综 述 与 专 论

马尔可夫修正的 BP 神经网络在 PM2.5 预测中的应用

陈广银, 蔡灏兢, 姜 欣

(昆山市环境监测站,江苏 昆山 215300)

摘要:本文以昆山某点位的监测数据及天气网的气象数据为基础,选取了影响 PM2.5 因素中的 10 个指标进行了相关性分析,结果表明 PM2.5 与 PM10 是高度相关,与 CO、SO_2、NO_2、O_3 显著相关,并依据分析结果对模型输入数据进行了降维。运用 BP 网络对序列 1-16 的 PM2.5 进行预测,结果显示其误差在-25 %~-26.9 %。将预测误差划分为 4 个状态,计算概率转移矩阵,并对序列 17、18 的 BP 预测结果进行修正,结果显示修正后的误差由 BP 网络的-14 %、-25 %降为-7.1 %、-8.3 %,预测准确度大大提高,表明基于马尔可夫-BP 神经网络模型在昆山 PM2.5 预测中具有一定的现实意义。

关键词:PM2.5;马尔可夫链; BP 模型;预测;修正

中图分类号:X831 **文献标识码**:A **文章编号**:1006-8759(2017)05-0008-04

APPLICATION OF CORRECTED BP NEURAL NETWORK IN PM2.5 PREDICTION BASED ON MARKOV

CHEN Guang-yin,CAI Hao-jing,JIANG Xin

(Kunshan Environmental Monitoring Station, KunShan, 215300 , China)

Abstract:ased on the meteorological data of a Kunshan´ monitoring site and Weather Network, the correlativity of 10 influencing factors on PM2.5 were analyzed. The results showed that PM2.5 and PM10 are highly correlated, positively correlated with CO, SO2, NO2, O3. According to the analytic result, model input data dimensionality was reduced. Using BP network to predict PM2.5 of the sequence 1-16, the results show that the error is from-25% to 26.9%. The prediction error is divided into 4 states, and the BP prediction results of 17、18 are revised by probability transfer matrix. Results show that the error by the BP network was from -14%、-25% downed to-8.3% 、-7.1% which was modified by Markov. Accuracy is greatly improved and it Indicates that the Markov-BP neural network model has a certain practical significance in Forecast of PM2.5 in Kunshan .

Key words: PM2.5;Markov chain ; Backpropagation Model; Prediction;Correction.

PM2.5 是指空气动力学等效直径等于或小于 2.5 微米的大气颗粒物, 其不仅降低大气能见度,显著减少日照,导致雾天增多,更对人体健康造成严重的危害[1]。DQ Rich 等[2]研究证实,PM2.5 会对呼吸系统和心血管系统造成伤害,导致肺癌、心血管疾病、出生缺陷甚至过早死亡。准确预测 PM2.5 的浓度对于人们预防其危害具有重要的现实意义,然而中国相关研究起步较晚,大多数研究侧重化学组成、来源解析以及其与气象因子之间的关系[3-4]。目前对 PM2.5 的预测多采用多元线性回归模型、自回归移动平均模型、时间序列,灰色系统等预测方法[5-8]。PM2.5 形成是非常复杂的化学、物

收稿日期:2017-06-20

理过程,表现出强烈的非线性特征,上述预测方法存在很大局限性，不能体现 PM2.5 与其影响因素之间的规律,拟合度较差。

神经网络具有非常强大的非线性映射能力,已有学者将其引入到了预测领域。1993 年,Božnar 等 [9] 首次运用神经网络预测 slovenia 地区的 SO_2 小时浓度，并与基于大气扩散模式的预测结果进行对比，表明神经网络预测结果更优。Perez 等[10]用神经网络预测了圣地亚哥 PM2.5 h 浓度。李作泳等[11]首次运用 BP 神经网络其建立某市 SO_2 质量浓度预测的模型，预则误差小于等于 25 %。王敏等[12]将 BP 神经网络用于城市 PM2.5 预测,并和普通克里格插值法进行了比较，显示 BP 预测 PM2.5 的优越性。虽然 BP 网络在预测 PM2.5 上有较大的优越性,但其极易陷入局部最优,影响预测准确度[13]。PM2.5 影响因素复杂,且其浓度呈现出某种随机波动的特征。马尔可夫模型适用于数据序列随机波动大的预测并被广泛应用到交通事故预测、降水预测中[14–15]。Dong M 等[16]在芝加哥地区用马尔可夫模型预测 PM2.5，结果显示模型可以准确预测未来 24 h 内的 PM2.5 浓度等级。甘茂林等 [17] 运用马尔可夫对 PM2.5 的浓度等级进行预则,结果显示该模型是有效的。龚明等[18]采用灰色系统对 PM2.5 预测，并用马尔可夫链对结果进行修正。本文综合运用了马尔可夫链和 BP 神经网络模型,对昆山市 PM2.5 的浓度值进行预测,以期改善传统的 BP 模型及 Markov 模型预测误差大的问题。

1 数据来源及预处理

历史气象数据来源于天气网，监测数据来源于昆山某监测点位(2016.4.1–5.31)共 61 天数据。用 MATLAB R2013a 将这些数据与 PM2.5 进行相关系数分析,结果如表 1:

表 1 PM2.5 相关系数分析

项目	最高温	最低温	天气状况	风向	SO_2	NO_2	PM10	CO	O_3
PM2.5 相关系数	0.29	−0.01	−0.25	−0.40	0.61	0.57	0.83	0.68	0.57

从上表可以看出，与 PM2.5 相关性系数 PM10>CO>SO_2>NO_2>O_3, 根据|r|<0.4 为低度线性相关;0.4≤|r|<0.7 为显著性相关;0.7≤|r|<1 为高度线性相关判断，昆山采样点处的 PM2.5 与 PM10 是高度相关,与 CO、SO_2、NO_2、O_3 显著相关,与气象条件低度相关。SO_2 是燃煤的特征污染物,CO、NO_2 是汽车尾气的特征污染物 [19]，说明昆山 PM2.5 的形成与机动车保有量及燃煤有着显著关联。气象参数也是影响 PM2.5 质量浓度的因素 [20],但本文相关分析结果显示 PM2.5 与气象参数相关性不大,通过分析原始数据发现,所选 61 天数据中的天气状态较为稳定,比如风力都是 3~4 级,而且仅代表市平均状况，而 PM2.5 的数据是一个点位所测,属局部特征,所以导致 PM2.5 与气象参数低度相关的结果。SO_2、NO_2、CO 是形成二次污染物前趋物质，其形成的一次污染物是 PM2.5 重要组成成分。通过以上分析,本文选取了与 PM2.5 相关系数较高的 SO_2、NO_2、CO、O_3、PM10、PM2.5 作为输入参数,次日 PM2.5 为输出参数。

2 马尔可夫及 BP 原理

2.1 马尔可夫模型

马尔可夫随机过程理论指出,系统将来所处的状态只与现在的状态有关，而与过去无关，马尔可夫预测根据系统状态间的转移概率预测系统未来的发展。转移概率反映了各随机因素对系统的影响程度及系统各状态间的内在规律性[21]。设某系统在时刻 t=n 有 k 个可能状态，即 Xn = 1,2,3 , …, k(n =0, 1, …), $a_i(n)$表示系统在时刻 t= n 处于状态 i 的状态概率,即 $a_i(n)=P(X_n=i)$,时刻 t=n+1 转移到状态 j 的概率为 p_{ij} (i, j =1 ,2, …, k)，即将来原状态只与现在的状态有关的条件概率 $p_{ij}=P(X_{n+1}=j|\ X_n=i)$,称 p_{ij} 为一步转移概率，$P=\{p_{ij}\}$为一步转移概率矩阵,简称转移概率和转移概率矩阵。K 步转移概率矩阵如下式(1)。

$$P^k=\begin{bmatrix} P_{11}^k & P_{12}^k & \cdots & P_{1j}^k \\ P_{21}^k & P_{22}^k & \cdots & P_{2j}^k \\ \vdots & \vdots & \vdots & \vdots \\ P_{j1}^k & P_{j2}^k & \cdots & P_{jj}^k \end{bmatrix} \quad (1)$$

其中 $p_{ij}\geqslant 0$, $\sum_{j=1}^{k}P_{ij}=1$ $i=1, 2, \cdots, k$。

$P_{ij}^k=P_{ij}^k=\frac{n_{ij}^k}{N_i}$, n_{ij}^k 为状态 Xi 经过 K 步变为 X_j 的次数,N_i 为状态 X_i 出现的总次数。设初始状态向量为 A^0,则经过 K 步转移后的向量为:

$$A^k=A^0*P^k \quad (2)$$

根据式(2)就可以由误差的初始状态及概率转移矩阵预测将来的误差变化趋势[22–23]。

2.2 BP 神经网络

人工神经网络是模仿生物神经系统的功能和结构发展起来的信息处理系统(图 1)。是由大量简单的神经元相互连接构成的复杂网络系统，其对非线性系统具有很强的模拟映射能力[24]。BP 神经网络的模型结构和权值通过学习过程得到。学习过程分为 2 个阶段：多层前馈阶段，即从输入层开始依次计算各层各节点的实际输入、输出；反向误差修正阶段，即根据输出层神经元的输出误差，沿原路反向修正各连接权值，使误差减少[25]。

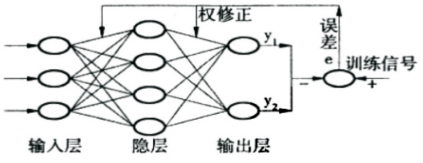

图 1　BP 神经网络原理

设输出层有 m 个神经元，BP 网络的实际输出是 y；期望输出是 y′；损失函数 ε 为：

$$\varepsilon = \frac{1}{2}\sum_{j=1}^{m}(y_j - y_j')^2 \quad (3)$$

每个权值的修正值为

$$\Delta\omega_{ij} = -\eta\frac{\delta\varepsilon}{\delta\omega_{ij}} = -\eta\frac{\delta\varepsilon}{\delta I_j}\frac{\delta I_j}{\delta\omega_{ij}} \quad (4)$$

通过调整权值 ω_{ij}，使损失函数最小，达到网络的最优化。式中：ω_{ij} 为输入单元 i 到隐含层单元 j 的权重；η 是学习速率，是中间第 j 个隐含层的传输函数。输入层到隐含层的函数采用 Logsig 型，隐含层到输出层的函数采用 Purelin 型[26]。

3　建模过程与修正结果

根据相关系数分析结果，采用 MATLAB R2013a 进行仿真实验，通过对模型参数进行多次调试，最终选定学习率为 0.03、训练精度 0.0001、最大训练次数 500、隐含层神经元个数 9，训练函数 trainlm；模型仿真结果如图 2：

网络将输入向量的 70 %作为训练网络，15 %用于模型的验证，15 %用于模型的预测，结果如图 2，得到了较为满意仿真结果。使用构建好的网络对序列样本 17、18 进行预测，BP 网络对序列 1–16 的预测值及与实际测量值的误差如表 2，将误差分为四个状态，E1:–30 %~–15 %、E2:–15 %~0 %、E3:0 %~15 %、E4:15 %~30 %。

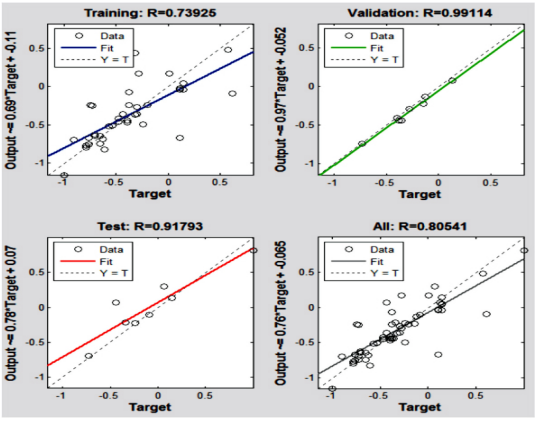

图 2 .BP 神经网络测试误差

表 2　马尔科夫模型状态

序列	实际测量值(mg/m³)	BP 预测值(mg/m³)	相对误差/%	状态
1	44	51	15.9	E4
2	56	51	–8.9	E2
3	30	25	–16.7	E1
4	24	30	25.0	E4
5	43	41	–4.7	E2
6	46	50	8.7	E3
7	45	41	–8.9	E2
8	47	45	–4.3	E2
9	23	24	4.3	E3
10	6	4.5	–25	E1
11	39	42	7.7	E3
12	43	42	–23	E2
13	69	61	–11.6	E2
14	51	52	2.0	E3
15	34	29	–14.7	E2
16	78	99	26.9	E4

根据表 2 误差状态划分，建立相应的状态转移概率矩阵，如下式：

一步状态概率转移矩阵：

$$P^1 = \begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{7} & \frac{2}{7} & \frac{3}{7} & \frac{1}{7} \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (5)$$

二步状态概率转移矩阵：

$$P^2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 6 & \frac{1}{2} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix} \quad (6)$$

根据状态转移矩阵对序列 17、18 的 BP 预测结果进行误差修正，序列 16 为初始状态，其向量为 A^0=[0 0 0 1]，乘以一步状态转移矩阵 P^1 得序列 17 的状态为 E2，对应的相对误差为(−15 %+0 %)/2=−7.5 %，则修正后的 BP 预则值 24x (1+7.5 %)=26。同理，初始向量乘以 P^2 得序列 18 的状态向量为[$\frac{1}{2}$ 0 $\frac{1}{2}$ 0]，既序列 18 的误差状态 E1 和 E3 的概率各占 50 %，则序列 18 的修正值为 29x(1+15 %)=33。

表 3 马尔科夫修正值与实际测量值的比较

序列	实际测量值(mg/m³)	BP 预测值(mg/m³)	相对误差/%	状态
17	28	24	−14	−7.1
18	36	29	−25	−8.3

4 结论

通过相关性分析发现，PM2.5 含量与 PM10 是高度相关，与 CO、SO_2、NO_2、O_3 含量显著相关，间接说明了昆山 PM2.5 的形成与机动车保有量及燃煤有着显著关联，通过对相关系数分析，对模型参数进行了降维。

PM2.5 的形成因素比较复杂，表现出较强的随机性，很难用一种数学模型进行准确预测。本文利用 BP 网络强大的非线性映射能力和马尔可夫随机性过程对 PM2.5 进行预测，结果显示其误差范围在−25 %~26.9 %。将误差划分为 4 个状态，通过概率转移矩阵对 BP 预则误差进行修正后，预测误差由 BP 网络的−14 %、−25 %降为−7.1 %、−8.3 %，准确度大大提高，说明经马尔可夫修正后的 BP 网络模型对 PM2.5 的预测具有一定的现实意义。

参与文献

[1] HaoJ,Kebin He,DuanL,etal.Air pollution and it´s control in China [J].Frontiers of Environmental Science&Engineering in China,2007,1 (2):129−142.

[2] DQ Rich,HM Kipen,W Huang,etal.Association between Changes in Air Pollution Levels during the Beijing Olympics and biomarkers of inflammation and thrombosis in healthy young adults [J].Journal of the American Medical Association,2012,307 (19):2068−2078.

[3]董海燕,古金霞,陈魁等.天津市区 PM2.5 中碳组分污染特征及来源解析[J].中国环境监测,2013,29(1):34−38 .

[4]朱倩茹,刘永红,徐伟嘉等.广州 PM2.5 污染特征及影响因素分析[J].中国环境监测,2013,29(2):15−21.

[5]胡玉筱,段显明.基于高斯烟羽和多元线性回归模型的 PM2.5 扩散和预测研究[J].干旱区资源与环境,2015,29(6):86−92.

[6]张玉丽,何玉,朱家明.基于多元线性回归模型 PM2.5 预测问题的研究[J].安徽科技学院学报, 2016,30(3):92−97.

[7] 李修成.城市范围内 PM2.5 时间预测研究[D].哈尔滨:哈尔滨工业大学,2016:1−60.

[8] 冯科展,解建军,张玫等.灰色模型在 PM2.5 预测中的应用[J].绵阳师范学院学报,2015,34(5):75−79.

[9] Bo?nar, M., Lesjak, M., Mlakar, P. A neural network−based method for the short time predictions of ambient SO 2 concentrations in highly polluted industrial areas of complex terrain[J]. Atmospheric Environment,1993, 27B:221−230.

[10] Perez P, Trier A, Reyes J. Prediction of PM2. 5 concentrations several hours in advance using neural networks in Santiago, Chile[J]. Atmos Environ,2000;34:1189−96.

[11] 李祚泳,邓新民.环境污染预测的人工神经网络模型[J].成都气相学院学报,1997,12(4):280−283.

[12]王敏,邹滨,郭宇等.基于 BP 人工神经网络的城市 PM2.5 浓度空间预测[J].环境污染与治,2013,35(9):65−70.

[13]刘春艳 ,凌建春,寇林元等.GA−BP 神经网络与 BP 神经网络性能比较[J].中国卫生统计,2013,30(2):173−180.

[14]张嘉琦. 道路交通事故死亡人数预测模型对比研究[J].中国安全科学学报,2016,26(9):46−49.

[15]廖捷,胡豪然,陈功. 叠加马尔科夫链在年降水量预测中的应用[J],安徽农业科学,2012,40(9):5532−5533.

[16] Yang D, Kuang Y, He D, etal. PM2. 5 concentration prediction using hidden semi−Markov model−based times series data mining[J]. Expert Syst Appl ,2009,36:9046−9055.

[17]甘茂林,吕王勇,符璐.基于离散参数马尔科夫链的 PM2.5 预测[J].安全与环境工程,2016,23(1):37−39.

[18]龚明,叶春明.基于修正灰色马尔科夫链的上海市 PM2.5 浓度预测[J].自然灾害学报,2016,25(5):97−104.

[19]宋宇,唐孝炎,方晨等.北京市大气细粒子的来源分析[J],环境科学,2002,23(6):11−16.

[20]孙语茉,张建.上海市 PM2.5 变化特征与气象因素相关性分析[J],社科学论,2015,152−154 .

[21]王永刚,吕学梅.民航事故征候的灰色马尔可夫预测[J],安全与环境学报,2008,8(1):163−165.

[22]白根川,夏建国,杨娟.马尔柯夫法在土地利用结构趋势预测中的应用[J],湖北农业学,2009,48(4):847−850.

[23]付长贺,邓甦.马尔科夫链在传染病预测中的应用[J], 沈阳师范大学学报,2009,27(1):28−30.

[24]周振民,刘荻.基于 Matlab 的人工神经网络用水量预测模型[J].中国农村水利水电,2007(4):45−47.

[25]毛政利,闫继涛,赖健清.基于 BP 人工神经网络的成矿预测模型[J]. 金属矿山,2009(7):66−68.

[26]张宏,马岩,李勇 .基于遗传 BP 神经网络的核桃破裂功预测模型[J].农业工程学报,2014,30(18):78−83 .