天气预报多元回归模型中模糊因子的集对分析

王国强[1],赵克勤[2],郑选军[1]
(1. 浙江省绍兴市气象台 浙江 绍兴 312000;2. 浙江省诸暨市联系数学研究所 浙江 诸暨 311811)

摘 要:天气预报模型中的预报因子一般都有较好的预报能力,因为它们是根据预报对象的特点、预报因子的物理意义和预报经验,使用一定的技术方法而精心筛选的. 但是预报因子在每次天气预报中的性能表现有清晰和模糊之分. 从集合预报和近邻估计两种方法的基本思路出发,定义了描述因子模糊性的统计量——变异系数,并用此来识别多元回归模型中的模糊因子;从集对分析的基本理论出发,推导了适合于多元回归分析的联系度公式,并借此来处理模糊因子. 在此基础上建立基于集对分析的天气预报多元回归模型,比传统模型明显地提高了天气预报准确率.

关键词:大气科学;天气预报;多元回归;模糊因子;集对分析
中图分类号:0212 **文献标识码**:A **文章编号**:1001 - 7119(2004)02 - 0151 - 05

Application of Set Pair Analysis to Fuzzy Predictors of Multiple Regression Weather Forcast Models

WANG Guo-qiang[1], *ZHAO Ke-qin*[2], *ZHENG Xuan-jun*[1]
(1. Shaoxing Meteorological Observatary, Shaoxing 312000, China; 2. Zhuji Institute of Contact Mathematics, Zhuji 311811, China)

Abstract: The predictors of weather forecast models have better forecast ability in general. Because they are specially selected according to the characteristics of predictands, physics meaning of predictors and forecast experience, and by means of certain technical methods. But sometimes when the situation has changed, predictors of multiple regression weather forcast models may appear in clear or fuzzy state in every forcasting. The fuzzy predictors can be identified and handled correctly by means of theory of Set Pair Analysis and method of variability coefficient. Forecast accuracys of multiple regression weather forcast models based on Set Pair Analysis are raised obviously.

Key Words: atmospheric science; weather forcasting; multiple regression; fuzzy predictor; Set Pair Analysis

0 引 言

天气系统是一种复杂的非线性开放系统,具有随机性、模糊性等不确定性. 人们通常采用多元回归模型进行天气预报,在多数情况下得到较为满意的预报结果,但也不乏预报失败的例子. 究其原因,除对天气系统自身运动机理没有完全搞清外,对天气系统变化的模糊性认识不足是重要的原因之一. 在预报工作中,对天气系统模糊性的认识集中体现在对回归模型中模糊因子的认识和处置上. 经过对众多预报失败个例的研究,我们发现传统回归模型中把模糊因子与其它预报性能良好的清晰因子相提并论,是导致预报失败的一个重要原因.

为此我们自 1996 年起,应用集对分析理论[1]和变异系数方法识别和处置天气预报回归模型中

收稿日期:2003 - 02 - 27
基金项目:浙江省气象局科研课题(200118)
作者简介:王国强,男,1944 年生,浙江宁波人,高级工程师.

的模糊因子,在此基础上建立了基于集对分析的天气预报多元回归模型.经几年来的预报实践证明,新模型比传统的天气预报多元回归模型的预报准确率明显提高[2].

1 天气预报多元回归模型中的模糊因子问题

传统的天气预报多元回归模型是一种成熟的数学模型,在建立这种多元回归模型时,一般需要大量收集与预报量有关的要素,需要细致分析预报量与预报因子的关系,从中初选出一批专业意义明确的待选因子,再用适当的数理统计方法精选出一组预报因子(所谓"最佳因子集合"),建立回归模型.因子集合中的每一因子一般具有良好的预测能力,因子之间也具有良好的合作预测能力.但在回归模型中因子集合的所谓良好性能,实际上是对于整个样本或次数众多的预测而言,是个整体的概念.对于样本中的某一个例或者应用中的某一次预测则并非完全如此,有时一个或几个因子的表现不好能导致一次预测的失败.

根据对 34 次天气预报失败的个例分析,其中由于一个预报因子表现模糊造成预报失败的情况占 2/3 以上,而且导致预报失败的因子并不是固定的.这种情况揭示了一个事实,即在用多元回归模型进行天气预报时,模型中各因子的相对重要性不断变化,一些因子在这次预报中表现出相对重要性,具有清晰的预报能力,而在另一次预报中的预报能力则显得模糊不清,甚至可能起负作用.我们称前面情况下的因子为清晰因子或预报性能清晰,称后面情况下的因子为模糊因子或预报性能模糊.这种现象给我们提出了一个有意义的问题:假设有一个多元回归模型,因变量为一维随机变量 Y,自变量为 m 维变量 X.在某次预测中有 p 个自变分量性能模糊,那么能否设法使这 p 个自变分量在这次预报中少发挥或者不发挥作用,而主要由其余 $(m-p)$ 个性能清晰的自变分量来决定模型的预测结论呢?同样地,在另一次预测中,若有另外 q 个自变分量性能模糊,能否使这 q 个自变分量少起作用或者不发挥作用,而由其余 $(m-q)$ 个性能清晰的因子去决定预报的结论,这就是天气预报多元回归模型中的模糊因子问题.

2 模糊性的定量描述和模糊因子的识别

事物的模糊性是客观存在的,对事物模糊性的数学表达有着一定的复杂性.为了对天气预报多元回归模型中因子的模糊性进行描述,可以先看看集合预报(ensemble forecasting)的观点和近邻估计(nearest neighbours estimation)的思想.

天气预报理论认为,从差别甚微的 n 个初始场出发,通过预报模式的积分,得到的 n 个预报结果应该差别"甚微".但大量的实践表明,这 n 个预报结果可能散发到较大区域,这是由天气预报的不确定性所引起.大气运动同时具有确定性和不确定性的双重特性,是人们在天气预报的长期实践中得到的辩证认识.大气运动的随机性,以及初始场和模式的不完全性导致了天气预报的不确定性.另外可以用不同的预报模式分别对于 n 个初始场作预报,并认为发散越小,预报模式的质量越好.发散的程度表示了预报不确定性的大小.这就是集合预报的观点[3].

近邻估计属于非参数回归(nonparametric regression)[4].近邻估计要在样本中为估计点 X 找到最为相近的 k 个个例,记为 X_i, Y_i(其中 $i=1,2,\cdots,k$),此时 k 个近邻的 X_i 是 k 个初始场,k 个近邻的 Y_i 是 k 个预测值.如果 k 近邻的 Y_i 分布集中,则表示用自变量 X 去预测 Y 的效果就好;反之如果 k 个 Y_i 的分布分散,则表示用 X 去预测 Y 的效果就差.k 个近邻的 Y_i 分布是集中或分散的程度称离散度,可以用标准差或方差来描述.考虑不同单位和不同数量级别的两组数字的离散度比较,则可采用标准差除以平均值所得到的变异系数(coefficient of variability)来表示,如果用 s 表示标准差,聊表示均值,则变异系数 C_v 表示为

$$C_v = \frac{s}{聊} \qquad (1)$$

变异系数是对因子模糊性的量度.利用(1)式可以对回归模型中每一因子的各个例分别做近邻分析,找出 k 个近邻 $(X_{(1)}, Y_{(1)})$, $(X_{(2)}, Y_{(2)})$, $(X_{(3)}, Y_{(3)})$, ……, $(X_{(k)}, Y_{(k)})$,对 K 个 Y 量 $Y_{(1)}$, $Y_{(2)}$, $Y_{(3)}$, ……, $Y_{(k)}$ 计算变异系数.在每一个例中,对 m 个因子计算出 m 个变异系数,变异系数取最大值并大于某一定值时,可认为该因子是模糊因

子.

这里以浙江省绍兴市气象台梅汛期雨量预报为例来说明模糊因子的识别过程. 模型中涉及样本的容量为 317,预报因子为 6 个,整个分析与判断过程分两步进行. 第一步,对由数值预报产品格点资料组合而成的因子进行天气学和统计学方面的分析加工;第二步,根据各个例中每个因子的 C_v 值来判断因子是否处于模糊状态. 表 1 列出了各因子各个例的变异系数 C_v,计算时取近邻数为 $k = 7$[5]. 对表中变异系数的进一步分析可知,当某因子的 C_v 值为 6 个因子 C_v 值中的最大,且它的 $C_v \geq 6.7$ 时,该因子的预报性能模糊. 如在个例 1 中因子 x_6 的变异系数 $C_v = 31.6$ 为 6 个因子中的最大值,并大于等于 6.7,表明因子此时 x_6 是模糊因子;而在个例 2、个例 3 和个例 4 中,可以看到因子 x_6、x_3 和 x_2 分别是模糊因子;分析个例 5 可以发现此时没有模糊因子,其余个例可类推.

表 1　各预报因子的变异系数

Table 1　The variability coefficients of every forecasting factor

个例序号	x_1	x_2	x_3	x_4	x_5	x_6
001	4.2	12.5	6.4	3.4	3.3	31.6
002	6.7	6.1	4.5	4.3	8.2	31.6
003	3.1	4.7	7.4	2.1	4.6	1.9
004	7.8	17.9	9.0	7.2	9.9	8.1
005	0.1	0.2	0.2	0.1	0.1	0.0
006	4.8	8.3	4.3	2.1	3.9	1.8
007	5.2	4.1	2.2	4.2	4.4	0.0
008	6.5	5.5	7.2	4.3	2.2	31.6
009	6.9	4.9	9.7	6.5	3.7	4.2
010	6.8	8.4	4.2	3.1	2.8	1.9
…	…	…	…	…	…	…
317	5.4	7.7	6.7	4.6	3.3	3.0

3　模糊因子的集对分析

不确定性是自然界和人类社会中普遍存在的一种客观现象. 为描述和分析这种不确定性,人们提出了研究模糊、随机、中介和信息不完全所致的不确定性的系统理论和方法,即集对分析(Set Pair Analysis,SPA). 该方法的基本思想就是对客观存在的种种不确定性予以承认,并把不确定性放入一个既确定又不确定的同异反系统进行辩证分析和数学处理.

所谓"集对"是指具有一定联系的两个集合所组成的对子. SPA 认为,对于集对中两个集合的若干特征可以作同异反系统分析,其联系度 $\mu(w)$ 可定量描述为

$$\mu(w) = a + b_i + c_j \qquad (2)$$

式中:同一度 $a = \frac{s}{n}$,差异度 $b = \frac{f}{n}$,对立度 $c = \frac{p}{n}$. 式(2) 表示在命题 w 下对某集对作分析,它们共有 n 个特性,其中 s 个特性为两个集合所共有,p 个特性为两个集合相对立,而 f 个特性则表现为既不对立又不同一,且有 $n = s + p + f$. 式中的 i 和 j 既是差异度和对立度的标记,同时又可给予赋值,以便对联系度进行计算. 计算时取 $j = -1$,而 i 的取值则根据命题的专业特点而定,其变化区间为 $[-1,1]$. 关于集对分析对不确定性问题的描述和计算可参阅文献[6].

为了表述简便起见,暂时把多元回归模型中的某一因子记为 $x_i (i = 1,2,\cdots,n)$,记 x_i 中的最大值为 $x_{\max}$,把第 i 个例的 x_i 与 $x_{\max}$ 组成集对. 对集对进行对比分析可知,当用 x_i 去预测 y_i 时,可以给出的预测 y_i 是一个概率分布,它的均值可作为实际使用中的预测值. 如果预测的不确定性较小,则同一度 $a = \frac{s}{n} = \frac{x_i}{x_{\max}}$,对立度 $c = \frac{p}{n} = \frac{(x_{\max} - x_i)}{x_{\max}}$,差异度 $b = \frac{f}{n} = 0$. 如果预测的不确定性较大,则同一度 $a = \frac{s}{n} = 0$,对立度 $c = \frac{p}{n} = 0$,差异度 $b = \frac{f}{n} = \frac{x_{\max}}{x_{\max}} = 1$. 例子中同一度、差异度和对立度的计算结果见表 2.

表 2　各个例中预报因子 x_1 的联系度分析

Table 2　The connection degree statistics of forecasting factor x_1 in every sample

个例序号	同一度 a	差异度 b	对立度 c
1	0.154	0.000	0.846
2	0.060	0.000	0.940
3	0.112	0.000	0.888
4	0.060	0.000	0.940
5	0.133	0.000	0.867
6	0.133	0.000	0.867
7	0.071	1.000	0.929
8	0.042	0.000	0.958
9	0.000	0.000	0.000
10	0.029	0.000	0.971
…	…	…	…
317	0.018	0.000	0.982

在多元回归预报模型的因子集合中,因子之间相互联系、相互制约,有机地组成一个整体. 在预报时如果发现一个或几个因子是模糊因子,说明他们在模型中的重要性已下降,甚至可能干扰模型作出正确预报结论,因而希望让这些因子在这次预报中失去作用,要达到此目的,显然不能简单地剔除这些因子,本节从集对分析的原理出发,用联系度公式导出解决这一问题的方法.

在联系度表达式中差异度 $b = f/n$,它表示在 n 个特征中有 f 个特征表现为既不同一又不对立,即有 f 个特征对预报量的预测持"含糊"态度,该因子与其勉强参与投票,还不如放弃"投票权",而把预报结论的决定权让给其它预报性能清晰的因子. 为此,这里令

$$i = \frac{\sum a}{\sum a + \sum c} + \frac{\sum c}{\sum a + \sum c} j \qquad (3)$$

如果回归因子模型的因子组合中有 m 个因子,则式中 $\sum$ 表示对 $(m-1)$ 个因子求和,其中不包括一个模糊因子. 在上式中求和与求平均等价.

(3) 式的含义是:当某因子的预报性能呈现模糊状态时,它的差异度的 f 个特征按一定比例分配给它的同一度和对立度,这个比值就是其它几个预报性能清晰因子的平均同一度与平均对立度之比. 把(3) 式代入前面(2) 式,可得

$$\mu = a + b\frac{\sum a}{\sum a + \sum c} + \left(c + b\frac{\sum c}{\sum a + \sum c}\right) j \qquad (4)$$

按 SPA 的规定,取 $j = -1$,故有

$$\mu = (a - c) + \frac{b(\sum a - \sum c)}{\sum a + \sum c} \qquad (5)$$

(5) 式是适用于多元回归预报模型的联系度表达式. 应用(5) 式对表 2 的数据进行计算,可得到各因子的联系度值,详见表 3. 用表 3 中各预报因子的值作为新的因子,就可以建立一种新的多元回归模型——基于集对分析的天气预报多元回归模型.

4 效果分析

用原因子 $x_1 \sim x_6$ 建立常规的回归预报模型与基于 SPA 的回归预报模型,进行对比分析. 前例中有 6 个预报因子 x_1、x_2、x_3、x_4、x_5、x_6,通过一系列的处理,可得到它们相应的映射为 μ_1、μ_2、μ_3、μ_4、μ_5、μ_6,映射主要应用公式(5) 以及因子的模糊性量度公式(1) 来进行的. 公式(5) 是针对多元回归模型的特点从集对分析的联系度公式推导而来. 如果注意一下整个推导过程,不难发现对于预报性能清晰的因子,映射只不过是线性变换;而对于预报性能模糊的因子来说,它在模型中的作用已由其它因子取代,映射自然是非线性变换. 映射前后因子与预报量的相关系数见表 4,可见大部分因子的相关系数有了明显提高. 对略有下降的因子 x_6,在映射时不再进行因子的模糊性判断和非线性变换,使它的相关系数维持在原来水平上. 两种预报模型的复相关系数列于表 4 末列,可以看到复相关系数也有了提高. 从表 4 可见,映射后无论是因子独立预报能力还是整个模型的综合预报能力都有了提高. 2002 年梅汛期(5 月 1 日～7 月 10 日) 按常规回归预报模型的准确率为 75.0 %,而基于 SPA 预报模型的准确率为 87.3 %.

表 3 预报因子的联系度 μ

Table 3 The connection degree values μ of all forecasting factors in 317 samples

个例	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
1	-0.692	-0.772	-0.684	-0.750	-0.836	-0.745
2	-0.880	-0.870	-0.791	-0.864	-0.734	-0.866
3	-0.776	-0.674	-0.588	-0.762	-0.792	-0.526
4	-0.880	-0.358	-0.762	-0.898	-0.806	-0.868
5	-0.734	-0.814	-0.814	-0.788	-0.804	-0.928
6	-0.734	-0.499	-0.548	-0.658	-0.844	-0.022
7	-0.585	-0.626	-0.510	-0.854	-0.832	-0.800
8	-0.916	-0.796	-0.644	-0.864	-0.774	-0.816
9	-0.754	-0.820	-0.822	-0.776	-0.794	-0.866
10	-0.942	-0.738	-0.548	-0.792	-0.562	-0.908
…	…	…	…	…	…	…
317	-0.964	-0.863	-0.894	-0.894	-0.980	-0.982

表 4 预报因子映射前后的相关系数和复相关系数的比较

Table 4 The changes of correlation coefficients and complex correlation coefficients before-to-after forecasting factors mapped

建立预报模型方法	x_1	x_2	x_3	x_4	x_5	x_6	复相关系数
常规回归预报模型	0.556	0.331	0.297	0.229	0.293	0.461	0.593
基于 SPA 的预报模型	0.557	0.458	0.366	0.276	0.348	0.437	0.648
效果评定	略提高	提高	提高	提高	提高	略下降	提高

集对分析有两个基本观点[7]:不确定性和确定性可以在同一个系统中进行分析和处理;不确定性与确定性在一定条件下可以相互转换. 两个观点已充分反映在上面的工作中. 下面再从一个具体的预测过程来剖析新模型如何对因子的不确定性进行分析,以及新模型如何实现不确定性与确定性之间的转换,从而改进了预报质量的过程. 通过剖析可以进一步考察这种改进机制的合理性.

表 5　SPA 对因子的订正

Table 5　The collections to forecasting factors by SPA means

		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	模型预测值
原因子 X (映射前)	变异系数 C_v	57.5	49.9	18.0	0.1	3.9	1.8	
	一元回归值	6.05	6.59	22.61	14.26	12.13	15.63	15.2
	因子评定	模糊	清晰	清晰	清晰	清晰	清晰	
新因子 μ (映射后)	因子值 μ	- 0.715	- 0.870	- 0.548	- 0.716	- 0.758	- 0.688	20.9
	一元回归值	14.24	6.59	22.61	14.26	12.13	15.63	

表 5 是以第 41 号个例为例所作的分析. 根据变异系数判断,第 1 因子是模糊因子,x_1 对预报量的估计为 6.05,以后出现的预报量实况为 25.6,预报量估计比实况明显偏小. 由于被识别为模糊因子,用公式(5)作非线性映射,得到新因子 μ_1(μ_1 是 x_2, x_3, x_4, x_5 和 x_6 的函数),用 μ_1 计算得到它对预报量的估计为 14.24,新因子的预报值比原因子预报值有了明显提高. 其它因子由于是清晰因子,它们的一元回归值没有变化. 原预报模型的预报值为 Y = 15.2,基于集对分析的模型的预报值为 20.9,后者比前者减少了误差 22.3 %. 可见通过模糊因子的识别和集对分析等一系列处理,预报误差有了明显减少,同时也可看出本文所给出的因子识别和处理机制是一般预报方法难以实现的.

5　结　语

5.1　天气预报实践证明,天气预报的不确定性客观存在,它取决于大气运动的随机性以及初始场和预报方法的不完全性. 在多元回归分析中我们用随机变量来表示预报量的不确定性,用模糊性来表示预报因子的不确定性. 为了定量地描述因子的模糊性,我们分析了集合预报和近邻估计两种方法的基本思路,从而定义了描述因子模糊性的统计量 —— 变异系数. 有了对模糊性的定义,就可以在多元回归模型中识别出模糊因子和清晰因子.

5.2　SPA 是研究自然界和人类社会中的不确定性,以及在一定条件下不确定性与确定性相互转化的规律的理论[8]. 从 SPA 的基本理论推导了适合于多元回归分析的联系度公式. 借助此公式可以很好地处理模糊因子,让它的作用由清晰因子来代替,使天气预报的不确定性部分地向确定性转化. 几年来绍兴市气象台在降水概率预报、雨量预报和大-暴雨预报方面的大量试验和业务运行表明,这种方法明显地提高了天气预报的准确率.

5.3　多元回归模型在理论和方法上已十分成熟,应用范围十分广泛,它不仅仅用于天气预报方面,也不仅仅用于预测方面. 回归分析对因子的筛选和组合有许多行之有效的技术,但是在选定一组"最佳因子组合"后,如何合理地使用这些因子,因子间又如何灵活地长短互补,至今尚未有更多的讨论. 天气预报实践表明,预报模型的预报失败个例往往由少数模糊因子的失误引起,为此合理地处理因子集合中的模糊因子,对提高天气预报准确率显得十分重要. 本文只是在改进传统的多元回归分析方法,以提高回归模型的预测准确率方面进行的一次初步探索.

参考文献:

[1]　赵克勤. 集对分析及其初步应用[M]. 杭州:浙江科技出版社,2000.9-15.

[2]　王国强. 集对分析 ——一种新的不确定性理论在 MOS 概率天气预报中的应用[J]. 浙江气象科技,1999,20(1):1-6.

[3]　李小泉. 美国国家气象中心中期预报时段内的集合预报[J]. 气象科技,1994,22(2):7-11.

[4]　陈希孺,柴根象. 非参数统计教程[M]. 上海:华东师范大学出版社,1993.272-286.

[5]　王国强,蒋延龙,陈红梅. 近邻估计 ——线性回归预报模型及其台风暴雨预报试验[J]. 气象科技,1999,27(4):25-29.

[6]　赵克勤. 集对分析对不确定性的描述和处理[J]. 信息与控制,1995,24(3):165-168.

[7]　赵克勤. 集对论 ——一种新的不确定性理论方法应用[J]. 系统工程,1996,14(1):18-23.

[8]　张　斌. 不确定性信息处理的集对论思想方法[J]. 模糊系统与数学,2001,15(2):89-93.