Applied Soft Computing 41 (2016) 15–21

Contents lists available at ScienceDirect

Applied Soft Computing

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

Efficient protocol for data clustering by fuzzy Cuckoo Optimization Algorithm

Ehsan Amiri [a,*], Shadi Mahmoudi [b]

[a] Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran
[b] Computer Engineering Department, Boukan Branch, Islamic Azad University, Boukan, Iran

ARTICLE INFO

Article history:
Received 25 April 2015
Received in revised form
22 November 2015
Accepted 7 December 2015
Available online 30 December 2015

Keywords:
Data clustering
Cuckoo Optimization Algorithm (COA)
Dataset
Fuzzy logic
Artificial intelligence

ABSTRACT

Data clustering is a technique for grouping similar and dissimilar data. Many clustering algorithms fail when dealing with multi-dimensional data. This paper introduces efficient methods for data clustering by Cuckoo Optimization Algorithm; called COAC and Fuzzy Cuckoo Optimization Algorithm, called FCOAC. The COA by inspire of cuckoo bird nature life tries to solve continuous problems. This algorithm clusters a large dataset to prior determined clusters numbers by this meta-heuristic algorithm and optimal the results by fuzzy logic. Firstly, the algorithm generates a random solutions equal to cuckoo population and with length dataset objects and with a cost function calculates the cost of each solution. Finally, fuzzy logic tries for the optimal solution. The performance of our algorithm is evaluated and compared with COAC, Black hole, CS, *K*-mean, PSO and GSA. The results show that our algorithm has better performance in comparison with them.

1. Introduction

Clustering is a main task in data analysis applications. Data clustering technique is introduced for grouping similar or dissimilar data in a given dataset. Grouping, decision-making, machine-learning situations, data mining, document retrieval, image segmentation and pattern classification are some important applications of clustering techniques [1]. Research in exact algorithms, heuristics and meta-heuristics for solving combinatorial optimization problems is increasingly relevant as data science grapples with larger and more complex data sets [2].

Hierarchical and partitional as mentioned in [3,4] are two categories of traditional clustering algorithms. In hierarchical clustering, the number of clusters need not be specified a prior but by partitional methods it should be determined. As a result, hierarchical methods cannot always separate overlapping clusters. Additionally, hierarchical clustering is static and points committed to a given cluster in the early stages cannot be moved between clusters [5]. While the partitional clustering divide in two categories: crisp clustering where each data point belongs to just one cluster and fuzzy clustering where every data point belongs to every cluster to some degree [6,7].

Xin-She and Deb in 2009 [8] proposed a Cuckoo Search (CS) algorithm. In 2011, Rajabioun [9] improved this algorithm and introduced a meta-heuristic algorithm called Cuckoo Optimization Algorithm (COA). COA has some priorities toward CS such as faster convergence, higher speed, high accuracy, local search ability along with general search, search with variable population (population extinction due to poor areas), ability to quickly solve optimization problems with high dimensions and so on. Fuzzy sets a theory that originally introduced by Zadeh [10,11] and it has been developed for expanded linguistic values. The linguistic values are terms which are used instead of numbers and fuzzy set theory [12]. Implementation of multi-criteria control strategies is enabling by rule based FLCs usage. When information is not complete, fuzzy logic is able to make real time decisions. Fuzzy logic systems are able to manipulate the linguistic rules in a natural way and they are particularly suitable in several context like clustering applications. This paper presents a new optimization approach to data clustering based on the COA and Fuzzy COA for decreasing clustering error rate. In fact, the dividing of clusters is dynamically regulated by a Fuzzy Logic Controller (FLC).

The rest of this paper is organized as follows: in Section 2, we have a brief discussion related with works on data clustering. Section 3, shows COA and we have a full description of it. Our work is introduced in Section 4 with some samples for more realization.

* Corresponding author. Tel.: +98 9384119312.
E-mail addresses: amireehsan@yahoo.com (E. Amiri), mahmoudi.Shadi@Yahoo.Com (S. Mahmoudi).

http://dx.doi.org/10.1016/j.asoc.2015.12.008
1568-4946/

Also, the performance of our proposal is evaluated with several benchmark datasets and the results are compared with some well-known works. Finally, Section 5 summarize and conclude our work.

2. Related works

Nowadays, many scientists work on data categoring in the clusters with different manners such as meta-heuristic algorithms that they are mostly used for solving optimization problems. In [13], the authors proposed an artificial bee colony clustering algorithm to optimally partition N objects into K clusters. Fathian et al. [14] proposed applications of honey bee mating optimization in clustering (HBMK-means). In [15], authors introduce a new hybrid algorithmic nature inspired approach based on the concepts of the Honey Bees Mating Optimization Algorithm (HBMO) and of the Greedy Randomized Adaptive Search Procedure (GRASP), for optimally clustering N objects into K clusters.

Genetic algorithms (GAs) [16–18] are another meta-heuristic methods. Bai [19] implemented a Master–Slave parallel genetic algorithm (PGA) with a Marsili and Giada log-likelihood fitness function to identify clusters within stock correlation matrices. Also, utilized the Nvidia Compute Unified Device Architecture (CUDA) programming model to implement the PGA and visualize the results using minimal spanning trees (MSTs). In [20], researcher proposed a new parameter estimation approach for the mixture normal distribution. The developed model estimates parameters of the mixture normal distribution by maximizing the log likelihood function using genetic algorithm (GA).

Ng et al. [21], presented a tabu search [22,23] based clustering algorithm to extend the K-means paradigm in order to categorical domains and domains with both numeric and categorical values. By using tabu search based techniques, their algorithm can explore the solution space beyond the local optimality aims to find a global solution of the fuzzy clustering problem.

Ant colony optimization (ACO) firstly proposed by Dorigo [24,25]. Shelokar et al. [26] presented an ant colony optimization methodology for optimally clustering N objects into K clusters. In [5], the authors proposed a novel algorithm called ant colony optimization with different favor (ACODF) for data clustering. Li et al. [27] proposed a two-stage framework for gene selection so that the modified ant system and improved ant colony system are used by the fuzzy logic control.

Lee and Geem [28] described a new Harmony Search (HS) meta-heuristic algorithm-based approach for engineering optimization problems with continuous design variables. Mahdavi and Abolhassani [29] proposed a novel Harmony K-means Algorithm (HKA) that deals with document clustering based on Harmony Search (HS) optimization method. Ni et al. [30] improved the PSO with a ring topology.

Hatamlou [31] proposed a new algorithm for data clustering by Black hole phenomena. The particle swarm optimization (PSO) algorithm was developed based on the swarm behavior, such as fish and bird schooling in nature [32]. The gravitational search algorithm (GSA) was constructed based on the law of gravity and the notion of mass interactions. In the GSA algorithm, the searcher agents are a collection of masses that interact with each other based on the Newtonian gravity and the laws of motion [33].

Saida et al. [34] presented a new algorithm for data clustering based on the cuckoo search optimization called CS. Cuckoo search is generic and robust for many optimization problems and it has attractive features like easy implementation, stable convergence characteristic and good computational efficiency.

Pei Honga et al. used of the general GGA representation and operators to reduce redundancy in the chromosome representation for attribute clustering. They compared the efficiency of the proposed approach with that of an existing approach [35].

Ozturk et al. [36] improved version of the discrete binary artificial colony algorithm (DisABC) and applied to the dynamic clustering problem. 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 some well-known computation based algorithms.

3. Cuckoo Optimization Algorithm

The Cuckoo Optimization Algorithm is based on the obligate brood parasitic behavior of some cuckoo species with the levy flight behavior instead of isotropic simple randomized hiking. These birds called brood parasite because of they even make nests and lays their eggs in the other host bird's nest. The cuckoo just should find a nest with the most similar eggs with its egg. It throws out one of the host bird eggs and lays its egg. Sometimes, the host bird detected cuckoo's egg and this time host bird throw out that.

The cuckoos look for the most suitable area to lay their eggs to maximize the survival rate [37]. when after the cuckoo chicks grow, they make a group and their society. Each group has its area. The best area is destination for other groups and they migrate to this area. Each group resides in an area nearest to the current best area. The egg laying radius calculates by considering the number of eggs each cuckoo will lay and also destination of each cuckoo has to current best area. Then, it lays eggs in some random nests inside its radius area. This process continues until the best position with maximum profit value is obtained and most of the cuckoos population is gathered around the same position [37].

The habitat array uses for keeping input variable values that these variables have floating point values. Eq. (1) shows the habitat array:

$$\text{habitat} = [x_1, x_2, x_3, \ldots, x_{N_{var}}] \tag{1}$$

The profit of a habitat is obtained by evaluation of profit f_p function at a habitat and Eq. (2) shows it:

$$\text{Profit} = f_p(\text{habitat}) = f_p(x_1, x_2, x_3, \ldots, x_{N_{var}}) \tag{2}$$

After that, for beginning the optimization, the algorithm generates a habitat matrix with $N_{var} * N_{pop}$ size and in each of these habits lay random egg number. In nature, each cuckoo lays between 5 and 20 eggs. These numbers are used as lower and upper limits for each cuckoo in each iteration. They also lay eggs within a maximum distance from their habits. This maximum range is called Egg Laying Radius (ELR). Eq. (3) shows this formula in which $\propto$ is an integer value for regularizing the maximum value of ELR, var_{hi} and var_{low} use in order to high limit and low limit.

$$ELR = \propto \times \frac{\text{The number of current cuckoo's eggs}}{\text{Total number of eggs}} \times (var_{hi} - var_{low}) \tag{3}$$

when the cuckoo's become mature, they stay in their groups and habits till the egg-laying time arrives. They migrate to new habits with more similarity of egg to the host birds. After that, the cuckoo groups are formed in a different area, the best society with maximum profit identifies and marks in a goal area for migrate other cuckoos.

The COA uses K-mean clustering technique for grouping cuckoos, after that the mean profit value determines for each group. Then the group with maximum profit finds and spots for new cuckoo's destination. This is obvious in this immigration some cuckoos have a deviation. Fig. 1 depicts this immigration. As this figure shows, after that goal area is determined, the cuckoos in another