Journal Homepage: -www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/3673
DOI URL: http://dx.doi.org/10.21474/IJAR01/3673

RESEARCH ARTICLE

DATA STREAM CLUSTERING ISSUES AND CHALLENGES-A SURVEY

B. Rupa and R. Soujanya.
Assistant Professor, Department of CSE, GRIET, Hyderabad.

Manuscript Info

Manuscript History

Received: 11 January 2017
Final Accepted: 10 February 2017
Published: March 2017

Key words:-
Data Mining, Data Stream, Clustering

Abstract

In recent years, advances in both hardware and software technology has allowed us to automatically record transactions and other information everyday at a rapid rate. Huge volumes of web, sensory and transactional data are continuously generated everyday as data streams, which need to be analyzed online as they arrive. Analysis of data streams have been researched extensively because of its emerging, imminent, and broad applications. One of the important method is clustering have been widely studied in the data mining community. Many existing data mining methods cannot be applied directly on streaming data because of the fact that the data needs to be mined in single pass. Furthermore, in data stream processing temporal locality is also quite important, because the essential patterns in the data may change and therefore, the clusters in the past history may no longer remain relevant to the future. In this paper we explore various issues and challenges on clustering data streams.

Introduction:-

Nowadays, many applications with huge amount of data which are caused limitation in data storage capacity and processing time. Traditional data mining algorithms are not suitable for this kind of applications so they should be changed or designed with new algorithms.

The main challenge is the *data-intensive* mining using a bounded number of resources to perform an analysis to an unlimited dataset. Furthermore, traditional Data Mining Algorithms work with a static dataset and the algorithm can afford to read the data several times on the other hand Stream Mining can afford the data reading once so the algorithms for this subfield of data mining are based on a single scan.

The data generation rates in somedata sources become faster than ever before. This fast generation of continuous streams of information has challenged our storage space, computation and communication capabilities in computing systems. Models and techniques have been proposedand developed over the past few years to address these challenges [3].

A data stream is a massive, infinite, temporally ordered, continuous and rapid sequence of data elements [1]. Research on data stream is motivated by emerging applications involving massive data sets such as customer click streams, supermarket, telephone records, stock market, meteorological research, multimedia data, scientific experiments data and engineering experiments data and sensor type of data. A new generation of mining

Corresponding Author: -B. Rupa.
Address: -Assistant Professor, Department of CSE, GRIET, Hyderabad.

algorithms are needed for real–time analysis & query response for these applications since the majority of conventional data mining algorithms can only be applied to static type of data sets that may be updated periodically in big chunks, but not in continuous streams of data. While data mining has become a reasonably well recognized field now, the data stream problem poses a no. of unique challenges which are not easily solved by traditional data mining methods.

Some of issues of data stream [2] like dynamic nature, Infinite size and high speed, unbounded memory requirements, Lack of global view, handling the continuous flow of data impose a great dispute for the researchers dealing with streaming data sets. Unlike traditional data sets, it is impossible to store an entire data stream or to scan through it multiple times due to its incredible volume. New methods may keep on growing in data streams at different times, to deal with this any data stream processing algorithm must continuously update their models to adapt to the changes.

Data streams pose special challenges to several data mining algorithms, not only because of the huge volume of the online data streams, but also because of the fact that the streaming data may show temporal association. Such temporal association may help disclose important data evolution characteristics, and they can also be used to develop efficient and effective mining algorithms [3]. Moreover, data streams require online mining, in which we wish to mine the data in a continuous fashion. Furthermore, the system needs to have the capability to perform an offline analysis as well based on the user interests. This is similar to an online analytical processing (OLAP) framework which uses the paradigm of pre-processing once, querying many times [5].

Rest of the paper is organized as follows. Section 2 presents types of data streams, its characteristics, applications. Overviews of data stream clustering techniques are presented in section 3. Research issues and challenges are discussed in section 4. Finally, section 5 summarizes this paper.

Data Streams: -

A data stream is a real-time, continuous, ordered sequence of items. It is not possible to control the order in which items appear nor is it feasible to locally store a stream in its entirety [7].There are two kinds of data streams, transactional data streams and measurement data streams [8].

Transactional data streams: These are the data streams which recorded the communication between data attributes, for example; purchasing item details in credit card, phone calls details of callers to dialed parties.

Measurement data streams: Consists of data from sensors on things of interest. These types of data streams which monitored the changes of entity states, for example; traffic information at router interfaces, weather forecasting data at weather stations and traffic data in sensor networks.

Characteristics [4] of data streams as follows:-

- Huge volumes of continuous data, possibly infinite
- Quick changing and requires high-speed, real-time response
- Data stream captures properly our needs of data processing today
- Random access is expensive—single scan algorithm
- Store only the summary of the data
- Most streaming data are at pretty low-level or multi-dimensional in nature, needs multi-level and multi-dimensional processing

Some of the applications evolving data streams:-

- Telecommunication calling records
- Transaction flows of credit card
- Network traffic flow and monitoring
- Stock exchange details in financial market
- Power supply and manufacturing details
- Sensor, monitoring and surveillance flows
- Video streams, RFIDs
- Security monitoring details
- Web logs and Web page click streams

- Massive amount of data sets

There are some of the main differences between Traditional data &Data Stream [9] shows in Table-1:

Table 1:-Traditional Vs Stream

	Traditional	**Stream**
No. of passes	Multiple	Single
Processing time	Unlimited	Restricted
Memory Usage	Unlimited	Restricted
Types of Results	Accurate	Approximate
Distributed	No	Yes

There are some of the differences between Traditional Data base Management System (DBMS) & Data Stream Management (DSMS) [10] shows in table-2:

Table 2:-Traditional DBMS Vs DSMS.

DBMS	**DSMS**
Random access	Sequential access
No real time services	Real time requirements
Unbounded Disk store	Bounded main memory
Access plan determined by query processor, physical data base design	Unpredictable because of variable data
Data at any granularity	Data at fine granularity
One time queries	Continuous queries

Most data mining and knowledge discovery techniques assume that there is a finite amount of data generated by an unknown, stationary probability distribution, which can be actually stored and evaluated in multiple steps by a batch mode algorithm. For data stream mining, however, the successful development of algorithms has to take into account the following restrictions [5]:

- Data objects arrive continuously;
- There is no control over the order in which the data objects should be processed;
- The size of a stream is (potentially) unbounded;
- After processing, data objects are discarded. In practice, one can store piece of the data for a given period of time, using a forgetting mechanism called to discard them later;
- The unknown data generation process is possibly non-stationary, i.e., its probability distribution may change over time.

Data Stream Clustering:-

Imagine an enormous amount of dynamic stream data, many applications involve the automated clustering of such data into groups based on their similarities. Although there are many powerful clustering algorithms for data sets in static form, clustering data streams involves other constraints on such algorithms, as any data stream model requires algorithm to make a single scan over the data, with restricted memory and limited processing time. Several algorithms have been developed for clustering data streams described as below:

STREAM –A k-median based Stream Clustering Algorithm is presented by Guha, Mishra, Motwani and O'Callaghan [11]. It consists of two phases and follows divide and conquer approach. In first phase, it divides the data streams in relevant buckets and then finds 'k' clusters in each bucket by applying k-median clustering algorithm. It stores cluster centers only and cluster centers are weighted based on the number of data points belongs to corresponding cluster and then ignore the data points. In second phase weighted cluster centers are clustered in small number of clusters. Though its space and time complexity is low but it cannot use to concept evolution in data.

CluStream [12] clustering evolving data streams is proposed by Aggarwal et al. It divides the clustering process in following two online component and offline components. The summary of data in the form of micro clusters is stored in online component. Micro-cluster is the temporal extension of clustering feature of BIRCH [11]. Summary statistics details of data are stored in snapshots form which gives the user flexibility to specify the

time interval for clustering of micro-clusters. Offline component apply the k-means clustering algorithm to cluster micro-clusters into larger clusters.

ClusTree [13] any time Stream Clustering is proposed by Kranen et a... It divides the clustering process in following two online and offline components. Online components are used to learn micro clusters. The micro clusters are arranged in hierarchical tree type of structure. Any variety of different offline components can be developed. It is a self adaptive algorithm and delivers a model at any time.

DenStream [14], a new approach for discovering clusters in an evolving data stream. The dense micro-cluster called as core-micro-cluster is introduced to summarize the clusters with different arbitrary shape, while the potential core-micro-cluster and outlier micro-cluster structures are projected to maintain and differentiate the potential clusters and outliers. A novel pruning strategy was designed based on these concepts, which guarantees the precision of the weights of the micro clusters with limited memory.

HPStream (clustering of high dimensional data streams) [15] is proposed by Aggarwal et al. This clustering technique uses a Fading Cluster Structure (FCS) to stores the summary of data streams and it gives more significance to recent data by fading the old data with moment in time. To handle high dimensional data selects the subset of dimensions by projecting on original high dimensional streaming data. Number of dimensions and dimensions are not same for each cluster. This is due to the fact that significance of each dimension in each cluster may differ. It is incrementally updatable and highly scalable on number of dimensions. Moreover it cannot find out the cluster of different arbitrary shapes and it requires domain knowledge for identifying the number of clusters and average number of projected dimension parameters.

E-Stream [16] is a data stream clustering technique which supports following five type of advancement in streaming data: Appearance of new cluster, Disappearance of an old cluster, Splitting of a large cluster, combining of two similar type of clusters and change in the behavior of cluster data itself. It uses a fading cluster structure with histogram to approximate the streaming data. Though its performance is better than HPStream clustering algorithm but it needs many parameters to be specified by user.

HUE-Stream [17] extends E-Stream which is described earlier, to support uncertainty in mixed data (heterogeneous data). A distance function with probability distribution is introduced in between two objects to support uncertainty in categorical attributes. To detect change in clustering structure, the proposed distance function is used to combine(merge) similar clusters and find the closest cluster of a given incoming data and proposed histogram management is used to split cluster in categorical data.

POD-Clus (Probability and Distribution-based Clustering) [18] is a model based clustering technique for streaming data. It is applicable to both clustering by example and clustering by variable scenario. For summarizing and updating the cluster information incrementally, it uses a cluster synopsis which comprises the mean, standard deviation, and number of points for each cluster. It maintains concept evolution by allowing new cluster appearance, splitting of a cluster, combining of two similar clusters and disappearance of a cluster.

Issues and Challenges: -

Data stream mining is an inspiring field of study that has raised many challenges and research issues to be addressed by the database and data mining communities.

The following are the some of the research issues and challenges [18, 19]:

Handling the continuous flow of data streams:this isa data management issue. Traditional database management systems are not capable of dealing with such continuous high data rate. Novel indexing approach, storage and querying methods are required to handle this non-stopping fluctuated flow of information streams.

Unbounded memory requirements due tononstop continuousflow of data streams: most of the machine learningmethods represent the main source of data mining algorithms. Most of machine learning methods require data to be resident in memory while executing the algorithm for analysis. Due to the huge amounts of thegenerated streams, it is absolutely a very importantconcern to deign space efficient techniques that can have only one scan or less over the incoming streaming data.

Required result accuracy:design a space and timeefficient techniques should be accompanied withacceptable result accuracy. Approximation algorithms as described earlier can assurance error limits.

Modeling changes of mining results over time: in some cases, the user is not interested in mining datastream results, but how these results are changed over time. If the number of clusters generated for example is changed, it might represent some changes in the dynamics of the arriving stream. Due to dynamic nature of datastreams using changes in the knowledge structuresgenerated would advantage many temporal-based application analysis.

Visualization of data mining results on small screensof mobile devices:conventional data mining visualization results on a desktop is still a research issue. Visualization in small screens of a PDA for example is a real challenge. Imagine a businessman is being streamed and analyzed data on his PDA. Such results should be efficiently visualized in a way that enables.

The specific challenges [20] in the context of data stream clustering scenario are as follows:
• Streams usually have massive volume, and it is often not possible to store the data explicitly on disk. Therefore, the data needs to be processed in a single look, in which all the summary data required for the process of clustering needs to be stored and maintained. The time needed to process each record must be small and constant. Otherwise, the model creation process would never be able to take up with the data stream.

• The patterns in the data stream may continuously evolve over time. From a stream mining perspective, this implies that the underlying clustering models need to be updated continuously. A usable model must be available at any time, because the end of data stream computation may never be reached, and an analyst may need results at any point in time.

• Different domains of data may pose different challenges to data stream clustering. For example, in a massive domain of discrete attributes, it is impossible to store summary representations of the data clusters efficiently without increasing the computational complexity of the problem significantly. Therefore, space-efficient methods need to be designed for massive domain clustering of data streams.

Another challenge that should be handled by data stream clustering algorithms is the ability of properly dealing with outliers, and also of detecting changes in the data distribution. The dynamic nature of evolving data streams, where new clusters often emerge while old clusters fade out, imposes difficulties for outlier detection. In general, new algorithms should provide mechanisms to distinguish between seeds of new clusters and outliers. Regarding the challenge of dealing with non-stationary distributions, the current and naive strategy employed by most available algorithms is to implicitly deal with them through window models.

Even though more robust change detection mechanisms have been implemented in generic frameworks, we believe future data stream clustering algorithms should explicitly provide mechanisms for performing change detection. Dealing with different data types imposes another challenge in data stream clustering. Different data types such as categorical and ordinal values are present within several application domains. In addition, complex data structures like DNA data and XML patterns are largely available, thus a more careful attention should be given to algorithms capable of dealing with different data types.

Conclusion: -

Now a days, continuous massive generation of stream data has lead to new trend in the field of data mining named as Data Stream Mining. So in this paper, we discussed various issues elevated by data streams and to be had an overview of various technology used for generating synopsis data structures from continuous generation of stream data. From our study we can conclude that streaming data evolves immense volumes of temporally data changing so, traditional techniques of data mining cannot be applied straightforwardly. Research in data stream mining is in early stage. If the problems caused by data streams are solved and if more efficient and interactive mining methods which are user friendly are developed, it is likely that in the near future stream mining will play vital role in business world, as it deals with many applications which involves mining from continuous data streams.
Bearing in mind that clustering data streams is a relevant and challenging task, we believe that much effort should be addressed to developing more sophisticate evaluation criteria, high-quality benchmark data, and a sound methodology for reliable experimental comparison of new data stream clustering algorithms.

References:-

1. A.Bifet,G.Holmes, R.Krikbyand B.Pfahringer, Data Stream Mining -A Practical approach,2011.
2. Madjid Khalilian , Norwati Mustapha, " Data Stream Clustering: Challenges and Issues", Proceedings of the International MultiConference of Engineers and Computer Scientists,2010Vol1,IMECS 2010,March 17-19,2010,HongKong.
3. Babcock B., Babu S., Datar M., Motwani R., Widom J. (2002). Models and Issues in Data Stream Systems, ACM PODS Conference.
4. S. Chakravarthy, Q. C. Jiang, "Stream Data Processing: A Quality of Service Perspective" 2009.
5. Domingos P., Hulten G. (2000). Mining High-speed Data Streams. ACM SIGKDD Conference.
6. Guha, Meyerson, Mishra, Motwani, and O'Callaghan. 2003. Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering 15, 515–528.
7. G. Hebrail, "Data stream management and mining", Mining Massive Data Sets for Security, F. Fogelman-Soulié et al. (Eds.), IOS Press, 2008.
8. N. Koudas and D. Srivastava, "Data Stream Query Processing, AT&T", Labs-Research, 2003.
9. B. Babcock, S. Babu, M. Datar, R. Motwani, J. Widom, "Models and Issues in Data Stream Systems", Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 1-16, 2002.
10. M. Lindeberg, "Design, Implementation, and Evaluation of Network Monitoring Tasks for the Borealis Stream Processing Engine", Master's Thesis, University of Oslo, May 2007.
11. L. callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-Data Algorithms for High-Quality Clustering," in Proceedings of IEEE International Conference on Data Engineering,
12. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the 29th international conference on Very large data bases - Volume 29, ser. VLDB '03. VLDB Endowment, 2003, pp. 81–92.
13. Kranen,Assent,Baldauf,Seidl, "Self Adaptive any time clustering",ICMD, 2009
14. F. Cao, M. Estery, W. Qian, A. Zhou, "Density-Based Clustering over an Evolving Data Stream with Noise", SDM, 2006.
15. 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 Thirtieth international conference on Very large data bases – Volume 30, ser. VLDB. VLDB Endowment, 2004, pp. 852–863.
16. K. Udommanetanakit, T. Rakthanmanon, and K. Waiyamai, "E-stream: Evolution-based technique for stream clustering," in Proceedings of the 3rd international conference on Advanced Data Mining and Applications, ser. ADMA '07. Berl in, Heidelberg: Springer-Verlag, 2007, pp. 605–615.
17. W. Meesuksabai, T. Kangkachit, and K. Waiyamai, "Hue-stream: Evolution-based clustering technique for heterogeneous data streams with uncertainty." in ADMA (2), ser. Lecture Notes in Computer Science, vol. 7121. Springer, 2011, pp. 27 –40.
18. P. P. Rodrigues, J. a. Gama, and J. Pedroso, "Hierarchical clustering of time-series data streams," IEEE Trans. on Knowl. and Data Eng., vol. 20, no. 5, pp. 615–627, May 2008.
19. L. Golab and M. T. Ozsu. Issues in Data StreamManagement. In SIGMOD Record, Volume 32, Number 2, June 2003.
20. G. Dong, J. Han, L.V.S. Lakshmanan, J. Pei, H.Wang and P.S. Yu. Online mining of changes from datastreams: Research problems and preliminary results, In Proceedings of the 2003 ACM SIGMOD Workshop onManagement and Processing of Data Streams. In cooperation with the 2003 ACM-SIGMOD International Conference on Management of Data, San Diego, CA, June 8, 2003.
21. C. Aggarwal. On Change Dignosis in Evolving Data Streams. InIEEE TKDE, 17(5), 2005.