# Chapter 14
# Mining Data Streams

**Prasanna Lakshmi Kompalli**
*Gokaraju Rangaraju Institute of Engineering and Technology, India*

## ABSTRACT

*In recent years, advancement in technologies has made it possible for most of the present-day organizations to store and record large streams of data. Such data sets which continuously and rapidly grow over time are referred to as data streams. Mining of such data streams is a unique opportunity and also a challenging task. Data stream mining is a process of gaining knowledge from continuous and rapid records of data. Due to increased streaming information, data stream mining has attracted the research community in the recent past. There is voluminous of literature which has been published in this domain over the past few years. Due to this, isolating the correct literature would be a grueling task for researchers and practitioners. While addressing a real-world problem, it would be more difficult to find relevant information as it would be hidden in data streams. This chapter tries to provide solution as it would be an amalgamation of all techniques used for data stream mining.*

## INTRODUCTION

Innovation in IT has given birth to huge amounts of data. Day to day activities like credit card transactions, web browsing, mobile usage, network traffic also generates huge amount of flowing data. Streams of data so produced contains valuable information. This information can be used in decision making, development of better quality products, finding new relations among existing items etc. As it is not conceivable to manually study all the data, automated techniques must be developed with good computational power for finding valuable and relevant information.

Data Mining deals with design of algorithms that help computers to identify valid and useful patterns, take quick and clever decisions based on empirical data. Data mining approaches work well with large amounts of static data stored in system but does not address the problem of continuously flowing data. Usually, a model created from training data of $i$ instances using data mining cannot be updated with the newly arriving data. For every newly arriving $(i+1)^{th}$ instance of data the whole training process must be repeated. So data mining techniques will be inefficient for addressing the problem of streams of flowing data. Mining data streams for gaining knowledge is a progressive discipline. The difference between Data mining and Data Stream Mining is illustrated in Table 1.

DOI: 10.4018/978-1-5225-4999-4.ch014

*Table 1. Differences between Data Mining and Data Stream Mining*

| Criteria | Data Mining | Data Stream Mining |
|---|---|---|
| Number of iterations | Multiple | Single |
| Processing Time | Limitless | Restricted |
| Memory Used | Limitless | Restricted |
| Access | Random | Sequential |
| Data | Persistent | Transient |

Data Stream mining is an archetype addresses the issues related to continuously arriving data. Handling and processing data steams require single examination of data, fast processing with minimum space utilization, availability of results on the request of user (Prasanna, 2015).

This chapter introduces the methodology and constraints of data stream mining, algorithms developed by well-known researchers. Processing streaming data also needs summarization. Several summarizing techniques used with streaming data are also discussed. It is hoped that the chapter will greatly help and provide a reference to researchers, practitioners and students interested in the emerging domain of data streams.

## BACKGROUND

Studying and formulating new techniques for data stream mining is a challenge encountered by researchers. In data streams data arrives like a stream and if not processed immediately it will be lost forever. Furthermore, assumption is made that data arrives rapidly so it is not practical to store this data in active storage media for longer period of time.

As technology has advanced the methods of collecting data continuously have been simplified. Extraction of knowledge from data streams is becoming a key task for several researchers (Prasanna, 2010). Mining data streams have multitudinous applications with many daunting research issues. Archetypal applications include monitoring of network, web mining, public health surveillance, telecommunication, and financial transactions etc., all these applications are characterized by mining data streams to discover knowledge, which are crucial for strategic decisions. As the entire data stream will not be available at once, processing needs a sequential strategy. The algorithms designed for working with data streams must sometime deal with data sizes which exceed the main memory of system. New methodologies are designed by extending the existing techniques of data mining by including the constraints posed by data streams. Before proceeding to further discussion, a study of the features related to data streams is presented.

### Features of Data Streams

The following features act as constraints over a model built for working with streaming data:

1. Voluminous data arrives continuously and rapidly. Hence not feasible to store it completely. Concise summary generated by modelling data streams can be stored;
2. Streams arrive fast, so each stream must be processed and discarded within limited amount of time;