第 24 卷 第 2 期 电子与信息学报 Vol.24 No.2
2002 年 2 月 JOURNAL OF ELECTRONICS AND INFORMATION TECHNOLOGY Feb. 2002

遗传 + 模糊 C-均值混合聚类算法 [1]

陈金山 韦 岗

(华南理工大学电信学院 广州 510641)

摘 要 本文提出了一种新的结合遗传算法 (GA) 和模糊 C-均值算法 (FCM) 的混合聚类算法 (HCA)。它通过对问题的解空间交替进行全局和局部搜索，达到快速收敛至全局最优解，较好地解决了 GA 在达到全局最优解前收敛慢和 FCM 算法容易陷入局部极小的问题。三组不同分布类型的数据聚类实验表明，该算法具有较好的通用性和有效性。

关键词 遗传算法，模糊 C-均值算法，聚类

中图号 TN911.72

1 引 言

聚类分析在众多的领域获得了广泛的应用，并取得了令人满意的效果 [1,2]。聚类分析就是对事物间的相似性进行区分和分类的数学方法。传统的聚类分析是一种硬划分，这种划分的界限是明确的。然而，实际上大多数对象并没有严格的属性，它们的性态和类属存在着中介性，适合进行软划分。由于模糊聚类分析具有描述样本类属中介性的优点，能更客观地反映现实世界，从而成为聚类分析研究的主流。已经提出的模糊聚类方法主要有：基于相似性关系和模糊关系法、基于模糊等价关系的传递闭包法、基于模糊图论的最大树法、基于目标函数法和基于求解等价矩阵与目标函数相结合的摄动模糊聚类法等。

基于目标函数的聚类由于具有设计简单、适用范围广，且可以转化为优化问题来借助经典数学的非线性规划理论求解等优点，成为聚类研究的热点。其中，受到广泛关注的是模糊 C-均值 (Fuzzy C-Means，FCM) 算法 [3,4]。FCM 算法实质上是初始聚类中心到聚类结果的映射，当初值确定后，聚类的结果就被唯一确定。由于目标函数存在许多局部极小点 [5]，而算法的每一步迭代都是沿着目标函数减小的方向进行。所以，若初始化落在一个局部极小点附近，就可能使算法收敛到局部极小值。这在聚类数比较大的情况下尤其突出。针对这种情况，采取的措施有：一是对不同的初始值多次执行该算法，然后从中选取最好的结果，这样不仅费时，而且也不能保证获得全局最优；二是在算法执行前，采用势函数法 [6] 或密度函数法 [7] 预先选择一个合理的初始聚类中心集 $V=\{v_1,v_2,\cdots,v_c\}$，这无疑增加了算法的计算量，而且势函数法和密度函数法都是以样本之间的距离为基础构造的，比较适合于球状分布的数据样本集，而对线状或其它不规则几何分布的数据样本集则可能得不到合理的初始聚类中心集；三是在 FCM 算法中引入全局寻优算法，如 Selim[8] 和 Asultan[9] 等人提出的模拟退火 + 模糊聚类算法，但由于模拟退火算法只有当温度下降足够慢时才能收敛到全局最优点，极大的运算时间限制了其实用性。

遗传算法 [10-12] 是一种无须进行敏度分析的随机全局优化算法，通过选种、杂交和变异，优胜劣汰，经过若干代的进化寻得最优解。本文正是将遗传算法的全局搜索能力和模糊 C-均值算法强大的局部搜索能力有机地结合起来，提出了遗传 + 模糊 C-均值混合聚类算法 (Hybrid Clustering Algorithm, HCA)，以此实现优化聚类算法的性能。第 2 节简要介绍 FCM 算法，第 3 节介绍 HCA 算法的关键问题及其实现，第 4 节为实验及结果讨论，最后是结论。

[1] 2000-04-06 收到，2000-12-27 定稿

国家自然科学基金 (69772027) 及霍英东青年教师基金资助项目

2 FCM 算法

特征空间 $X=\{x_1,x_2,\cdots,x_n\}$ 的模糊 C 划分可用模糊矩阵 $U=[u_{ij}]\in R^{cn}$ 表示，矩阵 U 的元素 u_{ij} 表示第 $j(j=1,2,\cdots,n)$ 个数据点属于第 $i(i=1,2,\cdots,c)$ 类的隶属度，u_{ij} 满足如下条件：

$$\forall j,\sum_{i=1}^{c}u_{ij}=1;\quad \forall i,j\ \ u_{ij}\in[0,1];\quad \forall i,\sum_{j=1}^{n}u_{ij}>0$$

Dunn[3] 定义了一个目标函数 $J_2(U,V)$：

$$J_2(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^2D_{ij}^2(x_j,v_i) \tag{1}$$

其中 $v_i\in R^n$ 为类别中心，$V=\{v_i|v_i\in R^n,i=1,2,\cdots,c\}$，$D_{ij}(x_j,v_i)$ 为数据点 x_j 到聚类中心 v_i 的欧氏距离。Bezdek[4] 将 Dunn 的算法推广到更一般的情况：

$$J_m(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^md_{ij}^2(x_j,v_i) \tag{2}$$

$d_{ij}^2(x_j,v_i)=(x_j-v_i)^TA(x_j-v_i)$，$d_{ij}$ 采用不同的距离定义，可用于不同几何特征的数据聚类分析，目标函数 $J_m(U,V)$ 为每个数据点到相应的聚类中心的加权距离平方和。FCM 算法是一个使目标函数 $J_2(U,V)$ 或 $J_m(U,V)$ 最小化的迭代收敛过程。

应用 Lagrange 乘数法求解在 u_{ij} 满足约束条件下，使 $J_m(U,V)$ 最小的优化问题，可以得到 U，V 的取值公式为

$$u_{ij}=\left[\sum_{k=1}^{c}\left[\frac{d_{ij}}{d_{kj}}\right]^{2/(m-1)}\right]^{-1},\quad v_i=\sum_{j=1}^{n}u_{ij}^mx_j\Big/\sum_{j=1}^{n}u_{ij}^m \tag{3}$$

(2), (3) 式中，$m(m>1)$ 为模糊指数，用来控制分类矩阵 U 的模糊程度。目前 m 的选择大都来自实验或经验，一般取 $1.1\le m\le 5$。当 $m\to 1$ 时，u_{ij} 取值范围由 [0,1] 变为 $\{0,1\}$，FCM 算法退化为 HCM(Hard C-Means) 算法。所以 HCM 算法实际上是 FCM 算法在 $m=1$ 时的特例。

3 HCA 算法

FCM 算法执行过程中，$J_m(U,V)$ 通过 U 与 V 的迭代沿着一子序列逐渐收敛到初始 $V(0)$ 附近的极值点或鞍点 [5]。由于 $J_m(U,V)$ 是一个多峰的复杂函数，因此，$V(0)$ 的选择就变得尤为重要。如果 FCM 算法寻优采取这样的方法：对不同的初始 $V(0)$ 多次执行该算法，然后选择其中最好的结果，就相当于在解空间中进行群体搜索。遗传算法也是在解空间进行群体搜索的，通过遗传操作，群体中的个体得到逐代优化，并逐步逼近最优解。遗传算法的这种全局优化特性，可不断“发现”新的更有希望的搜索区域。但遗传算法的局部搜索能力则不如启发式算法，而 FCM 算法的每步迭代都沿着使 J_m 减小方向进行，有很强的局部搜索能力。HCA 算

法正是结合了两种算法的优点而提出的。其基本思想是：使用 FCM 算法使群体中的每个个体快速趋向各自的极值点，通过遗传算子摆脱个体可能陷入的局部最优，重复进行这样的搜索，直到找到最优解。算法的基本框架如图 1 所示。

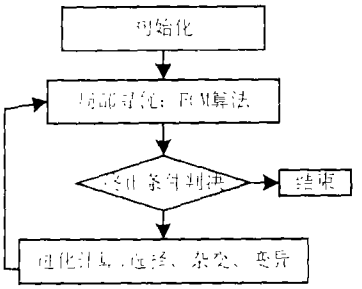

图 1 HCA 算法基本框图

HCA 算法的关键问题及其实现：

(1) 编码 编码的实质是在问题的解空间与算法的搜索空间之间建立一个映射。HCA 算法对聚类中心 $V=\{v_j|v_j\in R^n, j=1,2,\cdots,c\}$ 进行编码，v_j 表示一个个体 V_i 上的一个基因。由于 v_j 用实数表示，所以 HCA 算法采用实数编码。即每个个体用一个 $s=c\times n$ 维的实向量 $V_i=(v_{i1},v_{i2},\cdots,v_{ic})^T\in R^{sn}$ 表示，其中 $a_i\le v_{ik}\le b_i$，b_i、a_i 分别表示 v_{ik} 的上下界；$i=1,2,\cdots,n$；$k=1,2,\cdots,c$。

(2) 产生原始群体 以向量 $V_j(t)$ 表示第 t 进化代群体中第 j 个个体 IND(j) 的基因链所对应的设计变量向量，记第 t 代进化代中包含 n 个个体的群体向量 $V(t)=\{V_1(t),V_2(t),\cdots,V_n(t)\}^T$；随机产生 [0,1] 区间的浮点数逐个填充 n 个个体的基因链。由于产生的随机数服从均匀分布，因此原始群体遍历整个解空间，能充分地反映优化问题解的性态 [12]。

(3) 个体评价 由于优化问题是要找出满足约束条件，同时使目标函数最小的设计向量解，因此，我们建立个体适应值函数为

$$f(V_i(t))=\frac{1}{J_m(U_i(t),V_i(t))+\phi},\quad i=1,2,\cdots,n \tag{4}$$

其中 ϕ 为足够小的正数。

(4) 选种 为避免算法过早收敛和停滞，我们采用了线性排名选择策略：首先假设群体成员按适应值大小从好到坏排列为 $V_1,V_2,\cdots,V_n$，然后根据一个线性函数分配选择概率 P_i：

$$Pi=P(V_i)=[a-bi/(n+1)]/n,\quad i=1,2,\cdots,n \tag{5}$$

其中 a，b 为常数，a 和 b 的取值应满足 $\sum_{i=1}^{n}P_i=1$；对任意的 i，有 $P_i>0$，且 $P_1\ge P_2\ge\cdots\ge P_n$。易知，取 $1\le a\le 2$ 和 $b=2(a-1)$ 可满足上述要求。

选出当前最优个体 $V^*(t)=\arg\max\{f(V_k(t))|k=1,2,\cdots,n\}$。设 $V_0(t)$ 表示当前保留的最优个体，如果 $f(V^*(t))>f(V_0(t-1))$，则更新保留的最优个体，$V_0(t)=V^*(t)$；否则，$V_0(t)=V_0(t-1)$。

(5) 杂交 为了维持群体的多样性和避免算法过早收敛，我们采用近邻配对原则 [13]。这种配对方法不仅可避免较优模式过快地扩散，而且符合基因算法细粒度并行模型的要求，易于获得较大的并行度。

假设 $P(V_i)>P_c$(P_c 为杂交概率)，按近邻配对原则选出的两个父代个体为 Parent(1) 和 Parent(2)，其子代个体为 Child(1) 和 Child(2)，采用整体算术杂交：先生成 c 个 (0,1) 区间的随机数 $a_1,a_2,\cdots,a_c$，然后按 (6)，(7) 式进行杂交操作。

$$\text{Child}(1)\cdot\text{chrom}(j)=a_j*\text{Parent}(1)\cdot\text{chrom}(j)+(1-a_j)*\text{Parent}(2)\cdot\text{chrom}(j) \tag{6}$$

$$\text{Child}(2)\cdot\text{chrom}(j)=a_j*\text{Parent}(2)\cdot\text{chrom}(j)+(1-a_j)*\text{Parent}(1)\cdot\text{chrom}(j) \tag{7}$$

其中 $j=1,2,\cdots,c$。

(6) 变异　根据 (5) 式计算群体中每个个体的入选概率 $P(V_i)$，对于 $P(V_i) < P_m$(P_m 为变异概率，$P_m > 0$) 的个体 Parent(i)，随机选择其某一位基因 j 进行突变：

$$\text{Child}(i)\cdot\text{chrom}(j) = \text{Parent}(i)\cdot\text{chrom}(j) + \delta$$

其中 $\delta = \lambda N(0,1)$，$N(0,1)$ 为标准正态随机变量，$\lambda = K/t$，K 为一常数，t 为迭代的代数。

(7) 算法终止准则　当进化代数达到最大进化代数 G 或结果没有明显的改进时算法终止，即 $|\overline{J}(t) - \overline{J}(t-1)| < \varepsilon$($\varepsilon$ 为一小的正数) 或 $t = G$ 时，算法终止。其中 $\overline{J}(t)$ 定义为 $\overline{J}(t) = \frac{1}{n}\sum_{i=1}^{n} J_m(U_i(t), V_i(t))$。

保留最优个体的 CGA(Canonical Genetic Algorithm) 是全局收敛的 [14]，HCA 在选择操作后保留了最好解，所以 HCA 算法也是全局收敛的。

4 模拟实验与结果讨论

以下给出 HCA 算法和 FCM 算法对三组不同分布类型的二维数据集聚类的仿真实验，实验是在一台 CPU 为 PIII450 的微机上完成的。输入的数据已经归一化，数据集 1 有 140 个数据，数据集 2 有 300 个数据，数据集 3 有 350 个数据，三组数据集的分布情况见图 2, 4, 6。相应的 HCA 聚类结果见图 3,5,7。HCA 算法的有关参数的选择如表 1 所示。

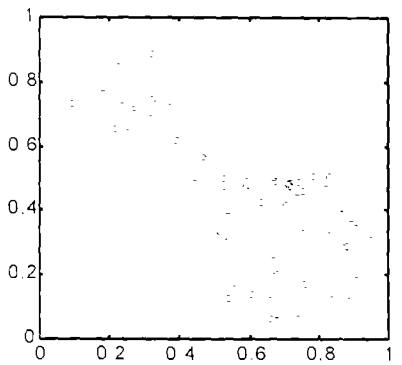

图 2　数据集 1 的样本空间分布图

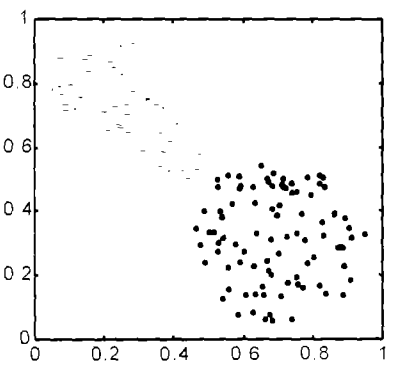

图 3　HCA 对数据集 1 的聚类结果

表 1　参数选择

	n	P_c	P_m	L	G	m	c	ε
数据集 1	30	0.9	0.01	3	150	2	2	10^{-5}
数据集 2	30	0.9	0.01	3	200	2	3	10^{-5}
数据集 3	30	0.9	0.01	3	200	2	4	10^{-5}

表 1 中参数的物理意义：n 为群体规模，P_c 为杂交概率，P_m 为变异概率，L 为 HCA 中局部寻优迭代次数，G 为最大迭代代数，m 为 FCM 算法的模糊指数，c 为聚类中心数，ε 为目标函数的误差精度。而 FCM 算法则采用 Dunn[3] 定义的算法，参数 (c,ε) 的选择与 HCA 算法相同。

对数据集 1，两种算法都运行 20 次，结果 FCM 算法只获得 15 次正确分类，而 HCA 算法获得 20 次正确分类。

我们再对数据集 2 进行同样的实验，结果 FCM 算法获得的正确分类是 13 次，而 HCA 算法则获得 20 次正确分类。

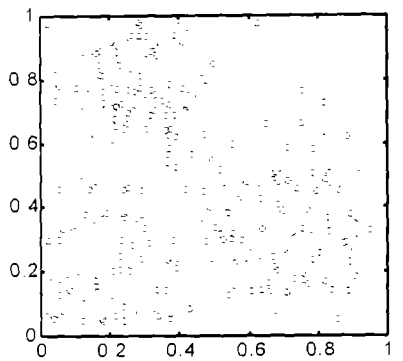

图 4 数据集 2 的样本空间分布图

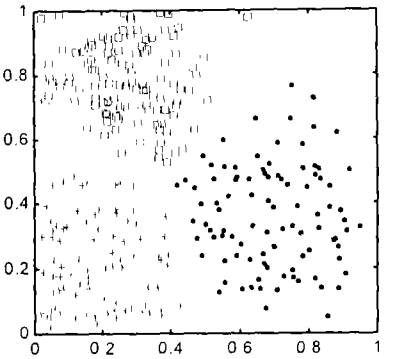

图 5 HCA 对数据集 2 的聚类结果

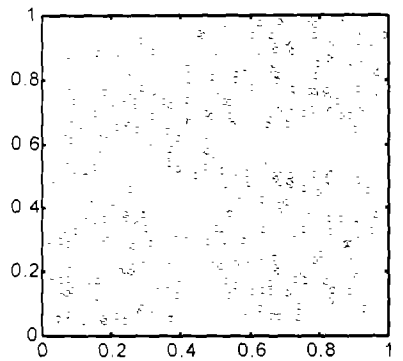

图 6 数据集 3 的样本空间分布图

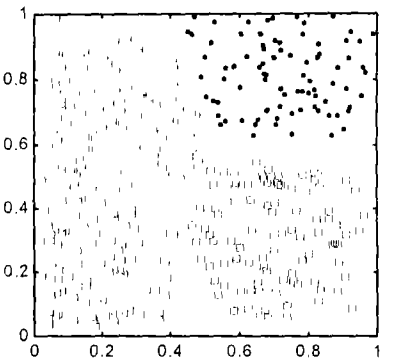

图 7 HCA 对数据集 3 的聚类结果

最后，我们对分布更不规则的数据集 3 进行同样的实验，结果 FCM 算法只获得 9 次正确分类，而 HCA 算法仍获得了 19 次正确分类。

表 2 给出两种算法的平均迭代次数及收敛时间。由表 2 可见，FCM 算法能较快地收敛，这是因为它容易陷入局部最小的结果。

表 2 平均迭代次数及收敛时间

	数据集 1		数据集 2		数据集 3	
	迭代次数	收敛时间 (ms)	迭代次数	收敛时间 (ms)	迭代次数	收敛时间 (ms)
FCM	12.3	135.8	16.05	176.5	22.1	244.1
HCA	13.4	189.3	17	238.9	25.3	332.6

从上述三组实验的结果可见，数据分布越规则、数据点越少，FCM 算法的聚类效果就越好，这与理论分析是一致的。而 HCA 算法的聚类效果受数据集的分布和数据点的多少影响较小，这说明 HCA 算法是稳健的。

5 结 论

HCA 算法利用遗传算法的全局搜索能力来摆脱 FCM 聚类运算时可能陷入的局部极小点，由 GA 得到的最优区域后，再用 FCM 算法快速收敛到最优点。HCA 算法结合了 GA 和 FCM

算法的优点，优化了聚类算法的性能。三组不同分布形式的数据聚类的实验表明，本文的算法是稳健的，并且能够较快地收敛到全局最优解。

参 考 文 献

[1] 高新波，谢维信，模糊聚类理论发展及应用的研究进展，科学通报，1999, 44(21), 2241-2251.

[2] Wu Youshou, Ding Xiaoqing, A new clustering method for Chinese character recognition system using artificial neural networks, Chinese J. of Electronics, 1993, 2(3), 1-8.

[3] J. C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 1973, 3(1), 32-57.

[4] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, Plenum Press, 1981, 43-93.

[5] J. C. Bezdek, R. Hathaway, M. Sabin, W. Tucker, Convergence theory for fuzzy C-means, Counterexample and repairs, IEEE Trans. on SMC, 1987, 17(5), 873-877.

[6] S. L. Chiu, Fuzzy model identification based on cluster estimation, J. Intelligent and Fuzzy Syst., 1994, 2(3), 267-278.

[7] D. Chaudhuri, B. B. Chaudhuri, A novel multiseed nonhierarchical data clustering technique, IEEE Trans. on SMC, 1997, 27(5), 871-877.

[8] S. Z. Selim, K. Alsultan, A simulated annealing algorithm for the clustering problem, Pattern Recognition, 1991, 24(10), 1003-1008.

[9] K. S. Asultan, S. Seltan, A global algorithm for the fuzzy clustering problem, Pattern Recognition, 1993, 26(9), 1357-1361.

[10] 贺前华，韦岗，陆以勤，基因算法研究进展，电子学报，1998, 26(10), 118-122.

[11] 潘正君，康立山，陈毓屏，演化计算，北京，清华大学出版社，1998, 1-43.

[12] 李强，周济，连续解空间的复合遗传算法，科学通报，1998, 43(24), 2662-2668.

[13] 张青富，彭伟，吴少岩等，遗传算法 + 正交设计：一种新的全局优化算法，第 4 届中国人工智能联合学术会议论文集，北京，清华大学出版社，1996, 127-133.

[14] G. Rudolph, Convergence analysis of canonical genetic algorithms, IEEE Trans. on NN, 1994, 5(1), 96-101.

A HYBRID CLUSTERING ALGORITHM INCORPORATING FUZZY C-MEANS INTO CANONICAL GENETIC ALGORITHM

Chen Jinshan　　Wei Gang

(*College of Electron. and Info., South China Univ. of Technology, Guangzhou* 510641, China)

Abstract　A new Hybrid Clustering Algorithm (HCA) that incorporates the fuzzy C-means into the canonical genetic algorithm is proposed in this paper. The HCA speeds up convergence before the genetic algorithm reach the global optima, and eliminates fuzzy C-means trapped local minima by performing global search and local search alternatively. The experiments for clustering three data sets with different distributions show that the HCA has better generalization and effectiveness.

Key words　Genetic algorithm, Fuzzy C-means, Clustering

陈金山：　男，1963 年生，博士后，主要研究方向有神经网络、模式识别、信号处理理论与 ATM 交换技术.

韦　岗：　男，1963 年生，教授、博士生导师，主要研究方向有神经网络、智能信息处理、信号处理、模式识别.