Engineering Applications of Artificial Intelligence 39 (2015) 235–244

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Fuzzy clustering of time series data using dynamic time warping distance

Hesam Izakian [a,*], Witold Pedrycz [a,b,c], Iqbal Jamal [d]

[a] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, T6G 2V4
[b] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[c] System Research Institute, Polish Academy of Sciences, Warsaw 00-716, Poland
[d] AQL Management Consulting Inc., Edmonton, AB, Canada

ARTICLE INFO

Article history:
Received 11 July 2014
Received in revised form
26 December 2014
Accepted 30 December 2014
Available online 17 January 2015

Keywords:
Clustering time series
Dynamic Time Warping (DTW)
Fuzzy clustering
Hybrid approach

ABSTRACT

Clustering is a powerful vehicle to reveal and visualize structure of data. When dealing with time series, selecting a suitable measure to evaluate the similarities/dissimilarities within the data becomes necessary and subsequently it exhibits a significant impact on the results of clustering. This selection should be based upon the nature of time series and the application itself. When grouping time series based on their shape information is of interest (shape-based clustering), using a Dynamic Time Warping (DTW) distance is a desirable choice. Using stretching or compressing segments of temporal data, DTW determines an optimal match between any two time series. In this way, time series exhibiting similar patterns occurring at different time periods, are considered as being similar. Although DTW is a suitable choice for comparing data with respect to their shape information, calculating the average of a collection of time series (which is required in clustering methods) based on this distance becomes a challenging problem. As the result, employing clustering techniques like K-Means and Fuzzy C-Means (where the cluster centers – prototypes are calculated through averaging the data) along with the DTW distance is a challenging task and may produce unsatisfactory results. In this study, three alternatives for fuzzy clustering of time series using DTW distance are proposed. In the first method, a DTW-based averaging technique proposed in the literature, has been applied to the Fuzzy C-Means clustering. The second method considers a Fuzzy C-Medoids clustering, while the third alternative comes as a hybrid technique, which exploits the advantages of both the Fuzzy C-Means and Fuzzy C-Medoids when clustering time series. Experimental studies are reported over a set of time series coming from the UCR time series database.

1. Introduction

Time series are commonly encountered in numerous application areas including finances, medicine, engineering, and environmental science. Considering high dimensionality and substantial volume of time series collected in different applications, extracting and visualizing available structure in this type of data is highly beneficial and exhibits numerous potential applications in data summarization, anomaly detection, etc.

In this study, we discuss and contrast a number of alternatives for fuzzy clustering of time series to reveal and visualize the available structure within this type of data. Fuzzy clustering is one of the widely used clustering techniques where, instead of assigning data to individual cluster, the Boolean-like nature of assignment is relaxed by assigning membership grades that assume values in the unit interval and quantify a strength of belongingness of a data point to the individual cluster. Fuzzy C-Means (FCM) (Bezdek, 1981) and Fuzzy C-Medoids (FCMdd) (Krishnapuram et al., 2001) are the two well-known and representative fuzzy clustering techniques. In both techniques, the objective is to form a number of cluster centers (prototypes) and a partition matrix so that a given performance index becomes minimized. FCMdd selects the cluster centers as some of the existing data points (medoids) whereas FCM generates a set of cluster centers using a weighted average of data. In both techniques, the intent is to minimize a weighted sum of distances between data points and cluster centers.

Selecting a distance function to evaluate similarities/dissimilarities of time series has a significant impact on the clustering algorithms and their final results produced by them. This selection may depend upon the nature of the data and the specificity of the application. In most partition-based time series data clustering techniques, the Euclidean distance is commonly used to quantify the similarities/

* Corresponding author: Tel: +1 780 7169026.
E-mail addresses: izakian@ualberta.ca (H. Izakian), wpedrycz@ualberta.ca (W. Pedrycz), iqbaljamal@aqlmc.com (I. Jamal).

http://dx.doi.org/10.1016/j.engappai.2014.12.015
0952-1976/

dissimilarities of time series (or their representation). However, in this technique, one compares the points of time series in a fixed order and cannot take into account existing time shifts. Furthermore, this distance function is applicable only when comparing equal-length time series. On the other hand, in most representation-based (feature-based) clustering techniques, the representatives of clusters cannot be reconstructed in the original time series domain and in such a way they are not useful for data summarization.

In this study, we use Dynamic Time Warping (Berndt and Clifford, 1994) (DTW) distance for clustering time series data. DTW is the most well known technique for evaluating similarity/dissimilarity of time series with respect to their shape information. It is a commonly encountered method and different versions of this technique have been reported in the literature (e.g., see (Salvador and Chan, 2007; Jeong et al., 2011; Bankó and Abonyi, 2012; Chu et al., 2002; Keogh and Pazzani, 1999)) for evaluating similarity/dissimilarity of time series. This distance function determines an optimal match between two time series by stretching or compressing some segments of the series. As the result, patterns occurring at different time instances of time series are considered as similar and consequently, this technique evaluates the similarity of time series with respect to their shapes. Moreover, this technique can express the distance between non-equal-length time series.

As mentioned earlier, DTW distance is a suitable measure to evaluate the similarities/dissimilarities of time series with respect to their shape information. However, there are some difficulties in applying this technique to partition-based clustering methods. Among several reasons behind this, time complexity of this technique is quadratic (Salvador and Chan, 2007) and calculating the average of a set of time series based on this measure is a challenging problem.

In this study, we propose and evaluate three alternatives for fuzzy clustering of time series data using DTW distance. These techniques cluster the time series data with respect to their shape information. Furthermore, the prototypes generated during the clustering process can be used for data summarization based on the shape information within the time series.

Using a DTW-based averaging technique reported in the literature (Petitjean et al., 2011), a Fuzzy C-Means (FCM) clustering is proposed. As the second method, a Fuzzy C-Medoids (FCMdd) clustering, that is free from calculating averages of time series, is examined to select a number of optimal cluster centers as well as an optimal partition matrix. As the third alternative, a hybrid of Fuzzy C-Means and Fuzzy C-Medoids technique is considered for clustering time series data. In all these techniques, time series are clustered based on shape similarities (using the DTW distance) and the cluster centers are in time series domain (not their representation). As a result, the centers can be considered as representatives of time series.

Partition-based clustering of time series data using DTW distance is a challenging problem that has been addressed in this study. Employing a DTW-based averaging technique in FCM is a novel idea presented in this work. Moreover, the proposed hybrid technique which exploits the merits of FCM and FCMdd for clustering time series data realizes a novel idea that has been proposed and investigated in this paper.

The study is structured as follows. In Section 2, we review the proposed methods for clustering time series data. In Section 3 the DTW distance along with an averaging technique based on this measure is briefly reviewed. Section 4 discusses different alternatives for fuzzy clustering of time series data using DTW, and Section 5 reports on the experimental studies. Finally, Section 6 concludes the paper.

2. Literature review

In this section, we briefly review some well-known similarity/dissimilarity measures of time series as well as some clustering techniques reported in the literature for this type of data.

Similarity measures used in time series data can be divided into three general categories including L_p-norm distances, elastic measures, and statistical techniques (Izakian et al., 2013). Selecting a similarity measure in time series data mining depends on the nature of data and the nature of the application itself. When comparing two time series based on a fixed mapping of their points is of interest, L_p-norm distances can be used. The most commonly used examples of L_p-norm are L_1 (Manhattan), L_2 (Euclidean), and L_∞ (Tchebyschev). These distances can be applied to compare time series in their original or a representation domain.

In Izakian et al. (2013), Izakian and Pedrycz (2014) and Izakian and Pedrycz (2014), authors presented an augmented version of Euclidean distance function for fuzzy clustering of time series data. The original time series as well as different representation techniques, including Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), and Piecewise Aggregate Approximation (PAA) were examined for clustering purpose. D'Urso and Maharaj (2009), transformed the time series data through their autocorrelation representation, and used the Euclidean distance to compare data in the new feature space. Then, a FCM technique was employed to cluster the transformed data. In Izakian and Pedrycz (2013), a clustering-based technique for anomaly detection in time series data was proposed. For detecting anomalies in the amplitude of time series, a Fuzzy C-Means clustering applied to the original representation of time series and the Euclidean distance function was employed as a dissimilarity measure. For the purpose of detecting anomalies in the shape of time series, first the data are transformed into an autocorrelation representation, and then the Euclidean distance was employed to compare time series in the transformed domain.

In Vlachos et al. (2003), time series data were represented using a Haar wavelet transform, and the K-Means algorithm along with the Euclidean distance employed to cluster data in the new feature space. In Maharaj and D'Urso (2011), time series are represented through a set of cepstral coefficients, and Euclidean distance is employed to quantify the dissimilarity of time series in the process of a Fuzzy C-Means clustering. Möller-Levet et al. (2003), represented time series data through piecewise linear functions, and proposed a short time series distance, measured as the sum of squared Euclidean distances between the corresponding slopes encountered in two time series. The Fuzzy C-Means algorithm was realized to cluster the data in the new feature space. In Nanda et al., (2010), the Euclidean distance was considered to cluster stock market time series using the K-Means, Fuzzy C-Means, and a self organization map for building a portfolio. The experimental results showed that K-Means could generate more compact clusters in comparison with the other clustering techniques.

Dynamic time warping distance (DTW) (Berndt and Clifford, 1994), longest common subsequence (LCSS) (Vlachos et al., 2002), and edit distance of real-number sequences (EDR) (Chen et al., 2005) are located in the elastic measures category. DTW helps to find an optimal match between two time series by stretching or compressing their segments, and evaluate the similarity of time series with respect to their shapes. LCSS employs the length of the longest subsequence occurring in two time series to quantify their similarity, and EDR takes into account the number of insert, delete and replace operations required to convert one sequence to another one to determine their similarity.

Authors in Niennattrakul and Ratanamahatana (2007) examined K-Means and C-Medoids algorithms for clustering time series data using dynamic time warping distance function. Experimental results indicated that the K-Means clustering cannot generate acceptable results when this distance function is considered (because of the problem of averaging time series based on this measure), and instead, C-Medoids technique may generate satisfactory results. In Keogh and Pazzani (1999), authors proposed a hierarchical clustering technique of time series data, and a DTW distance was considered to quantify the dissimilarity of time series. In Petitjean and Gançarski (2012), a DTW-based averaging of time series is proposed using a compact multiple