ISSN:0975-9646

A Quick Survey on Data Stream Mining

Poonam Debnath [#1], Santoshkumar Chobe [*2]

[#] *Department of Computer Engineering, University of Pune*
DYPIET, Pimpri, India

[*] *Asst. Professor of Computer Engineering Department*
DYPIET, Pimpri, University of Pune, India

***Abstract*— Traditional databases store sets of relatively static records without the concept of time, unless timestamp attributes are explicitly added. Limitations of traditional DBMSs in supporting streaming applications have been understood, prompting research to supplement existing technologies and build new systems to manage streaming data. At present a growing number of applications that generate massive streams of data need intelligent data processing and online analysis. The impending need for turning such data into useful information and knowledge augments the development of algorithms and frameworks that address streaming challenges. The pre processing, storage, querying and mining of such data sets are highly computationally challenging tasks. Mining data streams implies extracting knowledge structures represented in models and patterns in non stopping streams of information. In this paper, we present the theoretical foundations of data stream analysis and identify potential stream mining techniques.**

***Keywords*—Data streams, Data mining.**

I. INTRODUCTION

Recently a new class of emerging applications has become widely recognized: applications in which data is generated at very high rates in the form of transient data streams. In the data stream model, individual data items may be relational tuples, call records, web page visits, sensor readings, and so on. However, the continuous arrival of data in multiple, rapid, time varying, unpredictable and unbound streams open new elementary research problems. The rapid generation of continuous streams of information has posed a challenge for the storage, computation and communication capabilities in a computing system. The gigantic amounts of data arriving at high speed need application of semi-automated interactive techniques to perform real-time extraction of hidden knowledge. Typical data mining tasks include concept description, regression analysis, association mining, outlier analysis, classification, and clustering. These techniques find interesting patterns, tracing regularities and anomalies in the data set. However, traditional data mining techniques cannot be directly applied to the data streaming model. This is because most of them require multiple scans of data to mine the information, which is impractical for stream data. The amount of formerly happened events is usually immeasurable, so they can be either dropped after processing or archived separately in secondary storage. More importantly, the traits of the data stream can change over time and the evolving pattern needs to be recorded. Furthermore, the problem of resource allocation has to be considered in mining data streams. Due to the bulky volume and the high speed of streaming data, stream mining algorithms must handle the effects of system burden. Thus, how to accomplish optimum results under various resource constraints becomes a challenging task.

Fig. 1 illustrates the general data stream model. The cyclic process has three main steps occurring in a recursive format.

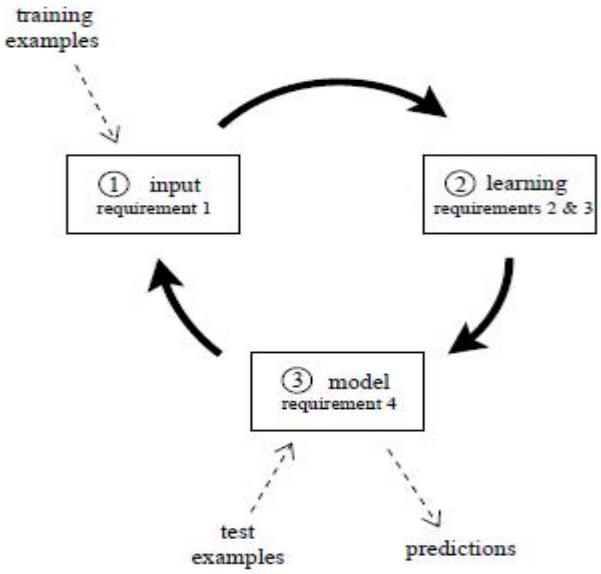

Fig. 1 Data Stream Model

- Requirement 1: Process a random sample at a time and examine it only once.
- Requirement 2: Decide a threshold limit for memory usage for a course of action, and do not exceed the limit.
- Requirement 3: Define time constraints for each process.
- Requirement 4: Predict the next incoming examples on the run.

II. FOUNDATIONS OF DATA STREAMS

The computational approaches of stream data mining mostly rely on statistics, complexity and computational theory. The real time nature of data streams and their high arrival rates impose high resource requirements on the system. In order to deal with resource constraints in a standard manner, many data summarization techniques have been used. They provide means to check only a subset of the whole data set or to transform the to an estimated smaller size data representation so that data mining

techniques can be employed. Also, computational theory techniques have been implemented to attain time and space efficient solutions. Summarization is often used for producing fairly accurate answers from databases. They blend data reduction and synopsis construction techniques. Summarization refers to transforming the data to a suitable form for stream analysis, which could be done by shortening the whole data set or choosing a subset of the incoming stream for analysis. When summarizing the data set, techniques such as sampling, sketching and load shedding are used. For selection of a subset from the data stream, synopsis data structures and aggregation functions are used.

A. Sampling

Sampling makes a probabilistic choice of stream elements under analysis. A bound for the error rate is usually given as a function of the samples per unit time. Few computation methods depend upon frequency counts over data streams previously performed by sampling. Very Fast Machine learning techniques use Hoeffding's bound to measure the size of the sample. Sampling techniques are used for clustering, classification and the sliding window model. The problem with sampling is the unknown dataset size. Management of data stream has to follow some special methods to find the error bounds. Fluctuating data rates are not addressed by sampling techniques. The relationship between data rate, sampling rate and error bounds should be inspected properly. It's not the correct choice for finding anomalies in surveillance analysis.

B. Sketching

Sketching involves construction of a summary of a data stream using a small amount of memory by vertically sampling. Usually it is applied for comparing data streams and in aggregate queries. Sketching techniques are suitable for distributed computation over multiple streams. The major downside of sketching is accuracy as it's tough to incorporate sketching algorithms on all kinds of data. Principal Component Analysis (PCA) would be a better solution if being applied in streaming applications.

C. Load Shedding

Load shedding refers to the process of eliminating a group of subsequent data streams during periods of overload. Its main function includes querying of the data streams for optimization as it is desirable to shack load to minimize the dive in accuracy. Load shedding occurs in two steps. Firstly, select target sampling rates for each query. In the second step, place the load shedders to realize the targets in the most efficient manner. It is not an ideal approach in time series mining because it drops chunks that might represent an important patter. Still, it has been successfully used in sliding window aggregate queries.

D. Synopsis Data Structures

Synopsis data structures symbolize information over data streams. Creating synopsis of data refers to the development of solution based on summarization techniques. Techniques used for construction of synopsis data structures are as follows:

1) Histograms: Histograms estimate the data in terms of one or more attributes of a relation by combining attribute values into buckets and approximating true attribute values and their frequencies in the data set based on a summary statistic maintained in each bucket. Histograms have been used extensively to capture data distribution in order to represent the whole data set by a small number of step functions. These procedures are widely used for stationary data sets. However conventional algorithms require super-linear time and space. This is due to the use of dynamic programming for optimal histogram construction. For most real-world database applications, there exist histograms which produce low-error ratio but their extension to the data stream case is a challenging task.

2) Wavelets: Wavelets are often used in database driven applications for hierarchical decomposition and summarization of the data sets. Wavelet coefficients are projections of the data set values onto a set of orthogonal vectors. The basic idea of the wavelet technique is that the higher order coefficients demonstrate the extensive trends in the data sets, whereas the localized trends are recorded by the lower order coefficients.

E. Sketches

Randomized description of wavelet techniques is known as sketching. Such methods are hard to implement as it is tricky to guess the interpretation based on sketch representations. Generalization of these techniques for a multi-dimensional case still remains an open problem.

F. Aggregation

Summarization of an incoming stream is generated using mean and variance method. If the input has high fluctuating distributions then the technique fails. It is often considered as a data rate adaptation technique in a resource-aware mining. Many synopsis methods such as wavelets, histograms, and sketches are not easy to use for the multi-dimensional data input. The random sampling method is often the only method of choice for high dimensional applications.

III. **ALGORITHMS**

The algorithms used for mining data streams are modified and enhanced version of basic data mining algorithms. Some of the widely accepted algorithms for data stream mining are:

A. Approximation algorithm

Approximation techniques used in algorithm design yield solutions with error bound and is approximate in nature. For dynamic tracking and providing absolution, these techniques are widely in an adaptive stream mining environment.

B. Sliding Window

The idea is to carry out a comprehensive analysis of the most recent data and over the old summarized data sets. The idea has been adopted in MAIDS. By imposing sliding window method on data streams approximations have become simple and due to its deterministic nature there is

no chance of bad random choices shall produce inaccurate approximations. The major advantage is that its focus is recent data.

C. Algorithm Output Granularity

It is a resource aware data set analysis approach used with irregular and high data rates applied to certain constraints. The process begins with mining the streams to adaptation of resources to merging the generated structures when memory is a bottleneck.

IV. CHALLENGES

The study of data stream mining has given birth to a few open issues that demand attention. Here is a short review of them:

- As real data might be irregular and unpredictable in nature, hence the algorithm should be able to manage the traffic by using optimal resources.
- An intelligent data preprocessing module in the algorithm can ensure high quality of end results.
- Due to use of limited resources for handling large amount of data one must ensure that the data structures are efficient to handle operations on the disk. I/O and indexing techniques are also critical aspects on the processing time.
- The technique should be intelligence to differentiate between noise and concept change in live stream.
- Visualization is also a concern, especially when the results are transmitted through wireless medium and viewed on mobile gadgets. Some additional efforts should be taken to complete the process with a limited bandwidth.
- Efficient querying mechanism is needed to modify process and retrieve the data at any point of time.

V. CONCLUSIONS

From all that have been discussed in this paper we can say that data streaming is still in its formative years and many aspects of the same demands attention.

We have discussed the characteristics and issues that a data streaming technique must address compulsorily

ACKNOWLEDGMENT

The authors would like to express thanks to the reviewers for helpful comments.

REFERENCES

[1] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In Proceedings of the Annual Symposium on Foundations of Computer Science. IEEE, November 2000.

[2] B. Babcock, M. Datar, R. Motwani, L.O'Callaghan: Maintaining Variance and k-Medians over Data Stream Windows, Proceedings of the 22nd Symposium on Principles of Database Systems, 2003.

[3] M. Charikar, L. O'Callaghan, and R. Panigrahy. Better streaming algorithms for clustering problems In Proc. of 35th ACM Symposium on Theory of Computing, 2003.

[4] P. Domingos and G. Hulten. Mining High-Speed Data Streams. In Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining, 2000.

[5] V. Ganti, J. Gehrke, and R. Ramakrishnan: Mining Data Streams under Block Evolution. SIGKDD Explorations 3(2), 2002.

[6] H. Wang, W. Fan, P. Yu and 1. Han, Mining Concept-Drifting Data Streams using Ensemble Classifiers, in the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Aug. 2003, Washington DC, USA.

[7] G. Wesolowsky, The Weber problem: History and perspective. Location Science, 1:5-23, 1993.

[8] M. Last, Online Classification of Nonstationary Data Streams, Intelligent Data Analysis, Vol. 6, No. 2, pp. 129-147, 2002.

[9] G. Cormode, S. Muthukrishnan What's hot and what's not: tracking most frequent items dynamically. PODS 2003: 296-306.

[10] A. Gilbert et al. Fast, small-space algorithms for approximate histogram maintenance. In Proc. of the 2002 Annual ACM Symp. on Theory of Computing, 2002.