ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

An Efficient Clustering Algorithm for Outlier Detection in Data Streams

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: Data mining is extensively studied field of research area, where most of the work is highlighted over knowledge discovery. Data stream is dynamic research area of data mining. A data stream is an enormous sequence of data elements continuously generated at a fast rate. In data streams, huge amount of data continuously inserted and queried, such data has very large database. The data stream is motivated by emerging applications involving massive data sets for example, consumer click streams and telephone records, bulky sets of web pages, multimedia datas, and financial transactions and so on. It raises new problems for the data stream community in terms of how to mine continuous arrival of high speed data items. Recently 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 techniques are highly helpful to cluster the similar data items in data streams and also to detect the outliers, so they are called cluster based outlier detection. 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 K-Means and Birch with CLARANS are used for clustering the data items and 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 analysing the experimental results, it is observed that the proposed BIRCH with CLARANS clustering algorithm performance is more accurate than the existing algorithm BIRCH with K-Means.

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

I. INTRODUCTION

Data mining is extensively studied field of research area. Extraction of interesting non-trivial, hidden and potentially useful patterns or knowledge from huge amount of data where most of the work is highlighted over knowledge discovery[1] .However, there are a lot of problem exists in huge database such as data redundancy, missing data, skewed data, invalid data etc. One of the major problem in data mining research increase in dimensionality of data gives rise to a number of new computational challenge not only due to increase in number of attributes. In recent years, we have observed that enormous research activity motivated by the explosion of data collected and transferred in the format of data streams. Data streams handles enormous amount of data being generated every day in a timely approach. One of the important characteristics in data stream is whenever the data information is unbounded there is no assumption length of ordering is maintained [1]. Data streams can be solved using the methodologies of data stream clustering, data stream classification, frequent pattern mining, sliding window, Association technique and so on[2]. Clustering is a prominent task in mining data streams, which group related objects into a cluster. The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects .A number of clustering algorithms have been introduced in recent years for data streams [9]. With the applicability of data streams, clustering data streams have acknowledged more attention in data mining research. The clustering can be considered the most important unsupervised learning problem; so, as every other problem of this category, it deals with finding a structure in a collection of unlabelled data. Outlier Detection over data stream is active research area from data mining that aims to detect object which have different behaviour, exceptional than normal object [3]. Depending upon different application domains these abnormal patterns are often referred to as outliers, anomalies, discordant observations, faults, exceptions, defects, aberrations, errors, noise, damage, surprise, novelty, peculiarities or impurity.

Over the years, a large number of techniques have been developed for building such models for outlier and anomaly detection. However, the real world data sets, data stream present a range of difficulties that bound the effectiveness of the techniques. The assumed behavior of outliers it does not belong to any cluster, or belong to a very small clusters, are otherwise forced to belong a cluster where they were different from other members of cluster. The clustering techniques are highly helpful to detect the outliers they are called cluster based outlier detection. The rest of this paper is organized as follows. Section 2 illustrates the review of literature. Section 3 discusses the Birch with K-means and Birch with Clarans clustering algorithms for detecting outliers in data streams. Experimental results are discussed in Section 4 and Conclusions are given in Section 5.

II. Literature Review

Pedro Pereira Rodrigues, et.al [11] proposed a new incremental algorithm for clustering streaming of time series. The ODAC system is Online Divisive-Agglomerative Clustering system continuously maintains a tree-like hierarchy of clusters that evolves with data. The system is designed and planed to process number of data streams that flow at high-rate. The system main features include update memory and time consumption that does not depend on the examples of number of stream. Moreover, the time and memory required to process in lower, whenever the cluster structure expands. Experimental results on real and artificial data assess the system processing qualities, signifying the competitive performance on data stream clustering time series, and also its deal with concept drift.

Irad Ben Gal, et.al [8] discussed about a several methods and techniques for outlier detection, how the outliers are distinguishing between uniform variate vs. multivariate techniques and parametric vs. nonparametric procedures. A variety of semi-supervised, supervised, and unsupervised techniques have been used for outlier detection and each has their own strengths and weaknesses have been discussed by an author in this research.

De Andrade Silva, J et.al [4] discussed many algorithms for clustering data streams based on the widely used k-Means have been discussed in this research. In this research the authors described an algorithmic framework that allows estimating k automatically from data as well as they assume the number of clusters as K is known and fixed a priori by the user. The proposed framework in this research is by using three state-of-the-art algorithms for clustering data streams are Stream LSearch, CluStream, and Stream KM++ combined with two algorithms for estimating the number of clusters, they are OMRK-Ordered Multiple Runs of k-Means and BKM- Bisecting k-Means .The authors had experimentally compares the results in both synthetic and real-world data streams. The statistical significance of analysis suggests that OMRK yields the best data partitions and BKM is more computationally efficient. The combination of OMRK with Stream KM++ leads to the best trade-off between efficiency and accuracy.

HosseinMoradiKoupaie, et.al [7] proposed cluster based outlier detection in data stream. In this research author prefer an incremental clustering algorithm to detect real outlier in stream data. K-means algorithm using Start Enter data with window with specify size .Clustering these data in window by K-means algorithm. Finding the sum of cluster that are small and faraway of other clusters as outlier .Report these data as online outlier and store in memory clustering these data in window by K means algorithm Add previous outlier in n previous window to this window ,finding some cluster that are small and faraway of other clusters as outlier. Finally they report these data as real outlier.

Hendrik Fichtenberger et.al,[6] proposed a data stream algorithm for the k-means problem called BICO (BIRCH Meets Core sets for *k*-Means Clustering),that combines the data structure of the SIGMOD test of time award winning algorithm birch with the theoretical concept of corsets for clustering problems. BICO computes high quality solutions in a time short and also bico computes a summary S of the data with a provable quality. For every center set C, S has the same cost as P up to a certain limit factor. In a data stream, the points arrive one by one in arbitrary order and limited storage space. In this research author compare BICO experimentally with popular BIRCH and Mac Queen Algorithm and also with approximation algorithms as Streamkm++ and Stream LS.

III. Methodology

In data streams we are going to apply the clustering technique for grouping the data items and also detecting the outliers. Clustering and Outlier detection is one of the important issues 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 BIRCH with K-Means and Birch 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:

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

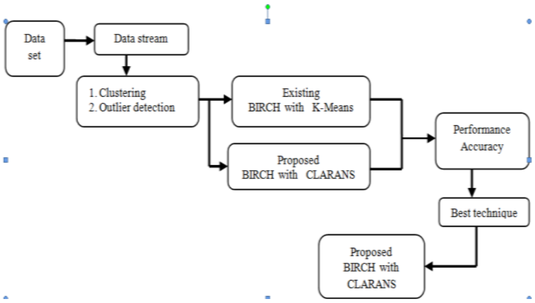

Fig 1: System Architecture of clustering algorithms for detecting outliers

A. Dataset

In order to compare the data stream clustering for detecting outliers, data sets were taken from UCI machine learning repository. All data sets have numeric attributes. Datasets namely Breast Cancer Wisconsin Dataset with 10 attributes, 699 instances and Pima Indian data setcontain8 attributes and 768 instances. Stream of data is an unbounded sequence of data. As it is not possible to store complete data stream, for processing we divide it into data chunks of same size. Chunk size is specified by the user which depends upon the nature of data divided into chunks of same size in different windows.

B. Clustering

Clustering is a prominent task in mining data streams, which group related objects into a cluster. A number of clustering algorithms have been introduced in recent years for data streams. Data stream mining is nothing but extracting the useful patterns from data streams. Data streams need to be processed as they arrive. The term clustering is employed by several research communities to describe the method of grouping unlabeled data. Clustering is used to improve the efficiency of the result by making groups of the data. The goal of a clustering algorithm is to group objects into meaningful subclasses. There are different types of clustering algorithms suitable for different types of applications they are followed by Hierarchical clustering algorithm, Partition clustering algorithm, Spectral clustering algorithm, Grid based clustering algorithm and Density based clustering algorithm.

C. Outlier Detection

Detecting outliers over data stream is active research area in recent years. Data are continuously coming in a streaming environment with a very fast rate and changing data distribution. The change of data distribution is called as concept drift. Outlier detection is varies according to different entities in different domains. Outlier detection terminology refers to task of finding outliers as per behaviour of data and distribution of data. There are different types of techniques used in data streams they are statistical Outlier Detection, Depth based outlier detection, Clustering Based Outlier Detection and so on. For our research we have used cluster based outlier detection as BIRCH with k-means and BIRCH with CLARANS.

D. BIRCH Clustering

BIRCH [15] is expanded into Balanced Iterative Reducing and Clustering using Hierarchies. It is a first algorithm proposed in the database area that addresses outlier's data points that should be regarded as noise and proposes a reasonable solution. There are two types of approaches in Birch clustering is defined as Probability-based approaches and Distance based approaches. Probability based approaches makes the assumption that probability distributions on separate attributes are statistically independent of each other. The probability-based tree is built to identify clusters. Distance-based approaches refer to global or semi-global methods at the granularity of data points. Assume that all data points are given in advance and can be scanned frequently. Ignore the fact that not all data points in the dataset are equally important. None of them have linear time scalability with stable quality. The Birch algorithm builds a dendogram called clustering feature tree while scanning the data set. BIRCH is local in that each clustering decision is made without scanning all data points. BIRCH develops the observation of the data space is not uniformly occupied, since every data point is a likely important for clustering. BIRCH makes full use of available memory to derive the finest possible sub clusters to ensure accuracy while minimizing I/O costs to ensure efficiency. Given N d-dimensional data points in a cluster: where i = 1, 2,…,N, the centroid, radius and diameter is defined as

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

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

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

R and D determine the properties of a single cluster. Next between two clusters, we define 5 alternative distances for measuring their nearest. Given the centroid of two clusters , the two different distances D0 and D1 are defined as

$$\vec{x}_{0_1} \, and \, \vec{x}_{0_2}$$

$$D0 = ((\vec{x}_{0_1} - \vec{x}_{0_2})^2)^{\frac{1}{2}}$$

$$D1 = | \vec{x}_{0_1} - \vec{x}_{0_2} | = \sum_{i=1}^{d} | \vec{x}_{0_1}{}^{(i)} - \vec{x}_{0_2}{}^{(i)} |$$

Given N_1 d-dimensional data points in a cluster: where i = 1,2,…N_1, and N_2 data points in another cluster where j = N_1+1,N_1+2,…,N_1+N_2, the average of inter-cluster distance as D_2, average intra-cluster as distance D_3 and variance increase distance D_4 of the two clusters are defined as Actually, D3 is D of the joined cluster.

$$D2 = \left(\frac{\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+N_2}(\vec{x}_i - \vec{x}_j)^2}{N_1 N_2}\right)^{\frac{1}{2}}$$

$$D3 = \left(\frac{\sum_{i=1}^{N_1+N_2}\sum_{j=1}^{N_1+N_2}(\vec{x}_i - \vec{x}_j)^2}{(N_1+N_2)(N_1+N_2-1)}\right)^{\frac{1}{2}}$$

$$D4 = \sum_{k=1}^{N_1+N_2}\left(\vec{x}_k - \frac{\sum_{l=1}^{N_1+N_2}\vec{x}_l}{N_1+N_2}\right)^2 - \sum_{i=1}^{N_1}\left(\vec{x}_i - \frac{\sum_{l=1}^{N_1}\vec{x}_l}{N_1}\right)^2 - \sum_{j=N_1+1}^{N_1+N_2}\left(x_j - \frac{\sum_{l=N_1+1}^{N_1+N_2}\vec{x}_l}{N_2}\right)^2$$

D0, D1, D2, D3, D4 are the measurement of two clusters. To determine the two clusters are closed are to be used. Figure 1 presents the overview of BIRCH it determines four phases, 1 .Loading, 2.Optional condensing, and 3.Global clustering and 4. Optional refining. The main task of Phase 1 is to scan all data and build an initial in-memory CF-tree using the given amount of memory and recycling space on disk.

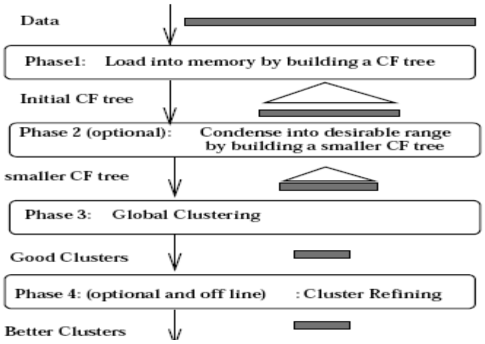
Data
Phase1: Load into memory by building a CF tree
Initial CF tree
Phase 2 (optional): Condense into desirable range by building a smaller CF tree
smaller CF tree
Phase 3: Global Clustering
Good Clusters
Phase 4: (optional and off line) :Cluster Refining
Better Clusters

Fig 2: Birch Overview [15]

This CF-tree tries to reflect the clustering information of the dataset in as much detail as possible subject to the memory limits. With crowded data points grouped into sub clusters, and sparse data points removed as outliers, this phase creates an in-memory summary of the data. After Phase 1, computations of subsequent in later phases will be: (1) fast because (a) no I/O operations are needed, and (b)the problem of clustering the original data is reduced to a smaller problem of clustering the sub clusters in the leaf entries; (2) accurate because (a) outliers can be eliminated, and (b) the remaining data is described at the finest granularity that can be achieved given the available memory; (3) less order sensitive because the leaf entries of the initial tree form an input order containing better data locality compared with the arbitrary original data input order once all the clustering information is loaded into the in-memory CF-tree, we can use an existing global or semi global algorithm in Phase 3 to cluster all the leaf entries across the boundaries of different nodes. Finally BIRCH does not assume that the probability distributions on separate attributes are independent.

E. K-means clustering

The k-means [12] algorithm is the best known partitioned clustering algorithm. The K-means clustering algorithm is a simple method for estimating the mean (vectors) of a set of K-groups. The most widely used K-Means among all clustering algorithms due to its efficiency and simplicity. Given a set of data points and the required number of k clusters and k is specified by the user, the k-means algorithm are partitions the data into k clusters based on a distance function. The k-means method partitions the data into k clusters, where as the k is supplied by the user. The algorithm partitions the objects so as to minimize with in cluster divergence, where a divergence is defined as the difference between an object and its centroid. The k-means algorithm is as follows

```
Algorithm k-means (k, D)
1 choose 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 clustering is a partitioning clustering algorithm. Clarans is expanded into clustering algorithm based on randomized search. Instead of comprehensively searching a random subset of objects, CLARANS [13] proceeds by searching a random subset of the neighbours of a particular solution, S. Thus the search for the best representation is not connected to a local area of the data. The CLARANS algorithm is directed by two parameters they are MAX neighbour, the maximum number of neighbours of S to assess and MAX sol, the numbers of local solutions are to be obtained. 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 neighbour R of S at random. Calculate the total swap contribution of the two neighbours.

4. If R has a lower cost, set R = S and go to Step 2. Otherwise increment j by one. If j ≤ MAX neighbour go to Step

5. When j > MAX neigh, 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. If i > MAX sol stop, otherwise go to Step 1.

IV. EXPERIMENT RESULTS

We have implemented our algorithms in MATLAB 7.10(R2010a). To evaluate two performance factors such as clustering accuracy and outlier detection accuracy are used for analysis. For that we have use two biological data set they are Pima Indian diabetes and breast cancer (Wiscosin). Detection rate refers to the ratio between the numbers of correctly detected outliers to the total number of actual outliers. False alarm rate is the ratio between the numbers of normal objects that are misinterpreted as outlier to the total number of alarms.

A. *Clustering Accuracy*

Clustering accuracy is calculated, by using two measures Precision and recall. The clustering algorithms BIRCH with K-MEANS and BIRCH with CLARANS for pima Indian diabetes and Wiscosin-breast cancer data set. Table 1 & Table 2 shows the clustering accuracy, precision and recall 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 +k means	Birch +clarans
detection rate	w1	32.00	36.80
	w2	33.00	35.70
	w3	31.00	33.00
false alarm rate	w1	50.00	45.00
	w2	36.06	33.76
	w3	40.00	35.00

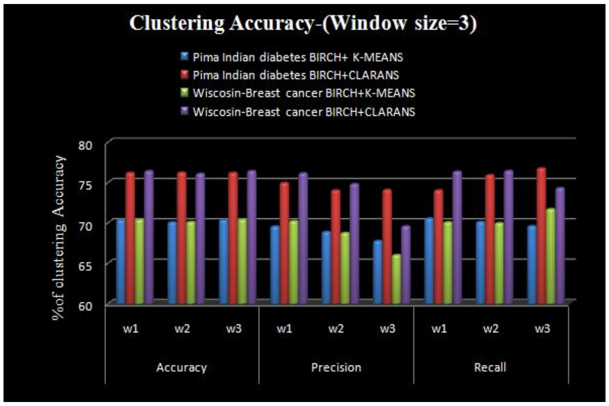

Fig 3: The Clustering Accuracy In Three Windows For Two Dataset

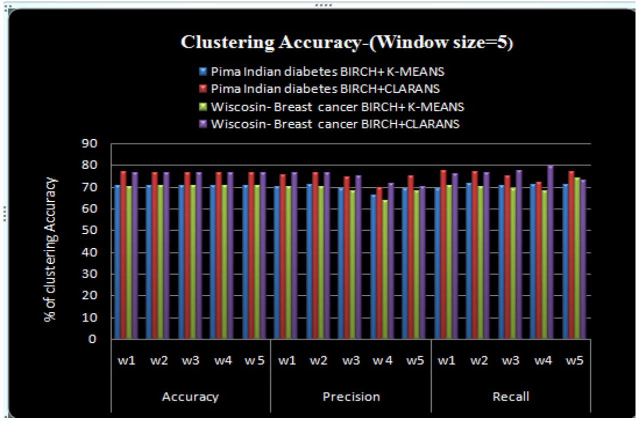

Fig 4: The Clustering Accuracy in Three Windows For Two Dataset

B. *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 K-MEANS and BIRCH 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.

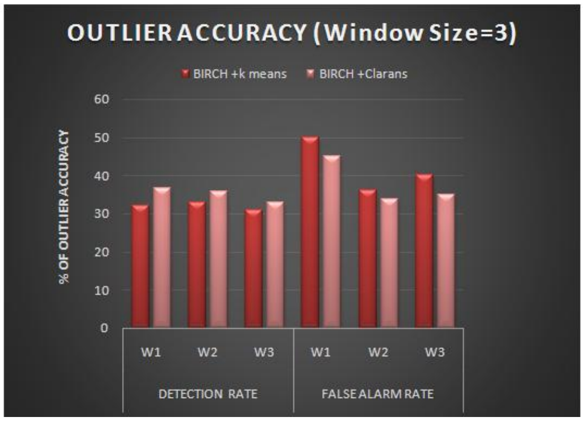

Fig 5: Detection Rate And False Alarm Rate In Three Windows-Pima Indian Diabetes

Table 4: Detection Rate And False Alarm Rate In Five Windows-Pima Indian Diabetes

Outlier accuracy	No. of windows	Birch +k means	Birch+clarans
Detection Rate	w1	32.40	33.82
	w2	39.80	41.28
	w3	33.50	36.69
	w4	24.57	25.78
	w5	33.60	34.00
False Alarm Rate	w1	44.44	33.82
	w2	50.00	41.28
	w3	51.12	36.69
	w4	27.00	25.78
	w5	44.00	34.00

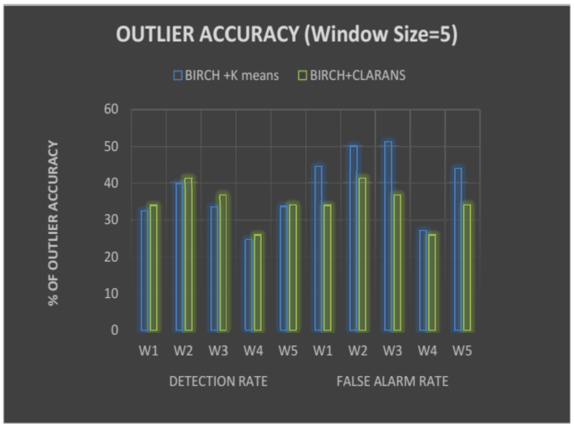

Fig 6: Detection Rate And False Alarm Rate In Five Windows-Pima Indian Diabetes

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

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 K-MEANS and Birch 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 +K-Means	Birch +Clarans
Detection Rate	w1	53.65	55.72
	w2	60.97	63.41
	w3	75.28	75.52
False Alarm Rate	w1	60.86	56.09
	w2	78.57	64.28
	w3	82.92	80.92

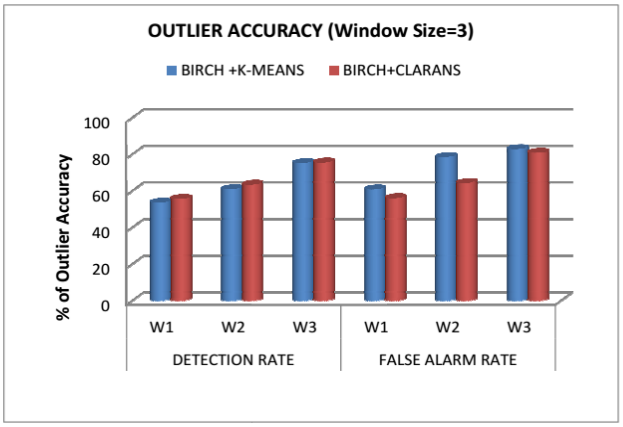

Fig 7: Detection Rate And False Alarm Rate In Five Windows - Breast Cancer (Wiscosin)

Table 6: Detection Rate And False Alarm Rate In Five Windows- Breast Cancer (Wiscosin)

Outlier Accuracy	**No. of Windows**	**Birch +K-Means**	**Birch+ Clarans**
Detection Rate	W1	54.08	56.02
	W2	51.06	53.60
	W3	61.61	64.00
	W4	75.75	77.70
	W5	73.47	75.47
False Alarm Rate	W1	59.52	52.30
	W2	63.63	62.50
	W3	73.80	72.00
	W4	83.33	78.00
	W5	73.17	72.00

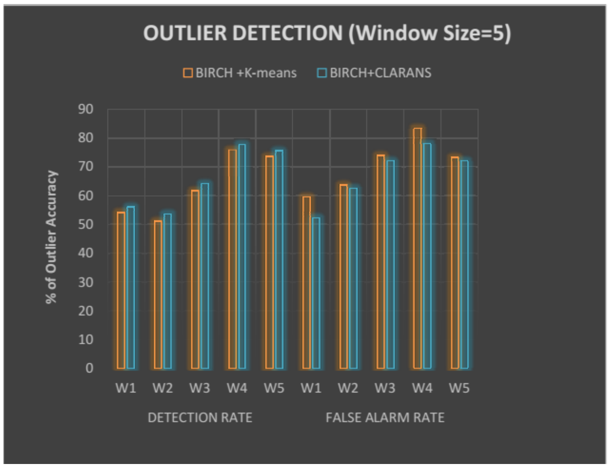

Fig 8: Detection Rate And False Alarm Rate In Five Windows - Breast Cancer (Wiscosin)

From the above graph, it is observed that Birch with clarans clustering algorithm performs better than Birch 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 birch with clarans clustering algorithm performs well because it contains high outlier detection accuracy when compared to Birch with K-means.

Table 1: The Clustering Accuracy In Three Windows For Two Dataset

Clustering Accuracy	**No. of windows**	**Pima Indian diabetes**		**Wiscosin- Breast cancer**	
		Birch+ K-Means	**Birch+Clarans**	**Birch+ K-Means**	**Birch+Clarans**
Accuracy	w1	70.31	76.17	70.38	76.39
	w2	70.03	76.2	70.08	76.06
	w3	70.31	76.17	70.38	76.39
Precision	w1	69.52	74.92	70.19	76.11
	w2	68.85	74	68.72	74.79
	w3	67.73	74.06	65.98	69.55
Recall	w1	70.52	74.01	70.04	76.33
	w2	70.09	75.89	69.94	76.41
	w3	69.57	76.73	71.67	74.29

Table 2: The Clustering Accuracy in Five Windows For Two Dataset

Clustering Accuracy	**No of Windows**	**Pima Indian Diabetes**		**Wiscosin Breast Cancer**	
		Birch+K Means	**Birch+Clarans**	**Birch+K Means**	**Birch+Clarans**
Accuracy	w1	70.27	70	76.62	76.42
	w2	70.32	70.21	76.12	76.59
	w3	70.32	70.21	76.12	76.59
	w4	70.32	70.21	76.12	76.59
	w5	70.39	70.5	76.31	76.25
Precision	w1	69.9	70	75.22	76.18
	w2	70.73	69.96	76.18	76.4
	w3	69.1	67.95	74.39	75.12
	w4	66.24	63.5	69.5	71.2
	w5	69.22	68.04	74.84	69.82
Recall	w1	68.91	70.26	77.31	75.87
	w2	71.24	69.96	76.88	76.46
	w3	70.29	69.07	74.88	77.3
	w4	70.81	68.17	72.14	79.21
	w5	70.8	73.65	76.64	72.76

V. CONCLUSION

In recent years, advances in hardware technology have allowed us to automatically record transactions of everyday life at a fast rate. It leads to large amounts of data which grow at an unbound less rate. Data streams are temporally ordered, massive, fast changing 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 analysed the clustering and outlier performance of BIRCH with CLARANS and BIRCH with K-Means clustering algorithm for detecting 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 clustering and outlier detection accuracy is more efficient in BIRCH with CLARANS clustering while compare to BIRCH 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, 2004, pp. 852-863.

[3] Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M.(2006), "A comparative study for outlier detection techniques in data mining", In Proc. 2006 IEEE Conf. Cybernetics and Intelligent Systems, pp. 1–6, Bangkok, Thailand.

[4] De Andrade Silva, J, Extending k-Means-Based Algorithms for Evolving Data Streams with Variable Number of Clusters. IEEE, Published in: Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on Volume: 2.

[5] Han.J and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2006.

[6] Hendrik Fichtenberger, Marc Gillé , Melanie Schmidt ,in Algorithms – ESA 2013 , Volume 8125, 2013, pp 481-492

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

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

[9] Madjid Khalilian, Norwati Mustapha, "Data Stream Clustering: Challenges and Issues", IMECS 2010.

[10] Mahnoosh kholghi, Mohammadreza Keyvanpour, "An analytical framework of data stream mining techniques based on challenges and requirements", IJEST, 2011.

[11] Pedro Pereira Rodrigues, João Gama, João Pedro Pedroso , "Hierarchical clustering of Time series Data Streams", IEEE Transactions on Knowledge and data engineering, May 2008 vol 20,no.5, pp. 615-627.

[12] T. Soni Madhulatha ,"overview of streaming-data algorithms, Department of Informatics, Alluri Institute of Management Sciences, Warangal, A.P. Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011.

[13]Sudipto Guha, Adam Meyerson , Nine Mishra and Rajeev Motwani, "Clustering Data Streams: Theory and practice," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, pp. 515-528, May/June 2003.

[14] Silvia Nittel , Kelvin T. Leung, "Parallelizing Clustering of Geo scientific Data Sets using Data Streams" Spatial Information Science & Engineering University of Maine &California.

[15] 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)

BIOGRAPHIES

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

Ms. P. Jothi 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.