文章编号: 1001-0920(2005)11-1261-04

一种基于克隆选择的聚类算法

罗印升[1,2], 李人厚[1], 张维玺[2]
(1. 西安交通大学 系统工程研究所, 西安 710049; 2. 江苏技术师范学院 电气与信息工程系, 江苏 常州 213001)

摘 要: 将克隆选择原理同典型的划分聚类方法结合起来, 提出一种克隆选择聚类算法. 该算法具有完成任意形状数据集聚类的能力, 可以自动确定簇的数目并得到簇的描述信息, 计算量小, 参数设置容易, 适用于具有实值连续属性的数据集. 基于模拟数据集和基准数据集分别进行实验, 结果表明该算法是有效的.

关键词: 克隆选择; 聚类算法; 簇分析; 数据集
中图分类号: TP18 **文献标识码**: A

Clustering Algorithm Based on Clone Selection Theory

LUO Yin-sheng[1,2], *LI Ren-hou*[1], *ZHANG Wei-xi*[2]
(1. Institute of System Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. Department of Electrical and Information Engineering, Jiangsu Teachers University of Technology, Changzhou 213001, China. Correspondent: LUO Yin-sheng, E-mail: yishluo@sohu.com)

Abstract: Based on clone selection theory and typical partition clustering approach, a new clustering algorithm is proposed. The algorithm is capable of completing clustering of datasets with arbitrary shape, and acquiring the number and description of clusters. Besides, the algorithm has smaller computational cost, and the parameter can be set easily. The algorithm is applicable to the datasets that have real-value and continuous attributes. The experiments with two simulation datasets and three benchmark datasets show the effectiveness of the algorithm.

Key words: Clone selection; Clustering algorithm; Cluster analysis; Datasets

1 引 言

簇分析(或称聚类)是指根据某种准则, 使待分析数据集中具有相似特性的对象成为一簇, 从而达到簇内对象间相似性最大, 不同簇间相似性最小[1]. 若用距离来度量对象之间的相异性, 则簇分析的目标就是使簇内对象之间的距离最小化, 而簇间距离最大化. 簇分析已广泛应用于生物学、医学、机器学习、图像分割与压缩、目标特征识别、信息恢复和数据挖掘[2].

聚类算法面临的主要问题有: 能否自动确定簇的数目; 是否对初始的参数和噪声敏感; 是否具有可扩放性和较快的速度; 处理混合数据的能力; 任意形状数据集的聚类和获得簇的信息描述. 为解决这些问题, 新的算法不断出现. 文献[3]提出一种解决任意形状数据集的聚类算法, 可自动获得簇的数目, 但其参数设置无参考基准而难以进行, 且无簇的信息描述. [4]提出一种蚁群聚类方法, 它具有全局最优、快速的特点, 但簇数目需要预先确定, 所有的数据均参与编码, 无处理任意形状数据集的能力.

本文将免疫系统中B 细胞克隆和突变机理与经典的划分聚类方法相结合, 以各数据之间的欧氏距离平均值作为设置数据划分和聚类中心融合的参考值, 以B 细胞的克隆、突变作为进化机制, 提出一种新的聚类算法, 称为克隆聚类算法(CCA). 它能自动获得簇的数目, 具有聚类任意形状数据集的能力, 并可获得簇的特征描述. 基于2 个模拟数据集和3 个基准数据集对CCA 算法进行了实验, 并与文献[4]的算法进行比较, 结果表明本文提出的算法具有较

收稿日期: 2004-11-09; **修回日期**: 2005-03-14

作者简介: 罗印升(1964—), 男, 陕西武功人, 博士生, 从事进化计算、智能控制等研究; 李人厚(1935—), 男, 浙江宁波人, 教授, 博士生导师, 从事智能控制、CSCW 等研究.

好的性能

2 概念与定义

2.1 数据集的描述

设待分析的数据集 X 由 n 个数据组成, 即 $X = \{X_1, X_2, \cdots, X_n\}$. 一个数据 $X_i \in R^d$ 称为一个模式(对象) 或者特征矢量, 它由 d 个测度组成, 即 $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,d}\}$, $x_{i,d}$ 称为特征或属性, 它的值称为属性值

2.2 相异性和准则函数

在簇分析中, 通常用定义在特征空间上的距离来度量两个模式之间的相异性[2]. 本文主要讨论定量及连续性的实数特征值, 因此使用 Euclidean 距离 E_d 来度量模式之间的相异性, 即

$$E_d(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2}. \quad (1)$$

准则函数用来判断聚类的质量, 用各模式和对应的聚类中心点之间的欧氏距离平方和之和, 即平方误差和作为准则函数[4], 记为 E,

$$E = \sum_{j=1}^{c} \sum_{i=1}^{n} \sum_{k=1}^{d} \| x_{i,k} - m_{j,k} \|^2, \quad (2)$$

其中 $m_{j,k}$ 是第 j 个中心点的第 k 个特征值. 这样准则函数的值越小划分越好.

2.3 数据划分与聚类中心点融合

为完成数据的聚类, 先将数据集划分为小的区域, 合并相近的区域; 然后重新划分, 使准则函数最小; 最后将最好的划分区域进行连接, 以确定簇的数目. 以数据集中所有模式之间欧氏距离平均值作为参考设定划分阈值, 阈值定义为

$$T = 2A \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_d(X_i, X_j) / (n(n-1)). \quad (3)$$

其中 A 是调节因子, $0 < A \leq 1$. 这样, 当某个模式与某个聚类中心点之间的距离小于 T 时, 它属于该中心点所代表的区域(以中心点为中心, 以 T 为半径的区域). 此时, 对应的两个区域至少有 1/3 重叠, 需要合并, 并重新计算中心点. 划分完成后, 若中心点之间的距离大于 T 而小于 $2T$, 则说明这两个小区域是相连的, 它们之间可以建立连接关系, 属于同一个簇

2.4 簇的描述

如何描述聚类算法最终所产生的簇是一个重要问题, 它对于获取知识、作出决策均有重要作用. 一般对簇的描述方法有重心、边界点的集合、聚类树中的节点及模式属性的逻辑表达式等. 为了对任意形状的数据集聚类所获得的簇描述也有效, 定义簇特征 (CF) 为三元组, 由簇中的模式数目 N, 中心点的数目 N_0(包括位置坐标) 及半径 R 组成, 即 $CF = \{N, N_0, R\}$.

3 克隆聚类算法

在CCA 中作为数据集的预处理, 需要计算所有模式之间的欧氏距离和的平均值, 以此作为设置数据集划分阈值 T 的参考. 因而, 用户只要在 $(0, 1]$ 的范围内选择合适的 A 值即可. CCA 的思想是以划分阈值 T 和准则函数作为数据集划分的依据和质量标准, 以 B 细胞的成熟和进化机理作为搜索更好划分的方法, 将数据集划分为若干个以聚类中心为代表点的区域, 包括合并相近的中心点并重新计算中心点, 然后根据中心点连接规则进行操作以确定簇的数目, 最后再次划分数据集获得簇的特征描述. 因此, CCA 不需要用户预先给定聚类数目, 参数的设置比较容易, 可以完成任意形状数据集的聚类. 算法的具体步骤如下:

Step 1: 计算数据集中各模式之间的欧氏距离平均值, 其计算式为

$$2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_d(X_i, X_j) / (n(n-1)). \quad (4)$$

Step 2: 两次划分数据集. 在数据集中, 按均匀分布随机选取 $NT(\sqrt{n})$ 个数据作为初始聚类中心点(NT 表示取整数), 划分数据集. 若模式与聚类中心点之间距离小于 T, 则该模式归入距离最小的聚类中心点; 若模式与所有聚类中心点之间距离均大于 T, 则将该模式作为新的聚类中心点. 所有的模式划分完毕后, 采用重心方法计算每个代表区域的重心, 此重心就作为新的聚类中心点坐标. 若两个中心点的距离小于 T, 则进行合并, 再次划分数据集, 并将准则函数值、中心点保存在最优划分单元中.

Step 3: 个体编码, 产生初始群体. 采用实数编码, 将 Step 2 中得到的中心点编码成一个个体, 以此个体作为母细胞进行克隆操作, 产生 4 个细胞

Step 4: 个体的增殖, 即克隆. 将 4 个细胞分为 4 组, 每组克隆产生 4 个子细胞组成一个群体

Step 5: 突变. 对 Step 4 中克隆得到的子细胞群体进行突变, 由于采用了实数编码, 变异采用高斯变异法. 为了保证变异的有效性, 变异后的中心点与原中心点之间的距离必须小于 T.

Step 6: 合并距离小于 T 的中心点

Step 7: 选出本次最优划分. 以变异、合并后的中心点为中心, 分别完成数据集的划分, 然后计算准则函数值, 在每一组群体中选择一个准则函数值最小的个体作为该组的母细胞. 所有组中选择准则函数值最小的划分为本次最优划分, 比较本次的最优划分准则函数值和最优划分单元中准则函数值, 在最优划分单元中保留较好的一个. 若最优划分 t 次反复后保持不变, 则继续执行; 否则, 转到 Step 4

Step 8: 确定簇的数目及每个簇中的模式数目. 以最优划分的中心点为中心, T 为半径, 划分数据集. 当中心点之间的距离大于 T 而小于 $2T$ 时, 建立连接关系, 它们属于同一簇. 不能建立连接关系的中心点代表了另外的簇, 这样便可自动获得簇的数目及每个簇中的模式数目.

Step 9: 输出簇的数目与描述, 算法结束.

4　仿真实验

4.1　模拟数据集

模拟数据集 1 和 2 分别如图 1 和图 2 所示, 用来验证 CCA 完成任意形状数据集聚类的能力, 自动确定簇的数目并获得簇的信息描述. 数据集 1 由 186 个二维数据组成, 构成了一个圆环和一个近似球形状的区域, 形成了两个簇. 数据集 2 由 286 个二维数据组成, 构成了两个相互包含的抛物线状的簇.

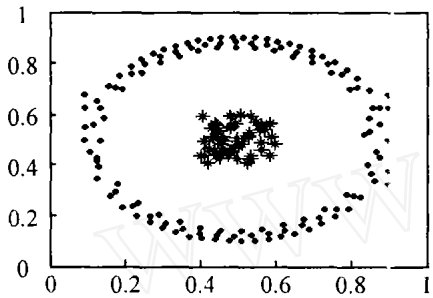

图 1　数据集 1

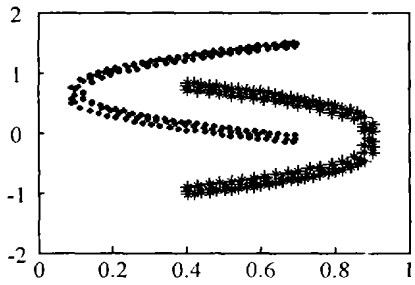

图 2　数据集 2

应用 CCA 分别对两个数据集进行实验, 结果如图 3 和图 4 所示. 算法最大反复次数设置为 400, t = 15. 对于数据集 1, CCA 的参数为 A = 0.6, 最终得到两个簇, 簇的信息描述为: 圆环簇 = {133, 8, 0.22}, 中心点的坐标为{(0.161 4, 0.674 0), (0.828 6, 0.314 3), (0.841 9, 0.648 1), (0.610 3, 0.140 3)}, {(0.617 7, 0.855 8), (0.159 2, 0.358 7), (0.362 9, 0.148 9), (0.361 1, 0.851 3)}; 近似球形簇 = {55, 1, 0.22}, 中心点的坐标为{(0.491 2, 0.502 1)}. 对于数据集 2, CCA 的参数为 A = 0.45, 算法最终得到两个簇, 左边抛物线形状簇的描述为{155, 15, 0.09}, 右边抛物线形状簇的描述为{131, 12, 0.09}, 中心点的坐标略.

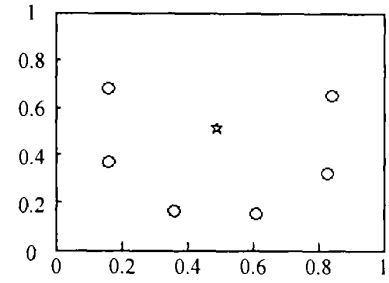

图 3　数据集 1 的聚类结果

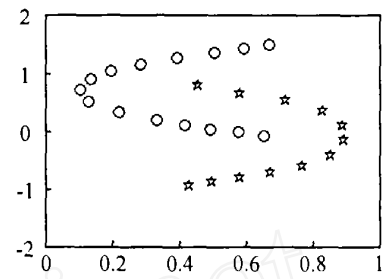

图 4　数据集 2 的聚类结果

4.2　基准数据集

本文使用的3个基准数据集来自站点[5], 它是一个开放的机器学习数据库. 这 3 个数据集分别是 Iris Plants Data, Wine recognition data 和 Thyroid gland data.

Iris 数据集由 150 个模式组成, 每个模式有 4 个数值属性(萼片长度, 萼片宽度, 花瓣长度和花瓣宽度), 可分为 Iris setosa, Iris versicolor 和 Iris virginica3 类. Iris setosa 数据类独立于另外两类, 后两类相关联.

Wine 数据集由 178 个模式组成, 每个模式有 13 个数值属性, 分为 3 类.

Thyroid 数据集由215 个模式组成, 每个模式有 5 个数值属性. 数据集划分为甲状腺机能正常、甲状腺机能亢进和甲状腺机能减退 3 类.

下面将CCA 同ACO 算法、GA 算法、SA 算法及 TS 算法[4]进行比较. 它们均属于典型的随机优化算法, 对随机优化算法而言, 其性能和初始解的产生有很大关系. 对此, 文献[4]的实验结果是每种算法对每个数据集独立运行 10 次, 然后得到聚类准则函数的最好值(F_{best}, 准则函数值最小)、平均值(F_{avg})、最差值(F_{worst})及准则函数评价次数的平均值. 对于同一数据集, 准则函数的评价次数和算法的执行时间通常是一致的, 因此CCA 算法也采用这些评价方法和性能指标.

对于 Iris 数据集, CCA 的参数 A = 0.2, 实验结果和文献[4]的结果如表 1 所示. CCA 能够自动确

定出 3 个簇, 具有较小的聚类准则函数评价次数和函数值. 3 个簇的描述分别为{50, 2, 0.5}, {53, 3, 0.5}, {47, 4, 0.5}. 由于后两个簇线性相关, 交织在一起, 由簇的描述信息可以看出, CCA 出现了错误划分. 文献[4]没有给出划分的结果, 但CCA 的划分结果较文献[6]的结果要好.

表 1 CCA 和文献[4]对 Iris 数据集的实验结果比较

方法	准则函数值			准则函数评价次数
	F_{best}	F_{avg}	F_{worst}	
ACO *	97.100 777	97.171 546	97.808 466	10 998
GA *	113.986 503	125.197 025	139.778 272	38 128
TS *	97.365 977	97.868 008	98.569 485	20 201
SA *	97.100 777	97.136 425	97.263 845	29 103
CCA	90.358 466	91.683 272	92.528 347	8 640

注: “ * ”表示其结果见文献[3], 下同.

对于 Wine 数据集, CCA 的参数 $A = 0.3$, 实验结果和文献[4]的结果如表 2 所示. CCA 能够自动确定出 3 个簇, 同样具有较小的聚类准则函数评价次数和函数值. 3 个簇的描述分别为{59, 3, 112}, {71, 4, 112}, {48, 3, 112}.

表 2 CCA 和文献[4]对 Wine 数据集的实验结果比较

方法	准则函数值			准则函数评价次数
	F_{best}	F_{avg}	F_{worst}	
ACO *	16 530.533 807	16 530.533 807	16 530.533 807	9 306
GA *	16 530.533 807	16 530.533 807	16 530.533 807	33 551
TS *	16 666.226 987	16 785.459 275	16 837.535 670	22 716
SA *	16 530.533 807	16 530.533 807	16 530.533 807	7 917
CCA	15 532.386 217	15 552.283 165	15 843.543 862	7 568

对于 Thyroid 数据集, CCA 的参数 $A = 0.4$, 实验结果和文献[4]的结果如表 3 所示. CCA 同样能够自动确定出 3 个簇, 也具有较小的聚类准则函数评价次数和函数值. 三个簇的描述分别为{150, 9, 7}, {35, 2, 7}, {48, 3, 7}.

表 3 CCA 和文献[4]对 Thyroid 数据集的实验结果比较

方法	准则函数值			准则函数评价次数
	F_{best}	F_{avg}	F_{worst}	
ACO *	10 111.827 759	10 112.126 903	10 114.819 200	25 626
GA *	10 116.294 861	10 128.823 145	10 148.389 608	45 003
TS *	10 249.729 17	10 354.315 021	10 438.780 449	29 191
SA *	10 111.827 759	10 114.045 265	10 118.934 358	28 675
CCA	9 124.754 312	9 156.683 701	9 283.458 703	13 184

5 CCA 的计算复杂性分析和 A 值的选取

算法的计算复杂性包括时间和空间复杂性, 这里主要分析CCA 的时间复杂性. 在Step 1 中, 时间复杂性为 $O\left[\frac{1}{2}(n(n-1))\right]$, 记为 $O(n^2)$, 与迭代次数 t 无关. 在 Step 2 中, 时间复杂性为 $O(\mathrm{NT}(\sqrt{n}) * n) = O(n^{3/2})$. 在 Step 4～ Step 7 中, 时间复杂性主要由计算准则函数产生, 记为 $O(N_0 * M * G)$. 其中, M 是小区域中包含数据点最多的数据点数目, G 是算法迭代次数. 中心点数目 N_0 的最大值取为 $\mathrm{NT}(\sqrt{n})$, M 的最大值取为 n, 则 $O(N_0 * M * G) = O(G * n^{3/2})$. 在 Step 8 中, 时间复杂性为 $O(N_0 * n)$, 最终的中心点数目 $N_0 \ll n$, $O(N_0 * n) = O(n)$. 综上, CCA 的时间复杂性为

$$O(n^2) + O(n^{3/2}) + O(G * n^{3/2}) + O(n) = O(n^2). \quad (5)$$

如果将 Step 1 归入数据预处理, 则 CCA 的时间复杂性为 $O(G * n^{3/2})$.

参数 A 值在 (0, 1] 内由大到小试探选取. 若取值太大, 则所有的数据可能归为一个簇; 若取值太小, 将会得到太多的簇数, 使计算量加大. 这样选取 A 值的过程较没有范围、没有方向的凑试方法有了很大改进.

6 结 论

本文将免疫系统中B 细胞克隆和突变机理与经典的划分聚类方法相结合, 以数据集中各数据之间的欧氏距离平均值作为设置数据划分与聚类中心融合的参考值, 以B 细胞的克隆、突变机制作为搜索更好划分的方法, 提出了一种新的聚类算法(CCA). 该算法具有聚类任意形状数据集的能力, 能够自动获得簇的数目和描述信息. 基于 2 个模拟数据集和 3 个基准数据集对该算法进行了实验, 结果表明算法是有效的.

参考文献(References)

[1] Han J W, Micheline Kamber. *Data Mining: Concept and Techniques* [M]. Vermont: Morgan Kaufmann Publishers, 2000.

[2] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review [J]. *ACM Computing Surveys*, 1999, 31(3): 264-323.

[3] Ma S, Wang T J, Tang S W, et al. *A New Fast Clustering Algorithm Based on Reference and Density* [M]. Berlin: Springer-Verlag, 2003.

[4] Shelokar P S, Jayaraman V K, Kulkarni B D. An Ant Colony Approach for Clustering[J]. *Analytica Chimica Acta*, 2004, 509(2): 187-195.

[5] University of California at Irvine. Machine Learning Repository of Data Sets [DB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1993.2/2004.9.

[6] Gautam Garai, Chaudhuri B B. A Novel Genetic Algorithm for Automatic Clustering [J]. *Pattern Recognition Letters*, 2004, 25(2): 173-187.