Information Processing Letters 116 (2016) 1–14

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Flower Pollination Algorithm with Bee Pollinator for cluster analysis

Rui Wang [a], Yongquan Zhou [a,b,*], Shilei Qiao [a], Kang Huang [a]

[a] College of Information Science and Engineering, Guangxi University for Nationalities, Nanning 530006, China
[b] Guangxi High School Key Laboratory of Complex System and Computational Intelligence, Nanning 530006, China

ARTICLE INFO

Article history:
Received 23 November 2014
Received in revised form 11 August 2015
Accepted 17 August 2015
Available online 20 August 2015
Communicated by S.M. Yiu

Keywords:
Flower pollination algorithm
Randomized algorithms
Discard pollen operator
Elite based mutation operator
Crossover operator
Clustering problem

ABSTRACT

Clustering is a popular data analysis and data mining technique. The k-means clustering algorithm is one of the most commonly used methods. However, it highly depends on the initial solution and is easy to trap into the local optimal. For overcoming these disadvantages of the *k*-means method, Flower Pollination Algorithm with Bee Pollinator is proposed. Discard pollen operator and crossover operator are applied to increase diversity of the population, and local searching ability is enhanced by using elite based mutation operator. Ten data sets are selected to evaluate the performance of proposed algorithm. Compared with DE, CS, ABC, PSO, FPA and *k*-Means, the experiment results show that Flower Pollination Algorithm with Bee Pollinator has not only higher accuracy but also higher level of stability. And the faster convergence speed can also be validated by statistical results.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Data clustering is the process of grouping together similar multi-dimensional data vectors into a number of clusters. Clustering algorithms have been applied to a wide range of problems, including exploratory data analysis, data mining [1], image segmentation [2] and mathematical programming [3]. Clustering techniques have been used successfully to address the scalability problem of machine learning and data mining algorithms, where prior to, and during training, training data is clustered, and samples from these clusters are selected for training, thereby reducing the computational complexity of training process, and even improving generalization performance [4,5].

Clustering algorithms can be grouped into two main classes of algorithms, namely hierarchical and partitional. The *k*-means clustering method [6] is one of the most commonly used partitional methods. However the results of *k*-means solving the clustering problems highly depend on the initial solution and it is easy to fall into local optimal solutions. For overcoming this problem, many scholars began to solve the clustering problem using meta-heuristic algorithms. Nikham et al. have proposed an efficient hybrid evolutionary algorithm based on combining ACO and SA (simulated annealing algorithm, 1989 [7] for clustering problem [8,9]. In 1991, A. Colorni et al. have presented an ant colony optimization (ACO) algorithm based on the behavior of ants seeking a path between their colony and a source of food. Then P.S. Shelokar and Y. Kao solved the clustering problem using the ACO algorithm [10,11]. J. Kennedy and R.C. Eberhart have proposed a particle swarm optimization (PSO) algorithm which simulates the movement of organisms in bird flock or fish school in 1995 [12]. The algorithm also has been adopted to solve this problem by M. Omran and V.D. Merwe [13,14]. Kao et al. have presented a hybrid approach according to combination of the *k*-means algorithm, Nelder–Mead simplex search and PSO for clustering analysis [15]. Kevin et al.

* Corresponding author.
E-mail address: yongquanzhou@126.com (Y. Zhou).

http://dx.doi.org/10.1016/j.ipl.2015.08.007
0020-0190/© 2015 Elsevier B.V. All rights reserved.

have used an evolutionary-based rough clustering algorithm for the clustering problem [16].

In this paper, a variant of the flower pollination algorithm is used to solve the data clustering problem. Flower Pollination Algorithm (proposed by Yang in 2012) [17] is a new population-based intelligent optimization algorithm by simulating flower pollination behavior in nature. And it has been extensively researched in last two years by scholars. Yang and Xingshi He have used FPA to solve multi-objective optimization problem in 2013 [18]. Marwa Sharawi has applied FPA for solving Wireless Sensor Network Lifetime global optimization in 2014 [19]. Osama Abdel-Raouf has used an improved FPA to solve Sudoku Puzzles in 2014 [20]. And FPA has been used to solve Large Integer Programming Problems by Ibrahim El-henawy in 2014 [21].

The remainder of this paper is organized as follows. Section 2 introduces the mathematical model of clustering; Section 3 describes the principle of basic flower pollination algorithm; while Section 4 specified implementation procedure of our proposed Flower Pollination Algorithm with Bee Pollinator (BPFPA); in Section 5, series of comparison experiments are conducted; result analysis will be given in Section 6; our conclusion and future works are described in Section 7.

2. The mathematical model of clustering

2.1. Mathematical definition of data clustering

The goal of data clustering is grouping data into a number of clusters, clustering problems can be seen in practice frequently. In this subsection, a mathematical definition of data clustering is presented. In order to explain the definition clearly, we supposed that there exists a data set $D = \{d_1, d_2, \ldots, d_n\}$. And each individual d_i $(i = 1, 2, \ldots, n)$ has many features. If the dimension is m, each individual can be shown as $d_i = (l_1, l_2, \ldots, l_m)$. Data clustering is a process which can classify the given data set D into a certain numbers of clusters $G_1, G_2, \ldots, G_K$ (assume K clusters) based on the similarity of individuals. And $G_1, G_2, \ldots, G_K$ should satisfy the following formulas:

1). $G_i \neq \emptyset, \quad i = 1, 2, \ldots, K.$

2). $G_i \cap G_j = \emptyset, \quad i, j = 1, 2, \ldots, K, \ i \neq j.$

3). $\bigcup_{i=1}^{K} G_i = \{d_1, d_2, \ldots, d_n\}.$

2.2. The principle of data clustering

In the clustering process, if the given data set D should be divided into K clusters $(G_1, G_2, \ldots, G_K)$, and each cluster must have one center c_j $(j = 1, 2, \ldots, K)$. It is supposed that $C = (c_1, c_2, \ldots, c_K)$ are the centers of $(G_1, G_2, \ldots, G_K)$. Where c_j is the center of subset G_j.

The main idea of clustering is to define K centers, one for each cluster. These centers should be placed in a crafty way, because different location will causes different result. Therefore, the better choice is to place them as far away from each other as possible. In this paper, we will use Euclidean metric as a distance metric. The expression is given as follows:

$$d(d_i, c_j) = \sqrt{\sum_{k=1}^{m} (d_{ik} - c_{jk})^2} \tag{1}$$

where d_i $(i = 1, 2, \ldots, n)$ is an individual in the given data set D, m is the number of individual features; c_j $(j = 1, 2, \ldots, K)$ is the center of jth subset. Because individual has m features, c_j can be presented by $(c_{j,1}, c_{j,2}, \ldots, c_{j,m})$. In order to confirm which subset d_i belongs to, the distances between d_i and c_j $(j = 1, 2, \ldots, K)$ should be calculated via (1). If the distance between d_i and c_{best} $(best = 1, 2, \ldots, K)$ is smaller than the distances between d_i and other centers (except c_{best}), we can make the decision that d_i should belongs to G_{best}. For example, if the value of $d(d_1, c_2)$ is smaller than $d(d_1, c_j)$, $(j \neq 2)$, we can draw the conclusion that d_1 should be distributed to G_1.

2.3. The performance evaluation function of data clustering

In this paper, data clustering problem is solved by population based algorithm (Flower Pollination Algorithm and Flower Pollination Algorithm with Bee Pollinator). For explaining the evaluation process explicitly, we suppose that given data set D should be divided into K subsets. And the dimension of individual of data set D is m. In order to optimize the coordinates of centers of K subsets, it is easily to find that the dimension of solution should be $K * m$. The individual in the population can be described as $s = (c_1, c_2, \ldots, c_K)$. A great classification should minimize the sum of distances value. So we should try to minimize the distance between individual d_i and the center (c_j) of subset it belongs to. Finally, the proposed algorithm aims at minimizing the objective function, which can be expressed as following:

$$f(D, C) = \sum_{i=1}^{n} \min\left\{\|d_i - c_k\| \mid k = 1, 2, \cdots, K\right\} \tag{2}$$

where $D = (d_1, d_2, \ldots, d_n)$ is the given data set, $C = (c_1, c_2, \ldots, c_K)$ is the centers of subsets $(G_1, G_2, \ldots, G_K)$.

3. Flower Pollination Algorithm (FPA)

Flower Pollination Algorithm (FPA) was founded by Yang in the year 2012. Inspired by the flow pollination process of flowering plants are the following rules [22,17]:

Rule 1: Biotic and cross-pollination can be considered as a process of global pollination process, and pollen-carrying pollinators move in a way that obeys Lévy flights.

Rule 2: For local pollination, a biotic and self-pollination are used.

Rule 3: Pollinators such as insects can develop flower constancy, which is equivalent to a reproduction probability that is proportional to the similarity of two flowers involved.

Download English Version:

https://daneshyari.com/en/article/427081

Download Persian Version:

https://daneshyari.com/article/427081

Daneshyari.com