

邓志成, 万金泉, 王艳, 等. 孤立森林算法优化的 XGBoost 对养殖废水处理水质预测建模[J]. 净水技术, 2024, 43(12): 96–102, 110.

DENG Z C, WAN J Q, WANG Y, et al. Modeling water quality prediction for aquaculture wastewater treatment by isolation forest algorithm-optimized XGBoost[J]. Water Purification Technology, 2024, 43(12): 96–102, 110.

孤立森林算法优化的 XGBoost 对养殖废水处理水质预测建模

邓志成¹, 万金泉^{1,*}, 王 艳¹, 朱 斌², 吴昌政¹, 吉世明²

(1. 华南理工大学环境与能源学院, 广东广州 510006; 2. 广东顺控自华科技有限公司, 广东佛山 528300)

摘 要 为解决水质软测量过程中数据失真问题, 研究采用孤立森林(isolation forest, IF)算法对水质传感器在线监测数据进行异常值处理, 使用递归特征消除(recursive feature elimination, RFE)优化模型变量选择, 采用 XGBoost 算法构建水质预测模型, 用于预测经处理后养殖鱼塘尾水出水化学需氧量(COD_{Cr})、总氮(TN)和总磷(TP)。试验表明, XGBoost 算法构建的生物净化池 COD_{Cr}、TN 和 TP 水质预测模型具有良好的预测性能, 各模型决定系数(R^2)分别达到了 0.837、0.804 和 0.878, 平均绝对误差(mean absolute error, MAE)分别为 0.679、0.087 和 0.036, 均方根误差(root mean square error, RMSE)分别为 0.700、0.105 和 0.044。同时, 使用 IF 算法对采集到的数据进行异常值识别与剔除后, 模型的 R^2 提升至 0.875、0.866 和 0.926, MAE 降低至 0.658、0.077 和 0.028, RMSE 降低至 0.681、0.099 和 0.035。研究对于发展水质智能软测量技术具有重要的指导价值。

关键词 机器学习 孤立森林 异常值检测 养殖尾水 水质预测

中图分类号: X714 文献标识码: A 文章编号: 1009-0177(2024)12-0096-08

DOI: 10.15890/j.cnki.jsjs.2024.12.011

Modeling Water Quality Prediction for Aquaculture Wastewater Treatment by Isolation Forest Algorithm-Optimized XGBoost

DENG Zhicheng¹, WAN Jinquan^{1,*}, WANG Yan¹, ZHU Bin², WU Changzheng¹, JI Shiming²

(1. College of Environment and Energy, South China University of Technology, Guangzhou 510006, China;

2. Guangdong Shunkong Zihua Technology Co., Ltd., Foshan 528300, China)

Abstract In order to solve the data distortion problem in the process of water quality soft measurement, this study adopted the isolated forest (IF) algorithm to process the abnormal value in the online monitoring data of water quality sensors, optimized the selection of model variables using recursive feature elimination (RFE). XGBoost algorithm was used to construct the water quality prediction model for predicting chemical oxygen demand(COD_{Cr}), total phosphorus (TP) and total nitrogen (TN) in the tailwater effluent of the treated farmed fish ponds. The experiments showed that the water quality prediction model for COD_{Cr}, TN and TP of the bio-purification pond constructed by the XGBoost algorithm had good prediction performance, and the coefficient of determination (R^2) of each model reached 0.837, 0.804 and 0.878, respectively, the MAE was 0.679, 0.087 and 0.036, and the RMSE was 0.700, 0.105 and 0.044, respectively. Meanwhile, after using IF algorithm to identify and remove the outliers of the collected data, the R^2 of the model was improved to 0.875, 0.866 and 0.926, the MAE decreased to 0.658, 0.077 and 0.028, and the RMSE decreased to 0.681, 0.099 and 0.035. This study has an important guiding value for the development of intelligent soft sensing technology of water quality.

Keywords machine learning isolation forest abnormal value detection aquaculture wastewater water quality prediction

[收稿日期] 2024-03-24

[基金项目] 广东省佛山市科技创新项目(2130218003140)

[作者简介] 邓志成(2000—), 男, 硕士, 研究方向为数据缺失及异常下的水质软测量, E-mail: 19828614518@163.com。

[通信作者] 万金泉(1965—), 男, 博士, 博士生导师, 研究方向为废水高效厌氧-好氧及高级氧化深度处理新理论与技术、废水处理系统的智能控制策略等, E-mail: ppjqwan@scut.edu.cn。

水质指标的快速准确获取是实现智慧水务的前提,随着人工智能技术的飞速发展,基于数据驱动的软测量技术能够借助快速、易测的辅助水质参数实时预测费时、难测的参数,可以解决人工检测耗时长,误差大且数据迟滞的问题。Huang 等^[1]于 2013 年首次提出了水质指标软测量的概念,并建立了一种基于混合模糊神经软测量模型,用于废水处理出水化学需氧量(COD_{Cr})的监测,结果表明,使用在线监测参数溶解氧(DO)和 pH 等,所建立的软测量模型就可以实现较高的 COD_{Cr} 精度测量。随后,Bagheri 等^[2]分别采用多层感知器和径向基函数(radial basis function, RBF)神经网络建立出水氨氮和总磷(TP)的预测模型,实现了出水氨氮浓度和总磷浓度的实时测量。Raduly 等^[3]基于典型的前馈神经网络建立了污水处理出水水质软测量模型,实现了对出水氨氮和出水总氮(TN)的同步测定。Huang 等^[4]提出了一种基于遗传算法的神经模糊系统的软传感器,用于实时生物废水处理过程中 COD_{Cr}、NO₃⁻ 和 PO₄³⁻ 浓度的预测,模拟结果表明,该模型具有良好的学习和泛化能力。刘博等^[5]利用 PCA-LSSVM 模型对厌氧塔出水挥发性脂肪酸(volatile fatty acid, VFA)进行软测量,通过使用主成分分析(principal component analysis, PCA)降低模型输入变量维数,随后基于 LSSVM 算法建立出水 VFA 软测量模型,结果表明,所建立的 PCA-LSSVM 软测量具有很好的仿真预测能力,最高决定系数(R^2)可达 0.999 6,实现出水 VFA 的准确预测。

但上述研究均面临建立软测量模型时原始数据失真共同的难题。软测量建模、模型训练和预测的精度都需要依赖于大量的基础实测原始数据,但快速水质在线测量电极会因污染、噪声信号等问题的存在而导致数据异常,这些异常值会影响所建立软测量模型的准确度。论文将适用于异常检测的孤立森林(IF)算法引入到 XGBoost 软测量预测模型中,通过对数据空间进行随机超平面拆分^[6],实现淡水鱼塘养殖尾水水质数据集中异常值的识别与剔除,以提高关键水质软测量模型预测精度。

1 试验

1.1 数据来源

研究使用的数据采集自广东佛山某养殖废水处理站点。通过在线水质传感器实时监测和人工取样

实验室检测 2 种方式,获取了 133 组由以下 12 个水质指标数据构成的数据集:沉淀池进水处的混合液悬浮固体(SS)、pH、温度(T)、COD_{Cr}、TP、TN、氨氮,藻菌池菌藻模块处的水中 DO 浓度,以及生物净化池出水处的 COD_{Cr}、TP、TN 和氨氮。其中 COD_{Cr}、TP、TN 和氨氮的数据为人工取样实验室检测所得,其余水质指标均为在线水质传感器实时读取的数据。

1.2 数据预处理

在建模过程中,由于不同特征变量的维度和维度单位之间存在较大差异,需要对数据进行归一化处理,从而减少和消除变量之间维度和量纲的影响,使得数据能够在相同的尺度上进行比较和分析。本研究采用最大-最小归一化,将处理后的数据限制为 $[0,1]$,其数学表达式如式(1)。

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中: x ——特征变量的原始数据;

x_{norm} ——特征变量归一化后的数据;

x_{\max} 和 x_{\min} —— x 中的最大值和最小值。

采用 R^2 、均方根误差(RMSE)和平均绝对误差(MAE)对训练好的模型进行性能评估,其数学表达式如式(2)~式(4)。

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (2)$$

$$R_{\text{RMSE}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (3)$$

$$M_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (4)$$

其中: R_{RMSE} ——RMSE 值;

M_{MAE} ——MAE 值;

y_i ——变量的第 i 个实际值;

\hat{y}_i ——第 i 个预测值;

\bar{y} ——变量的平均值;

N ——变量实际值的数量。

一般而言, R^2 是评估回归模型拟合优度的指标,用于衡量模型预测过程中总体数据的方差比和相关度,而 RMSE 和 MAE 是误差指标,用于量化模

型预测值与实际值的数值差距。

1.3 XGBoost 及递归特征消除 (RFE) 输入特征选择

XGBoost 是 Chen 等^[7]提出的一种可扩展的树机器学习系统。与普通的决策树算法相比,该算法的主要思想是通过不断地添加 1 棵新树来拟合前 1 棵树的预测结果与实际值之间的残差,从而形成 1 个新的模型,并将新模型作为下一步模型学习的基础,通过逐步减少预测错误,模型能够学习到数据之间的复杂关系,模型的最终预测结果由多棵树的累加得分所决定^[8]。XGBoost 在模型中内置了正则化参数,可以有效地防止模型发生过拟合,并且 XGBoost 算法对超参数的选择具有一定的容错性,即使参数设置不是最优的,也能获得不错的结果。

RFE 是一种特征变量选择算法,该算法通过逐步增加或移除特定特征变量从而获得能最大化模型性能的最优变量组合^[9],其伪代码如图 1 所示。在确定好模型算法后,使用所有特征变量训练模型,计算每个特征变量的重要性并进行排序,移除排名最低的特征,从而减小特征集合的大小。剩下的特征变量形成新的特征子集 S_i ,对每一个变量子集 S_i ($i=1, \dots, S$),使用减少后的特征集合重新训练模型,再次计算每个特征变量的重要性并进行排序。重复上述步骤,每次迭代都在减小的特征集合上重新训练模型,直到达到预定的特征数量或达到性能指标^[10]。

算法: 递归特征消除

- 1.1 使用所有特征在训练集上调优/训练模型
- 1.2 计算模型性能
- 1.3 计算变量重要性或排名
- 1.4 对于每个子集大小 S_i , $i=1, \dots, S$, 执行以下操作:
- 1.5 保留最重要的 S_i 个变量
- 1.6 [可选] 对数据进行预处理
- 1.7 使用 S_i 个预测变量在训练集上调优/训练模型
- 1.8 计算模型性能
- 1.9 [可选] 重新计算每个预测变量的排名
- 1.10 结束
- 1.11 计算 S_i 上的性能曲线
- 1.12 确定合适的预测变量数量

图 1 RFE 原理的伪代码实现

Fig. 1 Pseudo-Code Implementation of RFE Principle

1.4 IF

IF 是一种用于异常检测的机器学习算法,结构如图 2 所示,其基本原理是通过构建随机划分树来寻找异常值。算法通过随机选择数据特征和在特征范围内的随机划分点,逐步构建二叉树 (iTree),样本在树中的路径长度被用来度量其异常程度,较短的平均路径长度表示较高的异常程度。划分过程是随机的,正常样本需要多次的划分才能被孤立,相比正常样本,异常值的数量较少,特征值差异比较大,这使得异常值通常更容易被随机划分到较少的子空间,因此在树中具有较短的路径^[11]。通过随机构建多棵 iTree,并计算每个样本在这些树中的平均路径长度,最终将异常得分较高的样本视为异常值。与其他一些需要精细调整参数的算法相比,IF 算法通常不需要太多的参数调整,并且能够通过随机性来较好地处理含有噪声的数据。此外,IF 算法在处理不平衡数据集时表现良好,尤其是在数据集中异常值比例较低的情况下。

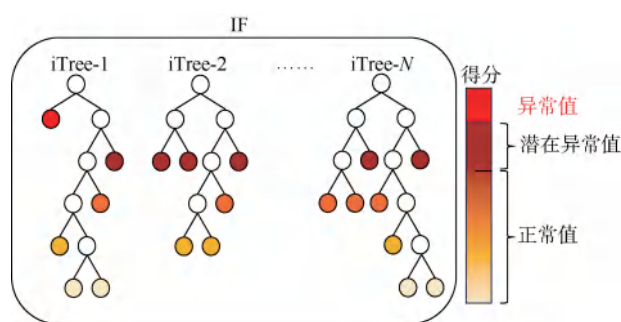


图 2 IF 结构

Fig. 2 Structure of IF

2 结果与讨论

2.1 水质数据分析

各水质指标的数值大小与特征如表 1 和图 3 所示。在水质参数分析过程中,将标准差与平均值的比值称作变异系数 (CV),用来反映数据离散程度的绝对值^[12]。除了 DO、进水 pH 和进水 TP 之外,大部分数据的 CV 都集中在 30% ~ 60%,甚至大于 60%。这表明这些数据是高度离散的,具有很强的非线性特征,这与实际废水处理过程中的水质特性是一致的^[13]。

2.2 水质预测模型性能分析

本研究选择 XGBoost 算法分别建立出水 COD_{Cr} 、TN 和 TP 的水质预测模型,由于进水 SS、进

表 1 各水质指标的单位、最大最小值、平均值、标准差和变异系数

Tab. 1 Unit, Maximum and Minimum, Mean, Standard Deviation and Coefficient of Variation of Each Water Quality Index

水质指标	单位	最小值	最大值	平均值	标准差	CV
DO	mg/L	3.44	8.58	5.01	0.945	18.9%
进水 pH 值	无量纲	6.48	8.08	7.52	0.164	2.18%
进水 SS	mg/L	5.00	347	48.7	48.7	104%
进水 SS	mg/L	1.00	9.20	4.58	2.15	46.9%
进水 COD _{Cr}	mg/L	13.2	145	26.7	13.5	50.7%
出水 COD _{Cr}	mg/L	3.43	25.4	5.68	2.26	39.7%
进水氨氮	mg/L	0.539	2.34	1.20	0.373	31.2%
出水氨氮	mg/L	0.040	0.720	0.174	0.121	69.9%
进水 TN	mg/L	1.50	7.78	3.56	1.14	32.0%
出水 TN	mg/L	0.445	2.50	1.13	0.378	33.5%
进水 TP	mg/L	0.363	1.36	0.638	0.142	22.3%
出水 TP	mg/L	0.090	0.47	0.270	0.089	33.0%

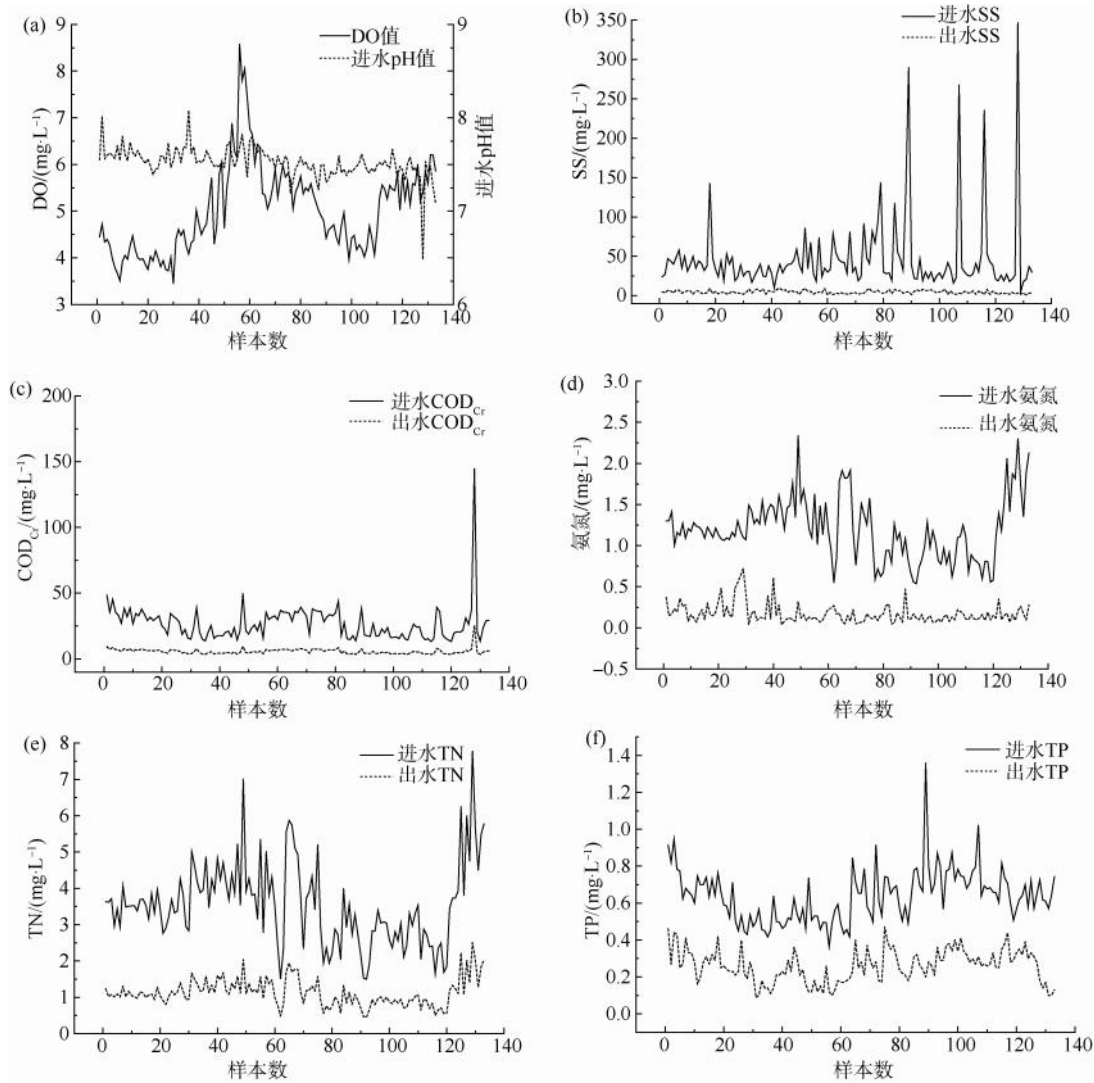


图 3 各水质指标的数值特征

Fig. 3 Numerical Characteristics of Each Water Quality Index

水 pH、进水 T 、进水 COD_{Cr} 、进水 TP、进水 TN、进水氨氮和 DO 均可作为模型的输入特征,为减少特征维数,降低模型计算成本的同时最大化模型预测性能,使用 RFE 算法进行模型的输入特征选择,使用 10 折交叉验证来测试具有最佳特征子集的模型,即数据集进行 10 次训练数据集和测试数据集的划分。对于每次运行,整个数据集被分为 10 个子样本,其中 9 个子样本用作训练数据来训练模型,其余子样本用作测试数据来评估模型。模型在 10 个测试数据集上的平均 R^2 被作为最终的模型性能指标。将模型 $R^2 > 0.8$ 作为选择最佳特征子集的迭代停止的性能指标。最终确定输入特征为 DO 和进水 SS、进水 COD_{Cr} 、进水 TP、进水 TN。由于数据集较小,为了保证有足够的数量用于模型训练和训练后模型性能检验,将数据集的数据按照 8 : 2 的比例随机划分为训练集(106 组)和测试集(27 组),为避免数据泄露导致的模型泛化能力的高估,测试集上特征数据的归一化将根据训练集上对应的特征数据的最大最小值进行。模型在训练集上采用 10 折交叉验证,不再单独设置验证集。为保证模型性能检验的统一性,同时考虑到所建模型在实际应用过程中异常值的不可避免性,故只考虑针对训练集数据使用 IF 算法进行异常值检测与剔除。IF 算法对训练集的数据进行拟合与异常得分计算,

随后对各特征数据的异常得分进行排序,按照默认设定的决策阈值确定出异常值数量,将异常得分较大的前 11 组数据确认为异常数据并对其进行标记,去除这些异常值后,新训练集有 95 组数据。模型在训练过程中,先使用随机搜索算法缩小模型超参数的寻优范围,然后使用网格搜索算法确定模型最优超参数组合,网格搜索范围和各模型的最佳超参数组合如表 2 所示。图 4 展示了异常值去除前后,各模型在测试集上的表现,具体的性能指标参数如表 3 所示。在对训练集去除异常值前,出水 COD_{Cr} 、TN、TP 模型的 R^2 分别达到了 0.837、0.804 和 0.878,MAE 分别为 0.679、0.087 和 0.036, RMSE 分别为 0.700、0.105 和 0.044,使用 IF 算法对训练集的数据进行异常值识别与剔除后,模型的 R^2 提升至 0.875、0.866 和 0.926,MAE 降低至 0.658、0.077 和 0.028, RMSE 降低至 0.681、0.099 和 0.035,模型 R^2 平均提升了 4.93%, MAE 平均降低了 0.013, RMSE 平均降低了 0.0113,模型性能得到明显提升。在所有关键出水水质预测模型中,出水 TP 模型的预测性能最好,这可能是因为与其他关键进出水水质指标相比,进水 TP 和出水 TP 的变异系数较低,模型的非线性特征相对较弱,输入特征与输出特征之间潜在关系易于模型捕获与学习。

表 2 各超参数的网格搜索范围和最优值

Tab. 2 Grid Search Range and Optimal Value of Each Hyperparameter

超参数	网格搜索范围			最佳超参数值		
	出水 COD_{Cr}	出水 TN	出水 TP	出水 COD_{Cr}	出水 TN	出水 TP
n_estimators	[60, 190]	[70, 160]	[75, 225]	140	120	160
max_depth	[3, 12]	[5, 16]	[5, 20]	8	12	13
subsample	[0.5, 1.0]	[0.5, 1.0]	[0.5, 1.0]	0.7	0.8	0.6
colsample_bytree	[0.5, 1.0]	[0.5, 1.0]	[0.5, 1.0]	0.8	0.75	0.5
learning_rate	[0.01, 0.11]	[0.03, 0.15]	[0.03, 0.15]	0.07	0.1	0.09
reg_lambda	[0, 3]	[0, 3]	[0, 3]	0.3	1.8	2.7
reg_alpha	[0, 3]	[0, 3]	[0, 3]	0	0.3	0.1

2.3 输入特征重要性分析

对于训练好的模型而言,通过计算每个输入特征在模型中被用来分裂的次数,可以得到各个特征的权重,用以衡量输入特征的重要性。通过分析具有较高重要性的特征变量,有助于理解输入特征之间和输入特征与模型输出之间的联系,增加模型的可解释性,帮助人类理解模型的内部机制以及对模型结果做出解释。图 5 展示了出水 COD_{Cr} 、出水 TN

和出水 TP 预测模型各输入特征的重要性。对于出水 COD_{Cr} 预测模型而言,进水 COD_{Cr} 、DO 和进水 TN 是最重要的 3 个特征变量;对于出水 TN 预测模型而言,进水 TN、DO 和进水 SS 是最重要的 3 个特征变量;对于出水 TP 预测模型而言,进水 TP、进水 COD_{Cr} 和进水 TN 是最重要的 3 个特征变量。进水 COD_{Cr} 、进水 TN 和进水 TP 在各模型的特征变量中具有最高重要性,这表示相较于其他输入特征,进水

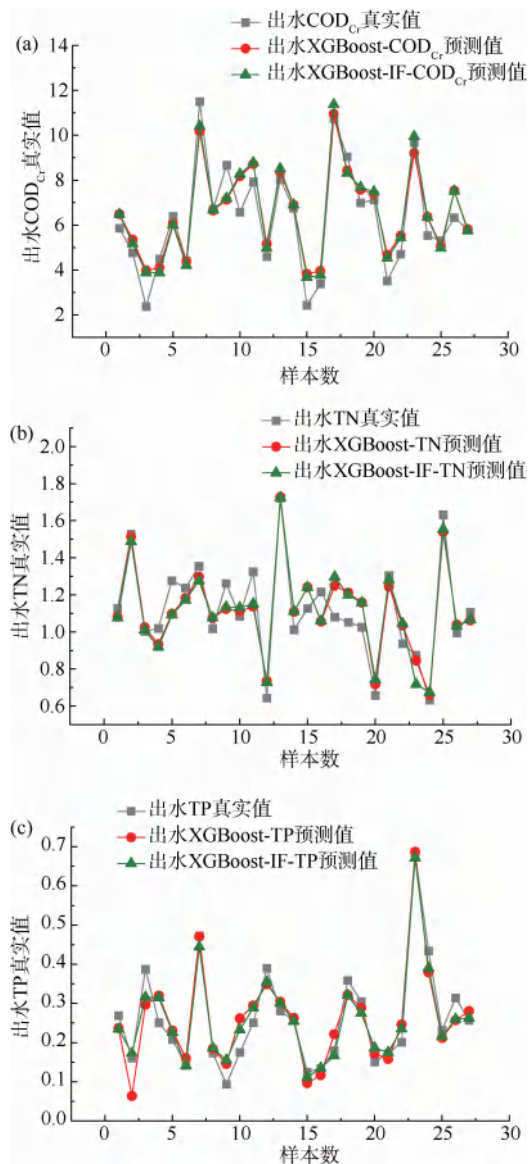


图 4 异常值去除前后模型在测试集上的表现
Fig. 4 Performance of Model on Test Set before and after Abnormal Value Removal

表 3 异常值去除前后模型在测试集上的性能表现

Tab. 3 Performance of Model on Test Set before and after Abnormal Value Removal

出水水质指标	XGBoost			IF-XGBoost		
	R^2	MAE	RMSE	R^2	MAE	RMSE
出水 COD_{Cr}	0.837	0.679	0.700	0.875	0.658	0.681
出水 TN	0.804	0.087	0.105	0.866	0.077	0.099
出水 TP	0.878	0.036	0.044	0.926	0.028	0.035

活性的强弱对于 TP 的去除至关重要。藻菌模块中细菌等微生物生长、繁殖和代谢所需的能量依赖于水中的有机碳和氮元素^[18], 鱼塘养殖尾水中 COD_{Cr} 和 TN 是维持微生物生命活动的关键要

COD_{Cr} 、进水 TN 和进水 TP 的数值大小对该模型的预测结果具有更高的影响力, 这是因为任何工艺对于污染物的去除都是有限的, 进水 COD_{Cr} 、进水 TN 和进水 TP 的浓度一定程度上决定了出水 COD_{Cr} 、出水 TN 和出水 TP 的浓度。

对于出水 COD_{Cr} 预测模型, 剩下的 2 个重要特征变量是 DO 和进水 TN。藻菌池 DO 的高低反映了池中藻菌的活性^[14], 与水中低 DO 条件相比, 高 DO 的条件会促进藻菌模块中好氧菌氧化分解水中有机物质的, 对出水 COD_{Cr} 的数值大小产生直接影响。同时, 有机物的氧化分解与 TN 中氨氮和亚硝酸氮的氧化还原过程会相互影响, 氨氮和亚硝酸氮被微生物氧化为硝酸盐的过程中会消耗氧气, 会与有机物的氧化分解存在一定的氧竞争, 从而对出水 COD_{Cr} 造成一定影响。

对于出水 TN 预测模型而言, 剩下的 2 个重要特征变量是 DO 和进水 SS。水中的 TN 主要由有机氮、氨氮、亚硝酸氮和硝酸氮组成, 其中氨氮和亚硝酸氮在藻菌池发生硝化反应时会被转化为硝酸氮, 同时消耗水中氧气, 而硝酸氮在生物净化池中通过植物根部的主动吸收能力实现部分去除^[15], 进而对出水 TN 造成影响。鱼塘养殖尾水中 SS 的主要成分包括鱼类粪便和未被消化吸收的养殖饲料, 这两者通常是水中 TN 的主要来源之一^[16-17], 因此进水 SS 浓度的高低在一定程度上间接反映了进水 TN 的高低。

对于出水 TP 预测模型而言, 剩下的 2 个重要特征变量是进水 COD_{Cr} 和进水 TN。在该鱼塘养殖废水处理站点中, 水中 TP 的去除主要依靠藻菌池中的藻菌模块, TP 中的有机磷通过细菌等微生物降解转化为无机磷, 无机磷则会被藻类吸收固定, 藻菌

素。同时, TN 中的亚硝酸盐和硝酸盐是藻类的生长和繁殖不可或缺的重要氮源。因此, 进水 COD_{Cr} 和进水 TN 通过影响菌藻模块从而间接影响出水 TP。

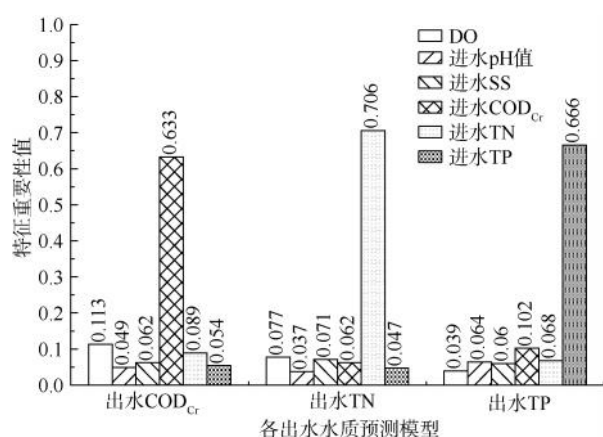


图5 各模型的输入特征重要性

Fig. 5 Importance of Input Features of Each Model

3 结论

本工作研究了 IF 去除异常值前后, 基于 XGBoost 算法建立的水质指标预测模型的性能变化, 主要得到以下结论。

(1) 使用 RFE 算法可以有效地消除特征之间的冗余, 降低输入特征维数, 在保证模型性能符合预期设定的情况下, 将输入特征从 8 个减少至 6 个, 降低了关键出水水质指标预测模型结构的复杂性, 增加了模型的可解释性。

(2) IF 算法可以有效地识别并剔除数据集中的异常值, 在对训练集进行异常值剔除后, 模型的预测性能得到明显提升, 优化后的出水 COD_{Cr}、出水 TN 和出水 TP 预测模型的 R^2 提升至 0.875、0.866 和 0.926, MAE 降低至 0.658、0.077 和 0.028, RMSE 降低至 0.681、0.099 和 0.035。异常值往往会导致数据特征的分布失真。在剔除这些异常值后, 模型能学习到更加接近真实数据的分布, 利于挖掘特征之间的潜在联系, 使得模型能够更准确地捕捉数据的本质特征。本研究所建立的出水 COD_{Cr}、出水 TN 和出水 TP 等关键出水水质指标预测模型可以为鱼塘养殖尾水及其他污水的处理排放提供一定数据指导。

尽管本工作利用实际水质数据建立了出水水质预测模型, 但所建立的模型在面对输入特征存在缺失值时可能会导致模型性能下降, 限制其实际应用能力。在未来的工作中, 需要探究输入缺失处理方法, 尽可能降低由缺失值导致的模型的预测性能下降。

参考文献

- [1] HUANG M, WAN J, HU K, et al. Enhancing dissolved oxygen control using an on-line hybrid fuzzy-neural soft-sensing model-based control system in an anaerobic/anoxic/oxic process [J]. Journal of Industrial Microbiology and Biotechnology, 2013, 40 (12): 1393–1401.
- [2] BAGHERI M, MIRBAGHERI S A, EHTESHAMI M, et al. Modeling of a sequencing batch reactor treating municipal wastewater using multi-layer perceptron and radial basis function artificial neural networks [J]. Process Safety and Environmental Protection, 2015, 93: 111–123. DOI: 10.1016/j.psep.2014.04.006.
- [3] RADULY B, GERNAEY K V, CAPODAGLIO A G, et al. Artificial neural networks for rapid WWTP performance evaluation: Methodology and case study [J]. Environmental Modelling & Software, 2007, 22(8): 1208–1216.
- [4] HUANG M, MA Y, WAN J, et al. A sensor-software based on a genetic algorithm-based neural fuzzy system for modeling and simulating a wastewater treatment process [J]. Applied Soft Computing, 2015, 27: 1–10. DOI: 10.1016/j.asoc.2014.10.034.
- [5] 刘博, 万金泉, 黄明智, 等. 基于 PCA-LSSVM 的厌氧废水处理系统出水 VFA 在线预测模型 [J]. 环境科学学报, 2015, 35(6): 1768–1778.
- [6] LIU B, WAN J Q, HUANG M Z, et al. A PCA-LSSVM model for on-line prediction of the effluent VFA in an anaerobic wastewater treatment system [J]. Acta Scientiae Circumstantiae, 2015, 35(6): 1768–1778.
- [7] SHAO C, DU X, YU J, et al. Cluster-based improved isolation forest [J]. Entropy, 2022, 24 (5): 611. DOI: 10.3390/e24050611.
- [8] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system [C]. San Francisco: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
- [9] LI J, AN X, LI Q, et al. Application of XGBoost algorithm in the optimization of pollutant concentration [J]. Atmospheric Research, 2022, 276: 106238. DOI: 10.1016/j.atmosres.2022.106238.
- [10] 罗红郊, 马晓琴, 孙妍, 等. 基于 RFE 特征选择的 PSO-SVM 用电量预测算法 [J]. 电子设计工程, 2023, 31(20): 172–176.
- [11] LUO H J, MA X Q, SUN Y, et al. PSO-SVM electricity consumption prediction algorithm based on RFE feature selection [J]. Electronic Design Engineering, 2023, 31(20): 172–176.
- [12] CHEN X, JEONG J C. Enhanced recursive feature elimination [C]. Cincinnati: Sixth International Conference on Machine Learning and Applications (ICMLA 2007), 2007.

(下转第 110 页)

- [12] SKLENICKOVA K, PECENKA M, ŘIHOVA AMBROZOVA J, et al. Influence of biodegradable polyurethane foam on biocenosis and sludge activity in reactors simulating low-load wastewater treatments[J]. *Journal of Water Process Engineering*, 2021, 44: 102455. DOI: 10.1016/j.jwpe.2021.102455.
- [13] 黎镛. 微生物载体对 MBBR 工艺性能及微生物群落结构的影响研究[D]. 长沙: 湖南大学, 2022.
- LI Y. Effects of microbial carriers on MBBR process performance and microbial community structure [D]. Changsha: Hunan University, 2022.
- [14] MAUCLAIRE L, SCHURMANN A, MERMILLOD-BLONDIN F. Influence of hydraulic conductivity on communities of microorganisms and invertebrates in porous media: A case study in drinking water slow sand filters[J]. 2006, 68, 100–108. DOI: 10.1007/s00027-005-0811-4.
- [15] 向钰. 不同溶解氧同步硝化反硝化脱氮效能及途径[D]. 重庆: 重庆大学, 2020.
- XIANG Y. Nitrogen removal capacity and pathways of the simultaneous nitrification and denitrification process under different oxygen conditions [D]. Chongqing: Chongqing University, 2020.
- [16] 杨龙, 詹旭, 周恩华, 等. 改性 HDPE 填料的制备及对 MBBR 脱氮除磷效果的强化[J]. *环境工程学报*, 2023, 17(5): 1674–1682.
- YANG L, ZHAN X, ZHOU E H, et al. Preparation of modified HDPE carrier and its strengthening effect on nitrogen and phosphorus removal of MBBR [J]. *Chinese Journal of Environmental Engineering*, 2023, 17(5): 1674–1682.
- [17] 袁维芳, 邱仲业, 王浩, 等. MBBR 工艺原理、填料特性及工程化应用进展[J]. *轻工科技*, 2023, 39(2): 78–81.
- YUAN W F, QIU Z Y, WANG H, et al. MBBR process principle, filler characteristics and engineering application progress[J]. *Light Industry Science and Technology*, 2023, 39(2): 78–81.

(上接第 102 页)

- [11] LIU J, WANG P, JIANG D, et al. An integrated data-driven framework for surface water quality anomaly detection and early warning [J]. *Journal of Cleaner Production*, 2020, 251: 119145. DOI: 10.1016/j.jclepro.2019.119145.
- [12] HANEMANN M, KANNINEN B. The statistical analysis of discrete-Response CV data [M]. Oxford: Oxford University Press, 2001.
- [13] HAN H G, ZHANG H J, LIU Z, et al. Data-driven decision-making for wastewater treatment process[J]. *Control Engineering Practice*, 2020, 96: 104305. DOI: 10.1016/j.conengprac.2020.104305.
- [14] 许华一, 李姗姗, 韦静, 等. 菌藻共生系统实现短程硝化工艺的藻供氧条件和影响因素分析[J]. *环境工程*, 2024, 42(5): 42–52.
- XU H Y, LI S W, WEI J, et al. Study on oxygen supply condition and influence of algal in partial nitrification process in a bacteria and algae symbiotic system [J]. *Environmental Engineering*, 2024, 42(5): 42–52.
- [15] 康军利. 人工湿地对二级出水中 TN、TP 去除效果的季节性研究[J]. *安全与环境学报*, 2012, 12(1): 40–44.
- KANG J L. Seasonal effects of the artificial wetlands on the removing of nitrogen and phosphorus content in secondary effluent [J]. *Journal of Safety and Environment*, 2012, 12(1): 40–44.
- [16] 刘梅, 原居林, 倪蒙, 等. “三池两坝”多级组合工艺对内陆池塘养殖尾水的处理[J]. *环境工程技术学报*, 2021, 11(1): 97–106.
- LIU M, YUAN J L, NI M, et al. Treatment of inland pond aquaculture tail water by multi-stage combined process of “three ponds and two dams” [J]. *Journal of Environmental Engineering Technology*, 2021, 11(1): 97–106.
- [17] LIU M, LIAN Q, ZHAO Y, et al. Treatment effects of pond aquaculture wastewater using a field-scale combined ecological treatment system and the associated microbial characteristics[J]. *Aquaculture*, 2023, 563: 739018. DOI: 10.1016/j.aquaculture.2022.739018.
- [18] WU B, RAN T, LIU S, et al. Biofilm bioactivity affects nitrogen metabolism in a push-flow microalgae-bacteria biofilm reactor during aeration-free greywater treatment [J]. *Water Research*, 2023, 244: 120461. DOI: 10.1016/j.watres.2023.120461.