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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Partitioning Clustering Algorithms for Data Stream Outlier Detection

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: Recently many researchers have focused on mining data streams and they proposed many techniques and algorithms for data streams. They are data stream classification, data stream clustering, and data stream frequent pattern items and so on. 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. The main objective of this research work is to perform the clustering process and detecting the outliers in data streams. In this research work, two partitioning clustering algorithms namely CLARANS and E-CLARANS (Enhanced Clarans) are used for clustering and detecting the outliers in data streams. Two performance factors such as clustering accuracy and outlier detection accuracy are used for observation. By examining the experimental results, it is observed that the proposed E-CLARANS clustering algorithm performance is more accurate than the existing algorithm CLARANS.

KEYWORDS: Data stream, Data stream clustering, Outlier detection, CLARANS, E-CLARANS

I. Introduction

A data stream is an unremitting, immediate, stream flow of sequence of items and it is not possible to control the order in which data item arrive, or not possible to store these entire data items. Some of the applications of areas in which data streams generated are sensor networks, traffic management, call detail records, blogging and twitter posts [1].Due to be short of resources where as this type of huge data, the modern data mining systems are not sufficient and equipped to deal with them. Data stream clustering is a well-known task in mining data stream, clustering is known as grouping related objects into a cluster. With the help of data stream clustering method [2], we can detect the outliers, and the outlier is nothing but it is an object that does not fulfil with the behaviour of normal data objects. Applications of outlier detection are web logs, fraud detection and click streams, communication of telecoms and web document. Clustering based outlier mining [14] methods are called as unsupervised in nature and its main objective is to find the outlier from the data stream using partitioning cluster based method. The object which does not belong to any cluster or belongs to a small cluster is affirmed as outlier, and the outlier detection process highly depends upon the clustering technique.

The remaining section of this paper is organized in the following way. Section 2 illustrates the review of literature. Section 3 describes how the CLARANS and E-CLARANS clustering algorithms are used to detect outliers in data streams. Section 4 discussed about the experimental results and Conclusions are given in Section 5.

II. Related work

In this paper [8] the author presented a clustering algorithm called CLARANS which is based on randomize search. The authors had developed two spatial data mining algorithms SD (CLARANS) and NSD (CLARANS). The experimental results and analysis indicated that both algorithms are effective, and can lead to discoveries that are difficult to obtain with existing spatial data mining algorithms. Finally, their experimental results showed that CLARANS is more efficient than existing clustering methods.

The paper [4**]** discussed a literature of several clustering procedures and multivariate outlier procedures. And also the features of multivariate outliers are also discussed, as well as the applications are highlighted in this survey. Finally the authors discussed about further research challenges on multivariate outliers.

In this paper [5] authors conversed about partitioning clustering based outlier detection for data streams. In this each and every data are entered into a specify size of window, and also they reported each and every data as outlier and also store the data. By using K means algorithm, they have been found small cluster, which is faraway to other clusters and termed as outlier.

In this paper [9] authors compared two partitioning clustering approaches namely CLARANS and FUZZY C MEANS. By measuring the clustering accuracy and outlier accuracy, the performance of clustering and outlier detection is better in CLARANS clustering algorithms.

III. METHODOLOGY

In data stream, the clustering technique is applied for grouping the data items and also detecting the outliers. Clustering and Outlier detection are most important problems in data streams. The main objective of this research work is to analyse the performance of the two partitioning clustering algorithms namely CLARANS and E-CLARANS for detecting the outliers. The system architecture of the research work is as follows as

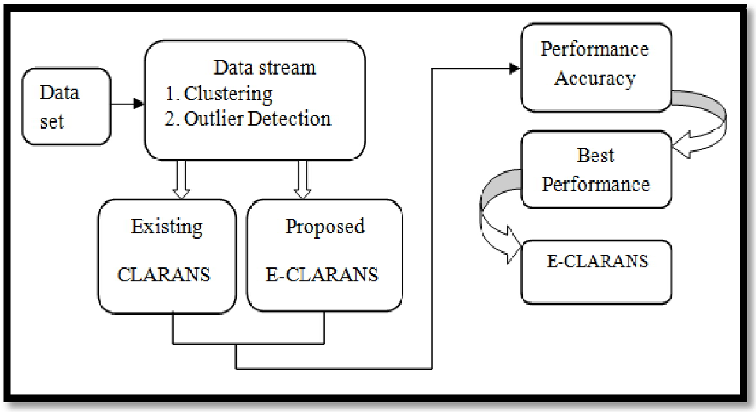

Figure 1: System Architecture

A. DATASET

Dataset which have been used in this research work is Pima Indian data set; it contains 768 instances and 8 attributes. This dataset is taken from UCI machine learning repository [3]. Data stream is an abundant flawless sequence of data and it is not possible to store the complete data stream, due to this reason we divide the data into chunks of same size in different windows.

B. CLUSTERING

Cluster analysis is used in a various number of applications; they are stock market analysis, data analysis, image processing and financial market analysis 14]. In data streams the clustering is one of the sub-process areas which are used to group the objects as well as it is used to detect the outliers efficiently and also clustering is one of the unsupervised action in data streams. The data stream clustering are different types of approaches they are distance based, grid based, partition based, hierarchical based and so on.

C. OUTLIER DETECTION

Outlier detection over streaming data is active research area from data stream mining that aims to detect object which have different [5] behaviour, exceptional than normal object. An outlier is an item that is notably unrelated or incompatible to other data object whereas weblogs click stream telecommunication, fraud detection, documents of web are the application areas of outlier detection in data streams. The other specified names of outlier detection are termed as noise, anomalies, indifferent, not catchable to the related object, and unknown. The clustering based outlier detection is a best technique to manage this problem. For our research we have used partitioning cluster based outlier detection algorithms CLARANS and E-CLARANS.

D.CLARANS

This method involves partitioning clustering algorithm in data streams [9]. First the data's are splitted into chunks of same size in different windows, after that consider each database(s) into data point (dp), partition of size=s/p, along with max neighbor of k=3. Then the minimum cost for each data point (dp) identifies the neighbor value, and it follows the condition i=1and j=1.Then the distance for each data point is calculated and also choose maximum distance (n) for each data points, if (s) has a lower cost, set current to(s), are increment j by 1.when j > max neighbor, compare the cost of current with minimum cost. If the cost value is less than (<) min cost, set minimum cost to current of cost value. Finally group the cluster, in order to satisfy the threshold value≤ min cost. Finally nodes are clustered and outliers are identified.

Algorithm 1: CLARANS

Input: Represent the database(S) into data point (dp) Partition size=S/P, with Max neighbor &K=3.
Output: Data point values are clustered & the outliers are detected
Procedure
1. Input the parameters num local and max neighbor. Initialize i to 1, for mincost to a large number.
2. Calculating the distance between each data points and also choose n of max distance at each of the data points.
3. Consider a random neighbor S of current, and calculate the cost differential of the two nodes.
4. If S has a lower cost, set current to S, increment j by 1 and If j is max neighbor, when j > max neighbor, compare the cost of current with min cost.
5. If the cost value is less than (<) min cost, set min cost to the cost of current value and set best node value to current node.
6. Finally group the cluster, to satisfy the threshold value≤ min cost. Then nodes are updated & cluster data and return outliers.
7. Else, Repeat the step 3 to step 5 up to best minimum cost (dmin), are found to other samples.
8. Return, Best cluster and detect the outliers efficiently.

E. E-CLARANS (Enhanced Clarans)

In E-CLARANS, first the data are splitted into chunks of same size in different windows, after that consider each database(S) into data point (dp), partition of size=s/p, along with max neighbor 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.Then calculate the distance for each data points and also choose maximum distance (n) for each data points. Set current to an arbitrary node in n: k, for each data point we have to set j to 1along with a random neighbor (s) of current value, and also calculate the cost differential of the two nodes. If (s) has a lower cost, set current to(s) is increment j by 1. when j > max neighbor, compare the cost of current with minimum cost. If the cost value is less than (<) min cost, set minimum cost to current of cost value. Finally group the cluster, in order to satisfy the threshold value≤ min cost. Then lastly nodes are clustered and detect outliers.

Algorithm 2: E-CLARANS

```
Input:  Represent the database(S) into data point (dp) Partition size=S/P, with Max neighbor &K=3.
Output: Data point values are clustered & the outliers are detected
Procedure
1. Input the parameters num local and max neighbor. Initialize i to 1, for mincost to a large number.
2. Calculating the distance between each data points and also choose n of max distance at each of the data points.
3. Set current to an arbitrary node in n: k and Set j to 1. Consider a random neighbor S of current, and calculate the cost differential of the two nodes.
4. If S has a lower cost, set current to S, are increment j by 1 and If j is max neighbor, when j > max neighbor, compare the cost of current with min cost.
5. If the cost value is less than (<) min cost, set min cost to the cost of current value and set best node value to current node.
6. Finally group the cluster, to satisfy the threshold value≤ min cost. Then nodes are updated & cluster data and return outliers.
7. Else, Repeat the step 3 to step 5 up to best minimum cost (dmin), are found to other samples.
8. Return, Best cluster and detect the outliers efficiently.
```

IV. EXPERIMENTAL RESULTS

We have implemented these two partitioning clustering algorithms 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. The different sizes of the window are 3 and 5.

A. CLUSTERING ACCURACY

TABLE I
THE CLUSTERING ACCURACY FOR CLARANS AND E-CLARANS
IN THREE AND FIVE WINDOWS

Clustering Accuracy	Window size	No. of windows	CLARANS	E-CLARANS
Accuracy	Window size=3	W1	87.50	90.00
		W2	87.50	90.00
		W3	87.60	90.00
	Window size=5	W1	87.01	89.00
		W2	87.09	89.00
		W3	87.09	90.00
		W4	87.09	90.00
		W5	87.50	90.00

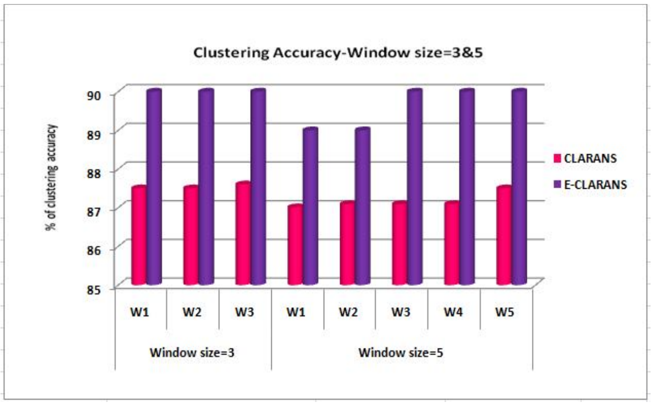

Figure 2: The clustering accuracy for CLARANS and E-CLARANS in three and five windows

From the above figure-2, it is observed that proposed E-CLARANS clustering algorithm performs better than CLARANS clustering algorithm.

B. Outlier Accuracy

Table II
The clustering accuracy for Clarans and E-clarans in three and five windows

Outlier Accuracy	Window size	No. of windows	CLARANS	E-CLARANS
Detection rate	Window size=3	W1	35.82	37.00
		W2	42.90	44.00
		W3	38.00	50.00
	Window size=5	W1	30.00	34.00
		W2	35.00	40.00
		W3	30.00	47.00
		W4	30.00	35.00
		W5	40.00	45.00
False alarm rate	Window size=3	W1	32.00	30.00
		W2	41.00	40.00
		W3	35.00	34.00
	Window size=5	W1	30.00	28.00
		W2	30.76	28.00
		W3	40.00	37.00
		W4	27.77	25.00
		W5	30.76	25.00

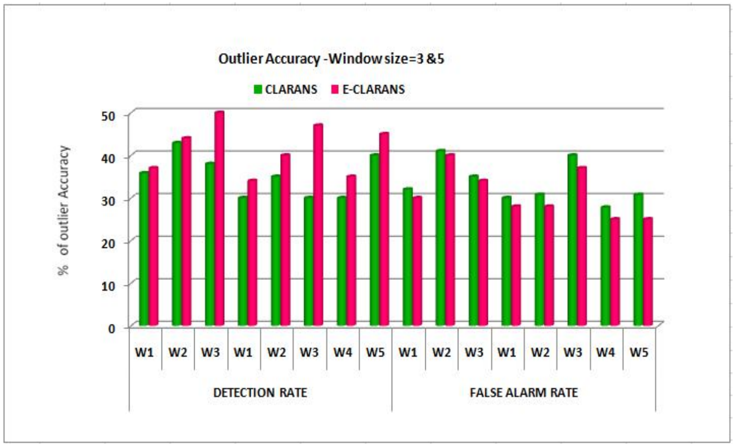

Figure 3: The outlier accuracy for CLARANS and E-CLARANS in three and five windows

From the above figure-3, it is observed that proposed E-CLARANS clustering algorithm performs better than CLARANS clustering algorithm.

V. Conclusion

Data streams are fast and limitless arrival of ordered and unordered data, by using of data streams clustering technique we can handle those data. Detecting outliers in data stream is one of the challenging research problems. In this paper, we have analysed the performance of CLARANS and E-CLARANS clustering algorithm for detecting the outliers. In turn to find the best clustering algorithm for outlier detection two performance measures are used. From the experimental results it is come to know that the outlier detection and clustering accuracies are more efficient in proposed E-CLARANS while compared to CLARANS clustering.

.

References

1. Aggarwal.C, Ed., “Data Streams – Models and Algorithms”, Springer, 2007.
2. Aggarwal.C.C, J. Han, J. Wang, and P. S. Yu,“A framework for clustering evolving data streams,” In Proc. of VLDB, pages 81-92, 2003.
3. C. J. Merz and P. M. Murph, UCI Repository of Machine Learning Databases Univ. of CA,Dept. of CIS, Irvine.
4. 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).
5. 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)Vol. 2, No. 3, Page: 25-34,2013.
6. J. Chandrika, Dr. K.R. Ananda Kumar, “Dynamic Clustering Of High Speed Data Streams”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.
7. Rajendra Pamula, Jatindra Kumar Deka,Sukumar Nandi “An Outlier Detection Method based on Clustering”, Second International Conference on Emerging Applications of Information Technology, 2011.
8. Raymond T. Ng and J. Han. Efficient and effective clustering method for spatial datamining, VLDB'94.
9. S. Vijayarani, P. Jothi, “A New Approach for Detecting Outliers in Data Streams”, International journal of engineering sciences & research Technology, ISSN: 2277-9655, Pg no: 3128-3133, November 2013.
10. Shifei Ding, Fulin Wu, Jun Qian, Hongjie Jia, “Research on data stream clustering algorithms” in Artificial Intelligence Review, springer 2013.
11. Sudipto Guha, Adam Meyerson, Nine Mishra and Rajeev Motwani, “Clustering Data Streams: Theory and practice,” IEEE Transactions onKnowledge and Data Engineering, vol. 15, no.3, pp. 515-528, May/June, 2003.
12. T. Soni Madhulatha, “overview of streaming-data algorithms”, Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November, 2011.
13. Yi-hong lu, Yan huang, “Mining DataStreams Using Clustering”, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics,vol.4, pp. 18-21,2005.
14. Yogita, Durga Toshniwal, “Clustering Techniques for Streaming Data–A Survey” in proc. Of the IEEE, 2012.

BIOGRAPHY

Dr. S.Vijayarani has completed MCA, M.Phil and Ph.D in Computer Science. She is 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, security issues, text mining and data streams. She has published papers in the 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 Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining and Data Streams.