

基于自适应遗传算法的椭圆聚类方法

任建斌, 杨 明, 孙凌燕

(中北大学 理学院, 山西 太原 030051)

摘 要: 提出一种利用自适应遗传算法对椭圆聚类进行检测和变量估计的新方法. 该算法利用自适应遗传算法对目标函数进行优化, 既提高了速度, 又避免了陷入局部极小值. 通过对几个同心和交叉椭圆的合成数据进行测试, 验证了该方法的有效性.

关键词: 模糊聚类; 椭圆聚类; 自适应遗传算法

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2010)04-0058-04

Fuzzy Clustering of Elliptic Ring-shaped Clusters Employing Adaptive Genetic Algorithms

REN Jian-bin, YANG Ming, SUN Ling-yan

(College of Sciences, University of North China, Taiyuan 030051, China)

Abstract: A new method, which is employing Adaptive Genetic Algorithms (AGA), is developed for the detection and parameter evaluation of elliptic ring-shaped clusters. The algorithm is based on optimization of an objective function employing AGA, through which we can not only speed up the clustering, but avoid being trapped into minimum. The algorithm is also validated through testing various forms of synthetic data, such as multiple intersecting ellipses, concentric ellipses, incomplete ellipses, and etc.

Key words: fuzzy clustering; elliptic ring-shaped clusters; adaptive genetic algorithms (AGA)

1 引言

聚类分析作为一种无监督学习模式, 被广泛用于模式识别的检测和多维数据之间内部结构的描述, 以及图像处理的各个方面, 如图像分割、边缘检测等等. 模糊聚类最著名的算法之一是 1973 年 Bezdek 给出的模糊 C-均值算法^[1]. 该算法及其派生算法已被成功运用于相交或不相交的紧凑聚类. 但是由于这些方法采用传统的最优化方法求解模型, 往往只能得到模型的局部极值点, 从而影响聚类的效果, 甚至使所采用的方法失效. 这一问题对于非凸集合性质的数据聚类 (如壳聚类等) 显得无能为力. 为此, 一些基于模糊 C-均值算法的目标函数的改进算法被提了出来, 如模糊 C-壳算法^[2] (FCS), 模糊 C-球壳聚类 (FCSS), 自适应模糊 C-壳聚类^[3]

(AFCS), 模糊 C-椭球面聚类 (FCES), 模糊 C-椭圆聚类^[4] (FCE) 等, 有针对地弥补了 FCM 算法的不足.

基于优化目标函数的模糊 C-椭圆聚类 (FCE) 是一种针对解决多个相交或同心椭圆 (或圆环) 聚类问题的比较好的方法. 然而该方法需借助于 FCM 和 FCR^[5] 两种算法进行初始值设定和对目标函数优化, 这样使得算法冗长且容易陷入局部极小值, 增加了计算的复杂度. 文中在此基础上, 引进现代全局优化方法之一的自适应遗传算法^[6-7] (AGA) 对目标函数进行优化, 解决了上述 FCE 算法的不足之处. 用相关数据进行测试, 表明新方法对球壳形数据有令人满意的聚类效果.

2 椭圆状数据的模糊聚类算法 FCE

模糊 C- 椭圆聚类算法(FCE)^[4] 是对目标函数进行优化,即聚类原型变量是目标函数的最优解.和传统表示椭圆不同的是,Gath 和 Hoory 采用三个简化变量来表示聚类原型,即两个焦点 $v_i^{(0)}$ 和 $v_i^{(1)}$,以及半径 $r_i = 2a_i$ (a_i 表示椭圆的长半轴).设球壳状分布的数据集 $X = \{x_1, x_2, \dots, x_n\}$ 中含有 c 个椭圆状的类,它们的聚类原型为 $(v_i^{(0)}, v_i^{(1)}, r_i), i = 1, 2, \dots, c$. U 是隶属度矩阵.算法的过程是对下面目标函数进行优化的过程:

$$J_{FCE}(U, V^{(0)}, V^{(1)}, R) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m D_{ij}^2, m \in (1, +\infty) \quad (1)$$

其中

$$D_{ij} = \|X_j - v_i^{(1)}\| + \|X_j - v_i^{(2)}\| - r_i = d_{ij}^{(1)} + d_{ij}^{(2)} - r_i \quad (2)$$

FCE 算法可描述如下:(1) 初始化:确定类别数 c , ($2 \leq c \leq n$) 模糊指数 m ,迭代阈值 ϵ ,设置迭代计数器 $t = 0$.

(2) 利用下式确定椭圆半径 r_i :

$$r_i^{(t)} = \frac{\sum_{j=1}^n (u_{ij})^m (d_{ij}^{(1)} + d_{ij}^{(2)})}{\sum_{j=1}^n (u_{ij})^m}, \quad \forall i, 1 \leq i \leq k \quad (3)$$

式中, $d_{ij}^{(1)}$ 和 $d_{ij}^{(2)}$ 分别是第 j 个模式到第 i 类的两个焦点的欧氏距离. u_{ij} 是第 j 个模式到第 i 类的隶属度.

(3) 利用下式更新焦点:

$$(v_i^l)^{(t)} = \frac{\sum_{j=1}^n (u_{ij})^m (x_j + (\|x_j - v_i^{1-l}\| - r_i) \frac{x_j - v_i^l}{\|x_j - v_i^l\|})}{\sum_{j=1}^n (u_{ij})^m}, \quad l = \{0, 1\} \quad (4)$$

(4) 更新隶属度矩阵 U :

$$u_{ij}^{(t)} = \frac{1}{\sum_{k=1}^c (\frac{D_{ij}}{D_{ik}})^{\frac{2}{m-1}}}, \quad \forall i, j, 1 \leq i \leq c, 1 \leq j \leq n. \quad (5)$$

(5) 若 $\|u^t - u^{(t+1)}\| < \epsilon$, 则停止循环,否则转向(3).

3 自适应遗传算法与 FCE 相结合的球壳聚类方法(AFCE)

自适应遗传算法是在标准遗传算法基础上的一种改进算法.在进化初期,遗传算子应确保种群在大范围内搜索,进行全局优化以避免过早收敛;进化后期,当逼近最优解时,种群就应该在局部范围内搜索,重点进化,尽可能提高解的精度.这就要求遗传算子必须根据种群的进化情况,随时调整进化策略.遗传算法的收敛性主要取决于作为其核心操作的交叉算子和变异算子,因为交叉提供了全局搜索能力,而变异则提供了局部搜索能力.据此,自适应遗传算法根据适应值和进化代数来调节个体的交叉率和变异率^[6]:

$$p_c = \begin{cases} p_{c-\max} - (\frac{p_{c-\max} - p_{c-\min}}{it_{\max}}) * it_{er}, & f' > f_{avg} \\ p_{c-\max}, & f' > f_{avg} \end{cases} \quad (6)$$

$$p_m = \begin{cases} p_{m-\min} + (\frac{p_{m-\max} - p_{m-\min}}{it_{\max}}) * it_{er}, & f' > f_{avg} \\ p_{m-\min}, & f' > f_{avg} \end{cases} \quad (7)$$

式中, p_c 表示交叉概率, $p_{c-\max}$ 表示最大交叉概率, $p_{c-\min}$ 表示最小交叉概率, p_m 为变异概率, $p_{m-\max}$ 表示最大变异概率, $p_{m-\min}$ 表示最小变异概率, $it_{\max}$ 为最大代数, it_{er} 为当前代数, f_{avg} 表示种群平均适应度值, f' 表示在要交叉的两个个体中较大的适应度值, f 表示要变异的个体适应度值.

这种交叉概率、变异概率调节方法在进化初期有较强的全局搜索能力和较弱的局部搜索能力,随着进化的进行,全局搜索能力逐渐减弱,局部搜索能力逐渐增强.此改进措施有助于保护优良个体的有效模式,便于找到全局最优解,防止“早熟”现象出现.该改进措施还有效保证了遗传操作所产生的新个体都是优良的,明确了进化的方向,加快了遗传算法的进化速度,避免了进化过程中个别个体退化现象.

基于以上考虑,将自适应遗传算法(AGA)与模糊 C- 椭圆聚类(FCE)有效结合起来,给出一种解决此类问题的新方法,即基于自适应遗传算法的椭圆聚类方法(AFCE) 在准备应用遗传算法求解问题时,要完成以下几个主要步骤^[7]:

(1) 确定问题的表示方案.即采用什么样的编码串表示问题,这将直接影响选择、杂交、变异等算子的设计,也影响算法的复杂性.对于模糊椭圆聚类,需要求解的变量有模糊隶属矩阵 U 以及聚类原

型(焦点 $v_i^{(0)}, v_i^{(1)}$ 和类半径 r_i), 由于待求解变量众多, 若对所有变量都进行编码势必影响聚类速度, 因此, 我们选择只对聚类原型进行编码, 且采用实数编码方案. 这样我们就可以很方便地根据聚类原型利用式(3) 计算相应的模糊隶属矩阵 U .

(2) 确定适应值度量方法. 本文要求的是问题中目标函数的最小值, 由于该值不小于 0, 因此可直接用目标函数 J_{FCE} 的倒数 $\frac{1}{J_{FCE}}$ 作为个体适应值.

(3) 确定控制算法的参数和变量. 包括种群规模 N 、杂交率 p_c 、变异率 p_m 、最大进化代数 T 等.

(4) 确定指定结果的方法和停止运行的准则. 在所给算例中, 停止准则简单地用最大进化代数来控制.

具体过程如下:

(1) 初始化: 给定聚类类别数 $c (2 \leq c \leq n)$ 和模糊指数 $m (1 < m < 2.5)$, 给定种群规模 N , 最大进化代数 T . 设置进化迭代计数器 $t = 1$;

(2) 随机产生 $2N$ 个个体作为初始种群, 其中个体 $P = (v_1^{(0)}, v_1^{(1)}, r_1, v_2^{(0)}, v_2^{(1)}, r_2, \dots, v_k^{(0)}, v_k^{(1)}, r_k)^T$, 用有限步的 FCE 算法更新每个个体, 然后计算每个个体的适应值 $F(p) = \frac{1}{J_{FCE}}$, 选择适应值较小的个体作为第代种群;

(3) 对当前种群中的每个个体, 以式(5) 计算交叉概率 p_c , 根据其值确定是否参与杂交运算, 并对所选出的染色体(个体) 进行相邻两个个体的杂交运算. 这里 $p_{c-\max}$ 取 0.9, $p_{c-\min}$ 取 0.1. 具体方法为: 设 P_1 和 P_2 是两个需要作杂交运算的染色体, 它们作算术杂交运算后得到的新染色体记为 P'_1 和 P'_2 , 则

$$\begin{cases} P'_1 = P_1 + (1 - r_1)(P_2 - P_1) \\ P'_2 = P_2 + (1 - r_2)(P_1 - P_2) \end{cases} \quad (8)$$

式中, r_1, r_2 是 $(0, 1)$ 中的随机数.

(4) 对所有染色体(包括当前种群中的个体和新产生的个体) 的每个基因位(个体分量), 以式(6) 计算变异概率 p_m , 根据其值确定是否参与变异运算. 这里 $p_{m-\max}$ 取 0.4, $p_{m-\min}$ 取 0.01. 对于给定的父亲 $P = (\xi_1, \dots, \xi_v)$, 若它的第 j 个基因 ξ_j 被选来变异, 则产生的新一代染色体为 $P' = (\xi_1, \dots, \xi_{j-1}, \xi'_j, \xi_{j+1}, \dots, \xi_v)$, 其中, ξ'_j 随机地按如下两种可能的取值变化.

$$\xi'_j = \xi_j + \Delta(t, \xi_j^U - \xi_j) \quad (9)$$

或

$$\xi'_j = \xi_j - \Delta(t, \xi_j - \xi_j^L) \quad (10)$$

式中, ξ_j^U 和 ξ_j^L 分别是 ξ_j 的上下界, t 是当前代数. 函数 $\Delta(t, \xi)$ 给出 $[0, \xi]$ 间的一个值, 使得 $\Delta(t, \xi)$ 随着 t 的增加而趋于 0. 例如可取为

$$\Delta(t, \xi) = \xi \cdot r \cdot \left(1 - \frac{t}{T}\right)^b \quad (11)$$

式中, r 是 $(0, 1)$ 中的随机数, T 是最大代数, b 是确定非均匀度的参数, 后面算例中取为 0.8, b 的大小决定了 $\Delta(t, y)$ 随着 t 的增加而趋于 0 的速度.

(5) 对(3)、(4) 中新产生的个体进行有限步的 FCE 运算以改善它们的适应值, 然后计算它们的适应值.

(6) 比较当前所有个体的适应值(包括父代个体), 选取适应值较大的个体作为第 $t+1$ 代种群. 令 $t \leftarrow t+1$.

(7) 当 $t > T$ 时, 终止进化, 输出输出最优个体. 否则转向(3) 继续迭代.

4 数据实验

实验 1: 这个实验展示了对没有噪声点干扰的四个交叉和同心椭圆的识别(见图 1). 图 1 中给出了四个椭圆的原始数据, 每个椭圆都由 100 个数据点构成. 图 1 中用 FCE 算法检测出了三个椭圆的大概位置, 由于交叉点的干扰, 使得识别出的椭圆中心都比原始数据有所偏移. 图 1 中用 AFCE 算法检测出了三个椭圆的准确位置. 从图中可以看出, 两种算法检测的结果有很大差异. 而 AFCE 算法由于采用了全局寻优和局部寻优相结合的方法, 使得聚类能

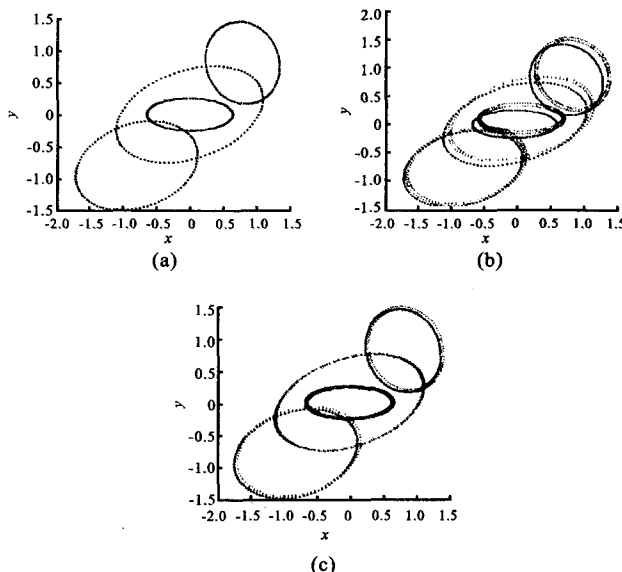


图 1 没有噪声点干扰的四个交叉和同心椭圆的识别

得到令人满意的效果。

实验2:这个实验展示了对四个有噪声点干扰的交叉椭圆的识别(见图2)。图2(a)中给出了四个椭圆的原始数据,每个椭圆都由200个数据点构成。

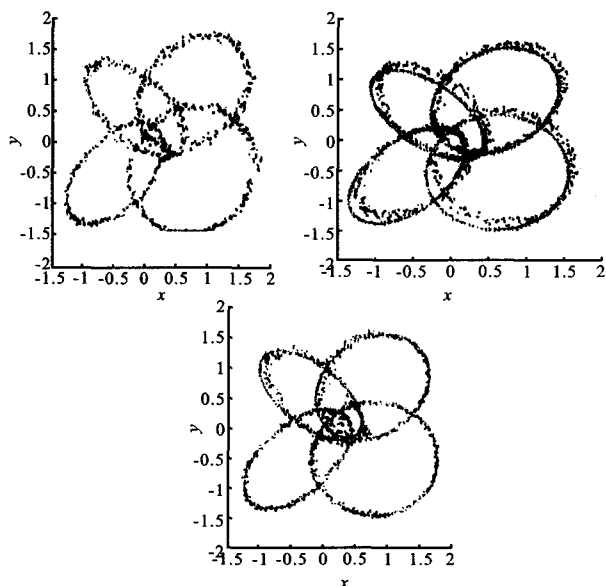


图2 有噪声点干扰的四个交叉和同心椭圆的识别

图2(b)中用FCE算法检测出了三个椭圆的大概位置。从图中可以看出,由于噪声点和交叉部分的干扰,使得识别出的椭圆都比原始数据有所偏移。图2(c)中用AFCE算法检测出了三个椭圆的准确位置。表1列出了对这三个椭圆各参数的估算值。

5 结束语

文中主要分析了模糊椭圆聚类算法FCE的缺点,以及产生这些缺点的原因所在,即FCE在初值确定和基于梯度法的交替寻优策略过度依赖FCM算法及FCR算法,这必然导致FCE算法在实践中往往得不到令人满意的聚类效果。基于此,我们将原FCE算法与遗传算法进行巧妙地结合起来,产生出所谓的基于自适应遗传算法与FCE相结合的模糊球壳聚类算法AFCE。数值试验表明,文中所提出的AFCE算法具有较好的抗噪性,它解决了FCE依赖FCR算法的问题,并且在优化过程中避免了传统方法易陷入局部极小值的缺点。经数据实验对比,证明了该方法的有效性。

表1 FCE和GFCE算法对图2所示数据的聚类结果

类 编 号	真值				FCE 聚类结果				GFCE 聚类结果			
	长轴	短轴	倾斜角	中心	长轴	短轴	倾斜角	中心	长轴	短轴	倾斜角	中心
1	0.95	0.55	45	(-0.6, -0.6)	1.0	0.55	45	(-0.6, -0.6)	0.95	0.55	45	(-0.6, -0.6)
2	0.95	0.47	-45	(-0.3, 0.4)	0.95	0.47	-45	(-0.3, 0.5)	0.98	0.47	-45	(-0.3, 0.4)
3	0.95	0.75	45	(0.6, 0.6)	0.75	0.59	45	(0.6, 0.6)	0.75	0.50	45	(0.6, 0.6)
4	0.95	0.95	-70	(0.6, -0.6)	0.95	0.58	-70	(0.5, -0.6)	0.95	0.65	-70	(0.6, -0.6)

参考文献:

- [1] Hartigan J A, Wong A. A k-means clustering algorithm [J]. Applied Statistics, 1979(28): 100-108.
- [2] Krishnapuram R, Frigui H, Nasraoui O. Quadric shell clustering and the detection of second degree curves [J]. Pattern Recognition Lett, 1992(14): 545-552.
- [3] DavE R N, Bhaswan K. Adaptive fuzzy C-shells clustering and detection of ellipses [J]. IEEE Trans. Neural Networks, 1992(3): 643-662.
- [4] Isak Gath, Dan Hoory. Fuzzy clustering of elliptic ring-shaped clusters [J]. Pattern Recognition Letters, 1995 (16): 727-741.

- [5] Man Y, Gath I. Detection and separation of ring-shaped clusters using fuzzy clustering [J]. IEEE Trans. Pattern Anal Mach. Intell, 1994(16): 855-861.
- [6] 黄友锐. 智能优化算法及其应用 [M]. 北京: 国防工业出版社, 2002.
- [7] 魏立梅, 谢维信. 模糊C-球壳聚类算法的研究 [J]. 电子与信息学报, 2001, 23(1): 37-44.
- [8] 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.

作者简介:

任建斌 男, (1977-), 硕士, 实验师. 研究方向为数据处理.