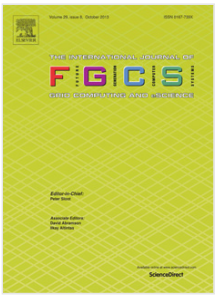
Accepted Manuscript



A survey of anomaly detection techniques in financial domain

Mohiuddin Ahmed, Abdun Naser Mahmood, Md. Rafiqul Islam

PII: S0167-739X(15)00002-3
DOI: http://dx.doi.org/10.1016/j.future.2015.01.001
Reference: FUTURE 2691

To appear in: *Future Generation Computer Systems*

Received date: 30 November 2013
Revised date: 28 August 2014
Accepted date: 6 January 2015

Please cite this article as: M. Ahmed, A.N. Mahmood, M.R. Islam, A survey of anomaly detection techniques in financial domain, *Future Generation Computer Systems* (2015), http://dx.doi.org/10.1016/j.future.2015.01.001

A Survey of Anomaly Detection Techniques in Financial Domain

Highlights

- Recently, in the financial sector, there has been renewed interest in research on detection of fraudulent activities.
- This paper presents an in-depth survey of various clustering based anomaly detection technique and compares them from different perspectives.
- In addition, we discuss the lack of real world data and how synthetic data has been used to validate current detection techniques.

A Survey of Anomaly Detection Techniques in Financial Domain

Mohiuddin Ahmed[a], Abdun Naser Mahmood[a], Md. Rafiqul Islam[b]

[a]*School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2600, Australia*
[b]*School of Computing and Mathematics, Charles Sturt University, Albury, NSW 2640, Australia*

Abstract

Anomaly detection is an important data analysis task. It is used to identify interesting and emerging patterns, trends and anomalies from data. Anomaly detection is an important tool to detect abnormalities in many different domains including financial fraud detection, computer network intrusion, human behavioural analysis, gene expression analysis and many more. Recently, in the financial sector, there has been renewed interest in research on detection of fraudulent activities. There has been a lot of work in the area of clustering based unsupervised anomaly detection in the financial domain. This paper presents an in-depth survey of various clustering based anomaly detection technique and compares them from different perspectives. In addition, we discuss the lack of real world data and how synthetic data has been used to validate current detection techniques.

Keywords:
Clustering, Fraud Detection, Anomaly Detection.

1. Introduction

Anomaly detection is an important aspect of data mining, where the main objective is to identify anomalous or unusual data from a given dataset. Anomaly detection is interesting because it involves automatically discovering interesting and rare patterns from datasets [1]. Anomaly detection has been widely studied in statistics and machine learning [2], where it is also known as outlier detection, deviation detection, novelty detection, and exception mining. Anomalies are important because they indicate significant

but rare events, and they can prompt critical actions to be taken in a wide range of application domains. For example, an anomaly in an MRI image may indicate the presence of a malignant tumour [3]. Similarly, abnormal behaviour in a credit card transaction could indicate fraudulent activities [4], an unusual traffic pattern in a network could mean that a computer is hacked or under attack, e.g., using worms and Denial of Service (DoS) attacks [5]. In this paper, we focus on anomaly detection research in the financial domain. The key contribution of this paper is it provides a structured and broad overview of extensive research on anomaly based fraud detection using clustering techniques, while providing insights into the effectiveness of these techniques in detecting anomalies. The issue of dearth of financial data is also discussed, along with a discussion on how to generate synthetic data for this domain. Although, there are a few survey papers on anomaly detection (see Table 1), however, there is a lack of focus on unsupervised techniques to detect fraud, which is important since new fraudulent attacks are being invented every day, and existing signature based techniques [6, 7] are unable to detect them.

Table 1 shows the set of methods, application domain and data covered by this survey and related surveys. Chandola et al [2] provided an extensive survey encompassing various techniques and application domains but did not focus on clustering based fraud detection, instead they discussed signature based fraud detection techniques. Andrei et al [8] also reviewed the fraud detection techniques using clustering but with limited details of the individual approaches. The survey also proposed a technique which was claimed to be clustering-based, while also using supervised learning methods, such as, Support Vector Machine [9]. Phua et al [10] also categorised, compared, and summarized a good number of published technical and review articles in automated fraud detection, however, newer attacks have emerged in the last 10 years since the report was published. Patcha et al [11] and Hodge et al [12] also presented various anomaly detection techniques based on supervised, unsupervised and clustering methods but not for fraud detection. Markou et al [13] and Beckman et al [14] also provided surveys on anomaly detection but only on supervised methods. Interestingly, none of these surveys discussed the issues of training and test data for anomaly detection which we have addressed in this survey.

Section 2 provides a background on fraud detection and financial losses

occurred due to fraud and, also discusses the fundamentals of clustering based anomaly detection. Section 3 and 4 includes discussion on partition and hierarchical clustering based fraud detection techniques. Section 5 includes miscellaneous types of clustering based fraud detection. In each section, we summarize the key features of the discussed techniques. Section 6 describes the data scarcity issue and how to address it using synthetic datasets. Section 7 concludes the paper.

Table 1: Comparison of this survey to other related surveys. ∗This Survey

Survey	**Supervised**	**Unsupervised**	**Clustering**	**Fraud Detection**	**Data Issue**
Ahmed[*]		√	√	√	√
Andrei [8]		√	√	√	
Chandola [2]	√	√		√	
Patcha [11]	√	√	√		
Bakar [15]	√	√			
Phua [10]	√	√	√	√	
Hodge [12]	√	√	√		
Markou [13]	√				
Beckman [14]	√				

2. Preliminaries

According to The American Heritage Dictionary [16], '*Fraud is defined as a deception deliberately practiced in order to secure unfair or unlawful gain*'. Fraud and financial crimes are a category of larceny that takes place when a person or entity embezzles money or property, or uses them in an illicit manner, with the intent to gain benefit from it. Davia et al [17] stated, '*Fraud always involves one or more persons who, with intent, act secretly to deprive another of something of value, for their own enrichment*'. Wells et al [18] stressed '*deception*' as the linchpin to fraud. Occupational fraud and abuse may be defined as: '*The use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organizations resources or assets.*' This definition encompasses a wide variety

of conducts by executives, employees, managers and principals of organizations. FBI has been treating such crimes as white collar crimes [19] and a number of other organizations are also dedicated to diminish such crimes.

According to the 2013/2014 global fraud report from Kroll [20], the level of fraud increased by every measure in the past 12 months. Overall, 70% of companies reported suffering from at least one type of fraud in the past year, up from 61% in the previous poll. Individual businesses also faced a more diverse range of threats on average, compared with the threats in 2012. Finally, the economic cost of these crimes mounted, increasing from an average of 0.9% of revenue to 1.4%, with one in ten businesses reporting a cost of more than 4% of revenue. The following Table 2 shows the summary of fraud profiles from different sectors.

Table 2: The summary of fraud profiles on different sectors [20]

Sector	**Level of Financial Loss**	**Investment on Anti-Fraud**
Natural Resource	Medium	High
Medical,Healthcare	High	High
Retail,Wholesale	Medium	Low
Financial Services	High	Medium
Manufacturing	High	Low
Information Technology	Medium	Low

2.1. **Areas of Fraud**

Fraud occurs when users misuse the resources provided by organization in an illegal way. However, organizations are interested in prompt action against fraudulent activities. Some of the most common applications of fraud detection are discussed below.

- **Credit card:** Everyday business, online shopping, electronic banking are largely dependent on credit cards. Meanwhile, the misuse of credit

or debit card is on the rise and a source of common fraud. The advancement of technology made it easier to produce counterfeit cards and conduct fraudulent transactions. There are a number of ways to fraud credit cards. Capturing card details using hidden device in an ATM (Automated Teller Machine) or shoulder surfing are examples of how credit cards can be fraudulently used. Figure 1 shows a forged card reading device installed over the original one. The device also has a pinhole camera which can record the video of a person entering PIN (Personal Identification Number) on the keypad. At a later time, the device is used to retrieve card details for illegal use.

Figure 1: Bankcard skimming device attached in an ATM [21]

Credit card frauds anomalies can be identified using transactional records and generally correspond to any of a number of indicators including unusually high payments, purchase of extremely unusual items, and high frequency of purchase. Customer profiling as well as detecting anomalies in consumer's transactional records are possible ways to detect fraud. The challenge associated with detecting fraudulent transactions are detecting them in an online and unsupervised manner, so that the new types of fraud are detected as soon as they happen.

- **Mobile Phone:** Mobile phone fraud includes a variety of scams that persuade consumers to buy various products. For example, offer of free

ring tones or chance to win fantastic prizes by replying to messages or calls. After responding to this mobile call and signing up to this service, one could be left facing a huge phone bill. To prevent the misuse of mobile phone account, it is necessary to detect any unusual usage pattern. The basic technique is to monitor the usage pattern and create customer profile of each of the account. Consequently, any deviation from the common usage pattern will issue an alarm to warn the customer to avoid probable scams. The challenge here is to create a representation of a normal user profile to identify unusual behaviours.

- **Insurance claim:** Insurance fraud is simply another form of theft which involves trick or deception. A recent study estimated that insurance fraud costs more $2 billion annually in Australia [22]. Automobile related fraud is a common type of insurance fraud. Criminal rings of illegal claimants manipulate the claim processing system by submitting forged documents. Typically, insurance investigators asses the claims and look for fraud on the submitted documents manually. However, detecting the illegal claims in an automated way is a key challenge for insurance fraud investigators.

- **Insider trading:** In recent times, stock market investment has received renewed interest from the general public. Insider trading is a criminal activity in the stock market, where profit is made by using inside information before it is made public. The inside information is of different forms and can affect the stock prices artificially [23]. In this domain, early detection is important to avoid individual or organizations from making illicit profits.

2.2. Types of Anomalies

Anomalies are referred to as patterns in the data that do not conform to a well-defined characteristic of normal behaviour. Anomalies are generated due to variety of abnormal activities, e.g., credit card fraud, mobile phone fraud, cyber-attacks etc. The underlying events are important and interesting to the data analyst. An important aspect of anomaly detection is the nature of anomaly. Anomalies can be categorized in the following ways.

- **Point Anomaly:** When a particular data instance deviates from the normal pattern of the dataset, it can be considered as a point anomaly. For a realistic example, we can consider expenditure on car fuel. If the usual car fuel usage of a person is five litres/day but if it becomes fifty litres in any random day then it is a point anomaly.

- **Contextual Anomaly:** When a data instance is behaving anomalously in a particular context, but not in other context, then it is termed as a contextual anomaly, or conditional anomalies. For example, the expenditure on credit card during a festive period, e.g., Christmas or New Year, is usually higher than the rest of the year. Although, the expenditure during a festive month can be high, it may not be anomalous due to the high expenses being contextually normal at that time. On the other hand, an equally high expense during a non-festive month could be deemed as a contextual anomaly.

- **Collective Anomaly:** When a collection of similar data instances are behaving anomalously with respect to the entire data set, then this collection is termed as collective anomaly. It might happen that the individual data instance is not an anomaly by itself, but due to its presence in a collection it is identified as an anomaly. For example, in a human Electro Cardio Gram (ECG) output, the existence of low values for a long period of time indicates underlying phenomenon corresponding to abnormal premature contraction [3], however, one low value by itself is not considered as anomalous.

2.3. Key Research Challenges in Anomaly Detection

The task of detecting anomalies by analysing deviations from normal behaviour pattern is non-trivial. Some of the key research challenges are as follows.

- There is a lack of effective general purpose anomaly detection techniques. An anomaly detection technique in one domain may not be suitable for other domains, since, both the normal and abnormal behaviour varies from domain to domain. For example, the technique[24]

used in stock exchange does not work well for network traffic analysis, although they both have transaction data.

- Data contains noise, which appears to be an anomaly, hence difficult to distinguish from an actual anomaly.
- Labelled data for training of models are scarcely available.
- Normal behaviour usually change, therefore, what is considered as normal now, may not be considered as normal in future. Thus, the current anomaly detection techniques may not work properly in future.
- Malicious adversaries try to make abnormal activities appear normal by imitating normal behaviour.

For these above-mentioned challenges, anomaly detection is not a trivial problem to solve. Most of the existing techniques are based on application domain and nature of data. Researchers from diversified disciplines have adopted various concepts to solve these problems. The general techniques are based on the concepts of classification, nearest neighbor, statistics, information theory and clustering [2]. Interestingly, all these concepts are covered by the existing surveys in various domains such as cyber security, image processing, medical and health systems, sensor networks etc. except the clustering based anomaly detection techniques in financial domain, which inspired us to investigate the concept of clustering for financial fraud detection. Next, we describe clustering, which is an useful unsupervised technique to identify normal and anomalous data.

2.4. Clustering Technique to Detect Anomaly

Due to the advancement of computing and the proliferation of data repositories, appropriate knowledge discovery approaches are required for extracting potentially useful information from data. In data mining, supervised and unsupervised learning techniques are two widely used knowledge discovery techniques.

Supervised learning is the machine learning task of inferring a function from labelled training data [25]. The training data consist of a set of training examples. In supervised learning, the training examples consist of an input object and a desired output value. A supervised learning algorithm learns

from the training data and creates a knowledge base which can be used for mapping new and unseen data. Classification refers to supervised learning algorithms, where pre-labelled data is required. Labelled data are rare and difficult to find. However, when pre-labelled data is available, the unseen data cannot be mapped which are not present in the labelled data, such as zero day attacks in intrusion detection domain.

Unsupervised learning tries to find hidden structure in unlabelled data, which distinguishes unsupervised learning from supervised learning [26]. Clustering refers to unsupervised learning algorithms, where pre-labelled data is not required to extract rules for grouping similar data instances [27]. Clustering process results in different partitioning of a given dataset based on different criteria used for clustering. There are different types of clustering techniques, but here we are going to discuss the ones that have been used in the literature for detecting fraudulent activities. A simple taxonomy of clustering is provided in Figure 2. Next we discuss about the basic steps of the clustering process as shown in Figure 3.

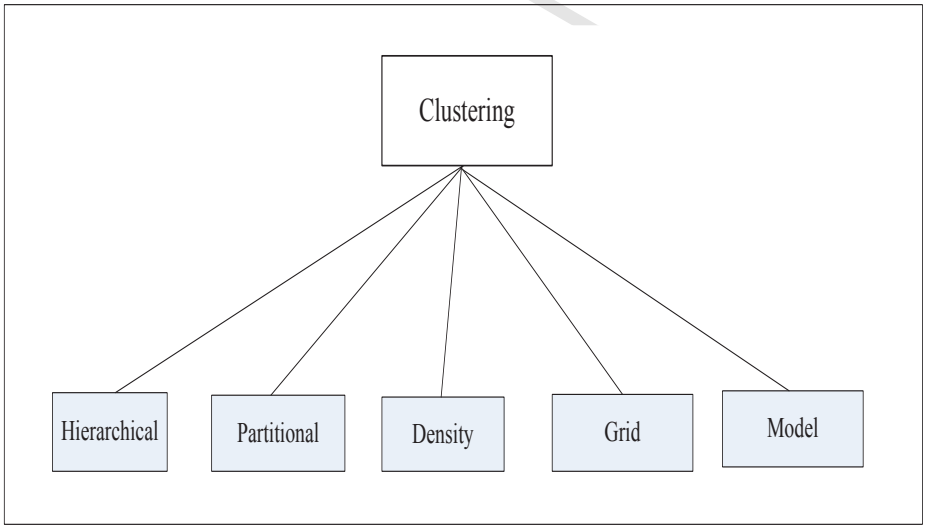


Figure 2: Taxonomy of clustering approaches

2.5. Basic Steps of Clustering

Figure 3 describes the clustering process and basic steps like feature extraction, clustering algorithm, validity and result interpretation are briefly explained below. From Figure 3, it is clear that clustering process requires input in the form of a dataset and its output is the discovered knowledge or a model.

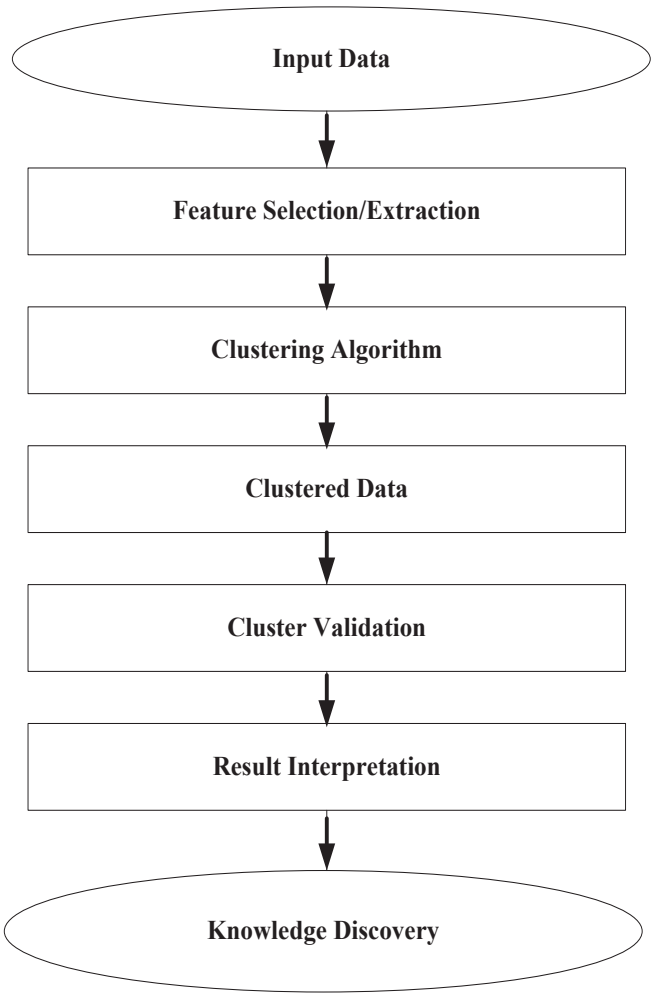


Figure 3: Steps in clustering

- **Feature Selection/Extraction:** Feature selection is often the first step in data pre-processing before clustering. Feature selection methods select a subset of features from the whole dataset, while in feature extraction some data transformations are required to generate new features from the original ones [28]. Feature selection or extraction plays an important role for effective clustering of the input dataset. It is expected that, features should be selected carefully, which are immune to noise and easy to extract and interpret. Doing this can greatly reduce the computational complexity and simplify the subsequent process.

- **Clustering Algorithm:** Now, let us formally define clustering. Given a set of input patterns, X = x_1, x_2,..,x_N. Clustering attempts to seek a k-partition of X, C = C_1, C_2....C_k ($K \leq N$), such that

  1. $C_i \neq \emptyset$, i = 1,.....,K; a cluster cannot be empty.

  2. $\bigcup_{i=1}^{k} C_i$ = X; the clustering process should cluster all the data points.

  3. $C_i \bigcap C_j = \emptyset$, i,j = 1,.....,$K$ and i $\neq$ j; each data object should belong to exclusively one cluster, however there are other clustering techniques e.g. fuzzy clustering, which extends the notion to associate each data object to more than one clustering using a membership function.

  4. $d(x_i, x_j)$ in C_i is minimized; distance between points inside a cluster, the intra cluster distance is minimized.

  5. $d(C_i, C_j)$ is maximized, distance between clusters, the inter-cluster distance is maximized.

- **Selection of Clustering Algorithm:** Selection of clustering algorithm is a vital step to cluster the underlying data. Similarity/dissimilarity measure also known as proximity measure quantifies how similar two data points are. A good clustering criterion leads to a partition that fits the data well. So, proximity measure and clustering criterion play a vital role in determining the accuracy of a clustering algorithm.

- **Cluster Validity:** Different clustering algorithms produce different sets of clusters; therefore, it is necessary to evaluate the clustering quality to decide which algorithm performs better. Since clustering is an unsupervised process, irrespective of clustering method, the final partition of the data set needs to be evaluated. Generally, there are

three categories of validity criteria, which are relative, internal and external [29]. External criteria evaluate the structure of resulting cluster with a prespecified structure which reflects a priori knowledge on clustering structure of the dataset. In contrast to external criteria, internal criteria evaluate the clustering structure without any prior knowledge. External and internal criteria are computationally intensive. Relative criteria compare the structure of resulting cluster with other clustering structure or the same clustering technique with different parameters on the dataset and find the best representation.

- **Result Interpretation:** The aim of clustering is to provide users with meaningful insights from the original data. However, it is the job of data expert or analyst to interpret the clustering solution and take further action. Clustering result interpretation for anomaly detection is discussed in more detail in Section 2.6

2.6. Key Assumptions in Clustering based Anomaly Detection

Since the goal of clustering is to group together similar data, it can be used to detect anomalous patterns in a dataset. There are three key assumptions when using clustering to detect anomalies. These are briefly discussed below.

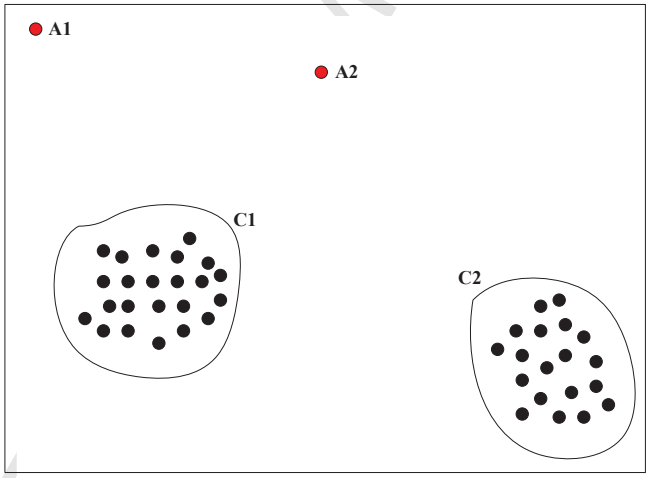


Figure 4: Example of anomaly based on A-1

- **Assumption 1 (A-1):** We can create clusters of normal data only, subsequently, any new data that do not fit well with existing clusters of normal data are considered as anomalies. For example, density based clustering algorithms do not include noise inside the clusters [30]. Here noise is considered as anomalous. For example, in the Figure 4, *C1* and *C2* are clusters containing normal instances and *A1,A2* are anomalies.

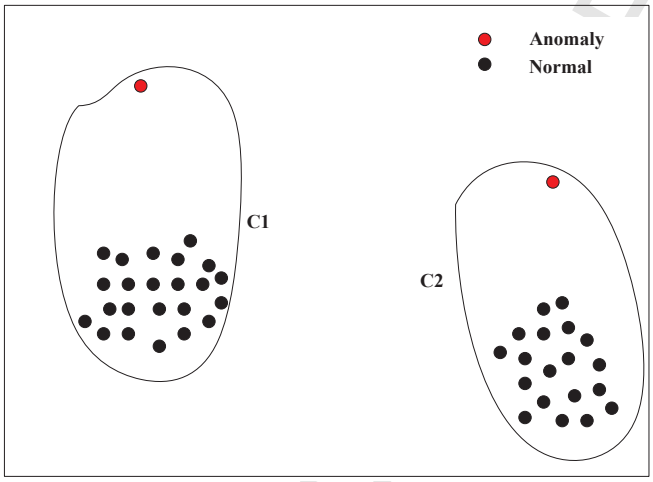


Figure 5: Example of anomaly based on A-2

- **Assumption 2 (A-2):** When a cluster contains both normal and anomaly data, it has been found that normal data lie close to the nearest cluster centroid but anomalies are far away from the centroids [31]. Under this assumption, anomalous events are detected using a distance score. For example, Svetlona et al [32] presented an outlier removal clustering algorithm (ORC) that provides outlier detection and data clustering simultaneously. Their proposed algorithm has two stages. First, the *k-means* clustering is applied and then *outlyingness factor*, o_i for each of the data point, x_i is calculated by taking the ratio of a point's distance to the centroid, C_{pi} and the maximum distance, d_{max} from the centroid to any other point, stated in equation (1). If outlying factor for any point is greater than a threshold T, it is considered as an outlier and removed from dataset. Their experimental data includes synthetic data and some map images. Mean Absolute Error (MAE) is

used to evaluate their algorithm performance and the value of parameter T is dependent on the dataset.

$$o_i = \frac{\|x_i - C_{p_i}\|}{d_{max}} \tag{1}$$

- **Assumption 3 (A-3):** In a clustering where there are clusters of various sizes, smaller and sparser can be considered as anomalous and thicker clusters can be considered normal. The instances belonging to clusters whose size and/or density is below a threshold are considered as anomalous. He et al [33] proposed a definition for cluster based local anomalies. According to their definition, all the data points in a certain cluster are considered as anomalies rather than a single point, as shown in Figure 6. The clusters *C1* and *C3* are considered as anomalous. They used some numeric parameters, i.e. α, β to identify Small Cluster (SC) and Large Cluster (LC). They used the SQUEEZER algorithm to cluster data, as it achieves both high quality of clustering and can handle high dimensional data. Then the FindCBLOF algorithm determines outlier factor of each individual record in dataset. CBLOF(t) for each record t is calculated following equation (2):

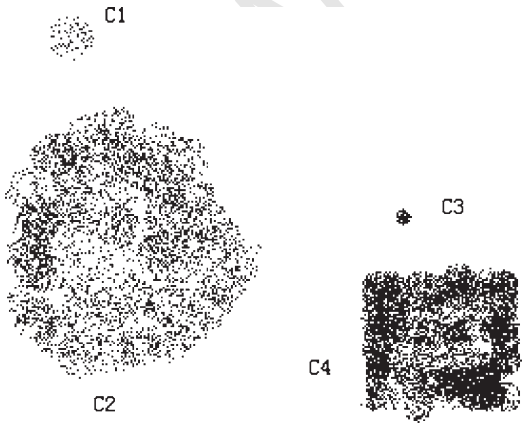


Figure 6: Anomalous clusters C1,C3; adapted from [33]

$$CBLOF(t) = \begin{cases} |C_i| * min(d(t, C_j)) & where\ t \in C_i, C_i \in SC\ and \\ & C_j \in LC\ for\ j = 1\ to\ b \\ |C_i| * (d(t, C_i)) & where\ t \in C_i\ and\ C_i \in LC \end{cases} \quad (2)$$

3. Partition based Clustering Techniques for Fraud Detection

Partitional clustering divides a set of data objects into a pre-specified number of clusters without any hierarchical structures. In this section the characteristics of different partition based clustering and recent fraud detection methods using such clustering will be discussed. Among the partition based clustering algorithms the popular *k-means* [34] clustering is widely used. PAM [35], CLARA [36], CLARANS [37] are also used as partitional clustering. *k-means* algorithm has a number of variants to optimize the clustering solutions, e.g., continuous *k-means* [38], compare means [39], trimmed *k-means* [40], *X-means* [41], *k-harmonic means* [42], *k-modes* algorithm [43], *k-prototype* [44], *k-probabilities* [45], however, we discuss the basic *k-means* in the next section..

Table 3: The main characteristics of partitional clustering algorithm

Algorithm	**Primary Data Type**	**Complexity**
k-means	Numerical	$\mathcal{O}(n)$
k-modes	Categorical	$\mathcal{O}(n)$
PAM	Numerical	$\mathcal{O}(k(n-k)^2)$
CLARA	Numerical	$\mathcal{O}(k(40+k)^2 + k(n-k))$
CLARANS	Numerical	$\mathcal{O}((kn)^2)$
FCM	Numerical	$\mathcal{O}(n)$

*n is the number of points in the dataset and k is the number of cluster

3.1. Basic k-means Algorithm

k-means is a well-known unsupervised clustering algorithm, also known to converge to a local minimum of the distortion measure i.e. the Euclidean distance. Basic *k-means* algorithm shows the basic steps of *k-means* clustering. In *k-means* clustering each object is assigned to precisely one of k

Basic *k-means* Algorithm
1. Select k points as initial centroids.
2. *Repeat*
3. Form k clusters by assigning each point to its closest centroid.
4. Recompute the centroid of each cluster.
5. *Until* Centroids do not change.

clusters. The algorithm takes an input parameter, k, which is known a priori by data expert and partitions a set of n objects into k-clusters. Once the clustering has been done, the resulting intra-cluster similarity is high but the inter-cluster similarity is low. The algorithm works as follows: at first, k-initial centroids are chosen randomly from the set of n objects. Each object is assigned to its closest cluster based on its Euclidian distance to the cluster centroid. The set of points assigned to a centroid is considered as a cluster. This is how k-clusters is formed. Next, the centroid of each cluster is updated based on the mean of the objects assigned to it. This process is repeated, so that each point is assigned to the nearest cluster based on the changes in the position of the cluster centroid. The process halts when no object changes the cluster or the centroids stop moving.

Table 4: The Symbols used in *k-means* algorithm

Symbol	**Description**
x	A vector representing an object
Capital C_i	The i^{th} cluster
Small c_i	Centroid of cluster C_i
k	The number of clusters
C_m	The mean of all points

To measure the quality of a clustering, *k-means* algorithm uses the Sum of Squared Error (SSE) and Total Sum of Squares (SST). Euclidian distance between each object and the centroid of the cluster to which it belongs, represents an error, and from this the total sum of squared errors is computed. SST is the squared total sum of distances between the mean of the dataset and all the points in dataset. Table 4 defines the notations used in the basic *k-means* algorithm. Given two different sets of clusters produced by *k-means*,

the clustering which has a lower SSE/SST, is considered better. SSE and SST are formally defined in equation (3) and (4)

$$SSE = \sum_{i=1}^{k} \sum_{C_i} dist(c_i, x)^2 \quad where \quad \forall x \in C_i \tag{3}$$

$$SST = \sum dist(C_m, x)^2 \quad where \quad \forall x \; in \; Dataset \; D \tag{4}$$

3.2. Recent Fraud Detection Techniques based on Partitional Clustering

Table 5: Key features of partitional clustering based fraud detection

Technique	Dataset	Tool	Reference
k-means	Refund transactions	WEKA	[46]
k-means	Claims data	WEKA,SAS	[47]
k-means	Transaction data	∗	[48]
X-means	Real transaction Histories	WEKA	[49]
Hybrid *k-means*	Financial Statement	∗	[24]
k-means	Purchasing data	ANOVA Analysis	[50]

∗ indicates not specified

Issa et al [46] proposed an anomaly detection method to identify fraudulent refunds by applying *k-means* algorithm to a refund transaction dataset from a telecommunication company. The dataset required pre-processing to improve the efficiency of the algorithm. As mentioned earlier in Section 3.1, the objective function of *k-means* is Sum of Squared Errors (SSE), which needs to be optimized, clearly the higher value of k, the lower the SSE, but after a certain point the decrease in SSE becomes insignificant and any increase in the value of k does not improve the clustering accuracy. Here, value of k was assigned as k = 2, 3,4,5,6. WEKA [51] is used to obtain basic *k-means* clustering results. The records which are far away from the rest are considered as suspicious and examined further by an expert from the telecommunication Company. No formal definition or an explanation of the detection method is provided in their work.

Thiprungsri et al [47] presented an approach to help auditors identify fraud when evaluating group life insurance claims. The basic idea to detect fraudulent claims is same as assumption 3 of clustering based anomaly detection (discussed in Section 2.6). Clusters with smaller population were considered as suspicious. The dominant characteristics of the suspicious clusters are: large beneficiary payment, huge interest amount and long processing time. The dataset has 40080 group life insurance claims paid out in 2009. *k-means* algorithm was applied using WEKA [51] and SAS [52] tool. The clusters that have smaller populations (less than 1% of the whole population) are considered as probable fraud. To verify the cluster analysis, whether it can really identify fraud or not, the suspicious clusters are selected for further investigations by experts.

Chang et al [49] used *X-means* [41] algorithm, which is advantageous over basic *k-means* algorithm to automatically determine the number of clusters. *X-means* algorithm uses the Bayesian Information Criterion (BIC) measure stated in equation (5) to determine the number of clusters in the underlying data. WEKA [51] tool was used for clustering and analyse the behaviour changes of online auction fraudsters in Yahoo!Taiwan [53].

$$BIC(M_j) = l_j(D) - \frac{P_j}{2} \log R \tag{5}$$

Where $l_j(D)$ is the log-likelihood of the data according to the j^{th} model and taken at the maximum likelihood point. P_j is the number of parameters in M_j and R refers to the size of dataset D. BIC is also known as Schwarz Criterion.

Nhien et al [48] presented a case study on anti-money laundering detection, where data mining and natural computing techniques were combined. Here *k-means* clustering was used because of its simplicity and the convex shape of the cluster. Here the value of k was fixed by anti-money laundering experts and suspicious/unsuspicious groups were created. The dataset consists of ten million transactions records from six funds administered by an anonymized bank's ten thousands customers.

Jans et al [50] proposed an approach to detect three purchase related fraud schemes: double payment of invoices, changing purchase order after

release and deviation of purchase order. The dataset used here is gathered from Enterprise Resource Planning (ERP) system. The data contains information about purchasing orders, goods receipt and invoices. Using *k-means* algorithm and analysing the results of ANOVA analysis [54], the high mean ratios of certain attributes within smaller clusters are treated as anomalies.

Deng et al [24] proposed a clustering model VKSOM combining *Self Organizing Map* and *k-means* clustering for fraud detection in financial statements. The model enjoys the benefit of unsupervised self-learning SOM [55]. Then *k-means* clustering is applied on the results of SOM. The VKSOM clustering model applies *Silhouette Index* to measure the validity of different clustering results. The experimental data includes financial statements of 100 listed companies in China during 1999-2006, where half of statements were Financial Fraudulent Statement (FFS).

Table 6: Comparison among the partitional clustering based fraud detection techniques

Reference	**Specific Domain**	**Assumptions of Fraud**
[**46**]	Telecommunication Company	Section 2.6 A-2
[**47**]	Life Insurance	Section 2.6 A-3
[**48**]	Bank	Frequency of change in activity
[**50**]	ERP	Section 2.6 A-3
[**24**]	Stock Exchange	Doubtful auditors report

Table 6 provides a brief comparison among the aforementioned fraud detection techniques based on partitional clustering. These techniques were developed for specific application domains which is clearly visible in Table 6. Furthermore, we could classify these techniques according to the clustering based anomaly detection assumptions discussed in Section 2.6. The insurance company and Enterprise Resource Planning (ERP) system fall under the assumption 3 (A-3), where the small clusters are considered as anomalous/fraudulent. Due to privacy, the datasets used in these techniques are not publicly available, which raises a constraint to compare the accuracy of detection among various techniques. Moreover, all of the techniques required involvement of experts to identify or confirm anomalies, however, the extent

of involvement varied among the different applications.

4. Hierarchical Clustering based Techniques for Fraud Detection

4.1. Hierarchical Clustering

These algorithms can be further divided into two types: agglomerative and divisive algorithm. Agglomerative clustering approaches are bottom up, as they start by considering each data point as a singleton cluster and iteratively merge it with the nearest data point (Figure 7). Divisive approaches are less popular as they need to consider $2^{N-1}-1$ possible two subset divisions for a cluster with N data points, which is computationally expensive [56]. Some representative hierarchical algorithm includes BIRCH [57], CURE [58], and ROCK [59].

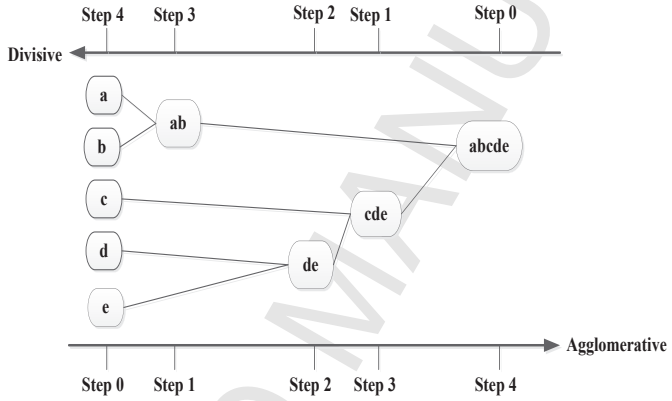


Figure 7: Hierarchical clustering

Among these most commonly used hierarchical clustering, BIRCH [57] has the lowest computational complexity. However, BIRCH is sensitive to the data ordering and lacks the ability to handle mixed data types i.e. combination of numerical and categorical data. In this category, CURE [58] is another hierarchical clustering algorithm which can recognize arbitrarily shaped clusters better than others, although having higher complexity than BIRCH. ROCK [59] has the maximum computational complexity with the advantage of handling categorical data.

Table 7: The main characteristics of hierarchical clustering algorithm

Algorithm	Primary Data Type	Complexity
BIRCH	Numerical	$\mathcal{O}(n)$
CURE	Numerical	$\mathcal{O}(n^2 \log(n))$
ROCK	Categorical	$\mathcal{O}(n^2 + nm_m m_a + n^2 \log(n))$

*n = number of points in dataset, k = number of cluster,
m_m = maximum number of neighbours of a point,
m_a = average number of neighbours of a point

4.2. Recent Fraud Detection Techniques based on Hierarchical Clustering

A Computational Model for Fraud Detection (CFDM) is proposed by Glancy et al [60] for detecting fraudulent financial reports. The manipulation of information in annual filings with the US Securities and Exchange Commission (SEC) is considered as fraud in this scenario. Both Expectation Maximization [61] and hierarchical clustering were used by this model on text documents for fraud detection. The quantitative model conducted repeated clustering on the text documents to create two clusters, where one cluster contains fraudulent information and another is normal. Using the sign test [62], clustering results were evaluated and only three false positives were present out of sixty nice sets of documents.

Torgo et al [63] proposed a method for utility based fraud detection using a hierarchical agglomerative clustering algorithm. The assumption on the fraudulent activity is that, the data corresponding to fraudulent activities will not group together with the normal data clusters. The outlier ranking, OR_h calculates the degree of being a fraudulent activity of each data instance using equation (7). For each of the merging steps in the hierarchical clustering i involving two groups of data ($g_{x,i}$ and $g_{y,i}$) the $of_i(x)$ is calculated using equation (6),

$$of_i(x) = max(0, \frac{|g_{y,i}| - |g_{x,i}|}{|g_{y,i}| + |g_{x,i}|}) \tag{6}$$

Where $g_{x.i}$ is the group to which data instance x belongs, and $|g_{x,i}|$ is the

group cardinality.

$$OR_h(x) = max_i(of_i(x)) \tag{7}$$

The data source was a national institute of statistics and contained the financial transactions of companies with foreign countries. The dataset covered a wide range of products containing 410 000 transaction reports over 8 months.

Another approach using the benefit of the dendrogram generated by hierarchical clustering algorithms is presented by Torgo et al [64]. The notion of fraud is the height in the dendrogram, where the data instances are merged in a cluster. It was assumed that, the fraudulent data instances will merge at the end of the clustering process and the normal instances will merge very early stage. The degree of being fraudulent instances, $OF_h(x)$ is calculated using equation (8)

$$OF_h(x) = \frac{h}{N} \tag{8}$$

where h is the level of the hierarchy H at which the data instance is merged and N is the number of training instances.

Table 8: Key features of hierarchical clustering based fraud detection techniques

Dataset	**Tool**	**Reference**
Annual Financial data	SAS Enterprise Miner	[60]
Foreign Trade Transaction	Not Specified	[63]
INTRASTAT data(INE)	R Package	[64]

Table 8 summarizes the hierarchical clustering based fraud detection techniques where the dataset and tools used are shown. Table 9 provides a comparison among the techniques based on the particular application domains of fraudulent activity, fraud definition and the characteristics of the data used. It is clear that, here in the case of hierarchical clustering also, different domains are targeted. Consequently, the assumptions on fraudulent behaviour vary. The datasets used are also publicly unavailable. Although

the techniques proposed by Torgo et al. [63, 64] focused on same domain, but emphasized on the resources and utilities to enhance the detection accuracy.

Table 9: Comparison among the hierarchical clustering based fraud detection techniques

Reference	**Specific Application**	**Assumptions of Fraud**	**Data Type**
[60]	SEC	Violation of SEC laws	Text
[63]	Foreign Trade	Higher utility requirements	Numerical
[64]	Foreign Trade	Information fabrication	Numerical

5. Miscellaneous Clustering Techniques for Fraud Detection

In this section, we discuss about the other approaches used for anomaly detection using clustering, namely, density based, latent and resolution based clustering.

Wang et al [65] presented a hybrid outlier mining algorithm to evaluate client moral risk in an insurance company. A resolution and density based outlier mining algorithm is introduced which is efficient for detecting fraudulent applications from the clients. The algorithm is a combination of resolution based algorithm (RB algorithm is a nonparametric outlier mining algorithm with greater accuracy) and density based algorithm [30] to detect top-n fraudulent activity without the need of input parameters.

A methodology called IFR^2 Framework [66] is utilized by Jans et al [67] to detect fraud in procurement process. The term IFR^2, stands for Internal Fraud Risk Reduction, which is a conceptual framework to reduce internal fraud risk. Emphasis is given on the application of the IFR^2 Framework for a real life database. The dataset used in this study was obtained from an international financial services provider which is one of the largest European financial institutions. In the proposed framework, LC clustering [68] is used to differentiate between fraud and normal activity. It is a variant of *k-means* clustering where probabilities are considered to define 'closeness' to each cluster centroid. LC clustering was preferred over the basic *k-means*

clustering due to it's cluster overlapping feature. Additionally, LC clustering algorithm has the ability to handle mixed data types and determines the number of clusters using statistical criteria.

Rui et al [69] presented a core decision tree algorithm to identify money laundering activities. The proposed algorithm is a combination of hierarchical clustering algorithm (BIRCH [57]) and partitional clustering algorithm (*k-means* [34]). BIRCH is an incremental algorithm which works efficiently for large databases; however, it cannot handle financial data very well. Unlike *k-means*, BIRCH requires more than one parameter such as (i) initial threshold (ii) page size (iii) outlier options (iv) memory size. As a result, proper parameter setting is a challenge for the financial data analysts. Additionally, BIRCH algorithm was initially developed to handle real image datasets. BIRCH is also sensitive to the data ordering and lacks the ability to handle mixed data types i.e. combination of numerical and categorical data which are commonly found in financial data. Synthetic dataset was used for experimentation but no anomalous data detection results were provided.

Panigrahi et al [70] used the popular density based clustering algorithm, DBSCAN [30] for credit card fraud detection. The assumption for fraudulent activity is quite straightforward. DBSCAN algorithm creates clusters from the underlying data and can separate noise from the original data. In this case, the noises are considered as credit card fraud. According to their detection scheme, a financial transaction is detected as anomalous when it does not belong to any cluster in the set C, where set C defines the set of clusters in database D. A degree of being fraud is given by $d_{outlier}$ (9), if $|N_\epsilon(p)| <$ MinPts; 0 otherwise, where, MinPts: Minimum number of points required in the ϵ neighborhood of each point to form a cluster.

$$d_{outlier} = 1 - \frac{\epsilon}{v_{avg}} \tag{9}$$

ϵ: Maximum radius of the neighborhood $N_\epsilon(p) = q \in D | dist(p, q) \leq \epsilon$. A simulator was developed to generate a synthetic transaction history that represents the behaviour of genuine cardholders as well as the fraudsters.

Jyotindra et al [71] proposed a hybrid approach for online fraud detection involving financial transactions. The hybrid approach is a combination of data mining, artificial intelligence and statistics. User profiles are cre-

ated using the current and past online activities. Density based clustering algorithm forms the clusters of the online transactions by the users. A new transaction is defined highly genuine, if it matches with past transactions. An online shopping organization volunteered the real credit card transaction dataset for this research with the condition of confidentiality.

Table 10: Key features of miscellaneous clustering based fraud detection techniques

Technique	**Dataset**	**Tool**	**Reference**
BIRCH,*k-means*	Synthetic	$*$	[69]
DBSCAN	Online Shopping Data	$*$	[71]
Resolution based	Policy Holder's data	JAVA	[65]
Latent Clustering	Purchasing order data	$*$	[67]
DBSCAN	Synthetic data	$*$	[70]

$*$ indicates not specified

Table 11: Comparison among the hierarchical clustering based fraud detection techniques

Reference	**Specific Application**	**Notion of Fraud**
[69]	Money Laundering	Recurrent transactions
[71]	Online Shopping	Suspicious behaviour
[65]	Insurance Company	Noisy Data
[67]	Procurement Process	Internal Fraud
[70]	Credit Card	Noisy Data

Table 10 and 11 displays the summary of different types of clustering techniques for fraud detection. There is no generalized technique since each of these techniques are developed targeting a specific domain. However, the clustering methodology may be similar. For example, Wang et al [65] and Panigrahi et al [70] both used DBSCAN clustering but for insurance company fraud and credit card fraud respectively. However, it is difficult to compare the detection accuracy due to the dataset unavailability and privacy issue.

Despite the lack of generality, the results from [65, 70] indicate that DBSCAN seems to work well in insurance and credit card fraud detection.

6. Scarcity of Publicly Available Real Data

There are very few publicly available real world datasets for conducting research in financial fraud detection [72]. However, obtaining real data from financial organizations for research purposes is extremely difficult due to privacy and competitive reasons. Synthetic data provides a solution for this data unavailability issue and gives a scope to work on fraud detection in a particular application domain [73]. However, the synthetic data generation process has the following challenges:

- Data has to reflect the normal and abnormal behaviour in a particular application domain. For example, while generating the synthetic data for credit card transactions, abnormal behaviour in the context of an insurance company are not taken into consideration. As a result, the fraud detection techniques are becoming context specific, i.e., credit card fraud detection techniques are not applicable for the insurance company fraudulent activities (also discussed in Section 3.2, 4.2 and 5). Consequently, synthetic data generation process is different for different application domains.

- Due to the above issue, synthetic data generation process requires assistance from an expert in the required domain. The fraud scenarios are designed by the expert.

- Accuracy of fraud detection varies from technique to technique and from domain to domain because one technique applied on synthetic data may result in many false positives and vice-versa because technique used successfully in one domain may result in lower accuracy in detecting fraud in another domain..

- Normal behaviour evolve over time, therefore, what is considered as normal now, may not be considered as normal in future. Thus, the basic assumptions or the definition of fraud will evolve too. Consequently, synthetic data generation process needs to take this change into account.

- There are also other factors that influence the synthetic data produced, i.e., human intervention, algorithmic limitations - such as the inability to generate true random numbers, etc.

- Figure 8 shows a comparison on the accuracy of different anomaly detection techniques (KNN [74], LOF [75], CBLOF [33] and LDCOF [76]) applied on real-world dataset and synthetic data. It is evident from Figure 8 that the performance of anomaly detection techniques on synthetic KDD cup 99 dataset [77] is better than real network traffic [78]. This shows that, we cannot rely solely on synthetic data to judge performance of anomaly detection techniques. We can also infer that, while developing anomaly detection techniques, synthetic data characteristics are considered to analyse normal and anomalous behaviour due to the real world data unavailability. However, while these techniques are applied on read world data, the performance is not acceptable. In addition, synthetic KDD cup dataset is unable to correctly represent the real-world scenario and due to this reason, both statistical and behavioral differences exist between synthetic and real world data.

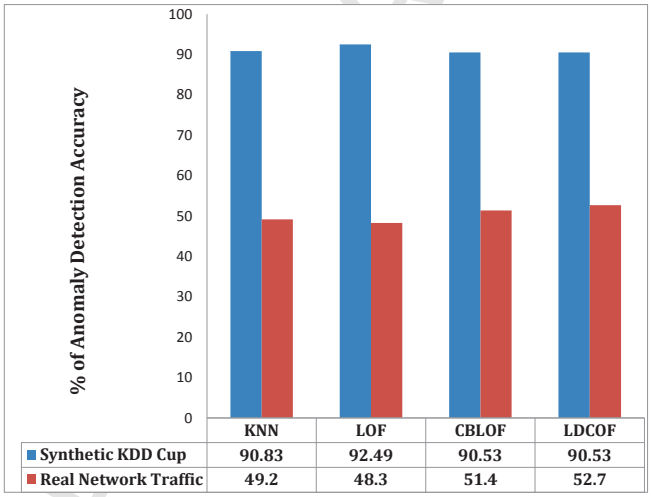


Figure 8: Accuracy comparison on Synthetic KDD Cup 99 [77] and Real Network Traffic [78]

6.1. Synthetic Data Generation Process

A methodology is developed by Barse et al [79] for generating synthetic data which is derived from real data. Figure 9 shows the various steps in Synthetic data generation. The initial step requires the data collection to represent the behaviour of the specific application. It is expected that, the data should contain both normal and anomalous data. To create a behaviour model from the collected data, in the second step data analysis is required. The third step contains the identification of parameters required to detect fraudulent activity. Based on the parameters identified in step three, both normal and fraud profiles are created in step 4. With the help of a simulator, the normal and fraud profile activities are simulated, thus the system is modelled in the last step.

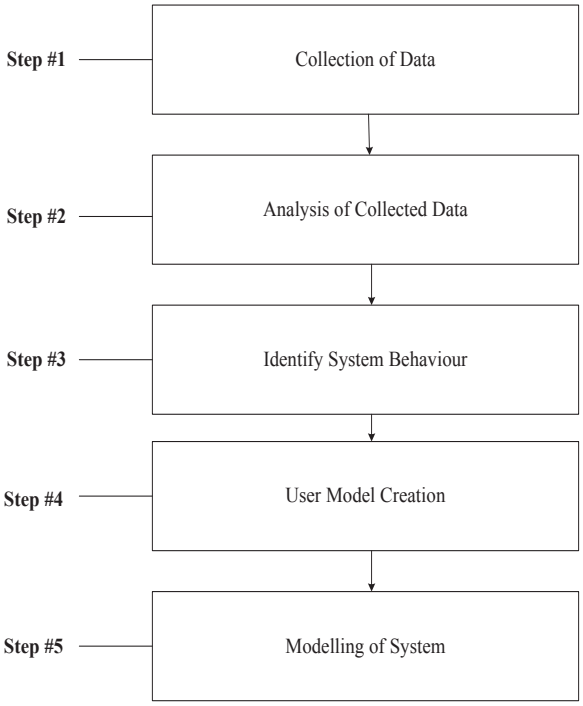


Figure 9: Synthetic Data Generation

Using the synthetic data generation process discussed above, Barse et al [79] described an example of synthetic data generation for Video-on-Demand (VoD) system. A brief description of the four fraudulent behaviours

in this VoD system is given below.

1. **Break-in fraud:** In this case, the legal user's identity is compromised and used with malicious intent.
2. **Billing fraud:** The access to the billing server is gained by the hacker and is being manipulated. For example, not paying for the ordered items.
3. **Illegal redistribution fraud:** A legal user distributes the contents which are not accessible by others such as non-members of the service.
4. **Failed logins:** With the intent to gain illegal access, guess password option is being tried several times.

Another approach was demonstrated by Fawcett et al [80]. In their study, they used email data for spam because of the availability in large amount. Spam is also known as Unsolicited Commercial Email (UCE) and Unsolicited Bulk Email (UBE), a frequent item in email communication. This type of data requires thorough feature selection and text processing operations. However, this option makes it difficult to model different types of fraudulent activities.

Table 12: Comparison of the synthetic data generation process

Dataset	**VoD [79]**	**UCE [80]**
Interchangeability	No	No
Process complexity	Complex	Simple
Privacy	Low	High
Type of fraud	Multiple	Single

Table 12 displays comparison between the two approaches of synthetic data generation discussed above. These approaches are from two different application domains, i.e., Video-on-Demand service and Unsolicited Commercial Email, however, reflects the fraudulent activities to conduct research to develop fraud detection prototype. Obviously, the notion of fraud is dissimilar to each other and different fraud detection techniques are not interchangeable with these two different types of data. From the generation

process it is clear that, the VoD system data requires more complex operation and processing than the spam (UCE) dataset. UCE dataset generation process follows only single types of fraud whereas the other one injects four different types of fraudulent activities. However, the privacy issue in email communication dataset is higher than the video on demand service.

7. Conclusion

Anomaly is a pattern in the data which does not conform to the expected behaviour which needs to be further analysed, such as fraudulent activity. Clustering is an unsupervised learning method, which is a powerful technique to identify normal and anomalous data. In this paper, a structured survey on clustering based fraud detection is presented and the issue of scarcity of real data is discussed. Different types of fraudulent activities are outlined and a brief report on recent global fraud is provided. A short description of different aspects of anomaly detection followed by clustering techniques is presented. The key assumptions of clustering based anomaly detection techniques and recent research on fraud detection using different types of clustering methods are also described. Finally, the unavailability of financial data for fraud detection research is explained and a methodology for synthetic data generation is discussed. We conclude that, a universal technique in the domain of fraud detection is yet to be found, due to the evolving change in context of normality and labelled data unavailability.

References

[1] M. Ahmed, A. N. Mahmood, J. Hu, Outlier detection, in: The State of the Art in Intrusion Prevention and Detection, CRC Press, USA, 2014, pp. 3–23.

[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (3) (2009) 15:1–15:58.

[3] J. Lin, E. Keogh, A. Fu, H. Van Herle, Approximations to magic: finding unusual medical time series, in: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, CBMS '05, IEEE Computer Society, Washington, DC, USA, 2005, pp. 329–334.

[4] R. J. Bolton, D. J. H, Unsupervised profiling methods for fraud detection, in: Proc. Credit Scoring and Credit Control VII, 2001, pp. 5–7.

[5] D. J. Marchette, Computer intrusion detection and network monitoring: a statistical viewpoint, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[6] E. Aleskerov, B. Freisleben, B. Rao, Cardwatch: a neural network based database mining system for credit card fraud detection, in: Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE, 1997, pp. 220–226.

[7] R. Brause, T. Langsdorf, M. Hepp, Neural data mining for credit card fraud detection, in: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, ICTAI '99, IEEE Computer Society, Washington, DC, USA, 1999, pp. 103–106.

[8] A. S. SABAU, Survey of clustering based financial fraud detection research, Informatica Economica 16 (1) (2012) 110–122.

[9] J. Wu, H. Xiong, J. Chen, Cog: local decomposition for rare class analysis, Data Mining and Knowledge Discovery 20 (2) (2010) 191–220.

[10] C. Phua, V. Lee, K. Smith, R. Gayler, A comprehensive survey of data mining-based fraud detection research, Artificial Intelligence Review (2005) 1–14.

[11] A. Patcha, J.-M. Park, An overview of anomaly detection techniques: existing solutions and latest technological trends, Comput. Netw. 51 (12) (2007) 3448–3470.

[12] V. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2) (2004) 85–126.

[13] M. Markou, S. Singh, Novelty detection: a review - part 2: neural network based approaches, Signal Processing 83 (2003) 2499–2521.

[14] R. J. Beckman, R. D. Cook, Outliers, Technometrics 25 (2) (1983) 119–149.

[15] Z. Bakar, R. Mohemad, A. Ahmad, M. Deris, A comparative study for outlier detection techniques in data mining, in: Cybernetics and Intelligent Systems, 2006 IEEE Conference on, 2006, pp. 1–6.

[16] American heritage dictionary, accessed: 2013-11-29.
URL http://www.ahdictionary.com

[17] P. C. J. W. Davia, H. R., J. Kastantin, in: Accountant's guide to fraud detection and control, John Wiley and Sons, 2000.

[18] J. Wells, in: Principles of fraud examination, John Wiley and Sons, 2005.

[19] FBI: White Collar Crime, accessed: 2013-11-29.
URL http://www.fbi.gov

[20] 2013/2014 Global Fraud Report, accessed: 2013-11-29.
URL http://www.kroll.com

[21] European ATM skimming machine your credit card's new worst enemy in Australian crime first, accessed: 2013-11-29.
URL http://www.couriermail.com.au/news

[22] Insurance Fraud Bureau of Australia, accessed: 2013-11-29.
URL http://www.ifba.org.au

[23] S. Donoho, Early detection of insider trading in option markets, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, ACM, New York, NY, USA, 2004, pp. 420–429.

[24] Q. Deng, G. Mei, Combining self-organizing map and k-means clustering for detecting fraudulent financial statements, in: Granular Computing, 2009, GRC '09. IEEE International Conference on, 2009, pp. 126–131.

[25] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of machine learning, The MIT Press, 2012.

[26] C. M. Bishop, Neural networks for pattern recognition, Oxford University Press, Inc., New York, NY, USA, 1995.

[27] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.

[28] A. Jain, R. P. W. Duin, J. Mao, Statistical pattern recognition: a review, Pattern Analysis and Machine Intelligence, IEEE Transactions on 22 (1) (2000) 4–37.

[29] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2-3) (2001) 107–145.

[30] M. Ester, H. peter Kriegel, J. S, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press, 1996, pp. 226–231.

[31] M. Ahmed, A. N. Mahmood, A novel approach for outlier detection and clustering improvement, in: Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on, 2013, pp. 577–582.

[32] V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, P. Fränti, Improving k-means by outlier removal, in: Proc. 14th Scandinavian Conference on Image Analysis (SCIA'05), 2005, pp. 978–987.

[33] Z. He, X. Xu, S. Deng, Discovering cluster based local outliers, Pattern Recognition Letters 2003 (2003) 9–10.

[34] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: L. M. L. Cam, J. Neyman (Eds.), Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, 1967, pp. 281–297.

[35] L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis (Wiley Series in Probability and Statistics), 1st Edition, Wiley-Interscience, 2005.

[36] R. T. Ng, J. Han, Efficient and effective clustering methods for spatial data mining, in: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, pp. 144–155.

[37] R. T. Ng, J. Han, Clarans: a method for clustering objects for spatial data mining, IEEE Trans. on Knowl. and Data Eng. 14 (5) (2002) 1003–1016.

[38] V. Faber, Clustering and the continuous k-means algorithm, Los Alamos Science 22 (1994) 138–144.

[39] S. Phillips, Acceleration of k-means and related clustering algorithms, in: D. Mount, C. Stein (Eds.), Algorithm Engineering and Experiments, Vol. 2409 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2002, pp. 166–177.

[40] J. Cuesta, J. Albertos, A. Gordaliza, A. Ramos, C. Bea, U. de Cantabria. Servicio de Publicaciones, Trimmed k-means: an attempt to robustify quantizers, Universidad de Cantabria, 1994.

[41] A. M. Dan Pelleg, X-means: extending k-means with efficient estimation of the number of clusters, in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 2000, pp. 727–734.

[42] L. Zhang, L. Mao, H. Gong, H. Yang, A k-harmonic means clustering algorithm based on enhanced differential evolution, 2013 Fifth International Conference on Measuring Technology and Mechatronics Automation 0 (2013) 13–16.

[43] A. Chaturvedi, P. E. Green, J. D. Caroll, k-modes clustering, Journal of Classification 18 (1) (2001) 35–55.

[44] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Discov. 2 (3) (1998) 283–304.

[45] D. Wishart, k-means clustering with outlier detection, mixed variables and missing values, in: M. Schwaiger, O. Opitz (Eds.), Exploratory Data Analysis in Empirical Research, Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg, 2003, pp. 216–226.

[46] H. Issa, M. Vasarhelyi, Application of anomaly detection techniques to identify fraudulent refunds, 2011, pp. 1–19.

[47] S. Thiprungsri, M. Vasarhelyi, Cluster analysis for anomaly detection in accounting data: an audit approach, in: The International Journal of Digital Accounting Research, Vol. 11, 2011, pp. 69–84.

[48] N. A. L. Khac, M. Kechadi, Application of data mining for anti-money laundering detection: a case study, in: Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, 2010, pp. 577–584.

[49] W.-H. Chang, J.-S. Chang, Using clustering techniques to analyze fraudulent behavior changes in online auctions, in: Networking and Information Technology (ICNIT), 2010 International Conference on, 2010, pp. 34–38.

[50] N. L. M. Jans, K. Vanhoof, Data mining for fraud detection: toward an improvement on internal control systems?, in: European Accounting Association - Annual Congress, 2010, pp. 1–27.

[51] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, H. W. Ian, The WEKA data mining software: an update, in: SIGKDD Explorations, Volume 11, Issue 1, 2009.

[52] SAS: business intelligence software, accessed: 2013-11-29.
URL `http://www.sas.com`

[53] Yahoo!Taiwan, accessed: 2013-11-29.
URL `http://www.tw.yahoo.com`

[54] Anova analysis, accessed: 2013-11-29.
URL `http://www.csse.monash.edu.au/ smarkham/resources/anova.htm`

[55] T. Kohonen, The self-organizing map, Proceedings of the IEEE 78 (9) (1990) 1464–1480.

[56] A. N. Mahmood, C. Leckie, P. Udaya, An efficient clustering scheme to exploit hierarchical data in network traffic analysis, IEEE Trans. on Knowl. and Data Eng. 20 (6) (2008) 752–767.

[57] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large databases, in: Proceedings of the 1996 ACM SIGMOD international conference on Management of data, SIGMOD '96, ACM, New York, NY, USA, 1996, pp. 103–114.

[58] S. Guha, R. Rastogi, K. Shim, Cure: an efficient clustering algorithm for large databases, in: Proceedings of the 1998 ACM SIGMOD international conference on Management of data, SIGMOD '98, ACM, New York, NY, USA, 1998, pp. 73–84.

[59] S. Guha, R. Rastogi, K. Shim, Rock: A robust clustering algorithm for categorical attributes, Information Systems 25 (5) (2000) 345 – 366.

[60] F. H. Glancy, S. B. Yadav, A computational model for financial reporting fraud detection, Decis. Support Syst. 50 (3) (2011) 595–601.

[61] T. Moon, The expectation-maximization algorithm, Signal Processing Magazine, IEEE 13 (6) (1996) 47–60.

[62] W. J. Conover, in: Practical nonparametric statistics, 3rd ddition, John Wiley and Sons, 1999.

[63] L. Torgo, E. Lopes, Utility-based fraud detection, in: 22nd International Joint Conference on Artificial Intelligence, 2011, pp. 1517–1522.

[64] L. Torgo, C. Soares, Resource-bounded outlier detection using clustering methods, in: Proceedings of the 2010 Conference on Data Mining for Business Applications, IOS Press, Amsterdam, The Netherlands, The Netherlands, 2010, pp. 84–98.

[65] W. Xiaoyun, L. Danyue, Hybrid outlier mining algorithm based evaluation of client moral risk in insurance company, in: Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on, 2010, pp. 585–589.

[66] N. L. Mieke Jans, K. Vanhoof, A framework for internal fraud risk reduction at it integrating business processes: the ifr2 framework, in: International Journal of Digital Account Research, Vol. 9, 2009.

[67] M. Jans, N. Lybaert, K. Vanhoof, Internal fraud risk reduction: Results of a data mining case study, International Journal of Accounting Information Systems 11 (1) (2010) 17 – 41.

[68] J. Magidson, J. K. Vermunt, Statistical innovations inc. (2002).

[69] R. Liu, X. long Qian, S. Mao, S. zheng Zhu, Research on anti-money laundering based on core decision tree algorithm, in: Control and Decision Conference (CCDC), 2011 Chinese, 2011, pp. 4322–4325.

[70] S. Panigrahi, A. Kundu, S. Sural, A. Majumdar, Credit card fraud detection: A fusion approach using dempstershafer theory and bayesian learning, Information Fusion 10 (4) (2009) 354 – 363.

[71] D. J. N. Dharwa, D. A. R. Patel, A data mining with hybrid approach based transaction risk score generation model (trsgm) for fraud detection of online financial transaction, International Journal of Computer Applications 16 (1) (2011) 18–25, published by Foundation of Computer Science.

[72] K. Bache, M. Lichman, UCI machine learning repository (2013).
URL `http://archive.ics.uci.edu/ml`

[73] E. Lundin, H. Kvarnström, E. Jonsson, A synthetic fraud data generation methodology, in: Proceedings of the 4th International Conference on Information and Communications Security, ICICS '02, Springer-Verlag, London, UK, UK, 2002, pp. 265–277.

[74] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, SIGMOD Rec. 29 (2) (2000) 427–438.

[75] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, SIGMOD Rec. 29 (2) (2000) 93–104.

[76] M. G. Mennatallah Amer, Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer, Shaker Verlag GmbH, Aachen, 2012, pp. 1–12.

[77] 1999 kdd cup dataset, accessed: 2013-12-21.
URL `www.kdd.ics.uci.edu`

[78] The full dataset derived from about a month's background traffic and simulated attacks, accessed: 2013-12-21.
URL `http://seit.unsw.adfa.edu.au/staff/sites/kshafi/Datasets`

[79] E. L. Barse, H. Kvarnström, E. Jonsson, Synthesizing test data for fraud detection systems, in: Proceedings of the 19th Annual Computer Security Applications Conference, ACSAC '03, IEEE Computer Society, Washington, DC, USA, 2003, pp. 384–394.

[80] T. Fawcett, 'In Vivo' Spam Filtering: a challenge problem for KDD, SIGKDD Explor. Newsl. 5 (2) (2003) 140–148.

Mohiuddin Ahmed is working in the arena of Data Mining and Network Security towards his PhD degree at the University of New South Wales, Canberra. He received his Bachelor of Science degree in Computer Science and Information Technology from Islamic University of Technology, Bangladesh, in 2011.

Mohiuddin Ahmed
School of Engineering and Information Technology
University of New South Wales
Canberra 2600, Australia
Mohiuddin.Ahmed@student.adfa.edu.au

Abdun Naser Mahmood received the BSc degree in Applied Physics and Electronics and the MSc degree in Computer Science from the University of Dhaka, Bangladesh, in 1997 and 1999, respectively. He completed his PhD degree from the University of Melbourne in 2008. He joined the University of Dhaka as a Lecturer in 2000, Assistant Professor in 2003, when he took a leave of absence for his PhD studies. Currently, he is working as a Lecturer at the University of New South Wales, Canberra with the School of Engineering and Information Technology. His research interests include data mining techniques for network monitoring and algorithm design for adaptive sorting and sampling.

Abdun Naser Mahmood
School of Engineering and Information Technology
University of New South Wales
Canberra 2600, Australia
a.mahmood@adfa.edu.au

Rafiqul Islam is a Lecturer of Computing in the School of Computing and Mathematics, Faculty of Business, Charles Sturt University. He earned BSc (Honours) and MSc (Research) from the department of Computer Science and Engineering, University of Dhaka, and PhD degree with the specialisation in Network Security from Deakin University, Australia. Dr. Islam is a member of the Institute of Electrical and Electronics Engineers (IEEE) and Fellow member of Association of Accounting Technician (AAT).

Rafiqul Islam
School of Computing & Mathematics
Charles Sturt University
Albury 2640, Australia
mislam@csu.edu.au

Photo of Abdun Naser Mahmood

Click here to download high resolution image

Photo of Mohiuddin Ahmed

Click here to download high resolution image

Photo of Rafiqul Islam

Click here to download high resolution image