DOI:10.20040/j.cnki.1000－7709.2026.20251581

融合特征工程与优化 VMD-LSBoost 的湖泊溶解氧浓度预测

范士军[1,2],李鹏程[3,4],张文秀[1,2],赵艳苓[1,2],单长河[4],苏永军[4],贾 悦[4]

(1. 任丘市水务局,河北 沧州 062550; 2. 任丘市水利建筑工程公司,河北 沧州 062550; 3. 河北师范大学地理科学学院,河北 石家庄 050010; 4. 河北水利电力学院河北省滨海湿地水资源调配及生态保护技术创新中心,河北 沧州 061001)

摘要:构建基于数据驱动的水质预测模型对湖泊污染预警具有重要意义,为克服输入特征缺乏理论基础及精度较低等缺点,首先应用K最近邻(KNN)算法、数据清除法、VMD变分模态分解数据处理技术对水质数据进行预处理;然后利用最大信息系数(MIC)算法确定各特征重要性程度并制定输入方案,导入LSBoost预测模型,同时为避免模型过拟合问题,提高模型泛化能力,利用霜冰优化算法(RIME)、变色龙优化算法(CSA)、大猩猩优化算法(GTO)等仿生算法优化模型超参数;最后将构建的32种深度学习模型应用于千岛湖DO浓度值预测中。结果表明,MIC算法优选输入特征,以降低冗余特征的干扰,LSBoost模型预测性能有所提升,在千岛湖小金山、三潭岛、大坝前水质监测站最优预测模型分别为MIC-LSBoost7、MIC-LSBoost4、MIC-LSBoost4,其决定系数(R^2)均大于0.81,预测误差(e_{MAPE})均小于4.05%。仿生算法对MIC-LSBoost模型超参数进行优化,构建湖泊DO浓度值最优预测模型,R^2 均大于0.890,e_{MAPE} 在3.072%以内,该模型可优选与预测指标相关特征,有效提高了模型捕捉DO浓度值演变规律的敏感性。

关键词:湖泊;溶解氧浓度;最大信息系数;仿生优化算法;水质预测模型

中图分类号:X832 **文献标志码**:A **文章编号**:1000-7709(2026)02-0078-06

Integrated Feature Selection and Optimized VMD-LSBoost for Lake Dissolved Oxygen Concentration Prediction

FAN Shi-jun[1,2], LI Peng-cheng[3,4], ZHANG Wen-xiu[1,2], ZHAO Yan-ling[1,2], SHAN Chang-he[4], SU Yong-jun[4], JIA Yue[4]

(1. Renqiu Water Affairs Bureau, Cangzhou 062550, China; 2. Renqiu Water Conservancy and Construction Engineering Company, Cangzhou 062550, China; 3. School of Geographic Sciences, Hebei Normal University, Shijiazhuang 050010, China; 4. Hebei Technology Innovation Center for Coastal Wetland Water Resources Allocation and Ecological Protection of Hebei University of Water Resources and Electric Engineering, Cangzhou 061001, China)

Abstract: Data-driven water quality prediction models are critical for lake pollution early warning. To address limitations such as insufficient theoretical basis for input features and low accuracy, water quality data are preprocesseed using K-nearest neighbor (KNN), data clearance, and variational mode decomposition (VMD) techniques. The maximal information coefficient (MIC) algorithm is employed to assess feature importance and formulate input schemes for the LSBoost prediction model. To mitigate overfitting and enhance generalization, bio-inspired optimization algorithms including frost ice optimization (RIME), chameleon swarm algorithm (CSA), and gorilla troop optimizer (GTO) are utilized to optimize hyperparameters of the model. The proposed framework is applied to predict dissolved oxygen (DO) concentrations in Qiandao Lake using 32 deep learning models. The results demonstrate that the input features selected by MIC reduced redundancy, improving LSBoost prediction performance. The optimal prediction models for the water quality moni-

收稿日期:2025-09-09,**修回日期**:2025-10-23,**接受日期**:2025-10-27

基金项目:国家科技支撑计划(31410005321403);沧州市科技计划自筹经费项目(20251105004);沧州市社会科学发展研究课题(2025376)

作者简介:范士军(1975-),男,高级工程师,研究方向为智慧水利、水利信息化,E-mail:1320329010@qq.com

通讯作者:李鹏程(1990-),男,博士研究生、讲师,研究方向为生态建设与灾害防治,E-mail:lipengcheng@hbwe.edu.cn

中国知网 https://www.cnki.net

toring stations in Xiaojinshan, Santan Island and in front of the Dam in Qiandao Lake are MIC-LSBoost7, MIC-LSBoost4 and MIC-LSBoost4, respectively. The coefficient of determination (R^2) was all greater than 0.81, and the prediction error (e_{MAPE}) was all less than 4.05%. The bionic algorithm optimized the hyperparameters of the MIC-LSBoost model to construct the optimal prediction model of the lake DO concentration. The R^2 was all greater than 0.890, and the e_{MAPE} was within 3.072%. This model can optimize the features related to the predictive indicators, effectively improving the sensitivity of the model in capturing the evolution law of DO concentration.

Key words: lake; dissolved oxygen concentration; maximum information coefficient; bionic optimization algorithm; water quality prediction model

1　引言

湖泊中溶解氧(DO)浓度变化反映了水体中植物、微生物等产生氧气(如光合作用)及消耗氧气(如有氧呼吸、硝化作用)过程的平衡,其演变情势预测对湖泊水质预警及水环境监管具有重要意义。基于数据驱动算法的水质预测方法表现出高精度、实时监测预警等优势。利用 CNN-BiLSTM 组合模型[1]、深度递归神经网络(DRNN)模型、支持向量机(SVM)[2]等模型预测水体 DO 浓度值,在处理特征数据和模型精度方面各有优劣势。与此同时,最大信息系数(MIC)算法对处理不同特征复杂非线性关系具有较强泛化性及容错性[3],最小二乘提升模型(LSBoost)在气液界面张力模拟[4]、微藻水热液化过程辅助建模[5]等领域展现了较高的预测精度及普适性,但在水质预测领域应用较少。为此,本文基于千岛湖 DO 浓度值,通过 K 最近邻(KNN)算法、数据清除法、VMD 变分模态分解数据处理技术等对水质数据组进行处理;以最小二乘提升(LSBoost)模型为基础,采用 MIC 算法进行特征优选,确定输入方案组合,利用新型仿生算法优化 LSBoost 模型,构建组合优化模型,利用评价指标体系分析,并确定湖泊 DO 浓度值最优预测模型。

2　数据来源与研究方法

2.1　研究区选取及数据来源

考虑到空间分布差异性、水动力特征等因素影响,选取千岛湖小金山、三潭岛及大坝前三个水质监测站,具体研究区及水质监测站分布见图 1。

基于 2021 年 1 月 1 日～2022 年 3 月 31 日千岛湖水质数据,监测数据获取间隔时间为 4 h,其中部分水质指标值缺失严重,经过筛选最终以叶电导率(EC)、氨氮(NH_3-N)、总氮(TN)、总磷(TP)、浊度(TB)、水温(W-T)、酸碱度(pH 值)、高锰酸盐指数(COD_{Mn})等 8 项指标作为特征值。

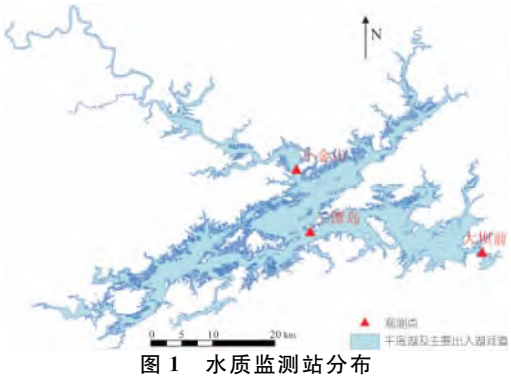


图 1　水质监测站分布

Fig. 1　Distribution of water quality monitoring stations

2.2　研究方法

2.2.1　数据处理方法

(1)缺失数据处理。为避免由于水质指标自身演变趋势失衡降低模型模拟精度,对连续 24 h 数据缺失的水质指标进行数据清除,不再进行插值处理。经过 KNN 反距离权重插值、数据清除法处理,千岛湖小金山、三潭岛及大坝前三个水质监测站分别确定 2 601、2 513、2 339 组水质数据。三个水质监测站各水质指标特性及影响因素存在较大差异,通过对不同站点水质规律进行研究及预测,可提高模型的稳定性及泛化能力。

(2)VMD(Variational Mode Decomposition)变分模态分解数据处理技术。VMD 技术通过变分优化框架将溶解氧浓度波动分解为多频带模态分量,可显著提升对河湖水质突发性污染事件的 DO 浓度预测稳定性[6],参考相关资料及调参,设定惩罚因子为 2 000,模态数 K 为 6,以三潭岛水质监测站为例,将溶解氧作为输入,分解的六个序列见图 2。

2.2.2　湖泊溶解氧浓度模型算法原理

(1)最大信息系数算法。最大信息系数(MIC)算法通过优化互信息(MI)方法,能够衡量 DO 浓度值与其他水质指标两变量之间关联程度的统计量,有效捕捉非线性关系,其计算公式为:

$$\mathrm{MIC}(X;Y)=\max_{i\cdot j<B}\frac{I(X;Y)}{\log_2[\min(X,Y)]} \quad (1)$$

式中,X、Y 为水质指标变量;B 为样本数量,两水质指标 MIC 值越接近 1,表示其关联性越强,

中国知网　https://www.cnki.net

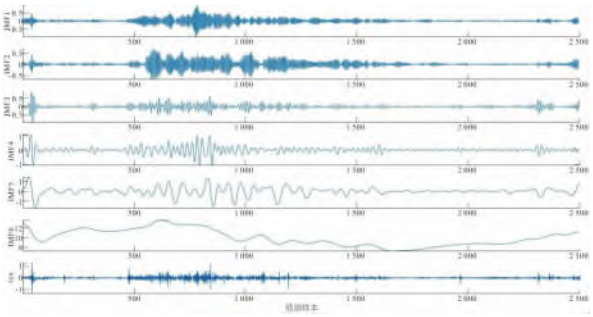


图 2 三潭岛水质监测站溶解氧时间序列分解

Fig. 2 Dissolved oxygen time series decomposition at Santan Island water quality monitoring station

MIC∈[0,1]。

(2)LSBoost 模型。LSBoost 模型通过逐步添加弱学习器来构建模型,这种方式使得算法对于异常值和噪声数据具有一定的鲁棒性。另外 LSBoost 算法需要调整超参数较少,可简化模型选择和调优过程。该算法主要缺点为易出现过拟合,需要多次试验及优化算法进行超参数调整等方法提高模型的泛化能力。其主要过程包括初始化模型、迭代拟合弱学习器以预测残差,通过最小二乘法优化残差拟合,最终将弱学习器加权组合成强学习器,具体步骤参见文献[7]。

(3)霜冰优化算法。霜冰优化算法(RIME)[8]通过模拟软霜的随机探索与硬霜的规律性穿刺机制,结合积极贪婪选择策略,在保证种群多样性的同时高效收敛于全局最优解。

(4)变色龙优化算法。变色龙优化算法(CSA)[9]通过模拟变色龙捕食时的搜索、锁定与位置更新行为,有效平衡全局探索与局部开发能力,在多模态优化问题中展现出强大的寻优性能与收敛效率。

(5)大猩猩优化算法。大猩猩优化算法(GTO)[10]通过模拟大猩猩的种群迁移和求偶行为来搜索最优解。在诸多标准基准函数和工程问题等复杂和高维的优化问题上,展现出更快收敛速度。

2.2.3 湖泊溶解氧浓度模型精度验证

构建组合模型,如 MIC-RIME-LSBoost7,表示结合 MIC 算法、RIME 算法及 LSBoost 模型,并采用方案 7,依次类推,合计 32 种模型。将每个滑动窗口后的第一个时间宽度(24 h)水质数据作为输出样本,单个窗口包含过去 96 h 数据,预测目标为窗口后 4 h 水质情况。将水质信息集按照 8∶1∶1 的比例分为训练集、测试集及验证集,通过计算平均绝对误差(e_{MAE})、相对误差(e_{MAPE})、均方误差(e_{MSE})、均方根误差(e_{RMSE})及决定系数(R^2)5 个基础指标及全指标综合系数(G_{GPI})[11],对比分析不同模型模拟精度。

3 结果与分析

3.1 模型超参数优化结果

经调参,RIME 算法中,种群大小为 50,随机数 $r_1=0.7$、$r_2=0.3$。CSA 优化算法种群大小为 40,维度为 2,感知猎物的概率 P_p 设为 0.5,学习因子 F 设置为 2.0。GTO 算法种群数量 $N=50$,迁移策略参数 p 用于控制大猩猩对未知位置的迁移策略,设为 0.03。LSBoost 模型主要超参数优化区间见表 1。

表 1 模型超参数优化区间

Tab. 1 Model hyperparameter optimization interval

模型超参数	学习率	子样本比例	采样率	弱分类器的数量	决策树最大深度
取值范围	[0.01,0.1]	[0.5,1.0]	[0.5,1.0]	[100,300]	[5,15]

3.2 基于 MIC 算法的特征优选分析

为提高湖泊 DO 浓度值预测精度,基于 MIC 算法计算各水质指标特征值并确定特征优选方案。图 3 为选定的三个水质监测站影响 DO 的特征值,根据特征值确定不同的输入变量组合,开展试验精度计算及对比分析。

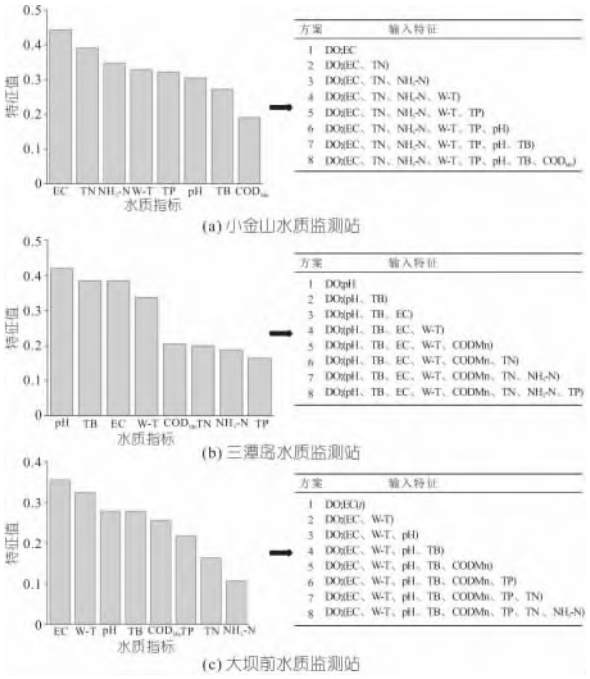


图 3 不同水质监测站特征优选结果及方案

Fig. 3 Preferred results for characterization of different water quality monitoring stations

3.3 不同模型模拟效果对比分析

中国知网 https://www.cnki.net

3.3.1　基于 MIC 及优化算法的 LSBoost 模型 DO 浓度值预测精度对比

(1)小金山水质监测站模型预测精度对比。图 4 为小金山水质监测站 DO 浓度值不同输入方案下不同模型精度。由图 4 可看出,随着输入参数增多,模型的精度先升高后降低,输入单一参数时精度最低,MIC-LSBoost7 模型精度达到最高,其 e_{MAE}、e_{MAPE}、e_{MSE}、e_{RMSE}、R^2 分别为 0.362 mg/L、3.581%、0.387 mg/L、0.622 mg/L、0.891,模型 G_{GPI} 为 0.154,输入 8 个参数时模型精度再次降低。MIC-LSBoost7 模型精度高于 LSBoost 模型,其余 7 种 MIC 特征优选模型精度低于 LSBoost 模型。相同输入方案情境下,经过仿生算法优化的 MIC-LSBoost 模型精度均高于 MIC-LSBoost 模型,3 种仿生算法优化的 MIC-LSBoost7 模型平均 G_{GPI} 达到 0.640,MIC-RIME-LSBoost7 模型精度最高,其 e_{MAE}、e_{MAPE}、e_{MSE}、e_{RMSE}、R^2 分别为 0.272 mg/L、2.667%、0.184 mg/L、0.428 mg/L、0.948,模型 G_{GPI} 为 0.709。不同输入方案情境下,各仿生算法优化效应表现不同,如方案 6 情境下,仿生算法优化效应 GTO>CSA>RIME,方案 7 情境下,仿生算法优化效应 RIME>GTO>CSA,因此通过预测效果对比,确定不同输入特征方案下的 DO 浓度值高精度预测模型是必要的。

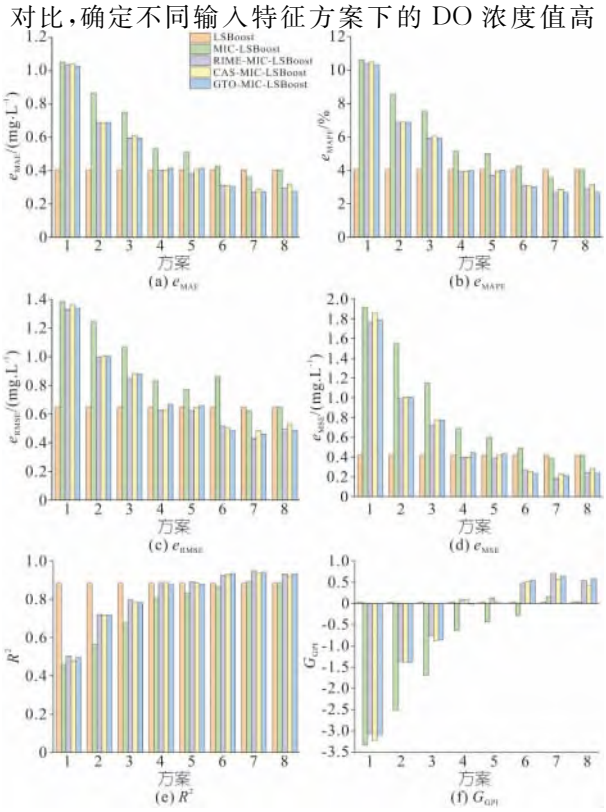


图 4　小金山水质监测站 DO 浓度值不同模型预测精度对比

Fig. 4　Comparison of the prediction accuracy of different models for DO at Xiaojinshan water quality monitoring station

(2)三潭岛水质监测站模型模拟精度对比。图 5 为三潭岛水质监测站 DO 浓度值不同输入方案下不同模型精度。由图 5 可看出,仅经过 MIC 特征优选的方案中,MIC-LSBoost4 模型精度达到最高,输入单一参数时模型精度最低。相同输入方案情境下,经过仿生算法优化的 MIC-LSBoost 模型精度均有所提升,其中经过 GTO 仿生算法优化的 MIC-GTO-LSBoost5 模型精度最高。由此,与小金山水质监测站 DO 浓度值预测结果有所差异,经过仿生算法优化的最高精度模型与 MIC-LSBoost 最高精度模型为不同输入方案。

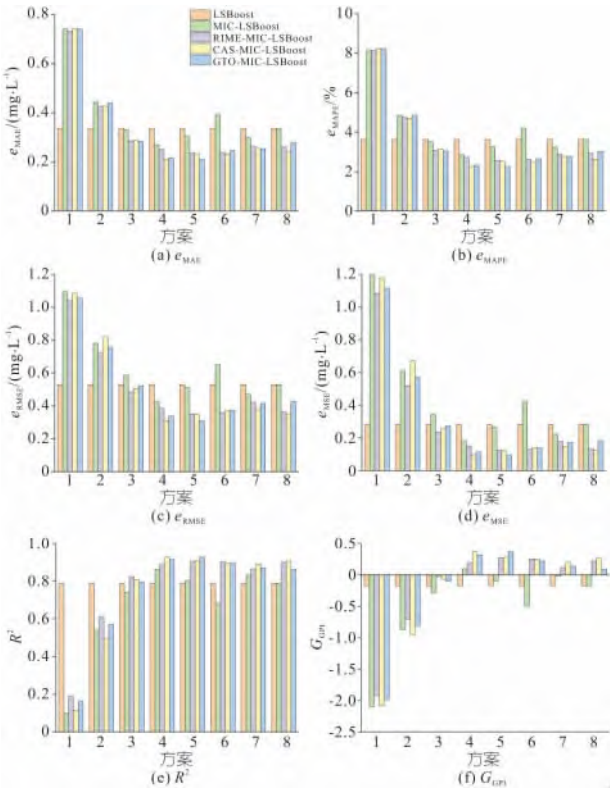


图 5　三潭岛水质监测站 DO 浓度值不同模型预测精度对比

Fig. 5　Comparison of prediction accuracy of different models for DO at Santan Island water quality monitoring station

(3)大坝前水质监测站模型模拟精度对比。图 6 为大坝前水质监测站 DO 浓度值不同输入方案下不同模型精度。由图 6 可看出,仅经过 MIC 特征优选的方案中,输入单一参数时模型精度最低,MIC-LSBoost4 模型精度达到最高。各方案中,经过 CSA 仿生算法优化的 MIC-CSA-LSBoost4 模型精度最高。在输入参数较少,为 1、2 时,RIME 算法优化效应高于 CSA 算法、GTO 算法,当输入参数较多,为 3～7 时,RIME 算法优化效应低于另两种仿生算法,各输入方案

中国知网　https://www.cnki.net

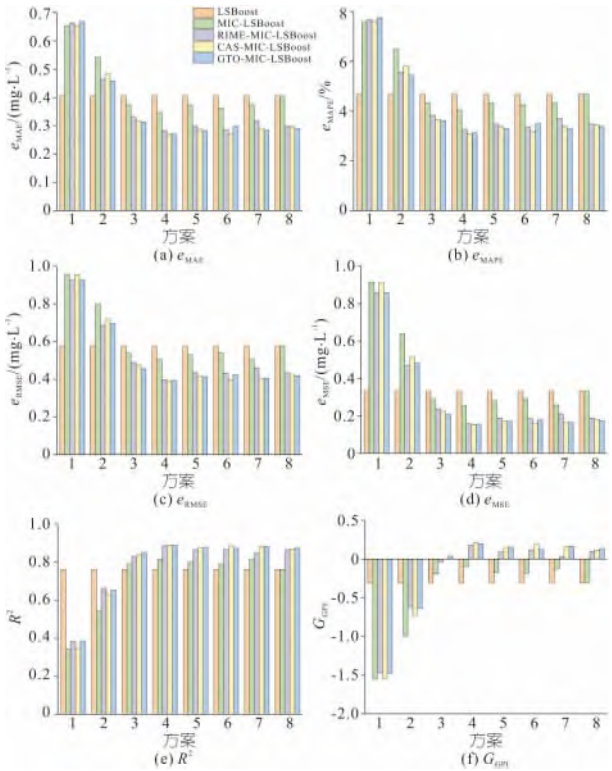


图 6 大坝前水质监测站 DO 浓度值不同模型预测精度对比

Fig. 6 Comparison of the prediction accuracy of different models for DO at the water quality monitoring station in front of the dam

中，CSA 算法与 GTO 算法优化效应较为接近。

3.3.2 LSBoost 模型预测性能及拟合效果分析

图 7 以小提琴、箱形组合图量化了各水质监测站 DO 浓度实测值及精度较高的各模型 DO 浓度预测值分布，其中箱线图中，从下到上分别表示 0.25、0.50、0.75 四分位数。根据小提琴图形状、宽度，观察 DO 浓度值数据密度及分布范围，对比各模型预测差异性，横坐标标注的预测模型自左

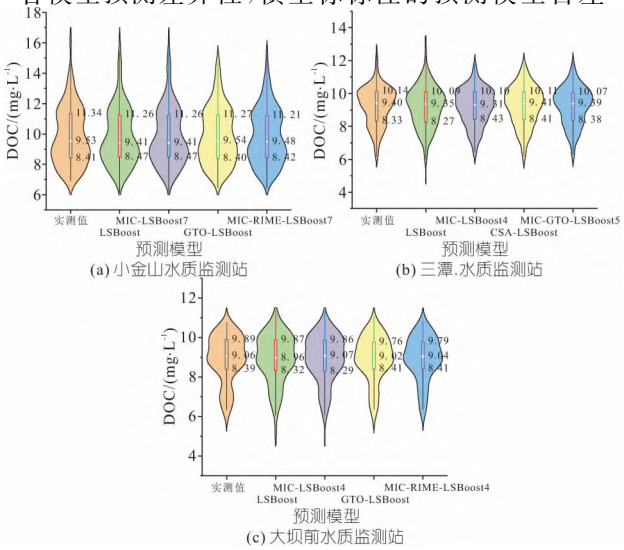


图 7 DO 浓度值预测效果小提琴及箱线组合图

Fig. 7 DO predictive effectiveness violin and box-line combinations

至右，模型预测性能逐渐提升，在三潭岛、大坝前水质监测站表现更为明显，箱线图中各模型四分位数支持了上述结论。该分布特征同样表明，经过 MIC 优选的 LSBoost 模型与仿生算法优化后的 LSBoost 模型，均可提升其预测性能，同时经过 MIC 优选及仿生算法优化后的 LSBoost 模型预测性能提升明显。

图 8 为 DO 浓度值预测效果相对误差图。图 8 中，模型 1～4 分别为 LSBoost 模型、MIC-LSBoost 精度最高模型、仿生算法优化 LSBoost 精度最高模型、MIC 优选与仿生算法优化 LSBoost 精度最高模型。由图 8 可看出，通过实测、预测 DO 浓度值相对误差时间序列走势，模型 1～4 中 DO 浓度值预测性能逐渐提升，相对误差振幅逐渐趋于平缓，拟合较大偏差值逐渐降低。经过特征优选及仿生算法优化的 LSBoost 模型，

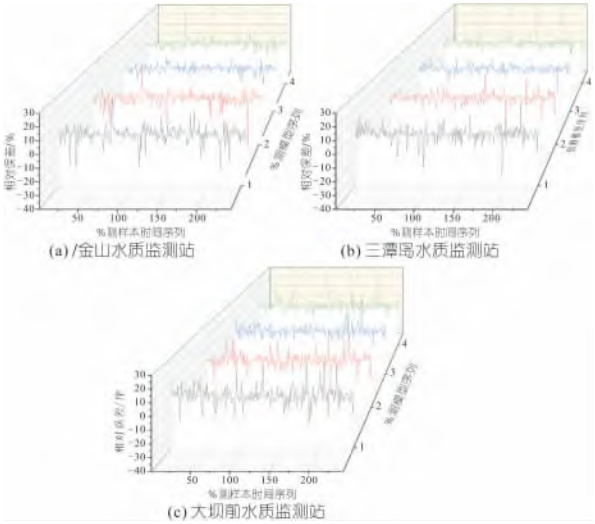


图 8 DO 浓度值预测效果相对误差图

Fig. 8 Relative error plot of the predicted effect of DO values

优选与预测指标相关的特征，提高了模型捕捉 DO 浓度波动趋势的性能，通过调整 LSBoost 模型学习率、子样本比例等超参数，有效提高了模型对 DO 浓度演变规律的敏感性，模型拟合效果进一步提升。

4 结论

a. 小金山、三潭岛、大坝前水质监测站 DO 浓度值预测中，仅经过 MIC 算法优选，最优预测模型分别为 MIC-LSBoost7、MIC-LSBoost4、MIC-LSBoost4，有效去除冗余特征值干扰，提高了 LSBoost 模型预测精度。

b. 不同输入方案下，经过 MIC 优选及仿生

中国知网 https://www.cnki.net

算法优化的模型均表现出了较好的拟合效果，MIC-RIME-LSBoost7、MIC-GTO-LSBoost5、MIC-CSA-LSBoost4 模型在小金山、三潭岛、大坝前水质监测站 DO 浓度值预测中，R^2 分别达到 0.948、0.928、0.890，e_{MAPE} 分别达到 2.667%、2.260%、3.072%，预测性能最优，可推荐作为千岛湖水质 DO 浓度值预测模型使用。

参考文献：

[1] WANG X Y, TANG X Y, ZHU M, et al. Predicting abrupt depletion of dissolved oxygen in Chaohu lake using CNN-BiLSTM with improved attention mechanism[J]. Water research, 2024, 261: 122027.

[2] MOGHADAM S V, SHARAFATI A, FEIZI H, et al. An efficient strategy for predicting river dissolved oxygen concentration: Application of deep recurrent neural network model[J]. Environmental monitoring and assessment, 2021, 193(12): 798.

[3] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting novel associations in large data sets [J]. Science, 2011, 334(6062): 1518-1524.

[4] BEHNAMNIA M, MOZAFARI N, DEHGHAN MONFARED A. Rigorous hybrid machine learning approaches for interfacial tension modeling in brine-hydrogen/cushion gas systems: Implication for hydrogen geo-storage in the presence of cushion gas [J]. Journal of energy storage, 2023, 73: 108995.

[5] WU W, HUANG J M, TSAI Y H. Machine Learning-Assisted Microalgae Biorefinery for Green Diesel Production[J]. Available at SSRN 4408666.

[6] 肖裕锋，张代青. 可解释性融合神经网络在巢湖总磷浓度预测中的应用[J]. 中国农村水利水电，2025(7)：44-51.(XIAO Y F, ZHANG D Q. Application of interpretable fusion neural network in the prediction of total phosphorus concentration in Chaohu lake[J]. China rural water and hydropower, 2025(7): 44-51.(in Chinese))

[7] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. The annals of statistics, 2001, 29(5): 1189-1232.

[8] SU H, ZHAO D, HEIDARI A A, et al. RIME: A physics-based optimization [J]. Neurocomputing, 2023, 532: 183-214.

[9] BRAIK M S. Chameleon Swarm Algorithm: A bio-inspired optimizer for solving engineering design problems [J]. Expert systems with applications, 2021, 174: 114685.

[10] ABDOLLAHZADEH B, SOLEIMANIAN GHAREHCHOPOGH F, MIRJALILI S. Artificial gorilla troops optimizer: A new nature-inspired metaheuristic algorithm for global optimization problems [J]. International journal of intelligent systems, 2021, 36(10): 5887-5958.

[11] 周晴晴，穆琳. 基于仿生优化算法构建混凝土拱结构应力预测模型[J]. 水电能源科学，2022，40(11)：141-145.(ZHOU Q Q, MU L. Stress prediction model of concrete arch structure based on bio-inspired optimization algorithm[J]. Water resources and power, 2022, 40(11): 141-145.(in Chinese))

(上接第 108 页)

[6] 王怀才，李北星，杨建波，等. 含粗骨料补偿收缩 UHPC 的制备与性能研究[J]. 新型建筑材料，2021，48(9)：41-45.(WANG H C, LI B X, YANG J B, et al. Preparation and properties of shrinkage-compensating ultra-high performance concrete containing coarse aggregate[J]. New building materials, 2021, 48(9): 41-45.(in Chinese))

[7] 陈宝春，季韬，黄卿维，等. 超高性能混凝土研究综述[J]. 建筑科学与工程学报，2014，31(3)：1-24.(CHEN B C, JI T, HUANG Q W, et al. Review of research on ultra-high performance concrete [J]. Journal of architecture and civil engineering, 2014, 31(3): 1-24.(in Chinese))

[8] 黄维蓉，杨玉柱，刘延杰，等. 含粗骨料超高性能混凝土的力学性能[J]. 硅酸盐学报，2020，48(11)：1747-1755.(HUANG W R, YANG Y Z, LIU Y J, et al. Mechanical properties of ultra-high performance concrete containing coarse aggregate[J]. Journal of the Chinese ceramic society, 2020, 48(11): 1747-1755.(in Chinese))

[9] 李北星，沈楚琦，李群胜. 集料粒径对超高性能混凝土力学性能与干燥收缩的影响[J]. 混凝土，2020(7)：55-58.(LI B X, SHEN C Q, LI Q S. Effect of aggregate particle size on mechanical properties and dry shrinkage of ultra-high performance concrete [J]. Concrete, 2020(7): 55-58.(in Chinese))

[10] YONG L, YU R, SHUI Z H, et al. Development of an environmental Ultra-High Performance Concrete (UHPC) incorporating carbonated recycled coarse aggregate[J]. Construction and building materials, 2023, 362: 129657.

[11] 中国工程建设标准化协会混凝土结构专业委员会. 超高性能混凝土试验方法标准：T/CECS 864-2021 [S]. 北京：中国建筑工业出版社，2021.(Committee of Concrete Structure, China Association for Engineering Construction Standardization. Standard for Test Method of Ultra-High Performance Concrete: T/CECS 864-2021[S]. Beijing: China Architecture & Building Press, 2021.(in Chinese))

[12] ZHAO Y S, QI Y, CHEN G F, et al. Synergistic effect of silica fume and ultra-fine glass microsphere on macro properties and hydration of coarse-aggregate UHPC[J]. Construction and building materials, 2023, 408: 133695.

中国知网 https://www.cnki.net