第 36 卷 第 6 期 电 子 科 技 大 学 学 报 Vol.36 No.6
2007 年 12 月 Journal of University of Electronic Science and Technology of China Dec. 2007

改进的模糊C-均值聚类方法

牛 强，夏士雄，周 勇，张 磊

(中国矿业大学计算机科学与技术学院 江苏 徐州 221008)

【摘要】 该文针对模糊C-均值算法容易收敛于局部极小点的缺陷，将遗传算法应用于模糊C-均值算法(FCM)的优化计算中，其中对传统遗传算法的编码方案、遗传算子约束条件及适应值函数等方面进行改进，提出了一种基于改进遗传算法的模糊聚类方法。实验表明，将改进的遗传算法与FCM算法结合起来进行聚类分析，可以在一定程度上避免FCM算法对初始值敏感和容易陷入局部最优解的缺陷，使聚类更合理，比单一使用FCM算法进行聚类分析的效果要好。

关 键 词 聚类； C均值算法； 模糊聚类； 遗传算法； 优化计算

中图分类号 TP301.6 **文献标识码** A

Improved Fuzzy C-Means Clustering Algorithm

NIU Qiang，XIA Shi-xiong，ZHOU Yong，ZHANG Lei

(School of Computer Science & Technology, China University of Mining and Technology Xuzhou Jiangsu 221008)

Abstract A method of fuzzy clustering based on genetic algorithms is proposed in this paper. This method applies the improved genetic arithmetic to optimization of the Fuzzy C-Mean (FCM) arithmetic. FCM arithmetic has the limitation of converging to the local infinitesimal point, in our method, some interrelated key technique problems, such as encoding method, genetic operators, restrict condition, fitness function for the traditional genetic algorithm, are further reformed. Experiment results show that the method can search global optimum partly so that the clustering results are better than those of only using the FCM.

Key words cluster; C-means algorithm; fuzzy clustering; genetic algorithm; optimization computation

模糊C-均值聚类算法(FCM)[1-2]是应用最为广泛的聚类算法之一，它具有算法简单、收敛速度快且能处理大数据集的优点，但是，也存在着很大的局限性[3-4]：聚类效果受初始时聚类中心的影响很大，算法采用梯度法求解极值，结果往往是局部最优，而得不到全局最优解[5-6]。

遗传算法[7]是一种非常有效的全局随机搜索和优化技术，有隐含并行性、鲁棒性和全局搜索等特点，已经广泛应用到众多领域。它仿效了遗传学中生物从低级到高级的进化过程，将进化操作应用于一群对搜索空间(或称参数空间)编码的基因串(或称染色体)中，在每一代，遗传算法同时搜索参数空间的不同区域，然后把注意力集中到解空间中期望值最高的部分。通过一群基因串一代又一代地繁殖和交换，遗传算法能搜索到多个局部极值，从而增加了找到全局最优解的可能性。尽管遗传算法是一种全局随机优化方法，但它也存在缺点。(1) 该算法完全依赖概率随机地进行寻优操作，虽然可以避免陷入局部极小，但受寻优条件的限制，一般只能得到全局范围内的次优解，很难得到最优解；(2) 通过参数的二进制编码字符串间接运算，人为地将连续空间离散化，导致计算精度与字符串长度、运算量之间的矛盾；(3) 由于遗传算法采用随机优化技术，所以要花费大量的时间，时间复杂度比较大。

针对以上存在的问题，本文提出一种改进的遗传算法(Improved GA，IGA)，采用实变量构成染色体进行交叉变异，并进一步将它与FCM算法相结合进行模糊聚类分析(IGA-FCM)。首先使用遗传算法求得全局最优解的近似解，然后以该近似解作为FCM算法的初始值，用FCM算法进一步求解，最终

收稿时间：2007－09－07

基金项目：国家自然科学基金(50674086)；高等学校博士学科点专项科研基金(20060290508)；江苏省社会发展科技计划(BS2006002)

作者简介：牛 强(1974－)，男，博士，主要从事数据挖掘与知识获取方面的研究；夏士雄(1961－)，男，教授，博士生导师，主要从事数据处理与信息融合方面的研究；周 勇(1974－)，男，博士，主要从事遗传算法与系统优化方面的研究；张 磊(1978－)，男，博士，主要从事智能数据分析与处理方面的研究.

得到全局最优解。

1 模糊C-均值聚类算法(FCM)算法

FCM是由文献[1]从硬C-均值算法(记为HCM)推广而来，已成为最常用和讨论较多的聚类算法之一。其描述如下：

令 $X=\{x_i,i=1,2\cdots,n\}$ 是一训练样本集， $X\subseteq R^p,c$ 为预定的类别数目， $v_i(i=1,2,\cdots,c)$ 为第 i 个聚类的中心， $u_{ik}(i=1,2,\cdots,c,\ k=1,2,\cdots,n)$ 是第 k 个样本对第 i 类的隶属度函数，且 $0\leqslant u_{ik}\leqslant 1$ 及 $0<\sum_{k=1}^{n}u_{ik}<n$，FCM的目标函数为：

$$J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m\parallel x_k-v_i\parallel^2 \tag{1}$$

式中 $U=\{u_{ik}\}$； $v=(v_1,v_2,\cdots,v_c)$； $m>1$ 为常数；其约束为：

$$\sum_{i=1}^{c}u_{ik}=1 \qquad \forall k=1,2,\cdots,n \tag{2}$$

在约束式(2)下优化式(1)得：

$$u_{ik}=\frac{(1/\parallel x_k-v_i\parallel^2)^{1/(m-1)}}{\sum_{j=1}^{c}(1/\parallel x_k-v_j\parallel^2)^{1/(m-1)}}$$

$$\forall i=1,2,\cdots,c,\ \ k=1,2,\cdots,n \tag{3}$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^m x_k}{\sum_{k=1}^{n}u_{ik}^m} \qquad \forall i=1,2,\cdots,c \tag{4}$$

模糊C-均值聚类算法是基于误差平方和目标函数准则，先给出初始方案，通过式(3)、(4)反复迭代，使得目标函数式(2)达到极小。

2 基于遗传算法的离散化算法

遗传算法(Genetic Algorithm，GA)是近年来迅速发展起来的一种有效的全局优化算法，它借用了生物遗传学中自然选择遗传和变异等作用机制，使得个体的适应性在优化过程中得以提高，其基本思路是：首先产生初始解群，然后按某种指标从解群中选取较优的个体，利用一些遗传算子(如交叉和变异等)对其进行运算，产生新一代的候选解群，重复此过程，直到满足某种收敛指标为止。与其他许多优化方法相比，遗传算法具有明显的全局搜索性以及对所求解问题的鲁棒性，因此被广泛应用于有关组合优化、模式识别、机器学习和图像处理等其他优化方法无法解决或难以解决的问题中，并已展示了其独特的优点。尽管遗传算法运用到工程优化领域已经越来越广泛，但是传统遗传算法在实际应用中还有待进一步改进[8-9]。本文给出一种改进的遗传算法，并将其运用于模糊聚类算法中。

2.1 编码设计

遗传算法最初都是采用二进制编码方式对问题候选解进行编码，为避免编码的复杂性，提高计算效率，本文采用实数(实值)编码策略，用特定实数编码方式，缩短染色体长度，提高了算法的收敛速度和全局寻优能力。每个染色体由可行解向量$\boldsymbol{R}$的元素列 $R_1,R_2,\cdots,R_n$ 表示，则相应染色体是 $V=R_1,R_2,\cdots,R_n$，其中定义整数pop_size作为染色体的个数，根据约束条件随机产生pop_size个初始可行染色体 $V_1,V_2,\cdots,V_{\text{pop\_size}}$。

遗传算法根据适应度值通过选择、杂交和变异三个遗传算子实现它的寻优过程，搜索能力由选择算子和杂交算子决定，变异算子则保证了算法能够搜索到问题空间的尽可能多的点，从而使其具有搜索全局最优的能力[10]。

2.2 选择操作

本文采取基于非线性排名的选择策略。选择过程是以旋转赌轮pop_size次为基础，每个旋转都为新种群选择一个染色体，选择过程如下：

对每个染色体 V_i 计算累积概率 q_i

$$\begin{cases}q_0=0\\ q_i=\sum_{j=1}^{i}\text{eval}(V_j), \qquad i=1,2,\cdots,\text{pop\_size}\end{cases} \tag{5}$$

从区间 $[0,\text{pop\_size}]$ 中产生一个随机函数 r。若 $q_{i-1}<r\leqslant q_i$，则选择第 i 染色体 V_i ($1\leqslant i\leqslant \text{pop\_size}$)。

2.3 交叉操作

本文采用整体算数杂交算子。设 (V_i',V_j') 为一个随机选择父代对，首先从 (0,1) 中产生一个随机数 c，然后按下列形式在 V_i' 和 V_j' 之间进行交叉操作，并产生两个后代 V_i'' 和 V_j''：

$$V_i''=cV_i'+(1-c)V_j' \tag{6}$$

$$V_j''=(1-c)V_i'+cV_j' \tag{7}$$

之后需要根据约束条件检验每一个后代的可行性，如果两个后代均可行，则用它们代替其父代，否则，保留其中可行的；然后产生新的随机数 c，重新进行交叉操作，直到得到两个可行的后代或者循环给定次数为止。

2.4　变异操作

本文的IGA在非一致性变异算子的基础上考虑了解的质量，提出了自适应变异算子，其具体描述为：首先等概率地生成二值随机数 $r(r\in\{0,1\})$，然后按式(8)进行变异：

$$v_i'=\begin{cases}v_i+\Delta(t,u_i-v_i) & r=0\\ v_i-\Delta(t,u_i-v_i) & r=1\end{cases} \tag{8}$$

式中　$\Delta(t,y)=y(1-r^{t^2})$，其中 r 为均匀分布在[0 ,1]上的随机数；$t=1-f(V)/f_{\max}$；$f(V)$ 为当前个体的适应值；$f_{\max}$ 为当前群体中的最大适应值；v_i' 为子代个体 V' 的第 i 个分量；v_i 为父代个体 V 的第 i 个分量；u_i、l_i 为 v_i 取值范围的上下界。从 $\Delta(t,y)$ 函数可以看出，在 y 值一定的条件下，随着 t 的减少(适应值增加)，其函数值接近0的概率就会增加，从而使得适应值大的个体在较小范围内搜索，而适应值小的个体在较大范围内搜索。这样变异算子能根据解的好坏自适应地搜索区域，从而较大地提高了算法的收敛速度和搜索能力。

2.5　适应度函数

当前群体中的每一个个体都对应着 C 个中心，对于这 C 个中心，按FCM算法，可以把数据集分为 C 个簇。定义聚类的适应度函数为：

$$f=\frac{1}{1+G_c} \tag{9}$$

式中　G_c 是FCM算法的聚类准则函数，即误差平方和准则函数：

$$G_c=J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m\|x_k-v_i\|^2 \tag{10}$$

2.6　算法描述

采用本文的遗传算法进行模糊聚类优化的主要步骤如下(数据集可以用 $X=\{x_1,x_2,\cdots,x_N\}$ 表示)：

步骤 1　产生初始群体及其编码。在集合中随机选取 S 组对象，每组 C 个点，代表了聚类的 C 个聚类中心点。根据类内对象距离最近，类间对象距离最远的原则，可以把剩下的 $(N-C)$ 个对象分别划分到 C 个簇里，这样，每一组都决定了一个聚类结果。用基因串 $V=(v_1,v_2,\cdots,v_i,\cdots,v_N)$ $(v_i=1,2,\cdots,C)$ 来表示某一聚类结果，当 $v_i=c(1\leqslant c\leqslant C)$ 时，表示第 i 个数据属于第 c 个簇。则这种编码方式的搜索空间有 C^N 个点。这 S 组对象就构成了初始种群。

步骤 2　计算适应度函数。

步骤 3　利用遗传算子，对当前一代的个体进行繁殖，从而产生后代。

步骤 4　由计算出的适应度淘汰父类中适应度较低的个体，然后接纳新的个体加入。

步骤 5　计算后代的适应度，并将适应度高的个体与父代中保留下来的个体合并成新的一代。

步骤 6　如果达到设定的繁衍代数，则返回最好的基因串，算法结束；否则，回到步骤3继续繁殖下一代。

3　实验测试

为了比较传统模糊聚类算法与基于改进型遗传算法的聚类算法(IGA-FCM)的性能，选择标准的IRIS数据作为测试样本点进行比较。IRIS数据由150个四维向量组成，每一个样本的四个分量分别表示IRIS 的 SepalLength 、 SepalWidth 、 PetalLength 和 PetalWidth。整个样本集包含三个IRIS种类Setosa、Versicolor和Virginica，每类各有50个样本。遗传算法参数设置：种群大小size=150；交叉概率 P_c=0.65，变异概率 P_m=0.05，最大运行次数maxrun= 100，两种算法仿真结果如表1所示。表中给出了30次运算中目标函数的最小值及其平均误差、达到最小值时的平均迭代次数、类内距离和类间距离。

表1　FCM与IGA-FCM比较

	FCM	IGA-FCM
类内平均距离	0.849±0.076	0.833±0.038
类间平均距离	3.351±0.243	3.314±0.209
目标函数最小值	0.662±0.154	0.632±0.118
平均迭代数	26	37

FCM有较好的适应值，但与本文方法相比并不显著。当考虑类间、类内距离时，本文优化方法能保证得到更加紧凑的类间和类内聚类结果，并具较小的误差。

图1描述了对IRIS数据的收敛性效果，FCM收敛较快(评价函数迭代28次)，但早熟于一个较大的目标函数值，而本文方法适应值较小，能以较快地(迭代43次)收敛于最小值。

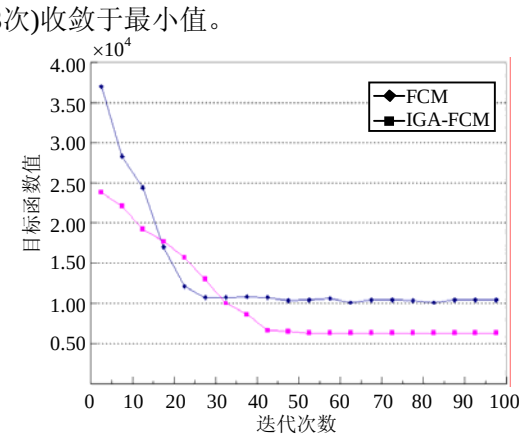

图1　算法收敛比较

(下转第1272页)