基于免疫遗传算法的模糊 C-均值聚类

孙 洋,罗 可
SUN Yang,LUO Ke

长沙理工大学 计算机通信与工程学院,长沙 410076
College of Computer and Communication Engineering,Changsha University of Science & Technology,Changsha 410076,China
E-mail:yangyang.sun@yahoo.com.cn

SUN Yang,LUO Ke.*C*-Means clustering based on immune genetic algorithm.Computer Engineering and Applications, 2009,45(23):152-153.

Abstract: In order to overcome the sensitive of FCM algorithm to the initial value,propose a FCM algorithm based on immune genetic algorithm.This algorithm uses the theory of immune system and the adjustment method of adaptive genetic operator(That is immune genetic algorithm) to improve FCM algorithm.And experiments have proved that this algorithm can effectively solve the premature convergence issues,guarantee the diversity of the population,and make clustering converge quickly and effectively to the global optimal solution.

Key words: clustering algorithm;Fuzzy *C*-Means(FCM);immune genetic algorithm;immune genetic FCM algorithm

摘 要:为了克服 FCM 算法对初值的敏感性,提出了一种基于免疫遗传算法的 FCM 算法。该算法利用免疫系统原理和遗传算子自适应调整的方法(即免疫遗传算法)来改进 FCM 算法。实验证明该算法能有效解决未成熟收敛的问题,保证了种群的多样性,使聚类问题最终快速、有效地收敛到全局最优解。

关键词:聚类算法;模糊 *C*-均值算法;免疫遗传算法;免疫遗传 FCM 算法
DOI:10.3778/j.issn.1002-8331.2009.23.042 **文章编号:**1002-8331(2009)23-0152-02 **文献标识码:**A **中图分类号:**TP311

1 引言

模糊聚类分析作为非监督机器学习的主要技术之一,建立了样本类属不确定性的描述,能够比较客观地反映现实世界[1],在数据挖掘、图像分割、适量量化、模式识别、模糊逻辑等诸多领域有着广泛的应用。在众多的模糊聚类算法中,应用最广泛且较成功的是 1974 年由 Dunn 提出并由 Bezdek 加以推广的模糊 *C*-均值(Fuzzy *C*-Means,FCM)算法[2]。该算法简单、收敛速度快且局部搜索能力强,但它对初始条件较为敏感,对不同的初始值有不同的聚类结果。由于该算法是基于梯度下降的算法,因此,常常不可避免地使目标函数陷入局部极值,甚至会出现退化解和无解的情况。基于遗传算法(GA)的聚类方法能够解决 FMC 的初值敏感性问题, 并有更多的机会获得全局最优解,但用 GA 仍会出现未成熟收敛现象,仍不能保证每次运行都得到全局最优解[3]。

将免疫机制引入遗传算法,有效地克服了标准遗传算法的早熟现象,并将免疫遗传算法和 *C*-均值算法有机结合,形成一种混合算法。根据聚类问题的实际情况设计遗传选择、交叉和变异算子,使得混合算法更快、更有效地收敛到全局最优解。

2 FCM 算法

FCM 算法是把 n 个数据 $x_i(i=1,2,\cdots,n)$分成 c 个模糊族,并求得每个族的类中心,使目标函数达到最小。FCM 算法目标函数为:

$$J_m(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{n}\sum_{j=1}^{c}(u_{ij})^m(d_{ij})^2 \tag{1}$$

这里 $\sum_{j=1}^{c}(u_{ij})=1,u_{ij}\in(0,1),\forall i,d_{ij}=\|x_i-x_j\|$。其中:$X=\{x_1,x_2,\cdots,x_n\}$为数据集,$m$ 为模糊加权指数,且 $1\leqslant m<\infty$,c 为聚类的类别数,且 $c\geqslant 2$,$\boldsymbol{U}=\{u_{ij}\}$表示隶属度矩阵,$u_{ij}$是第 j 类中的样本 x_i的隶属度,$\boldsymbol{V}=\{v_j\}$表示类中心矩阵。为使目标函数 J_m 达到最小,类中心和隶属度的更新如下:

$$v_j=\frac{\sum_{i=1}^{n}(u_{ij})^m\cdot x_i}{\sum_{i=1}^{n}(u_{ij})^m},j=1,2,\cdots,c \tag{2}$$

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\|x_i-v_j\|}{\|x_i-v_k\|}\right)^{\frac{2}{m-1}}\right]^{-1} \tag{3}$$

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60474070,No.10471036);湖南省科技计划项目(No.05FJ3074);湖南省教育厅重点项目(No.07A001)。

作者简介:孙洋(1984-),女,硕士生,主要研究方向为数据库技术、数据挖掘;罗可(1961-),男,教授,博士,研究方向为数据挖掘、计算机应用等。

收稿日期:2008-05-06 **修回日期:**2008-08-01

当 $d_{ij}=0$ 时，则 $u_{ij}=1, u_{ik}=0, k\neq j, i=1,2,\cdots,n$。

3　免疫遗传算法简介

3.1　免疫算法的基本原理

生物免疫系统中的克隆选择原理，描述了免疫系统对抗原激励做出免疫相应的基本特性[4]。

生物免疫系统对入侵生命体的抗原通过细胞的分裂和分化作用，自动产生相应的抗体来抵御，即免疫应答；在免疫应答过程中，部分抗体作为记忆细胞保存下来，当同类抗原再次侵入时，记忆细胞被激活并迅速产生大量抗体，使再次应答比初次应答更快更强烈，体现了免疫系统的记忆功能。同时，抗体与抗体之间根据浓度来相互促进和抑制，以维持抗体的多样性及免疫平衡，即抗体的浓度越高，则受抑制，浓度越低，则受促进，体现了免疫系统的自我调节功能[5]。

3.2　免疫遗传算法

免疫遗传算法是基于生物免疫机制提出的一种改进的遗传算法，该算法一般由抗原识别、初始抗体产生、适应度计算、记忆细胞分化、抗体的促进和抑制、抗体产生六个模块组成。此外，称 IGA 中个体为抗体。具体流程如图 1 所示。

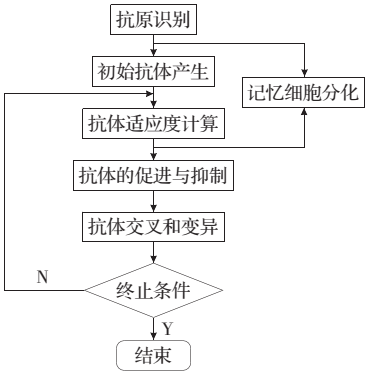

图 1　免疫遗传算法流程图

4　基于免疫遗传算法的 FCM 聚类算法

(1)编码和适应度函数

算法中的个体采用基于聚类中心的浮点数编码方式，每个抗体 S 由 c 个聚类中心组成，它可表示为长度为 $c\times l$ 的浮点码串.个体的适应度函数可定义为：

$$f=\frac{1}{1+J_m} \tag{4}$$

其中：J_m 也就是式(1)中的 J_m。

(2)基于免疫原理的选择操作

抗体浓度的定义：

$$p_d=\frac{\text{种群中相同或相似个体的数目}}{\text{种群规模}} \tag{5}$$

抗体的适应度概率的定义：

$$p_f=\frac{\text{种群中个体的适应度}}{\text{种群所有抗体的适应度之和}} \tag{6}$$

抗体的选择概率：

$$p=\alpha p_f+(1-\alpha)p_d \tag{7}$$

其中 $\alpha>0$ 是常数。

这种选择策略的优点：抗体被选择的概率不仅依赖于抗体适应度的大小，还要依赖于个体在种群中的规模，这样充分体现了免疫系统的特性，保证了种群的多样性。

(3)交叉和变异

文中采用一致交叉的方法，即在两个配对抗体的每个基因座上的基因都以相同的交叉概率 p_c 进行交换，从而形成两个新的个体。

变异采用均匀变异的方法，即分别用符合某一范围内均匀分布的随机数，以某一较小的变异概率 p_m 来替换个体编码串中各个基因座上的原有基因值。

算法流程如下：

步骤 1　初始种群产生

初始化，确定较小正常数 $\varepsilon(\varepsilon>0)$、交叉概率 p_c、变异概率 p_m 以及每代中 C-均值算法的迭代次数 L。置遗传代数 $t=1$，随即生成 n 个抗体，形成初始种群 P。

步骤 2　适应度计算

用公式(2)、(3)对种群中的每个抗体迭代 L 次，然后得到新种群 P_1，对 P_1 中的每个抗体分别运用公式(1)和式(4)计算 J_m 和 f 值。

步骤 3　终止条件判断

计算 $\bar{J}(t)=\frac{1}{n}\sum_{k=1}^{n}J_m(k)$，若 $\left|\bar{J}(t)-\bar{J}(t-1)\right|<\varepsilon$，指定最好的个体为算法的结果，否则转步骤 4。

步骤 4　记忆细胞分化

将种群 P_1 按 f 值降序排列，选取适应度最大的 N 个构成抗体子集 M(即选择记忆细胞，其规模是抗体总数的 20%)，当新的记忆细胞产生时抗体子集是满的时，就用高亲和力的抗体代替低亲和力的抗体。

步骤 5　抗体的促进或抑制

分别运用公式(5)~(7)计算 p_d、p_f 和 p，并以 p 为选择概率复制($M+P_1$)中的抗体，得到种群 P_2。

步骤 6　抗体产生

以交叉概率为 p_c 的一致交叉法和变异概率为 p_m 的均匀变异法作用于种群 P_2 得到种群 P_3，置 $t=t+1$，转步骤 2。

5　仿真实验

采用 C-均值算法、GA 算法、KGA 算法和该文算法分别进行仿真实验。遗传算法的交叉概率取为 0.70，变异概率取为 0.008，群体规模为 100，$L=5$，$q=1$，$\alpha=0.8$，$\varepsilon=10^{-4}$。实验采用文献[6]中的数据，共有两组：第一组是 Fisher 的 Iris 植物样本数据，由分别属于 3 种植物的 150 个样本组成，每个样本均为四维模式向量，代表植物的 4 种特征数据。用 4 种算法分别做了 3 次实验，每次实验迭代 50 次，结果见表 1。

表 1　Iris 数据实验结果

次数	C-均值算法	GA	KGA	该文算法
1	97.087 073	97.000 103	97.000 103	97.000 103
2	97.118 547	97.000 103	97.000 103	97.000 103
3	97.087 071	97.000 103	97.000 103	97.000 103

第二组数据是 300 个随机分布的四维随机向量，聚类数目是 8，不仅规模大于 Iris 数据，而且完全是随机分布的，没有明显的类别分界，存在大量的局部极优点，是很困难的优化问题。采用 4 种方法分别做 3 次实验，每次实验迭代 100 次，结果见表 2。

(下转 169 页)