AN ADAPTIVE GRID-BASED METHOD FOR CLUSTERING MULTI-DIMENSIONAL ONLINE DATA STREAMS

Toktam Dehghani
Department of Computer Engineering, Ferdowsi University Mashhad, Mashhad, Khorasan Razavi, Iran
dehghani.toktam@stu-mail.um.ac.ir
http://toktamdehghani.com

Mahmoud Naghibzadeh
Department of Computer Engineering, Ferdowsi University Mashhad, Mashhad, Khorasan Razavi, Iran
naghibzadeh@um.ac.ir
http://profsite.um.ac.ir/~naghibzadeh/

Mohammadreza Afsharisaleh
Department of Engineering, Islamic Azad University, Mashhad, Khorasan Razavi, Iran
mohammadrezaafsharisaleh@tooskahealth.com

Abstract:
Clustering is an important task in mining the evolving data streams. A lot of data streams are high dimensional in nature. Clustering in the high dimensional data space is a complex problem, which is inherently more complex for data streams. Most data stream clustering methods are not capable of dealing with high dimensional data streams; therefore they sacrifice the accuracy of clusters. In order to solve this problem we proposed an adaptive grid-based clustering method. Our focus is on providing up-to-date arbitrary shaped clusters along with improving the processing time and bounding the amount of the memory usage. In our method (B+C tree), a structure called “B cell tree” is used to keep the recent information of a data stream. In order to reduce the complexity of the clustering, a structure called “cluster tree” is proposed to maintain multi dimensional clusters. A Cluster tree yields high quality clusters by keeping the boundaries of clusters in a semi-optimal way. Cluster tree captures the dynamic changes of data streams and adjusts the clusters. Our performance study over a number of real and synthetic data streams demonstrates the scalability of algorithm on the number of dimensions and data without sacrificing the accuracy of identified clusters.

Keywords: data streams; data mining; clustering; grid-based clustering; high dimensional data streams.

1. Introduction

During the recent years, data streams have attracted attention in different applications of computer science, such as customer click streams, multimedia data, sensor data, network monitoring, telecommunication system, stock markets. A data stream is defined as a massive unbounded sequence of data elements continuously generated at a rapid rate [Park and Lee (2007)]. Management and processing of these online rapid unbounded streams raises new challenges because the traditional algorithms are usually not feasible to perform operations [Beringer and H"{u}llermeier (2003)]. Online data stream processing should satisfy the following requirements [Park and Lee (2007)]:

1. Each data element should be examined at most once to analyze a data stream.
2. Memory usage for data stream analysis should be confined finitely although new elements are continuously generated in a data stream.
3. Newly generated data elements should be processed as fast as possible to produce the up-to-date analysis result of a data stream.

Clustering refers to the process of grouping a collection of objects into "clusters" such that objects within the same class that are similar in a certain sense, and objects from different classes that are dissimilar[Berger and H"{u}llermeier (2003)]. Clustering of data streams has been studied in recent years but a few of these methods can effectively cluster large multi dimensional data streams. In this paper, we consider the problem of on-line