第 8 卷 第 2 期 地 震 学 报 Vol. 8, No. 2
1986 年 5 月 ACTA SEISMOLOGICA SINICA May, 1986

研究强震孕震过程的几种有序集群方法

王碧泉[1] 陈祖荫[2] 童国榜[3] 王春珍[1]

摘 要

本文用五种有序集群方法研究了华北地区和宁夏地区强震的孕震过程. 我们得到: 五种方法均能模拟地震活动交替的高潮、低潮现象; Fisher 方法和有序点群分析方法(采用离差平方和增量作为类间距离时)的效果均较好, 表明在各类中诸时间段的结构主要是团状分布而不是链状分布, 因此选择预报方法时考虑地震活动的这一特点是重要的; 对分移动窗口法和非参数拟合优度法的结果, 表明在相对平静期和显著活动期之间确实存在着客观的谷值; 图论方法则有助于进一步分出显著活动期中易发生强震的较小峰值.

五种方法中以 Fisher 方法的误识率 $\hat{\varepsilon}$ 最低, 而非参数拟合优度法则能帮助确定较为合理的分类. 此外对资料预先平滑也可改善结果.

一、引 言

强震的预测是人们关心的重要研究课题之一. 王碧泉等[1-5]从 1979 年起将模式识别方法用于研究强震前的中等地震活动特征和触发动力因子, 试图预测强震发生的时间. 凯依里斯-博洛克等[6] 和艾伦等[7] 在 1983 年也报告了将模式识别方法应用于预测强震发生时间方面的有关结果. 由于该方法能综合分析研究与强震的发生有关的大量的、各种各样的信息(特征), 因此, 它在地震预测中的应用已日益增多.

我们用模式识别中的一种方法——集群(或聚类)法曾得到[3], 有序集群比系统集群的结果好, 此结果从数学的角度有效地模拟了地震成组活动的现象. 模式识别是以特征分析为基础, 进而按一定的算法对诸样品进行分类的方法. 所谓有序样品的集群问题(简称有序集群问题)可以叙述如下: 对一批按一定次序排列的样品点 $X_1, X_2, \cdots, X_n$ 进行集群, 并要求每类的形式为

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

即在集群过程中不改变各点的排列顺序. 其中每个 X_i 可以是任意维特征空间中的点.

在模式识别中, 有基于各种数学方法的不同集群方法. 考虑到有序集群的结果较好, 本文中我们修改了以谷值寻找方法为基础的非参数拟合优度法、图论方法以及传统的点群分析方法, 使之适用于有序集群. 连同另外两种方法一起, 我们共应用了五种有序集群方法对我国华北地区及宁夏地区的样品进行了集群, 并对比了结果.

本文 1984 年 10 月 10 日收到, 1985 年 1 月 31 日收到修改稿.

1) 国家地震局地球物理研究所; 2)北京工业大学; 3)地质矿产部水文地质与工程地质研究所.

联系起来研究强震预测方法和强震的孕育过程是有益的. 模式识别方法除可对样品进行分类以预测强震发生的时间以外，还可以从数学的角度模拟地震的成组活动规律. 本文的另一目的是，在方法对比的基础上进一步揭示样品的结构和探讨强震的孕育过程.

二、特征的选取和资料的平滑

1. 样品和特征的选取

本文所研究的地区为华北地区（约为 109°—125°E, 31°—45°N 的范围）[1]和宁夏地区（102°—107°E, 31°—40°N）[3]. 所取用的地震活动、太阳黑子相对数、地球自转速度和加速度的资料与文献[2]中的资料相同.

本研究中每个样品为一个时间段. 试验表明，时间段的长度取为 3 年较好[3]. 现以 3 年为间隔，将 1836 年至 1982 年共分为 49 个时间段. 全部样品分为危险（D类）和安全（N类）两大类. D时间段是指在该段时间中曾发生 6.0 级以上的地震，N时间段则是指在该段时间中未曾发生 6.0 级以上的地震.

在下述几种模式识别方法的研究中，对每个时间段均选取以下 12 项特征（按文献[3]）:

x_1 本时间段前 15 年内 5.0—5.4 级地震次数;

x_2 同上，5.5—5.9 级地震次数;

x_3 同上，6.0—6.9 级地震次数;

x_4 同上，7.0 级以上地震次数;

x_5 同上，5.0—6.9 级地震次数;

x_6 同上，5.5—6.9 级地震次数;

x_7 本段前 16 年内，后 8 年与前 8 年的 5.0—5.4 级地震次数之比;

x_8 同上，5.5—5.9 级地震次数之比;

x_9 本段前 15 年内 x_2 与 x_1 数值之比;

x_{10} 太阳黑子相对数;

x_{11} 地球自转角速度;

x_{12} 地球自转角加速度.

其中 x_1 至 x_9 主要考虑了各级地震的频度、b 值以及频度随时间的变化；x_{10} 至 x_{12} 主要考虑了外力的触发作用. 对于上述特征曾进行过单个特征的检验、相关系数检验、典型相关分析、“R 型”聚类分析和线性映射，其结果表明了各特征的有效性[4].

2. 地震资料的平滑

我国有悠久的历史地震记载，为我们研究强震前的中等地震活动特征提供了丰富的资料. 但是历史地震目录在早期缺漏较多，以后随时间而日趋完整. 1900 年以后才有仪器记录地震，并根据仪器记录测定了地震的震级、震中等参量. 1957 年以后我国地震台网逐步建立起来，使能记录到的地震震级下限逐步降低，且震级、震中等参数的测定日益准确. 因此，某震级区间内地震频度的测定必受上述震级测定精度和记录完整情况的影

响. 例如，x_1，即 5.0—5.4 级地震频度的原始数据曲线中[图 1(a)]，在 1957 年以后 x_1 取值明显偏高可能就是上述原因的结果. 由于不同时期中资料精度和完整性不一致会影响模式识别的效果[1]，为此，对地震活动性特征进行了下述的 5 点 3 次平滑[8]. 平滑的目的在于消去数据中的噪声干扰，5 点 3 次平滑相当于利用 3 次最小二乘多项式的平滑，方法如下述：

对于任意 m 个等距点 $x_1, x_2, \cdots, x_m$ 上的数据 $y_1, y_2, \cdots, y_m$，令其平滑后的值为：

$$\bar{y}_1 = \frac{1}{70}(69y_1 + 4y_2 - 6y_3 + 4y_4 - y_5)$$

$$\bar{y}_2 = \frac{1}{35}(2y_1 + 27y_2 + 12y_3 - 8y_4 + 2y_5)$$

$$\bar{y}_i = \frac{1}{35}(-3y_{i-2} + 12y_{i-1} + 17y_i + 12y_{i+1} - 3y_{i+2})$$

$$i = 3, 4, \cdots, m-2$$

$$\bar{y}_{m-1} = \frac{1}{35}(2y_{m-4} - 8y_{m-3} + 12y_{m-2} + 27y_{m-1} + 2y_m)$$

$$\bar{y}_m = \frac{1}{70}(-y_{m-4} + 4y_{m-3} - 6y_{m-2} + 4y_{m-1} + 69y_m)$$

将全部资料分为三段（1899 年以前；1900年—1956 年；1957 年以后）进行了平滑，并将各段特征均值按第三段的均值予以调整.

图 1 示出了平滑前和平滑后的 x_1 和 x_5 值随时间变化的情况. 由图可见，平滑后的 x_1 和 x_5 值在整个时期中均在同一水平上波动，修正了原始数据中不同时期的频度相差较大

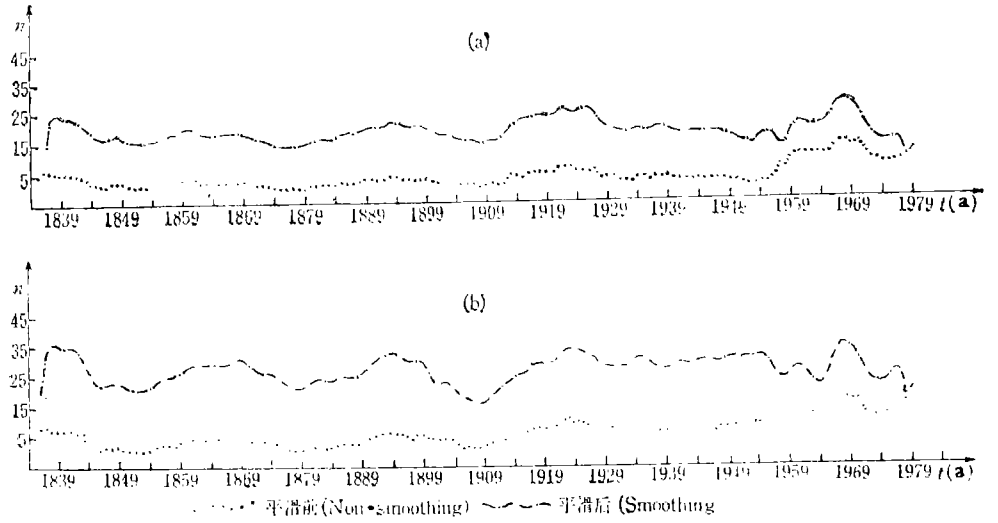

图 1　平滑前和平滑后的 x_i-t 图

(a)——x_1-t　(b)——x_5-t

Fig 1　Comparing non-smooth x_i-t with smooth x_i-t

(a)——x_1-t　(b)——x_5-t

的现象，而且还保留了其随时间的小的波动. 算法允许对同一批数据进行一次以上的平滑，直到得到较为符合实际情况的结果.

三、几种有序集群方法

1. Fisher 方法

Fisher 方法定义类

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

的直径为

$$d(i, j) = \sum_{l=i}^{j} (X_l - \bar{X}_{ij})^T (X_l - \bar{X}_{ij}) \tag{2}$$

其中 $\bar{X}_{ij}$ 为本类中各样品的均值. 集群的原则是使误差函数 $e[P(n, g)]$ 达到最小，即使

$$e[P(n, g)] = \min_{g \leqslant j \leqslant n} \{e[P(j-1, g-1)] + d(j, n)\} \tag{3}$$

其中 n 为样品总数，g 为类数，$e[P(n, g)]$ 表示 n 个样品分为 g 类时的误差，等等. 关于本方法的详细介绍参看文献[3][9].

本方法是最常使用的有序集群方法. 对于确定的类数 g 以及上述的集群准则，本方法可以得到全局最优解. Fisher 方法的两个不足之处是：1. 它的集群准则对于某些特定结构的样品集可能不适用[10]；2. 在实际问题中，往往难以事先确定类数，因而本方法通常不能指出最佳分类数.

2. 对分移动窗口方法[9]

对分移动窗口方法由 Webster 提出. 本方法要求事先确定一正整数 P_n，每次考察由 $2P_n$ 个有序样品构成的“窗口”

$$\{X_i, X_{i+1}, \cdots, X_{i+2P_n-1}\} \tag{4}$$

并计算前 P_n 个点与后 P_n 个点的均值之间的 Mahalanobis 距离平方：

$$d_M^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})^T \Sigma^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)}) \tag{5}$$

其中 $\bar{X}^{(1)}$ 与 $\bar{X}^{(2)}$ 分别表示窗口中前 P_n 点与后 P_n 点的均值，Σ 为全体样品的协方差矩阵. 当 d_M^2 超过某一阈值时，认为窗口的前后两部分应被分到不同的类. 不断移动窗口使之与各个样品点匹配便可找到所有分界点.

本方法在计算上远比 Fisher 方法简捷. 但是这一方法的结果显然会受到 P_n 值的影响. 当 P_n 较小时，所得的结果可能只反映样品集的局部性质；而 P_n 过大则可能漏掉合理的分界点.

3. 有序点群分析

对常用的点群分析或系统聚类方法[9]，我们予以修改以适应于有序集群的需要. 点群分析开始时假定 n 个样品各自代表一个独立的类，然后每次合并距离最近的两类，直到全体样品合为一类为止. 在用于有序样品时，要求每次被合并的两类必须取以下形式

$$\{X_P, X_{P+1}, \cdots, X_q\}, \{X_{q+1}, X_{q+2}, \cdots, X_r\} \tag{6}$$

即每次合并后应保持各样品序号的顺序不变.

点群分析的关键在于类间距离的选择. 如果用 $d(l, m)$ 表示点 X_l 与 X_m 间的距离,则两个类 ω_i, ω_j 间的距离 $d(\omega_i, \omega_j)$ 通常有以下几种定义方式:

(1) 最短距离 定义 ω_i 与 ω_j 的距离为两类中相距最近的两点间的距离,即

$$d(\omega_i, \omega_j) = \min_{\substack{X_l \in \omega_i \\ X_m \in \omega_j}} d(l, m) \tag{7}$$

(2) 最长距离 与(1)相反,定义 $d(\omega_i, \omega_j)$ 为两类中相距最远的两点间的距离,即

$$d(\omega_i, \omega_j) = \max_{\substack{X_l \in \omega_i \\ X_m \in \omega_j}} d(l, m) \tag{8}$$

(3) 平均距离 定义 $d^2(\omega_i, \omega_j)$ 为两类中各样品点两两之间距离平方的平均数,即

$$d^2(\omega_i, \omega_j) = \frac{1}{n_i n_j} \sum_{\substack{X_l \in \omega_i \\ X_m \in \omega_j}} d^2(l, m) \tag{9}$$

其中 n_i, n_j 分别是两类样品的个数.

(4) 重心距离 定义 $d(\omega_i, \omega_j)$ 为两类重心(均值)之间的距离.

(5) 离差平方和增量 设 ω_i, ω_j 和将它们合并后所得的新类 ω_{ij} 的离差平方和分别为 S_i, S_j 和 S_{ij}, 其中

$$S_i = \sum_{X_l \in \omega_i} (X_l - \bar{X}_i)^T (X_l - \bar{X}_i) \tag{10}$$

类似地可写出 S_j 和 S_{ij} 的表达式,并定义

$$d(\omega_i, \omega_j) = S_{ij} - S_i - S_j \tag{11}$$

为将两类合并后离差平方和的增量. 显然,增量愈小,则分类效果愈佳.

对于有序样品,有人还建议使用所谓

(6) 间隙距离 当 ω_i, ω_j 分别取(6)中所规定的形式时,定义

$$d(\omega_i, \omega_j) = d(q, q+1) \tag{12}$$

即前类末点与后类首点间的距离("间隙").

对于样品点间距离 $d(l, m)$ 本项研究采取两种定义方式,即(1)欧氏距离和; (2) 相似系数 r_{lm} (确切地说,取 $d(l, m) = 1 - r_{l,m}$).

以下两种算法都是我们对于一般集群算法进行修改的结果. 而且,它们都属于所谓非参数方法并以各点邻域中诸点的分类状况为基础.

4. 非参数拟合优度方法

原始意义下的拟合优度方法是一种对已知分类进行调整的算法[11]. 假设每个样品已分到 M 个类中的一个, 并记第 i 样品 X_i 的所在类为 $\omega_{k_i}(i = 1, 2, \cdots, N; 1 \leqslant k_i \leqslant M)$. 全体 ω_i 构成的矢量

$$\boldsymbol{\Omega} = [\omega_{k_1}, \omega_{k_2}, \cdots, \omega_{k_N}]^T$$

称为一个分类. 本方法的任务是求一个在某种意义下最优的 $\boldsymbol{\Omega}$. 为此首先需要确定一个集群准则 $J = J(\boldsymbol{\Omega})$ 和一个初始分类 $\boldsymbol{\Omega}(0)$, 并从 $\boldsymbol{\Omega}(0)$ 出发不断进行调整以使 J 达最小. 如果当前分类为

$$\boldsymbol{\Omega}(l) = [\omega_{k_1}(l), \omega_{k_2}(l), \cdots, \omega_{k_N}(l)]^T \tag{13}$$

则可试将第 i 点从现在的类 $\omega_{k_i}(l)$ 分到类 ω_j 中，此时 J 的变化量为

$$\Delta J(i, j, l) = J[\omega_{k_1}(l), \cdots, \omega_{k_{i-1}}(l), \omega_j, \omega_{k_{i+1}}(l), \cdots, \omega_{k_N}(l)] - J[\boldsymbol{\Omega}(l)] \tag{14}$$

若 $\Delta J(i, j, l) < 0$，则表明将 X_i 分到 ω_j 会使分类得到改善.

Koontz 等[12]提出的集群准则为

$$J(\boldsymbol{\Omega}) = \sum_{i<j}^{n}\sum f_R(X_i, X_j)[d_\omega(\omega_{k_i}, \omega_{k_j}) - d_X(X_i, X_j)]^2 \tag{15}$$

其中 $d_\omega(\omega_{k_i}, \omega_{k_j})$ 表示 ω_{k_i}, ω_{k_j} 两类间的距离而 $d_X(X_i, X_j)$ 表示 X_i, X_j 两点间的距离，$f_R(X_i, X_j)$ 是权. 这一准则的直观意义是明显的：任两点间的距离应与它们所在类间的距离尽量接近. 在进行了一系列简化之后，上述准则可以简化为以下形式：

$$J_{1R}(\boldsymbol{\Omega}) = \sum_{i<j}^{n}\sum f_R(X_i, X_j)[1 - \delta(\omega_{k_i}, \omega_{k_j})]^2 \tag{16}$$

而

$$f_R(X_i, X_j) = \begin{cases} 1 & d_X(X_i, X_j) \leqslant R \\ 0 & d_X(X_i, X_j) > R \end{cases} \tag{17}$$

$$\delta(\omega_{k_i}, \omega_{k_j}) = \begin{cases} 1 & \omega_{k_i} = \omega_{k_j} \\ 0 & \omega_{k_i} \neq \omega_{k_j} \end{cases} \tag{18}$$

这一准则的计算是极为简单的：对于任两点 X_i, X_j，首先，根据(17)和(16)，若 $d_X(X_i, X_j) > R$，则不必考虑（相距较远的点不会影响分类效果）；其次，根据(18)和(16)，若 $\omega_{k_i} = \omega_{k_j}$，也不必考虑(两点相距近而分在同一类时也不影响分类效果). 因此，J_{1R} 等于距离小于参数 R 而又不分在同一类的点对数目.

准则 J_{1R} 的变化增量是

$$D_{1R}(i, j, l) = \sum_{\substack{r=1 \\ r \neq i}}^{n} f_R(X_i, X_r)\delta[\omega_j, \omega_{k_r}(l)] \tag{19}$$

即对每个 X_i，考虑其 R 邻域中各点(不包括 X_i 自身)的类别并将它分到点数最多的类.

蒋代梅等[13]将 Koontz 的方法作了推广以使它成为一个完整的集群算法. 下面叙述我们在此基础上得到的有序集群方法.

首先，假定 n 个样品自成一类，并记此时的分类为

$$\boldsymbol{\Omega}^{(0)}(0) = (1, 2, \cdots, n)^T \tag{20}$$

其中第 i 分量表示点 X_i 所在的类号. 还需确定一个初始参数 $R^{(0)}$ 和步长 h.

在算法的任一步，设当前分类为 $\boldsymbol{\Omega}^{(k)}(l)$，其中 k 是外层迭代次数而 l 是内层迭代次数. 对于每个 $X_i (i = 1, 2, \cdots, n)$，求出序列

$$X_a, X_{a+1}, \cdots, X_i, \cdots, X_b \tag{21}$$

其中

$$\begin{aligned} &d(X_j, X_i) \leqslant R^{(k)} \qquad (j = a, a+1, \cdots, b) \\ &a = 1 \text{ 或 } d(X_{a-1}, X_i) > R \end{aligned} \tag{22}$$

$$b = n \text{ 或 } d(X_{b+1}, X_i) > R$$

换句话说，序列(21)表示在 X_i 的 $R^{(k)}$ 邻域[1]中与 X_i 的序号相邻接的全部样品点. 考虑其中每个点（X_i 自身除外)的类别并计算各类点数，将 X_i 重新分配到拥有点数最多的类，得到新的分类 $\Omega^{(k)}(l+1)$. 若条件

$$\Omega^{(k)}(l+1) = \Omega^{(k)}(l) \text{ 或 } \Omega^{(k)}(l+1) = \Omega^{(k)}(l-1) \tag{23}$$

都不满足，则重复上述步骤直到得到一个稳定的分类为止. 否则，认为本次内层迭代结束，重新调整各点类号. 如果全体样品已经合并为一类，则过程停止. 否则令 $R^{(k+1)} = R^{(k)} + h$, $\Omega^{(k+1)}(0) = \Omega^{(k)}(l+1)$，重新进行分类. 由于 $R^{(k+1)}$ 有所增加，重新分类的结果将使类数进一步减少.

试验表明，本方法可以处理各种类型的样品集，包括 Fisher 方法有效或无效的情况. 另外，本方法的内层迭代次数可以作为判断类数是否最佳的一个标志. 若条件 (23) 满足时，$l+1$ 大于某一阈值，则表示本次初始分类 $\Omega^{(k)}(0)$ 是一个比较稳定的分类，它需要经过较长时间迭代才能转移到新的情况.

5. 图论方法

Koontz[14] 等提出的图论方法同样根据每个样品点某邻域中各类点个数对其加以分类. 我们在此基础上提出的有序算法如下所述.

确定参数 $R^{(0)}$ 和步长 h. 在算法的任一步，对每个 X_i 求序列(21)，并令 N_i 表示序列(21) 中点的个数 (X_i 自身除外)而

$$g_{ij} = \frac{N_j - N_i}{d_{ij}} \tag{24}$$

其中 d_{ij} 是 X_i 与 X_j 的距离.

对每个 X_i:

(1) 若 $N_i = 0$，则 X_i 是一有向树的根(孤立点).

(2) 若 $N_i \neq 0$，对序列(21)中 X_i 的两个邻点求 g_{ij}，求出使 g_{ij} 达到最大的点 X_k，即

$$g_{ik} = \max_{j=i-1,i+1} g_{ij} \tag{25}$$

(3) 若 $g_{ik} < 0$，则 X_i 是根(密度最大的点).

若 $g_{ik} > 0$，则 X_i 的上一结点是 X_k.

若 $g_{ik} = 0$，则考虑 X_{i-1}, X_{i+1} 中使 $g_{ij} = 0$ 而且不是 X_i 的下一结点的所有点，记这些点之集为 π_i.

若 $\pi_i = \phi$，则 X_i 是根.

若 $\pi_i \neq \phi$，则 X_i 的上一结点为 $X_{k'}$:

$$d_{ik'} = \min_{j \in \pi_i} d_{ij} \tag{26}$$

由此可以得到一批有向树，每个树对应于一类. 树根是密度最大的点，以下各结点的密度依次减少. 当全体样品点已并为一类时停止计算，否则令 $R^{(k+1)} = R^{(k)} + h$ 并重新分

1) X_i 的 $R^{(k)}$ 邻域，指以 X_i 为中心，$R^{(k)}$ 为半径的圆(球，超球).

类.

四、结果的比较

1. 平滑的效果

同文献[3],我们仍以下列公式计算分类结果的误识率 $\hat{\varepsilon}$:

$$\hat{\varepsilon} = \sum_{i=1}^{2} P(\omega_i)\tau_i/n \tag{27}$$

其中 ω_1 和 ω_2 分别表示 D、N 两类样品，$P(\omega_i)$ 为第 i 类样品的先验概率，τ_i 为第 i 类样品中被错误分类的样品数(此处 $i=1$ 或 2)，n 为样品总数.

对华北和宁夏两地区,用上述诸种有序集群方法,分别对平滑前和平滑后的特征进行了模式识别. 对于宁夏地区,总的看来各种方法对平滑后特征的结果均比未平滑的好.以(27)式的 $\hat{\varepsilon}$ 来衡量,对平滑后的特征,样品分类后的误识率 $\hat{\varepsilon}$ 均较小. 图 2(b) 和图 2(c)

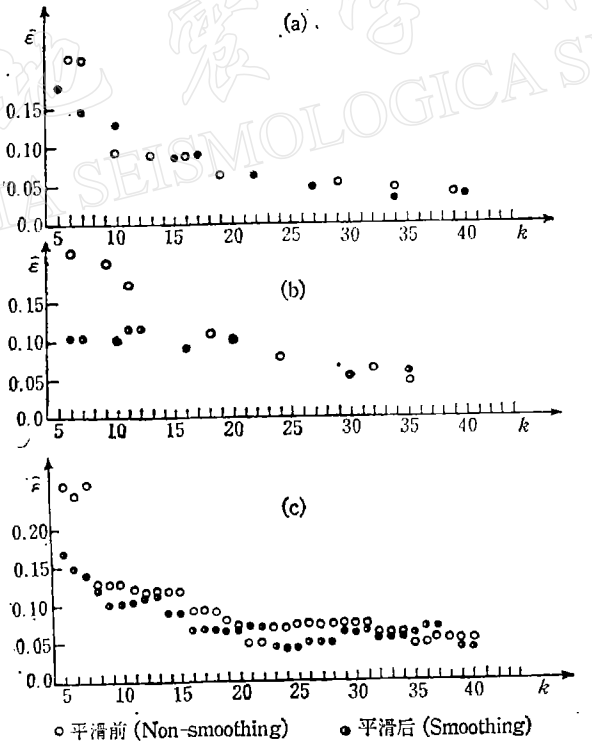

图 2 平滑前和平滑后 $\hat{\varepsilon}$ 的比较

(a)——非参数拟合优度方法(华北地区) (b)——非参数拟合优度方法(宁夏地区)
(c)——Fisher 方法(宁夏地区)

Fig 2 Comparing the error $\hat{\varepsilon}$ in result of non-smoothing with that of smoothing

(a)——nonparametric goodness of fit (North China) (b)——nonparametric goodness of fit (Ningxia region) (c)——Fisher's method (Ningxia region)

分别示出了非参数拟合优度方法和 Fisher 方法的 $\hat{\varepsilon}$，可见平滑后特征的分类效果好一些（$\hat{\varepsilon}$ 较小）．但对华北地区的资料平滑后，其分类效果的改进不如宁夏地区明显．多数方法中资料平滑后结果的误识率 $\hat{\varepsilon}$ 小一些，个别平滑后的 $\hat{\varepsilon}$ 略大．图 2(a) 举例示出华北地区一种方法的结果．

考虑到宁夏地区历史地震资料前期缺漏地震较华北地区多，以上多种方法的结果提示我们：对资料预先进行平滑可以改善结果．对类似于宁夏地区的资料，平滑后对结果的改善较为明显．

2. 非参数拟合优度方法和 Fisher 方法的结果

(1) 我们曾得到，由于 Fisher 的有序聚类法并不认为各样品是彼此孤立的，而是只对相邻的时间段才进行集群，因此能反映较长时期内地震活动随时间变化的规律，从而较好地模拟了华北地区地震活动交替的高潮和低潮现象[3]．本文所述的各种有序集群方法也都具有以上特点．取华北地区的资料并对特征进行平滑后（以下均取华北地区平滑后资料的结果），采用本文所述的各种有序集群方法，对样品进行了集群（分类）．图 3(b) 和

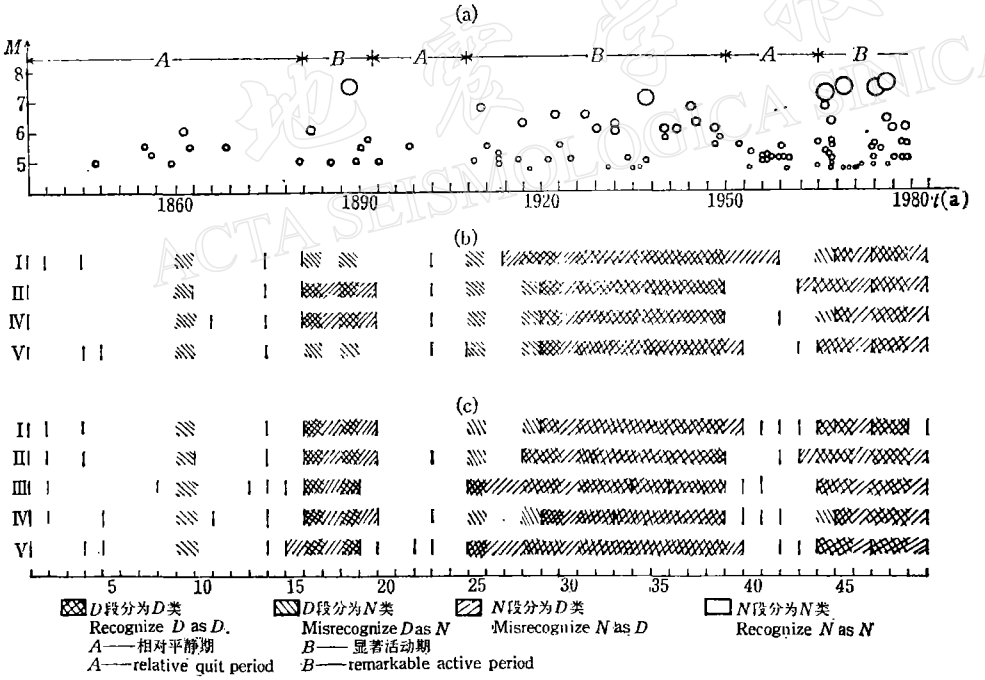

图　3

(a)——华北地区震级 (M)-时间 (t) 图　(b)——四种方法有序集群结果 ($K=10$)：*I*. 非参数拟合优度法；*II*. Fisher 方法；*III*. 图论方法．*V*. 对分移动窗口法，$P_n=2$．(c)——五种方法有序集群结果 ($K=15$)：*I*，*II*，*III*，*V* 同 (b)；*IV*. 有序点群分析法，$D_1=5$，$D_2=1$．

Fig. 3　(a)——Magnitude (M)-time (t), graph in North china. (b)——order clustering results of 4 methods ($K=10$): *I*. nonparametric goodness of fit; *II*. Fisher's method. *III*. graph theory; *V*. Webster's method ($P_n=2$). (c)——order clustering results of 5 methods ($K=15$): *I*, *II*, *III*, *V*. as (b); *IV*. order hierarchical clustering method ($D_1=5$, $D_2=1$)

表 1 各种有序集群结果的误识率 $\hat{\varepsilon}$

Table 1 Errors ($\hat{\varepsilon}$) of various order clust ering results (k——the number of groups)

No.	方法 \ K	$\hat{\varepsilon}$															
		7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	非参数拟合优度	0.146			0.128			0.086		0.087		0.087					0.061
2	Fisher 方法	0.089	0.089	0.089	0.100	0.092	0.092	0.092	0.087	0.092	0.100	0.087	0.087	0.087	0.087	0.085	0.085
3	图论方法		0.139			0.105	0.105	0.079		0.085	0.059						0.061
4	有序点群分析 ($D_1=5$, $D_2=1$)	0.102	0.113	0.113	0.095	0.095	0.095	0.095	0.095	0.095	0.095	0.082	0.082	0.082	0.074	0.074	0.074
5	对分移动窗口 ($P_g=2$)	0.163	0.163	0.126	0.100	0.100	0.100	0.111	0.111	0.111	0.111	0.111	0.079	0.066	0.072	0.072	0.059

表 2 非参数拟合优度法的迭代次数

Table 2 Number of iterations for nonparametric goodness of fit

外层迭代次数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
内层迭代次数	1	1	2	2	2	4	5	7	5	6	2	4	5	8	5	3	3
分类数 K	49	49	48	47	46	40	34	27	22	17	15	13	10	7	5	5	4

图 3(c) 分别示出了分类数 $K=10$ 和 $K=15$ 时的结果. 由图可见，各种方法所得结果比较一致. 当把全部样品分为 10 类时 ($K=10$) 就能较好地模拟地震交替的高潮和低潮活动的现象[图 3(b)]: 大多数情况下，图 3(b) 中以危险段居优势的一类或两类联成的大组与图 3(a) 中三个地震的显著活动期相对应，而以不危险段居优势的类所联成的大组则分别与三个相对平静期相对应. 当 $K=15$ 时，亦有类似的结果[图 3(c)]. 本结果与文献[3]的结果也是一致的.

(2) 表 1 中列出了当分类数 K 为 7 至 22 时各种方法所得结果的误识率 $\hat{\varepsilon}$. 由表可见，用 Fisher 方法分类的 $\hat{\varepsilon}$ 是比较低的，当 $K \geqslant 7$ 时 $\hat{\varepsilon}$ 就比较稳定，一般在 0.09 左右或小于 0.09. 而拟合优度法的 $\hat{\varepsilon}$ 则比 Fisher 法的 $\hat{\varepsilon}$ 略高一些，当 $K \geqslant 13$ 时，才有 $\hat{\varepsilon} < 0.09$. 但总的说来相差不大. 另外需要指出，拟合优度方法的第一类错误率(漏报强震)通常比 Fisher 法为低.

(3) 拟合优度法能够指出分类的相对稳定状态，即帮助确定较为合理的类数. 表 2 列出了分类数及其对应的内、外层迭代次数. 内层迭代次数表示从上一状态对应的分类数转到本状态所对应的分类数时所需要迭代的次数. 该数愈大则上一状态愈稳定，即上

一个分类数较为合理. 由表 2 可见，较大的内层迭代次数分别为 8，7，6，表明分类数为 10，34，22 都是比较合理的. 还可看到，对较大和较小的 K，内层迭代次数都较小，故 K 为中间一段数值时分类结果较为合理些. 考虑到这一点，表 1 中仅列出 K 为 7 至 22 间的 $\hat{\varepsilon}$ 以供比较，图 3 中举例示出 K 分别为 10 和 15 时的集群结果. 图中“D 段分为 D 类”和“N 段分为 N 类”表示正确分类，而另两种情况表示错分.

3. 图论方法、有序点群分析法和对分移动窗口法的结果

(1) 有序点群分析中以 $D_1=5$ 且 $D_2=1$ 的结果最好. 前已叙及，有序点群分析中有两个关于距离的参数：类间距离 $d(\omega_i, \omega_j)$ 和点间距离 $d(l, m)$. 它们均可采用不同的距离. 本文中以 D_2 和 D_1 的不同值分别表示取不同的点间距离和不同的类间距离：

$D_2=1$	欧氏距离	$D_2=2$	相似距离
$D_1=1$	最短距离	$D_1=2$	最长距离
$D_1=3$	平均距离	$D_1=4$	重心距离
$D_1=5$	离差平方和增量	$D_1=6$	间隙距离

表 3 中列出了取不同的距离时分类的误识率 $\hat{\varepsilon}$. 由表可见，当 $D_1=5$ 且 $D_2=1$ 时的结果最好，其中 $K\geqslant 10$ 后 $\hat{\varepsilon}$ 就稳定地 $\leqslant 0.95$，而且一般看来 $\hat{\varepsilon}$ 比取其他距离时为小；当 $D_1=2$ 且 $D_2=2$ 时所得结果次之，其 $\hat{\varepsilon}$ 一般看来也较小，且 $K\geqslant 12$ 后就有 $\hat{\varepsilon}\leqslant 0.92$；当 $D_1=6$ 且 $D_2=1$ 时 $\hat{\varepsilon}$ 最大，结果最差. 取其他距离时的结果居于其间.

前已叙及，$D_1=5$ 且 $D_2=1$ 表示当离差平方和增量最小时则合并两类，且取点间距离为欧氏距离. 这与 Fisher 的方法是比较接近的. 由表 1 可见，两者的 $\hat{\varepsilon}$ 也是差不多的.

$D_1=6$ 且 $D_2=1$ 表示合并两类的原则为间隙距离最短，且点间取欧氏距离. 此结果的 $\hat{\varepsilon}$ 最大表明，对本组样品而言，某类最后一个样品和下一类的第一个样品之间的距离不足够大，这可能正反映了地震活动的高潮期和低潮期不是截然分开的，即由相对平静期进入显著活动期时，地震活动是逐步增强的，反之地震活动也是逐步减弱的.

(2) 对分移动窗口法中以 $P_n=2$ 的结果略好. 表 4 中列出了 P_n（窗口的半长）取不同值时的 $\hat{\varepsilon}$. 由表可见，$P_n=1$ 时 $\hat{\varepsilon}$ 较大，而 $P_n=2$ 和 $P_n=3$ 的 $\hat{\varepsilon}$ 则相差不多. 由图 3(a) 可见，三个地震活动高潮期分别包括 4，14 和 5 个时间段，而 $P_n=2$ 和 3 时的窗口长度分别为 4 和 6，故能较好地模拟地震交替的高潮、低潮活动. 由于这里的窗口是移动选取的，因而取合适的窗口长度就能更好地反映某些较长时期的地震活动高潮(低潮)期.

(3) 将对分移动窗口(当 $P_n=2$ 时)、有序点群分析 ($D_1=5$ 且 $D_2=1$) 和图论方法的 $\hat{\varepsilon}$ 与 Fisher 方法的 $\hat{\varepsilon}$ 比较 (见表 1)，前三种方法的 $\hat{\varepsilon}$ 相差不多，其中对分移动窗口法的 $\hat{\varepsilon}$ 略高一些. 总的说来还是 Fisher 法的 $\hat{\varepsilon}$ 较小且较稳定. 较之其他三种方法，它从较小的 K 起就较稳定地逐步减小.

(4) 前已叙及，这三种方法同样能模拟地震活动交替的高潮、低潮现象. 图 3(c) 显示，这三种方法都是以较少的类数 (4 至 5 类)就模拟了三个相对活动期，其中对分移动窗口法较明显，它分别以第 5 类和第 10 类与前两个显著活动期对应，以第 14 类和第 15 类联合的一个大组与第三个显著活动期对应. 这一结果表明，这三种方法可以较好地对危

表 3 有序点群分析结果的误识率 $\hat{\varepsilon}$

Table 3 The error $\hat{\varepsilon}$ of order hierarchical clustering results

D_1	D_2	$\hat{\varepsilon}$ / K															
		7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	1	0.170	0.170	0.170	0.163	0.149	0.149	0.149	0.149	0.149	0.149	0.149	0.098	0.098	0.061	0.061	0.048
2	1	0.134	0.154	0.154	0.141	0.141	0.152	0.152	0.152	0.120	0.120	0.087	0.087	0.082	0.069	0.056	0.056
3	1	0.105	0.105	0.105	0.105	0.105	0.105	0.105	0.105	0.105	0.105	0.105	0.105	0.105	0.176	0.176	0.163
4	1	0.189	0.170	0.170	0.170	0.163	0.163	0.163	0.124	0.072	0.087	0.087	0.087	0.074	0.074	0.054	0.067
5	1	0.102	0.113	0.113	0.095	0.095	0.095	0.095	0.095	0.095	0.095	0.082	0.082	0.082	0.074	0.074	0.074
6	1	0.189	0.189	0.189	0.176	0.157	0.149	0.149	0.149	0.144	0.144	0.144	0.092	0.092	0.092	0.092	0.058
1	2	0.128	0.118	0.118	0.145	0.145	0.145	0.145	0.135	0.135	0.128	0.128	0.126	0.126	0.126	0.061	0.061
2	2	0.133	0.133	0.133	0.120	0.120	0.082	0.082	0.087	0.087	0.087	0.087	0.092	0.092	0.092	0.092	0.085
3	2	0.141	0.102	0.130	0.122	0.115	0.120	0.120	0.120	0.120	0.082	0.082	0.074	0.074	0.074	0.074	0.074

(D_1——the between-class distance; D_2——the between-sample distance; K——the number of groups)

表 4 对分窗口法结果的误识率 $\hat{\varepsilon}$

Table 4 The error $\hat{\varepsilon}$ of results of Webster's method

P_n	$\hat{\varepsilon}$ / K	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1		0.126	0.137	0.137	0.137	0.137	0.137	0.137	0.124	0.111	0.085	0.090	0.090	0.090	0.090	0.090	0.090
2		0.163	0.163	0.126	0.100	0.100	0.100	0.111	0.111	0.111	0.111	0.111	0.079	0.066	0.072	0.072	0.059
3		0.152	0.128	0.128	0.128	0.128	0.089	0.082	0.082	0.082	0.087	0.087	0.087	0.087	0.092	0.080	0.072

(K——the number of groups; P_n——the half length of window)

险样品 (D 类样品) 集群. 这三种方法对于相对平静期中的样品分类都较细也反映了这一事实.

对宁夏地区的资料，各种方法的集群结果与对华北地区资料的结果类似，不一一列举.

4. 对于强震孕震过程及其规律的讨论

综合上述各项结果,似可得到以下结论:

(1) 对各种有序集群结果的比较表明，以 Fisher 方法和采用离差平方和增量作为类间距离的有序点群分析效果较好. 这一事实说明，离差平方和在对地震时间段进行识别时是一项有效的工具. 或者说,在各个类中，诸时间段的结构主要是团状分布 (即围绕某一中心大体均匀地分布在一近于超椭球的区域内) 而不是链状分布. 这一点将为今后选择合理的预报方法提供依据.

(2)对分窗口法和拟合优度法(它也是一种谷值寻找方法)的结果说明,在相对平静期和显著活动期之间确实存在着客观的谷值. 在合理地选择方法的基础之上，及时而正确

地识别这些谷值，对于进行预报将是有意义的．饶有兴味的是，除去两种不同活动期的界限以外，前一方法对相对平静期中的谷值比较敏感而后者则对显著活动期内的谷值敏感．因此我们设想，这两种方法将可互相补充或校正．

(3) 与(2)不同，图论方法基本上是一种寻峰方法．这一方法的误识率也较低，表明在各类(主要是显著活动期)中同样有峰值，即在显著活动期中还可细分出易发生强震的时段．

(4) 上述三方面认识综合说明，强震孕震期间两类时段的分布状况是比较规则的，因而识别和预报是可能的．为了正确地进行预报，除去应该综合利用各种方法、精确地区分峰值与谷值外，还可根据对于两类错误代价的不同规定，选用适当的方法．

五、结　论

本文研究了华北地区和宁夏地区的资料，对特征进行了平滑，并对比了五种有序集群方法的结果，得到：

(1) 对特征进行平滑可以改善结果．当不同时期资料情况相差较大时（如宁夏地区的资料），对结果的改善较为明显．

(2) 五种方法均能模拟地震活动交替的高潮、低潮现象．

(3) Fisher 方法和采用离差平方和增量作为类间距离的有序点群分析方法的效果均较好，表明在各类中诸时间段的结构主要是团状分布而不是链状分布．选择预报方法时考虑地震活动的这一特点是重要的．

(4) 对分窗口法和拟合优度法的结果表明，在相对平静期和显著活动期之间确实存在着客观的谷值，且两种方法分别对平静期和活动期内的谷值较为敏感，可互为补充．

(5) 寻峰的图论方法结果表明，在显著活动期中同样有峰值，它有助于细分出其中易发生强震的时段．

(6) 5 种方法中以 Fisher 方法的误识率 $\hat{\varepsilon}$ 最低，而拟合优度法能指出分类的相对稳定状态，即帮助确定较为稳定的分类．

综上所述，加强资料的预处理(如平滑)可以改善结果，而根据强震孕震过程的规律选择合适的方法或综合多种方法，则有助于准确地进行预报．

参　考　文　献

[1] 王碧泉、杨锦英、王春珍，大震前地震活动的图象识别，地震学报，**4**，105—115，1982.

[2] 王碧泉、马秀芳，用模式识别方法研究强震发生的动力因子，地震学报，**5**，257—267，1983.

[3] 王碧泉、陈祖荫、王春珍，用聚类分析法研究强震的孕震过程，地震学报，**6**，121—128，1984.

[4] 陈祖荫、王碧泉、范　丰、王　斌，强震孕震过程中的特征选择，地球物理学报 **28**，588—598，1985.

[5] Wang Biquan, Chen Zuyin and Ma Xiufang, A nonparametric clustering method for order Samples and it's application to the analysis of earthquake data, *IEEE* (in print).

[6] Keilis-Borok V. I. and V. G. Kosobokov, Diagnosis of the increase of probability of an earthquake with magnitude 8 or more, Workshop on Pattern Recognition and Analysis of Seismicity, 5—16 December 1983.

[7] Allen, C., K. Hutton, V. I. Keilis-Borok, I. V. Kuznetsov, and I. M. Rotwain, Selfsimilar long-term Premonitory Seismicity Patterns in California and W. Nevada, Workshop on Pattern Recognition and

Analysis of Seismicity, 5—16 December 1983.
[8] 中国科学院沈阳计算技术研究所等,电子计算机常用算法,科学出版社, 1976.
[9] 方开泰、潘恩沛,聚类分析,地质出版社, 1982.
[10] 方开泰,几个有序样品的聚类方法,应用数学学报, **5**, 94—101, 1982.
[11] 福永圭之介,统计图形识别导论,陶笃纯译,科学出版社, 1978.
[12] Koontz, W. L. G. and K. Fukunaga, A nonparametric Valley-Seeking technique for cluster analysis, *IEEE Trans. Computers*, **C-21**, 1972.
[13] 蒋代梅、陈祖荫、陈传涓,以拟合优度为基础的两个算法及其在癌细胞自动识别中的应用,北京工业大学学报, **8**, 79—84, 1982.
[14] Koontz, W. L. G., A graph-theoretic approach to nonparametric cluster analysis, *IEEE Trans. Computers*, **C-25**, 1976.

SEVERAL ORDER CLUSTERING METHODS FOR THE STUDY OF SEISMOGENIC PROCESS OF STRONG EARTHQUAKES

WANG BIQUAN[1] CHEN ZUYIN[2] TONG GUOBANG[3] AND WANG CHUNZHEN[1]

Abstract

In this paper, the seismogenic process of strong earthquakes of North China and the Ningxia region is studied by using five order clustering methods. We found that all five methods can simulate the phenomena of alternatively high and low seimic activities; the better effects of both Fisher's method and order hierarchical clustering method (using incremental sum of squares of diviations as the between-class distance) show that the structures of time intervals in classes are mainly of cluster distribution and are not of chain distribution, thus it is important that this character of seismic activity should be considered in choosing the method of prediction. The results of Webster's method and nonparametric goodness of fit show that there are certainly objective valleyes between relatively queit periods and remarkably active periods. Then graph theory further helps us to distinguish the smaller peaks in which strong earthquakes are apt to occurr.

The error of Fisher's method is smallest among these five methods, while nonparametric goodness of fit can help make more reasonable classification. In addition, presmoothing of data can also improve the results.

1) Institute of Geophysics, State Seismological Bureau.
2) Beijing Polytechnical University.
3) Institute of Hydrogeology and Engineering Geology, Ministry of Geology and Mineral Resources, PRC.