DOI:10.20040/j.cnki.1000-7709.2023.20222020

基于进化算法优化的CNN-EA-ConvLSTM水质预测模型

王红晨[a],马 俊[b],陈博行[c]

(青海师范大学 a. 物理与电子信息工程学院; b. 高原科学与可持续发展研究院(物联网重点实验室); c. 计算机学院,青海 西宁 810016)

摘要:针对传统水质预测方法难以捕捉样本中时空特征的问题,提出建立基于CNN-EA-ConvLSTM水质预测模型,即首先通过卷积神经网络(CNN)对数据降维处理,提取样本特征,然后使用外部注意力机制探索样本间的隐藏信息,以卷积长短期记忆网络(ConvLSTM)进一步捕捉数据的空间特性,为使模型能达到最优效果,使用遗传算法优化模型中的超参数,最后以青海省的水质监测数据为样本对模型进行仿真验证。结果表明,该模型的平均绝对误差(M_{MAE})为0.063、均方根误差(R_{RMSE})为0.012、平均绝对百分比误差(M_{MAPE})为2.6%,与CNN-EA模型、CNN-LSTM模型相比M_{MAE}、R_{RMSE}、M_{MAPE}分别降低了18%和24%、14%和25%、16%和21%,模型可有效获取水质的时空特征,减弱不同样本间的影响,达到理想预测效果。

关键词:水质预测;卷积神经网络(CNN);外部注意力;ConvLSTM;遗传算法(GA)

中图分类号:TV211.1;TP391.9 **文献标志码**:A **文章编号**:1000-7709(2023)08-0073-04

1 引言

我国目前水质检测问题主要以水质监测和水质污染控制为主,对水质预测的研究缺少系统化,进而不能系统利用监测成果,导致无法快速发现污染事故。同时水环境系统的高度复杂性、指标的不确定性、非线性等特征致使需不断地分析和预测水质变化规律。随着人工智能的兴起,许多神经网络模型已应用于水质预测中[1],但传统的神经网络模型不适合处理有长期依赖关系的数据。传统的预测模型如CNN[2]、LSTM[3]、CNN-LSTM[4]、ATT-LSTM[5]模型虽能有效提取数据特征,准确预测水质变化,但传统的预测模型仅考虑样本的时间特性,难以捕捉样本的空间特性,无法深度学习数据的潜在隐藏信息。为了充分利用水质数据中的空间特性,提升预测准确率,本文将ConvLSTM[6]和外部注意力(External Attention)[7]机制引入到水质预测中,构建了CNN-EA-ConvLSTM水质预测模型,然后使用遗传算法(GA)对水质预测模型参数寻找最优解,并以青海省某地水质监测数据进行试验验证,确定该模型可有效提升预测精度、提高计算效率。

2 CNN-EA-ConvLSTM水质预测模型

CNN-EA-ConvLSTM水质预测模型由4部分组成:①先对输入的水质数据进行归一化处理,然后使用CNN模块将大量的水质数据降维为小数据量,有效保留其特征。②通过External Attention机制学习样本间的潜在联系,对CNN处理过的水质数据权值分配,提取特征。③使用ConvLSTM模块,与CNN模块和External Attention机制连接,建立CNN-EA-ConvLSTM水质预测模型。④采用GA优化算法优化水质预测模型的参数,提升模型的预测效率。具体步骤如下。

步骤1 选取数据,对各项水质数据进行归一化处理,并以8∶2的比例划分训练集与测试集。

步骤2 设置算法初始值,初始化种群,最大迭代次数为50,种群规模为20。

步骤3 对预测模型的隐藏层数、训练轮数、窗口数及学习率四个参数优化进行编码,初始化种群。

步骤4 计算遗传代数,解码优化参数。

步骤5 选取CNN-EA-ConvLSTM水质预测模型的误差函数作为适应度函数,计算个体适应度。

收稿日期:2022-09-28,**修回日期**:2022-10-24

基金项目:青海省自然科学基金项目(2021-ZJ-916)

作者简介:王红晨(1995-),女,硕士研究生,研究方向为深度学习、电子信息技术,E-mail:chen-xiaojiu@qq.com

通讯作者:马俊(1973-),男,博士、教授、博导,研究方向为无线电与智能系统、机器学习,E-mail:mjun7302@163.com

步骤 6 迭代寻优，进行选择、交叉、变异操作，满足终止条件则输出最优参数。

步骤 7 将步骤 6 输出的最优参数输入到 CNN-EA-ConvLSTM 模型建立水质预测模型。

2.1 卷积神经网络

CNN 模型是深度学习中常用的网络结构，由输入层、卷积层、激活函数层、池化层及全连接层共五部分组成。

卷积层后的特征图大小计算公式为：

$$N=(W-F+2P)/S+1 \tag{1}$$

式中，N 为卷积后产生的特征图大小；W 为输入矩阵大小；F 为卷积核大小；P 为填充值；S 为步长。

池化层后的特征图大小计算公式为：

$$N=[W+2P-D(F-1)-1]/S+1 \tag{2}$$

式中，D 为深度，即通道数。

卷积公式为：

$$V=\mathrm{conv}2(\boldsymbol{W},\boldsymbol{X},''\mathrm{valid}'')+b \tag{3}$$

式中，V 为 $\boldsymbol{W}$、$\boldsymbol{X}$ 的二维卷积；$\boldsymbol{W}$ 为卷积核矩阵；$\boldsymbol{X}$ 为输入矩阵；b 为偏置；valid 为卷积运算的类型。

2.2 External Attention 机制

External Attention 机制为一种外部注意力机制，见图 1。外部注意力机制使用两个线性层

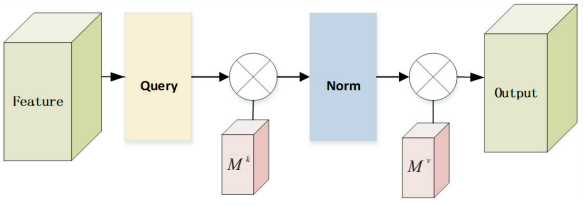

图 1 External Attention 机制结构图

Fig. 1 External Attention mechanism structure diagram

和一个归一化层代替现有的自注意力机制，具有线性复杂性，并考虑了不同特征图之间的潜在关系。计算公式简记为：

$$\boldsymbol{A}=\mathrm{Norm}(FM_K^T) \tag{4}$$

$$\boldsymbol{F}_{\mathrm{out}}=\boldsymbol{A}M_V \tag{5}$$

式中，$\boldsymbol{A}$ 为注意力矩阵；$M\in\mathbb{R}^{s\times d}$ 为引入的一个外部的 $S\times d$ 维空间记忆单元；$\boldsymbol{F}_{\mathrm{out}}$ 为注意力向量。

2.3 ConvLSTM 模型

ConvLSTM 模型是一种将 CNN 与 LSTM 在模型底层结合的卷积长短时期记忆神经网络，专门为时空序列设计的深度学习模块。ConvLSTM 的更新递归公式为：

$$\begin{cases} i_t=\delta(\boldsymbol{\omega}_i*[h_{t-1},x_t]+b_i) \\ f_t=\delta(\boldsymbol{\omega}_f*[h_{t-1},x_t]+b_f) \\ \widetilde{c_t}=\tanh(\boldsymbol{\omega}_c*[h_{t-1},x_t]+b_c) \\ c_t=f_t\odot c_{t-1}+i_t\odot\widetilde{c_t} \\ o_t=\delta(\boldsymbol{\omega}_o*[h_{t-1},x_t]+b_o) \\ h_t=o_t\odot\tanh(c_t) \end{cases} \tag{6}$$

式中，δ、tanh 均为激活函数，其中，i_t、f_t、o_t、c_t、$\widetilde{c_t}$、h_t 分别为输入门、遗忘门、输出门、单元状态、单元状态更新值和隐藏状态；$\boldsymbol{\omega}_i$、$\boldsymbol{\omega}_f$、$\boldsymbol{\omega}_o$、$\boldsymbol{\omega}_c$ 均为权重矩阵；b_i、b_f、b_o 分别为训练时输入门、遗忘门、输出门的偏置项；b_c 为单元状态更新后的偏置项；$*$ 为卷积；$\odot$ 为 Hadamard 乘。

2.4 优化算法

遗传算法(GA)为一种模拟自然进化过程搜索最优解的方法。该算法通过数学方式将问题的求解过程转换为类似生物进化的染色体基因的选择、交叉及变异过程。由于 CNN-EA-ConvLSTM 模型的超参数对模型预测有一定影响，因此使用 GA 优化算法对预测模型的隐藏层数、训练轮数、窗口数及学习率等参数寻优可有效提升模型的预测精度。

(1)适应度函数。适应度函数是驱动遗传算法的动力，用于区分全体中个体好坏的标准。本文选用水质预测模型的损失函数作遗传算法的适应度函数 $f(Y)$，公式为：

$$f(Y)=\frac{1}{N}\sum_{i=1}^{N}[y_i-\hat{y}_i(X)]^2 \tag{7}$$

式中，y_i、$\hat{y}_i$ 分别为输入数据的实际值、预测值。

(2)选择操作。选择过程采用轮盘赌法，从初始种群中找到适应度高的个体，计算其概率 P_i 为：

$$P_i=\frac{p_{\max}}{1-(1-p_{\max})^N}(1-p_{\max})^{N_i-1} \tag{8}$$

式中，$p_{\max}$ 为最佳染色体的概率；N_i 为染色体 i 的适应度值在种群中的序号。

(3)交叉和变异操作。交叉操作过程中为了保留亲本基因的相对顺序，选用有序交叉。变异操作过程中将个体序列打乱。该操作能使种群具有多样性，对挑选出来的个体进行均匀变异。交叉概率 P_c 和变异概率 P_m 采用随个体适应度值的变化而变化的取值方法，计算公式为：

$$P_c=\begin{cases} P_{c1}-\dfrac{(P_{c1}-P_{c2})(f'-f_{\mathrm{avg}})}{f_{\max}-f_{\mathrm{avg}}} & f'>f_{\mathrm{avg}} \\ P_{c1} & f'\leqslant f_{\mathrm{avg}} \end{cases} \tag{9}$$

$$P_m=\begin{cases} P_{m1}-\dfrac{(P_{m1}-P_{m2})(f'-f_{\mathrm{avg}})}{f_{\max}-f_{\mathrm{avg}}} & f''>f_{\mathrm{avg}} \\ P_{m1} & f''\leqslant f_{\mathrm{avg}} \end{cases} \tag{10}$$

式中，P_{c1}、P_{c2}、P_{m1}、P_{m2} 均为常数；$f_{\max}$ 为该代群体中最优个体的适应度；f_{avg} 为该代群体的平均适应度；f' 为要交叉的两个个体中较大的适应度；f'' 为要变异的个体适应度。

3 仿真验证

3.1　数据来源和环境配置

黄河发源于中国青海省巴颜喀拉山脉,流经青海、四川、甘肃、宁夏、内蒙古、陕西、山西、河南、山东 9 个省区。黄河流域面积 752 773 km^2,河长 5 464 km,年径流量为 $592\times10^8 m^3$。黄河流域地表水河流监测断面共有 216 个,其中干流断面 39 个、支流断面 177 个,黄河流域主要支流水质总体为轻度污染,主要污染指标为氨氮、总磷和化学需氧量,其中工业废水污染主要占 60%～70%,生活污水约占 30%,农业及其他污染占 10%以下。为预防水污染,我国推进流域控制单元精细化管理,明确考核断面,建立排放源(工业、生活、面源)和单元水质明确清晰的响应关系,将流域生态环境保护责任层层分解到各级行政区区域,建立完善责任体系。

选取黄河流域中青海省某地 2016 年 6 月～2021 年 1 月 5 520 组化学需氧量(COD)数据,数据采样频率为每 4 h 采集 1 次。使用前 80%组数据进行网络训练,后 20%组作为测试数据。硬件环境为 Interi5-1135G7,主频为 2.4 GHz;显卡 NVIDIA GeForce MX450,2GB 显存;内存 16 GB。模型使用 python3.6 进行编程实现,采用 Pytorch 作为深度学习框架,模型优化后的参数设置见表 1。

表 1　参数设置

Tab. 1　Parameter settings

Time Window	Learning Rate	nEpoch	Loss function	Number of hidden layers
7	0.000 74	883	M_{MSE}	35

3.2　模型评价指标

采用预测值与实际值之间的误差评估提出的水质预测模型性能,评估标准选择均方根误差(R_{RMSE})、平均绝对百分比误差(M_{MAPE})、平均绝对误差(M_{MAE})。计算公式见文献[8]。

3.3　仿真结果与分析

图 2 为在不同样本数量时模型精度变化。由图 2 可知,样本数量不同时模型精度有所变化,其中,使用一年以下数据建模时,所得各指标数据均较大,且指标变化不稳定。如 M_{MAE}、R_{RMSE} 指标先降后升,采用 1.5 年数据量建模时相较于一年以下的数据模型精度有明显提升,且逐渐趋于平稳。因此,建模时可采用 1.5 年最小样本数据量进行仿真验证。

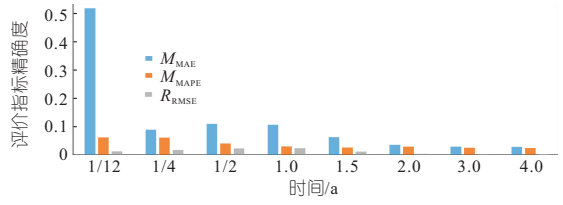

图 2　不同样本数量的精度对比

Fig. 2　Comparison of accuracy for different sample sizes

表 2 为指标相同、样本数量不同时模型的精度对比(选取为期 4 年的水质数据进行扩展试验)见表 2。由表 2 可知,以 M_{MAE} 指标为例,其值降低至 0.028 mg/L,与 1/12、1/4、1/2、1、1.5、2、3 年相比分别降低了 0.492、0.061、0.082、0.078、0.034、0.007、0.001,可见模型精度与样本数量密切相关,样本数量越多,精度越高。

表 2　不同样本数量预测结果精度分析

Tab. 2　Analysis of accuracy of prediction results with different sample sizes

样本数量/a	M_{MAE} /(mg·L^{-1})	M_{MAPE} /%	R_{RMSE} /(mg·L^{-1})
1/12	0.520	6.186	0.013
1/4	0.089	6.139	0.017
1/2	0.110	4.054	0.022
1.0	0.106	3.003	0.023
1.5	0.062	2.626	0.011
2.0	0.035	2.880	0.003
3.0	0.029	2.548	0.002
4.0	0.028	2.431	0.001

图 3 为不同方法下 COD 预测对比。由图 3 可知,CNN-LSTM 模型预测值与真实值之间波形有较大浮动,通过引入 External Attention 机制后,CNN-EA 模型有较明显的改善,然而本文构建的 GA-CNN-EA-ConvLSTM 预测模型与 CNN-EA、CNN-LSTM、CNN-EA-ConvLSTM 相比差异波动较小,与真实值更相似,模型的预测性能得到了有效提升。

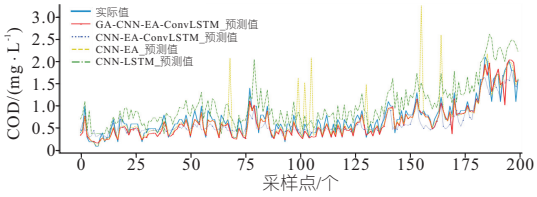

图 3　不同模型预测效果对比

Fig. 3　Comparison of prediction effects of different models

图 4 为同一指标因子、同一网络参数下几种网络模型损失函数的变化。由图 4 可知,本文模型与传统模型相比收敛速度平缓且更快,损失函

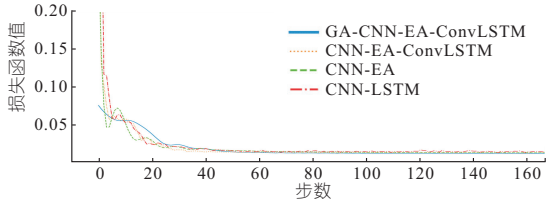

图 4　不同方法损失函数对比

Fig. 4　Comparison of loss functions of different methods

数值最趋近 0，损失函数值越小，预测精度越高，构建模型的鲁棒性更好。因此，本文方法相比于传统方法稳定性更好。

图 5 为验证本文模型在水质监测时外部环境改变引起样本序列变化的模型适应性及适用性。由于异常序列较少，为模拟异常情况，加入强度 0.5、持续时间 6、间隔 36 个时间点的倒 U 型曲线，灰色柱体为异常时刻。由图 5 可知，该模型能预测出异常位置的水质变化情况，预测精度高。综上所述，样本量和样本序列不同时模型均能预测准确，有较好的适用性及适应性。

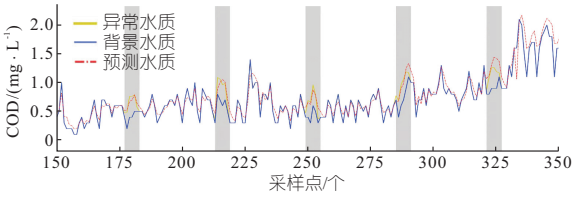

图 5 异常模拟预测结果

Fig. 5 Anomaly simulation prediction results

表 3 为不同模型的精度对比。由表 3 可知，指标因子为 COD 时，与 CNN-EA、CNN-LSTM、CNN-EA-ConvLSTM 模型相比，GA-CNN-EA-ConvLSTM 模型的 M_{MAE} 分别降低了 18%、24%、14%；M_{MAPE} 分别降低了 16%、21%、10%；R_{RMSE} 分别降低了 14%、25%、7%。由此可知，水质预测模型能充分学习不同样本之间的影响，并提取样本的时空特征。

表 3 不同模型预测结果精度分析

Tab. 3 Analysis of accuracy of prediction results of different models

模型	M_{MAE}/(mg·L^{-1})	M_{MAPE}/%	R_{RMSE}/(mg·L^{-1})
CNN-EA	0.077	3.1	0.014
CNN-LSTM	0.083	3.3	0.016
CNN-EA-ConvLSTM	0.074	2.9	0.013
GA-CNN-EA-ConvLSTM	0.063	2.6	0.012

4 结论

a. 进化算法优化的 CNN-EA-ConvLSTM 水质预测模型能对水质历史数据进行高精度追踪，提取数据的时间和空间特性，充分考虑样本间的潜在影响。仿真验证本文提出的方法误差最小，最接近真实数值。

b. 未来考虑将本文模型应用于水质异常检测方向，以预防水质污染。

参考文献：

[1] WANG Z, MAN Y, HU Y, et al. A deep learning based dynamic COD prediction model for urban sewage[J]. Environ. Sci.: Water Res. Technol., 2019, 5(12): 2210-2218.

[2] 余舒，杨志刚. 基于 DBSCAN 和 CNN 算法的重型车辆 NOx 排放预测模型[J]. 重庆交通大学学报(自然科学版)，2022，41(8)：134-141.

[3] 谌辰睿，谭传世，王兴霞，等. 基于 LSTM 神经网络的高拱坝混凝土温升阶段温度预测[J]. 水电能源科学，2022，40(6)：101-104，18.

[4] 王军，高梓勋，朱永明. 基于 CNN-LSTM 模型的黄河水质预测研究[J]. 人民黄河，2021，43(5)：96-99，109.

[5] 孙隽丰，李成海，曹波. 基于 TCN-BiLSTM 的网络安全态势预测[J/OL]. 系统工程与电子技术：1-11[2022-09-23]. https://kns.cnki.net/kcms/detail/11.2422.TN.20220922.0912.002.html.

[6] 杨鑫，张建云，周建中，等. 基于 ConvLSTM 网络的多源降雨融合方法[J]. 华中科技大学学报(自然科学版)，2022，50(8)：33-39.

[7] GUO M H, LIU Z N, MU T J, et al. Beyond self-attention: External attention using two linear layers for visual tasks[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(5): 5436-5447.

[8] 庞吉玉，张安兵，王贺封，等. 基于无人机多光谱影像和 XGBoost 模型的城市河流水质参数反演[J]. 中国农村水利水电，2023(3)：111-119.

CNN-EA-ConvLSTM Water Quality Prediction Model Based on Evolutionary Algorithm Optimization

WANG Hong-chen[a], MA Jun[b], CHEN Bo-hang[c]

(a. College of Physics and Electronic Information Engineering; b. Academy of Plateau Science and Sustainability (Key Laboratory of Internet of Things); c. College of Computer, Qinghai Normal University, Xining 810016, China)

Abstract: Aiming at the problem that the traditional water quality prediction method is difficult to capture the spatial and temporal characteristics of the sample, this paper proposes to establish a CNN-EA-ConvLSTM based water quality prediction model. The convolutional neural network (CNN) was used to reduce the dimensionality of the data and extract the sample features. Then the hidden information among samples was explored by the external attention mechanism. The convolutional long and short-term memory network (ConvLSTM) was further used to capture the spatial characteristics of the data. To achieve optimal results of the model, a genetic algorithm was used to optimize the parameters of the model. The water quality test data of Qinghai Province was used as a sample to simulate and validate the model. The results show that the mean absolute error (M_{MAE}) of the model is 0.063, the root mean square error (R_{RMSE}) is 0.012, and the mean absolute percentage error is 2.6%, which are respectively reduced by 18% and 24%, 14% and 25%, 16% and 21% compared with the CNN-EA model and CNN-LSTM model. Therefore, the model can effectively obtain the spatial and temporal characteristics of water quality, attenuate the influence of different samples, and achieve the ideal prediction effect.

Key words: water quality prediction; CNN; external attention; ConvLSTM; GA