Applied Soft Computing 12 (2012) 3421–3441

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Automatic fuzzy partitioning approach using Variable string length Artificial Bee Colony (VABC) algorithm☆

Zhi-gang Su*, Pei-hong Wang, Jiong Shen, Yi-guo Li, Yu-fei Zhang, En-jun Hu

Key Laboratory of Energy Thermal Conversion and Control of Ministry of Education, School of Energy and Environment, Southeast University, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 12 January 2011
Received in revised form 17 February 2012
Accepted 25 June 2012
Available online 11 July 2012

Keywords:
Fuzzy C-Means
Automatic clustering
Swarm intelligence
Artificial Bee Colony
Variable string length Artificial Bee Colony
Genetic Algorithm

ABSTRACT

Swarm intelligence based automatic fuzzy clustering is recently an important and interesting unsupervised learning problem. In this article, an automatic fuzzy clustering technique is proposed based on a novel version of Artificial Bee Colony (ABC) algorithm.

The idea of variable length genotypes is introduced to the ABC, and a novel version of ABC, called Variable string length Artificial Bee Colony (VABC) algorithm, is proposed. The VABC algorithm is derived from the ABC by redefining or modifying some operations in the ABC: the fixed length strings are represented by using variable length strings, the scheme for producing candidate solutions is modified, and some mutation operations are introduced. Use of VABC allows the encoding of variable number of clusters. This makes the VABC based Fuzzy C-Means clustering technique (VABC-FCM) not require a priori specification of the number of clusters. Moreover, the VABC-FCM has powerful global search ability under rational parameter setting. Some artificial data sets and real-life data sets are applied to validate the performance of VABC-FCM. The experimental results show that VABC-FCM can automatically evolve the optimal number of clusters and find proper fuzzy partitioning for these data sets when a rational validity index is adopted. Finally, the performance of VABC-FCM is compared with those of the Variable string length Genetic Algorithm based Fuzzy C-Means clustering (VGA-FCM), Particle Swarm Optimization algorithm based Fuzzy C-Means clustering (PSO-FCM), and Differential Evolutional algorithm based Fuzzy C-Means clustering (DE-FCM). The results show that the VABC-FCM outperforms VGA-FCM, PSO-FCM and DE-FCM in most of the cases.

1. Introduction

Clustering can be considered the most important *unsupervised learning* problem, so, as every other problem of this kind, it deals with finding natural partitioning of a data set such that data points within the same cluster are more similar than those within different clusters. The existing clustering algorithms can be simply classified into following two categories: hierarchical clustering and partitional clustering [40]. Fuzzy C-Means (FCM) [6] is a well-known partitional clustering technique that uses the principles of fuzzy sets to evolve a partition matrix for the unlabeled data points. However, FCM has three major limitations: (1) it often gets stuck at suboptimal solutions based on the initial configuration; (2) it requires the a priori specification of the number of clusters (c); and (3) it can detect only hyper spherical shaped clusters. In this paper, we focus on investigation of the first two issues and thus propose a novel fuzzy clustering technique.

To overcome the first limitation, the evolutionary algorithms and swarm intelligent techniques can be applied. Genetic Algorithm (GA) [10,32] is randomized search and optimization techniques guided by the principles of evolution and natural genetics. Genetic and other evolutionary algorithms such as Differential Evolution (DE) have been earlier used for clustering of data, see in [1,7,9,12,26,27,29,30,35], without claiming of completeness. Over the last decade, modeling the behavior of social insects such as birds, ants, and bees for the purpose of search and optimization has become an emerging area of swarm intelligence and successfully applied to cluster. Some ant colony algorithm based clustering techniques were presented [13,18,41] and heir performances were compared with GA, Tabu search (TS) and Simulated Annealing (SA) algorithm. The Particle Swarm Optimization (PSO), simulating bird flocking, was used for clustering in [17,25,31,36,47]. Honey-bees are among the most closely studied social insets. Their foraging behavior, learning, memorizing and information sharing characteristics have recently been one of the most interesting research areas in swarm intelligence [45]. Recently, Karaboga [20], Karaboga and Basturk [21] have described an Artificial Bee Colony (ABC)

☆ This work is supported by the National Natural Science Foundation of China (No. 51106025).

* Corresponding author. Tel.: +86 25 83794520; fax: +86 25 57714489.
E-mail address: suzhigang2008@163.com (Z.-g. Su).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.06.019

algorithm based on the foraging behavior of honey-bees for numerical optimization problems. They have compared the performance of the ABC algorithm with those of other well-known modern heuristic algorithms such as GA, DE and PSO for unconstrained optimization problems [21]. They remarked that the performance of ABC is better than those of the GA, DE and PSO in most of the cases. This motivates the ABC algorithm based K-means clustering techniques [23,48] and ABC algorithm based fuzzy clustering (ABC-FCM) [22]. The performance of ABC based clustering techniques was compared with the popular heuristics algorithm in clustering such as GA, SA, TS, and PSO [22,23,48]. It reveals that the ABC based clustering has very encouraging results in terms of quality of solution and the processing time required. Among all the above works, however, the number of clusters is assumed to be fixed a priori and/or the clusters are assumed to be crisp in nature.

In most of the real-life situations, the number of clusters in a data set is not known a priori. The real challenge in this situation is to be able to automatically evolve a proper value of c and to provide the appropriate partitioning of a data set. To overcome the second limitation, Maulik and Bandyopadhyay [28] attempt to automatically evolve the appreciate number of clusters as well as fuzzy partitioning of a data set. For this purpose, *Variable string length GA* (VGA), where different chromosomes in the population may encode different number of clusters, is used. Thus, the so-called Fuzzy-VGA clustering technique is proposed. In this method, clustering validity index such as Xie-Beni (XB index) of the partitioning encoded in a chromosome is used to measure its fitness value. In order to tackle the concept of variable string lengths, the crossover and the mutation operators are redefined accordingly. Following Maulik and Bandyopadhyay, Saha and Bandyopadhyay [37] propose a fuzzy genetic clustering technique based on a new Point Symmetry distance, called Fuzzy-VGAPS, which not only can automatically evolve the number of clusters but also can deal with the clustering for different shapes. Besides, there are few literatures on this issue until now. Therefore, it is necessary and interesting to propose novel methods on investigation of this issue.

As already remarked hereinbefore, the ABC algorithm is recently the most popular swarm intelligent algorithm for numerical function optimization and clustering. However, the conventional ABC based clustering techniques [22,23,48] cannot automatically determine the number of clusters. It therefore is interesting to propose a novel version of ABC algorithm holding the property such as that of the VGA. In other words, as a counterpart to VGA, there exists the Variable string length ABC (VABC), in which, different strings in the same population encode different number of clusters. In order to deal with the concept of variable string length, the original exploration scheme in ABC is modified. In addition, to avoid suboptimum and to accelerate the convergence, some mutation operations are introduced. With the VABC, a novel Fuzzy C-Means algorithm, which can automatically evolve the number of clusters, is thus proposed to partition unlabeled data points. The objective of this paper is to propose a novel fuzzy partitioning approach using Variable string length ABC. The two main contributions of this study are: the proposed VABC algorithm and the VABC based fuzzy partitioning approach.

The rest of this paper is organized as follows. Section 2 introduces the Variable string length ABC algorithm (VABC) after recalling the principles of the ABC algorithm. Section 3 presents the VABC based fuzzy partitioning approach (VABC-FCM). Section 4 designs a numerical experiment to analyze the influence of control parameters on the performance of VABC-FCM by using an artificial data set. In Section 5, some other artificial data sets and real-life data sets are applied to validate VABC-FCM, by comparing with other clustering approaches. The last section concludes this paper.

2. The ABC and VABC algorithms

In Section 2.1, the necessary foundation of ABC is recalled. Afterward, the Variable string length ABC algorithm, or called ABC with variable length strings, is proposed in Section 2.2.

2.1. Conceptions of the basic ABC algorithm

The ABC algorithm is a new swarm intelligence based optimizer proposed by Karaboga [20,21] for multivariable and multi-modal continuous function optimization. Inspired by the intelligent foraging behavior of honeybee swarm, the ABC algorithm classifies the foraging artificial bees into three groups, namely, employed bees, onlookers and scouts. A bee that is currently exploiting a food source is called an employed bee. A bee waiting in the hive for making decision to choose a food source is named as an onlooker. A bee carrying out a random search for a new food source is called a scout. In the ABC algorithm, each solution to the problem under consideration is called a food source and represented by a d-dimension real-valued string; whereas the fitness of the solution is correspond to the nectar amount of the associated food resource. Similar to the other swarm intelligence based approaches, the ABC algorithm is an iterative process. It starts with a population of randomly generated solutions or food sources. Then the following three steps are repeated until a termination criterion is met [21].

1. Send the employed bees onto the food sources and then measure their nectar amounts (i.e., the fitness).
2. Select the food sources by the onlookers after sharing the information of employed bees and determine the nectar amount of the food sources.
3. Determine the scout bees and send them onto the possible food sources.

A set of food source positions are randomly selected by the bees at the initialization stage and their nectar amounts are determined. Then, these bees come into the hive and share the nectar information of the sources with the bees, waiting on the dance area within the hive. At the second stage, after sharing the information, every employed bee goes to the food source area visited by her at the previous cycle since that food source exists in her memory, and then choose a new food source by means of visual information in the neighborhood of the present one. At the third stage, an onlooker prefers a food source area depending on the nectar information distributed by the employed bees on the dance area. As the nectar amount of a food source increase, the probability with which that food source is chosen by an onlooker increase, too. Therefore, the dance of employed bees carrying higher nectar recruits the onlookers for the food source areas with higher nectar amount. After arriving at the selected area, she chooses a new food source in the neighborhood of the one in the memory depending on visual information. Visual information is based on the comparison of food source positions. When the nectar of a food source is abandoned by the bees, a new food source is randomly determined by a scout bee and replaced with the abandoned one. In this model, at each cycle at most one scout goes outside for searching a new food source and the number of employed and onlooker bees are equal.

In the ABC algorithm, the position of a food source represents a possible solution of the optimization problem and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution. The number of the employed bees or the onlooker bees is equal to the number of solutions in the population. At the first step, the ABC generates a randomly distributed initial population of n solutions (food source positions). Each solution x_i ($i = 1, 2, \ldots, n$) is a d-dimensional vector, where the d is the number of optimization parameters. After initialization, the population