聚类分析在生态学中的应用

颜京松　　魏善武
（中国科学院南京地理研究所）　（西北高原生物研究所）

在研究某种生物种群动态或生产量时，要调查不同时期中一定区域内某种生物的数量，一般是通过抽样来进行。在非均匀分布情况下，为了既提高抽样的有效性，又合理地减少抽样数，降低工作量，可用分区（层）(stratified) 抽样法。再据各分区在所调查大面积区域内的比率，用加权均数求得该大区域内的该种生物数量。而对一大区的分区（层），可据某次随机抽样的资料用聚类分析法求出。它不仅可解决变量（或指标）的分类问题（如哪些生物在哪些小生境中的分布），且可解决多变量的总体的分类问题（如哪些采样站可归于一区）。

聚类分析不仅在种群生态学中，解决上述的分区问题，在群落生态学中也很有用。例如：研究群落的演替，比较不同空间或时间内区系的异同；用生物群落结构评价河流内不同河段污染情况等，可比较不同河段中群落一些属性的异同而予分类。在所研究的实体（如群落或小生境为数少时，可据直观或用较简单的指数，如 Jaccard 指数，Kulezynski 指数，相似商 (O. S)[1] 等来比较群落的异同。但若实体数及属性很多，用直观判断或上述诸指数难以比较时，则应考虑用其它方法。近十多年来，很多生态学家对应用多样性指数发生兴趣。但多样性所反映群落的信息仅是组成群落的生物种类数及个体数两种信息，它可使不同数量的分类单位所成的生物群和优势集中具有相同的多样性。另外，没有任一个共同种类的一些生物群也可具有相同的多样性。在此情况下，聚类分析和其它多元分析不仅可简化繁多复杂的生态学原始数据，且在分类中，系在实体中属性的同一性基础上进行比较，显示实体间的关系，进而将实体按属性分类。它可与多样性指数互为补充。

聚类分析可说明怎样进行分类，但一般不能说明该不该分类。聚类的合理与否，有赖于用具体的生态学知识与资料以及所欲研究的生态学问题予以分析和解决，所以聚类分析在生态研究中仅是方法之一，不能用数学代替生态学研究。

一、聚类分析的程序

聚类分析的主要过程，是对数据进行适当处理，建立数据矩阵或二向表（联列表），计算相似性测度（距离）再列相似矩阵，按一定规则聚类。兹分述如下：

（一）数据　在聚类分析中，选择哪些属性（指标）及每一属性的数据形式，应据具体情况及所需解决的生态学问题而定。例如，选取生物个体数或生物量作属性，将有很大差异。

表 1　m×n 阶数据矩阵

属　性	实体 1	2	…	n
1	x_{11}	x_{12}	…	x_{1n}
2	x_{21}	x_{22}	…	x_{2n}
⋮	⋮	⋮		⋮
m	x_{m1}	x_{m2}	…	x_{mn}

生态学中的数据一般都排成数据矩阵的形式，如表 1 所示。矩阵的 n 个列表示 n 个实体，m 个行代表属性。矩阵的元素 X_{ij} 表示第 j 个实体的第 i 个属性值。当数据是离散的情形，

1) Jaccard 指数 $=\frac{j}{a+b-j}$；Kulezynski 指数 $=\frac{j}{2}\times\left(\frac{1}{a}+\frac{1}{b}\right)$ Q. S. $=\frac{2j}{a+b}$，式中 a, b 为 A、B 生境内的种数，j 为 A、B 共有的种数。

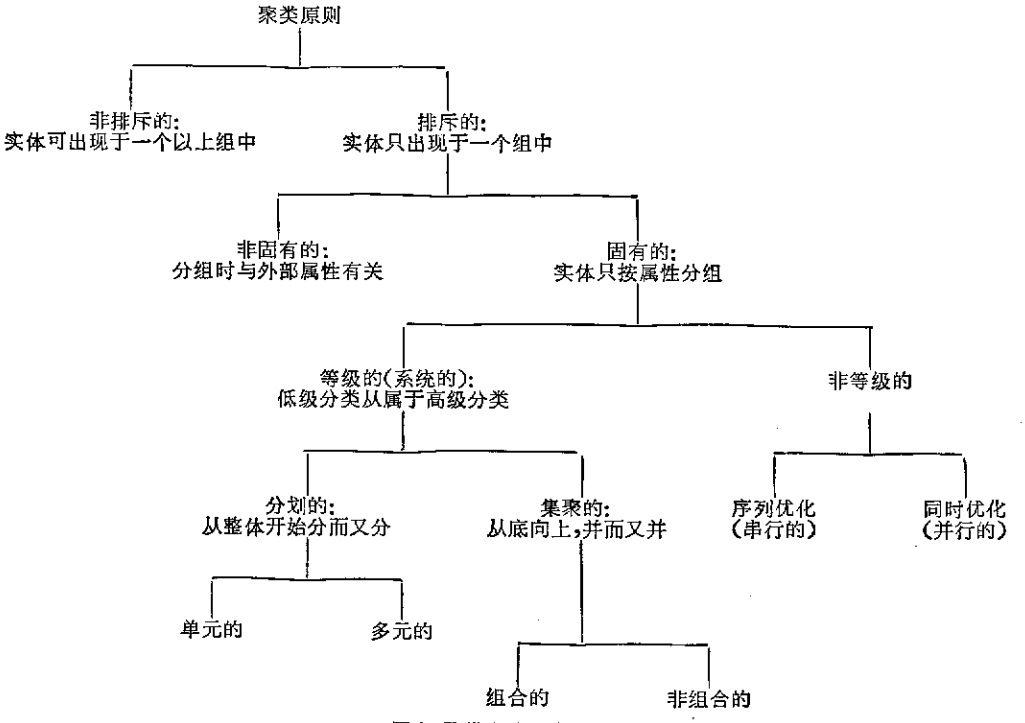

图 1 聚类方法分类

也可得到类似矩阵(联列表)。

在列数据矩阵时,实体和属性的选择,取决所要解决的问题。例如,以采样站或生物群,或群落为实体,以各相应生物种类的指标或影响生物的主要环境因素的指标作为属性列成矩阵,此为正相分析。它可解决这些采样站(或生物群或群落)的相似问题。反之,以生物种类为实体,而以采样站等为属性列成矩阵,此为反相分析。它可解决两种生物相遇的程度如何等。

(二) 相似性测度 为了分类,仅有数据是不够的,还需考虑实体间的关系。一般利用相似性测度表征实体间的相似或类似程度。主要的测度有: (1)相似系数;(2) 相关系数;(3) 欧氏距离;(4)信息含量测度;(5)概率测度。本文不在此详述。本文所用的测度是欧氏距离。

(三) 聚类方法 已有人提出了聚类方法的分类,于此不另赘言,今仅据 Boesch (1977) 的意见,将 Williams 的一种分类法在图 1 中列出。

应用最广的聚类方法是等级、集聚的组合法。利用组合法时,组与组,组与实体间的相似性可据数据矩阵计算,一旦算得相似(或距离)矩阵,就不必保留原始数据了。这个长处是非组合法所不具备的。表 2 中列出了八种组合法的递推公式。从近邻法的公式看出,组间的距离等于两组中最近成员的距离。对于其它方法亦可做类似的理解。有的文章对一些聚类方法做了较具体介绍 (方开泰, 1978)。可以用图 2 比较形象地说明聚类过程。对于近邻法, D_{hk} 等于 D_{hj}; 对于远邻法, D_{hk} 等于 D_{hj} 对于中线法, D_{hk} 等于三角形的中线,等等。

上面所述八个公式 (表 2), 可以用一个距

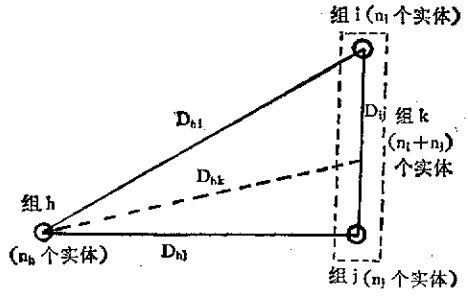

图 2 组 h 到组 i 和组 j 聚成的新组 k 的距离

表 2　八种组合聚类法一览表

近邻法	$D_{hk}=\min(D_{hi},\ D_{hj})$
远邻法	$D_{hk}=\max(D_{hi},\ D_{hj})$
组平均法	$D_{hk}^2=(n_i/n_k)D_{hi}^2+(n_j/n_k)D_{hl}^2$
简单平均法	$D_{hk}^2=\frac{1}{2}(D_{hi}^2+D_{hj}^2)$
形心法	$D_{hk}^2=(n_i/n_k)D_{hi}^2+(n_j/n_k)D_{hj}^2-[(n_in_j)/n_k^2]D_{ij}^2$
中线法	$D_{hk}^2=\frac{1}{2}D_{hi}^2+\frac{1}{2}D_{hj}^2-\frac{1}{4}D_{ij}^2$
可变法	$D_{hk}^2=\frac{1}{2}(D_{hi}^2+D_{hj}^2)+D_{ij}^2$
平方增量和法	$D_{hk}^2=[(n_k+n_i)/(n_h+n_k)]D_{hi}^2+[(n_k+n_j)/(n_h+n_k)]D_{hj}^2-[n_k/(n_h+n_k)]D_{ij}^2$

式中 n_i，n_j，n_k 和 n_h 分别为组 i，j，k 和 h 的实体数，D_{ij}^2 为组 i 和 j 距离平方。

表 3　组合聚类的参数值（引自 Boesch 1977）

方　法	α_i	α_j	β	γ	空间畸变
近邻法	1/2	1/2	0	$-\frac{1}{2}$	收　缩
远邻法	1/2	1/2	0	$\frac{1}{2}$	膨　胀
组平均	n_i/n_k	n_j/n_k	0	0	守　恒
简单平均	1/2	1/2	0	0	守　恒
形心法	n_i/n_k	n_j/n_k		0	守　恒
中线法	1/2	1/2	$-\frac{1}{4}$	0	守　恒
可变法	$(1-\beta)/2$	$(1-\beta)/2$		0	$\beta\cong 0$，守恒 $\beta>0$，收缩 $\beta<0$，膨胀
平均增量和	$\frac{n_h+n_i}{n_h n_k}$	$\frac{n_h+n_j}{n_h+n_j}$	$\frac{-n_h}{n_h+{}_n n_k}$	0	膨　胀

表 4　官厅水库中克拉伯水丝蚓 18 个采样站十次调查的数量（尾米$^{-2}$）

次＼站	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	3	67	96	82	2	3	6		21	22		2	3	9	1		2	1
2	8	18	70	1	5					4								
3	1	33		50	13					3		3	3	2	2	2		
4		152	143	3	15				53		4				6			
5	2	12	1	5										18				1
6		1	18		44	3					4	1		14	3	68	68	
7	1	12	28	67	51	29	15	65		4	4	6	1	3	3		3	33
8	1	76	94		2	36		1		8	2			61	7			
9	2	21	6	181	23	1	25	13	40	19	1	5	6	77	4		4	59
10	3	115	113	37	89	1			1	9		4				8		1

注：无数字处均是零。

离平方公式统一起来：

$$D_{hk}^2=\alpha_iD_{hi}^2+\alpha_jD_{hj}^2+\beta D_{ij}^2+\gamma|D_{hi}^2-D_{hj}^2|$$

如果采用相似性测度，则有如下的线性形式：

$$D_{hk}=\alpha_iD_{hi}+\alpha_jD_{hj}+\beta D_{ij}+\gamma|D_{hi}-D_{hj}|$$

对于上面两个公式相应于表 2 的八个递推公式的有关参数值可以从表 3 查出。

表 5　采样站间的欧氏距离矩阵

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	216.9																
3	234.0	77.0															
4	215.1	252.1	272.3														
5	112.5	182.3	200.4	194.5													
6	45.7	209.9	223.4	212.6	109.4												
7	28.5	217.7	236.4	191.6	107.7	45.7											
8	68.0	225.6	240.3	197.1	102.8	51.8	51.8										
9	68.5	181.0	205.6	185.5	120.7	82.5	59.0	90.6									
10	27.7	205.7	222.6	192.2	106.4	46.8	24.1	66.3	58.4								
11	11.4	217.1	235.5	216.8	112.0	42.5	27.7	62.4	88.4	31.0							
12	10.5	217.0	236.0	210.9	108.5	43.2	22.9	59.8	66.7	26.6	8.8						
13	9.9	218.9	238.0	211.7	114.0	46.1	24.0	64.5	65.6	26.5	9.8	6.6					
14	99.0	211.2	229.7	167.4	135.2	67.0	84.3	110.2	92.6	83.4	98.9	97.3	96.6				
15	13.1	213.2	232.1	213.9	111.9	39.6	26.6	63.3	63.1	28.5	6.7	10.8	10.3	93.8			
16	68.8	226.5	240.3	227.6	103.2	80.2	74.7	95.3	97.6	74.6	64.8	64.6	68.8	114.2	66.3		
17	68.6	229.7	242.7	224.4	107.6	78.8	72.3	92.5	95.4	73.6	64.3	67.3	68.1	102.6	65.7	9.9	
18	65.9	222.8	242.7	162.6	108.3	68.5	38.8	56.1	68.3	55.1	65.1	60.5	62.0	74.2	63.4	96.2	92.5

这种统一的公式，对于编制电子计算机程序非常有用。因为当变量数目很大、方法复杂时，庞大的计算量和繁复的运算必须依靠电子计算机才能解决问题。

（四）聚类步骤　今以官厅水库中克拉伯水丝蚓的分区为例说明聚类步骤如下：

（1）以 18 个采样点作为实体，以各采样点的十次调查的该动物密度（个体数·米$^{-2}$）为属性，列成原始数据矩阵（表 4）。

（2）计算实体间相似性或距离：可任择前述诸计算方法中的一种。在本例中，按欧氏距离公式算出了各实体（18 个采样站）之间的距离，列出的矩阵见表 5。由欧氏距离易知 $D_{ij} = D_{ji}$，所以得到的距离矩阵是一对称矩阵，因之表 5 给出的是一对角矩阵。

（3）定义并计算组间的相似性或距离。具体的递推公式列于表 2。

（4）逐步聚类。开始时每个实体自成一组，此时组间的距离就是实体间的距离。然后按第(3)条规定将相应的组合并为新组，再重新计算新组与每组的距离，再合并之，再计算新距离，……，一直到所有实体都被归类为止。

在本例中，从表 5 可见，组 12 和组 13 间的

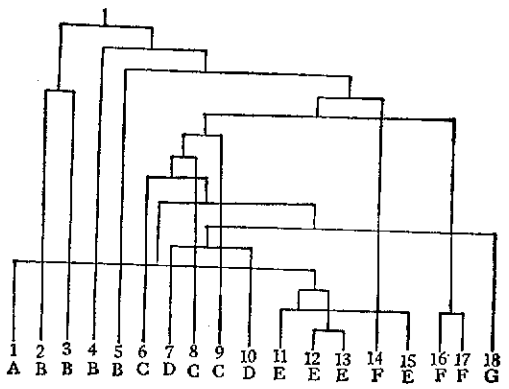

图 3　聚类图

距离（其实是实体 12 和 13 距离）$d_{12,13} = 6.6$ 最小，故将它们合并为新组 19，同理，由于

$$d_{11,15} = 6.7$$

为次最小，可将组 11 和 15 合并为新组 20；再计组 19，组 20 与其它组间的距离，发现

$$d_{19,20} = 8.8$$

最小，故将组 19 和组 20 合并为新组 21，再计算组 21 与其它组的距离，再合并，最后全部实体归为一个类。这个聚类过程可用一个树状的聚类图表示。

最后，从图 3 看出，18 个采样站可分为

A-G 七个类。就是说可将官厅水库内克拉伯水丝蚓的分布分成七个分区，其中A和G是根据水库的具体情况而定的

二、讨　论

聚类分析的分类结果是否合理，常取决于所选的属性指标及所用数据形式(及处理)是否符合生态学要求和标准。然后是选择相似性测度。假若研究者对实体间相似性概念主要是建立在优势种的丰度相似性基础上，则可在正相分析中应用定量相似测度。为此目的，Bray-Curtis 系数是较好的。但是，欧氏距离、相关和信息含量测度等加权测度可能更好。如果实体间相似概念建立于实体间所有生物种类的等加权 (equal weighting) 的基础上，并且希望说明实体间定量和定性差异时，则可选择应用 Canberra 系数，或在指标标准化以后，用前述的各种测度中的某一种。种间相似性的生态学标准与实体间相似性的标准不同，这时可以应用反相分析。

选定生态相似性的测度后，就要根据该相似性选取一种聚类方法，进行实体间的归类。前已述及，近来常用的是集聚的等级组合法，因其计算简单。其次，组平均法、可变法和平方增量和法也比较常用。

组平均法具有空间守恒性质，聚类结果与实际相似关系偏差很少。它适用于实体较少且需不使空间畸变的场合。

可变法的好处是可以不断改变聚类强度。因此当很多实体正在分类，而它们的相似性很复杂时，可变法比组平均法更有利。

平方和增量法也是一种较好的聚类法，特别当用欧氏距离作相似测度时，它比可变法更优越。

除前面介绍的组合聚类法外，还有逐步聚类法，有序聚类法等。聚类结果本身并未提供解释，必须依据群落构成种的一些生物学特性予以解释。

参 考 文 献

方开泰　1978　聚类分析。数学的认识与实践1：66—80。

Boesch D. F., 1977. Application of numerical classification in ecological investigation of water pollutions. Env. Res. Lab. office of Res. development U. S. Env. Protection Agency. 1—131.

Jaccard P., 1912. *New Phytol.* 11: 37—50.

Kulezynsky S., 1928. *Bull. Int. Acad. Pol. Sci. Lett., Suppl.* 2: 57—203.

Pielou E. C.(卢泽愚译) 1969 数学生态学引论. 科学出版社.

Serenson, T., 1948 *Biol. Skr.* (*K. Danske vidensk. Selsk.* N. S.) 5: 1—34.

Sneath P. H. A. & C. S. Greene, R. R. Sokal, 1973. Numerical taxonomy. The principles and practice of numerical classification. Freeman, San Franeisco. 573.

Williams W. T., 1971. *Ann. Rew. Ecol. Syst.* 2: 303—326.

Williams W. T. & W. Stephenson, 1973. *Jour. Kor. Biol. Ecol.* 11: 207—227.