



Dynamic clustering with improved binary artificial bee colony algorithm



Celal Ozturk*, Emrah Hancer, Dervis Karaboga

Erciyes University, Engineering Faculty, Computer Engineering Department, Kayseri, Turkey

ARTICLE INFO

Article history:

Received 16 July 2012

Received in revised form 26 June 2014

Accepted 30 November 2014

Available online 8 December 2014

Keywords:

Cluster analysis

Automatic clustering

Discrete optimization

Binary artificial bee colony algorithm

ABSTRACT

One of the most well-known binary (discrete) versions of the artificial bee colony algorithm is the similarity measure based discrete artificial bee colony, which was first proposed to deal with the uncapacitated facility location (UFLP) problem. The discrete artificial bee colony simply depends on measuring the similarity between the binary vectors through Jaccard coefficient. Although it is accepted as one of the simple, novel and efficient binary variant of the artificial bee colony, the applied mechanism for generating new solutions concerning to the information of similarity between the solutions only consider one similarity case i.e. it does not handle all similarity cases. To cover this issue, new solution generation mechanism of the discrete artificial bee colony is enhanced using all similarity cases through the genetically inspired components. Furthermore, the superiority of the proposed algorithm is demonstrated by comparing it with the basic discrete artificial bee colony, binary particle swarm optimization, genetic algorithm in dynamic (automatic) clustering, in which the number of clusters is determined automatically i.e. it does not need to be specified in contrast to the classical techniques. Not only evolutionary computation based algorithms, but also classical approaches such as fuzzy C-means and K-means are employed to put forward the effectiveness of the proposed approach in clustering. The obtained results indicate that the discrete artificial bee colony with the enhanced solution generator component is able to reach more valuable solutions than the other algorithms in dynamic clustering, which is strongly accepted as one of the most difficult NP-hard problem by researchers.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

With the rapid development of computer hardware and software, datasets with great capacities can be stored without more effort. However, these datasets cannot be used or specified by users without of any pre-process. One of the new interdisciplinary fields of computer science, data mining concerns with datasets by basically trying to extract meaningful data and summarizing it into useful information through clustering, feature extraction, statistical tests, etc. [1]. In this study, general motivation just focuses on clustering, which is one of the most appreciated subjects by the researchers and is used in many real-world applications such as bioinformatics, machine learning, image analysis, and pattern recognition and market analysis. In clustering, the main goal is to divide data into groups or clusters based on some similarity measures like distance, intervals within multidimensional data [2]. Through clustering, valuable information can be extracted from enormous quantities of data.

Clustering algorithms fall into two main categories, hierarchical and partitional algorithms. Hierarchical algorithms are based on the use of proximity matrix indicating the similarity between every pair of data points to be clustered and its result is “dendrogram representing the nested grouping of patterns and similarity levels at which groupings change and levels are created through bottom up or bottom down approaches [2]”. In agglomerative (bottom up) hierarchical algorithms, each data member is assigned to a unique cluster, then two clusters are found repeatedly according to the proximity matrix and finally they are merged. The basic agglomerative hierarchical algorithm has the following steps [3]: firstly, construct a similarity matrix which shows the difference between each pair of data. After that, assign $K=N$ where N is the number of data and K is the number of clusters. Then repeatedly find the nearest pair of distinct clusters, merge these clusters and decrement K , 1 by 1 while $K>1$. The process of merging clusters can be applied by different ways, but the well-known are single link and complete link. Single link algorithms are based on merging two groups which have the smallest distance between their closest members. In contrast, complete link algorithms are based on merging groups which have the smallest distance between their most distant members. As for the divisive hierarchical algorithms, all data members are

* Corresponding author. Tel.: +90 352207666x32581; fax: +90 3524375784.
E-mail address: celal@erciyes.edu.tr (C. Ozturk).

assigned to one cluster and then spitted a cluster at each stage until number of clusters is equal to the number of data points. Although the number of clusters is not predefined and the initial conditions do not affect the clustering process in hierarchical algorithms, these algorithms are not dynamic i.e. after a data point assigned to a cluster, its cluster cannot be updated, and the lack of information about global size and shape might cause overlapping clusters [4].

Contrary to hierarchical algorithms, partitional algorithms allow cluster members to be updated if it improves clustering performance. Partitional clustering attempts to decompose the data into a set of disjoint clusters using similarity criterion (e.g. square error function) which is tried to be minimized by assigning clusters to peaks in the probability density function, or the global structure [5]. Therefore, partitional clustering can be regarded as an optimization problem, which is also considered in this paper. In the usage of partitional clustering algorithms, the disadvantages of hierarchical algorithms are the advantages of partitional clustering algorithms, and vice versa [6].

Clustering can also be applied in two different modes: crisp (hard) and fuzzy (soft). Crisp clustering algorithms assume that each pattern should be assigned to only one cluster and the clusters are disjoint and non-overlapping. The most well-known example for the crisp clustering is the K-means algorithm. K-means [7] starts with K number of predefined clusters and then assigns each data member to its closest cluster. After the assignment, each cluster centroid is updated and this process is repeated until the termination criterion is satisfied. As in fuzzy clustering, a pattern may be assigned to all the clusters with a certain fuzzy membership function [2] (e.g. fuzzy C-means (FCM) [8]).

On account of the fact that K-means and FCM excessively depend on initial conditions, modifications have been proposed to improve the performance of the algorithms [9–12]. Moreover, evolutionary based clustering algorithms have been proposed in order to overcome local minima problem of these clustering approaches. Particle swarm optimization (PSO), proposed by Kennedy and Eberhart in 1995 [13,14], was applied to the clustering problems, and better performances were gained against K-means [15,16]. Omran and Al-Sharban [16] applied Baribones-PSO to image clustering problem. Wong et al. [17] proposed an improved version of the objective function, which was firstly proposed by Omran et al. [6]. Besides PSO, Hancer et al. [18,19] developed an artificial bee colony based brain tumour segmentation methodology from MRI images with the previously proposed objective function [6]. Ozturk et al. [20] determined the drawbacks of the objective functions in the literature and improved a new objective function satisfying well-separated and compact clusters. Moreover, the Ant colony optimization (ACO) was also applied to the clustering problem [21,22]. The detailed information can be found in [23–26].

It is clear that the number of clusters cannot be easily specified in many real world applications and datasets; therefore, the above mentioned algorithms requiring number of clusters as a parameter cannot be effectively employed. On behalf of these understanding, finding the “optimum” number of clusters in a data set has become an important research area. Proposed by Ball and Hall [27], ISODATA splits or merges clusters throughout the programme based on certain criteria in order to increase or decrease the number of clusters. However, ISODATA asks the user to specify the values of several parameters (e.g. the merging and splitting thresholds) and it can only merge two clusters under a user specified threshold [6]. Dynamic optimal cluster-peek (DYNOC) [28], similar to ISODATA, is based on maximizing the ratio of minimum inter-cluster distance to maximum intra cluster distance, but it also requires user specified parameters. Snob [29], Wallace’s programme for unsupervised classification of multivariate data, uses the minimum [message or description] length [encoding] (MML or MMD) principle to decide

upon the best classification of the data in order to assign objects to a cluster.

Evolutionary based algorithms have also been applied to the dynamic clustering problem, particularly in last decade. Omran et al. [4] proposed a PSO based dynamic image clustering (DCPSO), which was inspired by the ideas of Kuncheva and Bezdek [30]. In DCPSO, a cluster set (S) is first created and then binary PSO is applied to select cluster centroids from S . After that, the obtained cluster centroids from S within the best solution are refined by K-means. Das et al. [31,32] proposed differential evolution based algorithms (ACDE and AFDE) in which the parameters of F -scale and crossover rate are determined adaptively. In ACDE, each solution is represented by cluster centroids and associated activation values ([0,1]). Through evaluation, cluster centroids and their activators are updated simultaneously. Thus, it does not need to employ K-means to decrease the effects of initial conditions as in DCPSO. Kuo et al. [33] improved a hybrid PSO&GA algorithm to overcome the convergence problem of the PSO algorithm. However, the applied objective fitness function, based only on Euclidean distance, is not very convenient for dynamic clustering problem. Maulik and Saha [34] proposed a modified differential evolution clustering algorithm based on information of local and global best positions (MoDEAFC) to automatically extract information from remote sensing images. Rui et al. [35] employed DE and PSO sequentially for odd and even iterations and presented a comparative study on clustering validity indexes.

The main goal of this study is to demonstrate that the improved version of the discrete binary artificial colony algorithm (DisABC) [36] can be applied to the dynamic clustering problem. The novelty of the improved version of discrete artificial bee colony (referred as “IDisABC”) comprises two parts: modified random selection and modified greedy selection. These improved selection mechanisms are applied to search solution space intimately with the help of crossover and swap operators when the number of probable obtained outputs (M' vals) by dissimilarity calculation of two neighbourhood solutions is greater than one. In this way, the computational complexity of the algorithm is not affected so much. Moreover, the performance analysis and performance comparisons of the algorithms have been tested on benchmark problems in terms of the index quality, obtained number of cluster and correct classification percentage (CCP) by applying the static algorithms such as K-means and FCM in addition to the evolutionary computation based algorithms, including the DCPSO, GA and DisABC. It should be noticed that CCP is one of the most significant criterions to measure the quality of clustering; however, not many studies, especially related to dynamic clustering, reported the values of CCP.

The rest of the paper is organized as follows; Section 2 describes the ABC algorithm; Section 3 demonstrates the IDisABC algorithm; Section 4 defines the clustering problem; and Section 5 presents the comparative results of the state of the art algorithms with the proposed algorithm. Finally, Section 6 concludes the paper.

2. Artificial bee colony algorithm

A model of intelligent behaviours of honey bee swarm introduced by Karaboga in 2005 [37], the artificial bee colony (ABC) algorithm is a novel swarm intelligence based algorithm and has been applied to various problems such as in optimization of numerical problems [38], data clustering [39], neural networks training for pattern recognition [40], wireless sensor network deployment [41] and routing [42] and image analysis [19,43]. The ABC algorithm comprises of three phases: employed bee phase, onlooker bee phase and scout bee phase. In employed bee and onlooker bee phases, a new solution is produced in the neighbourhood of current solution via Eq. (1);

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات