

A Survey Paper on Data Stream Mining

Shazia Nousheen M
Department of Computer Engineering
BMS college of Engineering
Bangalore, India

Dr. Prasad G R
Department of Computer Engineering
BMS college of Engineering
Bangalore, India

Abstract— In recent times a huge volume of data is generated from various sources hence leading to wide spread interest on the emerging field of data stream mining. Data stream mining is the way of extracting meaningful knowledge from this huge volumes of information. There are various applications of this field hence introduction of new methods for data classification are widely researched to provide better standard of services. The aim of this survey is to provide a brief view on different classification techniques in data mining.

Keywords—Data streams; Data Streams mining; classification technique; concept drift;

I. INTRODUCTION

In the field of processing of information, Data mining ascribes to extracting 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 streams 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 the many evolving applications which involve huge volumes of data generated from various sensor data, data from supermarkets, telephone logs, and 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 data now 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 analyze 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.

The paper is organized as, Section I provides Introduction. Section II introduces the basic Methodologies for Processing Stream Data. Section III explains the

various Stream Mining Algorithms. Section IV describes various researches.

II. BASIC METHODOLOGIES

Since the data streams are huge in volume, it is difficult to store the data locally and analyze it. Therefore, usually there is a tradeoff between the accuracy of the analysis that is the result from the data analysis and the storage space. Further synopsis provides a summary of the data, this synopsis data structures 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. If n is the length of data stream instead of using $O(N)$ space to store the complete set of elements data structures makes use of $O(\log^k N)$, space, in the segment below we discuss various such techniques.

A. Random Sampling

The simplest method for construction of synopsis 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 instead multi dimensional representation is used similar to data points hence the synopsis 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 [5]. 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. A method called concise sampling is used [6].

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 decisions [7].

If the length of window is w , t is the time of arrival of new element the expiry of the element is $t+w$. Applications where the importance is more for recent data or in other words the analysis is sufficient on the recent most events to derive decisions, this method is used in fields like weather forecasting, stock market and sensor data. It also solves the problem of memory requirements as only a data of window of small size is maintained 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. Histograms

The frequency distribution of values in a stream of data is approximated by using histogram, which is a synopsis 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, and other such different kinds of queries. As the only thing to be determined is the set of buckets which lie in the range specified by the user. The query resolution can be further made efficient by deriving various strategies from histogram[8]. One such rule is the maintaining of 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[9] here the bucket size minimizes the frequency variance in every bucket, then further these histograms are used to approximately answer the queries, instead of sampling methods, but still application of histograms on data streams is a challenge.

D. Multiresolution Methods

A usual way to solve problems involving huge volumes of data is using reduction methods, the most popular of which is divide and conquer methods like structures. The advantage of this is it allows to balance between storage and accuracy and also allows understanding of stream data at multiple levels. For instance in binary tree each level of tree represents a different resolution. Similarly, two other methods are micro clusters and wavelets. In micro clusters [10] we can use Cluster Feature (CF) which is a data structure to form hierarchy of micro-clusters. It can be applied to multidimensional cases. Whereas, in case of wavelets we create a decomposition of characteristics of data into a collection of wavelet and basic functions. The wavelets are good for dealing with spacial and multimedia data.

E. 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 synopsis 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 [14] was created to enhance the execution of sketches on information stream query enhancement. In this method; we divide into segments the join attribute domain space and utilize it keeping in mind the end goal to process separate sketches of every segment. 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 [15].

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 point of advantage is this of techniques that 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 [16].

F. Randomized algorithms

Randomized algorithms, as arbitrary inspecting and sketching, are regularly used to manage enormous, high-dimensional information streams. The utilization of randomization regularly prompts easier and more effective algorithms in contrast with known deterministic calculations. In the event that a randomized calculation dependably gives back the right answer however the running times shift, it is known as a Las Vegas calculation. Conversely, a Monte Carlo calculation has limits on the running time however may not give back the right result. We for the most part consider Monte Carlo algorithms. One approach to think about a randomized calculation is just as a probability distribution over an arrangement of deterministic calculations.

III. DATA STREAM MINING ALGORITHMS

The data stream pattern has as of late risen in light of 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 resemblance. 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 –A k-median based Stream Clustering Algorithm is proposed by Guha, Mishra, Motwani and O'Callaghan [17]. It comprises of two stages and takes after partition and conquer approach. In first stage, it partitions 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 disposes of the data points. In second stage, weighted clustered centers are grouped in small number of groups. In spite of the fact that its space and time intricacy is low yet it can't adjust to evolution of concept in data.

CluStream [18] clustering developing 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 synopsis of information as micro-clusters. It is the transient augmentation of clustering highlight of BIRCH [11]. synopsis insights of information are put away in snapshots which gives the client adaptability to indicate the time interruption for grouping of micro-clusters. Offline component apply the k-means grouping to group micro-clusters into bigger segments.

ClusTree [19] Any time Stream Clustering is presented by Kranen et al. 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 [20] is presented by Aggarwal et al. for grouping of high dimensional information streams. It utilizes a Fading Cluster Structure (FCS) to stores 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 of the way that importance of every dimension in every group may contrast. It is incrementally updatable and exceptionally adaptable on number of dimensions. Be that as it may, it can't find the bunch of subjective shapes and require domain learning for indicating the quantity of groups and average number of anticipated measurements parameters

E-Stream [21] 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. In spite of the fact that its execution is superior to HPStream algorithm yet it requires numerous parameters to be indicated by client.

Hue Stream [22] augments E-Stream, which is portrayed before, so as to bolster instability in heterogeneous information. to support vulnerability in attributes a function which is the probability distribution of two elements is introduced. to identify difference in

structures of clusters the proposed function is utilized to combine clusters and to detect the cluster of a approaching incoming information and to partition cluster in definite data by using histogram management

POD Clus [23] is a model based bunching strategy for streaming information. POD(Probability and Distribution-based Clustering) It is pertinent to both bunching by illustration and grouping by variable situations. For creating synopsis of cluster data and upgrade it incrementally, it utilizes a bunch summary which involves the mean, standard deviation, and number of focuses for every group. It underpins evolution of concept by permitting new group appearance, splitting of a bunch, converging of two bunches and vanishing of a bunch.

B. Classification

There are a 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 this referring to stream data. 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 that is time varying in data streams, instead of conventional database frameworks, where just the present state is put away. This change in the way of the information takes the type 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 [24] presented by from Domingos and Hulten's proposes the spilling choice tree prompting which is called Hoeffding Tree. The name is gotten 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, 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 [25] presented by Domingos et al. makes a few changes to the Hoeffding tree calculation to enhance both rate and precision. It parts the tree utilizing the present best trait. Such a procedure has the property that its yield is (asymptotically) almost indistinguishable to that of a conventional learner. In spite of the fact 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, keeping in mind the end goal to have the most upgraded classifier.

Classification on Demand [26] presented by Aggarwal et al. have embraced the micro clusters presented in Clustream [18]. The grouping procedure is separated into two segments, first which performs synopsis of information and classification is performed by the second segment.

ANNCAD Algorithm [27] proposed by Law et al. is an incremental ordering calculation termed as Adaptive Nearest Neighbor 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 [28] idea presented by Wang et al. 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 of 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. As of late numerous researches about 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 this algorithms are portrayed below.

In [29] 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 [30] 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 the information streams inside the sliding window are 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 [31] 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 careful calculations, 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 itemsets.

In [32] to 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 a 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 [33] 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 itemset without re-scanning the connection in the information base.

IV. ISSUES IN RESEARCH

There are many challenges and research issues pertaining to study of extraction of information from data streams

- Handling the steady stream of information.
- Memory requirements for the huge volume of data.
- Maintaining the privacy of data in the processes of extraction from the stream of data.
- Data stream models to be integrated with adept learning or skilled learner.
- Presentation of the results from the mining of data.
- Adapting various changes with time in data.
- Drifting towards tools from the data stream algorithms.
- Scalability of the systems
- Transferring the results of extraction of information over a remote wireless system with a constrained data transfer capacity.
- As genuine information may be sporadic and might not be predictable in nature, thus the calculation ought to have the capacity to oversee the movement by utilizing ideal resources.
- Recovering the data at any point of instance of time and to modify process, querying mechanism must be optimized.
- The change of concept and to detect noises the model must be intelligent enough to differentiate both the changes.

V. CONCLUSION

There are many challenges In the present time, persistent generation of information streams has lead to the field of information mining termed as Data Stream Mining. So in this paper, we talked about different issues raised by information streams and introduced a review of different approach utilized for producing synopsis information structures from continuous information streams. There are various relative studies of the different techniques, yet it has not been observed that one single technique is better looked at than others. Issues like exactness, versatility, training time and numerous others add to picking the best strategy to classification of information for mining. The quest for best method for this still remains a topic to research on.

We likewise inspected a few algorithms created in Clustering, Classification and Association with respect to data mining, for managing the data stream. From our survey we can reason that data stream huge volumes of changing information So, old procedures of data mining can't be applied. the research in this field is in early stage. if the issues pertaining to this field are solved and if powerful and intelligent mining strategies which are easy to understand are produced, it is likely that sooner rather than later information stream mining will play an essential part in business world, as it manages numerous applications which includes mining from endless data streams. and research issues pertaining to study of extraction of information from data streams.

REFERENCES

- [1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition.
- [2] Charu C Agrawal, Data Streams: Models and Algorithms., Springer (Science +Business Media, LLC, 233 Spring Street, New York, NY 10013,USA), 2007.
- [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, 2010 Vol 1, IMECS 2010, March 17-19, 2010, Hong Kong.
- [5] Vitter J. S. (1985) Random Sampling with a Reservoir. ACM Transactions on Mathematical Software, Vol. 11(1), pp 37–57.
- [6] Gibbons P., Mattias Y. (1998) New Sampling-Based Summary Statistics for Improving Approximate Query Answers. ACM SIGMOD Conference Proceedings.
- [7] M.Datar, A.Gionis, P. Indyk, and R.Motwani. Maintaining stream statistics over sliding windows. SIAM Journal on Computing, 31(6):1794–1813, 2002.
- [8] Poosala V., Ganti V., Ioannidis Y. (1999) Approximate Query Answering using Histograms. IEEE Data Eng. Bull.
- [9] Jagadish H., Koudas N., Muthukrishnan S., Poosala V., Sevcik K., and Suel T. (1998) Optimal Histograms with Quality Guarantees. VLDB Conference.
- [10] Aggarwal C., Han J., Wang J., Yu P. (2003) A Framework for Clustering Evolving Data Streams. VLDB Conference.
- [11] Zhang, T., Ramakrishnan, R., Livny, M. Brich: An efficient data clustering method for very large data bases, In SIGMOD, Montreal, Canada, ACM (1996).
- [12] Keim D. A., Heczeko M. (2001) Wavelets and their Applications in Databases. ICDE Conference.
- [13] Guha S., Kim C., Shim K. (2004) XWAVE: Approximate Extended Wavelets for Streaming Data. VLDB Conference, 2004.
- [14] Dobra A., Garofalakis M., Gehrke J., Rastogi R. (2002) Processing complex aggregate queries over data streams. SIGMOD Conference, 2002.
- [15] Dobra A., Garofalakis M. N., Gehrke J., Rastogi R. (2004) Sketch-Based Multi-query Processing over Data Streams. EDBT Conference.
- [16] Ganguly S., Garofalakis M., Rastogi R. (2004) Processing Data Stream Join Aggregates using Skimmed Sketches. EDBT Conference.
- [17] 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,
- [18] 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.
- [19] Kranen, Assent, Baldauf, Seidl, "Self Adaptive any time clustering", ICMD, 2009
- [20] 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.
- [21] 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. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 605–615.
- [22] 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.
- [23] 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.
- [24] Domingos P. and Hulten G. (2000) Mining High-Speed Data Streams. In Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining.
- [25] Hulten G., Spencer L., and Domingos P. (2001) Mining Time-Changing Data Streams. ACM SIGKDD Conference.
- [26] Aggarwal C., Han J., Wang J., Yu P. S., (2004) On Demand Classification of Data Streams, Proc. 2004 Int. Conf. on Knowledge Discovery and Data Mining (KDD'04), Seattle, WA.
- [27] Law Y., Zaniolo C. (2005) An Adaptive Nearest Neighbor Classification Algorithm for Data Streams, Proceedings of the 9th European Conference on the Principles and Practice of Knowledge Discovery in Databases, Springer Verlag, Porto, Portugal.
- [28] Wang H., Fan W., Yu P. and Han J. (2003) Mining Concept-Drifting Data Streams using Ensemble Classifiers, in the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington DC, USA.
- [29] Chang, 2003] Joong Hyuk Chang, Won Suk Lee, Aoying Zhou; Finding Recent Frequent Itemsets Adaptively over Online Data Streams; ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining; August 2003.
- [30] [Lin, 2005] Chih-Hsiang Lin, Ding-Ying Chiu, Yi-Hung Wu, Arbee L. P. Chen; Mining Frequent Itemsets from Data Streams with a Time-Sensitive Sliding Window; SIAM Int'l Conf. on Data Mining; April 2005.
- [31] [Yang, 2004] Li Yang, Mustafa Sanver; Mining Short Association Rules with One Database Scan; Int'l Conf. on Information and Knowledge Engineering; June 2004.
- [32] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In In Proceedings of the 28th International Conference on Very Large Data Bases (VLDB), 2002.
- [33] Cormode, 2003] Graham Cormode, S. Muthukrishnan; What's Hot and What's Not: Tracking Most Frequent Items Dynamically; ACM Transactions on Database Systems; March 2005.