ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

Detecting Outliers in Data streams using Clustering Algorithms

Dr. S. Vijayarani[1] Ms. P. Jothi[2]

Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu , India[1]

M.Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India[2]

ABSTRACT: The data stream is a new arrival of research area in data mining where as data stream refers to the process of extracting knowledge structures from nonstop, fast growing data records. Emerging applications involved in data streams are motivated by many researches involving continuous massive data sets such as customer click streams, e-commerce, wireless sensor network, network monitor, telecommunication system, stock market and meteorological data. For handling this type of large data, the current data mining systems are not sufficient and equipped to deal with them, for this cause it leads to a numerous computational and mining challenges due to shortage of hardware limitations. Nowadays many researchers have focused on mining data streams and they proposed many techniques for data stream classification, data stream clustering and finding frequent items from data streams. Data stream Clustering and outlier detection provides a number of unique challenges in evolving data stream environment. Data stream clustering algorithms are highly used for detecting the outliers efficiently. 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 clustering algorithms 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. By analyzing the experimental results, it is observed that the proposed CURE with CLARANS clustering algorithm performance is more accurate than the existing algorithm CURE with K-Means.

***Keywords*:** Data stream, Data stream Clustering, Outlier detection, CURE, K-Means, CLARANS

I. 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. Data stream clustering is a sub-area of mining data streams, since the clustering algorithms arrange a dataset into several disjoint groups, such that points in the same group are related to each other and are unrelated to other groups, according to some similarity metrics. In order to use clustering in data streams, the requirements are [14] 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 [6] 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

Density based algorithms (DBSCAN) has been designed for the clustering of large noisy datasets on spatial data. DBSCAN introduced the concept of neighborhoods as a region of given radius in a sphere and including a minimum number of data points. Connected neighborhoods form clusters, thus leaving from the notion of spherical cluster. It produces clusters according to a density-based connectivity analysis. In order to perform clustering in data streams, we cannot save all the incoming data objects due to limited memory. The transactional clustering algorithm CLU-TRANS has been mainly used for two components they are online and offline transactions based on sliding window, in each sliding window had been assigned to the equal minimum clustering granularity to ensure that clustering time for the minimum granularity. 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. Outlier detection over streaming data is active research area from data mining that aims to detect object which have different behavior, exceptional than normal object. An outlier is an object [8] that is significantly dissimilar or inconsistent to other data object whereas telecommunication, fraud detection, web logs, click stream and web document are the application areas of outlier detection in data streams. There are many algorithms for outlier detection in static and stored data sets which are based on a variety of approaches like distance based outlier detection, density based outlier detection, nearest neighbor based outlier detection and clustering based outlier detection so on. The rest of this paper is followed as Section 2 illustrates the review of literature. Section 3 explains about the Cure with K-Means and CURE with CLARANS clustering algorithms used to detect outliers in data streams. Experimental results are discussed in Section 4 and Conclusions are given in Section 5.

II. LITERATURE REVIEW

Sudipto Guha, et.al [12] proposed a clustering algorithm called CURE and it is used for detecting outliers. CURE achieves by representing point per cluster its allow CURE to adjust well to the geometry of non-spherical shapes and the reduction helps to reduce the effects of outliers. The combination of random sampling and partitioning and the experimental results confirm that the quality of clusters produced by CURE is much better than those found by existing algorithms. Moreover, the authors expressed the partitioning and random sampling enable CURE to not only outperform existing algorithms but also to scale well for large databases without sacrificing the quality of cluster.

Elahi, M. Kun Li, et.al [7] discussed about a clustering based approach, it divide the stream into chunks and for cluster each chunk using k-mean in fixed number of cluster. In this research the author took the candidate outliers and mean value of every cluster for the next fixed number of data streaming chunks, to make sure that the detected candidate outliers are the real outliers. The mean value are used in the clusters of previous streaming chunk and the current chunk of mean values are taken to be the consideration, which is used to choose better outlierness for data stream objects. Quite a few experiments for different types of dataset confirm that the technique can find better outliers with low computational cost than the other existing distance based approaches of outlier detection in data stream.

Carlos Ordonez, et.al [5] talked about three variants of the K-means algorithm to cluster binary data streams. Variants are On-line K-means, Scalable Means, On-line K-means and Incremental K-means proposed a variant introduced that finds higher quality solution in less time. All variants were compared with real and synthetic data sets. The proposed Incremental K Means variant is faster than the already quite fast Scalable K-means and finds solution of comparable quality. The K-means variants are compared with respect to quality of speed and results. The proposed algorithms can be used to check the transactions. Further in this research the author discussed about the acceleration of Incremental K-means algorithm is not possible unless approximation, randomization or samplings are used.

Sharma.M, et.al [11] has conversed about the algorithm of k means for clustering of data streams and detection of outliers. The technique which has been used for outlier detection is based on distance as well as on time, on which they arrive in the cluster. The author takes into account the selection of k centers and variable size of buckets with the help of which space can be effectively utilized during clustering. Most traditional algorithms makes very difficult problem in clustering by reducing their quality for a better efficiency. In this research the author indicates a small increase of time, due to this cause the cluster can efficiently cluster the data without much loss of quality of data.

Thankran.Y, et.al [13] put forwarded an unsupervised outlier detection method for streaming data. This method is based on clustering as clustering is an unsupervised data mining task and it does not require labeled data. In this proposed method both density based and partitioning clustering method are combined to take advantage of both density and distance based outlier detection. It assigns weights to attributes depending upon their respective relevance in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

mining task and weights are adaptive in nature. Weighted attributes are helpful to decrease the effect of noisy attributes. The proposed method is incremental and adaptive to concept evolution.

III. METHODOLOGY

Clustering and Outlier detection is one of the important tasks in data streams. Outlier detection is based on clustering approach and it provides new positive results. The main objective of this research work is to perform the clustering process in data streams and detecting the outliers in data streams. In this research work, two clustering algorithms namely CURE with K-Means and CURE with CLARANS are used for clustering the data items and finding the outliers in data streams. The system architecture of the research work is as follows:

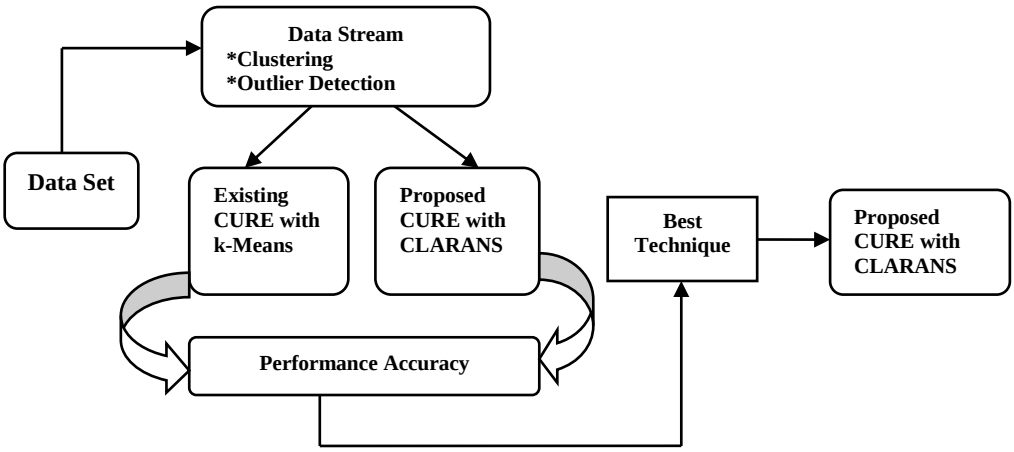

Figure 1: System architecture of clustering algorithms for outlier detection

A. DATASET

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

B. CLUSTERING

The clustering algorithm is used to group objects into significant subclasses and the clustering data streams are a sub area of mining data streams. The clustering algorithms for data streams should be adaptive in the sense that up to date clusters 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 clustering algorithm, 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. Cluster analysis is used in a number of applications such as data analysis, image processing, Stock market analysis etc.

C. OUTLIER DETECTION

Outlier detection has a wide range of applications such as insurance, fraud detection, intrusion detection, credit card and so on. It is further complicated by the fact that in many cases outliers have to be detected from a large volume of data growing at an unlimited rate [8].Traditional outlier detection algorithms cannot be applied to data stream efficiently since data stream is potentially infinite and evolving continuously. It has to be processed within a strict time constraint and limited space, thus outlier detection in data stream inflicts great challenges are followed as single scan. [9]. Outlier detection in data stream should be done very fast, preferably in single-scan and it is more complicated in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

data stream because it is difficult to build different data distribution basis in developing. Existing algorithms regard the model learned has the only distribution for detecting outliers and it also satisfy the evolving characteristics of data stream and may be too conservative to detect important outliers. So for that clustering based outlier detection is a best technique to manage this problem. For our research we have used cluster based outlier detection as CURE with K-means and CURE with CLARANS.

D. CURE CLUSTERING

CURE (Clustering Using Representatives) is an efficient data clustering algorithm for data streams that is more robust to outliers and identifies clusters having non-spherical shapes and wide variances in size. Cure is one of the hierarchical methods decompose a dataset into a tree-like structure. To avoid the problems with non-uniform sized or shaped clusters, it employs a hierarchical clustering algorithm that adopts a middle ground between the centroid based and all point extremes. In each iteration, it has a constant number c of well scattered points of a cluster are chosen and they are shrunk towards the centroid of the cluster by a fraction α. The scattered points after shrinking are used as representatives of the cluster. The clusters with the closest pair of representatives are the clusters that are merged at each step of CURE's hierarchical clustering algorithm. This allows CURE to correctly identify the clusters and makes it less sensitive to outliers [15]. CURE algorithm followed as

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.

The running time of the algorithm is O (n^2 log n) and space complexity is O (n). The algorithm cannot be straightly applied to large databases. So for this reason the random samplings are used to handle data sets and generally the random sample fits in main memory due to the random sampling there is a tradeoff between accuracy and competence. The basic idea is to partition the sample space into p partitions. In first pass the cluster are to be partition until the final number of clusters reduces to np /q for a few constant q ≥ 1. Run a second clustering pass on n/q partial clusters for all the partitions. For the second pass store the representative points for the merge. Merge requires the representative points of previous clusters before computing the new representative points for the merged cluster. Advantage of partitioning the input is used to reduce the execution times. The points for k clusters, the outstanding data points should be assigned to the clusters. Representative points for each of the k clusters are to be selected for fraction in order to choose the data point is assigned to the cluster containing the representative point closest to it.

E. K-MEANS CLUSTERING

K-Means clustering is a backbone method for detecting outliers. K-Means clustering requires iterative optimization of clustering centroids to gradually achieve better clustering results. This optimization process involves multiple data scans, which is infeasible in the context of data streams. K-mean describes that given dataset of n object divide into k cluster where k is desired number of cluster. A centroid is defined for each cluster in k-mean all data object are placed in cluster having centroid nearest to all data object. After processing all data object then k-mean centroid is calculated again and again. For each centroid it changes their location and its need to specify k number of cluster in advance. This process continues step by step until no centroid move. K-Means algorithm follows as

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

```
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
```

F. CLARANS CLUSTERING

CLARANS [15] 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 max-neighbours, where ever if the neighbor 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)the best node return resulting partition. CLARANS take a random dynamic selection of data at each step of process. Thus the same sample set is not used throughout in the clustering process. As a result better randomization source is achieved. CLARANS is accurately detecting outlier than CLARA and it is much less affected by increasing dimensionally and draw the sample of neighbours in each step of search this is benefit of confining the search localize area. 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).

IV. EXPERIMENTAL RESULTS

A. CLUSTERING ACCURACY

Clustering accuracy is calculated, by using two measures precision and recall. The clustering algorithms CURE with K-MEANS and CURE with CLARANS for Pima Indian diabetes and Wiscosin-breast cancer data set. Table I & Table II show the clustering accuracy, precision and recall in three windows and five windows.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

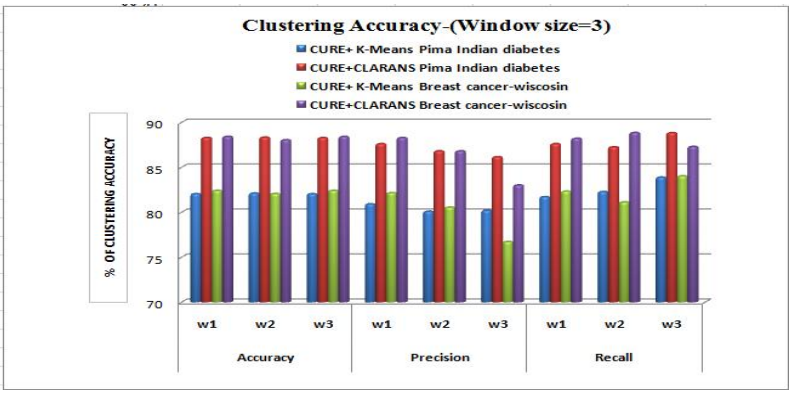

Figure 2: The clustering accuracy in three windows for two dataset

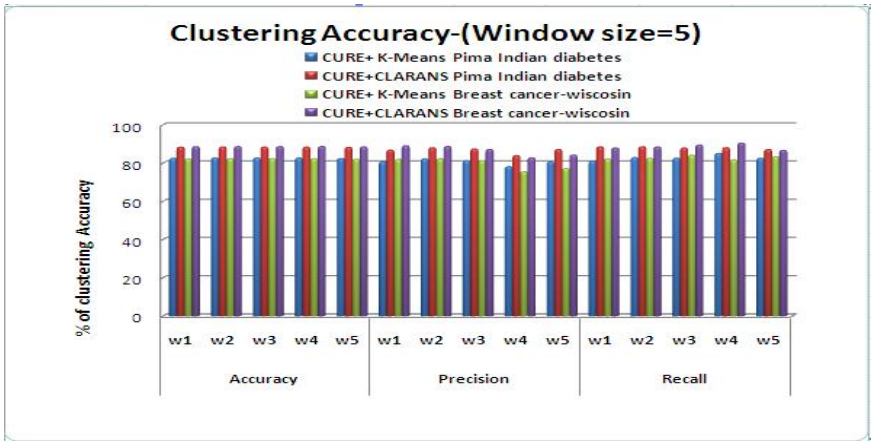

Figure 3: The clustering accuracy in five windows for two dataset

From the above graph, it is observed that CURE with CLARANS clustering algorithm performs better than CURE with K-Means clustering algorithms in Pima Indian Diabetes dataset and Breast cancer Wiscosin for both window size as five and three. Therefore the CURE with CLARANS clustering algorithm performs well because it contains high clustering accuracy when compared to CURE with K-Means.

B. OUTLIER ACCURACY

A. DETECTION RATE AND FALSE ALARM RATE FOR PIMA INDIAN DIABETES

Outlier detection accuracy is calculated, in order to find out number of outliers detected by the clustering algorithms CURE with K-MEANS and CURE with CLARANS for Pima Indian diabetes data set. Table III & Table IV shows the number of outlier detection rate and false alarm rate in three windows and five windows.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

TABLE III
DETECTION RATE AND FALSE ALARM RATE IN THREE WINDOWS-PIMA INDIAN DIABETES

Outlier Accuracy	No. of Windows	CURE +k Means	CURE +CLARANS
Detection rate	W1	36.72	39.48
	W2	36.12	37.54
	W3	33.80	35.00
False alarm rate	W1	38.10	34.42
	W2	30.00	26.08
	W3	27.91	26.68

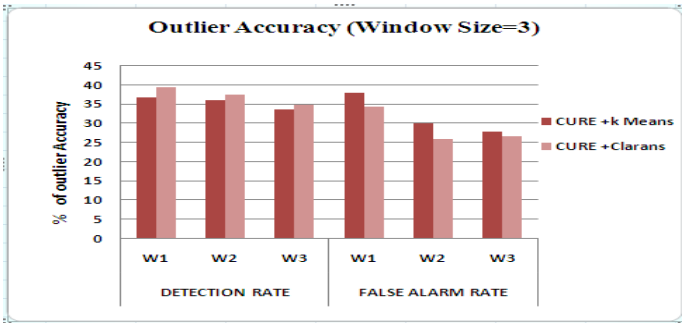

Figure 4: Detection rate and False alarm rate in three windows-Pima Indian diabetes

TABLE IV
DETECTION RATE AND FALSE ALARM RATE IN FIVE WINDOWS-PIMA INDIAN DIABETES

Outlier Accuracy	No. of Windows	CURE+K Means	CURE+ CLARANS
Detection rate	W1	33.89	37.79
	W2	44.03	45.22
	W3	37.61	39.84
	W4	25.54	27.96
	W5	35.50	36.44
False alarm rate	W1	38.88	32.22
	W2	38.00	37.03
	W3	34.00	22.22
	W4	23.00	19.11
	W5	35.55	30.00

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

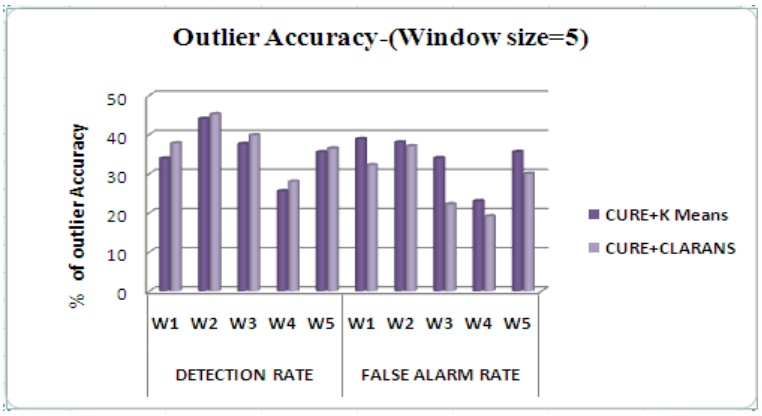

Figure- 5: Detection rate and False alarm rate in five windows-Pima Indian diabetes

From the above graph, it is observed that CURE with CLARANS clustering algorithm performs better than CURE with K-Means algorithms for detecting outliers in Pima Indian Diabetes dataset for both window size as five and three. Therefore the CURE with CLARANS clustering algorithm performs well because it contains high outlier detection accuracy when compared to CURE with K-Means.

B. DETECTION RATE AND FALSE ALARM RATE FOR BREAST CANCER (WISCOSIN)

Outlier detection accuracy is calculated, in order to find out number of outliers detected by the clustering algorithms CURE with K-Means and CURE with CLARANS for Breast cancer –Wiscosin data set. Table V & Table VI shows the number of outlier detection rate and false alarm rate in three windows and five windows.

TABLE V

DETECTION RATE AND FALSE ALARM RATE IN THREE WINDOWS -BREAST CANCER (WISCOSIN)

Outlier Accuracy	No. of Windows	CURE+ K-Means	CURE+CLARANS
Detection rate	W1	54.26	57.76
	W2	64.07	67.41
	W3	77.00	78.65
False alarm rate	W1	59.42	43.74
	W2	60.71	51.78
	W3	72.46	70.90

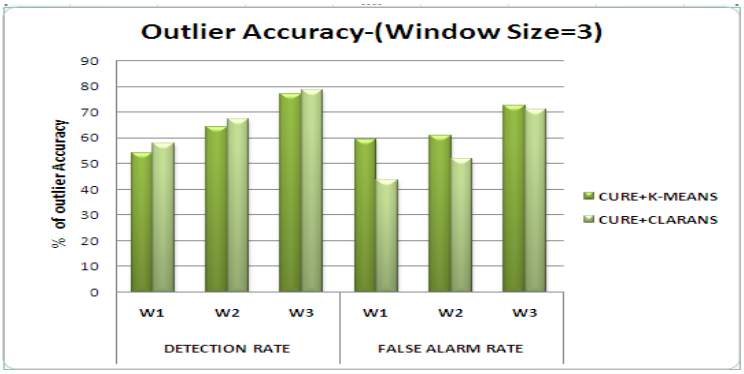

Figure-6: Detection rate and false alarm rate in three windows- Breast Cancer (Wiscosin)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

TABLE VI
DETECTION RATE AND FALSE ALARM RATE IN FIVE WINDOWS -BREAST CANCER (WISCOSIN)

Outlier Accuracy	No. of Windows	CURE +K-Means	CURE+CLARANS
Detection rate	W1	56.12	57.25
	W2	56.00	57.57
	W3	66.00	66.60
	W4	77.58	79.62
	W5	76.42	77.53
False alarm rate	W1	54.76	43.75
	W2	50.00	47.61
	W3	60.00	58.00
	W4	78.00	72.72
	W5	71.00	68.00

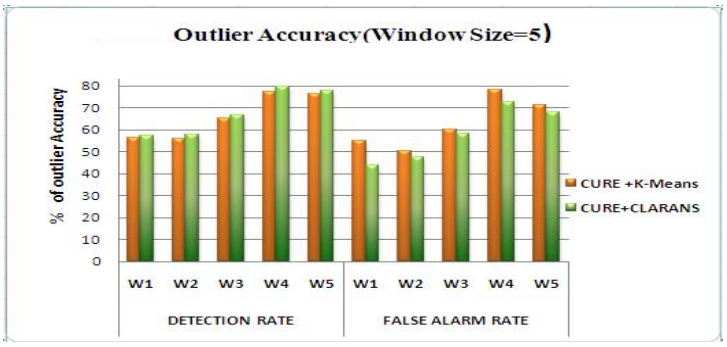

Figure-7: Detection rate and false alarm rate in five windows- Breast Cancer (Wiscosin)

From the above graph, it is observed that CURE with CLARANS clustering algorithm performs better than CURE with k-means algorithms for detecting outliers in both biological data set as Pima Indian diabetes and Breast Cancer (Wiscosin) in three windows as well as in five windows. Therefore the CURE with CLARANS clustering algorithm performs well because it contains high outlier detection accuracy when compared to birch with k-means.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

TABLE I THE CLUSTERING ACCURACY - THREE WINDOWS FOR TWO DATASETS

TABLE II

Clustering Accuracy	No. of Windows	CURE+ K-Means	CURE+CLARANS	CURE+ K-Means	CURE+CLARANS
		Pima Indian diabetes	Pima Indian diabetes	Breast cancer-Wiscosin	Breast cancer-Wiscosin
Accuracy	w1	82.03	88.28	82.40	88.41
	w2	82.10	88.32	82.05	88.03
	w3	82.03	88.28	82.40	88.41
Precision	w1	80.90	87.60	82.14	88.28
	w2	80.08	86.80	80.53	86.80
	w3	80.20	86.14	76.71	83.00
Recall	w1	81.70	87.60	82.32	88.20
	w2	82.26	87.25	81.11	88.84
	w3	83.88	88.82	84.02	87.29

TABLE II

THE CLUSTERING ACCURACY - FIVE WINDOWS FOR TWO DATASETS

Clustering Accuracy	No. of Windows	CURE+ K-Means	CURE+CLARANS	CURE+ K-Means	CURE+CLARANS
		Pima Indian diabetes	Pima Indian diabetes	Breast cancer-Wiscosin	Breast cancer-
Accuracy	w1	82.46	88.31	82.14	88.57
	w2	82.58	88.38	82.26	88.65
	w3	82.58	88.38	82.26	88.65
	w4	82.58	88.38	82.26	88.65
	w5	82.23	88.15	82.01	88.48
Precision	w1	80.71	86.81	81.9	88.99
	w2	82.17	87.96	82.19	88.66
	w3	81.22	87.36	81.13	87.09
	w4	78.11	83.86	75.32	82.70
	w5	80.66	87.07	77.19	84.18
Recall	w1	80.96	88.44	82.15	87.92
	w2	82.86	88.50	82.45	88.42
	w3	82.55	87.88	84.03	89.40
	w4	84.96	87.99	81.69	90.41
	w5	82.48	87.07	83.24	86.62

V. 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 and 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 outliers efficiently. In this paper we have analyzed the performance of CURE with K-Means and CURE with CLARANS

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

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 more efficient in CURE with CLARANS clustering while compare to CURE with K-Means with clustering.

REFERENCES

[1] C. Aggarwal, Ed, "Data Streams – Models and Algorithms", Springer, 2007.
[2] C. Aggarwal, J. Han, J. Wang, P.S. Yu, "A framework for projected clustering of high dimensional data streams", in Proceedings of the 30th VLDB Conference, Toronto, Canada, pp. 852-863, 2004.
[3] Bakar. Z, A. Mohemad, R .Ahmad, A. & Deris, M. M, "A comparative study for outlier detection techniques in data mining", IEEE Conf. Cybernetics and Intelligent Systems, Bangkok, Thailand, pp. 1–6,2006.
[4] D.Barbara, "Requirements for clustering data streams", ACM SIGKDD, Volume3 Issue 2, Pages 23-27 , January 2002.
[5] Carlos Ordonez, " Clustering Binary Data Streams with K-means", proceedings of ACM international conference on management of data, sigmod 1998.
[6]Chandrika.J, Dr. K.R. Ananda Kumar, "Dynamic Clustering Of High Speed Data Streams", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.
[7] Elahi, M. KunLi, Nisar, W. XinjieLv, HonganWang, "Fuzzy Systems and Knowledge Discovery", Fifth International Conference on Vol.5, and Vol .3, pp. 23-27, 2002.
[8] D. Hawkins, "Identification of outliers-Monographs on statistics and applied probability", First edition, pages-188, Springer published in 1980.
[9]Irad Ben-Gal, "Outlier Detection", Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer , Academic Publishers, 2005.
[10] L. O' Callaghan , " Clustering data streams theory and practice", IEEE transactions on knowledge and data engineering, Vol. 15, NO. 3,2003.
[11]Madjid Khalilian, Norwati Mustapha , "Data Stream clustering-Challenges and issues", Proceedings of the International Multi Conference of Engineers and Computer Scientists , Hong Kong ,Vol I,pp.17 - 19,March 2010.
[12]Sharma, M. Toshniwal, D, " Pre-clustering algorithm for anomaly detection and clustering that uses variable size buckets", Published in Recent Advances in Information Technology (RAIT), 1st International Conference on 15-17 March 2012.
[13] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "CURE: an efficient clustering algorithm for large databases", ACM LIBRARY, 1999.
[14]Thakran. Y, Toshniwal .D, " Unsupervised outlier detection in streaming data using weighted clustering", Intelligent Systems Design and Applications (ISDA), 2012.
[15]T. Soni Madhulatha, "overview of streaming-data algorithms", Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011.
[16]Yi-hong lu, Yan huang, "Mining DataStreams Using Clustering", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics,vol.4, pp. 18-21,2005.

Dr. S. Vijayarani

She has completed MCA, M.Phil and PhD in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security, bioinformatics and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. P. Jothi

She has completed M.Sc in Software Systems. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining and Data Streams.