Engineering Applications of Artificial Intelligence 73 (2018) 126–148

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Optimized gravitational-based data clustering algorithm

Mohammed Alswaitti, Mohamad Khairi Ishak, Nor Ashidi Mat Isa *

School of Electrical & Electronic Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia

ARTICLE INFO

Keywords:
Gravitational clustering
Centroid initialization
Nature-inspired algorithms
Exploitation and exploration balance
Clustering analysis

ABSTRACT

Gravitational clustering is a nature-inspired and heuristic-based technique. The performance of nature-inspired algorithms relies on the balance achieved between exploitation and exploration. A modification over a data clustering algorithm based on the universal gravity rule is proposed in this paper. Although gravitational clustering algorithm has a high exploration ability, it lacks a proper exploitation mechanism because of the impulsive velocity of agents that search the solution space, which leads to the huge step size of agent positions through iterations. This study proposes the following solutions to impose a balance between exploitation and exploration: (i) the dependence of the agent on velocity history is removed to avoid high velocity caused by accumulating previous velocities, and (ii) an initialization step of centroid positions is added using the variance and median initialization method with a predefined number of clusters. The initialization step eliminates the effects of random initialization and subrogates the exploration process. Experiments are conducted using 13 benchmark datasets from the UCI machine learning repository. In addition, the proposed algorithm is tested on two case studies using the electrical hotspots and cervical cell datasets. The performance of the proposed clustering algorithm is compared qualitatively and quantitatively with several state-of-the-art clustering algorithms. The obtained results indicate that the proposed clustering algorithm outperforms conventional techniques. Furthermore, the clusters obtained using the proposed algorithm are more homogeneous than those obtained using conventional techniques. The proposed algorithm quantitatively achieves better results than the other techniques in 9 out of 15 datasets in terms of accuracy, F-score, and purity.

1. Introduction

Clustering aims to extract natural groupings hidden in data to simplify these data into meaningful and comprehensible information. The resulting subgroups gather similar objects based on their features to form clusters. The search for clusters is unsupervised, and thus important in machine learning. Clustering has wide applications in the field of web analysis, business and marketing, education, scientific data exploration, and medical diagnostic.

Clustering algorithms have diverse categories. Each algorithm has its own working mechanism, ability to deal with certain types of data, advantages, and drawbacks. These algorithms can be broadly classified into partitional, hierarchal, and density-based clustering algorithms.

The simplest and the most used partitional clustering algorithm is K-means (Hartigan and Wong, 1979). It is a centroid-based clustering technique, where cluster objects are centered on their nearest representative according to a distance function (e.g., the Euclidean distance). The mean inside the cluster is computed, and the centroid updates its position to the position of the mean. The popularity of the K-means algorithm can be attributed to several reasons, such as its ease of implementation. Furthermore, K-means is a versatile algorithm in all aspects, that is, different approaches can be used for its initialization, distance function, convergence criterion, and so on. However, it has significant drawbacks that affect its performance, namely, the number of clusters K must be known in advance, it cannot detect overlapping clusters, it is sensitive to the initial positions of centroids and outliers, and its convergence to local minimum (Celebi et al., 2013).

Numerous studies have proposed different solutions to these problems to enhance the quality of clustering results obtained using the K-means algorithm. One of the competitive alternatives is the Fuzzy C-means algorithm (Bezdek et al., 1984). It is a soft clustering technique, where objects partially belong to all clusters with probabilities. A centroid is determined by averaging the value of all objects with different degrees specified by a membership function. The adopted fuzzy concept provides flexibility in centroid positioning and decreases the algorithm's sensitivity to the initialization. On the contrary, this approach obscures the boundaries between resulting clusters.

* Corresponding author.
E-mail addresses: mswaitti@gmail.com (M. Alswaitti), khairiishak@usm.my (M.K. Ishak), ashidi@usm.my (N.A.M. Isa).

https://doi.org/10.1016/j.engappai.2018.05.004
Received 9 January 2017; Received in revised form 18 April 2018; Accepted 11 May 2018

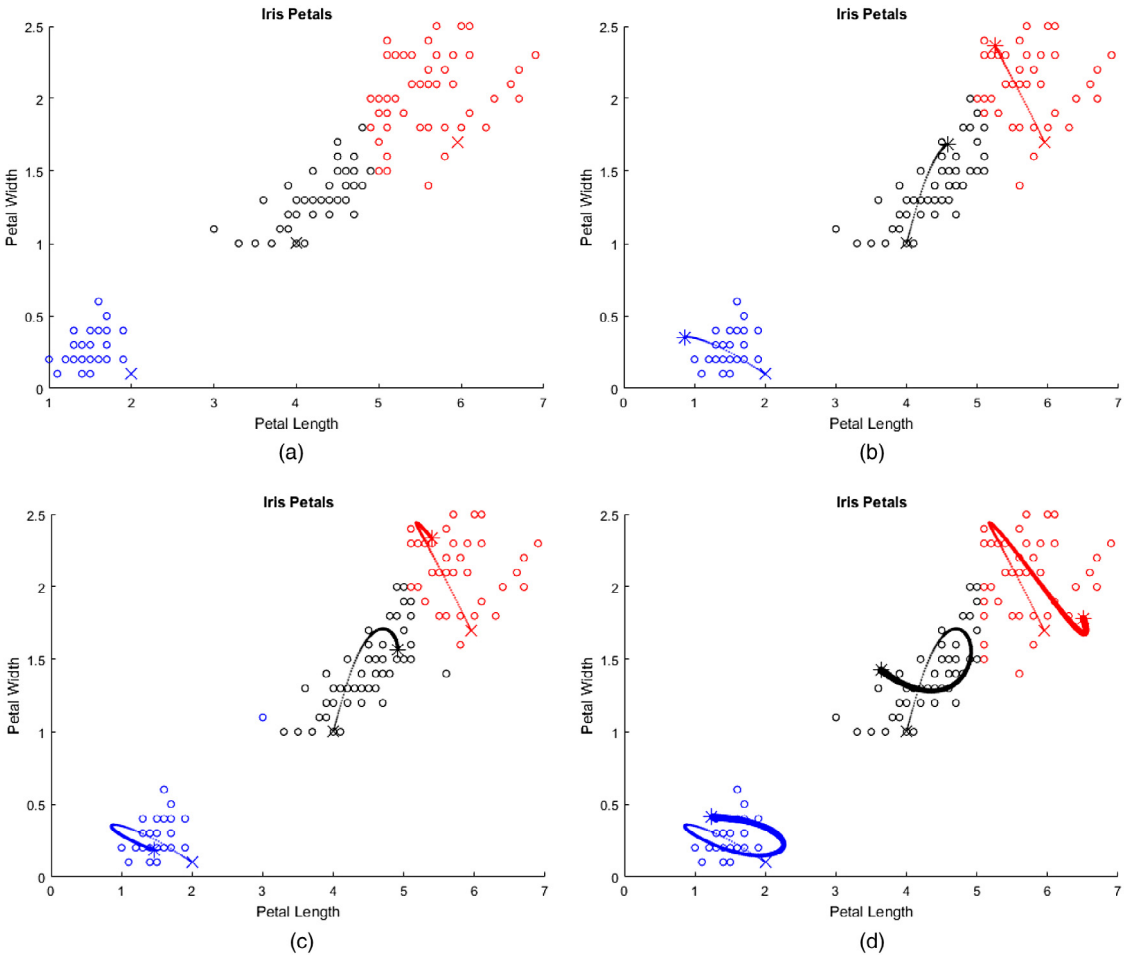

Fig. 1. Iris petals clustering by GC algorithm through iterations: (a) iteration 1, (b) iteration 50, (c) iteration 85, and (d) iteration 200.

In hierarchal clustering (Dabhi and Patel, 2016), two schemes have been adopted. The first is an agglomerative scheme in which each object represents a cluster at the beginning of the algorithm. Then, a merging process is applied until all objects are gathered in one cluster. The second is a divisive scheme, which is a top down approach that considers all objects to be in one cluster at the beginning of the algorithm. Then, a splitting process is applied until each object represents a distinct cluster. A tree diagram (dendrogram) can be used to represent the hierarchy of clustering obtained using both schemes; it allows exploring data on different levels of granularity. However, the complexity of hierarchal clustering algorithms is higher than that of partitional methods.

Connecting data objects based on their nearest neighbors is another approach called density-based clustering (Loh and Park, 2014). A cluster is defined as a connected dense component that grows in any direction the density leads to. This strategy allows detecting clusters with arbitrary shapes, and it is robust against outliers. However, density-based clustering algorithms are inadequate for clusters with varied densities and high dimensional data.

The plurality of clustering algorithms is a result of attempts to find the ultimate clustering algorithm, which seems intricate, because each clustering approach has its own advantages and disadvantages when applied to different datasets. Therefore, many modified versions have been proposed to handle the drawbacks of their "ancestors". In addition, other trends embraced ideas inspired by nature to propose clustering solutions.

Nature-inspired algorithms (evolutionary algorithms) have seized a competitive stature in solving clustering problems (Nanda and Panda, 2014). The prominence of bio-inspired computing is increasing due to its various applications in engineering (Kar, 2016). This type of intelligence is interpreted in distinct algorithms, such as the genetic algorithm (GA) (Holland, 1975; Razavi et al., 2015), the differential evolution (DE) (Cai et al., 2011; Liu and Guo, 2016; Storn, 1996), the particle swarm optimization (PSO) (Armano and Farmani, 2016; Kennedy and Eberhart, 1995; Thong and Son, 2016), recent swarm intelligence algorithms (Min et al., 2016), and the gravitational search algorithm (GSA) (Rashedi et al., 2009; Xiao et al., 2016).

Gravity theory has been utilized in numerous research areas to solve different problems. It has been used in the design of PID control systems (De Moura Oliveira et al., 2015), data classification (Rezaei and Nezamabadi-pour, 2015; Shafigh et al., 2013), optimization (Shen et al., 2015), training spiking neurons (Dowlatshahi et al., 2016), detection of cancer tumors in mammography images (Shirazi and Rashedi, 2016), the solution to open vehicle routing problem (Hosseinabadi et al., 2016), document clustering (Sadeghian and Nezamabadi-pour, 2015), and data clustering (Nikbakht and Mirvaziri, 2015).

Object attraction and merging are emulated based on gravity forces in data clustering algorithms, which employ gravity theory. Each data point is considered an object associated with a certain mass for an applicable process. Kumar and Sahoo (2014) conducted a detailed review of recent gravity-based algorithms and their applications in clustering and classification.

The commencement was by the unit attraction gravitational Markovian model presented by Wright (1977). In the model, objects obtain gravitational relations that cause movements toward each other. When objects are close, they join to form clusters that finally combine into one.

Yung and Lai (1998) employed the Markovian model of gravity-based clustering to segment red–green–blue (RGB) color images. Each