第10卷 第2期 西北地震学报 Vol.10,No.2
1988年6月 NORTHWESTERN SEISMOLOGICAL JOURNAL June,1988

系统聚类法在华北地区的地震分期和预测中的应用*

马淑田 王碧泉 王玉秀 杨锦英
（国家地震局地球物理研究所）
陈祖荫
（北京工业大学）

摘 要

本文用修改了的系统聚类法研究华北地区地震活动期，结果表明，当类间距离用离差平方和增量距离，且样品间用欧氏距离时，能较好地划分该地区的地震活动期。而样品间用相似系数时，不论类间距离是哪一种，所得结果均较差。本文还作了地震活动期的后验预测尝试，还通过逐次筛选方法对21个特征进行了选择。

一、引 言

国内外许多地震学者认为，在某些特定的地震区内，地震的活动有某种周期性，即地震活动的高低潮交替进行，因此许多人做了地震活动的分期工作，由于采用的方法和原则不同，即便对同一地震区，所得结果也不一定相同。王碧泉等[1]用模式识别中的有序集群方法，较定量地给出了对华北地区地震的分期，本文也是用有序集群方法，使用了更多的资料，试图对华北地区的地震进行分期和后验预测。

二、系统聚类方法

1.用以聚类的几种距离

（1）点间距离

m维空间中的两个样品点对应于两个矢量$\vec{V}_1$和$\vec{V}_2$，它们的坐标分别是(X_{11},X_{12}，…，X_{1m}）和（X_{21}，X_{22}，…，X_{2m}），则两样品点的相似距离和欧氏距离定义为：

相似距离（记为$ID_2=2$）

---

*地震学联合科学基金资助的课题

$$d = 1 - (\vec{V}_1 \cdot \vec{V}_2) / \sqrt{|\vec{V}_1| \cdot |\vec{V}_2|}$$

$$= 1 - \sum_{i=1}^{m} X_{1i} X_{2i} / \left[\left(\sum_{i=1}^{m} X_{1i}^2 \right) \cdot \left(\sum_{i=1}^{m} X_{2i}^2 \right) \right]^{1/2}$$

欧氏距离（记为$ID_2 = 1$）

$$d = |\vec{V}_1 - \vec{V}_2| = \left[\sum_{i=1}^{m} (X_{1i} - X_{2i})^2 \right]^{1/2}$$

（2）类间距离

若干个样品按某种分法被分成几类，两类间的距离为类间距离，类间距离有不同的定义。

最短距离（记为$ID_1 = 1$）：设有两类样品G_1（n_1个样品）和G_3（n_2个样品），G_1中的样品点与G_2中的样品点之间共有$n_1 \cdot n_2$个距离，在这$n_1 \cdot n_2$个距离中取最短的一个定义为G_1和G_2之间的距离。

最长距离（记为$ID_1 = 2$）：类似于最短距离，定义两类样品点间距离中最长者为类间距离。

平均距离（记为$ID_1 = 3$）：

$$d(G_1, G_2) = \left[\sum_{K=1}^{n_1 \cdot n_2} d_K(\vec{X}_i, \vec{X}_j) \right] / (n_1 \cdot n_2)$$

其中$d_K = |\vec{X}_i - \vec{X}_j|$（$i = 1, \cdots, n_1$，$j = 1, \cdots, n_2$）为两样品点的距离。$\vec{X}_i$、$\vec{X}_j$分别为$G_1$和$G_2$中的样品。

重心距离（记为$ID_1 = 4$）：分别求出两类样品各自的重心，再求两重心点间的距离。

离差平方和增量（记为$ID_1 = 5$）：设样品在G_1、G_2以及它们的并G_{12}中的离差平方和分别为：

$$S_1 = \sum_{i=1}^{n_1} \sum_{k_1=1}^{m} (X_{ik_1} - X_{k_1})^2$$

$$S_2 = \sum_{i=1}^{n_2} \sum_{k_2=1}^{m} (X_{ik_2} - X_{k_2})^2$$

$$S_{12} = \sum_{i=1}^{n_1+n_2} \sum_{k=1}^{m} (X_{ik} - X_k)^2$$

其中 $X_{k_1} = \sum_{i=1}^{n_1} X_{ik_1}/n_1$；$X_{k_2} = \sum_{i=1}^{n_2} X_{ik_2}/n_2$；$X_k = \sum_{i=1}^{n_1+n_2} X_{ik}/(n_1+n_2)$

（$K_1, K_2, \cdots, K = 1, 2, \cdots, m$）。

定义$d(G_1, G_2) = S_{12} - S_1 - S_2$为两类样品的离差平方和增量距离。

间隙距离（记为$ID_1 = 6$）：在有序样品分析中，相邻两类样品中前一类的最后一个样品点与后一类第一个样品点间的距离，定义为两类样品间的间隙距离。

2，聚类的方法[4、5]

本文使用修改了的“系统聚类法”，即“有序点群分析法”。其步骤是：

（1）各样品自成一类（这时有N=195类）。

（2）计算相邻各样品之间的距离（按以上介绍的某种定义），并将距离最近的两类并成一类，在并类时要求各样品的顺序号不变。因此在计算距离时只需计算顺序号相邻的两个样品的距离，而不必计算顺序号相间的各样品间的距离。

（3）计算相邻新类间的距离，再将距离最近的两类合并。同样，只需计算相邻两类间的距离。按这种顺序做下去，直到所有的样品归并为事先指定的类数为止。

三、特征的逐次筛选

本文使用的资料为华北地震活动区（107.5°—125.0°E；29.0°—44.0°N）1401年至1985年的地震目录，只取M≥5.0级的主震，有关资料详情见文献〔2〕。

本文以每三年为一段，即为一个样品。样品数用N表示，从1401年至1985年，当取3年为一时间段时，N=195。

每个样品用m个和它有关的因素（或称变量）来描述，这m种因素被称为m个特征。所有样品及其特征用一个m行N列的矩阵来描述，矩阵的每一列为一个m维矢量。在本文中有195个样品，即195个m维空间中的点。王碧泉等在文献〔2〕中研究了一组自相似地震活动函数，按此文献，我们取了21个特征，这些特征见表1。

表1

特征X_i	符号	特征内容
X_1	N	地震频度
X_2	E_1	地震能量
X_3	E_2	地震蠕变
X_4	E_3	震源面积
X_5	S_1	平均能量
X_6	S_2	平均蠕变
X_7	S_3	平均震源面积
X_8	F_1	中期地震活动
X_9	F_2	长期地震活动
X_{10}	F_3	超长期地震活动
X_{11}	D_{01}	短期相对于中期地震活动的偏离
X_{12}	D_{02}	短期相对于长期地震活动的偏离
X_{13}	D_{12}	中期相对于长期地震活动的偏离
X_{14}	D_{03}	短期相对于超长期地震活动的偏离
X_{15}	D_{13}	中期相对于超长期地震活动的偏离
X_{16}	D_{23}	长期相对于超长期地震活动的偏离
X_{17}	DMA	相对于极值的变化
X_{18}	Q	地震平静
X_{19}	A	地震活化
X_{20}	b	短期b值
X_{21}	b_M	中期b值

考虑到各特征的量纲不一致，会对结果产生不好的作用，所以首先将它们作了标准差标准化。由于21个特征〔2〕是自相似的，即它们是非独立的，因此，这些特征中有些可能对分类起的作用大些，有的小些，有的可能不起作用，而有的可能起相反的作用，所以我们试图从21个特征中选出一些对分类起较大作用的特征。

（1）在21个特征中依次去掉一个特征，保留其余20个进行运算，得到21个计算结果，并按文献〔4〕中计算误识率的公式计算，作出误识率随分类数的变化图。分析这些图形，选出误识率最小的图形，则该图所相应的特征被选出。例如当第5个特征不参加运算而用其余20个特征计算时得到的结果较好，误识率最小，则将除第5个特征以外的20个保留下来。

（2）从第一步得到的20个特征中依次取19个特征进行运算，选出误识率最小的19个特征。

（3）如此做下去，当误识率达某一数值时停止筛选，如此选出了14个特征。

四、结果与讨论

1.取21个特征，用前述聚类方法对195个样品聚类，比较分析结果认为 $ID_1 = 5$，$ID_2 = 1$ 时误识率较小，与实际的地震活动期较符合。除此以外，误识率有的较大，有的虽不大，但分类不稳定，与实际相差较大。一般来说当 $ID_2 = 2$ 时结果较差。由于结果的数据较多，不在本文列出，这里仅给出 $ID_1 = 5$，$ID_2 = 1$ 的部分分类结果。

2.由于 $ID_1 = 5$，$ID_2 = 1$ 组合时所得的分类结果比较稳定，图 1 示出了取全部21个特征时（m = 21）误识率 $\hat{\varepsilon}$ 随分类数K的变化。由图可见，除K = 10外，误识率都大于1.0。然后，按上述的逐次筛选方法，对这21个特征进行逐次筛选，最后选出了14个特征，它们是 X_2、X_3、X_5、X_6、X_7、X_8、X_9、X_{11}、X_{13}、X_{14}、X_{16}、X_{17}、X_{20}、X_{21}。这时的误识率 $\hat{\varepsilon}$ 随K的变化如图 1 所示，由图可见，当K大于13时，$\hat{\varepsilon}$ 均小于0.1，这样的误识率是比较小的（表 2 ）。因此从误识率的角度看其结果是可供参考的。

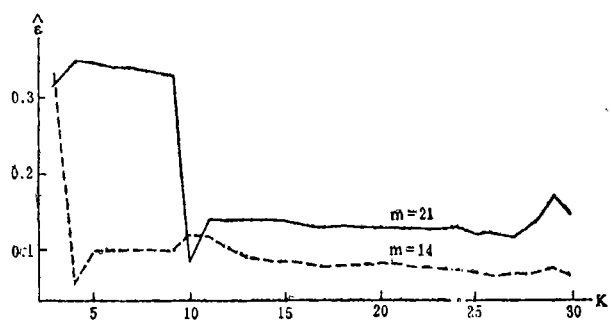

图 1　误识率随分类数K的变化图

Fig. 1 Error rates change with classification number

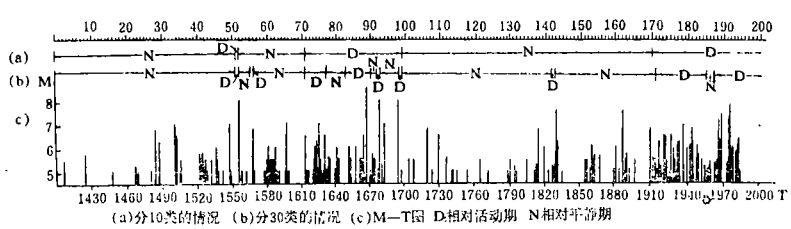

图 2　地震活动分期图

Fig. 2 Diagrams of divided seismic active periods

图 2、表 3、表 4 给出对地震活动的分期情况，从样品被分成 10 类（表 3 ）的情况来看，1401年至今可分为 3 个相对平静期和 3 个相对活动期；当分成30类时（表 4 ），有 9 个相对平静期和 9 个相对活动期，而且大的相对平静期中有小的相对活动期，反之亦然。从图 2（b）可以看出，本文给出的地震活动期与实际的地震活动期基本一致，因此笔者认为在使用适当的特征时，用系统聚类法对地震活动的分期结果是有意义的。

3.我们利用195个样品中的前193个样品进行分类，对最后两个样品（1980—1985年）进行预测。其方法是先计算出各类的重心及最后两个样品的重心，然后计算被预测类的重心与

表2

$\hat{\varepsilon}$ m / k	21	14
3	0.313	0.322
4	0.340	0.079
5	0.336	0.097
6	0.332	0.097
7	0.332	0.097
8	0.328	0.096
9	0.324	0.096
10	0.088	0.116
11	0.139	0.116
12	0.135	0.101
13	0.135	0.087
14	0.135	0.083
15	0.135	0.083
16	0.131	0.083
17	0.127	0.079
18	0.127	0.079
19	0.127	0.079
20	0.127	0.079
21	0.127	0.079
22	0.123	0.072
23	0.123	0.072
24	0.123	0.070
25	0.119	0.070
26	0.115	0.063
27	0.111	0.063
28	0.176	0.063
29	0.161	0.073
30	0.143	0.061
31	0.143	0.061
32	0.139	0.072

表3

类号	时 期	性质*
1	1401—1553	N
2	1554—1556	D
3	1557—1613	N
4	1614—1625	D
5	1626—1649	D
6	1650—1667	D
7	1668—1670	D
8	1671—1697	D
9	1698—1910	N
10	1911—1985	D

各类的重心点间的欧氏距离，比较这些距离的大小，如果被预测类与某类的重心距离近，则认为与该类是同类。表5给出了被分为10类和30类时的预测结果。当被分成10类时，被预测类与第10类最近，即被预测为D类；当被分成30类时与第26类最近，也被预测为D类，这与直接用195个样品分类时所得的结果是一致的。

根据计算结果，发现 $ID_2 = 2$ 与 ID_1 的任何值结合时所得的结果均较差，有的误识率较大。有的误识率虽不大，如 $ID_1 = 5$，$ID_2 = 2$ 时误识率可以小到0.04，可是所分的结果与实际情况相差较远，分30类时竟然把1415—1910年划分为N类，这可能是因为用相似系数需忽略矢量长度，只考虑其方向，而我们的资料不能忽略矢量长度的缘故。

有的时候并不是特征越多，对分类越有利。本文结果表明用21个特征所得的分类结果不如用筛选后的14个特征所得的分类结果好，这表明进一步作特征选择是必要的。

本文所用的资料，前一时期和后一时期的精度不一致，这就难免会影响分类结果，

* 发生一次以上M>6.0级地震的时间段为D类样品，否则为N类。全部195个样品中有48个D类样品，其余为N类样品，D类的先验概率为48/195。在某类中，若D类的概率>48/195，则该类性质为D，否则为N。

表4

类　号	时　期	性　质	类　号	时　期	性　质	类　号	时　期	性　质
1	1401—1466	N	11	1635—1649	N	21	1827—1829	D
2	1467—1553	N	12	1650—1667	D	22	1830—1853	N
3	1554—1556	D	13	1668—1670	D	23	1854—1856	N
4	1557—1565	N	14	1671—1676	N	24	1857—1910	N
5	1566—1568	D	15	1677—1679	D	25	1911—1913	N
6	1569—1613	N	16	1680—1694	N	26	1914—1955	D
7	1614—1622	D	17	1695—1697	D	27	1956—1961	N
8	1623—1625	D	18	1698—1748	N	28	1962—1973	D
9	1626—1631	D	19	1749—1784	N	29	1974—1976	D
10	1632—1634	N	20	1785—1826	N	30	1977—1985	D

预测结果　　表5

10　类			30　类								
类号	性质	欧氏距离	类号	性质	欧氏距离	类号	性质	欧氏距离	类号	性质	欧氏距离
1	N	5.051	1	N	5.488	11	N	4.882	21	D	3.444
2	D	9.901	2	N	4.799	12	D	4.399	22	N	5.379
3	N	4.075	3	D	9.901	13	D	27.360	23	N	5.966
4	D	3.840	4	N	4.992	14	N	2.942	24	N	4.910
5	D	3.235	5	D	4.132	15	D	6.199	25	N	5.294
6	D	4.399	6	N	4.129	16	N	4.796	26	D	2.013
7	D	27.360	7	D	4.155	17	D	10.371	27	N	2.027
8	D	3.945	8	D	6.258	18	N	5.837	28	D	4.326
9	N	5.438	9	D	3.161	19	N	5.775	29	D	6.562
10	D	1.807	10	N	6.633	20	N	5.250	30	D	2.434

因此需要对资料进行预处理。

虽然象1556年、1668年那样的大地震也能分出来，但它们所在类的时间太短，仅几年或十几年，危险期不可能这样短，这是有待研究和解决的问题。

（本文1986年10月25日收到）

参　考　文　献

〔1〕王碧泉、陈祖荫、童国榜、王春珍，研究强震孕震过程的几种有序集群方法，地震学报，待出版.
〔2〕王碧泉、杨锦英、王玉秀、陈锦标，自相似地震活动特征的提取，待出版.
〔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，Vol. 6，No. 2，1984.
〔4〕Wishart, W., An algorithm for hierarchical classification, Biometrics, Vol.25, No.1, 165—170, 1969.
〔5〕方开泰、潘思沛，聚类分析，地质出版社，1982.

DIVIDING SEISMIC ACTIVE PERIODS WITH MODIFIED HIERARCHICAL CLASSIFICATION METHOD

Ma Shutian, Wang Biquan, Wang Yuxiu and Yang Jinying
(*Institute of Geophysics, State Seismological Bureau*)
Chen Zuyin
(*Beijing Polytechnical University*)

Abstract

The Modified Hierarchical classification Method is used to divide seismic active periods of Northern China in this paper. The seismic active periods divided are better when the distances between classes are increment of square sum of deviation and the distances between samples are Euclidean Distance than that when the distances between classes are any kind of method mentioned in the paper and the distances between samples are resemblance coefficient. The methods of predicting seismic active peroids and selecting features are proposed, and they are used as well. After features are selected by our method the classification results can be improved and the posterior prediction made are consistent with actual state basically.