Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 159 (2008) 2428–2445
FUZZY sets and systems
www.elsevier.com/locate/fss

Kernelized fuzzy attribute C-means clustering algorithm

Jingwei Liu[a,*], Meizhi Xu[b]

[a]LMIB and Department of Mathematics, Beijing University of Aeronautics and Astronautics, Beijing 100083, PR China
[b]Department of Mathematics, Tsinghua University, Beijing 100084, PR China

Received 30 October 2006; received in revised form 17 March 2008; accepted 18 March 2008
Available online 26 March 2008

Abstract

A novel kernelized fuzzy attribute C-means clustering algorithm is proposed in this paper. Since attribute means clustering algorithm is an extension of fuzzy C-means algorithm with weighting exponent $m = 2$, and fuzzy attribute C-means clustering is a general type of attribute means clustering with weighting exponent $m > 1$, we modify the distance in fuzzy attribute C-means clustering algorithm with kernel-induced distance, and obtain kernelized fuzzy attribute C-means clustering algorithm. Kernelized fuzzy attribute C-means clustering algorithm is a natural generalization of kernelized fuzzy C-means algorithm with stable function. Experimental results on standard Iris database and tumor/normal gene chip expression data demonstrate that kernelized fuzzy attribute C-means clustering algorithm with Gaussian radial basis kernel function and Cauchy stable function is more effective and robust than fuzzy C-means, fuzzy attribute C-means clustering and kernelized fuzzy C-means as well.

Keywords: Fuzzy clustering; Fuzzy C-means; Attribute means clustering; Kernelized fuzzy C-means

1. Introduction

Based on fuzzy theory proposed by Zadeh [17], fuzzy clustering is a partition method that divides data points into groups (clusters) according to the membership grade or degree. Fuzzy C-means (FCM) is one of the most popular unsupervised fuzzy clustering algorithm, which is widely used in pattern recognition, image recognition, gene classification, *etc*. FCM was first derived from hard C-means algorithm by Ruspini, and then extended by Dunn. Finally, Bezdek extended a general FCM algorithm based on fuzzy weighting exponent m. Hence, Dunn type FCM is a special case of Bezdek type FCM with $m = 2$ [5,1]. Cheng extended the FCM algorithm by introducing stable function, and presented an attribute means clustering (AMC) algorithm, where FCM with $m = 2$ is a special case of AMC [2,3]. Furthermore, AMC algorithm is extended with exponential weight m and a Bezdek type AMC is proposed, called fuzzy fuzzy attribute C-means clustering (FAMC) algorithm [10]. FAMC takes AMC and FCM as its special case, respectively.

Recently, tremendous works focus on using kernel method [15,11–13,6,18–20,7,4], which first maps the data into high dimension space to gain high discriminant capability, and then calculates the measure of the samples in their original data space with Mercer kernel. This trend of kernalization method can be treated as modifying the distance measure of data samples with kernel function. Kernelized FCM (KFCM) is proposed by substituting the Euclidean distance with kernel function. The previous works show that KFCM performs better than FCM [18–20].

* Corresponding author. Tel.: +86 10 82317934.
E-mail addresses: jwliu@buaa.edu.cn, liujingwei03@tsinghua.org.cn (J. Liu).

0165-0114/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2008.03.018

Since FAMC is an extension of both FCM and AMC [10], we replace the Euclidean distance in FAMC with kernel distance and propose kernelized FAMC (KFAMC) algorithm in this paper. To demonstrate the performance of KFAMC, we compare the recognition rate of FCM, FAMC, KFCM, and KFAMC on standard Iris database and tumor/normal gene chip expression data. The experimental results show that FAMC has better performance than FCM, and KFAMC has better recognition performance and robustness than FCM, FAMC, and KFCM algorithm.

The rest of the paper is organized as follows: Section 2 briefly reviews FCM, AMC and FAMC. Section 3 discusses Kernelized FCM and Kernelized FAMC; Section 4 gives the updating scheme of FCM, FAMC, KFCM, and KFAMC; Section 5 introduces the fuzzy decision for pattern recognition; Section 6 introduces the experimental databases and reports the experimental results of comparison of classification accuracy and robustness with FCM, FAMC, KFCM, and KFAMC. And the discussion and conclusion are given in the last section.

2. Brief reviews of FCM, AMC, and FAMC

The general fuzzy C-means clustering algorithm was proposed by Bezdek based on fuzziness degree m [5,1]. By introducing the stable function, an iterative algorithm, AMC, was proposed by Cheng [2], where Dunn type FCM is a special case of AMC. Generalizing AMC algorithm, FAMC algorithm is proposed and it is an extension of both Bezdek type FCM and AMC [10].

2.1. General frame of fuzzy clustering

Suppose $\mathcal{X} \subset \mathbb{R}^d$ is any finite sample set, where $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$, and each sample is $x_n = \{x_{n1}, x_{n2}, \dots, x_{nd}\}$, $(1 \leqslant n \leqslant N)$. The category of attribute space is $\mathcal{F} = \{C_1, C_2, \dots, C_c\}$, where c is the cluster number. For $\forall x \in \mathcal{X}$, let $u_x(C_k)$ denote the attribute measure of x, where $\sum_{k=1}^{c} u_x(C_k) = 1$. Let $p_k = (p_{k1}, p_{k2}, \dots, p_{kd})$ denote the kth prototype of cluster C_k, where $1 \leqslant k \leqslant c$. Let u_{kn} denote the attribute measure of the nth sample belonging to the kth cluster, that is $u_{kn} = u_{x_n}(p_k)$, $U = (u_{kn})$, $p = (p_1, p_2, \dots, p_k)$. The task of fuzzy cluster analysis is to calculate the attribute measure u_{kn}, and decide the cluster which x_n belongs to according to the maximum cluster index $\arg\max_{1 \leqslant k \leqslant c} u_{kn}$.

2.2. Brief review of FCM

Bezdek type FCM is an inner product induced distance based least-squared error criterion non-linear optimization algorithm with constrains,

$$\begin{cases} J_m(U, p) = \sum\limits_{k=1}^{c} \sum\limits_{n=1}^{N} u_{kn}^m \|x_n - p_k\|_A^2 \\ s.t.\ U \in M_{fc} = \left\{ U \in \mathbb{R}^{c \times N} \mid u_{kn} \in [0,1], \forall n, k;\ \sum\limits_{k=1}^{c} u_{kn} = 1, \forall n;\ \ 0 < \sum\limits_{n=1}^{N} u_{kn} < N, \forall k \right\}, \end{cases} \tag{1}$$

where u_{kn} is the measure of the nth sample belonging to the kth cluster. $m \geqslant 1$ is the weighting exponent, also called fuzziness index or smoothing parameter. The distance between x_n and the prototype of kth cluster p_k is as follows:

$$\|x_n - p_k\|_A^2 = (x_n - p_k)^{\mathrm{T}} A (x_n - p_k), \tag{2}$$

The above formula is also called as Mahalanobis distance, where A is a positive matrix. When A is unit matrix, $\|x_n - p_k\|_A^2$ is Euclidean distance, we denote it $\|x_n - p_k\|^2$. Conveniently, we adopt Euclidean distance in the rest of the paper.

The parameters of FCM are estimated by updating $\min J_m(U, P)$ step by step according to the formulas below:

$$\begin{cases} p_k = \dfrac{\sum_{n=1}^{N} (u_{kn})^m x_n}{\sum_{n=1}^{N} (u_{kn})^m}, \\ u_{kn} = \dfrac{(1/\|x_n - p_k\|^2)^{1/(m-1)}}{\sum_{i=1}^{c} (1/\|x_n - p_i\|^2)^{1/(m-1)}} = \dfrac{(\|x_n - p_k\|)^{-2/(m-1)}}{\sum_{i=1}^{c} (\|x_n - p_i\|)^{-2/(m-1)}} \\ \quad n = 1, 2, \dots, N, \quad k = 1, 2, \dots, c. \end{cases} \tag{3}$$

when $m = 2$, FCM is the Dunn type FCM.