Applied Soft Computing 41 (2016) 192–213

Contents lists available at ScienceDirect

Applied Soft Computing

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

Automatic clustering using nature-inspired metaheuristics: A survey

Adán José-García, Wilfrido Gómez-Flores *

Technology Information Laboratory, Center for Research and Advanced Studies of the National Polytechnic Institute, Ciudad Victoria, Tamaulipas, Mexico

ARTICLE INFO

Article history:
Received 18 December 2014
Received in revised form 8 October 2015
Accepted 2 December 2015
Available online 31 December 2015

Keywords:
Cluster analysis
Automatic clustering
Nature-inspired metaheuristics
Single-objective and multiobjective metaheuristics

ABSTRACT

In cluster analysis, a fundamental problem is to determine the best estimate of the number of clusters; this is known as the automatic clustering problem. Because of lack of prior domain knowledge, it is difficult to choose an appropriate number of clusters, especially when the data have many dimensions, when clusters differ widely in shape, size, and density, and when overlapping exists among groups. In the late 1990s, the automatic clustering problem gave rise to a new era in cluster analysis with the application of nature-inspired metaheuristics. Since then, researchers have developed several new algorithms in this field. This paper presents an up-to-date review of all major nature-inspired metaheuristic algorithms used thus far for automatic clustering. Also, the main components involved during the formulation of metaheuristics for automatic clustering are presented, such as encoding schemes, validity indices, and proximity measures. A total of 65 automatic clustering approaches are reviewed, which are based on single-solution, single-objective, and multiobjective metaheuristics, whose usage percentages are 3%, 69%, and 28%, respectively. Single-objective clustering algorithms are adequate to efficiently group linearly separable clusters. However, a strong tendency in using multiobjective algorithms is found nowadays to address non-linearly separable problems. Finally, a discussion and some emerging research directions are presented.

1. Introduction

Cluster analysis is an unsupervised learning technique aimed at discovering the natural grouping of objects according to the similarity of measured intrinsic characteristics [1]. The two fundamental problems in automatic clustering are determining the optimal number of clusters and identifying all data groups correctly. In this sense, the number of combinations in assigning N objects into K clusters is[1]:

$$S(N,K)=\frac{1}{K!}\sum_{i=0}^{K}(-1)^{K-i}\binom{K}{i}i^{N}. \tag{1}$$

On the other hand, the search space size in finding the optimal number of clusters is[2]:

$$B(N)=\sum_{K=1}^{N}S(N,K). \tag{2}$$

Besides, the clustering (or grouping) problem of finding an optimal solution is NP-hard when $K>3$ [2]; hence, even for moderate-sized problems, the clustering task could be computationally prohibitive [3].

To limit the search space size, many clustering methods described in the literature assume a fixed number of clusters, which is unknown *a priori* in many clustering practices. To overcome this inconvenience, automatic clustering approaches aimed at finding the adequate number of clusters within the range $[K_{\min}, K_{\max}]$ have been developed.

The principal clustering techniques developed in the last 50 years were reviewed by Jain [4], who presented the evolution and trends in data clustering. Also, Xu and Wunsch [5] focused on algorithms for grouping data sets that are used in statistics, computer science, and machine learning. In the last decade, developments in automatic clustering have been strengthened [6–8]. In particular, nature-inspired metaheuristics have been applied to obtain satisfactory suboptimal solutions to the automatic clustering problem in an acceptable timescale [9]. These kinds of metaheuristics model the behavior of natural phenomena, which exhibit an ability to learn or adapt to new situations to solve problems in complex and changing environments [10].

Some review articles on clustering analysis that use nature-inspired metaheuristics have been published by Handl and Meyer [11], Sheikh et al. [12], Hruschka et al. [9], Rana et al. [13], Bong and Rajeswari [6], Nanda and Panda [14], and Alam et al. [15]. A review of ant-based and swarm-based clustering techniques was presented by Handl and Meyer [11]. A survey on genetic algorithms applied to clustering was summarized by Sheikh et al. [12]. Hruschka et al. [9] presented a brief summary of evolutionary algorithms and reviewed the initialization procedure, encoding scheme, crossover, mutation, and fitness evaluation for single and multiobjective cases. Bong and Rajeswari [6] investigated multiobjective nature-inspired clustering techniques applied to image segmentation. Recently, Nanda and Panda [14] surveyed some nature-inspired metaheuristics focused on the partitional clustering paradigm. Reviews on particle swarm optimization algorithms and their applications to data clustering were presented by Rana et al. [13] and Alam et al. [15].

* Corresponding author at: Technology Information Laboratory, CINVESTAV-IPN, Science and Technology Park TECNOTAM, Km. 5.5, Soto La Marina Road, Ciudad Victoria, Mexico. Tel.: +52 834 107 0220.
E-mail address: wgomez@tamps.cinvestav.mx (W. Gómez-Flores).

[1] $S(N,K)$ is known as the Stirling numbers of the second kind.
[2] $B(N)$ is known as the Bell numbers.

http://dx.doi.org/10.1016/j.asoc.2015.12.001
1568-4946/

Despite the relevance of these review articles, to the best of our knowledge, no review paper about nature-inspired metaheuristics for automatic clustering has been published. Therefore, we present an in-depth review of nature-inspired metaheuristics for automatic clustering that have been reported in the last two decades. This paper contributes in the following two main aspects: (i) it presents an up-to-date overview on single-solution, single-objective, and multiobjective metaheuristics applied to automatic clustering, and (ii) it provides a review of important aspects, such as encoding schemes, validity indices, data sets, and applications.

The outline of this paper is as follows. Section 2 describes the basic terms and concepts related to automatic clustering analysis. Section 3 presents single-solution metaheuristics that use a single agent or solution, which moves through the search space in a piecewise style. Section 4 reviews single-objective metaheuristics, in which a population of potential solutions cooperate to optimize a unique cost function. Section 5 presents multiobjective metaheuristics, which optimize distinct cost functions and consider a trade-off among them. Section 6 discusses relevant automatic clustering algorithms useful for solving specific data sets and applications. Finally, future tendencies and conclusions are given in Sections 7 and 8, respectively.

2. Basic preliminaries

2.1. Definitions

The following terms and notation are used throughout this paper:

- A *pattern* (or object) is a single data item represented by a vector of measurements $\mathbf{x} = \{x_1, x_2, \ldots, x_D\}^T$, where $x_i \in \mathbb{R}$ is a *feature* (or attribute), and D denotes the *dimensionality*.
- A *data set* is denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^D$, where N is the total number of patterns in the data set.
- A *cluster* (or group) can be defined as high-density regions separated by low-density regions within the feature space.
- *Clustering*, denoted as $\mathbf{C} = \{\mathbf{c}_k | k = 1, \ldots, K\}$, refers to the set of mutually disjoint clusters that partitions $\mathbf{X}$ into K groups.
- The *number of objects* in cluster $\mathbf{c}_k$ is denoted by $n_k = |\mathbf{c}_k|$.
- The *centroid of cluster* (or prototype) $\mathbf{c}_k$ is expressed as $\bar{\mathbf{c}}_k = 1/n_k \sum_{\mathbf{x}_i \in \mathbf{c}_k} \mathbf{x}_i$, whereas the *centroid of data set* $\mathbf{X}$ is $\bar{\mathbf{X}} = 1/N \sum_{\mathbf{x}_i \in \mathbf{X}} \mathbf{x}_i$.
- A *distance measure* is a metric (or quasi-metric) used to quantify the proximity between patterns.
- A *cluster validity index* uses a distance measure to quantitatively evaluate the obtained clustering.

2.2. Clustering techniques

The specialized literature on cluster analysis commonly classifies clustering techniques into partitional and hierarchical [1,4,14], which are detailed in the following subsections.

2.2.1. Partitional clustering

Partitional clustering can be performed in two different modes: hard (or crisp) and fuzzy. Hard clustering assumes that the membership between patterns and clusters is binary; thus, each pattern belongs to exactly one cluster. On the other hand, fuzzy clustering assigns different degrees of membership to the patterns for each cluster to build a non-binary relationship between them.

Hard clustering divides a data set directly into a prespecified number of clusters without a hierarchical structure [16], where a data set $\mathbf{X}$ is partitioned into K nonoverlapping groups $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$, such that the following three conditions should be satisfied:

- $\mathbf{c}_i \neq \emptyset$, $i = 1, \ldots, K$;
- $\bigcup_{i=1}^{K} \mathbf{c}_i = \mathbf{X}$;
- $\mathbf{c}_i \cap \mathbf{c}_j = \emptyset$, $i, j = 1 \ldots, K$ and $i \neq j$.

Perhaps the most fundamental algorithm related to hard clustering is the k-means algorithm, which attempts to minimize the sum-of-squared-error criterion [17,18]. Fifty years after its formulation, k-means is still popular and widely used because of its simplicity and low computational complexity [4]. However, a predefined number of clusters is required at the beginning of the algorithm, which is unknown in several real-world clustering applications. Hence, the k-means algorithm has been extended to automatically find the number of clusters; some of these extended approaches include the X-means [19] and the G-means algorithm [20].

Fuzzy clustering is an alternative definition given in terms of fuzzy sets, in which each pattern belongs to more than one cluster simultaneously, with a certain degree of membership $u_j \in [0, 1]$. The membership value of the ith pattern in the jth cluster should satisfy the following two conditions:

- $\sum_{j=1}^{K} u_j(\mathbf{x}_i) = 1$, $i = 1, \ldots, N$;
- $\sum_{i=1}^{N} u_j(\mathbf{x}_i) < N$, $j = 1, \ldots, K$.

The most well-known fuzzy algorithm is fuzzy c-means [21], which is essentially a fuzzy extension of the k-means method.

2.2.2. Hierarchical clustering

Hierarchical clustering algorithms produce a hierarchy of clustering called a dendrogram (or tree structure), which represents the nested grouping of the objects in a data set. The procedure builds N successive clustering levels, in which the current clustering is based on the solution obtained at the previous level. Therefore, hierarchical clustering does not require *a priori* knowledge about the number of clusters; however, the obtained groups are static because the objects assigned to a given cluster cannot move to another one.

Agglomerative and divisive approaches are the two main categories of hierarchical clustering, of which single-link and complete-link [4] algorithms are the most well-known.

2.3. Proximity measures

Clustering algorithms measure the proximity between objects to form groups [1]. The selection of the appropriate proximity measure is important because memberships are defined for every object in data set $\mathbf{X}$. Depending on the kind of proximity measure, different groupings can be created [22]. A proximity measure can be either a distance (dissimilarity) or a similarity between a pair of objects, between an object and a prototype, or between a pair of prototypes. The most common proximity measures used in the automatic clustering techniques described herein are detailed below.

- The Minkowski metric [16], or L_p-norm, is a dissimilarity measure defined as

$$\mathrm{d}_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{D} |x_i - y_i|^p \right)^{1/p}, \tag{3}$$

  where $\mathbf{x}$ and $\mathbf{y}$ are D-dimensional data vectors. Note that when $p = 2$, the Minkowski metric becomes the well-known Euclidean distance (or L_2-norm), which is denoted as $\mathrm{d}_e(\mathbf{x}, \mathbf{y})$. Two other common special cases of the Minkowski metric are the Manhattan distance (or L_1-norm), when $p = 1$, and the Chebyshev distance (or L_∞-norm), when $p \to \infty$, which is computed as

$$\mathrm{d}_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq D} |x_i - y_i|. \tag{4}$$

- The similarity between two vectors $\mathbf{x}$ and $\mathbf{y}$ can be measured by the cosine of the angle between them:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}, \tag{5}$$