Chemometrics and Intelligent Laboratory Systems 141 (2015) 107–124

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Time series clustering by a robust autoregressive metric with application to air pollution

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

[a] Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, P.za Aldo Moro, 5-00185 Rome, Italy
[b] Dipartimento di Scienze Politiche, LUISS Guido Carli, Viale Romania, 32-00197 Rome, Italy

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form 8 November 2014
Accepted 10 November 2014
Available online 18 November 2014

Keywords:
Autoregressive model
Robust fuzzy C-medoids clustering
Outliers
Environmental chemistry
Pollutant concentration
Nitrogen monoxide (NO) emissions

ABSTRACT

In this paper, following a fuzzy approach and adopting an autoregressive parameterization, we propose a robust clustering model for classifying time series. In particular, by adopting a fuzzy partitioning around medoids approach, the suggested clustering model is able to define the so-called medoid time series, which is a representative time series of each cluster, and the membership degrees of each time series to the different clusters. The robustness of the proposed clustering model is guaranteed by the adoption of a suitable robust metric for time series, i.e. the so-called exponential distance measure. In this way, the clustering model is able to tolerate the presence of outlier time series in the clustering process. In particular, it is capable of neutralizing and smoothing the disruptive effect of outlier time series, preserving the original clustering structure of the dataset, by assigning to outlier time series approximately the same membership degrees across clusters. To illustrate the usefulness and effectiveness of the suggested time series clustering model, a simulation study and an application to air pollution time series are carried out. Comparison with some existing clustering procedures suggested in the literature shows several advantages of the proposed model.

1. Introduction

The literature on time series clustering methods has increased considerably over the last two decades, with a large range of applications in many different fields, including environmental sciences, pharmaceutics, genetics, neurosciences, computational biology, biomedical sciences, finance, econophysics, and neuromarketing. In particular, in the experimental studies the usefulness and effectiveness of the time series clustering proves to be of particular interest. For instance, in *genetics* time series clustering has been used to group genes considering profiles of time expression from cDNA microarrays experiments; in *biomedicine*, to classify signals caused by particular illnesses connected to those of healthy people (i.e., EEG and EMG time series); in *neuroscience*, to classify fMRI (functional magnetic resonance imaging) time series; in *pharmaceutics*, to cluster drug effects attending to patients' time-response after drug intake; in chemometrics, to classify natural products according to chemical profiles recorded at different times (e.g. chemical composition of wines in different years); in environmetrics, for checking the performances of an environmental monitoring network based on a set of air pollutant emissions recorded in different times (air pollution time series) by a set of monitoring stations, it is useful to classify the stations in homogeneous clusters.

For some references on the application of time series clustering/classification on the above-mentioned and other experimental areas, see [1].

As we can see below, we focus our empirical attention on the clustering of environmental time series. For an overview of the literature on the theoretical aspects of time series clustering/classification and their applications in environmental sciences and other experimental fields, see Section 2.

In this paper, we propose a robust clustering model for classifying time series based on a suitable parametric representation of univariate time series, i.e. an autoregressive representation. In particular, following a fuzzy approach, the proposed robust clustering model is based on the partitioning around medoids procedure. The robustness of the clustering model is guaranteed by the adoption of a proper robust metric, i.e. the so-called exponential distance measure. In this way, our clustering model is able to tolerate the presence of outlier time series in the clustering process; by neutralizing and smoothing the disruptive effect of outlier time series, preserving as almost invariant the clustering structure of the dataset, by assigning to outlier time series almost the same membership degrees across clusters. The proposed model is suitable for clustering time series exhibiting persistence over time. Notice that the fuzzy approach also allow us to identify peculiar patterns of time series, like switching time series, i.e. time series showing a pattern typical of a given cluster during a certain time period, and a completely different pattern (representative of another cluster) in another time period [2,3].

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

http://dx.doi.org/10.1016/j.chemolab.2014.11.003
0169-7439/© 2014 Elsevier B.V. All rights reserved.

For detailed discussions on the benefits connected to autoregressive model-based clustering approach, fuzzy methodology, partitioning around medoids procedure see D'Urso et al. [2,3]. We observe that the fuzzy approach has been employed for clustering/classification of sequences in other research areas, for instance in computational biology [4–15].

Following the guidelines proposed in [16–23], we will document the proposed algorithm for clustering time series according to the following procedures in order to make its development clearer and more useful: (i) select the information to be extracted by time series; (ii) select a proper distance measure for comparing time series; (iii) introduce or develop a powerful algorithm to operate the clustering; (iv) properly evaluate the accuracy of the clustering model; (iv) establish resources for the development of the algorithm that are publicly available.

The paper is organized as follows. In Section 2, we present a review of the literature on time series clustering/classification methods, with particular attention to their application to environmental sciences. In Section 3, we introduce the fuzzy robust clustering model for time series. To illustrate the performances of the clustering model, and to assess its classification accuracy, we present and discuss the results of a simulation study in Section 4. As we focus our attention in particular on the usefulness and effectiveness of the time series clustering in environmental sciences, in Section 5 we utilize our clustering model for classifying air pollution time series. Conclusions are provided in Section 6.

2. Literature on clustering/classification of time series

In the literature several time series clustering methods have been suggested. For a survey on possible theoretical approaches, see, e.g., [1,24].

We remark that, following [1,24], time series clustering methods can be classified into three classes: 1) observation-based clustering: the time series clustering methods belonging to this approach are based on the actual time observations, i.e. observed time series or their transformations (see, e.g. [25,26]); 2) feature-based clustering: in this case, the methods are based on features derived for the time series (see, e.g., [27–31]); 3) model-based clustering: these methods are based on parameters estimates of model fitted to the time series (e.g. ARMA or ARIMA models) (see, e.g., [2,3]).

In this paper, we adopt the model-based approach. As we focus our attention in particular on the usefulness of time series clustering in environmental sciences in Section 2.1 we show a detailed overview on this field.

2.1. Literature of time series clustering/classification in environmental sciences

Clustering and classification methods have a crucial role in monitoring of the air quality [3]. In fact, “air quality monitoring is the main tool of local governments for the management and evaluation of air quality status. This practice follows technical regulation. An air monitoring network is usually composed of sites which measure atmospheric pollutants and weather variables. Classification of these monitoring stations is a method of network analysis and optimization. Classification highlights similarities among sites with respect to pollutant concentration levels and/or temporal profiles. Displaying groups on a map allows the identification of spatial patterns” [32]. Moreover, “in designing and maintaining a cost-effective monitoring network, it is important to recognize similarities and differences in the evolution of the variables sampled at different sites, in order to avoid or, at least, reduce redundancy. On the other hand, information collected by a redundant network, for example in the initial exploratory phase of a surveillance monitoring, could be extremely useful for partitioning a transition water body into homogeneous regions, for which different quality objectives may be established. With regard to the above issues, cluster analysis methods (or unsupervised classification) can play a very important role” [33].

In the literature, different clustering-based techniques have been proposed for analyzing air pollution. Sanchez Gomes and Ramos Martin [34] considered the *C*-means clustering method for identifying sources in Valladoid (Spain). Bohm et al. [35] used cluster analysis for detecting temporal patterns of ozone. Sanchez et al. [36], Dorling et al. [37] and Ruijgrok and Romer [38] considered pollutant data and wind data in the cluster analysis. Miranda et al. [39] analyzed the concentration in Mexico City utilizing the correlation coefficient and the Ward clustering method. Romo-Groger et al. [40] considered a similar analysis in Chile using the average linkage algorithm. Ludwig et al. [41] used cluster analysis for studying the daily ozone maxima in California. Lavecchia et al. [42] employed a complete linkage-based procedure on the monitoring network in Lombardia (Northern Italy) to evaluate similarities among ozone monitoring sites in terms of concentration levels and temporal trends. The authors compared the ozone patterns by using the Euclidean distance and the correlation coefficient in the clustering procedure. Wongphatarakul et al. [43] clustered sampled sites with similar characteristics by considering $PM_{2.5}$ chemical databases from seven sites around the world. By considering the Euclidean distance and the Ward clustering method, Ionescu et al. [44] classified estimated pollutant concentration fields obtained by utilizing the so-called *thin plate spline* functions, analyzing nitrogen dioxide data during peak episodes in Paris. Hierarchical clustering has been used to identify distinct sources of volatile organic compounds based on the grouping of the measured concentrations [45]. Moreover, hierarchical clustering could provide a description of regional chemical and transport processes associated with particular regimes and could provide information about the most relevant sources in the development of pollution episodes. Saksena et al. [46] clustered monitoring sites in Delhi on the basis of nine years of monthly average concentration data for three pollutants, i.e. nitrogen dioxide, sulfur dioxide and suspended particulate matter. They considered four agglomerative hierarchical clustering methods -i.e. average linkage, single linkage, complete linkage and centroid method- and Euclidean and Squared Euclidean distance. They observed that the most consistent results are obtained using the Euclidean distance and the average linkage method. A study of the data from Santiago's monitoring network was done by Silva and Quiroz [47] considering an index of multivariate effectiveness, based on Shannon information index. They found that air pollution data (CO, PM_{10}, O_3 and SO_2) from one of the stations (Parque O'Higgins) could be reproduced by using information from the other stations. In order to identify the representative stations for subsequent analysis of ozone concentration Gabusi and Volta [48] considered a hierarchical clustering approach for classifying Northern Italy measurement stations. Beaver and Palazoglu [49] used an aggregate solution of *k*-means clustering to characterize classes of ozone episodes occurring in the San Francisco bay. Cluster analysis based on the Pearson correlation coefficient is used by Gramsh et al. [50] on particulate matter and ozone data collected by Santiago de Chile's network to find city sectors with similar pollution behavior. Cluster analysis has been used to cluster back trajectories, in order to identify different classes of synoptic regimes over the duration of the trajectories [51,52]. Morlini [53] classified monitoring stations of ozone, sulfur dioxide, and carbon monoxide in Emilia Romagna region (Northern Italy) using a dynamic time-warping cost function as dissimilarity measure in average and complete linkage algorithms. In this framework, cluster analysis is used to classify fields obtained from observed data to identify “prototype” of spatial patterns. By considering a functional representation, Bengtsson and Cavanaugh [54] modeled the observed time series in a state space setup and classified the sites via hierarchical clustering methods relying on disparity measures based on Kullback information. Kim et al. [55] employed *k*-means clustering for classifying sites based on the temporal fluctuation of PM2.5. In order to identify city areas with similar air pollution behavior and to locate emission sources, Pires et al. [56,57]