# A Survey on Clustering Techniques based on Money Laundering Fraud Detection

**Vaibhavi Patel[1] Mikin Patel[2]**

[1,2]Department of Computer Engineering

[1,2]Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India - 388430

*Abstract*— Nowadays, banking frauds like money laundering becomes more and more sophisticated. Money laundering is a process by which legally obtained funds are given the appearance of having been legally obtained. There are various techniques for detecting the different types of frauds. These techniques include different data mining techniques. Data mining is a process of discovering the interesting and useful patterns and relationships in large volumes of data. Data mining techniques such as cluster analysis, classification, neural network, and prediction have been used to detect the banking fraud. Clustering analysis groups the similar objects into a cluster in such way that intracluster similarity of objects is high and intercluster similarity is low. The clustering result is calculated based on cluster quality and cluster performance. This paper presents survey on various clustering methods. These clustering techniques measure the similarity based on some criterion functions.

***Key words:*** Clustering techniques, Data mining, Money laundering

## I. INTRODUCTION

Money laundering (ML) is a serious problem for both economics and financial institutions [1] [7]. Every year, the huge amount of funds is generated from illegal activities like drug trafficking, people smuggling and these funds are mostly in the form of cash. The criminals who generate these funds need to bring them into the legal financial system without any distrust. The conversion of cash into forms makes it more practical. And thus, there is a gap between criminal activities and funds.

So basically, money laundering is not a single act process but, it is divided into three different processes: placement, layering, and integration.

1) Placement: This is the first and riskiest stage. When the person receives the huge amount of money it places it in bank account or distributes them over several accounts.
2) Layering: Then this amount is passed on various overseas bank accounts in the world where each bank follows strict secrecy codes or may be transferred to a single account. The purpose of this step is to hide the origin of illegal money. And it becomes hard to detect the origin of the money.
3) Integration: At this stage, the illegal money looks like legal money by investing this amount in legal business, or making purchase of luxury resources [7].

Data mining is a process of extracting and discovering the interestingness and useful knowledge from large amount of data. Data mining techniques like cluster analysis, classification, association analysis and regression are used to detect the money laundering fraud. These techniques and related algorithms are described in below as shown in figure.

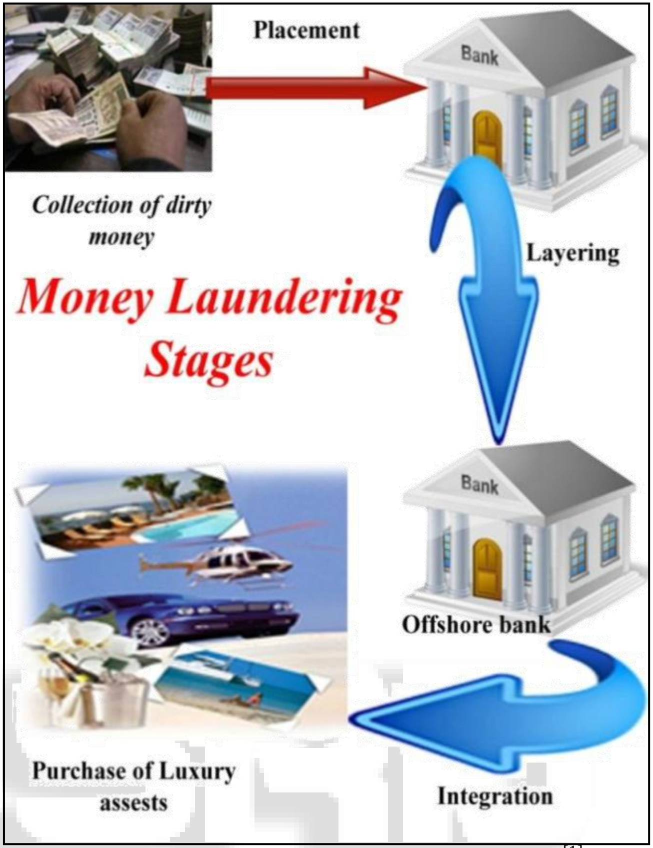

Fig. 1: Generic process of Money Laundering [1]

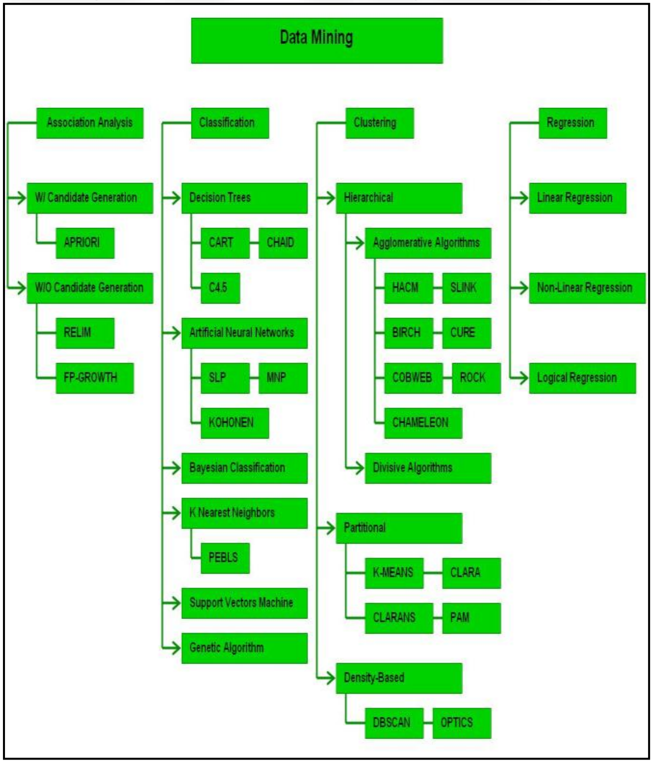

Fig. 2: Data mining techniques [9]

Various data mining applications are as below:

- Financial data analysis
- Retail industry
- Telecommunication industry
- Scientific applications
- Biological data analysis
- Intrusion detection

The rest of the paper is organized as follows: Section 2 contains the clustering techniques with their advantages and disadvantages. Section 3 finally concludes the paper.

## II. CLUSTERING TECHNIQUES

Clustering anlysis is a process of grouping the objects such that the objects within a group will be similar and dissimilar from the objects of other groups [4]. It is a unsupervised learning method which finds a stucture of unlabled data. Clustering measures the similarity between objects with the use of similarity funcitions. Clustering techniques are used to detect the banking fraud [3].

Clustering techniques broadly classified into five categories: partitioninig method, hierarchical mehtod, grid based clustering, density based clustering, model based clustering [11].

### A. *Partitioning Methods*

These methods partition the data in subset to check all possible subset systems is computaionaly not feasible [2]. It repositions the objects from one cluster to another. It is also used in top down method.

Basically it has two methods: k-mean, k-medoids

1) *Advantages*
   - Very simple
   - Easy to implement
2) *Disadvantages*
   - The number of clusters will be predefined.

### B. *Hierarchical Methods*

These methods build the hirarchy of the clusters. Thus, it connects the objects to make clusters based on the their distance. With different distances, dissimilar clusters are formed which are represented by using dendrogram [5] [6]. These methods further classified into two categories: agglomerative methods (i.e bottom up approach), divisive methods (i.e top down approach). And related algorithms are BIRCH, Chameleon, ROCK.

1) *Advantages*
   - Applicable to each attribute type
   - Easy to handle any form of distance or similarity
2) *Disadvantages*
   - Do not revisit the cluster once it has been created.

### C. *Grid based clustering*

These types of clustering methods focus on spatial data objects. This data models the geometric structure of the objects, their relationships,operations and properties. The dataset is quantized into cells which forms a grid stucture [5] [6]. And thus, the various operations are done on this newly formed grid cells. The example of grid based clustering is STING (STatistical INformation Grid based methods) algorithm which splits the data into grids and then starts its further working. Another example is CLIQUE (CLustering In QUEst) [6].

1) *Advantages*
   - Lower Processing time
   - Easy to design
2) *Disadvantages*
   - Limitation on the shape of clusters

### D. *Density based clustering*

Density based clusteing methods overcomes the disavantage of grid based mehtos i.e they allow to produce the arbitrary shapes of clusters [4]. In this method, the objects are grouped according to some specific density objective function. Density is the number of objects in a particular neighborhood of objects. The clusters are continuing to grow until the density in its neighborhood reaches certain threshold. Two types of this methods are: Density based connectivity clustering and density funcitons clustring. The examples of these methods are DBSCAN (Density Based Spatial Clustering of Application with Noise), OPTICS (Ordering Points to Identify the Clustering Structure), DENCLUE (DENsity based CLUstEring)[6].

1) *Advantages*
   - Capability to discover clusters of arbitrary shape.
   - Handle noisy data
2) *Disadvantages*
   - Clusters may be merged by a very narrow dense link

### E. *Model based clustering*

In model based methods, the data are typically clustered using some assumed mathematicl modelling structures. A model is hypothesized for each cluster and the best fit of data is foung for the given model [10]. The clusters may be located by constructing a density function which reflects the spatial distribution of data points. Practically, cluster will be mathematically represented by a parametric distribution (continuous or discrete). The examples of model based clustering are: EM algorithm (uses a mixture density model), Conceptual clustering (such as COBWEB), and neural network (like self organizing feature maps) [10].

1) *Advantages*
   - Flexible to choose the component distribution
   - It obtains the density estimation for each cluster
2) *Disadvantages*
   - Overfitting problem will be occurred.

## III. EVALUATION AND ASSESMENTS OF CLUSTERING TECHNIQUES

After performing the clustering algorithm, its clustering results is evaluated which is called as clustering validation.

### A. *Internal Evaluation*

In this evaluation, the clustering results is estimated based on the data which is clustered itself. It is well suited for the algorithm which produces clusters with high similarity within cluster and low similarity with other clusters. It is not suited for information retrieval applications. There are two methos for evaluating the quality of clusters: Dunn index, Davies-bouldin index [11].

### *B. Exrenal Evaluation*

In this type of method, clustering results are evaluated based on that data which are not used for clustering. The methods to measure thr quality of clusters are: Rand measure, F-measure [11].

## IV. CLUSTERING APPLICATIONS

- Pattern recognition
- Image analysis
- Bioinformatics
- Machine learning
- Voice mining
- Text mining

## V. CONCLUSION

From above discussion, it can be concluded that the clustering techniques are able to detect the banking frauds like money laundering. The clustering techniques are used in various ways i.e it is used as standalone algorithm, as hybrid algorithm or may be as combined algorithm helping in improving the clustering parameters.

## REFERENCES

[1] Mahesh Kharote, V.P. Kshirsagar, “Data Mining Model for Money Laundering Detection in Financial Domain”, International Journal of Computer Applications, Volume 85, No. 16,2014.
[2] L.V.Bijuraj, “Clustering and its Applications”, Proceeding of National Conference on New Horizons in IT, 2013, (169-172)
[3] Andrei Sorin SABAU, “Survey of Clustering based Financial Fraud Detection Research”, Informatica Economica, Volume 16, No. 1, 2012.
[4] Amandeep Kaur Mann, Navneet Kaur, “Survey Paper on Clustering Techniques”, IJSETR, Volume 2, Issue 4, April 2013, (803-806).
[5] Pradeep rai, Shubha Singh, “A Survey of Clustering Techniques”, IJCA, Volume 7, No. 12, October 2010.
[6] P. IndiraPriya, Dr. D.K. Ghosh, “A Survey on Different Clustering Algorithms in Data Mining Technique”, IJMER, Volume 3, Issue 1, 2013, (267-274).
[7] P. Umadevi, E. Divya, “Money Laundering Detection Using TFA System”, IET Publisher, 2012.
[8] http://en.wikipedia.org/wiki/Money_laundering
[9] http://www.dataminingarticles.com/img_bk/DMTechniques3.png
[10] http://home.deib.polimi.it/matteucc/clustering/tutorial_html/mixture.html
[11] http://en.wikipedia.org/wiki/Cluster_analysis#Applications