IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Credit Card Fraud Detection Using Hybrid approach

Siddhi Dave[1], Vidhi Sheth[2], Jay Vala[3]

Student, Dept. of Information Technology, G.H.Patel College of Engineering and Technology, Anand, India[1,2]

Assistant Professor, Information Technology,G.H.Patel College of Engineering and Technology,Anand,India[3]

Abstract: Fast Development in electronic trade has led to increase in the use of credit card payment mode. Frauds related to credit cards are also increase with usage of credit card payment mode. Data mining techniques are used to disclose fraudulent activity in credit card payment mode. Data mining process is to extract information from a dataset and transform it into an understandable structure for future use. Clustering method is used for dividing the objects in such way that objects in same group are more similar to each other than to those in other groups. In this paper, we are using hybrid approach that comprise of k-means clustering and that followed by Distance based techniques that detect outliers from dataset. The experimental results show the comparison of actual dataset accuracy and accuracy of proposed with dataset.

Keywords:Fraud detection, Credit card fraud, clustering, hybrid approach.

I. INTRODUCTION

In today's world credit cards are used as payment mode.Credit cards are used worldwide as they are accepted in millions of places at home and abroad. Due to increase in usage of credit cards, frauds related to it are also increase. Frauds are illegal and unauthorized use of account for personal gain and also misrepresentation of account information to obtain goods and/or services. So it is necessary to detect fraud. Data mining is a process of extracting hidden and useful information from the data and the knowledge discovered by data mining is previously unknown, potentially useful, and valid and of high quality[1].

Data mining techniques come in two main forms: supervised approach and unsupervised approach. Supervised mining techniques are appropriate when you have a specific target value you'd like to predict about your data. Frauds behavior can't be predicted so credit card fraud detection uses an unsupervised approach for detection of frauds. Unsupervised methods don't make use of labelled records (previously occurred frauds). It detects the changes in behavior or unusual Transactions [2].

In this work we are using Hybrid approach that comprise of Clustering algorithm and followed by Distance based techniques that detect outliers from dataset. Clustering is a division of data objects into groups of similar objects. Such groups are called clusters.

Objects possessed by same cluster tend to be similar, while dissimilar objects are possessed by different clusters. These clusters represent groups of data and provide simplification by representing many data objects by fewer clusters. And, this helps to model data by its clusters. Clustering is a method of unsupervised learning [3].

Different clustering methods can be classified into various Categories such as partitioning based methods, hierarchical methods, grid-based methods, density-based methods, model-based methods, methods for high dimensional data and constraint-based clustering. Among all these methods, this paper is aimed to explore one method – k-means – which is Partitioning based clustering method [3].

K-means algorithm is the most simplest and popular clustering algorithm among the others. The k-Means algorithm is used to decrease the complexity of grouping data. This algorithm is sensitive to the initial cluster centers which are randomly selected [4]. K-means algorithm is described below in proposed system. Detecting outliers has important applications in data cleaning as well as in the mining of abnormal points for fraud detection, stock market analysis, intrusion detection, marketing, network sensors. Finding anomalous points among the data points is the basic idea to find out an outlier. Distance based techniques use the distance function for relating each pair of objects of the data set. Distance based definition (these definitions are computationally efficient) represent a useful tool for data analysis [5]. In this work, we are using K-means clustering algorithm to find similar objects. To find out outliers we are using distance based method as per given threshold by user. Within a cluster get outliers, that are far from their cluster centroid.

II. HYBRID APPROACH

Hybrid Approach:
Steps of Hybrid Approach:-
Step 1: Input dataset D, Initialize number of cluster K.
Step 2: Randomly select centroid from dataset:

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

A: Assign each object to the cluster with the nearest Centroid.

B: Compute each centroid as the mean of the objects Assigned to it.
Step 4: Repeat until No changes in centroid.
Step 5: Calculate the distance of each point of cluster From centroid of cluster.
Step 6: Take threshold value T.
Step 7: If Distance > T than point is declared as "Outlier".

III. PROPOSED SYSTEM

Figure below is process flow of the desired system.

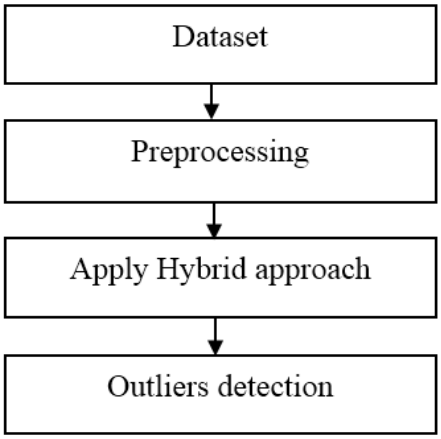

Fig. 1 Process flow of our system

1.Dataset:
We utilize standard German Credit Card Fraud dataset which is accessible on UCI Machine Learning Repository. Our dataset comprise of 22 attributes and 1000 instances. Out of 22 attributes we are using 8 attributes. The attributes are both numerical and categorical.

Attribute Name	Attribute Type
1.Status of existing checking account	Categorical
2.Payment Status of Previous Credit	Categorical
3 .Purpose	Categorical
4.Credit Amount	Numerical
5.Present employment since	Categorical
6.Sex & Marital Status	Categorical
7.Age (years)	Numerical
8.Occupation	Categorical

TABLE I:
DATA SET ATTRIBUTE DISCRIPTION

2. Preprocessing:
Data in real world is dirty, noisy, inconsistent, and incomplete. So we have to do preprocessing. We have done data cleaning and data reduction in weka tool.

3. Apply proposed algorithm as stated above.

4. Apply each step as stated above to detect outliers.

IV. EXPERIMENTAL RESULTS

	Case 1	Case 2	Case 3
Cluster 1 Outliers	64	85	85
Cluster 2 Outliers	6	6	14
Cluster 3 Outliers	100	100	100
Cluster 4 Outliers	57	57	57
Cluster 5 Outliers	42	42	42
Total Outliers	269	290	298

TABLE III:DESCRIPTION OF OUTLIERS

After applying proposed algorithm, we get 269,290 and 298 outliers out of 1000 instance. As mention in above three cases Threshold value are different as per user input. Number of Outliers is varying if the value of threshold is change. Number of outliers is depending on value of threshold.

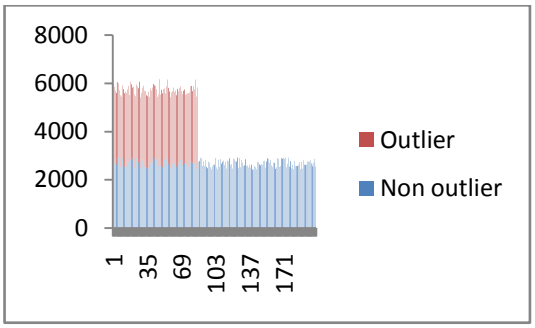

Fig 2 – Cluster 1 graph

As shown in above figure x- axis represents the number of instance in cluster1 and y – axis represents value of each Instance (E.g. first instance in cluster1 value is 2799). Total number of instance in cluster1 is 288. Red colour in graph represents number of outlier which is 85 and Blue colour represents number of non-outliers which is 203.

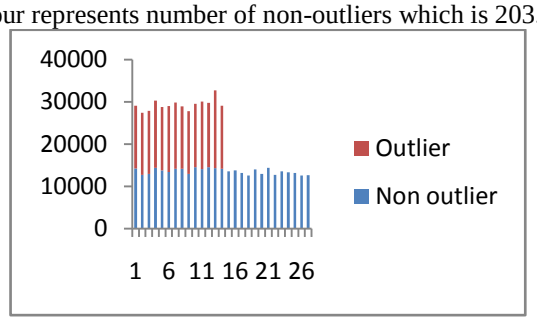

Fig3 – Cluster2 graph

As shown in above figure x- axis represents the number of instance in cluster2 and y – axis represents value of each Instance (E.g. first instance in cluster2 is 14277). Total number of instance in cluster2 is 41. Red colour in graph represents number of outlier which is 14 and Blue colour represents number of non-outliers which is 27.

DOI 10.17148/IJARCCE.2016.5568

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

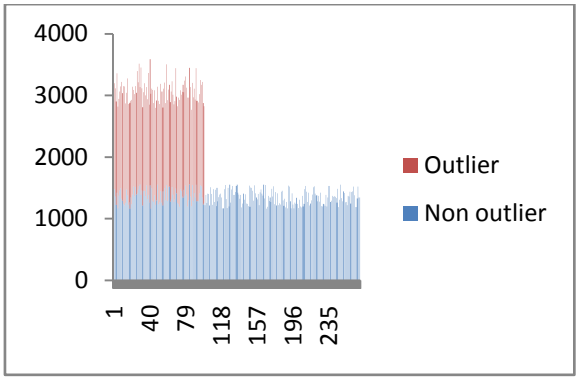

Fig4 – Cluster3 graph

As shown in above figure x- axis represents the number of instance in cluster3 and y – axis represents value of each Instance (E.g. First instance in cluster3 is 1272). Total number of instance in cluster3 is 370. Red colour in graph represents number of outlier which is 100 and Blue colour represents number of non-outliers which is 270.

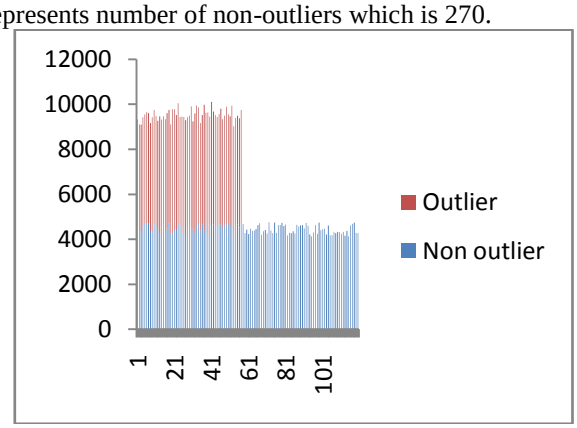

Fig5 – Cluster4 graph

As shown in above figure x- axis represents the number of instance in cluster4 and y – axis represents value of each Instance (E.g. First instance in cluster4 is 4482). Total number of instance in cluster4 is 177. Red colour in graph represents number of outlier which is 57 and Blue colour represents number of non-outliers which is 120.

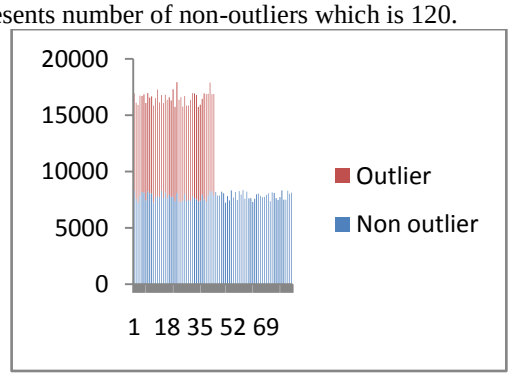

Fig6 – Cluster5 graph

As shown in above figure x- axis represents the number of instance in cluster5 and y – axis represents value of each Instance (E.g. first instance in cluster5 is 8299). Total number of instance in cluster5 is 124. Red colour in graph represents number of outlier which is 42 and Blue colour represents number of non-outliers which is 82.

The k- means clustering is used to form clusters. Five clusters are form. Distance based technique is used to find outliers from each clusters. Threshold value is different for each cluster.

V. CONCLUSION AND FUTURE WORK

Fraud cannot detect 100%. We get 99.33% accuracy. To detect the fraud accurately and efficiently, it is necessary that the real data should be available.The future work will be to use of different algorithm to increase accuracy. For this the different dataset should be available to improve credit card fraud.

REFRENCES

[1]. Ms. S. D. Pachgade and Ms. S. S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[2]. Ms. Amruta D. Pawar, Prof. Prakash N. Kalavadekar and Ms. Swapnali N. Tambe, "A Survey on Outlier Detection Techniques for Credit Card Fraud Detection", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. VI (Mar-Apr. 2014), PP 44-48.

[3]. Shalini S Singh and N C Chauhan, "K-means v/s K-medoids: A Comparative Study",National Conference on Recent Trends in Engineering & Technology, 13-14 May 2011.

[4]. Vaishali, "Fraud Detection in Credit Card by Clustering Approach", International Journal of Computer Applications (0975 – 8887) Volume 98– No.3, July 2014.

[5]. M. Knorr and R. T. Ng, "Finding intentional knowledge of distance-based outliers In VLDB '99: Proceedings of the 25thInternational Conference on Very Large Data Bases, pages 211–222, 1999.