Clustering Algorithms for Data Streams

Christos Makris
University of Patras, Greece

Nikos Tsirakis
University of Patras, Greece

INTRODUCTION

The World Wide Web has rapidly become the dominant Internet tool which has overwhelmed us with a combination of rich hypertext information, multimedia data and various resources of dynamic information. This evolution in conjunction with the immense amount of available information imposes the need of new computational methods and techniques in order to provide, in a systematical way, useful information among billions of Web pages. In other words, this situation poses great challenges for providing knowledge from Web-based information. The area of data mining has arisen over the last decade to address this type of issues. There are many methods, techniques and algorithms that accomplish different tasks in this area. All these efforts examine the data and try to find a model that fits to their characteristics in order to examine them. Data can be either typical information from files, databases and so forth, or with the form of a stream. Streams constitute a data model where information is an undifferentiated, byte-by-byte flow that passes over the time. The area of algorithms for processing data streams and associated applications has become an emerging area of interest, especially when all this is done over the Web. Generally, there are many data mining functions (Tan, Steinbach, & Kumar, 2006) that can be applied in data streams. Among them one can discriminate clustering, which belongs to the descriptive data mining models. Clustering is a useful and ubiquitous tool in data analysis.

BACKGROUND

Data Mining and Knowledge Discovery

Classic algorithms handle small amounts of data and face up performance problems when data are huge in capacity. For example, a sorting algorithm runs efficiently with some megabytes of data but could have difficulties in running for some gigabytes of data. Many methods such as clustering and classification have been widely studied in the data mining community. However, a majority of such methods may not be working effectively on data streams. This happens because data streams provide huge volumes of data and at the same time require online mining, in which we wish to mine the data in a continuous fashion. Generally, there are many specific problems with traditional algorithms. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. In addition, it gives new opportunities for exploring and analyzing new types of data and for analyzing old types of data with new ways. Data mining is an integral part of knowledge discovery in databases (KDD). These two terms are often used interchangeably (Dunham, 2003). Over the last few years, KDD has been used to refer to a process consisting of many phases, while data mining is only one of these phases. Below are some definitions of knowledge discovery in databases and data mining (Fayyad, Piatetsky-Shapiro, &Smyth, 1996a, 1996b).

- **Knowledge discovery in databases (KDD):** Is the process for finding useful information and patterns in data.

Knowledge discovery in databases is a process that involves five different phases which are listed bellow (Dunham, 2003):

1. Data selection
2. Data preprocessing
3. Data transformation
4. Data mining
5. Data interpretation/evaluation

Data mining attempts to autonomously extract useful information or knowledge from large data stores or sets. It involves many different algorithms to accomplish different tasks. All these algorithms attempt to fit a model to the data. The algorithms examine the data and determine a model that is closest to the characteristics of the data being examined. These algorithms consist of three parts:

- **Model:** The purpose of the algorithm is to fit to the data.
- **Preference:** Some criteria must be used to fit one model over another.

- **Search:** All algorithms require some technique to search the data.

There are many different methods used to perform data mining tasks. These techniques not only require specific types of data structures, but also imply certain types of algorithmic approaches. Data mining tasks are generally divided into two different categories.

- **Predictive tasks:** These tasks predict the value of a particular attribute based on the values of other attributes. Predictive tasks include classification, regression, time series analysis and prediction.
- **Descriptive tasks:** Here, the objective is to derive patterns or relationships in data. Descriptive tasks include clustering, summarization, association rules and sequence discovery.

CLUSTERING

Clustering and other mining techniques have grasped the interest of the data mining community. It is alternatively referred to as unsupervised learning or segmentation. In broad strokes, clustering is the problem of finding a partition of a data set so that, under some definition of "similarity," similar items are in the same part of partition and different items are in different parts. These parts are called clusters. Items can be any type of data in any form. The most "common used" form of items in clustering are vectors. These vectors consist of *d*-dimensions in the Euclidian or generally metric space, so we talk about clustering in many dimensions. Some techniques meet some limitations in the value of *d* and most of the times *d* is the metric which differentiates algorithms.

Sometimes it is useful to use a threshold in order to specify that all objects in a cluster must be sufficiently close to one another. A more enlightening definition could be (Dunham, 2003):

- We have a set of alike elements. Elements from different clusters are not alike.
- The distance between points in a cluster is less than the distance between a point in the cluster and any point outside it.

A proportional process in data bases is segmentation where alike records are grouped together. Clustering is one of the most useful processes of data mining for cluster recognition and to define patterns and trends over data. It is similar to classification except that the groups are not predefined, but rather extracted by the specific distribution of the data.

Classification of Clustering Algorithms

There are various classes of clustering algorithms depending, each time, on the method for the clusters definition. The most common division of them is as *hierarchical* or *partitional*. In the first class of algorithms, we have a nested set of clusters. In addition, each level of hierarchy has a separate set of clusters. At the lowest level, each item is in its own unique cluster and at the highest level all items belong to the same cluster. Hierarchical clustering does not have as input the desired number of clusters. In the second class of algorithms, we meet only one set of clusters. Apart from these two classes of clustering algorithms, there are some recent studies that look at *categorical* data and are targeted to *large databases*. Algorithms targeted to large databases may adopt memory constraints by either *sampling* the database or using data structures which can be *compressed* or pruned to fit into memory regardless of the size of the database. Another approach of clustering algorithms is to further classify them based on the implementation technique used.

Data Streams

Traditional data bases store static data without the sense of time apart from the case the time is a part of the data. As time elapses the data explode and there is a need of online processing and analysis of data from many applications. This need made current methods deficient and processed data took another form and name. Data streams are data which change continuously over the time in a fast rate. There are many types of applications like network monitoring, telecommunications data managements, clickstream monitoring, manufacturing, sensor networks and many others, where data have the form of streams and not finite sum of data. In these applications users make continuous queries in contrast with the classic queries (one-time queries).

Models

We consider an input stream with x_i items that arrives sequentially and describes a signal X, a one-dimensional function X: $[1....N] \rightarrow R^2$. Models differ on how x_i's describe X and can be divided into three categories (Muthukrishnan, 2003):

- **Time series model:** This model is suitable for time series data such as observing the traffic at an IP link for a predefined time horizon (e.g., every 5 minutes).
- **Cash register model:** This is the most popular data stream model. It is suitable for applications such as monitoring IP addresses that access a Web server and source IP addresses that send packages to a link.
- **Turnstile model:** This is the most general model and it is appropriate in order to study fully dynamic situations where there are inserts as well as deletes, but it is often hard to get interesting bounds in this model.