第33卷 第23期 计 算 机 工 程 2007年12月
Vol.33 No.23 Computer Engineering December 2007
·博士论文· 文章编号：1000—3428(2007)23—0016—03 文献标识码：A 中图分类号：TP391

基于克隆选择的快速动态聚类算法

张 旭[1,2]，郭 晨[2]
(1. 大连交通大学机械工程学院，大连 116028；2. 大连海事大学自动化与电气工程学院，大连 116026)

摘 要：为了在聚类数不确定的情况下实现聚类分析，通过借鉴生物免疫系统中的克隆选择原理并结合聚类有效性分析，提出了一种基于克隆选择的快速动态聚类算法。该算法可以根据样本数据自动确定聚类数目及中心位置，克服了传统聚类算法容易陷入局部极小值、对初始值敏感的缺点。通过引入新算子及适当选取聚类的初始中心，使算法的收敛速度明显提高，仿真实验结果表明了本算法的有效性。

关键词：克隆选择原理；聚类有效性分析；动态聚类

Fast Dynamic Clustering Algorithm Based on Clone Selection

ZHANG Xu[1,2], GUO Chen[2]
(1. School of Mechanical Engineering, Dalian Jiaotong University, Dalian 116028;
2. School of Automation and Electrical Engineering, Dalian Maritime University, Dalian 116026)

【Abstract】 In order to achieve cluster analysis with unknown number of clusters, this paper proposes a fast dynamic clustering algorithm based on clone selection, which is inspired by the clone selection principle of the vertebrate immune system and combines the cluster validity analysis. It not only adaptively determines the amount and the center's positions of clustering, but also avoids the local optima and the flaw about sensitive to the initialization. The convergence speed of this algorithm is improved obviously through introducing a new search operation and selecting appropriate initial clustering center. Experimental results indicate the validity of the proposed algorithm.

【Key words】 clone selection principle; cluster validity analysis; dynamic clustering

聚类分析是多元统计分析的方法之一，也是统计模式识别中非监督模式分类的一个重要分支。所谓聚类，就是将数据对象分组成多个类或簇，在同一个类中的对象之间具有较高的相似度，而不同类中的对象差别较大[1]。聚类分析主要解决两个问题：数据集中存在多少聚类簇以及这些簇的位置。如果分析前已知簇的数量就称为静态聚类(static clustering)；否则，要在分析过程中获得簇的数量就称为动态聚类[2]。

在传统的聚类方法中，基于目标函数的聚类算法由于把聚类问题归结为一个优化问题，具有深厚的泛函基础，从而成为聚类问题研究的主流[3]，模糊C-均值(FCM)算法就是其中最典型的一种，但由于FCM是采用梯度法求解极值，而梯度法的搜索方向总是沿着能量减少的方向，使得算法很容易陷入局部极小值。同时，FCM的分类数也需要事先人为指定。基于遗传算法(genetic algorithm, GA)的聚类方法尽管能以较高的概率收敛到全局最优点，但收敛速度较慢，而且还容易出现早熟[3]。

根据生物免疫理论中的克隆选择原理提出的克隆选择算法(clonal selection algorithm, CLONALG)[4]，由于其继承了生物免疫系统的众多属性，并具有自学习、自识别、自记忆的能力，因此它不仅能快速提供达到最优解的搜索范围，而且能得到全局最优的结果。

1 基于克隆选择的免疫聚类算法

1.1 模糊C-均值算法

模糊C-均值(fuzzy c-means, FCM)算法是目前应用最为广泛的模糊聚类算法。其将有限集合 $X=\{x_i \mid x_i \in R^P, i=1,2,\cdots,n\}$ 分成 c 类 $(1<c<n)$ 时，其分类矩阵为

$$\boldsymbol{U}=\{u_{ij} \mid i=1,2,\cdots,n,\ j=1,2,\cdots,c\}$$

其中，元素u_{ij}表示第i个数据点属于第j类的隶属度，且满足如下条件：

$$\begin{cases} u_{ij} \in [0,1], \forall i,j \\ \sum_{j=1}^{c} u_{ij}=1, \forall i \end{cases} \tag{1}$$

定义目标函数：

$$J_q(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^q d_{ij}^2(x_i,v_j) \tag{2}$$

其中，$q \in (1,\infty)$ 为模糊指数，用来控制分类矩阵的模糊程度，q 值越大，分类的模糊程度就越高；$v_j \in R^P$ 为类别中心，$\boldsymbol{V}=\{v_i \mid v_i \in R^P, i=1,2,\cdots,c\}$；$d_{ij}(x_i,v_j)$ 为数据点到类别中心的距离测度，一般采用欧式距离：

$$d_{ij}^2=(x_i-v_j)^{\mathrm{T}}(x_i-v_j) \tag{3}$$

应用 Lagrange 乘数法求解在满足式(1)约束条件下，使式(2)取最小的优化问题，可得 $\boldsymbol{U}$, $\boldsymbol{V}$ 的取值式为

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(d_{ij}/d_{ik}\right)^{\frac{2}{q-1}}},\ \forall i,j \tag{4}$$

基金项目：国家自然科学基金资助项目(60474014)；教育部高等学校博士学科点专项基金资助项目(20040151007)

作者简介：张 旭(1972－)，男，博士研究生，主研方向：免疫优化计算，数据挖掘，复杂机械的智能故障诊断；郭 晨，教授、博士生导师

收稿日期：2006-12-20 **E-mail：**fspuzx@newmail.dlmu.edu.cn

$$v_i=\frac{\sum_{i=1}^{n}u_{ij}^{q}x_i}{\sum_{i=1}^{n}u_{ij}^{q}},\quad \forall j \tag{5}$$

FCM 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(4)、式(5)反复迭代，使得式(2)达到极小。从算法过程来看，为使得目标函数值取得极小，采用了一阶导数的方法，使得隶属度的值依赖于训练样本的位置，其迭代求解结果通常是局部最优解，它和初始分类、训练样本的选择顺序密切相关。

1.2 基于克隆选择的免疫聚类算法

在基于克隆选择算法的聚类分析中，就是将聚类问题转化为一个在满足式(1)的约束条件下，使目标函数(抗体-抗原亲合度函数)取得最小的优化问题。算法中把要分类的数据对象视为抗原，把聚类中心看作是免疫系统中的抗体，数据对象的聚类过程就是免疫系统不断地产生抗体，识别抗原，最后产生出可以捕获抗原的最佳抗体过程。

(1)编码方式

算法中的抗体采用基于聚类中心的实数编码方案，根据各自取值范围，将其量化值编码成串。其形式为

$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_l)$

其中，$l=c\times p$，α 中前 p 个量化值代表第个 p 维聚类中心，依次类推。

(2)抗体-抗原亲合度函数

由聚类目标函数可知，目标函数越小，则聚类效果越好，而此时抗体-抗原亲合度应该越大。因此，可以借助目标函数构造亲合度，如下式：

$$f=\frac{1}{J_q(\boldsymbol{U},\boldsymbol{V})+1}=\frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{q}d_{ij}^{2}(x_i,v_j)+1} \tag{6}$$

(3)克隆算子

1)克隆操作。克隆 m 个最佳抗体 $c_1,c_2,\cdots,c_m$，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合度的测度成正比。

2)变异操作。采用高斯变异，即对于克隆变异算子作用到的每一个抗体的分量 $x_i(j)$，$i=1,2,\cdots,m;\ j=1,2,\cdots,p$。

$$x_i'(j)=x_i(j)+\eta_i\times N(0,1) \tag{7}$$

其中，$N(0,1)$ 是满足均值为 0，方差为 1 的正态随机变量；η_j 定义为抗体 x_i 的变异强度因子，随进化代数的增加而减少。

(4)C 搜索算子

除了克隆算子外，结合 FCM 的局部寻优能力定义了一个新的算子：C 搜索算子。对于每个记忆细胞，先按式(4)求出其对应模糊划分矩阵 $\boldsymbol{U}$，然后再按式(5)更新各中心，用更新后的 c 个中心构成新的个体。

1.3 算法描述

(1)设定聚类类数 c。设定终止条件 Sc，抗体总个体数为 P。

(2)随机产生个体数为 P 的抗体集合 $P(k)$，$k=0$。

(3)计算每个抗体的亲合度。

1)利用 $\{v_i,1\le i\le c\}$ 及式(3)计算 d_{ij}^2。

2)利用式(4)计算 $\boldsymbol{U}=[u_{ij}]_{c\times n}$。

3)利用 $\boldsymbol{U}$，d_{ij}^2 及式(2)计算目标函数 $J_q(\boldsymbol{U},\boldsymbol{V})$，进而由式(6)计算出每个个体的亲合度 f。

(4)抽取亲合度高的 m 个抗体进入记忆细胞。

(5)对每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。

(6)计算 D 中每个抗体的亲合度。

(7)选取 D 中亲合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。

(8)对新记忆细胞群进行抗体抑制，清除相似的记忆细胞，保留 m 个亲合度高的抗体组成新的记忆细胞群。如果满足终止条件 Sc，转向(9)，否则转向(5)。

(9)将记忆细胞中亲合度最高的抗体所对应的分类结果作为数据集 X 在分类数为 c 下的最佳聚类结果。

2 基于克隆选择的动态免疫聚类算法

2.1 聚类有效性分析

习惯上，把评价聚类结果的问题称为聚类有效性分析[5]。即需要分析，聚类的结果是不是最好的以及是不是可以信赖的。一般来说，聚类模式的衡量标准是聚类中心的模式尽量紧密，各聚类之间要尽量独立。文献[6]对目前常用的如*DB*, *XB*等聚类有效性指标作了比较，并提出了一种聚类效果更好的指标，定义为

$$I(c)=\left(\frac{1}{c}\times\frac{E_1}{J_c}\times D_c\right)^r \tag{8}$$

其中，$J_c=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\left\|x_k-v_i\right\|$；$D_c=\max_{i,j=1}^{c}\left\|v_i-v_j\right\|$；$\|\cdot\|$为欧几里德范数；$c$ 为聚类数，指数 $r\ge1$ 为实数，一般取 2；E_1对于给定的数据集是一个常数，主要起标准化作用，避免指标出现极小的数值；J_c 表示各类内的点与中心间的距离之和；D_c 表示各类中心间的最大距离。好的分类应该是类内的点集中紧凑，类与类的间距尽可能大，因此，I值越大说明聚类效果越好。

2.2 聚类初始中心的选择

结合聚类有效性分析进行动态聚类的通常方法是：依次改变给定的分类数目，用聚类算法求出对应的中心，再根据相应的聚类有效性判别准则，找出最符合指标的分类数。由于每次的初始聚类中心均为随机选取，相邻两次聚类过程间毫无联系，使得前次分类结果中的有效信息不能被充分利用。

在大量数据分析中发现，在依次增加聚类数目时，模糊聚类算法总是将类内聚最大的一类分成两类，而其他类(类成员及中心位置)基本不变。类内聚 S 由下式计算，其中，v_j，d_{ij} 意义与标准 FCM 相同，x_{ij} 为第 j 类的类成员，n 为第 j 类中成员的个数。

$$S_j=\sum_{i=1}^{n}d_{ij}^2(x_{ij},v_j) \tag{9}$$

因此，算法中将前次分类结果中类内聚最大的类先分成 2 类，再连同其它类中心组成新的初始中心，与随机选取的初始中心相比，收敛速度有了明显提高。

2.3 动态免疫聚类算法描述

(1)设定初始聚类类数 $c=2$。

(2)对数据集 X 按 1.3 节进行免疫聚类。

(3)将记忆细胞中亲合度最高的个体所对应的分类结果就作为数据集 X 的一种可能聚类结果。并根据式(8)计算出相应的聚类效果指标 I_c。

(4)如果 $c\le\sqrt{n}$ [7]，则 $c=c+1$，即增加聚类类数。否则转至(6)。

(5)对于记忆细胞集中的每一个记忆细胞，分别将其所对应的分类结果按式(9)计算出每一类的类内距，再将类内距最

大的一类用标准 FCM 算法分成两类,所得中心连同其它的类中心形成新的初始抗体。再补充一定随机抗体，组成初始抗体集，转至 1.3 节(3)。

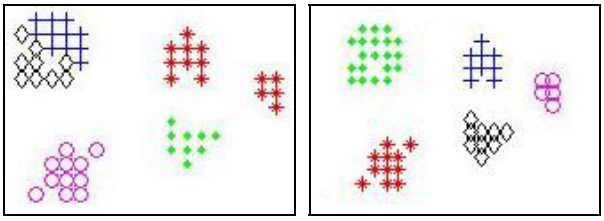
(6)在全部 $c(1<c\le\sqrt{n})$中，取得最大 I_c 值所对应的分类结果就是数据集 X 的最佳聚类结果。

3 仿真实验

3.1 静态聚类性能测试

为直观地显示实验结果，使用一组包含了 5 类的共 55 个数据点的二维原始数据集，分别用传统 FCM、基于遗传算法(GA)的模糊聚类算法及本文算法进行对比测试。本文聚类算法的参数设置如下：设初始抗体数 P=20，记忆细胞在抗体集所占比例为 10%，每个记忆细胞所产生的临时抗体数为 100，进化的终止条件为记忆细胞中最佳个体的亲合度连续 30 代未得到改善，或达到最大进化代数 100。基于 GA 的模糊聚类算法采用文献[8]的方法，参数设置如下：种群规模为 20，交叉率为 0.9，变异率为 0.05，采用数目为 2 的精英保留策略。

采用以上方法分别独立运行 30 次，由于传统的 FCM 对初始值敏感，因此其正确识别率较低，平均正确识别率只有 56.7%(如图 1(a))，基于 GA 的算法有 5 次发生早熟，没有得到最优解，而本文的基于克隆选择的免疫聚类算法全部达到全局最优(如图 1(b))。



(a)传统 FCM 分类结果　　(b)本文算法分类结果

图 1　分类结果

图 2 分别给出了基于 GA 的模糊聚类算法，基于标准克隆选择算法以及引入 C 搜索算子后的新克隆选择模糊聚类算法的平均收敛特性曲线。从中可以看出每一代中，基于克隆选择的模糊聚类算法得到的最优解都比遗传算法的要好，而且收敛速度更快；在引入 C 搜索算子后，新算法的收敛速度更有了明显的提高。

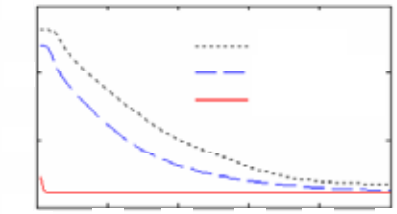


图 2　3 种算法的收敛曲线

3.2 动态聚类性能测试

首先使用在簇中心附近产生具有高斯分布的包含 8 个簇的三维数据集进行测试，每个簇包含 100 个样本。进行 10 次免疫模糊动态聚类，结果表明，本算法 100%地收敛到正确的聚类数，并均获得最佳聚类中心。

为了进一步验证算法的有效性，使用 UCI 机器学习库中的 IRIS 植物样本数据集，该数据集由分别属于 3 种植物的 150 个样本组成。对 IRIS 数据集进行 10 次免疫模糊动态聚类，结果表明，本算法不但可以正确给出其最佳聚类数 3，而且最优目标函数值 6 050.6，较文献[8]所给出的遗传 FCM 所得的 6 057.6 更小。

在收敛性能方面，由于标准克隆选择算法收敛时间相对较长，这里只比较带 C 搜索算子的新克隆选择算法在采用随机选取初始中心与依据上次聚类结果选取聚类中心两种方法的收敛性能。

表 1 为三维人工数据集和 IRIS 数据集在两种方法下平均收敛时间对照表，系统的配置为 P4 1.8 GHz, 256 MB 内存, 采用 WindowsXP 操作系统。从中可以看出，利用前次聚类结果确定初始中心比随机选取初始聚类中心平均收敛时间可缩短近 40%。

表 1　不同初始中心选取方式收敛性能对比

初始中心选取方式	三维人工数据集		IRIS 数据集	
	进化代数	收敛时间/s	进化代数	收敛时间/s
随机选取	323	358	403	527
依据前次	207	227	253	336

4 结束语

聚类问题在一定条件下可以归结为一个带约束的优化问题。基于生物免疫理论而提出的克隆选择算法作为一种鲁棒性很强的优化算法不仅能快速达到最优解的搜索范围，而且能得到全局最优的结果。将其应用到聚类分析之中，有效地克服了传统聚类算法容易陷入局部极小值，对初始化敏感的缺点。

结合聚类有效性准则提出一种快速动态免疫聚类算法。通过引入 C 搜索算子并充分利用动态聚类过程中前次聚类结果的有效信息，确定初始中心位置。大大缩短了算法的收敛时间。仿真实验表明，该算法不但可以根据数据自动确定聚类数目及中心位置，而且收敛速度快，具有广泛的应用前景。

参考文献

1 Han Jawei, Kamber M. Data Mining: Concepts and Techniques[M]. San Fransisco: Morgan Kaufmann, 2000.

2 Franti K. Dynamic Local Search for Clustering with Unknown Number of Clusters[C]//Proc. of the 16th IEEE International Conference on Pattern Recognition, Quebec, Canada. 2002: 240-243.

3 高新波. 模糊聚类算法的优化及应用研究[D]. 西安: 西安电子科技大学, 1999.

4 De Castro L N, Von Zuben F J. Learning and Optimization Using the Clonal Selection Principle[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(3): 239-250.

5 Pakhira M K, Bandyopadhyay A, Maulik U. Validity Index for Crisp and Fuzzy Clustering[J]. Pattern Recognition, 2004, 37(3): 487-501.

6 Maulik U, Bandyopadhyay S. Performance Evaluation of Some Clustering Algorithms and Validity Indices[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 24(12): 1650-1654.

7 于　剑, 程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学(E 辑), 2002, 32(2): 274-280.

8 Hall L O, Ozyurt I B, Bezdek J C. Clustering with a Genetically Optimized Approach[J]. IEEE Transactions on Evolutionary Computation, 1999, 3(2): 103-112.