



(四) 极差分割法

张 强

从这一讲开始，要比较系统地介绍一些较新的方法。从数学上说，它们是数理统计的多元分析中的分支；从气象上说，它们与“相似方法”非常接近。因此，每一种方法分三讲：先讲一种简便的手算的方法；第二次讲比较复杂一些的计算方法；第三次讲这种方法的一些灵活应用。没有条件进行较复杂计算的，可以跳过第二次讲的内容，直接看第三次。

气象上的相似方法是以过去已有的历史资料为基础，看今年的情况与历史上哪一年最接近，然后就按照最接近的年份来作预报。多元分析中有一些“聚类”的方法，就是把历史资料依资料本身的情况进行分类，这些“聚类”的方法在地质、生物分类、医疗诊断、……等方面都有广泛的应用。在气象上也可以用这种聚类的方法将历史资料进行分类，然后在分类的基础上去判别今年的情况与过去哪些年份组成的类最接近——相似，或者都不接近，今年是历史资料中没有出现过的新情况，这样就可以作出预报。由此可见，“聚类分析”的方法实质上是把相似方法从数学上更系统化、更定量化的一种数学工具。

分割法是聚类分析中的一个方法，下面用一个很简短的例子来说明。例如从1961—1970年上一年的太阳黑子的相对数如下表：

表 4.1

年 份	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
太阳黑子相对数	112	54	38	28	10	15	47	94	106	106

从这组资料本身来看，太阳黑子的多少是否有阶段性？如果我们不考虑时间的先后顺序，那就应将1961年、1969年和1970年归为一类，1962年与1967年归为一类。但这样归类就破坏了时间的顺序，看不出变化的阶段性。分割法所处理的问题就是这样一类与顺序有关的分类问题。它是将一长串资料分割成几段，即使只有10年数据，它的全部可能的分割方法就有512种，例如把这10年分成两段，就可以有9种分法：

$$\begin{aligned}
 (61) \quad & (62, 63, 64, 65, 66, 67, 68, 69, 70) \\
 (61, 62) \quad & (63, 64, 65, 66, 67, 68, 69, 70) \\
 & \vdots \\
 (61, 62, 63, 64, 65, 66, 67, 68, 69) \quad & (70)
 \end{aligned}$$

要把这10年分成三段，共有36种分法，全部写下来再逐个加以比较，哪一个分法是比较好的，就很烦了。实际上，究竟分成几段才合适，事先往往是不能预计的，要考察所有各种分割法中最好的一种，就需要比较。这样庞大的计算量，不仅手算不可能，使用电子计算机计算也有困难。下面我们介绍的极差分割法，明显地降低了计算量，使得手算也可以进行。

从前面三讲的介绍我们已经知道，描述一组数据内部不整齐的程度可以用这组数据的方差或标准差，这两个量手算是比较麻烦的，因此常常用极差来代替它。一组数据 x_1, \dots, x_n 的极差就是这组数据中的最大值与最小值之差，在统计上常用 R 表示极差，于是有

$$R = \max_{1 \leq i \leq n} x_i - \min_{1 \leq i \leq n} x_i \quad [R = (x_1, \dots, x_n \text{中最大值}) \text{ 减去 } (x_1, \dots, x_n \text{中最小值})]$$

例如表4.1中黑子相对数十个数据的极差 $R = 112 - 10 = 102$ ，头五年的极差是 $112 - 10 = 102$ ，后五年的极差是 $106 - 15 = 91$ ，1963—1967年这五年的极差是 $47 - 10 = 37$ ，……等等。很明显，一组数据的极差是很容易计算的，它的确也反映了这一组数据内部不整齐的程度。如果一组数据的极差是0（最大值与最小值一样），这组数据就都是同样的值。今后在各种手算的方法中，极差是常常要遇到的。

为了不使读者陷入冗长的计算，我们就只用1961—1966年这六个数据来介绍极差分割法。为了今后讨论方便，我们引入一些记号和概念：

1. 把 n 个有顺序的数据 x_1, \dots, x_n 分割成 k 段的一种分法，我们称它为对 x_1, \dots, x_n 的一个 k 分割，或简称为 k 分割；

2. 用符号 $S_n(k | \alpha_1, \alpha_2, \dots, \alpha_{k-1})$ 表示对 x_1, \dots, x_n 的下列这种 k 分割:

$(x_1, \dots, x_{\alpha_1})(x_{\alpha_1+1}, \dots, x_{\alpha_2})(x_{\alpha_2+1}, \dots, x_{\alpha_3}) \dots (x_{\alpha_{k-1}+1}, \dots, x_n)$ 。符号中 S 表示分割; n 标明是对 n 个数据的分割, 括号中第一个数 k 表示是一种 k 分割; 括号中竖线以后的 $k-1$ 个号码 $\alpha_1, \dots, \alpha_{k-1}$ 标明这种 k 分割的分法是:

- 第一段 从 x_1 起到 x_{α_1} 为止, α_1 是第一段最后一个数据的下标;
- 第二段 从 x_{α_1+1} 起到 x_{α_2} 为止, α_2 是第二段最后一个数据的下标;
- 第三段 从 x_{α_2+1} 起到 x_{α_3} 为止, α_3 是第三段最后一个数据的下标;
- ⋮
- 第 k 段 从 $x_{\alpha_{k-1}+1}$ 起到 x_n 为止。

因为第一段起始的一定是 x_1 , 最末一段的最后一个一定是 x_n , 所以毋需将它们再标出。例如

- $S_{10}(2 | 3)$ 就是 $(x_1, x_2, x_3) (x_4, x_5, x_6, \dots, x_{10})$;
- $S_{10}(3 | 6, 8)$ 就是 $(x_1, x_2, x_3, x_4, x_5, x_6) (x_7, x_8), (x_9, x_{10})$;
- $S_{10}(4 | 2, 3, 5)$ 就是 $(x_1, x_2) (x_3) (x_4, x_5) (x_6, x_7, x_8, x_9, x_{10})$;

如此等等。

对于表 4.1 中 1961—1966 年这六个数据, 先考虑全部可能的 2 分割:

$S_6(2 | 1) (112) (54, 38, 28, 10, 15)$ $S_6(2 | 2) (112, 54) (38, 28, 10, 15)$

$S_6(2 | 3) (112, 54, 38) (28, 10, 15)$ $S_6(2 | 4) (112, 54, 38, 28) (10, 15)$

$S_6(2 | 5) (112, 54, 38, 28, 10) (15)$ 每一个 2 分割都把这六个数据分成两段 (即两组数据), 分别计算这两段的极差, 用 R_1, R_2 表示, 于是有: (注意一个数据构成一组时, 它的极差自然是 0)

	R_1	R_2	R_1, R_2 中最大值
$S_6(2 1)$	0	$54 - 10 = 44$	44 *
$S_6(2 2)$	$112 - 54 = 58$	$38 - 10 = 28$	58
$S_6(2 3)$	$112 - 38 = 74$	$28 - 10 = 18$	74
$S_6(2 4)$	$112 - 28 = 84$	$15 - 10 = 5$	84
$S_6(2 5)$	$112 - 10 = 102$	0	102

我们希望分段后, 每一段内部数据之间的差异越小越好, R_1, R_2 反映了分段后各段内部的差异情况, 因此 R_1, R_2 越小越好, 即它们之中的最大的越小越好, 可见最好的 2 分割是 $S_6(2 | 1)$, 即分成这样两段:

$(\overset{1961\text{年}}{112}) (\overset{1962\text{年}—1966\text{年}}{54, 38, 28, 10, 15})$ 这样就找到了最优的 2 分割。

如果要找最优的 3 分割, 当然也可以将全部可能的 3 分割都写下来, 求出各段的极差, 然后把三段极差中最大的一个 $\max R_i$ 作为判别分割好坏的依据, 找到使 $\max R_i$ 达到最小值的那个 3 分割, 这样就可以求出最优的 3 分割。但这样比较麻烦, 即使是六个数据, 全部可能的 3 分割就有十个。现在我们不直接去考察全部的 3 分割, 而是考察 1961—1965 年这五个数据的最优 2 分割, 1961—1964 年这四个数据的最优 2 分割, ……等等。计算过程如下:

	R_1	R_2	$\max R_i$
$S_5(2 1) (112) (54, 38, 28, 10)$	0	44	44 *
$S_5(2 2) (112, 54) (38, 28, 10)$	58	28	58
$S_5(2 3) (112, 54, 38) (28, 10)$	74	18	74
$S_5(2 4) (112, 54, 38, 28) (10)$	84	0	84 最优的 2 分割是 $S_5(2 1)$;

	R_1	R_2	$\max R_i$
$S_4(2 1) (112) (54, 38, 28)$	0	26	26 *
$S_4(2 2) (112, 54) (38, 28)$	58	10	58
$S_4(2 3) (112, 54, 38) (28)$	74	0	74 最优的 2 分割是 $S_4(2 1)$;

	R_1	R_2	$\max R_i$
$S_3(2 1) (112) (54, 38)$	0	16	16 *
$S_3(2 2) (112, 54) (38)$	58	0	58 最优的 2 分割是 $S_3(2 1)$ 。

于是只要比较三个 3 分割, 就是在 $x_1, x_2, x_3, x_4, x_5; x_1, x_2, x_3, x_4; x_1, x_2, x_3$ 的最优 2 分割 $S_5(2 | 1), S_4(2 | 1), S_3(2 | 1)$ 基础上所形成的 3 分割。为了方便, 我们把已求得的最优 2 分割分别简记为:

$S_5(2 | 1) \rightarrow S_5^*(2), S_4(2 | 1) \rightarrow S_4^*(2), S_3(2 | 1) \rightarrow S_3^*(2)$ 。于是要比较的六个数据的 3 分割是:

	R_1	R_2	R_3	$\max R_i$
$S_6(3 S_5^*(2))(112) (54, 38, 28, 10) (15)$	0	44	0	44
$S_6(3 S_4^*(2))(112) (54, 38, 28) (10, 15)$	0	26	5	26
$S_6(3 S_3^*(2))(112) (54, 38) (28, 10, 15)$	0	16	18	18 *

$$S_6(3 | S_2(2))(112)(54)(38, 28, 10, 15) \quad 0 \quad 0 \quad 28 \quad 28$$

[注意：对于二个数据的2分割自然只有一种，因此用 $S_2(2)$ 表示，而且它一定也是最优的，同理，对 k 个数据的 k 分割只有一种，用 $S_k(k)$ 表示]

这样就找到了最优3分割是 $S_6(3 | S_3^*(2))$ ，记它为 $S_6^*(3)$ ——六个数据的最优3分割。同理，可以求出 $S_5^*(3)$ ， $S_4^*(3)$ ：

	R_1	R_2	R_3	$\max R_i$
$S_5(3 S_4^*(2))(112)(54, 38, 28)(10)$	0	26	0	26
$S_5(3 S_3^*(2))(112)(54, 38)(28, 10)$	0	16	18	18 *
$S_5(3 S_2(2))(112)(54)(38, 28, 10)$	0	0	28	28
$S_4(3 S_3^*(2))(112)(54, 38)(28)$	0	16	0	16
$S_4(3 S_2(2))(112)(54)(38, 28)$	0	0	10	10 *

$S_5^*(3)$ 是 $S_5(3 | S_3^*(2))$ ，即 $S_5^*(3) = S_5(3 | 1, 3)$ ， $S_4^*(3)$ 是 $S_4(3 | S_2(2))$ ，即 $S_4^*(3) = S_4(3 | 1, 2)$ 。完全类似，在 $S_6^*(3)$ ， $S_4^*(3)$ 的基础上求最优的4分割：

	R_1	R_2	R_3	R_4	$\max R_i$
$S_6(4 S_5^*(3))(112)(54, 38)(28, 10)(15)$	0	16	18	0	18
$S_6(4 S_4^*(3))(112)(54)(38, 28)(10, 15)$	0	0	10	5	10 *
$S_6(4 S_3(3))(112)(54)(38)(28, 10, 15)$	0	0	0	18	18
$S_5(4 S_4^*(3))(112)(54)(38, 28)(10)$	0	0	10	0	10 *
$S_5(4 S_3(3))(112)(54)(38)(28, 10)$	0	0	0	18	18

所以 $S_6^*(4) = S_6(4 | S_4^*(3)) = S_6(4 | 1, 2, 4)$ $S_5^*(4) = S_5(4 | S_4^*(3)) = S_5(4 | 1, 2, 4)$ 。

于是最优的5分割亦可以求出。

	R_1	R_2	R_3	R_4	R_5	$\max R_i$
$S_6(5 S_5^*(4))(112)(54)(38, 28)(10)(15)$	0	0	10	0	0	10
$S_6(5 S_4(4))(112)(54)(38)(28)(10, 15)$	0	0	0	0	5	5 *

所以 $S_6^*(5) = S_6(5 | S_4(4)) = S_6(5 | 1, 2, 3, 4)$ 。对六个数据考虑6分割就是 $S_6(6)$ 只有一种分法，每个数据单独构成一段。这样我们就全部找到了 $S_6^*(2)$ ， $S_6^*(3)$ ， $S_6^*(4)$ ， $S_6^*(5)$ 。这种算法从表面上似乎多算了一些 $S_5^*(2)$ ， $S_4^*(2)$ ， $S_3^*(2)$ ， $S_5^*(3)$ ， $S_4^*(3)$ ， $S_5^*(4)$ 等，但实际上大大地省略了比较的分割个数。如果读者认真地跟着计算一遍，就会发现许多 R_i 是相同的，只要算一次，以后用到时抄一下就行了。总结一下我们算得的结果是：

	$\max R_i$
$S_6^*(2): (112)(54, 38, 28, 10, 15)$	44
$S_6^*(3): (112)(54, 38)(28, 10, 15)$	18
$S_6^*(4): (112)(54)(38, 28)(10, 15)$	10
$S_6^*(5): (112)(54)(38)(28)(10, 15)$	5

这样求得的结果与我们的直观上的感觉也是非常符合的。

现在进一步对这种方法作几点说明：

1. 从这六年数的黑子相对数来看，究竟分成几段才合适呢？

我们可以画出一张图，把 $\max R_i$ 随分割的段数而变化的情况反应出来。如果不分割，即全部六个数据是一组，此时极差只有一个，就是 $112 - 10 = 102$ ；全部六个数据分成六段，相应的每一段极差是0，最大值自然也是0。因此有

分割数（即分割成几段）	1	2	3	4	5	6
$\max R_i$	102	44	18	10	5	0

画出图5.1，

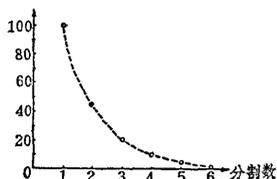


图 5.1

从图上看，曲线的转折点在分割数是3处，因此以分成三段比较合适。

2. 用这种方法找到的 $S_6^*(2)$ ， $S_6^*(3)$ ， $S_6^*(4)$ ， $S_6^*(5)$ 是不是能保证它的最优性？

从求的过程可以看到 $S_6^*(2)$ 一定是最优2分割，其余的 $S_6^*(3)$ ， $S_6^*(4)$ ， $S_6^*(5)$ 等不能保证它一定是最优的，但是可以有理由认为它们确实是比较好的。而且在下一讲中我们会看到：如果把极差改成方差，它们确实是最优的（下一讲在介绍算法的过程中就证明了最优性），所以现在也就称它们是最优分割。

3. 这种方法不只是处理一种气象要素随时间变化分段的方法，它可以同时处理很多气象要素。设第 1 年 S 种气象要素的值是 x_{1s} , $1=1, 2, \dots, n$; $S=1, 2, \dots, p$, 也即要考虑 n 年的资料, 每年有 p 个指标的值。例如考虑长江流域六站一月份的总的降水 $\Sigma R_{1月}$ (此处 $R_{1月}$ 指降水量); 一月份欧洲大型环流型 C 型的天数; 长江六站一月份总的平均温度 $\Sigma \bar{T}_{1月}$; 上一年太阳黑子的相对数这四个气象要素。以这四个指标的 10 年资料 (1961—1970 年) 为基础, 如何进行分割呢? 这时, 现有的资料就是下面的表 4.2:

表 4.2

年 份	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
$\Sigma R_{1月}$	209	109	3	423	93	264	168	260	473	149
C 型天数	16	9	25	14	17	3	14	17	6	6
$\Sigma \bar{T}_{1月}$	236	193	178	286	353	301	198	222	150	162
太阳黑子数	112	54	38	28	10	15	47	94	106	106

先将这些资料标准化, 即消除单位的影响, 使得每个气象要素的资料都在 0—1 这个范围内变化。用 x_{1s} 表示上面这些资料就得矩阵:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots \\ x_{101} & x_{102} & x_{103} & x_{104} \end{pmatrix} \begin{matrix} \rightarrow 61\text{年} \\ \rightarrow 62\text{年} \\ \vdots \\ \rightarrow 70\text{年} \end{matrix} \quad (10\text{年, 4个指标组成的资料})$$

$$\text{令 } y_{1s} = \frac{x_{1s} - \min_{1 \leq j \leq 10} x_{1j}}{\max_{1 \leq i \leq 10} x_{1s} - \min_{1 \leq i \leq 10} x_{1s}} \quad \begin{matrix} i = 1, 2, \dots, n \\ s = 1, 2, \dots, p \end{matrix} \quad \text{则 } y_{1s} \text{ 就是标准化以后的资料。}$$

算出 y_{1s} 后, 全部的计算过程和上面介绍的完全类似, 例如考虑 S_{10} (2 | 3) 相应的分割是:

$$\begin{matrix} \begin{pmatrix} Y_{11} & Y_{21} & Y_{31} \\ Y_{12} & Y_{22} & Y_{32} \\ Y_{13} & Y_{23} & Y_{33} \\ Y_{14} & Y_{24} & Y_{34} \end{pmatrix} & \begin{pmatrix} Y_{41} & Y_{51} & \dots & Y_{101} \\ Y_{42} & Y_{52} & \dots & Y_{102} \\ Y_{43} & Y_{53} & \dots & Y_{103} \\ Y_{44} & Y_{54} & \dots & Y_{104} \end{pmatrix} \\ \text{年份 } 61 & 62 & 63 & 64 & 65 & \dots & 70 \end{matrix}$$

$$\text{计算分段极差 } R_1, R_2, \text{ 此时 } R_1 = \max_{1 \leq s \leq 4} \left\{ \max_{1 \leq i \leq 3} y_{1s} - \min_{1 \leq i \leq 3} y_{1s} \right\} = \max_{1 \leq s \leq 4} R_{1s},$$

$$R_2 = \max_{1 \leq s \leq 4} \left\{ \max_{4 \leq i \leq 10} y_{1s} - \min_{4 \leq i \leq 10} y_{1s} \right\} = \max_{1 \leq s \leq 4} R_{2s},$$

R_{1s}, R_{2s} 就是对第 S 个指标分段后各段的极差, 反映各段总的差异情况就各用 R_{1s}, R_{2s} 的最大值 $\max_{1 \leq s \leq 4} R_{1s},$

$\max_{1 \leq s \leq 4} R_{2s}$ 。有了 $R_1, R_2, \max R_1$ 自然也就可以算出。这样就可以对多个气象要素同时进行分割。读者可以用本台

站的数据, 先选 5—6 年的资料, 完全仿照前面的算法算一遍, 熟悉一下整个计算过程和方法, 然后再来处理年代较长的资料。