Fraud Detection in Credit Card by Clustering Approach

Vaishali
M.Tech. (VLSI Design)
Banasthali University,
Rajasthan

ABSTRACT
Fraud is an unauthorized activity taking place in electronic payments systems, but these are treated as illegal activities. Fraud detection methods are continuously developed to defend criminals in adapting to their strategies. Fraud can be identified quickly and easily through fraud detection techniques. In this paper, clustering approach is used for credit card fraud detection. Data is generated randomly for credit card and then K-means clustering algorithm is used for detecting the transaction whether it is fraud or legitimate. Clusters are formed to detect fraud in credit card transaction which are low, high, risky and high risky. K-means clustering algorithm is simple and efficient algorithm for credit card fraud detection.

Keywords
Credit card, Fraud detection, Data generation, K-means clustering algorithm

1. INTRODUCTION
The credit card fraud detection technique used is outlier detection. An outlier is an observation which is different from others due to which the suspicion arises that it was generated by a different mechanism. Unsupervised learning comes under this model as the history of the data is not needed. It detects the observation that is different from the normal observations [1]. Outliers are used to detect the fraud. While in supervised method, the models are used to differentiate between fraudulent and non-fraudulent behavior to obtain the outlier.

Clustering have the application in the field of engineering and scientific disciplines like psychology, biology, medicine, computer vision, communication and remote sensing [2]. A set of pattern is observed by abstracting underlying structure in clustering. The patterns are clustered on the basis of more similar features than other pattern of group. Various clustering algorithms have been proposed to fulfil different requirements. Clustering algorithms are based on the structure of abstraction and are classified into hierarchical and partitional algorithms. Hierarchical clustering algorithms construct a hierarchy of partitions, which are represented as a dendrogram in which each partition is nested within the partition at the next level in the hierarchy [3]. Partitional clustering algorithms, with a specified or estimated number of non-overlapping clusters construct a single partition of the data in an attempt to recover natural groups which are presented in the data [4]. As the combinatorial optimization algorithms such as integer programming, dynamic programming and branch and bound methods have moderate number of data points and clusters so these algorithms are expensive.

K-means algorithm is the most simplest and popular clustering algorithm among the others [5]. 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.

2. K-MEANS CLUSTERING ALGORITHM
Clustering is a process of arranging data into groups of similar objects [6]. 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:

1. Hierarchical Agglomerative methods
2. Partitioning Methods
3. The Single Link Method (SLINK)
4. The Complete Link Method (CLINK)
5. The Group Average Method

K-means clustering algorithm is an unsupervised technique. Unsupervised technique 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.

K-means clustering algorithm is an algorithm that partitions or clusters N data points into K disjoint subsets Sj contains Nj data points to minimize the sum-of-squares criterion [7]

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2,$$

Where x_n - vector representing the n_{th} data point.

u_j - geometric centroid of the data points in S_j.

A global minimum of J is not achieved over the assignments by this algorithm. Discrete assignment is used rather than a set of continuous parameters, the "minimum" it reaches cannot be properly called a local minimum.

Celebi et al. [8] presented an overview of k-means initialization methods with an emphasis on their computational efficiency. Eight commonly used linear time initialization methods are compared on a large and diverse collection of real and synthetic data sets using various performance criteria. Finally, the experimental results using non-parametric statistical tests are compared. It is analyzed that popular initialization methods such as forgy, Macqueen, and maximin often performed poorly and there are significantly better alternatives to these methods that have comparable computational requirements.

Xiuchang and Wei [9] proposed a problem based on user behavior pattern analysis which has the insensitivity of

numerical value, strong noise, and uneven spatial and temporal distribution characteristics. The existing clustering methods, trajectory analysis methods, and behavior pattern analysis methods are analyzed, and clustering algorithm is combined into the trajectory analysis. It is obtained that the improved algorithm has a more advantage over the traditional K-Means algorithm. The result of the improved algorithm is 11.625%, while the result of traditional K- Means is 31.2%. It is obtained that the error rate of clustering data in the improved algorithm is lower than the traditional K-Means algorithm. Traditional clustering methods are compared on the basis of the test of the simulation data and actual data, the results demonstrate that the improved algorithm is more suitable for solving the trajectory pattern of user behavior.

K-Means clustering algorithm is popular due to following reasons:

1. K-Means clustering is popular due to its simplicity and it is easy to implement [8]. In every data mining software includes an implementation of it.
2. It is versatile as almost every aspect of the algorithm such as initialization, distance function, termination criterion etc. can be modified.
3. K-Means clustering has complexity in time. It has complexity of storage that is linear in N, D, and K. It has disk-based variants too that do not require all points to be stored in memory. It is guaranteed to converge at a quadratic rate.
4. It is invariant to data ordering, i.e., random shuffling of the data points.

3. EXPERIMENT

In this paper, data is generated randomly using Microsoft SQL Server Management Studio. Then K-means clustering algorithm is applied to detect fraud using Visual Studio 2012 software. K-means clustering algorithm is an unsupervised technique in which clusters problem is solved. The procedure of this algorithm follows a simple and easy way to classify a given data set through a certain number of clusters.

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.1. The data table is generated for different sets. Now, the data table includes transaction ID, transaction amount, transaction country, transaction date, credit card number, merchant category id, cluster id, is fraud and new transaction.

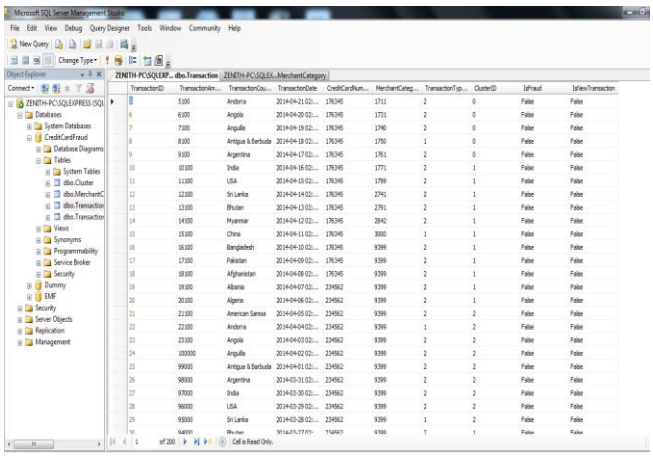

Figure 1 Data table used for detecting fraud

The data table details are:

Transaction ID: is automatically generated and incremented sequentially by the tool.

Transaction amount: is entered manually which cannot be generated automatically as amount is deposited or withdrawn.

Transaction country: By using credit card, the transaction is done all over the world in different countries. From any country, the transaction can be done.

Transaction date: The date on which the amount is deposited or withdrawn.

Credit card number: The different credit cards numbers are manually given for transaction. We used five credit card numbers manually.

3.1 Process for K-means clustering

The above generated data are used as source into the K-means clustering algorithm. The data column includes transaction ID, transaction amount, transaction country, transaction date, credit card number, merchant category, transaction type, cluster ID, isfraud and isnewtransaction. K- Means clustering algorithm is applied on this data using .NET language on Visual Studio 2012. Four clusters are formed i.e. low, high, risky and high risky to detect the transaction whether it is fraud transaction or legitimate transaction.

3.2 Pseudo code of new algorithm

The Pseudo code which we applied for credit card fraud detection is as:

1. Variable declaration
2. Validation
3. Making an entry of the input transaction to the database
4. Getting training data set
5. Converting list of transaction data objects to multi-dimensional
6. Apply clustering
7. Assigning cluster name/label
8. Commit transaction to database either as fraud or legitimate

Variable declaration: First, the variables used in this program are declared such as transaction amount, credit card number, new transaction, transaction date, merchant category id, transaction type id and transaction country.

Validation: It shows the validity of the details required for the transaction such as amount, country, credit card number, merchant category, transaction type, and date of transaction. Suppose the amount, name of country is entered and if u forget to enter the credit card number then it will not processed further and the dialog box will open with an instruction to enter the credit card number. As all the details are filled it will process further.

Making an entry of the input transaction to the database: The data table that is generated before is now entered into the database. It includes transaction country, transaction amount, transaction id, credit card number, merchant category id, transaction date.

Getting training data set: The data which is taken from the data table is now entered to get transaction data.

Converting list of transaction data object to multi-dimensional array: Here, the array is used so that the details will generate row wise. They are transaction amount, credit card number and transaction id etc.

Apply clustering: The k- means clustering is used by applying array, the four cluster data set (0, 1, 2, 3) are formed. As shown in the figure. The four clusters are formed that are represented by the four different colors orange, yellow, green and violet. Assuming that the orange color is for low cluster (i.e. cluster id 0), yellow color is for high cluster (i.e. cluster id 1), green color is for risky cluster (i.e. cluster id 2), violet for high risky cluster (i.e. cluster id 3).

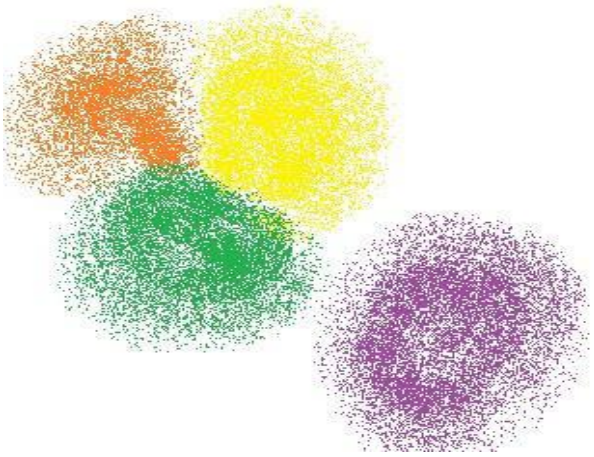

Figure 2 Clusters formation

Assigning cluster name/label: The four clusters which are formed are named/labelled as low cluster, high cluster, risky cluster and high risky cluster.

Commit transaction to database either as fraud or legitimate transaction: the current transaction details were taken and by using k means algorithm the fraud is detected. If it is fraud then the message will show 'fraud' or else it will show 'legitimate'.

4. RESULTS

Here, the results are shown that are found by using K-means clustering algorithm. In K-means algorithm, four clusters are formed as shown in Figure 2 above in experiment process which is low cluster, high cluster, risky cluster and high risky cluster as presented by colors (orange, yellow, green and violet). In this, the transaction on five credit card number is tested by applying K-means clustering. The results are as shown below:

1. When the amount is deposited in India by using the credit card number 176345, the transaction done belongs to low cluster and this is found to be legitimate transaction as shown in Figure 3. It means that though the transaction belongs to the low cluster (orange) there is risk of fraud. But it doesn't mean that it is fraud it can be legitimate transaction also. Legitimate transaction means that it is right and justifiable transaction.

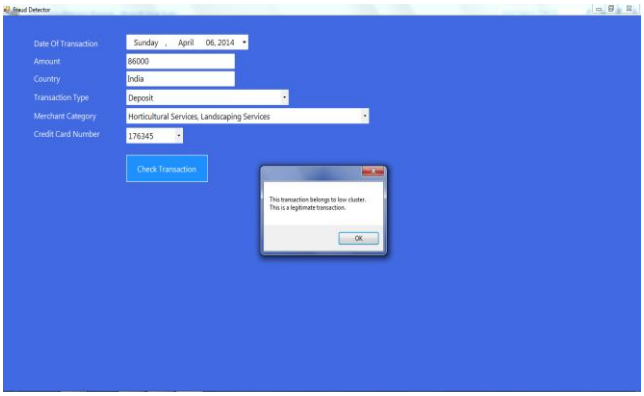

Figure 3 Low cluster and legitimate transaction

2. When the amount is deposited in Ukraine having the credit card number 176345, the transaction done belongs to low cluster and this is found to be fraud transaction as shown in figure 4. It means that as the transaction belongs to the low cluster (orange) there is risk of fraud and it is detected as fraud transaction.

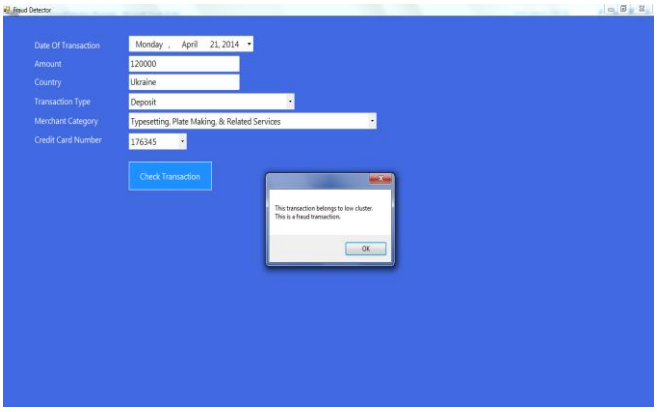

Figure 4 Low cluster and fraud transaction

3. Here, as the amount is withdrawn from Equador by using the credit card number 456723, the transaction done belongs to high cluster (yellow) and this is found to be fraud transaction as shown in figure 5. It shows that, as the transaction belongs to the high cluster (yellow) the risk of fraud is high. And it is fraud transaction.

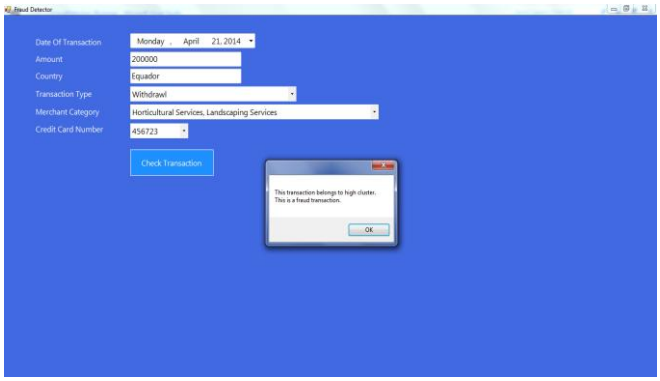

Figure 5 High cluster and fraud transaction

4. The amount is deposited from Sri Lanka by using the credit card number 234562, the transaction done belongs to risky

cluster (green) and this is found to be legitimate transaction as shown in figure 6. It shows that, as the transaction belongs to the risky cluster (green) the chance for fraud is more than the low and high. But then also it is not fraud transaction but legitimate transaction.

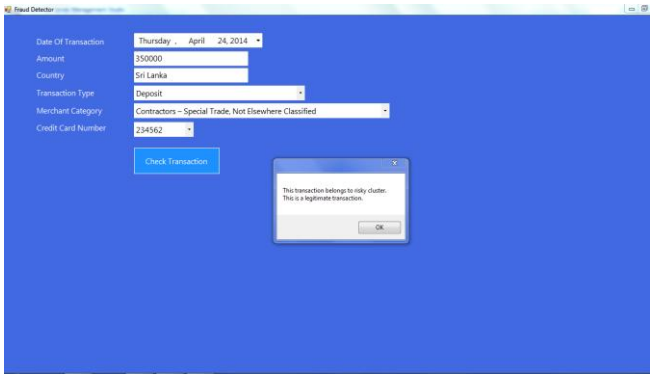

Figure 6 Risky cluster and legitimate transaction

5. The amount is deposited from India by using the credit card number 234562, the transaction done belongs to high risky cluster (violet) and this is found to be fraud transaction as shown in figure 7. It shows that, as the transaction belongs to the high risky cluster (violet) the chance for fraud is more than the low, high and risky cluster.

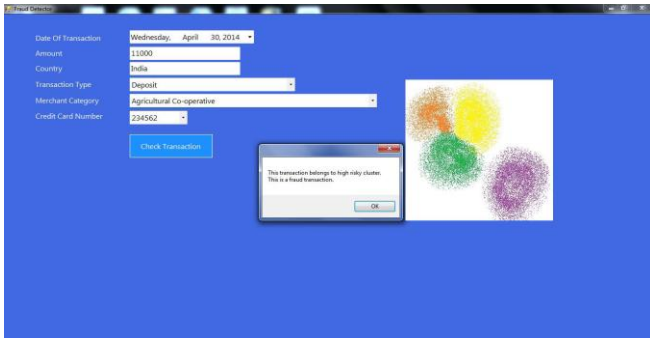

Figure 7 High risky cluster and fraud transaction

5. CONCLUSION AND FUTURE WORK

Fraud cannot be detected accurately 100%. Somehow, there were always some errors present while detecting the fraud. In this, we can see that no matter how much risky is the transaction but it is not necessary that chances of fraud are 100%. It is not like that the fraud is always present. May be the transaction is of high risk but we cannot say that whether it is fraud or legitimate transaction.

In these experiments, we found that the Clustering algorithm when implemented showed significant results. Most of the fraudulent activities could be correctly identified. However, there were quite a few non-fraudulent activities, which wrongly got detected as frauds. To detect the fraud accurately and efficiently, it is necessary that the real data should be available. It is found that the transaction is either fraud or legitimate through K-means clustering algorithm easily.

The future work will be to improve the fraud transactions by using simulated annealing. For this the real data is necessary thing to improve credit card fraud.

6. REFERENCES

[1] R.J. Bolton and D.J. Hand, **"**Unsupervised profiling methods for fraud detection", Department of Mathematics Imperial College London {r.bolton, d.j.hand}@ic.ac.uk

[2] A.Dharmarajan, T. Velmurugan, "Applications of partition based clustering algorithms: A survey", IEEE International Conference on Computational Intelligence and Computing Research 2013.

[3] S. V. Pons and J. R. Shulcloper, "Partition selection approach for hierarchical clustering based on clustering ensemble ", Springer-Verlag Berlin Heidelberg 2010

[4] S. H.,"Gene expression data knowledge discovery using global and local clustering", Journal of computing, volume 2, issue 3, march 2010.

[5] J.S. Mishra, S. Panda, and A. Kumar Mishra, "A novel approach for credit card fraud detection targeting the Indian market" IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 2, May 2013.

[6] S. Esakkiraj and S. Chidambaram, "A predictive approach for fraud detection using hidden markov model" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 1, January- 2013 C.

[7] V.S. Sunderam, G.D. Albada and P.M.A. Sloot, "Computational Science ICCS 2005".

[8] Celebi, Kingravi and Vela, "A comparative study of efficient initialization methods for the K-Means clustering algorithm", Expert systems with applications, 40(1): 200–210, 2013.

[9] Xiuchang and Wei, "An improved K-means clustering algorithm", Journal of networks, Vol. 9, No. 1, January, 2014.