文章编号: 1001-0920(2011)07-1074-05

基于减法聚类和快速紧密性函数的 SF-FCM

李 洪 波

(鲁东大学 信息科学与工程学院，山东 烟台 264025)

摘 要: 首先结合减法聚类和模糊 C-均值聚类各自的优点, 运用减法聚类自适应地确定模糊 C-均值聚类 (FCM) 的初始聚类数; 然后, 提出了改进的紧密性函数, 以此改进用于确定 FCM 聚类结构的有效性函数. 改进后的紧密性函数将对聚类结果贡献不大的数据予以剔除, 使得算法适应能力更强, 执行速度更快. 实验结果表明, 该快速紧密性函数是有效的, 而且计算速度更快.

关键词: 聚类；模糊 C-均值聚类；减法聚类；快速紧密性函数；自适应快速模糊 C-均值聚类

中图分类号: TP181；TP391.4 **文献标识码:** A

SF-FCM based on fast close function and subtractive clustering

LI Hong-bo

(Information Science and Engineer College，Ludong University，Yantai 264025，China. E-mail：fast_run_man@126.com)

Abstract: Firstly, the advantages of subtractive clustering and fuzzy C-means clustering (FCM) are combined to automatically determine the initial number of clusters in FCM. Then, an improved close function which is used in the function of validity to determine the cluster structure of FCM is proposed. The improved close function of cluster validity index ignores the data that have a faint impact on the result of clusters and leads to a stronger adaptability and a faster calculating. The experiment results show that the improved close function is effective and faster for computing.

Key words: clustering；fuzzy C-means clustering algorithm；subtractive clustering；fast close function；self-adaptive faster fuzzy C-means

1 引 言

聚类就是将物理对象或抽象对象组合成为由相似对象组成的多个类或簇的过程. 同一类中的对象尽可能相似, 不同类中的对象尽可能相异, 在这一过程中没有任何关于分类的先验知识指导, 仅以事物间的相似性作为类属划分的准则, 属于无指导分类的范畴[1]. 确定聚类是将对象集合划分成不相交的子集, 一个对象只属于一个聚类. 模糊聚类是将对象划分为若干模糊子集, 其允许对象对于不同的类有不同的隶属度. 模糊聚类由于能够描述样本类属的中介性, 能够客观地反映现实世界, 已逐渐成为聚类分析的主流. 目前, 比较成熟的聚类算法有基于模糊等价关系的传递闭包法、模糊 C-均值聚类 (FCM) 和减法聚类等. 基于目标函数的模糊聚类方法由于具有设计简单、适用范围广, 而且可以转化为优化问题而借助经典数学的非线性规划理论求解等优点, 已成为聚类研究的热点. 其中受到广泛关注的是模糊 C-均值 (FCM) 算法, 它是应用最为广泛且最为灵敏的一种算法.

FCM 的优点是用隶属度的方式表征数据点属于某类的程度, 如果初始的聚类中心和聚类数选择得好, 则可以达到较高的计算精度[2]. FCM 算法实质上是初始聚类中心到聚类结果的映射, 必须提供聚类中心个数, 当算法中初始参数确定后, 聚类的结果便被唯一确定了. 因此, 初始参数值的确定十分重要. 然而, 由于其初始聚类中心是随机生成的, 其收敛性能严重依赖于聚类的初始点, 在传统算法中隶属度函数及其参数的确定主要依靠人的经验, 往往需要反复试凑, 具有很大的主观性和不确定性, 收敛速度随初始点的不同有很大变化, 这便增加了模糊系统的应用难度, 降低了使用效率, 而且可能陷入局部极小点. 这样, 该算法对初始值特别敏感, 很容易陷入局部极小值或者

收稿日期: 2010-04-25；**修回日期:** 2010-11-01.

基金项目: 山东省自然科学基金项目(ZR2010GM013).

作者简介: 李洪波(1969−), 男, 副教授, 从事数据挖掘与商务智能的研究.

鞍点, 而得不到全局最优解; 使用这一聚类算法时, 必须事先指定数据集的聚类数, 然而聚类数 C 一般是很难预先知道的. 为此, 本文结合减法聚类和模糊 C-均值聚类提出一种改进型聚类算法, 并运用一个新的聚类有效性函数达到聚类算法的优化. 对 FCM 算法初始参数之一——聚类数 c 进行自适应确定, 并用减法聚类所得结果作为第 1 次初始化的参数, 对原有的 FCM 具有一定的优化效果.

2　标准 FCM 算法及已有的改进算法

2.1　标准 FCM 算法

标准 FCM 算法[3-4]通过优化目标函数得到每个样本点对类中心的隶属度, 从而决定样本点属于哪个聚类. 此算法是一种划分算法, 目标是使各个分类中的样本到聚类中心的加权距离平方和达到最小. 算法思想如下: 令 $X=\{x_i, i=1,2,\cdots,n\}\in R^s$ 是 s 维向量空间的一个特征向量, c 是预定的类别数目; X_1, $X_2,\cdots,X_c(2\leqslant c\leqslant n)$ 这 c 个子集组成特征向量集 X 的一个模糊划分; 用 u_{ki} 表示特征向量 x_k 属于子集 X_i 的隶属度, 用矩阵 R^{cn} 表示所有的实 $c\times n$ 阶矩阵集合. 约束条件为

$$\begin{aligned} &u_{ik}\in[0,1],\ 1\leqslant i\leqslant c, 1\leqslant k\leqslant n;\\ &\sum_{i=1}^{c}u_{ik}=1,\ 1\leqslant k\leqslant n;\\ &\sum_{k=1}^{n}u_{ik}>0,\ 1\leqslant i\leqslant c. \end{aligned}\tag{1}$$

令 $v=(v_1,v_2,\cdots,v_c)$ 是聚类中心, 其中 $v_i\in R^s$ 是类 i $(1\leqslant i\leqslant c)$ 的中心, 则 FCM 的目标函数可表示为

$$J_m(u,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m||x_k-v_i||^2,\tag{2}$$

式中 $||x_k-v_i||^2=(x_k-v_i,x_k-v_i)$, $1\leqslant m\leqslant+\infty$ 为模糊指数, 通常取 2. 在约束条件下优化目标函数, (u^*, v^*) 是 $J_m(u,v)$ 的局部极小值的必要条件为

$$\begin{aligned} &v_i^*=\frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m},\ i=1,2,\cdots,c;\\ &u_{ik}^*=\frac{1}{\sum_{j=1}^{n}\left(\frac{d_{ik}^*}{d_{jk}^*}\right)^{2/(m-1)}},\\ &\qquad i=1,2,\cdots,c,\ k=1,2,\cdots,n. \end{aligned}\tag{3,4}$$

FCM 算法是基于误差的平方和目标函数准则, 先给出初始方案, 通过式 (3) 和 (4) 反复迭代, 使目标函数 (2) 达到极小. 具体算法如下:

Step 1: 给定聚类数 c, 权指数 m 以及容许误差 $E_{\max}$ 的值, 此时终止条件 $e<E_{\max}$;

Step 2: 选择初始的聚类中心 $v=v_1,v_2,\cdots,v_c$;

Step 3: 用当前的聚类中心根据式 (4) 计算隶属度函数;

Step 4: 用当前的隶属度函数根据式 (3) 更新计算各类聚类中心;

Step 5: 重复 Step 3 和 Step 4 的运算, 直到满足终止条件.

当算法终止时, 便得到了每个类的聚类中心以及每个特征向量对于每个类的隶属度, 从而完成模糊聚类划分. 对于具有 m 个样本且聚类中心数为 c, 用标准 FCM 算法分割时, 如果迭代次数为 p, 则时间复杂度为 $O(pcm^2)$, 而且对初值敏感. 因此, 尽管聚类算法是无指导的算法, 但 FCM 算法要求聚类类别数的先验知识, 否则 FCM 算法会产生误导, 从而破坏了算法的无监督性和自适应性. 当给出的类别数的初值不正确时, 即使使用很好的聚类算法也不会得到最优的聚类结果.

2.2　标准 FCM 算法的已有改进算法

找到正确的聚类数具有非常重要的意义. 一旦得到了一种模糊划分, 便可以应用有效性函数判断这个划分是否反映了数据的真实结构. 有效性函数的有效性反映在紧致性和分离性两个方面, 紧致性表明类内样本的变差或分散的程度, 分离性表明类间的分离程度. Bezdek 针对这个问题提出了聚类有效性问题[5], 即确定数据集的聚类数问题, 并通过定义划分系数, 构造了第 2 个实用的有效性函数. 构造模糊聚类的有效性函数, 目的是更好地确定能够反映数据集真实结构的聚类数, 通常以数据集的紧致性和分离性作为验证聚类有效性的主要特征. 目前, 针对 FCM 算法, 人们已经从不同的角度提出了许多有效性函数, 其中 Xie 和 Beni 的有效性函数是利用隶属度和数据集构造的有效性函数. 用有效性函数确定初始聚类数的改进 FCM 聚类算法的迭代次数明显减少, 收敛速度加快且更稳定. 下面将改进的标准 FCM 算法简称为 IFCM. 根据文献 [6-7] 中的描述, 将 Xie 和 Beni 有效性函数运用于 FCM, 可得到改进的 FCM 算法描述定义如下:

聚类紧密性

$$\mathrm{Comp}=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}||x_k-v_i||^2;\tag{5}$$

聚类分离性

$$\mathrm{Sep}=(d_{\min})^2=\min||v_k-v_i||^2.\tag{6}$$

绝大多数有效性函数都集中刻画模糊划分的两个特征, 即紧致性和分离性. 其中: 紧致性度量是通过计算类内误差来体现类内的分散程度, 当它比较小

时, 意味着同一类中的数据差异较小; 分离性度量是通过计算模糊集之间的距离来体现类间的分离程度, 当它比较大时, 意味着不同类之间的数据差异较大. 紧致性用来表明类内样本的变差或分散程度, 分离性则表明类间的分离程度. 构造有效性函数的目的是寻找使得紧致性最小且分离性最大的最佳聚类数. 聚类有效性函数 S 定义为聚类的紧密性和分离性之比, 即 $S=\mathrm{Comp}/\mathrm{Sep}$.

3 减法聚类与标准 FCM 算法的比较

减法聚类是一种用于估计一组数据中聚类个数以及聚类中心位置的快速单次算法, 由减法聚类算法得到的聚类估计可用于初始化那些有重复优化过程的模糊聚类和模型识别算法[8]. 它假设每个数据都是一个潜在的聚类中心, 在周围数据点密度的基础上, 同时计算出每个数据点确定的聚类中心的可能性的量度. 具体算法如下:

Step 1: 选择最高可能性的点作为第 1 个聚类中心;

Step 2: 移除所有在第 1 个中心附近的数据点, 以确定下一个数据集群及中心位置;

Step 3: 重复上述过程, 直到所有数据都在半径之内.

由于聚类中心的候选集为数据点而非网格点, 减法聚类算法的计算速度非常快. 它的计算量与输入数据的维数无关, 仅与输入数据点的数目成简单的线性关系. 但是, 减法聚类算法所求出的聚类中心在原始的数据点上, 因此对于要求很高精度的场合, 减法聚类算法所求出的聚类中心将达不到精度要求. 目前已有先用减法聚类得到最佳聚类中心个数, 再将聚类中心个数用于 FCM 的应用[4]. 经聚类试验后发现, 聚类中心个数相同时, 经减法聚类后得到的聚类中心与经 FCM 迭代运算后得到的聚类中心相当接近.

4 基于快速紧密性函数的 SF-FCM

首先应用减法聚类得到合理的聚类数, 用于自适应地初始化 FCM 算法, 以得到合理的聚类数, 避免基于有效性函数 FCM 算法不必要的迭代; 然后基于有效性函数的紧密性公式 (5) 和分离性公式 (6) 对紧密性公式 (5) 进行改进, 以便更快速地确定聚类中心. 本文将基于改进紧密性函数的自适应快速 FCM 简称为 SF-FCM.

4.1 SF-FCM 算法的核心思想分析

由于 FCM 适应性强, 已经在图像分割和语音识别等多个领域得到广泛应用[9-14]. 但是, 应用 FCM 算法的前提是必须提供聚类中心个数, 且由于其初始聚类中心是随机生成的, 其收敛性能依赖于聚类的初始点, 收敛速度随初始点选择的不同有很大差异. 用传统试凑法确定聚类数 c, 即由领域专家根据鉴别聚类结果决定是否需要改动初始参数并重新进行聚类, 缺乏科学依据. 在实际问题中, 由于样本量大, 无法有效地确定聚类数 c, 采用 FCM 算法对大样本聚类时将耗费大量的空间和时间在反复确定聚类数 c 上, 并且调整过程很容易陷入局部最优.

目前已有先用减法聚类得到最佳聚类中心个数, 再将聚类中心个数用于 FCM 的应用[10]. 因此, 用减法聚类所得结果作为第 1 次初始化的参数, 从而自适应地确定初始聚类数 c, 并将 Xie 和 Beni 有效性函数运用于 FCM. 这样, 既可以自适应地达到聚类数的确定, 减少由用户确定聚类中心个数的盲目性, 又能快速确定聚类中心点.

因为一个好的模糊划分要求类内样本尽可能地相似, 而类间样本相差尽可能地大, 即要求有较小的紧致性度量和较大的分离性度量, 所以有效性函数的最小值应对应于最佳的聚类数. 即最小化 S 对应于最小化 FCM 的目标函数 J_m, 而且聚类独立性越高, 分离性 Sep 越高, Sep 越大, S越小. 这样, 最小化 S 便代表了一个最有效的最优划分, 由此可确定聚类数 c. 当 c 未达到最优解时, S 随着 c 的增加而减小, 趋向于紧密; 当 c 到达最优值而继续增大时, S 将会由最小值增大. 因此, 取 S 随 c 的增加而成为最小点的值作为聚类数. 由 S 的表达式可知, 当 c 变得很大, 接近于 n 时, 有效性函数存在单调递减的趋势, 容易使 c 的取值达到局部最小值. 因此, 新算法需要对 S 进行改进. 下面详细叙述新算法对 S 的改进.

根据文献 [6], 将新的聚类紧密性 Comp 定义为

$$\mathrm{Comp}=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}\delta_{ik}\frac{||x_k-v_i||}{u_{ik}^m}. \tag{7}$$

当 $u_{jk}>u_{ik}$ 且 $i\neq j$ 时, 函数 $\delta_{ik}=1$; 否则, $\delta_{ik}=0$. 这样可将距离聚类较远的对聚类的紧密性贡献不大的数据删除掉, 而将 u_{ik} 从式 (5) 的分子移到式 (7) 的分母上是为了加强紧密性. 随着 c 的增加, u_{ik} 变大, u_{ik} 越大意味着模式 x_i 越接近于聚类中心 v_i, 而 Comp 越小, 从而使 S 越小, 聚类越有效. 有效性函数 S 定义为

$$S=\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{c}\sum_{k=1}^{n}\delta_{ik}\frac{||x_k-v_i||}{u_{ik}^m}}{\min\limits_{i,k}||v_i-v_k||^2}. \tag{8}$$

由式 (7) 可以看出, 当其是距离聚类较远的对聚类的紧密性贡献不大的数据时, $\delta_{ik}=0$, 此时 $\mathrm{Comp}=0$, 从而可以将这个数据删除掉; 否则, $\delta_{ik}=1$, 对于紧密性可以有新的判断值. 而在式 (7) 中, 新的紧密性定义将某特征向量属于一个子集的隶属度 u_{ik} 从式 (5) 的分子移到式 (7) 的分母上了, 随着 c 的增加, u_{ik} 变

大, u_{ik} 越大意味着 x_i 越接近于聚类中心 v_i, 而 Comp 越小, 从而使 S 越小聚类越有效. 这样做可以加强紧密性的性能. 对于 S, Comp 原本就处于分子的地位, 其值性能的增加不仅不会对原有的目标函数有影响, 相反还会提高算法的优化速率.

综上所述, 利用式 (8) 定义的聚类有效性函数 S, 可以自适应地确定聚类中心的个数 c. 如此得到的值不仅不会影响原目标函数最小化的趋势, 而且不易陷入局部最小值.

4.2　SF-FCM 算法的核非形式化描述

算法处理过程如下:

Step 1: 给定参数模糊指数 m 和减法聚类参数, 调用减法聚类算法进行聚类. 将减法聚类得到的聚类中心 $\phi_i(i=1,2,\cdots,k)$ 赋给 SF-FCM 初始聚类中心作为第 1 次初始化的参数值, 即 $c=k, v_i(1)\approx\phi_i, i=1,2,\cdots,k$.

Step 2: 初始化聚类中心 v_i, $i=1,2,\cdots,c$. 设迭代次数 $p=0$, 计算各个数据到聚类中心的距离 d_{ik}, 计算隶属函数矩阵 $U^{(0)}=(u_{ik}^{(0)})$. 其中: $i=1,2,\cdots,c$; $k=1,2,\cdots,n$; u_{ik} 是矩阵 U 的第 i 行第 k 列元素, 代表第 k 个数据对第 i 个聚类中心的隶属度. 约束关系见式 (1).

Step 3: 重新计算 c 个聚类中心 v_i, $i=1,2,\cdots,c$.

Step 4: 重新计算隶属函数矩阵 U.

Step 5: 计算目标函数 $J_m^{(p)}(u,v)$, 如果 $|J_m^{(p)}(u,v)-J_m^{(p-1)}(u,v)|\leqslant\varepsilon$, 表示收敛, 则此聚类数迭代结束; 否则, $p=p+1$, 转 Step 3.

Step 6: 如果有效性函数 S 达到最小值, 则聚类过程结束; 否则, 聚类数 $c=c+l$, 转 Step 2. 最后取 S 随 c 的增加而成为最小点时 c 的值作为聚类数, 从而得到最优聚类数.

4.3　SF-FCM 聚类结果比较验证

用 Matlab 自带的数据 fcmdata 进行验证, 所得聚类结果比较如图 1～图 5 所示.

从图 1 和图 2 的比较可见, SF-FCM 算法的聚类中心与 IFCM 算法的聚类个数相同, 而聚类中心几乎一致, 说明 SF-FCM 的聚类质量与 IFCM 几乎相同. 但是 SF-FCM 的目标函数达到最小的过程时间比 IFCM 算法少, 而且所达到的目标函数值比 IFCM 算法得到的小很多. 而 IFCM 的目标函数达到最小的过程时间比 FCM 算法少, 而且所达到的目标函数值比 FCM 算法得到的小. 因此, SF-FCM 算法可以更快地达到与 IFCM 同样的效果, 从而找到各个聚类中心. 从图 5 不难看出, SF-FCM 算法收敛速度快且更加稳定, 减少了由用户确定聚类中心个数的盲目性.

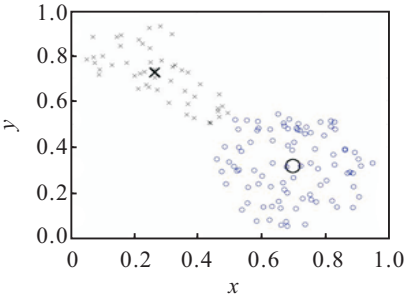

图 1　IFCM 算法聚类结果

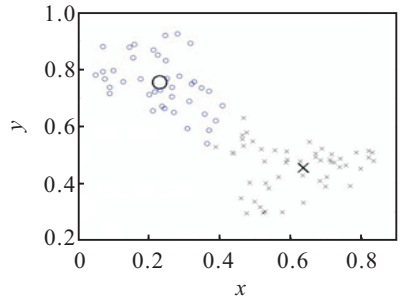

图 2　SFCM 算法聚类结果

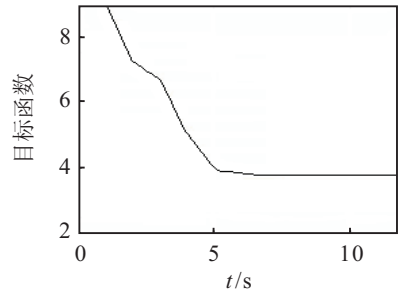

图 3　标准 FCM 目标函数变化曲线

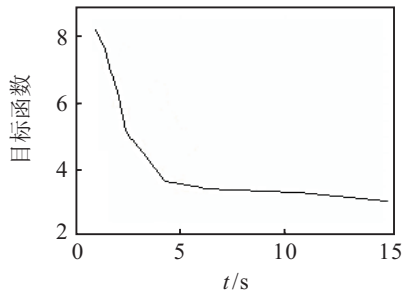

图 4　IFCM 的目标函数变化曲线

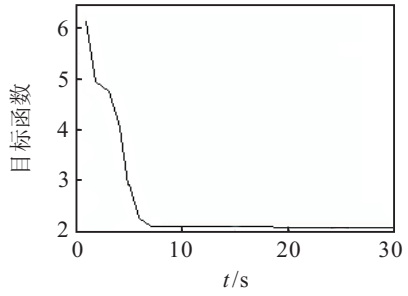

图 5　SF-FCM 目标函数变化曲线

5　SF-FCM 与其他算法性能的比较

聚类算法大致可分为分层聚类和划分聚类两类. 分层聚类方法中含有单连接聚类和全连接聚类, 单连接聚类算法适用于非同心、较分散而又类似链的

数据; 全连接聚类算法适用于同心聚类的数据. 划分聚类方法包括K均值算法、图论方法和FCM方法. K均值算法由于执行速度最快而备受欢迎, 图论方法的典型算法是MST算法. 这些算法的时间复杂度比较如表1所示. 在表1中: m为迭代次数, k为聚类数, n为数据规模, p为剔除无用数据之后的数据规模, $p\leqslant n$. 从表1不难看出, 时间复杂度是指最坏的运行情况, 即使这样, SF-FCM与FCM算法相当, 当SF-FCM的剔除数据较多时, 算法收敛速度明显加快, 时空性能远远优于单连接和全连接算法.

表 1 SF-FCM与其他算法时空复杂度

聚类算法	时间复杂度O	空间复杂度S
leader	kn	k
k-means	knm	k
ISODATA	knm	k
FCM	mkn^2	nk
SSFCM	mkn^2	nk
Shortest spanning path	n^2	n
Single-link	$n^2\log n$	n^2
Complete-link	$n^2\log n$	n^2

6 结 论

从聚类处理的方法看, 聚类分析有两种方法. 一是从数据对象的角度出发, 通过计算它们之间的相似程度(距离)形成合理的对象簇, 然后对这些簇进行概念分析和知识展示, 本文采用的聚类算法就是这种方法. 二是从描述数据对象的数据本身出发, 首先从这些数据中提取出基本概念; 然后对这些概念进行概化, 以形成更高层的概念; 最后将数据对象分到不同的高层概念中, 从而产生最终的聚类结果. 这是聚类分析方法前沿的研究课题, 有着广泛的应用背景. 另外, 基于语义环境的模糊聚类分析方法, 即半指导的聚类方法也值得进一步研究; 如何控制SF-FCM, 使其对噪声不敏感也是具有意义的工作.

参考文献(References)

[1] Krzysztof J Cios, Roman W Swiniarski, Witold Pedrycz, et al. Data mining — A knowledge discovery approach[M]. New York: Springer, 2007: 257-288.

[2] Tong Xiaojun, Li Hongxing, Chen Mianyun, et al. Distributivity and Zadeh's operators[J]. Kybernetes, 2006, 35(10): 1628-1635.

[3] Shitong W, Chung K F, Hongbin S, et al. Note on therelationship between probabilistic and fuzzy clustering [J]. Soft Computing, 2004, 8(5): 366-369.

[4] 同小军, 曾山, 欧军, 等. 两阶段模糊c-2均值聚类算法及其应用[J]. 华中科技大学学报: 自然科学版, 2008, 36(11): 71-75.
(Tong X J, Zeng S, Ou J, et al. Two-stage fuzzy c-mean cluster and its applications[J]. Huazhong University of Science and Technology: Natural Science Edition, 2008, 36(11): 71-75.)

[5] Bezdek J C. Cluster validity with fuzzy sets[J]. Cybernet, 1974(3): 58-74.

[6] 陈春明. 一种改进的模糊C-均值算法[J]. 情报探索, 2009, 13(7): 21-24.
(Chen C M. An improved fuzzy C-means algorithm[J]. Information Research, 2009, 13(7): 21-24.)

[7] 普运伟, 金炜东, 朱明, 等. 核空间中的Xie-Beni指标及其性能[J]. 控制与决策, 2007, 22(7): 829-835.
(Pu Y W, Jin W D, Zhu M, et al. Kernelized Xie-Beni index and its performance[J]. Control and Decision, 2007, 22(7): 829-835.)

[8] Lotfi A Zadeh, Berkeley C A. Fuzzy logic toolbox for use with Matlab version 2.2.1[M]. Natick: The Math Works, Inc, 2005: 135-237.

[9] 冯衍秋, 陈武凡, 梁斌, 等. 基于Gibbs随机场与模糊C-均值聚类的图像分割新算法[J]. 电子学报, 2004, 32(4): 645-647
(Feng Y Q, Chen W F, Liang B, et al. An new algorithm for image segmentation based on gibbs random field and fuzzy C-means clustering[J]. Chinese J of Electronics, 2004, 32(4): 645-647.)

[10] Tran D, Wagner M. Generalized fuzzy hidden Markovmodels for speech recognition[C]. 2002 AFSS Int Conf on Fuzzy Systems. Calcutta: Springer-Verlag GmbH, 2002: 345-479.

[11] 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): 73-75.
(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): 73-75.)

[12] Tong Xiaojun, Lin Yi, Tao Hongjiu. Relashionship betweenent ropy and similarity measure of fuzzy sets[J]. Kybernetes, 2006, 35(9): 1382-1392.

[13] Tong Xiaojun, Chen Mianyun, Li Hongxing. Pan-operations structure with non-idempotent pan-addition[J]. Fuzzy Sets and Systems, 2004, 145(3): 463-470.

[14] Tong Xiaojun, Chen Mianyun, Lin Yi. The structure of pan-addition operator with pre-determined pan-multiplication[J]. Information Science, 2006, 176(3): 321-331.