INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING, *SERIES B*
Volume 5, Number 1-2, Pages 113–122

DEVELOPMENT OF A WEIGHTED FUZZY C-MEANS CLUSTERING ALGORITHM BASED ON JADE

KANGSHUN LI, CHUHU ZHANG, ZHANGXIN CHEN, AND YAN CHEN

Abstract. To overcome the shortcomings of falling into local optimal solutions and being too sensitive to initial values of the traditional fuzzy C-mean clustering algorithm, a weighted fuzzy C-means (FCM) clustering algorithm based on adaptive differential evolution (JADE) is proposed in this paper. To consider the particular contributions of different features, a ReliefF algorithm is used to assign the weight for each feature. A weighted morphology-similarity distance (WMSD) based on ReliefF instead of the Euclidean distance is used to improve the objective function of the FCM clustering algorithm. Experimental results on the international standard Iris data and the contrast experimental results with other evolution algorithms show that the proposed algorithm has higher clustering accuracy and greater searching capability.

Key words. Fuzzy C-means, adaptive differential evolution, weighted morphology-similarity distance, clustering precision

1. Introduction

Clustering is to divide the given unknown-type sample data into some meaningful or useful clusters according to a particular standard in a way that the objects in the same cluster are very similar and the objects in different clusters are very different. The purpose of a clustering analysis that now covers the fields of statistics, biology and machine learning such as data mining and pattern recognition is to reveal the internal structure of the underlying data.

Unlike the traditional enforced division, the boundary of a fuzzy clustering analysis is not clear and it does not have the “either-or” property [1], which indicates the fact that data in real life is of intermediary nature. A fuzzy clustering method considers an affiliation relation between sample data and cluster centers, and this relation is extended from two values $\{0, 1\}$ to $[0, 1]$ in order to represent the fuzziness of data and reflect the uncertainty of real world matters better. Fuzzy C-means (FCM) clustering is one of the most widely applied fuzzy clustering algorithms. It was proposed by Dumn (1973) when he was promoting the hard C-means (HCM) [2], and was introduced to a clustering analysis by Bezdek (1981) [3]. The principles of the least squares and iterative gradient descent methods are applied to the classic FCM, which result in the traditional FCM that runs effectively only with spherical or ellipsoidal clustering and is extremely sensitive to noise and outliers [4]. In addition, it has the following defects: (1) Its results are strongly influenced by the initial cluster center and (2) because of using the gradient descent method, it is easily falling into local optimal solutions and cannot gain the global optimal solution. To overcome these defects, in recent years many researchers have focused on the direction of combining global intelligent algorithms with the FCM clustering. Paper [5] combined a global Genetic Algorithm(GA) with a local climbing technique, and an improved GA based on a weighted FCM clustering algorithm

Received by the editors February 12, 2014 and, in revised form, March 20, 2014.

This work is supported by the Guangdong Province Science and Technology Research Project with the Grant (No.2012A020602037) and the Research Project of Science and Technology of Education Department of Jiangxi Province of China with the Grant (No.GJJ12348).

was proposed in paper [6] by using a Gaussian mutation operator and multiple correlation coefficients based on a weighted Euclidean distance instead of the standard Euclidean distance. Both algorithms achieved good clustering results. An improved algorithm based on Particle Swarm Optimization (PSO) using field operation was proposed [7]; it also effectively obtained results on an Iris data set. Other researchers improved FCM using a distance correction factor [8] and a kernel-based learning approach [9]. FCM was analyzed based on different similarity estimation methods [10].

Differential Evolution (DE) was first proposed by Storn and Price (1996) [11, 12]. Compared with the general evolutionary algorithm, DE was proved to converge to the global optimal solution more quickly and be more stable. Compared with other intelligent algorithms, DE's main characteristic is its differential mutation. It is easy to implement, runs with high operating efficiency, and has only three parameters: population size NP, mutation factor F and crossover probability CR. However, the classical DE is sensitive to the mutation factor F and the crossover probability CR; the user must assign different F and CR to different questions. In order to overcome this disadvantage, a weighted fuzzy clustering algorithm based on an adaptive differential evolutionary algorithm (JADE) [13] is proposed in this paper. This algorithm utilizes the JADE global searching capability, and takes into account different contribution of each dimensional feature of a vector to the pattern classification. A weighted morphology-similarity distance based on ReliefF instead of the standard Euclidean distance is used to improve the objective function of the FCM clustering algorithm. Experimental results for this new algorithm compared with general methods show its advantage in terms of better clustering precision of the algorithm.

The organization of this paper is as follows. Section 2 gives a brief introduction to the standard FCM. The classical DE algorithm and the JADE algorithm will be depicted, respectively, in Sections 3 and 4. Section 5 outlines the weighted fuzzy clustering algorithm based on JADE. Section 6 gives the results of simulation and performance evaluation. Finally, conclusions are stated in Section 7.

2. Standard Fuzzy C-means Algorithm

FCM, in accordance with the principle of the least squares method, calculates an objective function mean squares error and iteratively optimizes the objective function to achieve the fuzzy classification of a data set. The basic idea is as follows: Dividing the data set $X = \{x_1, x_2, ..., x_n\} \in \mathbf{R}^n$ into C clusters, where $2 \leq C \leq n$, the clustering result is represented as an affiliation matrix $U = [u_{ik}]$ that satisfies: $u_{ik} \in [0, 1]$, $\sum\limits_{i=1}^{C} u_{ik} = 1 \ \forall k = 1, 2, ..., n$ and $0 < \sum\limits_{k=1}^{n} u_{ik} < n \ \forall i = 1, 2, 3, ..., C$. u_{ik} represents the x_k affiliation value to cluster i. Let $V = \{v_1, v_2, ..., v_C\}$ be the set of C cluster centers. FCM is implemented by minimizing the objective function $J_m(U, V)$ with the affiliation matrix U and the cluster center V:

$$J_m(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{C} (u_{ik})^m d_{ik}^2(x_k, v_i). \tag{1}$$

In formula (1), $m \in [1, \infty)$ is a weighted fuzzy exponent that controls the fuzzy degree of the integer partition. For $m = 1$, the fuzzy clustering will be degraded to the hard C-means clustering. $d_{ik}(x_k, v_i)$ is the Euclidean distance of the sample