Applied Soft Computing 28 (2015) 301–311

Contents lists available at ScienceDirect

Applied Soft Computing

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

Ant Colony Optimization based clustering methodology

Tülin İnkaya[a,*], Sinan Kayalıgil[b], Nur Evin Özdemirel[b]

[a] Uludağ University, Industrial Engineering Department, Görükle, 16059 Bursa, Turkey
[b] Middle East Technical University, Industrial Engineering Department, Çankaya, 06800 Ankara, Turkey

ARTICLE INFO

Article history:
Received 28 February 2014
Received in revised form 31 August 2014
Accepted 17 November 2014
Available online 13 December 2014

Keywords:
Clustering
Ant Colony Optimization
Multiple objectives
Data set reduction

ABSTRACT

In this work we consider spatial clustering problem with no a priori information. The number of clusters is unknown, and clusters may have arbitrary shapes and density differences. The proposed clustering methodology addresses several challenges of the clustering problem including solution evaluation, neighborhood construction, and data set reduction. In this context, we first introduce two objective functions, namely adjusted compactness and relative separation. Each objective function evaluates the clustering solution with respect to the local characteristics of the neighborhoods. This allows us to measure the quality of a wide range of clustering solutions without a priori information. Next, using the two objective functions we present a novel clustering methodology based on Ant Colony Optimization (ACO-C). ACO-C works in a multi-objective setting and yields a set of non-dominated solutions. ACO-C has two pre-processing steps: neighborhood construction and data set reduction. The former extracts the local characteristics of data points, whereas the latter is used for scalability. We compare the proposed methodology with other clustering approaches. The experimental results indicate that ACO-C outperforms the competing approaches. The multi-objective evaluation mechanism relative to the neighborhoods enhances the extraction of the arbitrary-shaped clusters having density variations.

1. Introduction

Cluster analysis is the organization of a collection of data points into clusters based on similarity [1]. Clustering is usually considered as an unsupervised classification task. That is, the characteristics of the clusters and the number of clusters are not known a priori, and they are extracted during the clustering process. In this work we focus on spatial data sets in which a priori information about the data set (the number of clusters, shapes and densities of the clusters) is not available. Finding such clusters has applications in geographical information systems [2], computer graphics [3], and image segmentation [4]. In addition, clusters of spatial defect shapes provide valuable information about the potential problems in manufacturing processes of semiconductors [5,6].

We consider spatial clustering as an optimization problem. Our aim is to obtain compact, connected and well-separated clusters. To the best of our knowledge, there is not a single objective function that works well for any kind of geometrical clustering structure. Therefore, we first introduce two solution evaluation mechanisms for measuring the quality of a clustering solution. The main idea behind both mechanisms is similar, and each mechanism is based on two objectives: adjusted compactness and relative separation. The first objective measures the compactness and connectivity of a clustering solution, and the second objective is a measure for separation. The difference between the two mechanisms is the degree of locality addressed in the calculations. The main advantage of these objectives is that the length of an edge is evaluated relatively, that is, it is scaled relative to the lengths of other edges within its neighborhood. This scaling permits us to evaluate the quality of the clustering solution independent of the shape and density of the clusters.

We implement the proposed solution evaluation mechanisms in a clustering framework based on Ant Colony Optimization (ACO). In order to find the target clusters, we use two complementary objective functions (adjusted compactness and relative separation) in a multiple-objective context. Hence, the output of ACO-C is a set of non-dominated solutions. Different from the literature, we are not interested in finding all non-dominated solutions or the entire Pareto efficient frontier. ACO-C has two pre-processing steps: neighborhood construction and data set reduction. Neighborhood construction extracts the local connectivity, proximity and density information inherent in the data set. Data set reduction helps reduce the storage requirements and processing time for the clustering task. Our experimental results indicate that ACO-C finds

[*] Corresponding author. Tel.: +90 224 2942605; fax: +90 224 2941903.
E-mail addresses: tinkaya@uludag.edu.tr (T. İnkaya), skayali@metu.edu.tr (S. Kayalıgil), nurevin@metu.edu.tr (N.E. Özdemirel).

http://dx.doi.org/10.1016/j.asoc.2014.11.060

the arbitrary-shaped clusters with varying densities effectively, where the number of clusters is unknown.

Our contributions to the literature are as follows:

1. The proposed solution evaluation mechanisms allow us to quantify the quality of a clustering solution having arbitrary-shaped clusters with different densities in an optimization context. The use of these evaluation mechanisms is not restricted to ACO. They can be used in other metaheuristics and optimization-based clustering approaches.
2. The proposed ACO-based methodology introduces a general, unified framework for the spatial clustering problem without a priori information. It includes the solution evaluation mechanism, extraction of local properties, data set reduction, and the clustering task itself.
3. ACO-C is a novel methodology for the clustering problem in which there is no a priori information, that is,
   - the number of clusters is unknown,
   - clusters may have arbitrary shapes,
   - there may be density variations within the clusters, and
   - different clusters may have density differences.

We provide the related literature in Section 2. Section 3 introduces the solution evaluation mechanisms. The details of ACO-C are explained in Section 4. Section 5 is about the empirical performance of ACO-C. First, we set the algorithm parameters using a full factorial design. Then, we compare ACO-C with some well-known algorithms. Finally, we conclude in Section 6.

2. Related literature

The clustering algorithms can be classified into partitional, hierarchical, density-based algorithms, and metaheuristics (simulated annealing, tabu search, evolutionary algorithms, particle swarm optimization, ACO, and so on). [1,7,8] provide comprehensive reviews of clustering approaches.

In this section, we present the related literature on the solution evaluation mechanisms and ant-based clustering algorithms.

2.1. Solution evaluation mechanisms

A good clustering solution has compact and connected clusters that are well-separated from each other. However, quantifying and measuring the clustering objectives (compactness, connectivity and separation) for a data set is not a trivial task. We review the solution evaluation mechanisms in the literature under four categories: partitional approaches, graph-based approaches, clustering validity indices, and multi-objective approaches.

Partitional approaches consider objective functions such as minimization of total variance/distance between all pairs of data points, or minimization of total variance/distance between data points and a cluster representative such as *k*-means [9,10] or *k*-medoids [11]. In these approaches, the number of clusters needs to be given as input, and the resulting clusters have spherical or ellipsoid shapes in general.

In order to handle the data sets with arbitrary-shaped clusters and density variations, graph-based approaches are proposed. Objective functions used are minimization of the maximum edge length in a cluster, maximization of the minimum/maximum/average distance between two clusters, and so on [12,13]. A typical complication for such objective functions is illustrated in Fig. 1(a). In Fig. 1(a) the maximum edge length within the spiral clusters is larger than the distance between these two clusters. In this case elimination of the longest edge causes division of the spiral clusters.

Another research stream in solution evaluation makes use of cluster validity indices. Validity indices are used to quantify the quality of a clustering solution and to determine the number of clusters in a data set [14,15]. In an effort to find the target clusters, some researchers use validity indices as objective functions in genetic algorithms [16–21]. However, most of the validity indices assume a certain geometrical structure in the cluster shapes. When a data set includes several different cluster structures, such as arbitrary shapes and density differences, these indices may fail. An example is provided in Fig. 1(b). The clustering solutions generated by DBSCAN [22] are evaluated using Dunn index [23] with different *MinPts* settings (within a range of 1–15). The number of clusters found with each setting is shown, e.g. 30 clusters are found when *MinPts* is set to two. Dunn index measures the minimum separation to maximum compactness ratio, so a higher Dunn index implies better clustering. Although the highest Dunn index (0.31) is achieved for the solutions with two and four clusters, the target solution has three clusters with a Dunn index of 0.09. Hence, Dunn index is not a proper objective function for such a data set.

Maulik and Bandyopadhyay [24] evaluates the performance of three clustering algorithms, namely *k*-means, single-linkage, and simulated annealing (SA) by using four cluster validity indices, namely Davies-Bouldin index, Dunn index, Calinski-Harabasz index, and index I. Compared to other validity indices, index I is found to be more consistent and reliable in finding the correct number of clusters. However, the four cluster validity indices are limited to extracting spherical clusters only. To handle different geometrical shapes, Bandyopadhyay et al. [25] uses a point symmetry-based distance measure in a genetic algorithm. The algorithm has difficulty in handling asymmetric clusters and density differences within a cluster.

Since a single objective is often unsuitable to extract target clusters, multi-objective (MO) approaches are considered to optimize several objectives simultaneously. To the best of our knowledge, VIENNA [26] is the first multi-objective clustering algorithm, which is based on PESA [27]. It optimizes two objective functions, total intra-cluster variance and connectedness. However, the algorithm requires the target number of clusters. One of the well-known MO clustering algorithms is the multi-objective clustering with automatic *k*-determination (MOCK) [28]. MOCK is based on evolutionary algorithms, and uses compactness and connectedness as two complementary objective functions. It can detect the number of clusters in the data set. The output of the algorithm is a set of non-dominated clustering solutions. However, it is capable of finding well-separated clusters having hyperspherical shapes. Improvements in this algorithm and its applications have been investigated [29,30]. Saha and Bandyopadhyay [31] also considers the clustering problem in a multi-objective framework. They optimize Xie-Beni (XB) index [32] and *Sym*-index [21] simultaneously, and introduce a multi-objective SA algorithm. This work is also limited to finding symmetric clusters. Saha and Bandyopadhyay [33] proposes several connectivity-based validity indices based on the relative neighborhood graph. In addition to *Sym*-index and index I, [34] uses one of the connectivity-based validity indices in [33] as the third objective. Adding this connectivity measure helps extraction of arbitrary shapes and asymmetric clusters.

There are additional solution approaches proposed for MO clustering such as differential evolution [35,36], immune-inspired method [37], and particle swarm optimization [38]. In these studies clustering objectives are either cluster validity indices such as XB index, *Sym*-index and FCM index, or compactness-connectivity objectives as in [28].