Data Stream Clustering Using Micro Clusters

Ms. Jyoti .S.Pawar[1], Prof. N. M.Shahane.[2]

[1] *PG student, Department of Computer Engineering K. K. W. I. E. E. R., Nashik Maharashtra, India*
[2] *Assistant Professor Department of Computer Engineering K. K. W. I. E. E. R., Nashik Maharashtra, India*

ABSTRACT

Data streams are massive, dynamic and unbounded. Due to these issues data stream clustering is challenging problem. Data stream are observed in network monitoring, critical scientific application, weather monito ring and astronomical applications, electronic business, stock trading etc. Data stream clustering puts additional constraints on clustering algorithms. Data streams must be processed in single pass with limited memory as well as with less processing time, but the streams can be highly dynamic. Most of the existing clustering algorithms are distance based and unable to handle the interwoven clusters and also it is impossible to save the data streams, because of infinite characteristic. Proposed work focuses on density based clustering algorithms using micro-clusters. The process is divided into two-phases, online and offline, micro clusters are created in online phase and final clusters are generated in offline phase.

Keyword : - *Data Streams, Density Based Clustering, Micro Cluster.*

1. INTRODUCTION

In recent years demands of data stream clustering increases rapidly .Data stream are observed in network monitoring, critical scientific application, weather monitoring and astronomical applications, electronic business, stock trading ,social networks ,sensor network etc. In these applications ,data stream arrives continuously and evolve significantly over time. There are many technologies available which facilitates us to record day to life transactions at rapid rate. Such process lead to large volume of continuous data. This data term as 'Data Stream'. Data streams are highly dynamic, massive and unbounded in nature. Due to these characteristics real-time data stream clustering is challenging problem. Data stream clustering puts additional constraints on clustering algorithms. Clustering in data stream environment needs some special requirements due to data stream's characteristics such as clustering in bounded memory and within limited processing time as well as with single pass over evolving data streams.

Data stream clustering is generally divided in two phases online and offline. Online phase summarized data into many micro clusters and then in offline phase micro clusters are merged and form macro cluster.[1]Reclustering is offline process hence its does not have limited time bound.

In literature various data stream clustering methods are discussed like hierarchical and partitioning which are use to create spherical-shape clusters. Density based clustering is one of the most important method to discover non-spherical shape and outliers. DENCLUE, DBSCAN, OPTICS, are density based clustering algorithm. These algorithm focuses on dense area of data points in data space and identify as cluster as they are separated by low density area .Another important method of clustering is grid based. Grid based clustering method has fast processing time and it is not depended on number of data points.

Micro-clustering is a one of the remarkable method in data stream clustering which use to compress data streams effectively as well as record the temporal locality of data[6] concept of micro -cluster was first introduced in[14] for large dataset and subsequently adapted in [5] for data streams.

Existing reclustering algorithm totally ignore the density of data between the micro-clusters. Due to this micro -Clusters which are close to each other might be merged although they are separated by low density. This problem is addressed by Tu and Chen[9]by introducing extension to D-stream algorithm.

The goals of this proposed system are-

- Develop an algorithm which works on single scanning of real data stream which is continuous and highly dynamic.
- The system should be capable to process data stream with low main memory and CPU usage.
- User should not require prior knowledge about nature of data stream, number of clusters.
- Data stream should be processed within reasonable time and form quality clusters.
- The system should discover arbitrary shaped clusters.
- The system should handle noise in data streams.
- The system should flexible and easily update the micro cluster in real time for better quality of clusters.

In this paper we developed and evaluate new method of data stream clustering using Micro-clusters to address this problem. Data stream clustering done in two phases online and offline. At online phase Micro-clusters are created and maintained, then in offline phase micro-clusters are reclustered or merged to form final cluster or Macro cluster. Shared density graph a new approach is used which captures original data density between micro-clusters during clustering.

This paper organized as follows, in section 2 we discussed literature related work. In section 3 introduces system architecture. In section 4 we analyzed the system.

2. RELATED WORK

Data streams can be cluster by using distance based methods, partitioning methods, hierarchical methods and density based methods.For each of these methods there are number of algorithms are available in data mining. Distance based algorithms are K-means. DBSCAN,OPTICS and DBCLASD are examples of density based algorithm. DENCLUE algorithm is based on density function. CHAMELEON is hierarchical based algorithm.

Data stream clustering algorithms are generally works in one phase. The data streams visits only once to the algorithm for classification or clustering, which is difficult to provide flexibility. To overcome this problem two phase data clustering algorithm introduced.

Martin Ester, Xiaowei Xu, Hans-Peter Jriegel and Jorg Sander proposed DBSCAN[1] algorithm in 1996.DBSCAN is data stream clustering algorithm based on density of nodes. Clusters are formed depending upon the data points which are closely packed to its neighbor. In this algorithm *ε-Neighbourhood* and *minPts* these two parameters are used. DBSCAN arbitrarily starts with first point which has not been visited. *ε-Neighbourhood* defines as the nodes which are within distance ε from given node. minPts are that data points which are consist of minimum number points of neighbours. The advantage of this algorithm is it does not require to specify number of clusters like in k-means. It also useful for finding arbitrary shaped clusters. DBSCAN is not independent of ordering of data points. This algorithm suffers from robustness. DBSCAN can produced low quality clusters if distance threshold ε is not properly choose. It have difficulties to work well when there is distance between data points is large.

DENCLUE another density based algorithm proposed by Hinneburg and Keim (KDD-98).DENCLUE can handle large amount of data set with noise. It uses statical density functions to find out arbitrary shaped clusters. It works significantly faster than DBSCAN. It uses grid cells and manage these cells as tree structure. It defines influence function which determines data points impact on its neighbourhood. It calculate sum of influence function to obtain density of data points within space. Center of cluster define the attraction between each density data points.

Xiaowei Xu, Martin Ester, Hans-Peter Jriegel and Jorg Sander proposed DBCLASD algorithm. DBCLASD refereed to Distribution Based Clustering of Large Spatial Databases. It is based on partitioning based method for clustering. DBCLASD does not require any input parameters. It is useful for determining arbitrary shape clusters

Charu C. Agggrwal, Jiawei Han, Jianyong Wang, Philip S. Yu proposed clustering framework for evolving data streams.[11].CluStream is one of the method from that framework which is first time introduced micro-cluster. Micro-cluster are data structure summarized the data streams instances which can be further easily updated and use for fast access. CluStream is first two phase algorithm. It has online and offline phase. In online phase data streams are processed and summarized and micro-clusters are created and also stored in main memory. In offline phase clusters are generated based on summarized information. In offline phase reclustering is done on obtained micro-clusters for that it uses weighted k-means algorithm. CluStream is based on k-means approach and discover only spherical clusters. This limitation of CluStream is overcome by Density based clustering methods. So density based clustering algorithms are two phases. It requires careful selection of input parameter.

Feng Cao,Martin Ester,Weining Qian and Aoying Zhou proposed Den-Stream an another clustering algorithm which is based on continuous data streams. Den-Stream algorithm is extension of DBSCAN algorithm and based on density-based. It adapted from CluStream framework. Den-Stream algorithm is extensions as rDen-Stream, C-Den-Stream ,D-Stream, MR-Stream. This algorithm introduced concept of micro-cluster and outlier. In Den-Stream concept of fading window introduced for micro clusters which are not updated over time. The main disadvantage of density based algorithm is that when density is changes widely.

MR-Stream clustering algorithm is an extension to the Den-Stream algorithm. MR-Stream data stream clustering algorithm works on multiple resolution data streams. In this algorithm space partitioning is used where data stream partition in tree like structures or in cells. In tree each node contains information about its parent as well as children node. This algorithm improves clustering quality and performance by determining timing of cluster generation.MR-Stream algorithm cannot work when data is high dimensional.

D-Stream is Density Grid based clustering algorithm. It is used for Real Time data streams. In this data points are mapped to the corresponding grids. Formation of clusters depends on attraction between adjacent grid cells and their density. The main advantage of D-stream algorithm has outlier handing technique.

3. PROBLEM FORMULATION

To design and develop a system for clustering data streams using Micro clusters

4. SYSTEM ARCHITECTURE

In this section we will discuss about architecture of data stream clustering using Micro-clusters. Data streams are massive, unbounded and dynamic. Data streams can not saved due to memory constraints. Clusters must be created within reasonable time. By considering all these characteristics and constraints of real time data streams system is designed as shown in fig 1.

Proposed work focuses on density based clustering algorithms using micro-clusters. The process is divided into two-phases, online and offline, micro clusters are created in online phase and final clusters are generated in offline phase

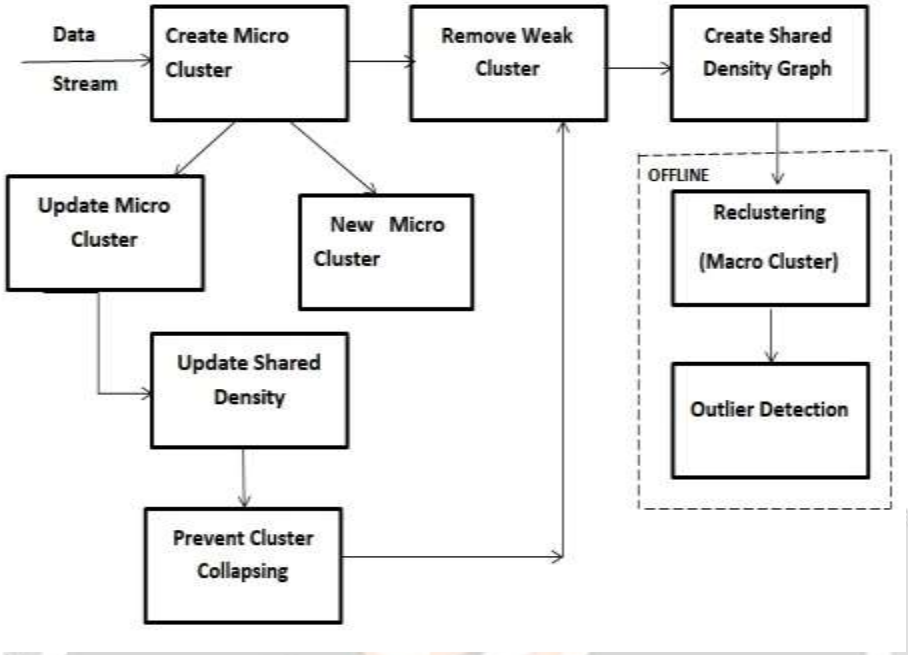

Fig -1: Architecture of System.

4.1. Creation of Micro clusters:

In this step Micro-clusters are stored as set of 'MC'. Each micro-cluster is represented by tuple(c,w,t) where c represents center of micro cluster, w is weight and time t represents last time micro-cluster updated. r is user specified radius of micro-cluster. Therefore, when data stream arrive it compare with its radius and puts into the proper micro-cluster.

4.2. New Micro cluster Creation:

New data point x is adding to existing micro-cluster or new micro-cluster is created. For this data point x determine all micro-clusters for which x can be added in their radius r. The distance between x and r is calculated by fixed radius nearest neighbor.

4.3.Updating Micro clusters:

After calculating fixed radius nearest neighbor if difference between radius r and new data point x is greater than 1 then that data point x is added to existing micro-cluster. Here weight of MC also updated. Weight of micro cluster is nothing but the total number of data points lies within that micro-cluster. After adding new data point x center of MC may be also updated. Update weight of each Micro cluster.

Clusters weights are decreases or faded after some time stamp by factor $2^{-\lambda}$ where $\lambda > 0$ .It is called as fading factor which is user specified.

Center of Micro clusters are also updated by moving towards new data points. This movement is controlled by Gaussian neighborhood function h().It is defined between two points p and q as

$$h(p,q)=\exp[-\|p-q\|^2/(2\sigma^2)]$$

Where $\sigma = r/3$

4.4.Update Shared Density:

Shared density capturing is new concept introduced to calculate dense area between micro-clusters. The shared density S between two micro-clusters i and j is estimated.

4.5. Prevent Clusters Collapsing:

To prevent collapsing of micro-clusters i and j care should be taken that the micro clusters should not come closer than r to each other. Calculate the distance between micro cluster i and j and update the time.

4.6. Remove Weak Clusters:

Weak clusters are defined as the weak micro-cluster entry in shared density graph between micro-cluster i and j that its weight is increases less than α in user specified clean up interval t_{gap} α is intersection factor between i and j micro-cluster.

Compare weight of each Micro cluster with w_{weak} and remove weak mc from MC and after that remove weak Shared Density S_{ij} from S.

4.7. Reclustering using Shared Density Graph:

As mentioned earlier Density based clustering is two phase process, online and offline. Reclustering is offline process. Noise threshold W_{min} and intersection factor α are user defined parameters. The Graph C is connectivity graph of shared density between only strong micro-clusters. To find connected component which are used to form final macro cluster connectivity graphs edges with a connectivity value greater than α i.e. intersection threshold are used.

5. EXPERIMENTS

In the experiments,3 datasets are used as benchmark dataset for performance evaluation. These datasets include KDD CUP-99,[7][8], Forest cover type[1], Sensor[10]

Table -1: Dataset Description

Dataset	No. of Instances	No of Features	Feature Type
KDD Cup-99	494020	41	Numeric
Forest Cover type	581012	54	Numeric
Sensor	919438	11	Numeric

Chart 2 shows the experimental results obtained over 10,000 data points from a stream from the KDD CUP-99 dataset. Proposed system compared with DSTREAM and CLUSTREAM As shown in 2 purity of micro cluster obtained by DBSTREAM is highest among other two algorithm of clustering.

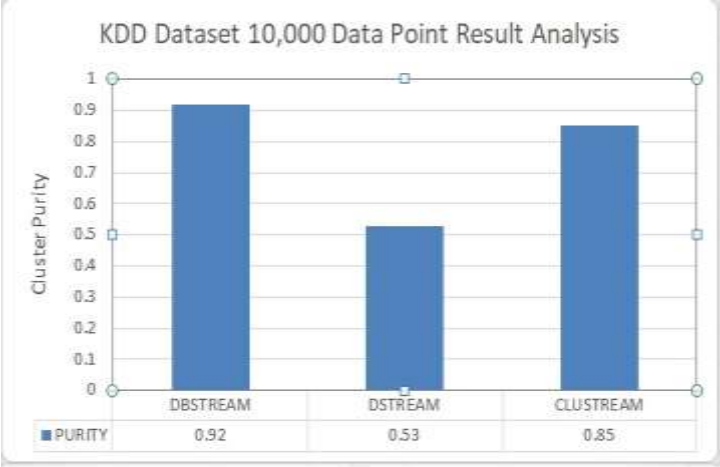

Chart -1: Purity of Micro clusters on KDD Cup-99 Dataset with 10000 data points.

A smaller WSS represent tighter clusters and thus a better clustering. However, WSS always will get smaller with an increasing number of clusters. Use these measures here for comparison chart 2 shows results obtained over 10,000 data points from a stream from the KDD CUP-99 dataset with respect to WSS.

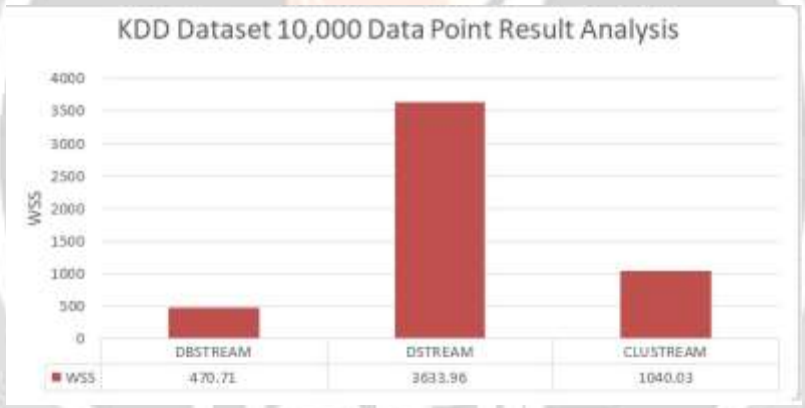

Chart -2 : WSS of Micro clusters on KDD Cup-99 Dataset with 10000 data points.

5. CONCLUSIONS

Traditional methods of clustering work effectively and efficiently to identify cluster on data stream. WE introduced data stream clustering algorithm using Micro-clusters. The algorithm process in two phases online and offline.In online phase micro clusters are maintained while in offline phase that are used for reclustering .New concept Shared density graph is also introduced which helps to gain quality clusters by removing weak clusters. Problems which are studied in existing system will be minimized by this approach. It is possible now to deal with real time data stream clustering within reasonable time and with limited memory. DBSTREAM algorithm performs better as compared to other clustering algorithm on real time dataset like KDD Cup-99.

5. ACKNOWLEDGEMENT

I would like to express my sincere thanks to my Guide Prof.N.M.Shahane and also Prof. Dr. S. S. Sane, Head of Department, Computer Engineering, K.K.W.I.E.E.R. for constant encouragement and support throughout our project, especially for the useful suggestions given during the course of project and having laid down the foundation for the success of this work and Prof. Dr. K. N. Nandurkar, Principal of K.K.W.I.E.E.R. for giving me this opportunity. I express my sincere thanks to all Professors of department for their unfailing inspiration.

I also acknowledge with a deep sense of reverence, my gratitude towards my parents, who has always supported me morally. At last but not the least, gratitude goes to all of my friends indirectly helped me to complete this project.

6. REFERENCES

[1] C. Aggarwal, Data Streams: Models and Algorithms, (series Advances in Database Systems). New York, NY, USA: Springer-Verlag, 2007.

[2] A. Amini and T. Y. Wah, "Leaden-stream: A leader density-based clustering algorithm over evolving data stream," J. Comput. Commun.,vol. 1, no. 5, pp. 26–31, 2013.

[3] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional data streams," in Proc.Int. Conf. Very Large Data Bases, 2004, pp. 852–863.

[4] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: Theory and practice,"IEEE Trans. Knowl. Data Eng., vol. 15, no. 3, pp. 515–528,Mar. 2003.

[5] L. Ertoz, M. Steinbach, and V. Kumar, "A new shared nearest neighbor clustering algorithm and its applications," in Proc. Workshop Clustering High Dimensional Data Appl. 2nd SIAM Int. Conf. Data Mining, 2002, pp. 105–115.

[6] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in Proc. SIAM Int. Conf.Data Mining, 2006, pp. 328–339.

[7] A. Hinneburg, E. Hinneburg, and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in Proc. 4th Int. Conf. Knowl. Discovery Data Mining,1998, pp. 58–65.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 1996, pp. 226–231.

[9] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams," in Proc. ACM Symp. Found. Comput. Sci., 12–14 Nov. 2000, pp. 359–366.
Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2007, pp. 133–142.

[10] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proc. Int. Conf. Very Large Data Bases, 2003, pp. 81–92.