Engineering Applications of Artificial Intelligence 64 (2017) 67–82

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A simplex method-based social spider optimization algorithm for clustering analysis

Yongquan Zhou [a,b,*], Yuxiang Zhou [a,c,d], Qifang Luo [a,b], Mohamed Abdel-Basset [e]

[a] College of Information Science and Engineering, Guangxi University for Nationalities, Nanning 530006, China
[b] Key Laboratory of Guangxi High Schools Complex System and Computational Intelligence, Nanning 530006, China
[c] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[d] Beijing Engineering Applications Research Center of High Volume Language Information Processing and Cloud Computing (Beijing Institute of Technology), Beijing 100081, China
[e] Faculty of Computers and Informatics, Zagazig University, Head of Department of Operations Research, Egypt

ARTICLE INFO

Keywords:
Clustering analysis
Social-spider optimization algorithm
Simplex method
Benchmark datasets
Meta-heuristic algorithm

ABSTRACT

Clustering is a popular data-analysis and data-mining technique that has been addressed in many contexts and by researchers in many disciplines. The *K*-means algorithm is one of the most popular clustering algorithms because of its simplicity and easiness in application. However, its performance depends strongly on the initial cluster centers used and can converge to local minima. To overcome these problems, many scholars have attempted to solve the clustering problem using meta-heuristic algorithms. However, as the dimensionality of a search space and the data contained within it increase, the problem of local optima entrapment and poor convergence rates persist; even the efficiency and effectiveness of these algorithms are often unacceptable. This study presents a simplex method-based social spider optimization (SMSSO) algorithm to overcome the drawbacks mentioned above. The simplex method is a stochastic variant strategy that increases the diversity of a population while enhancing the local search ability of the algorithm. The application of the proposed algorithm on a data-clustering problem using eleven benchmark datasets confirms the potential and effectiveness of the proposed algorithm. The experimental results compared to the *K*-means technique and other state-of-the-art algorithms show that the SMSSO algorithm outperforms the other algorithms in terms of accuracy, robustness, and convergence speed.
© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Data clustering describes the process of grouping data into a given number of clusters. The objective of data clustering is to gather data that share a high degree of likeness within a given cluster; this group of data will also be dissimilar from other data. Clustering algorithms have been applied to a wide range of fields and applications, such as data analysis, data mining (Ng and Han, 1994), image segmentation (Bhanu and Peng, 2000), and pattern recognition (Kamel and Selim, 1994) and outlier detection (Anaya-Sánchez et al., 2010; Gil-García and Pons-Porrata, 2010; Mahdavi et al., 2008; Friedman et al., 2007; Moshtaghi, 2011; Liao et al., 2008).

Clustering algorithms can be classified as either partitional or hierarchical (Han and Kamber, 2001). Partitional clustering algorithms divide data vectors into a predefined number of clusters by optimizing a given criterion. The *K*-means clustering algorithm (Zalik, 2008) is one of the most popular partitional clustering methods due to its simplicity and ease of use. However, its performance depends strongly on the initial cluster centers used and can converge to local minima. To overcome these problems, many methods have been proposed. Bezdeck proposed the Fuzzy C-Means (FCM) clustering algorithm (Bezdek, 1981), Zhang and Hsu introduced the K-Harmonic Means (KHM) algorithm (Zhang and Hsu, 2000), Wang and Zhou introduced the Flower Pollination Algorithm with Bee Pollinator for cluster analysis (Rui et al., 2016), Adán and Wilfrido introduced automatic clustering using nature-inspired metaheuristics (Adán and Wilfrido, 2015), Zhang and Zhou using grey wolf optimizer with powell local optimization for clustering analysis (Sen and Yongquan, 2015). However, traditional clustering methods have successfully solved the problem of data clustering with low-dimension, small data features. However, these algorithms have many shortcomings due to problems with noise and initialization, which

* Corresponding author at: College of Information Science and Engineering, Guangxi University for Nationalities, Nanning 530006, China.
E-mail addresses: yongquanzhou@126.com (Y. Zhou), analyst_mohamed@yahoo.com (M. Abdel-Basset).

http://dx.doi.org/10.1016/j.engappai.2017.06.004
Received 7 August 2016; Received in revised form 26 January 2017; Accepted 3 June 2017
0952-1976/© 2017 Elsevier Ltd. All rights reserved.

predefines the number of clusters and can cause convergence at local optima. These problems may also lead to low performance while analyzing high-dimension, large datasets. The efficiency and effectiveness of these traditional clustering algorithms are also often unacceptable. Thus, recently, many scholars have begun to use meta-heuristic algorithms to solve clustering problems because these methods are not sensitive to the initial values used and do not become trapped at local minima, even in large datasets. For example, in 2008, Nikam, Olamaie, Amiri et al. proposed an efficient hybrid evolutionary algorithm based on combining the ACO and SA in a clustering problem (Niknam et al., 2008a, b). Shelokar et al. introduced an evolutionary algorithm based on the ACO algorithm to solve clustering problems (Shelokar et al., 2004) in 2004. Merwe et al. presented the PSO algorithm to solve clustering problems (Merwe and Engelbrecht, 2003; Omran et al., 2005) in 2003. Yannis Marinakis et al. proposed a hybrid particle swarm optimization algorithm for clustering analysis (Marinakis et al., 2007) in 2007. In 2008, C. Ozturk and D. Karaboga used the ABC algorithm (Ozturk and Karaboga, 2008), and Wenping Zou et al. presented a cooperative artificial bee colony (CABC) algorithm for clustering analysis (Zou et al., 2010) in 2010.

The Social Spider optimization (SSO) algorithm (Cuevas et al., 2013; Cuevas and Cienfuegos, 2014) was proposed by Erik Cuevas in 2013 and was based on the simulation of cooperative behavior of social spiders. In the SSO algorithm, individuals emulate a group of spiders that interact to each other based on the biological laws of such a cooperative colony with two different search agents (i.e., spiders): males and females. Depending on the gender of a spider, each individual operates by a set of different evolutionary operators that mimic different cooperative behaviors that are typically found in such a spider colony. Different from most swarm intelligence algorithms, the SSO algorithm models each individual based on their gender. Thus, this algorithm realistically emulates the cooperative behavior of the swarms and incorporates computational mechanisms to avoid critical flaws that are common in other algorithms (e.g., premature convergence and incorrect exploration–exploitation balance). Thus, the SSO algorithm has been extensively researched and applied to various fields. Pereira, L. and Rodrigues et al. applied the SSO algorithm to train artificial neural networks and study Parkinson's disease identification (Pereira et al., 2014b). Pereira, D. R. and Rodrigues, D. et al. used the SSO algorithm to support the tuning of vector machines' parameters (Pereira et al., 2014a). Mirjalili, S. Z. and Saremi, S. applied the SSO algorithm to design evolutionary feed-forward neural networks (Mirjalili et al., 2015).

Thus, as the dimensionality of a search space and the amount of data increase, the problem of local optima entrapment and poor convergence rates persist. A primary obstacle in applying the SSO algorithm to complex problems has often been its high computational cost due to its slow convergence rate. In this study, the simplex strategy is applied to the original SSO algorithm to solve data-clustering problems by enhancing the algorithm's global and local searching abilities, thereby avoid becoming trapped at local optima and increasing the rate of convergence. The Simplex Method-Based Social Spider Optimization (SMSSO) algorithm uses the simplex method as a stochastic variant strategy to increase the diversity of a population while enhancing the local search ability of the original algorithm.

The remainder of this study is organized as follows. In the Section 2, a mathematical model of the clustering analysis problem is described. Section 3 presents the original SSO algorithm. The details of the SMSSO algorithm are then introduced in Section 4. The simulation and comparison of the proposed algorithm are presented in Section 5. Finally, remarks and conclusions are provided in Section 6.

2. Mathematical model of clustering analysis

Data clustering is the process of grouping data into a given number of clusters. The goal of data clustering is to collect data that share a high degree of likeness into the same cluster; this group of data will thus be dissimilar from the data in other clusters. The mathematical model of a clustering analysis (Ma et al., 2015) is described below.

2.1. Problem description

To clearly define the concepts used in this study, we assume that there is a dataset $D = \{d_1, d_2, \ldots, d_n\}$, where each datum d_i $(i = 1, 2, \ldots, n)$ has several attributes including the parameter m. Thus, each datum can be expressed as $d_i = (l_1, l_2, \ldots, l_m)$. The process of clustering is to group the dataset D into a K clusters $G_1, G_2, \ldots, G_K$ based on the similarity of each datum. Thus, G_i $(i = 1, 2, \ldots, K)$ should satisfy the following formulae:

$$G_i \neq \emptyset, i = 1, 2, \ldots, K. \tag{1}$$

$$G_i \cap G_j = \emptyset, i, j = 1, 2, \ldots, K, i \neq j. \tag{2}$$

$$\bigcup_{i=1}^{K} G_i = \{d_1, d_2, \ldots, d_n\}. \tag{3}$$

2.2. Clustering criteria

In the clustering process, if the given dataset D is grouped into K clusters $G_1, G_2, \ldots, G_K$, then each cluster must have a clustering center c_j $(j = 1, 2, \ldots, K)$. Thus, the vector of the clustering center can be expressed as $C = \{c_1, c_2, \ldots, c_K\}$. The primary idea of clustering is to find the best vector C that makes the data in the same cluster share a high degree of likeness while being dissimilar compared to the data from other clusters. In this study, we will use the Euclidian metric as a distance metric to represent the similarity of each data; this metric can be expressed as follows:

$$d(d_i, c_j) = \sqrt{\sum_{k=1}^{m} (d_{i,k} - c_{j,k})} \tag{4}$$

where m is the number of data attributes, $d_{i,k}$ is the kth attribute of the ith data in the given dataset D, $c_{j,k}$ is the kth attribute of the jth clustering center in the clustering center vector C, and $d(d_i, c_j)$ is the distance between the ith data and the jth clustering center. The distances between each datum and clustering center are first calculated using formula (4). Then, each datum is assigned to the nearest clustering center $c_{near}(c_{near} \in C)$. For example, if $d(d_i, c_j) < d(d_i, c_f)$, we assign d_i to the fth class.

2.3. Evaluation function of the data clustering

In this study, the SMSSO algorithm is proposed to solve data-clustering problems. To explain the evaluation process clearly, it is assumed that we want to divide the dataset D, where each datum d_i has m attributes, into K clusters. Additionally, each attribute of the dataset D is m. To optimize the coordinates of the centers of K clusters, the dimension of solution must be $K * m$. Each individual spider in the SMSSO algorithm represents one clustering center vector of the clustering problem and can be described as $C = \{c_1, c_2, \ldots, c_K\}$. A good classification should minimize the sum of distances required. Thus, we should only minimize the distance between each datum d_i and the center c_j $(j = 1, 2, \ldots, K)$ of the cluster to which it belongs. Finally, the proposed algorithm attempts to minimize the objective function, which is defined as follows:

$$f(D, C) = \sum_{i=1}^{n} Min\{\|d_i - c_k\| \mid k = 1, 2, \ldots, K\} \tag{5}$$

where D is the dataset, and C is the clustering center vector. Thus, solving clustering problems with the SMSSO algorithm requires identifying the optimal clustering center vector C to minimize the evaluation function.

3. Social spider optimization algorithm

The social spider algorithm optimization (SSO) (Cuevas et al., 2013; Cuevas and Cienfuegos, 2014) algorithm was proposed by Erik Cuevas in 2013 and was based on the simulation of cooperative behavior of

Download English Version:

https://daneshyari.com/en/article/4942646

Download Persian Version:

https://daneshyari.com/article/4942646

Daneshyari.com