International Journal of Approximate Reasoning 99 (2018) 12–38

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Robust fuzzy clustering of multivariate time trajectories

Pierpaolo D'Urso [a,*], Livia De Giovanni [b], Riccardo Massari [a]

[a] Department of Social Sciences and Economics, Sapienza University of Rome, Italy
[b] Department of Political Sciences, LUISS Guido Carli, Rome, Italy

ARTICLE INFO

Article history:
Received 11 November 2017
Received in revised form 27 February 2018
Accepted 7 May 2018
Available online 9 May 2018

Keywords:
Outlier time trajectory
Cross-sectional and longitudinal clustering
Dynamic time warping
Exponential distance
Robust fuzzy clustering
Partitioning around medoids

ABSTRACT

The detection of patterns in multivariate time series is a relevant task, especially for large datasets. In this paper, four clustering models for multivariate time series are proposed, with the following characteristics. First, the Partitioning Around Medoids (PAM) framework is considered. Among the different approaches to the clustering of multivariate time series, the observation-based is adopted. To cope with the complexity of the features of each multivariate time series and the associated assignment uncertainty a fuzzy clustering approach is adopted. Finally, to neutralize the effect of possible outliers, a robust metric approach is used, i.e., the exponential transformation of dissimilarity measures. The proposed models are robust extensions of the Fuzzy C-Medoids clustering algorithm for multivariate time series. With respect to the management of the time behaviour, four variants are proposed: the Cross-Sectional Fuzzy C-Medoids clustering model with exponential transformation (CS-Exp-FCMd) classifies the multivariate time series taking into account their respective instantaneous features; the Longitudinal Fuzzy C-Medoids clustering model with exponential transformation (L-Exp-FCMd) takes into account the evolutive (longitudinal) features; the Mixed Fuzzy C-Medoids clustering model with exponential transformation (M-Exp-FCMd) which consider simultaneously both the instantaneous and the longitudinal features in the clustering process; the Dynamic Time Warping-based Fuzzy C-Medoids model with exponential transformation (DTW-Exp-FCMd) uses the Dynamic Time Warping (DTW) distance. Three simulation studies show the clustering performance of the proposed models in presence of outliers, compared to their non-robust counterparts, and to other models proposed in the literature. An application on real-world data on the concentration of three pollutants in nineteen stations in the Metropolitan City of Rome shows the relevance of the robustness to outliers in the identification of the clusters.

1. Introduction

Multivariate time series, or multivariate time trajectories, are encountered in several research domains, like economics, finance, biomedical sciences, environmental studies, sociology, etc.

The empirical information embedded in multivariate time series presents a three-way structure "units × variables × times" [49]. For example, we can observe the same economic aggregates on UE countries over years, or a set of socio-economic indicators on a sample of workers observed for a period.

* Corresponding author.
E-mail addresses: pierpaolo.durso@uniroma1.it (P. D'Urso), ldegiovanni@luiss.it (L. De Giovanni), ricccardo.massari@uniroma1.it (R. Massari).

https://doi.org/10.1016/j.ijar.2018.05.002
0888-613X/

The task of clustering both univariate and multivariate time series has interested an increasing number of researchers and scholars especially in recent years (see, e.g., [73,66,70,34,32,43,61,60,62,42,41,59]), also due to the increasing availability of these data and of the techniques used to analyse them.

The aims of clustering time series are various. Clustering is a primary tool to discover patterns in large databases of time series [56]. The clustering task summarizes the information by means of clusters' prototypes. The graphical visualization and/or the descriptive analysis of the clusters' structure can help in easily detecting the main information in a dataset, such as regularities and anomalies. In particular, the discovery of anomalies could be of primary importance to avoid the disruptive effect of the presence of outliers [48].

The clustering of univariate and multivariate time series could be undertaken adopting different approaches (for a more detailed review, see [53,1,6,3]):

- Observation-based clustering: by means of this approach, clustering is based on a comparison of the observed time series, or a suitable transformation of them. This approach is useful in particular in the case when the time series are not very long [74,7,64,15,8–13,29,17–19,55,54,46,20,34,35,43,61,60,41].
- Feature-based clustering: this approach is particularly useful when one deals with long and noisy time series. Clustering is based on features extracted in the time domain, frequency domain, or wavelet decomposition of the time series [24,25,57,58,68].
- Model-based clustering: for these models, which are among the earliest works on time series clustering, it is assumed that a set of time series generated from the same model would most likely have similar patterns. The time series are clustered by means of parameter estimates or by means of the residuals of the fitted models [23,21,22,26,27,16].

In the present work we will adopt the observation-based clustering approach.

Following this approach, in the literature several distance measures and clustering models have been suggested for classifying time trajectories. Different dissimilarity measures for comparing and clustering (e.g., using a hierarchical clustering method) multivariate time trajectories have been proposed by Carlier [7], D'Urso and Vichi [29], D'Urso [17] and Coppi and D'Urso [8]. Košmelj [50] presented a two-steps procedure for calculating a dissimilarity for time-varying data with clustering aims. Košmelj and Batagelj [51] classify time-varying data considering an unsupervised approach in an exploratory framework. Sato and Sato [64] analyzed the fuzzy partitioning for three-way data with time occasions by following a cross sectional approach and solving a multicriteria optimization problem. Sato-Ilic and Sato [65] constructed a model to dynamically represent changing clusters embedding the concepts of conventional dynamic MDS or dynamic PCA into the additive clustering model. D'Urso [17] proposed different fuzzy clustering models for multivariate time trajectories based on their cross sectional and longitudinal features (i.e., position, variation and acceleration of the time trajectories). By considering the same time-dynamic features, D'Urso [19] proposed different robust fuzzy clustering models capable to neutralize the negative effects of possible outlier time trajectories in the clustering process. Vlachos et al. [69] developed a similarity measure based on the longest common subsequence to mine multivariate time trajectories. More efficient and less computational complex variants of Dynamic Time Warping (DTW) for classifying time series data were introduced by Keogh and Ratanamahatana [47] and Begum et al. [4]. Singhal and Seborg [66] proposed to adopt a C-Means algorithm implemented with similarity factors to cluster multivariate time series. Coppi and D'Urso [11] proposed fuzzy clustering models for time trajectories based on Shannon entropy. Fuzzy clustering models for time trajectories based on the Partitioning Around Medoids (PAM, [45]) approach have been proposed by Coppi et al. [12]. D'Urso and De Giovanni [20] introduced a Self-Organizing Map for multivariate time trajectories with an application on telecommunications data. Coppi et al. [13] defined different fuzzy clustering models for time trajectories with spatial information. Genolini and Falissard [34] introduced a C-Means based algorithm to cluster longitudinal data. Genolini et al. [35] expanded the clustering model to the case of multivariate longitudinal data. Petitjean et al. [61] proposed a procedure to averaging DTW-based dissimilarities for clustering time series within a C-Means approach. Izakian et al. [41] suggested a fuzzy clustering using DTW distance. Distance measures and fuzzy clustering models for imprecise time trajectories have been suggested by Coppi and D'Urso [9,10].

In time series clustering a relevant problem is related to the complexity of the feature space. In our opinion, a fuzzy approach is natural way to cope with the uncertainty of the assignment of each multivariate time series to each cluster. For instance, a multivariate time series, could have a dynamic pattern consistent with a given cluster for a certain time period and then a completely different dynamic more similar to another cluster. This switching behaviour, which is peculiar of time series objects, cannot be taken into account with a traditional "hard" approach, while it is easily described with a fuzzy approach.

Another issue is related to the definition of a prototype for each cluster. A prototype is an object which retains the main characteristics of its cluster. In the case of "static" cluster analysis, a natural choice is the (weighted) mean of the features of the objects belonging the cluster, a "centroid". In the case of time-varying cluster analysis, a centroid multivariate time series is more difficult to comprehend and to achieve, even if in the literature there are some proposals. For instance, Petitjean et al. [61] proposed a global way to average time series for clustering tasks. In our opinion, a more natural way to address this issue is to follow a PAM approach. Thus, the cluster's prototype is an observed representative multivariate time series, the "medoid". PAM is also convenient from a computational point of view. By adopting this approach it is possible to compute the distance matrix only once, since data do not change during the iterative clustering procedure.