Survey Paper on Clustering of High Dimensional Data Streams

C Kondaiah*1, Dr.P.Chandra Sekhar*2
[1] M.Tech Student, Department of CSE, JNT University, Ananthapur, AP.
[2] Department of CSE, SK University, Ananthapur, AP.

ABSTRACT:

The data stream problem has been studied extensively in recent years because thecollection of streaming data is very easy. So Clustering of streaming data is essential for classification and decision making. Yet, a lot of stream data is high dimensional in nature. Finding clustering in high dimensional data is a difficult task because of high dimensional data comprises hundreds of attributes.Density-based clustering algorithms treat clusters as the dense regions it's useful for the clustering of High dimensional data than conventional algorithms. Propose a new, high dimensional, projected data stream clustering method, called HPStream method. The method is implementing by combining a fading cluster structure, and the projection based clustering methodology.

Index Terms—*Clustering Data Streams, High Dimensional Data, projected clustering, High Dimensional Data Mining*

1. INTRODUCTION

Data streams are very important in recent years because of data streams constantly generate more data, to make this information/data understandable is important, it has to be processed.So clustering of data streams is important. However, a lot of stream data is high dimensional in nature. High-dimensional data is inherently more complex in clustering, classification, and similarity search.

Many applications of clustering are characterized by high dimensional data where each object is described by hundreds or thousands of attributes. Typical examples of high dimensional data can be found in the areas of computer vision applications, pattern recognition, and molecular biology, CAD (Computer Aided Design) databases. One example is the following basic facial recognition algorithm(Fig.1.)[7].

FIG 1:Facial Recognition

Let's look at the concrete example of a picture [8]:

FIG 2: Storage of Digital Cameraimage[8]

Suppose you have "**n**" images, each with a resolution of "**m**" pixels by "**k**" pixels. We can view each pixel within the image as a variable so that each of the n images resides in an m x k dimensional space. From there a training set of images is used to recognize new faces. Depending on the application and the images, we may be able to represent the training/new images with lower dimensions.

In the data stream domain, the clustering problem requires a process which can continuously determine the dominant clusters in the data without being dominated by the previous history of the stream. The high-dimensional case presents a special challenge

to clustering algorithms even in the traditional domain of static data sets. Some recent work on high-dimensional data uses techniques for projected clustering which can determine clusters for a specific subset of dimensions [1]. Of course, these subsets of dimensions may vary over the different clusters. Such clusters are referred to as projected clusters.

2. LITERATURE SURVEY

1. Evolution-Based Clustering of High Dimensional Data Streams with Dimension Projection:

Authors: Chairukwattana R., Kangkachit T., Rakthanmanon T., Waiyamai K

SE-Stream is an evolution-based stream clustering method that supports high-dimensional data streams. SE-Stream is able to monitor and detect a change in the clustering structure during the progression of data streams. SE-Stream reduces the complexity of stream processing by determining a suitable subset of dimensions of each active cluster to express cluster specific characteristics during the progression of data streams. With the elimination of redundant operations, SE-Stream is improved both in terms of cluster quality and execution time.

The authorsChairukwattana R., Kangkachit T., Rakthanmanon T., Waiyamai K in [4], SE-Stream algorithm is improved for effectively clustering over high-dimensional data streams. It implemented with E-Stream, the ability to monitor and detect a change of clustering structure is still continued. SE-Stream improves the dimension projection technique of its concentrates only on active clusters during the streams progression. Because of active clusters contain an enough number of members and they can be described by a small number of projected dimensions. It will also do, several redundant operations for determining cluster merge or split areeliminated. As result, the complexity of SE-Stream is decreased as well as its computation time.

2. Density-Based Clustering over an Evolving Data Stream with Noise

Authors: Feng Cao, Martin Ester,WeiningQian,Aoying Zhou

Clustering is an important task in mining developing data streams. Presentlylot of clustering algorithms for data streams. Present a dense Stream, a new approach for discovering clusters in an evolving data stream. The “dense” micro-cluster (named core-micro-cluster) is introduced to summarize the clusters with arbitrary shape, while the micro-clusters are merged created as macro clusters based on their similarity.

The authors Feng Cao, Martin Ester,WeiningQian,Aoying Zhou in [5], proposed Dense Stream, an effective and efficient method for clustering an evolving data stream. The method can discover clusters of arbitrary shape in data streams, and it is insensitive to noise. The structures of p-micro clusters and o-micro clusters maintain sufficient information for clustering, and a novel pruning strategy is designed to limit the memory consumption with aprecision guarantee. Our experimental performance evaluation over a number of real and synthetic data sets.

3. Density Micro-Clustering Algorithms on Data Streams: A Review

Authors: AminehAmini, Teh Ying Wah

Data streams are large, fast-changing, and in-finite. Applications of data streams can vary from important scientific and astronomical applicationsto crucial business and financial ones. Applications want algorithms to make a single pass with confined time and memory. Mining information streams are concerned with extracting knowledge systems represented in models and patterns in non-stopping data streams. Clustering is an outstanding task in mining data streams, which organization similar objects in a cluster. Several clustering algorithms have been introduced in recent years for data streams which can be primarily based on distance, they can only discover simplest spherical shapes. Therefore, density-based clustering algorithms are adopted for records streams with theability for not only discovering the arbitrary shape clusters but additionally for providing protection towards the outliers. In fact, in density-based clustering algorithms, dense regions of items within the information space are taken into consideration as clusters, which might be isolated by using low-density location (noise). However, in the clustering data streams, due to positive characteristics, it is impossible

to document all of the records. Microclusters are a method in movement clustering that maintains the compact data about the records objects in data streams. The micro cluster is a temporal extension of the clustering characteristic, which compresses the records correctly.

The authors Amineh Amini, Teh Ying Wah in [12], intend to study the amazing density-basedclustering algorithms on data streams the usage of micro-clusters. In this method, algorithm characteristics and examine their merits and limitations.Clustering data streams place additional constraints on clustering algorithms. Data streams require algorithms to make an unmarried pass over the information with bounded memory and constrained processing time, whereas the move may be especially dynamic and evolve through the years. Several clustering algorithms are added for data streams which are distance based and can't take care of the interwoven clusters. Besides that, saving the data streams is impossible, because of the limitless characteristic. Consequently, micro-cluster is introduced to a record of summary data. Here discover four density-based clustering algorithms using micro-clusters. These algorithms utilize the density-based clustering because of ability to discover any shape clusters and micro-clusters as a general summarization of incoming data streams for fixing data mining problems on streams. The algorithms are two phase, online and offline, in which the online phase maintains the micro clusters and offline phase generate the final clusters based on DBSCAN**.**

4. Clustering Data Streams Based on Shared Density between Micro-Clusters:

Authors: Michael Hahsler, Matthew Bolanos

Nowadays data is received automatically from many different kinds of equipment's like Smart Mobiles, sensor, Satellites are just a few of them. It has to be processed, so clustering of data streams is an important technic for data classification data and engineering (decision making). Microclustersare summarized the data stream in real-time with an online process from alarge number of data sets. Micro-clusters represent local density estimates by aggregating the information of many data points in a defined area[Fig 3.a]. On demand, a (modified) conventional clustering algorithm is used in a second offline step to re-cluster the micro-clusters into larger final clusters. For re-clustering, the centers of the micro-clusters are used as pseudo points with the density estimates used as their weights. Here describes DBSTREAM, the first micro-cluster-based online clustering component that explicitly captures the density between micro-clusters via a shared density graph.

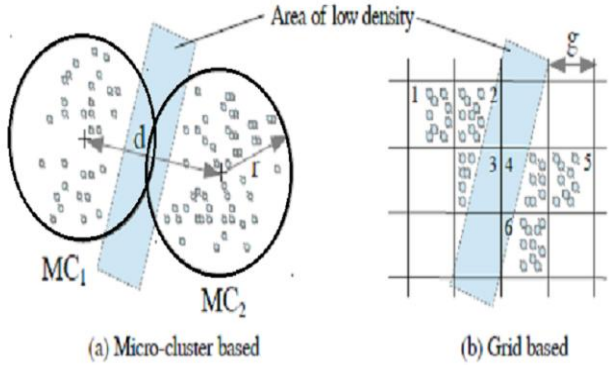

Fig 3. Problem with re-clustering when dense areas are separated by Small areas of low density with (a) micro clusters and (b) grid cells

The authors Michael Hahsler, Matthew Bolanos in [2], developed the first data stream clustering algorithm which expressed records the density in the area shared by micro-clusters and this information is used for re-clustering[Fig 3]. Here introduced the shared density graph [Fig 3]with the algorithms needed to maintain the graph in the online component of a data stream mining algorithm. Although they showed that the worst-case memory requirements of the shared density graph grow extremely fast with data dimensionality, complexity analysis and experiments reveal that the procedure can be effectively applied to data larger number of smaller micro-clusters to achieve comparable results.

5. A New Shared Nearest Neighbor Clustering Algorithm and its Applications:

Authors: LeventErtöz, Michael Steinbach, Vipin Kumar

Clustering depends critically on density and distance (similarity), but these concepts become increasingly more difficult to define as dimensionality increases. In the[5] authors offer definitions of density and similarity that work well for high dimensional data

(actually, for data of any dimensionality). In particular, authors use a similarity measure that is based on the number of neighbors that two points share, and define the density of a point as the sum of the similarities of a point's nearest neighbors. Authors then present a new clustering algorithm that is based on these ideas. This algorithm eliminates noise (low-density points) and builds clusters by associating non-noise points with representative or core points (high-density points). This approach handles many problems that traditionally plague clustering algorithms, e.g., finding clusters in the presence of noise and outliers and finding clusters in data that has clusters of different shapes, sizes, and density. authors have used our clustering algorithm on a variety of high and low dimensional data sets with good results, but authors of [5] present only a couple of examples involving high dimensional data sets: word clustering and time series derived from NASA Earth science data.

The authorsLeventErtöz, Michael Steinbach, Vipin Kumar in[6] have introduced a new clustering algorithm which combines a number of ideas to overcome many of the challenges traditionally plague clustering algorithms, e.g., finding clusters in the presence of noise and outliers and finding clusters in data that has clusters of different shapes, sizes, and density. In addition, our clustering approach works well for high dimensional data, where the concepts of distance and density are often ill-defined. To overcome the problems with distance in high dimensionality, LeventErtöz, Michael Steinbach, Vipin Kumar use a distance measure which is based on the number of neighbors that two points share. To handle problems with density, LeventErtöz, Michael Steinbach, Vipin Kumar define the density of a point as the sum of the similarities of a point's nearest neighbors. The key aspects of our clustering algorithm, besides its use of more effective notions of density and distance, are that it eliminates noise (low-density points) and builds clusters by associating non-noise points with representative or core points (high-density points).

6. Density-Based Clustering for Real-Time Stream Data:

Authors: Y. Chen and L. Tu.

Existing data-stream clustering algorithms such as Clustering Stream are based on k-means. These clustering algorithms are incompetent to find clusters of arbitrary shapes and cannot handle outliers. Further, they require the knowledge of k and user-specified time window. To address these issues, this paper proposes D-Stream, a framework for clustering stream data using a density-based approach. The algorithm uses an online component which maps each input data record into a grid and an offline component which computes the grid density and clusters the grids based on the density. The algorithm adopts a density decaying technique to capture the dynamic changes of a data stream. Exploiting the intricate relationships between the decay factor, data density, and cluster structure, our algorithm can efficiently and effectively generate and adjust the clusters in real time. Further, a theoretically sound technique is developed to detect and remove sporadic grids mapped to by outliers in order to dramatically improve the space and time efficiency of the system. The technique makes high-speed data stream clustering feasible without degrading the clustering quality. The experimental results show that our algorithm has superior quality and efficiency, can find clusters of arbitrary shapes, and can accurately recognize the evolving behaviors of real-time data streams.

The authors Y. Chen and L. Tu [10] propose D-Stream, a new framework for clustering real-time stream data. The algorithm uses an on- line component which maps each input data record into a grid and an offline component which computes the density of each grid and clusters the grids using a density-based algorithm. In contrast to previous algorithms based on k-means, the proposed algorithm can find clusters of arbitrary shapes, automatically determine the number of clusters, and is immune to outliers. The algorithm also proposes a density decaying scheme that can effectively adjust the clusters in real time and capture the evolving behaviors of the data stream. Further, a sophisticated and theoretically sound technique is developed to detect and remove the sporadic grids in order to dramatically improve the space and time efficiency without affecting the clustering results. The technique makes high-speed data stream clustering feasible without degrading the clustering quality.

III.CONCLUSION

In recent years, the management and processing of High Dimensional data streams has become a subject of dynamic research in numerous

fields of computer science such as, e.g., database systems, and data mining. Lot of research work has been carried in this field to develop an efficient clustering algorithm for High Dimensional data streams. High Dimensional data are frequently large and may contain outliers. Therefore, careful examination of the earlier proposed algorithms is necessary. In this paper we surveyed the current studies on High Dimensional Data clustering. These studies arestructured into many categories based upon clustering. Most clustering algorithms are not capable to make High dense clusters. In addition, this paper discusses about possible high dimensional data clustering algorithms. Finally, this study may promote the development of new High dimensional data mining methods, such as fading cluster structure and projection based clustering.

IV. REFERENCES

[1]. SunitaJahirabadkar, Parag Kulkarni., "Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 63– No.20, February 2013.

[2]. Michael Hahsler, MatthewBola˜nos., "Clustering Data Streams Based on Shared Density Between Micro-Clusters", IEEE Transactions On Knowledge And Data Engineering — Preprint, Accepted 1/17/2016.

[3]. Lance Parsons, EhteshamHaque, Huan Liu., "Subspace Clustering for High Dimensional Data: A Review",ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets: Volume 6 Issue 1, June 2004.

[4].Chairukwattana R., Kangkachit T., Rakthanmanon T., Waiyamai K., "Evolution-Based Clustering of High Dimensional Data Streams with Dimension Projection", Knowledge and Systems Engineering. Advances in Intelligent Systems and Computing, vol 245. Springer, 2014.

[5]. Feng Cao, Martin Ester,WeiningQian,Aoying Zhou., "Density-Based Clustering over an Evolving Data Stream with Noise", SIAM International Conference on Data Mining,2006.

[6].LeventErtöz, Michael Steinbach, Vipin Kumar., "A New Shared Nearest Neighbor Clustering Algorithm and its Applications".,Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining,(2002)

[7] 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.

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

[9] J. Gama, Knowledge Discovery from Data Streams, 1st Ed. London, U.K.: Chapman & Hall, 2010.

[10] 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.

[11] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, "Density-based clustering of data streams at multiple resolutions," ACM Trans. Knowl. Discovery from Data, vol. 3, no. 3, pp. 1–28, 2009.

[12] AminehAmini, Teh Ying Wah., "Density Micro-Clustering Algorithms on Data Streams: A Review", Proceedings of the international multiconference of Engineers and scientists 2011, vol1,IMESC, March 16-18-2011,Hong Kong

[13].http://en.wikipedia.org/wiki/Eigenface.

[14]. http://shoefer.github.io/intuitivemi/2015/07/19/data-numbers-representations.html.