最优子集的神经网络预报建模研究

陈　宁　金　龙　袁成松

（江苏省气象科学研究所，南京　210008）

提　要

作者尝试用最优子集方法进行神经网络长期预报模型的建模方法研究。结果表明，在很多情况下，由于最优子集方法比逐步回归方法能选取更好的预报因子，因此所构造的神经网络预报模型具有更好的拟合和预报效果，这为神经网络在长期预报的应用研究提供了新的思路和方法。

关键词：最优子集　逐步回归　神经网络

引　言

在现今的长期天气预报工作中，逐步回归方法是被较为广泛使用的预报方法之一。该方法计算简便快速，但是也存在一定的缺陷。当预报模型不合理或预报因子选取不适当时，预报效果比较差。我们知道长期天气过程变化具有很明显的非线性演变特征，而神经网络方法具有很强的处理非线性问题能力，但是由于其本身不能对预报因子进行筛选，而需要通过其它方法来选择预报因子，作者[1]曾用逐步回归方法来构造神经网络学习矩阵，进行预报建模研究，取得了较好的预报效果。在很多情况下，由于最优子集方法能够比逐步回归方法选取更好的因子，为此，本文尝试用最优子集方法来构造神经网络学习矩阵，进行预报建模研究，以期提高用神经网络方法制作长期预报的准确性。

1　最优子集方法和神经网络方法

我们知道，逐步回归方法在选入或剔除预报因子时，都是基于统计检验（F 检验），所以从理论上并不能以任何概率保证所挑选的自变量的“显著性”[2]。这样，挑选出的预报

因子集就有可能只是一个局部最优子集，而不是全局的最优。因而就不能充分体现模型特征，会对预报工作的准确性产生一定的影响。而最优子集方法正是针对上述问题提出的。该方法是对模型的所有子集进行筛选，即对所有因子进行各种组合，从中选出最优的方程。这种方法计算的工作量相当大，当有 P 个可供挑选的预报变量时，所有可能的回归模型就有 2^P-1 个，这样当 P 相当大时需检验的样本组合数就会大得惊人，这显然是要付出运算代价的。随着现在计算机的高速发展，以样本容量为14，筛选7个因子个数下的最优子集为例，需要进行 $C_{14}^{7}=3432$ 次相应的回归计算，利用现在较为普遍使用的pentium166来计算，所需时间不会超过半分钟，以这样的运算时间来解决一般长期预报的运算量是足够的。

在一般情况下，用回归方法建立预报方程时，入选的因子个数越多，方程的复相关系数会越高，方程的拟合效果也会越好，但是仅凭这一点是无法作为选择方程的准则。因此，要在众多的因子组合中真正筛选出最优子集，就需要采用统一标准的准则来衡量各个子集的优劣程度，常用的准则有修正复相关系数 $\overline{R}$、平均剩余平方和 $\hat{\sigma}^2$、平均预报均方差 S_P 等[3]，其中修正相关系数着眼于建模，后两个准则着眼于预测，它们对自变量个数的增加进行了较严厉的惩罚，公式如下：

$$\overline{R}^2=1-(1-R^2)\frac{n-1}{n-l-1} \tag{1}$$

$$\hat{\sigma}^2=\frac{S_{剩}}{n-l-1} \tag{2}$$

$$S_P=\frac{S_{剩}}{(n-l-1)(n-l-2)} \tag{3}$$

其中，$S_{剩}$ 为剩余平方和，n 为样本长度，l 为选取因子的个数，这样，$\hat{\sigma}^2$、S_P 最小，而 $\overline{R}^2$ 最大的回归模型即为最优子集预报模型。

一些学者研究发现，神经网络方法对具有较强非线性变化特点的平均降水预报问题比传统的逐步回归方法具有更好的预测能力[1]。同时，由于神经网络方法具有全息联想能力及很强的容错能力[4]，这对于实际的平均降水量预报是十分重要的。神经网络预测模型的大量参数是网络对输入的原始数据进行不断的学习训练来获得的。其学习训练过程，主要是通过调整网络模型输入层与隐含层及隐含层与输出层之间的各连接权系数及阈值，计算实际输出与期望输出的误差。当全部样本的输出误差小于设定的收敛误差时训练结束。根据这些确定的连接权系数和阈值，就可以得到神经网络的预报模型。其具体算法有很多，本文采用的是误差反传算法，该方法的数学原理、推导及学习算法详见文献[5]。

2　用最优子集方法构造神经网络学习矩阵

为了分析用最优子集方法构造神经网络学习矩阵建立预报模型的可能性和优越性，本文采用1952～1994年江淮（南京、南通、苏州、淮阴、盐城5站）6～8月汛期平均降水量作为预报量，首先通过计算普查预报量与前期海温场各网格点相关系数，取成片的大于4个格点以上的相关区作为一个预报因子，共获得14个预报因子。然后我们对这14个因子进行全部可能回归（2^{14}—1次），并根据式(1)、(2)和(3)分别计算出它们所对应的 $\overline{R}^2$、$\hat{\sigma}^2$ 和 S_P，挑选出对应不同因子个数下的最优子集；另外，为了便于比较，我们又采用逐步回归方法通过调整 F 值，也挑选出对应不同因子个数下的逐步回归预报模型。此时，样本长度取43（即用1952～1994年的资料建模，而将1995～1997年3年资料作为独立样本的预报检验）。两种方法取出的因子组合及有关参数见表1、2。通过对表1、2的比较分析我们看到：取1个因子时，两种方法同样选取 x_4 为最优；但是取2个因子时出现了不同的结果，最优子集方法选取 x_1、x_6，而逐步回归方法选取了 x_1、x_4，比较一下这两组

子集的复相关系数，$R_{16}=0.624$，而 $R_{14}=0.615$，很明显 $R_{14}>R_{16}$，即1、6号因子组合为最优，而1、4号因子组成的子集为局部最优；取3个因子时，出现了有意思的结果，最优子集应是 x_1、x_6、x_7 因子组合，而用逐步回归方法调整 F 值来选取3个因子时，F 取值不同，竟选取了两种不同的因子组合，当 $F=3$ 时，选取 x_1、x_4、x_{10} 3个因子，复相关系数 R 为0.660，$F=2$ 时，选取 x_1、x_6、x_7 这3个因子组合，此时复相关系数 R 为0.678；取4个因子时以 x_1、x_6、x_7、x_{12} 为最优，然而，不论怎么调整 F 取值，逐步回归方法始终取不到4个因子的组合；同样，取5个、6个因子时，虽然逐步回归方法选取出了相应的因子组合，但都不是最优的；在取8个因子以后，两种方法都选取出同样的因子组合，即逐步回归方法取到了相应因子个数的最优子集。通过以上分析，进一步说明了逐步回归方法在建立预报模型时，所建立的模型不一定是全局最优[6]，而可能是局部最优。另外，还需说明的是，在实际应用的逐步回归方法中，F 取值具有很大的偶然性，除非正好给出合适的 F 临界值，否则逐步回归方法难以得到最优模型，就会象表2中取3个因子组合时那样，把局部最优的子集当成全局最优的子集，影响模型最终的预报效果。文献[3]中曾采用著名的Hald水泥资料也得到过类似的计算效果。通过以上分析，不难发现，采用一般的逐步回归方程选择预报因子，不仅有可能会选择局部最优的因子子集，并且存在选择预报因子的不确定性；而采用最优子集方法，则可以很好地避免这些缺陷。这为我们试图采用最优子集方程构造更好的神经网络预报模型提供了依据。

表1　全部可能回归的最优子集（$m=14$）及相应的 $\overline{R}$、S_P 和 $\hat{\sigma}^2$

l	最优子集	R	$\overline{R}$	S_P	$\hat{\sigma}^2$
1	x_4	0.534	0.518	34732.6	1424036.0
2	x_1x_6	0.624	0.601	31087.0	1243480.0
3	$x_1x_6x_7$	0.678	0.648	28951.9*	1129126.0
4	$x_1x_6x_7x_{12}$	0.697	0.658	29001.8	1102068.0
5	$x_1x_6x_7x_{10}x_{12}$	0.707	0.659*	29731.2	1100053.0*
6	$x_1x_6x_7x_{10}x_{12}x_{14}$	0.713	0.655	30841.9	1110309.0
7	$x_1x_5x_6x_7x_{12}x_{13}x_{14}$	0.718	0.649	32181.5	1126353.0
8	$x_1x_5x_6x_7x_{10}x_{12}x_{13}x_{14}$	0.721	0.641	33694.4	1145609.0
9	$x_1x_5x_6x_7x_8x_{10}x_{12}x_{13}x_{14}$	0.723	0.630	35554.3	1173293.0
10	$x_1x_5x_6x_7x_8x_9x_{10}x_{12}x_{13}x_{14}$	0.726	0.620	37428.7	1197719.0
11	$x_1x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.727	0.606	39732.9	1231720.0
12	$x_1x_3x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.728	0.589	42307.0	1269210.0
13	$x_1x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.728	0.571	45165.3	1309793.0
14	$x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.728	0.551	48366.8	1354270.0

然而，为了进一步选择确定适宜的最优子集预报模型，我们对表1给出的不同因子个数上最优模型的修正复相关系数 $\overline{R}$、平均剩余平方和 $\hat{\sigma}^2$ 和平均预报均方差 S_P 的计算结果进行了分析。由表1可以看到，在这3个准则的数据栏中数据带有“*”的表示在该准则下此数据所对应的因子组合在各种不同因子个数的所有组合中是最优的。其中对于修正复相关系数 $\overline{R}$ 取5个因子时为最优，对于平均剩余平方和 $\hat{\sigma}^2$ 同样是取5个因子为最优，而对于平均预报均方差 S_P 则是取3个因子为最优。取5个因子，即取 x_1、x_6、x_7、x_{10}、

x_{12}时，所对应的S_P为29731.2，排在第3位；取3个因子，即取x_1、x_6、x_7时，所对应的$\overline{R}$、$\hat{\sigma}^2$分别排在对应准则的第5位和第4位；取4个因子，所对应的$\overline{R}$、S_P和$\hat{\sigma}^2$都分别排在对应准则的第2位；取6个因子时，所对应的$\overline{R}$、S_P、$\hat{\sigma}^2$都分别排在对应准则的第3位、第4位和第3位；综合考虑这3个准则，取5个因子，即取x_1、x_6、x_7、x_{10}、x_{12}时为全局最优子集，而逐步回归方法引入5个因子的顺序依次是x_4、x_1、x_{10}、x_7、x_{14}。为了更好地比较最优子集方法和逐步回归方法选取因子的优劣，我们把这两种方法取得的3、5、6个因子组合分别从样本中取出进行全回归逐步预报（逐步回归方法选取不到4个因子的组合，因此不好对4个因子的组合进行比较）。以1995～1997年3年样本作为独立样本的预报检验，结果见表3。从表3中可以明显看出，最优子集方法不论是拟合效果还是预报效果都比逐步回归方法好。综合以上的计算分析，不难看出，采用最优子集方程可以为我们提供更好的预报子集，而人工神经网络方法本身并不能提供选择预报因子的方法，因此，采用最优子集方法构造神经网络预报模型的学习矩阵，为提高预报模型的预报能力提供了可能。

表2　逐步回归方法选取的子集（$m=14$）及对应的复相关系数

l	逐　步　回　归	R
1	x_4	0.534
2	x_1x_4	0.615
3	$x_1x_6x_7$（$F=2$）	0.678
	$x_1x_4x_{10}$（$F=3$）	0.660
4	无法选取	—
5	$x_1x_4x_7x_{10}x_{14}$	0.701
6	$x_1x_6x_7x_8x_{10}x_{14}$	0.712
7	$x_1x_6x_7x_8x_{10}x_{11}x_{14}$	0.716
8	$x_1x_5x_6x_7x_{10}x_{12}x_{13}x_{14}$	0.721
9	$x_1x_5x_6x_7x_8x_{10}x_{12}x_{13}x_{14}$	0.723
10	$x_1x_5x_6x_7x_8x_9x_{10}x_{12}x_{13}x_{14}$	0.726
11	$x_1x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.727
12	$x_1x_3x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.728
13	$x_1x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.728
14	$x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}$	0.728

表3　最优子集方法和逐步回归方法选取不同因子组合的拟合效果和预报效果比较

	样本数	拟合效果/%		预报效果/%	
		最优子集	逐步回归	最优子集	逐步回归
3个因子	43	17.91	18.56	4.9	14.0
	44	17.56	18.33	−21.4	−24.2
	45	17.93	18.67	−12.0	−4.0
	绝对值平均	17.80	18.52	12.7	14.1
5个因子	43	16.47	17.06	5.8	11.4
	44	16.15	16.83	−25.1	−27.8
	45	16.66	17.32	−3.4	−8.9
	绝对值平均	16.43	17.07	11.4	16.0
6个因子	43	16.53	16.93	5.7	6.7
	44	16.21	16.64	−27.2	−27.0
	45	16.59	16.93	−5.0	−7.4
	绝对值平均	16.44	16.83	12.6	13.7

3　汛期降水预报模型的检验分析

为了对预报模型进行检验分析，将最优子集方法取出的 x_1、x_6、x_7、x_{10}、x_{12} 这5个因子作为网络学习矩阵的输入，而相应的实测平均降水量为期望输出，以此建立神经网络预报模型的学习矩阵。然后，把学习矩阵加载到网络的输入端，进行网络学习训练。其中网络的各项参数为：学习因子取0.9，动量因子取0.7，隐节点数为5，收敛误差取0.001。对网络进行5000次学习训练，当收敛误差趋于平稳时，训练结束。由确定的网络各项参数可以得到对历史样本的拟合数据以及相应的预报值。为了便于比较，我们把用逐步回归方法选取的5个因子，即 x_1、x_4、x_7、x_{10}、x_{14} 同样构造成神经网络学习矩阵，并进行相应的学习训练，得到相应的拟合数据和预报值。根据上述方法，对由最优子集方法和逐步回归方法这两种不同的方法分别构造的神经网络预报模型取不同样本长度的数据拟合进行比较（表略），可以看到，用最优子集方法构造神经网络学习矩阵得到的神经网络预报模型的平均拟合效果均为10.86%，优于用逐步回归方法构造学习矩阵得到的预报模型的平均拟合效果12.59%。为了进一步考察用最优子集构造神经网络学习矩阵得到的预报模型的实际预报能力，我们再对1995～1997年进行独立样本的预报检验。为了客观地对比分析，两种方法构造的神经网络模型所采用的网络结构、参数在3年的独立样本试验预报中与前面根据43个样本（1952～1994年）建立的预报模型一致，这样就使得独立样本的预报试验与实际预报是类似的。表4同时给出了1995～1997年，由两种方法分别建立神经网络预报模型的独立样本预报结果比较，由

表4　两种神经网络预报模型的独立样本预报检验

年份	实测	最优子集的ANN		逐步回归的ANN	
	/mm	预报/mm	相对误差%	预报/mm	相对误差%
1995	444.3	441.6	−0.6	332.7	−25.1
1996	607.5	487.0	−19.8	434.5	−28.5
1997	462.1	398.9	−13.7	443.2	−4.1
绝对值平均			11.4		19.2

表4可以看出，由最优子集方法构造神经网络学习矩阵得到的预报模型对这3年的独立样本预报精度令人满意。在3年的预报试验中，除1997年预报误差高于由逐步回归方法所建立的神经网络预报模型以外，1995、1996年两年预报误差均明显偏低，该方法3年预报结果的相对误差绝对值平均为11.4%，也低于由逐步回归方法建立的神经网络预报模型的预报相对误差的绝对值平均19.2%。以上的分析结果表明，由最优子集方法构造的神经网络学习矩阵所得到的神经网络预报模型无论是对历史样本的拟合效果还是独立样本的试验预报结果，均明显优于由逐步回归方法构造神经网络学习矩阵所得到的预报模型。此外，作者根据确定的神经网络预报模型对1998年汛期（6～8月）降水量趋势进行了预报，预报结果为452.1mm，而实况平均降水量586.0mm，预报值与实测值的相对误差为22.8%，这样的实际预报结果①，比逐步回

① 预报是实况出现前做出的，而检验是以后根据审稿修改意见要求进行的实际预报检验。

归神经网络预报模型对1995、1996两年的独立样本预报精度高，接近3年独立样本的预报平均相对误差，显示了一定的预报能力。但是，与最优子集的神经网络独立样本预报效果相比，仍表现出统计预报的通病，即拟合和独立样本预报效果一般均好于实际预报。当然，这只是一年的预报检验，有必要对预报模型不断地进行实际预报应用和改进提高。

4 小结

由于目前业务长期天气预报在很大程度上还依赖于统计预报方法，因此，将各种不同的统计预报方法相结合以探索新的长期预报方法和思路是十分有意义的。本文利用神经网络优良的自学习以及较强的的非线性映射能力，通过由最优子集方法来构造神经网络学习矩阵，建立预报模型，进行了有益的尝试。从试验结果来看，结果是令人满意的，这为神经网络预报方法在长期天气预报的应用研究提供了新的思路。然而，本文采用最优子集方法构造神经网络学习矩阵进行预报建模，其本质上仍然是从预报量与预报因子的线性关系去考察、选择预报因子，而如何从非线性角度来选择更好的预报因子建立神经网络的降水长期预报模型值得进一步深入研究。

参考文献

1 Jin Long(金龙) et al. Comparison of Long-term forecasting of June-August rainfall over Changjiang-Huaihe Valley. 大气科学进展(英文版). 1997,14(1):81～92.

2 俞善贤，汪铎. 试用最优子集与岭迹分析相结合的方法确定回归方程. 大气科学，1988,12(4):382～388.

3 周纪芗. 实用回归分析方法. 上海:上海科技出版社，1990.

4 靳蕃，范俊波，谭永东. 神经网络与神经计算机原理应用. 成都:西南交通大学出版社，1991.

5 Jin Long(金龙) et al. Study on mixed model of neural network for farmland flood/drought prediction. 气象学报(英文版)，1997,11(3):364～373.

6 施能，曹鸿兴. 基于所有可能回归的最优气候预测模型. 南京气象学院学报，1992,15(4):459～466.

Study of the Prediction Modeling of Neural Network Using the Optimal Subset

Chen Ning Jin Long Yuan Chengsong

(Jiangsu Institute of Meteorology, Nanjing 210008)

Abstract

A method about the neural network long-tem forecast is studied by using the optimal subset. The results show that the neural network model established by the optimal subset is found of better fitting and prediction accuracy, because the method using the optimal subset can select out better factors than stepwise regression. Thus it provides a new-line for the research of neural network prediction on long-term forecast.

Key Words: optimal subset stepwise regression neural network