Journal of King Saud University – Computer and Information Sciences (2016) xxx, xxx–xxx

King Saud University
Journal of King Saud University – Computer and Information Sciences
www.ksu.edu.sa
www.sciencedirect.com

Fuzzy clustering based on Forest optimization algorithm

Arash Chaghari, Mohammad-Reza Feizi-Derakhshi, Mohammad-Ali Balafar *

Department of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

Received 12 June 2016; revised 6 August 2016; accepted 20 September 2016

KEYWORDS
Fuzzy clustering;
Partition matrix;
Forest optimization;
Gradient method;
Clustering index

Abstract Clustering is one of the classification methods for data analysis and it is one of the ways of data analysis, too. There are various methods for fuzzy clustering using optimization algorithms such as genetic algorithm and particle swarm optimization algorithm that were specified. In this paper, the combination of one of the recent optimization algorithms called Forest optimization algorithm and one of the local search methods called gradient method are used to perform fuzzy clustering. The purpose of applying the gradient method is accelerating the convergence of the used optimization algorithm. To apply the proposed method, 4 types of real data sets are used. Cluster validity measures are used to obtain and verify the accuracy of the proposed method (FOFCM). By analyzing and comparing the results of the proposed method with the results of algorithms GGAFCM (fuzzy clustering based on genetic algorithm) and PSOFCM (fuzzy clustering based on particle swarm optimization algorithm), it has been shown that the accuracy of the proposed approach is significantly increased.

1. Introduction

Clustering is a classification way for data analysis, which is utilized to classify a set of data or patterns commonly multidimensional into different groups according to a predefined measure, in order that items in the same group are more almost the same than those in different groups. All the more particularly, the patterns that are generally s dimensional vectors are conveyed to c classes while certain sort of optimization criterion is minimized, and the patterns in the same class are more comparable than those in various classes at last. In recent decades, clustering plays the key role in different fields of science and engineering, such as data analysis, pattern recognition, machine learning, image segmentation, error detection and so on.

In general, clustering methods are divided into two general categories; crisp and fuzzy. The degree of the membership of each sample of the data is zero or one in crisp methods. In fact, crisp methods can be considered as a special case of fuzzy algorithms. In other words, the membership value of the sample that belongs to a cluster is one and its membership value for the rest of the clusters is zero. The advantage of crisp methods

* Corresponding author.
E-mail addresses: a.chaghari@tabrizu.ac.ir (A. Chaghari), mfeizi@tabrizu.ac.ir (M.-R. Feizi-Derakhshi), balafarila@tabrizu.ac.ir (M.-A. Balafar).
Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2016.09.005
1319-1578 © 2016 Production and hosting by Elsevier B.V. on behalf of King Saud University.

is its easiness and efficiency to implement. One of the famous algorithms in this area is the algorithm k-means (Forgy, 1965). Although these algorithms are widely used and have developed well, they are not appropriate for fuzzy data set. For this category of algorithms, it is assumed that the data set classes have nothing in common to one another and are completely separated from each other. On the other side of crisp methods, membership degree of the samples is set in the interval [0, 1] in fuzzy methods.

Bezdek developed a fuzzy clustering algorithm, the well-known fuzzy c-means (FCM) (Bezdek, 1973a). The algorithm is the fuzzy equivalence of the algorithm k-means. According to FCM usage, a lot of algorithms are presented to improve the accuracy of clustering. In the standard FCM algorithm and all the proposed methods for its improvement, the number of clusters should have already been set. In other words, in such circumstances, clustering problem can be defined as follows: n sample with s dimension should be in c cluster, so that each sample should be alleged in the corresponding cluster. So, there is an evaluation function that the cluster result is evaluated by and its purpose is to optimize the evaluation function by which, an optimal clustering is achieved.

Global optimization algorithms known as genetic algorithms (Bezdek and Hathaway, 1994; Maulik and Bandyopadhyay, 2000; Bandyopadhyay and Maulik, 2001), ant colony optimization (Dorigo et al., 1996), particle swarm optimization (Liu et al., 2005; De Falco et al., 2007) and chaos optimization (Li et al., 2008) are well-known algorithms to optimize fuzzy clustering. In other words, several researchers formulated the entire clustering task of FCM explicitly as an optimization problem and solved it using various metaheuristics viz., simulated annealing (Granelli et al., 1989; Victoire and Jeyakumar, 2005), variable neighborhood search (Li et al., 1997), genetic algorithms (Han et al., 2001; Victoire and Jeyakumar, 2005), tabu search (Walters and Sheble, 1993) and threshold accepting (Panigrahi et al., 2006) were suggested. Recently, Jayabarathi et al. (2005) applied DE after FC so that it can lead to a global optimum. DE was also used with FCM in several different ways. Gaing (2003) presented a real-coded modified DE based automatic fuzzy clustering algorithm which automatically evolves the number of clusters as well as the proper partitioning from a data set. Passino (2002) proposed an evolutionary-fuzzy clustering algorithm for automatically grouping the pixels of an image into different homogeneous regions. An improved variant of the DE was used to determine the number of naturally occurring clusters in the image as well as to refine the cluster centers. Mishra (2005) used DE to optimize the coordinates of the samples distributed randomly on a plane.

Researchers have tried to improve FCM by introducing excellent optimization methods to optimize the objective function of FCM, trying to avoid trapping into local minima. In Karaboga and Ozturk (2010), bee colony optimization algorithm is used and combined with the algorithm FCM, to cluster data. In the algorithm (Xiaoqiang and Jinhu, 2014) a combination of invasive weed optimization algorithm and clustering algorithm FCM is used so that clustering of the data is done. In the algorithm CPSFC (Li et al., 2012), a combination of particle swarm optimization algorithm, chaotic local search, and gradient method is used to provide good performance in capturing the global optimal fitness, thus getting the best clustering results.

The paper is organized as follows: in Section 2, basic concepts including standard FCM algorithm, Forest optimization algorithm, gradient method and validity indices of fuzzy clustering are mentioned. Section 3 describes the proposed method and in Section 4 the results of the implementation of the proposed method on the data set are shown. In Section 5, conclusions and future work are mentioned.

2. Basic concepts

In this section, the algorithm FCM, Forest optimization algorithm, and gradient method will be discussed. The noted meanings are prerequisite toward the proposed method. Also, for the proposed method evaluation, the evaluation measures will be described.

2.1. FCM algorithm

The main part of fuzzy clustering, is to determine similarity measure by which the distance between the patterns can be determined. In the algorithm FCM, the Euclidean distance is used as similarity measure. Fitness function that is used in FCM algorithm is defined as:

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{i,j})^m \|y_j - z_i\|_A^2 \tag{1}$$

where $Y = (y_1, y_2, \ldots, y_n)$ is the data set that the number of features or dimensions of each sample is equal to s. $Z = (z_1, z_2, \ldots, z_c)$ is the center of clusters. $U = [u_{i,j}]_{c\times n}$ is the partition matrix, $U_{ij} \in [0, 1]$ is interpreted to be the grade of membership of x_j in the ith cluster. Symbol $\|\cdot\|_A$ means norm of matrix A. If A equals the identity matrix, the phrase $\|y_j - z_i\|$ means the Euclidean distance from y_j to the ith cluster center. It is believed the minimization of J_m will produce the best cluster structure and the optimal cluster results.

The minimization of J_m can be reached by Lagrange multiplier method while the partition matrix U and cluster centers Z have expressions as follows:

$$\boldsymbol{u_{i,j}} = \left[\sum_{k=1}^{c}\left(\frac{\boldsymbol{d_{i,j}}}{\boldsymbol{d_{k,j}}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad \boldsymbol{1 \leqslant i \leqslant c;\ 1 \leqslant j \leqslant n} \tag{2}$$

By repeating Eqs. (2) and (3), the fitness function J_m tends toward its minimum value gradually. The algorithm FCM can be expressed as follows:

1. Set the cluster numbers c, set initial cluster centers $z_i^{(0)}$, $1 \leqslant i \leqslant c$, and set the tolerance ε to determine when to stop the algorithm.
2. Acquiring new values of u and z using Eqs. (2) and (3).
3. Calculating the value of the difference between the new cluster centers and the new degree of membership of the second phase of their previous values. If earned value is less than the threshold error ε or the number of iteration is equal to the maximum value, the algorithm will be terminated; otherwise, the second step is performed.

The FCM algorithm can be considered as a kind of local search. So, being located in local minimum and being sensitive to initial cluster centers, are the main problems of FCM