技术交流

用非线性分类筛选模式做短期气候预测的尝试

杨洪昌　杨志刚

（山东省气象局，济南　250031）

提　要

在相关普查的基础上，使用最优分割挑选因子，逐步对预报量进行分类，按照业务预报的需求，形成分类预报模式，取得了较好的试报、预报效果。

关键词： 相关普查　最优分割　分类预报模式

引　言

近几十年来，在经验短期气候预测方面进行了大量的实践、研究工作，形成了一些在预报中常用的方法。其中，在分类预报方面主要是基于 Fisher 或 Bages 准则的判别分析和基于各种相似系数的聚类分析。这些方法的主要缺点是在进行分类时所选的因子对预报量的所有样本均起作用。相关分析表明，在某种情况下，这一因子起主要作用；而在另一种情况下，另一因子起支配作用。为此，本文从最优分割出发建立分类筛选模式，并以山东 8 月降水量预报为例，说明此模式在短期气候预测中的应用。

1　方法步骤

相关分析是多元分析的基础，本文使用的分类筛选方法属于多元分析的内容之一。因此，进行分类筛选前需计算预报量 y 与前期影响因子 $x_l(l=1,2,\cdots,m)$ 的相关系数，并进行显著性检验。若检验结果在一定的显著水平下相关显著，则称 x_l 是 y 的预报因子。由于 y 与 x_l 有较好的相关关系，则必然存在 y 随 x_l 的增大而增大（或减小）的趋势。若将 x_l 按照递增（或递减）顺序排列，并将 y 按着相应的顺序排列，对其做 K 分割，在序列的容量为 n 时，一切可能的分法有[1]：

$$R'(n,K)=\begin{pmatrix} n-1 \\ K-1 \end{pmatrix}$$

对每一种分法，计算所分成的 K 类变差和。其中，第 p 种分法的变差和为：

$$V_p=\sum_{j=1}^{K}V_{pj} \qquad p=1,2,\cdots,\begin{pmatrix} n-1 \\ K-1 \end{pmatrix} \tag{1}$$

其中，$V_{pj}=\sum_{i=1}^{n_j}(y_{pij}-\bar{y}_{pj})^2$ 为第 p 种方法第 j 类的变差；$\bar{Y}_{pj}=\sum_{i=1}^{n_j}\frac{1}{n_j}y_{pji}$ 为第 p 种分法第 j 类的均值。

在 $\begin{pmatrix} n-1 \\ K-1 \end{pmatrix}$ 个 V_p 中 $\min(V_p)(p=1,2,\cdots,\begin{pmatrix} n-1 \\ K-1 \end{pmatrix})$ 所对应的 K 分割即为最优 K 分割。但其各类之间的差异与类内差异的比在统计上是否显著，还需要进行检验。

若我们将预报量序列 $y_i(i=1,2,\cdots,n)$ 分成 K 类，其各类的均值分别为 $\mu_1,\mu_2,\cdots,\mu_K$，所有样本的均值为 μ，则统计量

$$F_K=\frac{\sum_{j=1}^{K}n_j(\mu_j-\mu)^2/(K-1)}{\sum_{j=1}^{K}\sum_{i=1}^{n_j}(y_{ji}-\mu_j)^2/(n-K)} \tag{2}$$

服从自由度$(K-1,n-K)$的F分布，选取信度α，若$F_K>F_{K\alpha}$，则所做的K分割是显著的。

在做分类筛选时，首先通过相关分析选出预报量y的m个相关因子，将y按照每个因子的递增(或递减)顺序排列，得到y的m个序列，分别对各序列做二分割。若V_1、V_2分别为二分割第一类、第二类的变差，则$\min(V_1-V_2)_p(p=1,2,\cdots,\binom{n-1}{K-1})$对应的分割为最优二分割。这样，可得到$m$个最优二分割，分别对应于：$\min((V_1+V_2)_p)_l(l=1,2,\cdots,m)$，从中选取一个最小的，将其对应的最优二分割按照式(2)计算统计量F_2，若达到一定的显著水平，则将其对应的因子作为第一因子，并按该因子的递增(或递减)顺序将预报量分为两类；若F_2达不到一定的显著水平，表明从最优二分割的角度来说，按照最好的因子作出的分类，其两类之间的差异在统计上也是不显著的。这时需将预报量y的m个序列分别做最优三分割，四分割，…，直到F_K在选取一定的显著水平时显著为止。到此，第一步按照某一因子将预报量序列分为K类。。

对于所分得的各类，按照其它因子重复以上步骤，进行分类，直到预报量分成的各类类内差异较小，按照预报的要求没有必要再分下去时分类结束。

通过以上步骤，可实现逐步对预报因子筛选，对预报量进行分类，最后得到一个可用于业务预报的分类模式。

2 应用实例

选取山东省8月平均降水量作为预报量，据其1957～1987年的资料，通过计算其与上年11月至当年4月北半球500hPa高度，北太平洋表层海温的相关系数，相关显著标准为$\alpha=0.05$。规定相邻的相关显著格点数>3，格点相关系数绝对值最大值>0.5的区域为相关区。相关区内格点高度或海温距平和为预报因子，共得到8个预报因子，分别为：

x_1：上年11月45～60°N、115～140°W范围内10点500hPa高度距平和；

x_2：上年11月30～35°N、140～155°E范围内4点500hPa高度距平和；

x_3：当年1月10°N、60～70°W、90～110°W范围内5点500hPa高度距平和；

x_4：当年2月10～15°N、60～100°W范围内6点500hPa高度距平和；

x_5：当年4月75～80°N、120～150°W范围内6点500hPa高度距平和；

x_6：当年4月35～50°N、35～70°E范围内9点500hPa高度距平和；

x_7：当年3月5°N～10°S、80～160°W范围内22点海温距平和；

x_8：当年4月5°N～10°S、105～160°W范围内28点海温距平和。

① 将以上8个因子均按由小到大的顺序排列，山东省8月降水距平百分率分别按各因子相应的顺序排列。其按第一个因子的排列见表1(按其它因子的排列略)。

对表1中8月降水距平百分率序列做二分割，其最优二分割$S_{31}(2/2)$，即分成(71,69)(10,4,…,−58)两类，其对应的年代是：(1964,1974)(1985,1978,…,1957)。

同时，将8月降水距平百分率分别按照$x_2,x_3,\cdots,x_8$各因子的递增顺序排列并做最优二分割，结果发现，在按照上述8个因子递增顺序排列所做的最优二分割中，以按x_3递增顺序所做的最优二分割两类的变差和为最小。这表明，从最优二分割的角度来看，x_3是与山东省8月降水量关系最好的因子，故应选取x_3，将8月降水距平百分率分为两类。但在统计上是否显著，还需进行检验。

② 将8月降水距平百分率按x_3递增顺序排列，并做最优二分割，其结果见表2。

表1 山东省8月降水距平百分率按 x_1 递增顺序的排列结果

序号	年代	降水距平百分率/%	序号	年代	降水距平百分率/%
1	1964	71	17	1966	−43
2	1974	69	18	1982	4
3	1985	10	19	1981	−32
4	1978	4	20	1975	−4
5	1976	22	21	1987	29
6	1963	−4	22	1971	55
7	1961	0	23	1973	−22
8	1962	32	24	1970	−33
9	1969	−7	25	1960	−28
10	1959	−17	26	1968	−36
11	1984	25	27	1979	−52
12	1972	9	28	1958	7
13	1986	−41	29	1980	−57
14	1983	−64	30	1977	−33
15	1967	−11	31	1957	−58
16	1965	−4			

经计算，第一类的均值 $\mu_1=25.875$，变差为：$V_1=\sum_{i=1}^{8}(y_{1i}-\mu_1)^2=8188.875$；第二类均值 $\mu_2=-18.087$，变差为：$V_2=\sum_{i=1}^{23}(y_{2i}-\mu_2)^2=18209.826$；所有样本的均值为：$\mu=-6.742$（因为降水距平百分率是相对于1951～1980年的降水均值计算，故而 $\mu\neq 0$）；类间变差为：$\sum_{j=1}^{2}n_j(\mu_j-\mu)^2=11471.257$

$$F_2=\frac{\sum_{j=1}^{2}n_j(\mu_j-\mu)^2/(2-1)}{\sum_{j=1}^{2}V_j/(31-2)}=12.60$$

自由度(1,29)，$\alpha=0.01$ 时，$F_\alpha=7.60$

由于 $F_2>F_\alpha$，山东省8月降水距平百分率按 x_3 递增顺序所做的最优二分割在0.01的显著水平上两类的差异是显著的。因此，首先据 x_3 将山东省8月降水距平百分率分为两类，第一类的末样本对应于1964年，第二类的首样本对应于1968年，该两年 x_3 的平均值是−5，可作为分类的临界值。即 $x_3\leqslant -5$ 的年份属于第一类；$x_3>-5$ 的年份为第二类。

③ 对于以上得到的两类，分别按除 x_3 以外的其它因子的递增顺序排列，做最优分割，并进行显著性检验，逐步对其它因子进行筛选，对8月降水距平百分率进行分类。重复以上步骤，直到得到的各类内8月降水距平百分率的差异小到满足业务预报的要求时，分类筛选结束。附图为本例分类筛选结果。图中，每一类的上行数字是年代，下行数字是8月降水距平百分率；带“*”者为试报、预报年份。可见，在上述8个因子中，基于最优分割进行分类筛选，只有 x_3、x_8、x_6、x_5 是主要的，按照这4个因子，可逐步将山东省8月降水距平百分率分成6类。

表2 按照 x_3 分类所得的最优二分割

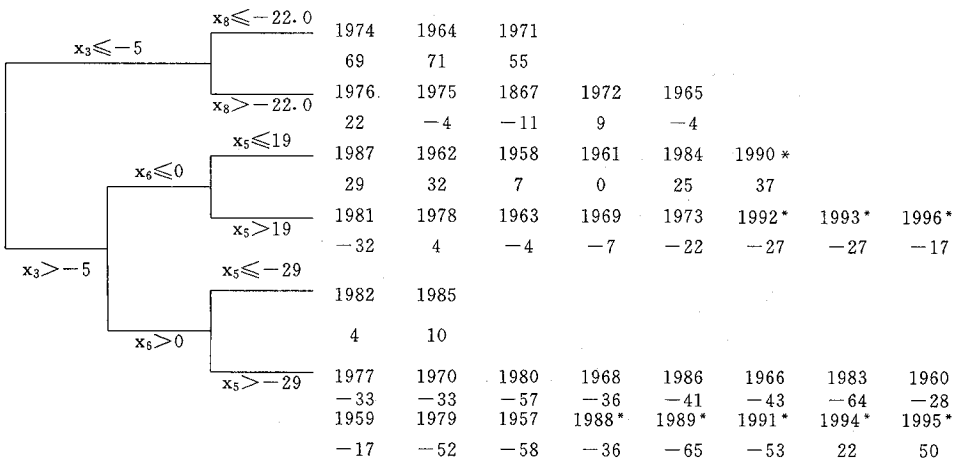
x_3 递增顺序号	降水距平百分率%	年代	x_3 递增顺序号	降水距平百分率%	年代
1	9	1972	17	−52	1979
2	69	1974	18	−32	1981
3	22	1976	19	−41	1986
4	−4	1975	20	7	1958
5	−4	1965	21	−17	1959
6	−11	1967	22	−28	1960
7	55	1971	23	0	1961
8	71	1964	24	−33	1970
9	−36	1968	25	−33	1977
10	−7	1969	26	29	1987
11	−4	1963	27	−22	1973
12	25	1984	28	−57	1980
13	32	1962	29	−58	1967
14	−43	1966	30	4	1982
15	4	1978	31	−64	1983
16	10	1985			

注：带有横线的为第一类，其余为第二类。

表3 山东省8月降水距平百分率分类

类别	均值/%	平均绝对偏差/%	最大绝对偏差/%	优势距平频率
1	65	7	10	+3/3
2	2	10	20	−5/5
3	19	12	19	+5/5
4	−12	12	20	−4/5
5	7	3	3	+2/2
6	−42	12	25	−11/11

表3给出了1957～1987年山东省8月降水距平百分率的分类统计，各类的平均绝对偏差均在12%以下，其最大绝对偏差≤25%。可见，已满足业务预报的要求。



附图 山东省8月降水距平百分率分类结果

用上述方法，对1988～1989年进行试报，趋势正确。在1990～1996年的预报实践中，用各类的优势距平做趋势预报、用各类的平均距平百分率做定量预报，除1994年和1995年外，其余5年其趋势预报均正确，定量预报与实况的距平百分率绝对差≤23%。

参考文献

1 张尧庭，方开泰. 多元统计分析引论. 北京：科学出版社，1982：445.

A Short-range Climate Prediction Using Non-linear Categorical Screen Model

Yang Hongchang Yang Zhigang
(Shandong Meteorological Bureau, Jinan 250031)

Abstract

On the basis of correlative general search, factors are selected by means of optimum category, and the predictands are classified gradually. According to the needs of operation a categorical prediction model is developed, and the results of the model are reasonable both for experimental and for actual forecast.

Key Words: correlative search optimum category categorical prediction model