第4卷 第1期 大 气 科 学 Vol. 4, No. 1
1980 年 3 月 SCIENTIA ATMOSPHERICA SINICA March, 1980

一种条件概率预报方法

汪 盛 章
(江西省赣州地区气象台)

提 要

本文提出一种计算较简便的非线性的概率预报方法——条件概率法. 本方法在单站长期天气预报中使用了66次，结果相当于将预报量的观测资料分为五级时，其平均绝对误差接近于一级.

近代，概率预报有了一定发展. 例如，有人采用数值预报的输出作为因子[5]，也有人在预报模型上作些研究[1—3]. 目前流行的概率预报方法，大多数采用回归模型，计算量较大. 这里提出一种计算比较简便的、非线性的条件概率法.

一、公 式

设已有编码资料

$$\{y_j, x_{ij}\} \quad (j = 1, 2, \cdots, N;\ i = 1, 2, \cdots, m;\ x_{ij} = 1, 2, \cdots, s_i;\ y_j = 1, 2, \cdots, s_0) \tag{1}$$

为多维马尔可夫链，其中 N 为样本容量，m 为因子总数，$s_i \geqslant 3$ 为因子 x_i 的编码分级总数，$s_0 \geqslant 3$ 为预报量 y 的编码分级总数，当 $x_1, x_2, \cdots, x_m$ 取某组固定值时（对应于第 j 个子样）预报 $y = k (k = 1, 2, \cdots, s_0)$ 的条件概率为

$$p_j\{y = k/x_1 x_2 \cdots x_m\}, \tag{2}$$

上述 s_0 个概率中若存在一个最大的概率，即有

$$\max_k P_j\{y = k | x_1 x_2 \cdots x_m\} = P_j\{y = l | x_1 x_2 \cdots x_m\} \tag{3}$$

时（其中 k 不固定，l 固定），我们就预报 $\hat{y} = l$.

令 E_i 表示单因子 x_i 取某固定值（对应于第 j 个子样）时对应预报量 $y = k$ 的随机事件，即

$$P_j\{E_i\} = P_j\{y = k | x_i\}, \tag{4}$$

令 E_i^c 为 E_i 的逆事件，则

$$P_j\{E_i^c\} = P_j\{y \neq k | x_i\} = 1 - P_j\{y = k | x_i\}. \tag{5}$$

对于选定的 m 个因子 $x_1, x_2, \cdots, x_m$ 取某组固定值时，或者是对应预报量 $y = k$，或

1978 年收到.

者是对应预报量 $y \neq k$，再无别的可能．因此，

$$P_i\{y = k|x_1 x_2 \cdots x_m\} = 1 - P_i\{y \neq k|x_1, x_2, \cdots, x_m\}. \tag{6}$$

定义 若有

$$P_i\{y \neq k|x_1 x_2 \cdots x_m\} = \sum_{i=1}^{m} P_i\{y \neq k|x_i\}, \tag{7}$$

则称 m 个随机事件 F_i^k 是整体弱独立的（注 1）．

对于齐次马尔可夫链，任一单因子 x_i 为某一固定值时，对应 $y = k$ 的概率与时间（或样本序列的次序）无关，即

$$P_i\{y = k|x_i\} = P\{y = k|x_i\}. \tag{8}$$

由 (5)—(8) 式，得

$$P_i\{y = k|x_1, x_2, \cdots, x_m\} = 1 - \prod_{i=1}^{m} (1 - P\{y = k|x_i\}). \tag{9}$$

$$(k = 1, 2, \cdots, s_0;\ j = 1, 2, \cdots, N)$$

于是，我们证明了下列结论：对于齐次马尔可夫链(1)，若各个因子 $x_i (i = 1, 2, \cdots, m)$ 分别取某一固定值时对应 $y = k$ 的 m 个事件 E_i^k 是整体弱独立的，则 (9) 式成立．

由 (6)、(7) 式可知，(9) 式给出了 m 个因子中至少有一个因子对应

$$y = k \quad (k = 1, 2, \cdots, s_0)$$

的概率．考虑到 m 个因子中至少有一个因子对应 $y = 1, 2, \cdots, s_0$ 的 s_0 个事件是相容的，故有

$$\sum_{k=1}^{s_0} P_i\{y = k|x_1 x_2 \cdots x_m\} \geqslant 1, \tag{10}$$

注意，对于不同的 j，(10)式左端的值往往是不同的．令

$$P_j^*\{y = k|x_1 x_2 \cdots x_m\} = \frac{P_j\{y = k|x_1 x_2 \cdots x_m\}}{\sum_{k=1}^{s_0} P_i\{y = k|x_1 x_2 \cdots x_m\}}, \tag{11}$$

则

$$\sum_{k=1}^{s_0} P_j^*\{y = k|x_1 x_2 \cdots x_m\} = 1. \tag{12}$$

由 (9) 式和 (11) 式，得

（注 1）按照经典独立性定义[4]，对于 m 个随机事件 $E_1^{k_1}, E_2^{k_2}, \cdots, E_m^{k_m}$，如果对于每一个 $r \leqslant m$ 以及任意两两不同的正整数 $k_1, k_2, \cdots, k_r \leqslant m$，等式

$$P\{E_{k_1}^{k_1} E_{k_2}^{k_2} \cdots E_{k_r}^{k_r}\} = \prod_{i=1}^{r} p\{E_{k_i}^{k_i}\} \tag{7-a}$$

成立，则称该 m 个事件是独立的．因此，对于符合 (7-a) 式独立意义的因子，必须满足 $\sum_{k=1}^{m} C_m^k$ 个等式，例如 $m = 4$ 时须符合 11 个等式；$m = 5$ 时，须符合 26 个等式．应该指出，其中两两独立的等式在统计预报实践中是很难实现的，而 (7) 式则要求较低．由此可知，“整体弱独立”比“独立”的条件弱得多。

$$P_j^*\{y=k|x_1x_2\cdots x_m\}=\frac{1-\prod_{i=1}^{m}(1-P\{y=k|x_i\})}{\sum_{k=1}^{s_0}\left[1-\prod_{i=1}^{m}(1-P\{y=k|x_i\})\right]}. \tag{13}$$

预报判据，当

$$\max_k P_j^*\{y=k|x_1x_2\cdots x_m\}=P_j^*\{y=l|x_1x_2\cdots x_m\} \tag{14}$$

成立时，预报 $\hat{y}=l$.

实际计算时，用条件频率来代替条件概率，

$$P\{y=k|x_i\}\doteq\begin{cases}n_{irk}/n_{ir\cdot}, & (n_{ir\cdot}>0),\\ 0, & (n_{ir\cdot}=0)\end{cases}$$
$$(i=1,2,\cdots,m;\ r=1,2,\cdots,s_i;\ k=1,2,\cdots,s_0) \tag{15}$$

其中

$$n_{ir\cdot}=\sum_{k=1}^{s_0}n_{irk} \tag{16}$$

为因子 $x_i=r$ 时的频数，n_{irk} 为 $x_i=r$ 时出现 $y=k$ 的频数.

对于因子组取某一组固定值的时刻，由(13)式或(9)式可得到 s_0 个条件概率. N 个子样，就有 $N\cdot s_0$ 个条件概率.(14)式的实质是在固定的时刻，按最大可能原理，对 s_0 个条件概率作出决策(预报). 我们也可以先对 $y=1$ 的 N 个条件概率、$y=2$ 的 N 个条件概率、……、$y=s_0$ 的 N 个条件概率分别与实况比较，定出 s_0 个临界值(它们往往是互不相同的)，然后再对固定时刻的 s_0 个值作出决策. 我们称前一种决策为单决策，后一种决策为复决策. 经验表明，复决策的效果比单决策稍好一些.

严格说来，应当对(13)式作复决策. 在我们的初步应用中，为了计算简单，对(9)式作复决策. 下面导出对应于(9)式的复决策规则. 对应于(13)式的复决策规则与前者类似，不再给出. 令 y_c 表示 y 的实况，设

$$P_{kj}=P_j\{y=k|x_1x_2\cdots x_m\}, \tag{17}$$

k 固定的 N 个条件概率 P_{kj} 可以分为二部分：其对应(同时刻的)实况 $y_c=k$ 的部分 P_{ki_1} 及其对应(同时刻的)实况 $y\neq k$ 的部分 p_{ki_2}. 令 $\min\limits_{(y_c=k)}P_{ki_1}$ 表示实况 $y_c=k$ 部分的最小值，$\max\limits_{(y_c\neq k)}P_{ki_2}$ 表示实况 $y_c\neq k$ 部分的最大值，则有

$$\min_{(y_c=k)}P_{ki_1}>\max_{(y_c\neq k)}p_{ki_2} \tag{18}$$

成立或不成立两种情形. 若(18)式成立，取

$$P_{kc}=\frac{1}{2}\left[\min_{(y_c=k)}p_{ki_1}+\max_{(y_c\neq k)}P_{ki_2}\right]\quad(k=1,2,\cdots,s_0) \tag{19}$$

为复决策的条件概率预报临界值.(18)式不成立时，情况比较复杂，限于篇幅，这里不讨论了.

复决策的预报判据. 当

$$\max(P_{kj}-P_{kc})/P_{kc}=(P_{l_j}-P_{l_c})/P_{l_c} \tag{20}$$

成立时，预报 $\hat{y}=l$. 在多数情况下，$P_{1c}, P_{2c}, \cdots, P_{s_0c}$ 相差不大，(20)式可以简化为下

列两个规则：

[规则1] 若 s_0 个条件概率都小于(或等于)其相应的临界值，则取其中的最大者，作出相应的预报；

[规则2] 若有 $r(1 \leqslant r < s_0)$ 个条件概率大于其相应的临界值，则在此 r 个条件概率中，取一个最大者，作出相应的预报.

二、应用效果

设有甲、乙两种预报工具(简称为甲、乙)，经过多次实践比较，甲平均误差为一级，乙平均误差为二级. 虽然它们都偏离实况，甲比乙要好. 为了反映上述客观情况，又要使评分计算尽可能简便，我们提出下列评分公式：

$$B_f = \left[1 - \frac{|y_c - \hat{y}|}{(s_0 - 1)}\right] \cdot 100\%, \tag{21}$$

其中 B_f 为得分，y_c 为实况，$\hat{y}$ 为预报结果，s_0 为 y_c 的分级总数，y_c、$\hat{y} = 1, 2, \cdots, s_0$. $B_f = 100\%$ 时，预报完全正确；$B_f = 0$ 表示预报与实况的级差为最大. 对于长期天气预报，我们认为 B_f 在75%(相当于将预报量分五级，误差一级)以上时，预报较好.

本方法应用于单站长期天气预报，已制作了三十四个预报工具，预报了六十六次，其中样本容量为19—24，因子选3—4个，历史概括率至少为90%. 用(21)式对66次预报分别评分，然后计算平均得分为

$$\bar{B}_f = \frac{1}{66}\sum_{i=1}^{66} B_{fi} = 76.7\%,$$

代入由(21)式得到的下式，令 $s_0 = 5$ 得

$$\overline{|y_c - \hat{y}|} = (1 - \bar{B}_f)(s_0 - 1) = 0.932.$$

这就是说，66次预报的平均绝对误差相当于将预报量分五级，误差接近一级.

三、讨 论

本方法计算比较简单，能较好地利用因子的非线性预报信息，初步应用的效果较好.

使用本方法的前提是要满足(3)、(7)、(8)和(15)式. 为此，我们设计了一套选因子的方法，有关这方面的详细内容，将另文讨论. 这里简要地说明一下.

如果所选的 m 个因子与预报量的关系不密切，就有可能导致(3)式不成立. 由于因子与预报量之间的客观规律是不知道的，我们只能以历史概括率的高低来衡量因子与预报量之间的关系是否密切. 为此，我们规定预报工具的合格率，即历史概括率的下限要达到90%.

由于(8)式和(15)式的要求，最好规定，预报工具使用一段时间后，就将新的资料输入预报工具，重新计算条件概率.

具体计算时，怎样严格判断 m 个事件 E_i^j 是整体弱独立的，能否将预报模型进一步改进以继续提高预报准确率等等问题尚需进一步研究。

（致谢：本文的修改，得到了周家斌、谢衷洁等老师的鼓励和帮助，特此致谢。）

参　考　文　献

[1] 王宗皓、李麦村等编著，天气预报中的概率统计方法，科学出版社，1974.
[2] 单站统计天气预报方法的研究，中国科学院大气物理研究所集刊第3号，1975，科学出版社.
[3] 王启鸣等，地震迁移的统计预报，数学学报，1(1974).
[4] M. 洛易甫著，梁文骐译，概率论(上册)，科学出版社，1966，10.
[5] Glahn, H. R. and Lowry, D. A., 1972, The use of model output statistics(MOS)in objective weather forecasting, *J. Appl. Meteor.*, **11**, 1203—1211.

A PREDICTION METHOD OF CONDITIONAL PROBABILITIES

Wang Sheng-zhang

(Ganzhou Meteorological Observatory, Jiangxi Province)

Abstract

In this paper, a simple and convenient method of nonlinear probability forecasts by arithmetic calculation is presented. This method has been tested 66 times in long-range weather forecast at a single station. The results show that the absolute mean deviation is a little less than one category when the observed value of the predictand is divided into five categories.