模糊软分类中最佳聚类数的确定①

诸克军，成金华，郭海湘
(中国地质大学管理学院，武汉 430074)

摘要：利用迭代自组织分析技术(ISODATA)和遗传算法(GA)嵌套构成遗传—迭代自组织分析技术(GA-ISODATA)共同执行模糊 C-均值的优化算法，该方法不仅能够在给定预分类数的前提下实现最佳分类，而且在完全不需要人工干预的环境下直接得到模糊 C-均值(FCM)中最佳分类数. 理论证明和实际应用都表明该方法是一种对复杂经济系统实行软分类的有效方法.

关键词：迭代自组织分析技术；遗传算法；最优分类；最佳分类数
中图分类号：O159　　**文献标识码**：A　　**文章编号**：1007-9807(2005)03-0008-07

0 引　言

ISODATA[1]聚类算法是由美国科学家[2,3] Bezdek 提出的一种效果理想的聚类分析方法，只要根据研究对象的具体情况确定预分类数 C 及收敛精度就可以实现对样本的最优划分. 然而，由于系统的高度复杂性和指标的模糊性，常常是事先无法确定预分类数. 即使对系统有一些了解，事先给定预分类数实际上是对算法的一种人工干预，很可能伤害了分类的科学性. 孙才志等[4]讨论了这种聚类算法的最佳聚类数可以通过数理统计中的 F 统计量来确定，并可运用信息论中的模糊熵的指标进行检验. 但 F 统计量的构造仍然是一个值得讨论的问题.

本文根据最优分类的含义构造一个新的适应度函数，运用遗传算法[5]与迭代自组织分析技术嵌套成遗传—迭代自组织分析技术(GA-ISODATA)共同实现模糊 C-均值的优化算法. 该方法不仅能够在给定预分类数的前提下实现最佳分类，而且在完全不知道预分类数的条件下直接得到模糊 C-均值(FCM)中最佳分类及相应的分类数. 由于不需要给定任何预分信息就可实现复杂系统的最优划分，补充和完善了 FCM 理论与 ISODATA 算法，为进一步研究模糊系统的规则数提供了思路. 最后，介绍该方法对我国 31 个地区的生产率进行分类的结果.

1 模糊-C 划分

定义 1　给定 $X = \{X_1, X_2, \cdots, X_N\} \subset R^n$，$R^n$ 表示实数 n 维向量空间，对于 $\forall k, 1 \leq k \leq N$，$X_k = (x_{k1}, x_{k2}, \cdots, x_{kn})^{\mathrm{T}} \in R^n$，其中，$x_{kj}(j = 1,2,\cdots,n)$ 是样本 $X_k(k = 1,2,\cdots,N)$ 的第 j 个属性值. 所谓 X 的一个模糊 C-划分是指

$$F_c = \left\{ U_{c\times N} \in M_{cN} \,\middle|\, \mu_{ik} \in [0,1],\ \forall i,k;\ \sum_{i=1}^{c}\mu_{ik} = 1,\ \forall k;\ 0 < \sum_{k=1}^{N}\mu_{ik} < N,\ \forall i \right\}$$
$$(i = 1,,2\cdots,c;\ k = 1,2,\cdots,N) \qquad (1)$$

其中，M_{cN} 是 $c \times N$ 阶矩阵的集合，μ_{ik} 表示样本 X_k 属于第 i 类的隶属度.

记 $V^{\mathrm{T}} = (V_1, V_2, \cdots, V_c)$ $(V_i \in R^n, i = 1,2,\cdots,c)$ 为聚类中心向量，Bezdek 的 FCM 算法的关键在对于给定的 c，选择隶属度 μ_{ik} $i = 1,2,\cdots,c$；$k = 1,2,\cdots,N$ 和 $V_i, i = 1,2,\cdots,c$ 使得误差函数

① 收稿日期：2003-12-29；修订日期：2005-03-18.
基金项目：国家自然科学基金资助项目(70273044).
作者简介：诸克军(1953—)，男，湖北松滋人，硕士，教授，博士生导师.

$$J_h(U,V,c) = \sum_{k=1}^{N}\sum_{j=1}^{c}\mu_{jk}^{h}d_{jk}^{2} = \sum_{k=1}^{N}\sum_{j=1}^{c}\mu_{jk}^{h}\|X_k - V_j\|^2, 1 \le h \le \infty \tag{2}$$

最小化. 这里 $d_{jk}^2 = \|X_k, V_j\|^2 = (X_k - V_j)^T \cdot (X_k - V_j)$, 且

$$V_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^h X_k}{\sum_{k=1}^{N}(\mu_{ik})^h}, \quad i = 1,2\cdots,c \tag{3}$$

Bezdek 证明了当 $h > 1$, 可用式(3) 及

$$\mu_{ik} = \left(\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{h-1}}\right)^{-1} \tag{4}$$

进行迭代运算, 该运算是收敛的[5~7]. 但 Bezdek 的 FCM 算法没有讨论最佳分类数 c[7~9]. 很明显, 当样本数量 N 比较小, 利用穷举法可以得到最佳分类数, 但当 N 很大时, 穷举几乎是不可能的. 本文主要研究如何确定 c、μ_{ij} 及 V_i 使得(2) 所表示的内中距最小(同类中的样本与其类中心离差的平方和)

$$d(c) = \sum_{i,j=1}^{c}\|V_i - V_j\|^2 \tag{5}$$

与式(5) 表示的类间距最大.

2　算法及其收敛性

2.1　遗传迭代自组织算法(GA-ISODATA)

定义 2　所谓一个模糊 c^* 划分是一个最佳分类(或者 c^* 是最佳分类数), 是指存在 $U^* = (\mu_{ij}^*)_{c\times N}$, $V^* = (V_1^*, V_2^*, \cdots, V_c^*)^T$, $V_i^* \in R^n$ 及自然数 c^*, $2 \le c^* \le N-1$, 使得对于任意的 U, V 和 c 有

$$J_h(U,V,c) + \frac{1}{c}d(c) \ge J_h(U^*,V^*,c^*) + \frac{1}{c}d(c^*) \tag{6}$$

成立.

定义 2 中 $2 \le c^* \le N-1$ 的条件是重要的. 事实上, 当 $c = N$ 时每个样本独立分为一类, 模糊划分成为硬划分, 此时, 隶属度矩阵 $U^{(N)} = I$, $V = X$ 显然有

$$J_h(U,V,c) \ge J_h(I,X,N) = 0;$$

$\forall c$ 且 $d(N) \ge d(c)$　$\forall c$

是一个绝对最优分类. 但这种划分并不是人们所希望的; 当 $c = 1$, 所有样本分为一类, 显然, 无意义. 因此, 最佳分类并不是绝对最优分类.

GA-ISODATA 算法步骤如下:

①随机产生长为 l 的二进制初始 n - 种群 $\overrightarrow{Y_0} = \{Y_1^0, Y_2^0, \cdots, Y_n^0\}$, $k = 0$ 开始;

②计算个体 $Y_i^0 (i = 1,2,\cdots,n)$ 对应的整数 $c_i^0 (i = 1,2,\cdots,n)$;

③对于每一个 c_i^0, 随机产生如式(1) 中的初始分类矩阵 $U_{c_i\times N} = (\mu_{ij})_{c_i\times N}$;

④根据式(3) 计算聚类中心 $V_s^0 (s = 1,2,\cdots,c_i^0)$;

⑤根据式(4) 迭代计算新的模糊矩阵 $U_{c_i\times N}^* = (\mu_{ij}^*)_{c_i\times N}$;

⑥对于预先给定的 $\varepsilon > 0$, 如果 $\max_{ij}[|\mu_{ij}^* - \mu_{ij}|] < \varepsilon$, 计算 V_s^*, 并转 ⑦; 否则返回 ④;

⑦根据式(2) 计算适应度函数值;

⑧$G(U^*,V^*,c_i) = \frac{Q}{J_h(U,V,c)} + \frac{1}{c}d(c)$ $(i = 1,2,\cdots,n)$, 这里, Q 是一个充分大的正数, 令 $G(U^*,V^*,c_i^*) = \min G(U^*,V^*,c_i)$ 称 c_i^* 为当前群体中的最优个体;

⑨保留当前群体中的最优个体(精英保留策略), 对种群实行遗传算法(采用适应度比例方法进行选择, 交叉概率 $p_c \in [0,1]$, 变异概率 $p_m \in (0,1)$) 得到新一代种群; $\overrightarrow{Y_{k+1}} = \{Y_1^{k+1}, Y_2^{k+1}, \cdots, Y_N^{k+1}\}$ 返回 ②直到式(5) 收敛.

2.2　GA-ISODATA 算法的收敛性

定理　如果变异概率 $p_m \in (0,1)$, 交叉概率 $p_c \in [0,1]$, 使用适应度比例方法进行选择并采用精英保留策略, 则 GA-FCM 算法一定收敛到全局最优解.

证明　因为 GA-FCM 算法的 ③—⑥就是 ISODATA 算法, 据 Bezdek 对于选定的 c, 只要 $h > 1$, 存在 $U_{c_i\times N}^* = (\mu_{ij}^*)_{c_i\times N}$ 和 $V_s^* (s = 1,2,\cdots,c)$ 使得

$$J_h(U^*,V^*,c) = \min J_n(U,V,s)$$

令适应度函数

$$f(c) = \frac{Q}{cJ_h(U^*, V^*, c) + \frac{1}{2c}d(c)} \quad (7)$$

其中:Q 是根据数量等级需要选定的正整数. 因为 $2 \leq c \leq N - 1$,所以,集合 $\{J_h(U^*, V_s^*, c)\}$ 和 $\{d(c)\}$ 的元素的个数都是确定而且是有限的,$f(c)$ 的最大值一定存在[10,11]. 因此,对 c,$2 \leq c \leq N$ 进行定理中所示的遗传操作,根据遗传算法收敛原理[2,12] GA-ISODATA 收敛到全局最优值.

3 我国各地区生产力水平的最佳软分类

3.1 样本和指标的选取

若实际产出为 y,有 n 个投入要素,则生产函数的一般形式为[13]

$$Y = F(x_1, x_2, \cdots, x_n; t) \quad (8)$$

它代表产出与投入要素之间的某种依存关系. 为了实现对我国 31 个省、市的生产力水平进行软分类,推广 C - D 函数(柯布 - 道格拉斯) 为如下形式[14]

$$Y = F(x_1, x_2, \cdots, x_n; t) = AK_t^{\alpha}L_t^{\beta}N_t^{\gamma} \quad (9)$$

式中:α、β、γ 是常数,并假设 $\alpha + \beta + \gamma = 1$(即产出的规模效益不变) 这里,$K$、$L$、$N$ 分别代表资本投入,劳动投入,土地(包括环境资源) 投入. 从式(9) 容易得到

$$A = YK_t^{-\alpha}L_t^{-\beta}N_t^{-\gamma} \quad (10)$$

式(10) 表示一个地区的技术进步与该地区的产出,资本投入,劳动投入,土地(包括环境资源) 投入具有一定的依存关系. 正是基于这种依存关系对我国 31 个省、市的技术进步进行软划分.

全国各地区 2001 年的指标数据如表 1:GDP 单位(亿元 / 万人),固定资产单位(亿元 / 万人),单位面积交通(千米 / 万平方千米) 就业人数单位(万人).

表 1 原始数据表

Table 1 The original data

地区	指标				地区	指标			
	GDP	固定资产值	交通	就业人口		GDP	固定资产值	交通	就业人口
北京	2.057 60	1.094 23	8 955.13	629.5	湖北	0.780 30	0.248 79	5 122.40	2 452.5
天津	1.832 80	0.702 19	9 541.97	410.5	湖南	0.603 90	0.178 03	3 744.90	3 438.8
河北	0.832 60	0.285 49	3 583.43	3 379.6	广东	1.368 10	0.44 77	5 684.21	3 962.9
山西	0.544 00	0.202 81	3 861.89	1 412.9	广西	0.466 00	0.136 93	2 663.92	2 543.4
内蒙古	0.650 30	0.211 88	656.25	1 013.3	海南	0.685 90	0.267 99	6 279.48	339.7
辽宁	1.200 10	0.338 86	3 509.17	1 833.4	重庆	0.565 00	0.225 07	4 098.85	1 624.0
吉林	0.755 30	0.260 76	2 406.38	1 057.2	四川	0.511 80	0.187 21	2 091.47	4 414.6
黑龙江	0.934 40	0.252 84	1 620.16	1 631.0	贵州	0.285 60	0.141 09	2 179.82	2 068.2
上海	3.067 40	1.242 03	13 524.41	692.4	云南	0.484 00	0.172 21	4 256.94	2 322.5
江苏	1.293 30	0.383 85	8 154.55	3 565.4	西藏	0.527 50	0.316 58	2 842.96	124.6
浙江	1.462 90	0.614 36	5 465.01	2 772.0	陕西	0.504 00	0.211 38	2 391.98	1 784.6
安徽	0.519 90	0.141 18	5 369.67	3 389.7	甘肃	0.416 50	0.178 78	1 296.35	1 187.2
福建	1.236 50	0.340 96	4 728.01	1 677.8	青海	0.575 40	0.375 43	346.60	240.3
江西	0.519 80	0.150 94	4 090.25	1 933.1	宁夏	0.530 00	0.339 4	1 820.42	278.0
山东	1.043 90	0.308 45	4 865.05	4 671.6	新疆	0.791 80	0.376 33	718.64	685.4
河南	0.590 30	0.161 60	4 444.43	5 516.6					

资料来源:中国统计年鉴 2002

3.2 程序框架

GA-ISODATA 算法的思路是将 ISODATA 算法嵌入 GA 算法中 ISODATA 执行的是任意分类数下的最优分类,而 GA 执行的是在整个空间搜索满足适应度最大时的分类数. 两者各司其责又优势互补. 计算程序框图如图 1:

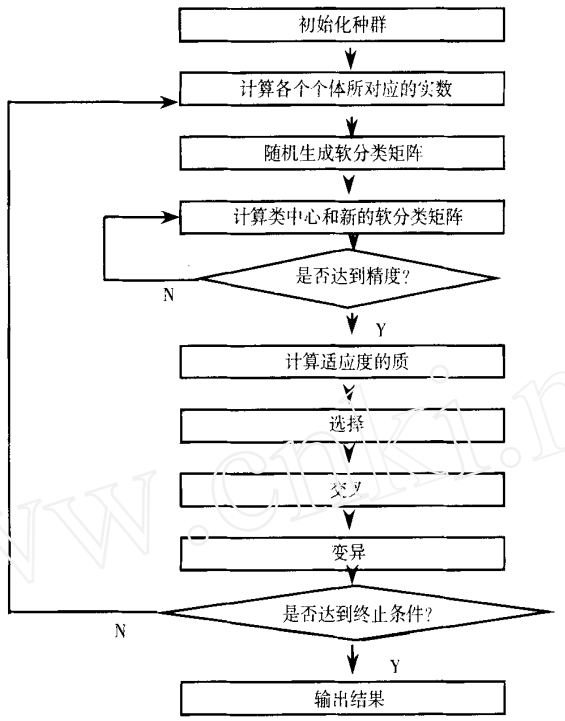

图 1　GA-ISODATA 算法程序框图

Fig. 1 The program flow of GA-ISODATA algorithm

3.3　分类结果

当 $\varepsilon = 10^{-7}$ 时, $h = 2$, $Q = 10^8$, 最佳分类数 $c = 5$, 模糊最优软分类矩阵为

表 2　最佳分类矩阵

Table 2 Matrix of the optimal cluster

地区	类别					地区	类别				
	1	2	3	4	5		1	2	3	4	5
北京	0.042 7	0.870 2	0.017 1	0.045 8	0.024 4	湖北	0.241 2	0.020 0	0.029 0	0.629 1	0.080 7
天津	0.010 1	0.968 9	0.004 4	0.010 6	0.006 0	湖南	0.334 6	0.017 8	0.060 3	0.330 7	0.256 6
河北	0.276 7	0.017 5	0.066 6	0.330 0	0.309 2	广东	0.944 9	0.005 9	0.005 7	0.030 2	0.013 2
山西	0.044 2	0.009 5	0.045 9	0.732 4	0.168 0	广西	0.023 1	0.003 4	0.036 1	0.064 0	0.873 3
内蒙古	0.007 3	0.002 4	0.928 8	0.015 1	0.046 5	海南	0.179 9	0.177 3	0.093 7	0.399 8	0.149 3
辽宁	0.048 5	0.008 3	0.052 6	0.529 7	0.360 9	重庆	0.021 3	0.004 0	0.014 8	0.905 7	0.054 3
吉林	0.033 9	0.009 5	0.307 8	0.124 2	0.524 6	四川	0.218 6	0.028 7	0.157 4	0.200 4	0.395 0
黑龙江	0.024 8	0.006 4	0.463 4	0.061 7	0.443 8	贵州	0.008 9	0.001 8	0.040 7	0.024 8	0.923 8
上海	0.103 5	0.681 6	0.052 9	0.095 2	0.066 7	云南	0.036 1	0.003 6	0.010 7	0.905 8	0.043 7
江苏	0.383 1	0.271 4	0.057 1	0.191 1	0.097 4	西藏	0.062 1	0.024 3	0.360 1	0.229 6	0.324 0
浙江	0.494 0	0.028 1	0.031 2	0.367 3	0.079 4	陕西	0.007 8	0.001 7	0.037 2	0.026 4	0.926 9
安徽	0.856 5	0.009 9	0.011 6	0.092 2	0.298 0	甘肃	0.007 7	0.002 3	0.894 0	0.018 4	0.077 5
福建	0.037 6	0.007 0	0.014 6	0.899 5	0.041 3	青海	0.019 0	0.007 6	0.847 3	0.038 2	0.087 9
江西	0.010 4	0.001 5	0.005 5	0.959 4	0.022 3	宁夏	0.023 6	0.008 8	0.743 2	0.065 6	0.158 8
山东	0.850 1	0.012 9	0.019 0	0.072 2	0.045 8	新疆	0.004 5	0.001 6	0.957 9	0.009 5	0.026 6
河南	0.630 2	0.035 1	0.057 6	0.151 8	0.125 3						

根据最大隶属度原则,其分类结果为表 3 所示:

表 3 各省市(自治区)生产力水平分类表

Table 3 The level of productivity cluster about each region of China

类 别	地 区 名 称
1 类	江苏、浙江、安徽、山东、河南、湖南、广东
2 类	北京、天津、上海
3 类	内蒙、黑龙江、甘肃、青海、宁夏、新疆
4 类	山西、辽宁、福建、江西、湖北、海南、重庆、云南
5 类	河北、吉林、广西、四川、贵州、西藏、陕西

最大适应度值为

$f(5) = 1.263\ 9 \times 10^{-8}$

GA-ISODATA 算法的最大特点是在事先不知道分类数的情况下自动进行最佳分类. 改进了软分类中需要给定预分数,然后实现最佳分类的做法. 根据表 1 的原始数据它将我国生产力分为发达地区(第 2 类);沿海及中部地区(第 1、4 类);以及西部边缘地区(第 3、5 类). 虽然,某些省份隶属相应类别的程度较低,例如湖南省、河南省对于相应类别的隶属程度分别只有 0.334 6 和 0.309 2(见表 2),不足以反映这些地区生产力的真实情况,这是由于指标体系的完整性和统计数据的准确性原因所造成的,完全可以通过增加统计指标和进一步核准统计数据来调整分类,使其更加符合客观情况.

4 结 论

所谓最优分类是指在给定分类数的前提下作出一个划分,使得类中距(式(2))最小. 本文所研究的是在不知道预分数的前提下,寻求一个最佳分类数,并且保持在这个分类数下按式(6)实现最小. 一般地讲,随着分类数的增加,内中距下降而内间距增加,表 4 列出了各个分类数对应的类中距、类间距以及适应度值. 图 2 描述了类中距与类间距的变化轨迹. 图 3 显示了在最佳准则下的最优值及对应的最佳类数. 因此,要求一个绝对最优分类数(同时满足内中距最小,内间距最大的分类数)是不可能的,事实上,只有当分类数等于样本数时才能同时达到最优. 但是可以根据自己的偏好选择权重实现内中距和类间距两者的调和满意解. 显然,最佳分类数的确定为描述模糊系统的模糊规则数提供了重要的依据.

表 4 类中距(ICD1),类间距(ICD2)和对应的适应度的值(FV)

Table 4 Inter-distance(ICD1), intra-distance(ICD2) and the value of fitness(FV)

	ICD1	ICD2	FV		ICD1	ICD2	FV
2	1.222 2e + 008	1.565 9e + 007	7.252 9e - 009	17	2.860 6e + 006	1.865 0e + 008	5.280 9e - 009
3	5.735 9e + 007	3.673 4e + 007	1.062 77e - 008	18	5.878 1e + 006	1.147 4e + 008	8.290 5e - 009
4	3.789 2e + 007	4.510 8e + 007	1.204 82e - 008	19	2.061 5e + 006	1.975 2e + 008	5.010 5e - 009
5	2.966 0e + 007	5.077 9e + 007	1.243 18e - 008	20	1.696 3e + 006	2.169 2e + 008	4.574 1e - 009
6	2.326 0e + 007	6.640 7e + 007	1.115 23e - 008	21	1.947 3e + 006	2.084 3e + 008	4.753 3e - 009
7	1.873 1e + 007	7.041 5e + 007	1.121 75e - 008	22	1.408 9e + 006	2.224 7e + 008	4.466 7e - 009
8	1.612 0e + 007	7.038 3e + 007	1.156 03e - 008	23	1.065 9e + 006	2.210 8e + 008	4.501 5e - 009
9	1.333 1e + 007	8.495 3e + 007	1.017 46e - 008	24	8.479 6e + 005	2.277 7e + 008	4.374 1e - 009
10	7.498 6e + 006	1.493 0e + 008	6.377 5e - 009	25	8.926 4e + 005	2.314 7e + 008	4.303 7e - 009
11	1.083 2e + 007	8.828 5e + 007	1.008 9e - 008	26	5.618 3e + 005	2.581 9e + 008	3.864 7e - 009
12	6.541 8e + 006	1.467 4e + 008	6.523 9e - 009	27	3.411 7e + 005	2.638 1e + 008	3.785 7e - 009
13	8.373 7e + 006	1.000 4e + 008	9.224 1e - 009	28	2.443 6e + 005	2.726 8e + 008	3.664 1e - 009
14	7.584 0e + 006	1.080 5e + 008	8.647 7e - 009	29	6.000 5e + 004	2.911 6e + 008	3.433 8e - 009
15	7.700 0e + 006	1.046 4e + 008	8.901 8e - 009	30	1.411 3e + 004	3.055 0e + 008	3.273 2e - 009
16	3.032 3e + 006	2.049 8e + 008	4.807 3e - 009				

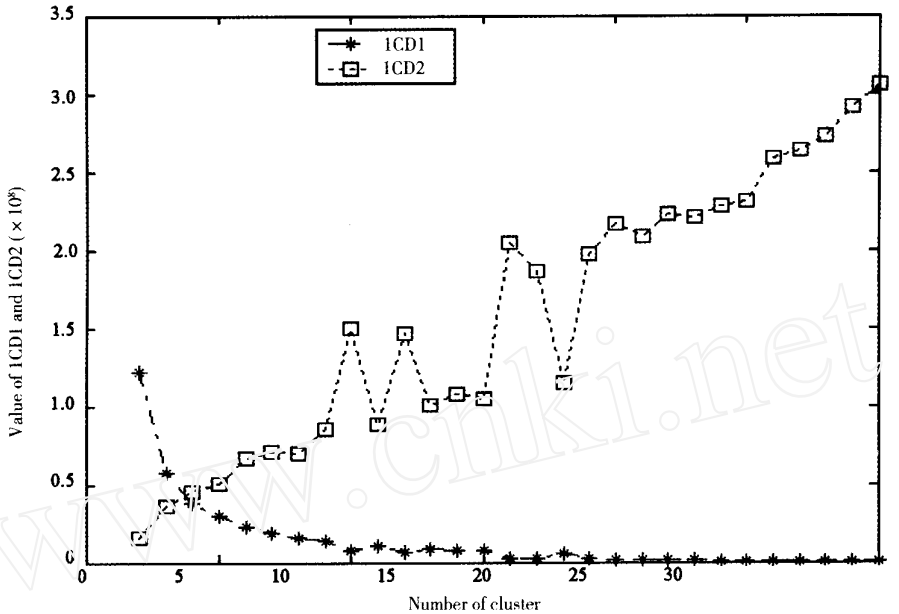

图 2　类中距(ICD1)与类间距(ICD2)变化轨迹

Fig. 2 The change track of inter-distance (ICD1) and intra-distance (ICD2)

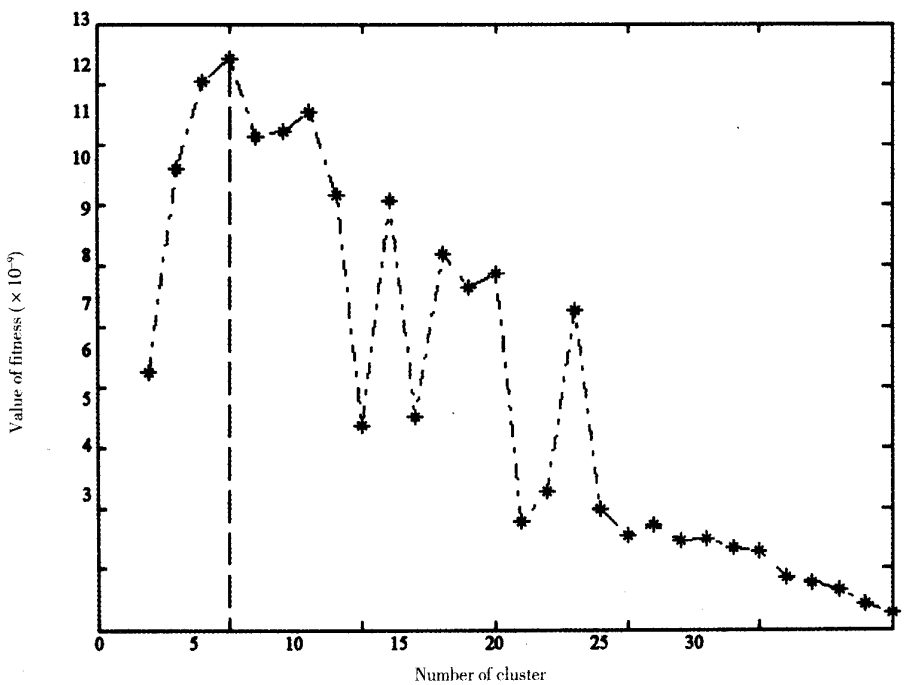

图 3　各个类别对应的适应度值

Fig. 3 The value of fitness corresponding to each cluster

参 考 文 献:

[1]郭嗣琮, 陈 刚. 信息科学中的软计算方法[M]. 沈阳: 东北大学出版社, 2001. 68—72.
Guo Sicong, Chen Gang. Method of Soft Computing in Information Science[M]. Shenyang: Press of Northeastern University, 2001. 68—72. (in Chinese)

[2]李敏强, 寇纪淞, 林 丹, 等. 遗传算法的基本理论与应用[M]. 北京: 科学出版社, 2002. 17—66.
Li Minqiang, Kou Jisong, Lin Dan, *et al*. Theory and Application of Genetic Algorithm[M]. Beijing: Press of Science, 2002. 17—66. (in Chinese)

[3]Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York: Plenum Press, 1981.

[4]孙才志, 王敬东, 潘 俊. 模糊聚类分析最佳聚类数的确定方法研究[J]. 模糊系统与数学, 2001, 15(1): 89—92.
Sun Caizhi, Wang Jingdong, Pan Jun. Research on the method of determining the optimal class number of fuzzy cluster[J]. Fuzzy Systems and Mathematics, 2001, 15(1): 89—92. (in Chinese)

[5]Bezdek J C, Hathaway R J. Local convergence of the fuzzy c-means a birtns[J]. Pattern Recognition Letter, 1986, 19(6): 237—246.

[6]Bezdek J C, Nikhil R Pal. Cluster Validation with Generalized Dunn's Indices[C]. Proceedings of the 2nd New Zealand Two-Stream International Conference on Artificial Neural Networks and Expert Systems, 1995. 190—193.

[7]Hathaway R J, Bezdek J C. Optimization of clustering criteria by reformulation[J]. IEEE Transactions Fuzzy Systems, 1995, 3(2): 241—245.

[8]Hall L O, Ozyurt B, Bezdek J C. Clustering with a genetically optimized approach[J]. IEEE Trans. on Evolutionary Computation, 1999, 3(2): 103—112.

[9]Suzuki J. A Further Result on the Markov Chain Model of Genetic Algorithms and its Application to a Simulated Annealing-like Strategy[C]. IEEE Trans. Sys., Man and Cybern., Part B: Cybernetics, 1998, 28(1): 168—175.

[10]吴 菲, 黄梯云. 遗传算法在基于二元决策数的模型迭代中的应用[J]. 管理科学学报, 1999, 2(2): 57—61.
Wu Fei, Huang Tiyun. The application of genetic algorithm in model selection[J]. Journal of Management Sciences in China, 1999, 2(2): 57—61. (in Chinese)

[11]程乾生, 王守章, 武连文. 基于遗传算法的多阶马氏链组合预测法[J]. 管理科学学报, 1998, 1(4): 27—33.
Cheng Qiansheng, Wang Shouzhang, Wu Lianwen. The combined prediction method based on genetic algorithm and multiple-order Markov chains[J]. Journal of Management Sciences in China, 1998, 1(4): 27—33. (in Chinese)

[12]张文修, 梁 怡. 遗传算法的数学基础[M]. 西安: 西安交通大学出版社, 2000. 157—159.
Zhang Wenxiu, Liang Yi. The Mathematical Base of Genetic Algorithm[M]. Xi'an: Press of Xi'an Jiaotong University, 2000. 157—159. (in Chinese)

[13]Solow R M. A contribution to theory of economic growth[J]. Quarterly Journal of Economics, 1956, 39(1): 222—232.

[14]Romer P. Endogenous technological change[J]. Journal of Political Economy, 1990, 98(5): 71—102.

Optimal number of clusters in fuzzy soft clustering

ZHU Ke-jun, CHENG Jin-hua, GUO Hai-xiang

School of Management, China University of Geosciences, Wuhan 430074, China

Abstract: This paper uses ISODATA and GA to construct the GA-ISODATA nestingly, in order to perform the optimization algorithms of the FCM at the same time. This method can not only complete the optimal partition on the promise of giving the number with pre-classification, but also directly get the optimal number with classification in FCM without people's engagement. Theory prove and practical application both demonstrate that this method is a valid way to realize soft-partition in complicated economic system.

Key words: GA-ISODATA; genetic algorithms; the best class; optimal class number