DOI:10.3969/j.issn.1674-6732.2019.02.004

基于自回归 GA－BP 神经网络的 AQI 预测

方正,张磊*,王玉琴,黄雅琨,徐静
(徐州工程学院机电工程学院,江苏　徐州　221018)

摘　要:基于徐州市 2013 年 12 月—2018 年 11 月的空气质量指数日均值,建立了时间序列自回归输入的 GA－BP 神经网络模型用于空气质量指数预测。结果表明,所建立的网络模型能够准确预测徐州市空气质量指数的变化趋势,其中夏季预测相对误差 18.23%,仿真均方根误差(RMSE)为 14.59;冬季预测相对误差 9.14%,仿真 RMSE 为 11.47。

关键词:自回归输入;GA－BP 神经网络;空气质量指数
中图分类号:TP183;X831　　**文献标志码**:B　　**文章编号**:1674－6732(2019)02－0022－04

An Autoregressive GA-BP Neural Network-based AQI Prediction

FANG Zheng, ZHANG Lei*, WANG Yu-qin, HUANG Ya-kun, XU Jing
(*School of Mechanical & Electrical Engineering, Xuzhou University of Technology, Xuzhou, Jiangsu* 221018, *China*)

Abstract: Based on the daily average of air quality index in Xuzhou from December 2013 to November 2018, this paper established a GA-BP neural network model with autoregressive input of time series for air quality index prediction. The experimental results showed that the established network model could accurately predict the change trend of Xuzhou air quality index. The relative error of summer forecast was 18.23%, the simulation RMSE was 14.59, the winter forecast relative error was 9.14%, and the simulation RMSE was 11.47.

Key words: Autoregressive inputs; GA-BP neural network; Air quality index

近年来,随着全球能量消耗,污染物排放的加剧,人们对空气质量指数(Air Quality Index, AQI)的关注程度与日俱增,一方面关注影响 AQI 的污染源,减少污染排放;另一方面研究对 AQI 的预测,以积极做好健康防护。

在 AQI 预测方面,多种先进智能算法被用于谋求精准预测。高帅等[1]提出将支持向量机(SVM)和飞蛾扑火优化(MFO)相结合的算法(MFO－SVM),张梦瑶等[2]基于改进加权马尔科夫链,李博群等[3]通过建立模糊时间序列模型,对一些地区的 AQI 进行实证分析和检验,获得了较为理想的预测结果。此外,基于小波分析[4]、主成分分析[5]、时间序列分析[6]的方法也被用于 AQI 预测。在诸多智能算法中,神经网络作为一个高度复杂的非线性动力学系统,也经常被用于 AQI 预测。例如文献[7－10]所建立的 AQI 预报模型都是基于神经网络完成的。然而,他们对神经网络输入的处理大多基于 AQI 的主要污染源,忽略了次要污染源的贡献;而若将 AQI 时间序列本身作为神经网络输入,虽可包含所有污染源数据信息,但同时会降低神经网络的泛化能力和预测精度,因此,基于自回归时间序列输入的神经网络在 AQI 预测方面的研究还不多见。

现使用 2013—2018 年徐州地区的 AQI 时间序列,引入自回归输入模式,采用遗传算法(Genetic Algorithm, GA)对神经网络初始权值进行最优化筛选,利用 BP 算法对网络参数反向微调,并对 AQI 的预测结果进行了验证。

1　数据预处理及网络输入端构建

1.1　数据预处理

从江苏省气象局网站收集整理徐州地区 2013 年 12 月—2018 年 11 月共 1 814 个 AQI 作为原始

收稿日期:2018－11－02;**修订日期**:2019－02－28
基金项目:大学生实践创新训练计划基金资助项目(xcx2018107)
作者简介:方正(1997—),男,本科在读。
* **通讯作者**:张磊　E－mail:triple－stone@foxmail.com

数据。原始 AQI 数据检验观测图见图 1。

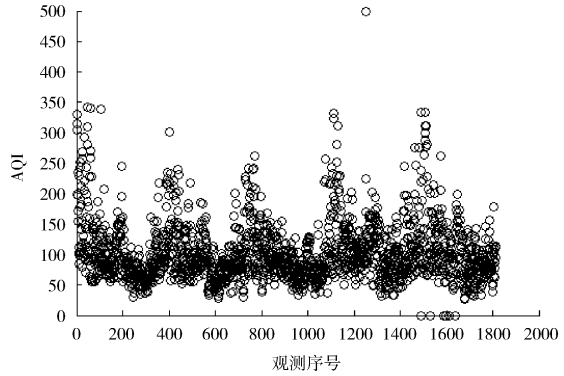

图 1　原始 AQI 数据检验观测图

由图 1 可见,时间序列存在 2 类“奇异”点:一类为脱离主数据群,距离主数据群较远点,如 AQI 值为 500 的一点,距离主数据群最大值直线位移超过 150;另一类是 AQI 值为“0”的点,该类点意味着空气质量出奇得好,没有一点污染。对于第一类奇异点,原数据组仅有 1 个数据,从数据预测的经验判断,该点破坏了原数据组的变化规律,会降低预测模型精度,故予以删除;对于第二类奇异点,尽管该类点距离主数据群很近,如图 1 所示数据序列号为 1600 附近的点,但基于生活经验的常识,将“AQI 为 0”视为不可能事件,故将原数据组的此类 14 个点也予以删除。为保证原数据组的时序完整性,所删除的 15 个点采用 cubic 插值补齐。

插值补齐后数据组数据范围为[26,343],极差为 317,仍较大,若直接作为神经网络的输入,容易造成系统震荡而出现数据欠拟合、无法收敛、预测精度降低、迭代过程中系统震荡过大导致不稳定等结果。采用神经网络常规数据处理方法,将原数据序列归一化到[– 1,1],后发现数据由于过于密集而无法反映原数据组的变化趋势,故不考虑神经网络对数据输入处理的常规归一化方法。

建立神经网络的输入为一维时间序列,量纲统一,不受常规神经网络因多维数据输入的量纲不统一,必须进行归一的限制。因此,考虑将原数据序列映射到某一区间$[x_1,x_2]$上,该区间$[x_1,x_2]$应达到使所建立神经网络的预测值与真实值的相对误差最小。取网络预测的均方误差(Mean Square Error, MSE)为目标函数,建立对最优区间搜索的表达式:

$$\min_{[x_1,x_2]} R = \mathrm{MSE}(Y) \tag{1}$$

式中:R——区间$[x_1,x_2]$内的均方根误差值,Y——所建立的神经网络。

上式的求解问题是个解空间搜索区域极大的最优极值求解问题,可采用粒子群(PSO)、蚁群等多种仿生智能算法求解。现讨论的核心是 GA – BP 神经网络的 AQI 预测性能,并非最优极值求解,对上式的求解问题做如下简化:

(1) 考虑计算成本,固定 $x_1 = 0$,将$[x_1,x_2]$双边求解简化为$[0,x_2]$区间的单边求解,其中 $0 < x_2 < 343$;

(2) 基于降低所要建立神经网络的复杂性,Y使用一般 BP 神经网络。

基于上述简化,为确定 x_2的较优解,将使用一般神经网络在 MATLAB 中编程求解。首先需要确定神经网络一维时间序列的输入端节点数。

1.2　自回归网络输入端构建

由于 AQI 数据为时间序列,不符合神经网络监督式训练模式,引入自回归模型框架,在输入端对原数据序列进行转换,以符合网络结构需求。用样本总体计算不同自回归阶数所对应的网络均方误差及测试样本相关度检验值见表 1。

表 1　控制变量情况下不同阶数网络测试结果

自回归阶数	MSE	预测相关度检验值
2	0.079 7	0.609 7
3	0.077 8	0.559 7
4	0.077 8	0.606 5
5	0.075 4	0.631 9
6	0.074 9	0.517 2
7	0.076 4	0.633 0
8	0.080 7	0.680 7
9	0.073 9	0.571 0
10	0.070 1	0.696 6
11	0.076 1	0.688 1
12	0.071 8	0.600 9
13	0.077 6	0.563 4
14	0.070 4	0.589 7
15	0.069 3	0.581 2

由表 1 可见,选择不同自回归阶数的 MSE 基本稳定在 0.07 左右,但相关度检验值存在较大差异。当自回归阶数为 10 时,MSE 比阶数 15 时高 0.000 8,但阶数 15 对于测试数据的相关度检验明显低于阶数 10,选择自回归阶数为 10。即以连续 10 d 的 AQI 数值为输入端,以第 11 d 的 AQI 作为标签数据,以此类推确立网络样本数据共 1 804

组。使用上述数据,建立输入端为 10 的一般 BP 神经网络,在 MATLAB 中编程,使用(1)式作为目标函数,搜索较优的 x_2 值为 224,故确定原时间序列的输入映射区间为[0,224]。图 2 为插值并映射到[0,224]后的数据序列。

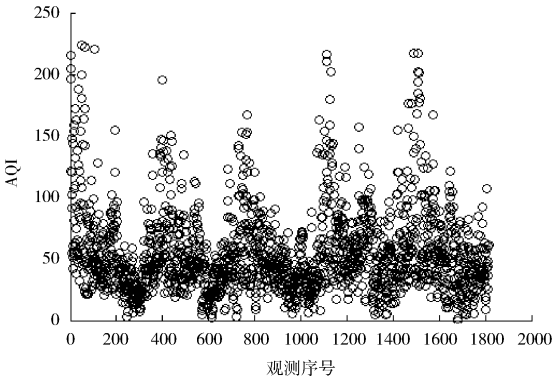

图 2　插值并映射后的 AQI 数据

2　GA－BP 神经网络预测模型

(1) 由上文可知设定输入端节点为 10;隐含层节点数目通过试错法选定,网络迭代开始前初始化各连接赋权 ω_{ij}^1、ω_{ij}^2 及阈值 θ,其中 θ 赋予[－0.2,0.2]的随机值,$i=1,2\cdots\cdots10$;$j=1,2\cdots\cdots1\ 804$。

(2) 确定隐含层及输出层的传递函数为 tansig 函数并归一化样本到范围[－1,1]。

(3) 将各隐含层权值进行遗传编码计算适应度并进行选择、交叉、变异操作,当种群适应度到达目标值时保留此时网络隐含层权值,设置其适应度函数与隐含层输出为:

$$fun = 1/\sum_{i=1}^{n} \mathrm{abs}(d_i - \xi_i) \tag{2}$$

$$\xi_j = \mathrm{tansig}(\sum_{i=1}^{m} \omega_{ij}^1 x_{ij} - \theta_i) \tag{3}$$

式中:fun——种群适应度;d_i——第 i 个网络输出层节点的期望输出;ξ_i——第 i 个网络输出层节点的仿真输出;ξ_j——第 j 个隐含层节点输出;ω_{ij}^1——第 i 个输入层节点到第 j 个隐含层节点的权值;x_{ij}——第 i 个输入端节点到第 j 个隐含层节点的样本值;θ_i——第 i 个输入层节点阈值。

(4) 计算 $\sigma_i = (d_i - \xi_i)\xi_i(1-\xi_i)$。　　(4)

式中:σ_i——第 i 个输出层节点校正误差;d_i——第 i 个网络输出层节点的期望输出;ξ_i——第 i 个网络输出层节点的仿真输出。

(5) 利用 ω_{ij}^2,ξ_i,ξ_j 计算隐含层的校正误差:

$$\sigma_{ij} = \xi_i(1-\xi_i)\xi_j\omega_{ij}^2。 \tag{5}$$

式中:σ_{ij}——第 i 个输入节点到第 j 个隐含层节点的修正误差;ξ_i——第 i 个网络输出层节点的仿真输出;ξ_j——第 j 个隐含层节点输出;ω_{ij}^2——第 i 个隐含层节点到第 j 个输出层节点的权值。

(6) 若其误差收敛于规定精度 ε 时进入停止学习,反之利用 ω_{ij}^2,ξ_i,ξ_j 和 θ 计算下一次学习中隐含层和输出层之间新的连接权值和神经元阈值,并对各个参数进行反向微调:

$$\theta(t+1) = \theta(t) + \gamma\{\eta(t)\sigma_i + \alpha[\theta(t) - \theta(t-1)]\},\eta(t) = \eta_0[1 - t/(T+M)] \tag{6}$$

$$\omega_{ij}^2(t+1) = \omega_{ij}^2(t) + \gamma\{\eta(t)\sigma_i\xi_j + \alpha[\omega_{ij}^2(t) - j(t-1)]\} \tag{7}$$

$$\omega_{ij}^1(t+1) = \omega_{ij}^1(t) + \gamma\{\eta(t)\sigma_{ij}\xi_i + \alpha[\omega_{ij}^1(t) - j(t-1)]\} \tag{8}$$

式中:$\theta(t+1)$——第 $t+1$ 次更新后的阈值;$\theta(t)$——第 t 次更新后的阈值;γ——学习率;$\eta(t)$——第 t 次更新后的梯度;σ_i——第 i 个输出层节点校正误差;α——动量系数,且 $\alpha \in (0,1)$;$\theta(t-1)$——第 $t-1$ 次更新后的阈值;η_0——初始梯度;t——学习次数;$j=1,2\cdots\cdots1\ 804$;T——总迭代次数;M——为防止分母为 0 的整数;$\omega_{ij}^2(t+1)$——第 $t+1$ 次更新后第 i 个隐含层节点到第 j 个输出层节点的权值;$\omega_{ij}^2(t)$——第 t 次更新后第 i 个隐含层节点到第 j 个输出端的权值;$\omega_{ij}^1(t+1)$——第 $t+1$ 次更新后第 i 个输入层节点到第 j 个隐含层节点的权值;$\omega_{ij}^1(t)$——第 t 次更新后第 i 个输入端到第 j 个隐含层节点的权值;σ_{ij}——第 i 个输入节点到第 j 个隐含层节点的修正误差;ξ_i——第 i 个网络输出层节点的仿真输出;ξ_j——第 j 个隐含层节点输出。

(7) 停止学习并保存训练好的各网络参数。算法整体流程见图 3。图中 k 为进化代数,K 为预设进化代数,ε 为网络误差,σ 为网络期望误差。

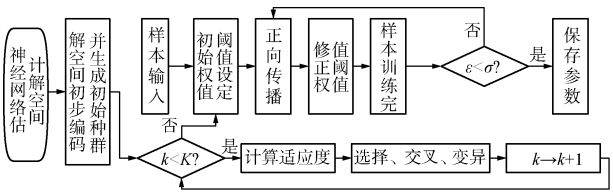

图 3　GA－BP 算法流程

3　仿真实验

为保证对网络泛化能力的测试,对于夏季,截取 2018 年 7 月前后共 50 d 样本作为测试集,训练样本总数 1 637 组;对于冬季,截取 2018 年 11 月底之前共 50 d 样本作为测试集,冬季测试时样本总数为 1 754 组。考虑到夏、冬季 AQI 所表现的范围并不相同,分别对夏、冬季进行连续的长期预测以展示所建立 GA – BP 神经网络模型的性能。

经过调试,所确定的 GA – BP 预测模型参数如下:输入端节点 10,隐含层节点 30,输出端节点 1,种群为 30,种群进化代数 10,交叉概率 0.4,变异概率 0.1,学习率 0.1,动量参数 0.9,最大迭代次数 100,在 MATLAB 环境下仿真输出的夏季和冬季的预测结果见图 4(a)(b)(RMSE 为均方根误差)。

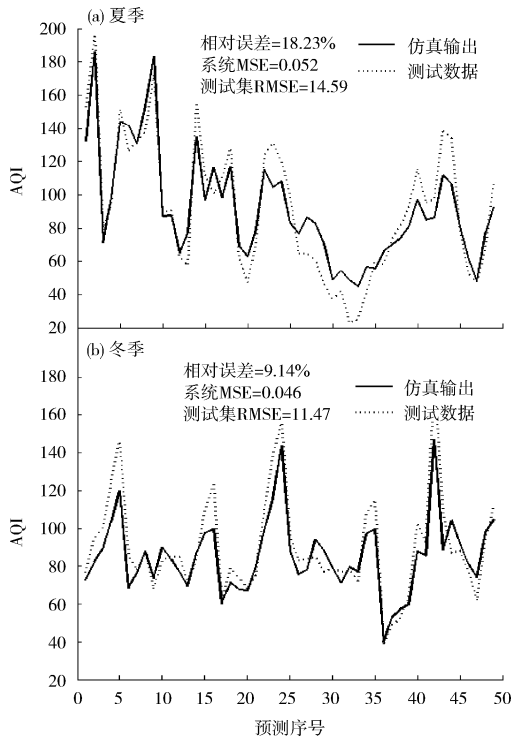

图 4　夏、冬季连续预测对比

由图 4 可见,计算夏、冬季 2 个预测系统所对应的 MSE 分别为 0.052 和 0.046,均接近于 0;其中夏季预测相对误差 18.23%,仿真 RMSE 为 14.59;冬季相对误差 9.14%,仿真 RMSE 为 11.47,故认为训练效果良好,2 个系统的 RMSE 相差不大,但可看出冬季相对于夏季其仿真结果要更加稳定一些;2 个系统虽在趋势追踪上效果良好,但对局部极大或极小值的追踪上效果相对要差一些。

4　结论

建立的网络模型能够准确预测徐州市空气质量指数的变化趋势,其中夏季预测相对误差 18.23%,仿真 RMSE 为 14.59;冬季预测相对误差 9.14%,仿真 RMSE 为 11.47。

[参考文献]

[1] 高帅, 胡红萍, 李洋,等. 基于 MFO – SVM 的空气质量指数预测[J]. 中北大学学报(自然科学版), 2018,39(4):372 – 379.

[2] 张梦瑶,黄恒君. 基于改进加权马尔科夫链的兰州市空气质量预测[J]. 兰州财经大学学报(自然科学版), 2018,34(3):111 – 117.

[3] 李博群,贾政权,刘利平. 基于模糊时间序列的空气质量指数预测[J]. 华北理工大学学报(自然科学版), 2018,40(3):79 – 86.

[4] 姚清晨,张红. 基于小波分析的太原市空气质量变化特征及预测[J]. 山西大学学报(自然科学版),2019,42(1):265 – 274.

[5] 姜新华, 薛河儒, 张存厚,等. 基于主成分分析的呼和浩特市空气质量影响因素研究[J]. 安全与环境工程, 2016, 23(1):75 – 79.

[6] 于萍. 时间序列分析在空气质量指数(AQI)预测中的应用[D]. 沈阳:辽宁师范大学, 2015.

[7] 夏晓玲, 尚媛媛, 宋丹. 基于 BP 神经网络的贵阳市空气质量指数预报模型[J]. 环境监控与预警, 2018, 10(3): 14 – 17.

[8] 薛兴钊. 基于 BP 神经网络的秦岭北麓中部空气质量预报研究[D]. 西安:西安建筑科技大学, 2014.

[9] 张鹏达. 基于 BP 神经网络的城市环境空气质量预测模型[J]. 自动化技术与应用, 2014, 33(1):9 – 11.

[10] 薛士琼. 基于 BP 神经网络的空气质量预测及可视化的实现[D]. 天津:天津大学, 2016.

栏目编辑　李文峻