Procedia Computer Science

Volume 80, 2016, Pages 461–471

ICCS 2016. The International Conference on Computational Science

Improving Multivariate Data Streams Clustering

Christian C. Bones[1], Luciana A. S. Romani[2], and Elaine P. M. de Sousa[1]*

[1] University of São Paulo, São Carlos, SP, Brazil. {chris, parros}@icmc.usp.br
[2] Embrapa Agricultural Informatics, Campinas, SP, Brazil.
luciana.romani@embrapa.br

Abstract

Clustering data streams is an important task in data mining research. Recently, some algorithms have been proposed to cluster data streams as a whole, but just few of them deal with multivariate data streams. Even so, these algorithms merely aggregate the attributes without touching upon the correlation among them. In order to overcome this issue, we propose a new framework to cluster multivariate data streams based on their evolving behavior over time, exploring the correlations among their attributes by computing the fractal dimension. Experimental results with climate data streams show that the clusters' quality and compactness can be improved compared to the competing method, leading to the thoughtfulness that attributes correlations cannot be put aside. In fact, the clusters' compactness are 7 to 25 times better using our method. Our framework also proves to be an useful tool to assist meteorologists in understanding the climate behavior along a period of time.

Keywords: Clustering, Data Streams, Data Mining, Fractal

1 Introduction

Extracting valid and useful knowledge from data streams is an important and costly task for many environmental science fields such as ecology, geology, atmospheric science, to name a few. An increasing number of devices and sensors have generated a huge amount of data incessantly, leading to new challenges and applications. For instance, sensors have been used to monitor the pollution in cities, the level of rivers and the meteorological conditions. Extracting valuable information from these flows of data could be helpful to avoid disasters, such as flooding or the extinction of fragile plants due to sudden changes in temperature.

In this scenario, clustering of data streams becomes an active research topic [2, 12, 7, 5, 18, 14, 17] with applications in several contexts. Clustering aims to group in the same cluster data streams that have similar properties and behavior over time, whereas data streams of different clusters must present dissimilar characteristics. For illustration, clustering techniques could be applied to cluster meteorological data stream sensors that have similar behavior along a period of time.

*The authors are grateful to CAPES, CNPQ and FAPESP for their financial support.

doi:10.1016/j.procs.2016.05.325

However, clustering data streams requires appropriate solutions for challenging issues, mainly: to capture and represent data evolution along the time; to deal with all the attributes of each stream, i.e., multivariate data streams; to take into account the correlation among the attributes; to read data only once; to provide answers as soon as the user demands them; to deal with outlier data streams. Therefore, in the past few years some methods have been proposed to process and analyze flows of data in real time [15, 14, 18, 17, 8, 16, 1], aiming to overcome some of these challenges. However, only few of them try to pull out valuable information and group the entire data streams based on their similar behavior over the time [14, 17, 8, 16]. Furthermore, most of these methods either do not support multivariate data streams [14, 17, 16] or only consider the similarity of attributes independently [15, 13, 8].

In this paper, we propose a new approach to cluster multivariate data streams, taking into account the entire data streams and their (dis)similar behavior along the time, bearing in mind the data streams evolution. Our approach also considers the correlation among the attributes of each data stream, aiming to improve the clusters' quality. We thus propose the framework eFCDS - Evolving Fractal-based Clustering of Data Streams. Its main module is a novel algorithm for clustering multivariate data streams, based on the continuous calculation of the attributes correlation by the fractal dimension. Also, it tracks the evolution of the data streams by checking cluster membership whenever a new value of fractal dimension is obtained. In other words, our method checks whether the data stream still belongs to the same cluster or it should be allocated in another one that better describes its behavior in that period of time. Another module of eFCDS checks whether an outlier data stream could be associated to one of the regular clusters without disrupt the clusters' formation rules. It also detects overlapping between the generated clusters and merge them when their union does not extrapolate a maximum standard deviation.

Finally, we apply our framework to climate data streams from meteorological sensors, which usually have more than one attribute (e.g. temperature and precipitation) and presumably there are some correlations among them. We conducted an experimental study on data from different Brazilian regions, provided by Agritempo[1]. Our results not only indicated that our approach can be useful to assist specialists in analyzing large amounts of climate data, which is relevant to current climate research, but also helps to identify regions with the same behavior along the time.

The rest of this paper is organized as follows. Section 2 presents background concepts and related work. Section 3 describes our approach to cluster data streams. Experimental results are discussed in Section 4 and Section 5 presents final remarks and future work.

2 Background and Related Work

In order to be considered a data stream the data collection must be generated continuously by one or more sources, i.e, $d_1, d_2, ...$ [11]. Moreover, each d_i could have more than one attribute, characterizing it as a multivariate data stream. Formally, let $\mathbb{S} = \{S_1, \ldots, S_n\}$ be a set of data streams sources where each $S_i = \left\{\vec{d_1}, \ldots, \vec{d_\infty}\right\}$ is a multivariate data stream. Also, each S_i is assumed to contain f attributes such that $\vec{d_i} = [a_1, \ldots, a_f]$ is the set of attributes.

Clustering in a data stream environment can be very costly [11], due to some basic requirements that must be presented in the algorithms [3]: (i) Representation of compact size; (ii) Quickly and incremental processing of new data items; (iii) Traceability of changes in groups; (iv) Quick and clear identification of outliers.

[1] www.agritempo.gov.br