第13卷第1期　　石　油　实　验　地　质　　V01.13, No.1
1991年3月　　EXPERIMENTAL PETROLEUM GEOLOGY　　Mar., 1991

对于有序聚类程序包ZONATION中几个算法的分析与应用实例

邢　雪　　童国榜　林锦璇

（北京工业大学应用数学系）　（地矿部水文地质与工程地质研究所）

本文对有序聚类程序色ZONATION（“分带”）中所用的三种算法(其中两个算法各含两种度量指标）进行了介绍与剖析，并以天津I 24孔钻井中的孢粉序列分析为例对方法进行了验证.

有序聚类（对于有序样品的聚类、邻接聚类）是一种附有约束条件的聚类分析问题。对于样本x = { X_1, X_2……, X_n }，其中诸样品是m维特征空间的点或向量，有序聚类在进行聚类分析时所得各类中样品的下标（顺序）必须相邻，即各类形式为

$$\{X_i, X_{i+1}, \cdots, X_j\}, \quad 1 \leqslant i \leqslant j \leqslant n。 \qquad (1)$$

有序聚类模型在很多实际问题中适用。在地学中，它特别适应于分带问题。这时，诸样品是取自钻井中的岩芯，它们对应于不同的深度，而样品的特征则是岩芯中的孢粉含量或百分比，也可能是其他内容。

在地质学领域中，近年来出现了许多专用程序色。其中既有庞大而精致的，也有小型而精炼的，有一些甚至是由一、两个人完成的。1985年，由植物学者H.J.Briks和统计学者A.D.Gordon联合研制了有序聚类程序色ZONATION（“分带”）〔1〕。我们手头的程序色是由美国华盛顿大学植物系和第四纪研究中心Matsuo Tsukada教授提供的。该程序色已由我们在IBM—PC型计算机上移植成功。

经典的有序聚类算法是Fisher在1958年提出的，国内通常译作最优分割法〔2〕。事实上，几乎任何一种聚类分析方法都可加以限制使之成为有序的，例如〔3〕，〔4〕也有一些专门用于有序样品的聚类方法，比较复杂的可以归结为非线性规划问题〔1〕。ZONATION选取了三种有序聚类方法，其中前两种是Fisher方法及其变型，但每种方法都包括两个不同的目标函数，第三种则是较少为人重视的有序最短距离法（约束单链法）。本文将对以上三种方法进行介绍和分析，并介绍它在孢粉分带中的一个应用实例。

一、算　法

以下仍然假设样本大小为n，特征空间维数为m。

1. Fisher 方法

本方法定义形式如式（1）所示的任一类的“直径”为

$$d(i,j)=\sum_{t=i}^{j}\sum_{k=1}^{m}(x_{kt}-\bar{x}_{ijk})^2 \qquad (2)$$

或

$$d(i,j)=\sum_{t=i}^{j}\sum_{k=1}^{m}x_{kt}\ln(x_{kt}/\bar{x}_{ijk}), \qquad (3)$$

其中$\bar{x}_{ijk}$表示该类均值X_{ij}的第k分量，而x_{kt}则表示样品X_t（$t=i, i+1, \cdots\cdots, j$）的

$$\bar{X}_{ij}=\sum_{t=i}^{j}X_t/(j-i+1),$$

第k分量。由（2）式定义的类直径称为离差平方和直径，而由（3）式定义的类直径则称为信息含量或熵意义下的直径。后一种定义是比较新的。更进一步，程序色允许对式（2）或（3）中和式的每一项乘以权因子w_k，以标明各特征的不同地位，或者起标准化的作用。

指定全体样品所需要分成的类数g，定义其中$i_1=1$，$j_g=n$，$1\leqslant i_l\leqslant j_l\leqslant n$。于是，

$$t(g,n)=\min\sum_{l=1}^{g}d(i_l,j_l), \qquad (4)$$

聚类问题便化为求所有可能的类直径之和的最小值问题。可以证明以下的递推公式：

$$\begin{aligned}t(g,n)&=\min_{g\leqslant i\leqslant n}\{t(g-1,i-1)+d(i,n)\},\\ t(2,n)&=\min_{2\leqslant i\leqslant n}\{d(1,i-1)+d(i,n)\},\end{aligned} \qquad (5)$$

于是，当g确定后，即可由此出发逐步得到合理的分类结果。

2.对分法

对分法的出发点与基本概念和Fisher法相同。不同之处在于分类的步骤：本方法每次将一类样品按Fisher法的原则（同样，可以选择离差平方和或熵两种类直径）分成两类。第一次分类将找到序号i_0，使

$$d(1,i_0-1)+d(i_0,n)=\min_{1\leqslant i\leqslant n}\{d(1,i-1)+d(i,n)\},$$

于是将全体样品分为两类G_1和G_2。第二步，将G_1和G_2各自分为两类，再判断是分裂G_1好还是分裂G_2好，从中找出更合理的方案，于是将全体样品分为三类。继续这个步骤直到达到所要求的类数为止。

当样本大小较大时，本方法可以避免计算各个t(j, i)，因此本方法的计算时间可能比Fisher法节省。

3.约束单链法

本方法相当于对常用的系统聚类法[5]进行有序约束的情形。值得注意的是，程序色中只采用一种类间距离，即最短距离。

程序色定义任两种样品 X_i 与 X_j 之间的距离为绝对值距离（“市区距离”）：

$$d_{ij}=\sum_{k=1}^{m}|x_{ki}-x_{kj}|,$$

其中 x_{ki} 为 x_i 的第k个分量。定义任两类 G_i，G_j 间的距离为两类中样品两两距离的最小值：

$$d(G_i,G_j)=\min_{\substack{Z_k\in G_i\\ Z_l\in G_j}}\{d_{kl}\}。$$

聚类开始前每个样品自成一类，然后每次合并距离最近（在下标相邻的前提下）的两类（点）。如果有两对点间距离都取当前最小值，便将它们同时合并。继续上述步聚直到达到所需的类数。

回顾上述三种方法，我们认为有两个事实值得注意。首先，在[1]中叙述的有序聚类方法还有很多，包括新颖而复杂的可变壁垒法。但是，程序色中只收纳了以上三种方法。似乎可以认为，研制者认为上述三种方法在一般情况下足以解决问题。或许他们还考虑了在微机上的可行性问题。其次，按照惯例，方法三中可以使用的类间距离定义不下六、七种之多，而最短距离法通常被认为效果较差的一种。但是，程序色中偏偏只使用最短距离法，而摒弃了其他距离，包括公认为效果最佳的离差平方和增量。关于这一点，我们的分析如下：最短距离适用于聚集成链状分布的类，而Fisher法适用于团状分布的类。研制者正是考虑到了这两种极端的情况。基于同样的理由，离差平方和增量正是由于效果与Fisher法相近而不被采用。我们的实践证实，对于本文中的实例，方法三的效果优于另二者。而对于另外一些数据[6]，结论恰好相反。

程序色的其他性能如下：（一）运行速度较快。（二）具有较充分的人机会话功能，例如加权，确定类数等。当使用者事先无法指定类数时，程序色将自动地认为类数为11（在多数情况下，这一数字已经足够多）。（三）输出方式包括屏幕显示，打印，枝状图，数字等等。

二、运行实例

原书中的例子样本大小较小。我们使用程序色对一个实例进行了试算，即对天津地区Ⅰ24孔井●的孢粉资料进行了分带以确定地层结构。

天津Ⅰ24孔井深183米，对应于地质上的第四纪全新统、上更新统及中更新统的上部地层。地层中含有丰富的孢粉化石。在不同的深度上取得了48个岩芯，其中45个孢粉统计数目较多。符合计算要求，用以作为聚类样品。对每样品选取20个特征，包括松（Pinus）、桦（Betula）、榆（Ulmus）、蒿（Artemisie）、藜科(Chenopodiaceae)禾本科（Gramineae）、香蒲（Typha）等各类植物孢粉含量。因此，聚类时 $n=45$，$m=20$。由于事先对类数无法做出估计，故取 $g=11$。计算结果如表1所示。表1右起第一列是地质工作者所鉴定的结果●，用作对照。

● 该孔孢粉化石由柯曼红同志鉴定

● 凌泽民等，天津市粘性土2工程地质特征及其对地面沉降的影响，1985。

1·对分法（以熵为度量）

表1的左起第2—11列显示了以熵为度量时的对分法聚类结果，第12列是最终结果，利用连续函数确定[7]。当连续函数K≥70%时，全部样品从上至下可分为五大类，依次记作A，B，C，D，E。

可以看出，与地质工作者的鉴定对照，A类精确地对应于时间Q_4，B，C两段对应于Q_3^3，D大体上对应于Q_3^2与Q_3^1，E大体上对应于Q_2^2，其间相差一个样品。

2.对分法（以离差平方和为度量）

同样，表的第13—22列显示了以离差平方和为度量时的对分结果，第23列为最终结果，仍取连续系数K≥70%，将全体样品划分为A′，B′，C′，D′和E′五大类。A′精确地对应于Q_4，其他几类着重刻划了井的上部各样品间的差异，而对下部样品的差异反映不够明显。因此，本方法对孢粉组合变化显著的部分（3—10，26—33）比较敏感。从这一点出发可以为孢粉组合演变的分析提供依据。

3.Fisher法（以离差平方和为度量）

限于篇幅，表中只举出了分为11类的情况，见第24例。由表可见，本方法对全井样品的划分比较匀称，除Q_3^3与Q_2^3外，其他几段间的边界附近都有相应的分界点。在3—10，25—32号样品间又各自细分为三类，反映了孢粉组合的变化。

4.有序最短距离法

本方法的聚类结果位于表1右起第3列，各类中带有星号的数字是聚类时的最短距离，它们显示了聚类的次序。如果用0.513，0.738，1.180作为界限，则全体样品被分为三大类，分别记作Ⅰ，Ⅱ，Ⅲ，其中类Ⅱ又可细分为四类。可以看到，本方法所得结果与地质人员的判断比较一致。这一事实表明，最短距离法在某些情形下是值得重视的。

纵观四种方法的聚类结果，以熵为度量的对分法和以最短距离为类间距离的有序点群分析法分出的类都比较大，反映了孢粉组合变化的较长周期性，也与地质工作者对地层划分的结果比较接近。与凌泽民等同志的工作相比，对1—3号样品（反映了对全新统下界的划分）、42—45号样品（反映对上更新统底界的划分，由于本次试算中删去了原有第32，36，46号样品，所以现在的41号样品就是原43号样品）的划分完全一致，其他结果也颇为接近。

与此不同，另外两种方法（以离差平方和为度量的对分法及Fisher法）则对孢粉组合变化显著的部分（3—10，25—32号样品）较为敏感。综上所述，在有序聚类中同时使用几种方法并对结果进行比较是有益的。

综合四种结果，我们将本批样品分为五个带，记作Ⅰ—Ⅴ，见表1右起第二列。不难看出，本次结果与最短距离法结果渊源最深。与地质学者的工作相比，差别在于Q_2^3的厚度有所增加，其他都是一致的。关于两者的差别，一种观点认为聚类结果更正确地

反映了孢粉组合的特征；另一种观点则认为Q_3^2的顶界部分孢粉组合的变化本来就不甚明显，因而是有待商榷的。

有序最短距离的枝状分类图见图1。全部四种方法在IBM—PC上的计算时间不超过30分钟。目前，我们正与地矿部合作研制一个更大的有序聚类程序包。

感谢Matsuo Tsukada教授最早向我们提供了程序包。感谢杨文红、李枝漫、陈祖荫与邢桥同志对我们工作的支持。

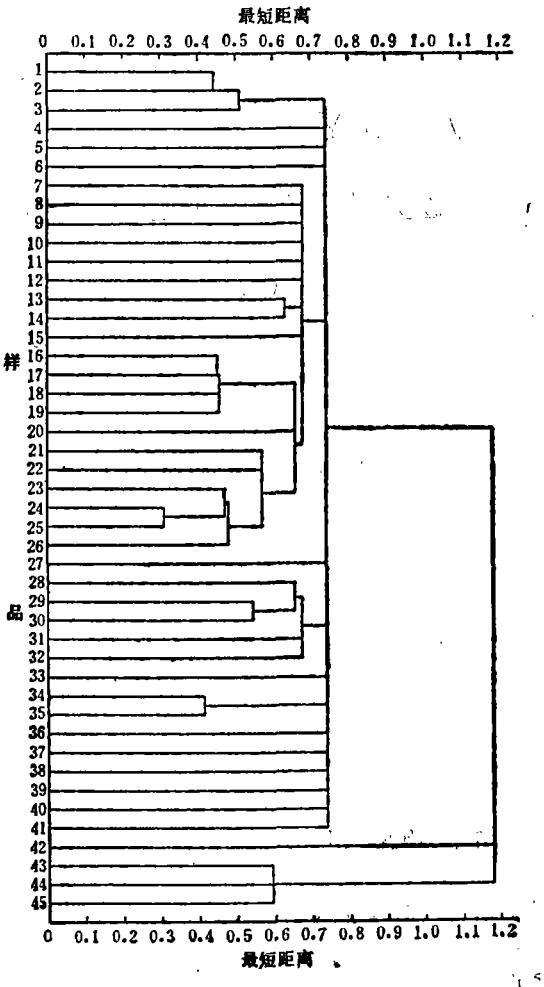

图1　45个样品有序聚类所得枝状图

参 考 文 献

[1] Brik, H.J., Gordon, A.D., Numerical Methods in Quaternary Pollen Analysis, ACADEMIC PRESS, 1985.

[2] Fisher, W.D., On Grouping for Maximum Homogeneity, J.Amaer.Statist.Assoc., Vol.53, 1985.

[3] 方开泰、潘恩沛，1982，聚类分析，地质出版社.

[4] 王碧泉、陈祖荫、童国榜等，研究强震孕震过程的若干有序集群方法，地震学报，1986.4.

[5] Wishart, W., An Algorithm for Hierarchical Classification, Biometrics, Vol.25, No.1, 1969.

[6] 童国榜，林锦璇，陈祖荫，化石孢粉的有序集群，石油实验地质，（待发表）1990.

ANALYSIS OF SEVERAL ALGORITHMS IN ZONATION USING SOFTWARE PACKAGE OF ORDERED CLUSTER

Xing Xue

(Mathematics Department of Beijing Industry University)

Tong Guobang **Lin Jinxuan**

(Institute of Hydrogeology and Engineering Geology, MGMR)

Abstract

The three algorithms (two algorithms contain two measuring indexes) in the zonation using software package of ordered clustering are introduced and analysed. The method is tested, taking the analysis of sporo-pollen from Tianjin hole I24 as an example.

表1 天津I-24孔孢粉带的计算结果

方法	对分算法																							动态算法	有序点群最短距离分类结果	孢粉带	时代**
分类准则	信息含量或熵										分类结果	离差平方和										分类结果	离差平方和				
分类数		3		5		7		9		11			3		5		7		9		11						
1											A											A′	1	I	I	Q_4
2																											
3																								0.513*			
4											B											B′ C′	2				
5																											
6											C											D′	3				
7																											
8																							4				
9																								I_1	I		
10																											Q_3^3
11																											
12																											
13																											
14																											
15																								0.675*			
16																											
17																							5				
18																											
19																											
20																								I_2	I		
21																											Q_3^2
22																											
23																											
24																											
25																						E′					
26																							6	0.659*			
27																											
28											D												7	I_3			
29																											
30																							8				
31																											
32																								0.676*			Q_3^1
33																									IV		
34																											
35																											
36																							9	I_4			
37																											
38																											
39																											
40																											
41																								0.738*			
42																							10				
43											E													II	V	Q_3^2
44																							11				
45																								1.180*			

* 聚类时的最短距离系数

** 1985凌泽民等，天津市粘性土工程地质特征及其对地面沉降的影响