ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Data Stream Mining – A Survey

R. Kalaivani[1], Dr. S.Vijayarani[2]

Ph.D Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, India[1]

Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, India[2]

ABSTRACT: At present a growing number of applications that generate massive streams of data which needs intelligent data processing and online analysis. The imminent need for turning such data into useful information and knowledge augments the development of algorithms and frameworks that address streaming challenges. The pre process, storage space, querying and mining of such data sets are very computationally difficult tasks. Time period observation systems, telecommunication methods, sensor devices and alternative dynamic environments are such cases. Mining data streams cares with extracting data structures drawn in model and patterns in continuous streams of information. In this paper, we have a tendency to gift the theoretical foundations of data stream analysis and establish potential directions of future analysis. Mining data stream techniques are being reviewed.

KEYWORDS: Data Stream mining; Data mining; Clustering mining; Classification mining; Association Mining

I. INTRODUCTION

In the field of processing of information, Data mining ascribes to extract useful information from huge volumes of data [1]. In the same way, mining Data Streams refers to extracting information from constant and continuous stream of data. The field of data stream is very recently recognized field [2]. A data stream is a huge, endless, temporally ordered, infinite and agile information [3].

In recent years, the interest in this field has increased resulting in large volumes of literature has been published in past few years. As well as researches have been carried out on data streams mainly motivated by many evolving applications which involve huge volumes of data generated from various domains, example, sensor data, data from supermarkets, telephone logs, data from satellites and various other sources. Traditional approaches are no longer good enough for mining data in today's environment, which require real time analysis and quick responses to queries as the data previously was static and was changing periodically but now data is continuous and rapidly changes hence many new mining algorithms are proposed. Data stream mining has become a mainstream field now. Since the traditional methods cannot solve the data stream issues there are various challenges to solve them some of them are [4] frequently changing dynamic nature, huge volume and speed with which data is generated, memory requirements, handling these continuous flow of data create a bigger problem and challenge for the researchers working on streaming data. In traditional data sets we could store the data and analyse it many times but this cannot be done with data streams due to huge volumes of data. Many new techniques keep evolving to deal; with these issues, the bottom line being that the algorithms must frequently update their models to accommodate the inconsistencies in the data. The main purpose of this survey is to study the various techniques and algorithms for mining data streams.

Information systems are additional advanced, even quantity of information being processed have increased and dynamic in nature, due to common updates. This streaming data has the subsequent characteristics.

- The information arrives always from data streams.
- No statements on data stream ordering may be created.
- The length of the data stream is limitless.

Efficiently and effectively capturing knowledge from data streams has become very critical; which include network traffic monitoring, web click-stream. There is a need of employment of semi-automated interactive techniques for the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

extraction of hidden knowledge and information in the real-time. Systems, models and techniques have been proposed and developed over the past few years to discuss these challenges [5].

More significantly, the traits of the data stream will perform modification over time and therefore the evolving pattern must be recorded. Moreover, difficulty of resource allocation needs to be careful in mining data streams. Because of the huge volume and therefore the high speed of streaming data, stream mining algorithms should handle the results of system burden. Thus, a way to accomplish optimum results from different resource constraints becomes a difficult task. Figure 1 shows the common data stream model. The cyclic method has three main steps happening in a recursive format.

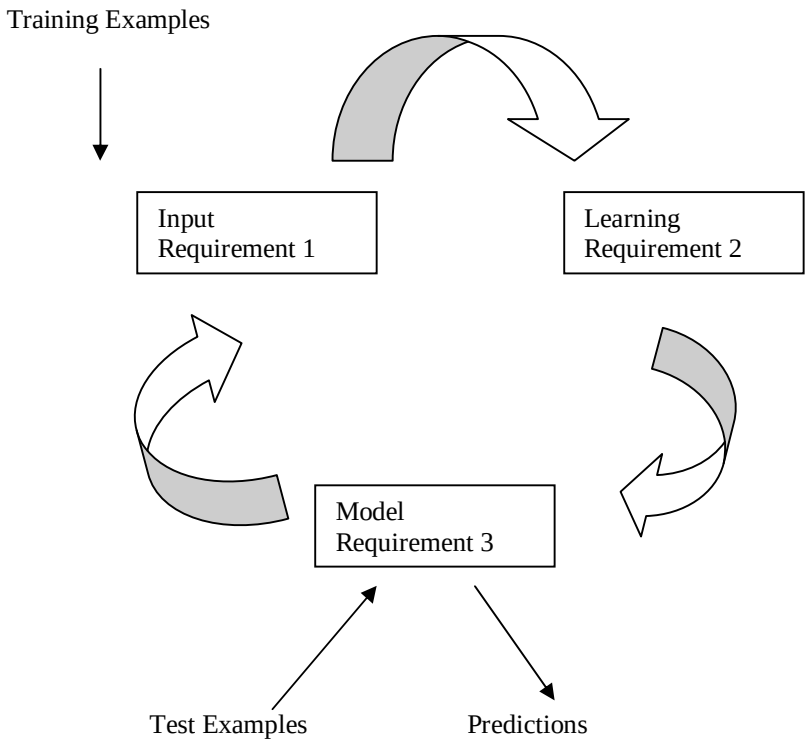

Fig. 1 Data Stream Model

- Requirement 1: method a random sample at a time and examine it one time.
- Requirement 2: choose an entrance limit for memory usage for a way of action, and don't exceed the limit.
- Requirement 3: identify time constraints for every method.
- Requirement 4: expect the subsequently incoming cases on the run.

The paper is structured as follows. Section 2 presents the related works. Section 3 discusses the methodologies for Stream Data Processing. Mining data stream techniques are analyzed in section 4. Some of the open research issues and challenges are discussed in section 5 and section 6 gives the conclusion of our work.

II. RELATED WORK

In case of data streams, the amount of distinct options or things that exist would be massive so this makes even the more number of cache memory or system memory are not appropriate for storing the whole stream data. The major drawbacks of data streams are the speed. Speed of information stream arrival is relatively higher than the speed of information store and process. In [6], authors discussed about the drawback of finding the top-k

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

frequent things during a data stream with stretchy sliding windows. The concept is to mine only the top-k frequent things rather than all the frequent things. However the crucial issue or limitation that evolves here is that the quantity of memory that's needed still for mining to find top-k frequent things remains a bounding issue.

In [7], authors specialized in developing a structure for categorizing dynamically growing data streams in view of the coaching and check streams for dynamic classification of datasets. The target is to extend a classification system during which coaching system will adapt to fast changes of the underlying data stream. The amount of memory accessible for mining stream data with one pass algorithms is extremely less and therefore there's probability for information loss. Additionally it's impossible to mine the information on-line as and once it seems due to difference in speed and some other vital factors. In [8] the authors discussed the method of finding most frequent items by using a hash based approach. The idea is to use 'h' hash functions and to build the hash table by using linear congruencies. Data streams are classified into two categories; they are offline data streams and online data streams.

In [9] the strategy of singular valued decomposition applied to search out the relationship between multiple streams. The idea of SVD was mainly applied to notice offline data streams. Clustering text data streams is one in all the topics that have evolved as necessary challenge for data processing researchers. The issue of spam detection, email filtering, clustering client behaviours, topic detection and identification, document clustering are a number of typical interest to data processing researchers. In [10], Liu et.al discussed on clustering text data streams. The concept is to increase the existing semantic smoothing model that works fine among fixed data streams for clustering dynamic data streams. The creators suggested two on-line clustering algorithms OCTS and OCTSM for clustering huge text data streams. A fabulous quantity of data is generated from internet instantly in different forms like public networks, data from sensors, face book and twitter. The rising data from internet is referred as Text message stream that is generated from different instant message applications and internets convey chat. This has become a main topic that has suit a hot theme of notice to the researchers operating within the area of data mining and includes a group of scope to figure to be contributed by the research community.

In [11], authors proposed the strategy of identifying the threads in dynamic data streams. The paper discussed three deviations of single pass clustering algorithm tracked by an original clustering algorithm which supported linguistic options. A technique of reducing the dimensionality of streaming data using a scalable supervised algorithm is proposed in [12]. The limitations of principal component analysis (PCA) and linear discriminant analysis (LDA) approaches are discussed. The writers illustrated the unsuitability of MMC for streaming data. A supervised progressive dimensionality reduction algorithm is planned to satisfy the requirements of streaming data set. In [13] the writers shown that the foremost cited end result certain is invalid. Table1 shows the comparative analysis of existing and proposed models.

S.No	Research paper name	Existing model	Proposed model
1	On Demand Classification of Data Streams [2]	The current model of the classification problem simply concentrates on methods for one-pass classification modelling of very large data sets	Classification of dynamic evolving data streams. Online classification and continuous adaption to the fast evolving data streams
2	Mining Top-K Frequent Items in a Data Stream with Flexible Sliding Windows [3]	Based on the properties of an item, the maxfrequency of an item is counted over a sliding window of which the length changes dynamically	Instead of reporting all frequent items, to only mine the top-k most frequent ones
3	Correlating synchronous and asynchronous data streams [5]	Identifying correlations between multiple data streams using singular value decomposition	Proposed an algorithm to maintain the SVD of multiple data streams and identify correlations between the streams
4	Clustering Text data	TF∗IDF scheme to represent the	Extend the semantic smoothing

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

	streams [7]	semantics of text data and often lead to poor clustering quality. Recently, researchers argue that semantic smoothing model is more efficient than the existing TF∗IDF scheme for improving text clustering quality	model into text data streams context firstly. Based on the extended model, they presented two online clustering algorithms online clustering of text streams (OCTS) and online clustering of text streams with merge (OCTSM) for the clustering of massive text data streams
5	Thread Detection in Dynamic Text Message Streams [8]	Beneficial for detecting the threads contained in the text stream for various applications, including information retrieval, expert recognition and even crime prevention. So far on this problem due to the characteristics of the data in which the messages are usually very short and incomplete	Proposed three variations of a single-pass clustering algorithm for exploiting the temporal information in the streams. An algorithm based on linguistic features is also put forward to exploit the discourse structure information

Table. 1 Comparative analysis of existing and proposed models

III. METHODOLOGIES

Since the data streams are huge in volume, it is difficult to store the data locally for analysis. Therefore, usually there is a trade-off between the accuracy of the analysis that is the result from the data analysis and the storage space. Further summary provides a summary of the data, this summary data structure that are smaller than base data sets. Hence, the output of the analysis is approximately correct. To counter these issues we need effective processing of data, efficient techniques and algorithms. When it comes to algorithms, it must be efficient in both time and space.

A. *Sampling:*

The simplest method for construction of summary in data streams is random sampling. Instead of considering the entire data, we can sample the data stream periodically, specialization of representation is not done and instead multi-dimensional representation is used i.e. similar to data points. Hence the summary can be used with many applications this is the main advantage of this method.

A method called reservoir sampling is used to choose s elements randomly, which are unbiased without replacement [14]. As to have a sample of unbiased data we had to know the length beforehand. Since it is not possible to know the length of data this modified approach is used.

The base idea of this method is very simple, a sample of size s is maintained which is referred as “reservoir”, from this reservoir a sample of size s can be obtained, when the reservoir itself is huge this generation of sample is very costly. To prevent this based on the elements that we have come across in the stream so far we obtain a true random sample by maintaining a set s candidates in the reservoir.

As the data keeps flowing, every new element we come across in the stream probably replaces an old element in the reservoir. The probability of this replacement is S/N where N is the elements so far in the stream and this method is called concise sampling. [15].

B. *Sliding Windows:*

Rather than sampling the data periodically we can use the concept of sliding window for analysis, the main motivation for this method is, instead of computation, to run on sample only. Recent data is taken in to consideration to make effective decisions [16].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

If the length of window is w, t is the time of arrival of new element the expiry of the element is t+w. This model is extremely helpful technique for complicated tasks like weather forecasting, stock market and sensor data analysis. In addition to this, it also reduces memory requirements because this technique needs a little window and in that window only data is stored. One recent on line software tool for analysis MAIDS uses this technique to obtain summary of data. In count based window most recent n elements are stored, in time based window we store the data which has arrived in the last T units of time.

C. ***Histogram*:**

The frequency distribution of values in a stream of data is approximated by using histogram, which is a summary data structure. A set of ranges are created by dividing the data along attributes sets and the count of each bucket is maintained.

The number of buckets in the histogram decides the space required. The data is divided into adjacent buckets, the width that is the range of bucket and the number of elements that is the depth depends on which rule is used for dividing the data.

Range queries can be easily answered using these methods. As the only thing to be determined is the set of buckets which lie in the range which is specified by the user. The query resolution can further be made efficient by deriving various strategies from histogram [17]. One such rule is to maintain the same range of each bucket called equal-width partitioning rule.

The disadvantage of this is the probability distribution function is not sampled properly. Alternatively there is another approach V-Optimal histograms, [18] here the bucket size minimizes the frequency variance in every bucket, then further these histogram are used to approximately answer the queries, instead of sampling methods, but still application of histograms on data streams is a challenge.

D. ***Sketches:***

Sketches are basically an expansion of the random projection procedure to the time series space. It can work in a single pass. The estimation of the frequency moments should be possible by summaries that are known as Sketches. These assemble a small space summary for a distribution vector (e.g., histogram) utilizing randomized linear projections of the basic information vectors. Sketches give probabilistic assurances on the nature of the approximate result. From a database point of view, the partitioning of sketches [19] was created to enhance the execution of sketches on information stream query enhancement. During this technique; we have a tendency to show intelligence partition be a part of attribute domain-space and use it in order to calculate separate sketches of every partition. The subsequent join assessment is figured as the sum of over all segments. This technique has likewise been examined in more detail for the issue of multi-inquiry evaluation [20].

One of the key preferences of sketch-based strategies is that they require space, which is sub linear in the information size being considered. Another advantage of this technique is it is conceivable to keep up sketches within the sight of cancellations. This is frequently unrealistic with numerous summation techniques, for example, random samples. One additionally intriguing trick for enhancing join size estimation is that of sketch skimming, it is portrayed in [21].

E. ***Aggregation:***

Summarizations of the received stream are produced with mean and variance. If the input streams have extremely irregular data distributions then this system fails. This could be used for merging offline data with on-line data that was studied in [22]. It is frequently considered as a data rate adaptation technique during a resource-aware mining.

Many abstract strategies like wavelets, histograms and sketches don't seem to be simple to use for the multi-dimensional cases. The sampling method is the only method used for handling high dimensional applications.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

IV. TECHNIQUES

The data stream pattern has recently emerged in response to the continuous information issue in data mining. Because of the persistent, unbounded, and rapid characteristics of data, there is an immense volume of information in both logged off and online information streams. New methods must be evolved to address the computational difficulties of data streams. Various procedures for extraction of information from data streams were proposed in concern with data mining.

A. *Clustering:*

Envision an enormous measure of dynamic stream information. Numerous applications require the computerized clustering of such information into segments depending on their alikeness. In spite of the fact that there are numerous effective grouping algorithms for static information sets, grouping or dividing data streams puts extra imperatives on such calculations, as any information stream model obliges algorithms to make a single pass over the information, with limited memory and constrained calculation time. A few calculations have been created for grouping information streams portrayed as beneath:

STREAM k-median based Stream Clustering Algorithm is discussed by Guha [23]. It includes of two stages and takes once divider and overcome approach. In first stage, it divides the information stream in buckets and after that discovers k clusters in every bucket by applying k-median grouping. It stores clusters and the cluster centers are weighted taking into account the quantity of information points belonging to the related cluster and afterward disposed the data points. In second stage, weighted cluster centers are grouped in small number of groups. While its space and time complexity is low but it cannot adapt to concept evolution in data.

CluStream [24] clustering algorithm for data streams is presented by Aggarwal et al. It partitions the grouping procedure in taking after two online component and components offline. Online segment stores the summary of information as micro-clusters. Summary insights of information are put away in snapshots which give the client adaptability to indicate the time interruption for grouping of micro-clusters. Offline component apply the k-means grouping to group microclusters into bigger segments.

ClusTree [25] Stream Clustering is presented by Kranen et a. It separates the grouping procedure in taking after two online and offline parts. Online part is utilized to learn micro cluster. The smaller scale bunches are ordered in various leveled tree structure. Any assortment of offline segments can be used. It is a self-versatile algorithm and conveys a model whenever needed.

HPStream [26] is presented by Aggarwal et al. for grouping of high dimensional information streams. It utilizes a Fading Cluster Structure (FCS) to store the outline of data and it gives more significance to recent information by blurring the old information with time. For taking care of high dimensional information it chooses the subset of measurements by projecting on unique high dimensional information stream. Number of dimensions and measurements are not same for every bunch. This is because; the importance of every dimension in every group may differ from one another. It is incrementally updatable and exceptionally adaptable on number of dimensions. But it can't find out the cluster of random shapes and require domain knowledge for specifying the number of clusters and average number of projected dimensions parameters.

E-Stream [27] is an information stream grouping system which bolsters taking after five sort of evolution in the data streams: Appearance of new bunch, Disappearance of an old bunch, Split of a huge group, converging of two identical groups and change in the conduct of group itself. It utilizes a blurring group structure with histogram to approximate the streaming information. Its execution is superior to HPStream algorithm yet it requires numerous parameters to be indicated by client.

B. *Classification:*

There are many strategies for the classification of static information. This is a two stage process comprising of model development from preparing data and arrangement where the model is utilized to foresee the class names of tuples from new information sets. In a conventional setting, the training information dwell in a generally static database so scanning can be carried out many times, yet in stream information, the information stream is fast to the point that capacity to store them and scanning it several times is infeasible. Another characteristic is time varying in data streams,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

instead of conventional database frameworks, where just the present state is stored. This change in the nature of the data takes the form of changes in the objective classification model after some time and is alluded to as concept drift. It is a vital thought when managing stream data. A few strategies have been proposed for stream information as demonstrated as follows.

Hoeffding Tree Algorithm [28] presented by from Domingos and Hulten's proposes the spilling choice tree prompting which is called Hoeffding Tree. The name is derived from the Hoeffding bound that is utilized as a part of the tree induction. The fundamental thought is, Hoeffding bound gives certain level of certainty on the selection of best attribute to divide the tree and thus we can develop the model in light of certain number of occurrences that we have seen. The principle point of preference of this algorithm is high precision with a small set of data samples, multiple scans of the same data are never done and it is incremental in nature. Aside from this the primary disadvantage of the algorithm is it can't deal with concept drift, in light of the fact that once a node is made, it can never show signs of change.

Fast Decision Tree [29] presented by Domingos et al. makes a few changes to the Hoeffding tree calculation to enhance both rate and precision. It splits the tree utilizing the present best attribute. Such a procedure has the property that its output is almost the same to that of a conventional learner. While that VFDT calculation works with small information streams, despite everything it can't deal with concept drift in data streams. To adjust this in information streams, VFDT algorithm was further formed into the Concept-adapting Very Fast Decision Tree calculation (CVFDT) it runs VFDT over sliding windows, which are fixed, the end goal is to have the most upgraded classifier.

Classification on Demand [30] presented by Aggarwal et al. have embraced the micro clusters presented in Clustream. The grouping procedure is separated into two segments, first which performs summary of information and classification is performed by the second segment.

ANNCAD Algorithm [31] is proposed by Law et al. is an incremental ordering calculation termed as Adaptive Nearest Neighbour Classification for Streams of data. The calculation utilizes Haar Wavelets Transformation for multi-determination information representation. A matrix based representation at every level is utilized. To address the issue of concept drift of information streams, an exponential fade variable is utilized to diminish the weight of old information in the grouping procedure. These calculations have accomplished precision over VFDT and CVFDT however the downside of this calculation is it can't deal with the sudden changes in concept drift.

Group based Classification [32] idea is presented by Wang et al. it is a system for carrying out extraction of information from streams of data with concept drift. It utilizes weighted classifiers to handle this problem. The thought is to prepare a group or set of classifiers from successive samples of the information stream. Every classifier is weighted and just the top K-classifiers are retained. The choices made by weighted votes of the classifier results in the output.

C. ***Association:***

There are usually two stages in algorithms for the association rule. The initial step is to find incessant item sets. In this progression, all continuous item sets that meet the threshold value are found and the second step is to infer association rules. In this progression, in light of the continuous item sets found in the initial step, the rules that meet the certainty basis are inferred. Nevertheless, customary association standard mining calculations are produced to take a shot at static information and, along these lines, can't be connected straight forwardly to mine association rules in stream information. Newly many researches are directed on the most proficient method to get frequently occurring elements, association rules and various patterns in the environment of stream of data. A segment of these algorithms are portrayed below.

In [33] Chang has proposed a technique for discovering late element sets adaptively over online information streams. It utilizes damped model which is likewise called as Time –Fading model in their calculation, which mines the frequent element sets in stream information. This model considers distinctive weights for new and old exchanges. This is appropriate for applications in which old information affects the results of extraction of data; however the impact diminishes over the long time.

In [34] to extract data from frequent item sets Lin has proposed a technique. For this it utilizes Sliding Window model as a part of their calculation. This model finds and keeps up continuous item sets in sliding windows. Just part of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

the information streams inside the sliding window is saved and processed when the information streams in. The measure of the sliding window may fluctuate as indicated by applications and resources in the system.

In [35] an algorithm is presented by Yang for extraction of short rules of association in a database. In this, they have utilized the calculations to produce the result of frequent item sets. In exact algorithms, the outcome sets comprise of the majority of the item sets that bolster estimations of which are more prominent than or equivalent to the threshold. The primary disadvantage of this calculation is it can just mine short item sets, which can't be connected to expansive item sets.

In [36] figure out the approximate frequency counts in streams of data, Manku has presented an algorithm lossy counting to store the item sets it utilizes lattices data structures. Accordingly it sorts the stream of input information in to appropriate sets of windows which are of pre-fixed size and computes every window consecutively. For every component in a window, it embeds an entry into a table, and monitoring of occurrences of the items, or if the component is as of now in the table, it redesigns its count of frequency. Toward the end of every window, the calculation expels components from entries in table which are of very less frequency or occurrence. The principle disadvantage of this calculation is space bound, scanning multiple times and past data influences the final result.

In [37] to keep track of number of frequent items occurring Cormode has presented an algorithm. A little or small information structures are maintained that monitor the exchanges on the connection, and at whatever point required rapidly yields or produces an output of every single hot elements from the item set without re-scanning the connection in the information base.

V. RESEARCH ISSUES AND CHALLENGES

Data stream mining could be an inspiring field of study that has raised several challenges and research issues that require to be self-addressed by the machine learning and data mining communities. Once developing mining techniques of this sort, there are additional issues that require to be considered than in usual databases. The following is a brief discussion of some crucial open research issues:

A. *Memory management:*

The first basic issue we want to think about is a way to optimize the memory area inspired by the mining algorithm. Memory management may be a specific challenge once process streams as a result of several real data streams are uneven in their rate of arrival and variation of data arrival rate over time. A stream mining algorithm with high memory price can have problem being applied in several things, like detector networks. Additional research has to be done in developing new summarization techniques for collecting valuable information from data streams. Absolutely addressing this issue within the mining algorithm will really improve the performance.

B. *Data pre-processing:*

Data pre-processing is a very important and time intense stage in the knowledge discovery method and should be taken into consideration once mining data streams. Coming up with a light-weight pre-processing techniques that can guarantee quality of the mining results is essential. The challenge is to automate such a method and combine it with the mining techniques.

C. *Compact data structure:*

Due to delimited memory size and also the large quantity of data streams coming endlessly, efficient and compact data structure is required to store, update and retrieve the collected information. Disappointment in developing such a data structure can mostly decrease the efficiency of the mining algorithm. Even though we have a tendency to store the information in disks, the extra I/O operations can increase the interval. Progressive maintaining of the data structure may be a necessity one since it’s impossible to rescan the whole input. Also, novel indexing, storage space and querying methods are needed to handle the frequent fluctuated flow of information streams.

The revise of data stream mining has born to some open challenges that demand attention. Here could be a short review of them:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

- As real data can be irregular and unpredictable in nature, therefore the algorithm should be capable to manage the traffic by using optimal resources.
- An intellectual data preprocessing module within the algorithm will make sure prime quality of finish results.
- Due to use of restricted resources for handling huge amount of data one should make sure that the data structures are capable to handle process on the disk. I/O and indexing methods also are essential aspects on the time interval.
- The method should be intellect to distinguish between noise and idea change in live stream.
- Visualization is additionally a priority, particularly once the results are transmitted through wireless medium and analyzed on mobile gadgets. Some further efforts should be taken to complete the method within a restricted bandwidth.
- Capable querying mechanism is required to change method and retrieve the data at any purpose of time.
- Optimize the memory, computation power whereas process huge data sets as several real data streams are uneven in their rate of arrival.
- High accuracy within the outcome generated whereas handling continuous streams of data.
- Transferring data mining results over a wireless network within a restricted bandwidth.
- Online Interactive process is required that helps user to change the parameters throughout processing period.

VI. CONCLUSION

The distribution of data stream development has necessitated the event of stream mining algorithms. In this paper we have mentioned many problems that are to be considered once planning and implementing the data stream mining technique. We have a tendency to even review a number of these methodologies with the existing algorithm such as clustering and classification.

We can conclude that almost this entire mining approach used one pass mining algorithms and only some of them even address the problem of drifting. From our study that data stream change huge volumes of temporally changing data so, usual techniques of data mining cannot be applied easily.

Research in data stream mining is in premature stage. If the problems created by data streams are resolved and if more successful and interactive mining techniques which are user friendly are to be developed, it is likely that within the close to future data stream mining can play an important role within the business world which involves mining from continuous data streams.

REFERENCES

1. J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, vol. 54, pp.6-7, 2006.
2. Charu C Agrawal, “Data Streams: Models and Algorithms”, Kluwer Academic Publishers, pp.2-5.
3. A. Bifet, G. Holmes, R. Krikby and B. Pfahringer, “Data Stream Mining-A Practical Approach”, 2011
4. Madjid Khalilian, Norwati Mustapha, “Data Stream Clustering: Challenges and Issues”, Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol1, IMECS 2010, pp.17-19, 2010.
5. B. Babcock, S. Babu, M Datar, R Motwani, andWidom, “Models and issues in data stream systems”,ACM Sigmod/PODS,pp.1-6, 2002.
6. Hoang Thanh Lam, Toon Calders, “Mining Top-K Frequent Items in a Data Stream with Flexible Sliding Windows”, Proceedings of ACM KDD, pp.283-292, 2010.
7. Charu C. Aggarwal, Jiawei Han and Philip S. Yu, “On Demand Classification of Data Streams”, Proceedings of ACM KDD, pp.503-508, 2004.
8. Cheqing Jin et.al,“Dynamically Maintaining Frequent Items over a Data Stream”, Proceedings of CIKM USA, 2003.
9. Sudipta Guha, D.Gunopulos and N. Kaudas, “Correlating synchronous and asynchronous data streams”, Proceedings of SIGKDD, 2003.
10. Yu.Bao.Liu et.al,“Clustering Text data streams”, Journal of computer science and technology, volume 23, issue 1, pp.112-128, 2008.
11. Dou Shen,Qiang Yang, Jian-Tuo-Sun and Zheng Chen,“Thread Detection in Dynamic Text Message Streams”, Proceedings of SIGIR USA, 2006.
12. Jun Yan et.al,“A scalable supervised algorithm for dimensionality reduction on streaming data”, Information Sciences an International Journal, Vol.176, pp. 2042-2065, 2006.
13. L.Rutkowski et.al, “Decision trees for mining data streams based on the McDiarmid’s bound”, IEEE Transactions on Knowledge and Data Engineering, 2012.
14. Vitter J. S,“Random Sampling with a Reservoir”, ACMTransactions on Mathematical Software, Vol. 11, pp. 37– 57, 1985.
15. Gibbons P, Mattias Y., “New Sampling-Based Summary Statistics for Improving Approximate Query Answers”, ACM SIGMOD International Conference on Management of Data, Vol.27, pp.331-342, 1998.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

16. M.Datar,A.Gionis, P. IndykandR.Motwani, "Maintaining stream statistics over sliding windows", SIAM Journal on Computing, Vol.31, pp.1794–1813, 2002.
17. Poosala V, Ganti V and Ioannidis Y, "Approximate Query Answering using Histograms", IEEE Data Eng. Bull, 1999.
18. JagadishH, Koudas N,Muthukrishnan S, PoosalaV, SevcikK, and Suel T. "OptimalHistogramswithQualityGuarantees", VLDB Conference, pp.275-286, 1998.
19. Dobra A, Garofalakis M, Gehrke J and Rastogi R,"Processing complex aggregate queries over data streams", SIGMOD Conference, pp.61-72, 2002.
20. Dobra A, Garofalakis M. N, Gehrke J and Rastogi R "Sketch- Based Multi-query Processing over Data Streams", EDBT Conference, 2002.
21. Ganguly S, Garofalakis M and Rastogi R. "Processing Data Stream Join Aggregates using Skimmed Sketches", EDBT Conference, Vol.2992, 2004.
22. C Aggarwal, J. Han, J. Wang, and P. S. Yu, "A Framework for Projected Clustering of High Dimensional Data Streams", Conference on VLDB, 2004.
23. L. callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-Data Algorithms for High-Quality Clustering," IEEE International Conference on Data Engineering, 2002.
24. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," Conference on VLDB, Vol.29, pp. 81–92, 2003.
25. Kranen,Assent, Baldauf and Seidl, "Self Adaptive Any Time Clustering",ICMD IEEE International Conference, pp.249-258, 2009.
26. C. C. Aggarwal, J. Han, J. Wang and P. S. Yu, "A framework for projected clustering of high dimensional data streams",Conference on VLDB, Vol.30, pp. 852–863, 2004.
27. K. Udommanetanakit, T. Rakthanmanon and K. Waiyamai, "Estream: Evolution-based technique for stream clustering" International conference on Advanced Data Mining and Applications, pp. 605–615, 2007.
28. Domingos P and Hulten G, "Mining High-Speed Data Streams", In Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining, pp.71-80, 2000.
29. Hulten G, Spencer L and Domingos P,"Mining Time- Changing Data Streams", ACM SIGKDD Conference, pp.97-106, 2001.
30. Charu C. Aggarwal, Jiawei Han and Philip S. Yu, "On Demand Classification of Data Streams", Proceedings of ACM KDD, pp.503-508, 2004.
31. Law Y, Zaniolo C,"An Adaptive Nearest Neighbor Classification Algorithm for Data Streams", Conference on the Principals and Practice of Knowledge Discovery in Databases, Vol.3721, 2005.
32. Wang H, FanW, Yu P and Han J,"Mining Concept- Drifting Data Streams using Ensemble Classifiers", ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp.226-235, 2003.
33. Joong Hyuk Chang, Won Suk Lee "Finding Recent Frequent Itemsets Adaptively over Online Data Streams", ACM SIGKDD, pp.487-492, 2003.
34. Chih-Hsiang Lin, Ding-Ying Chiu, Yi-Hung Wu and Arbee L. P. Chen,"Mining Frequent Itemsets from Data Streams with a Time-Sensitive Sliding Window", SIAM Int'l Conf. on DataMining, 2005.
35. Li Yang, Mustafa Sanver, "Mining Short Association Rules with One Database Scan", Int'l Conf. on Information and Knowledge Engineering, pp.392-395, 2004.
36. G. S. Manku and R. Motwani,"Approximate frequency counts over data streams", International Conference on Very Large Data Bases (VLDB), 2002.
37. Cormode, Graham Cormode and S.Muthukrishnan,"What's Hot and What's Not: Tracking Most Frequent Items Dynamically", ACM Transactions on Database Systems, Vol.30, pp.249-278, 2005.

BIOGRAPHY

R. Kalaivani has completed M.Sc in Computer Science. She is currently pursuing her Ph.D in Computer Science in the Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are Data Mining, Data Streams and big data.

Dr. S. Vijayarani MCA, M.Phil, Ph.D working as Assistant Professor in the Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues, text mining, web mining, information retrieval, data streams and big data. She has authored a book and published more than 70 research papers in the international journals and also presented papers in international and national conferences.