Swarm and Evolutionary Computation 10 (2013) 1–11

Contents lists available at SciVerse ScienceDirect

Swarm and Evolutionary Computation

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

Regular Paper

A robust clustering method based on blind, naked mole-rats (BNMR) algorithm

Mohammad Taherdangkoo [a,*], Mohammad Hossein Shirzadi [b], Mehran Yazdi [a], Mohammad Hadi Bagheri [c]

[a] Department of Communications and Electronics, Faculty of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
[b] Department of Industrial Engineering, University of Tehran, Tehran, Iran
[c] Center for Evidence-Based Imaging, Department of Radiology, Brigham & Women's Hospital, Harvard Medical School, Brookline, MA, USA

ARTICLE INFO

Article history:
Received 1 March 2012
Received in revised form
28 December 2012
Accepted 1 January 2013
Available online 14 January 2013

Keywords:
Data clustering
Meta-heuristic algorithm
K-means algorithm
BNMR algorithm

ABSTRACT

One of most popular data clustering algorithms is *K*-means algorithm that uses the distance criterion for measuring the correlation among data. To do that, we should know in advance the number of classes (*K*) and choose *K* data point as an initial set to run the algorithm. However, the choice of initial points is a main problem in this algorithm, which may cause that the algorithm converges to local optima. So, some other clustering algorithms have been proposed to overcome this problem such as the methods based on *K*-means (SBKM), Genetic Algorithm (GAPS and VGAPS), Particle Swarm Optimization (PSO), Ant Colony Optimization (Dynamic ants), Simulated Annealing (SA) and Artificial Bee Colony (ABC) algorithm. In this paper, we employ a new meta-heuristic algorithm. We called it blind, naked mole-rats (BNMR) algorithm, for data clustering. The algorithm was inspired by social behavior of the blind, naked mole-rats colony in searching the food and protecting the colony against invasions. We developed a new data clustering based on this algorithm, which has the advantages such as high speed of convergence. The experimental results obtained by using the new algorithm on different well-known datasets compared with those obtained using other mentioned methods showed the better accuracy and high speed of the new algorithm.

1. Introduction

Data Clustering is considered as an important issue and an essential pre-step for many fields such as data mining [1,13], math programming [18], scientific analysis, [5] and image segmentation [16,22]. Data clustering can provide a suitable way toward an ideal solution or even many lead directly to it. Data clustering aims at dividing a dataset into some classes without knowing any pre-information about the kind of relations exists between classes. There are many ways to cluster a dataset. One of the popular algorithms is the *K*-mean algorithm [9]. The algorithm tries to put the entire Dataset (i.e. *S*) into *K* clusters (i.e. $C_1,C_2,....,C_K$) with randomly selecting *K* data points (*K* data points as a set of primary centers). To do that, the clusters are formed such that the existing data in each cluster should have the minimum Euclidean distance to the center of that cluster. Hereby, following conditions should be satisfied:

$$\bigcup_{i=1}^{K} C_i = S, \tag{1}$$

where *S* is the entire dataset. Moreover, there should be at least one point in each cluster, such that:

$$C_k|_{k=1,\ldots,K} \neq \Phi, \tag{2}$$

where Φ is an empty set. Final condition is that there should not be any data point jointly existing in two different clusters, which can be expressed as follow:

$$C_k \cap C_j = \Phi|_{k \neq j}. \tag{3}$$

By settling these conditions, total Euclidean distance between the data points and cluster's centers is iteratively minimized such that the center of each cluster becomes the best represent of that cluster. Total Euclidean distance is defined as:

$$E = \sum_{j=1}^{K} \sum_{x_i \in C_j}^{N_j} \|x_i - Z_j\|, \tag{4}$$

where x_i is *i*th data point and belongs to the cluster C_j, Z_j is the center of C_j, *K* is the number of clusters and N_j is the number of

Abbreviations: GA, genetic algorithm; PSO, particle swarm optimization; ACO, ant colony optimization; SA, simulated annealing; ABC, artificial bee colony algorithm; BNMR, blind, naked mole-rats algorithm

* Corresponding author. Tel.: +98 9128243251.
E-mail addresses: mtaherdangkoo@yahoo.com, taherdangkoo@shirazu.ac.ir (M. Taherdangkoo), mh.shirzadi@ut.ac.ir (M. Hossein Shirzadi), yazdi@shirazu.ac.ir (M. Yazdi), mbagheri@partners.org (M. Hadi Bagheri).

2210-6502/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.swevo.2013.01.001

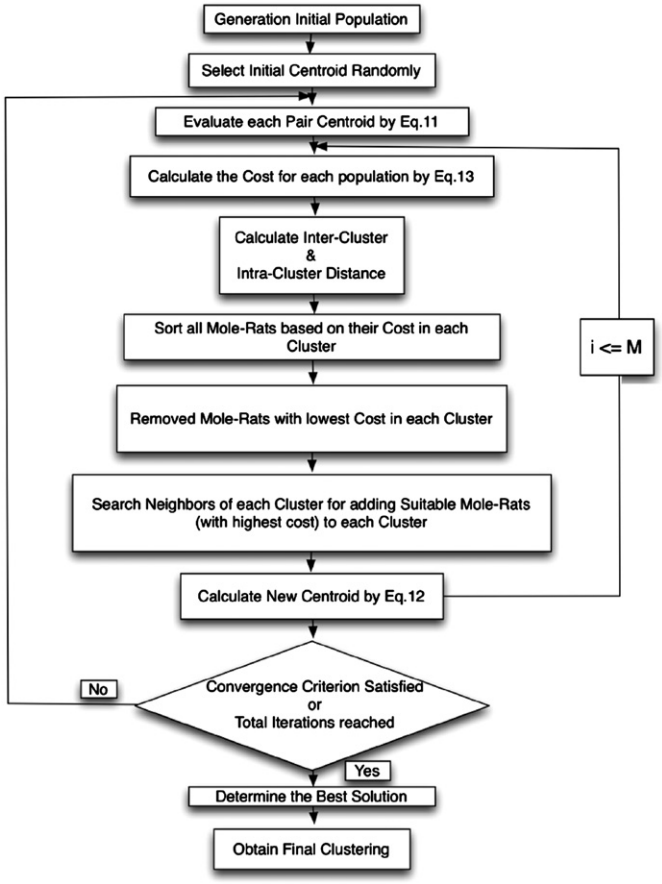

Fig. 1. Flowchart of BNMR algorithm for data clustering.

Table 1
The Characteristics of datasets.

Dataset Name	Number of objects (*N*)	Number of features (*d*)	Number of classes (*K*)
Vowel	871	3	6
Iris	150	4	3
Crude oil	56	5	3
Control chart	1500	60	6
Wood defects	232	17	13
Wine	178	13	3

data points in C_j. Selection of K data points as the initial centers of the clusters affects significantly the performance of the algorithm. Another fundamental problem in K-means algorithm is that if the entire dataset is large, the risk of convergence to local minimums will be decreased and eventually the response obtained after several repeats will not be the optimal response. To overcome these drawbacks, many clustering algorithms have been recently introduced [7,8,11,19,21]. For instance, [11] proposed K-Medoid algorithm that uses a different distance criterion to determine the best represent of each cluster. Other proposed methods used certain optimization algorithms in order to achieve optimal response (i.e. correctly classified data points). For instance, one of these optimization algorithms is the GA that has been widely employed for data clustering [2,3,10,12]. In [10], the responses were expressed as a string of bits and the algorithm starts using a population consisting of a series of initial responses and obtained ideal responses using the GA and continued its process until achieving a convergence to a global minimum. It should be noted that the optimization of the K-means algorithm using GA increases only the speed of reaching at the final response and improves a little the performance of K-means algorithm, although it solves somehow the problem of local minimum convergence.

Data clustering approach has been also extended by the algorithms based on the social behavior of ants known as ACO [6,20,23]. For instance, in [20] an algorithm was implemented using the increase of pheromone evaporation rate on data closer to the clusters' centers and achieved responses far better than previous algorithms. Recently, some researchers have used the PSO algorithm for data clustering [4,14,17]. For instance, in [14] data clustering was achieved by optimizing K-means algorithm based on PSO. The method demonstrated a performance much better than previous algorithms. In this method, k-means algorithm was used only in forming initial population of particles and then PSO classifies data points using the idea of K-means algorithm.

More recently, [24] has introduced a clustering algorithm using ABC algorithm that showed better performance than K-means algorithm. It proposed to avoid from local minimums using bee algorithm as an optimization procedure, however it took many iterations to achieve the optimal response and the entire process was very long. In this paper, we propose a robust optimization algorithm named BNMR algorithm to classify various types of dataset. By using this optimization algorithm in our proposed scheme for data clustering, we have clustered well-known datasets with the high speed and more precision than other previous well-known data clustering methods.

2. Blind, naked mole-rats (BNMR) algorithm

The BNMR algorithm was introduced in 2012 for the first time as an optimization method for numerical function optimization [27]. The algorithm designed based on the social behavior of blind naked mole-rats in a large colony. The research process starts from the center of the colony, where the queen and offspring live. Note that for simplification purposes, we have placed the employed moles and soldier moles in one single group, which is called the employed moles.

In the beginning with the production of the initial population of the blind naked mole-rats colony starts working in the whole problem space on a completely random way. Note that the population size is two times of the number of food sources and each of the food sources represents a response in the problem space. Let us define some parameters as follows:

$$\text{Members of BNMR} = [M_1, M_2, \ldots, M_N], \tag{5}$$

where N is the number of members related to the number of problem's unknown parameters. Initial production of food sources within the parameter's borders is defined as follows:

$$\begin{matrix} x_i = x_i^{Min} + \beta\left(x_i^{Max} - x_i^{Min}\right) \\ i = 1, \ldots, S \end{matrix}, \tag{6}$$

where x_i represents the ith food source and β is a random variable in the interval of [0, 1] and S represents the number of food sources.

Hereby, the food sources (responses) in the search process are considered as targets to be found by employed moles (i.e., finding the location of food sources and their neighbors, determining the volume of enrichment, discharging food sources and storing them in the kitchen room).

The random movement of employed moles starts from the center of the colony towards food sources and their neighbors.