

大气污染预测中提高 BP 网络泛化能力的方法

秦 侠, 雷 蕾, 姚小丽

(北京工业大学 环境与能源工程学院, 北京 100022)

摘 要: 为了通过预测大气环境的质量和发展变化, 来寻求有效地控制和改善环境质量的相应措施, 选用英国伦敦马里波恩监测站 $PM_{2.5}$ 的小时平均浓度监测资料, 采用贝叶斯归一化训练算法和提前终止法泛化改进的 BP 神经网络模型, 预报 $PM_{2.5}$ 的 24 h 内的各小时浓度. 结果表明, 采用本方法进行空气污染预报, 预测相对误差为 20%~49%, 提高了预报网络的泛化能力.

关键词: BP 网络; 泛化能力; $PM_{2.5}$; 污染预测

中图分类号: X 32

文献标识码: A

文章编号: 0254-0037(2007)08-0849-03

BP 网络(back-propagation network)为反向传播网络, 已经在许多学科得到广泛应用. 虽然国内外已有将神经网络法应用于大气污染预报的研究^[1-5], 但在目前的研究中, 比较多见的是根据影响大气污染物浓度的气象因素, 使用传统的 BP 神经网络建立预报模型^[6-7], 通常具有收敛速度慢、易陷入局部极小值等缺陷, 并且国内所用模型大都没有通过泛化改进^[7], 对于未通过泛化改进的网络模型, 当网络训练误差很小时, 如果将其应用于 1 个新的样本数据, 网络的训练误差会迅速加大, 从而使预报误差加大. 为了克服这些缺陷, 本文以 BP 网络预报 $PM_{2.5}$ 的 24 h 浓度为例, 采用贝叶斯归一化训练算法, 并在训练过程中加入 1 个验证集以寻找最佳训练次数, 当达到最佳训练次数后提前终止训练, 防止网络被过度训练.

1 $PM_{2.5}$ 大气污染预报模型的建立

本实例采用英国伦敦 Marylebone Road 监测站 2004 年和 2005-04-01—2005-07-31 的 $PM_{2.5}$ 的小时平均浓度监测值(共 $2 \times 122 \times 24$ 个数据点)进行研究. 用 2004 年数据训练网络, 训练好的网络用于 2005-04-01—2005-07-31 的 $PM_{2.5}$ 的第 2 天 24 h 各小时平均浓度预测.

1.1 输入向量和目标向量

为了使网络训练一开始就给各输入分量以同等重要的地位, 防止因净输入的绝对值过大而使神经元输出饱和, 以便于网络训练, 防止计算过程出现“过拟合”等问题, 本研究采用归一化方法对数据进行尺度变换, 归一化公式为

$$X'_k = \frac{x_k - x_{\min}}{x_{\max} - x_{\min}}$$

根据公式可将数据“归一化”, 即限定在 $[0, 1]$.

对于每一年, 都建立了 2 个矩阵: 输入向量矩阵 P 和目标向量矩阵 T . 每个矩阵都有 24 行 121 列, 其中第 i 行($i=1, 2, \dots, 24$)代表第 i 小时; 第 j 列($j=1, 2, \dots, 121$)代表 4 个月中的第 j 天; 目标向量矩阵的第 j 列为输入向量矩阵的第 $(j+1)$ 列, 表示目标值为第 2 天该小时 $PM_{2.5}$ 的平均浓度.

1.2 BP 网络的创建

本研究采用 3 层 BP 网络, 选取输入层节点数为 24(分别为一天内 24 h 的各小时 $PM_{2.5}$ 平均浓度); 输

收稿日期: 2006-08-24.

作者简介: 秦 侠(1969-), 女, 山西太原人, 副教授.

出层节点数为 1(为第 2 天该小时 $\text{PM}_{2.5}$ 的小时平均浓度);隐层节点数为 4;训练精度设置为 0.001,最大训练次数设置为 5 000.

对于学习算法,选用贝叶斯归一化学习算法.因为它可以修正典型前馈网络的误差性能函数,即在均方差的基础上,增加包含网络权值和阈值的均方值一项,并以自动的方式决定最佳误差性能调整率,再由误差性能调整率来分配均方差与该项的比例,这种修正的误差性能函数可使网络获得较小的权值和阈值,提高网络的泛化能力.

2 网络的训练及预报仿真结果

由于 Matlab 神经网络工具箱(neural network toolbox)提供了各种神经网络激活函数、训练函数及各种网络集成块等,通过对激活函数、网络函数等的调用,仅需写很少的源代码,即可完成必须的科学计算,所以本研究在 Matlab 环境下编写程序.以 2004 年矩阵 P 为输入向量,分别以 T 的第 i 行($i=1,2,\dots,24$)为目标向量进行训练,建立 $\text{PM}_{2.5}$ 24 h 的各小时浓度预报网络,用训练好的网络预测 2005 年输入向量,然后仿真输出结果,最后,绘制网络输出曲线,并与期望输出值曲线相比较,如图 1 所示.为量化 BP 网络的预测性能,将浓度的预测值与待预测目标值进行比较,计算 $\text{PM}_{2.5}$ 的 24 h 浓度平均相对误差,结果见图 2.

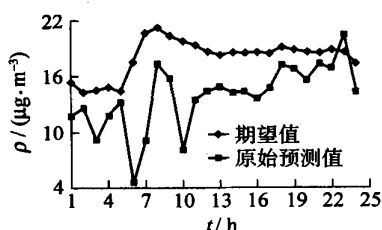


图 1 网络输出与期望输出曲线

Fig.1 Comparison of prediction value and expected value

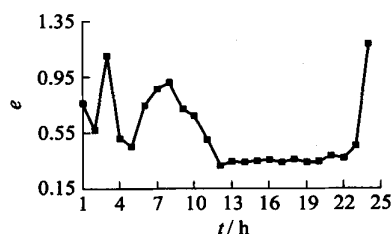


图 2 预测误差曲线

Fig.2 Prediction errors

图 1、图 2 表明,网络输出值和期望输出值之间的相对误差在 30%~118% 波动,结果较差,没有达到预测目的.虽然网络对训练样本的误差很小,但对新的待预测输入样本误差很大,这说明网络已经被训练得过度吻合,泛化能力很差,且网络收敛速度较慢,24 个网络全部收敛用时 319 s.

3 网络泛化能力的改进

从第 2 节的分析中可以看出, $\text{PM}_{2.5}$ 预报网络的建立方法有待改进.有研究表明^[8],在隐节点数一定的情况下,存在着 1 个最佳训练次数,在达到该训练次数之前,随着训练次数的增加,训练集误差和验证集误差同时下降;当超过这个训练次数,训练集误差继续减小而验证集误差开始上升,在此之前,若停止训练则训练不足;在此之后,若仍继续训练,则会训练过度,因此该训练次数为最佳训练次数.从第 2 节的分析中可以看出,预报网络对 2005 年的数据预测精度不高,原因在于网络被过度训练导致泛化能力较差.因此,找到这个最佳训练次数,达到后就停止训练网络,即达到最佳训练次数后使正常的训练提前停止,才能保证网络对新的输入样本具有较好的泛化能力.

作者将 2004 年原始数据样本集按 3:1 随机分成训练集(training)和验证集(validation).其中,训练集数据用来对神经网络进行训练,作初步的模型设定;验证集数据用来监控和调试模型权重,同时也用作模型评估.在样本集划分时,保证选择的随机性,是为了使分开的 2 部分数据的性质一致.在训练初始阶段,验证集形成的验证误差通常会随着网络训练误差的减小而减小;但是,当网络开始进入过度训练时,验证误差就会增大,当验证误差增大到一定程度时,即达到最佳训练次数时,网络训练就会提前停止,这时,训练函数会返回当验证误差取最小值时的网络对象.

将用提前终止法训练好的网络应用于 2005 年输入向量、仿真输出结果和绘制网络输出曲线, 并与期望输出值及未加验证集过度训练的网络输出值曲线相比较, 如图 3 所示, 计算得 $PM_{2.5}$ 的 24 h 浓度平均相对误差见图 4。

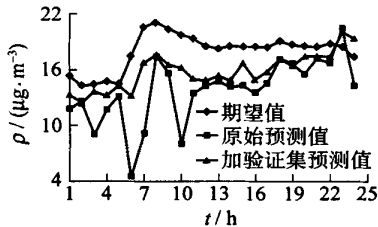


图 3 网络输出曲线对比

Fig. 3 Comparisons of prediction value

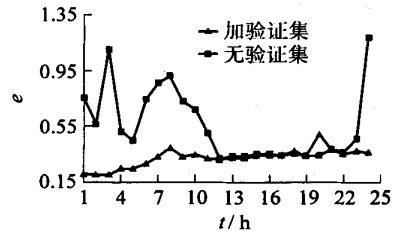


图 4 浓度相对误差

Fig. 4 Concentration relative errors

从图 3、图 4 可以看出, 加入验证集、找到最佳训练次数, 即提前终止训练后, 新的网络输出得到的曲线和期望输出值曲线很接近, 相对误差在 20%~49% 波动, 结果较好。24 个网络全部收敛用时 62 s, 收敛速度大大加快。因此, 用提前终止法改进网络是成功的, 用此法训练好的网络可用于实际大气污染预测。

4 结论

本文采用贝叶斯归一化训练算法训练网络, 并在训练过程中加入 1 个验证集以求找到最佳训练次数, 当达到最佳训练次数后, 提前终止训练, 防止网络被过度训练, 训练好的网络用于 $PM_{2.5}$ 的小时平均浓度预测, 预测相对误差为 20%~49%。这就解决了未改进的网络训练误差很小时, 1 个新的输入使对应的目标输出误差较大的问题, 即提高了 BP 网络的泛化能力。用此法训练好的网络可用于实际大气污染预测。

参考文献:

- [1] 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): 49-52.
MA Yan-jun, YANG Hong-bin, ZHANG Yun-hai. Study on prediction of atmosphere pollution concentration based on BP model[J]. Meteorological Monthly, 2003, 29(7): 49-52. (in Chinese)
- [2] 王灿星, 祁国伟, 何曦, 等. BP 神经网络用于大气中颗粒物(TSP)预测的研究[J]. 仪器仪表学报, 2003, 24(4): 530-539.
WANG Can-xing, QI Guo-wei, HE Xi, et al. Research on application of BP neural network to predict TSP pollution[J]. Chinese Journal of Scientific Instrument, 2003, 24(4): 530-539. (in Chinese)
- [3] 李少华, 董向元, 于静梅. 基于神经网络的大气污染预报方法的研究[J]. 东北电力学院学报, 2003, 23(4): 1-4.
LI Shao-hua, DONG Xiang-yuan, YU Jing-mei. Study on prediction of air pollution based on artificial neural network[J]. Journal of Northeast China Institute of Electric Power Engineering (Natural Science Edition), 2003, 23(4): 1-4. (in Chinese)
- [4] PÉREZ P, TRIER A, REYES J. Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile[J]. Atmospheric Environment, 2000, 34: 1189-1196.
- [5] HEO J S, KIM D S. A new method of ozone forecasting using fuzzy expert and neural network systems[J]. Science of the Total Environment, 2004, 325: 221-237.
- [6] 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.
WANG Jian, HU Xiao-min, ZHENG Long-xi, et al. Study on forecasting of air pollution based on BP mode[J]. Research of Environmental Sciences, 2002, 15(5): 62-64. (in Chinese)
- [7] 黄世芹. 改进神经网络在城市环境大气污染分季节预报中的应用[J]. 贵州气象, 2005, 25(3): 6-8.
HUANG Shi-qin. The application of improved BP neural networks in seasonal air pollution forecasting of urban area[J]. Journal of Guizhou Meteorology, 2005, 25(3): 6-8. (in Chinese)
- [8] 韩力群. 人工神经网络理论、设计及应用[M]. 北京: 化学工业出版社, 2004.

Methods to Improve the Generalization of BP Neural Network Applied in Air Pollution Forecasting

QIN Xia, LEI Lei, YAO Xiao-li

(College of Environmental and Energy Engineering, Beijing University of Technology, Beijing 100022, China)

Abstract: Atmospheric pollution prediction is helpful to find effective ways to control air pollution and improve air quality. BP neural network which was trained using Bayesian Regularization method and early stopping method was used to forecast the hourly concentration of $PM_{2.5}$. The data of $PM_{2.5}$ hourly concentration were obtained from monitoring site of Marylebone Road in London, UK. The prediction relative error goes from 20% to 49%. The results show that compared with networks which were trained using other methods, Bayesian Regularization method and early stopping method can improve the generalization ability of BP neural network.

Key words: back propagation network; generalisation; $PM_{2.5}$; pollution prediction

(上接第 833 页)

Two-Phase Sampling Method of Resident Trip

YANG Jun, LIU Xiao-ming, ZHANG De-xin, CHEN Jin-chuan

(Key Laboratory of Transportation Engineering, Beijing University of Technology, Beijing 100022, China)

Abstract: To solve problems in the sampling process of resident trip investigation, this paper applied mathematical statistics to studying the characteristics of stratified sampling method. A two-phase stratified sampling method and stratified principles have been discussed. The result shows that this method can save more manpower resources while keeping high investigation precision. It is suggested that two phase sampling method should be adopted in metropolitan area.

Key words: resident trip investigation; two phase sampling methods; stratified sampling method; sampling method