Amini A, Wah TY, Saboohi H. On density-based data streams clustering algorithms: A survey. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(1): 116–141 Jan. 2014. DOI 10.1007/s11390-013-1416-3

On Density-Based Data Streams Clustering Algorithms: A Survey

Amineh Amini, *Member*, *IEEE*, Teh Ying Wah, and Hadi Saboohi, *Member*, *ACM*, *IEEE*

Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya Kuala Lumpur 50603, Malaysia

E-mail: amini@siswa.um.edu.my; tehyw@um.edu.my; saboohi@siswa.um.edu.my

Received April 11, 2013; revised November 13, 2013.

Abstract Clustering data streams has drawn lots of attention in the last few years due to their ever-growing presence. Data streams put additional challenges on clustering such as limited time and memory and one pass clustering. Furthermore, discovering clusters with arbitrary shapes is very important in data stream applications. Data streams are infinite and evolving over time, and we do not have any knowledge about the number of clusters. In a data stream environment due to various factors, some noise appears occasionally. Density-based method is a remarkable class in clustering data streams, which has the ability to discover arbitrary shape clusters and to detect noise. Furthermore, it does not need the number of clusters in advance. Due to data stream characteristics, the traditional density-based clustering is not applicable. Recently, a lot of density-based clustering algorithms are extended for data streams. The main idea in these algorithms is using density-based methods in the clustering process and at the same time overcoming the constraints, which are put out by data stream's nature. The purpose of this paper is to shed light on some algorithms in the literature on density-based clustering over data streams. We not only summarize the main density-based clustering algorithms on data streams, discuss their uniqueness and limitations, but also explain how they address the challenges in clustering data streams. Moreover, we investigate the evaluation metrics used in validating cluster quality and measuring algorithms' performance. It is hoped that this survey will serve as a steppingstone for researchers studying data streams clustering, particularly density-based algorithms.

Keywords data stream, density-based clustering, grid-based clustering, micro-clustering

1 Introduction

Every day, we create 2.5 quintillion bytes of data; 90 percent of current data in the world has been created in the last two years alone. This data overtakes our capability to store and to process. In 2007, the amount of information created exceeded available storage for the first time. For example, in 1998 Google indexed 26 million pages, by 2000 it reached one billion, and in 2012 Google indexed over 30 trillion Web pages. This dramatic expansion can be attributed to social networking applications, such as Facebook and Twitter.

In fact, we have a huge amount of data generated continuously as data streams from different applications. Valuable information must be discovered from these data to help improve the quality of life and make our world a better place. Mining data streams is related to extracting knowledge structure represented in streams information. The research of mining data streams has attracted a considerable amount of researchers due to the importance of its application and the increasing generation of data streams[1-6].

Clustering is a significant class in mining data streams[5,7-11]. The goal of clustering is to group the streaming data into meaningful classes. Clustering data streams puts additional challenges to traditional data clustering such as limited time and memory, and further one pass clustering.

It is desirable for clustering data streams to have an algorithm which is able to, first discover clusters of arbitrary shapes, second handle noise, and third cluster without prior knowledge of number of clusters. There are various kinds of clustering algorithms for data streams. Among them, density-based clustering has emerged as a worthwhile class for data streams due to the following characteristics:

Firstly, it can discover clusters with arbitrary shapes. Partitioning-based methods are restricted to clusters structured on a convex-shaped. Discovery of clusters with a broad variety of shapes is very important for many data stream applications. For example, in the environment observations the layout of an area with similar environment conditions can be any shape.

Regular Paper
The work is supported by the University of Malaya Research under Grant No. RG097-12ICT.

Secondly, it has no assumption on the number of clusters. Most of the methods require previous knowledge of the domain to determine the best input parameters. Nevertheless, there is not a priori knowledge in a large amount of real life data.

Finally, it has the ability to handle outliers. For instance, due to the influence of different factors such as temporary failure of sensors in data stream scenario, some random noises appear occasionally. Detecting noise is one of the important issues specifically in evolving data streams in which the role of real data changes to noise over time.

There are different surveys recently been published in the literature for mining data streams. A number of them survey the theoretical foundations and mining techniques in data streams[2,12-14] as well as clustering as a significant class of mining data streams. Some of them review the well-known clustering methods in datasets[15-16]. Five clustering algorithms in data streams are reviewed and compared based on different characteristics of the algorithms in [17]. Furthermore, [18-20] review papers on different approaches in clustering data streams based on density. The work presented in [21] surveys existing clustering methods on data stream and gives a brief review on density-based methods. Different from them, this paper is a thorough survey of state-of-the-art density-based clustering algorithms over data streams.

Motivation. In real world applications, naturally occurring clusters are typically not spherical in shape and there are large amounts of noise or outliers in some of them. Density-based clustering can be applicable in any real world application. They can reflect the real distribution of data, can handle noise or outliers effectively, and do not make any assumptions on the number of clusters. Therefore, they are more appropriate than other clustering methods for data stream environments. Density-based method is an important data stream clustering topic, which to the best of our knowledge, has not yet been given a comprehensive coverage. This work is a comprehensive survey on the density-based clustering algorithms on data stream. We decouple density-based clustering algorithms in two different categories based on the techniques they use, which help the reader understand the methods clearly. In each category, we explain the algorithms in detail, including their merits and limitations. The reader will then understand how the algorithms overcome challenging issues. Moreover, it addresses an important issue of the clustering process regarding the quality assessment of the clustering results.

The remainder of this paper is organized as follows. In the next section, we discuss about the basic and challenges of clustering data streams as well as density-based clustering validation. Section 3 overviews the density-based clustering algorithms for data streams. Section 4 examines how the algorithms overcome the challenging issues and also compares them based on evaluation metrics. Finally, Section 5 concludes our study and introduces some open issues in density-based clustering for data streams.

2 Clustering Data Streams

Clustering is a key data mining task[5,7-11] which classifies a given dataset into groups (clusters) such that the data points in a cluster are more similar to each other than the points in different clusters.

Unlike clustering static datasets, clustering data streams poses many new challenges. Data stream comes continuously and the amount of data is unbounded. Therefore it is impossible to keep the entire data stream in main memory. Data stream passes only once, so multiple scans are infeasible. Moreover data stream requires fast and real time processing to keep up with the high rate of data arrival and mining results are expected to be available within short response time.

There are an extensive number of clustering algorithms for static datasets[15-16] where some of them have been extended for data streams. Generally, clustering methods are classified into five major categories[22]: partitioning, hierarchical, density-based, grid-based, and model-based methods (Fig.1).

A partitioning-based clustering algorithm organizes the objects into some number of partitions, where each partition represents a cluster. The clusters are formed based on a distance function like k-means algorithm[23-24] which leads to finding only spherical clusters and the clustering results are usually influenced by

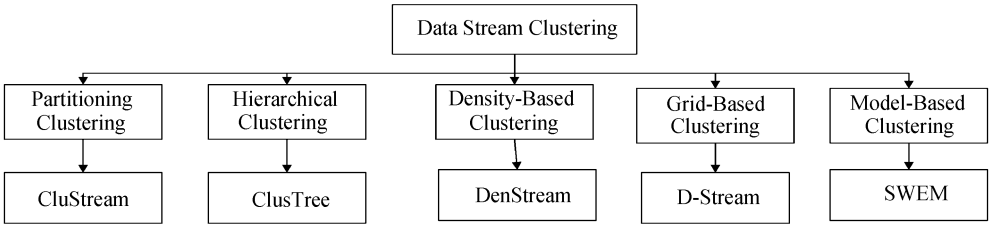

Fig.1. Data stream clustering algorithms[22].