Engineering Applications of Artificial Intelligence 61 (2017) 1–7

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

# A novel data clustering algorithm based on modified gravitational search algorithm

XiaoHong Han*, Long Quan, XiaoYan Xiong, Matt Almeter, Jie Xiang, Yuan Lan

*Key Laboratory of Advanced Transducers and Intelligent Control Systems, Ministry of Education China, Taiyuan University of Technology, Taiyuan, Shanxi, China*

ARTICLE INFO

*Keywords:*
Gravitational search algorithm
Learning algorithm
Collective behavior
Data clustering
Clustering Validation
Nature-inspired algorithm

ABSTRACT

Data clustering is a popular analysis tool for data statistics in many fields such as pattern recognition, data mining, machine learning, image analysis, and bioinformatics. The aim of data clustering is to represent large datasets by a fewer number of prototypes or clusters, which brings simplicity in modeling data and thus plays a central role in the process of knowledge discovery and data mining. In this paper, a novel data clustering algorithm based on modified Gravitational Search Algorithm is proposed, which is called Bird Flock Gravitational Search Algorithm (BFGSA). The BFGSA introduces a new mechanism into GSA to add diversity, a mechanism which is inspired by the collective response behavior of birds. This mechanism performs its diversity enhancement through three main steps including initialization, identification of the nearest neighbors, and orientation change. The initialization is to generate candidate populations for the second steps and the orientation change updates the position of objects based on the nearest neighbors. Due to the collective response mechanism, the BFGSA explores a wider range of the search space and thus escapes suboptimal solutions. The performance of the proposed algorithm is evaluated through 13 real benchmark datasets from the well-known UCI Machine Learning Repository. Its performance is compared with the standard GSA, the Artificial Bee Colony (ABC), the Particle Swarm Optimization (PSO), the Firefly Algorithm (FA), K-means, and other four clustering algorithms from the literature. The simulation results indicate that the BFGSA can effectively be used for data clustering.

## 1. Introduction

Data clustering is one of the most important and popular data analysis techniques, which involves the process of classifying an unlabeled dataset into clusters of similar objects. Each cluster consists of objects that are similar within the cluster and dissimilar to objects of other clusters (Barbakh et al., 2009; Jain, 2010; Berikov, 2014). Clustering has been applied in many applications such as web mining, text mining, image processing, stock prediction, signal processing, biology and other fields of science and engineering (Everitt et al., 2011; Bishop, 2006).

There are many clustering algorithms that have been proposed for clustering problems in the literature. Traditional clustering algorithms fall into two main categories: hierarchical algorithms and partitional algorithms (Everitt et al., 2001; Xu and Wunsch, 2005).

Hierarchical algorithms create a tree structure of clusters in the absence of any prior knowledge about the number of clusters (Nanda and Panda, 2014). These algorithms can be carried out through two modes: agglomerative mode or divisive mode. In agglomerative mode, each object is regarded as a separate cluster in the beginning and then two most similar clusters are merged at each step. This process reoccurs until termination criteria are satisfied. In divisive mode, all objects are considered as one cluster in the beginning and then each cluster is divided into two clusters until termination criteria are met.

In partitional algorithms, each cluster is assigned initially a centroid. Then based on the similarity between each object and each centroid, all objects will be classified into a corresponding cluster. The objective of these algorithms is to maximize the similarity within one cluster while minimizing connectivity among different clusters (Everitt et al., 2001; Xu and Wunsch, 2005).

Apart from traditional clustering algorithms, there are other clustering algorithms (Chang et al., 2009). Ensity-based methods (Ankerst et al., 1999; Ester et al., 1996) and nearest neighbor methods (Lu and Fu, 1978) are based on the idea that neighbor objects ought to belong to the same cluster. Bi-clustering algorithms (Madeira and Oliveira, 2004) make their clustering through row and column simultaneously. Multi-objective clustering algorithms (Dehuri et al., 2006) and clustering ensembles approaches (Hong et al., 2008) are

* Corresponding author.
*E-mail address:* jmqchs@sohu.com (X. Han).

http://dx.doi.org/10.1016/j.engappai.2016.11.003
Received 30 March 2016; Received in revised form 27 September 2016; Accepted 7 November 2016
Available online 23 February 2017

multi-objective clustering algorithms which optimize different characteristics of the dataset. Overlapping clustering algorithms are different from most of clustering algorithms, in which each object belongs to only one cluster. While in overlapping clustering, each object can belong to more than one cluster. Fuzzy C-means is one of the most popular overlapping clustering algorithms (Nanda and Panda, 2014).

In recent years, meta-heuristic algorithms are widely used to solve clustering problems (Nanda and Panda, 2014). From an optimization perspective, clustering problems can be formally considered as a particular kind of NP-hard grouping problem (Falkenauer, 1998). This type of algorithms includes searching for an optimal solution for clustering problems and reducing the risk of trapping in local optima. These algorithms include but not limited to genetic algorithms (GA) (Maulik and Bandyopadhyay, 2000), simulated annealing (SA) (Selim and Alsultan, 1991), Tabu search (Al-Sultan, 1995; Glover and Laguna, 1997), Artificial Bee Colony (ABC) (Karaboga and Ozturk, 2011a), Greedy Randomized Adaptive Search Procedure(GRASP) (Feo and Resende, 1989), Iterated Local Search(ILS) (Stutzle, 1999), Variable Neighborhood Search (VNS) (Mladenovic and Hansen, 1997), ant colony optimization (ACO) (Shelokar et al., 2004), Particle swarm optimization (PSO) (Chen and Ye, 2004; De et al., 2016), and so on.

Gravitational search algorithm (GSA) is one of the newest meta-heuristic optimization algorithms inspired by the Newtonian laws of gravity and motion (Rashedi et al., 2009). In GSA, an object in the search space attracts every other one with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. The GSA has been proved to be an excellent optimization method for different types of applications, including data clustering (Dowlatshahi and Nezamabadi-pour, 2014), fuzzy model identification (Li et al., 2012), classification (Han et al., 2014; Zhang et al., 2013; Li et al., 2015), economic emission load dispatch (Jiang et al., 2014), wind turbine control (Chatterjee et al., 2014), and power systems (Shuaib et al., 2015).

Motivated by the success of the GSA with variant optimization problems, this paper proposes a novel data clustering algorithm based on a modified GSA (called Bird Flock GSA, BFGSA). The aim of the modified GSA is to enhance the capability of GSA in exploration without reducing the capability in exploitation. In the proposed BFGSA, we introduce bird flock behavior into GSA, which is a collective response process of how birds flock together. The concept of collective behavior of birds is inspired by the concept previously proposed (Hereford and Blum, 2011; Netjinda et al., 2015), which was used to enhance the original PSO. In their proposed method, the original velocity and position are updated by new equations which simulates the collective behavior of starlings. In this paper, we integrate collective response process of birds into GSA to enhance its performance. In our work, we maintain the original velocity and position updating equation of GSA most of the time, except when the global best stops in a local optimum, the position of objects are updated by the mechanism called bird flock behavior.

The rest of this paper is organized as follows. The cluster analysis problem is discussed in Section 2. Section 3 provides a review of standard GSA. The description of the proposed clustering algorithm is presented in Section 4. In Section 5, the experiments of the proposed algorithm for data clustering are given. Finally, a brief conclusion is offered in Section 6.

## 2. The data clustering problem

Data clustering is a process of classifying a set of objects into groups in which similarity in the same group must be maximized and objects that belong to different groups must be dissimilar as possible (Nanda and Panda, 2014; Hruschka et al., 2009).

Mathematically, a data set with $N$ objects, each of which has $d$ attributes, is denoted by $X$={$X_1$, $X_2$, ..., $X_N$}$^T$, where $Xi$={$x_i^1$, $x_i^2$, ..., $x_i^d$} is a vector denoting the $i^{th}$ object and $x_i^j$ is a scalar denoting the $j^{th}$ attribute of $x_i$. The number of attributes is called the dimensionality of the data set. Let $\boldsymbol{X}_{n\times d}$ be the profile data matrix, with $n$ rows and $d$ columns. Given $\boldsymbol{X}_{n\times d}$, the goal of clustering is to classify $X$ into $K$ groups or clusters, $C_1$, $C_2$, ..., $C_k$, such that objects in the same cluster are as similar to each other as possible, while objects in different clusters are quite distinct. And also the following criteria should be satisfied (Nanda and Panda, 2014):

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

$$C_i \bigcap C_j = \varnothing, \quad i, j = 1, \ldots, K; i \neq j \tag{2}$$

$$C_i \neq \varnothing, \quad i = 1, \ldots, K \tag{3}$$

Eq. (1) and Eq. (2) show that all objects in datasets must be classified and each object must belong to only one cluster, and Eq. (3) indicates that each cluster must contain objects.

To find optimal cluster centers with meta-heuristic algorithms, the objective function should be minimized. In this paper, we use quantization error formula as objective function and the optimization problem can be defined as follows:

$$J_e = \frac{\sum_{j=1}^{N_c}\left[\sum_{\forall z_p \in C_{ij}} d(z_{p,m_j})/|C_{ij}|\right]}{N_c} \tag{4}$$

in which $|C_{ij}|$ indicates the number of cluster $C_{ij}$; $d$ indicates Euclidean distance between each data vector and the centroid. This Euclidean distance can be calculated by the following expression:

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2} \tag{5}$$

in which $k$ indicates the dimension; $N_d$ indicates the number of attributes of each data vector; $z_p$ indicates $p^{th}$ data vector and $m_j$ indicates centroid vector of cluster $j$. The cluster centroid vectors are recalculated through the following expression:

$$m_j = \frac{1}{n_j} \sum_{\forall z_p \in C_j} z_p \tag{6}$$

in which $n_j$ indicates the number of data vectors in cluster $j$ and $C_j$ indicates the subset of data vectors from cluster $j$.

## 3. The gravitational search algorithm

The gravitational search algorithm (GSA) is a newly developed stochastic population-base heuristic optimization algorithm based on the law of gravity and mass interactions, which was first introduced by Rashedi et al. Rashedi et al. (2009), Rashedi (2007), Rashedi et al. (2007). The algorithm provides an iterative process that simulates mass interactions, and develops through a multi-dimensional search space under the influence of gravitation. In GSA, all solutions are called agents or objects whose performances are evaluated by their masses; these agents or objects attract each other by gravitational force which causes a global movement of all objects towards objects with heavier masses (Rashedi et al., 2007, 2009; Rashedi, 2007).

Assumed there are $k$ objects, the position of the $i$th object is defined as Eq. (7):

$$X_i = (x_i^1, \ldots, x_i^d, \ldots, x_i^n),\ i = 1, 2, \ldots, k, \tag{7}$$

where $x_i^d$ denotes the position of $i$th object in the $d$th direction. The force exerting on the object $i$ from the object $j$ is defined as Eq. (8):

$$F_{ij}^d(t) = G(t)\frac{M_i(t) \times M_j(t)}{R_{ij}(t) + \varepsilon}[x_j^d(t) - x_i^d(t)], \tag{8}$$

where $M_j$ is the mass related to object$j$, $M_i$is the mass related to object $i$, $\varepsilon$ is a small constant, and $R_{ij}(t)$is the Euclidian distance between the object $i$and object $j$. $G$ is a function of the initial value $G_0$and iteration$t$,