文章编号: 1001-0920 (2014) 03-0465-05 **DOI:** 10.13195/j.kzyjc.2012.1778

增强型区间二型FCM算法

邱存勇, 肖 建, 韩 璐

(西南交通大学 电气工程学院，成都 610031)

摘 要: 不确定性存在于图像处理、模式识别等众多领域的实际应用中, 模糊C均值聚类(FCM)算法虽广泛应用于这些领域, 但其处理不确定性的能力较差. 引入区间二型模糊理论能有效提升算法处理不确定性的能力, 但相应地造成算法复杂度增加, 制约了区间二型FCM算法的推广应用. 鉴于此, 提出增强型区间二型FCM算法, 通过优化初始聚类中心和降型运算, 极大地减少了区间二型FCM算法的运算量, 并提升算法的收敛速度. 通过对随机和实际数据的实验比较验证了改进算法的有效性.

关键词: 模糊聚类；模糊C均值聚类；区间二型；降型

中图分类号: TP391.4 **文献标志码:** A

Enhanced interval type-2 fuzzy C-means algorithm

QIU Cun-yong, *XIAO Jian*, *HAN Lu*

(School of Electrical Engineering，Southwest Jiaotong University，Chengdu 610031，China．Correspondent：QIU Cun-yong，E-mail：chaucer_qcy@my.swjtu.edu.cn)

Abstract: There are various uncertainties in the applications in many fields such as image processing and pattern recognition. However, the fuzzy C-means(FCM) algorithm which is widely used in these fields cannot handle the uncertainties well. The introduction of interval type-2 fuzzy theory into FCM can bring such algorithm the ability of handling uncertainties, but the complexity of algorithm will increase accordingly. In order to reduce the complexity, an enhanced interval type-2 FCM algorithm is proposed. The initialization of cluster center and the process of type-reduction are optimized in this algorithm, which can greatly reduce the calculation of interval type-2 FCM and accelerate the convergence of the algorithm. The experimental results on random data and real data show the effectiveness of our proposed algorithm.

Key words: fuzzy cluster；fuzzy C-means；interval type-2；type-reduction

0 引 言

模糊聚类[1]因其无监督特性和易于理解的逻辑语言在图像处理、模式识别、计算机视觉等多领域有广泛的应用, 其中基于目标函数的模糊C均值聚类(FCM)算法[2-3]的理论最为完善, 应用也最为广泛[4-6]. 传统的FCM算法通过迭代更新样本到聚类中心的距离和隶属度实现聚类, 采用取值为[0,1]区间的一型模糊集合表述样本与各聚类中心的隶属关系. 在实际应用中, 聚类过程包含各种不确定性信息, 如距离测量、模糊化、样本获取、算法参数选择等过程中的不确定性因素, 以一型模糊为基础的FCM算法无法很好地处理这些不确定性. 事实上, Zadeh[7]在提出模糊理论时曾提出二型模糊和多型模糊的概念, 通过扩展一型模糊集合得出二型甚至多型模糊集以融入更多的不确定性信息. 之后, Mendel[8]对二型模糊理论进行深入研究, 证明了其理论的有效性和优越性. 目前二型理论已成功应用于很多领域[9-10], 其结论也证明二型模糊理论对于存在不确定性的系统具有优于一型的表现.

二型模糊理论在模糊聚类中的应用还处于起步阶段, 二型的引入使算法运算量呈指数倍增长, 导致二型在模糊聚类的应用受到限制. 已有文献提出用区间二型替代二型模糊集应用于聚类算法, 以提高算法的运算速度[11-12]. 二型模糊集由主、次隶属度函数构成, 通常次隶属度函数也是一种一型模糊集, 区间二型是二型的特殊情况, 其次隶属度函数取值为一特定值(常取1). 次隶属度函数的简化使区间二型模糊算法的复杂度降低, 因此, 区间二型模糊理论较二型理

收稿日期: 2012-11-28；**修回日期:** 2013-04-02.

基金项目: 国家自然科学基金项目(61134001, 51177137).

作者简介: 邱存勇(1986−), 男, 博士生, 从事模糊聚类、模式识别的研究；肖建(1950−), 男, 教授, 博士生导师, 从事模糊理论、控制理论等研究.

论得到更广的研究和应用. 文献[11]提出的区间二型模糊C均值(IT2FCM)算法将传统一型FCM扩展至区间二型, 针对参数存在的不确定性对模糊因子m进行扩展, 有效提升了算法处理不确定性的能力. 然而, 在实际聚类中处理的对象通常为大样本量数据, 应用IT2FCM算法虽然能得到较优的聚类结果, 但其运算速度依然不尽理想. 因此, 针对IT2FCM算法的优化, 提高算法的聚类效果和运算速度十分必要.

IT2FCM与FCM算法的主要区别在于IT2FCM有降型过程, 这是因为IT2FCM迭代后得到的是区间二型模糊集, 这种模糊集无法直接通过解模糊得到聚类结果. IT2FCM采用经典的KM (Karnik-Mendel)降型算法[13]对区间二型模糊集降型, 其迭代步骤多, 运算时间长. 文献[14]提出的EKM (Enhanced KM)算法, 结合模糊聚类的特点对IT2FCM算法的降型过程进行改进, 同时考虑IT2FCM随机选取聚类中心, 使得迭代次数增加并影响最终聚类结果. 本文结合降型算法的特点提出一种简易计算初始聚类中心的方法, 该方法得出的聚类中心可直接用于降型迭代, 有效提高算法收敛速度. 最后通过不同类型数据的实验比较验证了所提出算法的有效性.

1 知识回顾

1.1 区间二型模糊聚类算法

区间二型模糊聚类算法[11]步骤如下.

Step 1: 设定聚类数$c(2 \leqslant c \leqslant N$, N为样本总量), 模糊系数m_1和m_2, 设置迭代停止阈值ε, 初始化聚类中心$\boldsymbol{V}$.

Step 2: 按下式计算上、下隶属度函数:

$$\bar{\mu}_i(k) = \max\Big(\frac{1}{\sum\limits_{j=1}^{c}\Big(\frac{d_{ik}}{d_{jk}}\Big)^{2/(m_1-1)}}, \frac{1}{\sum\limits_{j=1}^{c}\Big(\frac{d_{ik}}{d_{jk}}\Big)^{2/(m_2-1)}}\Big),$$

$$\underline{\mu}_i(k) = \min\Big(\frac{1}{\sum\limits_{j=1}^{c}\Big(\frac{d_{ik}}{d_{jk}}\Big)^{2/(m_1-1)}}, \frac{1}{\sum\limits_{j=1}^{c}\Big(\frac{d_{ik}}{d_{jk}}\Big)^{2/(m_2-1)}}\Big). \quad (1)$$

其中: $d_{ik} = \|x_k - v_i\|_A$(一般$\|\cdot\|$取欧氏距离)为第k个样本x_k到第i个聚类中心v_i的距离, $\bar{\mu}_i(k)$和$\underline{\mu}_i(k)$分别为x_k到聚类中心v_i的上、下隶属度函数.

Step 3: 由KM算法迭代更新聚类中心$\boldsymbol{V}$的区间$[v_L, v_R]$.

Step 4: 取更新后聚类中心$\boldsymbol{V'} = (v_L + v_R)/2$, 若满足$\|\mathrm{V}' - \boldsymbol{V}\| < \varepsilon$则停止迭代, 否则令$\boldsymbol{V} = \boldsymbol{V'}$, 返回Step 2.

1.2 KM迭代算法

KM算法是由Karnik等[13]提出的用于对区间二型模糊集进行降型的迭代算法, 在区间二型模糊聚类中用于计算聚类中心的区间值$[v_L, v_R]$. 以计算v_L为例, 算法步骤如下.

Step 1: 升序排列样本$x_k = (x_{k1}, x_{k2}, \cdots, x_{kM})(k = 1, 2, \cdots, N)$, 如下所示:

$$\begin{aligned} &x_{11} \leqslant x_{21} \leqslant \cdots \leqslant x_{N1},\\ &x_{12} \leqslant x_{22} \leqslant \cdots \leqslant x_{N2},\\ &\quad\vdots\\ &x_{1M} \leqslant x_{2M} \leqslant \cdots \leqslant x_{NM}, \end{aligned} \quad (2)$$

其中M为样本的特征数.

Step 2: 取初始化隶属度函数$\mu_{ik} = (\bar{\mu}_{ik} + \mu_{ik})/2$, 其对应的第$i$类聚类中心为

$$v_i = \sum_{k=1}^{N} x_k\mu_{ik} \Big/ \sum_{k=1}^{N} \mu_{ik}. \quad (3)$$

Step 3: 由$x(s) \leqslant v_i \leqslant x(s+1)$找出切换点$s$, 对应区间值为

$$v_{iL} = \frac{\sum\limits_{k=1}^{s} x_k\bar{\mu}_{ik} + \sum\limits_{k=s+1}^{N} x_k\mu_{ik}}{\sum\limits_{k=1}^{s} \bar{\mu}_{ik} + \sum\limits_{k=s+1}^{N} \mu_{ik}},$$

$$v_{iR} = \frac{\sum\limits_{k=1}^{s} x_k\underline{\mu}_{ik} + \sum\limits_{k=s+1}^{N} x_k\bar{\mu}_{ik}}{\sum\limits_{k=1}^{s} \underline{\mu}_{ik} + \sum\limits_{k=s+1}^{N} \bar{\mu}_{ik}}. \quad (4)$$

Step 4: 若满足$v_{iL} = v_i$, 则停止迭代, 得到v_i的区间最小值v_{iL}; 若不满足, 则取$v_i = v_{iL}$, 返回Step 3. 类似可求得区间最大值v_{iR}.

KM算法通过迭代寻找最优切换点, 其迭代次数受初始值选择和样本数量的影响. 聚类算法的应用对象通常为大样本数据, 且IT2FCM算法的每一次迭代过程均需调用KM算法更新聚类中心, 因此找到提高IT2FCM中降型过程效率的方法是十分必要的.

2 增强型区间二型FCM算法

2.1 初始聚类中心的确定

传统聚类算法随机选择初始聚类中心, 易导致聚类过程收敛于局部极值点, 同时增加了算法的迭代次数. 文献[15-16]提出了计算初始聚类中心的方法, 但这些方法都是针对一型模糊聚类的, 当用于IT2FCM时其运算孤立于降型运算, 无法达到提升算法效率的效果. 下面介绍一种结合IT2FCM降型步骤的简易计算初始聚类中心的方法, 如图1所示.

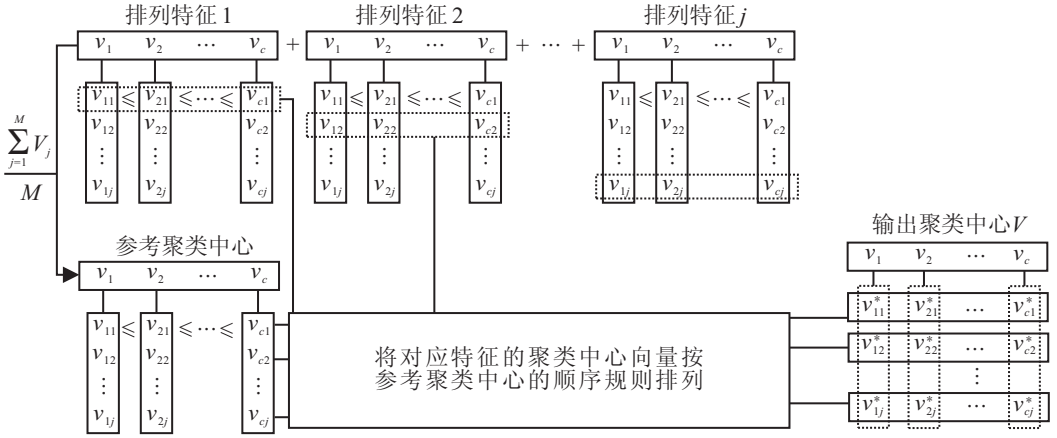


图 1　初始化聚类中心

Step 1: 升序排列样本数据的第 j 个特征值 ($j = 1, 2, \cdots, M$), 即 $x_{1j} \leqslant x_{2j} \leqslant \cdots \leqslant x_{Nj}$, 对应调整其余特征值位置, 记调整后样本为 X_j.

Step 2: 由式 (5) 计算样本 X_j 的聚类中心 V_j, 有

$$\boldsymbol{V}_j = [v_1, v_2, \cdots, v_i] = \left[\begin{array}{c} \dfrac{1}{[N/c]} \displaystyle\sum_{k=1}^{[N/c]} x_{kj} \\ \dfrac{1}{[N/c]} \displaystyle\sum_{k=[N/c]+1}^{2[N/c]} x_{kj} \\ \vdots \\ \dfrac{1}{N-(c-1)[N/c]} \displaystyle\sum_{k=(c-1)[N/c]+1}^{N} x_{kj} \end{array}\right]^{\mathrm{T}}. \tag{5}$$

其中: $[N/c]$ 为对 N/c 向下取整; $v_i = [v_{i1}, v_{i2}, \cdots, v_{ij}]^{\mathrm{T}}$ 为含有 j 个特征值的第 i 类聚类中心, $i = 1, 2, \cdots, c$.

Step 3: 对 Step 2 得出的聚类中心按某特征值进行排序, 如图 1 按特征 1 排序, 按图示方式计算参考聚类中心, 并将对应特征的聚类中心向量按参考聚类中心的排序规则进行排列, 得出初始聚类中心 V.

本文提出的初始化聚类中心方法是对多特征样本在不同特征下的聚类中心作加权平均, 与文献 [15-16] 方法相比, 本文方法以准确度的下降换取运算速度的提升, 但是由于初始化过程中嵌入了降型算法的运算, 在初始化后可直接跳过 KM 算法的 Step 1 和 Step 2, 能够明显提升算法的运算速度, 相比其他初始化方法更适用于区间二型模糊聚类算法.

2.2　改进降型运算

KM 算法由随机初始中心开始迭代过程, 迭代次数受随机中心选取和样本数量的影响, 文献 [11] 对 KM 算法进行分析, 提出了基于经验数据初始化方法的 EKM 算法. 虽然 EKM 有效提高了降型的运算速度, 但其依赖经验选取切换点的初始化方法没有考虑样本的自有特性, 难以通用于处理样本种类繁多的聚类分析. 本文以样本特征的加权平均初始化聚类中心, 对不同类型的样本均具有良好的适应性. 由于初始化过程引入了 KM 算法的排序过程, 在降型过程中可以直接代入初始中心值搜寻切换点, 有效地减少了算法的运算量. 另外, 考虑式 (4) 中 v_{iL}, 定义

$$\sum_{k=1}^{s} x_k \bar{\mu}_{ik} + \sum_{k=s+1}^{N} x_k \mu_{ik} = a, \ \sum_{k=1}^{s} \bar{\mu}_{ik} + \sum_{k=s+1}^{N} \mu_{ik} = b.$$

迭代过程中, 当切换点 s 变至 s' 时, KM 算法由式 (4) 重新计算 v'_{iL}. 对比 v_{iL} 和 v'_{iL} 的表达式, 由于切换点的改变幅度较小, 迭代存在大量的重复计算. 实际计算中只需找出 s 与 s' 之间的差异值即可, 如下所示:

$$a' = a + \mathrm{sgn}(s'-s) \sum_{k=\min(s,s')+1}^{\max(s,s')} x_k(\bar{\mu}_{ik} - \underline{\mu}_{ik}), \tag{6}$$

$$b' = b + \mathrm{sgn}(s'-s) \sum_{k=\min(s,s')+1}^{\max(s,s')} (\bar{\mu}_{ik} - \underline{\mu}_{ik}). \tag{7}$$

对应 v'_{iL} 表达式调整为

$$v'_{iL} = a'/b' = \frac{a + \mathrm{sgn}(s'-s) \displaystyle\sum_{k=\min(s,s')+1}^{\max(s,s')} x_k(\bar{\mu}_{ik} - \underline{\mu}_{ik})}{b + \mathrm{sgn}(s'-s) \displaystyle\sum_{k=\min(s,s')+1}^{\max(s,s')} (\bar{\mu}_{ik} - \underline{\mu}_{ik})}. \tag{8}$$

由式 (8) 可知, KM 算法的降型终止条件 $v_{iL} = v_i$ 等效于判定 $s' = s$, 由此设置降型算法终止条件为: 若满足 $s' = s$, 则停止迭代, 得到 v_i 的区间最小值 v_{iL}; 若不满足, 则取 $v_i = v_{iL}$, $a = a'$, $b = b'$, 返回 Step 3. 类似可求得区间最大值 v_{iR}, 其优化后的表达式为

$$v'_{iR} = \frac{a - \mathrm{sgn}(s'-s) \displaystyle\sum_{k=\min(s,s')+1}^{\max(s,s')} x_k(\bar{\mu}_{ik} - \underline{\mu}_{ik})}{b - \mathrm{sgn}(s'-s) \displaystyle\sum_{k=\min(s,s')+1}^{\max(s,s')} (\bar{\mu}_{ik} - \underline{\mu}_{ik})}. \tag{9}$$

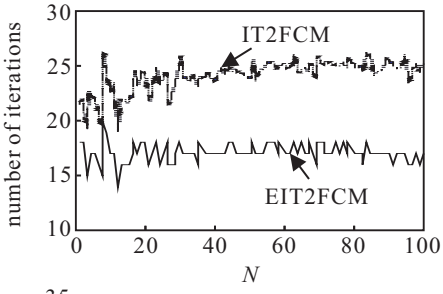
EIT2FCM 算法步骤与 IT2FCM 相同, 只是在迭代中采用第 2.1 节的方法初始化聚类中心, 而降型过程采用第 2.2 节的优化方法进行计算.

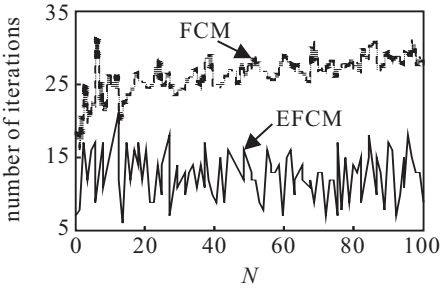
3 实验分析

本文分别对 FCM、EFCM(采用本文初始化方法的 FCM 算法)、IT2FCM 和提出的 EIT2FCM 进行对比, 通过记录其运行时间和迭代次数来验证算法的有效性.

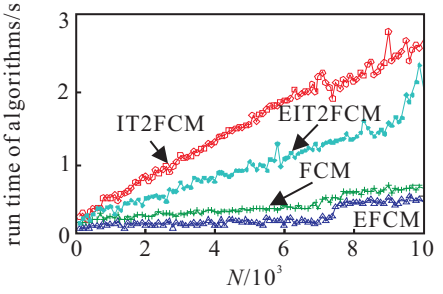
3.1 随机数据比较

选取不同样本量的随机数据进行实验, 取 $N = 100 : 10\,000$, 间隔 100, 特征数取 3. 4 种算法对每个样本 N 进行了 100 次蒙特卡洛实验, 对 100 次实验结果取均值, 各算法迭代次数与运行时间如图 2 所示.





(a) 迭代次数



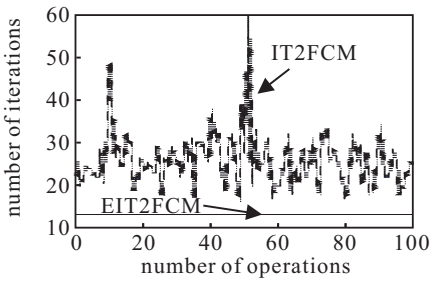
(b) 运行时间

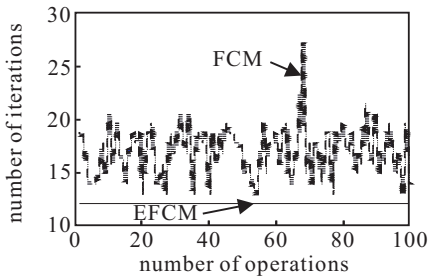
图 2 不同样本量下各算法迭代次数和运行时间比较

3.2 UCI 数据比较

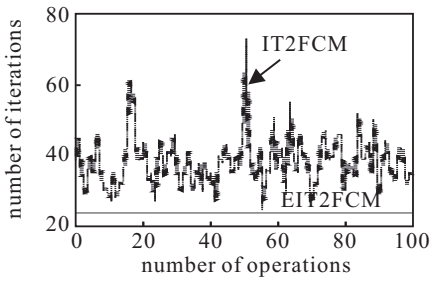
为了验证算法的有效性, 选取 UCI 数据库中较常用的真实数据 Iris 和 Wine 进行实验分析. Iris 和 Wine 分别为 4 个和 13 个特征数的数据, 图 3 和图 4 为 100 次实验记录的算法迭代次数和运行时间对比.

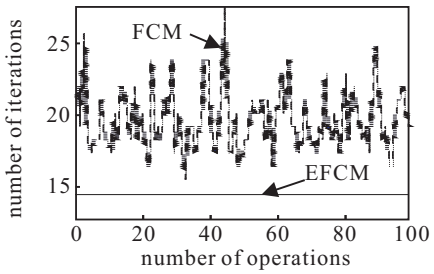
考虑有 M 个特征的样本 N, 单次 FCM 算法所需的运算量为 MN^2, 单次 IT2FCM 运算量为 $2MN^2 + MN(N-1) + TN^2$, 其中各项分别表示两次 FCM 运





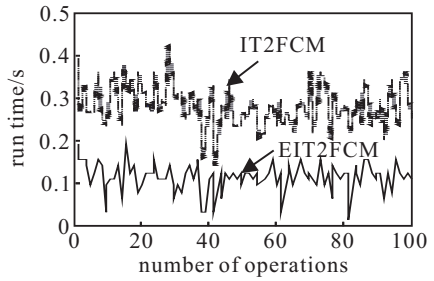
(a) Iris



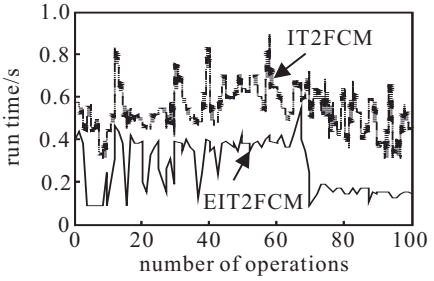


(b) Wine

图 3 Iris 与 Wine 数据各算法迭代次数比较



(a) Iris



(b) Wine

图 4 IT2FCM 与 EIT2FCM 运行时间比较

算、一次数据排序和一次降型运算, T 为 KM 降型算法所需迭代次数. 由于样本量 N 常为较大值, $N^2 \gg N \gg M$, 单次 IT2FCM 运算量可近似为 $3MN^2+TN^2$. 本文算法由于对降型过程进行了优化, 降型所需运算量由 TN^2 降为 $N+T(s'-s)^2$. 考虑 $N^2 \gg N \gg s'-s$, 单次 EIT2FCM 运算量可近似为 $3MN^2$. 若 IT2FCM 算法通过 R 次迭代收敛, 则其总运算量为 $3RMN^2+RTN^2$. 本文采用了初始化聚类中心的方法, 由图 2～图 4 可见, EIT2FCM 较 IT2FCM 迭代次数平均减少约 40%, 因此 EIT2FCM 运算总量可记为 $0.6R(3MN^2)$, 近似于 $2RMN^2$. 图 2 和图 4 的运行时间对比验证了 EIT2FCM 的运算效率明显优于 IT2FCM.

4 结　论

FCM 算法是图像处理、模式识别、计算机视觉等多领域广泛应用的经典算法, 然而其一型模糊隶属度函数无法处理聚类中存在的各种不确定性信息. 区间二型模糊理论的引入使算法处理不确定性的能力增强, 但是相应地算法复杂度增加, 使得区间二型模糊聚类算法的推广应用受到制约.

为了提高区间二型聚类算法的运算效率, 本文提出了一种增强型区间二型模糊 C 均值聚类算法 EIT2FCM. 针对传统模糊聚类算法随机初始化聚类中心导致算法迭代次数增加, 提出了一种加权平均初始化方法, 通过初始化聚类中心有效提高了算法的收敛速度. 另外, 由于区间二型模糊集在处理中需要特殊的降型过程, 不可避免地增加了区间二型算法的复杂度, 通过参考 EKM 算法对传统的降型步骤进行优化, 缩短了降型运算的时间. 实验结果显示, EIT2FCM 有效提高了运算效率, 较 IT2FCM 具有更快的收敛速度.

参考文献(References)

[1] Hopner F, Hoppner F, Klawonn F, et al. Fuzzy cluster analysis: Methods for classification, data analysis and image recognition[M]. Chichester: Wiley, 1999: 5-31.

[2] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy C-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2): 191-203.

[3] 高新波. 模糊聚类分析及其应用[M]. 西安: 西安电子科技大学出版社, 2004: 49-60.
(Gao X B. Fuzzy cluster analysis and its applications[M]. Xi'an: Xidian University Press, 2004: 49-60.)

[4] 朱喜林, 武星星, 李晓梅. 基于改进型模糊聚类的模糊系统建模方法[J]. 控制与决策, 2007, 22(1): 73-77.
(Zhu X L, Wu X X, Li X M. Modeling method of fuzzy inference system based on improved fuzzy clustering arithmetic[J]. Control and Decision, 2007, 22(1): 73-77.)

[5] Zhang D Q, Chen S C. Clustering incomplete data using kernel-based fuzzy C-means algorithm[J]. Neural Processing Letters. 2003, 18(3): 155-162.

[6] Ahmed M N, Yamany S M, Mohamed N, et al. A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data[J]. IEEE Trans on Medical Imaging, 2002, 21(3): 193-199.

[7] Zadeh L A. The concept of a linguistic variable and its application to approximate reasoning[J]. Information Sciences, 1975, 8(3): 199-249.

[8] Mendel J M. Uncertain rule-based fuzzy logic systems: Introduction and new direction[M]. New Jersey: Prentice Hall, 2001: 287-350.

[9] Mendel J M. Type-2 fuzzy sets and systems: An overview[J]. Computational Intelligence Magazine IEEE, 2007, 2(1): 20-29.

[10] 潘永平, 黄道平, 孙宗海. II 型模糊控制综述[J]. 控制理论与应用, 2011, 28(1): 13-23.
(Pan Y P, Huang D P, Sun Z H. Overview of type-2 fuzzy logic control[J]. Control Theory & Applications, 2011, 28(1): 13-23.)

[11] Hwang C, Rhee F C H. Uncertain fuzzy clustering: Interval type-2 fuzzy approach to C-means[J]. IEEE Trans on Fuzzy Systems, 2007, 15(1): 107-120.

[12] 于龙, 肖建, 周聪. 鲁棒区间类型 2 可能性 C 均值聚类[J]. 控制与决策, 2009, 24(4): 503-507.
(Yu L, Xiao J, Zhou C. Robust interval type-2 possibilistic C-means clustering[J]. Control and Decision, 2009, 24(4): 503-507.)

[13] Karnik N N, Mendel J M. Centroid of a type-2 fuzzy set[J]. Information Sciences, 2001, 132(1): 195-220.

[14] Wu D, Mendel J M. Enhanced karnik-mendel algorithms[J]. IEEE Trans on Fuzzy Systems, 2009, 17(4): 923-934.

[15] Yedla M, Pathakota S R, Srinivasa T M. Enhancing kmeans clustering algorithm with improved initial center[J]. Int J of Computer Science and Information Technologies, 2010, 1(2): 121-125.

[16] 刘笛, 朱学峰, 苏彩红. 一种新型的模糊 C 均值聚类初始化方法[J]. 计算机仿真, 2005, 21(11): 148-151.
(Liu D, Zhu X F, Su C H. A novel initialization method for fuzzy C-means algorithm[J]. Computer Simulation, 2005, 21(11): 148-151.)

(责任编辑: 郑晓蕾)