Neurocomputing 158 (2015) 144–154

Contents lists available at ScienceDirect

Neurocomputing

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

A dynamic shuffled differential evolution algorithm for data clustering

Wan-li Xiang [a,b,*], Ning Zhu [a], Shou-feng Ma [a], Xue-lei Meng [b], Mei-qing An [b]

[a] Institute of Systems Engineering, Tianjin University, Tianjin 300072, PR China
[b] School of Traffic & Transportation, Lanzhou Jiaotong University, Lanzhou, Gansu 730070, PR China

ARTICLE INFO

Article history:
Received 21 June 2014
Received in revised form
5 December 2014
Accepted 30 January 2015
Communicated by Swagatam Das
Available online 8 February 2015

Keywords:
Differential evolution
Shuffled scheme
Information exchange
Novel initial technique
Data clustering

ABSTRACT

In order to further improve the convergence performance of data clustering algorithms, a dynamic shuffled differential evolution algorithm, DSDE for short, is presented in this paper. In DSDE, mutation strategy DE/best/1 is employed, which can take advantage of the direction guidance information of best individual so as to speed up the corresponding algorithm. Meanwhile, inspired by shuffled frog leaping algorithm, a sorting scheme and a randomly shuffled scheme are used to divide a total population into two subpopulations during the evolving process. In this way, mutation strategy DE/best/1 is actually used in two subpopulations, respectively, which can effectively exchange information between two subpopulations and balance the exploitation ability of DE/best/1/bin. In addition, most popular data clustering algorithms suffer from the choice of initial clustering centers, which may cause a premature convergence. Here a novel initial technique, called the random multi-step sampling, is integrated into DSDE to overcome the shortcoming. Then an experiment tested on 11 well-known datasets has been carried out, and the related results demonstrate that DSDE significantly outperforms DE/rand/1/bin and DE/best/1/bin. Next, another comparison among DSDE and other four well-known data clustering algorithms is conducted. The related results also show that DSDE is superior to other four approaches including particle swarm optimization with age-group topology (PSOAG) in terms of objective function value, i.e., the sum of intra-cluster distance. In a word, all the experimental results confirm that the proposed algorithm DSDE can be considered as an excellent tool for data clustering.

1. Introduction

Clustering is an unsupervised learning method that divides a dataset into groups of similar objects by minimizing the similarity between objects in different clusters and maximizing the similarity between objects within the same cluster. When used on a set of objects, it is helpful for identifying some inherent structures of the set of objects. Thus, clustering analysis can play a key role in many research fields including data mining, machine learning, bioinformatics and so on. Especially, it can work well without any prior information, which is also the main difference to supervised/semisupervised classification. During the past five decades, many research works have been proposed for data clustering [6]. That is, the clustering problem has arisen many ways to cluster a dataset. These approaches mainly focus on complex network approach [1,27,29], *K*-means [4] and its enhanced variants [5,12,35,43], swarm intelligent algorithms [2,3,7,9,11,15-25,28,31,32,35-37,44], and others [8,10,13,14,26,30,33,34]. Among them, *K*-means algorithm is one of the most popular algorithms, and it is an unsupervised data clustering algorithm, which tries to divide an entire dataset (i.e., X) into K clusters (i.e., $C_1, C_2, \ldots, C_K$) through randomly choosing K data points as initial cluster centers. However, *K*-means algorithm is sensitive to the selection of initial points and may fail to cluster large scale datasets [12,36,43].

In order to overcome the shortcomings of *K*-means algorithm, many researchers have focused on swarm intelligence algorithms, which can perform parallel search in a complex search space so as to better avoid local minima trap. For example, Chuang et al. [2] proposed a chaotic particle swarm optimization (CPSO) for data clustering by replacing the random parameters of PSO with chaotic variables of the logistic map. Inspired by the crossover operator of genetic algorithm, Yan et al. [22] proposed a hybrid artificial bee colony algorithm (HABC) by introducing arithmetic crossover operation in original artificial bee colony algorithm. Hatamlou et al. [44] proposed a combined algorithm based on *K*-means algorithm and gravitational search algorithm (GSA). The related comparisons show that it is superior to both *K*-means algorithm and GSA. More recently, Jiang et al. [11] proposed a novel age-based particle swarm optimization, which is called the PSOAG. In PSOAG, a novel technique of keeping population diversity is introduced. Meantime, its superiority is obvious when compared with ant colony optimization algorithm (ACO), particle swarm

* Corresponding author.
E-mail address: xiangwl@tju.edu.cn (W.-l. Xiang).

http://dx.doi.org/10.1016/j.neucom.2015.01.058
0925-2312/

optimization (PSO), artificial bee colony algorithm (ABC) and differential evolution (DE) algorithm in terms of fitness function values and the clustering accuracy. In summary, the performances achieved by these modified variants based on various evolutionary computing techniques are better than before.

However, there still exists the shortcoming of premature convergence in those modified evolutionary algorithms to some extent. To be specific, the exploitation and exploration abilities of existing evolutionary algorithms have to be further balanced. Especially, the population diversity is necessary to be better kept during the evolution process, and a better mechanism of information sharing among different individuals are also needed to be designed. In order to further improve the clustering effectiveness of swarm based intelligent algorithms, a dynamic shuffled differential evolution, named DSDE, is proposed in view of the faster convergence speed of DE [41,42]. When compared with DE/rand/1/bin, DE/best/1/bin, ACO, ABC, PSO, and the recent PSOAG, the superiority of DSDE is obvious.

The rest of the paper is organized as follows. In Section 2, the problem formulation over data clustering is described. In Section 3, the traditional differential evolution algorithm is briefly described. Subsequently, a dynamic shuffled differential evolution, called the DSDE, is presented in detail in Section 4. Nextly, comprehensive experiments are conducted in Section 5 to validate the performance of DSDE. Finally, a conclusion is drawn in Section 6.

2. Problem formulation

For the global optimization based clustering problem [3,22,50], a clustering criterion, i.e., a degree of similarity, is needed, with which a global assignment can be found by minimizing the objective function of the sum of intra-cluster distance. As far as the similarity measurement is concerned, there are various similarity measurements such as Euclidian distance, Manhattan distance and Minkowski distance within some clustering researches. In general, the Euclidian distance is usually considered as the similarity measurement. In our study, the measurement of Euclidean distance is also used to calculate the distance of any two objects (o_i and o_j) within the cluster. Generally, it can be formulated as follows [22,36]:

$$D(o_i, o_j) = \| o_i - o_j \| = \sqrt{\sum_{m=1}^{d} (o_{im} - o_{jm})^2} \tag{1}$$

where d is the number of attributes of object, and o_i and o_j are two different objects within the cluster, respectively.

Based on the above similarity measurement, a partition clustering problem can be converted into a global optimization problem, which can be described as follows [3,11,21]:

$$\underset{Z,W}{\text{Minimize}} \quad f(Z,W) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} D(x_i, z_k) \tag{2}$$

$$\text{s.t.} \begin{cases} \sum_{k=1}^{K} w_{ik} = 1, \quad i = 1,2,\ldots,n; & \text{(a)} \\ w_{ik} \in \{0,1\}, \quad \forall i \in \{1,2,\ldots,n\},\ k \in \{1,2,\ldots,K\} & \text{(b).} \end{cases} \tag{3}$$

where n is the number of all the samples, and K is the number of given clusters, and x_i represents the coordinates of the ith object in current samples, and $w_{ik} = 1$ means that the ith object is grouped into the kth cluster, or else the ith object does not belong to the kth cluster, and $D(x_i, z_k)$ denotes the distance between the ith object x_i and the center of kth cluster z_k. It should be noted that $W = \{w_{ik} \mid i = 1,2,\ldots,n, k = 1,2,\ldots,K\}$, and $Z = \{z_k \mid k = 1,2,\ldots,K\}$.

What is more, the value of w_{ik} in Eq. (3) depends on a partition criterion of samples. Namely, given a sample set $X = \{x_1, x_2, \ldots, x_n\}$, determine a partition of all objects which satisfies the following equations [3]:

$$\begin{cases} \sum_{i=1}^{K} C_i = X; & \text{(a)} \\ C_i \cap C_j = \phi, \quad \forall i,j \in \{1,2,\ldots,K\} \wedge i \neq j; & \text{(b)} \\ C_i \neq \phi, \quad \forall i \in \{1,2,\ldots,K\} & \text{(c).} \end{cases} \tag{4}$$

where $C_i (i = 1,2,\ldots,K)$ is the objects set of ith cluster, and its members can be determined by the following [3]:

$$\begin{cases} C_i = \{x_k \mid \| x_k - z_i \| \leq \| x_k - z_p \|, x_k \in X\}, \quad p \neq i, p = 1,2,\ldots,K; & \text{(a)} \\ z_i = \dfrac{1}{|C_i|} \sum_{x_k \in C_i} x_k, \quad i = 1,2,\ldots,K & \text{(b).} \end{cases} \tag{5}$$

where $\| \cdot \|$ represents the Euclidean distance of any two objects in the sample set, and the mean of all objects within cluster i, z_i, denotes a new center of cluster i, which is often used in well-known clustering technique k-means.

3. Differential evolution algorithm

Differential evolution algorithm, first proposed by Storn and Price, is a simple yet powerful meta-heuristic algorithm for global optimization over continuous search space [41]. Since its invention in 1997, the DE algorithm has attracted many researchers to study the creative algorithm. For example, Liu and Lampinen proposed a fuzzy adaptive differential evolution algorithm by using a fuzzy controller to adapt the control parameters [47]. Teo proposed a parameterless differential evolution algorithm based on self-adaption [48]. Rahnamayan et al. proposed an opposition based differential evolution algorithm (ODE) [42], in which a novel opposition based learning technique is proposed. Gong et al. proposed a hybrid algorithm DE/BBO based on differential evolution and biogeography-based optimization [49].

Like other population-based intelligent algorithms, the first phase is to randomly generate an initial population $X (X = \{x_i \mid x_i = (x_{i1}, x_{i2}, \ldots, x_{iD}), i = 1,2,\ldots,N_p\})$ in DE. Subsequently, DE enters a loop composed of mutation, crossover, and selection operations.

3.1. Mutation

At this phase, a mutant vector v_i is generated by the following equation:

$$v_i = x_a + F \cdot (x_b - x_c) \tag{6}$$

where $i = 1,2,\ldots,N_p$, and $a,b,c \in \{1,2,\ldots,N_p\}$ are random integer number and $a \neq b \neq c \neq i$, and scale factor F is a real and constant number within $[0,2]$, which is employed to control the amplification of differential deviation $(x_b - x_c)$ [41].

3.2. Crossover

At the second phase, DE usually utilizes a binomial crossover operation to produce a trial vector $u_i = (u_{i1}, u_{i2}, \ldots, u_{iD})$ according to the following equation:

$$u_{ij} = \begin{cases} v_{ij} & \text{if } \text{rand}[0,1]_j < Cr \vee j = = j_{rand}, \\ x_{ij} & \text{otherwise}, \end{cases} \tag{7}$$

where $i = 1,2,\ldots,N_p$, and $j = 1,2,\ldots,D$, and $\text{rand}[0,1]_j$ is a real random number within [0,1], and $j_{rand} \in \{1,2,\ldots,D\}$ is a randomly generated integer, which is used to make sure that the trial vector u_i gets at least one component from the mutant vector v_i. In