DOI:10.16177/j.cnki.gxmzzk.2012.04.024

广西民族大学学报(自然科学版)
第18卷第4期 JOURNAL OF GUANGXI UNIVERSITY FOR NATIONALITIES Vol. 18 No. 4
2012年12月 (Natural Science Edition) Dec. 2012

主成分分析在线性模型与非线性模型的应用研究*

农吉夫
(广西民族大学 理学院,广西 南宁 530006)

数学

摘 要: 为了解主成分分析在线性模型与非线性模型预报中的应用效果,在2001—2011年热带气旋历史观测资料基础上,采用主成分分析方法,结合线性回归模型和神经网络模型,开展西北太平洋热带气旋的强度预报技术研究试验.根据提取的主要影响因子构造线性回归模型与BP神经网络的输入样本进行不同样本的台风强度预测.计算结果表明,主成分分析通过降低线性回归模型和BP神经网络模型的维数,减少自变量之间的复共线性,减小模型的预报平均绝对误差.

关键词: 主成分分析;线性模型;非线性模型;台风强度
中图分类号: O213 **文献标识码:** A **文章编号:** 1673—8462(2012)04—0030—05

0 引言

随着台风强度预报技术[1—3]的不断改进,可以考虑的影响台风强度变化的预报因子个数在不断增加,这些因子共同影响着台风强度变化,它们之间必然存在着复共线性关系.在统计建模过程中,很多学者采用最小二乘回归框架下的线性回归技术,在线性回归模型中,逐步回归方程一般入选因子的原则是根据F值自动从众多因子中剔除并入选构建预报模型的因子,但这些入选因子之间不可避免地也存在复共线性关系,而线性回归在所选因子之间相关性较显著时无法得到准确的结果,必然影响预报模型的预报能力[4].另外,在建模过程中,样本不同,所选的预报因子也会不同,这说明存在所选因子是否稳定的问题,也会影响回归方程的稳定性[5].

主成分分析法[6—8]作为系统降维和特征提取的基本方法,可将多个因子所组成的因子场,用数目较少的主成分代表原因子场的主要内容,且相互间是不相关的.另外,在对台风强度进行客观预报时,一般采用的是线性回归模型,使用非线性模型的较少,由于人工神经网络具有非线性、良好的容错性、自适应性以及其大规模并行处理的特征,现已被广泛应用到气象领域[9—10],笔者采用人工神经网络中的BP神经网络,建立非线性的台风强度预报模型.在建立BP神经网络时,如果输入因子过多或者因子间存在复共线性关系,也会影响预报模型的预报能力,故考虑在建立BP神经网络时同样对预报因子进行主成分分析,在系统降维的同时也消除因子之间的复共线性.

1 主成分分析与人工神经网络

1.1 主成分分析

主成分分析的基本思想是通过降维过程,将多个相互关联的数值指标转化为少数几个互不相关的综合指标的统计方法,即用较少的指标来代替和综合反映原来较多的信息,这些综合后的指标就是原来多指

* 收稿日期:2012-09-08.
基金项目:国家自然科学基金(11061005);广西教育厅科研项目(201204LX083).
作者简介:农吉夫(1975-),男(壮族),广西东兰人,广西民族大学理学院副教授,从事概率统计及气象预报建模研究.

标的主要成分. 它的处理方法是先利用主成分分析对多变量参数矩阵进行处理，由于主成分分析的实质是空间的坐标旋转，并不改变样本数据结构，得到的主成分是原变量的线性组合，而且两两不相关，能够最大限度地反映原变量所包含的信息，以一定标准选取前几个较重要的主成分之后，原来的多维问题大大简化. 具体的计算步骤如下：

1）对原始数据矩阵 $X_{(n\times p)}$ 标准化处理，得到新的数据矩阵 $Y_{(n\times p)}$；

2）建立标准化后的 p 个指标的相关系数矩阵 R；

3）计算相关矩阵 R 的特征值及相应的特征向量 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$，并使其从大到小排列，同时求得对应的特征向量 $u_1, u_2, \cdots, u_p$；

4）计算贡献率 e_m 和累计贡献率 E_m；

$$e_m = \frac{\lambda_m}{\sum_{i=1}^{p}\lambda_i}, E_m = \frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$$

5）确定主成分个数.

1.2　人工神经网络

人工神经网络[11-12]是通过数学方法对人脑若干基本特性进行模拟，是一种模仿人脑结构及其功能的非线性信息处理系统. 经过半个多世纪的发展，由于各种网络结构和算法系统的产生，已逐渐发展成较为完善的人工神经网络理论体系. BP 神经网络是该技术中应用最为广泛的一种，其特点有自适应、自组织、自学习的能力、非局域性和非凸性的突出优点，正是这些特点使得该方法已经解决了许多实际问题，其生命力也恰恰在于广泛的实用价值.

BP 神经网络通常由输入层、隐含层和输出层组成，其信息处理过程是由向前传播与向后学习两部分组成，网络学习规则是误差从输出层到输入层向后传播并修正样本的过程，学习目标是使网络的实际输出逼近目标样本，BP 神经网络的结构如图 1 所示.

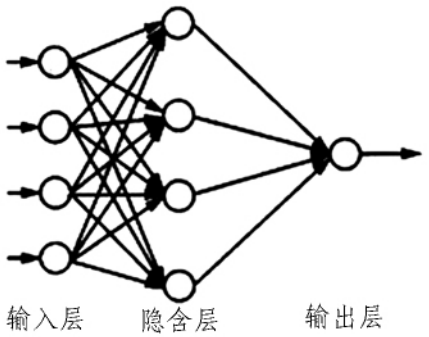

图 1　BP 神经网络结构

Fig. 1 BP neural network structure

2　台风资料

2.1　预报对象

笔者在进行台风强度预报建模试验时，以中国气象局编辑出版的 2001－2011 年共 11a 台风强度资料为基础，选取处于西北太平洋海域(180 E° 以西，0 ～ 45 N°)5－6 月份具有 48 h 以上生命史的台风个例作为预报研究对象，规定台风从进入该海域范围的第一点开始，每隔 6 h 作为一个统计样本. 在预报建模中，建立月各预报时效(12 h、24 h、36 h、48 h、60 h、72 h)的样本集，并选取每个样本集最后 30 个台风样本作为独立样本进行预报建模试验研究.

2.2　选择预报因子

利用气候持续法原理[13]初选得到包括起报时刻经度、纬度、台风中心最低气压、地面附近最大风速，起报时刻与前 12/24 小时的经(纬) 度差、风速差，过去 24 小时变压，前 24 小时地面附近最大风速变化，间隔 6/12 小时的前 12/24 小时经(纬) 向加速度，前 12/24 小时经(纬) 向移速、经纬向合成移速，前 12 小时至前 24 小时经(纬) 向移速、经纬向合成移速，前 6/12/24/36 小时所在经(纬) 度、气旋中心最低气压、地面附近最大风速等共 62 个可能对台风强度变化有影响的气候持续因子. 由于初选预报因子较多，如果将这些预报因子全部用于预报建模，势必会造成网络结构庞大，容易出现过拟合现象.

2.3　通过主成分分析选择预报因子

由于影响台风强度的因素较多，在建立预报模型时必须考虑多个预报因子对预报对象的影响，从统计预报的角度分析，众多预报因子往往存在复共线性关系，因子各自所带的噪声信息叠加，会造成信息重复、噪声增加，从而影响到预报模型的预报能力. 笔者主要利用主成分分析方法抑制多因子变量之间的严重复共线性关系，即要求输入的矩阵因子数量较少但预报信息却比较丰富，以提高多预报因子的预报能力.

把上述通过检验的待选因子的数据进行标准化，然后进行主成分分析，结果表明，待选因子主成分分析后，每个样本集前 10 ～ 13 个主成分累积方差贡献率均达 90% 以上，主要信息集中在这几个特征值较大的主分量上，且与台风强度相关性较好，以特征值相对较大且与台风强度相关性较好作为取舍因子标准，选择前 10 ～ 13 个主分量作为预报模型因子，各预报时效的主成分与累积贡献率如表 1 所示.

表 1　各预报时效利用主成分分析提取的主成分

Table 1 Principal component analysis of different forecast period

	12 h		24 h		36 h		48 h		60 h		72 h	
主成分	特征值	累积贡献率	特征值	累积贡献率	特征值	累积贡献率	特征值	累积贡献率	特征值	累积贡献率	特征值	累积贡献率
1	24.29	0.39	24.19	0.39	21.7	0.35	17.98	0.29	15.54	0.25	14.07	0.23
2	9.48	0.54	9.57	0.54	9.12	0.50	8.53	0.43	8.87	0.39	10.79	0.40
3	5.17	0.63	5.65	0.64	6.54	0.60	7.47	0.55	7.70	0.52	7.23	0.52
4	4.66	0.70	5.44	0.72	3.88	0.67	3.99	0.61	4.65	0.59	4.68	0.59
5	3.00	0.75	2.88	0.77	2.86	0.71	3.32	0.67	3.38	0.65	3.59	0.65
6	2.79	0.80	2.36	0.81	2.53	0.75	2.88	0.71	3.22	0.70	3.30	0.70
7	2.37	0.83	1.93	0.84	2.39	0.79	2.59	0.75	2.85	0.75	3.16	0.76
8	1.67	0.86	1.57	0.86	2.05	0.82	2.56	0.80	2.53	0.79	2.66	0.80
9	1.48	0.89	1.50	0.89	1.70	0.85	2.05	0.83	2.31	0.82	2.01	0.83
10	1.29	0.91	1.15	0.91	1.59	0.88	1.68	0.86	1.68	0.85	1.62	0.86
11					1.36	0.89	1.45	0.88	1.51	0.87	1.44	0.88
12					1.01	0.91	1.23	0.89	1.38	0.89	1.35	0.90
13							1.04	0.92	1.08	0.91		

数学

3　台风强度预报模型

3.1　主成分线性回归模型

把通过主成分选择预报因子的代入到线性回归模型中，得到 5 月各预报时效(12 h、24 h、36 h、48 h、60 h、72 h) 的主成分线性回归预报方程为

$$y_{12} = 28.27 - 0.478z_1 + 1.623z_2 + 0.422z_3 + 1.379z_4 + 1.79z_5 - 0.273394z_6 + 1.562z_7 + 0.147z_8 + 2.975z_9 + 0.481z_{10}$$

$$y_{24} = 29.21 + 0.265z_1 + 1.196z_2 - 1.236z_3 + 1.262z_4 - 0.809z_5 + 1.67z_6 + 2.029z_7 + 1.358z_8 + 2.12z_9 + 0.676z_{10}$$

$$y_{36} = 29.986 + 0.49z_1 - 0.475z_2 - 0.443z_3 + 0.399z_4 - 1.561z_5 + 0.889z_6 - 0.145z_7 + 0.577z_8 + 1.788z_9 + 2.117z_{10} + 1.661z_{11} + 0.896z_{12}$$

$$y_{48} = 30.686 + 0.057z_1 + 0.777z_2 + 0.757z_3 - 0.248z_4 + 0.592z_5 + 1.243z_6 + 0.253z_7 + 0.081z_8 + 0.988z_9 + 0.698z_{10} + 2.327z_{11} + 1.099z_{12} + 1.149z_{13}$$

$$y_{60} = 30.941 - 0.076z_1 + 1.314z_2 - 0.219z_3 - 0.684z_4 - 0.283z_5 - 1.001z_6 + 0.316z_7 + 0.477z_8 + 0.145z_9 + 0.485z_{10} + 2.178z_{11} - 1.091z_{12} + 1.351z_{13}$$

$$y_{72} = 30.932 + 0.699z_1 - 0.214z_2 + 0.108z_3 - 0.531z_4 - 0.301z_5 - 0.188z_6 + 0.287z_7 + 0.058z_8 - 0.117z_9 - 1.731z_{10} + 1.548z_{11} + 1.425z_{12}$$

其中以上各时效的预报方程中，$y_{12}, y_{24}, \cdots, y_{72}$ 分别是预报时效为 12 h，24 h，…，72 h 的预报量，$z_i (i = 1,2,\cdots,13)$ 是各预报方程通过主成分分析提取的主成分.

3.2　逐步回归模型

初选因子不进行主成分分析，对 5 月时效为 12 h 的台风强度建立逐步回归预报方程，取 $F = 2.0$ 筛选出统计检验显著的 11 个预报因子代入到逐步回归模型中，得到的预报方程为

$$y_{12} = -138.156 - 0.154x_1 + 1.277x_4 - 0.239x_8 - 0.107x_{15} + 0.0012x_{19} + 0.926x_{26} + 0.168x_{38} + 0.581x_{40} - 0.533x_{56} - 0.148x_{59} + 0.135x_{62}$$

其中 y_{12} 为 5 月时效为 12 h 的台风强度预报量，x_i 为经过逐步回归筛选得到的自变量.

同样的，对 5 月时效分别为 24 h、36 h、48 h、60 h、72 h 的台风强度建立逐步回归预报方程.

3.3　主成分神经网络模型

在建立 BP 神经网络预报模型时，采用一个三层的前馈网络，该神经网络模型的计算输出为

$$y = f\left(\sum_{i=1}^{p} b_i w_{ij} - \gamma_j\right)$$

上式中，b_i 是输入层到隐含层的激活值，w_{ij} 是连接权系数，初始时刻为一组给定的随机数，γ_j 是输出层单元阈值，$f(x)$ 取 Sigmoid 函数，其表达式为

$$f(x) = \frac{1}{1+e^{-x}}$$

隐层节点数的确定是神经网络设计中非常重要的一个环节，这一问题的复杂性，使得至今为止尚未找到一个很好的解析式，隐层节点数往往根据前人设计所得的经验和自己进行试验来确定. 一般认为，隐层节点数与求解问题的要求、输入输出单元数多少都有直接的关系. 而且，隐层节点数过少，则无法产生足够的连接权组合数来满足若干样本的学习；隐层节点数过多，则学习以后网络的泛化能力变差.

在大量实验结果的基础上，笔者发现隐层节点数

与输入和输出的信息量有关. 随着输入和输出的信息量的增加，隐层节点数也往往有所增加. 而在输入训练样本数不是太庞大的情况下，这个信息量则直接与输入网络的元素数和目标类别数有关. 因此，针对隐层结构的这样一些特点，这里给出一个经验公式作为参考：

$$S = 2M + N$$

其中 S 为隐层节点数，M 为输入矢量维数，N 为输出节点数.

4　预报试验结果对比分析

4.1　主成分分析在线性模型中的应用

为了解主成分分析应用到线性回归模型在台风强度预报中的效果，试验根据 3.1 节建立的台风强度主成分线性回归预报模型和 3.2 节建立的逐步回归模型，分别对 5 月各预报时效(12 h、24 h、36 h、48 h、60 h、72 h)的 30 个台风强度独立样本进行预报，并计算 30 个独立样本的预报平均绝对误差值(MAE)，并比较二者的预报结果，计算结果如表 2 所示，对比图如图 2 所示.

从表 2 可以看出，在使用独立样本的检验预报中，主成分线性回归模型的预报效果要好于逐步回归模型，每一预报时效误差均有一定减小.

表 2　四种预报模型 30 个台风强度独立样本预报效果对比

Table 2 The mean absolute error (MAE) of four kinds of forecast model

预报时效	12 h	24 h	36 h	48 h	60 h	72 h
样本总数	461	419	379	339	302	267
建模样本	431	389	349	309	272	237
MAE(逐步回归)	3.92	6.52	9.62	10.49	11.50	11.57
MAE(主成分回归)	3.27	5.82	8.47	10.25	11.31	11.53
MAE(神经网络)	3.26	6.34	9.56	9.80	11.30	12.36
MAE(主成分神经网络)	3.17	5.74	8.50	9.83	10.79	11.08

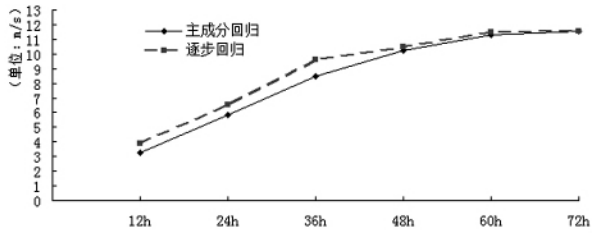

图 2　主成分回归与逐步回归对独立样本预报的平均绝对误差对比

Fig. 2 Mean absolute error of principal component regression and stepwise regression to prediction of independent sample

4.2　主成分分析在非线性模型中的应用

为了解非线性模型在台风强度预报中的预报效果，同时也了解主成分分析是否也可用于非线性模型，试验根据 3.3 节内容建立的台风强度非线性预报模型，即 PCA－BP－ANN 和 BP－ANN 两个预报模型，比较两者的预报结果. 在台风强度预报建模时，把通过主成分分析选择的预报因子作为输入节点(见表 1)，BP 神经网络模型的学习因子和动量因子分别统一取 0.9 和 0.7，收敛误差取 0.00001，训练次数取 5000 次，计算 30 个独立样本的预报平均绝对误差值(MAE)，并比较二者的预报结果，计算结果如表 2 所示，预报平均绝对误差如图 4 所示. 从图 4 可以看出，非线性的 BP 神经网络模型可以应用到台风强度的预报，并且利用主成分分析，可以在不减少因子信息量的情况下，解决 BP 神经网络结构过大、因子间复相关的问题，从而提高非线性预报模型的预报能力.

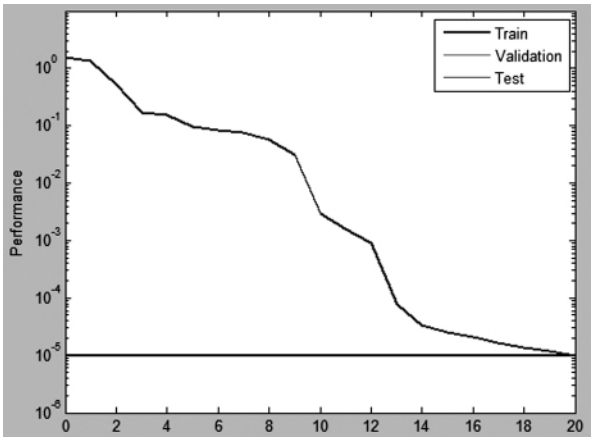

图 3　主成分神经网络训练误差变化图

Fig. 3 Training error change of principal component neural network

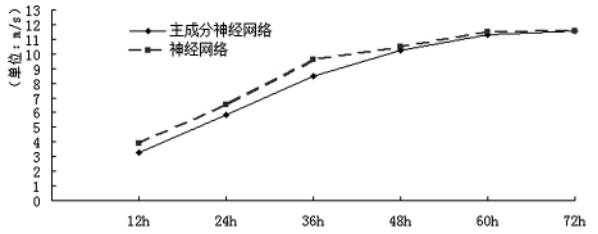

图 4　主成分神经网络与神经网络对独立样本预报的平均绝对误差对比

Fig. 4 Mean absolute error of principal component neural network and BP neural network to prediction of independent sample

5　结语

由于神经网络模型具有局部逼近的特征和较强的非线性映射能力，因此它能够较好地模拟具有较强非线性变化特点的台风强度预报问题. 基于主成分分析的 BP 神经网络简化了网络输入样本，消除了网络

输入之间的相关性,同时消除了输入因子的相关性并简化网络结构,大大提高网络的学习速率,降低了网络的输入层数,改善了程序执行效率,从整体上提高了网络的性能,在气象灾害预测研究应用取得了良好的效果.

[参　考　文　献]

[1]钮学新,藤卫平. 热带气旋强度变化预报的一种动力一统计预报方法[J]. 科学通报,1986,13,1003－1006.

[2]端义宏,余晖,伍荣生. 热带气旋强度变化研究发展[J]. 气象学报,2005,63(5),636－645.

[3]黄小燕,金龙,姚才,等. 夏季南海台风移动路径的一种客观预报方法[J]. 南京气象学院学报,2008,31(2),287－292.

[4]黄嘉佑. 气象统计分析与预报方法[M]. 北京:气象出版社,2004,121－140.

[5]陈佩燕,端义宏,余晖. 红外云顶亮温在西北太平洋热带气旋强度预报中的应用[J]. 气象学报,2006,64(4),474－484.

[6]黄嘉佑. 气象统计分析与预报方法[M]. 北京:气象出版社,2004.

[7]凌和良,桂发亮,楼明珠. BP神经网络算法在需水预测与评价中的应用[J]. 数学的实践与认识,2007,37(22),42－47.

[8]施能. 气象科研与预报中的多元分析方法[M]. 北京:气象出版社,2005.

[9]陈桦,程云艳. BP神经网络算法的改进及在MATLAB中的实现[J]. 陕西科技大学学报,2004,(4),45－47.

[10]农吉夫. 主成分分析与支持向量机的区域降水预测应用研究[J]. 广西民族大学学报:自然科学版,2009,15(2),89－93.

[11]董长虹. Matlab神经网络与应用[M]. 北京:国防工业出版社,2007.

[12]黎艳萍,王汝凉,郭炜伟,等. 具滞后的离散广义双线性系统的稳定控制分析[J]. 广西民族大学学报:自然科学版,2008,14(4):36－40.

[13]孔宁谦,陈润珍. 用统计动力方法作盛夏南海中北部热带气旋强度预报[J]. 广西气象,2006,27(1):4－5.

[责任编辑　苏　琴]
[责任校对　方丽菁]

数

学

Linear Model and Nonlinear Model based on Principal Components Analysis and Its Application

NONG Ji-fu

(*College of Science, Guangxi University for Nationalities, Nanning* 530006, *China*)

Abstract: In order to evaluate the potential forecast efficiency of principal components analysis (PCA) in linear model and nonlinear model, based on 2001－2011 historical tropical cyclone observation data, the PCA efficiency are investigated through multiple linear regression model and neural network model focusing on the northwestern Pacific Ocean tropical cyclone intensity prediction technology. According to these main factors, the input samples of linear regress model and BP neutral network are definite, and the models could be trained to predict tropical cyclone intensity. Result shows that PCA reduces the models dimension of linear regression and BP neural network, and weaken multi-collinearity among the independent variables, and the method based on PCA lessens average absolute error (MAE) of tropical cyclone intensity.

Key Words: principal components analysis; linear model; nonlinear model; tropical cyclone intensity