Applied Mathematics and Computation 243 (2014) 522–527

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Using independent component for clustering of time series data

Thelma Sáfadi

Department of Exact Sciences, Federal University of Lavras, Lavras, Minas Gerais, Brazil

ARTICLE INFO

Keywords:
Cluster analysis
Independent component analysis
Sea level

ABSTRACT

In this work we propose the use of independent component analysis for clustering time series. Considering different numbers of independent components, the complete linkage method was used to identify groups based on the estimated coefficients of the mixing matrix. The use of independent component analysis not only enables the clustering of time series as also provides us with information about the characteristics common to groups from the analysis of the components. The analysis is exemplified for time series of sea levels in different countries during the period of 26 years. The dendrogram obtained for 2 independent components showed four groups: one contains only Hong Kong, the second is formed by Malaysia and Thailand. The other two groups are formed by Australia, New Zealand and Brazil, Japan, Alaska, Singapore and Norway. We have shown that, using data sea level, the independent component analysis can reveal the underlying structure in the database and is a powerful tool for clustering of time series.

1. Introduction

Various algorithms have been developed to cluster different types of time series data. Almost all try to modify the existing algorithms for clustering static data in such a way that time series data can be handled or to convert time series data into the form of static data so that the existing algorithms for clustering static data can be directly used. The former approach usually works directly with raw time series data, thus called raw-data-based approach, and the major modification lies in replacing the distance/similarity measure for static data with an appropriate one for time series. The latter approach first converts a raw time series data either into a feature vector of lower dimension or a number of model parameters, and then applies a conventional clustering algorithm to the extracted feature vectors or model parameters, thus called feature- and model-based approach, respectively.

A survey of clustering time series is presented by Liao [12], this work presents methods for static and time series data. Almost all clustering algorithms requires a measure to compute the distance or similarity between two time series being compared. Depending upon whether the data are discrete-valued or real-valued and whether time series are of equal or unequal length, a particular measure might be more appropriate than another.

According to [15], the Microarray Time Series (MTS) data analysis methodologies can be divided into two classes. The first one assumes the observations on the expression at each time as independent variables, and so the usual methods such as hierarchical process [6] and the optimization [17] can be directly used to cluster genes with similar expression pattern. The second, performs clustering based on the set of parameters estimates from specific models, therefore, it is considered more interesting from the statistical viewpoint, since the temporal expression behavior can be taken into account in the clustering.

E-mail address: safadi@dex.ufla.br

http://dx.doi.org/10.1016/j.amc.2014.06.027
0096-3003/

Considering the model-based methods, and the difficulty of applying them to a large number of small series, Nascimento [13] proposed a Bayesian method that considers simultaneously an autoregressive panel data model fit and a hierarchical clustering of the parameter estimates from this model.

To examine the behavior of the volatilities of the main world stock market indices, Alencar and Sáfadi [1] analyzed the daily data for S&P500 (US), Shanghai Comp Index (China), FTSE100 (UK), CAC40 (France), DAX (Germany), S&P/TSX (Canada), Bovespa (Brazil), Merval (Argentina), Nikkei 225 (Japan) during the period from January 4th, 2008 to April 11th, 2011. There are several possible methods to cluster the volatilities, they consider two of them. The first method consider the comparison of estimates of the parameters in a APARCH model, which consider the baseline level of the volatility. The second method estimate the volatilities also using APARCH models and uses correlation coefficients to clusters these indices.

This paper intends to use independent component analysis technique for clustering of time series data. A hierarchical clustering is applied to the parameter estimates of the mixing matrix. The methodology is exemplified to the sea level data from 10 different countries during the period from 26 years. For the analysis we used the R software [14].

2. Independent component analysis

Blind source separation (BSS), a well-known problem, aims at recovering the sources from a set of observations. Applications include separating individual voices in cocktail party. BSS is a difficult task because we do not have any information about the sources and the mixing process. Independent component analysis, ICA, is a method tackling this problem by assuming that the sources are independent to each other [9], and finds the demixing matrix and corresponding independent signals from the observations with some criteria making the separated signals as independent as possible. Independent component analysis is a dimension reduction technique that uses the existence of independent factors in multivariate data and decomposes an input data set into statistically independent components. ICA can reduce the effects of noise or artifacts of the signal and is ideal for separating mixed signals.

ICA can also be contrasted with principal component analysis (PCA). Both ICA and PCA linearly transform the observed signals into components. The key difference however, is in the type of the components obtained. The goal of PCA is to obtain principal components which are uncorrelated. Moreover, PCA gives projections of the data in the direction of the maximum variance. The principal components are ordered in terms of their variances. In ICA however, we seek to obtain statistically independent components. PCA algorithms use only second order information. On the other hand, ICA algorithms may use higher order statistical information for separating the signals (see for example [4]). For this reason non-Gaussian signals (or at most, one Gaussian signal) are normally required for ICA algorithms based on higher order statistics [8,7,2].

Various ICA algorithms have been proposed. A group of Finnish scientists brought major contributions to the development of ICA. Karhunen et al. [10] studies allowed to interpret the ICA as a nonlinear extension of principal components analysis. This approach played a key role in understanding the ICA as a relevant issue in multivariate data analysis. Hyvärinen contributed to the development of criteria based on maximization of non-gaussianity, which is based on the algorithm FastICA (Fast Independent Component Analysis) [9].

ICA has been used successfully in various areas, for example, electroencephalographic (EEG), seed X-ray images and functional magnetic resonance imaging (fMRI) data. Leite et al. [11] applied independent component analysis (ICA) for automatic processing of radiographic images of 600 sunflower seeds. They used discriminant analysis (DA) to classify seed quality. The classification achieved an overall accuracy of 82%. The results showed that ICA and DA were effective in X-ray analysis to associate seed morphology and seedling performance.

Given a microarray data set $Y = (\mathbf{y_{ij}})_{m\times N} = (y_1, y_2, \ldots, y_m)^T$ (T means transpose) whose m rows are N-dimensional time series, each element y_{ij} in the matrix Y corresponds to the sea level value at time j for the ith series. We consider that the matrix Y is generated by mixing m mutually independent components expressed by

$$\mathbf{Y}_{m\times N} = \mathbf{A}_{m\times m}.\mathbf{S}_{m\times N}, \tag{1}$$

where $\mathbf{A}$ is the matrix of coefficients (a_{ij}) of the linear combination, named mixing matrix, while $\mathbf{S}$ is the matrix of independent component $\mathbf{s_j}$.

Aimed at size reduction, a number of $k < m$ of independent components (IC) can be selected by using principal component analysis (PCA) as pre-processing for ICA, so that

$$\mathbf{Y}_{m\times N} \approx \mathbf{A}_{m\times k}.\mathbf{S}_{k\times N}. \tag{2}$$

Each series y_i is decomposed into a linear combination of ICs (basis) given by

$$y_i = a_{i1}\mathbf{s}_1 + a_{i2}\mathbf{s}_2 + \cdots + a_{ik}\mathbf{s}_k, \tag{3}$$

for every $i = 1, 2, \ldots, m$, so that each series is represented by the coefficients of each independent component of the mixture.

Based on the estimates of rows $\mathbf{a_i}$ of the mixing matrix $\mathbf{A}$, a complete linkage method is performed to identify groups based on the independent components. The use of independent component analysis not only enables the clustering of time series as also provides us with information about the characteristics common to groups from the analysis of the components.