Procedia Computer Science

Volume 53, 2015, Pages 183–190

2015 INNS Conference on Big Data

A Time Series Clustering Technique based on Community Detection in Networks

Leonardo N. Ferreira[1] and Liang Zhao[2]

[1] Institute of Mathematics and Computer Science, University of São Paulo
Av. Trabalhador São-carlense, 400 CEP: 13566-590 - Centro, São Carlos - SP, Brazil.
leonardo@icmc.usp.br

[2] Department of Computing and Mathematics, University of São Paulo
Av. Bandeirantes, 3900 - CEP: 14040-901 - Monte Alegre - Ribeirão Preto - SP, Brazil.
zhao@usp.br

Abstract

Time series clustering is a research topic of practical importance in temporal data mining. The goal is to identify groups of similar time series in a data base. In this paper, we propose a technique for time series clustering via community detection in complex networks. First, we construct a network where every vertex represents a time series connected its most similar ones,. Similarity was calculated using different time series distance functions. Then, we applied a community detection algorithm to identify groups of strongly connected vertices in order to produce time series clusters. We verified which distance function works better with every clustering algorithm and compared them to our approach. The experimental results show that our approach statistically outperformed many traditional clustering algorithms. We find that the community detection approach can detect groups that other techniques fail to identify.

Keywords: Time series clustering, Unsupervised Learning, Complex networks, Community detection.

1 Introduction

Temporal data mining has received a lot of focus in the last years due to the ubiquity of this kind of data. Time series data clustering is a specific task with the goal of dividing a set of time series into groups, where similar ones are grouped in the same cluster [8]. This problem has been observed in many domains like climatology, geology, health sciences, energy consumption, failure detection, among others [20].

The two desired aspects when performing time series clustering is effectiveness and efficiency [19]. Effectiveness can be achieved by representation methods that should be capable of dealing with high dimensional data. Efficiency is obtained by using distance functions and clustering algorithms that can properly distinguish different time series in an efficient way. Keeping these two features in mind, many clustering algorithms have been proposed and they can be

Selection and peer-review under responsibility of the Scientific Programme Committee of INNS-BigData2015

doi:10.1016/j.procs.2015.07.293

broadly classified in two approaches: data-adaptation and algorithm-adaptation [20]. The former extracts features arrays from each time series data and applies conventional clustering algorithm. The latter use specially designed clustering algorithms to handle time series. The major modification is the distance function that should be capable of differing time series. In our best effort, we did not find any community detection algorithm in complex networks for time series clustering.

Complex networks form a recent research area interested in networks that have complex topology, are dynamically evolving in time and have large scale [4]. Many real world systems can be modeled by networks. One of the salient features found in many networks is the presence of community structure, which is represented by groups of densely connected vertices and, at the same time, with sparse connections between groups. Detecting these structures is interesting in many applications and it motivated the development of many community detection algorithms [13]. Different from traditional clustering algorithms, community detection algorithms use just the network structure to cluster nodes. Furthermore, considering the high dimension of these networks, community detection algorithms should be efficient and effective, what make them excellent for time series data clustering.

In this paper, we aim to apply network science to temporal data mining. We intend to verify the benefits of using community detection algorithms in time series data clustering. More specifically, we propose an algorithm based on 4 steps: (1) data normalization; (2) distance function calculation; (3) network construction, where every vertex represents a time series connected to its most similar ones using a distance function; (4) community detection. Experimental results show that this approach outperforms traditional clustering methods. The remainder of this paper is organized as follows. First, we present in Section 2 some background concepts and related works. In Sections 3 and 4 we present our approach and the experimental comparison respectively. Finally, we point some final remarks and future works in Section 5.

2 Background and Related Works

In this section, we review the two main concepts used in this paper: time series distance functions and community detection in networks. We also present some related works.

2.1 Time Series Distance Functions

We start by presenting the basic concept: time series. For simplicity and without loss of generality, we assume that time is discrete. A time series X is an ordered sequence of t real values $X = \{x_1, \ldots, x_t\}, x_i \in \mathbb{R}, i \in \mathbb{N}$. The main idea of clustering is to group similar objects. In order to discover which data are similar, several distance (or dissimilarity) functions were defined in the literature. In this paper, we use the terms "similarity" and "distance" in inverse concepts. In the case of time series distance measures, they can be classified in four categories [8]: shape-based, edit-based, feature-based, structure-based. Each of them will be following described.

- *Shaped Based Distance Functions*: The first and most used category of measures is based on the shape of the time series. These measures compare directly the raw data of two time series. The most common measures are the L_p norms that have the form:

$$d_{L_p}(X, Y) = \left(\sum_{i=1}^{t} (x_i - y_i)^p \right)^{\frac{1}{p}}, \tag{1}$$