Expert Systems With Applications 91 (2018) 170–186

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Density-based particle swarm optimization algorithm for data clustering

Mohammed Alswaitti[a], Mohanad Albughdadi[b], Nor Ashidi Mat Isa[a,*]

[a] School of Electrical and Electronic Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia
[b] University of Toulouse, IRIT/UPS Avenue de l'étudiant, 31400 Toulouse, France

ARTICLE INFO

Article history:
Received 18 May 2017
Revised 6 August 2017
Accepted 24 August 2017
Available online 1 September 2017

Keywords:
Particle swarm optimization
Swarm intelligence
Universal gravity rule
Kernel density estimation
Exploitation and exploration balance
Data clustering

ABSTRACT

Particle swarm optimization (PSO) algorithm is widely used in cluster analysis. However, it is a stochastic technique that is vulnerable to premature convergence to sub-optimal clustering solutions. PSO-based clustering algorithms also require tuning of the learning coefficient values to find better solutions. The latter drawbacks can be evaded by setting a proper balance between the exploitation and exploration behaviors of particles while searching the feature space. Moreover, particles must take into account the magnitude of movement in each dimension and search for the optimal solution in the most populated regions in the feature space. This study presents a novel approach for data clustering based on particle swarms. In this proposal, the balance between exploitation and exploration processes is considered using a combination of (i) kernel density estimation technique associated with new bandwidth estimation method to address the premature convergence and (ii) estimated multidimensional gravitational learning coefficients. The proposed algorithm is compared with other state-of-the-art algorithms using 11 benchmark datasets from the UCI Machine Learning Repository in terms of classification accuracy, repeatability represented by the standard deviation of the classification accuracy over different runs, and cluster compactness represented by the average Dunn index values over different runs. The results of Friedman Aligned-Ranks test with Holm's test over the average classification accuracy and Dunn index values indicate that the proposed algorithm achieves better accuracy and compactness when compared with other algorithms. The significance of the proposed algorithm is represented in addressing the limitations of the PSO-based clustering algorithms to push forward clustering as an important technique in the field of expert systems and machine learning. Such application, in turn, enhances the classification accuracy and cluster compactness. In this context, the proposed algorithm achieves better results compared with other state-of-the-art algorithms when applied to high-dimensional datasets (e.g., Landsat and Dermatology). This finding confirms the importance of estimating multidimensional learning coefficients that consider particle movements in all the dimensions of the feature space. The proposed algorithm can likewise be applied in repeatability matters for better decision making, as in medical diagnosis, as proved by the low standard deviation obtained using the proposed algorithm in conducted experiments.

1. Introduction

Unsupervised learning algorithms play an outstanding role in machine learning owing to their capabilities in exploring data without having any prior information about them, i.e., no labels are associated with these data. These algorithms aim to model the underlying structure or distribution in the data, which can be used for decision making and predicting future inputs, among others. Classic examples of unsupervised algorithms are represented in clustering and dimensionality reduction techniques (Ghahramani, 2004).

Clustering is a technique that extracts natural groupings hidden in data to simplify them into meaningful and comprehensible information. The resulting groups gather similar objects based on their features to form clusters. The applications of clustering techniques have been used in a wide range of different areas, including web analysis, business, marketing, education, data science, and medical diagnosis, among others.

In data clustering, each cluster is represented by a center and a similarity measure, such as the Euclidean distance, is then used to measure the similarity between a data point and the obtained centers. Finally, data points are assigned to the corresponding cluster associated with the nearest center to this data point. The

* Corresponding author.
E-mail addresses: mswaitti@gmail.com (M. Alswaitti), mohanad.albughdadi@enseeiht.fr (M. Albughdadi), ashidi@usm.my (N.A.M. Isa).

http://dx.doi.org/10.1016/j.eswa.2017.08.050
0957-4174/

more compact the data points are, the more obvious the patterns become. To achieve maximum compactness, centers should be located in the densest partitions of the cluster to reflect an accurate distribution of congregated data points inside that group and minimize the effect of the few extreme data points on the center positions.

Clustering algorithms can be classified broadly into different categories, namely, partitional, hierarchical, and density-based algorithms. Each category has its own working mechanism, capability to deal with certain types of data, advantages, and drawbacks (Saraswathi & Sheela, 2014).

One of the most widely used partitional clustering algorithms is the K-means (Hartigan & Wong, 1979), a centroid-based clustering technique, where objects in a cluster are centered around its nearest representative according to a distance function (e.g., Euclidean distance). These objects also have the same weight in updating the position of the centroid. Although the K-means is a versatile algorithm, it has some drawbacks that limit its usefulness. For instance, this algorithm lacks the capability to discover grouping in overlapping clusters. Furthermore, the K-means algorithm is sensitive to the initialization step where the initial positions of the centroids are specified (Celebi, Kingravi, & Vela, 2013). One of the competitive alternatives to the K-means algorithm is the fuzzy c-means algorithm (Bezdek, Ehrlich, & Full, 1984), a soft clustering technique in which objects partially belong to all clusters. A centroid is determined by averaging the value of all objects with different degrees specified by a membership function. The adopted fuzzy concept gives flexibility in centroid positioning and makes the algorithm less sensitive to initialization.

Two schemes have been adopted in hierarchical clustering (Dabhi & Patel, 2016). The first scheme is an agglomerative one, in which each object represents a cluster at the beginning of the algorithm. A merging process is then 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. A splitting process is then applied until each object represents one distinct cluster. A tree diagram (dendrogram) can be used to represent the hierarchy of clustering obtained using both schemes, providing the advantage of exploring data using different levels of the dendrogram with no prior information about the number of clusters. However, the complexity of the hierarchical clustering algorithms is higher than that of the partitional methods (Han, Pei, & Kamber, 2011).

Density-based clustering is another approach used to connect data objects based on their nearest neighbors (Loh & Park, 2014); it uses density to define a cluster as a connected dense component. This strategy allows for detecting clusters with arbitrary shapes and is robust against outliers. However, density-based clustering algorithms are not adequate for clusters with varied densities and high-dimensional data (Moreira, Santos, & Carneiro, 2005).

Other trends have adopted ideas inspired by nature by proposing population-based clustering algorithms, which have seized a competitive stature in solving clustering problems (Nanda & Panda, 2014). The prominence of bio-inspired computing is increasing due to its various applications in engineering (Kar, 2016). This kind of intelligence at the beginning was interpreted to distinct algorithms such as genetic algorithm (GA) (Holland, 1975), differential evolution (DE) (Storn, 1996), gravitational search algorithm (GSA) (Rashedi, Nezamabadi-pour, & Saryazdi, 2009), and particle swarm optimization (PSO) (Kennedy & Eberhart, 1995). Recently, numerous variations of swarm intelligence algorithms have been proposed for dynamic optimization (Mavrovouniotis, Li, & Yang, 2017) along with their applications to real-world problems.

One of the most popular swarm intelligence algorithms is the PSO, which is an optimization algorithm that has been widely used in data clustering because of its simplicity and scalability (Alam, Dobbie, Koh, Riddle, & Ur Rehman, 2014). However, several studies have shown that the PSO algorithm suffers from premature convergence to local optima (Liang, Qin, Suganthan, & Baskar, 2006; Rana, Jasola, & Kumar, 2011), which occurs as a result of the lack of diversity of PSO and the inability of particles to explore the entire feature space. This finding may be attributed to the imbalance between the exploitation and exploration processes (Matej, Liu, & Mernik, 2013). The exploration process allows for searching new solutions far from the current one in the search space. By contrast, the exploitation process aims at searching the nearby area of the current solution. Furthermore, a set of parameters (learning coefficients) are required to be tuned in the PSO algorithm, which may affect its performance (Alam et al., 2014; Silva Filho, M., Pimentel, Souza, & Oliveira, 2015). The values of these parameters are often set manually despite their effect on controlling the exploitation and exploration processes (Matej et al., 2013). Consequently, many attempts have been proposed to enhance the performance of the PSO clustering algorithm to solve the premature convergence and the tuning of learning coefficients (Lee, El-Saleh, & Ismail, 2014; Pei & Tong, 2016a; Filho et al., 2015).

In this paper, a new approach for data clustering is proposed based on the PSO algorithm combined with the kernel density estimation (KDE). The non-parametric properties of the KDE motivate its use for improving the balance between the exploitation and exploration processes. For instance, the KDE makes no model assumptions other than using a specific kernel, which enables it to model clusters with non-convex shapes. The KDE has no local minima and is not overly affected by outliers (Carreira-Perpinán, 2015). In the framework of the proposed algorithm, the universal gravity rule is also used to estimate multidimensional learning coefficients that consider the movement of the particles in all the dimensions of the feature space. In this context, most of the previous approaches assign a single value of the learning coefficients; this value assumes that the magnitude of the particle movement is equally affected in all the dimensions. The previous assumption does not consider the variety of data distribution in each dimension of the feature space.

The rest of the paper is organized as follows. A detailed background of the PSO algorithm, its utilization in cluster analysis, and improvements over the original PSO algorithm are provided in Section 2. The proposed algorithm is introduced in Section 3. Validation on datasets from the UCI using different measures and statistical analysis is conducted in Section 4. Finally, conclusions are drawn in Section 5.

2. Background

2.1. PSO algorithm

PSO is a population-based, heuristic, and evolutionary algorithm inspired by nature. It imitates bird flocking and fish schooling as a swarm, how they move, change their positions, and trajectories to find their destination. This theory was first exploited by Kennedy and Eberhart (1995) to solve nonlinear continuous problems. To find the optimal solution in the search space, particles (birds) are placed in a continuous movement, where the velocity of each particle represents the direction of the search, and the corresponding position is a candidate solution. In the traditional PSO, the initial velocities and positions of all particles are randomly selected. The next velocities and positions are updated through iterations until a global solution is found, which is determined according to a fitness function or an objective function throughout the search. The mathematical representation of the traditional PSO is as follows: