A Quick Review of Data Stream Mining Algorithms

Dr. Rajeev Tripathi & Dr. Santosh Kumar Dwivedi
Dept. of Computer Science, SRMGPC
Lucknow, India

Abstract: Limitations of traditional data in supporting streaming applications have been understood, prompting research to supplement existing technologies and build new systems to manage streaming data. In this paper we tried to present how data mining affects data streaming. To review the process for data stream from data, proper algorithm and model are analyzed. Various categories of page segmentation algorithms are mentioned here. In this paper, we present the theoretical foundations of data stream analysis and identify potential stream mining techniques.

1. Introduction

In recent years, melodramatic change has noticed in our ability to collect data from various sources, sensors, devices, in different format, form independent or connected applications. Scientist in real world are now thin king that potential source of “Big Data” is transforming world and giving new horizons to the Private and Government sectors and even in new generations of 2020, we will approximately leading with 35 zeta bytes of data storing capacities in data warehouse. Big data collector will collect data for m various domains such as health care domain, social media & networking, phone &TV, web data, financial data, genome data, retails, public service, environmental monitoring, scientific activity etc. Big Data is now used to convey large sorts of concept, as huge data quantity, real time data, social media data, and much more.

Data mining is a field which is concerned to understanding data patterns from huge datasets. We can say that the aim is to find out new patterns in data. A number of data mining techniques are there like classification, clustering, advanced neural networks, prediction and regression models used for different data mining approaches in various areas. 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.
2. Basics 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. 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.

2.1. Segmentation

Segmentation is a process that determines the elements of an image. The most important point which is necessary to locate the regions of the document where data is printed and distinguish is from figures and graphics. Text segmentation is the isolation of characters or words. Many segmentation algorithms in which segment words are used into isolated characters which are recognized individually. This process of segmentation is performed by isolating each connected component. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts. The problems in segmentation are divided into various categories: Extraction of touching and fragmented characters, distinguishing noise from text, skewing.

2.2. Clustering

Clustering is a data mining method that has not taken its real part in the works already quoted although, the most important algorithm of this method was very studied in the context of privacy preserving, which is k-means algorithm. Surveying privacy preserving k-means clustering approaches apart from other privacy preserving data mining ones is important due to the use of this algorithm in important other areas, like image and signal processing where the problem of security is strongly posed.

2.3. 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. 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.

2.4. 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. 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.

2.5. 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 multidimensional data input. The random sampling method is often the only method of choice for high dimensional applications.

3. Mining Algorithms for Entire Data Stream

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:

3.1. 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.

3.2. Sliding Window

The idea is to carry out a comprehensive analysis of the most recent data and over the old summarized data sets. 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.

3.3. 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.

3.4. Mining Algorithms for Entire Data Stream

Manku and Motwani proposed the first one-pass algorithm, Lossy Counting, to find all frequent item sets over a data stream [1]. Their algorithm is false-positive oriented in the sense that it does not allow false negatives, and has a provable bound on false positives.

3.5. Sliding Window

Chi et al. have studied the problem on mining closed frequent item sets over a sliding window of a data stream [12]. In particular, they assume the width of sliding window is not very large, therefore, the transactions of each sliding window could be held in the main memory. Clear, such assumption is very close to the problem setting of the incremental association rule mining [2]. But their focus is on how to maintain the closed frequent item sets in an efficient way.

3.6. Damped Window Model

Chang and Lee studied the problem to find recently frequent item sets over data streams using the damped window model. Specifically, in their model, the weight for an existing transaction in the data stream reduces by a decay factor, d, as a new transaction arrives.

4. 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:
- 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.
- As real data might be irregular and unpredictable in nature, hence the algorithm should be able to manage the traffic by using optimal resources.

5. Acknowledgment

This research paper is made possible through the help and support from everyone, including: parents, teachers, family, friends, and in essence, all sentient beings. Especially, please allow me to dedicate my acknowledgment of gratitude toward the following significant advisors and contributors: First and foremost, I would like to thank Dr. Vinay Mishra
for his most support and encouragement. He kindly read my paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper. Second, I would like to thank Dr. Vonodini Katiyar and Dr. Brijendra Singh to read my paper and to provide valuable advices. Finally, I sincerely thank to my parents, family, and friends, who provide the advice and financial support. The product of this research paper would not be possible without all of them.

5. References


[8] Shao, H., Zhao, H., Chang, G.: Applying Data Mining to Detect Fraud Behavior in Customs Declaration

[9] Nikunj shah “Fraud Detection: Data Mining & Audit Tools”


[12] H. Wang, W. Fan, P. Yu and I. 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.
