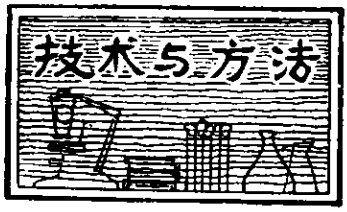

聚类分析在微生物数值分类上的应用

马俊才　　赵玉峰
（中国科学院微生物研究所，北京）

聚类分析是研究如何用数学方法将事物进行分类的学科。在聚类分析中人们把每个样品都看作是 n 维空间的一个点，按照不同的聚类方法定义样品组间的距离，利用聚类方法把距离小的样品归为一类，使同类间的样品尽量相似，不同类间的样品则相异。聚类分析可使原有样品归并成较少数目的类群或引出更加清晰的聚合趋势，以揭示样品之间的关系。

自数值分类法问世以来国外已作了大量研究，到 1973 年 Sneath 和 Sokal 对国际上常用算法的数学模型进行了系统的归纳[1]。在我国 1975 年由徐浩等首次引入并成功地应用于枯草杆菌的研究[2]。其后陆续有人开展这方面应用，使这一技术得到了较大发展[3-7]。本程序实现的是系统聚类法。

（一）系统的硬件配制

主机为 ACS8600（ALTOS 公司生产），配有一个 40 兆硬盘，一个 8 吋软盘，一台 17 兆磁带机，一台 SR6602 型六笔绘图仪，一台 FX-100 型打印机和两个 CRT 终端。

（二）方法和步骤

系统聚类常见的七种方法有其各自的聚类策略，1967 年 Lance 和 Williams 确定了一个适用于七种方法的通用公式：

$$D_{kr} = a * D_{ki} + b * D_{kj} + c * D_{ij} + d * |D_{ki} - D_{kj}|$$

式中 D_{kr} 表示样品组 k 与样品组 r 之间距离，样品组 r 是样品组 i 与样品组 j 的并组（即 $G_r = G_i \cup G_j$）。a、b、c、d 是随不同聚合策略所取的不同系数。各种方法的系数见表 1。程序流程见图 1。

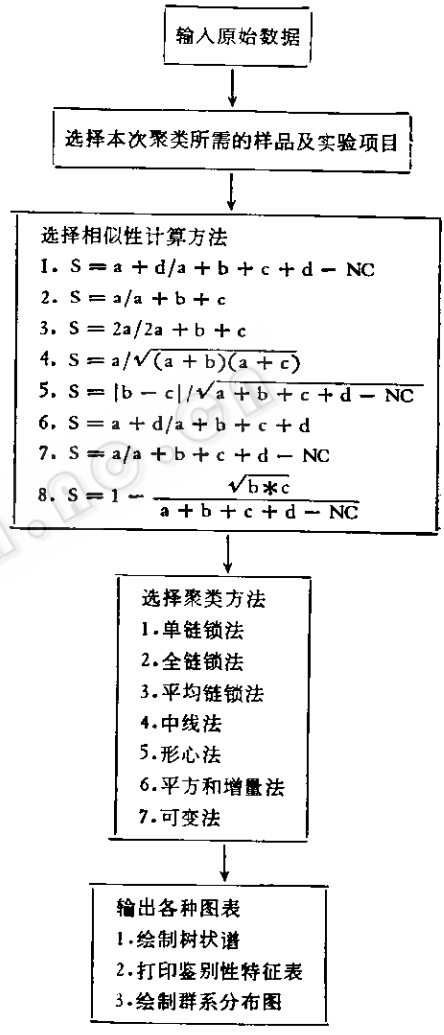

图 1　系统总框图

程序执行主要过程如下：

1. 输入所有菌株各项原始实验结果

$$\begin{pmatrix} x_{11} & x_{12} \cdots\cdots x_{1m} \\ x_{21} & x_{22} \cdots\cdots x_{2m} \\ \vdots & \\ x_{n1} & x_{n2} \cdots\cdots x_{nm} \end{pmatrix}$$

其中 x_{ij} 表示第 i 个样品第 j 项实验结果。

2. 确定本次计算的样品和实验项目：对已

表1 系统聚类法参数表

$D_{kr} = a*D_{ki} + b*D_{kj} + c*D_{ij} + d*|D_{ki} - D_{kj}|$

$(G_r = G_i \cup G_j)$

方法名称	a	b	c	d	特性
单链锁法 最近邻体法	1/2	1/2	0	−1/2	压缩,单调
全链锁法 最远邻体法	1/2	1/2	0	1/2	扩张,单调
平均链锁法 类平均法	$\frac{n_i}{n_{i+j}}$	$\frac{n_j}{n_{i+j}}$	0	0	保持,单调
中线法 中间距离法	1/2	1/2	−1/4	0	保持,非单调
形心法 重心法	$\frac{n_i}{n_{i+j}}$	$\frac{n_j}{n_{i+j}}$	$-\frac{n_i * n_j}{n^2_{i+j}}$	0	保持,非单调
平方和增量法 离差平方和法	$\frac{n_k + n_i}{n_k + n_{i+j}}$	$\frac{n_k + n_j}{n_k + n_{i+j}}$	$-\frac{n_k}{n_k + n_{i+j}}$	0	压缩,单调
可变法	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	<1	0	随β的取值而改变
加权平均链锁法	1/2	1/2	0	0	保持,单调

输入的数据，研究人员可按需要采用全部或选用部分样品、特征进行聚类。

3. 计算样品间相似性系数

按用户选用的相似性计算法计算各样品间的相似性系数，将其转换成相异性指标做成距离三角矩阵：

$$\begin{pmatrix} 0 & & & \\ d_{21} & 0 & & \\ d_{31} & d_{32} & 0 & \\ \vdots & & & \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

其中 d_{ij} 表示第 i 个样品与第 j 个样品间的距离。在样品和实验项目较多时运算时间过长，为便于管理，程序增加了暂停功能，程序员可随时中断程序运行，以后再从断点继续运算。

4. 选择聚类方式以确定公式中的 a，b，c，d。

5. 进行聚类[8]

a. 在距离三角矩阵中找一最小数 $D^{(1)}_{kl}$（设 $k < l$，因为 $D_{kl} = D_{lk}$）

b. 将 $G^{(1)}_k$ 和 $G^{(1)}_l$ 合并为 $G^{(2)}_k$，即：

$$G^{(2)}_k = G^{(1)}_k \cup G^{(1)}_l$$

$$G_i^{(2)} = G_i^{(1)}(i = 1, 2, \cdots\cdots, l-1, \text{且 } i \neq k)$$

$$G_i^{(2)} = G_{i+1}^{(2)}(i = l, \cdots\cdots, n-1)$$

c. 修改原来的距离三角阵为：

$$D_{ij}^{(2)} = D_{ij}^{(1)}(i, j = 1, 2, \cdots\cdots, l-1, \text{且 } i, j \neq k)$$

$$D_{ij}^{(2)} = D_{ij+1}^{(1)}(i = 1, \cdots l-1, i \neq k, j = l, \cdots, n-1)$$

$$D_{ij}^{(2)} = D_{i+1,j}^{(1)}(i = l, \cdots, n-1, j = 1, 2, \cdots, l-1, \text{且 } j \neq k)$$

$$D_{ij}^{(2)} = D_{i+1,j+1}^{(1)}(i = l, \cdots, n-1, j = l, \cdots, n-1)$$

$$D_{ik}^{(2)} = a * D_{ik}^{(1)} + b * D_{il}^{(1)} + c * D_{kl}^{(1)} + d * |D_{ik}^{(1)} - D_{il}^{(1)}|$$

$$(i = 1, 2, \cdots, l-1, \text{且 } i \neq k)$$

$$D_{i-1,k}^{(2)} = a * D_{ik}^{(1)} + b * D_{il}^{(1)} + c * D_{kl}^{(1)} + d * |D_{ik}^{(1)} - D_{il}^{(1)}|$$

$$(i = l+1, \cdots, n)$$

d. 对修改后的距离矩阵重复步骤5，直到所有样品都归为一类则聚类结束。

6. 打印输出各种图表： 包括原始数据表，归群前后的两个三角矩阵，鉴别性特征表，分类树状谱(四张树状谱图从略)及特定相似水平下的群系分布图等。

(三) 各种聚类方法的效果比较

我们用本系统先后对醋酸菌(148株132项实验)，根瘤菌(55株208项实验)，鱼类肠道细菌(103株112项实验)等数据进行聚类分析，都取得较好结果，现以醋酸菌为例将各类方法的效果逐一比较。

1. 单链锁聚合方式[7]。这是最早应用的聚类方式，其优点是聚合策略简单易于理解，该策略定义样品组间的距离为：两组中距离最小(相似性最大)的两样品间的距离。在两样品组合并后，这个并组与其余各样品组之间有更强的聚合力，从而促使这个并组更容易在一个较高的相似性水平上与其它组合并，即所谓空间压缩性质。根据计算结果，单链锁的整体聚合水平(相似性)高达77%，这种性质使得结果不能很好地表现较小的类群，减少了分类层次，降低了分辨力。

2. 全链锁聚合方式[7]。聚合策略与单链锁法相反，使两个样品组的并组较不易与其它样品组合并，阻碍了进一步聚合。这种加大组间距离的性质称作空间扩张的性质。计算结果全链锁整体聚合水平(相似性)仅为36%，这种性质使得类间相似性水平降低，增加分类层次提高了分辨力。在很多情况下，过多地增加分类层次会增加错分的机率降低结果的可信性。

3. 平均链锁法[7]。此法聚合策略是定义两样品组间的距离为两组间样品的平均距离。由于此法不具有空间收缩或扩张性质，是空间保持的，一些样品合并后，其并组与其它样品组间距离未被扩大或缩小，因而并组并不阻碍或促进聚合趋势。计算结果平均链锁的整体聚合水平（相似性）为63%，居于单链锁和全链锁之间。据平均链锁在68%水平下的群系分布图(略)，醋酸菌148株清晰地分成三个类群与传统分类结果基本一致。平均链锁法在聚合过程中具有单调性，其结果有较高的可信性。这正是该法在微生物学分类实验室得以广泛应用的原因。

4. 中线法[7]。该法聚合策略见图2左，中

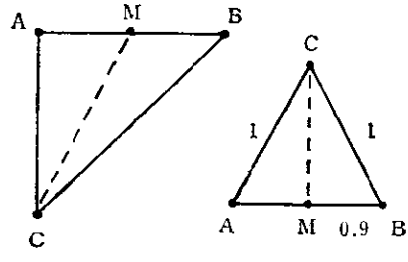

图2 中线法示意图

线法取并组 CU(A + B) 的距离 D_{CA+B} = 线段 CM 的长度。中线法虽具有空间保持的性质，却同时具有致命的弱点——聚类过程中的非单调性。所谓非单调性是指聚类过程中本级聚合水平(距离)小于上一级聚合水平，造成树状谱图形发生逆转如图2右所示。在等腰三角形 △ABC 中

$$CM = D_{CA+B} = \sqrt{1/2(D_{CA}^2 + D_{CB}^2) - 1/4D_{AB}^2}$$
$$= \sqrt{1/2 * (1+1) - 1/4 * 0.81}$$
$$= 0.893 < D_{AB}$$

即第二次C与A＋B的聚合距离反而比第一次A与B的聚合距离小，所以具非单调性。通常在公式中只要满足 $a + b + c \geqslant 1$ 则聚合过程是单调的。由于中线法的这一缺点，至今使用并不广泛。

5. 形心法[7]。聚合策略是定义两样品组间的距离为这两组形心间的距离。形心法具有空间保持的性质，但也具有非单调性，此法除用来对一些特殊数据处理外，用得也不广泛。

6. 平方和增量法[7]。聚合策略是设 Q_A 和 Q_B 分别为样品组A和样品组B的离差平方和，则

$$D_{AB} = Q_{AB} - Q_A - Q_B$$

$$D_{CA+B} = \frac{n_C + n_A}{n_C + n_{A+B}} D_{CA} + \frac{n_C + n_B}{n_C + n_{A+B}} D_{CB} + \frac{-n_C}{n_C + n_{A+B}} D_{AB}$$

它类似单链锁法具有空间压缩和单调聚合性质。

7. 可变法[7]。此法没有固定的聚合策略，它只要求公式中 $a = b, c < 1, a + b + c = 1, d = 0$，故可以写成：

$$D_{CA+B} = \frac{1-\beta}{2} D_{CA} + \frac{1-\beta}{2} D_{CB} + \beta \cdot D_{AB} \quad (\beta < 1)$$

式中 β 为聚合强度系数，当 $\beta > 0$ 时，聚合结果是空间压缩的；$\beta < 0$ 时，结果是空间扩张的；当 $\beta = 0$ 时，结果是空间保持的，此时可变法就成为通常所谓的加权的平均链锁群法。

聚类分析在微生物界作为一种新的分类方法，由于其短暂的发展历史，使得这种分析方法目前还存在着某些一时难以克服的缺点。尽管数值分类法目前还不可能完全取代已有很长发展历史的传统分类方式，但是可以确信：以电子计算机为工具，以数学手段为依据的现代分类方法，如果能够和积累起来的传统分类学知识相结合，无疑将成为分类工作者强有力的辅助工具。

参 考 文 献

[1] Sneath, p. H. A. and R. R. Sokal: Numerical Taxonomy: The Principles and Practice of Numerical Classification, W. H. Freeman and Company, 1973.

[2] 徐浩等：微生物学报，**15**(1)：31—36，1975。

[3] 卢运玉：微生物学报，**20**(1)：10—15，1980。

[4] 王大耜等：微生物学报，**21**(4)：385—401，1981。

[5] 金天如等：微生物学报，**25**(1)：13—18，1985。

[6] 方开泰等：聚类分析，地质出版社，1982。

[7] 阳含熙等：植物生态学的数量分析方法，科学出版社，1981。

[8] 蒋大宗等：数值诊断的统计方法，陕西科学技术出版社，1981。