第11卷 第4期
1991年12月

生 态 学 报
ACTA ECOLOGICA SINICA

Vol.11, No.4
Dec., 1991

最小码原理在分级式和非分级式多元聚类分析中的应用

高 琼
(中国科学院植物研究所，北京)

摘 要

植被数量生态研究中常用的多元聚类法(无论是分级式还是非分级式的)的目的是将一组具有多种属性变量(元)的个体按其属性的相似性和分异规律划分到某些具有代表性的类别。而在类别数的确定问题上，由于缺乏理论上的根据(指数量上的理论)，往往不可避免地带有主观性和盲目性。笔者以为聚类的类别数或聚类分析的模型结构应取决于原始数据的结构特征。应用计算理论中的最小码原理(The Minimum Description Length Principle)，笔者对聚类类别数和模型结构进行了数量上的优化选择，并将这一思想实现在一通用软件包FUZPAK中，实例分析表明优化结果较能反映原始数据的特征。

关键词：聚类，优化，模型，最小码原理。

一、引 言

植被数量生态研究的重要目的在于揭示植物群落中植物种的分布与环境生态因子之间的内在联系，从而找出植被分布的规律。研究对象的原始数据通常可表示为一矩阵(n行，m列)$[X]$，其元素x_{ji}表示n个个体(通常是样地或样方)中第j个个体的m个属性(通常是植物种)中第h个属性所取的值(如植物种在样方中的多盖度等)。所用的主要研究方法之一就是聚类分析。所谓聚类，就是将这 n 个个体按其属性的相似性和分异规律分别归于某些代表性的类别，从而实现对原始数据进行压缩、描述、概括和解释[1-3]。

聚类分析所用的具体计算方法很多，大致上可以分为两大类：其一为分级式聚类法[2]，如Cornell Ecological Program中的 TWINSPAN；其二为非分级式聚类法，具有代表性的非分级式聚类法有模糊ISODATA法和中心逐步修改聚类法等[3]。

无论是分级式还是非分级式聚类，人们通常要求分析结果给出最终类别结构或模型。在模糊ISODATA聚类中，表现为分类类别数的给定；在逐步修改中心的聚类中，表现为分异阈值的给定；在TWINSPAN 聚类中，表现为底层类别的最小个体数和分级级数的确定。所有如上问题中有一共同的基本问题，即给定数据矩阵$[X]$所代表的 n 个个体，应将其分为多少组或类别最合适？截至目前为止的聚类方法并未从理论上回答这四个问题，人们在确定最终分类上用的是一种试验—误差修改—再试验的方法，如用χ^2的方法确定TWINSPAN的最终分类[4]，用模糊熵的检验来确定ISODATA分析的类别数[3]等。而中心逐步修改聚类法的

注：本研究为国家自然科学基金资助项目。写作中承蒙张新时，袁嘉祖教授指点，在此谨致谢忱。
本文于1989年9月28日收到。

分异阈值则完全凭主观确定。χ^2检验确定TWINSPAN的最终分类需要反复合并，检验，可谓不胜其繁，且无一定之规可循。模糊熵检验ISODATA分类要求得到较小的模糊熵。而显然最小模糊熵不是选择类别的标准，因为当类别数据增加到同个体数相等时，各个体自成一类，模糊熵有最小值。显然这样的分类结果没有任何意义。

分类问题实质是一个建模问题，最终分类的确定可以看作是模型结构的确定。我们希望模型中尽可能包括原始数据中有规律的部分，而摈弃数据中随机的，无规律的部分。因此，过粗的分类容易损失数据中有规律的部分，而分类太细又易将随机的，无规律的因素反映在模型中。在此之间，必存在一最佳的模型结构，它能最大程度地反映数据中的规律性，而又能最大限度地摈弃数据中的随机、无规律部分。本文的目的就在于应用计算理论中的最小码原理[5-6]来优化聚类问题的模型结构。

二、基 本 原 理

最小码原理的理论基础在于贝叶斯推断理论和信息复杂度理论，现将贝氏概率公式记于下：

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)} \qquad (1)$$

式中$P(H_i|D)$表示给定的样本数据D在一组假设或理论H_j，$j=1,2,\cdots$，中，第i种假设H_i为母体支配机理的概率；$P(D|H_i)$为给定第i个假设H_i为母体支配机理时，样本数据D出现的概率；$P(D)$为数据出现的全概率，而$P(H_i)$为第i种假设出现的概率。贝叶斯推断理论认为，在一组假设中，如存在一H_i使$P(H_i|D)>P(H_j|D)$，$i\neq j$，$i,j=1,2,\cdots$，则H_i可以看作是产生数据D的母体的最佳模拟。

应用贝氏推断的最大困难在于$P(H_i)$的确定，这个量一般无法以客观，实验来确定，从而使推断失去了客观性。

现代信息理论认为，一事件出现的概率是与事件带有的信息量或复杂程度相联着的。而Rissanen[5-6]更进一步将事件的复杂度与需要在计算中描述这一事件的二进制代码联系起来。具体地，对（1）取对数的负值，就得到

$$-\log_2 P(H_i|D) = -\log_2 P(D|H_i) - \log_2 P(H_i) + B_1 \qquad (2)$$

式中B_1为一常数；右边第一项$-\log_2 P(D|H_i)$称作误差码长度(Error Code Length)；$-\log_2 P(H_i)$为模型码长度(Model Code Length)；$-\log_2 P(H_i|D)$为总描述长度或总码长(Total Code Length)。求最大$P(H_i|D)$的问题就转化为求最小的总码长的问题。概括地说，对一组给定的数据D和一系列假设或模型H_i，最小码原理指出，具有最小的描述码长的那一个模型或假设能最大限度地解释数据产生的机理和最大可能地消除随机因素对模型的影响。

为了得到最小码，要求(2)式右边各项都小，第三项为一常数，可以略去不计。这样就需要误差码和模型码都小。但这两项通常是矛盾的。简单的模型具有小的模型码，但会给出较大的误差码。反之，随着模型的复杂化，误差码会减少，但模型码又会增加。解最小码问题实际上是在模型码和误差码之间进行协调，使其和达到最小值。

三、常用中心式聚类最小码模式的建立

设具有m个属性变量的n个个体在C个聚类中心附近作m维正态分布，记个体向量$\{x_i\}=x_{ik}$，$k=1,2,\cdots,m$，$1\leqslant i\leqslant n$，以及聚类中心向量为$\{V_j\}=V_{jk}$，$k=1,2,\cdots,m$，$1\leqslant j\leqslant c$。若各属性变量相互独立，且x_{ik}和V_{jk}只能以有限位二进制码来表示，则以$\{V_j\}$为中心，$\{x_i\}$出现的概率为：

$$P(x_i|H_c,V_j)=\frac{\Delta V}{A}\exp\left[-\frac{1}{2}\sum_{k=1}^{m}\frac{(x_{ik}-V_{jk})^2}{2\delta_k^2}\right] \qquad (3)$$

式中H_c代表给定C类的这一模型；$P(x_i|H_c,V_j)$表示给定H_c和$\{V_j\}$，$\{x_i\}$出现的概率，ΔV为m维空间的微超立方体，其大小与$\{V_j\}$和$\{x_i\}$在计算机中的精度有关；δ_k^2为分布方差，A为归一化常数。

因此误差码的总长度为：

$$E_1=\frac{1}{2}\log_2(e)\sum_{i=1}^{n}\sum_{j=1}^{c}\delta_{ij}\sum_{k=1}^{m}[(x_{ik}-V_{jk})^2/\delta_k^2]+B_2 \qquad (4)$$

式中E_1为误差码长；δ_{ij}为1，如果个体i被分到第j类，否则为0，B_2为一常数。

值得讨论的是属性变量的相互独立性，一般来讲，两个属性变量一般不会相互独立的，这样分布密度函度中有协方差出现，但如对原始数据取坐标变换，如主分量分析，以主要轴的坐标为新的属性指标，就满足了属性变量间相互线性独立的要求。这样作还减少了聚类的计算量，因为主分量分析的目的在于降维，即以较少的新的属性指标来代替原来的较多的属性。而属性越少，计算量也越小。

对模糊ISODATA聚类法，聚类模型为一组聚类中心$\{V_j\}=\{V_{j1},V_{j2},\cdots,V_{jm}\}$，$1\leqslant j\leqslant c$。如果矩阵$[V]=V_{jk}$的每个元素以$N_b$个比特(Blts)在计算机中来表示，则模型码长为

$$M_1=N_b\cdot c\cdot m \qquad (5)$$

式中N_b决定了聚类中心的精度，N_b愈大，V_{jk}的精度就越高，模型的编码就愈长；M_1为总模型码长，它随聚类中心数c直线上升。

对于分级式聚类法，模型码中除有如上聚类中心的码长外，还必须反映各级分类的隶属情况，即模型码中须包含用来说明分类树状图所需的代码。以TWINSPAN为例，其树状图为一二进制树(Binary Tree)，每一枝一般有两个分枝。如图1所示，从树的根部到树梢，凡具有分枝的点被称为节点，而树稍则称作端点。每一个端点代表最终分类中的一个类别，因此具有一个聚类中心。每个节点中存有三个量，一是以该端点以下的所有个体为一类的聚类中心，以m个实数表示，另外两个是该节点下的两个分枝下的结点的地址。以C表示端点数，以S表示节点数(图1中，$C=5$，$S=4$)，则模型码可表示为：

$$M_1=N_b\cdot(C+S)m+2N_a\cdot S \qquad (6)$$

式中N_a是计算机中用来表示一个地址的码长。

总码长D_1是E_1和M_1之和。显而易见，E_1会随C的增加而下降，因为随着聚类类别数的增加，个体与聚类中心的距离会下降，但M_1的值则随C和S直线上升。在C和S的取值区

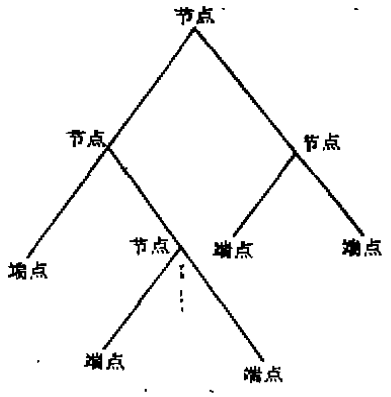

图 1 二进制树的节点和端点
Fig.1 Nodes and terminals of a binary tree

间，必有一最佳的C、S组合，使得总码D_l为最小，其对应的模型结构就是我们要求的最小码优化模型。

到此为止，要运用最小码原理，还需确定各属性变量分布方差δ_i^2。从理论上讲，δ_i^2应由各属性变量的随机程度而定，为此，我们定义综合类内平方和T_i为

$$T_i = \sum_{j=1}^{C}\sum_{j=1}^{n}\delta_{ij}(x_{ji}-V_{ji})^2/(n-c) \qquad (7)$$

一般来说，T_i包括确定和随机两部分变化量。当C从1开始增加，T_i中的确定部分逐渐减少。当C增加到一定值时，T_i中基本只剩下随机部分，其下降趋势迅速减缓，再增加C，T_i无明显变化，此时的T_i可以视为σ_i^2的估计量，在下面的计算中，若有

$$|T_i(i+1)-T_i(i)| \leqslant 0.1\,|T_i(2)-T_i(1)| \qquad (8)$$

则认为$\sigma_i^2 = T_i(i+1)$。式中$T_i(i)$表示以i为类别数时所得的T_i值。

误差码中的常数B_1、B_2与优化问题无关，实际计算时可舍去。

四、计 算 步 骤

从前述原理，不难得出具体分析计算的步骤如下（其中非分级式的模糊ISODATA算法部分已被实施于软件 FISOD (FUZPAK之一)中)：

1. 读入数据矩阵〔X〕，进行有关预处理，如归一化处理，对数变换[2]和排序降维等处理。

2. 对非分级式聚类(如模糊ISODATA)，令类别数C从1开始逐渐增加，对每一C值作相应聚类分析。对分级式聚类分析(如TWINSPAN)，按常规作逐步增加的两分(或多分)分割，每一次两分分割都增加一个节点和一个端点。然后按平均或加权平均的方法算出端点下的类别的中心，再按式(7)求出各属性变量的T_i。根据式(8)的标准确定 σ_i^2 的估计量，直到所有的σ_i^2都估计完毕。

3. 对每一C值，求出M_l和E_l算出D_l。继续增加C值，直到 D_l 从下降转为上升，且模型码M_l大于前所算得的最小的D_l值($D_l$min)为止。$D_l$min所对应的C、S即所求的模型结构参数。

五、实例分析和讨论

笔者以文献〔2〕中的PARK GRASS DATA为一实例用上述方法进行了植物种的聚类分析。该数据计有38个样方，44个植物种。对此数据作主分量分析，得前五轴特征值之和为

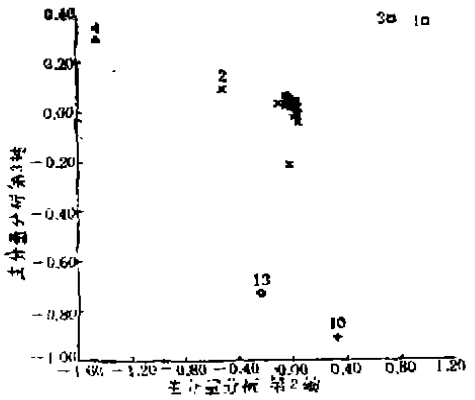

图 2　分析举例的排序和优化聚类结果
Fig.2　The ordination and optimized classification result of the sample analysis

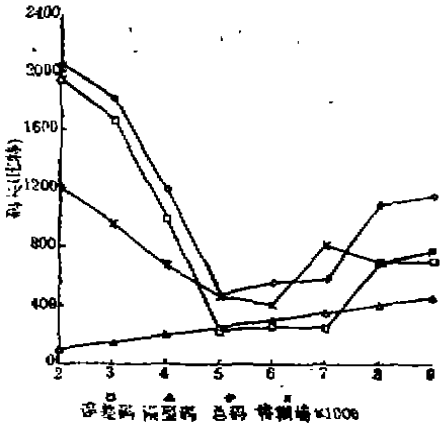

图 3　分析举例的最小码优化过程
Fig.3　The optimization process of the sample analysis

总变化量的93.5%。据原作者分析，第一轴反映了各植物种的平均产量，第二、三轴坐标如图2所示。原作者指出图2中离散程度较大的6个种：1。*Agrostic tennis*，2。*Alopecurus Pratensis*，3。*Anthoranthum odoratum*，4。*Arrhenatherum elatius*，10。*Festuca rubra*，13。*Holcus lanatus* 为优势种，其多度在有关样方中占绝对优势。以前五轴坐标为新的属性变量进行最小码优化模糊ISODATA聚类分析，其优化过程如图3所示。

图中显示了E_f，M_f和D_f以及模糊熵H_f随类别数变化的情况。如前所述E_f随 C 增加而下降，反略有波动，而M_f(在$N_s=10$，$m=5$的情况下)则随C 直线上升，D_f 的最小值出现在$C=5$，而H_f的最小值出现在$C=6$。因此以 $C=5$或$C=6$可望得到较为合理的分类结果。

将$C=5$的模糊ISDATA 聚类结果进行最终归类，归类方法是在软分划矩阵[3]〔R〕中，实行"硬化"。〔R〕的元素r_{ij}表示第 j 个个体（在此例中是植物种）对第i个类别的隶属度。在〔R〕中的每一列(对应于一个植物种)中，找出最大 的 r_{ij}，其所对应的行(代表一个类别)所代表的类别就是该个体应归的类别。按此原则归类，得到的结果如图2所示：4,10,13 号植物种各自成一类，而1，3两植物种成一类，2号种与其余各次要种归于一类。基本上将6个主要植物种区别开来。说明以最小码原理作为聚类优化模型处理是比较合理的。

参 考 文 献

〔1〕Gauch, H. Jr., 1982, Multivariate analysis in community ecology, Cambridge university press, Cambridge, U. K. 173—210.

〔2〕Digby, P. G. N. and R. A. Kempton, 1987, Multivariate analysis of ecological commuuities, Chapman and Hall, New York, 55—69, 124—148.

〔3〕贡嘉祖，冯普臣，1986，《模糊数学及其在林业上的应用》，第137—146页，中国林业出版社，北京。

〔4〕Jarvis, Devra, 1991, TWINSPAN cLassification of plant communities in Sichuan, China, Ph. D. Thesis, University of Washington, U. S. A.

〔5〕Rissanen, J., 1978, Modeling by shortest data description, Automatica, 14:465—471.

〔6〕Rissanen, J., 1983 A universal prior for integers and estimation by minimum description length, The annals of statisfics 11(2):416—431.

THE MINIMUM DESCRIPTION LENGTH PRINCIPLE AS APPLIED TO HIERACHICAL AND NONHIERACHICAL CLUSTERING ANALYSIS

Gao Qiong

(Institute of Botany, Academia Sinica, Beijing)

Many existing clustering methods rely on the experience and insight into the original data of the analyst to determine the final model structure featured by number of groups (C) either directly (i. e., the fuzzy ISODATA analysis) or indirectly (i. e., the TWINSPAN classification). Hence certain arbitrariness and subjectivity are inevitably involved in the result of analysis when the individuals consisting of the data are not significantly clustered and when the analyst has difficult to gain insight into the data. the principle of minimum description length (MDLP) is a outgrowth of Bayesian inference MDLP transforms the probabilities in Bayesian's formula into their respective description lengths, thus, maximizing the probability of a particular model out of a set of models with different complexities (DL) which is sum of the model description length (ML) and the error description length (EL) given the model is the true theory behind the data. The clustering analysis with different C can be regarded as a set of candidate models with complexities or ML proportional to C. EL is measured by the sum of squares of the Euclidean distance of all individuals with respect to their respective clustering centers. A great C often gives smaller EL but greater ML. Minimizing DL is essentially a compensation between EL and ML in light of the theory of Bayesian probability.

Key words: clustering, models, the minimum description length principle.