Engineering Applications of Artificial Intelligence 45 (2015) 415–428

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A data clustering approach based on universal gravity rule

Abbas Bahrololoum, Hossein Nezamabadi-pour *, Saeid Saryazdi

Department of Electrical Engineering, Shahid Bahonar University of Kerman, P.O. Box 76169-133, Kerman, Iran

ARTICLE INFO

Article history:
Received 22 February 2015
Received in revised form
23 July 2015
Accepted 24 July 2015
Available online 15 August 2015

Keywords:
Data clustering
Law of gravity
Nature inspired algorithm
Clustering analysis
Data classification

ABSTRACT

In this paper, a new robust data clustering algorithm inspired by Newtonian law of gravity is proposed. The proposed algorithm not only reduces the effects of noise and outliers but also, it is not sensible to the initial positions of the centroids. In the proposed method, data points and the cluster centroids are considered as fixed celestial objects and movable objects, respectively. The celestial objects apply a gravity force to the movable objects and change their positions in the feature space and therefore, the best positions of the cluster centroids are determined by employing the law of gravity. To evaluate the performance of the proposed algorithm, a comparative experimental study with some well-known clustering algorithms, using three visual datasets as well as several benchmark datasets from UCI, is performed. The experimental results confirm the effectiveness and the efficiency of the proposed clustering algorithm.

1. Introduction

Generally, there is no common terminology on definition of data clustering, but most researchers describe a cluster through its internal homogeneity and external separation, i.e., data points in the same cluster should be similar (or related) to each other, and different from (or unrelated to) the data points in other clusters. Both similarity and dissimilarity should be examinable in a clear and meaningful way. These algorithms are usually unsupervised, and they are mostly used in the fields of machine learning, data mining, pattern recognition, image analysis and bioinformatics (Jain et al., 1999; Hammouda, 2011).

The general part of all clustering algorithms is to find the representatives of clusters, i.e. cluster centroids for compact clusters. A clustering algorithm decides, each input data belongs to which cluster, (i.e. is the closest to which centroid). In some of the clustering techniques, the algorithm tries to partition data points into a given number of clusters, e.g. K-means (Hartigan and Wong, 1978) and fuzzy C-means (Bezdek et al., 1984). In some cases the number of clusters is not known a priori. Such an algorithm starts by finding the largest cluster first, next goes to find the second, and so on (Yager and Filev, 1994; Chiu, 1995; Wu and Yang, 2002).

The well-known K-means algorithm is one of the most used algorithms due to its efficiency and simplicity in data clustering where, it measures the distance between clusters' representatives (centroids) and data points to partition data into K clusters. In most cases, the Euclidean distance is used as the dissimilarity measure. To find the best position of the representatives, the K-means algorithm minimizes a cost function of data variations around the centroids. However, the initial state may cause the algorithm to trap into a local optimum, therewith affecting the quality of the final solution. Many studies have been made to overcome this drawback of the K-means algorithm, particularly by using fuzzy set theory and evolutionary algorithms (Taherdangkoo and Bagheri, 2013; Niknam and Amiri, 2010).

Fuzzy C-means (FCM) algorithm is an improvement over K-means clustering. In this algorithm, each data point belongs to a cluster by a degree of membership. Similar to the K-means algorithm, FCM relies on minimizing a cost function of the dissimilarity measure to find centroids. The FCM uses the concept of fuzzy set theory to handle the uncertainty associated with the data to be clustered (Gustafson and Kessel, 1979; Pal et al., 2005; Zarandi et al., 2009).

The Mountain clustering algorithm calculates a mountain function (density function) at every possible position in the data space, and chooses the position with the greatest density value as the center of the first cluster. Then, it removes the effects of the first cluster mountain function and finds the second cluster center (Moertini, 2002). This process is repeated until a desired number of clusters is found. The subtractive clustering is similar to the mountain clustering, except that it uses data point positions to calculate the density function, thus reduces the number of calculations significantly (Kim et al., 2005). This means that the computation depends on the problem size instead of the problem dimension.

Each clustering algorithm has its advantages and disadvantages in clustering of different types of data points. Therefore, many

* Corresponding author. Tel.:/fax: +98 34 33235900.
E-mail address: nezam@uk.ac.ir (H. Nezamabadi-pour).

http://dx.doi.org/10.1016/j.engappai.2015.07.018
0952-1976/

clustering methods have been proposed to rectify disadvantages of these algorithms. Also, some researchers tried to suggest new algorithms inspired by nature. In this paper, a nature inspired clustering algorithm is introduced by employing Newtonian law of gravity. Gravity based clustering algorithms are not new and their history returns to the 70s. In the following, first related works are reviewed and then the proposed work is described.

1.1. Related works

Recently, more and more attention has been focused on using nature based inspired algorithms to solve clustering problems (Nanda and Panda, 2014). Moreover, there are many clustering algorithms which have been made by hybridizing different types of evolutionary algorithms with K-means algorithm to overcome the disadvantage of K-means. Evolutionary algorithms such as genetic algorithm (GA) (Maulik and Bandyopadhyay, 2000), ant colony optimization (ACO) (Kashef and Nezamabadi-pour, 2014), particle swarm optimization (PSO) (Niknam and Amiri, 2010), gravitational search algorithm (GSA) (Rashedi and Nezamabadi-Pour, 2013; Dowlatshahi and Nezamabadi-pour, 2014) , are usually nature inspired. Some of these methods are reviewed in Yazdani et al. (2014).

Various applications of the gravity theory in solving problems of different areas have been considered including image edge detection (Sun et al., 2007; Deregeh and Nezamabadi-pour, 2014), data classification (Shafigh et al., 2013; Rezaei and Nezamabadi-pour, 2015), optimization (Soleimanpour-Moghadam et al., 2014), and data clustering (Sanchez et al., 2014; Wright, 1977; Yung and Lai, 1998). Gravity-based clustering algorithms simulate the process of the attraction and merging of objects by their gravity forces. To realize data clustering, these algorithms consider each data point as an object and assign a mass to it.

The first version of gravitational clustering algorithm was proposed by Wright (1977). This algorithm uses a Markovian model for the gravitational clustering. It is an incremental algorithm that updates the position of each data point in each iteration. How the objects are joined is determined by the continuous motion of all objects in the system according to gravitational forces. In this method, objects do not converge together but rather converge to "equilibrium" positions (Wright, 1977).

Yung and Lai (1998) present a Markovian model of clustering based on gravitational concepts. The model is used for color image segmentation in RGB color space. In the clustering process, each pixel is considered as an object with the unity mass. All objects apply gravitational force to each other. The Markovian model of gravitational attraction between two objects i and j is defined as follows:

$$F(m_i, m_j) = -G\frac{m_i m_j}{\|X_i - X_j\|^3}(X_i - X_j) \tag{1}$$

where X_i and X_j present the locating D-dimensional vectors of objects i and j, respectively, m_i and m_j are the mass of them, G is the universal gravitational constant and $\|.\|$ is a Euclidean norm function where $\|X_i - X_j\| = \sqrt{\sum_{l=1}^{D}(x_{il} - x_{jl})^2}$. This force causes the objects (data points) to move in the space. Two objects that move to the same location are merged and form a new object that its mass is considered as the summation of the two merged masses.

A clustering method based on the notion of attraction force between each pair of data points has been presented by Kundu (1999). The clusters are formed by allowing each data point to move slowly under the resultant effect of all forces exerted to it, and by merging two data points when they become close to each other. When two or more data points (clusters) are merged, the sum of their masses becomes the mass of the resulting cluster; therefore, the mass of each cluster equals the number of its data instances (Kundu, 1999).

In the work done by Gomez et al. (2004), a modified version of Newton's law of gravity is used. This modification simplifies the calculation of gravitational force, where Eq. (2) is used for moving a data point according to the gravitational law. The value of the universal gravitational constant, G, is reduced after each iteration, which serves to eliminate the big crunch effect of all the data points. This is used as a mechanism that does not end up with only one cluster

$$X(t+1) = X(t) + G\,\vec{d}\,f\left(\frac{\tilde{d}}{\vec{d}}\right) \tag{2}$$

where $\vec{d} = X_i - X_j$ is the vector direction, f is a decreasing function such as $f(\alpha) = 1/\alpha^3$ and X is the position of the data point, $\vec{d}$ is the vector direction of such data point in time t, $\tilde{d}$ is the rough estimate of maximum distance between the closest points such as $2\sqrt{D}/\sqrt{3}n^{\frac{1}{D}}$ for n data points in the D-dimensional [0,1] Euclidean space and $\|.\|$ is the norm function (Gomez et al., 2004).

Long and Jin (2006) proposed a simplified gravitational clustering (SGC) method for multi-prototype learning based on minimum classification error. It simulates the movement of objects according to the gravity force and checks for possible merging. The gravitational clustering is simplified by ignoring velocity and multi-force attraction. The pair objects are merged based on the following equations:

$$\begin{aligned} \{X_i, X_j\} &= \arg\min_{i,j}\|X_i - X_j\|^{\frac{m_i m_j}{2}} \\ m_{merged} &= m_i + m_j \\ X_{merged} &= \frac{X_i m_i + X_j m_j}{m_i + m_j} \end{aligned} \tag{3}$$

The approach proposed by Zhang is based on a simplified version of Newtonian gravitational forces and Newtonian motion of objects, as shown in Eqs. (4) and (5), respectively (Zhang and Hongshan, 2010).

$$V_i(t + \Delta(t)) = V_i(t) + G\,m_i\,\Delta(t)\frac{\vec{d}}{\vec{d}^3} \tag{4}$$

$$X_i(t + \Delta(t)) = X_i(t) + V_i(t)\,\Delta(t) + G\,m_i\,\Delta(t)^2\frac{\vec{d}}{\vec{d}^3} \tag{5}$$

where $\Delta(t)$ is a small discrete time interval and V is the velocity of object i.

The value of the universal gravitational constant, G is reduced at each iteration (Zhang and Hongshan, 2010). In the GRIN algorithm, an incremental hierarchical clustering technique based on the gravity theory, is presented to construct clustering dendrograms. An incremental clustering algorithm refers to the abstraction of distribution of data instances generated by the previous run of the algorithm. The mass of a cluster is the number of its data instances. GRIN algorithm works in two phases: initial phase, and incremental phase. In both phases, it invokes the gravitational agglomerative hierarchical clustering algorithm (Chen et al., 2005).

Ilc and Dobnikar (2012) presented a gravitational based method for clustering, known as GSOM, in which, each data point is viewed as a mass object. In GSOM, the basic idea is used and integrated with SOM, considering the connections between neurons (Ilc and Dobnikar, 2012).

Rashedi et al. (2009) have been proposed a stochastic population-based metaheuristic, called Gravitational Search Algorithm (GSA), based on Newtonian law of gravity and the laws of motion. Originally, GSA is designed for solving continuous optimization problems. GSA, like most metaheuristics has flexible abilities in different application such as clustering (Hatamlou et al., 2012).

Rashedi and Nezamabadi-Pour (2013) have been proposed a clustering algorithm based on the theory of gravity. In the