Information Sciences 309 (2015) 138–162

Contents lists available at ScienceDirect

Information Sciences

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

Bias-correction fuzzy clustering algorithms

Miin-Shen Yang [a,*], Yi-Cheng Tian [a,b]

[a] Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan
[b] Center for General Education, Hsin Sheng College of Medical Care and Management, Longtan, Taiwan

ARTICLE INFO

Article history:
Received 6 July 2013
Received in revised form 13 February 2015
Accepted 8 March 2015
Available online 14 March 2015

Keywords:
Cluster analysis
Fuzzy clustering
Fuzzy c-means (FCM)
Initialization
Bias correction
Probability weight

ABSTRACT

Fuzzy clustering is generally an extension of hard clustering and it is based on fuzzy membership partitions. In fuzzy clustering, the fuzzy c-means (FCM) algorithm is the most commonly used clustering method. Numerous studies have presented various generalizations of the FCM algorithm. However, the FCM algorithm and its generalizations are usually affected by initializations. In this paper, we propose a bias-correction term with an updating equation to adjust the effects of initializations on fuzzy clustering algorithms. We first propose the so-called bias-correction fuzzy clustering of the generalized FCM algorithm. We then construct the bias-correction FCM, bias-correction Gustafson and Kessel clustering and bias-correction inter-cluster separation algorithms. We compared the proposed bias-correction fuzzy clustering algorithms with other fuzzy clustering algorithms by using numerical examples. We also applied the bias-correction fuzzy clustering algorithms to real data sets. The results indicated the superiority and effectiveness of the proposed bias-correction fuzzy clustering methods.

1. Introduction

Clustering is a method for determining the cluster structure of a data set such that objects within the same cluster demonstrate maximum similarity and objects within different clusters demonstrate maximum dissimilarity. Numerous clustering theories and methods have been evaluated in the literature (see Jain and Dubes [10] and Kaufman and Rousseeuw [11]). In general, the most well-known approaches are partitional clustering methods based on an objective function of similarity or dissimilarity measures. In partitional clustering methods, the k-means (see MacQueen [14] and Pollard [20]), fuzzy c-means (FCM) (see Bezdek [2] and Yang [23]), and possibilistic c-means (PCM) algorithms (see Krishnapuram and Keller [12], Honda et al. [8], and Yang and Lai [24]) are the most commonly used approaches.

Fuzzy clustering has received considerable attention in the clustering literature. In fuzzy clustering, the FCM algorithm is the most well-known clustering algorithm. Previous studies have proposed numerous extensions of FCM clustering (see Gath and Geva [4], Gustafson and Kessel [5], Hathaway et al. [6], Honda and Ichihashi [7], Husseinzadeh Kashan et al. [9], Miyamoto et al. [15], Pedrycz [17], Pedrycz and Bargiela [18], Wu and Yang [22], Yang et al. [25], and Yu and Yang [26]). Regarding the generalization of FCM clustering, Yu and Yang [26] proposed a generalized FCM (GFCM) model to unify numerous variations of FCM. However, initializations affect FCM clustering and its generalizations. In this paper, we evaluated a bias-correction approach by using an updating equation to adjust the effects of initial values and then propose the bias-correction fuzzy clustering methods.

* Corresponding author.
E-mail address: msyang@math.cycu.edu.tw (M.-S. Yang).

http://dx.doi.org/10.1016/j.ins.2015.03.006
0020-0255/

The rest of this paper is organized as follows. Section 2 presents a brief review of the FCM and GFCM algorithms. Section 3 presents the procedures involved in deriving the bias-correction fuzzy clustering algorithms. In these procedures, a bias-correction term is first assessed using an updating equation. The bias-correction FCM (BFCM), inter-cluster separation (ICS), and Gustafson and Kessel (GK) algorithms are then proposed. The bias-correction term is used as the total information for fuzzy c-partitions so that the proposed BFCM, GK, and ICS algorithms can be used to adjust gradually the effects of poor initializations. Section 4 presents comparisons between different clustering algorithms. In the comparisons, the number of optimal clustering results, error rates and root mean squared errors (RMSEs) are used as performance evaluation criteria. Numerical and real data sets are used to demonstrate the effectiveness and usefulness of the proposed bias-correction algorithms. Finally, conclusions and discussion are stated in Section 5.

2. Fuzzy clustering algorithms

Let $X = \{x_1, \ldots, x_n\}$ be a set of n data points in an s-dimensional real Euclidean space. Let c be a positive integer greater than one. The FCM objective function [2,23] is expressed as follows:

$$J_m(\mu, a) = \sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\|x_k - a_i\|^2 \tag{1}$$

where $m > 1$ is the weighting exponent, $a = \{a_1, \ldots, a_c\}$ is the set of cluster centers, and the membership μ_{ik} represents the degree to which the data point x_k belongs to a cluster i with

$$\mu = [\mu_{ik}]_{c\times n} \in M_{fcm} = \left\{\mu = [\mu_{ik}]_{c\times n}\middle|\sum_{i=1}^{c}\mu_{ik} = 1,\ \mu_{ik} \geqslant 0,\ 0 < \sum_{k=1}^{n}\mu_{ik} < n\right\}$$

The FCM algorithm is developed with the objective of obtaining a partition matrix $\mu = [\mu_{ik}]_{c\times n}$ and a set $a = \{a_1, \ldots, a_c\}$ of cluster centers to minimize the objective function $J_m(\mu, a)$. By Lagrange multiplier, the necessary conditions for the minimum of $J_m(\mu, a)$ are the following updating equations:

$$a_i = \frac{\sum_{k=1}^{n}\mu_{ik}^{m}x_k}{\sum_{k=1}^{n}\mu_{ik}^{m}} \tag{2}$$

$$\mu_{ik} = \frac{\|x_k - a_i\|^{\frac{-2}{m-1}}}{\sum_{j=1}^{c}\|x_k - a_j\|^{\frac{-2}{m-1}}} \tag{3}$$

According to Eqs. (2) and (3), the FCM algorithm can be described as follows:

FCM algorithm

Step 1: Fix $2 \leqslant c \leqslant n$ and fix any $\varepsilon > 0$
Give an initial $a^{(0)}$ and let $t = 0$.
Step 2: Compute the membership $\mu^{(t+1)}$ with $a^{(t)}$ using Eq. (3).
Step 3: Update the cluster center $a^{(t+1)}$ with $\mu^{(t+1)}$ using Eq. (2).
Step 4: Compare $a^{(t+1)}$ to $a^{(t)}$ in a convenient matrix norm $\|\cdot\|$.
IF $\|a^{(t+1)} - a^{(t)}\| < \varepsilon$, STOP
ELSE $t = t + 1$ and return to step 2.

The FCM algorithm is the most commonly used clustering algorithm. Numerous generalizations of the FCM algorithm exist. Yu and Yang [26] proposed a unified model, called the generalized FCM (GFCM). The GFCM objective function is expressed as follows:

$$J_m^h(\mu, a) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left[\mu_{ik}^{m}h_i(d(x_k, a_i)) - \frac{\gamma}{c}\sum_{j=1}^{c}h_0\big(d(a_i, a_j)\big)\right] \tag{4}$$

where $\sum_{i=1}^{c}\mu_{ik} = f_k$ for $f_k \geqslant 0$; $\gamma \geqslant 0$ are constant weights; $h_i(x),\ i = 0, 1, \ldots, c$ are continuous functions of $x \in [0, +\infty)$ satisfying its derivative $h_i'(x) > 0$ for all $x \in [0, +\infty)$, and $\sqrt{d(x_k, a_i)}$ is the distance between the data point x_k and the cluster center a_i. The GFCM framework enables modeling numerous FCM variants. By Lagrange multiplier, the necessary conditions for a minimum of $J_m^h(\mu, a)$ are obtained as follows: