A Study of Clustering Based Algorithm for Outlier Detection in Data streams

Dr. T. Christopher
Assistant Professor, PG& Research Department of Computer Science, Government Arts College, Coimbatore.
chris.hodcs@gmail.com

T. Divya
M.Phil Scholar,PG& Research Department of Computer Science,Government Arts College, Udumalpet.
divyathirumurthi@gmail.com

---ABSTRACT---------------------------------------

Recently many researchers have focused on mining data streams and they proposed many techniquesand algorithms for data streams. It refers to the process of extracting knowledge from nonstop fast growing data records. They are data stream classification, data stream clustering, and data stream frequentpattern items and so on. Data stream clustering techniques are highly helpful to cluster the similar data items in datastreams and also to detect the outliers, so they are called cluster based outlier detection. Outlier Detection is a fundamental issue in Data Mining. It has been used to detect and remove unwanted data objects from large dataset. The clustering techniques are highly helpful to detect the outliers called cluster based outlier detection.The data stream is a new emerging research area in Data Mining. It refers to the process of extracting knowledge from nonstop fast growing data records.

Keywords:Data stream, Data stream Clustering, Outlier detection.

1.INTRODUCTION

Data mining is broadly studied field of research area, where most of the work is highlighted over knowledge discovery, in that data stream is one of the research areas in data mining because data stream data are massive, fast changing, unlimited, continuous flow and infinite. Applications of data streams can vary from scientific and astronomical applications to important business and financial ones therefore, real-time analysis and mining of data streams have attracted substantial amount of researches [5]. One of the major problems in data mining research is increase in dimensionality of data gives rise to a number of new computational challenges. In recent years, it is observed that enormous research activity actuated by the explosion of data collected and transferred in the format of data streams. A data stream is a continuous, real time, river flow sequence of data items and it is not possible to control the order in which data item arrive or it is not feasible to locally store a stream in its entirety. The applications of data streams are generated like transactions, ATM data, credit card operations and popular web sites logs has been led to motivate by the study of data stream. Various algorithms for mining a stream data do not fit in primary memory due to lack of resources where as this type of large data, the current data mining systems are not sufficient and equipped to deal with them. In order to use clustering in data streams, the requirements are to be generated for overall high-quality clusters without seeing the old data, high quality, efficient incremental clustering algorithms and analysis in multi-dimensional space. There are several types of clustering techniques are useful for outlier detection. The hierarchical algorithms create a hierarchical decomposition of the objects and they are either agglomerative bottom-up or divisive top-down. Agglomerative algorithms start with each object, and successively merge groups according to a distance measure, where as the clustering may stop when all objects are in a single group or at any other point the user wants and these methods called as greedy bottom-up merging. Divisive algorithms follow the reverse approach, it starts with single group of all objects and successively split groups into smaller ones, until all object falls into single cluster, are to be preferred. Partitioning algorithm constructs various partitions for the data elements and then evaluates them by some criteria Data stream clustering methodologies are highly helpful to detect outliers and outlier detection is one of the data mining tasks and it is otherwise called as outlier mining.

An outlier detection, streaming data is one of the active research area from data mining that aims to detect object which have different actions exceptional than normal object. An outlier is an object that is significantly dissimilar or inconsistent to other data object where as click stream, fraud detection, web logs, and web documents are the application of outlier detection in data streams area. There are many algorithms for outlierdetection in static and stored data sets which are based on a variety of approaches like nearest neighbour based,density based outlier detection, distance based outlier detection and clustering based outlier detection.

2.LITERATURE REVIEW

S. Vijayarani and P. Jothi [1] discussed about two clustering algorithms namely BIRCH with K-means and CURE with K-means which are used for clustering the data items and finding the outliers in data streams. To analyze the experimental result, two performance factors are used such as Clustering Accuracy, Outlier Detection Accuracy.

In this paper the proposed CURE with K-means algorithm has given good performance results when compared with the algorithm BIRCH with K-means clustering algorithm.

S. Vijayarani and P. Jothi [2] discussed about data stream clustering algorithms which are highly used for detecting the outlier efficiently. This paper has focused on clustering process and detecting outliers in data streams.
In this paper two clustering algorithm namely CLARANS and E-CLARANS are used for finding the outliers in data streams. Finally by analysing the result it is found that
E-CLARANS algorithm performance is more accurate than the existing algorithm CLARANS.

S. Vijayarani and P. Jothi [3] discussed about data stream clustering algorithms which are highly used for detecting the outlier efficiently. This paper has focused on two clustering algorithm namely CURE with K-means and CURE with CLARANS are used for finding the outliers in data streams. Different sizes and types of data sets and two performance factors such as clustering accuracy and outlier detection accuracy are used for analysis.
In this paper proposed CURE with CLARANS clustering algorithm performance is more accurate than the existing algorithm CURE with K-means.

S. Vijayarani and P. Jothi [4]discussed about outlier detection in data stream clustering algorithm is highly needed and the clustering process in data streams and detecting the outliers in data streams. Different sizes and types of data sets and two performance factors such as clustering accuracy and outlier detection accuracy are used for analysis
In this paper proposed CURE with CLARANS clustering algorithm performance is more accurate than the existing algorithm BIRCH with CLARANS.

S. Vijayarani and P. Jothi [5] discussed about data stream clustering techniques are highly helpful to cluster the similar data items in data streams and also detect the outliers, so they are called cluster based outlier detection in data streams. Different sizes and types of data sets and two performance factors such as clustering accuracy and outlier detection accuracy are used for analysis
In this paper proposed BIRCH with CLARANS clustering algorithm performance is more accurate than the existing algorithm BIRCH with K-means.

D. Joice and K. Lakshmi and K. Thilagam [6] discussed about data stream is a new emerging research area in data mining.
In this paper is to perform the clustering process in data streams and to detect the outliers in high dimensional data using the existing clustering algorithms like K-means, CLARA, CLARANS and CURE. The experimental result of this paper shows that CURE clustering algorithm yields best performance compared to other algorithms.

3. METHODOLOGY

Clustering and Outlier detection is one of the important tasks in data streams. Outlier detection is based on clusteringapproach and it provides new positive results. The main objective of this research work is to perform the clusteringprocess in data streams and detecting the outliers in data streams. In this research work, two clustering algorithms namely CURE and K-Means are used for clustering the data items and finding the outliersin data streams.

3.1Dataset

In order to compare the data stream clustering for detecting outliers, data sets were taken from UCI machinelearning repository. Datasets namely Breast Cancer Wisconsin Dataset with 699 instances, 10 attributes and PimaIndian data set contain 768 instances and 8 attributes. These two biological data sets have numeric attributes whichhave been used in this research work. Data stream is an unbounded sequence of data as it is not possible to storecomplete data stream, for this purpose we divide the data into chunks of same size and each chunk size is specified bythe user which depends upon the nature of data and finally we divided the data into chunks of same size in differentwindows.

3.2 Clustering

The clustering algorithm is used to group objects into significant subclasses and the clustering data streams are a subarea of mining data streams. The clustering algorithms for data streams should be adaptive in the sense that up to dateclusters are obtainable at any time, taking new data items into account as soon as they arrive. There are different types
of clustering algorithms are fitting for different types of applications they are chased by Hierarchical clusteringalgorithm, Partition clustering algorithm, Density based clustering algorithm and Grid based clustering algorithm. Clustering is defined as an unsupervised problem. There are no predefined class label exists for the data points. Clusteranalysis is used in a number of applications such as data analysis, image processing, Stock market analysis etc.

3.3 Outlier Detection

Outlier detection has a wide range of applications such asfraud detection, intrusion detection, and credit cardanalysis. It is further complicated by the fact that in manycases outliers have to be detected from a large volume ofdata growing at an unlimited rate. Traditional outlierdetection algorithms cannot be functional to data streamefficiently, since the data stream is potentially infinite andevolving continuously. It has to be routed within an exacttime constraint and limited space, thus outlier detection indata stream imposes great challenges are followed. Thecluster based outlier detection is a best technique tosupervise this problem.

3.4 Related Algorithms

3.4.1 K-Means

The K-means algorithm is the best known partitioned clustering algorithm. It is a simple method for estimating the mean (vector) of set K groups. The most widely used K-means among all clustering algorithms is due to its efficiency and simplicity. The K-means algorithm is as follows

Algorithm k-means (k, D)

```
1 chooses k data points as the initial cancroids (cluster centers)
2 repeat
3 for each data point x ∈ D do
4 compute the distance from x to each centered;
5 assign x to the closest centered
// a centered represents a cluster
6 end for
7 re-compute the centered using the current cluster memberships
8 until the stopping criterion is met
```

3.4.2 CURE

CURE stands for Clustering using Representatives Algorithm. CURE is an efficient data clustering algorithm for large databases. It is processed using hierarchical methods to decompose a dataset into tree like structures. It uses two clustering approaches namely Partitioning clustering algorithm and Hierarchical clustering algorithm.

1. When applied to Partitioning clustering algorithm, the sum of squared errors is appeared in large differences in sizes or geometrics of different clusters.
2. When applied to Hierarchical clustering algorithm, it measures the distance between(dmin, dmean) work with different shapes of clusters. But the running time is high when n is very large.

So, to avoid this problem of non uniform sized(or) shaped clusters of CURE hierarchical algorithm, the centroid points of clustering are merged at each step. This enables CURE to correctly identify the clusters and makes its sensitive to outliers. The running time of the algorithm us O(n2 log n) and space complexcity is O(n). The CURE Algorithm is us follows,

Algorithm CURE

CURE (no. of points, k)
Input: A set of points S
Output: k clusters
1. For every cluster u (each input point), in u. mean and u.rep store the mean of the
points in the cluster and a set of c representative points of the cluster initially c = 1
since each cluster has one data point. Also u. closest stores the cluster closest to u.
2. All the input points are inserted into a k-d tree T.
3. Treat each input point as separate cluster, compute u. closest for each u and then
insert each cluster into the heap Q.
4. While size (Q) > k.
5. Remove the top element of Q (say u) and merge it with its closest cluster u. closest (say v) and compute the new representative points for the merged cluster w. Also
remove u and v from T and Q.
6. Also for all the clusters x in Q, update x. closest and relocate x.
7. Insert w into Q.8. Repeat.

3.4.3 CLARANS Clustering

CLARANS [3] clustering algorithm is nothing but it is used for randomized search and CLARANS is abbreviated as Clustering Large Application Based upon Randomized Search. Ng and Han proposed a new algorithm in 1994 called CLARANS. It uses random search to generate neighbours by starting with arbitrary node and randomly check maxneighbours,where ever if the neighbour represent better partition the process continue with new node otherwise local minimum is found and algorithm restart until num local minima is found (value of num local is=2 recommended)thebest node return resulting partition. CLARANS take a random dynamic selection of data at each step of process. Thusthe same sample set is not used throughout in the clustering process. As a result better randomization source isachieved. CLARANS is accurately detecting outlier than CLARA and it is much less affected by increasingdimensionally and draw the sample of neighbours in each step of search this is benefit of confining the search localizearea. CLARANS algorithm followed as

1. Randomly choose k mediod
2. Randomly consider the one of mediod swapped with non mediod
3. If the cost of new configuration is lower repeat step 2 with new
solution
4. If the cost higher repeat step 2 with different non mediod object
unless limit has been reached
5. Compare the solution keeps the best
6. Return step 1 unless limit has been reached (set to the value of 2).

3.5 Outlier DetectionAccuracy

3.5.1 Accuracy

Outlier detection accuracy is calculated in order to find out the number of outliers detected by the clustering algorithms.

3.5.2 Detection Rate

Detection rate refers to the ratio between the number of correctly detected outliers to the total number of outliers, the detection rate is calculated using the formula,

$$d' = \frac{\mu_{S-\mu_N}}{\sqrt{1/2((\sigma_S^2 + \sigma_N^2))}}$$

The above formula provides the separation between the means of the signal and the noise distributions compared against the standard deviation of the noise distribution. The distributed signal and noise with mean and the standard deviation are represented as μ_S and σ_S, and μ_N andσ_N.

3.5.3 False alarm Rate

False alarm rate refers to the ratio between the numbers of normal objects that are misinterpreted as outlier to the total number of alarms. The other name for it is False Detection Rate. In order to calculate the false alarm rate the formula is,

FDR = FP/(TP + FP)= 1-PPV

The above formula uses False Positive (FP), True Positive (TP) and Positive Predictive (PPV) values to find the false alarm rate.

3.6 Clustering Accuracy

Clustering accuracy is calculated using three measures i.e., Accuracy, Precision and Recall.

3.6.1Accuracy

The accuracy determines how close the measurement comes to the true value of the quantity. So, it indicates the correctness of the result. The accuracy is calculated by using the above formula,

ACC=(TP+TN)/(P+N)

Where True Positive(TP), True Negative(TN),Positive(P) and Negative(N) values are used to calculate the clustering accuracy.

3.6.2 Precision

PPV=TP/(TP+FP)

The relative precision indicates the uncertainty in the measurement as a fraction of the result. The precision is calculated by using the formula,

Where True Positive(TP) and False Positive(FP) values are used to find out the clustering accuracy of precision values.

3.6.3 Recall

The recall relates to the test‘s ability to identify a condition correctly. The recall tests have few type II errors. The recall is calculated using the formula,

TPR=TP/P=TP/(TP+FN)

Where True Positive(TP), False Negative(FN) and Positive(P) values are used to found the recall values.

4. CONCLUSION

Data streams are dynamic ordered, fast changing, massive, limitless and infinite sequence of data objects. Data streams clustering technique are highly helpful to handle those data. The outlier detection is one of the challenging areas in data stream. By using data stream hierarchical clustering and partition clustering are helpful to detect the outlier efficiently.

In this paper we have analysed the performance of CURE with K-Means and CURE with CLARANS clustering algorithm for detecting the outliers. In order to find the best clustering algorithm for outlier detection several performance measures are used. From the experimental results it is observed that the outlier detection accuracy is moreefficient in CURE with CLARANS clustering while compare to CURE with K-Means with clustering.

References

[1] Dr. S. Vijayarani and Ms. P. Jothi, Hierarchical and Partitioning ClusteringAlgorithms for Detecting Outliers in Data Streams, *International Journal of Advanced Research in Computer and Communication Engineering*, Volume. 3,Issue 4, April 2014.

[2] Dr. S. Vijayarani and Ms. P. Jothi, Partitioning Clustering Algorithms For Data Stream Outlier Detection, *International Journal of Innovative Research in Computer and Communication Engineering*, Volume. 2, Issue 4, April 2014.

[3] Dr. S. Vijayarani and Ms. P. Jothi, Detecting Outliers in Data streams using Clustering Algorithms, *International Journal of Innovative Research in Computer and Communication Engineering*, Volume. 2, Issue 8, October 2013.

[4] Dr. S. Vijayarani and Ms. P. Jothi,Comparative Analysis of Clustering Algorithms for Outlier Detection in Data Streams, *International Jounal of engineering sciences& Research Technology*, October 2013.

[5] Dr. S. Vijayarani and Ms. P. Jothi, An Efficient Clustering Algorithm For Outlier Detection In Data Streams, *International Journal of Advanced Research in Computer and Communication Engineering*, Volume. 2, Issue 9, September 2013.

[6] D.Joice, K. Lakshmi and K. Thilagam, Comparison Of Cluster Based Algorithms For Outlier Detection In High Dimensional Dataset, *Karpagam Journal of computer science*, Volume 8, issue 3, April 2014.