ISSN 2321 3361 © 2017 IJESC

Research Article Volume 7 Issue No.7

Survey on Micro Clustering Data Streams using Agglomerative Approach

M.Kavitha[1], R.Baby[2]
Assistant Professor[1], M.Phil Research Scholar[2]
Department of Computer Science
Tiruppur Kumaran College for Women, Tirupur, Tamilnadu, India

Abstract:
Data stream mining is an active research area that has recently evolved to mine knowledge from large amounts of continuously generated data. In this context, surveys of several data stream clustering methodologies have been proposed to perform micro clustering. Data stream clustering imposes several challenges to be rectified, such as dealing with non-stationary, unbounded data that arrive in an online fashion. The intrinsic nature of stream data requires the development of algorithms capable of fast and incremental processing of data objects, suitably addressing time and memory limitations. Most of the micro clustering algorithms are object based. In this research paper, we have presented a small survey of the data stream clustering algorithms. This clustering approach has much advantage over many areas such as market analysis (stock market), detection of crime etc,. In addition, this work presents an overview of the usually-employed experimental methodologies. This context may help future researchers on the data stream clustering.

Keywords: DataStream Clustering, Density-based Clustering Micro-clustering, Stream Algorithm.

I. INTRODUCTION:

In recent years, upcoming trends in hardware technology have allowed us to automatically record transactions and other pieces of information of everyday life at a fast rate. Such processes generate large amounts of online data which grow at an unlimited rate. These kinds of online data are called as data streams. The statistical information can be controlled about in the form of micro-clusters. Object-based data stream clustering algorithms can classified into two steps. First one is data abstraction step (that is online component) and second one clustering step(that is offline component). The online abstraction step summarizes the data stream with the use of particular data structures for dealing with space and memory constraints of stream applications. These data structures summarize the stream in order to secure the concept of the original objects without the need of actually storing them..For summarizing the continuously-arriving stream data and, at the same time, for giving greater importance to up-to-date objects, a powerful approach is object based data stream clustering. After completing the data abstraction step, data stream clustering algorithms acquire a data partition via an offline clustering step (offline component). The offline component is used in relation with a wide variety of inputs (e.g., time horizon, and number of clusters) to provide a quick understanding of the broad clusters in the data stream. Data stream clustering algorithms should adopt anomaly detection mechanisms that are able to differentiate between actual anomalies and cluster evolution, considering that the data stream distribution may vary over time. Applications of data streams will include mining data generated by sensor networks, meteorological analysis, stock market analysis, and computer network traffic monitoring etc,. For data stream mining, data objects should arrive continuously, there should be no control over the order in which the data objects should be processed.

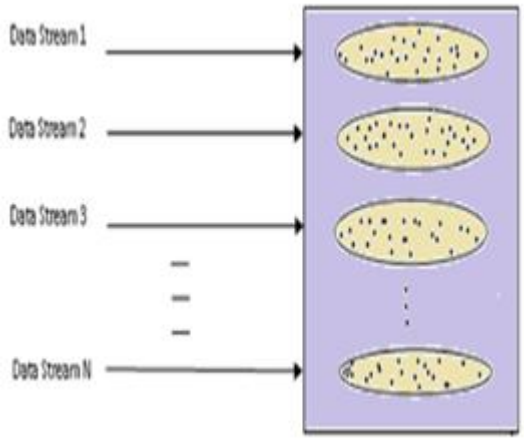

Figure.1. Clustering Data Streams

II. A FRAMEWORK FOR CLUSTERING EVOLVING DATA STREAMS:

Charu C. Aggarwal, Jiawei Han..en[1] has proposed significantly various philosophy for data stream clustering. The idea is partition the clustering process into an online component which periodically saves detailed summary statistics. The online component is utilized by the analyst who can use a wide variety of inputs (number of clusters, time horizon) in order to give a quick understanding of the broad clusters in the data stream. The problems of efficient choice, storage, and use of this statistical data for a fast data stream turns out to be quite tedious. For this purpose, we use the concepts of a pyramidal time frame in

coincidence with a micro-clustering approach. Our performance experiments over a number of real and synthetic data sets elaborates the electiveness, efficiency and insights provided by our approach.

III. DATA STREAM CLUSTERING: A SURVEY

Jonathan A. Silva, Elaine R.Faria,.. undergone a detailed survey of data stream clustering algorithms, by giving a thorough discussion of the important design components of state-of-the-art algorithms. In this survey lot of references is given that eloborates the uses of data stream clustering in various domains, such as sensor networks, network intrusion detection etc,. Information about software packages and data repositories is also available for helping researchers and practitioners. Lastly, some important problems and open questions that can be subject of future research are discussed.

IV. DENSITY-BASED CLUSTERING OVER AN EVOLVING DATA STREAM WITH NOISE:

Feng Cao, Martin Ester..en[3] has proposed Clustering is an important task in mining including data streams. In this work, they present *Den Stream*, a new method for discovering clusters in an evolving data stream. The "dense" micro-cluster is introduced to summarize the clusters with arbitrary shape, while the potential core-micro-cluster and outlier micro-cluster structures are proposed to maintain and differentiate the potential clusters and outliers. A novel pruning strategy is created based on these concepts, which guarantees the precision of the weights of the micro-clusters with limited memory.

V. DENSITY-BASED CLUSTERING FOR REAL-TIME STREAM DATA:

Yixin Chen, Li Tu [4] has proposed Existing data-stream clustering algorithms such as CluStream which are based on k-means. The algorithm contains an online component which matches each input data record into a grid and an offline component which calculates the grid density and then it clusters the grid based on the calculated density. The algorithm then uses a density decaying technique to capture the dynamic changes of a data stream.. Further, a theoretically sound technique is created to detect and reject sporadic grids mapped to by anomalies in order to dramatically increase the space and time efficiency of the system.. The experimental results figure out that their algorithm has superior quality and efficiency, can find clusters of arbitrary shapes, and can accurately recognize the evolving behaviors of real-time data streams.

VI. AN EFFICIENT APPROACH TO CLUSTERING IN LARGE MULTIMEDIA DATABASES WITH NOISE

Alexander Hinneburg, Daniel A. Keim[6] has proposed the effectiveness and efficiency of the existing algorithms, however, is somewhat limited, while clustering in multimedia databases needs clustering high-dimensional feature vectors and since multimedia databases often contain large amounts of noise. In the work, they introduce a new algorithm to clustering in multimedia databases known as DENCLUE (Density- based Clustering). The basic idea of their new approach is to develop the overall point density empirically as the sum of incense functions of the data points. The benefits of their new approach it has the advantage that is better properties of clustering in data sets (with large amounts of noise), it does not deny a solid mathematical description of arbitrarily shaped clusters in high-dimensional data sets.

VII. A FRAMEWORK FOR PROJECTED CLUSTERING OF HIGH DIMENSIONAL DATA STREAMS:

Charu C. Aggarwal, Jiawei Han..en [8] has proposed the nature of stream data makes it essential to use algorithms which needs only one pass over the data. Recently, single-scan, stream analysis methods have been proposed in this con- text. However, a lot of stream data is high- dimensional in nature. High-dimensional data is intrinsically more difficult in clustering, classification, and similarity search. In the work, they propose a new, high- dimensional, projected data stream clustering method, called HPStream. The approach incorporates a fading cluster structure, and the projection based clustering methodology.. their performance study with both real and synthetic data sets demonstrates the efficiency and effectiveness of their proposed framework and implementation methods.

VIII. UNSUPERVISED CLUSTERING IN STREAMING DATA:

Dimitris K. Tasoulis, Niall M. Adams..en [9] has proposed tools for automatically clustering streaming. In the work they present an extension of conventional kernel density clustering to a spatio-temporal setting, and also develop a novel algorithmic scheme for clustering data streams. Experimental results describe both the high efficiency and other benefits of this new approach.

IX. TEMPORAL STRUCTURE LEARNING FOR CLUSTERING MASSIVE DATA STREAMS IN REAL-TIME:

Michael Hahsler, Margaret H. Dunham[11] proposed a work which describes one of the first attempts to model the temporal structure of massive data streams in real-time using data stream clustering.. In the work they represents a new framework called Temporal Relationships Among Clusters for Data Streams which allows us to understand the temporal structure during clustering a data stream. they identify, organize and describe the clustering operations which are used by state-of-the-art data stream clustering algorithms. Then they show that by designing a set of new operations to transform Markov Chains with states representing clusters dynamically, they can efficiently capture temporal ordering information.

X. SOSTREAM: SELF ORGANIZING DENSITY-BASED CLUSTERING OVER DATA STREAM:

Charlie Isaksson, Michael Hahsler..en[12] has proposed in the work they propose a data stream clustering algorithm, called Self Organizing density based clustering over data Stream (SOStream). This algorithm has several novel features. Instead of applying a fixed, user defined similarity threshold or a static

grid, SO Stream detects structure within quick evolving data streams by automatically adapting the threshold for density-based clustering. It also adopts a novel cluster updating rule which is inspired by competitive learning methods created for Self Organizing Maps (SOMs).

X. ESSENTIALS OF THE SELF-ORGANIZING MAP:

Teuvo Kohonen[13] has proposed the self-organizing map (SOM) is an automatic data-analysis method. It is widely used to clustering problems and data exploration in industry, finance, natural sciences, and linguistics. The most extensive applications, exemplified in the work, can be found in the management of massive textual databases and in bioinformatics. The SOM is related to the classical vector quantization (VQ), which is applied extensively in digital signal processing and transmission. These models are automatically related with the nodes of a regular (usually two-dimensional) grid in an orderly fashion such that more similar models become automatically related with nodes that are adjacent in the grid, whereas less similar models are placed farther away from each other in the grid. This organization, a kind of similarity diagram of the models, makes it possible to gain an insight into the topographic relationships of data, especially of high-dimensional data items.

XI. INTRODUCTION TO STREAM: AN EXTENSIBLE FRAMEWORKFOR DATA STREAM CLUSTERING RESEARCH WITH R:

Michael Hahsler, Matthew Bolognas [14] has proposed the data streams are ordered and probably confounded sequences of data points made by a typically non-stationary data generating process. Common data mining tasks related with data streams include clustering, classification and frequent pattern mining. New algorithms for these types of data are developed regularly and it is important to evaluate them thoroughly under standardized conditions. In this research they introduce stream, a research tool that contains modeling and simulating data streams.. The main benefit of stream is that it seamlessly colloborates with the large existing infrastructure given by R. In addition to data handling, plotting and easy scripting capabilities, R also provides existing algorithms and allows users to interface code written in many programming languages familiar among data mining researchers (e.g., C/C++, Java and Python). In this work they describe the architecture of stream and focus on its use for data stream clustering research. stream was developed with extensibility in mind and will be extended in the future to cover additional data stream mining tasks like classification and frequent pattern mining.

XII. EFFICIENT ONLINE EVALUATION OF BIG DATA STREAM CLASSIFIERS:

Albert Bifet, Gianmarco..en [15]has proposed the evaluation of classifiers in data streams is fundamental so that poorly-performing models can be found, and improved better-performing models. This is an increasingly relevant and important task as stream data is generated from more sources, in real-time, in large quantities, and is now considered the biggest source of big data. Both researchers and practitioners need to be able to effectively evaluate the performance of the methods they adopt. However, there are many challenges for evaluation in a stream. Current frameworks for calculating streaming and online algorithms are able to give predictions in real-time, but as they use a prequential setting, they build only one model, and are thus not able to compute the statistical significance of results in real-time. In the work they propose a new evaluation methodology for big data streams. This methodology addresses unbalanced data streams, data where change occurs on various time scales, and the question of how to divide the data between training and testing, over multiple models.

XIII. CONCLUSION:

Clustering data including data streams becomes an substantial technique for data and knowledge engineering projects. A data stream is an ordered and unbounded sequence of data points. Such streams of constantly arriving data are generated for many types of applications include GPS data from smart phones, web clickstream data, computer network monitoring data, tele-communication connection data, etc. This survey produces a short summary of the data stream clustering problems. This research work will be definetly useful for the future researchers on data stream clustering.

XIV. 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]. J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, 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.

[3]. 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.

[4]. Y. Chen and L. Tu, "Density-based clustering for real-time stream data," New York, NY, USA: ACM, 2007, pp. 133–142.

[5]. L. Tu and Y. Chen, "Stream data clustering based on grid density and attraction," ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–27, 2009.

[6]. A. Hinneburg, E. Hinneburg, and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," AAAI Press, 1998, pp. 58–65.

[7]. S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: Theory and practice," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, pp. 515–528, 2003.

[8]. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional data streams," in Proceedings of the International Conference on Very Large Data Bases (VLDB '04), 2004, pp. 852–863.

[9]. D. Tasoulis, N. Adams, and D. Hand, "Unsupervised clustering in streaming data," in IEEE International Workshop on Mining Evolving and Streaming Data.

[10]. D. K. Tasoulis, G. Ross, and N. M. Adams, "Visualizing the cluster structure of data streams," in Advances in Intelligent Data Analysis VII, ser.

[11]. M. Hahsler and M. H. Dunham, "Temporal structure learning for clustering massive data streams in real-time,".

[12]. C. Isaksson, M. H. Dunham, and M. Hahsler, "Sostream: Self organizing density-based clustering over data stream," in Machine Learning and Data Mining in Pattern Recognition, ser.

[13]. T. Kohonen, "The self-organizing map," Neurocomputing, vol. 21, pp. 1–6, 1998

[14]. M. Hahsler, M. Bolanos, and J. Forrest, stream: Infrastructure for Data Stream Mining, 2015, R package version 1.2-2.

[15]. A. Bifet, G. de Francisco Morales, J. Read, G. Holmes, and B. Pfahringer, "Efficient online evaluation of big data stream classifiers," .