

熊兴俊. 基于 BP 神经网络的大气污染物浓度预测模型[J]. 智能计算机与应用, 2024, 14(8): 129–132. DOI: 10.20169/j.issn.2095-2163.240821

基于 BP 神经网络的大气污染物浓度预测模型

熊兴俊

(台山市环境监测站, 广东 台山 529200)

摘要: 针对目前大气污染物浓度预测精度不高的问题, 本文提出了一种大气污染物浓度预测模型。对大气污染物浓度时间序列进行相空间重构, 采用 C-C 法计算延迟时间和嵌入维数, 在此基础上构建了网络结构为 5-11-1 的 BP 神经网络大气污染物浓度预测模型。采用某市 SO₂ 浓度监测数据进行仿真分析, 并将 BP 神经网络的污染物浓度预测模型结果与 SVM 模型和 ELM 模型进行对比, BP 神经网络对测试集的预测结果的均方根误差和平均相对误差分别为 0.298 和 4.35%, 预测精度更高, 验证了本文所提污染物浓度预测模型的正确性和实用性。

关键词: 大气污染物; 浓度预测; BP 神经网络; 相空间重构

中图分类号: TP183

文献标志码: A

文章编号: 2095-2163(2024)08-0129-04

A prediction model for atmospheric pollutant concentration based on BP neural network

XIONG Xingjun

(Taishan Environmental Monitoring Station, Taishan 529200, Guangdong, China)

Abstract: In response to the current problem of low accuracy in predicting the concentration of atmospheric pollutants, this paper proposes a model for predicting the concentration of atmospheric pollutants. A phase space reconstruction was performed on the time series of atmospheric pollutant concentration, and the C-C method was used to calculate the delay time and embedding dimension. Based on this, a BP neural network prediction model for atmospheric pollutant concentration with a network structure of 5-11-1 was constructed. Using SO₂ concentration monitoring data from a certain city for simulation analysis, the results of the pollutant concentration prediction model of the BP neural network were compared with the SVM model and ELM model. The root mean square error and average relative error of the prediction results of the BP neural network on the test set were 0.298 and 4.35%, respectively, with higher prediction accuracy, verifying the correctness and practicality of the pollutant concentration prediction model proposed in this paper.

Key words: atmospheric pollutant; concentration prediction; BP neural network; phase space reconstruction

0 引言

随着工业化进程的不断加快, 能源消耗量逐年上升, 大量二氧化硫、一氧化氮等空气污染物被排放到大气中, 出现严重的大气污染问题^[1-2]。由于大气污染物的排放量不能很好的预测, 导致部分企业减排措施无法很好地落实^[3-4]。因此对大气污染物浓度预测问题进行研究, 建立高精度大气污染物浓度预测模型对于大气污染治理和保护生态环境具有重要意义。

从数学的角度看, 大气污染物浓度预测实际上

是一个回归问题, 因此一些智能算法被应用到污染物浓度预测。文献[5]将联邦学习算法和深度神经网络算法相结合, 建立了大气污染物浓度预测模型, 对 SO₂、CO、O₃、NO₂、PM_{2.5}、PM₁₀ 等 6 种大气污染物进行了预测, 取得了良好的效果; 文献[6]为了提高大气污染物浓度预测精度, 采用互补集成经验模态分解对样本数据进行分解, 得到多个模态分量, 在此基础上采用双向门控循环单元算法建立了大气污染物浓度预测模型, 并采用徐州市相关数据对模型的有效性进行了验证; 文献[7]认为大气污染物排放数据是一个混沌时间序列, 对污染物排放数据进

行了相空间重构,采用改进萤火虫算法(IGSO)对支持向量机(SVM)的惩罚参数和核参数进行了优化,建立了基于IGSO-SVM大气污染物浓度预测模型。现有大气污染物浓度预测模型普遍存在预警精度不高、预测效果不稳定的问题,因此高精度大气污染物浓度预测模型还有待进一步研究。

本文根据大气污染物浓度的混沌时间特性,对大气污染物浓度时间序列进行相空间重构,构建了基于BP神经网络的大气污染物浓度预测模型,采用某市SO₂浓度监测数据进行仿真分析,对所提模型的正确性和有效性进行了验证。

1 相空间重构

1.1 相空间重构理论

20世纪70年代初,Packard研究发现,某个系统在动力学方面的特征可以通过相空间进行分析,提出了相空间重构法^[8],为解决混沌时间序列问题开辟了一条新的途径。

1981年,荷兰数学家Takens在相空间重构理论研究的基础上,提出了Takens定理。Takens定理需满足以下两个定义:

定义1 令系统中有两个度量空间,分别为 (N, ρ) 和 (N_1, ρ_1) ,且满足映射 $\varphi: N \rightarrow N_1$,其中 φ 满映射, $\rho(x, y) = \rho_1(\varphi(x), \varphi(y)) (\forall x, y \in N)$, (N, ρ) 和 (N_1, ρ_1) 为等距同构的两个度量空间。

定义2 当 (N_1, ρ_1) 和 (N_2, ρ_2) 两个度量空间与子空间 (N_0, ρ_2) 等距同构时, (N_1, ρ_1) 可以嵌入 (N_2, ρ_2) 中。

Takens定理:令集合 M 是 d 维空间中的紧流形, $\varphi: M \rightarrow M_1$, φ 为光滑的微分同胚, $y: M \rightarrow R^d$, y 具有二阶连续导数, $\varphi(\phi, y): M \rightarrow R^{2d+1}$, 其中 $\varphi(\phi, y) = (y(x), y(\phi(x)), y(\phi^2(x)), \dots, y(\phi^{2d}(x)))$, 则称 $\varphi(\phi, y)$ 是集合 M 到集合 R^{2d+1} 的一个嵌入。

利用Takens定理,可将长度为 N 的时间序列 $x(t)$ 在嵌入维数为 m 的空间进行重构, $X(t)$ 为相空间中包含 m 个分量的一个相点,重构后的序列如式(1)所示:

$$X(t) = \{x(t), x(t + \tau), \dots, x(t + (m - 1)\tau)\} \quad (1)$$

其中, M 为进行相空间重构后的相点的数量; m 为嵌入维数; τ 为延迟时间; $t = 1, 2, \dots, M, M = N - (M - 1)\tau$; N 为时间序列的长度。

$X(t)$ 的矩阵形式如式(2)所示:

$$X(t) = \begin{bmatrix} x(1) & x(1 + \tau) & \cdots & x(1 + (m - 1)\tau) \\ x(2) & x(2 + \tau) & \cdots & x(2 + (m - 1)\tau) \\ \vdots & \vdots & & \vdots \\ x(M) & x(M + \tau) & \cdots & x(M + (m - 1)\tau) \end{bmatrix} \quad (2)$$

从Takens定理可以看出,当时间序列不受噪音干扰且时间长度无穷大时,嵌入维数 m 和延迟时间 τ 可任意取值。但在实际问题中,上述条件几乎不可能满足,因此嵌入维数 m 和延迟时间 τ 要合理选取,否则会影响混沌时间序列相空间重构的质量,影响预测效果。

1.2 C-C法

目前,嵌入维数 m 和延迟时间 τ 的计算方法有很多,应用最广泛的是C-C法^[9]。C-C法同时考虑了延迟时间 τ 和嵌入维数 m ,根据时间序列的关联积分得出相关性,然后构造统计量,通过延迟时间与统计量的关系图获得最佳延迟时间 τ 和嵌入时间窗宽度 τ_w ,最后根据 τ, m 和 τ_w 之间的等量关系确定最佳嵌入维数。

令时间序列 $x = \{x_i | i = 1, 2, \dots, N\}$,相空间重构的嵌入维数和延迟时间分别为 m 和 τ ,重构所得的序列为 $X = \{X_i\}$,关联积分公式(3):

$$C(m, N, r, \tau) = \frac{2}{M(M-1)} \sum_{1 \leq i < j \leq M} \theta(r - \|X_i - X_j\|), \quad r > 0 \quad (3)$$

其中, r 为领域半径; N 为时间序列长度; $\theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$

W.A. Brock等人研究指出,独立分布的时间序列 $x = \{x_i | i = 1, 2, \dots, N\}$ 分成 τ 个没有交集的子序列后,如果 m 和 τ 固定,则应该满足 $C(m, N, r, \tau) = C^m(1, N, r, \tau)$,因此可以构造统计量如式(4)所示:

$$S(m, N, r, \tau) = C(m, N, r, \tau) - C^m(1, N, r, \tau) \quad (4)$$

统计量 $S(m, N, r, \tau)$ 表示非线性时间序列的相关性,因此可以以此为基础确定最佳延迟时间 τ_d 和嵌入时间窗宽度 τ_w 。

根据C-C法,可以得到最佳嵌入维数的计算公式(5):

$$m = \frac{\tau_w}{\tau_d} + 1 \quad (5)$$

其中: τ_w 为嵌入时间窗宽度, τ_d 为最佳延迟时间。

研究表明,相空间重构的延迟时间 m 和嵌入维数 τ 之间是有一定的联系的, C-C法同时计算 m 和

τ_d 比单独计算出的结果准确性更高,因此本文采用 C-C 法同时确定 m 和 τ 的值。

2 基于 BP 神经网络的大气污染物浓度预测模型

2.1 BP 神经网络

BP 神经网络的核心思想是误差的反向传播。1969 年,误差的反向传播理论就已被提出,但并不成熟,直到 20 世纪 80 年代,数学家 Rumelhart 等深入研究该理论后提出了 BP 神经网络^[10]。BP 神经网络的原理实际上就是预测值与实际值的误差反向传递过程,在误差传递过程中,利用误差对权值和阈值求导,根据导数的负梯度方向对二者进行调整,使得预测结果满足精度要求。

2.2 模型的建立

研究表明,大气污染物浓度是随时间不断变化的,污染物浓度时间序列具有混沌属性,可以对其进行相空间重构^[11]。在此基础上利用 BP 神经网络建立大气污染物浓度预测模型,建模步骤如下:

(1) 获取大气污染物浓度数据,形成样本集,将其划分为训练集和测试集;

(2) 对样本数据进行归一化处理;

(3) 对训练集样本进行相空间重构,利用 C-C 法求取延迟时间和嵌入维数;

(4) 根据嵌入维数确定模型输入量,即输入层神经元的个数,利用 BP 神经网络建立预测模型;

(5) 根据预测误差最小确定各预测模型隐含层神经元的数目及最佳矩阵集,通过交互验证法获得隐含层节点数最优值,本文选择均方误差作为适应度函数,计算公式(6):

$$f = \frac{1}{M} \sum_{i=1}^M (y_i - y_i^*)^2 \quad (6)$$

其中, M 为训练集样本容量; y_i 为 i 时刻的污染物实际浓度; y_i^* 为 i 时刻的污染物预测浓度。

(6) 利用训练好的 BP 神经网络预测模型对测试集样本进行预测。

2.3 模型的评价

采用均方根误差 ε_{RMSE} 和平均相对误差 ε_{MAPE} 对 BP 神经网络模型的预测效果进行评价,见公式(7)和公式(8):

$$\varepsilon_{RMSE} = \sqrt{\frac{1}{M} \sum_{i=1}^M (y_i - y_i^*)^2} \quad (7)$$

$$\varepsilon_{MAPE} = \frac{1}{M} \sum_{i=1}^M \frac{|y_i - y_i^*|}{y_i} \times 100\% \quad (8)$$

3 算例分析

采用华南某市 2020 年 4 月 16 日至 2020 年 4 月 30 日共 15 天的 SO_2 浓度监测数据组成样本数据进行仿真分析, SO_2 浓度监测数据的采集频率为 1h/次,共获得 600 个样本数据。对 600 个样本数据进行划分,前 540 组作为训练集,后 60 组作为测试集。

利用 C-C 法求得最佳延迟时间 $\tau_d = 3$, 嵌入时间窗宽度 $\tau_w = 10$, 根据公式(12)可得最佳嵌入维数 $m = 5$, 即输入层神经元的个数为 5。

设置输入层到隐藏层的传递函数为 tansig 函数,隐含层到输出层的训练函数为 trainlm 函数,学习率为 0.1%,精度为 10^{-8} ,采用训练集数据进行训练,通过交互验证法确定隐含层的最优个数,不同隐含层神经元个数的训练误差见表 1。由表 1 可知,隐含层神经元个数为 11 时,模型的适应度值最小,由此可以确定 BP 神经网络预测模型的网络结构为 5-11-1。

表 1 不同隐含层神经元个数的训练误差

Table 1 Training error of the number of neurons in different hidden layers

隐含层神经元个数	适应度值
3	8.10×10^{-9}
4	7.60×10^{-9}
5	6.80×10^{-9}
6	8.60×10^{-9}
7	9.60×10^{-9}
8	1.04×10^{-9}
9	1.36×10^{-9}
10	1.19×10^{-9}
11	5.50×10^{-9}
12	7.40×10^{-9}
13	8.20×10^{-9}

利用训练好的 BP 神经网络模型对测试集进行预测,二氧化硫浓度预测结果如图 1 所示,可见 BP 神经网络模型的预测值与实际值变化趋势基本一致。

为了对比分析,采用支持向量机(SVM)和极限学习机(ELM)对测试集样本进行预测,BP 神经网络、SVM 和 ELM 的预测误差见表 2。对比表 2 中的数据可知,BP 神经网络对测试集预测的均方根误差和平均相对误差相比 SVM 和 ELM 均明显下降,可见基于 BP 神经网络的大气污染物浓度预测模型能够显著降低预测过程中的波动性,提高预测精度。

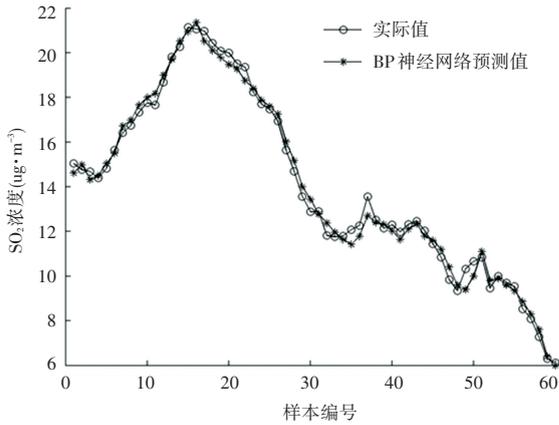


图1 二氧化硫浓度预测结果

Fig. 1 Prediction results of sulfur dioxide concentration

表2 三种模型的预测误差

Table 2 Prediction errors of three models

模型	$\varepsilon_{RMSE} / (\text{mg} \cdot \text{m}^{-3})$	$\varepsilon_{MAPE} / \%$
BP神经网络	0.298	4.35
SVM	0.595	7.29
ELM	0.482	6.48

4 结束语

本文根据大气污染物浓度时间序列的混沌属性,对大气污染物浓度时间序列进行相空间重构,采用C-C法计算延迟时间和嵌入维数,在此基础上构

建了5-11-1的BP神经网络大气污染物浓度预测模型。采用某市SO₂浓度监测数据进行仿真分析的结果表明,本文所提BP神经网络预测模型能够显著降低预测过程中的波动性,提高预测精度。

参考文献

- [1] 张骏,周磊,王文梁,等. 基于Prophet模型的空气污染物浓度预测方法[J]. 中国环境监测,2022,38(6):59-68.
- [2] 凌德森,王晓凯,朱涛. 基于自适应时序分解的空气污染物浓度预测[J]. 测控技术,2023,42(1):83-91,98.
- [3] 王彤彤,严华. 基于EMD和时空图神经网络的污染物浓度预测研究[J]. 现代计算机,2021,27(34):29-35.
- [4] 郑霞,胡东滨,李权. 基于小波分解和SVM的大气污染物浓度预测模型研究[J]. 环境科学学报,2020,40(8):2962-2969.
- [5] 钱栋炜,崔阳光,魏同权. 基于深度神经网络与联邦学习的污染物浓度预测二次建模[J]. 计算机科学,2022,49(S2):932-936.
- [6] 吴子伯,崔云霞,曹炜琦,等. 基于CEEMD-BiGRU模型的徐州市大气污染物浓度预测[J]. 环境工程,2022,40(9):9-18.
- [7] 李萍,倪志伟,朱旭辉,等. 基于改进萤火虫优化算法的SVR空气污染物浓度预测模型[J]. 系统科学与数学,2020,40(6):1020-1036.
- [8] 张琴,林欣如,向阳,等. 基于最优变分模态分解与相空间重构的LSTM船舶横摇运动预测[J]. 船舶工程,2023,45(12):68-74.
- [9] 赵自阳,王红瑞,赵岩,等. 基于混沌理论的汉江上游安康站1950—2014年逐月降水量特征[J]. 长江科学院院报,2021,38(7):137-142.
- [10] 孙宝磊. 基于BP神经网络的大气污染物浓度预测[D]. 昆明:昆明理工大学,2017.
- [11] 付学良,杨洋,李纯子. 基于混沌—支持向量机的大气污染物浓度预测模型[J]. 电子世界,2013(4):78-79.