事件概率的离散似然估计

刘健文　王炳仁

（空军气象研究所）

一、引　言

应用统计方法时，我们经常需要从基于一组相同预报因子的大量可能事件中选择其中之一来作预报。但预报因子中包含的信息量往往有限，难以对之作出精确的描述。对这类问题气象上认为最合适的处理是给出各种可能事件的概率分布。

计算概率分布我们目前更多的是采用经典的回归方法。由于其假设前提是变量遵从于正态分布，因此计算结果一般趋向于平均状况，对突变性天气往往缺乏预报能力。本文介绍的离散似然函数方法，是近年国外在非线性统计预报方面的一个新尝试。

二、方　法

离散似然函数（Discrete Likelihood Function）方法主要基于 Bayes 逆概率定理。假定我们观测到一组当前天气状态，问“这组当前天气状态演变为有雨过程的可能性如何，或者相反地，演变为无雨过程的可能性又如何？”则由 Bayes 定理，这两种可能性可以通过倍乘每一先前状态的条件概率，得到全部观测项目的联合概率而求出。倍乘的作用将会使与我们所着眼事件（比如雨）不一致的先前状态对事件的似然性产生显著影响。

设存在 G 个相互独立、完备的事件（预报量）。预报因子个数为 R，用矢量 $\underline{x}\equiv[x_1,\cdots,x_R]$ 来表示。根据贝叶斯定理，事件的后验概率为：

$$B(g\mid\underline{x})=\frac{f(\underline{x}\mid g)\cdot q(g)}{\sum_{g=1}^{G}f(\underline{x}\mid g)\cdot q(g)} \quad (1)$$

$$(g=1,2,\cdots,G)$$

式中 $q(g)$ 是组 g 的先验概率。$f(\underline{x}\mid g)$ 是给定 g 时观测值 $\underline{x}$ 的联合概率密度函数，即似然函数。

似然函数 $f(\underline{x}\mid g)$ 可以表示成 $\underline{x}$ 中各个元素条件概率的乘积：

$$\begin{aligned}f(\underline{x}\mid g)=&f(x_1\mid g)\cdot f(x_2\mid g,x_1)\cdot\\&\cdots\cdot f(x_r\mid g,x_1,\cdots,x_{r-1})\cdot\\&\cdots\cdot f(x_R\mid g,x_1,\cdots,x_{R-1})\\&(g=1,2,\cdots,G)\end{aligned} \quad (2)$$

假设每个 x_r 有离散的分布（连续预报因子可以通过积分进行离散化）。用 Z 表示 S_r 个离散构造，则（2）式右端可以写成下列形式：

$$\left\{\begin{aligned}&f(x_1\mid g)=P(x_1=Z_{1,i}\mid g)\\&\qquad\qquad i=1,\cdots,S_1\\&f(x_2\mid g,x_1)=P(x_2=Z_{2,i}\mid g,x_1)\\&\qquad\qquad i=1,\cdots,S_2\\&\qquad\vdots\\&f(x_r\mid g,x_1,\cdots,x_{r-1})=P(x_r=\\&\quad Z_{r,i}\mid g,x_1,\cdots,x_{r-1})\\&\qquad\qquad i=1,\cdots,S_r\\&\qquad\vdots\\&f(x_R\mid g,x_1,\cdots,x_{R-1})=P(x_R=\\&\quad Z_{R,i}\mid g,x_1,\cdots,x_{R-1})\\&\qquad\qquad i=1,\cdots,S_R\\&\qquad\qquad(g=1,\cdots,G)\end{aligned}\right. \quad (3)$$

这里 P 表示条件概率，它可以用其他方法估计，如 REEP 等。对一般项 r，这些估计值的形式为：

$$\begin{aligned}\hat{P}(x_r\mid g,x_1,\cdots,x_{r-1})=&A_{0r}+A_{1r}x_1\\&+\cdots+A_{r-1}x_{r-1}\end{aligned} \quad (4)$$

因此，由每个特定的Z表示的x均有唯一的离散值。如果全部Z转化成非0即1的J个伪变量，则(4)可写成最终形式：

$$\hat{P}(Z_{j=1}|g,z_1,\cdots,z_{j-1})=A'_{0j}+A'_{1j}Z_1+\cdots+A'_{j-1}Z_{j-1} \quad (5)$$
$$(j=1,\cdots,J;g=1,\cdots,G)$$

G中每一组的方程个数J等于原有x的离散类型总数减去R，因为x的划分是完备的，每个x中有一个Z计算冗余。

对具体预报，预报因子的观测值为$Z=[Z_1,Z_2,\cdots,Z_j,\cdots,Z_J]$。给定$g$时这种观测值的估计似然函数为：

$$\hat{f}(\underline{Z}|g)=\hat{P}(Z_1=z_1|g)\cdot\hat{P}(Z_2=z_2|g,Z_1)\cdot\cdots\cdot\hat{P}(Z_j=z_j|g,Z_1,\cdots,Z_{j-1})\cdot\cdots\cdot\hat{P}(Z_J=z_J|g,z_1,\cdots,Z_{J-1}) \quad (6)$$
$$(g=1,2,\cdots,G)$$

在这里，一般项$P(Z_j=z_j|g,Z_1,\cdots,Z_{j-1})$是给定$g$和$Z_1,\cdots,Z_{j-1}$时实际观测$Z_j=z_j$的概率估计。其中假定经过类似REEP的处理后条件概率仍为独立。

先验概率可以从样本的频数$n_g(g=1,\cdots,G)$中确定，即：

$$q(g)=n_g/N \quad (g=1,\cdots,G) \quad (7)$$

这里 $N=\sum_{g=1}^{G}n_g$

这样我们就可以求出期望的后验概率。对给定的观测值$\underline{Z}$，用(6)和(7)求出$f(\underline{Z}|g)$和$q(g)$的近似值，代入(1)，得到：

$$\hat{B}(g|\underline{Z})=\frac{\hat{f}(\underline{Z}|g)\cdot\hat{q}(g)}{\sum_{g=1}^{G}\hat{f}(\underline{Z}|g)\cdot\hat{q}(g)} \quad (8)$$
$$(g=1,\cdots,G)$$

三、计算实例

我们以一个简单的例子，来说明计算和预报的全过程。

表1是某次试验的数据集。预报对象有3组($G=3$)。预报因子为x_1、x_2、x_3，其中x_1、

表1 试验数据集（每个x中的冗余已经略去）

类型 i	预报因子 x_1 z_1	x_1 z_2	x_2 z_3	x_2 z_4	x_3 z_5	预报量 $g=1$ n_{1i}	$g=2$ n_{2i}	$g=3$ n_{3i}	合计 N_i
1	0	0	0	0	0	2	3	10	15
2	0	0	0	0	1	3	2	10	15
3	0	0	0	1	0	2	12	3	17
4	0	0	0	1	1	14	3	3	20
5	0	0	1	0	0	2	2	10	14
6	0	0	1	0	1	14	3	3	20
7	0	1	0	0	0	14	2	3	19
8	0	1	0	0	1	3	11	3	17
9	0	1	0	1	0	2	12	3	17
10	0	1	0	1	1	2	11	2	15
11	0	1	1	0	0	3	2	10	15
12	0	1	1	0	1	2	3	10	15
13	1	0	0	0	0	2	12	2	16
14	1	0	0	0	1	3	2	10	15
15	1	0	0	1	0	2	12	2	16
16	1	0	0	1	1	14	3	3	20
17	1	0	1	0	0	2	2	10	14
18	1	0	1	0	1	14	3	3	20
（合计）						100	100	100	300

x_2 划分为 3 级，x_3 划分为 2 级，即 $S_1=S_2=3, S_3=2$。每组有 100 个观测值，即 $n_1=n_2=n_3=100, N=300$。观测值分属各组共有 18 种可能的组合。

从表 1 中统计出各组的平方和及叉积矩阵，再通过 REEP 法计算，则可确定方程(5)。

解回归方程，对每组得到以下结果：

第一组：

$$\begin{cases}\hat{P}(Z_1=1|1)=0.37000\\ \hat{P}(Z_2=1|1,Z_1)=0.41270\\ \quad -0.41270\,Z_1\\ \hat{P}(Z_3=1|1,Z_1,Z_2)=0.43243\\ \quad +0.00000\,Z_1-0.24013\,Z_2\\ \hat{P}(Z_4=1|1,Z_1,Z_2,Z_3)=0.71696\\ \quad +0.00000\,Z_1-0.43658Z_2\\ \quad -0.65796\,Z_3\\ \hat{P}(Z_5=1|1,Z_1,Z_2,Z_3,Z_4)\\ \quad =0.60222+0.00000\,Z_1\\ \quad -0.42706\,Z_2+0.26630\,Z_3\\ \quad +0.27856\,Z_4\end{cases}\tag{9}$$

第二组：

$$\begin{cases}\hat{P}(Z_1=1|2)=0.34000\\ \hat{P}(Z_2=1|2,Z_1)=0.62121\\ \quad -0.62121\,Z_1\\ \hat{P}(Z_3=1|2,Z_1,Z_2)=0.20000\\ \quad -0.05294\,Z_1-0.07805\,Z_2\\ \hat{P}(Z_4=1|2,Z_1,Z_2,Z_3)=0.72660\\ \quad -0.19234\,Z_1-0.08843\,Z_2\\ \quad -0.63301\,Z_3\\ \hat{P}(Z_5=1|2,Z_1,Z_2,Z_3,Z_4)\\ \quad =0.39502-0.10456\,Z_1\\ \quad +0.29410\,Z_2+0.14180\,Z_3\\ \quad -0.17230\,Z_4\end{cases}\tag{10}$$

第三组：

$$\begin{cases}\hat{P}(Z_1=1|3)=0.30000\\ \hat{P}(Z_2=1|3,Z_1)=0.44286\\ \quad -0.44286\,Z_1\\ \hat{P}(Z_3=1|3,Z_1,Z_2)=0.33333\\ \quad +0.10000\,Z_1+0.31183\,Z_2\\ \hat{P}(Z_4=1|3,Z_1,Z_2,Z_3)=0.26038\\ \quad +0.04478\,Z_1+0.10711\,Z_2\\ \quad -0.31960\,Z_3\\ \hat{P}(Z_5=1|3,Z_1,Z_2,Z_3,Z_4)\\ \quad =0.52995+0.15463\,Z_1\\ \quad +0.16775\,Z_2-0.29875\,Z_3\\ \quad -0.13069\,Z_4\end{cases}\tag{11}$$

如果取观测值为 $z=[00000]$的组合，则由(6)有：

$$\hat{f}(\underline{Z}|g)=\hat{P}(Z_1=0|g)\cdot\hat{P}(Z_2=0|g,Z_1)\cdot\hat{P}(Z_3=0|g,Z_1,Z_2)\cdot\hat{P}(Z_4=0|g,Z_1,Z_2,Z_3)\cdot\hat{P}(Z_5=0|g,Z_1,Z_2,Z_3,Z_4)\quad(g=1,2,3)\tag{12}$$

将(9)、(10)、(11)代入，得到：

$$\begin{cases}\hat{f}(\underline{Z}|1)=(0.63000)\cdot(0.58730)\cdot(0.56757)\cdot(0.28304)\cdot(0.39778)\\ \quad =0.02364\\ \hat{f}(\underline{Z}|2)=(0.66000)\cdot(0.37879)\cdot(0.80000)\cdot(0.27340)\cdot(0.60498)\\ \quad =0.03308\\ \hat{f}(\underline{Z}|3)=(0.70000)\cdot(0.55714)\cdot(0.66667)\cdot(0.73962)\cdot(0.47005)\\ \quad =0.09039\end{cases}$$

由(7)估计的先验概率为：

$$\begin{cases}q(1)=100/300=0.33333\\ q(2)=100/300=0.33333\\ q(3)=100/300=0.33333\end{cases}\tag{14}$$

把(13)和(14)代入(8)，得到：

$$\hat{B}(1|\underline{x})=\frac{(0.02364)\cdot(0.33333)}{(0.02364)\cdot(0.33333)+(0.03308)\cdot(0.33333)+(0.09039)\cdot(0.33333)}=0.16070$$

$$\hat{B}(2|\underline{x})=\frac{(0.03308)\cdot(0.33333)}{(0.02364)\cdot(0.33333)+(0.03308)\cdot(0.33333)+(0.09039)\cdot(0.33333)}=0.22487$$

$$\hat{B}(3|\underline{x})=\frac{(0.09039)\cdot(0.33333)}{(0.02364)\cdot(0.33333)+(0.03308)\cdot(0.33333)+(0.09039)\cdot(0.33333)}$$
$$=0.61444$$

同样地，我们可求出另外17种组合的后验概率。所有结果列于表2。

我们可以把上述从似然函数求出的后验概率直接用于预报。对于任一组观测数据，从表2得到各类别的预报概率，再由最大概率或临界阀值等预报原则，即可得出具体的类型预报。

表2 18种组合类型的计算结果

项目 / 类型	$\hat{B}(1\|x)$	$\hat{f}(x\|1)$	$\hat{B}(2\|x)$	$\hat{f}(x\|2)$	$\hat{B}(3\|x)$	$\hat{f}(x\|3)$
1	0.161	0.024	0.225	0.033	0.614	0.090
2	0.225	0.036	0.136	0.022	0.640	0.102
3	0.105	0.018	0.658	0.113	0.237	0.041
4	0.691	0.133	0.169	0.032	0.141	0.027
5	0.135	0.020	0.143	0.021	0.722	0.106
6	0.700	0.131	0.130	0.022	0.170	0.032
7	0.670	0.125	0.218	0.041	0.113	0.021
8	0.161	0.027	0.525	0.090	0.295	0.029
9	0.200	0.032	0.691	0.111	0.109	0.018
10	0.159	0.027	0.705	0.119	0.136	0.023
11	0.239	0.039	0.052	0.009	0.710	0.115
12	0.206	0.031	0.280	0.041	0.514	0.076
13	0.151	0.024	0.612	0.096	0.238	0.037
14	0.230	0.036	0.252	0.039	0.519	0.081
15	0.101	0.018	0.769	0.137	0.130	0.023
16	0.738	0.133	0.102	0.018	0.160	0.029
17	0.150	0.020	0.236	0.031	0.614	0.081
18	0.637	0.131	0.116	0.024	0.248	0.085

四、应用及讨论

离散似然函数（以下简称DLF）方法目前在美国主要用于两个方面的业务预报。一是技术发展实验室（TDL）向全国发布的降水类型条件概率（POPT）18小时预报，对象是雨、冻雨、雪，它使用模式输出统计结果；另一是俄亥俄州的异常天气（NOWX）3小时预报，它仅用单站地面观测资料。此外，正尝试用于短时天气预报系统GEM中。

表3给出了在POPT和NOWX预报中

表3 DLF和其他方法的预报效果对比（引自文献[2]）

项目 / 方法 \ 变量性质		不独立		独立	
	统计量	布赖尔得分	报对数	布赖尔得分	报对数
POPT	REEP	0.112	15153	0.095	15389
	LOGIT	0.111	15137	0.095	15389
	DLF	0.107	15177	0.097	15108
	UDLF	0.105	15183	0.094	15402
NOWX	REEP	0.292	19603	0.270	24079
	DLF	0.285	19677	0.262	24271
	OPTIMUM	0.281	19745	0.265	24214

DLF 方法和 REEP、LOGIT、OPTIMUM 等其他方法的可比结果。统计量为布赖尔得分和报对数。DLF 和 REEP 使用 0—1 型预报因子，LOGIT 的因子为连续形式。预报以最高概率为选择准则。表中的 UDLF 代表无偏 DLF，因为 DLF 的估计不一定是无偏的，作为补偿，DLF 中的回归方程可以用全部原有因子来决定，且 DLF 概率作为独立变量，这样可以确保最终的 DLF 概率对非独立变量也是无偏的。

从表 3 结果中可以看出，无论是对独立样本还是非独立样本，DLF 方法在两个预报项目上均改善了布赖尔得分和准确率，特别是独立样本的 NOWX 预报，提高幅度较大。业务应用表明，DLF方法对于突变性天气，如锋面过境及伴随的天气、云幂和能见度的变化等的预报能力较诸如回归这样的经典统计方法要强。

在设计上，DLF 方法比直接回归的 REEP 方法复杂一些，但比有相似之处的多元判别分析(MDA)简单。不过，它的结果要优于 REEP 和 MDA(REEP 和 MDA 在预报效果上被认为是等效的)。REEP 的缺点在于相互作用式中，预报因子间缺乏相加性，相比之下 DLF 利用了所有预报因子一阶相互作用的信息。试验证实如果在 REEP 中考虑所有一阶相互作用项，则其与 DLF 在布赖尔得分上几乎一致。在处理上，REEP 采用了组频率对元素施加了权重的协方差矩阵，而 DLF 则由于先验概率照顾了组频率差异可以等效地对待每一个矩阵。

跟 0—1 预报的布尔处理相同，DLF不需要计算所有布尔之间的叉积和，这是一大优点。此外，气象上目前主要还是把资料按季节、时间和天气条件的不同进行分类，如相似环流型等，但这种划分大部分并没有很好地改进预报；而 DLF 在数据分层时不是按预报因子的初始条件，而是按预报量条件，体现出其抓住了一定类型的非线性性。

如何把 DLF 这一非线性统计预报方法更有效地运用于预报业务，目前还在进一步探索之中。它在数据分析中无疑是反映相互作用信息的一个有价值的工具。总的看来，这一以似然函数为基础的事件选择方法可以使似然性在分类预报上发挥最合适的作用。

参 考 文 献

[1] J. R. Bocchieri, A new operational system for forecasting precipitation type, Mon. Wea. Rev., 107, 1979.

[2] R. G. Miller, Estimating event probabilities by discrete likelihood functions. Sixth Conference on Probability and Statistics in Atmospheric Sciences, Banff, Amer. Meteor. Soc., 93-97, 1979.

[3] R. G. Miller, Very short range statistica forecasting of automated weather observations Ninth Conference on Probability and Statistics in Atmospheric Sciences, Banff, Amer. Meteor Soc., 388-394, 1985.

[4] M. M. Tatsouka, The relationship between canonical correlation and discriminant analysis, and a proposal for utilizing qualitative data in discriminant analysis, Graduate school of Education. Havard University Thesis Proposal, 47pp.