

分类逐步筛选相似预报

中央气象局研究所 赵 漥

$$S_1^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y}_1)^2$$

$$S_2^2 = \frac{1}{n-m} \sum_{i=m+1}^n (y_i - \bar{y}_2)^2$$

相似方法是目前天气预报，特别是中长期预报中广泛采用的方法之一。选相似有各种办法，我们采用了分类逐步筛选相似法，现介绍如下：

一、方法

设有 p 个因子： x_1, x_2, \dots, x_p ，预报量为 y ，各有 n 次观测样本。问题是有了 $X^\circ = (x_1^\circ, x_2^\circ, \dots, x_p^\circ)'$ 后，找出 $X_i = (x_{1i}, x_{2i}, \dots, x_{pi})'$ ($i = 1, 2, \dots, p$) 中与其相似的样本 ('表示转置)。

(一) 相似标准

怎样才算相似，这是最基本的一个问题，我们认为，相似程度的确定应从大气运动的本质特征入手，建立恰当的数量表征，可称为相似指数，以 g 记之。 g_{ij} 表示样本 X_i 与 X_j 的相似程度。应有 $|g_{ij}| \leq 1$ 。当 $g_{ij} = 1$ 时， X_i 与 X_j 相等； $g_{ij} = -1$ 时， X_i 与 X_j 完全相反。相似的标准即给出临界值 $g_0 \geq 0$ ， $g_{ij} \geq g_0$ ，样本 X_i 与 X_j 相似， $g_{ij} \leq -g_0$ ，样本 X_i 与 X_j 相反。

(二) 多因子综合

相似过程的确定，往往需要考虑多方面的因子，例如环流、海温、太阳活动、前期降水、气温分布等等。预报员一般从当前最突出的特点出发（例如作1977年预报时，1976年大范围持续低温、北半球极涡的异常活动及太阳黑子低值年是最引人注目的），找出一些相似年份，然后，为了过滤样本，再增加一些因子，使相似年份集中在极少数几个样本上。分类逐步筛选相似的方法就是模拟这样的思路做的。

首先，根据 X° 的特点，把 P 个因子分成 k 类，每一类有 P_k 个因子 ($k = 1, 2, \dots, k$)， $\sum_{k=1}^k P_k = P$ 。下面，

我们把因子的值记成 x_{ki} ， k 表示类别， $i = 1, 2, \dots, P_k$ 表示第 k 类中第 i 个因子， $j = 1, 2, \dots, n$ 表示样本数。

入选因子的顺序从 $k = 1$ 开始，对 k 值相等的各因子，分别计算 t_i 值：

对每一个因子 x_{ki} ，给出值 x_{pi} （可取各因子的均值或其它值），设

$$x_{ki1} \geq x_{ki2} \geq \dots \geq x_{kim} \geq x_{pi} \geq x_{kim+1} \geq \dots \geq x_{kin}$$

$$t_i = \sqrt{\frac{|\bar{y}_1 - \bar{y}_2|}{m S_1^2 + (n-m) S_2^2}} = \sqrt{\frac{m(n-m)(n-2)}{n}}$$

$$\text{其中 } \bar{y}_1 = \frac{1}{m} \sum_{j=1}^m y_j \quad \bar{y}_2 = \frac{1}{n-m} \sum_{i=m+1}^n y_i$$

取 t_i 值最大者作为第一入选因子。其意义是对 x_{ki} 的两种不同取值范围，预报量 y 取值的差异最大。也可以将式中 x_{ki} 与 y 互换， t_i 值即反映了对预报量 y 的不同取值范围，该因子取值的差异。更一般地可考虑依 x_{ki} 大小排列的 y 的最优分割⁽¹⁾。

对入选因子 x_{ki} 计算 $g(x_{ki}; x_i^\circ)$ ，使 $g \geq g_0$ 成立的 j 为相似样本，称为一级相似；使 $g \leq -g_0$ 成立的 j 为相反样本，同样，称为一级相反样本。舍去不相似（反）的样本，把相似（反）的样本作为全部样本考虑其它因子，重复上述过程，引入第二个因子。从一级相似（反）样本中又进一步选出相似（反）样本，称为二级相似（反）。第一类因子全部入选后，顺序考虑第二类，第三类……直至最后只剩下一、二个样本为止。如这时已引入1个因子，则称这些因子为1级相似（反）样本。这就是分类逐步筛选相似的过程。

(三) 稳定性

随着相似级别的提高，相似样本不断地减少，预报量 y 出现的情况也不同，若 y 值忽大忽小，预报时就难以判断；若 y 值的变化（分布）比较稳定，对预报的价值就大些。因此也可以在分类逐步筛选相似预报中，引入对不同相似级别中 y 值分布稳定性概念及判据。

(四) 多时刻资料的使用

大气处在不断运动的过程中，各要素随时间有很强的承继性，要从运动的即连续的而不是静止的角度考虑相似。因此，除了考虑样本间的相似外，有必要考虑样本对时间差分的相似，即连续变化过程的相似。

(五) 预报取值

求出相似及相反的样本后，预报时以相似样本为主，参考相反样本，相反样本与相似样本的 y 取值差别大些为好。另外，相似还不是相等，相似样本间还有差别，预报时需作一些订正。

二、预报实例：

预报量 y 为长江中下游五站（上海、南京、芜湖、九江、汉口）5—8月平均降水量。选择了11个因子：
 x_1 ：当年1月极涡强度， x_2 ：当年1月东亚槽位置，
 x_3 ：当年1月东亚槽强度， x_4 ：当年1月副高强度，
 x_5 ：当年1月副高面积指数， x_6 ：当年1月亚欧经向环流指数， x_7 ：当年1月全国温度等级， x_8 ：前一年年平均太阳黑子相对数， x_9 ：前一年 y 值， x_{10} ：前一年 y 值减前第二年 y 值， x_{11} ：前第二年 y 减前第三年 y 值。共使用了1954—1974年共21年资料。

规定

$$g(x_{i1}, x_{im}) = \begin{cases} 1 & \\ 0 & \\ -1 & \end{cases}$$

$$g_0 = 0$$

即距平符号相同即算相似，否则为相反， x_{pi} 取为 x_i 对上一级相似样本的平均。

用1975年、1976年资料作了检验：

把11个因子分为四类， x_4, x_5, x_6, x_7 为第一类， x_1, x_2, x_3 为第二类， x_8, x_{10}, x_{11} 为第三类， x_9 为第四类。1975年相似入选因子（按入选顺序）为 x_5, x_6, x_7, x_4, x_2 ，相似于1957、1974年，y在667—738毫米间，相反入选因子为 x_5, x_4, x_7 ，与1958、1960、1966年相反，这三年y在365—501毫米间，1975年实况为714毫米。

1976年与1968、1972年相似，y在404—478毫米间，

与1960、1966年相反。实况为492毫米。

试做了1977年预报。规定 x_1, x_7, x_8 为一类， x_2, x_3, x_5, x_6 为二类， x_9, x_{10}, x_{11} 为三类。相似入选因子 x_8, x_1, x_7 ，与1955、1956、1963年相似，y在580—785毫米间。相反入选因子 x_8, x_7, x_1 ，与1975、1971年相反，y在510—738毫米间。上述相似与相反有矛盾，主要考虑相似。预报y正常稍多。

三、讨论

(1) 相似方法较简单、直观。能表达一般线性模式不能表达的关系，对极值也有一定的预报能力。

(2) 因子引入的先后顺序对选出的相似样本有决定性的影响。由于分类逐步筛选相似法采取事先分类这样一个步骤，可以使预报员充分发挥天气分析的潜力。

(3) 可以对不同因子规定不同的 g_0 ，使其尽量符合天气学原理。