G Model
ASOC 3026 1–11
ARTICLE IN PRESS

Applied Soft Computing xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

Hybrid differential evolution algorithm for optimal clustering

Q1 Josef Tvrdík [a],*, Ivan Křivý [b]

[a] Centre of Excellence IT4Innovations, Division of University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, 30. dubna Street 22, 701 03 Ostrava, Czech Republic
[b] Department of Computer Science, University of Ostrava, Czech Republic

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form 14 May 2015
Accepted 8 June 2015
Available online xxx

Keywords:
Optimal clustering
Adaptive differential evolution
k-means algorithm
Hybrid search
Heuristic rearrangement of cluster centers
Numerical comparison

ABSTRACT

The problem of optimal non-hierarchical clustering is addressed. A new algorithm combining differential evolution and *k*-means is proposed and tested on eight well-known real-world data sets. Two criteria (clustering validity indexes), namely TRW and VCR, were used in the optimization of classification. The classification of objects to be optimized is encoded by the cluster centers in differential evolution (DE) algorithm. It induced the problem of rearrangement of centers in the population to ensure an efficient search via application of evolutionary operators. A new efficient heuristic for this rearrangement was also proposed. The plain DE variants with and without the rearrangement were compared with corresponding hybrid *k*-means variants. The experimental results showed that hybrid variants with *k*-means algorithm are essentially more efficient than the non-hybrid ones. Compared to a standard *k*-means algorithm with restart, the new hybrid algorithm was found more reliable and more efficient, especially in difficult tasks. The results for TRW and VCR criterion were compared. Both criteria provided the same optimal partitions and no significant differences were found in efficiency of the algorithms using these criteria.

1. Introduction

Q3

Cluster analysis (or clustering) is an important exploratory technique used for splitting a collection of objects into relatively homogeneous groups (called clusters) based on object similarities. Clustering is completely data driven process denoted as unsupervised machine learning.

Clustering problem can be defined formally as follows. Let $\mathcal{O}$ be a set of n objects to be grouped. The aim of the clustering algorithm is to find such a partition $\mathbf{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ fulfilling the conditions:

1 $\mathcal{C}_l \neq \emptyset$ forall $l = 1, 2, \ldots, k$
2 $\mathcal{C}_l \cap \mathcal{C}_m = \emptyset$all $l \neq m$
3 $\cup_{l=1}{}^{k}\mathcal{C}_l = \mathcal{O}$
4 objects belonging to the same cluster are as similar to each other as possible, while the objects belonging to different clusters are as dissimilar as possible.

Each partition fulfilling conditions 1–3 is called *feasible* partition. The count of different feasible partitions is given by Stirling number of the second kind $S(n, k)$

$$\begin{aligned} S(n,k) &= \frac{1}{k!}\sum_{l=0}^{k}(-1)^l \binom{k}{k-l}(k-l)^n \\ &= \frac{1}{k!}\sum_{l=1}^{k}(-1)^{k-l}\binom{k}{l} l^n. \end{aligned} \qquad (1)$$

Let us suppose that each object of $\mathcal{O}$ is characterized by p real-valued attributes. Data matrix $\boldsymbol{Z}$ of size $n \times p$ is composed of n row vectors $\boldsymbol{z}_i$, where each element z_{ij} represents the j-th real-valued attribute of the i-th object. Then each feasible partition can be evaluated by a function (criterion) that quantifies the goodness of the partition based on the similarity or dissimilarity of the objects. The function reflects the vague condition 4. The solution of clustering problem is to find such a partition that optimizes the function.

It is known that the clustering problem is hard, Brucker [1] has shown that the clustering problem is NP-hard when $k > 3$. The count of feasible partitions is very large, e.g. for popular test problems like *Iris* is $S(150, 3) \approx 6 \times 10^{70}$ and for *Glass* even $S(214, 6) \approx 4 \times 10^{165}$. This implies that exhaustive search is not applicable in the solution of clustering problem. Thus, heuristics have to be used. There are two main approaches how to solve the clustering problem,

Q2 * Corresponding author. Tel.: +420 597 092183.
E-mail addresses: josef.tvrdik@osu.cz (J. Tvrdík), ivan.krivy@osu.cz (I. Křivý).

http://dx.doi.org/10.1016/j.asoc.2015.06.032
1568-4946/© 2015 Published by Elsevier B.V.

Please cite this article in press as: J. Tvrdík, I. Křivý, Hybrid differential evolution algorithm for optimal clustering, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.06.032

hierarchical and non-hierarchical. The non-hierarchical (partition) clustering algorithms try to decompose the data sets directly into a set of disjoint clusters via optimizing a chosen function. Such functions are also called the clustering validity indices or criteria.

Because of difficulty of the clustering problem, the historically first clustering algorithms inclusive of the most popular *k*-means algorithm are iterative. Their weakness consists in the fact that the solution found by them is often suboptimal only. The iterative clustering algorithms are described in many textbooks, for an overview see, e.g. [2]. Several important iterative clustering algorithms are experimentally compared in Hamerly and Elkan [3].

The goal of our paper is to propose a novel hybrid algorithm combining differential evolution and *k*-means algorithm and apply it to non-hierarchical clustering. The improved hybrid algorithm is derived from our results [4,5], where a preliminary versions of hybridization were introduced and proved to be more efficient than non-hybrid differential evolution.

The rest of the paper is organized as follows. Applications of evolutionary algorithms to clustering are surveyed briefly in Section 2. Two criteria (clustering validity indices) used in experimental comparison are defined in Section 3. The basic frame of DE algorithm is described in Section 4 and encoding of partition for DE in Section 5. New heuristic rearrangement of cluster centers is proposed in Section 6. Section 7 presents the proposed hybrid algorithm in detail. The organization of experiments and experimental setting are described in Sections 8–10. The experimental results are shown and discussed in Section 11 and the paper is closed by concluding remarks in Section 12.

2. A brief literature overview on evolutionary algorithms in clustering

Application of evolutionary algorithms or swarm intelligence to optimal clustering seems to be a natural choice how to solve this algorithmically difficult clustering problem. The paper by Paterlini and Krink [6] can be considered a pioneering work in this field. They compared the performance of floating point encoded genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO). The experimental results showed that DE outperformed the other algorithms in the comparison, which is in accordance with [7], where DE was also the best performing in experimental comparison with GA and PSO on a set of standard benchmark optimization problems.

Das et al. [8] applied an improved DE to clustering including the search of optimal number of clusters. DE was also used in fuzzy-clustering algorithm [9]. Appropriate setting of DE control parameters for clustering problems (especially parameter controlling the exponential crossover) was also studied in [10]. A comprehensive survey of evolutionary algorithms for clustering up to 2009 was published by Hruschka et al. [11].

Hybridization of DE by using *k*-means algorithm for a local search appeared in several papers. Very similar approach was suggested independently in [12] and [4]. A hybrid DE variant with a special rearrangement of the rank of cluster centers after completing each generation [5] proved to be substantially more efficient than standard DE. The efficiency of *k*-means hybridization in evolutionary algorithms was studied by Naldi et al. [13] and found to be beneficial. Karaboga and Ozturk [14] used Artificial Bee Colony (ABC) algorithm in classification of objects and compared it with PSO and other classification algorithms on 13 test problems. They found that ABC can be successfully applied to clustering for the purpose of classification. Abraham et al. [15] provided a very comprehensive survey on swarm intelligence algorithms for data clustering. A survey of evolutionary and swarm intelligence algorithms for clustering can be also found in [16] and [17].

3. Criteria of optimal partitioning

There are several optimizing criteria convenient for comparing the degree of optimality over all possible partitions, see e.g. [2,18]. In order to preserve the possibility of comparison with the results presented in literature [6], two criteria are used in experimental tests of the proposed algorithm.

Trace within criterion (hereafter TRW), proposed by Friedman and Rubin [19], is based on minimizing the trace of pooled-within groups scatter matrix ($\boldsymbol{W}$) defined as

$$\boldsymbol{W} = \sum_{l=1}^{k} \boldsymbol{W}_l . \tag{2}$$

$\boldsymbol{W}_l$ is the variance matrix of attributes for the objects belonging to cluster C_l given by

$$\boldsymbol{W}_l = \sum_{j=1}^{n_l} (\boldsymbol{z}_j^{(l)} - \bar{\boldsymbol{z}}^{(l)})(\boldsymbol{z}_j^{(l)} - \bar{\boldsymbol{z}}^{(l)})^T , \tag{3}$$

where $\boldsymbol{z}_j^{(l)}$ is the vector of attributes for the *j*-th object of cluster C_l, $\bar{\boldsymbol{z}}^{(l)} = \left(\sum_{j=1}^{n_l} \boldsymbol{z}_j^{(l)}\right)/n_l$ the vector of means (centroids) for cluster C_l, and $n_l = |C_l|$. Thus, TRW is simply

$$\mathrm{TRW} = \mathrm{tr}(\boldsymbol{W}). \tag{4}$$

The between groups scatter matrix can be expressed analogously in the form

$$\boldsymbol{B} = \sum_{l=1}^{k} n_l (\bar{\boldsymbol{z}}^{(l)} - \bar{\boldsymbol{z}})(\bar{\boldsymbol{z}}^{(l)} - \bar{\boldsymbol{z}})^T , \tag{5}$$

$\bar{\boldsymbol{z}} = \left(\sum_{i=1}^{n} \boldsymbol{z}_i\right)/n$ being the vector of means for all objects, $n = \sum_{l=1}^{k} n_l$. It can be easily proved that the total scatter matrix $\boldsymbol{T}$, defined as $\boldsymbol{T} = \sum_{i=1}^{n} (\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})^T$, meets the equality $\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}$.

Variance ratio criterion (VRC) is based on maximizing the following ratio

$$\mathrm{VRC} = \frac{\mathrm{tr}(\boldsymbol{B})/(k-1)}{\mathrm{tr}(\boldsymbol{W})/(n-k)}. \tag{6}$$

For given $k > 1$, VCR criterion can be expressed as $VCR = c_1/TRW - c_2$, where $c_1 = (n-k) \times tr(\boldsymbol{T})/(k-1)$ and $c_2 = (n-k)/(k-1)$. It shows that the global optimum point is equivalent for both criteria. From this point of view, there is no reason to use the VCR criterion if *k* is given. However, when the heuristic search of the global optimum point is used, the shape of one function (landscape) may be more convenient than the other.

4. Differential evolution algorithm

The differential evolution (DE), introduced by Storn and Price [20], has become one of the most frequently evolutionary algorithms used for solving the continuous global optimization problems [21]. When considering the minimization problem, for a real function $f(\boldsymbol{x}) \to \mathbb{R}$, where $\boldsymbol{x}$ is a continuous variable (vector of length *d*) with the domain $D \subset \mathbb{R}^d$, the global minimum point $\boldsymbol{x}^*$ satisfying condition $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$ for $\forall \boldsymbol{x} \in D$ is to be found. The domain *D* is defined by specifying boundary constraints, $D = \prod_{j=1}^{d} [a_j, b_j],\ a_j < b_j,\ j = 1, 2, \ldots, d$.

The initial population of *N* points is generated at random, uniformly distributed in *D*, each point in *D* is considered as a candidate of the solution and then the population is evolving generation by generation until the stopping condition is met. Next generation *Q*

Please cite this article in press as: J. Tvrdík, I. Křivý, Hybrid differential evolution algorithm for optimal clustering, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.06.032