Information Systems 53 (2015) 16–38

Contents lists available at ScienceDirect

Information Systems

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

Time-series clustering – A decade review

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi*, Teh Ying Wah

*Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya (UM),
50603 Kuala Lumpur, Malaysia*

ARTICLE INFO

Article history:
Received 13 October 2014
Accepted 27 April 2015
Available online 6 May 2015

Keywords:
Clustering
Time-series
Distance measure
Evaluation measure
Representations

ABSTRACT

Clustering is a solution for classifying enormous data when there is not any early knowledge about classes. With emerging new concepts like cloud computing and big data and their vast applications in recent years, research works have been increased on unsupervised solutions like clustering algorithms to extract knowledge from this avalanche of data. Clustering time-series data has been used in diverse scientific areas to discover patterns which empower data analysts to extract valuable information from complex and massive datasets. In case of huge datasets, using supervised classification solutions is almost impossible, while clustering can solve this problem using unsupervised approaches. In this research work, the focus is on time-series data, which is one of the popular data types in clustering problems and is broadly used from gene expression data in biology to stock market analysis in finance. This review will expose four main components of time-series clustering and is aimed to represent an updated investigation on the trend of improvements in efficiency, quality and complexity of clustering time-series approaches during the last decade and enlighten new paths for future works.

1. Introduction

Clustering is a data mining technique where similar data are placed into related or homogeneous groups without advanced knowledge of the groups' definitions [1]. In detail, clusters are formed by grouping objects that have maximum similarity with other objects within the group, and minimum similarity with objects in other groups. It is a useful approach for exploratory data analysis as it identifies structure(s) in an unlabelled dataset by objectively organizing data into similar groups. Moreover, clustering is used for exploratory data analysis for summary generation and as a pre-processing step for other data mining tasks or as a part of a complex system.

With increasing power of data storages and processors, real-world applications have found the chance to store and keep data for a long time. Hence, data in many applications is being stored in the form of time-series data, for example sales data, stock prices, exchange rates in finance, weather data, biomedical measurements (e.g., blood pressure and electrocardiogram measurements), biometrics data (image data for facial recognition), particle tracking in physics, etc. Accordingly, different works are found in variety of domains such as Bioinformatics and Biology, Genetics, Multimedia [2–4] and Finance. This amount of time-series data has provided the opportunity of analysing time-series for many researchers in data mining communities in the last decade. Consequently, many researches and projects relevant to analysing time-series have been performed in various areas for different purposes such as: subsequence matching, anomaly detection, motif discovery [5], indexing, clustering,

* Corresponding author. Tel.: +60 196918918.
E-mail addresses: saeed@um.edu.my (S. Aghabozorgi),
shirkhorshidi_ali@siswa.um.edu.my,
Shirkhorshidi_ali@yahoo.co.uk (A. Seyed Shirkhorshidi),
tehyw@um.edu.my (T. Ying Wah).

http://dx.doi.org/10.1016/j.is.2015.04.007
0306-4379/© 2015 Elsevier Ltd. All rights reserved.

classification [6], visualization [7], segmentation [8], identifying patterns, trend analysis, summarization [9], and forecasting. Moreover, there are many on-going research projects aimed to improve the existing techniques [10,11].

In the recent decade, there has been a considerable amount of changes and developments in time-series clustering area that are caused by emerging concepts such as big data and cloud computing which increased size of datasets exponentially. For example, one hour of ECG (electrocardiogram) data occupies 1 gigabyte, a typical weblog requires 5 gigabytes per week, the space shuttle database has 200 gigabytes and updating it requires 2 gigabytes per day [12]. Consequently, clustering craved for improvements in recent years to cope with this incremental avalanche of data to keep its reputation as a helpful data-mining tool for extracting useful patterns and knowledge from big datasets. This review is opportune, because despite the considerable changes in the area, there is not a comprehensive review on anatomy and structure of time-series clustering. There are some surveys and reviews that focus on comparative aspects of time-series clustering experiments [6,13–17] but none of them tend to be as comprehensive as we are in this review. This research work is aimed to represent an updated investigation on the trend of improvements in efficiency, quality and complexity of clustering time-series approaches during the last decade and enlighten new paths for future works.

1.1. *Time-series clustering*

A special type of clustering is time-series clustering. A sequence composed of a series of nominal symbols from a particular alphabet is usually called a temporal sequence, and a sequence of continuous, real-valued elements, is known as a time-series [15]. A time-series is essentially classified as dynamic data because its feature values change as a function of time, which means that the value(s) of each point of a time-series is/are one or more observations that are made chronologically. Time-series data is a type of temporal data which is naturally high dimensional and large in data size [6,17,18]. Time-series data are of interest due to their ubiquity in various areas ranging from science, engineering, business, finance, economics, healthcare, to government [16]. While each time-series is consisting of a large number of data points it can also be seen as a single object [19]. Clustering such complex objects is particularly advantageous because it leads to discovery of interesting patterns in time-series datasets. As these patterns can be either frequent or rare patterns, several research challenges have arisen such as: developing methods to recognize dynamic changes in time-series, anomaly and intrusion detection, process control, and character recognition [20–22]. More applications of time-series data are discussed in Section 1.2. To highlight the importance and the need for clustering time-series datasets, potentially overlapping objectives for clustering of time-series data are given as follows:

1. Time-series databases contain valuable information that can be obtained through pattern discovery. Clustering is a common solution performed to uncover these patterns on time-series datasets.
2. Time-series databases are very large and cannot be handled well by human inspectors. Hence, many users prefer to deal with structured datasets rather than very large datasets. As a result, time-series data are represented as a set of groups of similar time-series by aggregation of data in non-overlapping clusters or by a taxonomy as a hierarchy of abstract concepts.
3. Time-series clustering is the most-used approach as an exploratory technique, and also as a subroutine in more complex data mining algorithms, such as rule discovery, indexing, classification, and anomaly detection [22].
4. Representing time-series cluster structures as visual images (visualization of time-series data) can help users quickly understand the structure of data, clusters, anomalies, and other regularities in datasets.

The problem of clustering of time-series data is formally defined as follows:

Definition 1:. Time-series clustering, given a dataset of n time-series data $D=\{F_1,F_2,..,F_n\}$, the process of unsupervised partitioning of D into $C=\{C_1,C_2,..,C_k\}$, in such a way that homogenous time-series are grouped together based on a certain similarity measure, is called time-series clustering. Then, C_i is called a cluster, where $D=\cup_{i=1}^{k} C_i$ and $C_i \cap C_j = \varnothing$ for $i \neq j$.

Time-series clustering is a challenging issue because first of all, time-series data are often far larger than memory size and consequently they are stored on disks. This leads to an exponential decrease in speed of the clustering process. Second challenge is that time-series data are often high dimensional [23,24] which makes handling these data difficult for many clustering algorithms [25] and also slows down the process of clustering [26]. Finally, the third challenge addresses the similarity measures that are used to make the clusters. To do so, similar time-series should be found which needs time-series similarity matching that is the process of calculating the similarity among the whole time-series using a similarity measure. This process is also known as “whole sequence matching” where whole lengths of time-series are considered during distance calculation. However, the process is complicated, because time-series data are naturally noisy and include outliers and shifts [18], at the other hand the length of time-series varies and the distance among them needs to be calculated. These common issues have made the similarity measure a major challenge for data miners.

1.2. *Applications of time-series clustering*

Clustering of time-series data is mostly utilized for discovery of interesting patterns in time-series datasets [27,28]. This task itself, fall into two categories: The first group is the one which is used to find patterns that frequently appears in the dataset [29,30]. The second group are methods to discover patterns which happened in datasets surprisingly [31–34]. Briefly, finding the clusters of time-series can be advantageous in different domains to answer following real world problems:

Anomaly, novelty or discord detection: Anomaly detection are methods to discover unusual and unexpected patterns which happen in datasets surprisingly [31–34]. For example,