Clustering of Data Streams: A Second Look

Vasudha Bhatnagar

Department of Computer Science
University of Delhi, Delhi, India
vbhatnagar@cs.du.ac.in

Sharanjit Kaur

Acharya Narendra Dev College
University of Delhi, Delhi, India
skaur@cs.du.ac.in

Revolution in digitized technologies has made it possible to acquire data on-line in the form of data streams, which are continuous and infinite in nature. Multiple applications varying from critical scientific applications to business and financial applications generate transient data. Since streaming data is ordered sequence of continuously growing unlabeled data instances, it is not feasible to apply traditional data mining techniques to reveal hidden, useful and novel patterns.

Potency to consolidate and capture natural structures from unlabeled data has made clustering a popular choice in stream mining. Single scan of data, bounded memory usage, constant per-point processing time and capturing data evolution are the key challenges during clustering of streaming data.

Clustering of data streams has gained popularity because it aids summarization of data characteristics without the pre-requisite of class labels. Applications like detecting network usage pattern in telecommunication, tracking evolution in various phenomena in seismic and metrological studies, network intrusion detection, monitoring spread of diseases etc. require clustering of data streams. Assignment of a data point directly to a cluster is imprudent for streaming data due to limited storage and computing capabilities. Instead, a rich synopsis structure is adopted for summarizing statistics of incoming data points. Subsequently, consolidated information in synopsis is mined for inherent patterns with multifarious applications in data mining.

The objective of this tutorial is to present a comprehensive overview of common approaches used for clustering streams with *emphasis on synopsis selection.* Tutorial begins with a discussion on important issues related to stream clustering, followed by a critical analysis of three main approaches. Some contemporary and well-known algorithms for each approach are discussed. Explanation of algorithms is supported by exemplary applications. Finally, *a formulation of a generic architectural framework for stream clustering algorithms* for better understanding of the issues is discussed.

Bio: Vasudha Bhatnagar did her masters in Computer Applications from University of Delhi, Delhi, India in 1985. She worked in Centre for Development of Telematics from 1985 - 1989 as a software engineer in Operating System and Traffic group. She completed doctoral studies from Jamia Millia Islamia, New Delhi, India in 2001. She is currently an Associate Professor in the Department of Computer Science, University of Delhi, Delhi, India. Her broad area of interest is Intelligent Data Analysis. She is particularly interested in developing process models for Knowledge Discovery in Databases, and algorithms for classification and clustering.

Bio: Sharanjit Kaur did her Ph.D. from Department of Computer Science, University of Delhi, Delhi, India in February, 2011. She started teaching in 1995 in Acharya Narendra Dev College (University of Delhi), Delhi and is currently Associate Professor there. Her research interest spans the area of stream mining and databases.

International Conference on Management of Data
COMAD 2011, Bangalore, India, December 19–21, 2011