International Journal of Applied Metaheuristic Computing
Volume 13 • Issue 1

A Modified Cuckoo Search Algorithm for Data Clustering

Preeti Pragyan Mohanty, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, India*
Subrat Kumar Nayak, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, India

ABSTRACT

Clustering of data is one of the necessary data mining techniques, where similar objects are grouped in the same cluster. In recent years, many nature-inspired clustering techniques have been proposed, which have led to some encouraging results. This paper proposes a modified cuckoo search (MoCS) algorithm. In this work, an attempt has been made to balance the exploration of the cuckoo search (CS) algorithm and to increase the potential of the exploration to avoid premature convergence. This algorithm is tested using 15 benchmark test functions and is proven as an efficient algorithm in comparison to the CS algorithm. Further, this method is compared with well-known nature-inspired algorithms such as ant colony optimization (ACO), artificial bee colony (ABC), particle swarm optimization (PSO), particle swarm optimization with age group topology (PSOAG), and CS algorithm for clustering of data using six real datasets. The experimental results indicate that the MoCS algorithm achieves better results as compared to other algorithms in finding optimal cluster centers.

KEYWORDS

Cuckoo Search Algorithm, Data Clustering, Intra-Cluster Distance

1. INTRODUCTION

Clustering is a method of grouping an enormous amount of data into different groups. The data in the same group exhibit similar properties, while the data in different groups exhibit different properties. This technique is used for finding patterns among the data in each group. Over the past few years, clustering has played a key role in various fields of research such as image analysis, machine learning, data mining, pattern recognition, information retrieval, statistics, biology, medical sciences, market research, etc. (Ahalya & Pandey, 2015).

The traditional clustering algorithms are mostly classified into two prime types, i.e., hierarchical clustering and partitional clustering (Leung et al., 2000; Xu & Tian, 2015). The outcome of the hierarchical clustering approach is a tree-like structure representing the clustering process, where the dataset is partitioned into different groups. In this type of clustering, the data objects present in one group cannot be reassigned to another group (Armano & Farmani, 2016). Moreover, this clustering can be performed even if the number of groups is not known. The major disadvantage of this technique is that it fails to separate the overlapping groups as the information about the shape and size of groups

DOI: 10.4018/IJAMC.2022010101 *Corresponding Author

is not known. The partitional clustering approach partitions the dataset into a set of groups such that data present in each group are different. In partitional clustering, either the number of clusters is known prior to the partitioning or it is unknown and needs to be predicted for an unknown dataset. This paper deals with the partitional clustering problem, where the number of clusters of a dataset is known beforehand. The partitional clustering approach aims to optimize some dissimilarity criteria such as minimizing the intra-cluster distance between the objects in one group and maximizing the inter-cluster distance between different groups.

Generally, partitional clustering algorithms consist of centroid-based algorithms. One of the popular centroid-based algorithms is the *k*-means algorithm proposed by Stuart Lloyd in 1957 (Lloyd, 1982). The main aim of the *k*-means clustering algorithm is to partition the objects into *k* groups by randomly choosing *k* number of data objects as initial centroids. Though this algorithm is easy to implement, still it gets stuck at the local minimum as it is sensitive to the initial position of the groups.

The nature-inspired metaheuristic algorithms such as Firefly Algorithm (FA) (Yang, 2008), Ant Colony Optimization (ACO) (Dorigo, 1992), Cuckoo Search (CS) (Yang & Deb, 2009), Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995), Glowworm Swarm Optimization (GSO) (Krishnanand & Ghose, 2005), Artificial Bee Colony (ABC) (Karaboga et al., 2005), etc. have been able to improve upon the disadvantages of *k*-means algorithm. These algorithms are used to solve optimization problems that do not have a specific satisfactory solution and generate near-optimal results (Nesmachnow, 2014). In these algorithms, a population of candidate solutions is randomly generated and the best population for the next generation is selected based on some fitness values. These algorithms simultaneously optimize these candidate solutions to generate a globally optimized solution (Jiang et al., 2013).

Many researchers have proposed many variations of nature-inspired metaheuristic algorithms for clustering, which have resulted in inefficient solutions. It is impossible to cover all the algorithms proposed for clustering. In this paper, we confine our literature to some of the swarm intelligence algorithms proposed for clustering. PSO is a swarm intelligence algorithm inspired by a flocking of birds or swarm of fish. In this algorithm, the particles move towards their best previous positions and towards the best particle in order to achieve the global optimal solution. The PSO based clustering approach proposed by Cura (2012) follows this technique and achieves better outcomes on the basis of objective function values, error rate and computation time when compared to other swarm-intelligence techniques. The Ant Colony Optimization (ACO) for clustering the data into *k* clusters was proposed by Shelokar et al. (2004). This method involves distributed agents that imitate the way real ants go for searching their food from their nest. This algorithm was tested on different datasets yielding better computational solutions in less time. Zhang et al. (2010) used the ABC algorithm for clustering. This algorithm involves employed bees, onlooker bees and scout bees who get involved in the process of searching for their food. In this paper, Deb's method is adapted for selecting a food source instead of the greedy search process for selecting the food source. This method gives encouraging results when tested with different datasets. A recently developed algorithm and a variation of the PSO algorithm, called Particle Swarm Optimization with Age Group topology (PSOAG) is used for clustering of data (Jiang et al., 2013). In this method, an approach for maintaining population diversity is defined. The PSOAG algorithm proves as an effective algorithm for data clustering as it has better intra-cluster distance, better clustering accuracy and low computation time as compared to PSO, ACO, ABC, and DE algorithm.

Saida et al. (2014) used the CS algorithm for clustering of data. CS algorithm is based on an aggressive breeding strategy of cuckoos laying their eggs in the nest of other birds. The cuckoo birds go for searching their food with the help of a random walk method called the Levy flight approach. The Levy flight approach is a random walk in step lengths with a high tailed probability distribution. Also, these cuckoo eggs when discovered by the host birds are either thrown away or the nest is abandoned depending on a certain probability. Although the CS algorithm is good at exploring the search space, still it is slow in exploiting the solutions (Long et al., 2014). Hence, a Modified Cuckoo Search (MoCS)