International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 4143–4149

An Effective Data stream mining Algorithm for Inliers and Outlier Detection in Micro, macro clusters

[1]L.Ramesh, [2]S.Gopinathan

[1]Ph.D.,Research Scholar,Department of Computer Science, University of Madras, Chennai, 600025,India
[2]Professor, Department of Computer Science, University of Madras, Chennai,600025,India.

Abstract

In the fast growing world applications are generating data in enormous volumes called data streams. Data stream is imaginably large, continual, rapid flow of information and in data mining the important tool is called clustering, hence data stream clustering (DSC) can be said as active research area. Recent attention of data stream clustering is through the applications that contain large amounts of streaming data. Data stream clustering is used in many areas such as weather forecasting, financial transactions, website analysis, sensor network monitoring, e-business, telephone records and telecommunications. In this, paper proposes data stream mining Algorithm improvement which deals with inlier and outlier detection using the proposed approach. The performed experiments show the effectiveness of data stream mining Algorithm in detecting and dealing with inlier and outliers from the stream, when compared potentially of the proposed approach to be used in other micro cluster based data stream algorithms.

Key Words: *data stream Clustering, inlier, outlier detection.*

1. INTRODUCTION

One of the most promising tasks for DS, clustering organizes stream objects into clusters in accordance with their similarity, being that objects of the same cluster present high similarity when compared to objects from other clusters [2]. Within the DS scenario, there exist a number of requirements that need to be considered for this task [3]: i) capacity of discovering clusters with different formats, ii) adaptation to dynamic changes in which clusters can evolves, new clusters can appear and old one disappear, iii) ability to detect outliers, iv) flexibility in returning clusters to the user at the requested moment, v) memory management, and vi) definition of the input parameters. A majority of DS clustering algorithms possess two phases: i) online (or data abstraction), which continually maintains a statistical summary of the objects of the DS, and ii) offline (or macro clustering), which performs a macro-clustering of the static summary through use of classic clustering algorithms, e.g., k-means [4] and DBSCAN [5]. The detection

ISSN: 2233-7857 IJFGCN

of outliers in DS is challenging, in both online and offline phases. By performing a simple sweep across the data, the outliers may be confused with objects of a new cluster or the evolution of existing ones.

The insertion of outliers into the data abstraction structure, named micro-cluster may cause performance loss of the algorithm in the online phase, by influencing the evolution of the micro-clusters, and in the offline phase, by causing the loss of valid information used for the analysis. Therefore, it is necessary that the algorithms are capable of following the evolution of the data, in a manner that maintains a consistent cluster [1], [3]. This research paper proposes a new approach to address outlier detection in the micro clusters of the DS clustering algorithms. In the proposed approach, potential outliers, i.e. those objects that could not be added to the model, are inserted into micro clusters analysis. In order to demonstrate the efficiency of the proposed approach, we propose the data stream mining algorithm, which is an extension of the CluStream [1], for dealing with outliers using an auxiliary memory. However, this approach can be applied to other micro-cluster based DS algorithms.

2. RELATED LITERATURE REVIEW

It is used to mining local outliers by building spars In this research [2] a novel outlier cluster detection Algorithm (ROCF) without top-n parameter. ROCF can detect the outlier cluster that are hard to detect out by other distance based or density based outlier detection algorithms. The proposed algorithm can detect the outlier and outlier cluster without parameter to specify the number of outliers or percentage of outliers in a database. In this research the experimenter has [3] analyzed the clustering and outlier performance of BIRCH with CLARANS and BIRCH with K-Means clustering algorithm for detecting outliers. By using DSH clustering and partition clustering which are assistive to detect the outliers efficiently.

The researcher [4] introduces the notion of the local outlier factor LOF, which captures exactly this relative degree of isolation. It demonstrate that out heuristic appears to be very promising in that it can identify meaningful local outliers and shows that the approach of finding local outliers is efficient for datasets where the nearest neighbor queries are supported by index structures and still practical for very large datasets.

In this research the experimenter has [7] Presents the novel data stream outlier detection algorithm SODRNN. It is carried out on both artificial and real data sets show that the proposed method is efficient and effective. It is improve the performance of this algorithm in high-dimensional data stream environment. In this research [12] a new algorithm named DBCOD that unifies density-based clustering and outlier detection. DBCOD needs just an input parameter Validates the algorithm with extensive experimental evaluation on different shape, large scale and high-dimensional databases.

The researcher [8] generalize local outlier factor of object and propose a framework of clustering based outlier detection. The theoretical analysis and the experimental results show that FCBOD has better performance than clustering based outlier detection methods. In this research experimenter [9] proposed a hierarchical clustering based global outlier detection method. It isolates degree of the data sets by the

ISSN: 2233-7857 IJFGCN

hierarchical clustering tree and the distance matrix, and the determine the number of outliers to delete. The algorithm is applicable to multi-level data and large data sets.

3. BLOCK DIAGRAM OF PROPOSED WORK

The following figure represents the framework of our proposed work for an effective data stream mining algorithm for inlier and outlier detection in micro clusters. As shown figure there are various components involves in proposed framework.

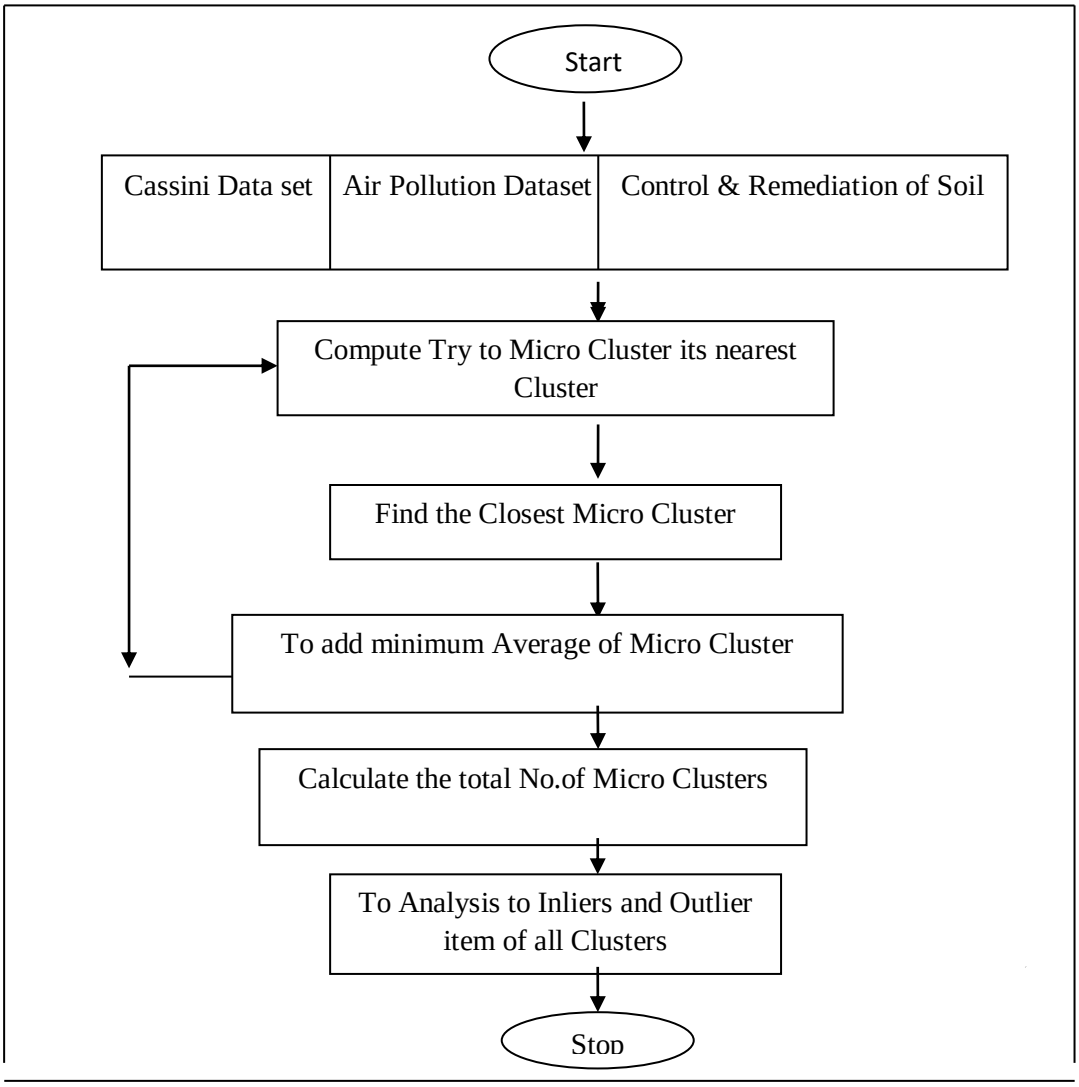


Fig1: Block Diagram of Proposed Work

In the figure 1, shown the block diagram of proposed work. The framework for an effective data stream mining algorithm for inlier and outlier detection in micro clusters Input Dataset: The process in sectioned into three dataset like Cassini, Air pollution dataset and Control and Remediation of soil, Creation of micro cluster is a lot of individual data indicates that are close one another to add nearest cluster.

ISSN: 2233-7857 IJFGCN

Subsequent of data separation in the process that new micro cluster, to add minimum average of micro cluster for the calculate inlier and outlier detection for all micro clusters.

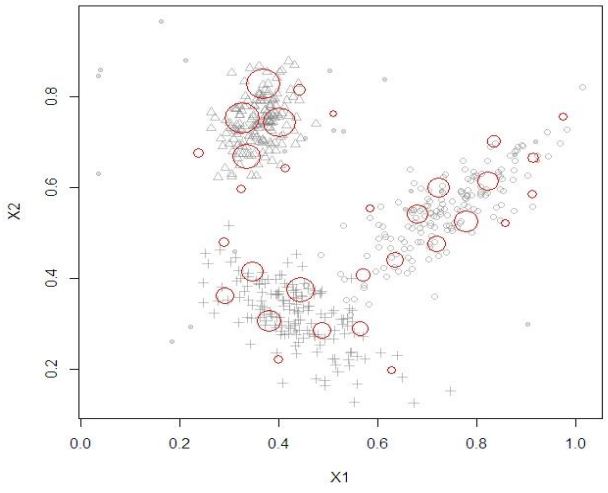
4. ALGORITHM FOR PROPOSED SYSTEM

(DATA STREAM MINING ALGORITHM OUTLIER DETECTION)

1: Initialization
2: Try to micro cluster(Mc) its nearest cluster(nc)
3: for (r=0.05) the new cluster
4: find the closest micro cluster
5: Add Mc←average minimum of radius (dist, r=0.05)
6: if dist (old item, new item) < maximum limit then insert (new item)
7: else
8: add (old item, max)// added new Micro Cluster (Mc)
9: end if
10: To Calculate the total number of micro clusters
11: To Analysis to inliers item and outlier item for all micro clusters
12: end for
13: end update

5. RESULT AND DISCUSSION

Fig 2 :Data Stream Mining Algorithm for Cassini Dataset



Total No. of Attributes:7

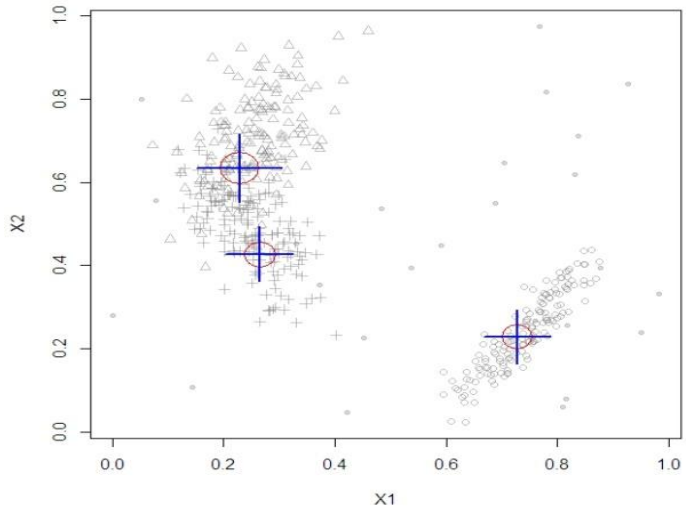
Total No. of Instances: 294

ISSN: 2233-7857 IJFGCN

Name of Data Stream Mining Algorithm	Inliers item	Outlier item	No.of Clusters
DBStream	283 items	3 items	8 clusters
DenStream	287 items	7 items	4 clusters
D-Stream	292 items	2 items	3 clusters
D-Stream with Attraction	281 items	13 items	6 clusters
Clustream	285 items	9 items	5 clusters

In Table 1, the experiments were performed on total No.of Attributes 7 and total No.of Instances 294 the input parameters used for data stream mining algorithm were implemented on MOA. In the DBStream clusters are occurred 283 inlier item, 3 outlier item and 8 clusters, DenStream clusters are occurred 287 items inlier, 7 outlier detection and 4 clusters, D-Stream clusters are occurred 292 inlier items,2 outlier detection and 3 clusters, D-Stream with Attraction clusters are occurred 281 items, 13 items inlier , 6 outlier detection and 6 clusters and finally Clustream are occurred 285 items , 9 item inlier , 5 outlier detection are performed.

Fig 3: Data Stream Mining Algorithm for Air Pollution Dataset



Total No. of Attributes:10

Total No. of Instances: 549

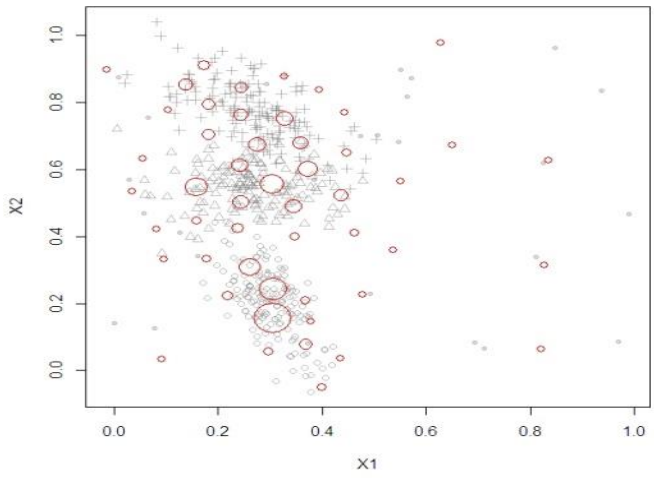
Name of Data Stream Mining Algorithm	Inliers item	Outlier item	No.of Clusters
DBStream	548 items	1 items	6 clusters
DenStream	540 items	9 items	5 clusters
D-Stream	545 items	4 items	3 clusters
D-Stream with Attraction	529 items	20 items	8 clusters

ISSN: 2233-7857 IJFGCN

Clustream	543 items	6 items	4 clusters

In Table 2, the experiments Air pollution dataset were performed on total No.of Attributes 10 and total No.of Instances 549 the input parameters used for data stream mining algorithm were implemented on MOA. In the DBStream clusters are occurred 548 inlier item, 1 outlier item and 6 clusters, DenStream clusters are occurred 540 items inlier, 9 outlier detection and 5 clusters, D-Stream clusters are occurred 545 inlier items,4 outlier detection and 3 clusters, D-Stream with Attraction clusters are occurred 529 items inlier, 20 items inlier , 8 outlier detection and finally Clustream are occurred 543 items , 6 item inlier , 4 outlier detection are performed.

Fig 4:Data Stream Mining Algorithm for Control and Remediation of soil Dataset



Total No. of Attributes:10

Total No. of Instances: 85

Name of Data Stream Mining Algorithm	Inliers item	Outlier item	No.of Clusters
DBStream	78 items	7 items	3 clusters
DenStream	76 items	9 items	4 clusters
D-Stream	83 items	2 items	3 clusters
D-Stream with Attraction	80 items	5 items	8 clusters
Clustream	79 items	6 items	7 clusters

In Table 3, the experiments Control and Remediation of soil dataset were performed on total No.of Attributes 10 and total No.of Instances 85 the input parameters used for data stream mining algorithm were implemented on MOA. In the DBStream clusters are occurred 78 inlier item, 7 outlier item and 3 clusters, DenStream clusters are occurred 76 items inlier, 9 outlier detection and 4 clusters, D-Stream clusters are occurred 83 inlier items,2 outlier detection and 3 clusters, D-Stream with Attraction clusters

ISSN: 2233-7857 IJFGCN

are occurred 80 inlier items, 5 items outlier , 8 clusters and finally Clustream are occurred 79 inlier items , 6 item outlier , 7 clusters detection are performed.

6. CONCLUSION

Based on the experiments presented in this research paper, it is possible to verify that the proposed method for inlier and outlier detection and validation is effective in most of the studied scenarios. Exceptions occurred in some datasets; the resulted micro-clusters presented significantly less outliers than the original D-Stream.

REFERENCES

[1] XujunZhao,JifuZhang,Xiao Qin " Expert System with Applications -LOMA , A local outlier mining algorithm based on attribute relevanceanalysis,p.p. 272-280,2017.

[2].Jinlong Huang, QingshengZhu,Lijun Yang, DongDongCheng,Quanwang Wu : conference on Knowledge- Based Systems,p.p.1-9,2017.

[3] S.Vijayarani,Ms. P Jothi "An Efficient Clustering Algorithm forOutlier Detection in Data Streams", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue9,3657-3665,September 2013.

[4].Markus M. Breunig, Hans-peter Kriegel, Raymond T. Ng,Jorg Sander: LOF : Identifying Density-Based Local Outliers.International Confer-ence on Management of Data, 2000.

[5].Bryan Perozzi,LemanAkoglu, Patricia Iglesias Sanchez,EmmanuelMuller : Focused Clustering and Outlier Detection in Large AttributedGraphs,2000.

[6]. S.Ganapathy, N.Jaisankar, P.Yogesh and A.Kannan,An Intelligent Sys-tem for Intrusion Detection Using Outlier Detection, IEEE-InternationalConference on Recent Trends in Information Technology,p.p-119-123,2011

[7].ZhongPing Zhang, YongXinLiang " A Data Streams Outlier Detection Algorithm Based on reverse K nearest Neighbours",International Journal of Adavanced Science and Technology- 2011.

[8].Sheng-Yi Jiang, Ai-Min Yang : Framework of Clustering-Based OutlierDetection, International Conference on Fuzzy Systems and KnowledgeDiscovery,475-479,2009.

[9]. Bin-meiLIANG : A Hierarchical Clustering Based Global OutlierDetection Method.

[10]. Harshada C. Mandhare "A Comparative study of Cluster Based OutlierDetection, Distance Based Outlier Detection and Density Based OutlierDetectionTechniques," International Conference on Intelligent Computingand Control Systems p.p.931-935,2017.

[11] K. Thangavel, A. KajaMohideen,Semi Supervised K-means Clusteringfor Outlier Detection in Mammogram Classification,p.p.68-72.

[12]Pamula, Rajendra, Jatindra Kumar Deka, and Sukumar Nandi. "An outlier detection method based on clustering." In Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on, pp. 253-256. IEEE, 2011.

ISSN: 2233-7857 IJFGCN