Expert Systems with Applications 37 (2010) 8679–8684

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Ant clustering algorithm with *K*-harmonic means clustering

Hua Jiang *, Shenghe Yi, Jing Li, Fengqin Yang, Xin Hu

College of Computer Science, Northeast Normal University, Changchun, Jilin 130117, China

ARTICLE INFO

Keywords:
Clustering
K-means
K-harmonic means clustering
Ant clustering algorithm

ABSTRACT

Clustering is an unsupervised learning procedure and there is no a prior knowledge of data distribution. It organizes a set of objects/data into similar groups called clusters, and the objects within one cluster are highly similar and dissimilar with the objects in other clusters. The classic *K*-means algorithm (KM) is the most popular clustering algorithm for its easy implementation and fast working. But KM is very sensitive to initialization, the better centers we choose, the better results we get. Also, it is easily trapped in local optimal. The *K*-harmonic means algorithm (KHM) is less sensitive to the initialization than the KM algorithm. The Ant clustering algorithm (ACA) can avoid trapping in local optimal solution. In this paper, we will propose a new clustering algorithm using the Ant clustering algorithm with *K*-harmonic means clustering (ACAKHM). The experiment results on three well-known data sets like Iris and two other artificial data sets indicate the superiority of the ACAKHM algorithm. At last the performance of the ACAKHM algorithm is compared with the ACA and the KHM algorithm.

1. Introduction

Clustering is a popular data analysis method and plays an important role in data mining. So far it has been widely applied in many fields, like web mining, pattern recognition, machine-learning, spatial database analysis, artificial intelligence, and so on.

The existing clustering algorithms can be simply classified into the following two categories: hierarchical clustering and partitional clustering (Jain, Murty, & Flynn, 1999). The classic *K*-means algorithm (KM) is the most popular clustering algorithm due to its simplicity and efficiency.

Though the *K*-means algorithm is widely used to solve problems in many areas, KM is very sensitive to initialization, the better centers we choose, the better results we get. Also, it is easily trapped in local optimal (Khan & Ahmad, 2004). Recently much work was done to overcome these problems.

Simulated annealing (SA) algorithm was proposed to find the equilibrium configuration of a collection of atoms at a given temperature, and it is always used to solve the combinatorial problems. Simulated annealing heuristic was used with *K*-harmonic means to overcome local optimal problem (Güngör & Ünler, 2007).

Tabu search (TS) is a search method used to solve the combinatorial optimization problems, and the algorithm TabuKHM (Tabu *K*-harmonic means) was developed in 2008 (Güngör & Ünler, 2008).

A hybrid technique was proposed by Kao, Zahara, and Kao (2008). It is based on the *K*-means algorithm, Nelder–Mead simplex search, and particle swarm optimization (*K*–NM–PSO). The *K*–NM–PSO searches for cluster centers of an arbitrary data set as does the KM algorithm, but it can effectively find the global optima.

Particle swarm optimization (PSO) is a popular stochastic optimization technique developed by Kennedy and Eberhart, and a new hybrid algorithm based on PSO and KHM was proposed (Yang & Sun, 2009).

Moreover, some other hybrid heuristic methods like genetic simulated annealing or tabu-search with simulated annealing were ever used with clustering algorithm to solve local optimal problem (Chu & Roddick, 2003; Huang, Pan, Lu, Sun, & Hang, 2001).

In this paper we propose a new algorithm using the Ant clustering algorithm with *K*-harmonic means clustering (ACAKHM). This paper is organized as follows. Section 2 describes the clustering algorithms and gives prominence to the *K*-harmonic means clustering. Section 3 introduces the Ant clustering algorithm. In Section 4, our new algorithm, Ant clustering algorithm with *K*-harmonic means clustering, is presented. Section 5 explains the data sets and the experimental results. Finally, Section 6 summarizes the main conclusion of this study.

2. Clustering

Clustering is an unsupervised learning procedure and there is no a prior knowledge of data distribution (Liu, 2006). It is the

* Corresponding author. Fax: +86 0431 84536331.
E-mail address: jiangh289@nenu.edu.cn (H. Jiang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.06.061

process of organizing a set of objects/data into groups called clusters, and the objects within one cluster are similar as much as possible according to a predefined criterion which is always defined with similarity measure.

There are two categories of clustering, the hierarchical clustering and the partitional clustering. The hierarchical clustering can be either agglomerative or divisive. The process of the hierarchical clustering is grouping a set of objects with a sequence of partitions, either from singleton clusters to a cluster including all individual or vice verse. In this paper we pay more attention to the partitional clustering.

2.1. The partitional clustering

The partitional clustering can be described as follows (Xu, 2005):

Given a set of input patterns $X = \{x_1, \ldots, x_i, \ldots, x_N\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T \in R^d$ and x_i is the feature (attribute, dimension, or variable) of the data.

Partitional clustering attempts to seek a K-partition of $X, C = \{C_1, C_2, \ldots, C_k\}$, $(k \leqslant N)$, and

(1) $C_j \neq \phi, \quad j = 1, \ldots, k;$
(2) $\bigcup_{j=1}^{k} C_j = X, \quad j = 1, \ldots, k;$
(3) $C_i \cap C_j = \phi, \quad i, j = 1, \ldots, k, \quad i \neq j.$

The above partitional clustering is a kind of hard partitional which means each pattern only belongs to one cluster. However, in most cases the pattern may be allowed to belong to two or more clusters. This is known as fuzzy clustering which characteristic is that a pattern belongs to all clusters with a degree of membership $m(c_j/x_i)$, where c_j is the center of cluster C_j.

$m(c_j/x_i)$ presents the degree of membership of the object x_i belongs to the cluster j. And it satisfies the following constraints:

$$\sum_{j=1}^{k} m(c_j/x_i) = 1, \ \forall i; \quad \sum_{i=1}^{N} m(c_j/x_i) < N, \ \forall j.$$

K-means and K-harmonic means are both center-based clustering algorithms. Particularly, K-means algorithm was first presented over three decades ago. It makes use of minimizing the total mean-squared distance from each point of the data set to the point of closest center. It is hard partitional clustering. On the contrary, the K-harmonic means is fuzzy clustering, and it was presented recently by Zhang, Hsu, and Dayal (1999, 2000) and modified by Hammerly and Elkan (2002). It minimizes the harmonic average from all points in the data set to each center. It will be explained in detail in the following section.

2.2. The K-harmonic means clustering

The K-harmonic means was proposed by Zhang et al. (1999, 2000) and modified by Hammerly and Elkan (2002). It is a center-based clustering algorithm. The difference between KM and KHM is that the KM algorithm gives equal weight to all of the data points and the KHM algorithm every time gives dynamic weight to each data point with a harmonic average. The harmonic average assigns a large weight to a data point that is not close to any centers and a small weight to the data point that is close to one or more centers. Because of this principal, the KHM algorithm is less sensitive to the initialization than the KM algorithm.

Before we introduce the K-harmonic means clustering, we explain some notations used in the procedure of clustering at first (Güngör & Ünler, 2008; Hammerly & Elkan, 2002; Yang & Sun, 2009):

x_i: ith data point, $i = 1, \ldots, N$.
c_j: jth cluster center, $j = 1, \ldots, k$.
$KHM(X, C)$: The objective function of the KHM algorithm.
$m(c_j/x_i)$: The grade of membership value of the point x_i belongs to cluster j.
$w(x_i)$: The grade of influence value of the point x_i to the position of center c_j in the next iteration.

The detail of the K-harmonic means clustering algorithm is shown as follows:

1. Initialize the KHM algorithm by choosing the initial centers randomly.
2. Calculate objective function value according to

$$KHM(X, C) = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}}, \tag{1}$$

   p is an input parameter and it was proved that KHM works better with the value of $p > 2$.
3. Calculate the membership of each data point x_i to each center c_j according to

$$m(c_j/x_i) = \frac{\|x_i - c_j\|^{-p-2}}{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}, \quad m(c_j/x_i) \in [0, 1]. \tag{2}$$

4. Calculate the weight of each point according to

$$w(x_i) = \frac{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}{\left(\sum_{j=1}^{k} \|x_i - c_j\|^{-p}\right)^2}. \tag{3}$$

5. Calculate the new center location with the membership and weight of each point according to

$$c_j = \frac{\sum_{i=1}^{N} m(c_j/x_i) \cdot w(x_i) \cdot x_i}{\sum_{i=1}^{N} m(c_j/x_i) \cdot w(x_i)}. \tag{4}$$

6. Repeat steps 2–5 until it reaches the predefined number of iterations or until the objective function $KHM(X, C)$ does not change significantly.
7. Assign the point x_i to the cluster j with the biggest $m(c_j/x_i)$.

Due to $m(c_j/x_i)$, the KHM algorithm is particularly useful when the boundaries of the clusters are not well separated and ambiguous. Also, the KHM algorithm is less sensitive to the initialization than the KM algorithm.

3. Ant clustering algorithm

The standard Ant clustering algorithm (ACA) was proposed by Lumer and Faieta (1994), and it closely mimics the ant behavior described in Ant-based clustering written by Deneubourg et al. (1991). The idea of the Ant-based clustering is gathering the corpses and sorting the larval of ants. The principle of gathering or sorting is the positive feedback of the behavior of the ants. The ACA technique provides a relevant partition of data without any knowledge of initial cluster centers, which is the merit of this technique. Given that agent ants perform random walks on a two-dimensional grid on which the objects are scattered randomly, and the size of grid is dependent on the number of objects. The agent ants are allowed to move throughout the grid, picking up and dropping the objects influenced by the similarity and density of the objects within the agent ant's immediate current neighborhood, as well as the agent ant's state (whether it is or is not loading an object) (Handl & Meyer, 2007).

The probability of picking up an object will be increased with low density neighborhoods, and decreased with high similarity