一种基于动态遗传算法的聚类新方法

何 宏,谭永红
(上海师范大学信息与机电工程学院,上海 200234)

摘 要: 如何确定聚类数目一直是聚类分析中的难点问题.为此本文提出了一种基于动态遗传算法的聚类新方法,该方法采用最大属性值范围划分法克服划分聚类算法对初始值的敏感性,并运用两阶段的动态选择和变异策略,使选择概率和变异率跟随种群的聚类数目一致性变化,先进行不同聚类数目的并行搜索,再获取最优的聚类中心.七组数据聚类实验证明该方法能够实现数据集最佳划分的自动全局搜索,同时搜索到最佳聚类数目和最佳聚类中心.

关键词: 聚类分析; 遗传算法; 动态选择; 变异
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2012)02-0254-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.02.008

A Novel Clustering Method Based on Dynamic Genetic Algorithm

HE Hong, TAN Yong-hong
(*College of Information, Mechanical and Electronic Engineering, Shanghai Normal University, Shanghai* 200234, *China*)

Abstract: How to determine the number of clusters is always a difficult problem in data cluster analysis. Therefore, a novel dynamic genetic clustering algorithm (DGCA) is proposed in this paper. The DGCA adopts a maximum attribute range partition method to overcome the sensitiveness to initial values of cluster centers for clustering algorithms. Furthermore, the two-stage dynamic selection and mutation operations are used in the DGCA to make selection probability and mutation probability vary with the consistency of the number of clusters in the population. Firstly the parallel search in different numbers of clusters is carried out. Then the optimal search for the best cluster centers is conducted. Numerical experiments on seven data sets show that the proposed DGCA can realize the global search for the best partition and find the optimal values for both the number of clusters and the cluster centers.

Key words: cluster analysis; genetic algorithm; dynamic selection; mutation

1 引言

聚类分析是一种寻求数据自然聚集结构的重要方法,如何正确地确定聚类的数目是决定聚类质量的关键,也是数据聚类的难点问题[1].选择聚类数目的方法可以分为两大类[2]:局部确定法和全局确定法.其中局部确定法适合分层聚类算法,由于聚类过程中需要对不同类是否合并或拆分进行判断,使得聚类数目的确定过程变得比较繁琐.而全局确定法是根据聚类指标和不同聚类数目的聚类结果人工确定聚类数目,通常适合划分聚类法,其工作量会随数据量的增大而迅速增加.并且许多划分聚类法受聚类初值分布影响较大,从而导致最佳聚类数目的错误结果[3].为解决这些问题,目前有些科研人员将遗传算法与划分聚类方法相结合,根据遗传算法全局寻优的特点,用聚类结果有效性评价指标确定适应度函数,从而实现自动确定最佳聚类数目.文献[4]就是将遗传算法与一种分离合并算法结合用以估计聚类数目 k.文献[5]则结合多目标遗传算法与 K-means 算法确定类别数目.文献[6]用混合小生境遗传算法寻找出最佳 k.而文献[7]采用人工免疫网络结合遗传算法获取 k.此外,文献[8]利用基于决策树的混合聚类方法确定聚类数目.文献[9]则提出了细菌进化算法自动寻找最佳 k.然而,这些方法都没有同时考虑聚类结果受聚类初值分布影响的问题,并且选取的聚类指标复杂,计算量大.

因此,本文将提出一种基于动态遗传算法的聚类新方法(Dynamic Genetic Clustering Algorithm,简称为 DGCA),该方法按照最大属性值范围划分法选择聚类中心初值,同时运用两阶段动态选择和变异方法,使选择率和变异率随种群聚类数目一致性变化,先在不同聚类数

收稿日期:2010-10-15;修回日期:2011-08-28
基金项目:国家自然科学基金(No.60971004,No.61171088);上海市自然科学基金(No.10ZR1422400);上海教委科研创新重点项目(No.09ZZ141);上海师范大学重点学科项目(No.DZL811);上海师范大学原创与前瞻性预研项目(No.DYL201006)

目下并行聚类,自动搜索出最佳的聚类数目,然后再获取最优的数据划分.最后将 DGCA 应用于 7 个数据集中验证了其聚类的有效性.

2　动态遗传聚类算法

设数据样本为 M 个,每个样本数据 a_i 含有 P 个属性,则 $\boldsymbol{a}_i=\{v_1,v_2,\cdots,v_P\}$,样本点集合 $\boldsymbol{A}=\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_M\}$ 的数据空间为 $\boldsymbol{S}^{M\times P}$,若需要将 M 个样本分成 k_i 类,即有类集合 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_{k_i}\}$,其中各类中心点构成的数据阵为 $\boldsymbol{Z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_{k_i}\}$.根据文献[10],若 k_i 未知,其取值可为集合 K 中任意数,且有 $K=[k_{\min},k_{\max}]$,其中 $k_{\min}=2$,$k_{\max}=round(\sqrt{M})$.寻找数据样本 M 的最佳聚类划分模式 $\boldsymbol{C}_b$,就是在 $k_{\max}$ 个划分模式中获取 M 的最佳聚类中心数目 k_b 和最佳聚类中心 $\boldsymbol{Z}_b$.动态遗传聚类算法 DGCA 将遗传算法与 K-means 相结合,聚类过程中 K-means 在相同聚类数目下局部寻优,遗传算法在不同聚类数目下全局寻优,变化的选择率和变异率指导算法搜索的深度和广度.并先在 K 中搜索 k_b,再找出 $\boldsymbol{Z}_b$,从而获得数据空间 $\boldsymbol{S}$ 的最佳划分模式 $\boldsymbol{C}_b$.

2.1　编码方法

为兼顾编码长度和数据信息两方面的因素,本文采用聚类中心的方式编码.设种群中有 N 个个体,每个进化个体 $\boldsymbol{X}_i$ 代表一组聚类中心,即 M 的一种数据划分模式.聚类中心的个数 k_i 变化,$k_i\in K$,每个中心有 P 维实数编码的属性,则个体 $\boldsymbol{X}_i$ 为变长度实数编码,且编码长度 $L_i=P\times k_i$.

2.2　初始种群的产生

实验证明选择好的类中心初值对划分聚类的质量影响很大[11].本文中样本之间的比较采用欧式距离,考虑到属性值的取值差异将会直接影响样本距离的大小,因此,本文采用最大属性值范围划分法,即对取值范围最大的属性值 p_m 进行分层聚类,设个体 $\boldsymbol{X}_i$ 代表 k_i 个聚类中心,p_m 的取值范围 R_m 为 $[p_{m1},p_{m2}]$,用分层聚类法将 R_m 划分成 k_i 个区间,然后再在每个区间中随机选取数据,从而选择对应的样本作为聚类中心初值.相对于随机选择法容易出现毫无意义的初值、距离优化法对阈值设置敏感、密度估计法计算量太大[11,12]等不足,本文的方法不仅充分运用了数据的信息,而且计算量也比较小.

2.3　适应度函数

聚类结果有效性的验证方法可以分为内部验证指标和外部验证指标[13].本文采用内部验证指标作为 DGCA 不同聚类类数目下聚类结果评价的寻优目标函数,并采用外部验证指标进行 DGCA 与其他聚类方法的比较.实验证明有些内部验证指标在 k 取不同数值下的变化曲线存在拐点,这些拐点往往是最佳的聚类数目[14,15],因此,本文选取这些内部指标中的 Calinski-Harabasz index(简称为 CH index)作为个体的适应度函数.

CH index 是基于类内部的聚合度和类外部分离度而定义的聚类验证指标.设内部的聚合度用类内数据间的平方距离和为

$$WGSM=\sum_{i=1}^{k}\sum_{x\in\boldsymbol{C}_i}d(\boldsymbol{x},\boldsymbol{z}_i)^2 \tag{1}$$

类外部分离度为

$$BGSM=\sum_{i=1}^{k}n_i d(\boldsymbol{z}_i,\boldsymbol{z}_{tot})^2 \tag{2}$$

其中 $\boldsymbol{z}_{tot}$ 为所有数据点的中心,n_i 为第 i 类中的样本个数,则 CH index 可以表示为

$$CH_k=\frac{BGSM}{k-1}\cdot\frac{M-k}{WGSM} \tag{3}$$

对于聚类问题,$WGSM$ 越小越好,而 $BGSM$ 越大越好,所以 CH_k 越大越好,其最大值就是最佳划分 $\boldsymbol{C}_b$.

2.4　遗传操作设计

2.4.1　两阶段动态选择

为避免在局部最优的 k 值下获得错误的聚类划分,DGCA 采用两阶段动态选择.当 k 未知时,聚类算法首先获取 k_b,然后才在 k_b 个子空间内寻找到 $\boldsymbol{Z}_b$.设种群聚类数目的一致性为 $k_{con}=\frac{n_{sk}}{N}$,其中 $0<k_{con}\leqslant 1$,n_{sk} 为当代种群中具有相同聚类数目的个体数的最大值.个体 $\boldsymbol{X}_i$ 的适应度为 $f(\boldsymbol{X}_i)$,则 DGCA 中 $\boldsymbol{X}_i$ 的选择概率为

$$P_s(i)=\begin{cases}\dfrac{f^{\alpha}(\boldsymbol{X}_i)}{\sum_{j=1}^{N}f^{a}(\boldsymbol{X}_j)}, & 0<k_{con}<1,\text{对不同的 } k_i \text{ 进行选择}\\[2ex] \dfrac{f(\boldsymbol{X}_i)}{\sum_{j=1}^{N}f(\boldsymbol{X}_j)}, & k_{con}=1,\text{对不同的 } \boldsymbol{Z}_i \text{ 进行选择}\end{cases} \tag{4}$$

其中动态选择因子 $\alpha=e^{-k_{con}}$,且 $e^{-1}<\alpha<1$.

第一阶段,当 $0<k_{con}<1$ 时,以搜索 k_b 为主,设 G 代个体 $\boldsymbol{X}_i$ 所表示的聚类数目为 k_i,$\boldsymbol{X}_i$ 的选择概率为 k_i 的选择概率,即若选择 $\boldsymbol{X}_i$ 的重复数为 N_i,则类数目 k_i 重复 N_i 个,而不是 $\boldsymbol{X}_i$ 重复 N_i 个.这样每个 k_i 下,都有类中心不同的个体并行进化,避免了固定的初始类中心选择不佳带来的影响.并且 k_i 选择概率随 k_{con} 变化而变化,设种群中适应度为 $f(\boldsymbol{X}_i)$ 的个体数为 N_s,而由 N_s 个体组成的集合为 $\boldsymbol{S}_s$,$\boldsymbol{S}_s=\{\boldsymbol{X}_j\mid f(\boldsymbol{X}_j)=f(\boldsymbol{X}_i),\boldsymbol{X}_j\in\boldsymbol{S}\}$.则 k_i 的选择概率可以写为

$$p_s(i)=\left(N_s+\frac{\sum_{X_j\notin S_s} f^{\alpha}(\boldsymbol{X}_j)^{-1}}{f^{\alpha}(\boldsymbol{X}_i)}\right)$$

$$=\left[N_s+\sum_{X_j\notin S_s}\left(\frac{f(\boldsymbol{X}_j)}{f(\boldsymbol{X}_i)}\right)^{\alpha}\right]^{-1} \quad (5)$$

由上式可知,当 k_{con}逐渐增大为 1 时,α 则逐渐减至 e^{-1},若 $f(\boldsymbol{X}_i)$较小,即 $f(\boldsymbol{X}_i)<f(\boldsymbol{X}_j)$,$P_s$ 则变大;反之个体 $f(\boldsymbol{X}_i)$较大,P_s 则变小.即当搜索过程中种群的个体所表示的聚类数目逐渐趋于一致时,优胜 k_i 的选择概率随 k_{con}的增大而减小,优胜 k_i 不会快速占领种群,增加了对最佳聚类数目搜索的深度.k_{con}为 1 时,种群将收敛于 k_b.

第二阶段,当 k_{con}为 1 时,种群个体所表示的聚类数目均为 k_b,但代表的聚类中心 $\boldsymbol{Z}_i$ 不同.选择操作为进行个体 $\boldsymbol{X}_i$ 重复选择,并使具有最佳聚类中心的个体选择概率最大,以最后确定 $\boldsymbol{Z}_b$.此外,两阶段的个体选择都采用轮盘赌的方法和精英保留策略,优胜的个体将直接保留到下一代参与继续进化.

2.4.2 同类并行交叉

DGCA 中的选择操作指导算法向最优聚类数目进化,是粗聚类,为了使每一次个体交叉都具有意义,交叉应在同一聚类数目 k 内进行细进化.因此,DGCA 是按照代表相同 k 的个体分成子群体,然后同一子群体中的个体进行交叉,k 不同的个体间不交叉,各子群并行交叉操作,并采用固定交叉率的 1 点交叉,两个个体 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 前 q 个编码互换,q 的取值为$[0,\min(L_i,L_j)]$范围内的任意整数,其中 L_i 和 L_j 分别为 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 的编码长度.

2.4.3 两阶段动态变异

为了使搜索能够遍历整个数据空间,DGCA 的变异率 p_m 将根据种群 k_{con}的大小而进行动态改变.当 $0<k_{con}\leqslant 0.9$ 时,算法重点搜索 k_b,p_m 随着 k_{con}的增大而减小,且对 $\boldsymbol{X}_i$ 表示的 k_i 进行变异,变异后的 k_i' 在 K 中随机取值;当 $0.9<k_{con}\leqslant 1$ 时,为算法从 k_b 的搜索自然过渡到 $\boldsymbol{Z}_b$ 搜索,个体 $\boldsymbol{X}_i$ 表示的 k_i 逐渐趋于一致,不对 k_i 变异,而是对聚类中心 $\boldsymbol{Z}_i$ 进行变异,且每个个体变异率随适应度函数的增加而减小,即 p_m 根据聚类效果的优劣而不同,进行均匀分布的双向变异,所以有:

$$p_m(i)=\begin{cases}0.1(1-k_{con}), & 0<k_{con}\leqslant 0.9 \text{ 对 } k_i \text{ 进行变异}\\ e^{-\frac{f(\boldsymbol{X}_i)}{\sum_{j=1}^{N} f(\boldsymbol{X}_j)}}, & 0.9<k_{con}\leqslant 1 \text{ 对 } \boldsymbol{Z}_i \text{ 进行变异}\end{cases} \quad (6)$$

在 $0.9<k_{con}\leqslant 1$ 时,设 $x_{ij}(j=1,2,\cdots,L_i)$为个体 $\boldsymbol{X}_i$ 的第 j 个数位值,x_{ij}'为 x_{ij}双向变异以后的值,则 x_{ij}'以概率 $p_m(i)$按照下式取值:

$$x_{ij}'=\begin{cases}x_{ij}+T_{ij}(x_{j\max}-x_{ij}), & T_{ij}\geqslant 0\\ x_{ij}+T_{ij}(x_{ij}-x_{j\min}), & T_{ij}<0\end{cases} \quad (7)$$

其中 $x_{j\max}$和 $x_{j\min}$分别为个体 $\boldsymbol{X}_i$ 第 j 个数位取值范围的最大值和最小值,T_{ij}为在$[-1,1]$区间均匀分布任意值.

2.5 DGCA 算法的基本流程

DGCA 算法的基本流程如下:

(1)初始化:

①输入样本数 M,迭代次数 G_m,交叉率 P_c,种群数 N,聚类中心个数行向量 $\boldsymbol{K}=\{k_1,k_2,\cdots,k_N\}$;

②用最大属性值范围划分法产生 N 个个体;

(2)计算个体适应度:

①由 K-means 方法获取新的聚类中心;

②根据式(3)计算个体适应度函数;

③将最佳个体保留至下一代,并参与进化.

(3)由遗传操作产生新一代个体:

①按照个体适应度进行两阶段动态选择操作;

②不同 k 的子群内个体进行并行交叉操作;

③根据 k_{con}进行两阶段的动态变异.

(4)终止条件判别:满足终止条件否?满足输出聚类结果;不满足则返回步骤 2 继续进化.

3 实验仿真

为验证 DGCA 聚类的有效性,实验中将其与 K-means 和标准遗传 K-means 的聚类方法(Standard Genetic K-means Clustering Algorithm,简称为 SGKC)相比较,其中 SGKC 选用标准遗传算法的选择和变异操作,并用随机选择聚类中心初值的方法,其他环节与 DGCA 相同.分别选择 4 个人工数据集和 3 个真实数据集进行实验,其中人工数据集如图 1 所示,分别代表聚类常用的 Ruspini 标准数据集[16]、有重合的数据集 Atestdata1、大类中含有子类的数据集 Atestdata2、多类的数据集 Atestadat3.3

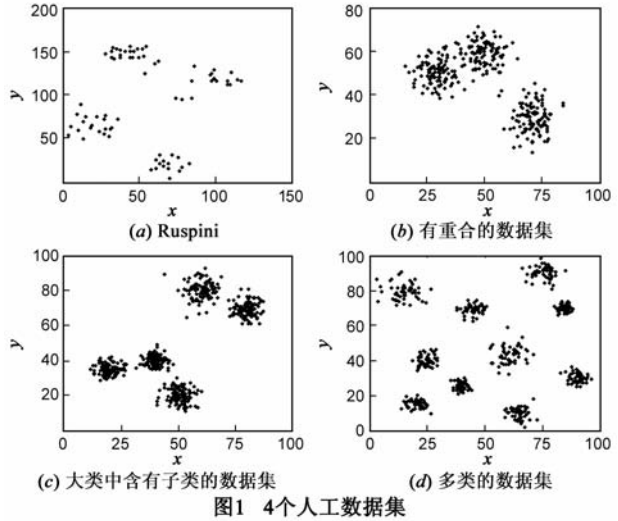

图1 4个人工数据集

个真实数据集的数据分别为 Iris data、Glass data、Wisconsin Breast Cancer[17](简称为 WBCancer).

3.1 多次运行结果比较

多次运行实验中设种群数 N 为 60,进化代数 G_m = 200,交叉率 P_c = 0.8,SGKC 的变异率 P_m = 0.05,三种方法分别对 7 组数据进行聚类,并运行 100 次,记录 DGCA 和 SGKC 的聚类数目的正确率 R_c(R_c =(搜索到正确聚类数目的次数/总运行次数)×100%),全局最佳适应度函数均值 $GBestF_m$,如表 1 所示.表 1 中的结果显示在数据集 k 未知的情况下,DGCA 除 glass 数据集以外,聚类数目的正确率 R_c 均为 100%,而 SGKC 的 R_c 均小于 100%,Atestdata3 的 R_c 甚至小于 50%.以 Iris 数据集为例,DGCA 和 SGKC 的 R_c 分别为 100%和 99%,与现有自动聚类方法相比较,文献[6]中 HNGA 只能将其划分为两个大类,文献[8]中 HGA 的 R_c 为 98.67%,而文献[9]中 BEA 的 R_c 仅为 88%.这说明 DGCA 和 SGKC 将变长度编码遗传算法与 K-means 相结合,可以利用遗传算法的优点,通过全局搜索自动确定数据集的聚类数目,而 DGCA 对聚类数目搜索的准确率更高.

表 1 DGCA 和 SGKC 聚类数目结果比较

数据集名	M	P	k_c	聚类方法	R_c(%)	$GBestF_m$
Ruspini	75	2	4	SGKC	99	425.122
				DGCA	100	425.327
Atestdata1	300	2	3	SGKC	97	1125.349
				DGCA	100	1131.847
Atestdata2	500	2	5	SGKC	82	3961.526
				DGCA	100	4027.469
Atestdata3	500	2	10	SGKC	48	3516.714
				DGCA	100	3635.817
Iris	150	4	3	SGKC	99	559.780
				DGCA	100	560.087
Glass	214	9	6	SGKC	90	123.708
				DGCA	94	123.830
WBCancer	683	9	2	SGKC	72	924.306
				DGCA	100	954.706

此外,将 K-means 中的 k 设为已知的实际聚类数目 k_c,而 DGCA 和 SGKC 进化过程中同时搜索最佳聚类数目和聚类中心.记录三种方法获得的聚类结果有效性的外部评价指标 Rand Index 均值 RI_m 和 Ajusted Rand Index 均值 ARI_m[18],如表 2 所示.两个指标的取值范围均为[0,1].只有在聚类结果与已知的正确分类完全一致的情况下,RI 和 ARI 值为 1,其值越小说明聚类结果越差.表 2 的数据显示尽管 K-means 的 k 已知,但 SGKC 除了 Glass 和 WBCancer 数据集略低于 K-means 以外,其它数据集 SGKC 的 RI_m 和 ARI_m 都高于 K-means 的结果.而 DGCA 的 RI_m 和 ARI_m 都比 K-means 和 SGKC 大,且 4 个人工数据的聚类指标均为 1.并且在表 1 中 DGCA 获得的全局最佳适应度函数均值 $GBestF_m$ 均比 SGKC 的结果大,这说明 DGCA 由于加入了两阶段动态选择和动态变异方法使搜索路径更加有方向性,同时采用的最大属性值范围法也能够克服聚类算法对聚类初值的敏感性,所得聚类结果更加准确.

表 2 DGCA、SGKC 和 K-means 聚类结果评价指标比较

数据集名	聚类方法	RI_m	ARI_m
Ruspini	K-means	0.9288	0.8065
	SGKC	0.9998	0.9995
	DGCA	1	1
Atestdata1	K-means	0.9356	0.8671
	SGKC	0.9809	0.9566
	DGCA	1	1
Atestdata2	K-means	0.9205	0.7865
	SGKC	0.9960	0.9870
	DGCA	1	1
Atestdata3	K-means	0.9565	0.7906
	SGKC	0.9942	0.9655
	DGCA	1	1
Iris	K-means	0.8490	0.6711
	SGKC	0.8794	0.7293
	DGCA	0.9879	0.9572
Glass	K-means	0.6922	0.2460
	SGKC	0.6723	0.2342
	DGCA	0.8686	0.7896
Wisconsin Breast Cancer	K-means	0.9254	0.8493
	SGKC	0.9148	0.8288
	DGCA	0.9986	0.9784

3.2 单次运行结果比较

为了观察动态选择和动态变异方法对 k_b 和 $\mathbf{Z}_b$ 搜索的实际情况,以 Atestdata3 为例,将 SGKC 和 DGCA 分别运行一次,算法各参数同多次运行实验,记录 DGCA 和 SGKC 分别在 G = 1、50、150、200 最佳个体所代表的聚类中心聚类分布情况,如图 2 所示.并同时记录 DGCA 和 SGKC 的全局最佳适应度函数 GBestF、全局最佳个体的聚类数目 GBestk、种群一致性 k_{kcon} 随进化代数 G 的变化曲线,如图 3 所示.从图 2 中可以看出,由于 DGCA 采用了最大属性值范围划分法选取聚类中心的初值,所以,进化初期 DGCA 选取的初始聚类中心更加符合数据的实际分布情况.在整个搜索过程中,SGKC 无法克服标准遗传算法易陷入局部最小值的缺陷,虽然在 G = 104 时种群的 k_{con} 为 1,但实际上种群所获得的 GBestk 从 G = 15 开始到进化结束就一直为 12,由于收敛于错误的聚类数目,最后获得 GBestF 仅为 3535.422,运行结束后 SGKC 的 RI 和 ARI 分别为 0.9901 和 0.9415.这说明标准遗传操作同时搜索 k_b 和 Z_b 的效果较差.而 DGCA 在两阶段的动态选择和变异的作用下,0 < k_{con} < 1 时主要搜索最佳聚类数目,图 2 和图 3 显示,尽管进化初期 DGCA 的 Bestk = 12,但是由于 k_i 的选择

概率和变异率都随 k_{con} 变化而变化，使得 GBestF 和 GBestk 的曲线也随着 k_{con} 波动在不断变化，在 $G=91$ 时搜索到 GBestk = 10，避免了 k 收敛于局部最优．在精英保留策略的作用下，$G=126$ 时，$k_{con}=1$，种群在个体的 k_i 一致的情况下进行 $\boldsymbol{Z}_b$ 的搜索，个体的选择概率随适应度函数改变，并采用双向变异，最终在 $G=153$ 时获得 GBestF = 3635.817，DGCA 聚类结果评价指标 *RI* 和 *ARI* 均为 1，聚类结果的正确率达 100%．

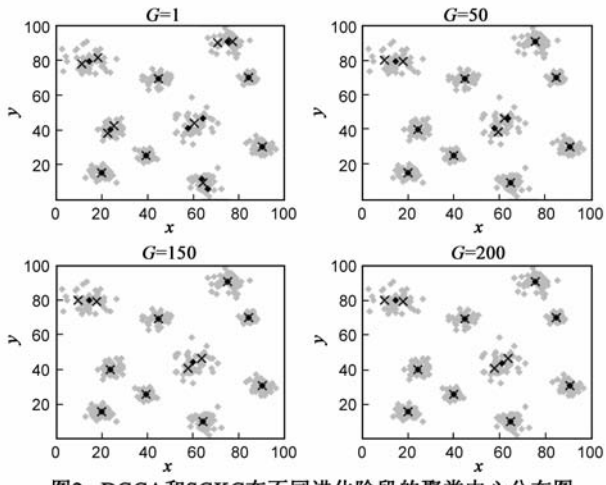

图2 DGCA和SGKC在不同进化阶段的聚类中心分布图
注：黑色圆点为DGCA搜索到的聚类中心；+字号为SGKC搜索到的聚类中心．

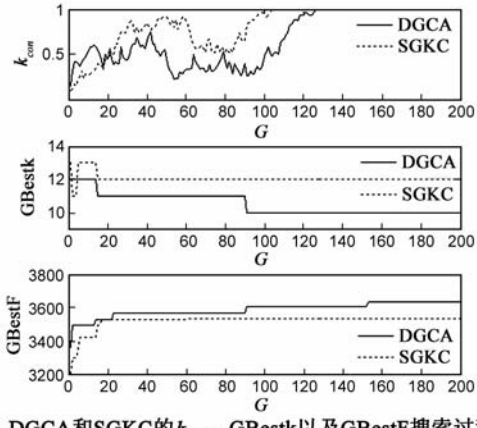

图3 DGCA和SGKC的k_{con}、GBestk以及GBestF搜索过程的变化曲线

4 结论

在数据特征知识匮乏的情况下，如何确定聚类数目是正确聚类的前提．本文提出了一种基于动态遗传算法的自动聚类新方法，该方法可以实现对同一数据集进行不同聚类数目的并行聚类，采用两阶段的动态选择和动态变异操作，使选择概率和变异率跟随种群聚类数目的一致性变化而变化，先重点搜索最佳聚类数目，而后获取最佳聚类中心．同时采用的最大属性值范围划分法克服了划分聚类算法对初始值的敏感性．通过人工数据集和真实数据集测试实验证明 DGCA 能够实现球状数据集、有重合的数据集、大类中含有子类的数据集、多类的数据集的最佳聚类划分的全局搜索，同时获取最佳聚类数目和聚类中心．由于没有普遍适用的聚类算法，因此，未来的研究是如何将这种动态遗传算法应用于对含有异常数据或不同形状数据集的自动聚类．

参考文献

[1] A K Jain. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651 – 666.

[2] A D Gordon. Classification[M]. Chapman & Hall/CRC, Boca Raton, FL, 2 Edition, 1999. 163 – 175.

[3] Lin Yu Tseng, Shiueng Bien Yang. A genetic clustering algorithm for data with non-spherical-shape clusters[J]. Pattern Recognition, 2000, 33: 1251 – 1259.

[4] Othman R M, Deris S, Illias R M, et al. Automatic clustering of gene ontology by genetic algorithm[J]. International Journal of Information Technology, 2006, 3(1): 37 – 46.

[5] Liu Y M, özyer T, Alhajj R, et al. Integrating multi-objective genetic algorithm and validity analysis for locating and ranking alternative clustering[J]. Informatica, 2005, 29: 33 – 40.

[6] Weiguo Sheng, Stephen Swift, Leishi Zhang, et al. A weighted sum validity function for clustering with a hybrid niching genetic algorithm[J]. IEEE Transactions on System, Man, and Cybernetics-Part B: Cybernetics, 2005, 35(6): 1156 – 1167.

[7] 钟将，吴中福，等．基于人工免疫网络的动态聚类算法[J]．电子学报，2004，32(8)：1268 – 1272．
Zhong Jiang, Wu Zhongfu, et al. A Novel dynamic clustering algorithm based on artificial immune network[J]. Acta Electronica Sinica, 2004, 32(8): 1268 – 1272. (in Chinese)

[8] Sung-Hae Jun. A hybrid genetic algorithm and new criterion for determining the number of clusters[J]. International Journal of Soft Computing, 2006, 1(4): 313 – 318.

[9] Swagatam Das, Archana Chowdhury, Ajith Abraham. A bacterial evolutionary algorithm for automatic data clustering[A]. Evolutionary Computation, CEC′09 IEEE Congress on[C]. IEEE, 2009. 2403 – 2410.

[10] N R Pal, J C Bezdek. On cluster validity for the fuzzy c-means model[J]. IEEE Transactions Fuzzy System, 1995, 3(3): 370 – 379.

[11] A Juan, E Vidal. Comparison of four initialization techniques for the K-medians clustering algorithm[J]. Lecture Notes in Computer Science, 2000, 1876: 842 – 852.

[12] Fuyuan Cao, Jiye Liang, Liang Bai. A new initialization method for categorical data clustering[J]. Expert Systems with Applications, 2009, 36: 10223 – 10228.

[13] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis. Clustering validity checking methods: Part II[J]. ACM SIGMOD

Record,2002,31(3):19 – 27.

[14] S Saitta, B Raphael, IFC Smith. A comprehensive validity index for clustering[J]. Intelligent Data Analysis, 2008, 12:529 – 548.

[15] Sandro Saitta, Benny Raphael, Ian F. C. Smith. A bounded index for cluster validity[J]. Lecture Notes in Computer Science, 2007, 4571:174 – 187.

[16] E H Ruspini. Numerical methods for fuzzy clustering[J]. Inform Sci, 1970, 2:319 – 350.

[17] Frank, A & Asuncion, A. UCI machine learning repository [DB/OL]. http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science, 2010-4-26.

[18] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis. On clustering validation techniques[J]. Journal of Intelligent Information Systems, 2001, 17(2/3):107 – 145.

作者简介

何　宏　女,1973 年生于新疆克拉玛依市.副教授,博士,硕士生导师.研究方向为计算智能理论及其在系统建模、控制、优化和数据挖掘中的应用.

E-mail:heh@shnu.edu.cn

谭永红　男,1958 年生于广西桂林.教授,博士,博士生导师.研究方向为系统建模、生物医学信号处理、信息融合.

E-mail:tany@shnu.edu.cn