

基于季节性分解与长短期记忆网络的水质动态预警

许博文¹, 毕敬¹, 苑海涛², 王功明¹, 乔俊飞¹

(1. 北京工业大学信息学部, 北京 100124;

2. 北京航空航天大学自动化科学与电气工程学院, 北京 100191)

摘要: 随着地表水水质恶化日益严重, 有效的水质预警预测技术对于水资源的可持续发展与应急响应机制实施至关重要。长短期记忆网络在水质时间序列预测问题中被广泛使用, 但是仅使用长短期记忆网络进行水质预测并不能解决各种复杂因素造成的水质序列的不规则波动问题。为了解决该问题, 提出一种数据驱动的水质预测混合模型, 该模型将基于局部加权回归散点平滑(Loess)的季节与趋势分解(STL)算法与基于编解码的长短期记忆网络(LSTM-ED)结合。首先通过 STL 的加法模型将水质时间序列分解为 3 个子序列, 然后利用多元 LSTM-ED 神经网络对子序列进行预测, 通过叠加将数据恢复为实际值, 最后通过季节性分段的拉依达准则进一步判断水质是否存在异常并做出预警。实验结果表明, 与单一的 LSTM、LSTM-ED 以及基于序列分解的 LSTM-ED 模型相比, 所提出的模型能显著地提高水质时间序列预测的精度和可靠性, 并为水质动态预警提供有效的数据支持。

关键词: 季节性分解; 长短期记忆网络; 水质动态预警; 异常检测

中图分类号: TP399

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202145

Dynamic water quality warning with seasonal decomposition and long short-term memory network

XU Bowen¹, BI Jing¹, YUAN Haitao², WANG Gongming¹, QIAO Junfei¹

1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

2. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

Abstract: Surface water quality is increasingly deteriorated in recent years, and therefore, high-quality early warning and prediction of water quality are essential for sustainability of water resources and emergency response mechanisms. Long short-term memory (LSTM) network is widely applied in the existing literature on the prediction of water quality time series. However, only applying LSTM for the prediction of water quality time series cannot well address irregular fluctuations in the water quality series caused by multiple complex factors. To solve this problem, a data-driven prediction model for the water quality time series was proposed, named STL-LSTM-ED, which was composed of seasonal-trend decomposition using locally weighted scatterplot smoothing (STL) and LSTM based on encoder-decoder (LSTM-ED). Compared with several typical models of LSTM, LSTM-ED, and a sequence decomposition method based on LSTM, the proposed STL-LSTM-ED can significantly improve the prediction accuracy and reliability of the water quality time series, and also provide the effective data support for dynamic warning of water quality.

Key words: seasonal decomposition, long short-term memory network, dynamic warning of water quality, anomaly detection

收稿日期: 2021-01-14; 修回日期: 2021-03-12

通信作者: 毕敬, bijing@bjut.edu.cn

基金项目: 水体污染控制与治理科技重大专项 (No.2018ZX07111005); 国家自然科学基金资助项目 (No.61802015, No.62073005, No.62173013)

Foundation Items: The Major Science and Technology Program for Water Pollution Control and Treatment of China (No.2018ZX07111005), The National Natural Science Foundation of China (No.61802015, No. 62073005, No.62173013)

1 引言

水作为地球上珍贵的资源,它决定着人类和动植物的生存,水对于生态系统和社会都至关重要。然而,随着工业化进程的不断加快,工业废水、农业施药和生活污水等水质污染事件引发的水质异常现象频繁发生。为了实现水资源的可持续利用,加强水环境监管工作俨然成为我国政府迫在眉睫的重要任务。水环境监管工作可以有效解决水资源短缺以及水质污染等问题,其中水质预测及水质动态预警是水环境监管工作的重要研究内容之一。据世界卫生组织的报告显示,到2025年,全球将有半数人口居住在水资源紧张的地区。受污染的水可以传播多种疾病,如腹泻、霍乱、痢疾、伤寒和脊髓灰质炎等。受污染的饮用水估计每年造成超过48.5万例腹泻死亡。水质异常现象每年都在急剧增加,因此水质监测和水质动态预警刻不容缓。

物联网技术的发展以及越来越多的水质监测传感器的部署,使得获取高频的水质监测数据变得可行。水质监测和预警系统可降低水资源被污染的风险,水质指标的急剧变化往往暗示着水质污染的发生,及时对水质异常进行预警对于水资源的管理至关重要。通过大数据分析技术可以准确地预测水质的变化,反映实际的水质情况,从而为相关部门的联合调控与应急预警提供重要依据。由于采集的水质数据是时间序列数据,因此有效地利用大量与水质时间序列相关的参数之间的关系,可以提高水质预测及动态预警的准确性^[1]。关于时间序列数据预测的方法有很多,但是应用在水质预测领域中的方法并不多,这也成为研究的难点之一。同时,为了保证预警的可靠性与时效性,需要模型具有较强的预知能力,这体现在预测模型里即需要有较大的预测步长。预测步长越大,预留给管控者的调控时间越多。研究发现,受季节性影响,不同季节的水质指标值合理波动范围具有较大的差异,为此异常检测的手段应考虑季节性的影响,进行局部的异常分析。水质监测系统是一个非线性动态系统^[2],为了有效控制和治理水质污染事件,实现水资源的可持续利用,本文提出了基于季节性分解与长短期记忆(long short-term memory, LSTM)网络的水质动态预警方法(STL-LSTM-ED),从而获取更准确的预测结果。在此基础上,探索了季节性对异常检测方法的影响,采用按季节性分段的滑动窗口方式

对预测结果进行异常检测。本文提出的方法可以提供更加准确的预测效果,使得预警频率更加合理,从而大幅度降低错误预警导致的人力物力等资源的浪费。

综上所述,本文的主要贡献有以下3个方面:

(1) 本文给出LSTM与基于编解码的长短期记忆网络(long short-term memory based on encoder-decoder, LSTM-ED)在水质预测上的效果;

(2) 本文探索将基于局部加权回归散点平滑(locally weighted scatterplot smoothing, Loess)的季节与趋势分解(seasonal-trend decomposition using Loess, STL)模型与经验模态分解(empirical mode decomposition, EMD)模型应用于LSTM-ED模型时对水质预测效果的影响,最终确定提出的STL-LSTM-ED模型充分结合了时间序列的周期性,从而有效地提高水质预测的精度;

(3) 在此基础上,本文提出按季节性分段的滑动窗口方式来计算局部预警阈值,从而实现动态预警,并有效地避免因季节性变化导致的预警。

本文的结构如下:第2节讨论了与水质预测及与动态预警相关的研究;第3节介绍了水质动态预警方法;第4节介绍了一系列基于LSTM预测模型的对比实验和结果分析以及季节性对水质异常检测的影响;第5节对全文进行了总结并讨论了如何进一步改进该模型。

2 相关工作

目前,传统的时间序列预测方法可以利用数据之间的线性关系来捕捉其线性动态发展趋势。传统方法以其简单易行的优势在时间序列预测领域得到了广泛的应用。例如,Oliveira P J等人^[3]采用自回归综合移动平均(autoregressive integrated moving average, ARIMA)模型预测区域需水量,实验证明,该模型可以获得较高的预测精度。但是,由于ARIMA模型非常依赖时间序列数据的稳定性,不适用于本文具有复杂非线性关系的数据。Parmar K S等人^[4]采用统计分析和时间序列预测模型对水质分级和评价管理进行了研究。Cao Q等人^[5]提出了基于最小二乘支持向量机(least squares support vector machine, LS-SVM)和马尔可夫链的预测模型。水环境参数具有随机性和复杂性等特点,且需要大量的水质参数和环境参数,因此难以建立周期较长的水质模型。

单一预测模型在某些数据集上的预测精度有限^[6], 需要做一些尝试来提高预测的精准性。多个模型结合的方式可以得到更高的预测精度^[7]。因此, 一些研究者提出了混合预测模型。张建晋等人^[8]构建了面向季节性时空数据的预测式循环网络, 并应用在城市计算中, 用于季节性时间序列模型的预测。Tseng F M 等人^[9]结合了季节性自回归综合移动平均 (seasonal and autoregressive integrated moving average, SARIMA) 和反向传播神经网络的混合模型来预测季节性时间序列数据。Zhou Z J 等人^[10]提出了一种基于灰色和 ARIMA 的混合建模和预测方法。Khashei M 等人^[11]使用 ARIMA 与人工神经网络 (artificial neural network, ANN) 的组合模型对时间序列数据进行了预测。Zhang L 等人^[12]探索了小波神经网络 (wavelet neural network, WNN), 将小波变换和神经网络结合进行短期水质预测。Shi B 等人^[13]将小波分析方法和人工神经网络模型结合, 进行高频地表水水质预测。

水质数据分解可以消除周期性因素的影响, 反映水质的真实规律和趋势, 提高水质预测的准确性。目前存在很多模型, 如基于小波分解的预测模型与支持向量机^[14]、基于经验模态分解的长短期记忆网络预测模型^[15]、结合 ARIMA 模型和小波分解的混合模型^[16]、基于长短期记忆网络模型和高斯混合模型^[17]、基于变分模式分解和 LSTM 的模型^[18]等。上述分解模型可以对序列进行分解并提高预测精度, 但应用以上模型进行水质预测时, 并没有充分利用这些时间序列数据的特性。

而 STL 模型充分考虑了水质数据的特征, Chen D W 等人^[19]将 STL 应用于同样具有季节性特征的地铁客流量数据的预测, 将其分解为季节、趋势和残差分量。然而, 该工作只预测了趋势子序列, 并

没有充分利用分解后的全部特征。

3 水质动态预警方法

本文工作提出的水质动态预警方法将 STL 的加法模式与 LSTM-ED 神经网络模型结合, 从而提高预测性能, STL-LSTM-ED 模型结合异常检测模型的结构如图 1 所示。该混合算法主要包括以下 3 个步骤。

步骤一: 水质序列分解。使用 STL 算法将水质时间序列分解为趋势、季节周期、残差 3 个子序列, 将子序列合并, 从而得到一个三维时间序列。

步骤二: 水质序列预测。将步骤一中的三维时间序列输入多元 LSTM-ED 模型中进行训练, 将得到的三维预测结果进行矩阵相加, 得到一维时间序列。然后计算该序列与步骤一中待分解的时间序列的损失函数值, 通过降低该损失函数值得到最优的超参数。图 1 中 c 表示融合了编码器全部信息的特征输出。

步骤三: 水质序列预警。通过按季节性分段的滑动窗口方式, 利用拉依达准则将步骤二中的预测结果进行局部异常检测, 分析得出预警点。

3.1 基于 Loess 的季节与趋势分解

STL 是由 Cleveland R B 等人^[20]提出的时间序列分解算法。它将一个包含 N 个点的周期性时间序列 Y , 分解成趋势分量 T 、季节周期分量 S 和残差分量 R 。

$$Y_v = T_v + S_v + R_v, \quad v=1, \dots, N \quad (1)$$

趋势分量表示时间序列低频率和长时间的变化; 季节周期分量表示时间序列周期性频率的变化; 残差分量则表示原时间序列除去趋势分量和季节周期分量后剩余的值。STL 算法基于局部加权回归散点平滑算法进行周期子序列的平滑, 并通过内、外循环进行趋势拟合与季节周期分量的计算以

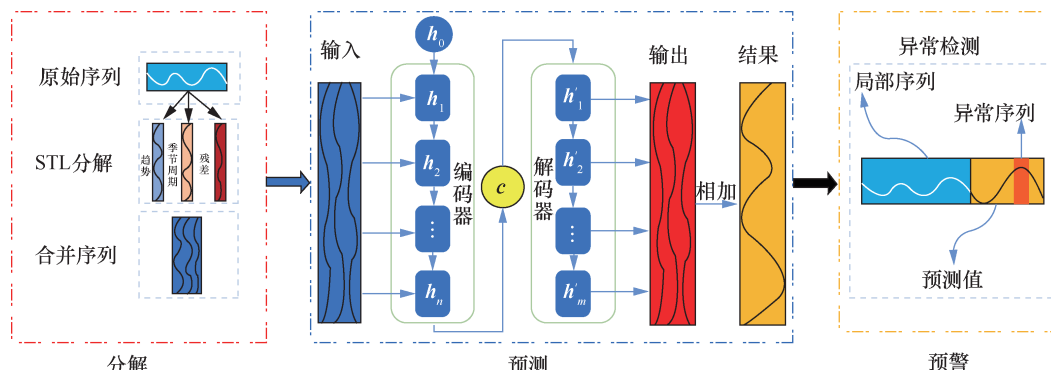


图1 STL-LSTM-ED模型结合异常检测模型的结构

及鲁棒性权重的更新。

Loess 是对二维散点图进行平滑的常用方法。内循环部分主要负责对时间序列的周期子序列进行平滑，对趋势分量和季节周期分量进行更新；外循环部分主要负责计算鲁棒性权重，避免数据中的离群值对计算结果产生影响。

3.2 长短期记忆网络

LSTM^[21]是一种循环模型，已被用于许多序列学习任务，如手写识别、语音识别和情感分析等。LSTM 由输入门、输出门和遗忘门组成。图2展示了由门单元组成的 LSTM 单元内的信息流。LSTM 的计算步骤如式(2)~式(10)所示。

步骤一：将前一时刻隐藏层的输出 H_{t-1} 和当前状态的输入 X_t 合并。

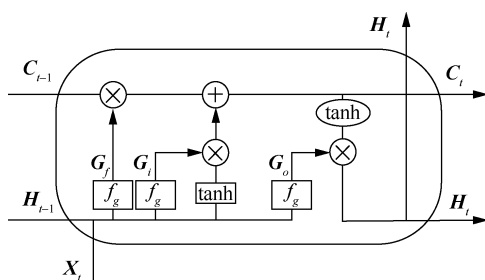


图2 LSTM单元内的信息流

$$I_t = H_{t-1} + X_t \quad (2)$$

步骤二：计算第一个 tanh 的输出向量 Y_t 。

$$Y_t = f_c(W_c \cdot I_t + b_c) \quad (3)$$

其中， f_c 为激活函数，一般采用 tanh 函数， W_c 为连接权重值， b_c 为偏置量。

步骤三：计算对应单元输入门、遗忘门和输出门的输出向量。

$$G_i = f_g(W_i \cdot I_t + b_i) \quad (4)$$

$$G_f = f_g(W_f \cdot I_t + b_f) \quad (5)$$

$$G_o = f_g(W_o \cdot I_t + b_o) \quad (6)$$

其中， G_i 、 G_f 和 G_o 分别是输入门、遗忘门和输出门的输出向量。权重 W_i 、 W_f 、 W_o 和偏置 b_i 、 b_f 、 b_o 是 3 个门单元对应的参数。 f_g 为门单元的激活函数 Sigmoid，其输出值在 0 和 1 之间。

步骤四：计算长期记忆的新单元状态向量 C_t 。

$$Y'_t = G_i \cdot Y_t \quad (7)$$

$$C'_t = G_f \cdot C_{t-1} \quad (8)$$

$$C_t = Y'_t + C'_t \quad (9)$$

其中， Y'_t 为 G_i 与式(3)中输出向量 Y_t 的逐元素乘积，即哈达玛积； C'_t 为上一时刻细胞状态向量 C_{t-1} 与遗忘门 Sigmoid 输出的哈达玛积；将 Y'_t 与 C'_t 相加得到 C_t ， C_t 表示 LSTM 单元在 t 时刻的细胞状态，是长期记忆的新细胞状态载体，当被重新使用时会反馈给 LSTM 单元。在这一步骤中，长期记忆的细胞状态会获得新的信息，但会忘记一些旧信息。

步骤五：计算 LSTM 单元的输出向量 H_t 。

$$H_t = f_c(C_t) \cdot G_o \quad (10)$$

其中， H_t 为 LSTM 单元的输出向量， f_c 为与步骤二中相同的激活函数。激活函数可以稳定 LSTM 单元多次重用后的输出值。输出门的输出向量 G_o 可以控制 LSTM 单元是否应该产生输出。此外，在此步骤中，单元状态向量 C_t 不受 f_c 的影响，因此可以很容易地保留 LSTM 单元的原始输出 Y'_t ，以便下次重用。

3.3 基于编解码的长短期记忆网络

LSTM-ED 模型^[22]近年来常被用于序列到序列的学习任务，如机器翻译^[23]以及 LSTM 变种的应用^[24]。LSTM-ED 模型如图3所示，其工作原理如下：使用基于 LSTM 的编码器将多变量输入序列映射为固定维向量表示；译码器是另一个 LSTM 网络，它使用该向量表示来产生目标序列。图3中， $X = \{x^1, x^2, \dots, x^{n_x}\}$ 是输入序列， C_t 是编码器在第 t 步的中间状态，其中， $C_t \in \mathbb{R}^m$ ， m 是编码器的神经元的数量；解码器将 C_t 解码成目标序列 $Y = \{y^1, y^2, \dots, y^{n_y}\}$ 。

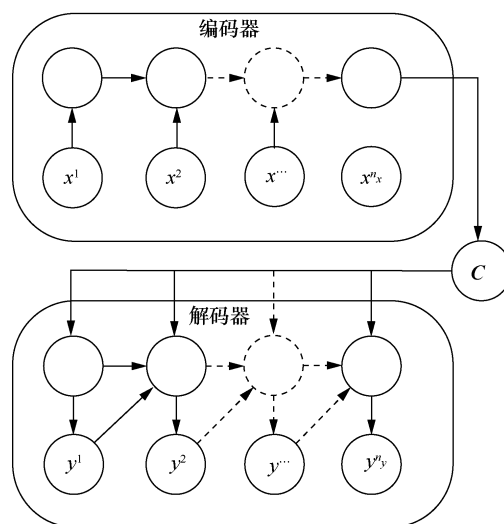


图3 LSTM-ED模型

3.4 局部异常检测

考虑到水质变化受季节影响,本文提出一种基于季节性分解与长短期记忆网络的水质动态预警方法,通过计算选定季节范围内的水质时间序列进行异常检测,即局部异常检测。基于此,有效地降低了季节性导致的水质变化趋势异常波动。采用基于拉依达准则的离群点检测方法,通过滑动窗口方式计算局部均值 μ 与方差 σ^2 ,在正态分布的假设下,区域 $\mu+3\sigma$ 包含了99.7%的数据,如果某个值与均值 μ 的距离超过了 3σ ,这个值将被定义为异常值。其中设置滑动窗口的长度为 m ,截取的序列为 $X(m)=\{x_1, x_2, \dots, x_m\}$ 。 μ 与 σ^2 的定义如式(11)和式(12)所示。异常值的定义方式如下:如果 $x > \mu + 3\sigma$ 或 $x < \mu - 3\sigma$,则该值为异常值;否则该值为正常值。

$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)} \quad (11)$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)^2 \quad (12)$$

4 仿真与分析

4.1 数据集及实验设置

本文利用美国地质勘探局(United States Geological Survey, USGS)国家水信息系统(national

water information system, NWIS)发布的加利福尼亚州纽曼圣华金河支流默塞德河水体的水环境数据。时间跨度为2012年5月至2020年8月,采样周期为15~60 min。数据集共约280 000个数据,训练集、验证集和测试集的比例为8:1:1。训练集用于拟合模型,验证集用于在训练过程中对模型进行校正,测试集用于评估训练模型的预测精度和鲁棒性。本文利用水质指标值做预测。pH值、溶解氧与水温3个指标的原始数据如图4所示,从整体趋势上看,呈现较强的周期性。由溶解氧与水温的物理特性可知,水温越高,溶解氧越低,二者具有一定的相关性。

4.2 评价指标

为了验证和评价所提出的STL-LSTM-ED模型的预测能力,本文利用3种常用的误差评价指标来计算预测值与实际值的差距。评价指标包括均方误差(mean square error, MSE)、均方根误差(root mean square error, RMSE)、平均绝对百分比误差(mean absolute percentage error, MAPE)。各评价指标的计算式如下:

$$MSE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (14)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (15)$$

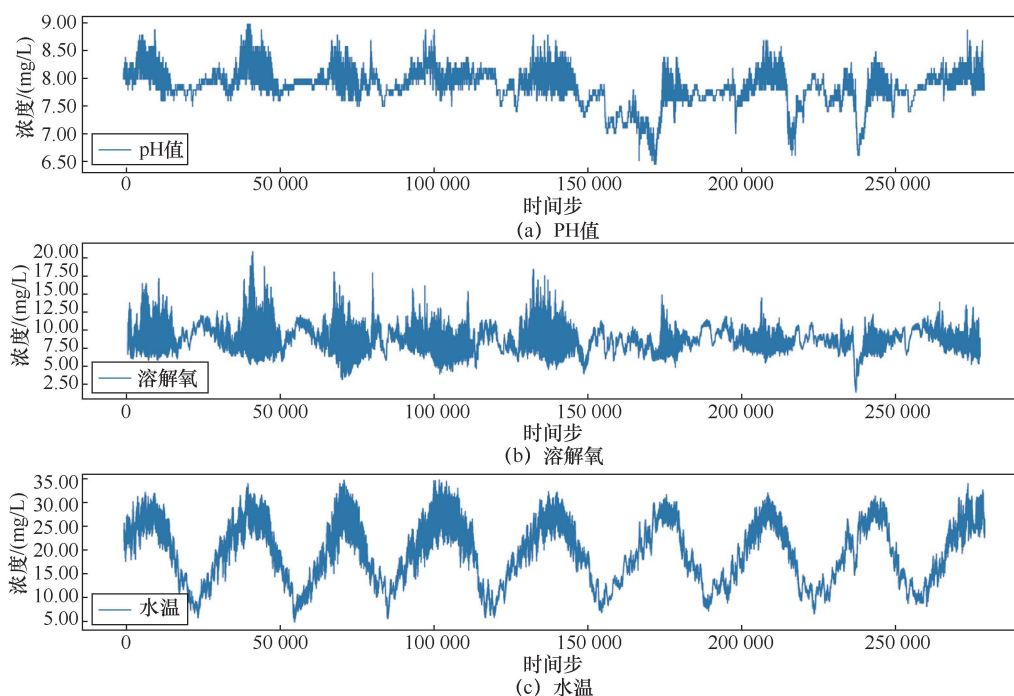


图4 pH值、溶解氧与水温3个指标的原始数据

其中, n 是样本个数, $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$ 是预测结果序列, $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ 是实际序列。

同时, 为了验证季节性分段在异常检测上的效果, 引入 F1 分数 (F1-Score) 对模型进行评估。F1-Score 的计算过程如式 (16)~式 (18) 所示, 评价标准参数说明见表 1。其中, Precision 表示精准率, Recall 表示召回率。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (16)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (17)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

表 1 评价标准参数说明

真实情况	预测为正例	预测为反例
正例	TP	FN
反例	FP	TN

4.3 基于 LSTM-ED 的预测模型

LSTM-ED 模型通常被用于解决 Seq2Seq (sequence to sequence) 问题。LSTM-ED 模型可以实现序列到序列的输出, 即直接输出多步的预测结果。而在采用递归滚动预测方法的 LSTM 模型中, 使用预测结果进行下一步的预测, 因此导致了误差累计。本文通过多步预测实验将 LSTM 与 LSTM-ED 模型进行了对比。LSTM 与 LSTM-ED 模型的多步预测精度见表 2。从表 2 可以看出, LSTM-ED 模型均取得了较好的效果。LSTM 在溶解氧指标上的预测效果如图 5 所示, LSTM-ED 在溶解氧指标上的预测效果如图 6 所示。本文中 LSTM 神经网络的参数设定保持一致。其中 LSTM 的隐藏层层数为 2, 隐藏层维度为 32, 损失函数使用均方误差, 随机失活率为 0.2。此外, 本文使用了 Adam 优化器。

4.4 基于序列分解的预测模型

EMD 是 Huang N E 等人^[25]提出的一种自适应处理技术, 常被用于分解非线性非平稳时间序列。信号的自适应体现在它根据时间尺度特征对信号进行分解。EMD 不需要预先设置基函数或分解尺度, 可以有效减少人为干预。EMD 的基本原理是将原始信号自适应地分解为一系列振荡函数, 其中包括有限个具有不同特征尺度的固有模态函数 (intrinsic mode function, IMF) 和一个残差序列。固有模态函数满足以下两个条件:

表 2 LSTM 与 LSTM-ED 模型的多步预测精度

步长	LSTM			LSTM-ED		
	MAPE	MSE	RMSE	MAPE	MSE	RMSE
1	10.71	0.91	1.15	6.92	0.56	0.79
2	10.96	0.94	1.18	7.07	0.57	0.8
3	11.34	0.97	1.22	7.21	0.59	0.81
4	11.77	1.01	1.27	7.28	0.62	0.83
5	12.2	1.05	1.32	7.35	0.64	0.83
6	12.63	1.09	1.37	7.55	0.69	0.86
7	13.02	1.13	1.42	7.62	0.71	0.87
8	13.37	1.16	1.46	7.69	0.73	0.89
9	13.69	1.19	1.5	8.17	0.76	0.92
10	13.97	1.22	1.53	8.49	0.78	0.93
11	14.2	1.24	1.56	8.91	0.83	0.95
12	14.41	1.26	1.59	9.23	0.87	0.99
13	14.58	1.28	1.61	9.49	0.9	1.02
14	14.73	1.29	1.63	9.88	0.92	1.05
15	14.85	1.3	1.65	10.32	0.94	1.08
16	14.95	1.31	1.66	10.77	0.97	1.1
17	15.04	1.32	1.67	11.51	0.99	1.13
18	15.1	1.33	1.68	11.89	1.04	1.16
19	15.15	1.33	1.69	12.21	1.13	1.19
20	15.19	1.34	1.69	12.65	1.19	1.21
平均	13.59	1.18	1.49	9.11	0.82	0.97

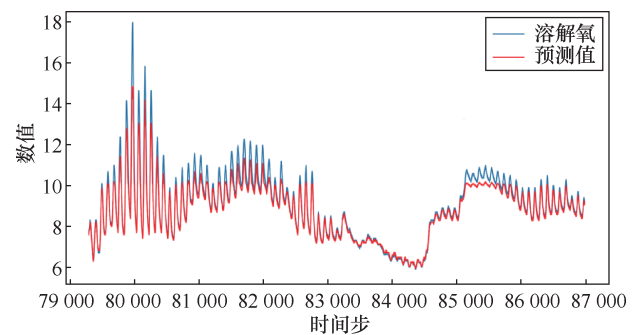


图 5 LSTM 在溶解氧指标上的预测效果

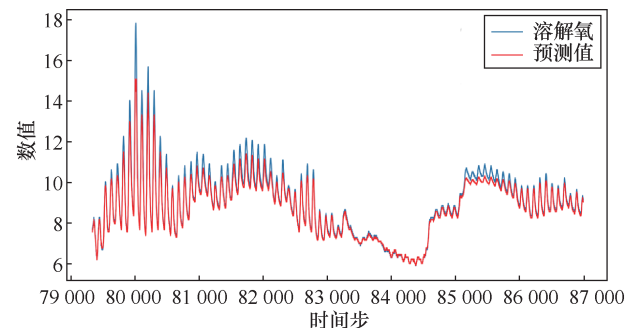


图 6 LSTM-ED 在溶解氧指标上的预测效果

(1) 在整个数据长度内, 局部极值点的个数和过零点的个数必须相等或最多相差 1;

(2) 在任何时刻, 局部极大值定义的包络和局部极小值定义的包络的均值都为 0。

在数据集分割上, 先进行 EMD 分解会导致数据泄露问题, 后进行 EMD 分解无法保证训练集与测试集的子序列个数相同。因此, 在对比实验中, EMD 训练集分解了 14 个子序列, 验证集分解了 8 个子序列, 测试集分解了 11 个子序列, 将多余的子序列合并为残差序列。最终的数据集为 8 个子序列, 然后输入 LSTM-ED 网络进行训练。在上述实验基础上, 对比了基于序列分解的 LSTM-ED 模型在水质时间序列预测上的效果。EMD 和 STL 对水质时间序列的分解结果分别如图 7、图 8 所示。基于序列分解的 LSTM-ED 模型的多步预测精度见表 3。从表 3 可以看出, STL-LSTM-ED 模型取得了较好的预测效果。EMD-LSTM-ED 在溶解氧指标上的预测效果如图 9 所示, STL-LSTM-ED 在溶解氧指标上的预测效果如图 10 所示。

4.5 基于拉依达准则的水质异常检测模型

在拉依达准则中, 可以根据实际数据来确定具体将几个标准差作为异常判别的阈值, 因此可以将该准则扩展为 $k\sigma$ 法。为了验证季节性分段异常检测的效果, 按照先验知识对上述水质数据集中的异常进行人工标记。设定周期长度参数 Period, 即

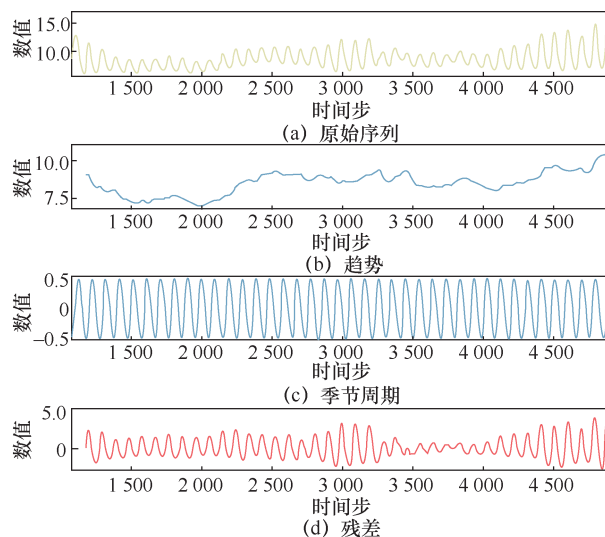


图 8 STL 对水质时间序列的分解结果

需要输入的数据的长度。在实验过程中, 首先验证了不同 k 值下的异常检测效果, 实验结果见表 4。当 $k=4$ 时, F1-Score 取得最大值, 为 0.516, 模型评估指标 AUC 为 0.67。拉依达准则方法中季节分段长度与 k 的 F1-Score 值热力图如图 11 所示。

在该水质数据上, 当 $k=5$, Period=11 000 时, F1-Score 达到最大值, 为 0.594, 此时召回率为 0.462, 精准率为 0.833, AUC 为 0.73。实验表明, 增加了季节性分段异常检测的效果优于整体数据的异常检测, 因此, 将季节性应用于异常检测具有捕捉局部异常的特性。

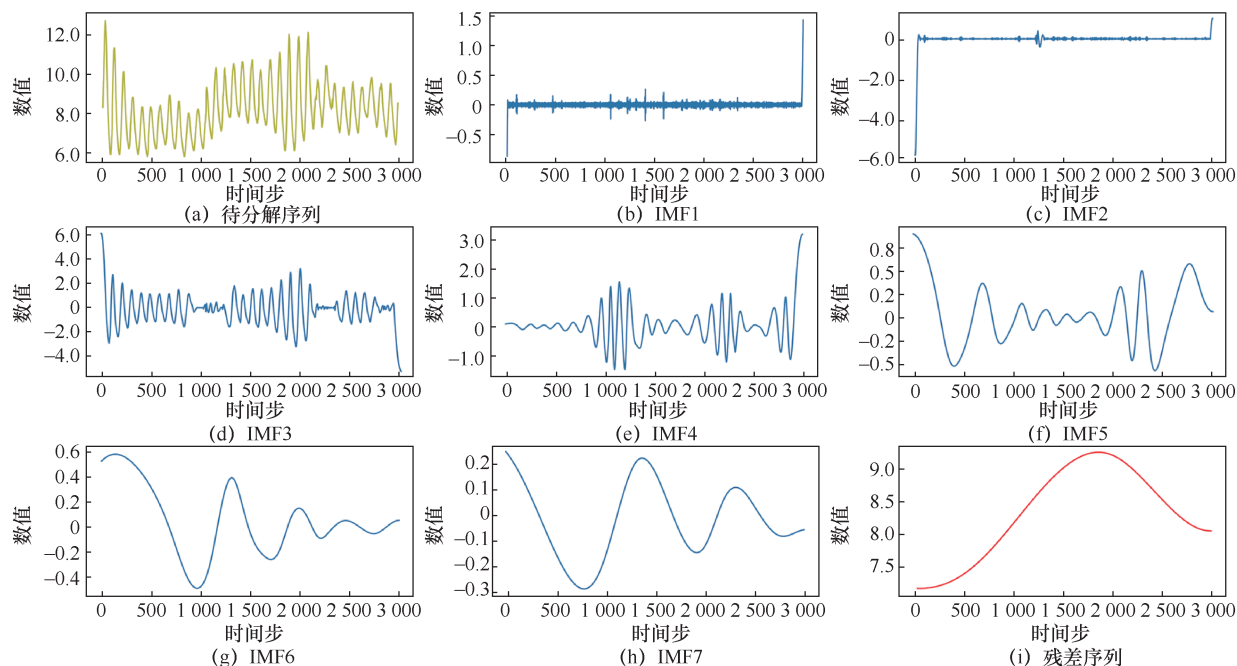


图 7 EMD 对水质时间序列的分解结果

表3 基于序列分解的 LSTM-ED 模型的多步预测精度

步长	EMD-LSTM-ED			STL-LSTM-ED		
	MAPE	MSE	RMSE	MAPE	MSE	RMSE
1	3.21	0.28	0.46	3.73	0.33	0.43
2	3.4	0.3	0.47	3.97	0.35	0.45
3	3.69	0.32	0.5	4.22	0.37	0.48
4	4.05	0.35	0.53	4.48	0.39	0.51
5	4.44	0.39	0.57	4.75	0.41	0.54
6	4.86	0.43	0.62	5.01	0.44	0.57
7	5.29	0.47	0.66	5.27	0.46	0.6
8	5.73	0.5	0.71	5.53	0.48	0.63
9	6.16	0.54	0.76	5.79	0.5	0.65
10	6.59	0.58	0.8	6.04	0.52	0.68
11	7.02	0.62	0.85	6.28	0.55	0.71
12	7.43	0.66	0.9	6.52	0.57	0.74
13	7.83	0.7	0.94	6.74	0.59	0.76
14	8.21	0.73	0.99	6.96	0.6	0.79
15	8.58	0.76	1.03	7.16	0.62	0.81
16	8.93	0.8	1.07	7.36	0.64	0.84
17	9.26	0.83	1.11	7.73	0.67	0.87
18	9.57	0.86	1.15	7.97	0.7	0.9
19	9.86	0.88	1.18	8.22	0.73	0.93
20	10.13	0.91	1.22	8.48	0.76	0.96
平均	6.71	0.60	0.83	5.31	0.46	0.60

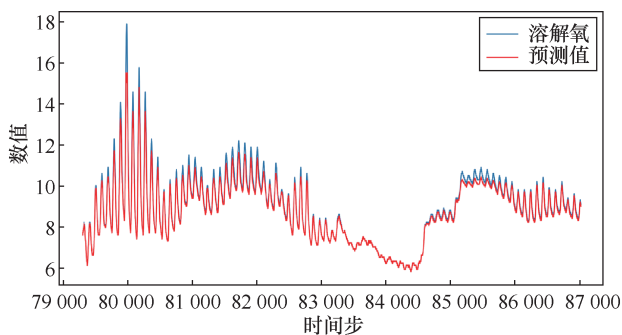


图9 EMD-LSTM-ED 在溶解氧指标上的预测效果

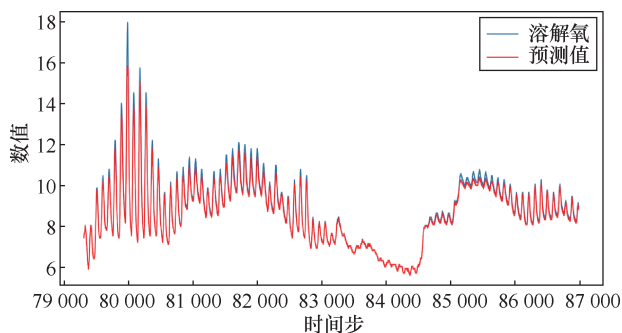
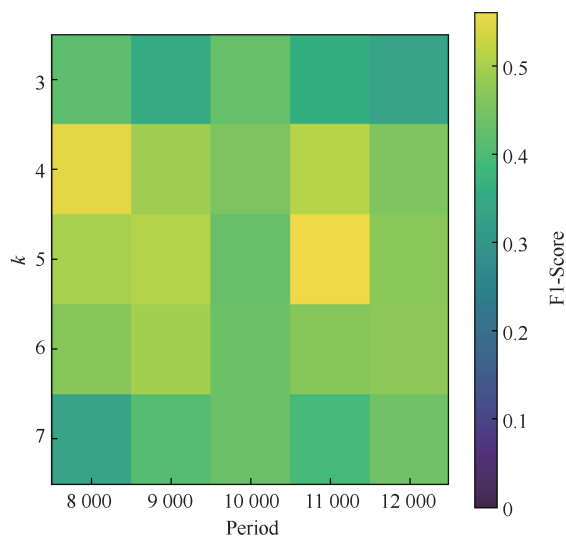


图10 STL-LSTM-ED 在溶解氧指标上的预测效果

表4 不同 k 值对异常检测效果的影响

k 值	F1-Score	AUC
3	0.494	0.62
4	0.516	0.67
5	0.501	0.64
6	0.472	0.60
7	0.463	0.57
8	0.455	0.56

图11 拉依达准则方法中季节分段长度与 k 的 F1-Score 值热力图

5 结束语

预测是水质预警的重要前提, 预测的可靠性与精准性能够降低频繁预警导致的人力物力的开销, 同时避免了低灵敏度预警导致的管控滞后。本文提出了一种基于季节性分解与长短期记忆网络的水质动态预警方法 STL-LSTM-ED, 该方法将 STL 分解模型与基于编解码的长短期记忆网络结合, 用来预测水质时间序列。该方法采用 STL 进行季节性分解, 使得神经网络更加专注学习复杂的特征, 因此采用 LSTM-ED 模型比单一的 LSTM 有更好的预测精度。同时本文对比了 EMD 模型。相比于分解出 3 个子序列的 STL, EMD 会分解出更多的子序列, 这会导致预测误差叠加, 使得预测的准确率下降。研究结果表明, STL-LSTM-ED 模型在各项指标中均优于其他模型。在未来工作中, 笔者将进一步探索该模型的泛化能力, 并将其应用于不同的水质监测站点的数据, 从而进一步优化模型并提高实时预测与实时预警分析的能力。

参考文献:

- [1] JIN T, CAI S B, JIANG D X, et al. A data-driven model for real-time water quality prediction and early warning by an integration method[J]. *Environmental Science and Pollution Research International*, 2019, 26(29): 30374-30385.
- [2] 权利敏, 杨翠丽, 乔俊飞. 基于 CFNN 的污水处理过程溶解氧浓度在线控制[J]. *智能科学与技术学报*, 2020, 2(3): 261-267.
QUAN L M, YANG C L, QIAO J F. CFNN-based online control for dissolved oxygen concentration of wastewater treatment processes[J]. *Chinese Journal of Intelligent Science and Technology*, 2020, 2(3): 261-267.
- [3] OLIVEIRA P J, STEFFEN J L, CHEUNG P. Parameter estimation of seasonal ARIMA models for water demand forecasting using the harmony search algorithm[J]. *Procedia Engineering*, 2017, 186: 177-185.
- [4] PARMAR K S, BHARDWAJ R. Water quality management using statistical analysis and time-series prediction model[J]. *Applied Water Science*, 2014, 4(4): 425-434.
- [5] CAO Q, LIU B X, CHENG X. Notice of retraction: forecasting model based on LS-SVM and Markov chain for the water quality of Poyang Lake[C]//*Proceedings of 2011 7th International Conference on Natural Computation*. Piscataway: IEEE Press, 2011: 2070-2073.
- [6] LE X H, HO H V, LEE G, et al. Application of long short-term memory (LSTM) neural network for flood forecasting[J]. *Water*, 2019, 11(7): 1387.
- [7] AYADI A, GHORBEL O, OBEID A M, et al. Outlier detection approaches for wireless sensor networks: a survey[J]. *Computer Networks*, 2017, 129: 319-333.
- [8] 张建晋, 王韞博, 龙明盛, 等. 面向季节性时空数据的预测式循环网络及其在城市计算中的应用[J]. *计算机学报*, 2020, 43(2): 286-302.
ZHANG J J, WANG Y B, LONG M S, et al. Predictive recurrent networks for seasonal spatiotemporal data with applications to urban computing[J]. *Chinese Journal of Computers*, 2020, 43(2): 286-302.
- [9] TSENG F M, YU H C, TZENG G H. Combining neural network model with seasonal time series ARIMA model[J]. *Technological Forecasting and Social Change*, 2002, 69(1): 71-87.
- [10] ZHOU Z J, HU C H. An effective hybrid approach based on grey and ARMA for forecasting gyro drift[J]. *Chaos, Solitons & Fractals*, 2008, 35(3): 525-529.
- [11] KHASHEI M, BIJARI M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting[J]. *Applied Soft Computing*, 2011, 11(2): 2664-2675.
- [12] ZHANG L, ZOU Z H, SHAN W. Development of a method for comprehensive water quality forecasting and its application in Miyun reservoir of Beijing, China[J]. *Journal of Environmental Sciences*, 2017, 56: 240-246.
- [13] SHI B, WANG P, JIANG J P, et al. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies[J]. *Science of the Total Environment*, 2018, 610/611: 1390-1399.
- [14] SUN Y X, LENG B, GUAN W. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system[J]. *Neurocomputing*, 2015, 166: 109-121.
- [15] CHEN L, CHI Y G, GUAN Y Y, et al. A hybrid attention-based EMD-LSTM model for financial time series prediction[C]//*Proceedings of 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. Piscataway: IEEE Press, 2019: 113-118.
- [16] CONEJO A J, PLAZAS M A, ESPINOLA R, et al. Day-ahead electricity price forecasting using the wavelet transform and ARIMA models[J]. *IEEE Transactions on Power Systems*, 2005, 20(2): 1035-1042.
- [17] WU D, JIANG Z K, XIE X F, et al. LSTM learning with Bayesian and Gaussian processing for anomaly detection in industrial IoT[J]. *IEEE Transactions on Industrial Informatics*, 2020, 16(8): 5244-5253.
- [18] NIU H L, XU K L, WANG W Q. A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network[J]. *Applied Intelligence*, 2020, 50(12): 4296-4309.
- [19] CHEN D W, ZHANG J H, JIANG S X. Forecasting the short-term metro ridership with seasonal and trend decomposition using loess and LSTM neural networks[J]. *IEEE Access*, 2020, 8: 91181-91187.
- [20] CLEVELAND R B, CLEVELAND W S, MCRAE J E, et al. STL: a seasonal-trend decomposition [J]. *Journal of official statistics*, 1990, 6(1): 3-73.
- [21] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- [22] HABLER E, SHABTAI A. Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages[J]. *Computers & Security*, 2018, 78: 155-173.
- [23] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. *arXiv preprint*, 2014, arXiv:1409.0473.
- [24] 倪维健, 孙宇健, 刘彤, 等. 基于注意力双向循环神经网络的业务流程剩余时间预测方法[J]. *计算机集成制造系统*, 2020, 26(6): 1564-1572.
NI W J, SUN Y J, LIU T, et al. Business process remaining time prediction using bidirectional recurrent neural networks with attention[J]. *Computer Integrated Manufacturing Systems*, 2020, 26(6): 1564-1572.

- [25] HUANG N E, SHEN Z, LONG S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903-995.

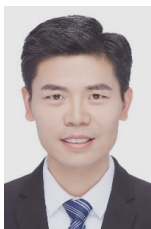
[作者简介]



许博文（1996—），男，北京工业大学信息学部硕士生，主要研究方向为异常检测、水质预警等。



苑海涛（1986—），男，博士，北京航空航天大学自动化科学与电气工程学院副教授，IEEE 高级会员，现任中国体视学学会理事。主要研究方向为云计算、边缘计算、数据中心、深度学习、智能优化、时间序列特征建模与预测等。



王功明（1987—），男，博士，北京工业大学信息学部助理研究员，主要研究方向为污水处理过程特征建模与水质预测、神经自组织结构设计与优化、离散事件动态系统安全策略优化等。



毕敬（1979—），女，博士，北京工业大学信息学部副教授、博士生导师，主要研究方向为时空大数据特征建模与分析、计算智能、深度学习、数据中心节能等。



乔俊飞（1968—），男，博士，北京工业大学信息学部教授、博士生导师，主要研究方向为智能计算、自组织控制以及污水处理过程优化运行控制等。