This article can be cited as S. M. H. Hasheminejad, M. Vosoughian and M. Zamini, AB2C: Artificial Bee Colony for Clustering, International Journal of Artificial Intelligence, vol. 18, no. 2, pp. 51-73, 2020.

AB2C: Artificial Bee Colony for Clustering

Seyed. Mohammad. Hossein. Hasheminejad[1], Marziyeh. Vosoughian[1], Mohamad. Zamini[2],
[1]Department of Computer Engineering, Alzahra University, Tehran
{SMH.Hashemienjad, M.Vosoghian}@alzahra.ac.ir
[2]Department of Information technology, Tarbiat Modares University, Tehran
M.zamini@modares.ac.ir

Abstract

Clustering is one of the most challenging unsupervised techniques. Methods that are used for clustering, are not able to meet all the needs of issues simultaneously. Clustering large datasets with many features are difficult which provide high computational complexity. Artificial bee colony algorithm (ABC) is an efficient, effective and robust technique in clustering with fewer control parameters in swarm intelligence, which emulates the foraging manner of honeybees. In this paper, the aim is to find an algorithm with suitable iteration to achieve the optimal solution. Therefore, the improved artificial bee colony algorithm, called AB2C (Artificial Bee Colony for Clustering), to converge faster has used the best solution obtained by bees to update their population and K-means algorithm to initialize the population as guided. The proposed AB2C is applied to some UCI data sets and comparing the obtained results with CPSOII and other clustering techniques such as CPSOI, CGA, and other ABC-based clustering reveals that the proposed AB2C yields promising results.

Keywords: Clustering, bee colony algorithm, evolutionary algorithms, meta-heuristic algorithms

Computing Classification System: G.1.6, I.5.2, G.3, I.2.8, I.2.1

1. INTRODUCTION

Clustering algorithms are aimed to split the dataset into various groups with the same properties of the underlying structure of data. An unlabeled dataset is given and tried to make some groups with similar properties or criterion(Li and Tang, 2018). Therefore, the output groups are called partition, where all the input samples are labeled with its corresponding cluster label.

The term swarm is used for a colony of animals like ants and bees which do social manner in a self-organized and decentralized way(Karaboga et al., 2014). Each agent in these swarms has a stochastic behavior without supervision. The need for effective optimization algorithms with the ability to solve complex real-world optimization problems resulted in the development of many evolutionary algorithms (EAs) like particle swarm optimization (PSO) (Kennedy and Eberhart, 1995), differential evolution (DE) (Storn and Price, 1997), genetic algorithm (GAs) (Holland, 1992), ant colony optimization (ACO) (Dorigo and Gambardella, 1997) and artificial bee colony (ABC) algorithm (Karaboga and Basturk, 2007) which are based on the Darwinian theory of evolution. Evolutionary algorithms which is used as a solution, rely on a hypothesis that newer generations of a species due to suitable mutations, crossover and subsequent selection will be usually of a better quality than their ancestors. However, in spite of many successful technical applications genetic algorithms show some lacks as it is a theory and it

needs a lot of corrections from the biological point of view. Also, generating new populations does not guarantee automatically also convergence to a solution. Lastly, only a part of knowledge will be advanced from parents to their descendants by the inheritance process(Vaščák, 2012).

The main goal of evolutionary algorithms is to avoid local optima trap real-world optimization problems. The reason we want to optimize EAs performance is to better exploitation of some multi-dimensional and complex optimization issues.

ABC algorithm was first developed by (Karaboga, 2005) for global optimization problems and since then it has been used widely in different fields like neural networks(Badem et al., 2017), clustering(Kuo and Zulvia, 2018, Ilango et al., 2018, Banharnsakun, 2017, Zhang et al., 2010, Saoud, 2019), image processing(Mondal et al., 2018, Ansari et al., 2017), numerical function optimization(Cui et al., 2017), scheduling problems(Sundar et al., 2017, Xue et al., 2017), self-adaptive ABC algorithms (Xue et al., 2017, Karaboga and Kaya, 2016), sensor networks (Mann and Singh, 2017), data mining (Karaboga and Ozturk, 2011, Hancer et al., 2018) and also protein structure optimization(Li et al., 2017, Zheng et al., 2017). To increase the exploitative tendency of ABC in a solution-level to memorize the entire solutions, Kiran and Babalik (Kiran and Babalik, 2014) added a memory board which saves the solutions whose qualities are better than the average fitness value. A self-adaptive modification rate (MR) on the basis of successful update probability to generate suitable parameters was proposed to further enhance the convergence rate of the proposed algorithm. (Masoud et al., 2013) presented Combinatorial Particle Swarm Optimization(CPSOII) for dynamic clustering, which finds the best number of clusters and categorizes data objects simultaneously. In this paper, for better clustering data objects of previous paper, we have proposed an improved bee colony clustering. CPSOII is the improved version of CPSO which needed to determine the number of clusters in advance. Zhu and Kwong (Zhu and Kwong, 2010) proposed a gbest-guided ABC (GABC) to better the search efficiency by applying the information of global best (gbest) solution. To improve exploitation of ABC Gao and Liu (Gao and Liu, 2012) implemented ABC/best/1 search equation with a new random initialization method.

In recent years, new versions of the ABC algorithm for various applications, are provided. (Ozturk et al., 2015) ABC algorithm with discrete GA operators is used for dynamic clustering. (Karaboga and Gorkemli, 2014) algorithm qABC with regard to the neighborhood for each bee is introduced, the position of each bee on the basis of its neighbors will be updated. This algorithm was to imitate the real onlooker bees' behavior.

In (Zhang et al., 2010), the ABC algorithm is used for data clustering. In this study, the centroids of clusters are only identified and the intra-cluster distance between identified clusters is evaluated. In (Yan et al., 2012) crossover operator GA in the ABC algorithm to exchange food source information between bees is used. (Ozturk et al., 2015) has improved ABC algorithm by introducing a new objective function to determine the suitability of food sources. In another work proposed by (Karaboga and Kaya, 2016) an Adaptive and Hybrid Artificial Bee Colony (aABC) algorithm to train ANFIS by utilizing arithmetic crossover rate and adaptivity

coefficient. Results showed by using an arithmetic crossover, aABC algorithm gains the rapid convergence feature. In (Saoud, 2019), the ABC algorithm is used for hierarchical network clustering. The goal of this study is to find community structure in real world networks.

In the proposed algorithm, called AB2C (Artificial Bee Colony for Clustering), to faster converge has been used of the best answer to update the population, and K-means algorithm to initialize the population as directed. The proposed AB2C is applied to some UCI data sets and the results are presented.

2. CLUSTERING

In the clustering, there is no label on the dataset for training to understand the patterns, hence clustering is more difficult than supervised classification methods(Zamini and Hasheminejad, 2019). In fact, in supervised learning labels are like a clue to grouping data objects as a whole but in the case of clustering, deciding whether an instance belongs to a group or not is difficult. There are several parameters or features, which should be considered for clustering. Furthermore, the curse of dimensionality should be added to the consideration. Another important factor is high dimensionality that leads to high computational cost and consistency of the algorithms. One solution for these problems is using feature selection methods (Saxena et al., 2010). Jain (Jain, 2010) depicted that the principal goal of clustering is to achieve natural grouping of a set of instances, points, or objects based on some inherent similarity. The measure of distance between patterns is a common approach to define similarity. The lower the distance, the higher similarity between two instances. The distance measures will be defined in table 1. Also Fig. 1 the illustrated the classification of clustering approaches(Fraley and Raftery, 1998).

According to (Kaufman and Rousseeuw, 2009), in data mining and statistics there are two main clustering approaches called, hierarchical and partitioning. Hierarchical clustering is an analysis method, which tries to find a hierarchy of clusters and partitioning is the assumption that clusters are exclusive groups of data characterized by the small within-cluster. For hierarchical clustering generally, it is divided into top-down (divisive) and bottom-up (agglomerative) approach. Furthermore, (Li and Tang, 2018) suggested the following four additional procedure of clustering assumptions: Density-based methods, Spectral clustering algorithms, Subspace clustering algorithms, affinity propagation algorithms. In the assumption of density-based clustering methods like FLAME (Fu and Medico, 2007), OPTICS (Ankerst et al., 1999) and DBSCAN (Ester et al., 1996) are sets of density-connected points separated by low-density areas. According to (Von Luxburg, 2007) graph cut technologies are used in spectral clustering and the assumption is that clusters are joint components on similarity graphs fitting data. In Subspace clustering algorithms (Vidal, 2011) assume that clusters have multiple subspaces with low-dimension. In the affinity propagation algorithm (Frey and Dueck, 2007) the assumption is that each cluster has its exclusive instance, which can be found through the procedure of passing iterative messages.