Mathematical Biosciences 245 (2013) 76–85

Contents lists available at ScienceDirect

Mathematical Biosciences

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

A biased random-key genetic algorithm for data clustering

P. Festa *

Department of Mathematics and Applications, University of Napoli FEDERICO II, Napoli, Italy

ARTICLE INFO

Article history:
Available online 26 July 2013

Keywords:
Computational biology
Molecular structure prediction
Clustering
Combinatorial optimization

ABSTRACT

Cluster analysis aims at finding subsets (*clusters*) of a given set of *entities*, which are homogeneous and/or well separated.

Starting from the 1990s, cluster analysis has been applied to several domains with numerous applications. It has emerged as one of the most exciting interdisciplinary fields, having benefited from concepts and theoretical results obtained by different scientific research communities, including genetics, biology, biochemistry, mathematics, and computer science.

The last decade has brought several new algorithms, which are able to solve larger sized and real-world instances. We will give an overview of the main types of clustering and criteria for homogeneity or separation. Solution techniques are discussed, with special emphasis on the combinatorial optimization perspective, with the goal of providing conceptual insights and literature references to the broad community of clustering practitioners.

A new biased random-key genetic algorithm is also described and compared with several efficient hybrid GRASP algorithms recently proposed to cluster biological data.

1. Introduction

Cluster analysis aims to group data such that the most similar *objects* belong to the same group or cluster, and dissimilar objects are assigned to different clusters. The objects are also called *entities* or *patterns* and are usually represented as a vector of measurements or a point in a multidimensional space.

Cluster analysis has been applied to several domains, such as natural language processing [57] (where large vocabularies of words of a given natural language must be clustered w.r.t. corpora of very high size), galaxy formation [61] (a study has been conducted on the formation of galaxies by gas condensation with massive dark halos), image segmentation [60] (where the segmentation is achieved by searching for closed contours of the elements in the image), and biological data [5,35,41,43,20].

Starting from one of the pioneering paper of Rao, which appeared in 1971 [45], more recent surveys on clustering algorithms and their applications can be found in [27,28].

In cluster analysis, the criterion for a clustering to be optimal strongly depends upon the specific application in which it is to be used. In the general case, the cluster task can be mathematically formulated as a constrained fractional non-linear 0–1 programming problem and there are no computationally efficient procedures for solving such a problem. In some variants and special cases the problem becomes computationally tractable, as deeply discussed in [45].

The scope of this paper is to provide an overview of the main types of clustering and criteria for homogeneity or separation. Special emphasis is given to the most efficient metaheuristic techniques that can be applied to cluster data and a new biased random-key genetic algorithm is also described and compared with several efficient hybrid GRASP algorithms recently proposed to cluster biological data.

The remainder of this paper is organized as follows. In Section 2, the cluster analysis task is formulated and the most used distance measures between the various entities are described. In Section 3, properties and state-of-the-art solution approaches are discussed. A new biased random-key genetic algorithm is proposed in Section 4 and computational results are presented in Section 5 demonstrating empirically that the new described algorithm results in better quality solutions. Concluding remarks are given in the last section.

2. Problem formulation and distance measures definition

Cluster analysis involves the problem of finding a partition of a given set of entities into a pre-assigned number of mutually exclusive clusters.

Formally, we are given.

- ◇ a set of N objects (entities, patterns) $\mathcal{O} = \{o_1, \ldots, o_N\}$;
- ◇ a set of M of pre-assigned clusters $\mathcal{S} = \{S_1, \ldots, S_M\}$;

* Tel.: +39 081675605; fax: +39 081675605.
E-mail address: paola.festa@unina.it

0025-5564/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2013.07.011

♢ a function $d : \mathcal{O} \times \mathcal{O} \mapsto \mathbb{R}$ that assigns to each pair $o_i, o_j \in \mathcal{O}$ a "metric distance" or "similarity" $d_{ij} \in \mathbb{R}$ (usually, $d_{ij} \geqslant 0$, $d_{ii} = 0, d_{ij} = d_{ji}$, for $i, j = 1, \ldots, N$)

and the task consists in assigning the objects in $\mathcal{O}$ to some cluster in $\mathcal{S}$. The assignment is done while optimizing some distance criteria in such a way that the greater is the similarity (or proximity, homogeneity) within a cluster and the greater is the difference between clusters, the better or more distinct is the clustering [25,27].

Pattern proximity (similarity) is usually measured by a distance function defined on pairs of patterns.

A data object $o_i, i = 1, \ldots, N$, can be formalized as the following numerical vector

$$\overrightarrow{A}_i = \{a_{ij} \mid 1 \leqslant j \leqslant L\},$$

where

- a_{ij} is the value of the jth feature for the ith data object and
- L is the number of features.

Then, the proximity d_{ij} between two objects o_i and o_j is measured by a proximity function d of corresponding vectors $\overrightarrow{A}_i$ and $\overrightarrow{A}_j$.

Several different scientific communities have used and discussed a variety of distance measures (see, for example [6,25,28]). Some of them are listed in the following.

Euclidean distance The Euclidean distance is maybe the most popular metric for measuring the distance between two data objects.

Given two objects o_i and $o_j \in \mathcal{O}$, their Euclidean distance in L-dimensional space is defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{L} (a_{ik} - a_{jk})^2} = \|\mathbf{a}_i - \mathbf{a}_j\|_2. \tag{1}$$

It has an intuitive meaning and it is usually used to evaluate the proximity of objects in two or three-dimensional space. In general, it works well when the data set has " compact" or " isolated" clusters [36].

Pearson's correlation coefficient An alternate measure is the Pearson's correlation coefficient, which measures the similarity between the *shapes of two patterns* (*profiles*).

Given two objects o_i and $o_j \in \mathcal{O}$, their Pearson's correlation coefficient is defined as

$$d_{ij} = \frac{\sum_{k=1}^{L} [(a_{ik} - \mu_{o_i}) \cdot ((a_{jk} - \mu_{o_j}))]}{\sqrt{\sum_{k=1}^{L} (a_{ik} - \mu_{o_i})^2} \cdot \sqrt{\sum_{k=1}^{L} (a_{jk} - \mu_{o_j})^2}}, \tag{2}$$

where μ_{o_i} and μ_{o_j} are the mean value for $\overrightarrow{A}_i$ and $\overrightarrow{A}_j$, respectively.

This correlation coefficient views each object as a random variable with L observations and measures the similarity between two objects by calculating the linear relationship between the distributions of the two corresponding random variables. One drawback of the Pearson's correlation coefficient is that it assumes an approximate Gaussian distribution of the patterns and may not be robust for non-Gaussian distributions, as experimentally shown by Bickel [10].

City-block or Manhattan. City-block or Manhattan distance simulates the distance between points in a city road grid. It measures the absolute differences between two object attributes.

Given two objects o_i and $o_j \in \mathcal{O}$, their City-block or Manhattan distance is defined as

$$d_{ij} = \sum_{k=1}^{L} \mid a_{ik} - a_{jk} \mid . \tag{3}$$

Cosine or uncentered correlation. Cosine or uncentered correlation is a geometric correlation defined by the angle between two objects.

Given two objects o_i and $o_j \in \mathcal{O}$, their cosine or uncentered correlation is defined as

$$D_{ij} = \frac{\sum_{k=1}^{L} a_{ik} \cdot a_{jk}}{\sum_{k=1}^{L} a_{ik}^2 \sum_{k=1}^{L} a_{jk}^2}. \tag{4}$$

Note that,

- the larger is the value of D_{ij}, the lower is the angle between the objects;
- $D_{ij} \in [-1, 1] : D_{ij} = -1$ implies that the angle between vectors representing o_i and o_j is a right angle; while $D_{ij} = 1$ implies that the angle between o_i and o_j is 0;
- $d_{ij} = 1 - \mid D_{ij} \mid$.

3. A review of the most popular clustering techniques

According to Jain et al. [27] (see the taxonometric representation of clustering methods in Fig. 1), state-of-the-art clustering algorithms can be mainly divided into two families: partitioning and hierarchical algorithms.

A partitioning method partitions the set of data objects into non-overlapping clusters such that each data object belongs to exactly one cluster. Instead, in a hierarchical approach a cluster is permitted to have subclusters and the result of the clustering task is a set of nested clusters that can be organized in a tree. Each node of the tree corresponds to a cluster and it is the union of its children (subclusters). Clearly, the leaves have no subclusters and the root node represents the cluster containing all the objects.

3.1. Hierarchical clustering algorithms

The most popular hierarchical clustering algorithms are the single-link algorithm [54], complete-link [32], and minimum-variance algorithms [30]. The single-link and the complete-link approaches differ in how they define the similarity between a pair of clusters: in the single-link approach, this distance is the minimum of the distances between all pairs of patterns drawn from the two clusters (one pattern from the first cluster, the other from the second); in a complete-link algorithm, this distance is the maximum of all pairwise distances between patterns in the two clusters. In either case, two clusters are merged to form a larger cluster based on minimum distance criteria.

3.2. Partitioning clustering algorithms

The most popular partitioning clustering algorithms are the squared error algorithms (among them the most famous is the k-means method [38]), graph-theoretic algorithms [62], and mixture-resolving and mode-seeking algorithms [25].

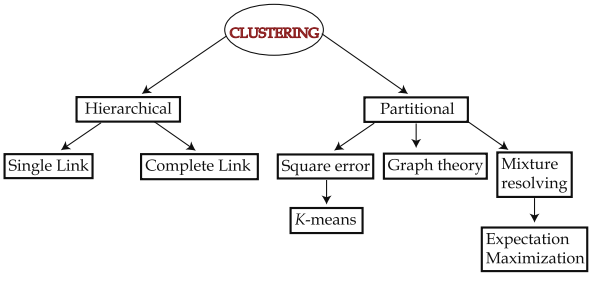

Fig. 1. A taxonomy of clustering methods.