Clustering Techniques for Streaming Data–A Survey

Yogita
Electronics and Computer Engineering Department
Indian Institute of Technology Roorkee
Roorkee, India
thakranyogita@gmail.com

Durga Toshniwal
Electronics and Computer Engineering Department
Indian Institute of Technology Roorkee
Roorkee, India
durgafec@iitr.ernet.in

***Abstract*-Nowadays many applications are generating streaming data for an example real-time surveillance, internet traffic, sensor data, health monitoring systems, communication networks, online transactions in the financial market and so on. Data Streams are temporally ordered, fast changing, massive, and potentially infinite sequence of data. Data Stream mining is a very challenging problem. This is due to the fact that data streams are of tremendous volume and flows at very high speed which makes it impossible to store and scan streaming data multiple time. Concept evolution in streaming data further magnifies the challenge of working with streaming data. Clustering is a data stream mining task which is very useful to gain insight of data and data characteristics. Clustering is also used as a pre-processing step in over all mining process for an example clustering is used for outlier detection and for building classification model. In this paper we will focus on the challenges and necessary features of data stream clustering techniques, review and compare the literature for data stream clustering by example and variable, describe some real world applications of data stream clustering, and tools for data stream clustering.**

Keywords— **Streaming Data; Data Stream Mining; Clustering.**

I. Introduction

Data streams are temporally ordered, fast changing, massive, and infinite sequence of data objects [1]. Unlike traditional data sets, it is impossible to store an entire data stream or to scan through it multiple times due to its tremendous volume. New concepts may keep evolving in data streams over time. Evolving concepts require data stream processing algorithms to continuously update their models to adapt to the changes.

Data streams are ubiquitous. These can be found in many application domains from online financial transaction to medical systems and space research centers, where satellites are continuously generating streaming data. And even new applications are emerging day by day due to significant growth in computer processing speed and spread of computer networks. So there is a need of effective and efficient data mining techniques for streaming data which can handle the challenges associated with streaming data. Data mining techniques for streaming data includes: clustering, classification, frequent pattern mining and outlier detection which can be used to mine patterns from streaming data. In this paper, we are focusing on clustering. Because it is helpful to gain insight of data and data characteristics and can be used as a pre-processing step with other data mining techniques. For an example cluster based outlier detection [2] and building a classification model using the cluster features.

Data stream clustering discovers clusters in the streaming data. Data stream clustering is very different from traditional clustering:

- For traditional clustering data sets are static, but the data stream is dynamic in nature.
- Because of massive size of data stream, it is not possible to store data streams in memory and scan it multiple times. But for traditional clustering data sets are store in memory and can be scanned multiple times.
- The clustering results of data streams change over time, but not so for traditional clustering.

There are following two different approaches for data stream clustering: Data stream clustering by example and data stream clustering by variable that are described in section 3rd and 4th.

The rest of this paper is organized as follows: Section 2 discusses challenges and requirements of data stream clustering. Section 3 describes and reviews the literature on data stream clustering by example. Section 4 describes and reviews the literature on data stream clustering by variable. Section 5 Compare and summarize the work on data stream clustering. Section 6 presents applications and tools of data stream clustering. Section 7 concludes the paper.

II. Challenges and Necessary Features of Data Stream Clustering Technique

In this section we will discuss the challenges of data stream clustering present in front of data mining community and requirements of a good data stream clustering techniques.

A. Challenges of Data Streams

- **Infinite Size and High Speed –** Data streams are of infinite size and continuously flow at very high speed of data. Due to this it is impossible to store data streams and processing of data streams is computationally very expensive.
- **Dynamic Nature** – Streaming data behaviour keeps on changing over time. Clustering model developed on

978-1-4673-4529-3/12/$31.00 ©2012 IEEE