计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(6):38−47 [doi: 10.15888/j.cnki.csa.008506]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于小波去噪和 LSTM 的 Seq2Seq 水质预测模型①

袁梅雪[1], 魏守科[1,2], 孙　铭[1], 赵金东[1]

[1](烟台大学 计算机与控制工程学院, 烟台 264005)
[2](北京迪普迅智能信息技术有限公司, 北京 100089)
通信作者: 魏守科, E-mail: shouke.wei@gmail.com

摘　要: 建立水质模型预测水质变化是保障饮用水安全、人类健康和维持生态平衡的关键. 本文提出了基于小波分解去噪和 LSTM 的双层双向 Seq2Seq 混合模型 (W-Bi2Seq2Seq) 来预测水质的变化. 使用 Daubechies5 (db5) 小波将数据集分解为低频序列和高频序列, 高频序列作为噪声去除, 仅保留低频信号用作所提出模型的输入. 选取了烟台市门楼水库的 4 项水质指标数据 (pH、氨氮、电导率和浊度) 用于模型的训练, 验证和测试. 所提出的小波双层双向模型 (Bi2) 与小波单层单向模型 (Uni1)、小波单层双向模型 (Bi1)、小波双层单向模型 (Uni2)、传统的 LSTM 模型以及基于小波分解的 LSTM 模型 (W-LSTM), 进行比较实验. 其实验结果显示, 在训练过程中, 4 个 Seq2Seq 模型都具有很好的性能, 都能够很好拟合 4 项水质指标的历史数据集. 然而, 测试结果表明, Bi2 在预测精度和泛化能力方面优于其他 5 个模型, 并且显著提高复杂度较高的水质数据的预测精度.

关键词: 水质预测; 小波去噪; Daubechies5; LSTM; Seq2Seq 模型; 小波分析; 深度学习; 门楼水库

引用格式: 袁梅雪,魏守科,孙铭,赵金东.基于小波去噪和 LSTM 的 Seq2Seq 水质预测模型.计算机系统应用,2022,31(6):38–47. http://www.c-s-a.org.cn/1003-3254/8506.html

Seq2Seq Water Quality Prediction Model Based on Wavelet Denoising and LSTM

YUAN Mei-Xue[1], WEI Shou-Ke[1,2], SUN Ming[1], ZHAO Jin-Dong[1]

[1](School of Computer and Control Engineering, Yantai University, Yantai 264005, China)
[2](Deepsim Intelligent Information Technology Co. Ltd., Beijing 100089, China)

Abstract: Building water quality models to predict variations in water quality is essential for drinking water safety, human health, and ecological balance. In this study, a bidirectional two-layer hybrid Seq2Seq model based on wavelet decomposition denoising and long short-term memory (LSTM), i.e., a W-Bi2Seq2Seq model, is proposed to predict changes in water quality. The Daubechies5 (db5) wavelet is used to decompose datasets into low-frequency series and high-frequency ones. The high-frequency series are removed as noise, and only the low-frequency signal is kept and used as the input to the proposed model. Data series of four water quality indices (pH, NH3-N, conductivity, and turbidity) are collected from the Menlou Reservoir in Yantai, Shandong Province for model training, verification, and testing. The proposed wavelet-based bidirectional two-layer model (Bi2) is compared with the wavelet-based unidirectional one-layer model (Uni1), wavelet-based bidirectional one-layer model (Bi1), wavelet-based unidirectional two-layer model (Uni2), traditional LSTM model, and LSTM model based on wavelet decomposition (W-LSTM). The experimental results show that all the four Seq2Seq models have favorable performance in fitting the historical datasets of the four indices during the training process. Nevertheless, the testing results indicate that Bi2 is superior to the other five models in terms of prediction accuracy and generalization ability and significantly improves the prediction accuracy on water quality data

① 基金项目: 山东省自然科学基金 (ZR2020MF148)
收稿时间: 2021-08-17; 修改时间: 2021-09-26; 采用时间: 2021-10-09; csa 在线出版时间: 2022-05-26

with high complexity.

Key words: water quality prediction; wavelet denoising; Daubechies5 (db5); long short-term memory (LSTM); Seq2Seq models; wavelet analysis; deep learning; Menlou Reservoir

水是生命体最重要的组成部分, 是生命繁衍的基本条件. 随着经济的迅速发展, 工业和生活排放废水量增加, 大量未处理污水排入河流或地下水中, 不仅导致水体使用功能大幅下降, 还加剧了水资源匮乏问题[1]. 据相关文献[2,3] 的研究, 世界上只有很少的部分河流未受污染的影响. 水污染也是造成一些发展中国家疾病和死亡的重要原因之一[4]. 联合国发表的资料表明, 全球有 11 亿人缺乏安全饮用水, 每年有 500 多万人死于与水有关的疾病[5]. 水质的恶化, 已经构成制约和引发一个地区或城市经济发展甚至社会不安定的重要因素[6]. 因此, 建立水质模型预测水质变化是保障饮用水安全和人类健康的关键.

水质数据通常是按时间顺序排列的时间序列数据.循环神经网络 (RNN) 是一种适合于时间序列数据预测的方法[7,8]. 如 Kumar 等人[7] 对河流月流量数据进行了预测研究, 并将 RNN 与前馈神经网络进行了比较, 预测效果较好. Jia 等人[8] 使用 RNN 对湖泊温度和水质数据进行建模, 并通过与 ANN 模型的对比, 证明 RNN 在时间序列数据预测中具有更高的准确性. 然而, RNN 模型存在梯度消失、梯度爆炸和对长距离序列数据的信息依赖性较差等问题. 为了解决这些问题, Hochreiter[9] 提出了 LSTM 模型, 并证明了 LSTM 在预测时间序列数据方面具有独特的优势, 与 RNN 相比有效地提高了预测精度. Hu 等人[10] 对 ANN 和 LSTM 模型对降雨径流量的预测进行了比较, 结果表明 LSTM 模型具有更好的仿真性和更高的智能性. Hu 等人[11] 和 Liu 等人[12] 使用 LSTM 分别对海水养殖区的海水质量和长江的饮用水质量进行了研究, 证明 LSTM 能更准确地反映水质变化的发展趋势. 然而, 对于波动范围大的数据, 单一的 LSTM 模型难以确保预测的准确性[13].

为了提高模型的泛化能力和预测精度, Vinyals 等人[14] 提出了用于时间序列预测的序列对序列 (Seq2Seq) 模型. Seq2Seq 是一种具有编解码结构的网络模型, 不限制输入序列和输出序列的长度, 使模型更加灵活. 同时, 引入注意机制, 减少了早期序列信息的压缩, 进一步提高了远程信息依赖能力. Xiang 等人[15] 运用 Seq2Seq 模型估算每小时降雨径流量, 结果表明其预测精度优于试验中所有其他对比模型. Kao 等人[16] 以台湾石门水库流入量为预测对象, 证明了 Seq2Seq 模型的可靠性. 然而, 目前 Seq2Seq 模型在水质数据预测中的研究还处于起步阶段.

此外, 其他研究将 LSTM 和小波分解相结合, 以提高单一 LSTM 模型的精度. 孙铭等人[17] 建立了水质小波分解和 LSTM 时间序列预测模型 (W-LSTM), 与传统的 LSTM 模型相比, 此模型具有更高的预测精度和泛化能力. Barzegar 等人[18] 提出了一种用于多尺度湖泊水位预测的混合 CNN-LSTM 深度学习和边界校正最大重叠离散小波变换 (DWT) 模型, 成功地提高了湖泊水位预测的精度. Du 等人[19] 提出了一种 DWT、主成分分析 (PCA) 预处理技术和 LSTM 结合的混合模型, 用于需水量预测, 与其他参照预测模型的结果比较, 证明了其所提出混合模型的优越性. Xie 等人[20] 提出了一种结合 LSTM 和 DWT 的深度学习方法 (WA-LSTM) 来预测长江 6 个代表性河段的日水位, 结果表明该方法在应用中稳定可靠.

基于以上研究, 本文提出了一种更为先进的小波 (Wavelet) Seq2Seq 模型 (W-Seq2Seq), 通过小波分解去噪和 LSTM 双层双向 Seq2Seq 模型 (BiSeq2Seq) 相结合来的方法来预测水质变化. 通过与其他 5 种不同结构的模型的对比实验, 验证了所提出方法的有效性.

1 模型算法

1.1 离散小波变换

傅里叶变换是信号处理中广泛使用的分析工具, 它将时域信号转换为频域信号; 但傅里叶变换在时域上缺乏辨别能力[21]. 小波变换的发展解决傅里叶变换时域信息丢失的现象, 小波利用一系列带通滤波器用于将原始时域信号分解为二维时频信息, 这大大提高了局部信号的性能, 并提高了模型的抗噪声性能[22,23].

小波变换是一种数据分解和重构的方法, 使用低通滤波器和高通滤波器将原始数据分解为低频小波系数 cA_n 和高频小波系数 $cD_1, \cdots, cD_n$ [24].

小波变换包括连续小波变换 (CWT) 和离散小波变换 (DWT) 两种. 其中, CWT 基小波$\psi(t)$变换公式为:

$$\psi_{ab}(t)=\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{1}$$

其中, a 为尺度因子 ($a>0$), b 为位移因子 ($b\in R$). 小波变换尺度通过调整 a 和 b 的值, 把实现时间序列信号分解高频时间和低频时间系数.

CWT 公式如下所示:

$$WT_f(a,b)=\int_{-\infty}^{+\infty}f(t)\overline{\psi_{ab}(t)}dt \tag{2}$$

其中, $WT_f(a,b)$ 代表连续小波系数, $f(t)$ 表示原始数据, $\overline{\psi_{ab}(t)}$表示$\psi_{ab}(t)$的共轭函数.

连续小波变换的缺点在于连续小波变换会计算所有尺度的小波系数, 不仅浪费资源, 也产生大量冗余数据. 因此, 通常使用 DWT. DWT 是连续小波变换在尺度和位移上以 2 的幂离散化而得到; 其基小波$\psi_{ab}(t)$变换公式:

$$a=a_0^j, b=ka_0^jb_0 \tag{3}$$

$$\psi_{jk}(t)=a_0^{-\frac{j}{2}}\psi(a_0^{-j}t-kb_0) \tag{4}$$

其中, $a_0>0$, $b_0\in R$, $\forall j,k=0,1,2,\cdots,m\in Z$, $\psi_{jk}(t)$表示小波变换的基小波.

DWT 公式如下所示:

$$WT_f(j,k)=\int_{-\infty}^{+\infty}f(t)\overline{\psi_{j,k}(t)}dt \tag{5}$$

其中, $WT_f(j,k)$ 表示离散小波系数, $f(t)$ 表示原始数据, $\overline{\psi_{jk}(t)}$表示$\psi_{jk}(t)$的共轭函数.

分别重构低频和高频小波系数, 得到低频信号 rA_n 和高频信号 $rD_1,\cdots,rD_n$; 其中低频信号表示近似信息, 高频信号表示详细信息.

最后, 低频信号和高频信号相加, 实现原始信号的重构, 其重构公式如下.

$$f(t)=cA_nl(\psi_{ik}(t))+\sum_{n=1}cD_nh(\psi_{ik}(t)) \tag{6}$$

其中, $f(t)$ 表示重构信号, $l(\psi_{jk}(t))$表示低通滤波器, $h(\psi_{jk}(t))$表示高通滤波器.

小波分解的两个重要任务是选择最优小波和确定分解层数. 常用的小波有 Haar、Daubechies、Sym、Bior、Coif、Morlet、Mexican Hat 和 Meyer. 本文选取 Daubechies5 (db5) 作为基小波, 其原因 db5 是 dbN 小波族中常用的小波之一, 比较适合相对平滑数据集的分解[17].

根据文献 [25], 小波变换的最大分解层数用式 (7) 计算得出.

$$level_{\max}=floor(\log_2(L/(wl-1))) \tag{7}$$

其中, $level_{\max}$ 表示最大分解层数, $floor$ 表示向下取整函数, L 用来表示数据长度, wl 表示小波分解低通滤波器的长度.

1.2 Seq2Seq 模型

Seq2Seq 模型由编码器和解码器两部分组成, 每部分相当于一个独立的 LSTM 模型. 不同之处在于编码器将时间序列数据作为输入 (每个 LSTM 对应于一个时间步长), 生成指定长度的向量 C 作为输出. 向量 C 由编码器中最后一个 LSTM 的隐藏层状态和单元状态组成. 在解码器中, 向量 C 解码为隐藏层状态和单元状态作为输入时, 每个 LSTM 单元将产生预测结果作为输出. 实验对两层双向 Seq2Seq 模型 (Bi2) 结果与一层单向 Seq2Seq 模型 (Uni1)、一层双向模型 (Bi1) 和两层单向模型 (Uni2) 的结果进行了对比.

Uni1 是最基础的 Seq2Seq 模型, 其中编码器中只有一个正向层. 对于 Bi1, 与 Uni1 的区别在于其编码器包含两个独立的循环结构 (LSTM), 一个正向, 一个反向. 与 Uni1 的编码器类似, Bi1 的前向结构用于计算隐藏层信息和单元状态, 而反向结构用来反向读取序列数据 (从 n 到 1), 并计算反向结构生成的一组隐藏层和单元状态. 正向结构产生的隐层和单元状态与反向结构不同, 最后通过连接两种结构的相应部分得到向量 C. 反向结构允许网络先学习后续数据, 并根据反向数据调整其参数, 这可能有助于网络获得前向 LSTM 不具备的依赖性.

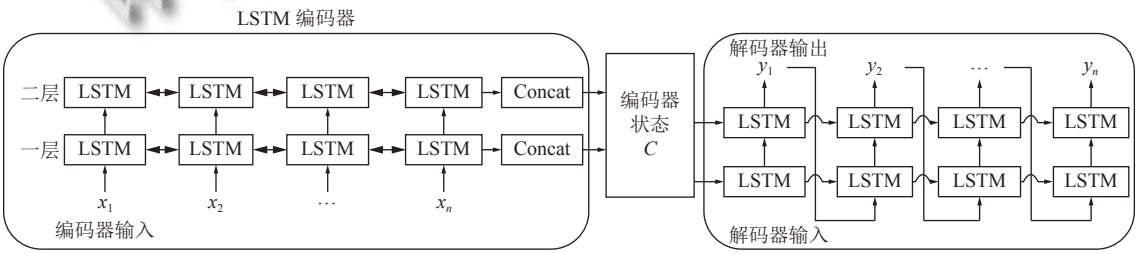

图 1 两层双向 Seq2Seq 模型 (Bi2)

在 Uni2 中, 两个层堆叠在一起, 其中编码器和解码器的第一层 LSTM 的输出被传递到第 2 层作为其输入. 理论上, 具有更多隐藏层的深层 Seq2Seq 体系结构可以有效地学习复杂模式, 并逐步建立输入序列数据的更高

级别表示. Bi2 的编码器和解码器与 Uni2 类似的方式堆叠, 而区别在于第 1 层的正向和反向结构的输出分别传递给第 2 层的正向和反向结构 (图 1). 在理论上, 由于具有更多的隐藏层和前后向层, Bi2 比其他 3 种 Seq2Seq 模型结构具有更强大的能力, 学习较复杂的模式.

1.3 数据处理

1.3.1 均值平滑

运用均值平滑方法, 通过取缺失数据或异常值左右相邻值的平均值来替换数据集中的缺失值和异常值, 如式 (8) 所示.

$$x_t = \frac{x_{t-1} + x_{t+1}}{2} \tag{8}$$

其中, x_t 是数据 t 时刻缺失或异常值的替代值, x_{t-1} 是时间 t–1 上的数据值, x_{t+1} 是时间 t+1 上的数据值.

1.3.2 标准化

为加快模型训练收敛速度, 提高预测精度, 通常将数据集归一化为 [−1, 1] 或 [0, 1] 之间的值. 本文使用了最大-最小归一化方法, 其计算方法为式 (9).

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \times (max_{\text{new}} - min_{\text{new}}) + min_{\text{new}} \tag{9}$$

其中, x_{norm} 表示归一化后的数据; x 为原始数据; $x_{\max}$ 和 $x_{\min}$ 分别代表原始数据的最大值和最小值; max_{new} 和 min_{new} 表示范围的上限和下限, 分别等于 1 和 0.

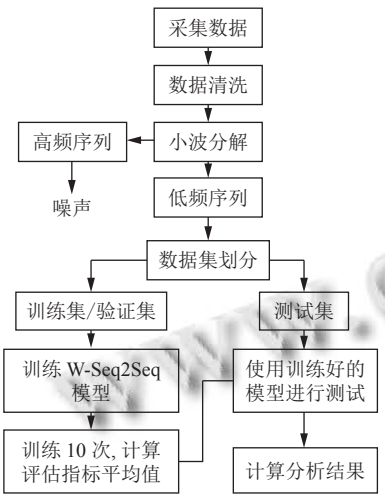

图 2 W-Seq2Seq 模型原理流程图

1.4 W-Seq2Seq 模拟流程

小波分解去燥和 Seq2Seq 结合的混合预测建模方法 (W-Seq2Seq) (图 2), 使用小波将数据集分解为高频信号和低频信号; 其中仅保留低频信号作为原始数据的近似值进行建模, 而高频信号细节作为噪声去除. 所提出的 W-Seq2Seq 建模方法不仅继承了 Seq2Seq 模型强大的预测精度, 而且通过有效降低数据噪声, 为高频复杂数据集提供了一个更优化、更通用的操作系统.

2 研究数据

2.1 数据描述

本文的研究数据来自山东省烟台市福山区门楼镇以西 2 000 m 处青岩河下游的门楼水库, 该水库面积约为 14.65 平方公里, 是烟台市区主要水源地, 烟台市区 70% 以上工业生产和居民生活用水都来自这里[26]. 水库总库容 2 亿立方米, 是一座防洪、灌溉、发电、饲鱼、参观、游览的综合性水库. 然而, 随着当地经济的发展, 水库水质也在逐年恶化, 富营养化已成为水库所面临的主要生态问题. 因此, 控制水库污染, 防止水质进一步恶化已迫在眉睫.

2.2 数据样本

选择 pH、氨氮 (NH_3-N)、电导率和浊度 4 项水质指标, 用无线传感器每 3 s 自动采集数据一次. 采样时间为 2020 年 3 月 6 日至 2020 年 4 月 24 日, 每项指标共采集数据 1 440 000. 最后, 采取每 1 200 个数据的平均值, 把数据转化为小时数据来分析和建模, 每项指标的小时数据集有 1 200 个值. 4 项指标的可视化图清楚地显示了 pH 值和电导率数据系列中存在异常值 (图 3), 使用式 (8) 的平均平滑法替换数据集中的异常值. 清洗后的数据集作为最后实际使用数据, 其中每项指标数据的前 1 080 个值 (90%) 用于模型训练, 最后 120 个值 (10%) 用于模型的测试.

数据集划分及其统计分析如表 1 所示, 浊度测试数据集的平均值为 152.195、最小值为 100.42、最大值为 232.974、第 25 个百分位数、50 个百分位数和 75 个百分位分别为 120.897、145.721 和 178.031, 都远大于训练数据集中相应数据, 这意味着测试数据拥有序列的最大值, 且变化幅度较大, 这表明训练数据可能难以训练一个模型来准确预测测试数据值. 然而, 氨氮和电导率的测试数据集和训练数据集之间没有太大差异, 表明训练数据集足以训练一个模型来准确预测测试数据值.关于 pH 值, 统计分析结果和数据可视化结果显示数据集具有高频特征, 所以需要更复杂的模型来准确预测其测试数据.

3 实验分析

3.1 实验环境

实验中使用的计算机环境配置如下: Windows 10

(64 位), 采用 Intel Core I5-6500 中央处理器, CPU 频率为 3.2 GHz, 内存为 4 GB. 编程语言采用 Python 3.6; 科学计算库采用 Numpy 1.18.5、数据分析库采用 Pandas 1.1.0、数据可视化库采用 Matplotlib 3.3.0. 机器学习库采用 TensorFlow 2.0、集成开发环境 (IDE) 是 PyCharm Professional Edition 2020.1.1.

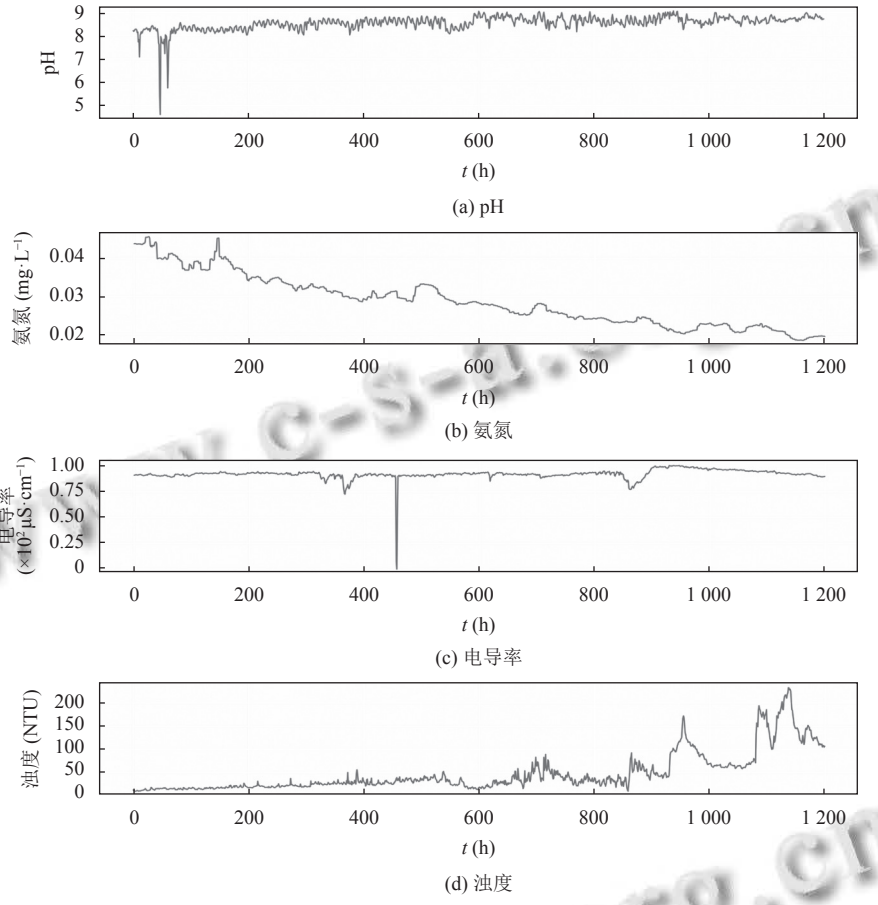

图 3 清洗前的数据集曲线图

表 1 清理后数据的统计分析结果

实验指标	数据集	数量	均值	标准差	最小值	25%	50%	75%	最大值
pH	Total	1 200	8.555	0.250	7.087	8.402	8.584	8.729	9.064
	Training	1 080	8.536	0.255	7.087	8.371	8.561	8.715	9.064
	Testing	120	8.724	0.102	8.506	8.660	8.719	8.782	8.990
氨氮 (mg/L)	Total	1 200	0.029	0.007	0.018	0.023	0.028	0.033	0.046
	Training	1 080	0.030	0.006	0.020	0.024	0.029	0.033	0.046
	Testing	120	0.020	0.001	0.018	0.019	0.020	0.021	0.023
电导率(×102 μS/cm)	Total	1 200	0.923	0.034	0.726	0.910	0.922	0.935	1.003
	Training	1 080	0.922	0.036	0.726	0.909	0.922	0.934	1.003
	Testing	120	0.925	0.016	0.893	0.916	0.924	0.941	0.952
浊度 (NTU)	Total	1 200	48.399	43.126	10.159	20.832	31.422	59.542	232.974
	Training	1 080	36.866	24.385	10.159	20.134	28.893	43.826	172.276
	Testing	120	152.195	35.723	100.42	120.897	145.721	178.03	232.974

注: 25%、50%和75%分别表示第25个百分位数、第50个百分位数和第75个百分位数

3.2 模型评估指标

选取 *MSE*, *RMSE* 和 *MAPE* 作为模型训练和预测精度的评价指标. 3 个评价指标的计算方法用式 (10)、式 (11) 和式 (12) 表示.

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y_t - \overline{y_t})^2 \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \overline{y_t})^2} \tag{11}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \overline{y_t}}{y_t}\right| \tag{12}$$

其中, N 表示数据长度, y_t 表示实际数据, $\overline{y_t}$ 表示预测数据, y_{mean} 表示实际数据的平均值.

3.3 小波分解结果

每项指标的长度为 1 200, 所以根据式 (7), db5 小波对各指标序列最大分解层数为 7 层. 然而, 根据实际经验最大分解层的一半通常是最佳分解层数[19]. 因此在实验中选择了 4 个分解层.

表 2 显示了实际数据和 db5 小波分解的四阶低频信号 (rA_4) 之间的差异. 氨氮的 rA_4 比其他 3 个指标的误差更小, 如 *MSE*≈1.65E–07, *RMSE*≈0.00046 和 *MAPE*≈0.0085. 电导率的 rA_4 和实际数据值之间的误差也非常小, *MSE*≈0.000082, *RMSE*≈0.0091 和 *MAPE*≈0.0091. 相对氨氮和电导率, pH 值的 rA_4 值和实际数据值之间的误差较大, *MSE*≈0.018, *RMSE*≈0.133, 但是 *MAPE* (≈ 0.012) 表明 pH 的 rA_4 值代表实际数据集的准确率也非常高, 约为 98.8%.对于浊度, rA_4 的 *MSE* (60.38) 和 *RMSE* (7.77) 非常大, 而 *MAPE* (< 0.109) 表明浊度 rA_4 代表其实际数据集的准确率也达到 89.1%. 图 4 显示了 4 个指标的实际数据集与其分解后的 rA_4 之间的比较结果, 进一步表明了低频信号 rA_4 在降低数据噪声影响的同时很好地保持了原数据变化的趋势.

表 2 db5 小波分解的四阶低频信号 (即近似值) 与每个指标的实际数据值之间的误差

实验指标	*MSE*	*RMSE*	*MAPE*
pH	0.01769940	0.1330390	0.0122330
氨氮	1.6525E–07	0.0004605	0.0084520
电导率	0.00008220	0.0090714	0.0061145
浊度	60.3803271	7.7704779	0.1089212

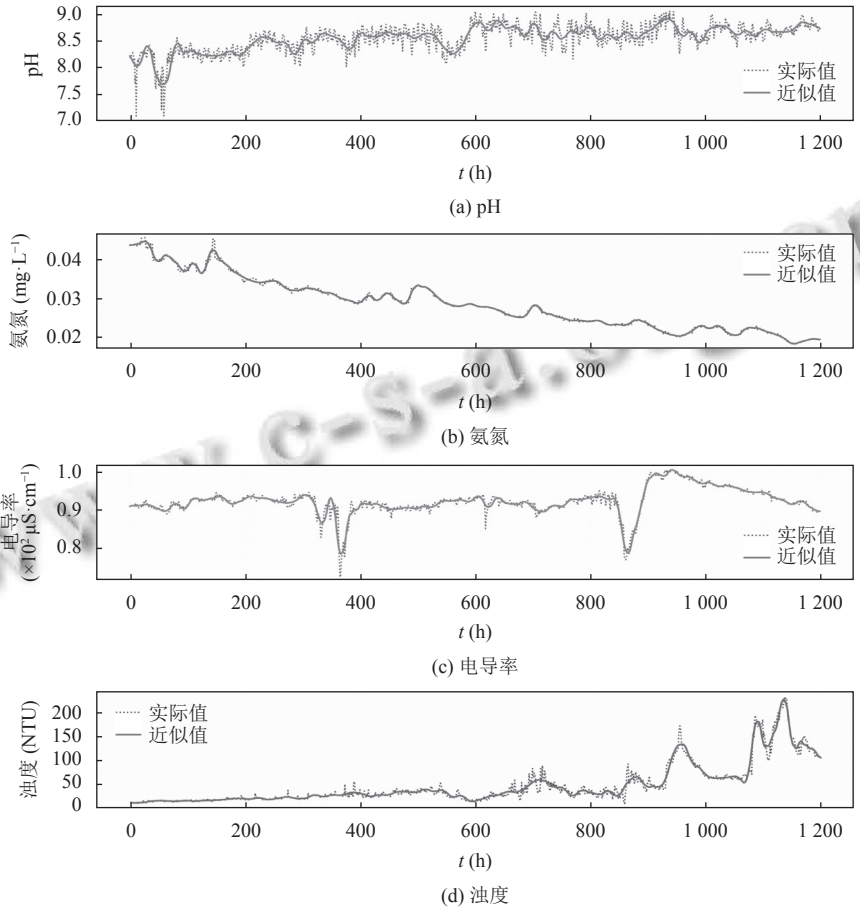

图 4 db5 小波分解近似值曲线 (rA_4) 和实际数据集比较图

3.4 训练结果

此项研究中, Bi2 模型和其他 5 个比较模型 (Uni1、Bi1、Uni2、LSTM、W-LSTM) 都在相同的实验平台和环境下运行. 为了避免实验中的随机因素, 每个模型均训练 10 次, 每次运行 100 轮次 (epoch), 最后取结果的平均值. 在训练过程中, 训练数据集的 10% 进一步用于模型验证, 表 3 总结了 4 种 Seq2Seq 模型和传统 LSTM 模型, 以及孙铭等人[17] 提出的基于小波分解的 LSTM 模型 (W-LSTM) 的训练评估对比结果.

对于 pH 值, 训练的比较结果显示 Bi2 模型在训练和验证数据集上具有最小误差值 (*MSE*≈0.0002, *RMSE*≈0.014 和 *MAPE*≈0.00002), 表明此模型对历史数据的拟合性能最好. 对于氨氮, Bi1 具有最佳的拟合和验证误差, 其 *MSE*、*RMSE* 和 *MAPE* 分别约为 0.0001, 0.01 和 0.0001. 关于电导率, 与 Bi2 相比, 虽然 Uni1 具有略小的训练误差 *MSE* (≈0.0025) 和 *RMSE* (≈0.0501); 但 Bi2 具有较小的训练误差 *MAPE* (≈0.000083) 和验证误差 *MSE* (0.0037)、*RMSE* (0.0608)和 *MAPE* (0.000093), 并且与 LSTM 和 W-LSTM 模型相比, Bi2 具有更小的误差值. 此结果表明, 在训练过程中 Bi2 具有最佳的模拟性能. 对于浊度, 在训练和测试集上 Uni1 具有相对较小评估误差, 但是总体而言, Bi2 和其他两个 Seq2Seq 模型的评估误差也非常小, 如 *MSE* (0.000187–0.000467)、*RMSE* (0.0137–0.0216) 和 MAPE (0.000016–0.000064), 此外 4 种 Seq2Seq 模型与 LSTM 和 W-LSTM 相比总体上误差较小. 训练集和验证集上的评估比较结果表明 4 个 Seq2Seq 模型都具有非常好的性能, 能够高精度地拟合 4 项水质指标的历史数据集, 与其他模型相比具有更好的性能.

表 3 4 种 Seq2Seq 模型和 LSTM、小波 LSTM (W-LSTM) 的训练评估对比结果

模型	实验指标	训练集			验证集		
		MSE	*RMSE*	*MAPE*	*MSE*	*RMSE*	*MAPE*
Uni1	pH	0.000428	0.0207	0.000062	0.000467	0.0216	0.000064
	氨氮	0.001900	0.0436	0.000137	0.002200	0.0469	0.000863
	电导率	0.002500	0.0501	0.000106	0.004400	0.0663	0.000113
	浊度	0.002600	0.0510	0.000045	0.002000	0.0447	0.000037
Bi1	pH	0.000222	0.0149	0.000044	0.000356	0.0189	0.000051
	氨氮	0.000128	0.0113	0.000096	0.000142	0.0119	0.000136
	电导率	0.006700	0.0819	0.000113	0.007700	0.0877	0.000106
	浊度	0.002700	0.0520	0.000045	0.002100	0.0458	0.000037
Uni2	pH	0.000284	0.0169	0.000019	0.000451	0.0212	0.000024
	氨氮	0.001700	0.0412	0.000098	0.001700	0.0412	0.000081
	电导率	0.006000	0.0775	0.000098	0.006700	0.0819	0.000101
	浊度	0.004200	0.0648	0.000053	0.004400	0.0663	0.000054
Bi2	pH	0.000187	0.0137	0.000016	0.000204	0.0143	0.000018
	氨氮	0.000459	0.0214	0.000102	0.000469	0.0217	0.000117
	电导率	0.002600	0.0510	0.000083	0.003700	0.0608	0.000093
	浊度	0.004100	0.0640	0.000051	0.004000	0.0632	0.000050
LSTM	pH	0.009508	0.0975	0.009418	0.009750	0.0954	0.009625
	氨氮	0.001297	0.0321	0.002566	0.001041	0.0301	0.001954
	电导率	0.0073212	0.0812	0.000241	0.007514	0.0800	0.000201
	浊度	0.004531	0.0702	0.000064	0.004512	0.0695	0.000054
W-LSTM	pH	0.004518	0.0672	0.006866	0.004865	0.0686	0.006546
	氨氮	0.001043	0.0310	0.023454	0.001035	0.0328	0.023465
	电导率	0.006121	0.0721	0.000102	0.006152	0.0698	0.000098
	浊度	0.004310	0.0665	0.000059	0.004598	0.0674	0.000052

注: 部分数据来自于文献[27]

3.5 测试结果

6 种模型的测试结果的对比如表 4 和图 5 所示, 具体而言, 在 pH 值测试数据集上, Bi2 具有最佳的预测性能, 预测精度极高, 其 *MSE*、*RMSE* 和 *MAPE* 分别为 7.33E–04、0.027 1 和 0.002 5 (表 4). 然而, 其他 5 个对比模型 Uni1、Bi1、Uni2、LSTM 以及 W-LSTM 的 *MSE*、*RMSE* 和 *MAPE* 测试误差较大, 表明其他 5 个对比模型的预测性能相对较低. 模型的预测对比曲线图 (图 5(a)) 进一步显示, Bi2 对 pH 的预测曲线与实际数据变动趋势吻合较好; 而 Uni1 的预测值图却是一条直线, 表明该模型的预测能力不足. 虽然 Bi1、Uni2、LSTM 和 W-LSTM 都可以预测到 pH 指标的波动趋势, 但它们放大了其波动幅度 (图 5(a)).

就氨氮而言, Uni2 和 Bi2 都具有非常好的预测性能, 因其较小测试误差, *MSE* (3.30E–07 和 4.16E–07)、*RMSE* (5.75E–04 和 6.45E–04) 和 *MAPE* (0.019 7 和 0.023 9) (表 4). Uni1 和 Bi1 虽也预测到氨氮的趋势, 但这两个模型的误差远大于 Uni2 和 Bi2 (图 5(b)), LSTM 和 W-LSTM 与 Uni1 和 Bi1 相比, 误差较大, 不能很好地预测氨氮的趋势. 至于电导率, 测试评估结果显示 Bi2 具有最小的 *MSE* (6.84E–05)、*RMSE* (0.008 3) 和 *MAPE* (0.007 4), 表明 Bi2 具有最好的预测性能; 而 Uni1 和 Bi1 的预测值曲线几乎是线性的, 其分别低估和高估了氨氮测试数据曲线的波动 (图 5(c)).

就浊度而言, *MSE*、*RMSE* 和 *MAPE* 测试评估误差较大, 表明 6 个模型对浊度的整体预测精度不如对其他 3 个指标的预测精度高 (表 4). 因为浊度测试数据集含有整个数据集中的最大值, 且变化浮动较大, 所以模型可能难以准确预测测试数据集, 没有捕捉到测试数据集的波动 (图 5(d)). 然而, 与其他 5 个模型相比, Bi2 的 *MSE* 和 *RMSE* 评估误差较小; 表明 Bi2 也有良好的预测结果. 较小的 *MAPE* (0.114 2) 误差, 进一步表明 Bi2 模型的预测准确率为 88.6% (表 4). 评估比较曲线图直观地显示了 Bi2 较好的预测性能, 其能够捕捉测试数据的波动行为; 而其他 5 个模型预测结果仅为平滑的递减曲线, 不能很好地预测测试数据 (图 5(d)).

4 结论

本文提出一种新颖的小波分解去噪和双层双向 Seq2Seq 的混合水质预测模型 (W-Bi2Seq2Seq). 小波分解的结果证实, 最大分解层数的一半是最佳分解层数, 即在这种情况下的四阶低频信号 (rA4): (1) 是实验实际数据集的最佳近似值; (2) 降低数据复杂性和噪声对实验数据影响的有效方法; (3) 提高模型的泛化能力.

表 4　4 种小波 Seq2Seq 模型和 LSTM、小波 W-LSTM 的测试评估结果

模型	实验指标	*MSE*	*RMSE*	*MAPE*
Uni1	pH	0.003 80	0.061 800	0.006 1
	氨氮	7.88E–07	8.88E–04	0.037 8
	电导率	5.03E–04	0.022 400	0.020 6
	浊度	4 728.200	68.762 00	0.380 7
Bi1	pH	0.007 900	0.089 300	0.007 6
	氨氮	1.03E–06	0.001 000	0.042 7
	电导率	5.79E–04	0.024 100	0.022 4
	浊度	6 405.300	80.033 00	0.489 1
Uni2	pH	0.006 500	0.080 300	0.007 1
	氨氮	3.30E–07	5.75E–04	0.019 7
	电导率	1.48E–04	0.012 180	0.011 1
	浊度	4 276.900	65.398 00	0.346 7
Bi2	pH	7.33E–04	0.027 100	0.002 5
	氨氮	4.16E–07	6.45E–04	0.023 9
	电导率	6.84E–05	0.008 300	0.007 4
	浊度	420.910 0	20.516 00	0.114 2
LSTM	pH	0.116 726	0.339 200	0.036 4
	氨氮	8.81E–05	0.003 429	0.056 7
	电导率	6.57E–04	0.045 812	0.028 7
	浊度	6 541.00	90.125 00	0.517 3
W-LSTM	pH	0.075 691	0.027 511	0.026 1
	氨氮	4.25E–05	0.001 642	0.026 2
	电导率	5.06E–04	0.021 978	0.018 3
	浊度	4 678.400	67.548 92	0.396 8

注: 部分数据来自文献[27]

所采用的小波双层双向模型 (Bi2) 的评估结果与小波单层单向模型 (Uni1)、小波单层双向模型 (Bi1)、小波双层单向模型 (Uni2)、LSTM 模型以及 W-LSTM 模型的结果进行比较. 训练评估结果表明, 提出 4 种 Seq2Seq 模型整体上优于 LSTM 和 W-LSTM 模型, 并且对不同复杂程度的水质数据都有良好的拟合能力. 然而, 测试比较结果表明, Bi2 与其他 3 种 Seq2Seq 模型相比, 在预测复杂性程度较高的水质数据时更具优势. 因为其复杂的建模结构, 能够显著提高模型的预测精度和泛化能力.

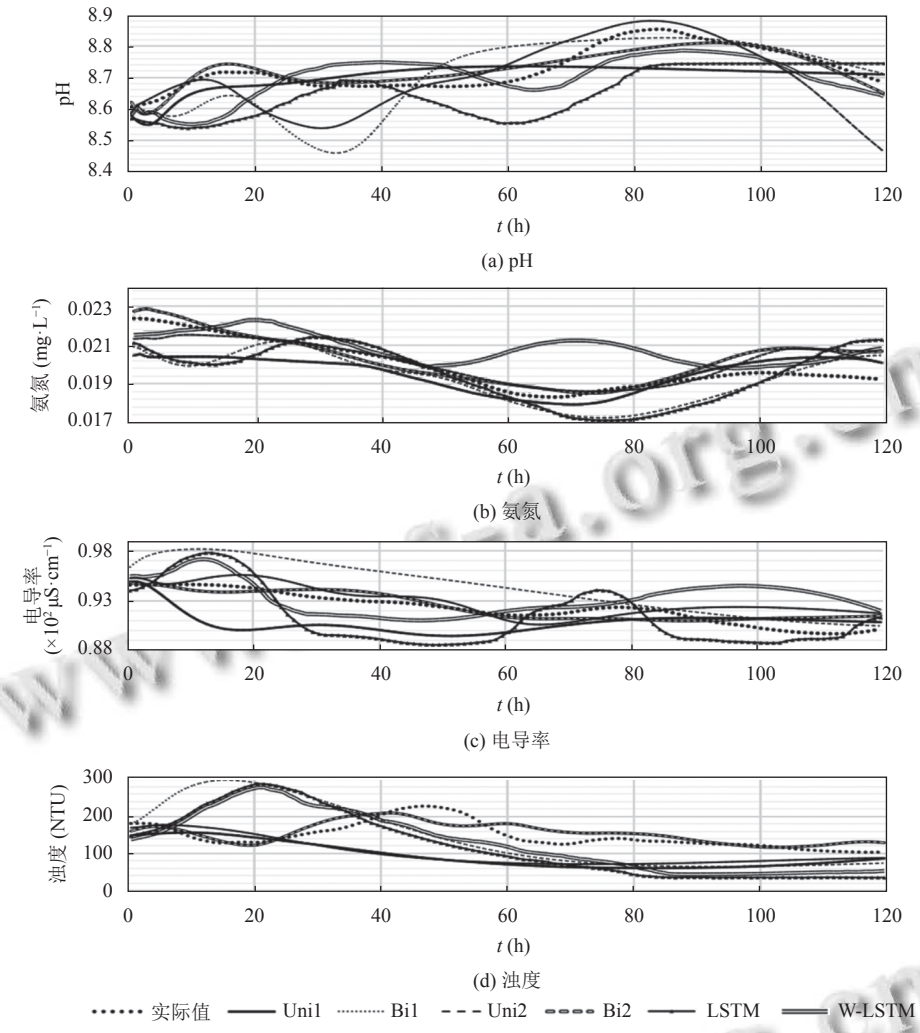

图 5 4 种 Seq2Seq 模型、LSTM、W-LSTM 模型预测结果对比曲线

参考文献

1 余周, 胡志伟, 吴佳, 等. 我国水污染现状、危害及处理措施研究. 环境与发展, 2019, 31(6): 61, 63.

2 Woznicki SA, Nejadhashemi AP, Ross DM, *et al*. Ecohydrological model parameter selection for stream health evaluation. Science of the Total Environment, 2015, 511: 341–353. [doi: 10.1016/j.scitotenv.2014.12.066]

3 Chen YM, Xia JH, Cai WW, *et al*. Three-phase-based approach to develop a river health prediction and early warning system to guide river management. Applied Sciences, 2019, 9(19): 4163. [doi: 10.3390/app9194163]

4 Wang Q, Yang ZM. Industrial water pollution, water environment treatment, and health risks in China. Environmental Pollution, 2016, 218: 358–365. [doi: 10.1016/j.envpol.2016.07.011]

5 张静, 孙晓杰. 水资源环境与人类健康相关性研究. 建材与装饰, 2016, (48): 124–125.

6 高荣伟. 我国水资源污染现状及对策分析. 资源与人居环境, 2018, (11): 44–51. [doi: 10.3969/j.issn.1672-822X.2018.11.009]

7 Kumar DN, Raju KS, Sathish T. River flow forecasting using recurrent neural networks. Water Resources Management, 2004, 18(2): 143–161. [doi: 10.1023/B:WARM.0000024727.94701.12]

8 Jia XW, Karpatne A, Willard J, *et al*. Physics guided recurrent neural networks for modeling dynamical systems: Application to monitoring water temperature and quality in lakes. arXiv: 1810.02880, 2018.

9 Hochreiter S. Untersuchungen zu dynamischen neuronalen Netzen[Master's thesis]. Munich: Technical University of Munich, 1991.

10 Hu CH, Wu Q, Li H, *et al*. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation.

Water, 2018, 10(11): 1543. [doi: 10.3390/w10111543]
11 Hu ZH, Zhang YR, Zhao YC, *et al*. A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. Sensors, 2019, 19(6): 1420. [doi: 10.3390/s19061420]
12 Liu P, Wang J, Sangaiah AK, *et al*. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability, 2019, 11(7): 2058. [doi: 10.3390/su11072058]
13 Lin SL, Huang HW. Improving deep learning for forecasting accuracy in financial data. Discrete Dynamics in Nature and Society, 2020, 2020: 5803407.
14 Vinyals O, Bengio S, Kudlur M. Order matters: Sequence to sequence for sets. 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2015.
15 Xiang ZR, Yan J, Demir I. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resources Research, 2020, 56(1): e2019WR025326.
16 Kao IF, Zhou YL, Chang LC, *et al*. Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting. Journal of Hydrology, 2020, 583: 124631. [doi: 10.1016/j.jhydrol.2020.124631]
17 孙铭, 魏守科, 王莹洁, 等. 基于小波分解的 LSTM 水质预测模型. 计算机系统应用, 2020, 29(12): 55–63. [doi: 10.15888/j.cnki.csa.007695]
18 Barzegar R, Aalami MT, Adamowski J. Coupling a hybrid CNN-LSTM deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting. Journal of Hydrology, 2021, 598: 126196. [doi: 10.1016/j.jhydrol.2021.126196]
19 Du BG, Zhou QL, Guo J, *et al*. Deep learning with long short-term memory neural networks combining wavelet transform and principal component analysis for daily urban water demand forecasting. Expert Systems with Applications, 2021, 171: 114571. [doi: 10.1016/j.eswa.2021.114571]
20 Xie ZQ, Liu Q, Cao YL. Hybrid deep learning modeling for water level prediction in Yangtze River. Intelligent Automation & Soft Computing, 2021, 28(1): 153–166.
21 郭彤颖, 吴成东, 曲道奎. 小波变换理论应用进展. 信息与控制, 2004, 33(1): 67–71. [doi: 10.3969/j.issn.1002-0411.2004.01.015]
22 Mustière F, Bolić M, Bouchard M. Speech enhancement based on nonlinear models using particle filters. IEEE Transactions on Neural Networks, 2009, 20(12): 1923–1937. [doi: 10.1109/TNN.2009.2033367]
23 刘凯, 李文权, 赵锦焕. 短时公交客流小波预测方法研究. 交通运输工程与信息学报, 2010, 8(2): 111–117. [doi: 10.3969/j.issn.1672-4747.2010.02.021]
24 Zhao JD, Wei SK, Wen XB, *et al*. Analysis and prediction of big stream data in real-time water quality monitoring system. Journal of Ambient Intelligence and Smart Environments, 2020, 12(5): 393–406. [doi: 10.3233/AIS-200571]
25 Lee GR, Gommers R, Waselewski F, *et al*. PyWavelets: A Python package for wavelet analysis. Journal of Open Source Software, 2019, 4(36): 1237. [doi: 10.21105/joss.01237]
26 王国群, 王怡, 门世臣. 同心曲——山东烟台门楼水库增容建设纪事. 中国水利, 1991, (4): 43–44.
27 孙铭. 基于深度学习算法的水质预测模型研究 [硕士学位论文]. 烟台: 烟台大学, 2021.