热带亚热带植物学报 1996, 4(4): 18－23
Journal of Tropical and Subtropical Botany

以平均距离为基础的两种多元等级聚合新策略

张金屯
(山西大学生命科学系，太原 030006)

摘要 本文基于扩展的 Lance 和 Williams 的多元聚合模型之上，提出两种新的聚合策略：新组内平均距离法和加权平均距离法。这两个方法既考虑了组间距离，同时也考虑组内的同质性，聚类结果更好。本文用山西中条山荆条灌丛的数据作为这两种方法的应用实例，结果较好地描述了群落类型间的关系，生态意义明确。这两个方法分辨力较强，适合群落低级单位的分类，宜于群落生态关系的研究。

关键词 数量分类；聚合分析；植被分析；新组内平均距离法；加权平均距离法

TWO POLYTHETIC-AGGLOMERATIVE CLUSTERING STRATEGIES

Zhang Jintun
(Dept. of Biological Science, Shanxi University, Taiyuan 030006)

Abstract Based on extended Lance and Williams' model, two polythetic agglomerative clustering strategies, i.e. the Average distance of new cluster and the Weighted average distance, are introduced in this paper. As an example, vegetation data of shrub, *Vitex negundo* var. *heterophylla*, collected from Zhongtiao mountains in Shanxi Province were analysed with these two new methods. The Results illustrate that the two methods are effective and successful in describing the relationships between vegetation clusters. They have some advantages compared with fuzzy equivalence clustering and group averaging, and are partically suitable for vegetation classfication at association level.

Key words Numerical classification; Cluster analysis; Vegetation analysis; Average distance of new cluster; Weighted average distance

多元等级聚合方法是在数量分类中占有重要位置的一类方法。在生物学、生态学、医学、地学等学科中得到了广泛的应用[1－3]。该类方法首先要计算距离矩阵 D，然后根据一定的原则和程序，将样方一一合并，直到合成为一组。由于聚合策略的不同而导致产生不同的方法。聚合策略指的是如何定义一个样方或一个样方组与新形成样方组间的距离，比如最近邻体法将一个样方和一个样方组间的距离定义为该样方与这组中最近一个样方间的距离，两个样方组间的距离定义为两个组中最近的两个样方间距离[4－5]，这类方法最初是基于同一个模型之上[6]。

$$D_{C,\ A+B}=\alpha_A D_{CA}+\alpha_B D_{CB}+\beta D_{AB}+\gamma\,|D_{CA}-D_{CB}| \qquad (1)$$

1995－12－25 收稿；1996－10－03 修回

这里 $D_{C,A+B}$ 表示样方(或样方组)C 与样方组 A+B 之间的距离系数，D_{CA}，D_{CB}，D_{AB} 分别表示样方 C 和 A、样方 C 和 B、样方 A 和 B 之间的距离系数。α_A，α_B，β 和 γ 都是常数，它们取不同的值就代表不同的聚合策略，而形成不同的方法。

以模型(1)为基础的方法都是以寻求组间距离最小化，而没有考虑组内关系[7]。近年来一些学者扩展了模型(1)，使其既要考虑组内的距离，也要参考组内样方间的同质性，以求得到更合理的聚类结果。模型(1)可以扩展为[8]：

$$D_{C,A+B}=\alpha_A D_{CA}+\alpha_B D_{CB}+\beta D_{AB}+\gamma|D_{CA}-D_{CB}|+\lambda_C D_C+\lambda_A D_A+\lambda_B D_B \tag{2}$$

这里 D_C，D_A 和 D_B 分别表示样方组 C、A 和 B 组内各样方对之间的平均距离或平方和或方差。如果 C、A 和 B 只有一个样方，则 D_C，D_A 和 D_B 就等于 0，λ_C，λ_A 和 λ_B 是常数。这样 α_A，α_B，β，γ，λ_C，λ_A，λ_B 的取值不同，将代表不同的聚合策略。

这一模型更为通用，可以证明以前的方法都符合该模型，只是它们的 λ_C，λ_A，λ_B 均等于 0。

基于模型(2)，一些学者引入新组内平方和法及新组内方差法等聚合新方法，并在一定范围内得到了应用。这里我们引入两个以平均距离为基础的多元聚合新策略而产生两种新方法，它们是新组内平均距离法(Average distance of new cluster)和加权平均距离法(Weighted average distance)。我们以山西省中条山地区的荆条灌丛数据作为这两种新方法的分析例子。

1 方法

1.1 新组内平均距离法

该方法把样方和一个样组间的距离定义为组内平均距离，要使得合并后新组内的平均距离最小，如果两个样方或两个样方组 A 和 B 合并为一个新组 A+B，使得：

$$D_{A+B}=\min\{DIS(A+B)\} \tag{3}$$

则它们应最先合并。所以，该方法也可以叫做新组内平均联结法。

要计算组内平均距离，首先要计算组内距离和，即：

$$D_{AB}=\frac{1}{b_{A+B}}\sum_{j,k\in A+B} d_{jk} \tag{4}$$

$$D_A=\frac{1}{b_A}\sum_{j,k\in A} d_{jk} \tag{5}$$

这里 D_A 和 D_{AB} 分别代表样方组 A 和样方组 A+B 的组内距离和。

$b_A=\binom{n_A}{2}$ 为样方组 A 内的样方对的数目；$b_{AB}=\binom{n_A+n_B}{2}$，$n_A$ 和 n_B 分别是样方组 A 和 B 所含的样方数。d_{jk} 表示样方 j 和 k 间的欧氏距离。同理得：

$$D_{CA}=\frac{1}{b_{CA}}\sum_{j,k\in C+A} d_{jk} \qquad D_C=\frac{1}{b_C}\sum_{j,k\in C} d_{jk} \qquad D_{CB}=\frac{1}{b_{CB}}\sum_{j,k\in C+B} d_{jk} \qquad D_B=\frac{1}{b_B}\sum_{j,k\in B} d_{jk}$$

不难看出，当样方组 C 和样方组 A+B 合并为新组 C+A+B 时，新组内的距离和为：

$$\sum_{j,k\in C+B+A} d_{jk}=b_{CA}D_{CA}+b_{CB}D_{CB}+b_{AB}D_{AB}-b_C D_C-b_A D_A-b_B D_B \tag{6}$$

用新组内距离和除以 b. 就得到新组 C+A+B 的组内平均距离，即：

$$D_{C,A+B}=\frac{b_{CA}}{b.}D_{CA}+\frac{b_{CB}}{b.}D_{CB}+\frac{b_{AB}}{b.}D_{AB}-\frac{b_C}{b.}D_C-\frac{b_A}{b.}D_A-\frac{b_B}{b.}D_B \quad (7)$$

这里 $b.=\binom{n_t}{2}$，$n_t=n_C+n_A+n_B$ 为新组 C+B+A 的样方总数。(7)式是模型(2)的特殊形式。

1.2 加权平均距离法

加权平均距离法要使得新组内平均距离增加最小，如果两个样方合并为一组 A+B，只要：

$$D_{A+B}=\min\{DIS(A+B)-1/2\,DIS(A)-1/2\,DIS(B)\} \quad (8)$$

则样方组 A 和 B 应最先合并，加权平均距离法忽略了样方组的大小，因此，样方组 A+B 的平均距离为 $DIS(A+B)=D_{AB}+1/2\,D_A+1/2\,D_B$ (9)

同理：

$$DIS(C+A)=D_{CA}+1/2\,D_C+1/2\,D_A \quad (10)$$

$$DIS(C+B)=D_{CB}+1/2\,D_C+1/2\,D_B \quad (11)$$

样方组 C+A+B 的平均距离为 $DIS(C+A+B)=1/b.\,[b_{CA}DIS(C+A)+b_{CB}DIS(C+B)+b_{AB}DIS(A+B)-b_CD_C-b_AD_A-b_BD_B]$ (12)

样方组 C 和 A+B 合并前的平均距离为 $DIS(C,A+B)=1/2D_C+1/2DIS(A+B)$ (13)

将(9)－(11)式代入(12)和(13)式后，并由(12)式减去(13)式，就得到样方组 C 和 A+B 合并后的平均距离增加量：

$$D_{C,A+B}=\frac{b_{CA}}{b.}D_{CA}+\frac{b_{CB}}{b.}D_{CB}+\left(\frac{b_{AB}}{b.}-\frac{1}{2}\right)D_{AB}-\frac{b_C+n_An_B}{2b.}D_C-\frac{b.-2b_A-2n_Cn_B}{4b.}D_A-\frac{b.-2b_B-2n_Cn_A}{4b.}D_B \quad (14)$$

(14)式也是模型(2)的特殊形式。

以上两个方法的模型系数列入表 1 中。

表 1 两种聚合方法的聚合策略（系数）

Table 1 Clustering strategies (coefficients) of the two agglomerative methods

方法名称 Methods	模型系数 Model coefficients						
	α_A	α_B	β	γ	λ_C	λ_A	λ_B
新组内平均距离法 Averaging distance of new cluster	$\frac{b_{CA}}{b.}$	$\frac{b_{CB}}{b.}$	$\frac{b_{AB}}{b.}$	0	$-\frac{b_C}{b}$	$-\frac{b_A}{b.}$	$-\frac{b_B}{b}$
加权平均距离法 Weighted average distance	$\frac{b_{CA}}{b.}$	$\frac{b_{CB}}{b.}$	$\frac{b_{AB}}{b.}-1/2$	0	$-\frac{b_C+n_An_B}{2b.}$	$-\frac{b.-2b_A-2n_Cn_B}{4b.}$	$-\frac{b.-2b_B-2n_Cn_A}{4b.}$

2 植被数据

本文数据用山西中条山的荆条(*Vitex negundo* var. *heterophylla*)灌丛数据，研究地位于东经

115°15′－112°00′，北纬 35°00′－35°25′，海拔 500－1000m。年均温 13.3℃，7 月份均温 26.1℃，元月份均温 -0.8 ℃，年降水量 667.6mm。群落种类组成丰富。灌木层荆条占绝对优势，还有黄刺梅(*Rosa xanthina*)、小叶鼠李(*Rhamnus parvifolia*)、红叶(*Cotinus coggygria* var. *cinerea*)等。下层以蒿类(*Artemisia* spp.)，白羊草(*Bothriochloa ischaemum*)，披针苔(*Carex lanceolata*)，翻白草(*Potentilla discolor*)等为主。关于群落的生态环境和组成的详细描述见参考文献[11]。原始数据经简缩后由 30 个植物种和 41 个样方构成矩阵。该数据曾用模糊等价聚类(Fuzzy equivalence clustering)和组平均法(Group averaging)进行过分类[11]，结果 41 个样方被分为 11 个组，代表 11 个群丛，其中包括 3 个过渡性明显的群丛[11]。

3　结果分析和讨论

以上数据用本文的两种新方法分析、计算在山西大学生命科学系生态室 AST386 微机上完成。两种方法的聚合树状图见图 1。

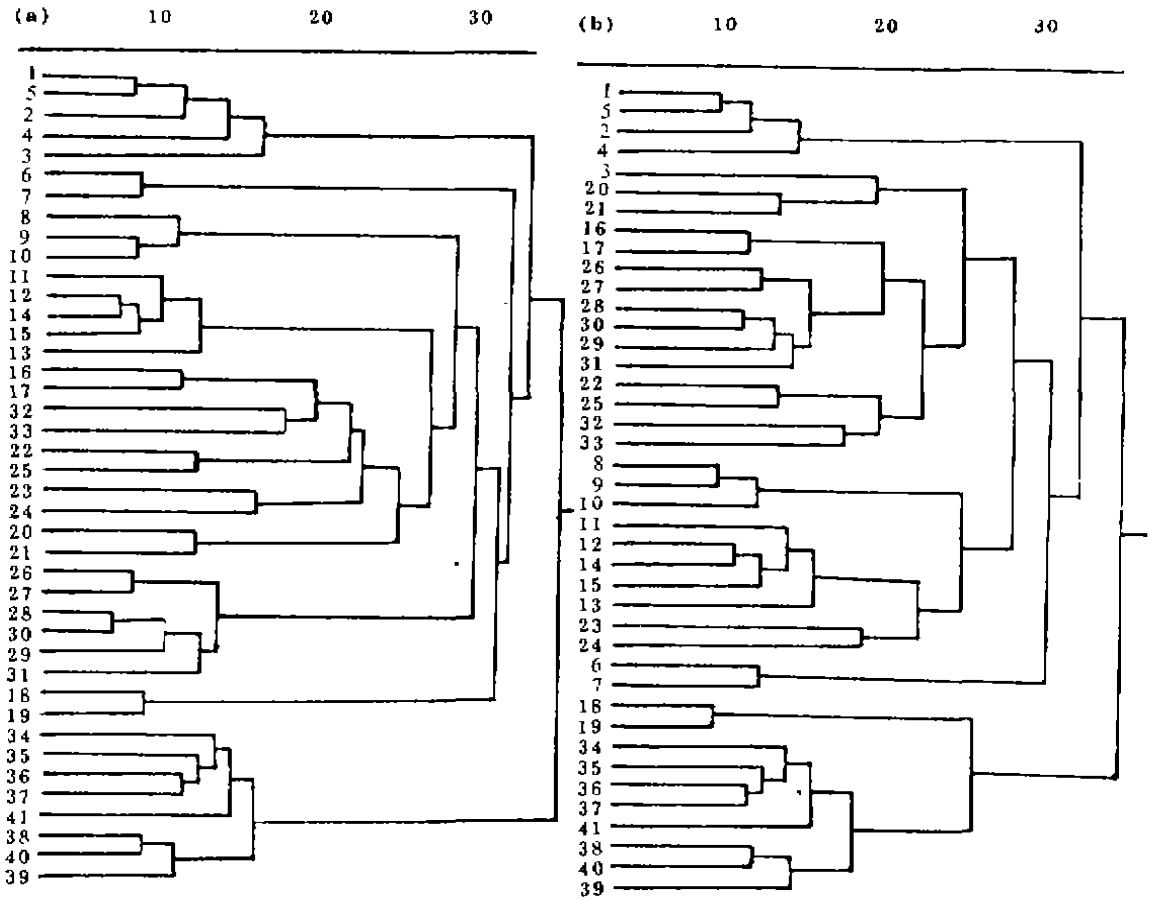

图 1　两种方法的聚合结果图　(a)新组内平均距离法　(b)加权平均距离法

Fig. 1　Dendrograms of (a) Average distance of new cluster and (b) Weighted average distance

新组内平均距离法分类结果(图1a)与模糊等价聚类和组平均聚类结果基本一致，将41个样方分为12个群丛，即{1－5}，{6－7}，{8－10}，{11－15}，{16－17}，{32－33}，{22，25}，{23－24}，{20－21}，{26－31}，{18－19}，{34－41}。因此，可以说新组内平均距离法较好地反映了群落类型之间的关系，分类结果具有明确的生态意义。不同的是它将样方组{22－25}分为两组:{23－24}，{22，25}。从DCA排序图(图2)上可以看出，样方组{22－25}分布较为松散，说明组内样方间的相似性或同质性较低．即组内差异较大。所以，该组分为两组也不为怪，说明新组内平均距离法分辨力更强。生态分析表明，水分和海拔高度是制约各群丛分异和分布的主要因子。本文研究地位于中条山主峰以西，来自东面的湿润气流受主峰阻隔，雨量少于主峰以西地区，因此群丛多属中旱生类型，如群丛{6－7}，{8－10}，{16－17}等为典型旱生类型，而群丛{1－5}，{11－15}，{32－33}等为中旱生类型。在海拔较高的地方，水分条件优越，分布着覆盖度大，生长繁茂的群丛，草本层以喜湿的披针苔草为主，如群丛{22，25}，{23－24}，{34－41}等。在海拔较低的地方，多分布着中旱生类型，如群丛{6－7}，{16－17}等。在同一地方，海拔影响更明显。详细的生态解释见参考文献[11]。

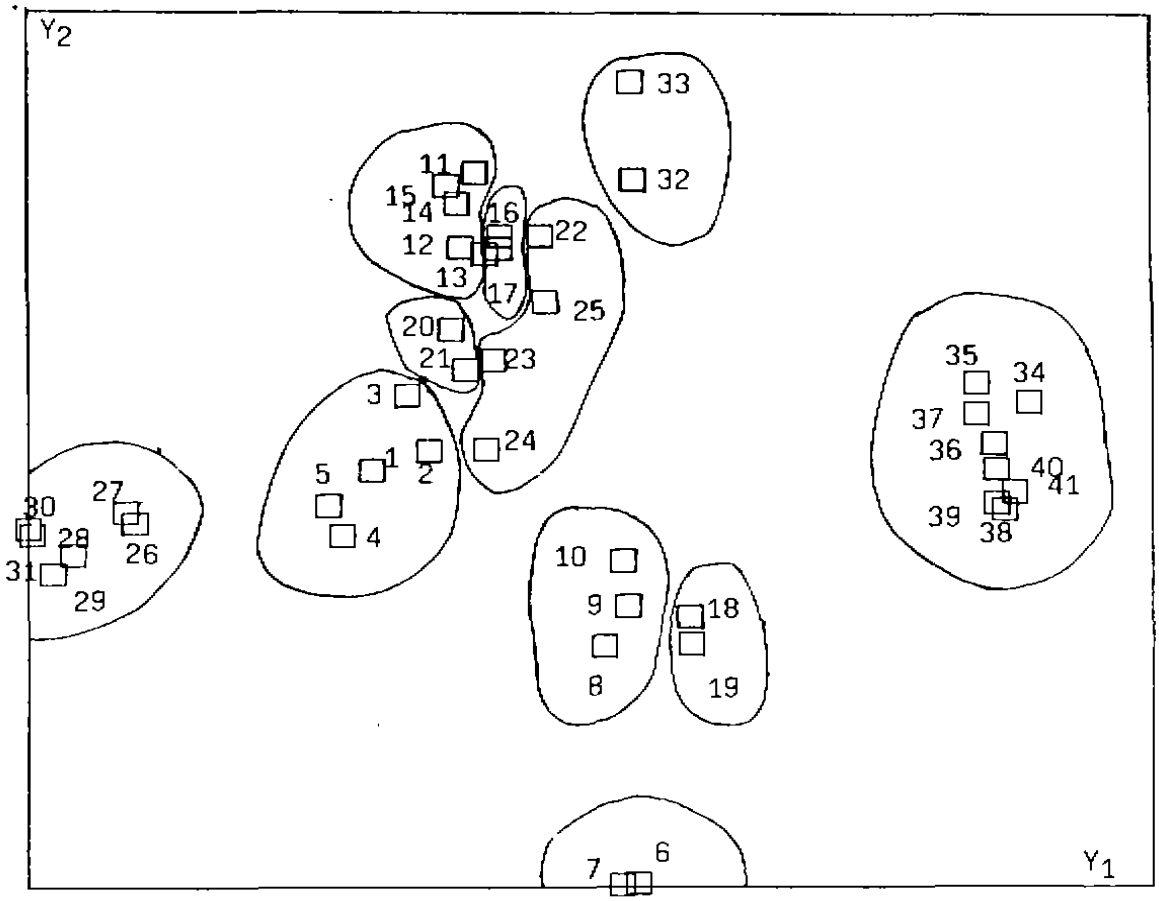

图2 41个样方的Detrended Correspondence Analysis 排序图

Fig 2 Detrended Correspondence Analysis ordination plot of 41 quadrats

加权平均距离法分类结果(图1b)也是将41个样方分为12组，{1－2，4－5}，{3，20－21}，{16－17}，{26－31}，{22，25}，{32－33}，{8－10}，{11－15}，{23－24}，{6－7}，{18－19}，{34－41}。该方法与模糊聚类，组平均法结果基本一致，不同的也是将样方组{22－25}分为了两组，这一点同新组内平均距离法结果吻合。另外该方法将样方3与样方组{20－21}合为一组，这是因为样方3具有过渡性，兼有样方组{1－2，4－5}和样方组{20－21}的特征，这一点可以从41个样方的除趋势对应分析(DCA)排序图上看出来(图2)。很明显样方3介于这两个样方组的中间。

从本文分析来看，新组内平均距离法和加权平均距离法都是有效的植被数量分类方法。理论上讲，这两个方法基于模型(2)之上，比基于模型(1)的多元聚合方法更具优越性，因为，这两个方法既考虑了组间的距离关系，同时也考虑了组内的同质性，这样聚类的结果，使得同组内更相似，而组间距离更明显，有利于最后的解释。本文所用的数据量不算大，有些优点难于表现出来。这两个方法的结果与模糊等价聚类的结果基本一致，与组平均法也基本吻合，但分辨力较强，在低级分类中具有优越性，在群落的详细研究中较为适宜。新组内平均距离法和加权平均距离法具备多元等级聚合分类的基本特征，它们是空间保持，也是单调的，在多元聚类方法中是较好的方法。模型(2)比模型(1)更具有代表性，它包含了模型(1)，并且扩展了聚合途径。很可能基于这一模型之上，会有更多的新方法出现，这方面需要进一步研究。

这两个新的聚合方法可以使用各种类型的数据。在生态学中，二元数据，各种数量数据或综合指标均可使用。在其它学科中也是如此。

参考文献

1　Greig-Smith P. Quantitative Plant Ecology. 3rd Edition. London. Blackwell Scientific Publications. 1983

2　Orloci L. Multivariate Analysis in Vegetation Research. 2rd ed. Junk, The Hague, 1978

3　阳含熙，卢泽愚. 植物生态学的数量分类方法. 北京：科学出版社. 1981

4　张金屯. 植被数量生态学方法. 北京：中国科学技术出版社. 1995

5　张金屯. 植被数量分析方法的发展. 当代生态学博论. 北京：中国科学技术出版社. 1992　249－265

6　Lance G N. Williams W T. A general theory of classificatory sorting strategies I. Hierarchical systems. Comput J. 1967, 9:373－380

7　Podani J. New combinatorial clustering methods. Vegetatio. 1989. 81:61－77

8　Podani J. Multivariate Analysis in Ecology and Systematics. SPB Academic Publishing. The Hague. 1994

9　Sneath P H A. Sokal R R. Numerical Taxonomy. 2nd ed. Freeman. San Francisco. 1973

10　Anderberg M R. Cluster Analysis for Application. Wiley. New York. 1973

11　张金屯. 模糊聚类在荆条灌丛分类中的应用. 植物生态学与地植物学丛刊. 1985. 9(4):306－313