A NOVAL APPROACH TO PREPARE DATA SET USING DATA STREAM MINING

S. Brintha Rajakumari  
Research scholar,  
Department of CSE,  
brintha.ramesh@gmail.com,  
Bharath University, Chennai, India.

Dr. T. Nalini  
Professor,  
Department of CSE,  
dmalinichidabaram@gmail.com,  
Bharath University, Chennai, India.

ABSTRACT: A data stream is an emerging research area and also a challenging problem in present days. Streaming is a technique for transferring data from one place to another. A data stream is a continuous, real time, uninterrupted sequence of coherent data. The paper presents the overall study about data stream and its process model and structure used for data set preparation in data mining analysis.

KEYWORDS: Big Data, Data Streams, Data Mining.

I. INTRODUCTION

Data Stream means the continuous flows of ordering data. A goal of data mining is to discover useful patterns from a given data set or data stream. Example of data stream includes ATM transaction, computer network traffic, phone conversation, and Web Searches and Sensor data. Now a day’s stream mining is an important and attractive area for database and data mining researchers. Mining data streams are a challenging area in the data mining research because new data enter into the database and older data is slow with rapid speed. The main goal of stream mining is the discovery of useful patterns and rule from the data set. Data streams emerging class of applications in recent years, which is often continuous, uncontrolled, high speed and data sharing as time changes [1].

Data streams can be classified into two types, they are offline data streams and online data streams. Web log reports, queries used to retrieve information from data warehouse are examples for offline data streams. Network transactions, sensor data, network traffic monitoring, intrusion detection and credit card fraud detection are examples of online data streams. Data Stream Mining is a process of extracting knowledge structure of continuous, rapid data records [2]. Mining data stream requires high-speed, real time processing in order to maintain a high speed data arrival and results must be attracted within short response time.

In several emerging applications [5], data takes the form of continuous data streams, as opposed to finite stored datasets such as web-server logs, stock tickers, network traffic measurements, click streams, data feeds from sensor networks, and telecom call records. Stream processing is different from computation over traditional stored data set in two important aspects: the sheer volume of a stream over its lifetime could be huge, and queries require timely answers; response times should be small. As a consequence, it is not possible to store the stream in its entirety on secondary storage and scan it when a query arrives.

The paper is organized as follows. Section 2 describes the related work; section 3 describes the definition of data stream mining process model. The architecture of data stream mining discussed in section 4. In Section 5 describes the issues in data streams and finally concluded the paper in section 6.

II. RELATED WORK

Paper [2] discussed about the research issues for applications wherever data arrive from a single stream as well as multiple streams. In this paper [3] presented online active learning in dynamic problems with potentially adversarial concept drifts. It shows the problem using real UGC data from a news portal at Yahoo, that active learning is powerful for reducing labeling efforts in dynamic problems. The paper [4] introduced Hoeffding trees, a method for learning from the high-volume data streams that are increasingly common. Hoeffding trees allow learning in very small constant time per example, and have strong security of a high asymptotic relationship to the corresponding batch trees.

The author presented algorithms for computing frequency counts exceeding a user-specified threshold over Data streams in [5]. Algorithm scans easily be deployed for streams of single to n items like those found in IP network monitoring. On paper [6] proposed an efficient clustering algorithm for analyzing categorical data stream algorithms only require input stream and their storage space depends on the structural parameters of the graphs, the approximation guarantee, and the confidence probability. The paper [8] present ActMiner, which addresses four most important challenges to data stream classification.
such as infinite length, concept drift, concept evolution and limited labeled data.

### III. DATA STREAM MINING PROCESS MODEL

Data mining extracts useful information from large dataset and links with information retrieval, pattern recognition, statistics and machine learning. Data collected from many resources such as medical, financial, scientific, marketing and computer machine data. Now a day’s web server, sensor, internet routers and web servers generate large volume of data. The data size, complexity of data types and user interaction with mining processes are the challenges in data mining.

Stream data mining gives useful patterns and also many issues in the system. They are the types of queries one wants to answer on a data stream, filtering a data stream select element with property x from the stream, counting number of dissimilar elements in the most recent k elements of the stream, estimating average and standard deviation of last k elements, Finding frequent elements, mining query streams such as google wants to identify what queries are more frequent today than yesterday, mining click streams such as yahoo needs to know which of its pages are receiving an odd number of hits in the past hour, mining social network news feeds such as look for trending topics on twitter, facebook sensor networks, many sensors feeding into a central manager telephone call records such as data feeds into customer bills as well as settlements between telephone companies, IP packets monitored at a switch such as gather information for optimal routing and detect denial-of-service attacks.

### V. ARCHITECTURE OF DATA STREAM SYSTEM

Everyday millions of transactions records, searches and call records. Several applications Data feeds from sensor applications, log records or click-streams in web tracking, financial tickers, manufacturing processes, email messages, call detail records in telecommunications, performance measurements in network monitoring and traffic management generate data Streams. The architecture of the data stream system is in the figure 2. All the stream input stored into the various storage devices and processed output stream generated by the query processor will be stored in the output buffer.
The massive Online Analysis tool includes a collection of offline, online methods and tools for data stream mining evaluation. MOA is linked to WEKA such as the Waikato Environment for Knowledge Analysis, which is an open source tool containing implementations of a wide range of machine learning methods written in Java. The main benefits of Java are portability applications which can run on any platform with an appropriate Java virtual machine contains strong well developed support libraries.

VI. CONCLUSION

Mining data streams is a challenging area in the data mining research. The main goal of stream mining is the discovery of useful patterns and rule from the data set. In this paper discussed the issues that need to be considered when designing a stream data mining technique and presented the data stream model and structure used for mining analysis that will be useful to prepare the data set for data mining analysis.

REFERENCES


[4] Pedro Domingos, Geoff Hulten, “Mining High Speed Data Streams”.


[6] Zengyou He, Xiaofei Xu, Shengchun Deng, Joshua Zhexue Huang, “Clustering Categorical data Streams”.


[8] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham “Classification And Novel Class Detection In Data Stream With Active Mining”, Springer- Verlag Berlin Heidelberg, pp.311-324 2010.