Expert Systems With Applications 55 (2016) 184–193

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multiobjective clustering analysis using particle swarm optimization

Giuliano Armano, Mohammad Reza Farmani*

Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, I-09123, Cagliari, Italy

ARTICLE INFO

Keywords:
Clustering
Multiobjective
Particle swarm optimization

ABSTRACT

Clustering is a significant data mining task which partitions datasets based on similarities among data. This technique plays a very important role in the rapidly growing field known as exploratory data analysis. A key difficulty of effective clustering is to define proper grouping criteria that reflect fundamentally different aspects of a good clustering solution such as compactness and separation of clusters. Moreover, in the conventional clustering algorithms only a single criterion is considered that may not conform to the diverse and complex shapes of the underlying clusters. In this study, partitional clustering is defined as a multiobjective optimization problem. The aim is to obtain well-separated, connected, and compact clusters and for this purpose, two objective functions have been defined based on the concepts of data connectivity and cohesion. These functions are the core of an efficient multiobjective particle swarm optimization algorithm, which has been devised for and applied to automatic grouping of large unlabeled datasets. A comprehensive experimental study is conducted and the obtained results are compared with the results of four other state-of-the-art clustering techniques. It is shown that the proposed algorithm can achieve the optimal number of clusters, is robust and outperforms, in most cases, the other methods on the selected benchmark datasets.

1. Introduction

It is well known that huge amounts of data are currently being stored and collected in databases, and that this quantity continues to grow rapidly. Valuable information, still hidden in data, should be revealed to improve the decision-making process in organizations. Data mining consists of all methodologies that apply data analysis techniques to discover previously-unknown valid patterns and relationships in large datasets. These methods include a number of technical approaches, such as classification, data summarization, dependency network finding, regression, anomaly detection, and clustering (Han & Kamber, 2000). As for clustering, it is the process of partitioning data into groups with the desired properties that data in each group should be similar, while data from different groups should be dissimilar. Different areas, such as data mining, machine learning, biology, and statistics, include the roots of data clustering (Cheng, Yang, & Cao, 2013; Kao, Zahara, & Kao, 2008; Leung, Zhang, & Xu, 2000; Nguyen & Cios, 2008; Qiu, Xu, Gao, Li, & Chi, 2016; Saha, Alok, & Ekbal, 2016; Sahoo, Zuo, & Tiwari, 2012; Thong et al., 2015).

Generally speaking, hierarchical and partitional clustering encompass most of the existing clustering methods. Hierarchical clustering results in a tree in which each internal node embodies other nodes (i.e., its children), until leaves are encountered (Leung et al., 2000). Hierarchical clustering algorithms do not need to know in advance the number of clusters and are independent from the initial conditions. On the other hand, they are typically "greedy", meaning that objects that belong to a cluster cannot be reassigned to other clusters in the clustering process. Moreover, due to lack of information about the global shape or size of the clusters, these algorithms may not be able to separate overlapping clusters (Jain, Murty, & Flynn, 1999). Also partitional clustering typically decomposes a dataset into a set of disjoint clusters. Many partitional clustering algorithms try to minimize some measure of dissimilarity for objects that belong to the same cluster while maximizing the dissimilarity for objects that belong to different clusters. Summarizing, the main drawbacks of hierarchical algorithms usually become advantages for partitional algorithms, and vice versa (Frigui & Krishnapuram, 1999).

Swarm intelligence (SI) is an innovative subcategory of artificial intelligence, inspired by the intelligent behavior of insect or animal groups in nature, including ant colonies, bird flocks, fish schools, bee colonies, and bacterial swarms (Kennedy & Eberhart, 2001). In recent years, SI methods like swarm-based clustering algorithms have been successfully used to deal with clustering problems

* Corresponding author. Tel.: +39 3426325885.
E-mail addresses: armano@diee.unica.it (G. Armano), mohammad.farmani@diee.unica.it, farmani60@gmail.com (M.R. Farmani).

http://dx.doi.org/10.1016/j.eswa.2016.02.009
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

(Abraham, Das, & Roy, 2008; Bharne, Gulhane, & Yewale, 2011; Das, Abraham, & Konar, 2008; Grosan, Abraham, & Chis, 2006; Jiang, Li, Yi, Wang, & Hu, 2011; Omran, Salman, & Engelbrecht, 2006). For this reason, the research community has recently given them special attention, mainly due to the fact that swarm-based approaches are particularly suited to perform exploratory analysis and also because many issues are still open in this field (Abraham et al., 2008).

In this paper, we confine ourselves to the application of particle swarm optimization (PSO) to clustering. Similar to other SI methods, PSO is inspired by a phenomenon that occurs in nature –i.e., the social behavior of bird flocking or fish schooling (Poli, Kennedy, & Blackwell, 2007). Two PSO-based clustering methods are reported in Rana, Jasola, and Kumar (2011): the first method is used to find the centroids for a user-specified number of clusters and the second method is aimed at extending PSO with K-means (used to seed the initial swarm). It is shown that the latter algorithm has better convergence, compared to the classical version of K-means. Yang et al. propose a hybrid clustering algorithm based on PSO and K-harmonic (KHM) means (PSOKHM) (Yang, Sun, & Zhang, 2009). They show that the PSOKHM algorithm increases the convergence speed of PSO, is capable of escaping from local optima, and has better performance than PSO and KHM clustering on seven datasets. A multiobjective PSO and simulated annealing clustering algorithm (MOPSOSA) is proposed in Abubaker, Baharum, and Alrefaei (2015). This method simultaneously optimizes three different objective functions, which are used as cluster validity indexes for finding the proper number of clusters (and the clusters) according to the given dataset. Euclidean distance, point symmetry and short distances are considered validity indexes in MOPSOSA. The method obtains more promising results in comparison with other conventional clustering algorithms. Several other PSO-based clustering algorithms have been proposed in the literature (for a comprehensive review about PSO-based clustering the interested reader may consult (Cura, 2012; Izakian & Abraham, 2011; Kalyani & Swarup, 2011; Sarkar, Roy, & Purkayastha, 2013; Tsai & Kao, 2011)). However, they mostly consider a single function as the objective of the clustering problem and, to the best of our knowledge, all recent works on multiobjective clustering do not apply the concept of Pareto optimal solutions (Kasprzak & Lewis, 2001).

In this paper, a multiobjective clustering particle swarm optimization (MCPSO, hereinafter) framework is proposed, which obtains well-separated, connected, and compact clusters, regardless from the expected optimal number of clusters and their characteristics. MCPSO is also able to automatically determine the optimal number of clusters. To achieve these goals, two conflicting objective functions are defined, based on the concepts of *connectivity* and *cohesion*, and MCPSO uses them to find a set of non-dominated clustering solutions, called Pareto front. A simple decision maker is then used to select the best solution among Pareto solutions. A comparison of the MCPSO performance against those obtained using four state-of-the-art clustering algorithms has also been made. As selected datasets are in fact labeled, we have been able to measure the average “accuracy” on clusters, assuming that each cluster actually accounts for a unique label. The accuracy measured on the results of clustering, together with the required computational time, are used as performance metrics in the comparative analysis.

The rest of this paper is organized as follows. In Section 2, swarm intelligence and multiobjective optimization are defined. The proposed MCPSO algorithm and the clustering objective functions are described in detail in Section 3. A comprehensive set of experimental results are provided in Section 4. Section 5 reports conclusions.

2. Multiobjective optimization and swarm intelligence

In the area of metaheuristics, swarm intelligence (SI) belongs to the group of approaches that apply the self-organized and decentralized characteristics of natural or artificial phenomena to deal with complex optimization problems. In particular, the behavior of natural individuals who relate to each other and to their environment plays a significant role in designing SI algorithms. Many of these algorithms have been introduced in recent years and have been successfully applied to different kinds of problems and framed in several application fields (Kennedy & Eberhart, 2001). Although these algorithms have been mainly used with single objective optimization models, in our view their robust and population-based nature make them good candidates for multiobjective optimization problems.

In general, a multiobjective optimization problem requires the simultaneous satisfaction of different and often conflicting objectives. These objectives are characterized by functions that may be dependent or not. A multiobjective optimization problem is characterized by the need of finding a vector of n decision variables $V = [v_1, v_2, \ldots, v_n]$ which concurrently: a) satisfies m equality $h_i(V) = 0, i = 1, \ldots, m,$ b) satisfies p inequality $g_j(V) \le 0, j = 1, \ldots, p$ constraints, and c) optimizes (i.e. minimizes or maximizes) a vector of k objective functions $F(V) = [f_1(V), f_2(V), \ldots, f_k(V)]^T$. It is worth mentioning that, in general, each objective function achieves its optimum at a different point in the space of decision variables and that no combination of them exists able to simultaneously optimize all the components of the objective vector (Marler & Arora, 2004). Pareto optimality is one of the concepts that has been used to address this type of problems (Kasprzak & Lewis, 2001). Considering a minimization problem, a decision vector $V^* \in V$ is called Pareto optimal (non-dominated) solution if and only if no $V' \in V$ exists such that $f_i(V') < f_i(V^*)$ for at least one $i = 1, \ldots, k$ and $f_i(V') \le f_i(V^*)$ for the remaining cases.

Multiobjective particle swarm optimization (MOPSO) (Coello & Lechuga, 2002), Multiobjective Ant Colony Optimization (MOACO) (Angus & Woodward, 2009), Multiobjective Artificial Bee Colony (MOABC) (Omkar, Senthilnath, Khandelwal, Naik, & Gopalakrishnan, 2011), Multiobjective Differential Evolution (MODE) (Robič & Filipič, 2005), and Multiobjective Artificial Immune Systems (MOAIS) (Coello & Cortés, 2005) are some of the main multiobjective SI methods that have been proposed for solving various theoretical and practical problems.

3. Multiobjective clustering with particle swarm optimization

In this section, we describe the MCPSO method. As already pointed out, it is based on the particle swarm optimization algorithm (Kennedy & Eberhart, 2001), in a multiobjective setting. MCPSO consists of two main phases: optimization and decision making. Two conflicting objective functions are defined, based on *connectivity* and *cohesion* with the aim of obtaining well-separated, compact, and connected clusters. The optimization phase results in a set of optimal solutions for the given clustering problem, called Pareto solutions (Kasprzak & Lewis, 2001). These solutions represent trade-offs among conflicting objectives. In particular, each Pareto solution is a partition with a different number of embedded clusters. This collection of solutions is used by MCPSO to automatically determine the optimal number of clusters. As any of the Pareto solutions can be considered optimal, a simple decision maker is used to select the best solution among Pareto solutions, based on a trade-off between two objectives.