二维空间中硬聚类算法影响力因子的作用研究*)

金 健 黄国兴 梁道雷
(华东师范大学信息科学技术学院 上海 200062)

摘 要 经典硬聚类算法 HCM (hard c-means)完全基于欧氏距离,针对其无法较好应对各簇规模差异较大的情况,提出在每个欧氏距离项上加入一个影响力因子,使基于距离的标准转变为更通用的基于角度的标准的方法(HCMef算法)。用该算法对二维空间中两类分布密度基本一致,样本数对比分别为 1000∶1000、1000∶5000 和 1000∶10000,正态分布且类边界从较模糊到较清晰的不同数据进行试验。结果显示,HCMef 方法可以很好地找到聚类中心的标准设定值,在各种情况下都有很明显优势,表现出很强的稳定性。表明该方法在二维两类情况下的可行性,并值得做进一步推广研究。

关键词 HCM,聚类,影响力因子

Study on Influence of Effectiveness Factor in HCM Algorithm in 2-D Space

JIN Jian HUANG Guo-Xing LIANG Dao-Lei
(School of Information Science and Technology, East China Normal University, Shanghai 200062)

Abstract The classic hard c-means (HCM) is totally based on Euclid Distance, and it cannot cope with situations of different cluster sizes. Method of attaching an effectiveness factor to each distance item (HCMef) is proposed, transforming the criterion based on distance into the more general criterion based on angle. Two-cluster data sets in 2-D space, normally distributed, with the similar density, the cluster boundaries from vague through clear, and the contrast of cluster population 1000 ∶ 1000, 1000 ∶ 5000 and 1000 ∶ 10000 respectively, are experimented. The results show that HCMef can find the pre-established cluster centers faster and more precisely. The advantages of HCMef are obvious under various situations. The feasibility of HCMef in 2-D space with 2 clusters are verified and the further study is worth performing.

Keywords HCM, Clustering, Effetiveness factor

1 引言

聚类总体来说就是要对空间中掺杂在一起的点集利用算法进行归类,使得类内点彼此更相似,而类间的点彼此不相似。模糊C均值算法 (fuzzy c-means, FCM)是一种基于目标函数优化的分配型聚类算法,是硬聚类算法(HCM)的扩展版本。在FCM的聚类结果中,每个点并不完全属于某个特定的类,而是对某个特定的类有个隶属程度。这更适用于现实应用中不具有明确分界的样本,因此 FCM 算法已成功而广泛地应用到各个领域,例如:分类学、特征分析、模式识别、图像处理、医学及人工神经网络等。

然而,由于经典 FCM 算法完全基于欧氏距离,空间中的某个未知点是否属于某个类,完全由点的位置和聚类中心的位置来决定,而没有考虑类的规模[1]。如果将地球和月亮紧挨着放在一起,则按照经典 FCM 算法,地球上很大一部分土地将被划归月亮所有。而两者的聚类中心也将发生偏移,小类的中心将偏向大类[2~5]。因此,经典 FCM 算法是基于各类样本规模相似的假设的。尽管 FCM 算法的模糊性在一定程度上弱化了聚类中心的偏移及因此带来的误差,但是这种误差在很多场合下还是令人难于接受。

目前已有很多学者提出了 FCM 的改进算法,用于解决上述问题[1]。FCM 算法的模糊性,使得对改进算法的聚类结果评价的客观性难以较好把握。这表现为在评价 FCM 算法的聚类结果时,难以确定客观的聚类中心,尤其是在类别分界不太明显的情况下。并且,即使在已经确定了聚类中心的情况下,要评判聚类结果时还是要把结果去模糊化。例如,将某个样本的隶属度最大值归类于相应的聚类,即,将此最大值看作了 1,而其它值看作为 0。而 FCM 算法的特殊情况——硬聚类算法(HCM),其聚类结果就是 0-1 值[2,6]。在 HCM 算法中,由于其结果的 0-1 特征,确定测试数据的客观的聚类中心将变得更加容易,更有利于对聚类结果的评价。

该方法试图对 HCM 算法进行改进测试,先搞清楚改进项的特性,将更有利于将其应用到 FCM 算法中,而不至于掩盖或模糊化某些重要特性。在文[1]中,作者介绍了带影响力因子的硬聚类算法 HCMef,并对一维空间中两类等间隔分布的小数据集进行了试验。结果显示,在两类靠得比较近,分界不明显时,HCMef 的聚类正确率更高;当两类离得较远时,HCM 和 HCMef 都能正确分类,但 HCMef 所需迭代次数更少,更快地找到了标准的聚类中心。本文在此基础上,继续对此算法进行研究。并将维数扩展到二维平面,采用 1000~10000 不等的较大数据集,且数据分布采用正态分布的随机数据。以验证该算法的可行性及影响力因子在算法执行过程

*)华东师范大学 211 重点项目资助。金 健 博士研究生,研究方向为知识发现与数据挖掘;黄国兴 博士生导师,研究方向为智能信息系统、知识发现与数据挖掘;梁道雷 博士研究生,研究方向为知识发现与数据挖掘。

中对聚类结果的作用。

2 硬聚类算法(HCM)

聚类问题可描述为:设 $X=[x_1^T \quad x_2^T \quad \cdots \quad x_Q^T]^T$,$x_q \in \mathcal{R}^P$ 为 P 维特征空间中的 Q 个样本,其中 x_q 表示第 q 个样本。设 $B=\{\beta_1,\beta_2,\cdots,\beta_C\}$ 是由 C 个类组成的集合,每个类都代表着该类所具有的特征,每个 β_c 都由一组样本组成[7]。

2.1 经典硬聚类算法(HCM)

传统 HCM 算法的训练过程也就是使目标函数最小化的过程,其中目标函数为:

$$\min J_1=\sum_{q=1}^{Q}\sum_{c=1}^{C}\mu_{c,q}^1 d_{c,q}^2 \tag{1}$$

$$s.t.\ \sum_{c=1}^{C}\mu_{c,q}=1 \quad \forall q\in\{1,2,\cdots,Q\} \tag{2}$$

(1)式中,$\mu_{c,q}$ 表示第 q 个样本是否属于第 c 类,其中:

$$\mu_{c,q}=\begin{cases}1 & \text{若第 } q \text{ 个样本属于第 } c \text{ 类}\\ 0 & \text{否则,}\end{cases}$$

$$U=\begin{bmatrix}\mu_{1,1} & \cdots & \mu_{1,q} & \cdots & \mu_{1,Q}\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ \mu_{c,1} & \cdots & \mu_{c,q} & \cdots & \mu_{c,Q}\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ \mu_{C,1} & \cdots & \mu_{C,q} & \cdots & \mu_{C,Q}\end{bmatrix}_{C\times Q}$$

$d_{c,q}$ 是第 q 个样本和第 c 类的聚类中心之间的距离或者两者的相似度,可以根据需要选择不同的度量函数。本文选用的是常用的欧氏距离,即:

$$d_{c,q}=\| x_q-v_c \|_2 \tag{3}$$

其中,$v_c=[v_{c,1} \quad v_{c,2} \quad \cdots \quad v_{c,P}]$ 是第 c 类的聚类中心。

(2)式保证每个样本只能属于一个类。

算法步骤如下[8]:

① 初始化类别数 C,中止迭代误差阈值 ε,最大迭代次数 K,当前迭代次数 $k=0$,初始聚类中心 $U^{(k)}$(通常采用随机数),并利用(4)式计算相应的 $V^{(k)}$;

$$v_c=\frac{\sum_{q=1}^{Q}\mu_{c,q}x_q}{\sum_{q=1}^{Q}\mu_{c,q}} \quad \forall c \tag{4}$$

② $k=k+1$;

③ 利用(3)式计算 $d_{c,q}$ 并利用(5)式更新隶属度矩阵 $U^{(k)}$;

$$\mu_{c,q}=\begin{cases}1 & \text{if } d_{c,q}=\min\limits_{i=1,2,\cdots,C} d_{i,q}\\ 0 & \text{otherwise}\end{cases} \quad \forall c,q \tag{5}$$

④ 利用(4)式计算相应的聚类中心 $V^{(k)}$;

⑤ 判断;

设误差函数为

$e=\| V^{(k)}-V^{(k-1)} \|_2$

若 $e<\varepsilon$,或当前迭代次数 $k>K$,则中止迭代;

否则,转②。

2.2 带影响力因子的硬聚类算法(HCMef)

从传统的 HCM 算法中我们可以发现:$\mu_{c,q}$,即 x_q 是否被归类到 β_c,完全由样本到聚类中心的距离决定。然而,该思想的可行性是建立在每个类的规模(包括密度及样本的空间分布)必须相似的假设条件上的。考虑图 1。

图中有两类边界为圆形的样本,其中心分别为 O_1 和 O_2,半径分别为 r_1 和 r_2。计算点 Ω 到两个聚类中心 O_1 和 O_2 的隶属关系,当满足 $r_1+s'<r_2$ 时,基于距离的传统 HCM 算法会将 O_2 中的点 Ω 划分到 O_1 中。

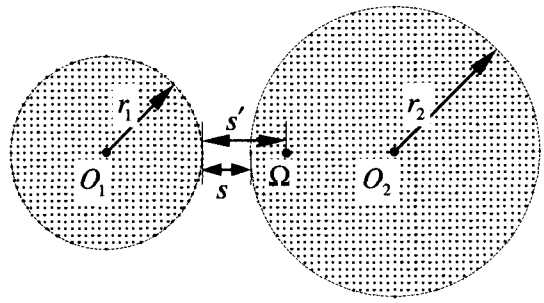


图 1 两不同规模分类分布示意图

针对上面提到的问题,我们相信 Ω 是属于 O_1 还是 O_2,不应只由距离决定,还应考虑分类自身的规模[9]。每个类对类内样本都有某种作用力(见图 2)。即使较短的距离也可能有较弱的作用力。

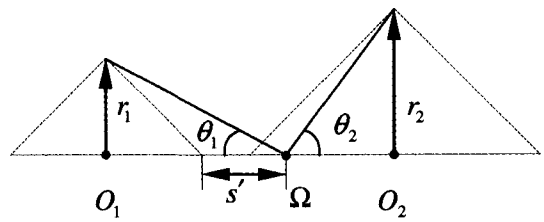


图 2 不同规模的分类简化图

修改目标函数为:

$$\min J_1=\sum_{q=1}^{Q}\sum_{c=1}^{C}\mu_{c,q}^1(\omega_c d_{c,q})^2 \tag{6}$$

仍采用(2)式作为约束条件。这里我们为每个类赋予一个影响力因子 ω_c。可以推导出新的隶属矩阵计算方法:

$$\mu_{c,q}=\begin{cases}1 & \text{if } \omega_c d_{c,q}=\min\limits_{i=1,2,\cdots,C}\omega_i d_{i,q}\\ 0 & \text{otherwise}\end{cases} \quad \forall c,q \tag{7}$$

聚类中心仍使用(5)式计算。

(6)式中,$\omega_c d_{c,q}$ 取代了原来的 $d_{c,q}$。如果类内样本以相同的密度均匀分布,且影响力因子定义为:

$$\omega_c=f(\beta_c)=\frac{1}{|\beta_c|} \tag{8}$$

其中,$|\beta_c|$ 表示第 c 类 β_c 中的样本数,则

$$\omega_c d_{c,q}=\frac{d_{c,q}}{|\beta_c|} \tag{9}$$

(9)可以理解为图 2 中的 $\frac{r}{d}$,即 θ_1 和 θ_2 的正切值。这样,原来基于距离的方法就转变为基于角度的方法。(注:这里可以根据不同条件下的不同数据集选用不同的函数作为影响力因子。)

HCMef 算法与 HCM 算法类似[1]:

① 初始化 $C,\varepsilon,K,k=0,\omega_c^{(k)}$ 和 $U^{(k)}$,并利用(5)式计算相应的聚类中心 $V^{(k)}$;

② $k=k+1$;

③ 利用(3)式计算 $d_{c,q}$ 并利用(7)式更新隶属度矩阵 $U^{(k)}$,再利用(5)式计算相应的聚类中心 $V^{(k)}$,根据当前的 $U^{(k)}$ 得到分类集合 β_c,用(8)式更新影响力因子 $\omega_c^{(k)}$;

④ 判断

若误差 $e<\varepsilon$,或当前迭代数 $k>K$,则中止迭代;

否则,转②。

3 实验

3.1 数据集

现实应用中的数据集往往含有不可意料的噪声，它们适合在测试算法的健壮性时使用，而在测试算法的有效性时会掩盖算法所带来的真实效果。为了能更清楚地查看算法所带来的效果，本文实验采用的实验数据集为人造数据。

文[1]中已经初步得到结论，即，当两类的边界较明显时，HCM 和 HCMef 均能正确分类，只是 HCMef 效率更高些。当两类样本靠得很近，其边界连人眼也很难分辨时，很难对算法聚类效果进行客观评价，对它进行聚类测试的意义显然不大。鉴于以上原因，本实验中的测试数据集只考虑边界由模糊到清晰的三种情况。

为了考察两类数据在不同规模对比度下的聚类效果，本文采用了三种不同的规模对比度，分别为 1000∶1000、1000∶5000 和 1000∶10000。其中，1000∶5000 表示同一数据集中两类样本的样本数分别为 1000 和 5000。由于类别规模应该是类内样本密度和类内样本的空间分布广度的统一体，本文为简便起见，将两类样本的分布密度设成基本一致。这样，考虑类的规模时只需统计类内样本总数即可。样本数据集采用正态分布的随机数，其边界形状呈圆形，因此，若两类样本规模对比为 $Q_1 : Q_2$，则其正态分布的方差为 $\sqrt{Q_1} : \sqrt{Q_2}$。

测试样本数据集共三组，每组有三种不同的边界分布情况。具体参数如表 1 所示(样本分布具体情况见图)。

表 1 三组数据集的参数设置

组号	规模对比	x,y 维 (类 1)		x,y 维 (类 2)	
		均值	方差	均值	方差
1.1	1000∶1000	0	1	2	1
1.2	1000∶1000	0	1	3	1
1.3	1000∶1000	0	1	4	1
2.1	1000∶5000	0	1	4	$\sqrt{5}$
2.2	1000∶5000	0	1	5	$\sqrt{5}$
2.3	1000∶5000	0	1	6	$\sqrt{5}$
3.1	1000∶10000	0	1	5	$\sqrt{10}$
3.2	1000∶10000	0	1	7	$\sqrt{10}$
3.3	1000∶10000	0	1	9	$\sqrt{10}$

表 1 中的类别 2 的均值的取值是经查看数据集的分布图后人工选取的，目的是要使两类的边界具有不同的模糊程度。

3.2 实验方法

在确定了实验数据以后，HCM 和 HCMef 算法需要设置一些初始参数：聚类数 C，中止迭代误差 ε，最大迭代次及 HCMef 特有的影响力因子初始值 $\omega_c^{(0)}$。另外，算法在执行前还要设定一个初始聚类中心 $V^{(0)}$。

本实验的目的是要看 HCMef 与 HCM 的聚类结果的差异，而如何自动找到最佳聚类数目并非本文研究重点，所以本实验中聚类数事先给定[10]，即，$C=2$。

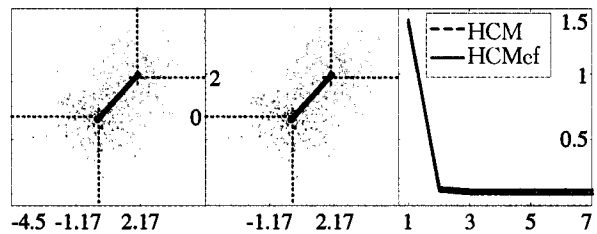
实验中设定的最大迭代次数为 100，为了便于两算法的比较，使它们均有 100 次迭代结果，这里的中止迭代误差阈值设得足够小。

影响力因子在各个类中是通过其值的对比度来发挥作用的，经典的 HCM 算法可以看作 HCMef 的一个特例，即，各类影响力因子恒取 1。HCMef 算法的影响力因子取值为该类中的样本数的倒数，$\omega_c=\frac{1}{|\beta_c|}$。为了防止该值过大或过小，对该取值作适当变换，即令 $\omega_c=\frac{Q/2}{|\beta_c|+Q/2}$。

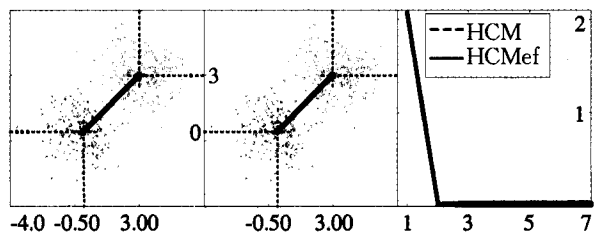
初始隶属度矩阵在满足(2)式的约束下取 0-1 随机数。先确定 0-1 隶属度矩阵后再计算相应的聚类中心，可以防止先取随机聚类中心时带来的定义域过广阔问题。

3.3 实验结果及分析

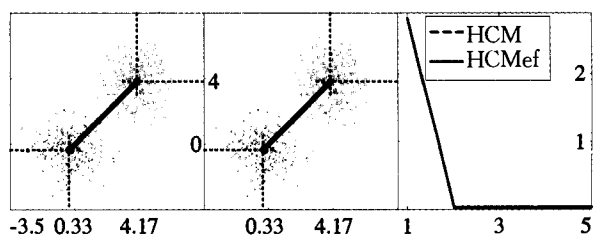
三组数据集的聚类结果如图 3、4、5 所示。为了方便观察，图中样本点只随机显示了总样本数的 1/5。左边两图分别是两类样本(分别用＋和×表示)的分布情况及 HCM 和 HCMef 算法搜索聚类中心的过程，虚线的交点分别表示产生数据集时设定的理想聚类中心；最右图显示的是 HCM 和 HCMef 在搜索聚类中心过程中，随着算法迭代次数的增加，其找到的聚类中心与理想聚类中心的欧氏距离。



(a) 数据集 1.1 实验结果



(b) 数据集 1.2 实验结果



(c) 数据集 1.3 实验结果

图 3 HCM 和 HCMef 搜索聚类中心过程及误差曲线(数据集 1)

第一组数据集中，两类样本的规模相同。从图 3 的测试结果看，HCM 和 HCMef 的效果差别不明显，两者最终收敛到几乎相同的聚类中心。

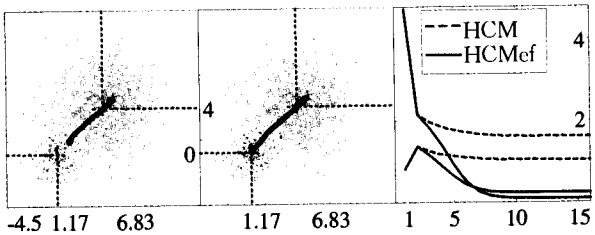
第二组数据集中，两类样本规模之比为 1∶5。图 4 结果显示，两类样本的边界从模糊到清晰的三种情况，HCMef 均能找到离理想聚类中心较近的点；而 HCM 算法找到的结果不甚理想，尤其表现在两类边界较模糊时。

第三组样本规模差别较大，达到 1∶10。当两类靠得很近时，HCM 算法找到的聚类中心较大程度地偏离了理想聚类中心；即使在类别边界较为清晰的数据集 3.3 上面，HCM 算法找到的聚类中心仍有较大偏差。然而，在各种情况下，HCMef 均达到了较高的收敛精度。

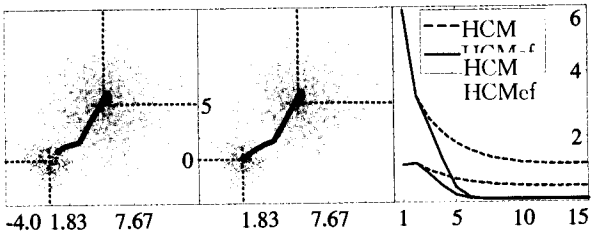
表 2 描述了 HCM 和 HCMef 对各数据集收敛到最终状态时的误差，及到达该最终状态所需的迭代次数。

从表 2 可以看出，在数据规模不同的情况下，HCMef 除

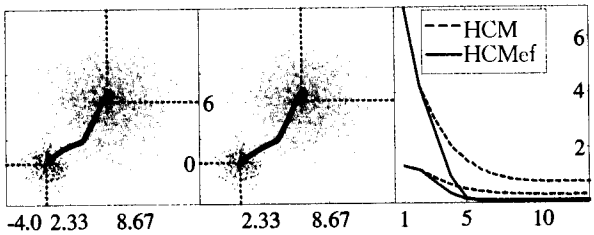
了具有较高的收敛精度外，其收敛速度也明显高于 HCM，当数据规模对比度较大时表现得尤为明显。



(a) 数据集 2.1 实验结果

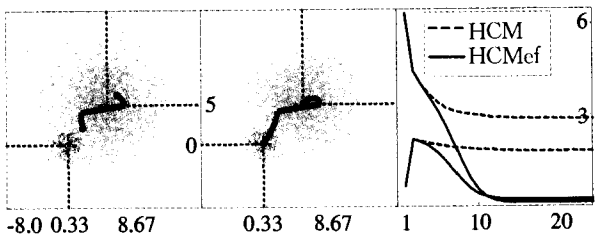


(b) 数据集 2.2 实验结果

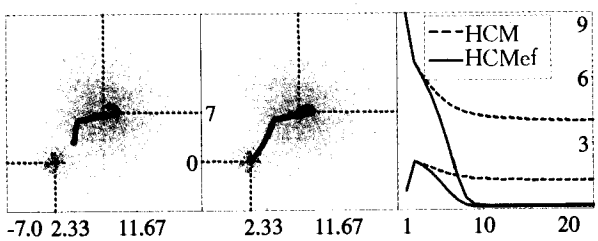


(c) 数据集 2.3 实验结果

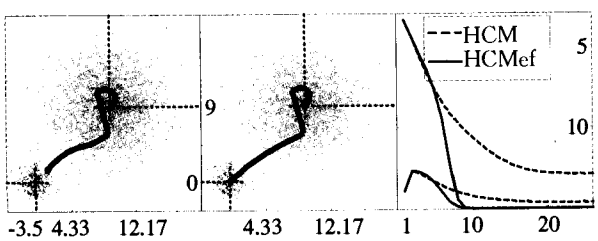
图 4 HCM 和 HCMef 搜索聚类中心过程及误差曲线(数据集 2)



(a) 数据集 3.1 实验结果



(b) 数据集 3.2 实验结果



(c) 数据集 3.3 实验结果

图 5 HCM 和 HCMef 搜索聚类中心过程及误差曲线(数据集 3)

表 2 最终收敛误差(精度为 0.0001)及迭代次数

数据集	HCM				HCMef			
	d_1	K_1	d_2	K_2	d_1	K_1	d_2	K_2
1.1	0.0899	4	0.1145	4	0.0878	5	0.1153	5
1.2	0.0336	5	0.0135	5	0.0341	4	0.0116	4
1.3	0.0343	3	0.0308	3	0.0343	3	0.0308	3
2.1	1.6061	14	1.0278	14	0.0906	12	0.2279	12
2.2	0.4752	13	1.1634	13	0.0936	9	0.0739	9
2.3	0.2928	11	0.7436	11	0.0898	7	0.0537	7
3.1	2.8706	22	1.8036	22	0.126	16	0.2274	16
3.2	4.0465	21	1.2907	21	0.1587	13	0.1098	13
3.3	2.1319	24	0.378	24	0.0386	11	0.0398	11

注：d_c 表示类别 c 的最终状态与其理想聚类中心的误差(欧氏距离)；K_c 表示第一次达到最终状态所需迭代次数。

小结 使 HCM 对不同规模的数据集具有较好的适应性和稳定性，并且推广到 FCM 算法中，可以更大程度上扩大聚类的现实应用范围。本文在二维数据平面上验证了 HCMef 算法在二类随机数据集情况下的聚类效果。在多于二类的情况下，各类别之间均具有相互作用力，即，每个类别与其它类别的影响力对比度不同。如何使该影响力因子具有自适应性并自动调节对不同类别的作用力问题，还有待进一步研究。

参 考 文 献

1 Wang Jing-Jing, Jin Jian, Huang Guo-Xing, et al. Study on Influence of Effectiveness Factor in HCM Algorithm [J]. Journal of Computer Science and Technology. (submitted)

2 Bensaid M, Hall L O, Bezdek J C, et al. Partially supervised clustering for image segmentation [J]. Pattern Recognition, 1996,29: 859～871

3 刘小芳，曾黄麟，吕炳朝. 部分监督加权模糊 C-均值算法的聚类分析[J]. 计算机仿真，2004，22 (3)：114～116

4 钮永莉，陈水利. 模糊 C 均值算法的改进[J]. 模糊系统与数学，2004，18 (增刊)：304～308

5 Miyamoto S, Takata T. A Method of Fuzzy c-Means Using a Quadratic Term for Fuzzification and Control of Cluster Sizes. In: Proc. of 3rd Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty, Osaka Oct. 2000. 2001. 126～131

6 Hall L O, Ozyurt B, Bezdek J C. Clustering with a genetically optimized approach [J]. IEEE Transactions on Evolutionary Computation, 1999,3:103～112

7 Frigui H, Krishnapuran R. A robust competitive clustering algorithm with applications in computer vision [J]. IEEE Transactions on Pattern Analysis, 1999,21:450～465

8 Jang J-S R, Sun C-T, Mizutani E. Neuro-Fuzzy and Soft Computing [M]. Prentice Hall, Upper Saddle River, NJ, USA,1997

9 宋相法，李声威，陈国强，等. 基于改进的 Fuzzy C-means 聚类算法的纹理分割[J]. 河南大学学报(自然科学版)，2005，35 (1)：69～71

10 Gokcay E. A New Clustering Algorithm for Segmentation of Magnetic Resonance Images [D]. A Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy, UNIVERSITY OF FLORIDA