COMPARISON OF CLUSTER BASED ALGORITHMS FOR OUTLIER DETECTION IN HIGH DIMENSIONAL DATASET

Joice D, Lakshmi K[1], Thilagam K[2]

ABSTRACT

Outlier Detection is a fundamental issue in Data Mining. It has been used to detect and remove unwanted data objects from large dataset. Clustering is the process of grouping a set of data objects into classes of similar data objects. 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.

The main objective of 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 shows that CURE clustering algorithm yields best performance compared to other algorithms.

Keywords : *Outlier Detection, Data Stream, Data Stream Clustering.*

I. INTRODUCTION

Data Mining is an extensive studied field of research area. Data Mining is mining of knowledge from large amount of data. There are lot of problem exists in large databases such as data redundancy, missing data, invalid data etc., one of the major problem in data stream research area is in handling high dimensional datasets. Outlier Detection is a branch of Data Mining, which refers to the problem of finding objects in large dataset that vary from other data objects.

Outlier Detection is a fundamental issue in Data Mining research area. It has been used to detect and remove unwanted anomalie objects from large dataset. Outlier Detection is the necessary step in variety of practical applications such as intrusion detection, health system monitoring and criminal activity detection in e-commerce and can also be used in scientific research for data analysis and knowledge discovery in the fields of the chemistry, biology, astronomy etc.,

II. LITERATURE REVIEW

S. Vijayarani and P. Jothi [7] discussed about two clustering algorithms namely BIRCH with K-Means and BIRCH with CLARANS 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 and Accuracy.

[1]Research Scholar and Assistant Professor, Karpagam University, E-mail : joyjanifa@gmail.com,

[2]Research Scholar and Assistant Professor, Karpagam University,lakhsk2012@gmail.com,
E-mail : thilagam@gmail.com

In this paper the proposed BIRCH with CLARANS clustering algorithm has given good performance results when compared with the other algorithm BIRCH with K-Means clustering algorithm.

T. Soni Madhulatha [6] discussed about the study of a particular data mining task and clustering which can be used as the first step in many knowledge discovery processes. She explained that the data streaming is processed using few clustering algorithms namely K-Means, CLARANS, BIRCH, DBSCAN, LSEARCH, CURE and STING.

These clustering techniques are applied to time series data because it has the characteristics of processing very high dimensional dataset. This paper provides a broad survey of the most basic techniques, and an overview of elementary clustering techniques which is most commonly used. μ_n *and* μ_p

S. Guha et.al [3] proposed a new clustering algorithm called CURE. This algorithm is more robust to detect the outlier's and to identify clusters having non-spherical shapes and wide variances in size to handle large database. CURE employs a combination of random sampling and partitioning methods for analyzing processes.

The performance of CURE has been effectively explained and compared with other algorithm like BIRCH and MST (Minimum Spanning Tree). The Results show that the quality of clusters produced by CURE is much better than those found by existing algorithms.

S. Vijayarani and P. Jothi [9] discussed about the data stream clustering algorithms which are highly used for detecting the outliers efficiently. This paper has focused on clustering process and detecting outliers in data streams.

In this paper two clustering algorithms namely CURE with K-Means and CURE with CLARANS are used for finding the outliers in data streams. Finally by analyzing the result it is found that CURE with CLARANS clustering algorithm performance is more accurate than the algorithm CURE with K-Means.

Shruthi Aggarwal and Prabhdip [8] described about the large number of partition based clustering algorithms namely K-Means, CLARA, CLARANS, ECLARANS and CLATIN which are used for outlier detection processes.

This paper has described about the comparative study of different clustering algorithms. The main objective of this paper is to improve the outlier detection in terms of time complexity and efficiency.

III. Methodology

Clustering techniques are applied for grouping of data items in high dimensional data streams and detecting outliers. Clustering and Outlier detection are one of the important issues in data streams. Outlier detection based on clustering approach provides new positive technical results. The main objective of this research work is to perform the clustering process in data streams and detecting the outliers in high dimensional data. In this research work the clustering algorithms namely K-Means, CURE, CLARA, CLARANS are used for clustering the data items and finding outliers in high dimensional data streams.

A. DATA SET

To compare the performance among different existing clustering algorithms, dataset is taken from UCI Machine Learning Repository. The dataset related to Echocardiogram has been used to compare the performance. The system architecture diagram of the proposed method is shown in Figure 1.

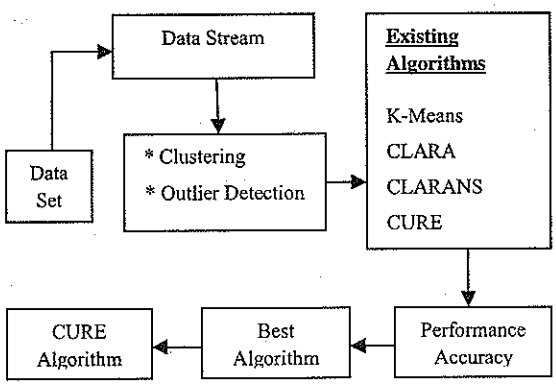

Figure 1: The System Architecture of Clustering Algorithms for Outlier Detection

The Echocardiogram biological dataset has numerous important attributes which have been used in this research work. Data stream is an unbounded large sequence of data. But it is impossible to store complete data stream. So for this purpose dataset is divided into chunks of same size and categorized into different windows (W1, W2 and W3).

B. CLUSTERING

The clustering algorithms are used to group objects into significant subclasses. There are different types of clustering algorithms used for different types of applications namely Hierarchical clustering algorithm, Partition clustering algorithm, Density based clustering algorithm and Grid based clustering algorithm. But clustering is defined as an unsupervised outlier detection problem. Clustering algorithm is used in number of applications such as, Data analysis, Image processing and Stock market analysis etc. This research work includes some best clustering algorithms like K-Means, CLARA, CLARANS and CURE algorithm to find an anomaly detection process. But the CURE algorithm yields good performance in detecting outliers from high dimensional dataset.

IV. RELATED ALGORITHMS

A. 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 centroids (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 using till the stopping criterion is met

B. CLARA

CLARA stands for **C**lustering a **L**arge **A**pplications Algorithms. The focus is on clustering large number of objects rather than small number of objects in high dimensions. It works by clustering a sample from the dataset and then assigns all objects in the dataset. This algorithm relies on the sampling approach to handle large datasets. CLARA draws a small sample from the dataset

1. Draw a sample from the n objects and cluster it into k groups.
2. Assign each object in the dataset to the nearest group.
3. Store the average distance between the objects and their respective groups.
4. Repeat the process five times, selecting the clustering with the smallest average distance.
5. While assign a large number of objects to group.

and applies the PAM Algorithm. The CLARA Algorithm is as follows,

C. CLARANS

CLARANS stands for **C**lustering a **L**arge **Ran**dom **S**ubset Searching Algorithm. CLARANS proceeds by searching a random subset of neighbors for a particular solution. This algorithm used two parameters for calculating solutions namely MAXneigh, the maximum number of neighbors of S to access and MAXsol, the maximum number of local solutions. The CLARANS Algorithm is as follows,

1. Set S to be an arbitrary set of k representative objects.
   Set i =1
2. Set j = 1.
3. Consider a neighbor R of S at random. Calculate the total swap contribution of the two neighbors.
4. If R has a lower cost, set R = S and go to Step 2. Otherwise increment j by one. If $j \leq$ MAXneigh goto Step 1.
5. When j > MAXneigh, compare the cost of S with the Best solution found so far.
   If the cost of S is less, record this cost and the representation. Increment i by one.
6. If I > MAXsol stop, otherwise goto Step 1.

D. CURE

CURE stands for **C**lustering **u**sing **Re**presentatives 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 is **O (n2 log n)** and space complexity is **O (n)**. The CURE Algorithm is us follows,

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.

But this CURE algorithm cannot be directly applied to large databases. So before applying the following enhancements has to be made,

1. Random Sampling
2. Partitioning for speedup
3. Labeling data on disk

To handle large datasets, do random sampling and draw a sample dataset. The random sample datasets are stored in main memory. The basic idea is to partition the sample space into P partitions. The remaining data points should also be assigned to the labeling data on disk. The main advantage of partitioning the input is to reduce the execution time.

V. Experiment Results

All the algorithms are implemented using MATLAB (R2010a). Two performance factors are considered namely Outlier Detection Accuracy and Clustering Accuracy for result analysis. For the experimental purpose a very high dimensional Echocardiogram biological dataset is chosen.

A.OUTLIER DETECTION

ACCURACY

Outlier detection accuracy is calculated in order to find out the number of outliers detected by the clustering algorithms namely K-Means, CLARA, CLARANS and CURE from Echocardiogram dataset into three windows. This is calculated using Detection Rate and False Alarm Rate.

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,

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

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 deviations are represented as μ_S and σ_S, and μ_N and σ_N.

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 ucalculate 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.

Table 1 and Figure 1 shows the outlier accuracy of Detection Rate and False Alarm Rate in three windows of Echocardiogram dataset. Dataset is compared against the existing clustering algorithms namely K-Means, CLARA, CLARANS and CURE.

Table 1: Outlier Accuracy in Three Windows of Echocardiogram Dataset

Outlier Accuracy	No. of Windows	K-Means	CLARA	CLARANS	CURE
Detection Rate	W1	54.26	57.76	64.26	67.76
	W2	64.07	67.41	74.07	77.41
	W3	77.00	78.65	87.00	88.65
False Alarm Rate	W1	69.42	59.70	53.40	43.74
	W2	70.71	61.70	60.75	51.78
	W3	82.46	80.90	72.46	70.90

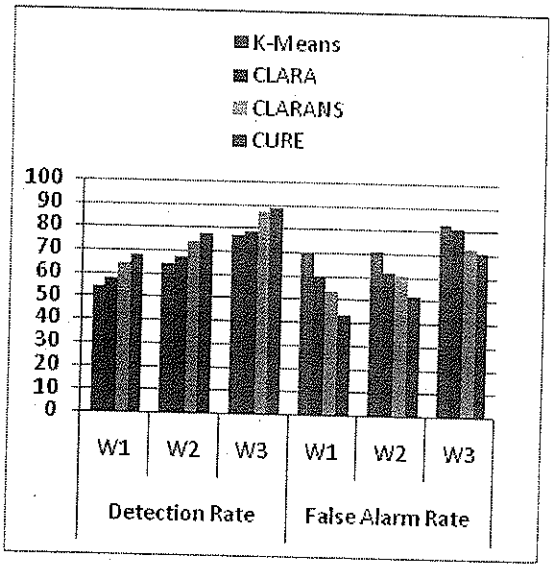

Figure 1: Outlier Accuracy in Three Windows of Echocardiogram Dataset

From the above graph, it is observed that CURE clustering algorithm performs better than other clustering algorithms such as K-Means, CLARA and CLARANS for detecting outliers in biological dataset of Echocardiogram in three windows. The CURE clustering algorithm performs well since it contains high Outlier Detection accuracy and low False Alarm rate when compared to other clustering algorithms such as K-Means, CLARA and CLARANS.

B. CLUSTERING ACCURACY

Clustering accuracy is calculated using three measures i.e., Accuracy, Precision and Recall. The clustering algorithms namely K-Means, CLARA, CLARANS and CURE are applied for echocardiogram dataset to find the clustering accuracy in three windows.

ACCURACY

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.

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.

RECALL

The recall relates to the test's ability to identify a condition correctly. The recall test 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.

The above Table 2 and Figure 2 shows the clustering accuracy in terms of Precision and Recall in three windows of Echocardiogram dataset.

Table 2 : Clustering Accuracy in Three Windows of Echocardiogram Dataset

Clustering Accuracy	No.of Windows	K-Means	CLARA	CLARANS	CURE
Accuracy	W1	82.4	84.51	86.3	88.41
	W2	82.05	84.16	86.92	88.03
	W3	82.4	84.51	86.3	88.41
Precision	W1	82.14	84.25	86.17	88.28
	W2	80.53	82.64	84.79	86.8
	W3	76.71	78.82	81.99	83
Recall	W1	82.32	84.43	86.19	88.2
	W2	81.11	83.22	86.73	88.84
	W3	84.02	86.13	85.1	87.29

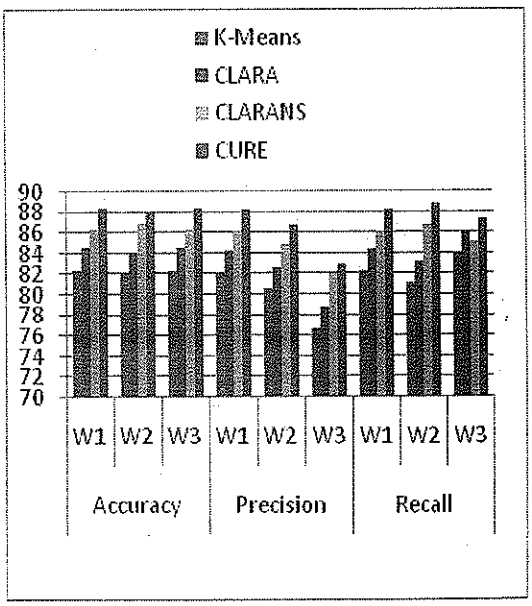

Figure 2 : Clustering Accuracy in Three Windows of Echocardiogram Dataset

From the above graph, it is observed that CURE clustering algorithm performs better than other clustering algorithms namely K-Means, CLARA and CLARANS in detecting outliers from the biological dataset of Echocardiogram in three windows. The CURE clustering algorithm performs well since it contains high Clustering Accuracy when compared to other clustering algorithms of K-Means, CLARA and CLARANS.

VI. Conclusion

The Outlier Detection is one of the challenging area in Data Mining using different data streams. By using large datasets, hierarchical clustering and partitioning clustering are helpful to detect the anomalies very efficiently. In this paper, the clustering and outlier performance are analyzed in K-Means, CLARA, CLARANS and CURE clustering algorithm for detecting outliers. To find out the best clustering algorithm for detecting outliers some important performance measures are used. From the Experimental results it is observed that the Clustering and Outlier Detection Accuracy is more efficient in CURE clustering when compare to other clustering algorithms such as K-Means, CLARA and CLARANS.

It is concluded that in handling high dimensional dataset it is necessary to choose a proper algorithm according to the size of the dataset. If the dataset is low dimensional, use K-Means algorithm. If the dataset is middle dimensional use CLARA or CLARANS algorithm. If the data set is very high dimensional better go for CURE algorithm. Once the correct algorithm is chosen properly then clustering process and outlier detection will become easier. In future combinations of existing Algorithms will be experimented in order to produce even more accuracy in results.

References

1. A Comprehensive *Overview of Basic Clustering Algorithms* Glenn Fung June 22, 2001.
2. Streaming-Data Algorithms For *High-Quality Clustering Liadan O'Callaghan,* NinaMishra, Adam Meyerson, Sudipto Guha, Rajeev Motwani.
3. S. Guha, R. Rastogi, and K. Shim. CURE: An *Efficient clustering algorithm for large databases. In Proc.* SIG-MOD, pages 73-84, 1998.
4. R.T. Ng and J. Han. Efficient and effective clustering methods for spatial Data Mining In Proc. VLDB, pages 144-155, 1994.
5. C. Aggarwal, Ed., Data Streams – Models and Algorithms, Springer, 2007.
6. T. Soni Madhulatha, "overview of streaming-data algorithms, Department of Informatics, Allure Institute of Management Sciences, Warangal, A.P. Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011.
7. 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.1, Issue-8, October 2008.
8. Shruti Aggrwal and Prabhdip Kaur, *"Survey of Partition Based Clustering Algorithm Used for Outlier Detection"*, International Journal for Advance Research in Engineering and Technology, volume 1, Issue V, June 2013.

9. 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 Vol. 2, Issue 9, September 2013.

10. D.Napoleon and S.Pavalakodi, *"A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set"*, International Journal of Computer Applications (0975 – 8887) Volume 13– No.7, January 2011.

11. Hossein Moradi Koupaie, Suhaimi Ibrahim, Javad Hosseinkhan , *"Outlier Detection in Stream Data by Clustering Method"* International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 3, 2013, Page: 25-34, ISSN: 2296-1739.

12. Irad Ben-Gal, *"outlier detection", Department of Industrial Engineering Tel*-Aviv University Ramat-Aviv, Tel-Aviv 69978, Israel.

Author's Biography

Joice. D, has completed MCA in Vivekananda College of Information and Management Studies, Thiruchengode. She is currently pursuing her M.Phil in Computer Science at Karpagam University, Coimbatore. Her field of interest is Data Mining.

Lakshmi. K, MCA M.Phil, working as an Assistant Professor in the Department of Computer Applications at Karpagam University, Coimbatore. She is having 10 years of teaching experience. She has published good number of papers in National and International Journals. Her research areas of interest are Data Mining and Computer Networks especially in Adhoc, MANET and VANET.

Dr. K. Thilagam, working as an Assistant Professor in the Department of Computer Applications at Karpagam University, Coimbatore. She is having 15 years of teaching experience. She has published 11 papers in National and International Journals. Her area of interest are Image Processing, Computer Network and Data Mining.