计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(12):55−63 [doi: 10.15888/j.cnki.csa.007695]

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

基于小波分解的 LSTM 水质预测模型①

孙　铭[1], 魏守科[1,2,3], 王莹洁[1], 赵金东[1], 袁梅雪[1]

[1](烟台大学 计算机与控制工程学院, 烟台 264005)
[2](山东琢瑜清泉智能软件科技有限公司, 烟台 264005)
[3](北京迪普迅智能信息技术有限公司, 北京 100089)
通讯作者: 魏守科, E-mail: sigmundwei@163.com

摘　要: 水是人类和其它生命体所依赖的不可缺少的资源, 建立水质预测模型预测水质状况具有重要的社会经济和生态环保价值. 本文建立了基于小波分解的长短期记忆网络 (LSTM) 时间序列预测模型 (W-LSTM), 运用 Daubechies5 (db5) 小波将水质数据分解为高频率和低频率信号, 再将这些信号作为 LSTM 模型的输入, 来训练模型预测水质数据. 利用安徽阜南王家坝流域采集到的 4 项水质指标 (pH 值、DO、CODMn、NH_3N) 对该模型进行训练、验证和测试, 并与传统 LSTM 神经网络模型的训练和预测结果进行比较. 结果显示所提出的方法在多种评价指标上均优于传统 LSTM 模型, 表明了该方法具有较高的预测精度和泛化能力, 是一种更有效的模拟预测手段.

关键词: 水质预测; 小波分解; LSTM 神经网络; 王家坝流域

引用格式: 孙铭,魏守科,王莹洁,赵金东,袁梅雪.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. http://www.c-s-a.org.cn/1003-3254/7695.html

Prediction Model of Water Quality Based on Wavelet Decomposition and LSTM

SUN Ming[1], WEI Shou-Ke[1,2,3], WANG Ying-Jie[1], ZHAO Jin-Dong[1], YUAN Mei-Xue[1]

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

Abstract: Water is an indispensable source of human being and other living species, thus it has significant value of social economy and ecosystem to establish water quality prediction model. This study developed a W-LSTM time series model to predict water quality based on wavelet decomposition and LSTM. Daubechies5 (db5) wavelet was used to decompose water quality data series into high frequency and low frequency signals, and these signals were used as the inputs of LSTM model to train the model to predict water quality data. Four water quality indices (pH, DO, CODMn, and NH_3N) collected from the Wangjiaba River basin in Funan, Anhui Province, China were used to train, validate, and test the model. The training and prediction results of the model were compared with these results of the traditional LSTM neural network model. The results show that the proposed model is superior to the traditional LSTM model in a variety of evaluation indicators. It is proved that this method has higher prediction accuracy and generalization ability and it is a more effective modeling and prediction approach.

Key words: water quality prediction; wavelet decomposition; LSTM; Wangjiaba River basin

① 基金项目: 烟台市科技计划 (2018YT06130844, 2019YT06130885)
Foundation item: Science and Technology Plan of Yantai City (2018YT06130844, 2019YT06130885)
收稿时间: 2020-04-17; 修改时间: 2020-05-15; 采用时间: 2020-05-28; csa 在线出版时间: 2020-11-30

水是人类和其它生命体赖以生存的重要资源, 由于过去工业废水和生活污水未经处理而排放到水体, 导致河流湖泊水体的严重污染, 从而严重破坏了水体的生态环境、生物多样性及其生态功能和服务功能[1]. 据相关研究, 全世界只有小部分河流没有受到水污染影响[2,3], 在一些发展中国家, 水污染是导致疾病和死亡的主要原因之一[4], 仅在中国范围内, 每年因水污染导致约 1.9 亿人次患病, 其中 6 万人死于肝癌、胃癌等疾病[5], 有数据统计, 自 1980 年以来太湖水域频繁发生藻华, 导致长江三角洲地区约 41 种鱼类、65 种浮游动物和 16 种大型植物消失[6]. 因此, 建立准确有效的水质预测模型, 意义重大.

目前, 对于水质的模拟预测方法主要有灰色动态模型群, 混沌理论, 小波神经网络和 BP 人工神经网络等. 如: 李如忠等[7] 利用灰色系统理论构建了一个由 6 个灰色模型组成的灰色动态模型群, 并且利用该模型对水体中的氨氮浓度进行预测, 最终结果取这 6 个 GM 模型的平均值, 消除了 GM 模型本身的不稳定性, 取得了不错的预测效果; 徐敏等[8] 利用混沌理论和相空间重构思想对于水体中溶氧量进行了分析, 结果表明水质具有混沌性, 看似水质变化是无规律的, 但其在短期内具有一定的内在规律可以探寻和预测, 利用混沌相空间模型对水质进行了短期预测, 也取得了一定的成果; 陈建秋等[9] 使用小波神经网络来对水质进行长期预测, 预测精度较高, 证明了其方法的可行性; RI 和侯德刚等[10] 提出 BP 神经网络对水质进行预测, 其中化学需氧量 (COD)、pH 值等数值较接近真实值, 其他指标的预测值的误差也与真实值相差不大, 取得了非常好的预测效果.

水质数据通常是按照时间先后顺序排列的, 较前述文献模拟预测方法, 循环神经网络 (Recurrent Neural Network, RNN) 更加适合处理这种时间序列数据. 如: Jia 等[11] 使用 RNN 对湖泊温度和水质数据进行建模, 通过与 ANN 模型对比证明 RNN 对时间序列数据预测具有更高的精度和准确性; Kumar 等[12] 对河流月流量数据进行预测研究, 将 RNN 与前馈神经网络进行对比试验, 结果表明 RNN 能够以更少的时间代价取得更好的预测效果. 然而, RNN 网络模型存在梯度弥散、梯度爆炸以及对序列数据中长距离依赖信息能力差的问题[13], 而 LSTM 拓展了 RNN 能够更好地解决上述问题, 有效地提高了预测准确度. LSTM 也在许多领域都取得了不错的进展, 比如在自然语言处理方面, 胡新辰[14] 利用 LSTM 解决语义关系分类问题取得了重要成果; 在股票运作方面, 孙瑞奇[15] 基于 LSTM 并利用拟牛顿法原理改变网络模型的学习速率, 证明了 LSTM 能够很好地预测股市的变化; 在空气质量预测方面, 张冬雯等[16] 利用 LSTM 更精确地对 Delhi 和 Houston 两地的空气质量 AQI 指数做出了预测; 在降雨径流量预测方面, Hu 等[17] 通过对比 ANN 和 LSTM两种模型的预测结果, 表明 LSTM 模型具有更好的仿真性和更高的智能性. 上述多个研究都表明 LSTM 对时间序列数据的预测方面具有得天独厚的优势. 然而, 利用 LSTM 对水质时间序列进行预测的文献资料相对较少. 如: 刘晶晶等[18] 采用 K-Similarity 方法对地表水水质数据进行降噪, 利用 LSTM 神经网络预测降噪后水质数据变化, 研究表明相较于 BP 神经网络和 RNN, LSTM 对水质序列数据有更好的预测能力; Hu 等[19] 和 Liu 等[20] 使用 LSTM 分别研究了海产养殖区的海水水质和扬子江水源地的饮用水水质, 他们实验结果都表明 LSTM 能够更准确地反映水质变化的发展趋势, 证明了 LSTM 预测水质的可行性和有效性. 但是, 传统的神经网络模拟预测方法对于序列波动变化较大并存在长期趋势的时间序列, 其预测结果并不理想[21,22]. 本文提出基于小波分解的 LSTM 时间序列模拟预测方法 (W-LSTM), 运用小波将水质数据分解为高频和低频信号, 作为 LSTM 模型的输入, 来训练模型预测水质数据. 同时, 将模型预测结果与传统 LSTM 神经网络的结果进行对比, 验证该方法的有效性.

1 W-LSTM 算法原理

1.1 小波变换原理

傅里叶变换是信号处理领域应用极广的一种分析手段, 它可以将时域信号转换成频域信号, 但是傅里叶变换在时域中没有辨别能力[23]. 小波变换正是针对傅里叶变换的不足之处发展而来, 利用小波和一族带通滤波器对原时域函数进行分解, 将信号分解为二维的时频信息, 极大地增强了局部信号的表现能力, 提高了模型的抗噪性[24].

小波变换是一种数据分解、重构方法, 该方法首先分别利用低通滤波器和高通滤波器将原始数据分解成低频小波系数 cA_n 和高频小波系数 cD_1, $\cdots$, cD_n. 其中, 低频小波系数还可以再做进一步的分解, 此过程可

以迭代数次, 直至达到最大分解次数.

小波变换可以分为连续小波变换 (CWT) 和离散小波变换 (DWT). 为了提高连续小波变换处理复杂问题的能力, CWT 对基小波$\psi(t)$进行改造, 如下式:

$$\psi_{ab}(t)=a^{-\frac{1}{2}}\psi\left(\frac{t-b}{a}\right) \tag{1}$$

其中, a 为伸缩因子 ($a>0$), b 为平移因子 ($b\in R$), 通过调整 a 和 b 的值来够控制小波变换的尺度, 从而达到高频处时间细分, 低频处频率细分, 实现自适应时频信号分析的要求[25].

连续小波变换公式如下:

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

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

然而, 连续小波变换会计算所有尺度上的小波系数, 这一耗时的过程也会产生许多冗余数据. 因此, 在实际过程中通常使用离散小波变换. 离散小波变换是对连续小波变换在尺度和位移上按照 2 的幂次进行离散化所得. 将$\psi_{ab}(t)$函数中 a 和 b 的计算方法如式 (3) 所示:

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

其中, $a_0>0$, $b_0\in R$, $\forall j$, $k=0,1,2,\cdots,m\in Z$, 则函数$\psi_{jk}(t)$的计算方法如式 (4) 所示:

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

离散小波变换公式如下:

$$Wf(j,k)=\int_{-\infty}^{+\infty}f(t)\overline{\psi_{jk}(t)}\mathrm{d}t \tag{5}$$

其中, $Wf(j,k)$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.

将原始数据进行分解之后, 再分别对低频小波系数和高频小波系数进行重构. 低频小波系数和高频小波系数重构后得到低频信号 rA_n 和高频信号 $rD_1,\cdots,rD_n$. 其中, 低频信号表示逼近信息, 高频信号表示细节信息.

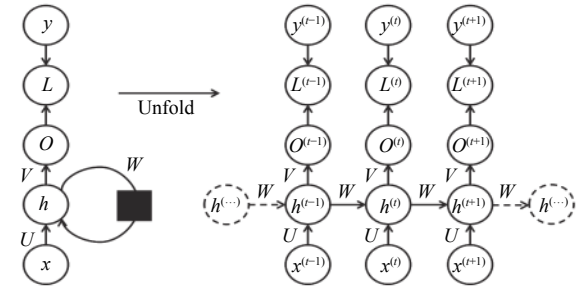
最后, 将所有低频信号和高频信号相加实现数据还原. 重构与还原公式如下:

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

其中, $f(t)$ 表示还原之后的数据, $l(\psi_{ik}(t))$表示低通滤波器, $h(\psi_{ik}(t))$表示高通滤波器.

1.2 LSTM 原理

RNN 擅长处理以时间序列数据作为输入的预测问题, 其原因在于 RNN 的网络结构可以处理时间序列数据之间的相关性. RNN 结构如图 1 所示.



(a) RNN 基本结构图 (b) RNN 展开结构图

图 1 RNN 原理结构图

图 1(a) 为 RNN 的基本结构图, 包括输入层 x、隐藏层 h、输出层 o, 在隐藏层 h 上有一个循环操作, 同时 RNN 在所有时刻的线性关系参数 U、W、V 都是共享的, 极大地减少了参数训练量. 图 1(b) 为 RNN 展开结构图, 可以看到 RNN 通过权值 W 实现隐藏层之间的依赖关系.

然而, 在实际使用时发现 RNN 存在诸如梯度消失、梯度爆炸以及长距离依赖信息能力差等问题, 为了解决这些问题, 引入了 LSTM. LSTM 在主体结构上与 RNN 类似, 其主要的改进是在隐藏层 h 中增加了 3 个门控 (gates) 结构, 分别是遗忘门 (forget gate)、输入门 (input gate)、输出门 (output gate), 同时新增了一个名为细胞状态 (cell state) 的隐藏状态.

图 2 展示了 LSTM 隐藏层的内部结构, 其中 $f(t)$、$i(t)$、$o(t)$ 分别表示 t 时刻遗忘门、输入门、输出门的值, $a(t)$ 表示 t 时刻对 $h(t-1)$ 和 $x(t)$ 的初步特征提取.

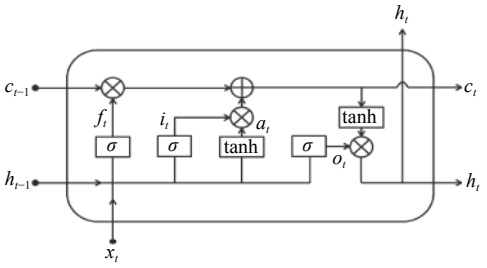


图 2 LSTM 隐藏层结构原理

$$f(t)=\sigma(Wfh_{t-1}+Ufx_t+bf) \quad (7)$$

$$i(t)=\sigma(W_ih_{t-1}+U_ix_t+b_i) \quad (8)$$

$$a(t)=\tanh(W_ah_{t-1}+U_ax_t+b_a) \quad (9)$$

$$o(t)=\sigma(W_oh_{t-1}+U_ox_t+b_o) \quad (10)$$

其中, x_t 表示 t 时刻的输入, h_{t-1} 表示 t–1 时刻的隐层状态值, W_f、W_i、W_o 和 W_a 分别表示遗忘门、输入门、输出门和特征提取过程中 h_{t-1} 的权重系数, U_f、U_i、U_o 和 U_a 分别表示遗忘门、输入门、输出门和特征提取过程中 x_t 的权重系数, b_f、b_i、b_o 和 b_a 分别表示遗忘门、输入门、输出门和特征提取过程中的偏置值, tanh 表示正切双曲函数, σ表示激活函数 Sigmoid.

$$\tanh(x)=\frac{1-e^{-2x}}{1+e^{-2x}} \quad (11)$$

$$\sigma(x)=\frac{1}{1+e^{-x}} \quad (12)$$

遗忘门和输入门计算的结果作用于 $c(t-1)$, 构成 t 时刻的细胞状态 $c(t)$.

$$c(t)=c(t-1)\odot f(t)+i(t)\odot a(t) \quad (13)$$

其中, ⊙为 Hadamard 积[26,27]. 最终, t 时刻的隐藏层状态 $h(t)$ 由输出门 $o(t)$ 和当前时刻的细胞状态 $c(t)$ 求出.

$$h(t)=o(t)\odot\tanh(c(t)) \quad (14)$$

2 W-LSTM 模型

2.1 W-LSTM 网络模型

LSTM 神经网络对预测时间序列数据具有较强的优势, 但对于复杂度和变化频率较高的数据, 单一 LSTM 预测方法很难获取数据的变化规律, 使得模拟和预测结果欠佳. 而小波分解能将原始数据中不同频段的信息进行分解, 极大地降低数据复杂度, 再分别对这些数据进行预测从而提高预测精度. 本文将上述两种方法结合提出基于小波分解的 LSTM 时间序列预测模型 (W-LSTM). 其训练、预测流程如图 3 所示. 步骤如下:

(1) 对采集到的水质指标数据使用均值平滑法降噪, 然后归一化.

(2) 4 项样本数据统一划分为前 435 组作为训练数据, 后 45 组作为测试数据.

(3) 使用训练数据作为样本输入用于训练 W-LSTM 神经网络模型, 对模型进行如下两步操作:

① 选取“db5”作为基小波, 并对数据进行 3 阶小波分解, 获取低频信号 rA_3 和高频信号 rD_1、rD_2、rD_3.

② 使用 LSTM 分别对 rA_3、rD_1、rD_2、rD_3 进行预测.

不断调整参数, 直到获取目标 loss 或者达到最大训练次数. 最终生成 W-LSTM 神经网络模型.

(4) 使用测试数据作为 W-LSTM 模型输入样本, 输出模型预测准确度并与对比试验模型进行误差比较.

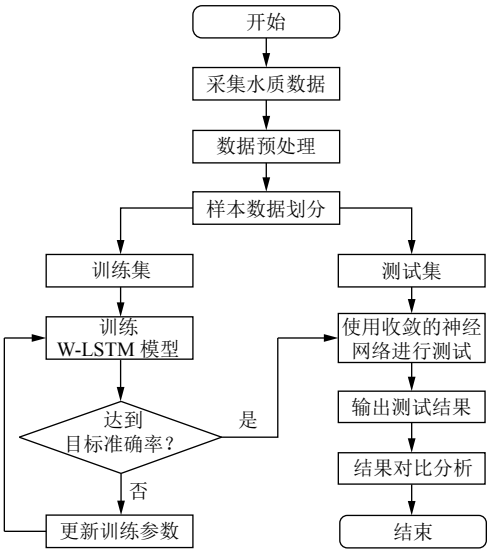


图 3 基于 W-LSTM 模型的水质预测流程图

2.2 数据样本

本文以安徽阜南王家坝水库的水质数据作为研究对象, 该水库位于安徽省阜阳市阜南县王家坝国家湿地公园, 湿地占地面积约为 6761.71 公顷, 作为当地市民主要的供水水库, 其水质健康显得十分必要. 根据国家地表水质环境质量标准[28], 选取 pH 值、溶解氧含量 (DO)、高锰酸盐指数 (CODMn) 和氨氮含量 (NH_3N) 指标作为实验数据. 所有指标数据的采集时间均为 2018 年 03 月 01 日到 2019 年 06 月 23 日, 每 24 小时采集一次, 数据一共 480 组, 取前 435 组作为训练数据, 后 45 组数据作为测试数据.

对数据样本简单分析, 查看是否有缺失值、异常值等情况, 如表 1 所示.

表 1 数据样本统计分析 (DO 值、CODMn 值和 NH_3N 值的单位为: mg/L)

实验参数	均值	标准差	最大值	最小值	缺失数量
pH	7.558 354	0.563 744	8.53	0	2
DO	7.386 375	1.992 297	13.2	0	1
CODMn	5.349 792	1.852 043	19.1	0	1
NH_3N	0.847 396	0.509 247	3.74	0.17	0

2.3 数据预处理

2.3.1 数据清洗

在数据采集和测量的过程中由于仪器设备故障、不当的人为操作以及其他不可控因素的干扰, 采集到的数据不可避免的会导致一些数据丢失和数据录入失真的情况, 如果直接使用这些含有噪声的数据开展实验研究, 不仅耗费人力物力资源, 还会产生不准确的实验结论, 从而误导日后的研究工作. 因此, 在实验开始之前, 首先要对数据进行清洗. 观察实验数据和表 1 后发现仅存在几处数据缺失的情况, 正常录入的数据没有发现明显噪声. 采用均值平滑法将数据缺失部分的数据补充完整. 均值平滑法是利用缺失数据左右相邻两处的数据, 取平均值来替代缺失数据, 如式 (15) 所示:

$$x_a = \frac{x_{a-1} + x_{a+1}}{2} \tag{15}$$

其中, x_a 为 a 时刻的缺失数据, x_{a-1} 为 a–1 时刻的正常数据, x_{a+1} 为 a+1 时刻的正常数据.

2.3.2 数据归一化

为了加快模型的收敛速度同时提升模型的预测精度, 需要对数据进行归一化处理, 将数据转换成 [0, 1] 之间的数值. 本文使用 max-min 归一化方法, 其计算方法如式 (16) 所示:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{16}$$

其中, x_{norm} 表示归一化之后的数据, x 表示未归一化的数据, $x_{\max}$、$x_{\min}$ 分别表示所有数据中的最大值和最小值.

2.4 离散小波变换流程

使用离散小波分解数据时应注意两点. 第一、需要确定基小波的种类. 常用的基小波有 Haar 小波、db 小波、sym 小波、bior 小波、coif 小波、Morlet 小波、mexicanHat 小波以及 Meyer 小波. 他们都是一个小波族, 每个小波族中包含众多具体的小波. 最佳小波的选择没有明确的标准, 但实际上无论选哪种小波作为基小波差别也不很大. 本文选择 Daubechies5 (db5) 作为基小波 (db5 是 db 小波族中常用的小波之一, 如图 4 所示), 原因是 db5 更适用于分解比较平滑的数据集, 而我们采集的水质数据整体上比较平滑. 第二, 需要确定分解层数, 利用式 (17)[29] 可以计算出数据的最大分解层数为 5 层, 但是根据经验选最大分解层数的一半即可, 所以最终确定分解层数为 3 层.

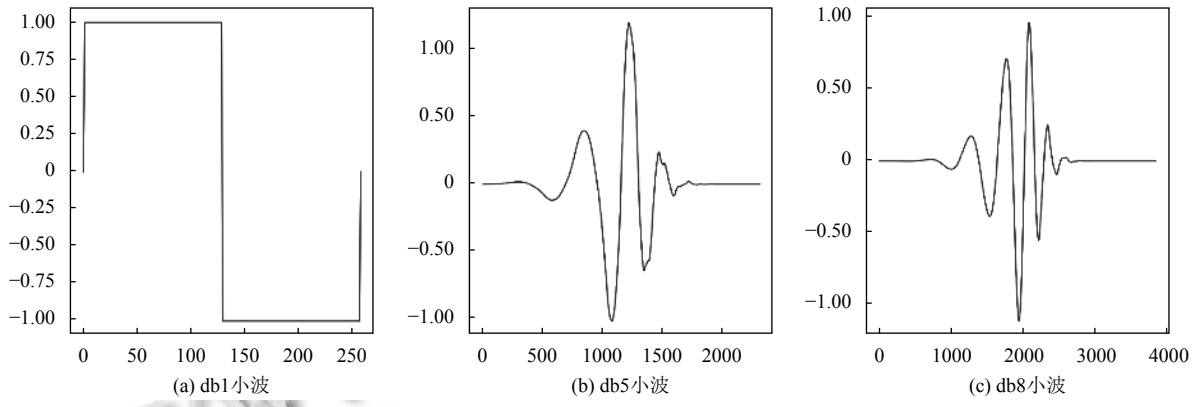


图 4 db 小波示意图

$$L = \ln(n_d/(lw-1))/\ln \tag{17}$$

其中, lw 表示小波分解低通滤波器的长度, n_d 表示数据长度.

2.5 模型评估

本文选择 4 种评价指标作为判断模型预测效果优劣的依据, 其分别是均方误差 (Mean Squared Error, *MSE*)、均方根误差 (Root Mean Squared Error, *RMSE*)、平均绝对误差 (Mean Absolute Error, *MAE*) 和平均百分比误差 (Mean Absolute Percentage Error, *MAPE*), 其计算方法如式 (18)~式 (21) 所示.

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y}_t)^2 \tag{18}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y}_t)^2} \tag{19}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|(y_t - \bar{y}_t)| \tag{20}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \bar{y}_t}{y_t}\right| \quad (21)$$

其中, N 表示总数据量, y_t 表示真实值, $\bar{y}_t$表示预测值.

3 实验结果分析

3.1 实验平台与环境

实验所使用的计算机配置如下: 处理器为英特尔 Core i5-8250U, CPU 频率为 1.8 GHz, 内存为 8 GB, 操作系统为 Windows 10 (64 位); 程序设计语言为 Python 3.7, 数值计算、分析库为 Numpy 1.17.1, Pandas 0.25.2, 机器学习库为 Tensorflow 1.14.0, 数据可视化库为 Matplotlib 3.1.1; 集成开发环境为 PyCharm Community Edition 2018.3.1.

3.2 结果分析

为了更好地验证所提出模型的精确性, 选取传统的 LSTM 神经网络与该模型对比实验. 两种模型均在相同的实验平台和环境下进行. 均采用自适应矩估计 (adaptive moment estimation) 进行优化, 损失函数选择 *MSE*、*RMSE*、*MAE*、*MAPE* 4 种方式进行评价. 为尽量避免实验中产生偶然因素, 每组实验各进行 10 次.

3.2.1 小波分解

以 pH 数据为例直观展示 3 阶小波分解的结果, 如图 5 所示.

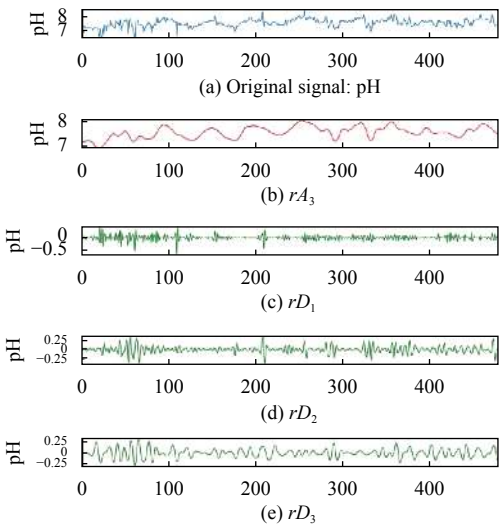


图 5 pH 数据的原始值及其 3 阶小波分解得到的低频数据 rA_3 和高频数据 rD_1、rD_2、rD_3

然而, 将经过小波重构之后各个频段的数据信号相加还原, 这一过程与原数据确实存在一定的误差. 表 2 展示了本次实验中 4 项指标重构后与原始数据的误差值. 可发现其最大误差为 6.70e–16, 最小的误差为 8.33e–17, 此误差值对实验结果影响非常小, 可以忽略不计.

表 2 还原数据与原始数据的误差值

实验参数	误差值
pH	6.37e–16
DO	8.71e–16
CODMn	6.70e–16
NH_3N	8.33e–17

3.2.2 训练结果

本文中 W-LSTM 模型和 LSTM 模型的调节参数包括 batch_size(批量大小), window_size(窗口大小), num_units(节点数量), Learning_rate(学习率), steps(训练步长). 在保证网络快速收敛的同时又具有较高的预测精度, 经过多次实验测试与参数调整, 模型达到最优结果. 表 3 展示了实验相关参数的最终配置结果.

表 3 W-LSTM 和 LSTM 的网络参数以及收敛速度

模型参数	W-LSTM				LSTM
	*rA*3	*rD*3	*rD*2	*rD*1	原始数据
batch_size	30	30	40	40	20
window_size	30	30	30	30	60
num_units	32	64	128	256	128
learning_rate	0.0003	0.0003	0.0005	0.0007	0.0001
steps	5000	3500	3000	2000	4000
收敛速度(s)	48	49	67	143	203
总收敛速度(s)			307		203

实验中反映 W-LSTM 和传统 LSTM 两种模型训练拟合情况的各项评估指标值记录在表 4 中. 结果显示两种模型对 pH 的拟合情况基本一致, 且相较于其他 3 项实验参数拟合精度最高, *MSE* 均低于 0.0008, 这与 pH 数据的值域变化较小有关; 而 DO、CODMn 和 NH_3N 传统 LSTM 模型训练拟合结果却都略优于 W-LSTM, 3 项参数在 MSE 上分别减小了 0.0066、0.0073 和 0.002, 究其原因, 不难发现 W-LSTM 模型将原数据分解为低频 rA_3 和高频 rD_1、rD_2、rD_3 4 项值, 并对它们分别拟合, 拟合过程的增多不可避免地会增大误差, 最终导致同样量级的训练过程会呈现不同的拟合效果, 同样地, 拟合过程增多会降低模型训练收敛的速度, 其所耗时间必定高于传统 LSTM 模型. 表 3 结果显示 W-LSTM 经过 4 个模型训练过程, 总收敛时间比传统 LSTM 模型耗时多约 100 s.

为了更加直观的表现各项数据的拟合情况, 将 W-

LSTM 与传统 LSTM 的拟合情况进行对比如图 6 所示. 从图中可以观察到两种模型都充分学习了训练数据的特性, 拟合情况良好, 并且没有过拟合的情况发生, 能够达到训练要求, 证明实验的有效性.

表 4　4 项指标 W-LSTM 和 LSTM 模拟训练拟合精度评估结果

实验参数	W-LSTM				LSTM			
	MSE	*RMSE*	*MAE*	*MAPE*	*MSE*	*RMSE*	*MAE*	*MAPE*
pH	0.000795	0.028208	0.013851	0.001859	0.000776	0.027864	0.012363	0.001666
DO	0.012872	0.113457	0.081328	0.011719	0.006262	0.079133	0.041823	0.006132
CODMn	0.040112	0.200279	0.071837	0.011687	0.032834	0.181201	0.06726	0.010809
NH_3N	0.002985	0.054639	0.037843	0.054266	0.001027	0.032043	0.017319	0.025662

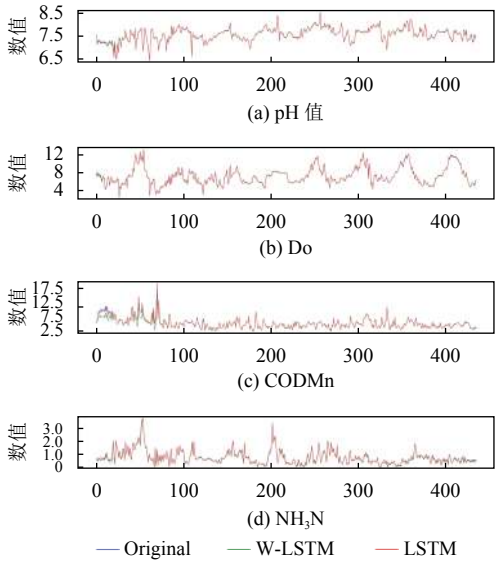


图 6　W-LSTM 和 LSTM 模型 4 项指标训练结果对比图

3.2.3　测试结果

本文的实验数据其最大频率为 240 Hz, 对其做 3 阶 DWT, 则 rA_3 表示频段小于 30 Hz 的分量数据, rD_3、rD_2 和 rD_1 分别表示频段 30~60 Hz、60~120 Hz、120~240 Hz 的分量数据. 理论上相较于原始数据, 分量数据的复杂度更低, 所以对分量数据进行预测的准确度也更高, 通过分量数据获得的全频率上的预测结果准确度也更高. 其中, 高频数据来自原始数据变化较快的部分, 反映信号细节变化特征, 低频数据来自原始数据变化较慢的部分, 低频信号比较平滑, 反映信号的变化趋势.

表 5 为 W-LSTM 和传统 LSTM 模型在 10 次预测中各项指标的均值对比情况.

从表 5 中可以明显看出, W-LSTM 模型在水质时间序列指标数据预测方面优于传统 LSTM 模型. 在 *MSE*、*RMSE*、*MAE* 和 *MAPE* 4 项评估指标中, W-LSTM 比传统 LSTM 的预测精度在 pH 数据上分别提高了 35.1%、18.9%、28.3% 和 28.3%; 在 DO 数据上分别提高了 62.3%、35.0%、34.6% 和 31.3%; 在 CODMn 数据上分别提高了 27.9%、15.4%、17.6% 和 15.4%; 在 NH_3N 数据上分别提高了 53.8%、32.3%、35.8% 和 44.7%. 究其原因小波变换能够对数据的整体趋势和细节信息的分层把握能力, 加上 LSTM 模拟预测时间序列数据上的优势, 保证了 W-LSTM 不仅能够更清晰的了解数据的整体走势, 还能更精确的预测数据的细节变化. 这为 W-LSTM 在时间序列数据预测方面提供了更强的能力, 而且其效果更优于传统 LSTM. 观察表 5 中 W-LSTM 模型的预测情况不难发现, 在多项指标上 pH 和 NH_3N 的结果精度较高, 而 DO 和 CODMn 的结果精度相对较低. 其主要原因是 pH 和 NH_3N 数据的标准差较小 (表 1), 数据离散程度较低, 所以期望获得的预测精度越高; 而 DO 和 CODMn 数据的标准差相对较大 (表 1), 数据离散程度相对较高, 致使期望获得的预测精度稍有逊色.

表 5　W-LSTM 模型和传统 LSTM 模型在 3 次预测中各项评估指标均值结果

实验参数	W-LSTM				LSTM			
	MSE	*RMSE*	*MAE*	*MAPE*	*MSE*	*RMSE*	*MAE*	*MAPE*
pH	0.075684	0.275105	0.202511	0.026124	0.116627	0.339245	0.282599	0.036455
DO	0.82769	0.90471	0.7109	0.084609	2.194613	1.391566	1.08754	0.123123
CODMn	0.663747	0.811614	0.646335	0.144818	0.920293	0.959266	0.784275	0.171192
NH_3N	0.041036	0.201751	0.15782	0.396233	0.088827	0.298038	0.245853	0.716348

图 7 进一步展示了 W-LSTM 和传统 LSTM 模型对 pH、DO、CODMn 和 NH_3N 4 项水质指标的预测对比结果, 可以看出 W-LSTM 相较于传统 LSTM 模型的预测情况, 在总体趋势上与原数据更为一致, 同

时对某些细节信息例如峰值处也有更加精确的预测表现.

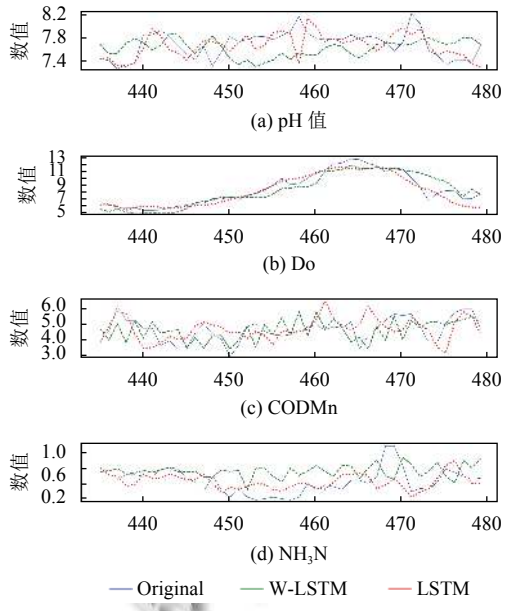


图 7 W-LSTM 和 LSTM 模型 4 项指标预测结果对比图

4 结论

本文提出了基于小波分解的 LSTM 时间序列预测模型 (W-LSTM), 对水质指标数据进行模拟预测实验. 结果发现, 使用 db5 小波对水质数据进行分解与重构过程的误差非常小, 表明离散小波变换具有完全重现原始数据的能力, 保证实验的有效性. 其次, 传统 LSTM 模型预测水质数据的结果在整体趋势上通常不能很好地表现出来, 而 W-LSTM 最大优势在于对整体趋势的判断以及对细节的把握, 实现了对时间序列数据的精确预测. 最后, 通过对低频数据预测的观察与分析还可以从宏观上了解数据的未来走势, 从而更好地指导工作展开.

以王家坝水库水质数据作为研究时间序列数据的切入点, 本文通过实验分析证明 W-LSTM 能够显著提高水质数据预测的精度. 然而, 试验仅运用了一个水域的部分水质数据, 研究结论是否具有通用性仍有待大量试验验证. 因此, 未来将 W-LSTM 模型应用于更多场景, 以研究和验证此方法的通用型.

参考文献

1 HaRa J, Mamun M, An KG. Ecological river health assessments using chemical parameter model and the index of biological integrity model. Water, 2019, 11(8): 1729. [doi: 10.3390/w11081729]

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 Tao T, Xin KL. Public health: A sustainable plan for China's drinking water. Nature, 2014, 511(7511): 527–528. [doi: 10.1038/511527a]

6 Guan BH, An SQ, Gu BH. Assessment of ecosystem health during the past 40 years for Lake Taihu in the Yangtze River Delta, China. Limnology, 2011, 12(1): 47–53. [doi: 10.1007/s10201-010-0320-6]

7 李如忠, 汪家权, 钱家忠. 基于灰色动态模型群法的河流水质预测研究. 水土保持通报, 2002, 22(4): 10–12. [doi: 10.3969/j.issn.1000-288X.2002.04.003]

8 徐敏, 曾光明, 苏小康. 混沌理论在水质预测中的应用初探. 环境科学与技术, 2004, 27(1): 51–54. [doi: 10.3969/j.issn.1003-6504.2004.01.024]

9 陈建秋, 张新政. 基于小波神经网络的水质预测应用研究. 2006 中国控制与决策学术年会论文集. 天津, 中国. 2006. 723–726.

10 RI SI, 侯德刚, 张振家, 等. 基于 BP 人工神经网络的生化处理水水质预测. 现代化工, 2009, 29(12): 66–68, 70. [doi: 10.3321/j.issn:0253-4320.2009.12.016]

11 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 preprint arXiv: 1810.02880, 2018.

12 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]

13 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述. 计算机应用, 2018, 38(S2): 1–6, 26.

14 胡新辰. 基于 LSTM 的语义关系分类研究 [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2015.

15 孙瑞奇. 基于 LSTM 神经网络的美股股指价格趋势预测

模型的研究 [硕士学位论文]. 北京: 首都经济贸易大学, 2016.
16 张冬雯, 赵琪, 许云峰, 等. 基于长短期记忆神经网络模型的空气质量预测. 河北科技大学学报, 2020, 41(1): 67–75. [doi: 10.7535/hbkd.2020yx01008]
17 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.
18 刘晶晶, 庄红, 铁治欣, 等. K-Similarity 降噪的 LSTM 神经网络水质多因子预测模型. 计算机系统应用, 2019, 28(2): 226–232. [doi: 10.15888/j.cnki.csa.006756]
19 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.
20 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.
21 Wei SK, Yang H, Song JX, *et al*. A wavelet-neural network hybrid modelling approach for estimating and predicting river monthly flows. Hydrological Sciences Journal, 2013, 58(2): 374–389.
22 Wei SK, Zuo DP, Song JX. Improving prediction accuracy of river discharge time series using a wavelet-NAR artificial neural network. Journal of Hydroinformatics, 2012, 14(4): 974–991.
23 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]
24 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]
25 梁百川. 小波变换理论及应用. 舰船电子对抗, 1998, (5): 1–10.
26 樊顺厚. 广义 Hadamard 积. 天津纺织工学院学报, 2000, 19(4): 6–7.
27 薛长峰. 矩阵的 Hadamard 乘积. 盐城工学院学报 (自然科学版), 2003, 16(3): 38–39, 52.
28 国家环境保护总局, 国家质量监督检验检疫总局. GB 3838-2002 地表水环境质量标准. 北京: 中国环境科学研究院, 2002.
29 樊计昌, 刘明军, 王夫运, 等. 浅析小波最大分解层. 科技导报, 2008, 26(10): 40–42. [doi: 10.3321/j.issn:1000-7857.2008.10.012]