Engineering Applications of Artificial Intelligence 72 (2018) 54–66

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Evolutionary static and dynamic clustering algorithms based on multi-verse optimizer

Sarah Shukri [a], Hossam Faris [a], Ibrahim Aljarah [a,*], Seyedali Mirjalili [b], Ajith Abraham [c]

[a] King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan
[b] Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD 4111, Australia
[c] Machine Intelligence Research Labs (MIR Labs), Auburn, WA, United States

ARTICLE INFO

Keywords:
Optimization
Metaheuristics
Multi-verse optimizer
Clustering
Swarm Intelligence
Data mining
Machine learning
Evolutionary Computation

ABSTRACT

Clustering based on nature-inspired algorithms is considered as one of the fast growing areas that aims to benefit from such algorithms to formulate a clustering problem as an optimization problem. In this work, the search capabilities of a recent nature-inspired algorithm called Multi-verse Optimizer (MVO) is utilized to optimize clustering problems in two different approaches. The first one is a static clustering approach that works on a predefined number of clusters. The main objective of this approach is to maximize the distances between different clusters and to minimize the distances between the members in each cluster. In an attempt to overcome one of the major drawbacks of the traditional clustering algorithms, the second proposed approach is a dynamic clustering algorithm, in which the number of clusters is automatically detected without any prior information. The proposed approaches are tested using 12 real and artificial datasets and compared with several traditional and nature-inspired based clustering algorithms. The results show that static and dynamic MVO algorithms outperform the other clustering techniques on the majority of datasets.

1. Introduction

Clustering is a data mining task that aims to classify data instances which are usually represented as vectors of data points into a set of groups (clusters) based on predefined similarity metrics. Traditionally, similarity metrics are represented by distance measurements between the different data points in different clusters (Donalek, 2011). Based on the similarity measures, data instances within the same cluster are expected to be more similar than any instances from other clusters. In contrast, the instances from different clusters are expected to be more dissimilar (Jain et al., 1999). Clustering can be used broadly in many applications such as information retrieval (Jardine and van Rijsbergen, 1971), image processing (Reddy et al., 2007), document categorization (Fung et al., 2003), pattern recognition (Baraldi and Blonda, 1999), and web log clustering (Xie and Phoha, 2001).

Clustering is considered as one of the most challenging problems in data mining. Clustering applies unsupervised learning methods by analyzing and extracting useful information and patterns of the target data without having any prior knowledge about the data labels. Therefore, the general challenging problems that face clustering tasks can be summarized in two points: first, clustering deals with unlabeled data where the actual class is unknown (Donalek, 2011). Second, determining the right number of the clusters without any prior information might be difficult and challenging. Usually, experts in the domain are consulted to help in determining the right number of the clusters.

In general, clustering techniques are classified into two main categories: hierarchical and partitional. In the hierarchical clustering, the clustering process is divided into two main classes: agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative clustering, the data points are merged together from different clusters into a single cluster. On the other hand, divisive clustering based methods split a large cluster into smaller clusters using similarity based measures (Donalek, 2011; White and Frenk, 1991; Karypis et al., 1999; Bandyopadhyay and Coyle, 2003).

In contrast to the hierarchical clustering, partitional clustering methods (e.g. *k*-Means algorithm) divide data points into predefined groups/clusters. The cluster representative centroid is then calculated for each group. After that, all clusters are evaluated using some similarity based measures such as the sum of squared errors (SSE), which is calculated based on the distances between the data points and cluster centroids. The main objective of the partitional clustering methods is

* Corresponding author.
E-mail addresses: hossam.faris@ju.edu.jo (H. Faris), i.aljarah@ju.edu.jo (I. Aljarah), seyedali.mirjalili@griffithuni.edu.au (S. Mirjalili), ajith.abraham@ieee.org (A. Abraham).

https://doi.org/10.1016/j.engappai.2018.03.013
Received 18 May 2017; Received in revised form 15 March 2018; Accepted 17 March 2018

to minimize the SSE measure (Lee and Antonsson, 2000; Jarboui et al., 2007; Xu and Wunsch, 2005). Partitional clustering methods have been widely applied to diverse domains including image segmentation (Marroquin and Girosi, 1993), intrusion detection (Jianliang et al., 2009), and information retrieval (Bellot and El-Bèze, 1999). Partitional clustering methods have some advantages such as low computation complexity and simplicity. However, they suffer from major drawbacks such as the local optimum problem and the sensitivity to the initial centroids. Therefore, many new algorithms have been developed in literature to overcome such drawbacks. Some of these new methods benefit from the capabilities of nature-inspired metaheuristics.

Nature-inspired metaheuristics are search algorithms mainly inspired by some phenomena in nature such as the biological systems, swarming behavior, physical or chemical systems. These algorithms have gained an increased popularity in the recent years due their superior efficiency and reasonable run time as compared to exact methods when solving large-scale and challenging real-world problems (Yang, 2011). One of the well-regarded recent nature-inspired algorithms is the Multi-Verse optimizer algorithm (MVO). MVO was first implemented and proposed by Mirjalili et al. (2016). MVO is inspired by the abstract concepts of the popular multi-verse theory. Since its proposal, MVO has been effectively used and applied in several applications such as engineering applications (Trivedi et al., 2016; Jangir et al., 2017) and machine learning problems (Faris et al., 2017, 2016; Aljarah et al., 2018; Mafarja et al., 2017).

In this paper, the MVO algorithm is employed to tackle the challenges of clustering problems. Two approaches based on MVO are proposed and developed. The first is a static approach, in which the number of clusters has to be predefined before starting the clustering process. The second version is a dynamic clustering approach, which automatically detects the number of clusters without any prior knowledge about the nature of the dataset. The proposed approaches benefit from the search capabilities and powerful operators of the MVO algorithm to alleviate the drawbacks of the partitional clustering algorithms. To verify the effectiveness of the proposed approaches, they will be evaluated based on real and artificial datasets with different levels of difficulty. They will also be compared to a set of popular clustering approaches in the literature.

The remainder of this paper is organized as follows: Section 2 reviews a wide range of related works in the area of nature-inspired clustering based algorithms. Section 3 describes the preliminaries of the clustering problem and the MVO algorithm. In Section 4 and Section 5, the details of the proposed static and dynamic approaches of the MVO based clustering algorithms are given, respectively. The evaluation measures that are used to evaluate the proposed clustering approaches are listed in Section 6. The experiments and results are presented in Section 7. Finally, Section 8 summarizes the conclusions and future directions of this work.

2. Related works

Data clustering is an important and critical problem in data mining. As such, many algorithms were proposed in the literature to solve clustering problems. The K-Means algorithm is considered as one of the most popular partitioning clustering algorithms (MacQueen et al., 1967). This algorithm aims to partition a given dataset into a pre-defined number of groups (k) by minimizing the distances between the data points in the same group and maximizing the distance between the data points in different groups. Another traditional clustering algorithm is called Furthest First (FF) (Kumar et al., 2013). FF is an enhanced version of K-Means algorithm that assigns each new centroid at furthermost location from the existing centroids. This technique speeds up the clustering process compared to the original K-Means algorithm.

Most of the traditional clustering techniques suffer from several drawbacks including the need to determine the number of the groups as an input parameter, which is not practical in real-world problems, the possibility to fall in a local optimum, and the sensitivity to the initial centroids. To tackle these issues, many nature-inspired algorithms were introduced. For instance, Maulik and Bandyopadhyay (2000) proposed a genetic algorithm-based clustering (CGA) in 2000. The optimization capability of genetic algorithm was used to find the best centroids. The experimental results proved the superiority of the CGA algorithm over the K-Means algorithm. Another variant of the genetic algorithm called grouping genetic algorithm (GGA) in Agustıet al. (2012) was used to solve the clustering problem. GGA applied the concepts of grouping encoding and evolutionary operators (mutation, and crossover) on the clustering. The performance of the algorithm in different clustering problems obtained competitive results. Other works that implemented GA based clustering can be found in Scheunders (1997), Doval et al. (1999), Rao et al. (2016) and Ding and Fu (2016).

In Van der Merwe and Engelbrecht (2003), Particle Swarm Optimization (PSO) was applied on clustering analysis. In this work, the capability of the PSO and its advantages such as avoiding the premature convergence to local optima is used to refine the centroids. In 2008, Ahmadyfard and Modares (2008) combined the PSO and K-Means algorithm. The motivation of this hybrid was to benefit from the fast convergence of the PSO in the initial generations, and K-Means, which is faster to reach global solution. Parallel versions of PSO clustering were introduced in Aljarah and Ludwig (2012, 2013a, c). Hassanzadeh and Meybodi (2012) presented a new clustering approach based on the Firefly algorithm. The Firefly algorithm was combined with K-Means algorithm to optimize the best values of the best centroids of the predefined clusters.

In Lee and Antonsson (2000), evolutionary-based clustering approach was proposed. The authors proposed a new evolutionary algorithm to optimize the number of the best centroids along with the best set of groups. The proposed approach designed to deal with 2D spatial data due to the limitation of graphical representation for multi-dimensional data. Each cluster's centroid was represented as a pair of (x, y), and each gene represented n number of centers ordered in an ascending way. Their evaluation measures showed promising results.

In Shelokar et al. (2004), a clustering based Ant colony optimization algorithm was proposed to solve clustering problems by simulating the movement technique performed by ants to group similar instances in the same cluster. The performance of the proposed algorithm showed better results compared to other stochastic algorithms.

In Aljarah and Ludwig (2013b), a clustering algorithm (CGSO) based on the nature-inspired algorithm called Glowworm Swarm Optimization (GSO) was introduced. The CGSO was adjusted to fit the data clustering problem and to locate multiple optimal centroids within the search space. The authors used three fitness functions to evaluate the performance of the algorithm. Their conducted experiments showed that the algorithm outperforms four of the most popular algorithms in most of the used datasets.

In Ozturk et al. (2015), the authors introduced a new nature-inspired clustering based on binary artificial bee colony. The main objective was to introduce a new solution generation mechanism for dynamic clustering by considering new similarity measures in an efficient way. More examples of nature-inspired based clustering approaches can be found in Omran et al. (2005), Senthilnath et al. (2011), Hatamlou (2013), Aljarah and Ludwig (2012) and Al-Madi et al. (2014).

In most of the existing nature-inspired based clustering algorithms, the number of predefined clusters is needed in advance to solve clustering problems. However, in several practical applications, the number of the clusters is unavailable before exploring the data set. Another issue is that some of the nature-inspired clustering algorithms suffer from slow convergence and poor clustering quality. In this paper, two approaches are proposed to take advantage of MVO optimization efficiency to tackle the clustering problem to enhance the speed of the convergence and solve the problem of determining the number of clusters in advance.