

基于改进人工免疫方法的混合模糊聚类算法^①

苏锦旗¹, 张文字¹, 薛 昱²

¹(西安邮电大学 经济与管理学院, 西安 710061)

²(北京信息控制研究所, 北京 100089)

摘 要: 为了解决传统模糊聚类算法对初始值敏感、目标函数易陷入局部极小值等问题, 将模糊核聚类方法与人工免疫算法相结合, 提出了一种基于改进人工免疫方法的混合模糊聚类算法. 算法通过借鉴生物免疫系统中的克隆选择原理和记忆机制, 自动确定聚类数目及中心位置, 同时还集成了模糊c均值搜索算子用于加快收敛速度. 仿真实验结果表明了本算法在收敛性、收敛速度和分类性能的有效性.

关键词: 人工免疫; 模糊聚类; 克隆变异; 收敛性和有效性

Hybrid Fuzzy Clustering Method Based on the Modified Artificial Immune Theory

SU Jin-Qi¹, ZHANG Wen-Yu¹, XUE Yu²

¹(College of Economics and Management, Xi'an University of Posts & Telecommunications, Xi'an 710061, China)

²(Beijing Institute of Information and Control, Beijing 100089, China)

Abstract: The fuzzy clustering algorithm is sensitive to the initial center and is easily convergence to local optimum. Inspired by the clone selection principle and memory mechanism of the vertebrate immune system, a new hybrid clustering method based on the modified artificial immune theory is presented. It not only adaptively determined the amount and the center's positions of clustering, but also avoided the local optima and the flaw about sensitive to the initialization. Also, the K-means algorithm is used as a search operator in order to improve the convergence speed. Experimental results indicate the validity and convergence of the proposed algorithm.

Key words: artificial immune; fuzzy clustering; clone and mutation; the validity and convergence

随着计算机技术的发展, 聚类应用范围已经扩展到了图像分割、模式识别、信号压缩和特征提取等各个领域^[1,2]. 传统的聚类分析是硬划分, 即把每个待辨识的对象严格划分到某类中. 而实际上大多数对象并没有严格的属性, 它们在性态和类属方面存在着中介性, 因此适合进行软划分. 模糊集理论的提出为这种软划分提供了有力的分析工具, 人们用模糊的方法来处理聚类问题, 并称之为模糊聚类方法. 它是目前比较流行的一种聚类算法^[3].

模糊聚类算法具有算法简单、速度快、局部搜索能力强等优点^[4]. 然而算法的执行结果对初始条件较为敏感, 选取不同的初始值可能会导致不同的聚类结果. 此外, 这种算法是基于目标函数的聚类算法, 一

般采用梯度法求解极值, 易陷入局部极小值, 甚至出现退化解和无解的情况^[5]. 基于遗传算法(GA)的聚类方法能解决对初始值敏感的问题, 获得全局最优的几率更大^[6], 但当样本数目、类别和维数较多时, GA 常常会产生早熟现象, 仍无法保证每次运行都获得全局最优解^[7].

针对上述聚类方法的不足, 这里借鉴生物免疫系统中的免疫细胞克隆和记忆、亲合力成熟等机理, 将模糊核聚类方法与人工免疫算法相结合, 对样本特征进行优化, 并通过人工免疫算法对模糊聚类算法进行改进, 提出了基于人工免疫原理的模糊聚类算法.

1 模糊c-均值算法

模糊c-均值算法(FCM)是一种主流的迭代下降聚

①基金项目:国家自然科学基金(71173248);陕西社科基金(13Q081);西安邮电大学青年教师科研基金 (ZL2012-30)

收稿时间:2014-01-29;收到修改稿时间:2014-03-04

类方法^[8], 其核心思想是通过迭代方法寻找聚类中心和隶属度函数, 使得目标函数达到最小^[9], 从而把数据对象划分到不同的簇中, 使生成的簇尽可能的紧凑和独立.

$$J = \sum_{k=1}^c \sum_{i=1}^N u_{ik}^p \|x_i - c_k\|^2 \quad (1)$$

该算法过程是通过反复移动各个簇中心, 以极小化簇集内的总度量(如距离、相似度)来完成的. 具体地说, 设样本为 $X_i (i=1, 2, \dots, n)$. $c_k (k=1, 2, \dots, c)$ 为每个类的中心, u_{ik} 是第 k 个样本对于第 i 类的隶属度函数. 在满足 $\sum_{i=1}^c u_{ik}=1$ 的条件下, 根据 lagrange 乘数法, 可得目标函数式(1)取得极小值的必要条件:

$$u_{ik} = \frac{\left(\frac{1}{\|x_i - c_k\|} \right)^{\frac{1}{p-1}}}{\sum_{j=1}^c \left(\frac{1}{\|x_i - c_j\|} \right)^{\frac{1}{p-1}}} \quad (2)$$

$$c_k = \frac{\sum_{i=1}^N u_{ik}^p x_i}{\sum_{i=1}^N u_{ik}^p} \quad (3)$$

2 基于人工免疫原理的模糊聚类算法基础

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响应的基本特性^[10]. 在基于克隆选择原理的免疫算法中, 抗原对应于问题的目标函数, 抗体对应于目标函数的优化解. 算法先根据抗体的适应值对解进行评价和选择, 然后通过记忆细胞保留局部最优解, 以保持解的多样性, 再基于抗体之间的亲和度逐步改善优化过程, 最终获得问题的全局最优解.

本文所提出的基于改进人工免疫方法的混合聚类算法, 对以往一般免疫算法步骤进行改进, 采用了记忆细胞和抑制细胞分化机制和由抑制细胞进行的抑制抗体产生的机制. 它防止复制记忆细胞, 从而提高了对问题的解的搜索效率.

1) 聚类问题的编码方案

由聚类目标函数可知, 聚类的目标就是要获得数据集 X 的一个模糊划分矩阵 W 和聚类的原型 P . 在基于人工免疫原理的聚类算法中, 令一组聚类原型 P 为一个抗体, 把原型中的 c 组特征连接起来, 根据各自的取值范围, 就可以将其量化(用二进制串表示)编码成一个抗体.

$$A_i(0) = \underbrace{\{\zeta_1, \zeta_2, \dots, \zeta_m, \dots, \zeta_j, \dots, \zeta_{(k-1)m+1}, \zeta_{(k-1)m+2}, \dots, \zeta_{km}\}}_{\text{编码}(p_1)} \underbrace{\dots}_{\text{编码}(p_k)} \quad (i=1, 2, \dots, N) \quad (4)$$

式中参数集依据每个原型 $p_i (1 \leq i \leq k)$ 取值. 由于我们处理的是混合特征, 因而在抗体中, 除了具有数值参数以外, 还应具有类属参数, 而类属特征是无序的, 因此可以直接编码到抗体中化.

2) 抗体亲和力计算

抗体的适应值函数(即抗原与抗体之间的亲和力函数)可以由目标函数变换得到.

$$f = \frac{1}{J+1} \quad (5)$$

其中 J 为均方误差函数. J 值小的个体(即类内离散度和小), 相应的适应值就大, 聚类效果越好. 抗体与抗体之间的亲和力反映了抗体之间的相似程度, 当抗体相似时, 目标函数越小, 聚类效果就越好, 其亲和力较大, 反之, 则较小, 采用如下公式:

$$f(A) = \frac{1}{1+C(W, P)} = \frac{1}{1 + \sum_{i=1}^k \sum_{l=1}^n w_{il}^2 (d(x_l, p_i))^2} \quad (6)$$

3) 记忆细胞和搜索算子

抗体浓度的计算可以采用下式:

$$c_v = \frac{\sum_{w=1}^N s_{v,w}}{N} \quad (7)$$

$$s_{v,w} = \begin{cases} 1 & B_{v,w} \geq T \\ 0 & \text{otherwise} \end{cases}$$

式中, T 为设定的阈值; N 为种群规模, 即抗体集合中所含的抗体数. 这样, 在考虑某种抗体浓度时, 满足一定程度近似的抗体可看作同一种抗体. 当某种抗体 c_v 的浓度超过阈值 T 时, 这表明某种抗体在抗体种群中占了绝对优势(即达到了一个相对的最优点), 产生

记忆细胞 m , 以记录代表此局部最优解的抗体.

本文还利用模糊聚类分析的局部寻优定义了一个新的算子, 即C搜索算子. 对于每个记忆细胞, 按式(2)求出其对应模糊划分矩阵 U ; 然后再按式(3)更新中心, 用更新后的中心构成新的个体.

4) 克隆变异操作

a) 克隆操作. 克隆 m 个最佳抗体 $c_1, c_2, \dots, c_m$, 产生一个临时的克隆群体 C , 每个记忆细胞的克隆规模与其亲和度的测度成正比.

b) 变异操作. 变异运算按照基因位进行, 每一个基因位上的浮点数, 以变异概率发生随机变异, 发生变异的基因位被另一在变异范围内均匀分布的随机数取代. 其中变异范围与个体的适应值成反比.

3 算法描述

本文将克隆选择算法的高效全局优化搜索特点和模糊聚类具有局部搜索能力强的特点结合, 提出克隆模糊聚类算法. 利用克隆选择算法求解数据集的模糊聚类需要解决三个问题: 将聚类问题的解编码到抗体中; 构造抗体-抗原亲和度函数来度量每个抗体对聚类问题的响应程度; 选择克隆算子, 确定操作参数的取值, 以确保快速收敛到最优解. 在该算法中, 基于免疫机理, 将待分类的数据对象视为抗原, 把聚类中心看作抗体. 数据对象的聚类过程即免疫系统不断产生抗体, 识别抗原, 最终产生出可以捕获抗原的最佳抗体过程.

本算法主要包括抗体选择、克隆、超变异及再选择等操作, 其求解聚类问题的具体步骤描述如下:

1) 抗原识别及抗体编码. 在人工免疫算法中, 抗原一般指所求解问题的目标函数. 这里将待分类的数据对象作为抗原, 并对聚类中心矩阵 V 进行编码, 即将 k 组表示聚类中心的特征连接起来, 根据各自的取值范围, 将其量化值编码成二进制基因串, 作为对应的抗体.

2) 初始化. 在克隆选择算法中, 产生第一代一定数量的个体构成的群体(即祖先群体)通过选择、克隆、变异、替代等操作产生下一代群体, 群体一代一代的进化, 直到找到最优解. 初始种群的规模一般是根据求解的实际问题和经验事先确定的. 在算法的求解过程中, 祖先群体一般是通过随机的方法生成的, 这样可以保证搜索空间中的每个可能解在初始群体中有相同的出现机会. 本文算法随机选取数据集中要求的聚类数目的点数形成染色体即初始个体, 设定算法各参数, 随机产生一组聚类中心作为初始抗体群体, 种群规模为 N , 按照种群规模的要求产生初始的种群.

3) 亲和力计算. 用式(6)来构造抗原-抗体亲和力. 聚类问题的最优结果对应于目标函数的最小值, 此时

的亲和力最大.

4) 克隆和变异操作. 算法对克隆操作产生的临时种群进行变异操作, 变异的概率一般定为常数. 变异算子按照染色体的基因位进行, 每一染色体上的聚类中心按照变异概率随机变异; 若变异后染色体的适应度函数小于变异前染色体的适应度函数, 则变异后个体取代变异前个体, 反之保持原染色体不变, 经过变异生成临时种群 c^* . 这里抽取亲和度高的 m 个抗体进入记忆细胞, 对每个记忆细胞进行克隆操作, 产生临时抗体集合 C , 再对临时抗体集合 C 进行变异操作, 产生新抗体集 D .

5) 抗体亲合度计算. 抗体-抗原的亲和度指群体中的一个个体的适应度函数, 他是用来区分群体中个体好坏的标准和算法是否继续演进的判断条件. 在克隆选择算法中, 对目标函数值的使用是通过评价个体的适应度来体现的. 本文将优化算法用于求解聚类问题, 求解使总的类间离散度之和最小, 适应度函数的值越小, 表示解最优. 这里计算 D 中每个抗体的亲和度, 选取 D 中亲和度高的抗体加入记忆细胞集, 并对每个记忆细胞按 c 搜索算子进行更新.

6) 迭代终止条件判断. 在克隆选择算法中, 终止算法的方式通常有两种: 设置一定的迭代次数, 当达到最大迭代次数时算法终止; 若在规定的迭代时间内, 算法的最优解仍没有大的变化, 则终止算法. 本文终止条件设定为亲和力所能达到的阈值或迭代次数, 若满足则终止迭代, 确定当前种群中的最佳个体作为算法的最优解; 否则继续.

7) 将记忆细胞中亲和度最高的抗体所对应的分类结果作数据集 X 的一种可能聚类结果. 在全部聚类数 $c(1 < c \leq n)$ 中, 求得最小 J 值所对应的分类结果就是数据集的最佳聚类结果.

4 仿真实验

仿真数据集包括两组, 一组为实际数据集, 一组为合成数据集. 实际数据集为 Fisher 的 Iris 植物样本数据^[11], 由分别属于 3 种植物的 150 个样本组成, 其中每个样本为 4 维向量, 代表植物的 4 种特征数据, 聚类数目为 3. 合成数据集为 400 个随机分布的 4 维随机向量, 聚类数目为 8. 本文算法的参数设置如下: 亲和力阈值 $T=0.95$; 记忆细胞数量 $M=90$, 种群规模 $N=25$. 算法迭代 80 次结束.

算法输入: 数据集、聚类数目、遗传 FCM 算法和克隆 FCM 算法的种群数目、克隆 FCM 算法的记忆细胞数目、变异概率等;

算法输出: 目标函数值、聚类时间、迭代次数.

对模糊聚类算法(FCM)、基于遗传算法的 FCM 和基于克隆免疫的 FCM 三种不同算法进行 20 次实验,

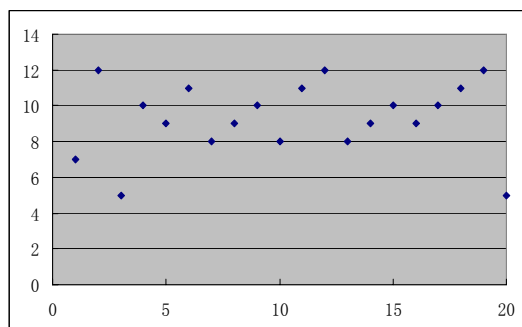
每种算法均可以找到最优解,但找到最优解的次数却不同.通过表 1,我们可以看出由于随机选取初始值,FCM 算法在 20 次实验中只有 4 次找到了最优解,证明了 FCM 算法对初始值敏感,且容易陷入局部最优解,但算法计算简单,所花费的时间很少;而基于 GA 算法和基于克隆算法由于从全局的角度搜索种群的最优解,克服了 FCM 算法对初始值敏感且容易陷入局部最优的缺点,所以在 20 次实验中均可以找到最优解,但由于算法是基于群体的计算,所以花费了大量的时间.

对比遗传算法和克隆算法,我们发现克隆算法所花费的时间要多于遗传算法,这是因为克隆算法的选择算子、克隆算子、变异算子要比遗传算法的交叉算子、变异算子复杂,所需的计算量较大,所以算法所花费的时间略多于遗传 K-均值.

表 1 三种不同算法结果对比表

算法	最优聚类次数	平均目标函数值	算法平均所需时间
FCM	4	82.7	0.45s
遗传算法	20	73.5	8.3s
克隆算法	20	60.7	8.5s

对于有明显类别界限的数据集,通过图 1 可以看出克隆算法在 20 次实验中,均能在 13 次迭代之内收敛到最优解,进一步证明了算法的收敛性.

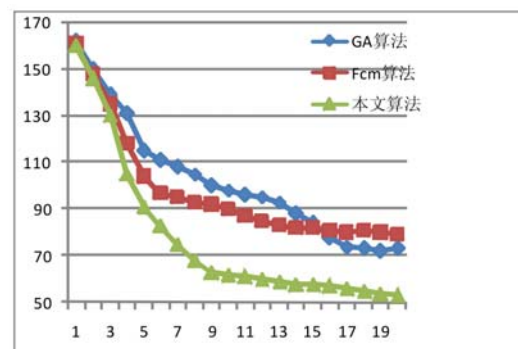


(注:横坐标表示第 n 次实验,纵坐标表示收敛时的迭代次数)

图 1 数据集 1 聚类时收敛的迭代次数

图 2 是克隆 FCM 算法和遗传 FCM 算法在一次实验中的收敛曲线对比图,图中横坐标代表迭代次数,纵坐标为每次迭代运算所获得的目标函数值 J 的最优值(即最小值).我们可以观察到两种改进的算法大约在迭代 10 次的时候即可以找到最优解.本文克隆 FCM 算法由于采用克隆算子不断地增加种群中优良个体的数目,从而加快了找到最优解的速度,体现在图中即克隆 FCM 算法的目标函数的值的下降曲线明显快于遗传 FCM 算法.

由图 2 可知,本文算法的收敛速度明显高于遗传算法,而且在每一次迭代过程中,本文的算法比模糊 c-均值算法及遗传算法的目标函数要小的多,说明其聚类效果明显优于模糊 k 均值算法和遗传算法.



(注:横坐标为迭代次数,纵坐标为目标函数值)

图 2 基于人工免疫的聚类算法与 FCM 及基于 GA 的聚类算法收敛曲线

5 结语

本文提出了一种基于改进人工免疫方法的混合聚类算法,其将模糊 c 均值算法作为一个搜索算子,在模拟进化算法的基础上,快速、有效地搜索到全局最优解.实验结果表明,所提算法在性能上优于模糊 c 均值算法,亦能扩展到其他聚类性能指标可表示为聚类中心函数的聚类问题上.

参考文献

- 1 Han JW, Kambermicheline, 范明, 孟小峰译. 数据挖掘概念与技术(第 2 版). 北京:机械工业出版社, 2007.
- 2 章成志. 主体聚类及其应用研究. 北京:国家图书馆出版社, 2013.
- 3 Chiu SL. Fuzzy Model Identification Based on Cluster Estimation. *Journal of Intelligent Fuzzy Systems*, 1994, 2(3): 267-268.
- 4 Zhi XB, Fan JL, Zhao F. Fuzzy linear discriminant analysis-guided maximum entropy fuzzy clustering algorithm. *Pattern Recognition*, 2013, 46(6): 1604-1615.
- 5 支小斌. 自适应判别降维模糊聚类算法研究[博士学位论文]. 西安:西安电子科技大学, 2013.
- 6 王丙景, 高茂庭. 一种基于遗传算法的聚类集成方法. *计算机工程与应用*, 2013, 49(8): 164-168.
- 7 Chuang YC, Chen CT. A study on real-coded genetic algorithm for process optimization using ranking selection, direction-based crossover and dynamic mutation. *2011 IEEE Congress of Evolutionary Computation, CEC*. 2011. 2488-2495.
- 8 Pakhira MK, Bandyopadhyay A, Maulik U. Validity index for crisp and fuzzy clustering. *Pattern Recognition*, 2004, 37(3): 487-501.
- 8 Bezdek JC, Pal SK. *Fuzzy Models for Pattern Recognition*. Piscataway, USA: IEEE Press, 1999.
- 9 Prabhjot K, Soni AK, Anjana G. A robust kernelized intuitionistic fuzzy c-means clustering algorithm in segmentation of noisy medical images. *Pattern Recognition Letters*, 2013, 34(2): 163-175.
- 10 de Castro LN, Von Zuben FJ. Learning and optimization using the clonal selection principle. *IEEE Trans. on Evolutionary Computation*, 2002, 6(3): 239-251.
- 11 Blake CL, Merz CJ. UCI repository of machine learning data-bases. <http://www.ics.uci.edu/mlearn/MLRepository.html>