PCA-BP 神经网络在 SO_2 浓度预报中的应用

于文革[1,2]　王体健[1]　杨　诚[2]　孙　莹[2]

(1. 南京大学大气科学系, 210093; 2. 辽宁省丹东市气象台)

提　要: 将基于主成分分析(PCA)的 BP 神经网络预报方法引入大气污染预报,建立 SO_2 浓度预报模型。结果表明:应用主成分分析对数据进行前处理,以原始预报因子的主成分作为 BP 神经网络的输入,降低了数据维数,消除了样本间存在的相关性,大大加快了 BP 神经网络的收敛速度。对模型进行预报验证,预报值与实际值之间的绝对误差为 0.0098,预报值与实际值的相关系数达到 0.885,得到较好的预报效果。并且比一般的 BP 神经网络模型具有较高的拟合和预报精度。

关键词: 主成分分析　BP 神经网络　大气污染　SO_2 浓度预报

Application of PCA-BP Neural Network to SO_2 Concentration Forecast

Yu Wenge[1,2]　Wang Tijian[1]　Yang Cheng[2]　Sun Ying[2]

(1. Atmospheric Science Department of Nanjing University, Nanjing 210093;
2. Dandong meteorological Observatory, Liaoning Province)

Abstract: Based on principal components analysis (PCA), the BP (Back Propagation) neural network forecast method is introduced in air pollution prediction and the SO_2 concentration prediction model is established. The results indicate that by applying the principal component analysis in the data pre-processing and taking the principal components of primitive predictor as the input of neural network, it can reduce the dimension of data, eliminate the correlation between the samples, and largely speed up the convergence rate. The verification of forecast model shows that the absolute error between the forecasts and the real value is 0.0098, and the correlation coefficient between them reaches 0.885. The PCA-BP model has a fit accuracy better than the common BP model.

Key Words: principal components analysis　BP neural network　air pollution　SO_2 concentration forecasting

收稿日期:2007 年 11 月 27 日;　修定稿日期:2008 年 2 月 3 日

引 言

随着大气污染问题日趋严重，城市空气污染作为一个主要的环境问题正迅速凸现出来[1]。开展大气污染预报业务，为公众为政府决策提供有价值的参考信息，具有十分重要的意义。自2001年6月47个环保重点城市联合发布环境空气质量预报至今，相继有许多城市开展大气污染预报业务。随着气象业务技术体制改革的不断推进以及业务领域的不断拓展，大气污染预报越来越受到各级气象和环保部门的重视，并将其纳入常规预报业务。近20年来，大气污染预报模式的研究得到了很大的发展。中国气象科学研究院大气物理研究所徐大海、朱蓉[2]建立的城市空气污染数值预报系统(CAPPS)，在对大气平流扩散方程积分基础上建立箱格预报模式，进行空气污染潜势预报；马雁军[3]、刘罡[4]、王俭[5]、周秀杰[6]等将人工神经网络引入到大气污染预报，并取得较好效果。

近些年来，人工神经网络(Artificial Neural Network，ANN)技术得到较大发展，尤其是BP神经网络，广泛用于大气科学领域[7-11]。但在实际应用中发现，当输入样本较多且具有多重共线性时，会降低网络的训练速度和效率，影响预报精度。为此，本文引入主成分分析方法对样本进行预处理，以达到数据降维，去除相关性的目的。结合BP神经网络，将主成分作为输入层节点，以减少网络训练时间，提高网络训练精度，并将其应用于大气污染物浓度预报。

1 资料来源及处理

研究所用SO_2浓度数据由丹东市环境监测中心提供，包括城区4个环境空气质量自动监测站2005年、2006年采暖季(12月至次年3月)SO_2浓度日均值。同步气象资料来源于丹东市气象台，包括气温、气压、风速等常规地面观测资料。用2005年的数据作为主成分分析和神经网络训练学习样本，建立神经网络模型，并采用2006年的数据进行预报验证。

计算4个空气自动站SO_2浓度的平均值，以代表丹东城区日平均SO_2浓度，将其作为预报量，将气温、气压、风速、相对湿度、蒸发等气象要素作为预报因子，进行相关分析，结果如表1所示。

由表1可知，除最小相对湿度、日照时数和平均气压，其余9个因子的相关系数检验显著水平均达到0.01。

表1 SO_2浓度与气象因子相关系数

预报因子	相关系数	预报因子	相关系数
平均气温	−0.396**	最大风速	−0.429**
最高气温	−0.319**	相对湿度	0.271**
最低气温	−0.440**	最小相对湿度	0.142
日照时数	−0.152	平均气压	0.147
蒸发量	−0.574**	混合层厚度	−0.340**
平均风速	−0.515**	前日SO_2浓度	0.290**

注：**表示显著水平达到0.01

2 SO_2浓度预报模型的建立

基于PCA-BP神经网络SO_2浓度预报模型的建立，关键在于预报因子的主成分分析以及BP神经网络输入模型的确定和训练数据的选取。下面以丹东城区SO_2日平均浓度预报为例，介绍预报模型的建立方法。

2.1 基于PCA的数据处理

将表1中相关系数检验显著水平达到0.01的预报因子构成9维随机向量$\boldsymbol{X}=(X_1,X_2,\cdots,X_9)$，应用SPSS统计软件对其进行主成分分析[12]，得出特征值和其对应的特征向量(见表2、表3)。

表 2　方差分解主成分提取分析

主成分	特征值	方差贡献比/%	累积方差贡献比/%
1	3.447	38.305	38.305
2	2.932	32.579	70.883
3	1.046	11.619	82.502
4	0.836	9.293	91.795
5	0.328	3.639	95.435
6	0.164	1.824	97.259
7	0.133	1.474	98.733
8	0.105	1.165	99.899
9	0.009	0.101	100.000

表 3　特征值对应的特征向量(载荷)

因子	主成分			
	1	2	3	4
X_1(前日 SO_2 浓度)	0.026	−0.115	0.813	−0.563
X_2(平均气温)	0.416	0.366	0.036	−0.011
X_3(最高气温)	0.399	0.356	−0.014	−0.127
X_4(最低气温)	0.397	0.362	0.091	0.075
X_5(蒸发量)	−0.090	0.471	−0.288	−0.452
X_6(平均风速)	−0.320	0.362	0.276	0.322
X_7(最大风速)	−0.373	0.318	0.217	0.135
X_8(相对湿度)	0.401	−0.115	0.314	0.542
X_9(混合层厚度)	−0.315	0.358	0.156	0.204

由表 2 可以看出,前 4 个主成分的累积方差贡献比为 91.795%>85%。据此,提取 4 个主成分,结合表 2 的主成分系数构成如下 4 个主成分:

$$\begin{aligned}Z_1 = {} & 0.026X_1 + 0.416X_2 + 0.399X_3 + \\ & 0.397X_4 - 0.090X_5 - 0.320X_6 - \\ & 0.373X_7 + 0.401X_8 - 0.315X_9 \quad (1)\end{aligned}$$

$$\begin{aligned}Z_2 = {} & -0.115X_1 + 0.366X_2 + 0.356X_3 + \\ & 0.362X_4 + 0.471X_5 + 0.362X_6 + \\ & 0.318X_7 - 0.115X_8 - 0.358X_9 \quad (2)\end{aligned}$$

$$\begin{aligned}Z_3 = {} & 0.813X_1 + 0.036X_2 - 0.014X_3 + \\ & 0.091X_4 - 0.288X_5 + 0.276X_6 + \\ & 0.217X_7 + 0.314X_8 + 0.156X_9 \quad (3)\end{aligned}$$

$$\begin{aligned}Z_4 = {} & -0.563X_1 - 0.011X_2 - 0.127X_3 + \\ & 0.075X_4 - 0.452X_5 + 0.322X_6 + \\ & 0.135X_7 + 0.542X_8 + 0.204X_9 \quad (4)\end{aligned}$$

从主成分系数可以看出,平均气温、相对湿度在第一主成分上有较高载荷,第二主成分中蒸发量有较高的载荷,前日 SO_2 浓度在第三和第四主成分上有较高的载荷。

2.2　PCA-BP 神经网络模型的建立

利用式(1)～(4)求得的 4 个主成分(Z_1,Z_2,Z_3,Z_4)作为 BP 神经网络输入层神经元矩阵,预报日 SO_2 浓度标准化后(FSO_2)作为其输出层神经元,利用 MATLAB 提供的神经网络工具箱[13](Neural Network Toolbox,简称 NNbox)构建 3 层 BP 神经网络模型,如图 1 所示。其中输入层为 4 个神经元,输出层为 1 个神经元。隐含层节点数的选择至今尚未找到一个很好的解析式,往往根据前人的经验和反复试验来确定。在相同总体误差情况下,选择收敛最快的;或用不同的隐层节点数训练网络,最佳隐层节点数由网络训练的最小系统误差确定出。

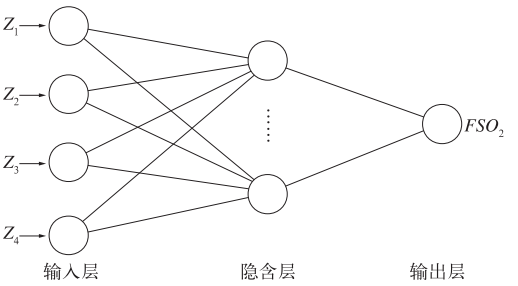

图 1　SO_2 浓度预报 BP 神经网络模型

将 PCA 处理后的 2006 年采暖季数据作为 BP 神经网络的训练样本,设隐层节点数初值为 5,通过改变隐层节点数对网络进行训练,训练结果如表 4 所示。

表 4　不同隐层节点数的 BP 神经网络训练结果

隐层节点	5	6	7	8	9	10	11	12
系统误差	0.2158	0.1403	0.1090	0.0719	0.0583	0.0235	0.0492	0.0425

由表 4 可以看出,当隐层节点数为 10 时,BP 神经网络的系统误差最小,由此建立输入层为 4 个神经元,输出层为 1 个神经元,隐含层为 10 个神经元的 BP 神经网络,并利

用该网络训练输出的隐含层权值 w_1 和阈值 b_2 及输出层权值 w_2 和阈值 b_2，应用 MATLAB 的仿真函数 simuff()建立采暖季 SO_2 预报模型：

$$FSO_2 = \text{simuff}(P, w_1, b_1, 'tansig', w_2, b_2, 'Purelin') \quad (5)$$

其中：FSO_2 为预报日 SO_2 浓度(标准化)；P 为主成分矩阵；tansig 为正切 S 型传递函数；Purlin 为线性传递函数。

利用式(5)对 2005 年采暖季 SO_2 浓度进行拟合检验，并对 2006 年采暖季 SO_2 浓度进行预报验证。

3 模型拟合及预报效果检验

为了对比分析，建立了一般 BP 神经网络模型(数据未进行 PCA 处理)。利用 2005 年采暖季数据进行拟合检验。图 2 为一般 BP 神经网络模型和 PCA-BP 神经网络模型 SO_2浓度拟合值与实际值的对比，经计算二者的拟合误差分别为 0.005 和 0.002，拟合值与实际值相关系数分别为 0.912 和 0.985。通过对比分析不难发现，基于 PCA-BP 神经网络模型比一般的 BP 神经网络模型拟合结果更为精确。由此可以说明，基于 PCA-BP 神经网络模型对 SO_2 浓度具有良好的拟合能力。

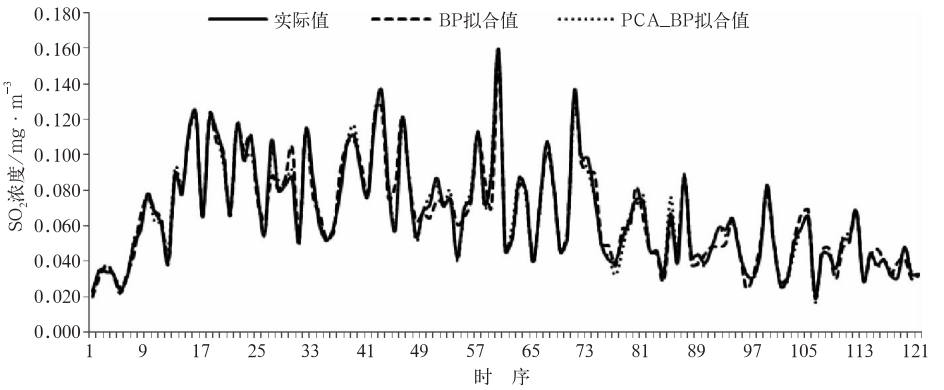

图 2 PCA-BP 和 BP 神经网络模型 SO_2 浓度拟合值与实际值对比

利用 2006 年采暖季数据进行预报验证，图 3 为应用一般的 BP 神经网络模型和 PCA-BP神经网络模型SO_2浓度预报值与实

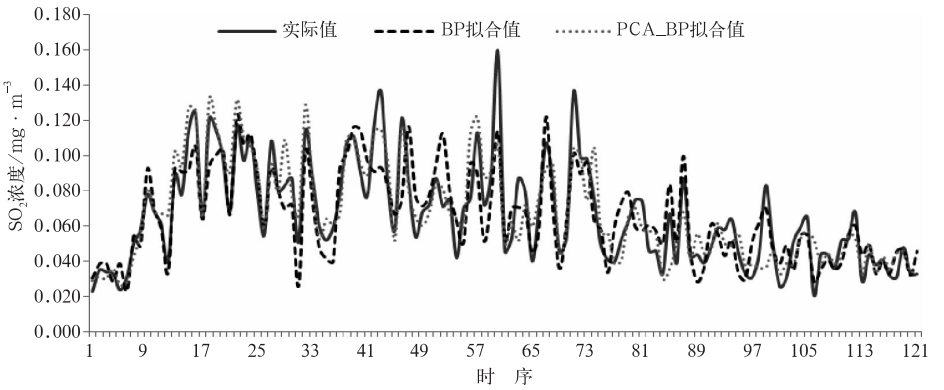

图 3 PCA-BP 和 BP 神经网络模型 SO_2 浓度预报值与实际值对比

际值对比,经计算二者的平均绝对误差为分别为 0.0123 和 0.0098,预报值与实际值的相关系数分别为 0.820 和 0.885。通过比较,可以看出,基于 PCA-BP 神经网络模型对采暖季 SO_2 浓度具有良好的预报能力,其性能优于一般的 BP 神经网络模型,且预报结果更为准确。如果充分考虑影响采暖季 SO_2 浓度的各种因子(如高空观测气象要素等),选择合理的隐层节点数,便能得到较好的预报效果。

4 结　论

(1) 将主成分分析引入 BP 神经网络的前处理,以原始预报因子的主成分作为网络的输入,降低了数据维数,消除了样本间存在的相关性,大大加快了 BP 神经网络的收敛速度,并且比一般的 BP 神经网络模型具有较高的拟合和预报精度。

(2) 将 PCA-BP 神经网络模型应用于丹东城区采暖季 SO_2 浓度预报,预报值与实际值之间的绝对误差为 0.0098,预报值与实际值的相关系数达到 0.885。

(3) 应用 2005 年一个采暖季的资料建立 PCA-BP 神经网络模型,并对 2006 年采暖季 SO_2 浓度进行预报试验,虽然得到了比较好的预报效果,但不排除偶然性的存在。随着观测资料的积累,将对模型做适当调整,以期建立更加稳健的预报模型。

(4) 鉴于 PCA-BP 神经网络模型对对采暖季 SO_2 浓度具有良好的预报效果,可推广应用于其他大气污染物的浓度预报,但该方法是否适用,而且是否存在季节性的差异,能否应用于大气污染物浓度的逐时预报,还有待进一步研究,并将撰文详述。

参考文献

[1] 张继娟,魏世强. 我国城市大气污染现状与特点[J]. 四川环境,2006,25(3):104.

[2] 徐大海,朱蓉. 大气平流扩散的箱格预报模式与污染潜势指数预报[J]. 应用气象学报,2000, 11 (1):1-12.

[3] 马雁军,杨洪斌,张云海. BP 神经网络法在大气污染预报中的应用研究[J]. 气象,2003,29(7):49.

[4] 刘罡,李听,胡非. 大气污染物浓度的神经网络预报[J]. 中国环境科学,2000,20(5):429-431.

[5] 王俭,胡筱敏,郑龙熙,等. 基于 BP 模型的大气污染预报方法的研究[J]. 环境科学研究,2002,15(5):62-64.

[6] 周秀杰,苏小红,袁美英. 基于 BP 网络的空气污染指数预报研究[J]. 哈尔滨工业大学学报,2004,36(5):582-585.

[7] 张承福. 人工神经网络在天气预报中的应用研究[J]. 气象,1994,20(6):43-47.

[8] 周曾奎,韩桂荣,朱定真,等. 人工神经网络台风预报系统[J]. 气象,1996,22(1):18-21.

[9] 汤子东,郑世芳,奚秀芬. BP 人工神经元网络在春季降水量预报中的应用[J]. 气象,1997,23(8):34-37.

[10] 施丹平. 人工神经网络方法在降水量级中期预报中的应用[J]. 气象,2001,27(6):40-45.

[11] 段婧,苗春生. 人工神经网络在梅雨期短期降水分级预报中的应用[J]. 气象,2005,31(8):31-36.

[12] 米红,张文璋. 实用现代统计分析与 SPSS 应用[M]. 北京:当代中国出版社,2000,10.

[13] 丛爽编. 面向 MATLAB 工具箱的神经网络理论与应用(第 2 版)[M]. 北京:中国科学技术大学出版社,2003.5.