A Study on Outlier Detection using Partition Clustering Approach

K. Merlin Jeba[1] Dr.V.Srividhya[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

Abstract— Extraction of the hidden knowledge is the data mining task. While in extraction unwanted data occurred due to the relevant extraction mechanism. Those unwanted data called outliers. Detecting the outlier is an extremely important task in a wide variety of application Domains. Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data. Detecting the outliers, data mining uses many approaches. While clustering mechanism works effectively in the data mining approaches. For our study here use the clustering based outlier detection mechanism. Using the partitioning based clustering algorithm, it analyzes the outlier data well. This paper provides a study about the various partition based clustering approach to analyzing the outliers well.

Key words: Data Mining, Clustering, Partition cluster, study and Effective outlier detection

I. INTRODUCTION

Data mining is the method of haul out pattern from data. It can be used to discover patterns in data but is often carried out only on a model of data. The mining process will be unsuccessful if the samples are not the good expression of the superior body of the data [1]. The discovery of a particular outline in a particular set of data does not unavoidably mean that pattern is found in another place in the larger data from which that model was drawn. An imperative part of the method is the corroboration and legalization of patterns on other models of data. Verification is done by avoiding the scrutiny the unwanted things in our data. Here not needed data is referred to as outliers. Detecting the outliers is a wonderful in data mining task, referred to as outlier mining. Outliers are stuff that does not comply with the general actions of the data. By definition, outliers are rare happening and hence represent a small segment of the data [2]. Outlier uncovering has the direct submission in a wide variety of domains such as mining for incongruity to detect network intrusions, fraud uncovering in mobile phone industry and newly for detecting terrorism-related activities. Outlier uncovering is very essential of any modeling train. A failure to detect outliers or their ineffective management can have serious ramifications on the might of the inferences drained from the exercise [3]. There is large digit of techniques are available to execute this task, and often collection of the nearly all opposite technique poses a big defy to the practitioner. There is no regular technique for outlier detection. Some of the outlier detection practice is:

- Distance based outlier detection
- Clustering based outlier detection
- Density based outlier detection
- Depth based outlier detection

Each of these techniques has its own recompense and disadvantages. In general, in all these processes the technique to detect outliers consists of two steps. The first make out an outlier around a data set using a set of inliers (normal data). In the second step, a data request is analyzed and acknowledged as the outlier when its attributes are dissimilar from the attributes of inliers. All these systems assume that all normal instances will be like, while the anomalies will be unlike. Outliers, being the most excessive explanation, may include the sample lowest amount or sample maximum or both depending on whether they enormously high or low. However, the sample minimum or example greatest is not always outliers because they may not be strangely remote from other comments. Many numerical techniques are sensitive to the occurrence of outliers. Inspection for outliers should be a standard part of any data analysis. Outlier detection is a very imperative research work in the field of data mining. For discover the outlier this paper shows the survey of group based outlier detection method. The clustering-based process assumes that the normal data objects fit into great and dense clusters, whereas outliers belong to small or light clusters or do not belong to any clusters [4]. Clustering-based methods detect outliers by investigative the relationship among substance and clusters. Intuitively, an outlier is an object that belongs to a small and isolated cluster or does not belong to any cluster. This leads to three general methods to clustering-based outlier detection, judge an object. In the 1st method, does the object go to any cluster? If not, then it is acknowledged as an outlier. In 2nd method, is there a large detachment between the object and the cluster to which it is contiguous? If yes, it is an outlier. In final, is the object part of a small or sparse cluster? If yes, then all the objects in that cluster are outliers [5]. Like this in outlier will detect in the cluster. For perceive the outlier this paper provide the various method of the clustering algorithm for detecting the outlier in a successful manner.

II. LITERATURE REVIEW

Dr. S. Vijayarani1, Ms. P. Jothi [6] find the data stream is a new arrival in the research area in data mining whereas data stream refers to the process of extracting knowledge structures from unlimited and fast growing data records. For handling this type of stream data, the recent data mining methods are not sufficient and equipped to deal with them, for this reason, it leads to a numerous computational and mining challenges due to the shortage of hardware limitations. Nowadays many researchers have focused on mining data streams and they proposed many techniques for data stream classification and clustering, as well as mining frequent items from data streams. Data stream clustering and outlier detection provide a number of unique challenges in evolving data stream environment. Data stream clustering algorithms are highly used for sense the outliers in an efficient manner. The main purpose of this research work is to perform the clustering development and detecting the outliers in data streams. In this research work, two types of

clustering algorithms namely BIRCH with K-Means and CURE with K-means is used for finding the outliers in data streams. Two performance factors such as clustering exactness and outlier detection accuracy are used for observation. Through examining the experimental fallout, it is observed that the CURE with K-Means clustering algorithm performance is more accurate than the BIRCH with K-Means algorithm.

Ms.P.Jothi Dr. S. Vijayarani [7], focused on mining data streams and they proposed many techniques and algorithms for data streams. They are data stream classification, data stream clustering, and data stream frequent outline items and so on. Data stream clustering techniques are highly helpful to cluster the similar data items in information streams and also to detect the outliers, so they are called cluster based outlier detection. The main objective of this research work is to perform the clustering process and detecting the outliers in data streams. In this research work, two partitioning clustering algorithms namely CLARANS and E-CLARANS (Enhanced Clarans) are used for clustering and detecting the outliers in data streams. Two performance factors such as clustering accuracy and outlier detection accuracy are used for observation. By examining the experimental results, it is observed that the proposed ECLARANS clustering algorithm performance is more accurate than the existing algorithm CLARANS.

Sivaram.K, and Saveetha.D [8] analyze Outlier detection is an extremely important task in a wide variety of application Domains. Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data. It has many uses in applications like fraud detection, network intrusion detection and clinical diagnosis of diseases. In these algorithms, outliers are only by-products of clustering algorithms and they cannot rank the priority of outliers. In this paper, a proposed method based on clustering approaches for outlier detection is presented. The algorithm first performs partition clustering using one of the algorithms PAM/CLARA/CLARANS/CLATIN. The algorithm produces a set of clusters and a set of medoids (cluster centers). Small clusters are then determined and considered as outlier clusters. The rest of outliers (if any) are then detected in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each one of the points in the same cluster. The performance of the four algorithms on outlier detection efficiency was compared. The main objective is to detect outliers while simultaneously perform clustering operation.

Janpreet Singh and Shruti Aggarwal [9] analyze the data mining in various fields, due to the nature of extracting useful data from a collection of databases or data warehouses, data mining is used, with various algorithms and techniques to extract useful data from the databases. Clustering is the technique of extracting useful data from databases, but with the extraction of the object from the dataset an unwanted data also comes that is known as the outlier. To detect outlier there are various methods. In recent years outlier detection techniques are used in various field and applications such as in credit card fraud detection etc. Due to the increase of data on the web outlier detection has become an important part of the data mining. So to detect outlier from different data sets different outlier detection techniques are used with different clustering algorithms. The most popular clustering algorithm is the k-mean algorithm and is widely used to cluster the data set and for outlier detection and can be improved according to need to detect outliers.

III. PARTITIONS BASED CLUSTERING ALGORITHM USED FOR OUTLIER DETECTION

Partition based clustering create k partition of data set with n data object. It is an iterative relocation technique is used to get better the clustering by moving up the object from one group to another. Partition based clustering is represented by centroid or media. They use iterative way to produce the clustering. In partition algorithm, given n objects, these methods make k partitions of the data, by assigning objects to groups, with each partition representing a cluster [10]. Generally, each cluster must contain at least one object; and each object may belong to one and only one cluster, although this can be relaxed. The present study analyzes the use of K-Means, PAM, CLARA and CLARANS.

A. K-Means Clustering algorithm

K-means is a well-known partitioning based clustering technique that attempts to find a user specified the number of clusters represented by their centroids. K-Means clustering algorithm is simplest and widely used clustering technique. In this algorithm, the number of clusters K is specified by user means classifies instances into the predefined number of cluster. The first step of K-Means clustering is to choose k instances as a center of clusters. Next assign each instance of a dataset to nearest cluster [11]. For instance, assignment, measure the distance between the centroid and each instances using Euclidean distance and according to minimum distance assign each and every data points into the cluster. K –Means algorithm takes less execution time when it applied on small dataset [12]. When the data point increases to maximum then it takes maximum execution time. It is fast iterative algorithm but it is sensitive to outlier and noise. This is used to find the outlier in the data.

B. Partitioning Around Medoid (PAM)

PAM is developed by Kaufman and Rousseau in 1987. The algorithm chooses k- medoid initially and then swaps the medoid object with no medoid as a result quality of cluster is improved. The PAM algorithm forms clusters by examining all objects that are not medoids. This imposes an expensive computation cost of O(k(n-k)2) in each iteration [13]. Algorithm works well with small dataset but does not work well with the large dataset. However, it is very robust when to compare with k-mean in the presence of noise or outlier. PAM procedure is given in Figure 1, where k is the number of clusters, n is the number of objects in the datasets, S is the set of objects to be clustered, sj is an object $\in$ S, R denotes the set of objects $\in$ S selected of medoids, rj, rc $\in$ R, d is the dissimilarity function.

C. Clustering Large Applications (CLARA)

This algorithm deal with larger data sets, a sampling-based method, called Clustering LARge Applications (CLARA) was developed by Kaufman and Rousseeuw (1990). CLARA draws multiple samples of the data set, applies

PAM on each sample, and returns its best clustering as the output. The complexity of each iteration now becomes O(ks2 + k(n-k)), where s is the size of the sample, k the number of clusters, and depends on the sampling method and the sample size [14]. CLARA cannot find the best clustering if any sampled medoid is not among the best k medoids. When the sample size is small, CLARA's efficiency in clustering large data sets comes at the cost of clustering quality. Therefore, it is difficult to determine the sample size. This provided of size 40 + 2k gave satisfactory results. However, this is only valid for a small k.

D. CLARANS

CLARANS algorithm mixes both PAM and CLARA by searching only the subset of the dataset and it does not confine itself to any sample at any given time. One key difference between CLARANS and PAM is that the former only checks a sample of the neighbors of a node. But, unlike CLARA, each sample is drawn dynamically in the sense that no nodes corresponding to particular objects are eliminated outright. In other words, while CLARA draws a sample of nodes at the beginning of a search, CLARANS draws a sample of neighbors in each step of a search. This has the benefit of not confining a search to a localized area. This method involves partitioning clustering algorithm in data streams [15, 16]. First, the data's are split into chunks of the same size in different windows, after that consider each database(s) into data point (DP), the partition of size=s/p, along with the max neighbor of k=3. Then the minimum cost for each data point (dp) identifies the neighbor value, and it follows the condition i=1and j=1.Then the distance between each data point is calculated and also choose the maximum distance (n) for each data points, if (s) has a lower cost, set current to(s), are increment j by 1 then it returns the best cluster and detects the outliers efficiently.

E. Compare Table for Analyses of Clustering

Algorithm	Advantage	Disadvantage
K-Means algorithm	Fast iterative to find outliers	Sensitive to outlier and noise
Partitioning Around Medoid (PAM)	It is robust in presence of noise and outlier, Shows maximum outliers, less false alarm rate	More iteration for finding the outliers
Clustering LARge Applications (CLARA)	Find multiple outliers , more accuracy, less false alarm rate, higher detection rate	Quality based on the cost.
CLARAN	Search the entire outlier by splitting the data, find multiple records, less false rate.	Expensive

Table 1: Compare analyses of Clustering algorithm.

IV. CONCLUSION

In data mining, outlier detection is the important task to remove the unwanted data in the dataset. Outlier detection is a task that finds the data which is dissimilar or inconsistent with respect to the remaining data. Detecting the outlier DM use many techniques like classifier outlier, cluster outlier, etc, While clustering mechanism works effectively in the data mining approaches. From our study here use the clustering based outlier detection mechanism. Here the partitioning based clustering algorithm it analyzes the outlier data well. This paper shows the partition based clustering approach to analyzing the outliers. For that here used 4 algorithm K-Means, PAM, CLARA, and CLARANS. From our knowledge CLARANS works better when compared to another algorithm.

REFERENCES

[1] Arun K Pujari: Data Mining Techniques, Universities Press (India) Private Limited 2001.
[2] S.Vijayarani S.Nithya, "An Efficient Clustering Algorithm for Outlier Detection"
[3] Jiang, S., And An, Q. (2008) Clustering-Based Outlier Detection Method, Fifth International Conference on Fuzzy Systems and Knowledge Discovery.
[4] Ramaswami, S., R. Rastogi and K. Shim,Efficient Algorithm for Mining Outliers from Large Data Sets. Proc. ACM SIGMOD, 2000, pp. 427-438.
[5] Ng, R. and Han, J. (1994) Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Conf. Very Large Databases, Pp. 144–155..
[6] Dr. S. Vijayarani, Ms. P. Jothi, "Hierarchical and Partitioning Clustering Algorithms for Detecting Outliers in Data Streams"
[7] Ms.P.Jothi and Dr. S. Vijayarani **"**Partitioning Clustering Algorithms for Data Stream Outlier Detection"
[8] Sivaram.K, Saveetha.D, "AN EFFECTIVE ALGORITHM FOR OUTLIER DETECTION".
[9] Janpreet Singh and Shruti Aggarwal, "Survey on Outlier Detection in Data Mining"
[10] Shruti Aggrwal, Prabhdip Kaur, "Survey of Partition Based Clustering Algorithm Used for Outlier Detection"
[11] H.S Behera, Rosly Boy,Lingdoh, Diptendra Kodama Singh "An Improved hybridized k-means clustering algorithm for high dimensional data set & it's performance analysis" International Journal of Computer Science and Engineering (IJCSE).
[12] K.Yoon, O.Kwon and D.Bae, "An approach to outlier Detection of Software Measurement Data using the Kmeans Clustering Method", First International Symposium on Empirical Software Engineering and Measurement, Madrid.,pp:443-445, 2007
[13] Pam Garima Singh, Vijay Kumar, "An Efficient Clustering and Distance-Based Approach for Outlier Detection".
[14] Elahi, M. K., Li ;Nisar, Wasif2 ; Xinjie, Lv1 ; Hongan, Wang (2009). Detection of local outlier over dynamic data streams using efficient partitioning method. 2009
[15] Mahfouz, M.A., and Ismail, M.A. (2009) Fuzzy relatives of the CLARANS algorithm with application to text clustering, World Academy of Science, Engineering and Technology, Vol. 49, Pp. 334-341.
[16] Ng, R. and Han, J. (2002) CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE Transactions on Knowledge and Data Engineering. Vol.14, No.5.

