International Journal of Grid and Distributed Computing
Vol. 9, No. 11 (2016), pp.285-296
http://dx.doi.org/10.14257/ijgdc.2016.9.11.24

A Hybrid Clustering Algorithm for Outlier Detection in Data Streams

S. Vijayarani[1] and P. Jothi[2]

[1]Assistant Professor, [2]Research Scholar, Department of Computer Science
Bharathiar University, Coimbatore
[1]vijimohan_2000@yahoo.com, [2]gifjot.88@gmail.com

Abstract

In current years, data streams have been gradually turn into most important research area in the field of computer science. Data streams are defined as fast, limitless, unbounded, river flow, continuous, stop less, massive, tremendous unremitting, immediate, stream flow, arrival of ordered and unordered data. Data streams are divided into two types, they are online and offline streams. Online data streams are mainly used for real world applications like face book, twitter, network traffic monitoring, intrusion detection and credit card processes. Offline data streams are mainly used for manipulating the information which is based on web log streams. In data streams, data size is extremely huge and potentially infinite and it is not possible to lay up all the data, so it leads to a mining challenge where shortage of limitations has occur in hardware and software. Data mining techniques such as clustering, load shedding, classification and frequent pattern mining are to be applied in data streams to get useful knowledge. But, the existing algorithms are not suitable for performing the data mining process in data streams; hence there is a need for new techniques and algorithms. The main objective of this research work is to perform the clustering process in data streams and detecting the outliers in data streams. New hybrid approach is proposed which combines the hierarchical clustering algorithm and partitioning clustering algorithm. In hierarchical clustering, CURE algorithm is used and enhanced (E-CURE) and in partitioning clustering, CLARANS algorithm is used and enhanced (E-CLARANS). In this research work, the two algorithms E-CURE and E-CLARANS are combined (Hybrid) for performing a clustering process and finding the outliers in data streams. The performance of this hybrid clustering algorithm is compared with the existing hybrid clustering algorithms namely BIRCH with CLARANS and CURE with CLARANS. The performance factors used in this analysis are clustering accuracy and outlier detection accuracy. By analyzing the experimental results, it is observed that the proposed hybrid clustering approach E-CURE with E-CLARANS performance is more accurate than the existing hybrid clustering algorithms.

Keywords*: Data streams, Clustering, Outlier detection, CURE, BIRCH, CLARANS, E-CURE, E-CLARANS*

1. Introduction

Data mining is wide spread studied field of research area, where information are interesting, non-trivial, hidden and potentially useful patterns or knowledge from huge amount of data (Margaret H. Dunham, 2006). Significant problems exists in huge databases are invalid data, data redundancy, missing data, distorted data, unbounded data etc. One of the most important problems occurs in data mining research is increase in dimensionality of data and it gives rise to a number of new computational challenges. In recent years, we have monitored that enormous research activity motivated by the explosion of data collected and transferred in the format of data streams.

ISSN: 2005-4262 IJGDC

Data streams handles bulk amount of data being generated every day and to be updated in timely approach. Some of the application areas of data streams are astronomy, telecommunications industry, business companies, commercial banks, stock exchanges and news organizations. In addition to this, the evolving nature of data streams are defined in most of the applications such as detection of intrusions or abnormal behavior in wireless sensor networks, meteorological research, super markets and medical field (Muthukrishnan, 2003). The characteristics of data stream as well as its elements also evolve over time. This type of data streams is referred to as temporal locality and adds an inherent temporal component to the data stream mining process. Stream elements should thus be analyzed in a time-aware manner to accommodate the changes in stream characteristics.

Data stream properties are unbounded and can generate an infinite amount of data and also the high storage rate, it is impractical to accumulate an entire data streams or to scan through it multiple times. Hence, it makes many challenges in storage in announcement capabilities of computational methods. Because of high volume and speedy arrival of upcoming data, it is desirable to use partially routine interface techniques to extract embedded knowledge from data. Different types of data stream techniques are data stream clustering, data stream classification, sliding window in data streams, frequent pattern in data streams, network intrusion and detection in data streams, association in data streams, concept drift in data streams, and outlier detection in data streams (Aggarwal. C.C *et.al.*, 2004).

Clustering is an outstanding task in mining data streams, which groups related objects into a cluster and the clustering techniques are mainly categorized as partitioning and hierarchical. Data is organized into several groups in which each group contain similar objects. Number of clustering algorithms has been introduced in recent years for data streams. Data stream clustering can be considered as unsupervised learning problem, it deals with finding a structure in a collection of unlabelled data (Aggarwal. *et.al.*, 2004).

Hierarchical clustering algorithms recursively nested clusters either in agglomerative method by starting with each data point in its own cluster and merging with similar pair of clusters successively to form a cluster hierarchy or divisive top-down approach method. This method is defined with all the data points in one cluster and recursively, dividing each cluster into smaller clusters. The most popular agglomerative clustering procedures includes the single linkage and complete linkage methods. Single linkage methods are called as nearest neighbor. In this method, the distance between two clusters are corresponds to the shortest distance between any two members in the two clusters. Next the complete linkage is termed as furthest neighbor. This is the oppositional approach to single linkage and it assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters. In complete linkage there are two approaches, the first approach is regular linkage, it is nothing but the distance between couple of clusters is termed as the average distance between all pairs of the two clusters members. Second approach is termed as centroid, the geometric center centroid of each cluster is computed, and the distance between the two clusters is equal to the distance between the two centroids.

Another type of clustering approach is partitioning clustering. The major idea of partitioning based algorithm is, first to partition the data space and then prune partitions as soon as it determine the outliers. Partitioning based algorithm is subject to the pre-processing step in which data space is split into cells and data partitions, together with the minimum bounding of data partitions. Given a database of objects, a partition clustering algorithm constructs k partitions of n data, where each cluster optimizes a clustering decisive factor, such as the minimization of the sum of squared distance from the mean within each cluster. Partitioning clustering algorithms tries to locally improve the condition and also it computes the values of the similarity or distance, order the results, and select the best one that optimizes the criterion.

Data stream clustering methodologies are highly helpful to detect outliers and outlier detection is one of the important data mining tasks and it is also called as outlier mining. There are many algorithms for outlier detection (Hodge.V *et. al*, 2003) in static and stored data sets which are based on a variety of approaches like distance and density based outlier detection, nearest neighbor based outlier detection method and clustering based outlier method. The clustering-based methods assume that the normal data objects belong to large and crowded clusters, while outliers belongs to sparse or small clusters or do not belong to any clusters. The clustering based methods are used efficiently to detect outliers by examining the relationship between objects and clusters.

Outlier detection can also be defined as finding the unrealistic data or static data (Barnett.V *et. al*, 1994) that is of no use with rest of the data set. Depending upon different application domains these abnormal samples are often referred to as outliers, peculiarities anomalies and observations in discordant, errors, faults, exceptions, defects, irregularity, noise, damage, surprise, novelty and impurity. As the technology advances, lot of applications will produce time sequenced, vertiginous and massive data streams, such as network flow monitor, E-commerce, and wireless communication. These are considered to be high dimension domain; hence the possibilities of outliers are very common in these areas. With the growth of increase in the size of the dataset, the process of determining outliers becomes more tedious and complex. Efficient detection of outliers diminishes the risk of making poor decisions based on untrue data, and aids in preventing, identifying, and repairing the effects of spiteful or damaged behavior. Hence, clustering based outlier detection in data stream is highly required.

The remaining section of this paper is organized as follows; Section 2 reviews the related works. Section 3 describes the proposed and the existing hybrid clustering algorithms in data streams. Experimental results are discussed in Section 4 and conclusions are given in Section 5.

2. Related Works

DikshaUpadhyay, *et.al.*, [2010] discussed about data stream mining, the streamed data can be high dimensional, various dimension reduction techniques were applied prior to perform clustering. The comparative analysis of various stream mining procedures and dimension reduction techniques are discussed.

Hossein Moradi Koupaie, *et.al.*, [2013] presented cluster based outlier detection in data stream. In this research paper, author used K-means algorithm for entering a data into specified size of window and also they report each and every data as outlier and then storing the data. By using K-means algorithm, they identified small cluster, which is faraway to other clusters and termed as outlier.

Shifei Ding, *et.al.*, [2013] compared typical data stream clustering algorithms which are proposed in recent years. Algorithms used for analysis are BIRCH algorithm, Local Search algorithm, Stream algorithm and CluStream algorithm. The authors summarized about the latest research achievements in this field and introduced some new strategies to deal with outliers and noise data.

Luis Torgo, *et.al.*, [2010] discussed a methodology for the application of hierarchical clustering methods to perform the task of outlier detection. This methodology is tested on the official statistics data and the foreign trade transactions data. Authors used an outlier ranking method and this method has achieved better results on this particular application.

Sudipto Guha, *et.al.*, [2003] discussed a clustering algorithm called CURE and it is used for detecting outliers. The combination of random sampling and partitioning and the experimental results confirmed that the quality of clusters produced by CURE is much better than those found by existing algorithms. Moreover, authors expressed the partitioning and random sampling enables CURE to not only perform existing algorithms but also to scale well for large databases without sacrificing the quality of cluster.

Ren, J.W, *et.al.*, [2009] suggested an approach for heterogeneous data streams, which divided the stream in chunks. Following that, each chunk is grouped and the same clusters are kept in matching cluster positions. Quantity of contiguous cluster situations and the illustration degree are calculated to create the final outlier references such as potential outliers. The experimental findings had given better scalability and higher investigation accuracy.

S.D.Pachgade, *et.al.*, [2012] proposed an outlier detection algorithm which is used to group the data in to number of clusters. Due to diminution of dataset size, the computation time reduced significantly in nature. By using threshold value from the user, outliers are detected from cluster; they have proved the hybrid approach of clustering takes less computation time.

S.Vijayarani, *et.al.*, [2013] proposed a new hybrid approach which combines the hierarchical and partitioning clustering algorithms. In this hybrid approach, BIRCH (Hierarchical clustering) algorithm and CLARANS (Partitioning clustering) algorithm are used for outlier detection in data streams. This performance is compared with the existing algorithm named as BIRCH with K-Means. By measuring the clustering accuracy and outlier accuracy, the outlier detection performance is better in hybrid approach, *i.e.*, BIRCH with CLARANS clustering algorithm.

S.Vijayarani, *et.al.*, [2013] discussed the performance of CURE with CLARANS and CURE with K-Means clustering algorithms. CURE with CLARANS algorithm belongs to the hybrid type which combines the hierarchical clustering and partitioning clustering algorithms. Based on the performance factors, authors concluded that the proposed hybrid algorithm has produced good results than the existing algorithm, *i.e.*, CURE with K-Means. S.Vijayarani, *et.al.*, [2013] have compared the results of two hybrid approaches namely BIRCH with CLARANS and CURE with CLARANS. By measuring the clustering accuracy and outlier accuracy, the performance of clustering and outlier detection is better in CURE with CLARANS clustering algorithms.

3. Problem Objective and Contribution

A data stream is an ordered sequence of objects X_1, X_2, $X_{3,\ldots,}X_n$. Data stream clustering methods are very efficient to detect outliers. Intuitively, an outlier is an object that belongs to a small and remote cluster, or does not belong to any cluster. The main work of this research is illustrated as clustering based outlier detection in data streams. From the literature, it has been seen that most of the research work is implemented using hierarchical and partitioning clustering algorithms alone. Both the existing clustering algorithms in data streams not detected the outliers in efficient manner because the cluster centroid values are not updatable.

To overcome this problem, this research work has proposed new hybrid algorithms for performing the clustering process in data streams and also detecting the outliers in data streams. New hybrid approach is proposed which combines the hierarchical clustering algorithm and partitioning clustering algorithm. In hierarchical clustering, CURE algorithm is used and enhanced (E-CURE) and in partitioning clustering, CLARANS algorithm is used and enhanced (E-CLARANS). The performances of these algorithms are compared with the existing algorithms. The clustering accuracy and outlier detection accuracy is efficient in newly proposed hybrid algorithm E-CURE with E-CLARANS.

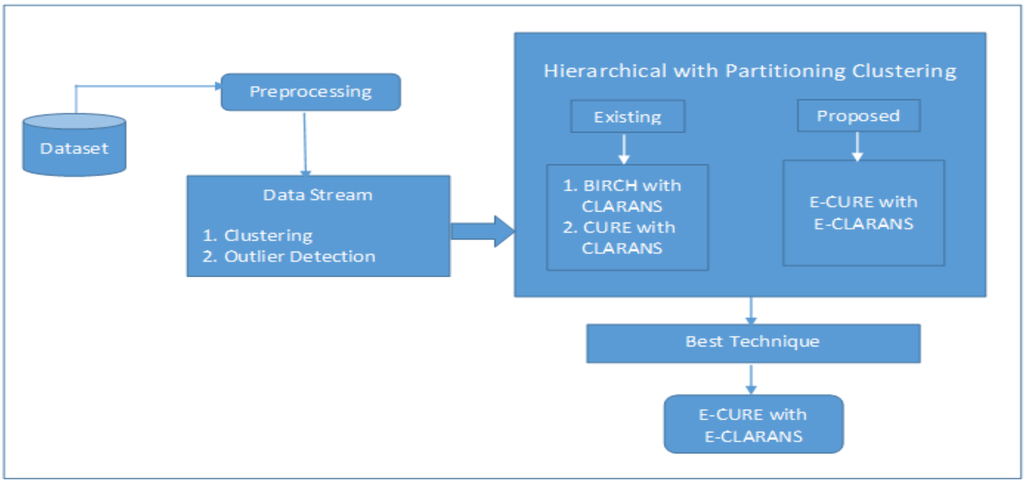

Figure 1. System Architecture

3.1 Data Set

In order to perform the clustering and outlier process in data streams, this research work used health data set, i.e. Pima Indian diabetes data set which contains 8 attributes and 768 instances. This dataset is collected from UCI repository. In data streams the data arrival is continuous, to perform data mining tasks the arrived data is divided into different partitions. With this assumption, we divide the Pima Indian diabetes data into several chunks. Each chunk has the same number of instances and numbers of chunks used in this work are three and five.

3.2 Preprocessing

The real world health data sets are highly vulnerable to missing and inconsistent data. Such datasets are of low quality and leads to low quality mining results because quality of mining results depends upon the quality of data. Pre-processing techniques includes handling missing data, normalization and aggregation. In this research work we have used data pre-processing techniques to improve data quality. In order to find the missing value Euclidean distance method is used and normalization is used for avoiding the duplication of data items.

3.3 Outlier Detection

3.3.1 BIRCH with CLARANS: In this method hierarchical and partitioning clustering algorithm is combined and it is called hierarchical with partitioning clustering algorithm in data streams. First the data items are divided into separate chunks, after that consider a original data sample S which has X_1, X_2,..X_n data items and n is number of data points dp. For the original data point (dp) the centroid c, radius r, diameter d of each data point are identified. By using Euclidean distance the distance between the data points and the centroids are calculated. By using manhattan distance, finding the present value of data points, and determines how these data points are closer to the group of individual point D_1, D_2 and D_3…D_n. Next step of this process calculates, least square and the sum of the n1 data point is closer to other data point n2, in this we have to find the least square (LS) and sum of square (SS) of data point using clustering feature is cf_1 and cf_2. For updating the data point (dp), calculate the minimum cost average value (c1+c2)/2. After performing the cluster updation process, the next step is to verify whether the threshold value is less than the centroid value then it groups the updated clusters. The pseudo code of this algorithm is given in Table 1.

Table 1. BIRCH with CLARANS

Input: Represent the data sample, S is X_1, X_2,…X_n

Output: Data point values are clustered and the outliers are detected.

Procedure:

1. Consider a original data sample S which has X_1, X_2,..X_n data items and n is number of data points (dp).
2. From the original data point (dp) find the centroid, radius r, diameter d of each data point.
3. By using Euclidean distance find the distance of each data point which is closer to the centroid.
4. By using Manhattan distance, find the present value of data point (dp), which is closer to other group of individual point D_1, D_2, D_3.. D_n.
5. The n1 data point is closer to other data point n2, find the least square (LS), and also find the sum of square (SS) of dp and also find the data points clustering feature cf_1, cf_2.
6. Verify the threshold value; if it is less than the centroid value then it groups the updated clusters.
7. Else, go to the step 3.
8. Return, Best cluster and detect the outliers efficiently.

3.3.2 CURE with CLARANS: First the data items are divided into chunks. From the database(s) we have to select the data point (dp) and then the data items are partition with size of (s/p), along with maximum neighbor and partitioned of k=3. Then the minimum cost for each data point (dp) is identified the neighbor value, and it follows the condition i=1and j=1. For each data point value, consider and set the current arbitrary node is Gn1rs=$\sum^{n}_{i=1}$ [dp*min cost]/n along with n:k. Next step of this process calculates the cluster centroid values d(x,y) which is calculated using Euclidean distance. For updating the data points, two conditions are to be verified. First the cluster centroid value has to check along with minimum cost value (dmin) and threshold value (α), and the second condition is cluster centroid value has to check along with arbitrary node of Gn_1rs and the minimum cost of each data point value, groups the updated clusters. Then the Figure 4.4.2 represented the entire process.

Table 2. CURE with CLARANS

Input: Represent the database(s) into data point (dp), partition size=s/p, with
Maxneighbor k=3.

Output: Data point values are clustered &the outliers are detected.

Procedure:

1. Consider the sample input database(S) is X_1, X_2…X_n into data point (dp) partition size=s/p, with Max neighbor &k=3.
2. Find the minimum cost for (i to 1000) with the average of neighbor value and also satisfy the condition of i=1 and j=1.
3. If S has a lower cost, set current to S, are increment j by 1 and If j is maximum neighbor, when j > max neighbor, compare the cost of current with minimum cost. Then set the current of an arbitrary node is $Gn_1rs=\sum^{n}_{i=1}$ [dp *min cost]/n and also consider n: k.
4. Calculate the cluster centroid values d(x,y) using distance function (i.e.) Euclidean distance.
5. If the cluster centroid value ≤ dmin (i.e.) minimum cost /α (i.e.) Threshold value is checked along with $Gn_1rs \leq$ min cost.
6. Partially updated the cluster data and return outlier data.
7. Else, Repeat the step 4 to step 5 up to best dmin, and Gn_1rs are found to other samples.
8. Return, Best cluster and detect the outliers efficiently.

3.3.3 E-CURE with E-CLARANS: The important and enhanced step in E-CURE with E-CLARANS algorithm is, first the data items are divided into chunks, after that from the database(s) we have to select the data point (dp) and then the data items are partitioned with size of (s/p), along with maximum neighbor and partition of k=3.Then the minimum cost for each data point (dp) of neighbor value it has to follows the condition of i=1and j=1. Then for each data point value, we have to set the current of an arbitrary node is Gn1rs=$\sum^{n}_{i=1}$ [dp*min cost]/n along with n:k. Next step of this process calculates the cluster centroid values d(x,y) calculated using Euclidean distance. After calculating the centroid values using Euclidean distance, the kullback libber divergence probability function is used to update and group the clustered data in a relevant manner.

$$D_{KL}(P\,\|\,Q)=\sum_i \ln\left(\frac{P(i)}{Q(i)}\right)P(i)$$

Where P(i) is the probability of the d(x,y) and Q(i) is the probability of the another data point d(x,y). First the cluster centroid value has to check along with minimum cost value (dmin) and threshold value (α), and the second condition is cluster centroid value has to check along with arbitrary node of Gn_1rs and the minimum cost of each data point value, and then it groups the updated clusters.

Table 3. E-CURE with E-CLARANS

```
Input:  Represent the database(S) into data point (dp), partition of size=s/p, with Max neighbor
         and k=3.

Output: Data point values are clustered and the outliers are detected

Procedure:

1. Consider the sample input database(S) is $X_1,X_2,\ldots X_n$ into data point (dp) partition size=s/p, with Max
    neighbor and k=3.
2. Find the minimum cost for i=1 for (i to 1000) with the average of neighbor value and also
    satisfy the condition of i=1 and j=1.
    2.1 If S has a lower cost, set current to S, are increment j by 1 and If j is max neighbor,
         when j >maxneighbor,
    2.2 Compare the cost of current with mincost.
3. Then set the current of an arbitrary node is $Gn_1rs=\Sigma^n{}_{i=1}[dp*\text{min cost}]/n$ and also consider n: k.
4. Calculate the cluster centroid values d(x,y) using distance function (i.e.) Euclidean distance.
    4.1 The micro cluster are used to calculate by kullback libber divergence probability
         function $D_{KL}(P\,||\,Q)=\sum_i \ln\left(\frac{P(i)}{Q(i)}\right)P(i)$ helps to update and group the clustered data.
    4.2 P(i) is the probability of the d(x,y) Q(i) is the probability of the another data point d(x,y).
5. If the cluster centroid value ≤ dmin (i.e.) minimum cost /α (i.e.) Threshold value and also the value is
    checked along with $Gn_1rs \leq$ min cost.
6. Partially updated the cluster data and return outlier data.
7. Else, Repeat the step 4 to step 5 up to best dmin, and $Gn_1rs$ are found to other samples.
8. Return, Best cluster and detect the outliers efficiently.
```

4. Performance Evaluation

To test the proposed and existing methods, a series of performance studies were conducted. The evaluation was performed on PC Intel Pentium processor, 2GB RAM, OS Windows 7 Ultimate 32-bit. Algorithms are implemented in MATLAB7.10 (R2010a). For evaluation; two performance factors such as clustering accuracy outlier detection accuracy are used. The clustering accuracy and the outlier detection accuracy is calculated by using two measures they are detection rate and false alarm rate.

4.1 Clustering Accuracy

From Figure 2, it is observed that newly proposed E-CURE with E- CLARANS algorithm's performance is better than BIRCH with CLARANS and CURE with CLARANS clustering algorithms.

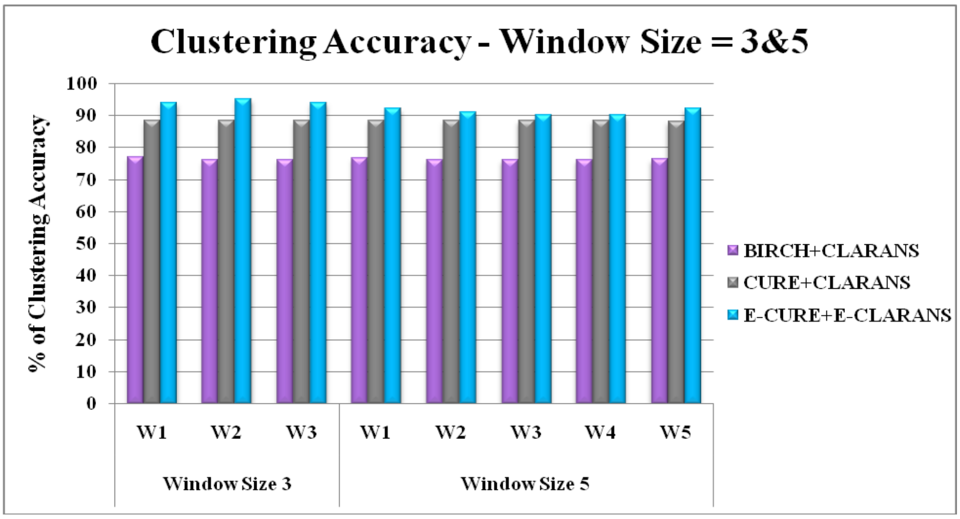

Figure 2. Clustering Accuracy

4.2. Outlier Accuracy

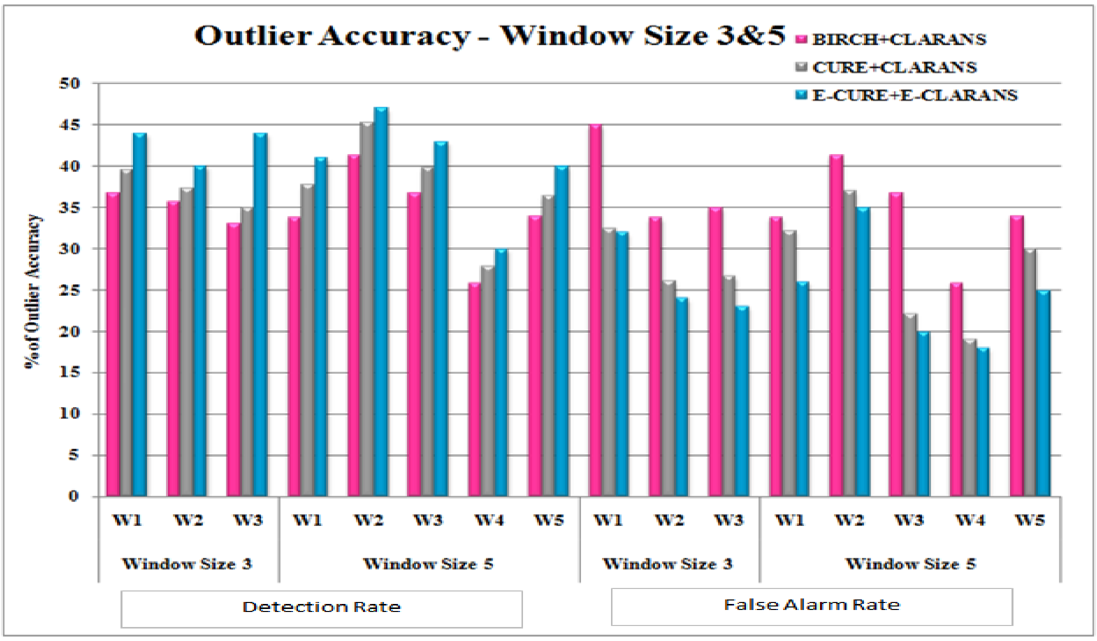

Figure 3. Outlier Accuracy

From the above Figure-3 it is observed that newly proposed hierarchical with partitioning hybrid clustering algorithm i.e. E-CURE with E-CLARANS clustering algorithms performance is better than existing algorithms.

4. Conclusion

Data stream is one of the data mining or knowledge discovery task, data streams are rapid and unlimited arrival of ordered and unordered data. Using data mining techniques in data streams gives several research issues and challenges. This research work used hierarchical and partitioning clustering algorithms alone. The proposed hybrid clustering, the centriod values are updated to form better cluster and detected the outliers efficiently by using kullback libber divergence probability function. The performance of this hybrid clustering algorithm is compared with the existing hybrid clustering approaches namely BIRCH with CLARANS and CURE with CLARANS. The performance factors used in this analysis are clustering accuracy and outlier detection accuracy. By examining the experimental results, it is observed that the proposed hybrid clustering approach E-CURE with E-CLARANS has produced more accurate results than the existing hybrid clustering algorithm. In future, the performances of these algorithms are to be tested with different data sets.

References

[1] Aggarwal.C.C, J. Han, J. Wang, and P. S. Yu,(2003)"A framework for clustering evolving data streams," In Proc. of VLDB, pages 81-92.

[2] Aggarwal.C.C, J. Han, J. Wang, and P. S. Yu,(2004)"A framework for projected clustering of high dimensional data streams," In Proc. of VLD pages 852-863.

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

[4] Chandrika.J, Dr. K.R. Ananda Kumar, (March 2012) "Dynamic Clustering Of High Speed Data Streams", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1.

[5] D. Hawkins,(1980) "Identification of outliers-Monographs on statistics and applied probability", First edition, pages-188, Springer.

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

[7] DikshaUpadhyay, Susheel Jain, Anurag Jain,(2010)"Comparative Analysis of Various Data Stream Mining Procedures and Various Dimension Reduction Techniques" ,International Journal of Advanced Research in Computer Scienc, IJARCS All Rights Reserved 179 ISSN No. 0976-5697.

[8] Elahi, M. KunLi, Nisar,W.XinjieLv, HonganWang, (2002), "Fuzzy Systems and

[9] Knowledge Discovery", Fifth International Conference on Vol.5, and Vol .3, pp. 23-27.

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

[11] Hodge.V and J. Austin, (2003) "A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, Vol. 22, pp. 85-126.

[12] HosseinMoradiKoupaie,Suhaimi Ibrahim,JavadHosseinkhani,(2013"Outlier Detection in Stream Data by Clustering Method", International Journal of Advanced Computer Science and Information Technology (IJACSIT)Vol. 2, No. 3, 2013, Page: 25-34, ISSN: 2296-1739.

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

[14] JurgenBeringer and EykeHullermeier, (2006), "Online clustering of parallel data streams," Data Knowl. Eng, 58:180–204.

[15] LiadanO'Chalaghan, Nina Mishra, Adam Meyerson, SudiptoGuha, Rajeev Motwani, (2002), "Streaming data algorithms for high quality clustering", Proceedings of the 18th international conference on data engineering,pp.685– 694.

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

[17] MaciejJaworski, PiotrDuda, Lena Pietruczuk,(2012) "Artificial Intelligence and Soft Computing",Lecture Notes in Computer Science Volume,7268, pp 82-91.

[18] Margaret H.Dunham,(2006) "Data Mining Introductory and Advanced Topics".

[19] Muthukrishnan, S. (2003), "Data streams: algorithms and applications",In Proc. 2003 , Annual ACMSIAM Symp. Discrete Algorithms (SODA'03), pages 413–413, Baltimore, MD,Jan. 2003.

[20] ProdipHore, Lawrence O. Hall, and Dmitry B. Goldgof, (2008)"Creating Streaming Iterative Soft ClusteringAlgorithms, IEEE, Page(s):1 –5, E-ISBN: 978-1-4244-2352-1,Print ISBN:978-1-4244-2351-4.

[21] R. J. Hathaway and J. C. Bezdek,(2006) "Extending Fuzzy and Probabilistic Clustering to Very Large Data Sets," Journal of Computational Statistics and Data Analysis, pages 215-234.

[22] Ramaswamy S., Rastogi R., Kyuseok S, (2000) ," Efficient Algorithms for Mining Outliers from Large Data Sets, Proc. ACM SIDMOD Int. Conf. on Management of Data.

[23] Raymond T. Ng and J. Han, "Efficient and Effective clustering method for spatial data mining", VLDB'94.

[24] Ren, J. W., Qunhui ; Zhang, Jia ; Hu, Changzhen,(2009)"Efficientoutlier detection algorithm for heterogeneous data streams. 6th International Conference on Fuzzy Systems and Knowledge Discovery,FSKD.

[25] S. D. Pachgade, S. S. Dhande, (June 2012), "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277, Volume 2, Issue 6.

[26] S. Vijayarani, P. Jothi, September (2013) "An Efficient Clustering Algorithm for Outlier Detection in Data Streams", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 9, ISSN (Print) : 2319-5940,ISSN (Online) : 2278-1021.

[27] S. Vijayarani, P. Jothi, October (2013), "Comparative Analysis Of Clustering Algorithms For Outlier Detection In Data Streams", International Journal of Engineering Sciences & Research Technology.

[28] S. Vijayarani, P. Jothi, November (2013), "A New Approach for Detecting Outliers in Data Streams", International journal of engineering sciences & research Technology, ISSN: 2277-9655, Pg no: [3128-3133].

[29] S. Vijayarani, P. Jothi, October (2013)"Detecting Outliers in Data streams using Clustering Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 8,. ISSN (Online): 2320-9801, ISSN (Print): 2320-9798.

[30] Sharma, M. Toshniwal, D,(2012) " Pre-clustering algorithm for anomaly detection and clustering that uses variable size buckets", Published inRecent Advances in Information Technology (RAIT), 1st International Conference on 15-17.

[31] Shifei Ding, Fulin Wu, Jun Qian, HongjieJia, (2013), "Research on data stream clustering algorithms"inArtificial Intelligence Review, Springer.

[32] Shruti Aggrwal1, Prabhdip Kaur,(2013) "Survey of Partition Based Clustering Algorithm Used for Outlier Detection", International Journal For Advance Research In Engineering And Technology.

[33] SudiptoGuha, Adam Meyerson, Nine Mishraand Rajeev Motwani, May/June (2003) "Clustering Data Streams:Theory and practice," IEEE Transactions onKnowledge and Data Engineering, vol. 15, no.3, pp. 515-528.

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

[35] Thakran.Y,Toshniwal.D,(2012)"Unsupervised outlier detection in streaming data using weighted clustering", Intelligent Systems Design and Applications (ISDA).

[36] Yang.J, 2003 "Dynamic clustering of evolving streams with a single pass," In Proc. of ICDE, pages 695-697.

[37] Yi-honglu, Yan huang, 2005 "Mining DataStreams Using Clustering", Proceedings of the Fourth International Conference on Machine Learning andCybernetics,vol.4, pp. 18-21.

[38] Zhang, T., Raghu, R., Miron, 1996.L.BIRCH: An Efficient Data ClusteringMethod for Very Large Databases, ACM SIGMOD Record, vol. 25(2), 103-11.

Authors

S. Vijayarani, MCA, M.Phil, Ph.D working as Assistant Professor in the Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues, text mining, web mining, information retrieval, data streams and big data. She has authored a book and published more than 70 research papers in the international journals and also presented papers in international and national conferences.

P. Jothi has completed M.Sc and M.Phil in Computer Science in the Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining and Data Streams. She has published papers in international journals and conferences.