AI 与智慧水务

李祥, 孙韶华, 刘帅, 等. 基于优化 XGBoost 算法的黄河下游引黄水库富营养化预测[J]. 净水技术, 2025, 44(9): 157-165.
LI X, SUN S H, LIU S, et al. Prediction of diversion reservoir cutrophication in Yellow River downstream based on optimized XGBoost algorithm [J]. Water Purification Technology, 2025, 44(9): 157-165.

基于优化 XGBoost 算法的黄河下游引黄水库富营养化预测

李 祥,孙韶华,刘 帅,马中雨,王明泉,宋武昌,陈发明,李桂芳,贾瑞宝*
(山东省城市供排水水质监测中心,山东济南 250011)

摘 要 **【目的】** 为预测研究黄河下游引黄水库富营养化风险,助力提升水源水质监测预警和应急响应能力,文章开展了麻雀搜索耦合极限梯度提升算法(SSA-XGBoost)的模型构建研究和应用。**【方法】** 文章以济南市 2 座典型引黄水库为研究对象,利用其 2013 年—2022 年的水质历史数据和同期气象数据,针对其高总氮、高氮磷比、高藻等水质污染特征,采用计算效率高、预测性能优秀的极限梯度提升算法(XGBoost)并使用麻雀搜索算法(SSA)对其 4 个超参数进行寻优,以影响富营养化的水体理化性质、营养盐以及太阳光照等关键因子,如水温、pH、溶解氧、高锰酸盐指数、浑浊度、总磷、总氮、氨氮、硝酸盐、氮磷比、7 d 日照时数均值、7 d 太阳辐射总量均值 12 个指标为模型输入变量,以表征藻类生物量的重要指标叶绿素 a 为输出变量,构建了适于济南市 2 座引黄水库的水体富营养化预测预警模型。**【结果】** SSA-XGBoost 富营养化预测模型的均方根误差(RMSE)为 4.25 μg/L,平均绝对误差(MAE)为 3.19 μg/L,拟合优度(R^2)为 0.77,其预测精度优于 BP 神经网络模型和支持向量机模型,叶绿素 a 等级预测准确率可达 85%以上,对 2 座引黄水库叶绿素 a 预测结果影响最大的变量为 pH,其次为硝酸盐,再者为高锰酸盐指数。**【结论】** 总体上,SSA-XGBoost 富营养化预测模型精度较高、性能良好、实用性强,模型的构建研究和应用将为 2 座引黄水库藻类风险预测预警提供参考和依据。

关键词 引黄水库 SSA-XGBoost 富营养化 叶绿素 a 预测预警
中图分类号: TU991 **文献标志码**: A **文章编号**: 1009-0177(2025)09-0157-09
DOI: 10.15890/j.cnki.jsjs.2025.09.020

Prediction of Diversion Reservoir Eutrophication in Yellow River Downstream Based on Optimized XGBoost Algorithm

LI Xiang, SUN Shaohua, LIU Shuai, MA Zhongyu, WANG Mingquan, SONG Wuchang, CHEN Faming, LI Guifang, JIA Ruibao*
(*Shandong Province City Water Supply and Drainage Water Quality Monitoring Center, Jinan 250011, China*)

Abstract [**Objective**] In order to study and predict the risk of eutrophication in the water of the Yellow River diversion reservoir in the lower reaches of the Yellow River, and further improve the monitoring, early warning, and emergency response capabilities of water source quality, a model construction research and application of the sparrow search coupled extreme gradient boosting algorithm (SSA-XGBoost) were carried out. [**Methods**] Taking two typical Yellow River diversion reservoirs in Jinan City as the research objects, using their water quality historical data and meteorological data from 2013 to 2022, targeting their high total nitrogen, high nitrogen phosphorus ratio, high algae and other water pollution characteristics, the XGBoost algorithm with high computational efficiency and excellent predictive performance was adopted, and the sparrow search algorithm was used to optimize their four parameters. The key

[收稿日期] 2024-09-27
[基金项目] 国家重点研发计划(2021YFC3200904);山东省重大科技创新工程项目(2020CXGC011406);国家水体污染控制与治理科技重大专项(2017ZX07502002)
[作者简介] 李祥(1985—),男,主要从事水质监测预警研究的工作,E-mail:sdsflixiang@163.com。
[通信作者] 贾瑞宝(1968—),男,主要从事水质检测、水处理等研究的工作,E-mail:jiaruibao68@126.com。

factors affecting eutrophication, such as physical and chemical properties, nutrients, and solar radiation, including water temperature, pH, dissolved oxygen, permanganate indicator, turbidity, total phosphorus, total nitrogen, ammonia nitrogen, nitrate, nitrogen phosphorus ratio, 7 day average sunshine hours, 7 day average total solar radiation, and other 12 indices were used as input variables for the model. A prediction and early warning model for water eutrophication suitable for two Yellow River water diversion reservoirs in Jinan City had been established. [**Results**] The root mean square error (RMSE) of the SSA XGBoost eutrophication prediction model was 4.25 μg/L, the average absolute error (MAE) was 3.19 μg/L, and the goodness of fit (R^2) was 0.77. The prediction accuracy was better than that of the BP neural network model and support vector machine model, and the accuracy of chlorophyll a level prediction could reach over 85%. The variable that had the greatest impact on the chlorophyll a prediction result of the two Yellow River diversion reservoirs was pH, followed by nitrate, and then permanganate index. [**Conclusion**] Overall, the SSA XGBoost eutrophication prediction model has high accuracy, good performance, and strong practicality. The construction and application research of the model will provide reference and basis for algae risk prediction and early warning in two Yellow River water diversion reservoirs.

Keywords diversion reservoir of Yellow River　SSA-XGBoost　eutrophication　chlorophyll a　prediction and early warning

水体富营养化是水生生物对水中氮、磷等主要营养盐浓度增加的响应及变化过程,后期可能会引起蓝藻水华和水体缺氧等水质恶化现象[1]。作为当前水环境研究的热点问题之一,国内外学者[2-4]开展了大量的研究,主要集中在藻类水华发生的机理、影响因素以及藻类生长预测等方面。藻类生长及水华的发生具有复杂性、非线性、时变性等特点,对于水华预测,传统的机理模型虽然取得了较好的效果,但也由于计算过程复杂、影响因素众多,而存在一定的使用困难。近年来,随着人工智能和大数据技术的快速发展,模拟智能算法在水华预测建模中显出较强的优势,其中,BP 神经网络和支持向量机(SVM)等算法应用最为广泛[4-7]。BP 神经网络虽具有较强的非线性映射能力、容错性和自适应等特性,但受原理上的一些局限,模型的精度和泛化能力受到一定程度的影响。SVM 虽具有全局最优、避免维数灾、泛化能力强、小样本等优点[7],但也存在大规模训练样本计算速度慢、对模型参数和核函数较为敏感等缺点。极限梯度提升算法(XGBoost)因其具有支持并行处理、最优解收敛速度快、避免过拟合等优点,在预测性能和计算效率上表现优秀,自问世以来已在工业互联网、医疗卫生等领域获得广泛应用[8]。但目前其在湖库富营养化预测方面的应用还较少,鉴于此,本研究将 XGBoost 算法引入到水体富营养化预测预警中并探索其应用价值。叶绿素 a 浓度作为表征藻类生物量的重要指标,在水体富营养化评价中起关键作用,能够用于判断湖库水体藻类浓度状态及水华的发生情况[7]。本研究以黄河下游济南市两座引黄水库为研究对象,以水体叶绿素 a 浓度作为富营养化程度的表征,基于 XGBoost 算法并使用麻雀搜索算法(sparrow search algorithm,SSA)对其进行优化改进,以期进一步提高 XGBoost 算法模型计算效率和精度。最后通过与 BP 神经网络模型和 SVM 模型进行比较,筛选出最优的济南市 2 座引黄水库水体富营养化预测预警模型,模型的构建和研究将为水体富营养化的预测预警及风险管理提供参考和依据。

1　材料与方法

1.1　研究区概况

济南市南依泰山,北跨黄河,位于黄河下游地区,属于暖温带大陆性季风气候,季风明显,四季分明,多年平均气温为 14.3 ℃,多年平均降水量为 665.7 mm。该市的水库承担了全市 75%的城市用水,水库按水源可分为引黄水库和山区水库 2 种[9],为保障工农业生产和生活饮用水发挥了关键作用。其中,引黄水库主要包括鹊山水库和玉清水库 2 座中型水库,占地面积分别为 7.4 km^2 和 6.1 km^2,库容分别为 4.600×10^7 m^3 和 4.850×10^7 m^3,设计供水能力均达 4.0×10^5 m^3/d。济南黄河段为“地上悬河”,2 座引黄水库均位于黄河沿岸,属于地上围坝平原水库,库区内地势平缓,水体相对较浅(<10 m),黄河水经过引黄闸提取、清污和沉沙后进入水库,流域内的地表径流不直接排入水库。2 座水库库区周边均分布着较密集的耕地,鹊山水库库区范围内常住人口超过 22 000 人,玉清水库还兼具生态补源功能,向小清河、腊山河进行生态补水,近年来为改善水库周边环境,周边新建了多

个湿地公园。2 座引黄水库同属于黄河下游的浅型水库,蓄集了黄河水预沉后的低浊度高营养盐澄清水,水体相对较浅,容易导致藻类疯长,水体富营养化风险较高,2010 年前后曾发生过蓝藻和硅藻水华。济南市还是南水北调东线工程的重要枢纽城市,自 2013 年通水以来,在调水周期内虽有少部分长江水会进入玉清水库,但从全年来看,其水质整体变化不大,依然具有典型的低浑浊度、高总氮、高藻等水质特点[10]。

1.2 藻类生长影响因子与数据源

藻类生长受水生态环境中水质理化性质以及营养盐等多种环境因子的影响,理化因子主要包括水温、pH、溶解氧、高锰酸盐指数、浑浊度等,营养盐因子主要包括总磷、总氮、氨氮、硝酸盐、氮磷比等,营养盐在藻类光合作用中参与藻细胞的合成[2,11]。本研究收集 2013 年—2022 年济南市供排水监测中心检测的两水库水质数据,每组水质数据除上述指标外还应包含叶绿素 a。2 座水库水质平均检测频次约为每月 1 次,采样点在水库进水口、出水口、水库中央、岸边等附近均有分布。因两水库冬季水温较低,叶绿素 a 含量也较低,剔除掉冬季及指标数据不全的数据组,共整理得到 185 组数据。各环境因子数值特征及其与叶绿素 a 含量的 Pearson 相关性分析结果分别如表 1 和表 2 所示。根据国家“十二五”水专项“南水北调山东受水区饮用水安全保障技术研究与综合示范”课题研究及水质检测结果,在南水北调调水期内虽有少部分长江水进入玉清水库进行了掺混,引起水库内硫酸盐和氯化物等指标浓度发生变化,但整体上,掺混后玉清水库还是以黄河水为主,且影响藻类生长的理化性质及氮、磷等营养盐因子。由表 1 可知,2 座引黄水库具有高氮、低磷、高氮磷比的水质特点,总氮均值为 2.72 mg/L,总磷均值为 0.021 mg/L,氮磷比均值为 172.5。研究[12]指出,当氮磷比大于 10 时,磷可以考虑为藻类生长的限制性营养盐。相关性分析结果显示,叶绿素 a 与温度、pH、高锰酸盐指数、氨氮、总磷、浑浊度呈极显著正相关($P<0.01$),相关系数为 0.248~0.594,这类因子可能为影响藻类生长的主要限制性因素;与总氮、硝酸盐、氮磷比呈极显著负相关($P<0.01$),相关系数为-0.576~-0.333,这是因为两水库中总氮主要为硝酸盐氮且氮磷比数值较高,在其高浓度或数值范围内是藻类生长的限制性因子[13];与溶解氧呈显著负相关($P<0.05$),相关系数为-0.194,表明水体中叶绿素 a 含量越高,则溶解氧含量越低。根据相关性分析结果,引黄水库叶绿素 a 含量与以上环境因子均具有显著或极显著的相关性,本研究将以上 10 个环境因子作为水体富营养化预测模型的输入因子。

藻类生长离不开光合作用,因此,本研究将直接作用于藻类生长和水华暴发的太阳光照因子也纳入水体富营养化叶绿素 a 预测预警模型的输入因子指标体系[7,14-15]。通过地理遥感生态网科学数据注册与出版系统(www.gisrs.cn)获取了济南市气象站的太阳光照数据(区站号为 54823),主要包括日照时数(SSD)和太阳辐射总量(I)2 个指标因子。按照水质检测日期从太阳光照数据集中提取 2 个指标因子对应的数据,2 个指标数据受天气影响较大并综合考虑藻类生长周期,本研究在构建回归预测模型时,将水样采集日期近 7 d 两指标因子的均值作为输入因子[7],近 7 d 两指标因子均值特征如表 1 所示。

表 1 各变量的数值特征

Tab. 1 Numerical Characteristics of Each Variable

项目	均值	最大值	最小值	标准差
温度/℃	17.6	28.8	1.0	7.1
pH 值	8.36	8.70	7.71	0.16
溶解氧/(mg·L^{-1})	8.8	12.3	6.4	1.2
高锰酸盐指数/(mg·L^{-1})	2.85	6.42	1.92	0.59
氨氮/(mg·L^{-1})	0.14	0.32	<0.02	0.07
总磷/(mg·L^{-1})	0.021	0.093	<0.010	0.010
总氮/(mg·L^{-1})	2.72	4.77	0.77	0.81
氮磷比	172.5	926.0	26.6	145.6
硝酸盐/(mg·L^{-1})	2.07	3.86	0.49	0.78
浑浊度/NTU	3.82	12.80	0.50	2.36
叶绿素 a/(μg·L^{-1})	10.1	60.2	1.5	9.0
7 d SSD 均值/h	6.35	11.7	1.67	2.25
7 d I 均值/[MJ·(m^2·d)$^{-1}$]	16.4	26.1	5.46	5.11

注:当变量数值小于检出限时,按检出限的 1/2 进行统计分析。

1.3 XGBoost 的原理及方法

1.3.1 XGBoost 的原理

XGBoost 的目标是创建 K 回归树,以使树组的预测值尽可能接近真值,并具有最大的泛化能力。其算法的关键是利用损失函数的二阶泰勒展开,并

表 2　指标间相关性分析结果
Tab. 2　Results of Correlation Analysis between Indicators

指标	温度	pH	溶解氧	高锰酸盐指数	氨氮	总磷	总氮	硝酸盐	浑浊度	氮磷比	叶绿素 a
温度	1	/	/	/	/	/	/	/	/	/	/
pH	0.261**	1	/	/	/	/	/	/	/	/	/
溶解氧	−0.748**	−0.176*	1	/	/	/	/	/	/	/	/
高锰酸盐指数	0.119	0.489**	−0.073	1	/	/	/	/	/	/	/
氨氮	−0.024	0.417**	−0.059	0.357**	1	/	/	/	/	/	/
总磷	−0.052	0.121	−0.007	0.411**	0.272**	1	/	/	/	/	/
总氮	−0.155	−0.208*	0.101	−0.317**	−0.031	−0.074	1	/	/	/	/
硝酸盐	−0.153	−0.306**	0.164*	−0.364**	−0.174*	−0.178*	0.871**	1	/	/	/
浑浊度	0.181*	0.329**	−0.211*	0.135	0.319**	0.127	−0.208*	−0.340**	1	/	/
氮磷比	0.003	−0.152	−0.024	−0.258**	−0.181*	−0.600**	0.498**	0.507**	−0.075	1	/
叶绿素 a	0.248**	0.594**	−0.194*	0.501**	0.425**	0.445**	−0.440**	−0.576**	0.408**	−0.333**	1

注:** 表示极显著相关($P<0.01$);* 表示显著相关($P<0.05$)。

加入正则化降低模型复杂度,避免过拟合,柴敬等[8]对其原理进行了阐述。

1.3.2　SSA 的原理

SSA 在 2020 年由 Xue 等[16]提出,它是一种新的群体智能优化算法,其灵感来源于麻雀觅食和反捕食行为,SSA 将麻雀种群划分为发现者和加入者,通过模拟麻雀的觅食过程,来搜索最优解,搜索空间得到有效减少,搜索效率也获得提高,因其具有寻优能力强、搜索速度快等优点,目前在机器学习领域的应用也越来越多[17]。SSA 基本原理和实现流程如下。

(1)麻雀位置初始化。

作为一种仿真试验,SSA 算法定义虚拟的麻雀位置与麻雀的食物,麻雀位置用矩阵 $\boldsymbol{X}$ 表示,如式(1)。

$$\boldsymbol{X}=\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \tag{1}$$

其中:n——麻雀的种群数量,个;

d——待优化变量的维数;

$x_{n,d}$——第 n 个麻雀在待优化变量的 d 维空间中的坐标位置。

种群中具有更高适应度值的生产者在搜寻过程中优先获取食物。种群中的生产者带领整个群体搜寻食物,所有麻雀的适应度可以用式(2)中的矩阵 $\boldsymbol{F}_x$ 表示,$\boldsymbol{F}_x$ 的每行值代表每个个体的适应值。

$$\boldsymbol{F}_x=\begin{bmatrix} f([x_{1,1} \quad x_{1,2} \quad \cdots \quad x_{1,d}]) \\ f([x_{2,1} \quad x_{2,2} \quad \cdots \quad x_{2,d}]) \\ \vdots \\ f([x_{n,1} \quad x_{n,2} \quad \cdots \quad x_{n,d}]) \end{bmatrix} \tag{2}$$

(2)更新发现者位置,更新方法策略如式(3)。

$$X_{i,j}^{t+1}=\begin{cases} X_{i,j}^{t}\cdot\exp\left(\dfrac{-i}{\alpha\cdot E_{\max}}\right), & R_2<S_{\mathrm{T}} \\ X_{i,j}^{t}+Q\cdot L, & R_2\geqslant S_{\mathrm{T}} \end{cases} \tag{3}$$

其中:t——当前迭代次数;

$X_{i,j}^{t}$——迭代 t 次时第 j 维的第 i 个麻雀;

$X_{i,j}^{t+1}$——迭代 $t+1$ 次时第 j 维的第 i 个麻雀;

$E_{\max}$——最大迭代次数;

α——(0,1]的均匀随机数;

Q——符合正态分布的随机数;

L——$1\times d$ 的单位列向量;

R_2——预警值,$R_2\in[0,1]$;

S_{T}——安全值,$S_{\mathrm{T}}\in[0.5,1.0]$。

如果 $R_2<S_{\mathrm{T}}$,周围没有捕食者,发现者进行全局搜索;如果 $R_2\geqslant S_{\mathrm{T}}$,麻雀发现捕食者,需快速转移到安全地带。

(3)更新加入者位置,更新方法策略如式(4)~式(5)。

$$X_{i,j}^{t+1}=\begin{cases}Q\cdot\exp\left(\dfrac{X_{w,j}^{t}-X_{i,j}^{t}}{i^{2}}\right), & i>\dfrac{N}{2}\\ X_{b,j}^{t+1}+\left|X_{i,j}^{t}-X_{b,j}^{t+1}\right|\cdot\boldsymbol{A}^{+}\cdot L, & i\leqslant\dfrac{N}{2}\end{cases}\tag{4}$$

$$\boldsymbol{A}^{+}=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\cdot\boldsymbol{A}^{\mathrm{T}})^{-1}\tag{5}$$

其中:$X_{w,j}^{t}$——第 t 轮更新时麻雀在 j 维度中的最差位置;

$X_{b,j}^{t+1}$——第 $t+1$ 轮更新时麻雀在 j 维中的最优位置;

$\boldsymbol{A}$——元素为-1 或 1 的 $1\times d$ 矩阵。

(4)更新预警麻雀位置,更新方法策略如式(6),预警麻雀占整个种群的比例一般为 10%~20%。

$$X_{i,j}^{t+1}=\begin{cases}X_{i,j}^{t}+k\cdot\left(\dfrac{\left|X_{i,j}^{t}-X_{w,j}^{t}\right|}{(f_i-f_w)+\delta}\right), & f_i=f_g\\ X_{b,j}^{t}+\beta\cdot\left|X_{i,j}^{t}-X_{b,j}^{t}\right|, & f_i>f_g\end{cases}\tag{6}$$

其中:k——[-1,1]内的随机数;

β——步长控制参数;

δ——非常小的常数,设为 1×10^{-8},避免分母为 0;

f_i——当前麻雀的适应度值;

f_g——全局最好的适应度值;

f_w——全局最差的适应度值。

当 $f_i>f_g$ 时,麻雀位于种群的边缘;当 $f_i=f_g$ 时,麻雀位于种群中心[17]。

1.3.3 模型构建及参数调优

将温度、pH、溶解氧、高锰酸盐指数、浑浊度、总磷、总氮、氨氮、硝酸盐、氮磷比、7 d SSD 均值、7 d I 均值 12 个影响水体叶绿素 a 含量的指标作为模型的输入因子,将叶绿素 a 作为模型的输出因子,构建基于 XGBoost 算法的水体富营养化预测预警模型。构建模型时,数据集的划分以及超参数的设置是影响模型性能的关键,按照 8∶2 的比例将所有数据集划分为训练集和测试集,再将训练集平均分成 5 个子集,采用具有可对模型进行评估、减小过拟合、解决数据不平衡等作用的 5 折交叉验证法(轮流将训练集中的 4 个子集作为训练子集,剩余 1 个子集作为验证集),以每次训练评估指标的平均值作为最终的模型性能评估指标。XGBoost 模型的超参数较多,其中,最大迭代次数、树的最大深度、学习率、随机有放回抽样对模型的拟合性和鲁棒性影响较大,且可取值范围较广,难以根据经验选取最优值,如采用网格搜索,则计算量过大,影响计算效率[17]。为此,本研究采用 SSA 算法对 XGBoost 模型的这 4 个参数进行寻优,模型耦合构建具体流程可分为 8 步,具体如下。

步骤 1:确定输入输出因子,划分数据集,初始化 XGBoost 模型 4 个超参数。

步骤 2:初始化 SSA 麻雀参数,设定种群数量、最大迭代次数以及预警麻雀比重等,定义适应度函数,将麻雀种群划分为捕食者和加入者。

步骤 3:运行 XGBoost 模型,计算交叉验证均方误差均值并输出,SSA 计算适应度并排序。

步骤 4:更新捕食者位置。

步骤 5:更新加入者位置。

步骤 6:更新预警麻雀位置,获得新位置后,如果新位置分布比之前更好,则进行更新。

步骤 7:迭代过程,重复步骤 3~步骤 6,直到 SSA 最大迭代次数。

步骤 8:输出耦合模型最优参数和预测结果[15]。

利用 Matlab R2022a 软件进行模型构建,SSA 算法初始化麻雀种群数量为 50,发现者所占比例为 20%,预警值为 0.8,最大迭代次数为 100,以建模过程中 XGBoost 交叉运行 5 次的均方误差均值作为 SSA 的适应度函数。根据经验并查阅相关文献[8,17-19],XGBoost 模型的树的最大深度、学习率、最大迭代次数、随机有放回抽样 4 个超参数寻优数值分别为 1~20、0.01~0.20、50~300、0.1~1.0。通过不断的寻优、训练和测试,确定了 4 个超参数的最优值,各参数的含义及最佳取值如表 3 所示。

2 结果与分析

2.1 模型结果与评估

采用均方根误差(RMSE)、平均绝对误差(MAE)以及拟合优度(R^2)回归模型的评估指标对模型进行评估,式(7)~式(9)分别为 3 个评估指标的计算公式。RMSE 与 MAE 则反映了模型预测值与真实值之间的差异,其数值越小说明预测的结果越好;R^2 用于衡量模型对数据信息量的捕捉,反映了信息拟合度的好坏,其数值越接近 1 模型表现越好。为更好地评估测试 SSA-XGBoost

表 3　XGBoost 最佳参数与默认值
Tab. 3　Xgboost Best Parameters and Default Values

参数名称	最优值	参数含义
max_depth	5	树的最大深度,值越大越容易过拟合
learning_rate	0.06	学习率,每次迭代更新权重时的步长
n_estimators	210	基学习器个数(最大迭代系数),基学习器通常为树模型
subsample	0.75	训练使用的数据随机采样的比例

模型的性能,本研究还分别采用 SVM 和 BP 神经网络算法构建了预测模型,SVM 模型采用了表现优秀的径向基核函数,模型的惩罚因子(C)、损失系数(ε)、核函数参数(g)分别为 0.6、0.08、0.15;BP 神经网络模型采用单隐含层结构,训练函数为随机梯度下降算法,激活函数为 sigmoid,性能函数是 MSE,最大迭代次数为 200,学习率为 0.05。利用测试集 37 组数据对 3 种模型的性能进行测试,经计算,SSA-XGBoost 模型的 RMSE 和 MAE 分别为 4.25 μg/L 和 3.19 μg/L,R^2 为 0.77;BP 神经网络模型 RMSE 和 MAE 分别为 5.41 μg/L 和 4.67 μg/L,R^2 为 0.69;SVM 模型 RMSE 和 MAE 分别为 5.21 μg/L 和 4.43 μg/L,R^2 为 0.73。总体上,3 个模型的预测效果均较好,但综合比较,SSA-XGBoost 模型的表现最为优秀,精度也最高。3 个模型预测值与真实值的测试比对结果如图 1 所示。3 个模型的预测值与真实值的变化趋势基本一致,其中,SSA-XGBoost 模型的预测值与真实值吻合度最好,SSA-XGBoost 模型实用价值较高。

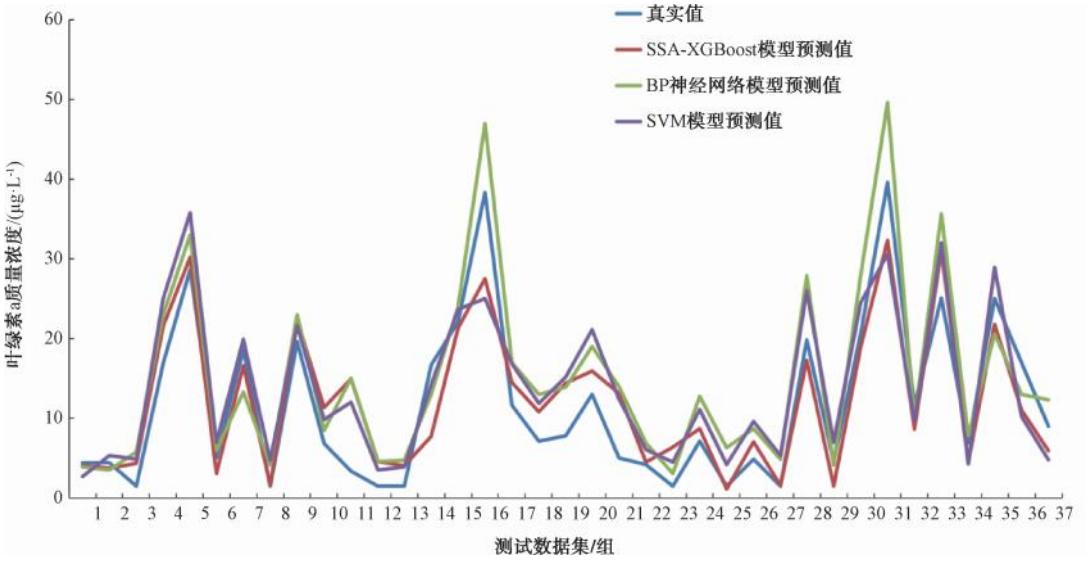

图 1　水体富营养化预测模型测试效果比对
Fig. 1　Comparison of Testing Results of Prediction Models for Water Body Eutrophication

$$R_{\mathrm{RMSE}}=\sqrt{\frac{\sum_{i=1}^{m}(y_i-y_i')^2}{m}} \tag{7}$$

$$M_{\mathrm{MAE}}=\frac{\sum_{i=1}^{m}|y_i-y_i'|}{m} \tag{8}$$

$$R^2=1-\frac{\sum_{i=1}^{m}(y_i-y_i')^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2} \tag{9}$$

其中:R_{RMSE}——RMSE 值;
M_{MAE}——MAE 值;
m——测试集样本的数量;
y_i——第 i 个样本的叶绿素 a 真实值,μg/L;
y_i'——第 i 个样本的叶绿素 a 预测值,μg/L;
$\bar{y}$——测试集所有样本叶绿素 a 的实测均值,μg/L。

2.2 模型敏感性分析

水体富营养化影响因素众多,且不同水体的主导影响因素各不相同,为进一步明确 SSA-XGBoost 模型 12 个输入变量对预测结果的影响程度,对输入变量进行敏感性分析。依次对测试集中每个输入变量的数值进行步长为 10% 的调整,数值总体为 -30%~30%,每次验证集只调整 1 个输入变量的数值,其他变量的值保持不变,代入模型后进行回归预测并计算新的预测值与原始预测值的 RMSE,计算结果如表 4 所示。同样步长的调整造成的 RMSE 越大,该变量对预测结果的影响就越大。由表 4 可知,对叶绿素 a 预测结果影响最大的变量为 pH,其次为硝酸盐,再者依次为高锰酸盐指数、溶解氧、氨氮、总氮和 7 d *I* 均值,温度、总磷和 7 d SSD 均值影响较小。硝酸盐、氨氮、总氮、总磷为营养盐因子,7 d *I* 均值、温度和 7 d SSD 均值与太阳光照有关,总体上,营养盐因子的重要程度要大于太阳光照因子,说明太阳光照不是 2 座引黄水库影响藻类生长的主要限制因子,营养盐才是主要驱动因子。另一方面,NO_3^- 的敏感性要大于氨氮、总氮和总磷,这也与王文杰等[10]的研究发现,某些藻类生长主要是利用硝酸盐氮源的结论基本一致。

表 4 输入变量敏感性分析
Tab. 4 Sensitivity Analysis of Input Variables

变量	RMSE					
	-30%	-20%	-10%	10%	20%	30%
温度	1.10	1.02	0.98	0.42	0.47	0.54
pH	4.11	3.30	2.60	1.90	3.90	4.89
溶解氧	1.36	1.01	0.74	0.97	1.88	2.13
高锰酸盐指数	1.62	1.45	1.36	1.52	1.78	2.32
氨氮	2.51	2.17	1.22	0.30	1.12	1.30
总磷	0.83	0.72	0.61	0.17	0.20	0.28
总氮	1.54	1.15	1.12	1.06	1.33	1.16
硝酸盐	2.21	2.03	1.83	1.88	1.90	2.82
浑浊度	1.13	0.77	0.45	0.32	0.45	0.76
氮磷比	1.09	0.82	0.39	0.71	0.74	0.85
7 d SSD 均值	0.82	0.66	0.30	0.12	0.11	0.15
7 d *I* 均值	1.43	1.23	0.86	0.70	0.89	1.00

2.3 模型应用性验证

为进一步验证 SSA-XGBoost 模型的预测精度和应用性,对应国际公认的叶绿素含量分级,采用高阳俊等[20]对叶绿素 a 的分级标准,具体如表 5 所示。利用模型对随机抽取的 20 个样本的水体富营养化程度进行等级预测,等级预测结果如表 6 所示,水体富营养化等级的预测准确率为 85%以上,有 3 组数据的预测等级产生偏差,但偏差也仅限于相邻等级之间,有 1 组实测浓度等级为Ⅰ级,预测等级为Ⅱ级;有 2 组实测等级为Ⅱ级,预测等级为Ⅲ级,整体上,模型预测等级精度较高,实用性较强。

表 5 基于叶绿素 a 的富营养化分级标准[20]
Tab. 5 Eutrophication Classification Standard Based on Chlorophyll a[20]

营养分级	标准分级	叶绿素 a/(μg·L^{-1})
贫营养	Ⅰ	<1.6
中营养	Ⅱ	1.6~10
轻富营养	Ⅲ	10.0~26.0
中富营养	Ⅳ	26.0~64.0
重富营养	Ⅴ	64.0~160.0
极端富营养	劣Ⅴ	>160

表 6 水体富营养化模型预测等级结果
Tab. 6 Prediction Level Results of Water Body Eutrophication Model

实测值/(μg·L^{-1})	预测值/(μg·L^{-1})	实际等级	预测等级
4.4	4.05	Ⅱ	Ⅱ
4.45	3.72	Ⅱ	Ⅱ
1.5	4.38	Ⅰ	Ⅱ
16.9	21.56	Ⅲ	Ⅲ
28.6	30.19	Ⅳ	Ⅳ
5.01	3.08	Ⅱ	Ⅱ
18.6	16.57	Ⅲ	Ⅲ
1.5	1.57	Ⅰ	Ⅰ
19.6	22.32	Ⅲ	Ⅲ
6.8	11.35	Ⅱ	Ⅲ
22.1	21.30	Ⅲ	Ⅲ
38.3	27.48	Ⅳ	Ⅳ
11.6	14.47	Ⅲ	Ⅲ
7.1	10.81	Ⅱ	Ⅲ
13	15.90	Ⅲ	Ⅲ
4.2	4.47	Ⅱ	Ⅱ
7.1	8.72	Ⅱ	Ⅱ
1.5	1.08	Ⅰ	Ⅰ
4.9	7.02	Ⅱ	Ⅱ
19.8	17.27	Ⅲ	Ⅲ

注:当叶绿素 a 实测值小于实验室 3 μg/L 的检出限时,按检出限的 1/2 进行统计分析。

3　结论

研究收集整理了黄河下游 2 座引黄水库 2013 年—2022 年的水质历史数据及气象数据,以 12 个影响水体叶绿素 a 含量的指标作为模型的输入因子,采用具有泛化能力强、计算性能高等特点的 XGBoost 算法,并耦合 SSA 算法对其超参数进行寻优,构建了适于济南市 2 座引黄水库的富营养化预测预警 SSA-XGBoost 模型。模型评估结果表明,SSA-XGBoost 模型性能优于 BP 神经网络和 SVM 模型,模型 R^2 为 0.77,RMSE 为 4.25 μg/L,MAE 为 3.19 μg/L,性能较好,精度较高。模型敏感性分析表明,对 2 座引黄水库叶绿素 a 预测结果影响最大的变量为 pH,其次为硝酸盐,再者依次为高锰酸盐指数、溶解氧、氨氮、总氮和 7 d I 均值,温度、总磷和 7 d SSD 均值影响相对较小。模型应用性验证结果表明,叶绿素 a 等级预测准确率可达 85%以上,模型实用性较强。

SSA 算法通过智能搜索策略,在效率、全局优化能力和适应性上优于网格搜索,尤其适合 XGBoost 这类参数多、范围广的模型调优。本研究耦合 SSA 与 XGBoost 2 种智能机器学习算法构建模型,不仅弥补了 XGBoost 算法参数优化时人工试算与网格搜索低效性的弊端,还保留了其算法计算效率高、泛化能力强的优点。但本研究也有一定的局限性,一方面,本研究侧重于数据本身进行价值提取,在藻类生长机制等方面未做深入研究,影响藻类生长的因素较多,除本研究筛选出的主要环境因子与气象因子外,其生长还受水体换水周期、流速、水位等水文水力条件以及水中微量元素和生物因素等的影响,藻类生长及水华形成不同阶段的主要影响因子也会不同,富营养化暴发机制复杂,因此,模型输入因子存在一定的不确定性[21]。另一方面,我国幅员辽阔,不同区域湖泊水库的气候条件、水文、水体水质特征可能存在较大差异,模型的泛化能力会受到影响,如本研究针对浅型水库构建预测模型,可能无法预测深水湖泊的富营养化。因此,对于我国南方热带、深水、贫营养或极端富营养化且与本研究 2 座引黄水库氮磷营养结构差异较大的湖库,模型适用性可能不强或失效,需重新进行训练。在模型算法特点方面,数据训练集和测试集的划分具有一定的随机性,建模时虽采用了交叉验证法进行训练和评估,但模型依然会受到数据随机性和噪声的影响。

预测模型具有一定的时空适宜性,应用该方法时,应针对某个特定的水体,挖掘数据价值以构建适宜模型,对于我国北方很多地区,冬季藻类叶绿素 a 含量通常较低,建模时可考虑剔除。后续,为进一步提高模型预测精度和实用性,一方面,可继续优化预测模型的输入因子指标体系,如研究增加水体磷酸盐、透明度、几日前叶绿素 a 浓度、降雨量和风速等指标,并根据藻类生长周期及其时空分布特点,分季节构建更具有时空适宜性和针对性的预测模型,对于藻类多发季节,为尽量捕捉水质叶绿素 a 浓度变化特征,可适当增加各指标的检测频率。另一方面,XGBoost 等机器学习算法本质是从大量的数据中学习规律,随着实验室水质等数据量的不断积累,后续还需增加建模所用数据量,并对模型进行不断地优化和验证。本研究采用耦合算法构建的 SSA-XGBoost 水体富营养化回归预测模型,可为黄河下游引黄水库水源水质风险预警提供技术支撑,并为水厂工艺的调整提供有效的信息支持。

参考文献

[1] 侯伟, 孙韶华, 贾瑞宝. 中国北方山区水库与引黄水库富营养化特征[J]. 中国环境监测, 2016, 32(2): 58-63.
HOU W, SUN S H, JIA R B. Eutrophication and water characteristics of mountain and Yellow River reservoirs in Northern China[J]. Environmental Monitoring in China, 2016, 32(2): 58-63.

[2] 王震, 邹华, 杨桂军, 等. 太湖叶绿素 a 的时空分布特征及其与环境因子的相关关系[J]. 湖泊科学, 2014, 26(4): 567-572.
WANG Z, ZOU H, YANG G J, et al. Spatial-temporal characteristics of chlorophyll-a and its relationship with environmental factors in Lake Taihu [J]. Journal of Lake Sciences, 2014, 26(4): 567-572.

[3] WANG L, WANG X Y, JIN X B, et al. Analysis of algae growth mechanism and water bloom prediction under the effect of multi-affecting factor [J]. Saudi Journal of Biological Sciences, 2017, 24(3): 556-562.

[4] 刘旭华, 宋振宏, 刘华民, 等. 基于 HEA 算法的红山水库富营养化模拟和预测研究[J]. 内蒙古大学学报(自然科学版), 2023, 54(1): 53-60.
LIU X H, SONG Z H, LIU H M, et al. Simulation and prediction of eutrophication in Hongshan Reservoir by hybrid evolutionary algorithm [J]. Journal of Inner Mongolia University (Natural Science Edition), 2023, 54(1): 53-60.

[5] JIA W J, CHENG J, HU H Z. A cluster-stacking-based approach to forecasting seasonal chlorophyll-a concentration in coastal waters [J]. IEEE Access, 2020(8):99934-99947.

[6] 张虎军, 宋挺, 朱冰川, 等. 太湖蓝藻水华暴发程度年度预测[J]. 中国环境监测, 2022, 38(1): 157-162.
ZHANG H J, SONG T, ZHU B C, et al. Annual forecast of the extent of cyanobacteria bloom in Taihu Lake [J]. Environmental Monitoring in China, 2022, 38(1): 157-162.

[7] 张成成, 陈求稳, 徐强, 等. 基于支持向量机的太湖梅梁湾叶绿素 a 浓度预测模型[J]. 环境科学学报, 2013, 33(10): 2856-2861.
ZHANG C C, CHEN Q W, XU Q, et al. A chlorophyll-a prediction model for Meiliang bay of Taihu based on support vector machine [J]. Acta Scientiae Circumstantiae, 2013, 33(10): 2856-2861.

[8] 柴敬, 王润沛, 杜文刚, 等. 基于 XGBoost 的光纤监测矿压时序预测研究[J]. 采矿与岩层控制工程学报, 2020, 2(4): 60-67.
CHAI J, WANG R P, DU W G, et al. Study on time series prediction of rock pressure by XGBoost in optical fiber monitoring [J]. Journal of Mining and Strata Control Engineering, 2020, 2(4): 60-67.

[9] 侯伟, 陈燕, 孙韶华, 等. 黄河下游典型水库浮游植物群落结构及其与环境因子的关系[J]. 水资源与水工程学报, 2018, 29(2):32-68.
HOU W, CHEN Y, SUN S H, et al. Phytoplankton community dynamics and its relationship with key environmental factors in typical reservoirs, lower reaches of Yellow River [J]. Journal of Water Resources and Water Engineering, 2018, 29(2): 32-68.

[10] 王文杰, 姚旦, 赵辰红, 等. 氮磷营养盐对四种淡水丝状蓝藻生长的影响[J]. 生态科学, 2008, 27(4):202-207.
WANG W J, YAO D, ZHAO C H, et al. The effects of nitrogen and phosphorus nutrients on the growth of four freshwater filamentous blue-green algae[J]. Ecological Science, 2008, 27(4):202-207.

[11] 杨大兴. 碧流河水库富营养化特征分析及影响因素研究[J]. 水资源开发与管理, 2024(5): 27-31.
YANG D X. Analysis of eutrophication characteristics and influencing factors of Biliuhe Reservoir [J]. Water Resources Development and Management, 2024(5): 27-31.

[12] 顾启华. 富营养化水体中藻类水华成因分析与研究[D]. 天津: 天津大学, 2006.
GU Q H. Analysis and study on cause of algal-bloom in eutrophic aqueous environment[D]. Tianjin: Tianjin University, 2006.

[13] 孙越峰, 秦艳杰, 李洪波, 等. 辽河口海域叶绿素 a 的时空分布特征及其影响因素[J]. 环境保护科学, 2020, 46(2): 44-48.
SUN Y F, QIN Y J, LI H B, et al. Spatial and temporal distribution characteristics and influencing factors of chlorophyll a in the Liaohe Estuary [J]. Environmental Protection Science, 2020, 46(2): 44-48.

[14] 杨婷. 气象因子与太湖蓝藻水华的响应关系研究[D]. 南京: 南京信息工程大学, 2012.
YANG T. Study on coupling relationship between meteorological factors and dynamic algal bloom in Taihu Lake[D]. Nanjing: Nanjing University of Information Science and Technology, 2012.

[15] 黄宇波, 范向军, 杨霞, 等. 水文气象因素对香溪河藻类动态的影响[J]. 中国农村水利水电, 2023(1): 1-7.
HUANG Y B, FAN X J, YANG X, et al. The effects of hydro-meteorological factors on the algae dynamics in Xiangxi River [J]. China Rural Water and Hydropower, 2023(1): 1-7.

[16] XUE J K, SHEN B. A novel swarm intelligence optimization approach: Sparrow search algorithm [J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.

[17] 安惠伦. 基于 PCA-SSA-XGBoost 模型的拱坝应力预测研究[D]. 天津: 天津大学, 2021.
AN H L. Research on arch dam based on PCA-SSA-XGBoost model [D]. Tianjin: Tianjin University, 2021.

[18] 李蕙萱, 吴瑞溢. 利用 XGBoost 和 SVR 算法的地铁站客流量模型研究[J]. 三明学院学报, 2019, 36(6): 56-64.
LI H X, WU R Y. The research on the models of metro traffic passenger flows based on the XGBoost and SVR algorithms [J]. Journal of Sanming University, 2019, 36(6): 56-64.

[19] 朱明, 王春梅, 高翔, 等. XGBoost 在卫星网络协调态势预测中的应用[J]. 小型微型计算机系统, 2019, 40(2): 2561-2565.
ZHU M, WANG C M, GAO X, et al. Application of XGBoost in the prediction of satellite network coordination situation [J]. Journal of Chinese Computer Systems, 2019, 40(2): 2561-2565.

[20] 高阳俊, 曹勇, 赵振, 等. 基于叶绿素 a 分级的东部湖区富营养化标准研究[J]. 环境科学与技术, 2011, 34(12): 218-220.
GAO Y J, CAO Y, ZHAO Z, et al. Study on eutrophication criteria based on chlorophyll a grading in east Lake Area[J]. Environmental Science & Technology, 2011, 34 (12): 218-220.

[21] 葛裕豪. 查干湖富营养化演变规律及其影响因素分析[D]. 哈尔滨: 黑龙江大学, 2024.
GE Y H. Analysis of eutrophication evolution and influencing factors in Chagan Lake[D]. Harbin: Heilongjiang University, 2024.