Information Sciences 326 (2016) 227–242

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Time series clustering via community detection in networks

Leonardo N. Ferreira [a,*], Liang Zhao [b]

[a] Institute of Mathematics and Computer Science, University of São Paulo Av. Trabalhador São-carlense, 400 CEP: 13566–590 - Centro, São Carlos - SP, Brazil

[b] Department of Computing and Mathematics, University of São Paulo Av. Bandeirantes, 3900 - CEP: 14040–901 - Monte Alegre - Ribeirão Preto - SP, Brazil

ARTICLE INFO

Article history:
Received 13 February 2015
Revised 8 June 2015
Accepted 18 July 2015
Available online 3 August 2015

Keywords:
Time series data mining
Time series clustering
Complex networks
Community detection

ABSTRACT

In this paper, we propose a technique for time series clustering using community detection in complex networks. Firstly, we present a method to transform a set of time series into a network using different distance functions, where each time series is represented by a vertex and the most similar ones are connected. Then, we apply community detection algorithms to identify groups of strongly connected vertices (called a community) and, consequently, identify time series clusters. Still in this paper, we make a comprehensive analysis on the influence of various combinations of time series distance functions, network generation methods and community detection techniques on clustering results. Experimental study shows that the proposed network-based approach achieves better results than various classic or up-to-date clustering techniques under consideration. Statistical tests confirm that the proposed method outperforms some classic clustering algorithms, such as *k*-medoids, diana, median-linkage and centroid-linkage in various data sets. Interestingly, the proposed method can effectively detect shape patterns presented in time series due to the topological structure of the underlying network constructed in the clustering process. At the same time, other techniques fail to identify such patterns. Moreover, the proposed method is robust enough to group time series presenting similar pattern but with time shifts and/or amplitude variations. In summary, the main point of the proposed method is the transformation of time series from time-space domain to topological domain. Therefore, we hope that our approach contributes not only for time series clustering, but also for general time series analysis tasks.

1. Introduction

Time series data mining has received a lot of attention in the last years due to the ubiquity of this kind of data. One specific task is clustering with the goal to divide a set of time series into groups, where similar ones are put in the same cluster [12]. Such kind of problems has been observed in many application domains like climatology, geology, health sciences, energy consumption, failure detection, among others [35].

The two desired aspects when performing time series clustering are effectiveness and efficiency [34]. 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 those can be broadly classified into two approaches:

* Corresponding author. Tel.: +55 3592251907.
E-mail addresses: leonardo@icmc.usp.br, ferreira@leonardonascimento.com (L.N. Ferreira), zhao@usp.br (L. Zhao).

http://dx.doi.org/10.1016/j.ins.2015.07.046
0020-0255/

data-adaptation and algorithm-adaptation [35]. The former extracts feature arrays from each time series and then applies a clustering algorithm in its original form. The latter uses specially designed clustering algorithms to directly handle time series. In this case, the major modification is the distance function, which should be capable of distinguishing time series.

Complex networks form a recent and interesting research area. Here, a complex network refers to a large scale network with non-trivial connection pattern [4]. Many real-world systems can be modeled by networks. One of the salient features 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 such structures is interesting in many real applications. For this reason, many community detection algorithms have been developed [14] and such algorithms present a powerful mechanism for general data mining tasks. A brief review of community detection techniques will be given in the next section.

In the original form of time series, only the local relationship among neighbor data samples can be easily identified, while long distance global relationship remains unknown in general. On the other hand, time series analysis, such as time series clustering, classification or prediction, requires not only local information, but also global knowledge to capture the pattern formation of given time series. Network (graph) is a powerful mechanism, which is able to characterize relationship between any pair or any groups of data samples. Therefore, the transformation from time series to network representation is hopefully to present an alternative way for time series analysis. From the technical view point, network-based clustering techniques also present attractive advantage, as described below. Up to now, majority of existing time series clustering techniques in literature use k-means, k-medoids or hierarchical clustering algorithms in their original forms or modified versions. The common feature of these algorithms is that they try to break data samples into clusters in such a way such that the partition optimizes a criterion defined by a given distance function. As a consequence, these techniques can just find clusters of a specific shape already determined by the distance function. For example, k-means with the Euclidean distance function can only produce Gaussian distributed clusters. On the other hand, it has been shown that network-based clustering techniques can capture arbitrary cluster shapes. This is because the network-based techniques identify connectivity patterns of the input data and such patterns can be any shape in the Euclidean space. Finally, many community detection techniques have been proposed and some of them have even linear time complexity when the constructed network is sparse [31]. This feature also makes them attractive to 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 including 4 steps of processing: 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, where each community represents a time series cluster. In summary, this paper presents the following contributions:

- The main contribution is the proposal of using community detection in complex networks for time series clustering. For this purpose, we transform time series from time-space domain to topological domain. Since network is a general representation, which has ability to characterize both local and global relationship among nodes (representing data samples), our approach is useful not only for time series clustering, but also for other kinds of time series analysis tasks. To our knowledge, applying community detection techniques for time series clustering has not been reported in the literature;
- Extensive numerical study has been conducted in this paper. Specifically, we study, in the time series clustering context, combinations of time series data sets, time series distance functions, network construction methods and community detection algorithms. In comparison to other time series clustering algorithms, experimental results and statistical tests show that the network-based approach present better results.
- Last but not least, the proposed method presents some desired features when applied to real clustering problems. It can effectively detect shape patterns presented in time series due to the topological structure of the underlying network constructed in the clustering process. At the same time, other techniques studied in this paper fail to identify such patterns. Moreover, the proposed method is robust enough to group time series presenting similar pattern but with time shifts and/or amplitude variations.

The remainder of this paper is organized as follows. Firstly, we present in Section 2 some background concepts and related works to this paper. In Sections 3 and 4 we present our approach and the experimental results, respectively. Finally, we point some final remarks and future works in Section 5.

2. Background and related works

In this section, we review the three main components of time series clustering used in this paper: time series distance measures, clustering algorithms and community detection in networks.

2.1. Time series distance measures

We start by presenting the basic concept: time series. For simplicity and without loss of generality, we assume that time is discrete.

Definition 1 (Time Series)**.** A time series X is an ordered sequence of t real values $X = \{x_1, \ldots, x_t\}, x_i \in \mathbb{R}, i \in \mathbb{N}$.