

多因子两值预报效能的初步探讨

张文坚

(浙江省气象科学研究所)

提 要

到目前为止,对多因子两值预报,已有不少方法,如编码法、REEP法等。但诸方法并没有涉及两值预报的内在规律。本文从概率论角度对它进行了研究,提出了预报效能的概率估计法。

针对简单投票法,我们导出了数学表达式,用它们探讨了多因子预报比单因子预报优越的条件。发现多因子预报效能取决于1)因子数;2)各因子的历史拟合率;3)各因子历史拟合率之间的差异程度;4)预报原则等四方面因素。还发现历史拟合率小于0.5的因子是起负贡献的。

一、多因子预报效能的概率估计

在实际工作中发现,预报工具的历史拟合率高者,不一定实际效果好;历史拟合率低者,不一定实际效果差。预报工具的历史拟合率不能恰当地表示该工具的预报效能,那么,如何客观地来估计预报效能呢?下面从概率论的角度进行探讨。

设事件 $A_1, A_2, A_3 \cdots A_m$ 相互独立,则有关系:

$$P(A_1, A_2, A_3 \cdots A_m) = P(A_1)P(A_2)P(A_3) \cdots P(A_m) \quad (1)$$

对于一个预报对象而言,每个因子对它作预报,有二个互不相容的事件:报准和报不准。它们的概率是该因子报准的概率与报不准的概率。而不同因子的事件串可看作是一个个相互独立的事件。多因子预报就是多个事件的交,这些事件的不同的交可看作一系列新事件,它们的概率取决于参加交的各个事件的概率。这些新事件由预报原则区分为两类:报准事件和报不准事件。由上可知,简单地考虑,多因子预报的效能可以由两方面决定,即(1)各因子的预报概率——历史拟合率或历史错误率(姑且这样说)。(2)预报原则。现举例说明。

设有三个因子 x_1, x_2, x_3 , 对某预报对象作预报,它们的历史拟合率分别为 $P(A)$ 、 $P(B)$ 、 $P(C)$, 历史错误率为 $P(\bar{A})$ 、 $P(\bar{B})$ 、 $P(\bar{C})$, 其数值为:

$$P(A) = 0.6 \quad P(B) = 0.7 \quad P(C) = 0.8$$

$$P(\bar{A}) = 0.4 \quad P(\bar{B}) = 0.3 \quad P(\bar{C}) = 0.2$$

假设这三个因子互相独立,用简单投票法(以后称投票法)综合,则综合后报准的概率可以

估计为

$$\begin{aligned} P &= P(ABC) + P(AB\bar{C}) + P(A\bar{B}C) + P(\bar{A}BC) \\ &= P(A)P(B)P(C) + P(A)P(B)P(\bar{C}) + P(A)P(\bar{B})P(C) \\ &\quad + P(\bar{A})P(B)P(C) = 0.788 \end{aligned}$$

$P = 0.788$ 可作为该工具预报效能的估计值.

二、多因子预报的正确率问题

一般认为多因子预报无疑比单因子预报优越, 即预报效能比单因子高. 但从上例的计算中发现并非如此. 下面就以投票法为例来探讨多因子预报效能的内在关系. 分两种情况来研究.

1. 各因子历史拟合率一致时

设有 m 个因子相互独立, 各因子的历史拟合率均为 p , 历史错误率为 p' , 考虑因子为奇数, 用投票法综合后的预报效能可估计为

$$P_m = p^m + C_m^1 p^{m-1} p' + C_m^2 p^{m-2} p'^2 + \cdots + C_m^{\frac{m-1}{2}} p^{\frac{m+1}{2}} p'^{\frac{m-1}{2}} \quad (2)$$

要想知道多因子综合后的预报效能比单因子有所提高, 只须证明 $P_m \geq p$, 或者把上式化为 $P_m = p + \Delta p_m$ ($\Delta p_m \geq 0$) 的形式. 可以证明 (2) 可写成 $P_m = p + \Delta p_m$ 的形式, 且有如下的关系式:

$$P_m = p + pp'(p - p')(C_m^0 + C_m^1 pp' + C_m^2 p^2 p'^2 + \cdots + C_m^{\frac{m-3}{2}} p^{\frac{m-3}{2}} p'^{\frac{m-3}{2}}) \quad (3)$$

(证明过程见附录). 从 (3) 式可知只须 $p - p' \geq 0$, 即 $p \geq 0.5$, 就能使 $\Delta p_m \geq 0$. Δp_m

表 1

Δp \ m \ p	3	5	7	9	11
0.55	0.0248	0.0431	0.0583	0.0714	0.0831
0.60	0.0480	0.0826	0.1102	0.1334	0.1535
0.65	0.0683	0.1148	0.1502	0.1783	0.2013
0.70	0.0840	0.1369	0.1740	0.2012	0.2218
0.75	0.0938	0.1465	0.1794	0.2011	0.2157
0.80	0.0960	0.1421	0.1667	0.1804	0.1884
0.85	0.0893	0.1234	0.1379	0.1444	0.1474
0.90	0.0720	0.0914	0.0973	0.0991	0.0997
0.95	0.0428	0.0488	0.0498	0.0500	0.0500
$\Delta P_{\max}$	0.0962	0.1467	0.1798	0.2039	0.2218
$\Delta P_{\max}$ 时 P	0.7845	0.7595	0.7396	0.7248	0.7003

可看作比单因子预报效能 p 增加的部分。此条件在二值预报上一般都能成立。(3)式适用于 $m \geq 3$ 的情况。如 $m = 7$ 时,形式如下:

$$P_7 = p + pp'(p - p')(C_1^0 + C_3^1 pp' + C_5^2 p^2 p'^2)$$

下面列出不同 p 值时的 Δp 值与 Δp 的极大值, m 取 3、5、7、9、11, 见表 1。根据表 1 作图 1, 从图 1 可看出:

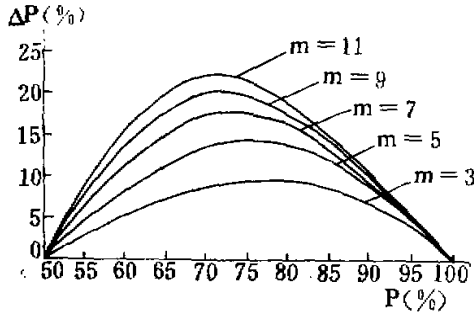


图 1 不同因子数 m 时 p (各因子相同) 与 Δp 之关系。

(1) 随着因子数的增加, Δp 也增加。

(2) 因子数相同, 不同的 p 值, Δp 也不同, 中间大, 两边小, 服从一种偏态分布。

(3) 随着因子数的增加, Δp 的极大值向 p 的低端漂移。

(2)、(3) 给予我们一个启示, 可以根据不同的因子情况来选择综合因子的个数, 能最大限度地提高预报效能。

实际上各因子的历史拟合率不会完全

一样, 在各因子 p 值相差不太大的情况下, 可以用算术平均值来代替 p 查表, 看看选几个因子较为有利。同时, 应权衡一下增加综合因子个数有利, 还是改变预报原则有利。当然因子个数不能任意增加, 因为因子增加后, 组合数也增加, 不稳定性也就增加, 到一种组合只有一个个例时, 方法本身不是反映预报对象的共性, 而是反映它的个性, 也就失去了运用的价值。投票法组合数最少, 为 $m + 1$ 种, 其它两值预报方法一般为 2^m 种, 因此投票法可以综合更多的因子, 因增加因子使预报效能受益, 它能提供的最大限度的贡献可以作为其它预报方法的一个参照标准。凡因改变预报原则在预报效能上所受的益超过投票法时, 就可采用改变预报原则的办法, 否则采用增加因子的办法。

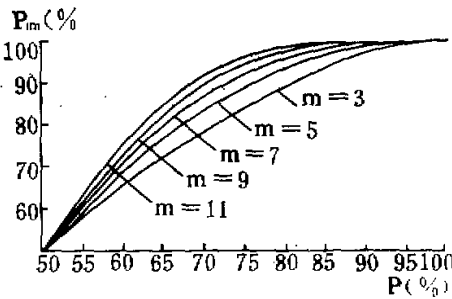


图 2 不同因子数 m 时 p (各因子相同) 与 P_m 之关系。

随着因子的增加, 预报效能有明显的提高, 但并不是要求无限地增加因子, 有 p 到 11 个 p 值在 0.7 以上的因子就能使预报效能达到 0.9 以上, 如 p 值在 0.75 以上, 综合后预报效能可超过 0.95。图 2 是将表 1 的资料稍加处理后作成, 图中曲线呈抛物线, 可看出 9 个因子与 11 个因子的曲线在 p 大于 0.75 以上是很接近的了。

2. 因子历史拟合率不一致

现考虑如下情况, 设有 m 个因子, K 个因子的历史拟合率为 $p + x$, $m - K$ 个因子的历史拟合率为 p 。此时, 如考虑最简单 $K = 1$ 的情况, 不难得到下式

$$P_m = (p+x)[C_{m-1}^0 p^{m-1} + C_{m-1}^1 p^{m-2} p' + \cdots + C_{m-1}^{\frac{m-1}{2}} p^{\frac{m-1}{2}} p'^{\frac{m-1}{2}}] \\ + (1-p-x)[C_{m-1}^0 p^{m-1} + C_{m-1}^1 p^{m-2} p' + \cdots + C_{m-1}^{\frac{m-3}{2}} p^{\frac{m-1}{2}} p'^{\frac{m-1}{2}}] \quad (4)$$

(4)式整理后为

$$P_m = C_m^0 p^m + C_m^1 p^{m-1} p' + \cdots + C_m^{\frac{m-1}{2}} p^{\frac{m+1}{2}} p'^{\frac{m-1}{2}} \\ + x C_{m-1}^{\frac{m-1}{2}} p^{\frac{m-1}{2}} p'^{\frac{m-1}{2}} \quad (5)$$

利用附录结果可得到下式为

$$P_m = (p+x) + \Delta p_m \\ = (p+x) + pp'(p-p')(1 + C_1^1 p p' + C_2^2 p^2 p'^2 + \cdots + C_{\frac{m-3}{2}}^{\frac{m-3}{2}} p^{\frac{m-3}{2}} p'^{\frac{m-3}{2}}) \\ - x(1 - C_{\frac{m-1}{2}}^{\frac{m-1}{2}} p^{\frac{m-1}{2}} p'^{\frac{m-1}{2}}) \quad (6)$$

要使 $\Delta p \geq 0$, 须满足不等式

$$x \leq \frac{pp'(p-p')(1 + C_1^1 p p' + C_2^2 p^2 p'^2 + \cdots + C_{\frac{m-3}{2}}^{\frac{m-3}{2}} p^{\frac{m-3}{2}} p'^{\frac{m-3}{2}})}{1 - C_{\frac{m-1}{2}}^{\frac{m-1}{2}} p^{\frac{m-1}{2}} p'^{\frac{m-1}{2}}} = \tilde{p}_m \quad (7)$$

表 2

$\tilde{p}$ p	3	5	7
0.55	0.049	0.068	0.084
0.60	0.092	0.126	0.152
0.65	0.125	0.166	0.196
0.70	0.145	0.186	0.216
0.75	0.150	0.186	0.207
0.80	0.141	0.168	0.182
0.85	0.120	0.137	0.144
0.90	0.088	0.096	0.099
0.95	0.047	0.050	0.050

(6)式为 x 的一次方程, 如果 $K=2$, 可得到 x 的二次方程, 依此类推. 下面列出 $K=1$ 时, 不同 p 值 $m=3, 5, 7$ 的 $\tilde{p}_m$ 值, 见表 2, 并由表 2 作图 3.

从图 3 可看出其图形与因子历史拟合率一致时极其相似.

三、预报实例

以缙云站 9 月雨量为例, 所用资

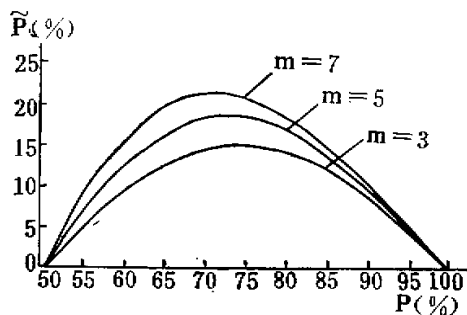


图 3 不同因子数 m 时 p (各因子相同) 与 $\tilde{p}$ 之关系.

们的组合数为 $2^3=8$, 比 5 个因子的投票法的组合数 $5+1=6$ 要大。考虑到最优因子 x_1 , 并考虑使三个因子的互相关概率小, 选得 x_1, x_3, x_5 三个因子进行综合, 见表 3。这两种方法的拟合率均为 $18/20=0.90$, 1980 年试报也都报准。

从这个长期预报的例子来看, 因子数起了主要作用, 因此似乎用投票法好一些。

四、讨论与结语

本文的宗旨是探索一种概率预报方法, 它不需要人为地求临界值, 对历史和将来的情况都只求一种可能性, 这符合统计预报的原理。因为一种统计预报工具, 即使拟合率达到百分之百, 它的预报效能也不可能达到百分之百。

在作多因子预报效能的概率估计时, 使用了单因子的拟合率, 这是不可避免的。但对我们研究多因子预报却带来了好处, 因为因子是客观存在的, 剩下的问题只需要研究采用什么方法使预报效能达到最大。本文通过对投票法的解析, 认为预报效能取决于四方面的因素:

(1) 因子数 m

在因子互相独立的前提下, 增加 m 对提高预报效能有明显的好处。如在 p 值均为 0.70 时, 从 3 个因子增加到 5 个因子可提高预报效能 0.05。由于投票法组合数少, 有利于增加 m 而受益。但是, 过量的因子也是没有必要的。

(2) 各因子的 p 值

因子的 p 值是四个因素中最主要的因素。但对于不同数量的因子和处在不同 p 值时, p 的作用也不一样, 从表 1 就可见一斑。

(3) 因子间 p 值的差异程度

这个问题以前一直没有引起重视。本文的分析结果表明, 因子之间 p 值的差异程度越小越好, 否则会影响综合预报的效果。另外从(3)式告诉我们 p 值小于 0.5 的因子是起负贡献的, 在综合时应该去掉这种因子。亦即 p 值小于 0.5 的因子不适宜使用多因子综合方法, 但可以先浓缩因子, 使之 p 值超过 0.5, 再进行多因子综合。

(4) 预报原则

优秀的预报原则的确定并非容易, 因为它受到前面 3 个因素的综合影响, 这还有待于深入研究。

本文仅对 0、1 两值预报问题进行探讨, 其实对多值问题也有帮助。如在因子历史概括率一致时, 对于三值问题, (3) 式只须把 p 看作 p_1 , p' 看作 $p_2 + p_3$, 就得到了预报第一级出现可能性的公式, 以此类推, 四值问题、五值问题等也可适用。要使综合后提高预报效能, 仍要求预报级的因子 p 值大于 0.5。

本文提出的多因子两值预报效能取决于四个方面因素的结论只是分析了投票法后得

到的,是否普遍适用于其它两值预报尚待进一步弄清.

致谢: 本所王爱平同志作了部分计算工作,在此表示感谢.

参 考 文 献

- [1] 王宗皓、李麦村,大气预报中的概率统计方法,科学出版社,1978.
[2] 复旦大学数学系主编,概率论与数理统计,上海科学技术出版社,1962.

附录 关于(3)式的证明

$$P_m = p^m + C_m^1 p^{m-1} p' + C_m^2 p^{m-2} p'^2 + \dots + C_m^{\frac{m-1}{2}} p^{\frac{m+1}{2}} p'^{\frac{m-1}{2}} \quad (1)$$

$$P_m = p + p p' (p - p') (C_1^1 + C_3^1 p p' + C_5^1 p^2 p'^2 + \dots + C_{\frac{m-3}{2}}^{\frac{m-3}{2}} p^{\frac{m-3}{2}} p'^{\frac{m-3}{2}}) \quad (2)$$

(当 $m \geq 3$ m 为奇数)

求证(1)=(2)

1. $m=3$ 时,(1)式为

$$\begin{aligned} P_3 &= p^3 + C_3^1 p^2 p' = p^3 - p^2 p' + 3p^2 p' = p - p p' + 2p^2 p' \\ &= p - p p' (1 - 2p) = p + p p' (p - p') = (2) \end{aligned}$$

证毕

2. $m=5$ 时

因为从(2)式

$$\begin{aligned} P_5 - P_3 &= p + p p' (p - p') (1 + C_3^1 p p') - [p + p p' (p - p')] \\ &= C_3^1 p^2 p'^2 (p - p') \end{aligned}$$

故只须从(1)式证明 $P_5 - P_3 = C_3^1 p^2 p'^2 (p - p')$ 即可. 从(1)式

$$\begin{aligned} P_5 - P_3 &= p^5 + C_5^1 p^4 p' + C_5^2 p^3 p'^2 - (p^3 + C_3^1 p^2 p') \\ &= p^5 + 5p^4 p' + 10p^3 p'^2 - (p^3 + p^2 p' + 3p^2 p') \\ &= p^5 + 5p^4 p' + 10p^3 p'^2 - (p^3 + p^2 p' + p^2 p' + p^2 p'^2 \\ &\quad + 3p^2 p' + 3p^2 p'^2 + 3p^2 p'^2 + 3p^2 p'^2) \\ &= p^5 + 5p^4 p' + 10p^3 p'^2 - (p^3 + 5p^2 p' + 7p^2 p'^2 + 3p^2 p'^3) \\ &= 3p^3 p'^2 - 3p^2 p'^3 \\ &= C_3^1 p^2 p'^2 (p - p') \end{aligned}$$

证毕

3. $m=n$ 时

从(2)

$$P_n - P_{n-1} = C_{n-1}^{\frac{n-1}{2}} p p' (p - p') p^{\frac{n-3}{2}} p'^{\frac{n-3}{2}}$$

从(1)式

$$\begin{aligned} P_n - P_{n-1} &= (p^n + C_n^1 p^{n-1} p' + C_n^2 p^{n-2} p'^2 + C_n^3 p^{n-3} p'^3 \\ &\quad + C_n^4 p^{n-4} p'^4 + \dots + C_n^{\frac{n-1}{2}} p^{\frac{n+1}{2}} p'^{\frac{n-1}{2}}) \\ &\quad - (p^{n-1} + C_{n-1}^1 p^{n-2} p' + C_{n-1}^2 p^{n-3} p'^2 + C_{n-1}^3 p^{n-4} p'^3 \\ &\quad + C_{n-1}^4 p^{n-5} p'^4 + \dots + C_{n-1}^{\frac{n-1}{2}} p^{\frac{n-1}{2}} p'^{\frac{n-1}{2}}) \\ &= (p^n + C_n^1 p^{n-1} p' + C_n^2 p^{n-2} p'^2 + C_n^3 p^{n-3} p'^3 \\ &\quad + C_n^4 p^{n-4} p'^4 + \dots + C_n^{\frac{n-1}{2}} p^{\frac{n+1}{2}} p'^{\frac{n-1}{2}}) \\ &\quad - (p^{n-1} + 2p^{n-1} p' + p^{n-2} p'^2 + C_{n-2}^1 p^{n-1} p' \\ &\quad + 2C_{n-2}^1 p^{n-2} p'^2 + C_{n-2}^2 p^{n-3} p'^3 + C_{n-2}^3 p^{n-2} p'^2 \\ &\quad + 2C_{n-2}^3 p^{n-3} p'^3 + C_{n-1}^1 p^{n-4} p'^4 + C_{n-1}^2 p^{n-3} p'^3 \\ &\quad + 2C_{n-1}^2 p^{n-4} p'^4 + C_{n-1}^3 p^{n-5} p'^5 + C_{n-1}^4 p^{n-4} p'^4 \\ &\quad + 2C_{n-1}^4 p^{n-5} p'^5 + C_{n-1}^5 p^{n-6} p'^6 + \dots + C_{n-1}^{\frac{n-1}{2}} p^{\frac{n+1}{2}} p'^{\frac{n-1}{2}}) \\ &\quad + 2C_{n-1}^{\frac{n-1}{2}} p^{\frac{n+1}{2}} p'^{\frac{n-1}{2}} p' + C_{n-1}^{\frac{n-1}{2}} p^{\frac{n+1}{2}} p'^{\frac{n-1}{2}} \\ &= (p^n - p^n) + (C_n^1 - 2 - C_{n-1}^1) p^{n-1} p' + (C_n^2 - C_{n-1}^2) \end{aligned}$$

$$\begin{aligned}
& -2C_{n-2}^1 - 1)p^{n-2}p'^2 + (C_n^1 - C_{n-2}^1 - 2C_{n-2}^2 \\
& - C_{n-1}^1)p^{n-3}p'^3 + (C_n^2 - C_{n-2}^2 - 2C_{n-1}^2 \\
& - C_{n-2}^3)p^{n-4}p'^4 + \dots + (C_{\frac{n-3}{2}}^{\frac{n-3}{2}} - C_{\frac{n-3}{2}}^{\frac{n-3}{2}} - 2C_{\frac{n-3}{2}}^{\frac{n-5}{2}} \\
& - C_{\frac{n-2}{2}}^{\frac{n-1}{2}})p^{\frac{n+3}{2}}p'^{\frac{n-3}{2}} + (C_{\frac{n-1}{2}}^{\frac{n-1}{2}} - 2C_{\frac{n-2}{2}}^{\frac{n-3}{2}} \\
& - C_{\frac{n-1}{2}}^{\frac{n-5}{2}})p^{\frac{n+1}{2}}p'^{\frac{n-1}{2}} - C_{\frac{n-2}{2}}^{\frac{n-3}{2}}p^{\frac{n-1}{2}}p'^{\frac{n+1}{2}}
\end{aligned}$$

$\therefore C_{\frac{n-1}{2}}^{\frac{n-1}{2}} = C_{\frac{n-1}{2}}^{\frac{n-1}{2}} - \frac{n-1}{2} = C_{\frac{n-2}{2}}^{\frac{n-3}{2}}, (C_{\frac{n-3}{2}}^{\frac{n-3}{2}} - C_{\frac{n-2}{2}}^{\frac{n-5}{2}} - 2C_{\frac{n-2}{2}}^{\frac{n-5}{2}} - C_{\frac{n-1}{2}}^{\frac{n-3}{2}}) = 0$
 类似各项均为 0, 故上式

$$\begin{aligned}
& = C_{\frac{n-2}{2}}^{\frac{n-3}{2}}p^{\frac{n+1}{2}}p'^{\frac{n-1}{2}} - C_{\frac{n-2}{2}}^{\frac{n-3}{2}}p^{\frac{n-1}{2}}p'^{\frac{n+1}{2}} \\
& = C_{\frac{n-2}{2}}^{\frac{n-3}{2}}pp'(p - p')p^{\frac{n-3}{2}}p'^{\frac{n-3}{2}}
\end{aligned}$$

证毕

A PRELIMINARY STUDY OF THE FORECASTING EFFICIENCY OF THE MULTI-FACTOR BI-VALUE METHOD

Chang Venjan

(Institute of Meteorology, Zhejiang Province)

Abstract

There are already many methods for the Multi-factor Bi-value forecast, such as Coding method, REEP method, etc. But all of these methods are not related to the intrinsic law of Bi-value forecast. This paper investigates it from the viewpoint of probability, and a method for probability estimate of forecasting efficiency is proposed.

The mathematical expressions have been obtained using the simplified voting method. With the conditions under these expressions which the Multi-factor forecast is superior to the Mono-factor forecast are discussed. It is found that the efficiency of Multi-factor forecast depends upon four elements: (1) number of factors; (2) historical fitting rate of each factor; (3) the difference between these fitting rates; and (4) principle of forecast, etc. In addition, it is found that when a factor's historical fitting rate is smaller than 0.5, it gives a negative effect.