IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-1 E-ISSN: 2347-2693

# Novel approach for data stream clustering through micro-clusters shared Density

Prashant V. Desai[1*], Vilas S. Gaikawad[2]

[1*]Department of Computer Engineering, RajshreeShahu School of Engineering and Research, JSPM NTC, Pune, INDIA
[2]Department of Computer Engineering,RajshreeShahu School of Engineering and Research JSPM NTC, Pune, INDIA

**Available online at: www.ijcseonline.org**

Received: 29/12/2016 Revised: 08/01/2017 Accepted: 17/01/2017 Published: 31/01/2017

***Abstract:*** **Clustering is the process of organizing objects into groups whose members are similar in some way and is very important technique in data mining as it has its applications spread extensively, e.g. marketing, biology, pattern recognition etc. So summarize the data stream in the real life with the online process is called as micro-cluster but it shows the density when we are combining the data in the one place. In the offline process we are using the modification clustering algorithm to re-clustering into larger cluster. For that the center of micro-cluster point as the pseudo point with density randomly calculates their weight. That density information area of micro-cluster is not preserved the online process. So used DBSTREAM, the first micro-cluster based on online clustering component capture the density between micro-cluster via shared density graph. We develop and evaluate a new method to address this problem for micro-cluster-based algorithms. The density information in this graph is then exploited for re-clustering based on actual density between adjacent micro-clusters. For that shared density graph improves clustering quality over other popular data stream clustering methods which require the creation of a larger number of smaller micro-clusters to achieve comparable results.**

***Keywords- Data mining, data stream clustering, density-based clustering.***

## *I.* INTRODUCTION

Clustering is an important technique of exploratory data mining, which divides a set of objects (instances or patterns) into several groups (also called clusters) in such a way that objects in same group are more similar with each other in some sense than with the objects in other groups. We introduced an extension to the grid-based D-Stream algorithm. the concept of attraction between adjacent grids cells and showed its effectiveness. In that we use shared density graph which captures the density of the original data between micro-clusters during clustering. So we used DBSTREAM, the first micro-cluster-based online clustering component that explicitly captures the density between micro-clusters via a shared density graph. we develop and evaluate a new method to address this problem for micro-cluster-based algorithms. Reclustering represents the algorithm's offline component which uses the data captured by the online component. In this system advantage since it implies that we can tune the online component to produce less micro-clusters for shared-density reclustering. This improves performance and, in many cases, the saved memory more than offset the memory requirement for the shared density graph.

In this paper Re-clustering approaches completely ignore the data density in the area between the micro-clusters and thus might join micro-clusters which are close together but at the same time separated by a small area of low density. To overcome this problem D-Stream algorithm shared density graph which explicitly captures the density of the original data between micro-clusters during clustering and then show how the graph can be used for reclustering micro-clusters. To the best of our knowledge, investigate using a shared-density-based reclustering approach for data stream clustering.

Why Data Mining?

The amount of the data seems to increase rapidly every single day for the majority domains related to information processing, and the need to find a way to mine and get knowledge from databases is still crucial. Data Mining (DM) defines the automated extraction procedures of hidden predictive information from databases. DM problems addressed by intelligent systems are: pattern recognition, prediction, classification, clustering etc.

There are two major issues that will affect the image data mining process.

• The similarity matching process.

• The generality of the application area, that is, the breadth of usefulness of data mining from a practical point of view.

In computer science, data stream clustering is defined as the clustering of data that arrive continuously such as telephone records, multimedia data, financial transactions etc. Data stream clustering is usually studied as a streaming algorithm and the objective is, given a sequence of points, to construct a good clustering of the stream, using a small amount of memory and time.

The problem of data stream clustering is defined as: Input: a sequence of n points in metric space and an integer k.

Output: k centers in the set of the n points so as to minimize the sum of distances from data points to their closest cluster centers.
This is the streaming version of the k-median problem.

Clustering based on density (local cluster criterion), such as density-connected points or based on an explicitly constructed density function
Major features:

- Discover clusters of arbitrary shape
- Handle noise
- One scan
- Need density parameters

To improve performance for micro-cluster-based algorithms used the concept of a shared density graph which detail captures the density of the original data between micro-clusters during clustering and then show how the graph can be used for reclustering micro-clusters to reduce the density.

The data stream model has recently attracted attention for its applicability to numerous types of data, including telephone records, Web documents, and click streams. For analysis of such data, the ability to process the data in a single pass, or a small number of passes, while using little memory, is crucial.

## II. RELATED WORK

DBSTREAM is closely related to DBSCAN [12] with two important differences. Similar to DenStream [11], density estimates are calculated for micro-clusters rather than the epsilon neighbourhood around each point. This reduces computational complexity significantly. The second change is that DBSCAN's concept of reachability is replaced by connectivity. Reachability only reflects closeness of data points, while connectivity also incorporates interconnectivity introduced by CHAMELEON [13]. In general, the connectivity graph developed in this paper can be seen as a special case of a shared nearest neighbor graph where the neighbors shared by two MCs are given by the points in the shared area. As such it does not represent k shared nearest neighbors but the set of neighbors given by a fixed radius. DBSTREAM uses the most simple approach to partition the connectivity graph by using  as a global threshold and then finding connected components.However, any graph partitioning scheme, e.g., the ones used for CHAMELEON or spectral clustering, can be used to detect clusters. Compared to D-Stream's concept of attraction which is used between grid cells, DBSTREAM's concept of connectivity is also applicable to micro-clusters. DBSTREAM's update strategy for micro cluster centers based on ideas from competitive learning allows the centers to move towards areas of maximal local density, while DStream's grid is fixed. This makes DBSTREAM more flexible which will be illustrated in the experiments by the fact that DBSTREAM typically needs fewer MCs to model the same data stream.

## III. PROBLEM STATEMENT

Re-clustering approaches completely ignore the data density in the area between the micro-clusters and thus might join micro-clusters which are close together but at the same time separated by a small area of low density. To overcome this problem D-Stream algorithm shared density graph which explicitly captures the density of the original data between micro-clusters during clustering and then show how the graph can be used for reclustering micro-clusters. To the best of our knowledge, investigate using a shared-density-based reclustering approach for data stream clustering.

## IV. MOTIVATION

The data stream model has recently attracted attention for its applicability to numerous types of data, including telephone records, web documents and click streams. For analysis of such data, the ability to process the data in a single pass, or a small number of passes, while using little memory, is crucial
As lots of data are combined together because of that density are increasing in the micro-clustering in online system. So we think density-based clustering is a well-research area and we can only give a very brief overview here.
The widely used practice of viewing data stream clustering algorithms as a class of one- pass clustering algorithms is not very useful from an application point of view. The intrinsic nature of stream data requires the development of algorithms capable of performing fast and incremental processing of data objects, suitably addressing time and memory limitations.
The data stream model has recently attracted attention for its applicability to numerous types of data, including telephone records, Web documents, and click streams. For analysis of such data, the ability to process the data in a single pass, or a small number of passes, while using little memory, is crucial.

## V. EXISTING SYSTEM

In existing system k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriority. The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as BabyCenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centre.

## VI. SYSTEM ARCHITECTURE

The figure illustrates the system flow of the existing framework of automatic clustering on Density Metrics, which consists of the following steps. First we calculate the object density and density based distance. And then find clusters with their centers.

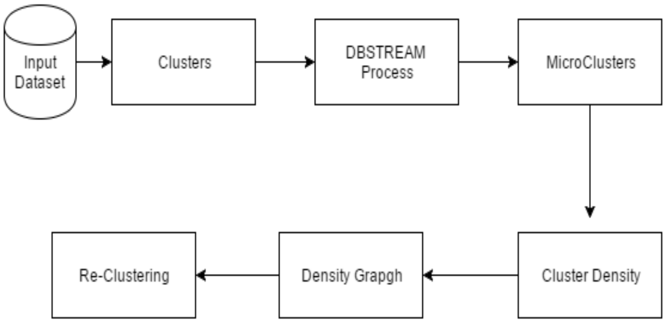

Fig 1.System Architechture

## VII. ALGORITHM FOR EXISTING SYSTEM

```
MC                                        Δ Set of MCs
mc ϵ MC has elements mc =(c, w,t)  Δcenter, weight,
                                   last update time
S   Δ weighted adjacency list for shared density graph
sij  ϵ  S has an additional field t. Δ time of last update
t                              Δ current time step
Require: User-specified parameters
r                              Δclustering threshold
גΔfading factor
tgap                           Δcleanup interval
Wmin                           Δminimum weight
α                              Δintersection factor

function UPDATE(x)   Δnew data point x
    N ← findFixedRadiusNN (x, MC, r)
    if |N| < 1 then         Δcreate new MC
        add (c = x, t = t, w = 1) to MC
    else                    Δupdate existing MCs
        for each i ϵ  N  do
            $mci[w] \leftarrow mci[w]^2 + 1$
            mci[c] ←  mci[c] + h(x, mci[c])(x - mci[c])
            mci[t] ← t              Δupdate shared density
            for each j ϵ N where j > i do
                sij  ←  sij 2 + 1
                sij [t] ←  t
            end for
        end for        Δprevent collapsing clusters
        for each (I , j) ϵ  N x N and j > i do
            if  dist(mci[c] , mcj [c]) < r  then
                revert mci[c], mcj [c]  to previous positions
            end if
        end for
    end if
    t ← t + 1
end function
```

## VIII. PROPOSED SYSTEM

The purpose of the project is to improve performance for micro-cluster-based algorithms used the concept of a shared density graph which detail captures the density of the original data between micro-clusters during clustering and then show how the graph can be used for reclustering micro-clusters to reduce the density.

## IX. CONCLUSION

The data stream clustering algorithm which detailed manner records the density in the area shared by micro-clusters and uses this information for reclustering. We have introduced the shared density graph together with the algorithms needed to maintain the graph in the online component of a data stream mining algorithm.

## REFERENCES

[1]. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for clustering evolving data streams,” in Proceedings of the International Conference on Very Large Data Bases (VLDB ’03), 2003, pp. 81–92.

[2]. F. Cao, M. Ester, W. Qian, and A. Zhou, “Density-based clustering over an evolving data stream with noise,” in Proceedings of the 2006 SIAM International Conference on Data Mining. SIAM, 2006, pp. 328–339.

[3]. Y. Chen and L. Tu, “Density-based clustering for real-time stream data,” in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007, pp. 133–142.

[4]. J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. a. Gama, “Data stream clustering: A survey,” ACM Computing Surveys, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013.

[5]. A . Bifet, G. de Francisci Morales, J. Read, G. Holmes, and B. Pfahringer," Efficient Online Evaluation of Big Data Stream Classifiers." 2015

[6]. H. Kremer, P. Kranen, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer." An Effective Evaluation Measure for Clustering on Evolving Data Streams." 2011

[7]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'1996), 1996.

[8]. Komkrit Udommanetanakit, Thanawin Rakthanmanon, and Kitsana Waiyamai." Estream: Evolution-based technique for stream clustering. " 2007

[9]. Dimitris K. Tasoulis, Niall M. Adams, David J. Hand." Unsupervised clustering in streaming data." 2006

[10]. Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, Member, IEEE, and Liadan O'Callaghan. "Clustering data streams: Theory and practice."2013

[11]. F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in Proceedings of the 2006 SIAM International Conference on Data Mining. SIAM, 2006, pp. 328–339.

[12]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'1996), 1996, pp. 226–231.

[13]. G. Karypis, E.-H. S. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," Computer, vol. 32, no. 8, pp. 68–75, Aug. 1999. [Online]. Available:http://dx.doi.org/10.1109/2.781637

[14]. Enakshmi Nandi1, Debabrata Sarddar "A Modified MapReduce-K-Means Clustering Based Load Distribution Method for Wireless Sensor Network in Mobile Cloud computing" ,2016

[15]. Pritika goel, "An Improved Load Balancing Technique in Weighted Clustering Algorithm" ,2016.Available: http://www.ijcseonline.org/pdf_paper_view.php?paper_id=972&18-IJCSE-01678.pdf

## Authors Profile

*Mr. Desai Prashant* pursed Bachelor of computer Science from University of Shivaji University, India in 2006 . *He is currently pursuing Master of Engineering from*Rajshree Shahu School of Engineering and Research, JSPM NTC, Pune, India. His main research work focuses on Data Mining.

***Mr. Vilas S. Gaikwad*** (F8001741) received the BE Degree in Computer Science & Engineering from the Dr. BAMU Aurangabad in 2010, the M.Tech degree in Computer Science &Engineering from Walchand College of Engineering (An autonomous Institute), Sangli in 2012. He is currently a Assistant Professor in the Department of Computer Engineering, Rajshree Shahu School of Engineering and Research, JSPM NTC, Pune, India.His research area include Image Processing, Data Mining and Computer Network.