神经网络模型在 O_3 浓度预测中的应用

沈路路，王聿绚，段雷*
(清华大学环境科学与工程系，北京 100084)

摘要：O_3 是近地面大气中一种重要的二次污染物. 本研究采用神经网络多层感知器(Multi-Layer Perceptron)和多元线性回归2种模型，以广州万顷沙站2006年的气象观测数据为输入，对该站 O_3 的1 h平均峰值浓度进行提前1 d的预测，并比较了2种模型的预测效果. 模型的输入参数为前1d O_3 的最高1h平均浓度和第二天的气象参数(温度、湿度、风速、风向、气压和光照). 为了降低神经网络的复杂度以提高模型的泛化能力，采用了OBS(Optimal brain surgeon)方法对神经网络模型进行了修剪. 结果表明，经过修剪后的神经网络预测结果的准确指数(agreement index)为92.3%，RMSE为0.042 8 mg/m^3，R-square为0.737，重污染事件(1 d中 O_3 峰值浓度超过0.20 mg/m^3)的预报准确率为77.0%. 为了进一步提高重污染事件发生概率大小的预报效果，采用了神经网络分类器对臭氧的污染级别进行预测，该处理后重污染事件预报准确率可以达到83.6%. 综合比较神经网络模型和多元线性回归模型的拟合效果后发现，神经网络模型在 O_3 峰值预报中具有明显优势，本研究建立的神经网络模型具有臭氧污染预测预警的实用价值.

关键词：神经网络模型；多层感知器；O_3 污染预测；多元线性回归；预报模型
中图分类号：X51 文献标识码：A 文章编号：0250-3301(2011)08-2231-05

Application of Artificial Neural Networks on the Prediction of Surface Ozone Concentrations

SHEN Lu-lu, WANG Yu-xuan, DUAN Lei
(Department of Environmental Science and Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Ozone is an important secondary air pollutant in the lower atmosphere. In order to predict the hourly maximum ozone one day in advance based on the meteorological variables for the Wanqingsha site in Guangzhou, Guangdong province, a neural network model (Multi-Layer Perceptron) and a multiple linear regression model were used and compared. Model inputs are meteorological parameters (wind speed, wind direction, air temperature, relative humidity, barometric pressure and solar radiation) of the next day and hourly maximum ozone concentration of the previous day. The OBS (optimal brain surgeon) was adopted to prune the neutral work, to reduce its complexity and to improve its generalization ability. We find that the pruned neural network has the capacity to predict the peak ozone, with an agreement index of 92.3%, the root mean square error of 0.042 8 mg/m^3, the R-square of 0.737 and the success index of threshold exceedance 77.0% (the threshold O_3 mixing ratio of 0.20 mg/m^3). When the neural classifier was added to the neural network model, the success index of threshold exceedance increased to 83.6%. Through comparison of the performance indices between the multiple linear regression model and the neural network model, we conclud that that neural network is a better choice to predict peak ozone from meteorological forecast, which may be applied to practical prediction of ozone concentration.

Key words: neural network; multilayer perceptron; prediction of O_3 contamination; multiple linear regression; prediction model

O_3 是大气中一种非常重要的氧化剂，而且是光化学烟雾的一种重要指示物质[1] 但是 O_3 不同于其他污染物，因为它并不是由人为活动直接排放的，而是来自于大气中复杂的光化学反应，其反应主要前体物为氮氧化物(NO_x)和非甲烷烃(NMHC)[2] O_3 作为一种重要的二次污染物，已经受到人们的广泛关注[3] 近地面 O_3 浓度的升高对人体健康和植物都具有不利影响. 流行病调查显示，在 O_3 浓度较高时期，特别是光化学烟雾污染发生时期，往往伴随着人群死亡率的大幅度增长[4,5] 高浓度的 O_3 对植物的生长也有明显的危害作用，特别是在植物的生长发育时期[6] 出于对人们健康因素的考虑，提前预报 O_3 峰值浓度具有重要意义.

尽管 NO_2、CO 等前体物质排放量的改变会影响每天 O_3 的浓度，但是研究发现，气象条件通过影响光化学反应进程，从而影响到 O_3 的生成和分解，对 O_3 浓度的生成也具有重要的作用[7] 由于 NO_x 和 NMHC 的排放与工业生产和城市活动密切相关，对一个特定的城市或城市群而言，这些前体物的排放量具有一定的、可估计的变化范围和变化规律，而臭

收稿日期：2010-09-30；修订日期：2010-12-06
基金项目：国家高技术研究发展计划(863)项目(2006AA0A307)
作者简介：沈路路(1987～)，男，硕士研究生，主要研究方向为大气化学模式，E-mail：sllhappy0729@163.com
* 通讯联系人，E-mail：lduan@tsinghua.edu.cn

氧的逐日和逐时变化则主要是由于气象条件变化及其带来的传输途径的变化而引起的. 因此,如果能找到气象条件和 O_3 浓度之间的统计关系,就有可能预报出 O_3 浓度[8] 多元线性回归是常用的寻找输入参数(比如气象条件)和 O_3 浓度之间统计关系的方法之一[9,10] 多元线性回归需要进行变量形式的转换,可以在一定程度上达到模拟 O_3 生成的非线性过程的目的[11] . 但是研究显示,线性模型在高浓度 O_3 的预测效果不太理想[12] . 此外,广义加法模型[13] ,时间序列分析[14] ,模糊系统模型[15] ,神经网络模型[2,14,16,17] 等也被应用到研究中,其中神经网络模型得到最广泛的关注,并在实际预测中得到了应用[12] .

本研究在广州万顷沙站 2006 年气象观测数据的基础上,用多元线性回归和神经网络的多层感知器模型(MLP)探索通过气象条件预报 O_3 浓度的方法,并对 2 种预测方法的结果进行比较,这为 O_3 浓度的实际预报方法进行了理论上的探索.

1 数据处理

本研究中用到的是万顷沙站 2006 年气象和臭氧浓度观测数据. 在时间序列中,数据连续缺失3 h 以下(含 3 h)的采用线性插值法进行填补,缺失3 h 以上的不予处理. 通过该处理方法后,万顷沙站气象数据总共缺失 327 h,占全年总数的 3.7%. 臭氧小时浓度超过我国环境空气质量二级标准时(即 1 个标准大气压下及 25℃ 时的臭氧小时浓度不超过 0.20 mg/m^{3}或者 102 ppbv)在本研究中被定义为臭氧重污染. 结果显示,万顷沙站的重污染小时为 204 h,比例为 2.33%.

2 全年信息提取

本次预报工作主要是通过前 1 d 的 O_3 浓度和第 1 d 的气象条件去预报第 2 d 的 O_3 最高浓度,特别是预测重污染发生的可能性. 在实际运行中,第 2 d 的气象条件可以通过天气预报模式获得,因此在本研究中将第 2 d 的气象条件作为臭氧统计预报模式的输入参数. 在 O_3 重污染预报中,本研究以天(24 h)为单位对观测数据进行处理,分别提取以下变量形式[如果该天 08:00 ~ 19:00 缺失数据 3 h(含 3 h)以上的,该天以缺失处理]. 经过处理后的全年万顷沙站有效数据有 353 个. 为了预报臭氧,本研究提取的变量形式如表 1.

表 1 O_3 浓度预报提取的不同变量形式

Table 1 Different variable formats used in the prediction of O_3 concentration

变量	变量的提取形式	提取形式的英文前缀	变量提取的时间范围
温度、风速、风向、相对湿度、气压	最大值	Max	08:00 ~ 19:00
	最小值	Min	
	极差	cha = max − min	
	均值	Mean	
	标准差	Std	
光照	中午均值	midday	11:00 ~ 14:00
风速、相对湿度	对 O_3 生成具有重要影响的一段时间的均值	mean_important	10:00 ~ 16:00
前 1 d 臭氧	前 1 d 的 O_3 峰值浓度		

3 研究方法

3.1 软件

SAS 9.1: Copyright ⓒ 2002–2004, SAS Institute Inc., Cary, NC, USA. All rights reserved. Produced in the United States of America.

3.2 模型

本研究主要采用了多元线性回归模型和神经网络模型.

3.2.1 多元线性回归模型

多元线性模型中考虑的变量包括 yesterday_O_3、max_temperature、min_temperature、mean_important_speed、max_humidity、min_humidity、midday_solar、temp_cos、temp_sin 和 mean_pressure,其中 temp_cos 和 temp_sin 是白天平均风向的余弦和正弦值. 模型采用 stepwise 过程筛选变量,最后保留在模型中的变量必须在 0.95 置信度下通过 t 检验.

3.2.2 神经网络模型

在人工神经网络中,多层感知器(Multi-Layer Perceptron)是运用最广泛的一种模型[18] 一个典型的神经网络由 3 部分组成:输入层(input layer)、隐含层(hidden layer)和输出层(output layer). 对于一个含有 n_I 个输入参数 X,n_H 个隐含层单元的神经网络,神经网络的输出值 y_p 如式(1).

$$y_p = \sum_{j=1}^{n_H} W_j h\left(\sum_{i=1}^{n_I} w_{ji} x_i + b_j\right) + B \qquad (1)$$

式中,h 是激活函数,b_j 和 B 是神经网络的偏差(Bias),w_{ji}是第 j 个输入层单元到第 i 个隐含层的权重,W_j 是第 j 个隐含层单元到输出层的权重. 利用一系列数据对神经网络进行训练,寻找最佳权重 W 使得目标函数 E 最小. 在本研究中,神经网络的目标函数采用式(2)计算.

$$E = \frac{1}{2n}\sum_{i=1}^{n}(y - y_p)^2 \tag{2}$$

式中,y 为实际观测值,而 y_p 为预测值. 本研究中,神经网络模型的组合函数为线性函数,激活函数为 tanh 函数,训练算法为 Levenberg-Marquardt 算法.

3.2.3　神经网络模型的 OBS 修剪

为了降低神经网络模型的复杂程度,本研究中采用了 OBS(Optimal brain surgeon)[19] 方法对神经网络的权进行修剪. 该修剪方法将参数的重要性定义为将参数值设为 0 所带来的成本[20],也就是目标函数的增加值. 本研究对目标函数 $E(w)$ 进行泰勒展开:

$$\begin{aligned}\Delta E &= E(w + \Delta w) - E(w)\\ &\approx \left(\frac{\partial E}{\partial w}\right)^T \Delta w + \frac{1}{2}[\Delta w]^T H \Delta w\end{aligned} \tag{3}$$

式中,$H = \frac{\partial E}{\partial w \partial w^T}$为 Hessian 矩阵. 笔者假定神经网络经过训练后已经达到目标函数的局部最小,因此$\frac{\partial E}{\partial w}$近似等于 0,而$\frac{1}{2}[\Delta w]^T H \Delta w$ 近似为目标函数的变化. 如果删掉权重 w_j,那么认为该权重被认为 0,同时其他权重也会发生相应的变化,以使目标函数的增加值最小,表示如下:

$$e_j^T \partial w + w_j = 0 \tag{4}$$

式中,e_j 矩阵的第 j 个元素为 1,其他元素均为 0. 根据 Lagrange 方法可知,神经网络的修剪就是求函数 L 的极小值:

$$L = \frac{1}{2}[\Delta w]^T H \Delta w + \lambda(e_j^T \partial w + w_j) \tag{5}$$

联立方程(3) ~(5),解得:

$$L_j(w) = \frac{w_j^2}{2[H^{-1}]_{jj}} \tag{6}$$

式中,H^{-1}为 Hessian 矩阵的逆矩阵,而 $[H^{-1}]_{jj}$为其第 jj 个元素,$L_j(w)$ 表示删掉第 j 个权重引起目标函数的增加值. 对神经网络的修剪过程如下:①训练神经网络,使之达到局部最优;②利用 OBS 方法,得到各个权重重要性的排序;③尝试修剪掉重要性最低的那个权重;④重新训练该神经网络,如果简化后的神经网络性能与原神经网络不相上下,则保留该神经网络结构,然后回到第 2 步重新修剪;如果神经网络性能明显下降,则停止修剪.

3.3　模型拟合评价参数

为了评价模型的拟合效果,本研究采用了以下评价参数.

平均偏差:

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)$$

均方根误差:

$$\mathrm{RESE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$$

相关系数:

$$R^2 = \frac{\left[n\sum_{i=1}^{n}(O_iP_i) - \sum_{i=1}^{n}O_i\sum_{i=1}^{n}P_i\right]^2}{\left[n\sum_{i=1}^{n}O_i^2 - (\sum_{i=1}^{n}O_i)^2\right]\left[n\sum_{i=1}^{n}P_i^2 - (\sum_{i=1}^{n}P_i)^2\right]}$$

准确系数:

$$d = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(|P_i - \overline{P}| + |O_i - \overline{O}|)^2}$$

准确预报系数:

$$\mathrm{SITE} = \frac{n_1}{n_1 + n_2}$$

错误预报率:

$$\mathrm{ERTE} = \frac{n_3}{n_3 + n_4}$$

式中,O_i 为观测值,P_i 为预测值,n_1、n_2、n_3 和 n_4 分别表示准确预报的重污染天数,未能准确预报的重污染天数,将非重污染天预报成重污染的天数,准确预报的非重污染天数.

4　过程与结果

4.1　神经网络预测模型

与多元线性回归模型相比,神经网络模型拟合参数相对较多,修剪掉神经网络模型中不重要的权重以提高模型的泛化能力非常必要. 本研究每次随机抽取 70% 的数据用作训练数据集,剩下的 30% 用作验证数据集,修剪方法为 OBS 方法. 将该过程反复进行 20 次,综合判断各个权重的重要性,保留重要权重.

研究中还发现,修剪掉隐含层单元-输出层单元的权重会导致 O_3 峰值的模拟效果明显变差. 因此在

修剪过程中，跳过对隐含层单元-输出层单元的修剪，从而保证了在降低神经网络复杂程度的同时，也提高了峰值 O_3 的模拟效果. 经过反复训练和比较后，神经网络模型采用 1 个隐含层 5 个隐含单元的结构. 经过修剪后，保留了 51 个权重中的 21 个，删掉了 58.8% 的权重. 神经网络的输入参数包括第 2d 最高温度、最低温度、最高湿度、最低湿度、中午平均光照、10:00 ~ 16:00 的平均风速以及前 1 d 的 O_3 峰值浓度.

修剪过程只改变神经网络的结构，而不保留任何权重值. 为了检验模型效果，笔者按照时间序列，每次从数据集中抽取 5 d 作为检验(test)数据集，剩下的 348 d 作为训练(train)数据集. 采用线性模型和修剪后的神经网络模型进行拟合，运行后比较 2 种模型的拟合结果如表 2 和表 3 所示.

比较 2 种模型的 RMSE、R^2 和 d 发现，神经网络模型拟合效果明显好于线性模型. 神经网络模型的重污染预报率为 77.0%，比线性模型的 60.7% 高出了 16.3%. 神经网络模型全年拟合结果对比如图 1 所示.

表 2 线性模型和神经网络模型拟合结果对比

Table 2 Comparsion of the fitting results between linear regression and artificial neural networks

模型	MBE/mg·m^{-3}	RMSE/mg·m^{-3}	R^2	d
线性模型	0.001 3	0.047 29	0.676	0.892
神经网络模型	0.000 57	0.042 86	0.737	0.923

表 3 线性模型和神经网络模型 O_3 重污染拟合结果对比

Table 3 Comparsion of threshold exceedance between linear regression and artificial neural networks

模型	n_1	n_2	n_3	n_0	SITE/%	ERTE/%
线性模型	37	24	10	282	60.7	3.42
神经网络模型	47	14	12	280	77.0	4.11

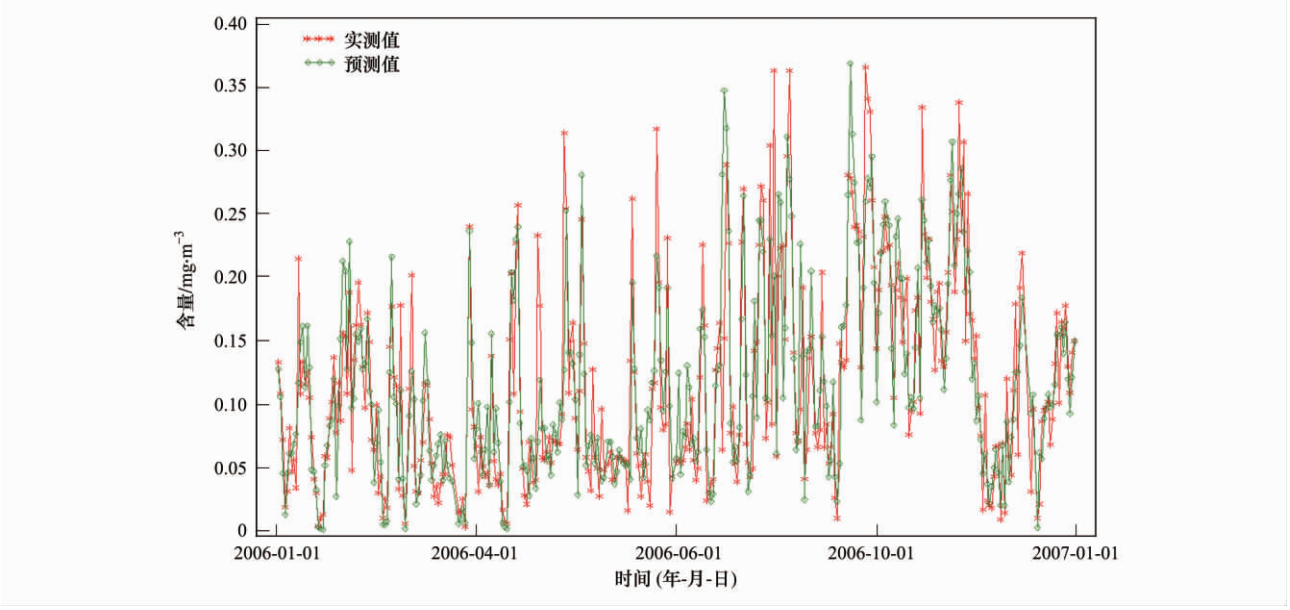

图 1 神经网络预测值与实际值的时间序列对比

Fig. 1 Fitting results between linear regression and artificial neural networks

4.2 神经网络分类器预测臭氧污染级别

通过神经网络分类器，笔者能够实现对 O_3 峰值浓度污染级别的预测，并能给出落入不同污染级别的概率，这样更加有助于普通公众以正确的态度去理解预测的 O_3 污染程度. 为了实现该过程，笔者需要将目标值由具体的 O_3 浓度(数值变量)转换为污染级别(分组变量). 本次模拟采用的分类标准如表 4 所示.

表 4 神经网络分类器 O_3 浓度分类标准

Table 4 Ozone pollution categories used in the neural classifier

污染级别	Ⅰ	Ⅱ	Ⅲ
O_3/mg·m^{-3}	0 ~ 0.10	0.10 ~ 0.20	> 0.20
数量/d	186	106	61

神经网络采用一个隐含层 4 个隐含层单元的结构，每次从数据集中随机抽取 70% 的数据用作 train 数据集，剩下的 30% 用作 valid 数据集，修剪方法为 OBS(optimal brain surgeon). 将该过程反复进行 20 次，综合判断各个权重的重要性，保留重要权重. 经过修剪后的权重保留了 46 个权重中的 24 个，修剪掉 47.8% 的权重. 按照时间序列，每次从数据集中抽取 5 d 作为 test 数据集，剩下的 348 d 作为 train

数据集,用修剪后的神经网络模型进行拟合.拟合后的神经网络分类准确率为为 79.6%;其中重污染天 61 d,预报出 51 d,准确率为 83.6%;非重污染天错误预报 16 d;错误预报率为 5.5%.

5　结论

本研究探索了通过统计分析预测 O_3 浓度的方法,采用了神经网络模型和多元线性回归模型.结果显示,神经网络在重污染预报方面性能远远优过多元线性回归模型,其中通过 OBS 方法修剪神经网络在提高神经网络性能中起到了重要作用.神经网络分类器进一步提高了重污染预测的准确率,并为将该方法运用到实际 O_3 污染级别预测中奠定了基础.

致谢: 本研究所需的数据由广东省环境监测总站提供,在此表示感谢.

参考文献:

[1] Freijer J I, Van Eijkeren J C H, Van Bree L. A model for the effect on health of repeated exposure to ozone [J]. Environmental Modelling and Software, 2002, **17**(6): 553–562.

[2] Abdul-Wahab S A, Al-Alawi S M. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks [J]. Environmental Modelling & Software, 2002, **17**(3): 219–228.

[3] Lefohn A S, Foley J K. Establishing relevant ozone standards to protect vegetation and human health: expose/dose response considerations [J]. Air and Waste, 1993, **43**(1): 106–112.

[4] Schlinka U, Dorlingb S, Pelikan E, *et al*. A rigorous inter-comparison of ground-level ozone predictions [J]. Atmospheric Environment, 2003, **37**(23): 3237–3253.

[5] Hoek G, Schwartz J D, Groot B, *et al*. Effects of ambient particulate matter and ozone on daily mortality in Rotterdam [J]. Archives of Environmental Health, 1997, **52**(6): 455–463.

[6] Al-Alawi S M, Abdul-Wahab S A, Bakheit C S. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone [J]. Environmental Modelling & Software, 2008, **23**(4): 396–403.

[7] Lengyel A, Heberger K, Paksy L, *et al*. Prediction of ozone concentration in ambient air using multivariate methods [J]. Chemosphere, 2004, **57**(8): 889–896.

[8] Elkamel A, Abdul-Wahab S, Bouhamra W, *et al*. Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach [J]. Advances in Environmental Research, 2001, **5**(1): 47–59.

[9] Krupa S, Nosal M, Ferdinand J A, *et al*. A multi-variate statistical model integrating passive sampler and meteorology data to predict the frequency distributions of hourly ambient ozone (O_3) concentrations [J]. Environmental Pollution, 2003, **124**(1): 173–178.

[10] Barrero M A, Grimalt J O, Canton L. Prediction of daily ozone concentration maxima in the urban atmosphere [J]. Chemometrics and Intelligent Laboratory Systems, 2006, **80**(1): 67–76.

[11] Abdul-Wahab S, Bouhamra W, Ettouney H, *et al*. A statistical model for predicting ozone levels in the Shuaiba Industrial Area in Kuwait [J]. Environmental Science and Pollution Research, 1996, **3**(4): 195–204.

[12] Dutot A L, Rynkiewicz J, Steiner F E, *et al*. A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions [J]. Environmental Modelling & Software, 2007, **22**(9): 1261–1269.

[13] Davis J M, Speckman P. A model for predicting maximum and 8 h average ozone in Houston [J]. Atmospheric Environment, 1999, **33**(16): 2487–2500.

[14] Ballester E B, Valls G C I, Carrasco-Rodriguez J L, *et al*. Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks [J]. Ecological Modelling, 2002, **156**(1): 27–41.

[15] Lin Y Q, Cobourn W G. Fuzzy system models combined with nonlinear regression for daily ground-level ozone predictions [J]. Atmospheric Environment, 2007, **41**(16): 3502–3513.

[16] Sousa S I V, Martins F G, Pereira M C, *et al*. Prediction of ozone concentrations in Oporto city with statistical approaches [J]. Chemosphere, 2006, **64**(7): 1141–1149.

[17] Wang W J, Lu W Z, Wang X K, *et al*. Prediction of maximum daily ozone level using combined neural network and statistical characteristics [J]. Environment International, 2003, **29**(5): 555–562.

[18] Jain A K, Mao J C, Mohiuddin K M. Artificial neural networks—a tutorial [J]. Computer, 1996, **29**(3): 31–44.

[19] Hassibi B, Stork D G, Wolff G J. Optimal brain surgeon and general network pruning [C]. 1993 IEEE International Conference on Neural Networks, Vols 1–3: 293–299.

[20] 张俊妮. 数据挖掘与应用 [M]. 北京: 北京大学出版社, 2009. 115–116.