Applied Soft Computing 23 (2014) 319–332

Contents lists available at ScienceDirect

Applied Soft Computing

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

Hybrid meta-heuristic optimization algorithms for time-domain-constrained data clustering

Mª Luz López García [a], Ricardo García-Ródenas [a,*], Antonia González Gómez [b]

[a] Departamento de Matemática Aplicada, Escuela Superior de Informática, Universidad de Castilla la Mancha, 28012 Ciudad Real, Spain
[b] Departamento de Matemática Aplicada a los Recursos Naturales, E.T. Superior de Ingenieros de Montes, Universidad Politécnica de Madrid, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 5 January 2013
Received in revised form 4 June 2014
Accepted 25 June 2014
Available online 3 July 2014

Keywords:
Time series clustering
Segmentation of multivariate time series
Nelder–Mead simplex search method
Particle swarm optimization
Genetic algorithms
Simulated annealing

ABSTRACT

This paper addresses the question of time-domain-constrained data clustering, a problem which deals with data labelled with the time they are obtained and imposing the condition that clusters need to be contiguous in time (the time-domain constraint). The objective is to obtain a partitioning of a multivariate time series into internally homogeneous segments with respect to a statistical model given in each cluster.

In this paper, time-domain-constrained data clustering is formulated as an unrestricted bi-level optimization problem. The clustering problem is stated at the upper level model and at the lower level the statistical models are adjusted to the set of clusters determined in the upper level. This formulation is sufficiently general to allow these statistical models to be used as black boxes. A hybrid technique based on combining a generic population-based optimization algorithm and Nelder–Mead simplex search is used to solve the bi-level model.

The capability of the proposed approach is illustrated using simulations of synthetic signals and a novel application for survival analysis. This application shows that the proposed methodology is a useful tool to detect changes in the hidden structure of historical data.

Finally, the performance of the hybridizations of particle swarm optimization, genetic algorithms and simulated annealing with Nelder–Mead simplex search are tested on a pattern recognition problem of text identification.

Introduction

The problem of time-series segmentation means partitioning a time series in *K*-time segments that are internally homogeneous [1]. Time-series segmentation has been applied in a wide range of fields such as signal analysis [2,3], industrial process monitoring [4–6], time series DNA micro-array analysis [7], loading identification for stable operation of thermal power units [8], automatic segmentation of traffic patterns [9,10], human motion analysis [11], geophysics environmental research [12], among others. The desired goals depend on the specific application and aim to locate stable periods of time, to identify changing points, or simply to express the original time series in a compact way.

One of the most widely used methods for dealing with time-series segmentation problems is cluster analysis. Clustering methods refer to the process of dividing a set of objects into groups so that members of the same group are similar to each other and different from members of the rest of the groups. The clustering analysis problem can be formulated as one of optimizing a loss (or merit) function subject to a set of constraints. This approach is very versatile and has allowed additional information to be added to the clustering process by adding new constraints or modifying the objective function. This formulation allows the inclusion of information about the shape of the clusters, the distribution of data and the presence of noise and outliers. In this paper we consider constrained time-dependent clustering analysis. This approach enriches the cluster analysis by adding the time at which the data is obtained to the set of constraints and taking into account the fact that neighbouring observations in time may belong to the same cluster.

K-means and Fuzzy *c*-Means (FCM) clustering methods have been adapted to solve multiple clustering analysis problem variants [13,14]. These methods do not incorporate a time structure to

* Corresponding author. Tel.: +34 926295300.
E-mail addresses: Marialuz.lopez@uclm.es (M.L. López García), Ricardo.Garcia@uclm.es (R. García-Ródenas), antonia.gonzalez@upm.es (A.G. Gómez).

http://dx.doi.org/10.1016/j.asoc.2014.06.046
1568-4946/

Table 1
Application of hybridization strategies to cluster analysis.

Algorithm	Purpose/feature	Reference
PSO + Nelder–Mead + *K*-Means(*K*-NM-PSO)	To speed up the convergence	[19]
K-harmonic mean + PSO (PSO*K*HM)	Robustness against initialization	[20]
PSO + Simulated Annealing (PSO-SA)	To find global optima	[21]
K-harmonic mean + PSO (PSO*K*HM)	Robustness against initialization	[20]
K-means + Simulated Annealing + Contraction Factor PSO + Taguchi method (*K*SACPSO)	Comparative study of hybridization strategies	[22]
Fuzzy adaptive PSO + Ant colony + *K*-means (FAPSO-ACO-*K*)	To improve partitioning quality	[23]
Chaotic ant swarm (CAS)	To improve the efficiency of PSO-clustering	[24]
PSO + *K*-means (*K*PSO); PSO with Contraction Factor (CFPSO)	Comparative study of hybridization strategies	[25]
PSO + *K*-harmonic means	Robustness against initialization	[26]
PSO	To deal with known or unknown numbers of clusters	[27]
Hybrid PSO-SA BIClustering algorithm	To extract bi-clusters of gene expression data	[28]
Ant Colony Optimization with Tabu search	Applied to the Information Retrieval (IR) system	[29]
SA + Variable Neighbourhood Search (VNS)	To handle large scale and high dimensional data	[30]
Naked mole-rats (BNMR) algorithm	To speed up the convergence	[31]
PSO with a heuristic search algorithm	To improve the quality of the solution	[32]
Intelligent ant-colony optimization (ACO)	To improve agglomerative hierarchical clustering	[33]
Ant-based clustering	Applied to constrained clustering problems	[34]

obtain the clustering. Steinley and Hubert [15] adapt the *K*-means algorithm to situations where observations are put in order (for instance, by the time instant they are collected). These authors propose a two-stage procedure, in the first stage an initial object order is identified through an auxiliary quadratic assignment optimization heuristic and in the second stage, dynamic programming recursion is applied to optimally subdivide the object set subject to the order constraints. This paper shows that the use of an order-constrained *K*-means strategy improves interpretability and cluster recovery. These solutions are very similar, in terms of sum of square errors (SSE), to the solution obtained without time structure.

Łęski and Owczarek [2] adapt Fuzzy *c*-Regression Models (FCRM) described in Hathaway and Bezdek [16] to this problem and propose Time-Domain-Constrained Fuzzy *c*-Regression Models (TDC-FCRM), and an insensitive version for outliers of this method, the so-called ε-Insensitivity in Time-Domain-Constrained Fuzzy *c*-Regression Models (εTDCFCRM).

In this paper we deal with the general problem of time-domain-constrained clustering analysis in which the observations obtained in each subinterval are generated by an arbitrary statistical model, but not necessarily a linear regression model as the references above consider.

A promising field of research in clustering is the hybridization of evolutionary algorithms. Evidence of this is the literature review by Abul Hasan and Ramakrishnan [17] which contains more than 125 items where these algorithms are applied to cluster analysis. Hybrid algorithms try to make full use of the merits of various optimization techniques to obtain an efficient method for finding global optima over wide areas. So-called Particle Swarm Optimization has been successful in solving clustering problems [18]. The advantages of PSO algorithms can be summarized as follows: they avoid local optima, do a search in the entire solution space, are robust with respect to initialization parameters, viable, efficient with a smaller computational burden and it is simple to choose the correct parameter values. Table 1 presents some hybrid algorithms, paying special attention to methods based on PSO, that are available in the literature. An exhaustive review is beyond the scope of this paper but it can be found in Abul Hasan and Ramakrishnan [17] and Rana et al. [18]. The main motivations of these methods are similar to optimization algorithms and they are essentially three: (i) to speed the process up, (ii) to improve the quality of the solution and (iii) to adapt the algorithm to a specific application.

Fan and Zahara [35] propose the hybridization of the Nelder–Mead (NM) method and PSO for unconstrained optimization. They prove that the hybrid algorithm NM + PSO is a strategy that achieves an excellent trade-off between computational burden and quality of the solution. In Kao et al. [19] NM + PSO algorithm is applied to cluster analysis and *K*-means algorithm is used as an initialization process. This hybrid algorithm is called *K*-NM-PSO. In the numerical experiments reported it is shown that *K*-NM-PSO carries out a smaller number of iterations than PSO, NM-PSO, *K*-PSO and *K*-means for a similar level of error.

In Vakil Baghmisheh et al. [36] a hybrid PSO is proposed, the so-called PS-NM algorithm, to the problem for crack detection in cantilever beams. This hybridization strategy is different from that given in Fan and Zahara [35] and Kao et al. [19]. These authors, at each iteration, initialize NM with the simplex defined by the $N+1$ best particles and update the rest of the population using a modified PSO method. The solution obtained by the NM algorithm replaces the $N+1$th best solution in the next iteration, while PS-NM applies the NM method to the best solution found by the PSO.

In this paper we applied the hybridization framework given in Espinosa-Aranda et al. [37] to the time-domain-constrained data clustering. This framework is general enough to be applicable to the hybridization of algorithms that have been successful in cluster analysis such as PSO, simulated annealing and genetic algorithms. One iteration of this method applies several iterations of population-based algorithm until achieving n_c improvements of the objective function. Then, a fixed number of iterations of the NM algorithm is carried out, starting from the best found solution. The proposed algorithm combines both the global search ability of a population-based algorithm and the local search efficiency of NM. The parameter n_c plays the role of a threshold that guarantees being in a new neighbourhood of solutions that can then be locally explored with NM. These algorithms can be described in the exact optimization framework given in García et al. [38], where the algorithm used in the *column generation phase* (in this case a population-based algorithm) is accelerated by means of the algorithm for solving the so-called *restricted master problem* (in this case NM). This algorithmic class has achieved remarkable results in network flow optimization problems [39] and global optimization [37].

The contributions of this paper can be summarized as follows:

- A general formulation of time-domain-constrained data clustering is described based on unconstrained bi-level optimization. This formulation is enough to consider any statistical model describing data inside each time cluster.
- An assessment of a class of algorithms, based on hybridization of a global optimization algorithm such as PSO, genetic algorithms or simulated annealing and the NM algorithm for the proposed model. The advantage of hybridization is illustrated by a hard problem of pattern recognition. An efficient hybridization