第30卷 第2期
2009年4月

华南农业大学学报
Journal of South China Agricultural University

Vol. 30, No. 2
Apr. 2009

遗传 K 均值方法在品种资源分类中的应用

徐永春[1]，张森文[2]

(1 华南农业大学 工程学院，广东 广州 510642;2 暨南大学 应用力学研究所，广东 广州 510632)

摘要:采用实数编码方式,对聚类的中心矩阵进行编码,通过数组变换将染色体与相应聚类中心的数组进行匹配,通过轮赌选择和自适应的交叉、变异操作及均值小生境的种群优化对聚类中心的编码进行更新迭代,最终得到稳态的聚类误差函数和划分效果最好的聚类中心. 然后通过对某基地的甘蔗品种进行分析、比较,分析的误差函数结果显示,Ringa *K*-Means 改进的聚类效果明显优于传统的 *K*-Means 方法及 Sga-*K*-Means 方法的聚类效果.

关键词:遗传算法; *K* 均值聚类; 实数编码; 甘蔗品种; 小生境

中图分类号:TP306.1　　**文献标识码**:A　　**文章编号**:1001-411X(2009)02-0097-04

·The Application on Variety Clustering of Genetic-Algorithm-*K*-Means

XU Yong-chun, ZHANG Sen-wen

(1 College of Engineering, South China Agricultural University, Guangzhou 510642, China;
2 Applied Mechnical Institute, Jinan University, Guangzhou 510632, China)

Abstract: This paper proposed one kind of *K*-Means analysis method based on the genetic algorithm by the average value niche. The real number method was used to encode the clustering center, and the chromosome was matched with the clustering central array correspondingly throuth array transformation. The method was used to update the clustering central with selection, crossover, mutation and the average value niche population optimization. The stable state of the error function and the best dividing clustering center was obtained. The experimental result demonstrated that the Ringa-*K*-Means method was obviously better than the traditional *K*-Means method and the Sga-*K*-Means method.

Key words: genetic algorithm; *K*-Means; float encoding; sugar cane variety; niche

聚类算法是指利用数的方法研究和处理给定对象的分类. 算法的目的是使同一类中对象的特性尽可能地相似,不同类对象间的特性差异尽可能地大,且所有的聚类方法都有自己的聚类准则. 而聚类算法的聚类准则通常可分为基于距离、基于密度及基于连接的3种方式. 在传统的品种资源分析中,外界条件因子的筛选结果与分类的结果有着直接关系,通常传统方法主要采用主成分因子或者逐步回归的方法来选择其显著水平的外界因子,在完成外界因子的筛选后,最后再进行品种优劣的分类. 由此在传统的品种资源分类过程中,不可避免地出现初始点和人为主观因素的干扰,为提高分类的客观性,就需要采用更为合理、智能的分类初选方案.

本文根据 *K*-Means 动态聚类算法的特点,结合品种资源数据的特点,采用实数编码方式对聚类问题的解(聚类中心)进行编码,并介绍了解码方式、适应度函数以及所采用的选择、交叉和变异操作. 为了说明本文方法的可行性和优越性,最后采用某农场的若干甘蔗品种资源数据进行实例验证,并与其他算法进行对比.

收稿日期:2008-06-28

作者简介:徐永春(1974—),男,工程师,博士;通讯作者:张森文(1944—),男,教授;E-mail:tzhsw@jnu.edu.cn

1 *K*-Means 聚类算法

K-Menas 聚类作为动态聚类方法之一，其动态聚类的主要思想是将样本集合分为 k 类，将 n 个向量 x_j $(1,2,\cdots,n)$ 分为 c 个组 $G_i(i=1,2,\cdots,c)$，并求每组的聚类中心，使得非相似性（或距离）指标的价值函数（或目标函数）达到最小. 当选择欧几里德距离为组 j 中向量 x_k 与相应聚类中心 c_i 间的非相似性指标时，价值函数可定义为：[1]

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k,x_k \in G_i} \|x_k - c_i\|^2 \right), \quad (1)$$

这里，$J_i = \sum_{k,x_k \in G_i} \|x_k - c_i\|^2$ 是组 i 内的价值函数. 这样 J_i 的值依赖于 G_i 的几何特性和 c_i 的位置. 一般来说，可用一个通用距离函数 $d(x_k,c_i)$ 代替组 i 中的向量 x_k，则式(1) 总价值函数可表示为：

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k,x_k \in G_i} d(x_k - c_i) \right). \quad (2)$$

则 *K*-Means 算法处理步骤一般如下：

①从 $\{x_1,x_2,\cdots x_n\}$ 中随机选择 k 个点 $c_1,c_2,\cdots,c_n$ 作为 k 个聚类集合的中心点；

②以 $c_1,c_2,\cdots,c_n$ 为中心点进行集合划分，划分的原则是：如果 $\|x_i - c_j\| < \|x_i - c_m\|, m=1,2,\cdots,k; j=1,2,\cdots,k; i=1,2,\cdots,n$ 且 $j \neq m$. 则将 x_i 划分到集合 G_j 中；

③根据各集合中的点计算新的中心点 $c_1^*,c_2^*,\cdots,c_k^*$；

$$c_i^* = \frac{1}{n_i} \sum_{x_j \in G_i} x_j, i = 1,2,\cdots,k, \quad (3)$$

其中，n_i 为 G_i 中点的个数；

④如果 $c_i^* = c_i, i=1,2,\cdots,k$，则计算结束，当前中心点为聚类划分的结果，否则令 $c_i^* = c_i$，返回步骤②.

K-Means 算法主要通过迭代搜索获得聚类的划分结果，虽然 *K*-Menas 算法运算速度快，内存开销小，比较适合于大样本量的情况，但是聚类结果受初始凝聚点的影响很大，不同的初始点选择会导致截然不同的结果；并且当按最近邻归类时如果遇到 2 个凝聚点距离相等的情况，不同的选择也会造成不同的结果，因此 *K* 均值动态聚类法具有因初始中心的不确定性而存在较大偏差的情况[3].

2 改进遗传 *K*-Means 聚类算法研究

对于传统 *K*-均值算法，由于初始聚类中心点的选择很大程度地影响聚类效果，如果有 2 个或多个“种子”点无意中跑到一个类内，则其聚类结果很难区分[3]. 为了提高聚类误差函数的精确，因此提出采用前端方法如遗传算法等智能化算法求出好的初始聚类中心，再进行最终的 *K*-Means 函数聚类，获得稳定的聚类结果.

遗传算法是一种通过模拟自然进化过程搜索最优解的方法，其显著特点是隐含并行性和对全局信息的有效利用能力，只需少量结果就可反映探索空间较大的区域，便于实时处理，而且具有较强的稳健性，可避免陷入局部最优. 在此本文提出一种基于实数编码的均值小生境的遗传聚类算法 Rinaga *K*-Means 算法，避免 *K*-Means 受初始分类的影响及聚类误差函数陷入局部极值的可能；其次由于数据的属性值大小不一致，在数据聚类之前，首先对原始数据进行标准化，将各属性数据压缩在(0,1)区间内，提高数据分类运算的普遍性. Rinaga *K*-Means 的步骤如下.

1)编解码方式：遗传算法中的进化过程是建立在编码机制基础上的，编码对于算法的性能，如搜索能力等影响很大. 常用的编码方式有浮点数编码和二进制编码，相比之下，实数编码计算量简单，内存空间数据转换快，因此本文采用实数编码方式实现数据样本与遗传染色体的转换[2,5].

聚类问题的解是各聚类的中心，对于样本空间中 m 个点 $\{x_1,x_2,\cdots,x_m\}$ 的 L 个属性聚类问题，遗传算法中的每个染色体包括 $L \times k$ 个对应的实数分量，其中 k 为初始聚类中心的个数.

2)种群初始化：从待分类的点集中随机选择 k 个点作为问题的一个解并进行编码. 在这里由于初始中心为 $L \times k$ 数组，在遗传迭代过程中，不容易控制其染色体进化，因此需要进行数据变换，将 $L \times k$ 数组转变为一维数组，其染色体长度为 $L \times k$，然后重复进行 N 次染色体初始化，形成为 N 个 $L \times k$ 长度大小的初始种群.

3)适应度函数：在处理过程中，对于每个染色体采用与 *K*-均值算法相同的方式进行聚类的划分和重新计算各聚类的中心，然后依式(2)用每个聚类中的点与相应聚类中心的距离作为判断聚类划分质量的准则函数 J，J 越小，表示聚类划分的质量越好[4].

遗传算法的目的是搜索使 J 值最小的聚类中心，由于在 *K*-Means 计算过程中，距离函数过大，因此对适应度函数进行线性修正，即扩大 10^5 倍(×

10^5),提高适应度函数的存活率. 适应度函数最终数学描述为:

$$f = 10^5/(1 + J). \tag{4}$$

4)选择操作:选择操作体现了遗传算法中的"适者生存"原则,适应度越高的个体,参与后代繁殖的概率应该越高. 本文则采用轮盘赌方法进行个体的选择.

5)交叉操作:在交叉过程中,进行交叉的父代染色体均为一维的 k 个初始中心点,如 $A_1A_2B_1B_2$ 和 $C_1C_2D_1D_2$ 2 条父代染色体,其组成由 2 个初始中心点组成 1 组染色体,因此,在交叉操作过程中主要通过交换 2 个父个体的中心点的数据(初始中心点如 A_1A_2 和 C_1C_2 交换)来产生新的子个体($A_1A_2D_1D_2$). 本文采用单点交叉,对于编码长度为 $L \times k$ 的个体,首先按照 P_c 概率随机生成交叉点位置(cp),然后交换 2 个父个体中位于 cp 右侧的部分,从而生成 2 个新的子代个体.

6)变异操作:子代个体除了继承父代个体的信息外,还会按一定的概率发生变异,这体现了生物遗传的多样性. 本文则使用自适应变异概率 P_m 对子代 $L \times k$ 位染色体的某变异点进行变异操作.

7)自适应交叉、变异算子:由于随着迭代次数的增加,优化种群逐渐趋于集中,此时变异、交叉算子应降低其参数值,因此,选用如下公式(5)和(6)作为 Rinaga K-Means 聚类的自适应交叉和变异算子参数的更新机制.

交叉算子自适应公式:

$$P'_c = \begin{cases} P_{c1} \times \sqrt{1 - (t/t_{\max})}, & P'_c > P_{c2} \\ P_{c2}, & P'_c \leqslant P_{c2} \end{cases}, \tag{5}$$

式中,交叉系数 P_{c1}、P_{c2} 分别为 0.8、0.5,t 为遗传迭代次数,$t_{\max}$ 为最大遗传代数. 这种自适应交叉概率能够保证在迭代初期,交换率较大,从而加快进化的速度,避免遗传算法陷入迟钝状态,同时能够保证在迭代后期,交换率较低,并逐步减小至一常量,从而保证平滑收敛,同样能够保持搜索空间的连续性,增大找到全局最优解的可能性.

$$P'_m = \begin{cases} P_{m1} \times \exp(-\lambda \times t/t_{\max}), & P'_m > P_{m2} \\ P_{m2}, & P'_m \leqslant P_{m2} \end{cases}, \tag{6}$$

式中,变异系数 P_{m1}、P_{m2} 分别为 0.1、0.05,λ 为一常数,根据种群分布大小取值,一般为[5,10]. 这种自适应变异概率能够保证在迭代初期,个体性能较差时,变异率较大以造成足够的扰动,扩大解空间,而随着迭代次数的增加,变异率逐步减小至常量,从而保证平滑收敛. 显然,自适应变异率的这种变化是符合优胜劣汰规律的.

8)均值小生境种群优化:根据式(3)计算的初始中心距离 d_i 进行均值小生境计算,其共享函数为:

$$S_i = \sum_{j \in N}^{c} Sh(d_{ij}), \tag{7}$$

$$\text{而 } Sh(d_{ij}) = \begin{cases} 1 - \dfrac{d}{\sigma_{\text{share}}}, & d \leqslant \sigma_{\text{share}} \\ 0, & d \leqslant \sigma_{\text{share}} \end{cases}, \tag{8}$$

其中,N 为初始种群数,小生境半径参数采用均值方法实现,可提高种群适应度自然分布概率[2].

$$\sigma_{\text{share}} = \text{mean}(\text{sum}(\|di - d_{j+1}\|)), \tag{9}$$

其中,$i = 1,2,\cdots,N$.

在计算出了群体中各个染色体的共享度之后,依据下式来调整各个染色体的适应度:

$$f'(x_i) = \frac{f(x_i)}{S_i}, (i = 1,2,\cdots,N). \tag{10}$$

在迭代过程中,每次得到聚类划分后,返回到步骤 3)后,用校正后的聚类中心替换原来的聚类中心进行迭代.

9)遗传迭代终止规则:终止条件设计中,给定迭代次数或误差精度来终止遗传算法的运行. 在每次进行选择、交叉和变异操作之后,记录当前子代中最好适应度值与上一次记录的子代最好适应度值的误差值 e,然后判断 e 是否小于误差解. 满足终止条件后,则最后适应度最高的个体为最优聚类中心.

以获取的最佳初始中心点,进行 K-Means 聚类函数最终迭代运算,获得最后的最佳分类结果.

3　遗传 K 均值聚类实例

采用某农场 2 组甘蔗品种数据资源进行比较、分析,来验证遗传 K 均值方法的适用性. 在分析过程中,由于品种资源各项属性数据量纲、大小不一致,对于优选分析处理的结果容易存在偏差. 为了消除各个变量量纲之间的差异,首先对文献[7]原始属性数据采用下式进行标准化处理:

$$\overline{x_i} = \frac{x_i - x_{i\min}}{x_{i\max} - x_{i\min}}, \quad i = 1,2,\cdots,k, \tag{11}$$

式中:$x_{i\min}$、$x_{i\max}$ 分别表示第 i 个影响因子 x_i 在 i 组训练样本中的最小值与最大值,$\overline{x_i}$ 表示归一化之后的影响因子值.

然后分别采用 K-Means、Sga K-Means 及 Ringa K-

Means 3 种方法进行分析,在分析过程中的基本参数设置:聚类中心 k 为 4,遗传主要参数 $n=80$、$P_c=0.8$、$P_m=0.1$,迭代次数(Maxgen)为 50,最终聚类运算比较,得 2 组甘蔗品种的聚类方案结果(表 1、2).

表 1 第 1 组不同模型聚类结果

Tab. 1 Clustering results using different models with the first example

分类	*K*-Means	Sga-*K*-Means	Rinaga-*K*-Means
一类	台 16	台 25	93/159
	粤 c29	台 16	台 25
	89/1626	89/113	
二类	台 25	94/128	台 16
	94/128	93/159	94/128
		台 10	台 10,粤 c29
三类	89/525	79/177	89/525
	79/177	华 1	89/1626
	90/7907	90/7909	粤糖 89/240
	粤糖 89/240	粤 c29	
四类	93/159	粤糖 89/240	华 1
	台 10	89/525	79/177
	华 1		90/7909
			89/113
误差适应度	105.36	144.83	170.14
优选时间/s	0.34	461.06	445.13

表 2 第 2 组不同模型聚类结果

Tab. 2 Clustering results using different models with the second example

分类	*K*-Means	Sga-*K*-Means	Rinaga-*K*-Means
一类	巴西 45	新台糖 23	新台糖 22
	桂糖 94-116	巴西 45	新台糖 10
	新台糖 16	新台糖 22,台优 2	
	新台糖 10	桂糖 94-116	
二类	新台糖 22	粤糖 93-159	粤糖 93-159
	桂糖 94-119		新台糖 20
	桂糖 93-102		
三类	新台糖 23	新台糖 16	新台糖 23,台优 1
	台优 1,台优 2	新台糖 10	巴西 45,台优 2
四类	粤糖 93-159	台优 1	桂糖 94-116
	新台糖 20	桂糖 94-119	桂糖 94-119
		桂糖 93-102	桂糖 93-102
		新台糖 20	新台糖 16
误差适应度	121.44	184.4	218.86
优选时间/s	0.818	615.187	378.07

从表 1、2 的结果可知,*K*-Means 误差适应度最小,误差最大,其分类效果最差. 改进的 Rinaga-*K*-Means,在误差函数及运行曲线上明显优于 Sga-*K*-Means 的模型,这主要是由于 Sga-*K*-Means 算法的不足所引起的,在优解进化过程中,出现早熟或局部跳出优解的可能性,导致后代迭代出现次优解的可能,因此在 *K*-Means 改进模型中,采用改进的 Rinaga-*K*-Means 算法能够较好地完成 *K*-Means 初始中心点的搜索工作,从而提高品种的优选分类效率. 其次相对于传统的层次或者关联方法而言[7],Rinaga-*K*-Means 则更为客观,不受传统的层次或关联方法人为主观因素影响的条件下获得评分结果,能更客观提高分类效果

4 结论

通过对甘蔗品种资源数据进行 *K*-Means, Sga-*K*-Means 和 Ringa-*K*-Means 3 种不同聚类分析算法比较,说明 Ringa-*K*-Means 利用改进的 Ringa 算法设计 *K*-Means 聚类函数的初始聚类中心,所获得的结果更具有稳健性、且搜索全面,从而降低了避免陷入局部最优的可能性,也说明 Rinaga-*K*-Means 算法具有良好的聚类质量和综合性能稳态的效果,可适用于智能化的品种资源优选工作和其他的数据挖掘分析.

参考文献:

[1] 朱明. 数据挖掘[M]. 合肥:中国科学技术大学出版社, 2002: 5.

[2] 王小平,曹立明. 遗传算法——理论、应用与软件实现[M]. 西安:西安大学出版社,2002:1.

[3] 傅景广,许刚,王裕国. 基于遗传算法的聚类分析[J]. 计算机工程,2004, 30(4): 122-124.

[4] GOLBERG D E. Genetic algorithms in search otimization and machine-learning MA[M]. [s. l]: Addison Welery, 1989.

[5] KRISHNA K, MURTY M N. Genetic k-means algorithm [J]. IEEE Trans on System,Man and Cybernetics, 1999, 29(3): 433-4393.

[6] SRINIVAS P. Adaptive probabilities of crossover and mutation in genetic algorithm[J]. IEEE Trans on Systems, Man and Cybernetics, 1994,24(4): 656-667.

[7] 王维赞, 朱秋珍,邓展云. 灰色关联度分析在甘蔗新品种评价中的应用[J]. 广西农业科学,2002(1):7-11.

【责任编辑 周志红】