Expert Systems with Applications 37 (2010) 4761–4767

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An artificial bee colony approach for clustering

Changsheng Zhang [a,*], Dantong Ouyang [b], Jiaxu Ning [c]

[a] College of Information Science & Engineering, Northeastern University, Shenyang 110819, PR China
[b] Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun 130012, PR China
[c] Institute of Grassland Science Northeast Normal University, PR China

ARTICLE INFO

Keywords:
Clustering
Meta-heuristic algorithm
Artificial bee colony
K-means

ABSTRACT

Clustering is a popular data analysis and data mining technique. In this paper, an artificial bee colony clustering algorithm is presented to optimally partition *N* objects into *K* clusters. The *Deb's* rules are used to direct the search direction of each candidate. This algorithm has been tested on several well-known real datasets and compared with other popular heuristics algorithm in clustering, such as GA, SA, TS, ACO and the recently proposed K–NM–PSO algorithm. The computational simulations reveal very encouraging results in terms of the quality of solution and the processing time required.

1. Introduction

Clustering is an important problem that must often be solved as a part of more complicated tasks in pattern recognition, image analysis and other fields of science and engineering. Clustering procedures partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some predefined criteria. The existing clustering algorithms can be simply classified into the following two categories: hierarchical clustering and partitional clustering (Sander, 2003.). Hierarchical clustering operates by partitioning the patterns into successively fewer structures. Since it is not the subject of this study we will not mention it in detail. Partitional clustering procedures typically start with the patterns partitioning into a number of clusters and divide the patterns by increasing the number of partitions. The most popular class of partitional clustering methods is the center-based clustering algorithms.

K-means has been used as a popular center-based clustering method due to its simplicity and efficiency, with linear time complexity. However, *K*-means has the shortcomings of depending on the initial state and converging to local minima (Selim & Ismail, 1984). In order to overcome these problems, many heuristic clustering algorithms have been introduced. A genetic algorithm based method to solve the clustering problem was proposed by Mualik and Bandyopadhyay (2002) and experimented on synthetic and real-life datasets to evaluate its performance. Krishna and Murty (1999) proposed a novel approach called genetic *K*-means algorithm for clustering analysis which defines a basic mutation operator specific to clustering called distance-based mutation. It has been proved that GKA converge to the best-known optimum through using the theory of finite Markov chain. A simulated annealing approach for solving the clustering problem is proposed by Selim and Al-Sultan (1991). The parameters of the algorithm were discussed in detail and it has been proved theoretically that a clustering problem's global solution can be reached. Sung and Jin (2000) proposed a tabu search based heuristic for clustering. Two complementary functional procedures, called packing and releasing procedures were combined with the tabu search.

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 clustering. An ant colony clustering algorithm for clustering is presented by Shelokar, Jayaraman, and Kulkarni (2004). The algorithm employs distributed agents who mimic the way real ants find a shortest path from their nest to food source and back. Its performance was compared with GA, tabu search, and SA. The particle swarm optimization which simulates bird flocking was used for clustering by Kao, Zahara, and Kao (2008). In order to improve its performance further, the PSO algorithm is hybridized with *K*-means and Nelder–Mead simplex search method. Its performance is compared with GA (Murthy & Chowdhury, 1996) and KGA (Bandyopadhyay & Maulik, 2002) algorithm.

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 (Teodorovic et al., 2006). Recently, Karaboga and Basturk (2008) have described an artificial bee colony (ABC) 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 genetic algorithm, differential evolutional algorithm and particle swarm

* Corresponding author. Tel.: +86 0431 85166487.
E-mail address: zcs820@yahoo.com.cn (D. Ouyang).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.11.003

optimization algorithm for unconstrained optimization problems. In this work, ABC algorithm is extended for solving clustering problems. The performance of the algorithm has been tested on a variety of data sets provided from several real-life situations and compared with several other proposed clustering algorithms. This paper is organized as follows. In Section 2, we discussed the clustering analysis problems. The ABC algorithm and the ABC algorithm adapted for solving clustering problems are introduced in Section 3. Section 4 will present experimental studies that show that our method outperforms some other methods. Finally, Section 5 summarizes the contribution of this paper along with some future research directions.

2. The clustering problem

Let $\boldsymbol{O} = \{o_1, o_2, \ldots, o_n\}$ be a set of n objects and let $\boldsymbol{X}_{n\times p}$ be the profile data matrix, with n rows and p columns. Each ith objects is characterized by a real-value p-dimensional profile vector $x_i (i = 1, \ldots, n)$, where each element x_{ij} corresponds to the jth real-value feature ($j = 1, \ldots, p$) of the ith object ($i = 1, \ldots, n$).

Given $X_{n\times p}$, the goal of a partitional clustering algorithm is to determine a partition $G = \{C_1, C_2, \ldots, C_k\}$ (*i.e.*, $C_g \neq \Phi,\ \forall g;\ C_g \cap C_h = \Phi,\ \forall g \neq h;\ \cup_{g=1}^{k} C_g = O$) such that objects which belong to the same cluster are as similar to each other as possible, while objects which belong to different clusters are as dissimilar as possible. For this, a measure of adequacy for the partition must be defined. A popular function used to quantify the goodness of a partition is the total within-cluster variance or the total mean-square quantization error (MSE) (Güngör & Ünler, 2007) which is defined as follows:

$$\text{Perf}\ (O, G) = \sum_{i=1}^{n} \text{Min}\left\{ \|o_i - C_l\|^2 | l = 1, \ldots k \right\} \tag{1}$$

Where $\|o_i - C_l\|$ denotes the similarity between object o_i and center C_l. The most used similarity metric in clustering procedure is Euclidean distance which is derived from the Minkowski metric.

$$d(x, y) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r} \Rightarrow d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{2}$$

In this study we will also use Euclidean metric as a distance metric. The clustering problem is to find the partition G^* that has optimal adequacy with respect to all other feasible solutions $G = \{G^1, G^2, \ldots, G^{N(n,k)}\}$ (*i.e.*, $G^i \neq G^j,\ i \neq j$) where

$$N(n, k) = \frac{1}{k!} \sum_{g=0}^{k} (-1)^g \binom{k}{g} (k - g)^n$$

It is the number of all feasible partitions. It has been shown that the clustering problem is NP-hard when the number of clusters exceeds three (Brucker, 1978).

3. Artificial bee colony based clustering

3.1. Honey bee modeling (Karaboga & Basturk, 2008)

The minimal model of forage selection that leads to the emergence of social intelligence of honey bee swarms consists of three essential components: food sources, employed foragers and unemployed foragers, and two leading modes of the behavior, recruitment to a nectar source and abandonment of a source, are defined (Karaboga, 2005). A food source value depends on many factors, such as its proximity to the nest, richness or concentration of energy and the ease of extracting this energy. The employed foragers are associated with particular food sources, which they are currently exploiting or are "employed". They carry with them information about these food sources and share this information with a certain probability. There are two types of unemployed foragers, scouts and onlookers. Scouts search the environment surrounding the nest for new food sources, and onlookers wait in the nest and find a food source through the information shared by employed foragers.

In ABC algorithm (Basturk & Karaboga, 2006; Karaboga & Basturk, 2008), the colony of artificial bees consists of three groups of bees: employed bees, onlookers and scouts. A food source represents a possible solution to the problem to be optimized. The nectar amount of a food source corresponds to the quality of the solution represented by that food source. For every food source, there is only one employed bee. In other words, the number of employed bees is equal to the number of food sources around the hive. The employed bee whose food source has been abandoned by the bees becomes a scout.

As other social foragers, bees search for food sources in a way that maximizes the ration E/T where E is the energy obtained and T is the time spent for foraging. In the case of artificial bee swarms, E is proportional to the nectar amount of food sources discovered by bees. In a maximization problem, the goal is to find the maximum of the objective function $F(\theta), \theta \in R^P$. Assume that θ_i is the position of the ith food source; $F(\theta_i)$ represents the nectar amount of the food source located at θ_i and is proportional to the energy $E(\theta_i)$. Let $P(c) = \{\theta_i(c) | i = 1, 2, \ldots, S\}$ (c: cycle, S: number of food sources being visited by bees) represent the population of food sources being visited by bees.

As mentioned above, the preference of a food source by an onlooker depends on the nectar amount $F(\theta)$ of that food source. As the nectar amount of the food source increases, the probability with the preferred source by an onlooker bee increases proportionally. Therefore, the probability with the food source located at θ_i will be chosen by a bee can be calculated as

$$P_i = \frac{F(\theta_i)}{\sum_{k=1}^{S} F(\theta_k)} \tag{3}$$

After watching the dances of employed bees, an onlooker bee goes to the region of food source located at θ_i by this probability and determines a neighbor food source to take its nectar depending on some visual information, such as signs existing on the patches. In other words, the onlooker bee selects one of the food sources after making a comparison among the food sources around θ_i. The position of the selected neighbor food source can be calculated as $\theta_i(c + 1) = \theta_i(c) \pm \phi_i(c)$. $\phi_i(c)$ is a randomly produced step to find a food source with more nectar around θ_i. $\phi(c)$ is calculated by taking the difference of the same parts of $\theta_i(c)$ and $\theta_k(c)$ (k is a randomly produced index) food positions. If the nectar amount $F(\theta_i(c + 1))$ at $\theta_i(c + 1)$ is higher than that at $\theta_i(c)$, then the bee goes to the hive and share her information with others and the position $\theta_i(c)$ of the food source is changed to be $\theta_i(c + 1)$, otherwise $\theta_i(c)$ is kept as it is.

Every food source has only one employed bee. Therefore, the number of employed bees is equal to the number of food sources. If the position θ_i of the food source i cannot be improved through the predetermined number of trials "*limit*", then that food source θ_i is abandoned by its employed bee and then the employed bee becomes a scout. The scout starts to search a new food source, and after finding a new source, the new position is accepted to be θ_i. Every bee colony has scouts that are the colony's explorers. The explorers do not have any guidance while looking for food. They are primarily concerned with finding any kind of food source. As a result of such behavior, the scouts are characterized by low search costs and a low average in food source quality. Occasionally, the scouts can accidentally discover rich, entirely unknown food sources. In the case of artificial bees, the artificial scouts could have the fast discovery of the group of feasible solutions as a task.