ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# Comparative Analysis of Clustering Algorithms for Outlier Detection in Data Streams

**Dr. S.Vijayarani*1, Ms. P. Jothi2**

[*1] Assistant Professor, Department of Computer Science, School of Computer Science and Engg., Bharathiar University, Coimbatore, India

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

vijimohan_2000@yahoo.com

## Abstract

Nowadays, data mining has become one of the most popular research areas in the field of computer science, because data mining techniques are used for extracting the hidden knowledge from the large databases. In data mining, most of the work is emphasized over knowledge discovery and data stream mining is becoming an active research area in this domain. A data stream is a similar to river, it means continuous and massive sequence of data elements are in and out generated at a rapid rate and the analysis of data stream has been recently attracted attention over in data mining research community. When the amount of data is very huge, it leads to a numerous computational and mining challenges due to shortage of hardware and software limitations. Data mining techniques are newly proposed for data streams they are highly helpful to mine are data stream clustering, data stream classification, frequent pattern technique, sliding window techniques and so on. For outlier detection data stream clustering algorithm is highly needed. This 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 BIRCH with CLARANS and CURE with CLARANS are used for finding the outliers in data streams. Different types, sizes 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 CURE with CLARANS clustering algorithm performance is more accurate than the BIRCH with CLARANS.

**Keywords**: Data stream, Data stream Clustering, Outlier detection, BIRCH, CURE, CLARANS

## Introduction

Data mining is defined as the non-trivial extraction of hidden, potentially and earlier unknown useful information about data, due to the advances and production of information technologies and application of number of databases, as well as their complexity and dimension grows rapidly. However, there are lot of problems exists in huge database such as data redundancy, skewed data, missing data, incorrect data etc [1]. 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 [2]. 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. A model which is developed from data stream for clustering is an appropriate method for handling huge volume of updatable data.[2]Data stream clustering is a prominent task in mining data streams, the clustering can be considered the most important unsupervised learning problem; clustering is known as grouping related objects into a cluster. By using clustering method we can detect outliers, so it has become one of the data mining tasks and so it is called as outlier mining. An outlier is an object that does not fulfill with the behavior of normal data objects. The definition proposed by a D.M. Hawkins [4], an outlier is an observation, which deviates so much from other observations as to arouse suspicions that it was generated

ISSN: 2277-9655
Impact Factor: 1.852

by a different mechanism. Another definition proposed by Barnett and Lewis [14] that an outlier is an observation which appears to be inconsistent with the remainder of that set of data and there are many applications of outlier detection they are telecommunication, fraud detection, web logs, click stream and web document and so on. Due to the enormous applications, outliers are becoming more interesting factor. Outliers have various names they are anomalies, errors, noise, damage, novelty, surprise, peculiarities or contaminants[5].Clustering based outlier mining methods are called as unsupervised in nature and the main objective is used to find the outlier from the data stream using hierarchical cluster based method and partioning cluster based method. The cluster based outlier detection approaches are defined as it uses a cluster based technique for detecting outlier where it finds closely related objects. The object which does not belong to any cluster or belongs to a small cluster is declared as outlier, and the outlier detection is highly depends upon clustering usage for detecting outliers.

The remaining portion of the paper is organized as follows; section 2 illustrates the review of literature. Section 3 describes 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.

## Literature Review

**Dajun Wang et.al,** [3] conversed about an efficient hierarchical agglomerative clustering (HAC) algorithm based on vertical and horizontal dimension reduction algorithms. In this research the authors described a unique similarity based on effort to identify outliers among high dimensional time series objects in financial markets and also the similarity between two assets decreases in the portfolio and the benefits of diversification increase. Finally, the authors proposed a unique similarity measurement, definition and calculation based on the time-value function and discloses a series of experiment results illustrating the effectiveness of the framework and as well as outlier detection can be used to monitor portfolio diversification and therefore mitigate risk.

**Luis Torgo, et.al** [7] put forth a methodology for the application of hierarchical clustering methods to the task of outlier detection and the methodology is tested on the problem of cleaning official statistics data and the objective is to detect erroneous foreign trade transactions in data collected by the Portuguese Institute of Statistics (INE). The method is based on the output of standard hierarchical agglomerative clustering algorithms, where as resulting requires no significant additional computational costs. In this study the authors compared the proposal state of the art outlier ranking method (LOF) and show that the method achieves better results on this particular application and the experiments results are also competitive with previous results on the same data. At last, the outcome of the experiments raises important questions concerning the method currently followed at INE concerning items with small number of transactions.

**Rajendra Pamula, et.al** [10] presented the clustering based method to hold outliers, the k-means clustering algorithm to divide the data set into clusters and the points which are lying near the centroid of the cluster are not probable candidate for outlier and it can prune out such points from each cluster. To calculate a distance based outlier score for remaining points and the computations are needed to calculate the outlier score reduces considerably due to the pruning of some points. Based on the outlier score it's declared the top n points with the highest score as outliers. The local distance-based outlier factor to measure the degree to which an object deviates from its neighborhood. The precision of detecting outliers of the method is at per or higher than the existing methods though we pruned out some of the points. The experimental results using real data set demonstrate the number of computations is fewer then the proposed method performed better than the existing method.

**G. S. David Sam Jayakumar, et.al**[6] implemented a new method of clustering was proposed based on multivariate outlier detection and several clustering procedures available in the literature, the proposed technique gives a particular idea to cluster the sample observations in a survey study based on the multivariate outliers. The feature of the proposed clustering technique was elaborately discussed and the authors also highlighted the application of the technique in a survey research. Based on the results de-rived, the proposed technique gives more insights to the researcher to cluster the sample observation at 5% and 1% significance level. Finally the authors inform an idea for further research by conducting simulation experiments for testing relationship between the significance level and the number of outlier clusters extracted.

## Methodology

In data streams, clustering techniques are applied for grouping the data items and also detecting the outliers. Clustering and Outlier detection is one of the important problems 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

ISSN: 2277-9655
Impact Factor: 1.852

work, two clustering algorithms namely BIRCH with CLARANS 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:

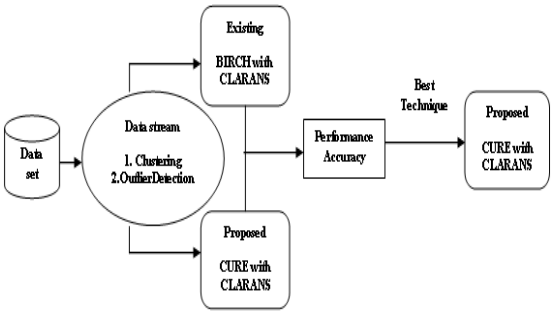


**Figure 1: System architecture of clustering algorithms for outlier detection**

### 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. The chunk size is specified by the user which depends upon the nature of data and then we divided the data into chunks of same size in different windows.

### B. Clustering

The clustering algorithm is used to grouping objects into important subclasses and the clustering data streams is a sub division of mining data streams[15]. The clustering algorithms are different types of applications and they are followed by hierarchical clustering algorithm, partition clustering algorithm, density based clustering algorithm and grid based clustering algorithm. Cluster analysis is used in a number of applications such as data analysis, image processing, stock market analysis etc. Cluster analysis or clustering is the assignment of a set of observations into subsets called clusters so that observations in the same clusters are related in some sense [1]. It is a useful procedure for the discovery of data distribution and patterns in the original data and the goal of clustering technique is to find out both the sparse and dense and area in a data set. It is a method of unsupervised learning and a common technique for statistical data analysis used in many areas, such as data mining, machine learning, image analysis, pattern recognition and bioinformatics so on.

### C. Outlier Detection

Outliers are objects that do not comply with the general behavior of the data. By definition, outliers are rare occurrences and hence represent a small portion of the data. Outlier detection has direct applications in a wide variety of domains such as mining for anomalies to detect network intrusions, fraud detection in mobile phone industry and recently for detecting terrorism related activities [5].Outlier detection is less straight forwarding in two dimensional spaces, because visual inspection is less valuable and the order statistics are lacking and traditional multivariate outlier-detection methods are based on the calculation of the generalized squared Mahalanobis distances for each data point. Unfortunately, outliers greatly increase the covariance matrix and can therefore effectively mask their own existence. To offset this covering problem, Rousseau introduced the robust minimum volume ellipsoid (MVE) method for detection of outliers in multidimensional data subsets to find the subset that minimizes the volume occupied by the data [8] and the clustering based outlier detection is a best technique to identify the outliers.For our research we have used cluster based outlier detection as BIRCH with CLARANS and CURE with CLARANS.

### D. BIRCH

BIRCH [15] is abbreviated as balanced iterative reducing and clustering using hierarchies is an unsupervised algorithm used to perform hierarchical clustering over particularly huge data sets. An advantage of Birch is its ability to dynamically and incrementally cluster, multi dimensional metric and cluster incoming data points in an effort to create the best feature clustering for a given set of resources time and memory constraints. In addition, Birch is clustering algorithm proposed in the database area to handle noise data points that are not part of the underlying pattern effectively. The each clustering choice is made without scanning all data points and currently existing clusters and it utilizes the observation that data space is not usually uniformly occupied and not every data point is evenly significant. It use of available memory to derive the best possible sub clusters while minimizing Input and output costs and it is also an incremental method that does not require the whole data set in advance. Given a set of N d-dimensional data points, the clustering feature CF of the set is distinct as the triple CF=[N, LS, SS] LS is the linear sum, SS is the square sum of data points. Clustering features are ordered in a CF tree, and it is a height balanced tree along with two parameters branching factor B and threshold T. Each non-leaf node contains the most B entries of the form [$CF_i$, $child_i$], where childi is a pointer to its $i^{th}$ , child node $CF_i$ and the clustering feature representing and they are related to sub cluster.

ISSN: 2277-9655
Impact Factor: 1.852

$$\vec{x}_0 = \frac{\sum_{i=1}^{N} \vec{x}_i}{N}$$

$$R = \left(\frac{\sum_{i=1}^{N} (\vec{x}_i - \vec{x}_0)^2}{N}\right)^{\frac{1}{2}}$$

$$D = \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (\vec{x}_i - \vec{x}_j)^2}{N(N-1)}\right)^{\frac{1}{2}}$$

In the algorithm in the first step it scans all data and builds an initial memory CF tree using the given amount of memory and in the second step it scans all the leaf entries in the initial CF tree to rebuild a minor CF tree, while removing outliers and grouping crowded sub clusters into better ones. In step three an existing clustering algorithm is used to cluster all leaf entries where here an agglomerative hierarchical clustering algorithm is functional directly to the sub clusters represented by their CF vectors. It also provides the elasticity of allowing the user to specify either the desired number of clusters or the desired diameter threshold for clusters and after this step a set of clusters is finded that captures major distribution pattern in the data. However there forced to exist insignificant and localized inaccuracies which can be handled by an optional step 4 and in step 4 the centroids of the clusters are used to formed in step and redistributed the data points to its closest seeds to obtain a new set of clusters. The step 4 also provides us with an option of discarding outliers and t hat is a point which is too far from its closest seed can be treated as an outlier.

**E.CURE**

CURE [11] is a hierarchical clustering technique where each partition is nested into the next partition in the sequence and CURE is an agglomerative algorithm where disjoint clusters are successively merged until the number of clusters reduces to the desired number of clusters. CURE is strong to outliers and identifies clusters with wide variances and non-spherical shapes in size. Each cluster is represented by a fixed number of well scattered points and it is positioned between centroid based ($d_{avg}$) and all point ($d_{min}$) in tenses. A constant number of well scattered points is used to capture the shape and extend of a cluster and the points are shrunk towards the centroid of the cluster by a factor α.These well scattered and shrunk points are used as representative of the cluster. This allows CURE to correctly identify the clusters and makes it less sensitive to outliers. In CURE, instead of using a single centroid to represent a cluster, a constant number of delegate two points are chosen to represent a cluster. Similarity between two clusters is calculated and the closest pair of the representative points belonging to different clusters then new representative points for the merged clusters is determined by selecting a constant number of well scattered points from all the data points and shrinking them towards the centroid of the cluster according to a shrinking part. The cure is capable of finding clusters of arbitrary sizes and shapes, as it represents each cluster into multiple representative points and shrinking the delegate points towards the centroid helps CURE in avoiding the problem of noise and outliers present in the single link method. The popular value of the decrease factor in CURE is dependent upon cluster sizes and shapes, amount of noise in the data and in some agglomerative hierarchical algorithms, the comparison between two clusters is detained by the aggregate of the similarities that is interconnectivity among pairs of items belonging to different clusters. Hence, many such schemes normalize the aggregate similarity between a pair of clusters with respect to the expected inter-connectivity of the clusters involved. The cure algorithm follows 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.

**F.CLARANS**

CLARANS [12] is abbreviated as Clustering Large Application Based upon Randomized Search and it use random search, to generate neighbors by starting with arbitrary node and randomly check max-neighbors. CLARANS are similar to PAM and CLARA while it starts with the selection of mediod at randomly and it describes the neighbor dynamically and it checks max neighbor for swapping and if the pair is negative then it chooses another medoid set or otherwise it chooses current selection of medoids as local optimum and restarts with the new selection of medoids randomly and it stops the process until returns the best. If the neighbor represent as better partition the process continue with

ISSN: 2277-9655
Impact Factor: 1.852

new node otherwise local minimum is found and algorithm restart until num local minima is found the value of num local is=2 recommended then the best node return resulting partition. CLARANS take a random dynamic selection of data

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).

at each step of process and thus the same sample set is not used throughout in the clustering process and CLARANS is accurately detecting outlier than CLARA and it is much less affected by increasing dimensionally and draw the sample of neighbors in each step of search this is benefit of confining the search localize area.

## Experimental Results

The algorithms used in this research work are implemented in MATLAB 7.10 (R2010a). In order to evaluate the performance of the algorithms, the two factors namely clustering accuracy and outlier accuracy are used.

**Clustering Accuracy**

Clustering accuracy is calculated, by using two measures Precision and recall. The clustering algorithms BIRCH with CLARANS and CURE with CLARANS for two data sets are Pima Indian diabetes and Wiscosin-breast cancer data set. Table 1 & Table 2 show the clustering accuracy, precision and recall in three windows and five windows.

**Table 1: The clustering accuracy in three windows for two dataset**

| Clustering Accuracy | No. of Windows | BIRCH+CLARANS | CURE+CLARANS | BIRCH+CLARANS | CURE+CLARANS |
|---|---|---|---|---|---|
| | | Pima Indian diabetes | Pima Indian diabetes | Breast cancer-Wiscosin | Breast cancer-Wiscosin |
| Accuracy | w1 | 76.17 | 88.28 | 76.39 | 88.41 |
| | w2 | 76.20 | 88.32 | 76.06 | 88.03 |
| | w3 | 76.17 | 88.28 | 76.39 | 88.41 |
| Precision | w1 | 74.92 | 87.60 | 76.11 | 88.28 |
| | w2 | 74.00 | 86.80 | 74.79 | 86.80 |
| | w3 | 74.06 | 86.14 | 69.55 | 83.00 |
| Recall | w1 | 74.01 | 87.60 | 76.33 | 88.20 |
| | w2 | 75.89 | 87.25 | 76.41 | 88.84 |
| | w3 | 76.73 | 88.82 | 74.29 | 87.29 |

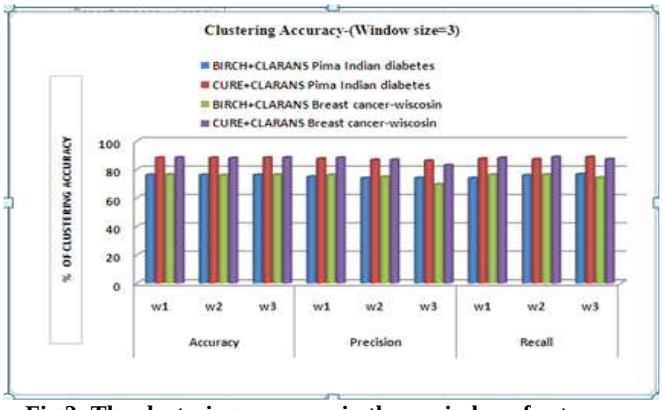


**Fig 2: The clustering accuracy in three windows for two dataset**

ISSN: 2277-9655
Impact Factor: 1.852

**Table 2: The clustering accuracy in five windows for two dataset**

| Clustering Accuracy | No. of Windows | BIRCH +CLARANS | CURE+CLARANS | BIRCH +CLARANS | CURE+CLARANS |
|---|---|---|---|---|---|
| | | Pima Indian Diabetes | Pima Indian diabetes | Breast cancer-Wiscosin | Breast cancer-Wiscosin |
| Accuracy | w1 | 76.62 | 88.31 | 76.42 | 88.57 |
| | w2 | 76.12 | 88.38 | 76.59 | 88.65 |
| | w3 | 76.12 | 88.38 | 76.59 | 88.65 |
| | w4 | 76.12 | 88.38 | 76.59 | 88.65 |
| | w5 | 76.31 | 88.15 | 76.25 | 88.48 |
| Precision | w1 | 75.22 | 86.81 | 76.18 | 88.99 |
| | w2 | 76.18 | 87.96 | 76.40 | 88.66 |
| | w3 | 74.39 | 87.36 | 75.12 | 87.09 |
| | w4 | 69.50 | 83.86 | 71.20 | 82.70 |
| | w5 | 74.84 | 87.07 | 69.82 | 84.18 |
| Recall | w1 | 77.31 | 88.44 | 75.87 | 87.92 |
| | w2 | 76.88 | 88.50 | 76.46 | 88.42 |
| | w3 | 74.88 | 87.88 | 77.30 | 89.40 |
| | w4 | 72.14 | 87.99 | 79.21 | 90.41 |
| | w5 | 76.64 | 87.07 | 72.76 | 86.62 |

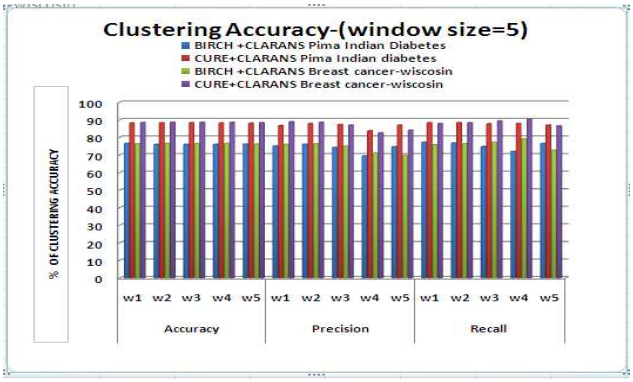


**Figure 2: The clustering accuracy in five windows for two dataset**

From the above graph, it is observed that CURE with CLARANS clustering algorithm performs better than BIRCH with CLARANS 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 BIRCH with CLARANS.

**Outlier accuracy**

***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 BIRCH with CLARANS and CURE with

CLARANS for Pima Indian diabetes data set. Table 3 & Table 4 show the number of outlier detection rate and false alarm rate in three windows and five windows.

**Table 3: Detection rate and false alarm rate in three windows-Pima Indian diabetes**

| Outlier Accuracy | No. of Windows | BIRCH+CLARANS | CURE +CLARANS |
|---|---|---|---|
| **Detection rate** | **W1** | **36.80** | **39.48** |
| | **W2** | **35.70** | **37.54** |
| | **W3** | **33.00** | **35.00** |
| **False alarm rate** | **W1** | **45.00** | **34.42** |
| | **W2** | **33.76** | **26.08** |
| | **W3** | **35.00** | **26.68** |

ISSN: 2277-9655
Impact Factor: 1.852

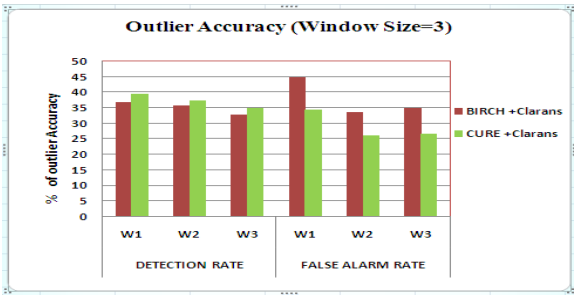


**Figure 3: Detection rate and False alarm rate in three windows-Pima Indian diabetes**

**Table 4: Detection rate and False alarm rate in five windows-Pima Indian diabetes**

| Outlier Accuracy | No. of Windows | BIRCH+CLARANS | CURE +CLARANS |
|---|---|---|---|
| Detection rate | W1 | 33.82 | 37.79 |
| | W2 | 41.28 | 45.22 |
| | W3 | 36.69 | 39.84 |
| | W4 | 25.78 | 27.96 |
| | W5 | 34.00 | 36.44 |
| False alarm rate | W1 | 33.82 | 32.22 |
| | W2 | 41.28 | 37.03 |
| | W3 | 36.69 | 22.22 |
| | W4 | 25.78 | 19.11 |
| | W5 | 34.00 | 30.00 |

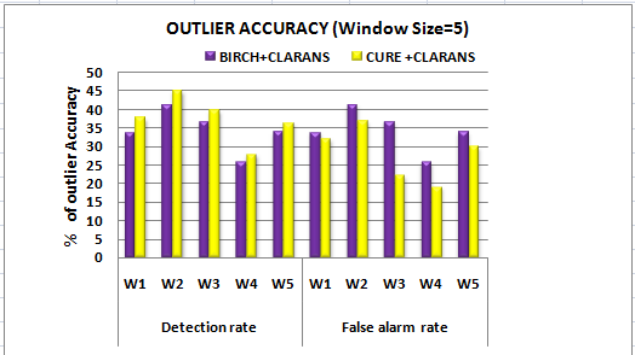


**Figure 4: Detection rate and false alarm rate in five windows-Pima Indian diabetes**

From the above graph, it is observed that CURE with CLARANS clustering algorithm performs better than BIRCH with CLARANS 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 BIRCH with CLARANS.

**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 BIRCH with CLARANS and CURE with CLARANS for Breast cancer –Wiscosin data set. Table 3 & Table 4 show the number of outlier detection rate and false alarm rate in three windows and five windows.

**Table 5: Detection rate and false alarm rate in three windows-Breast cancer (Wiscosin)**

| Outlier Accuracy | No. of Windows | BIRCH+CLARANS | CURE +CLARANS |
|---|---|---|---|
| Detection rate | W1 | 55.72 | 57.76 |
| | W2 | 63.41 | 67.41 |
| | W3 | 75.52 | 78.65 |
| False alarm rate | W1 | 56.09 | 43.74 |
| | W2 | 64.28 | 51.78 |
| | W3 | 80.92 | 70.90 |

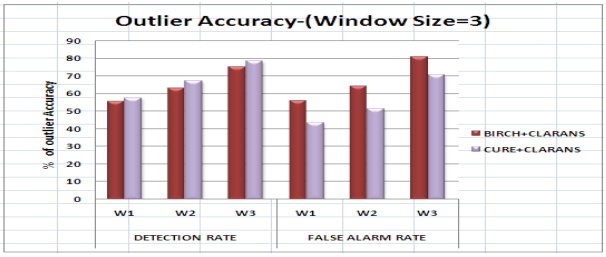


**Figure 5: Detection rate and false alarm rate in three windows- Breast Cancer (Wiscosin)**

**Table 6: Detection rate and false alarm rate in five windows- Breast cancer (Wiscosin)**

| Outlier Accuracy | No. of Windows | BIRCH+CLARANS | CURE +CLARANS |
|---|---|---|---|
| Detection rate | W1 | 56.02 | 57.25 |
| | W2 | 53.60 | 57.57 |

ISSN: 2277-9655
Impact Factor: 1.852

| | | | |
|---|---|---|---|
| | W3 | 64.00 | 66.60 |
| | W4 | 77.70 | 79.62 |
| | W5 | 75.47 | 77.53 |
| False alarm rate | W1 | 52.30 | 43.75 |
| | W2 | 62.50 | 47.61 |
| | W3 | 72.00 | 58.00 |
| | W4 | 78.00 | 72.72 |
| | W5 | 72.00 | 68.00 |

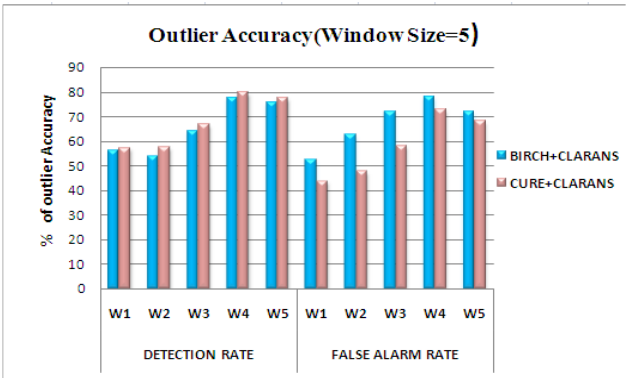


**Figure 6: Detection rate and false alarm rate in five windows- Breast Cancer (Wiscosin)**

From the above graph, it is observed that CURE with CLARANS clustering algorithm performs better than BIRCH with CLARANS 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 CLARANS.

## 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 outliers efficiently. In this paper we have analyzed the performance of BIRCH with CLARANS and CURE with CLARANS clustering algorithm for detecting the outliers. In order to find the best clustering algorithm for outlier detection two performance measures are used. From the experimental results it is observed that the outlier detection accuracy and clustering accuracy is more efficient in CURE with CLARANS clustering while compared to BIRCH with CLARANS clustering.

## References

[1] 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, 2004, pp. 852-863.

[3] DajunWang , Fortier,P.J, Michel,H.E.,Mitsa,T.FidelityInvestments, "Hierarchical Agglomerative Clustering Based T-outlier Detection" MassachusettsUniv,ICDM Workshops 2006.

[4] D.M. Hawkins, "Identification of Outliers", London: Chapman and Hall, 1980.

[5] Irad Ben-Gal, "Identification of outliers", Department of Industrial Engineering, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel D. Barbara, "Requirements for clustering data streams," ACM SIGKDD.

[6] G. S. David Sam Jayakumar and Bejoy John Thomas, "A New Procedure of Clustering Based on Multivariate Outlier Detection", Journal of Data Science 11(2013).

[7] Luis Torgo, Carlos soares, "Resource-bounded Outlier Detection using Clustering Methods", proceedings of the 2010 conference on data mining for business applications.

[8] M. L. Prasanthi , A.Krishna Chaitanya , Dr. N. Sambasiva Rao ,"Detection of Outliers and Change points in a Data Stream of Bio Informatics Data", International Journal of Engineering Research & Technology (IJERT)Vol. 1 Issue 6, August – 2012,ISSN: 2278-0181.

[9] Madjid Khalilian, Norwati Mustapha "Data Stream clustering-Challenges and issues", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2010 Vol I,IMECS 2010,March 17 - 19, 2010, Hong Kong.

[10] Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi "An Outlier Detection Method based on Clustering", Second International

ISSN: 2277-9655
Impact Factor: 1.852

Conference on Emerging Applications of Information Technology, 2011.

[11] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, CURE: an efficient clustering algorithm for large databases, ACM LIBRARY, 1999.

[12] T. Soni Madhulatha, Advanced Computing: An International Journal (ACIJ), "overview of streaming-data algorithms", Vol.2, No.6, November 2011.

[13] Thakran Y, Toshniwal .D, "Unsupervised outlier detection in streaming data using weighted clustering", Intelligent Systems Design and Applications (ISDA), 2012. vol. 3, pp. 23-27, 2002.

[14] V. Barnett and T. Lewis, "Outliers in Statistical Data", New York: John Wiley Sons, 1994.

[15] YI-HONG LU1, YAN HUANG, "Mining DataStreams Using Clustering", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21.

[16] Zhang, T, Raghu, R, Miron, L." BIRCH:An Efficient Data Clustering Method for Very Large Databases", ACM SIGMOD Record, vol. 25(2), 103-114 (1996).