk=1;
Do
OldMSE = MSE;
MSE1 = 0;
For j=1 to k
mj=0; nj=0;
endfor
For i = 1 to n
For j = 1 to k
Compute squared Euclidean distance d2 (xi, mj);
endfor
Find the closet centroid mj to xi;
mj = mj+xi;
nj = nj+1;
MSE1 = MSE1 + d2(xi, mj);
endfor
For j = 1 to k
nj = max(nj, 1);
mj = mj/nj;
endfor
MSE = MSE1; while (MSE < OldMSE)

Before the k-means algorithm converges, distance and centroid computations are done while loops are executed a number of times, say 1, where the positive integer 1 is known as the number of k-means iterations. The precise value of 1 varies according to the initial starting cluster centroids even on the same dataset. So the time complexity of the algorithm is O (nkl), where n is the total number of objects in the dataset, k is the required number of clusters we identified and 1 is the number of iterations, $k \leq n$, $1 \leq n$.

3.2 Limitations of K-Means algorithm:

- It is computationally very costly as it involves several distance calculations of each data point from all the centroids in every iteration.
- The final cluster results greatly depend on the selection of initial centroids which causes it to converge at local optimum.

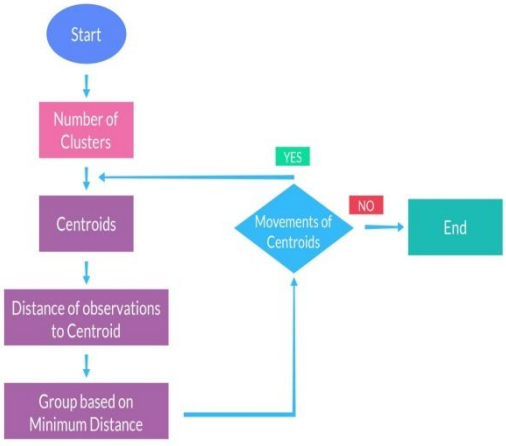

Fig. 3 K-Means algorithm

3.3 Working Methodology:

An Authorized User performs an online transaction then sending profile is matched into our database and if it matches then the transaction is performed successfully and then user is notified that transaction is done successfully.

If an Unauthorized User tries to perform an online transaction and if the sending profile doesn't matches into the database then access is blocked to that user and system failure occurs. HMM traces the IP address of the organization from where unauthorized user was trying to gain transaction and it also sends notification on authorized user mobile number and raises the alarm to admin system.

IV. HMM Model

An HMM is a double embedded stochastic process much more complicated process as compared to a traditional Markov model. The HMM uses the price range: 1. High 2. Medium 3.Low as prediction.

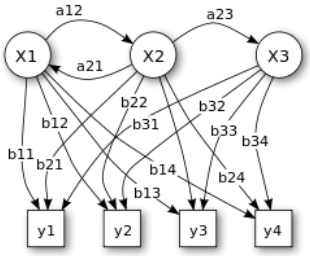

Fig. 4. HMM Architecture

4.1 Techniques and algorithm used:

To record the credit card transaction dispensation process in conditions of a Hidden Markov Model (HMM), it creates original deciding inspection symbols in our representation. We quantize the purchase values x into M price ranges V1, V2,…., VM, form the study symbols by the side of the issuing bank. The genuine price variety for each symbol is configurable based on the expenditure routine of personal cardholders. HMM determine these prices range. Dynamically, by using clustering algorithm (like K clustering algorithm) every card holder transactions price values are changed. It uses cluster Vk for clustering algorithm as k ¼ 1, 2…. M, which can be represented both observations on price value symbols as well as on price value range.

In this prediction process it considers mainly three price value ranges such as (1) low – l (2) Medium – m (3) High – h. So set of this model prediction symbols is V {l,m,h}, so V ¼ f as l (low), m (medium), h (high) which makes M ¼ 3.

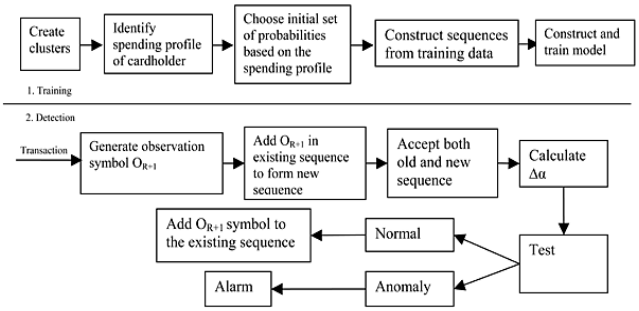

Fig.5 Process Flow for Proposed FDS

Advantages:

1. The detection of the fraud is found much faster than the existing system.
2. In case of the existing system even the original card holder is also checked for fraud detection. But in this system no need to check the original user as we maintain a log.
3. We can find the most accurate detection using this technique.
4. This reduces the tedious work of an employee in the

Application:

1. Provide easy and well security to Online Shopping.
2. Detect Frauds and trace the location from where the transaction has been made.

Result:

This solution can used banks, organizations and governmental centers for transaction management that prevent losses of this misuse or fraudulent behaviors. This algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the

banks and a series of anti-fraud strategies can be adopted to prevent banks from great losses before reduce risks. The steps to be performed to reduce fraud access are providing New card about their contact details, own login and password. In security information, it will get the information detail and its store's in database. If the card lost then the security information form arises, it has set of question where user has to answer it correctly to move to the transaction. After transaction, the information are verified by verification seeking party, the verification information being given by a third, verifying party, based on confidential information in the possession of the initiating party. In verification process will seeks card number is correct the relevant process will be executed. If the number is wrong, mail will sent to the user saying the card no has been block and they can't do the further transaction.

V. Conclusion

The Enhanced Algorithm is easy to implement and it proves to be a better method to determine the initial centroids to be used in the k-means clustering algorithms. As the end clustering result of the k-means clustering methods are highly dependent on the selection of initial centroids, so they should be a systematic method to determine the initial centroids which makes the k-means algorithm to converge in global optima and unique clustering results.

This requirement is fulfilled by the algorithm. Besides solving the problem of non-unique results, are the algorithm is also widely applicable to different types to problems. The problems with uniform as well as the problems with non-uniform distribution of data points are better addressed by our algorithm. Our algorithm tries to enhance the k-means clustering algorithm by eliminating one of its drawbacks. But still lots of work needs to be done to enhance the k-means algorithm to a greater extent.

K-Means can be applied on the numerical data only. But day to day life we encounter scenarios with a combination of both numeric and categorical data values. So future work can be carried out in the direction of making the k-means algorithm applicable for mixed type of data.

The algorithm is easy to implement and it proves to be a better method to determine validity of credit card. Algorithm is used as a validity criterion for a given set of numbers. Almost all credit card numbers are extracted following this validity criterion…also called as the Luhn check or the Mod 10 check. It went without saying that the Luhn check is also used to verify a given existing card number.

If a credit card number does not assured this check, it is not a valid number. So future work carried out in the direction of making the k-means algorithm applicable for different length of credit card numbers.

In this, we used an HMM in detection of credit card fraud. We modeled the sequence of transactions in credit card processing using an HMM. We have used clusters that are generated by using k-means clustering algorithms as our observation symbols.

VI. Future Enhancement

Efficient credit card fraud detection system is an utmost requirement for any card issuing bank. Credit card fraud detection has drawn quite a lot of interest from the research community and a number of techniques has been proposed to counter credit card fault. The Fuzzy Darwinian Fraud detection systems improve the system accuracy. Since the fraud detection great of Fuzzy Darwinian Fraud detection systems in terms of true positive is 100% and shows good results in detective fraudulent transactions. The neutral network based CARD WATCH show good accuracy in fraud detection and processing speed is also high, but it is limited to one-network per customer. The fraud detection hidden Markov model is very low compare to other methods. The hybridized algorithm name BLAH-FDS identifies and detects fraudulent transactions using sequence algorithm tool. The processing speed of BLAST-SSAHA is fast enough to enable on-line detection of credit card fraud. BLAH-FDS can be effectively used to counter frauds in other domains such as telephone communication and banking fraud detection. The ANN and BNN are used to detect cellular phone fraud, network intrusion. All the techniques of credit card fraud detection discussed in this survey paper have its own strength and weaknesses. Such a survey bill enables us to build a hybrid approach for identifying fraudulent credit card transactions.

References

[1]. Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), "Credit card fraud and detection techniques: a review", Banks and Bank Systems, Volume 4, Issue 2, 2009 .

[2]. Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, "A review of Fraud Detection Techniques: Credit Card", International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012 .

[3]. Vladimir Zaslavsky and Anna Strizhak," credit card fraud detection using selforganizing maps", information & security. An International Journal, Vol.18,2006.

[4]. L. Mukhanov, "Using bayesian belief networks for credit card fraud detection," in Proc. of the IASTED International conference on Artificial Intelligence and Applications, Insbruck, Austria, Feb. 2008, pp. 221– 225.

[5]. Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun K. Majumdar, "CreditCard Fraud Detection Using Hidden Markov Model" IEEE, Transactions On Dependable And Secure Computing, Vol. 5, No 1. , January-March 2008

[6]. V. Bhusari, and S. Patil, "Study of Hidden V. Bhusari, and S. Patil, "Study of Hidden Markov Model in Credit Card Fraudulent Detection", International Journal of Computer Applications (0975 – 8887) Volume 20– No.5, April 2011.
[7]. Ghosh, S., and Reilly, D.L., 1994. Credit Card Fraud Detection with a Neural-Network, 27th Hawaii International l Conference on Information Systems, vol. 3 (2003), pp. 621- 630.
[8]. Syeda, M., Zhang, Y. Q., and Pan, Y., 2002 Parallel Granular Networks for Fast Credit Card Fraud Detection, Proceedings of IEEE International Conference on Fuzzy Systems, pp. 572-577 (2002).
[9]. Rabiner, L. R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, vol. 77, no. 2, Feb. 1989, pgs 257 - 285. There is a lot of notation but verbose explanations accompany.
[10]. Dodge, Charles and Jerse, Thomas A. Computer Music: Synthesis, Composition, and Performance, 2nd ed., Shirmer Books: New York, 1997. pgs 361-368 are of particular interest - composing music with markov chains.
[11]. 3.Grimmett,G.R. and Stirzaker, D.R. Probability and Random Processes, 2nd ed., Clarendon Press: Oxford, 1994. Chapter 6 is on Markov chains, intended for advanced undergrads.
[12]. Isaac, Richard. The Pleasures of Probability, Springer: New York, 1995. Chapter 16 is on Markov Chains, intended for undergrads.
[13]. Navigating Through the Risks of Credit Card Processing (Savvy Business Owner's Guide) paperback – April 30, 2010 by Bill Pirtle.
[14]. Data Mining Application for Cyber Credit-Card Fraud Detection System by John Akhilomen.
[15]. Advances in K means Clustering: A ... Book by Junjie Wu
[16]. Hidden Markov Models for Time Series: An Introduction Using R, 2nd Edition, by Walter Zucchini, Iain L. Macdonald, and Roland Langrock. Monographs on Statistics and Applied Probability 150, Published by CRC Press, 2016.
[17]. Data Clustering: Theory, Algorithms, ... Book by Chaoqun Ma, Guojun Gan, and Jianhong Wu originally published: 2007.