Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 1, Number 10; October, 2014 pp. 33-37

http://www.krishisanskriti.org/jbaer.html

A Fast Fraud Detection Approach using Clustering Based Method

Surbhi Agarwal[1], Santosh Upadhyay[2]

[1]*M.tech Student, Mewar University, Chittorgarh Rajasthan*
[2]*GCET, Gr.Noida, Uttar Pradesh*

Abstract: **Due to rapid advancement in credit card transactions, credit card fraud has become increasingly rampant in recent years. Owing to levitate and rapid escalation of E-Commerce, cases of credit card fraud allied with it are also intensifying which results in trouncing of billions of dollars worldwide each year. Nowadays, fraud is one of the major causes of great financial losses, not only for merchants, but individual clients are also affected. Many data mining techniques have evolved in detecting various credit card fraudulent transactions. Out of the available data mining techniques, clustering has proven itself a constant applied solution for detecting fraud. Clustering process groups the data in such a way so that highly similar data come under one group. An outlier is an instance (record, transaction, etc) that does not conform to a well-defined and general pattern of the expected behavior in a certain dataset. In this paper, clustering and outlier detection techniques are used as a hybrid approach to find these mischief activities. Using clustering, the data sets are partitioned and outlier detection is used to find the fraudulent data. In proposed approach, two techniques are combined to efficiently find the outlier from the data set. The experimental results using real dataset demonstrate that proposed method takes less computational cost and performs better than the distance based method. Proposed algorithm efficiently prunes of the inliers and save huge number of extra calculations.**

Keywords: **Credit card fraud detection, E-Commerce, clustering, outlier detection, data mining.**

1. INTRODUCTION

In modern world most of the people are using credit cards. It is the most popular payment mode. Detecting frauds in this online transaction is a very difficult task. So, there is a need to develop a model to detect these frauds, in the business area and also in the academics.

Fraud in this context is best described by Phua et. al.[1] as "*the abuse of a profit organization's system without necessarily leading to direct legal consequences"* (Phua, V. Lee, Smith, & Gayler, 2005). Clustering helps in grouping the data into similar clusters that helps in uncomplicated retrieval of data. Cluster analysis is a technique for breaking data down into related components in such a way that patterns and order becomes visible. Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. In recent years, conventional database querying methods are inadequate to extract useful information, and hence researches nowadays are focused to develop new techniques to meet the raised requirements. It is to be noted that the increase in dimensionality of data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of attributes. Outlier detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the database.

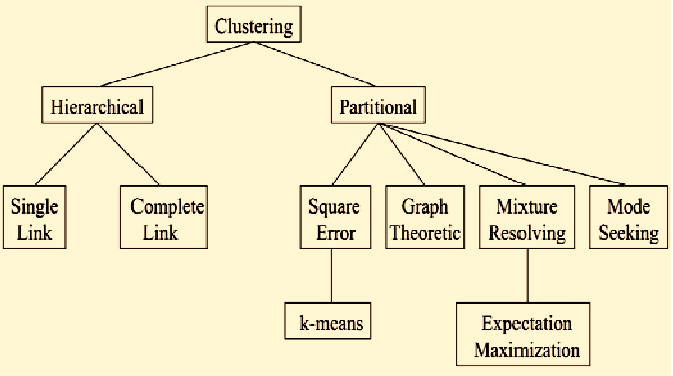

Fig. 1. Clustering Taxonomies [6]

2. LITERATURE REVIEW

Credit card fraud detection has drawn a lot of research interest and a number of techniques, with special emphasis on data mining and neural networks. J.Daud Pathan et al.[2] describes the "Credit card fraud detection system using Hidden Markov Model and k-clustering". In this paper HMM is used to model the sequence of operation in credit card transaction processing. If an incoming credit card transaction is not accepted by the trained HMM with sufficiently high probability, it is considered to be fraudulent. S.Raheja et al.[3] describes the

"Credit Card Fraud Detection By Improving K-Means".In this paper the series of operation is modeled in credit card transaction processing using a K-Means and LUHN algorithm and how it can be used for detecting frauds is also covered. Luhn algorithm is instructed with the behavior of card holder. If an incoming credit card transaction is not accepted by the k-mean with sufficient high probability, it is considered to be fraudulent then confirmation is given by Luhn algorithm.

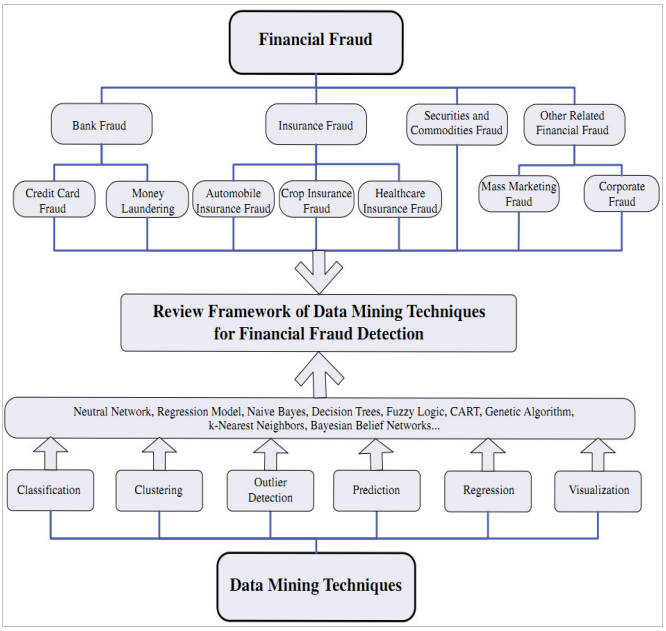

Fig. 2. Financial Fraud Detection Review Framework

L.V. Bijuraj[4] describes "Clustering and its applications". The paper covered all the types and methods of clustering by giving certain examples. In this paper, use of clustering in text mining has been discussed. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. J. S. Mishra et al.[5] describes "A Novel Approach for Credit Card Fraud Detection Targeting the Indian Market". In this paper it is shown how a fraud can be reported instantly while the fraudulent transaction is in process. Andrei Sorin SABAU [6] describes "Survey of clustering based financial fraud detection research". This paper surveys clustering techniques used in fraud detection over the last ten years. Fig. 2 shows various data mining techniques and types of financial fraud. Ms. S. D. Pachgade et al.[7] describes "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach".Proposed Method for outlier detection uses hybrid approach. Purpose of approach is first to apply clustering algorithm that is k-means which partition the dataset into number of clusters and then find outliers from the each resulting clusters using distance based method. M. A. Vasarhelyi et al. [8] describes "Application of Anomaly Detection Techniques to Identify Fraudulent Refunds". This paper describes classification-based and clustering-based anomaly detection techniques and their applications.

As an illustration, the paper applies K-Means, a clustering-based algorithm, to a refund transactions dataset from a telecommunication company, with the intent of identifying fraudulent refunds. N. L. Khac et al.[9] decribes "Application of Data Mining for Anti-money Laundering Detection: A Case Study". Dr. R. Dhanapal et al.[10] describes the "Analysis of credit card fraud detection methods". In this paper, three methods to detect fraud are presented. Firstly, clustering model is used to classify the legal and fraudulent transaction using data clusterization of regions of parameter value. Secondly, Gaussian mixture model is used to model the probability density of credit card user's past behavior so that the probability of current behavior can be calculated to detect any abnormalities from the past behavior. Lastly, Bayesian networks are used to describe the statistics of a particular user and the statistics of different fraud scenarios. A.Kundu et al.[11] describes "Credit card fraud detection using hidden markov model". In this paper, the sequence of operations in credit card transaction processing is modeled using a Hidden Markov Model (HMM) and how it can be used for the detection of frauds. An HMM is initially trained with the normal behavior 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, it is ensured that genuine transactions are not rejected. Detailed experimental results are presented to show the effectiveness of the approach and comparing it with other techniques available in the literature.

3. CLUSTERING BASED FFD SURVEY

As a result of the research methodology, 10 articles were selected for inclusion. They have been grouped based on application domain, clustering technique and case study dataset. Papers are ordered on publishing year and clustering technique in Table 1.

4. PROPOSED WORK

A Fraud Detection System (FDS) runs at a credit card issuing bank. Each incoming transaction is submitted to the FDS for verification. FDS receives the card details and the value of purchase to verify whether the transaction is genuine or not. The types of goods

That is bought in that transactions are not known to the FDS. It tries to find any outlier in the transaction based on the spending profile of the cardholder, amount, month, day (Monday as first day, Tuesday as second day, and so on.), shipping address, and billing address, etc. If the FDS confirms

the transaction to be malicious, it raises an alarm, and the issuing bank declines the transaction. The concerned card holder may then be contacted and alerted about the possibility that the card is compromised. In this section, we explain how K-means clustering can be used for Credit card fraud detection.

5. SYSTEM ARCHITECTURE

We have proposed our work in two phases-

- **Phase I**- clustering using K-means

Table 1. Surveyed articles

Author	Year	Application Domain	Clustering Technique	Dataset
J.Daud Pathan et al.[2]	2014	Credit Card Fraud	HMM and K-Clustering	Online Transaction
S.Raheja et al.[3]	2014	Credit Card Fraud	Modified K-Means, Luhn algorithm	Credit Card Numbers
L.V.Bijuraj[4]	2013	Clustering Applications	K-Means, DBSCAN clustering	Protein & fat contents
J.S.Mishra et al.[5]	2013	Credit Card Fraud	HMM	Online payment
A.Sorin SABAU [6]	2012	Financial Fraud	K-means, hierarchical clustering	Internal and external data against organization
Ms. S.D.Pachgade et al.[7]	2012	Medical Diagnosis	Cluster Based and distance based	Cancer dataset, liver disorder dataset
M.A.Vasarhelyi et al.[8]	2011	Refund Fraud	K-means	Refund transaction data
N.L. Khac et al. [9]	2010	Money Laundering Fraud	K-means	Transaction Data
Dr. R.Dhanapal et al.[10]	2009	Credit card fraud	Clustering model, Probability density estimation, Bayesian networks	Training Data
A.Kundu et al.[11]	2008	Credit Card Fraud	HMM	Credit card issuing bank

Phase II- outlier or anomaly detection by Distance-based approach

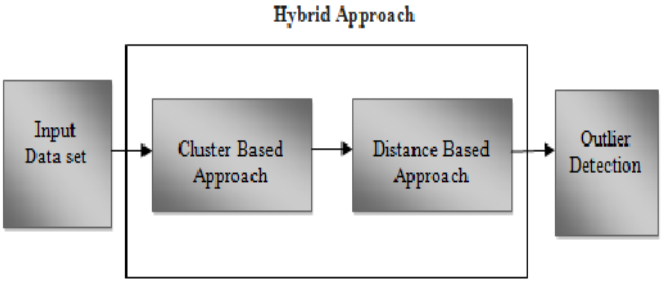

Fig. 3. System Architecture

5.1 Input Data Set

Collecting dataset from different transactions

5.2 Cluster Based Approach

Clustering is a popular technique used to group similar data points or objects in groups or clusters. Clustering is an important tool for outlier analysis. Cluster based approach is here act as data reduction. First, clustering technique is used to group the data having similar characteristics. And then calculate the centroids for each group.

5.3 Distance Based Approach

Distance based technique is used to calculate maximum distance value for each cluster. If this maximum distance is greater than some threshold which is given by the user then it will declare as outlier, otherwise as a real object or inliers. Distance-based methods are much more flexible and robust. They are defined for any data type for which we have a distance measure and do not require a detailed understanding of the application domain

5.4 Outlier Detection:

Outlier detection is an extremely important task in a wide variety of application domains. Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data or which are far away from their cluster centroids.

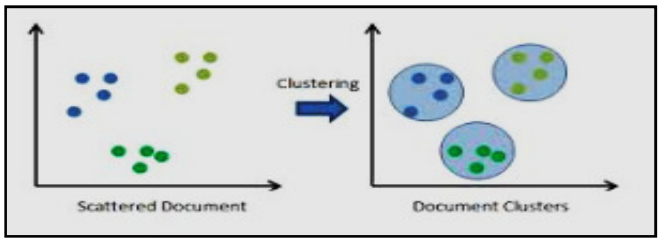

Fig. 4. Clustering of scattered data

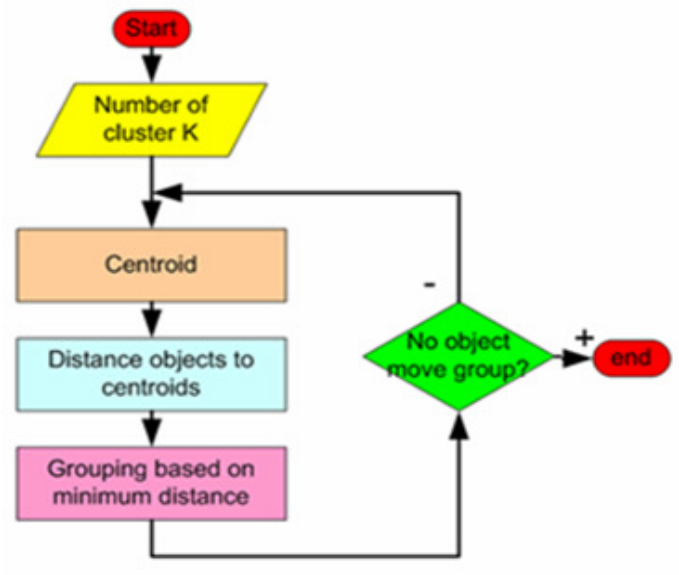

Fig. 5. K-means clustering process

6. THE K-MEANS CLUSTERING AND DISTANCE-BASED ALGORITHM PROCESS

K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating globular clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative. K-means clustering and outlier detection is combined to generate a hybrid approach. The steps for hybrid approach are-

The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

For each data point:

1. Calculate the distance from the data point to each cluster.
2. If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
3. Repeat the above step until a complete pass through all the data points' results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. Calculate threshold value for each cluster-

   i) Find distance of nearest centroid from NT point, say dNT.

   A) Finding nearest cluster-

- Determine all k-distances between NT and k-cluster centroids.
- Find the cluster which has minimum distance between its centroid and NT point and that will be the nearest cluster.

   ii) Find radius of nearest cluster.

   iii) Radius value is set as threshold value by the user.

Compare threshold value with dNT. If new transaction is greater than threshold then it will declare as "outlier".

Let NT be the new transaction occurred.
Th be the threshold value set by the user.
Cd be the max distance among centroids.
if NT<Th
then "normal"
elseif NT<Cd
then "normal"
else
"outlier"

7. RESULT

We use synthetic data to evaluate our work.

7.1 Performance Measure

Two indicators, Detection Rate and False Alarm Rate, are used to measure the accuracy of the method. The Detection Rate shows the number of true outliers that have been successfully detected divided by the total number of outliers present in the dataset. The False Alarm Rate is defined as the number of normal instances incorrectly labelled as outliers divided by the total number of normal instances. A good method should provide a high Detection Rate together with a low False Alarm Rate. ROC (Receiver Operating Characteristics) curves are plotted depicting the relationship between False Alarm Rate and Detection Rate for one fixed test set combination. ROC curves are a way of visualizing the trade-offs between detection and false alarm rate.

7.2 Dataset Preparation

The dataset was obtained by simulating a large number of different transactions while purchasing. The goal was to produce a good training set for learning methods that use labelled data. As a result, the proportion of transactions to normal ones in the dataset is very large as compared to data that one would expect to observe in practice.Unsupervised anomaly detection algorithms are sensitive to the ratio of outliers in the dataset. If the number of outliers is too high, each outlier will not show up as anomalous.

7.3 Results

We have calculated Detection Rate and False Alarm Rate for all the samples. The results shown in table 2 and the corresponding ROC (Receiver Operating Characteristics)

curves shown in figure 6 is a representation of most frequently found result.

Table 2. Result

Number of Clusters	Detection Rate (%)	False Alarm Rate (%)
250	98.4	16.688
160	92.0	13.251
110	89.8	10.614
50	87.2	5.651
19	83.0	1.86
8	60.6	0.481

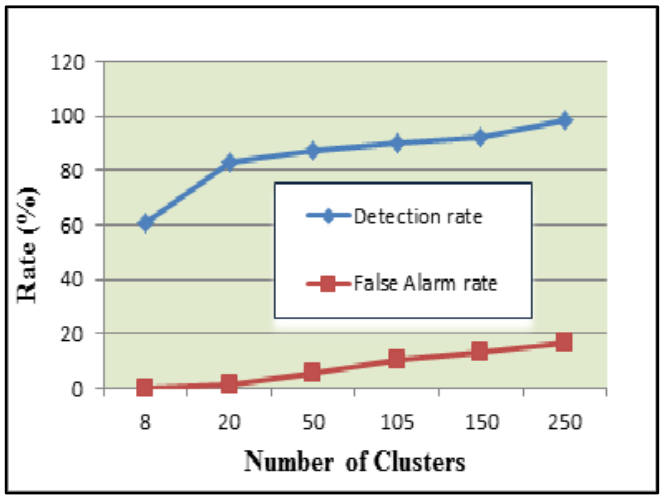

Fig. 6. Graph between number of clusters and rate

8. CONCLUSION

This paper aims to detect outliers which are grossly different from or inconsistent with the remaining dataset. Existing outlier detection methods are ineffective on scattered real-world datasets due to implicit data patterns and parameter setting issues. We proposed an efficient outlier detection method. We first group the data (having similar characteristics) into number of clusters. Due to reduction in size of dataset, the computation time is reduced considerably. Then we take threshold value from user and calculate outliers according to given threshold value for each cluster. We get outliers within a cluster. Hybrid approach takes less computation time.

Approach only deals with numerical data, so future work requires modifications that can make applicable for text mining also. The approach needs to be implemented on more complex datasets. Future work requires approach applicable for varying datasets.

REFERENCES

[1] Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 1–14. Retrieved November 26, 2010, from http://arxiv.org/pdf/1009.6119.

[2] Mohd A.Z. Khan, J.D. Pathan and A.H.E. Ahmed, "Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering", IJARCCE, vol. 3, Pages 5458-5461, February 2014.

[3] M. Singh, Aashima, S. Raheja, "Credit Card Fraud Detection by Improving K-Means", IJETR, Vol-2, Issue-5, May 2014.

[4] L. V. Bijuraj, "Clustering and its Applications", NCNHIT 2013.

[5] J. S. Mishra, S. Panda, A. K. Mishra, "A Novel Approach for Credit Card Fraud Detection Targeting the Indian Market", IJCSI, Vol. 10, Issue 3, No 2, May 2013.

[6] A. Sorin SABAU, "Survey of clustering based Financial Fraud Detection Research", Informatica Economica vol. 16, no. 1/2012.

[7] Ms. S. D. Pachgade, Ms. S. S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", IJARCSSE, Vol 2, Issue 6, June 2012.

[8] M. A. Vasarhelyi, H.Issa, "Application of Anomaly Detection Techniques to Identify Fraudulent Refunds", 2011.

[9] Nhien An Le Khac and M. T. Kechadi, "Application of Data Mining for Anti-money Laundering Detection: A Case Study", ICDMW, pp. 577-584, 2010.

[10] V.Dheepa, Dr. R. Dhanapal, "Analysis of Credit Card Fraud Detection Methods" IJRTE, vol 2, No. 3, November 2009.

[11] A. Srivastava, A. Kundu, S. Sural, A. K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model", IEEE Transactions on Dependable and Secure Computing, vol. 5, NO. 1, January-March 2008.