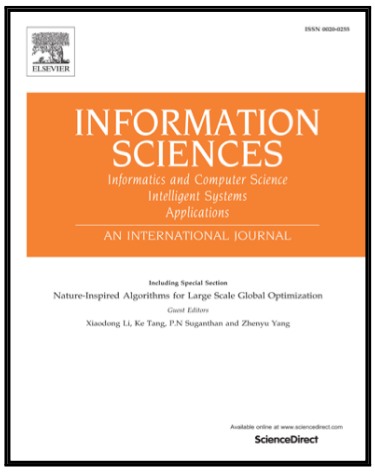

Simultaneous optimisation of clustering quality and approximation error for time series segmentation

Antonio Manuel Durán-Rosal, Pedro Antonio Gutiérrez,
Francisco José Martínez-Estudillo, César Hérvas-Martínez

PII: S0020-0255(16)31361-5
DOI: 10.1016/j.ins.2018.02.041
Reference: INS 13446

To appear in: *Information Sciences*

Received date: 19 October 2016
Revised date: 9 January 2018
Accepted date: 17 February 2018

Please cite this article as: Antonio Manuel Durán-Rosal, Pedro Antonio Gutiérrez, Francisco José Martínez-Estudillo, César Hérvas-Martínez, Simultaneous optimisation of clustering quality and approximation error for time series segmentation, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.02.041

Simultaneous optimisation of clustering quality and approximation error for time series segmentation

Antonio Manuel Durán-Rosal[*,a], Pedro Antonio Gutiérrez[a], Francisco José Martínez-Estudillo[b], César Hérvas-Martínez[a]

[a]*Department of Computer Science and Numerical Analysis, University of Córdoba, Rabanales Campus, Albert Einstein building, 14071, Córdoba, Spain*
[b]*Department of Quantitative Methods, Universidad Loyola Andalucía, Escritor Castilla Aguayo 4, 14004 Córdoba, Spain*

Abstract

Time series segmentation is aimed at representing a time series by using a set of segments. Some researchers perform segmentation by approximating each segment with a simple model (e.g. a linear interpolation), while others focus their efforts on obtaining homogeneous groups of segments, so that common patterns or behaviours can be detected. The main hypothesis of this paper is that both objectives are conflicting, so time series segmentation is proposed to be tackled from a multiobjective perspective, where both objectives are simultaneously considered, and the expert can choose the desired solution from a Pareto Front of different segmentations. A specific multiobjective evolutionary algorithm is designed for the purpose of deciding the cut points of the segments, integrating a clustering algorithm for fitness evaluation. The experimental validation of the methodology includes three synthetic time series and three time series from real-world problems. Nine clustering quality assessment metrics are experimentally compared to decide the most suitable one for the algorithm. The proposed algorithm shows good performance for both clustering quality and reconstruction error, improving the results of other mono-objective alternatives of the state-of-the-art and showing better results than a simple weighted linear combination of both corresponding fitness functions.

Key words: Time series segmentation, multiobjective optimisation, clustering, evolutionary computation

1. Introduction

Time series are an important class of temporal data objects collected chronologically. The corresponding databases are often large, high in dimensionality and require continuous updating. Thus, their intrinsic characteristics make them difficult to analyse. In this context, dimensionality reduction, similarity measurement, segmentation, visualisation and mining methods (such as hidden pattern discovery, clustering, classification or rule discovery) are part of time series research [16, 25, 35].

The segmentation task aims at creating an accurate approximation of the time series, by reducing its dimensionality while retaining the essential features. The objective of this task is to minimise the reconstruction error of a reduced representation with respect to the original time series. Segmentation tasks do not only reduce storage space but also increase the performance of data mining techniques. According to the literature review, current time series compression techniques require expert understanding of the time series, and appropriate threshold values need to be adjusted in order to reduce information loss. The number of segments is the key for time series segmentation, because, in real problems, the size of the segments must be different.

Unequal-length time series segmentation can be explored by using two approaches. The first approach aims to locate specific points of the time series, which are treated as turning points connecting different segments. Pratt and Fink [29] proposed a turning point selection method, which characterises these points as those with a maximal or minimal value with respect to nearby data. Man and Wong [26] suggested another approach, where a function is used

[*]Corresponding author
Email address: i92duroa@uco.es (Antonio Manuel Durán-Rosal)