

DOI:10.3876/j.issn.1000-1980.2020.06.009

三种机器学习模型在太湖藻华面积预测中的应用

吴娟¹,朱跃龙²,金松¹,杨涛²,冯钧²,吴志勇²,薛涛¹,姜悦美¹

(1.太湖流域管理局水文局(信息中心),上海 200434; 2.河海大学水文水资源学院,江苏 南京 210098)

摘要:基于2014—2018年太湖气象水文水质数据与卫星遥感数据,分别采用支持向量机(SVM)、长短记忆神经网络(LSTM)、极端梯度提升树(XGBoost)模型模拟全太湖、贡湖、南部沿岸区、中西北湖区的蓝藻水华(简称藻华)面积。结果表明:(a)XGBoost全太湖与分区藻华面积回归模型模拟效果较好,其次是SVM、LSTM回归模型;不同时间尺度下SVM、XGBoost回归模型对全太湖藻华面积模拟结果偏小,但有效模拟了藻华的发展趋势。(b)XGBoost分类模型在全太湖、中西北湖区模拟准确率较高,优于SVM、LSTM分类模型;在贡湖、南部沿岸区,3种分类模型准确率均较高。(c)以当天、提前1d的气象水文水质因子作为全太湖与分区藻华面积模型输入,XGBoost回归与分类模型模拟精度较高、稳健性较好,预测应用情景较好。

关键词:机器学习;蓝藻水华模拟;支持向量机;长短记忆神经网络;极端梯度提升树;太湖

中图分类号:X171 **文献标志码:**A **文章编号:**1000-1980(2020)06-0542-10

Area prediction of cyanobacterial blooms based on three machine learning methods in Taihu Lake

WU Juan¹, ZHU Yuelong², JIN Song¹, YANG Tao², FENG Jun²,
WU Zhiyong², XUE Tao¹, JIANG Yuemei¹

(1. Bureau of Hydrology Information Center of Taihu Basin Authority, Shanghai 200434, China;
2. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China)

Abstract:Based on atmospheric-hydrological data and satellite remote sensing data from 2014 to 2018, the support vector machine (SVM), long short-term memory model (LSTM), extreme gradient boosting (XGBoost) model were applied to predict the cyanobacterial bloom area in fields including the whole region, Gonghu Bay, southern coastal region and central north-western region of Taihu Lake. The results demonstrated that the XGBoost regression model had better accuracy than SVM and LSTM regression model in the whole region and subdivided regions. Compared with the observed cyanobacterial bloom area, simulated areas of SVM regression model and XGBoost regression model were lower in the Taihu Lake under different time scales, while the development tendency of cyanobacterial blooms was effectively simulated. In addition, the XGBoost classification model had better accuracy than SVM and LSTM classification model for the whole region and the central north-western region of Taihu Lake. Three classification models had high accuracy in the Gonghu Bay and the southern coastal region of Taihu Lake. Finally, taking the atmospheric-hydrological data and water quality data of the same day and one day advanced as model inputs, the XGBoost regression model has high accuracy and robustness in cyanobacterial bloom area simulation, which had a promising application prospect for the cyanobacterial bloom prediction.

Key words: machine learning; cyanobacterial bloom simulation; support vector machine; long short-term memory model; extreme gradient boosting model; Taihu Lake

太湖(119° 52′ 32″E ~ 120° 36′ 10″ E、30° 55′ 40″N ~ 31°32′58″N)是我国第三大淡水湖,水面面积约

基金项目:国家重点研发计划(2018YFC0407900)
作者简介:吴娟(1987—),女,工程师,硕士,主要从事水文预报、水资源与水环境研究。E-mail:wujuan3412@163.com
引用本文:吴娟,朱跃龙,金松,等.三种机器学习模型在太湖藻华面积预测中的应用[J].河海大学学报(自然科学版),2020,48(6):542-551.
WU Juan, ZHU Yuelong, JIN Song, et al. Area prediction of cyanobacterial blooms based on three machine learning methods in Taihu Lake [J]. Journal of Hohai University (Natural Sciences), 2020, 48(6):542-551.

2338 km²,是典型的大型浅水湖泊。近几十年来,太湖蓝藻水华(以下简称“藻华”)呈现非线性、复杂性等特点^[1]。藻华不完全是本地蓝藻短时快速生长暴发所致^[2],而是已经成为优势种群的蓝藻在合适条件下,受物理(温度^[3]、光照^[4]、水动力^[5]等)、化学(氮、磷^[6]等)^[7]、生物(与其他藻类的竞争)^[8]因素等共同影响,在较短的时间里改变水平与垂直位置形成“暴发”^[9],因此,藻华本质是藻类生物量在水体中逐渐增加、上浮、聚集、迁移的过程^[10]。藻华扩张分为时间扩张、空间扩张和生物量扩张。时间扩张用藻华持续时间、发生时间等表征,空间扩张用基于遥感影像的藻华面积表征,生物量扩张用叶绿素 a、浮游植物总生物量等表征^[11]。由于生物量统计主要以特定湖区采样监测、计数鉴定方法为主,受制于采样站点的位置与频次,获取资料困难,因此,本文以研究藻华时间、空间扩张为主。

研究表明,气候变化、水体富营养化是太湖藻华增多的重要原因^[12]。大面积偶发性太湖藻华受气象因子影响较大,而大面积频发性藻华主要受营养盐空间分布影响^[13]。太湖处于较高的营养盐水平,超过藻华发生的阈值^[14],但季节性营养盐抑制过程^[15]对藻华强度有一定的限制作用^[16],足以维持水华持续发生^[17]。杭鑫等^[18]认为气温、风、降水等气象因子对太湖藻华影响较大,适宜的气象因子为藻华暴发提供了有利条件。秦伯强等^[19]研究表明:藻华面积形成过程较复杂,蓝藻细胞生长阶段,营养盐浓度、温度、光照等因素为藻华提供了物质基础;在蓝藻水华暴发阶段,藻华面积空间分布则受蓝藻细胞团浮力作用与水动力湍流作用的共同影响,风速决定藻华垂向分布,风向和风速共同决定其水平分布。

蓝藻生态动力学模型因过程复杂、参数众多^[20],导致应用在预测中存在一定的困难,而机器学习算法(如神经网络等)为太湖藻华预测提供了新的理论和方法^[21]。机器学习是一门研究计算机模拟或实现人类活动的学科^[22],也是知识发现、数据挖掘的重要基础^[23]。通过挖掘数据背后的深度价值和内在联系^[24],机器学习具有高度的非线性与灵活性^[25],应用于短期降水预报^[26]、径流预测^[27]、藻华预警^[28],有效提升了预报精度和效率。张艳会等^[29]采用 BP 神经网络和模糊理论,建立了藻华发生的模糊风险评价方法;于家斌等^[30]提出基于长短期记忆(LSTM)循环神经网络(RNN)藻华预测模型(LSTM 模型预测精度较高,对样本具有较好的适应性);罗晓春等^[31]采用随机森林机器学习算法分析同期气象因子与藻华综合指数的关系,定量评估影响藻华的主要气象因子的贡献率。上述研究为太湖藻华的预测预警与风险管理提供了参考。

太湖湖面宽广,一般采用野外观测和数值模拟研究湖流导致藻华迁移和堆积。风向分布具有不均匀性,并随着风速减小差异有所扩大。在低风速条件下,无法形成稳定环流,蓝藻容易在太湖表面积聚,而此时各湖区之间风向差异较大,需要借助蓝藻生态动力学模型深入研究。本研究侧重采用全太湖、分湖区建模的思路,基于支持向量机(SVM)、长短记忆神经网络(LSTM)、极端梯度提升树(XGBoost)模型,快速预判各湖区藻华面积,旨在为太湖藻华预测预警、风险管理提供技术支撑。

1 数据与方法

1.1 数据来源

按照岸线、水质和水下地形等因素,太湖可分为竺山湖、梅梁湖、贡湖等^[32],见图 1。近年来,中分辨率成像光谱仪(MODIS)数据因其光谱分辨率高、观测周期短等特点,广泛应用于太湖藻华动态监测^[33]。考虑到太湖藻华每年 4 月开始零星暴发,12 月以后基本无大范围藻华暴发,因此,选取 4—11 月的 EOS/MODIS 影像数据(排除了云层、天气质量等对影像质量影响较大的数据)共 761 景,时间间隔从 1 d 到 15 d 不等,藻华面积来源于中国科学院南京地理与湖泊研究所解译成果。本研究采用的藻华面积为绝对面积,指当像元内存在水华时,只计算像元内水华覆盖部分的面积。太湖内共布设 33 个监测点,覆盖了全部水域,监测项目包括 pH、总磷、总氮、高锰酸盐指数、氨氮、叶绿素 a、蓝藻密度,仅每月上旬监测 1 次,根本无法满足藻华面积建模要求,因此本研究只能采用环湖 10 个水质站、4 个气象站时间分布较精细的数据进行建模。水质指标

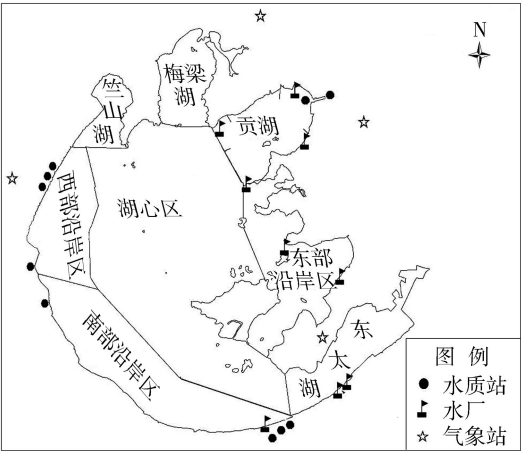


图 1 太湖分区与观测站示意图

Fig.1 Sub-regions in Taihu Lake and observation stations

包括水温、pH、电导率、溶解氧、氨氮、总磷、总氮、浊度、高锰酸盐指数,气象指标包括气温、风速、风向、降水。本研究的资料序列均为2014—2018年。由于东部沿岸区与东太湖没有MODIS藻华资料,而贡湖、南部沿岸区均有水源地,因此,本研究采用全太湖、贡湖、南部沿岸区、中西北湖区(西部沿岸区+湖心区+竺山湖+梅梁湖)分别建模。南部沿岸区代表站包括幻溇、濮溇、汤溇,中西北湖区代表站包括长丰涧、大港、洪巷、官渎、社渎。

1.2 研究方法

太湖藻华受物理、化学、生物因素等共同影响,机理十分复杂,呈非线性变化特征。本研究需要先采用Pearson相关系数去除不相关或重复变量,提取对各湖区藻华面积影响较大的主要气象水文水质因子。从时间、空间尺度上考虑到不同湖区水质差异较大,MODIS成像时间为上午10:30左右,因此,从时间尺度上分成3个模型,模型1采用当天0:00—8:00之前的气象、水文、水质数据预测当天的藻华面积,模型2采用提前一天0:00—23:00的数据预测当天的藻华面积,模型3采用提前两天0:00—23:00的数据预测当天的藻华面积。由于SVM、LSTM、XGBoost模型既可解决分类问题,也可解决回归问题,并且处理非线性问题有一定的优越性,在水质评价与预测研究中有一定的应用情景,因此,本研究分别开展两类算法的应用研究,共建立24个模型。训练集、验证集为2014—2018年序列数据,其中随机抽取70%的样本(533组)用于训练,剩下30%的样本(228组)作为验证集。

1.2.1 数据预处理

由于输入变量可能不在同一个数量级上,数据之间的差异性可能对模型学习能力存在一定的不利影响,为了减少数值差异、保证模型参数稳定收敛,需要进行归一化处理,确保各变量处于相同的量级。

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

式中: X ——输入变量; X' ——变量 X 归一化后的值; $X_{\min}$ 、 $X_{\max}$ —— X 的最小值、最大值。

1.2.2 SVM

SVM是通过核函数,将低维输入空间中线性不可分的点映射成高维特征空间中线性可分的点,通过划分超平面,使所有的点到分类超平面的距离最大化^[34]。其中,高维空间中距离分类超平面最近的那些点所对应的低维空间点被称为支持向量^[35]。核函数的作用是接受两个低维空间中的输入向量,计算出经过某种变换后二者在高维空间中的向量内积值,具体见文献^[36]。

常用的核函数有线性、多项式、径向基、sigmoid核函数。经对比,本研究采用误差较小、分类准确率较高的径向基函数作为核函数^[37]。SVM采用径向基核函数,相应参数为惩罚因子 c 和核参数 g 。为确定最佳的参数值,将 c 、 g 分别取以2为底的指数离散值,代入 V 折交叉验证法中,选取平均误差最小的 c 、 g 值;采用网格搜索法,减小取值范围和间隔,确定最佳参数,根据计算,最佳参数为 $c=0.76$ 、 $g=0.38$ 、 $V=5$ 。

1.2.3 LSTM

RNN属于深度学习方法^[38],用于处理非线性时间序列,但在实际建模中,存在梯度消失及梯度爆炸问题^[39]。为此,Hochreiter等提出了LSTM,将RNN中隐含层中的神经元替换为记忆体,实现信息保留与长期记忆。通过记忆细胞进行状态信息存储,门结构负责细胞状态的更新与保持^[40]。每个记忆体包含一到多个记忆细胞和3种“门”,包括“遗忘门”、“输入门”和“输出门”。LSTM训练过程是寻找最优参数,使模型收敛、损失函数达到最小^[41]。算法见文献^[42]。LSTM网络层中隐藏神经元数目是影响准确率的重要参数,为寻求模型的全局最优解,采用基于梯度下降、具有相当鲁棒性的ADAM自适应学习率优化算法;经计算,初始学习率取0.001,隐藏神经元数为256,可使全局最优。

1.2.4 XGBoost

XGBoost模型为树结构,是对梯度提升决策树算法的改进,采用加法学习模型进行优化^[43]。通过不断迭代生成新的树,加入子树使模型不断逼近样本分布,将多个分类准确率较低的树模型组合成一个准确率较高的模型^[44]。通过在Xgboost损失函数中加入正则项,既控制模型的复杂度,也降低模型的方差,避免模型过拟合^[45]。算法见文献^[46]。XGBoost模型主要参数包括eta、subsample、max_depth。eta指学习速率,可以提高模型的鲁棒性;subsample指训练的实例样本占总体样本的比例;max_depth为树的最大深度,用来避免过拟合。经计算,最佳参数为eta=0.15,sub-sample=0.2,max_depth=8。

1.3 模型评价

常用评价指标包括:纳什效率系数(N_{SE})、相对偏差(P_B),公式如下:

$$N_{SE} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{2}$$

$$P_B = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)}{\sum_{i=1}^n y_i} \times 100\% \tag{3}$$

式中: i ——时刻; n ——数据长度; $\hat{y}_i$ ——模拟数据; y_i ——实测数据; $\bar{y}$ ——实测数据平均值。

N_{SE} 取值范围为负无穷至 1,越接近 1 表示模拟效果越好、可信度越高; N_{SE} 接近 0,表示模拟结果接近观测值的平均值水平,即总体结果可信,但过程模拟误差大。 P_B 绝对值越小,模拟效果越好。

Pearson 相关系数(r)是描述两个随机变量线性相关的统计量,计算公式如下:

$$r = \frac{\sum_{i=1}^n (W_i - \bar{W})(J_i - \bar{J})}{\sqrt{\sum_{i=1}^n (W_i - \bar{W})^2} \sqrt{\sum_{i=1}^n (J_i - \bar{J})^2}} \tag{4}$$

式中: W_i 、 J_i ——随机变量; $\bar{W}$ 、 $\bar{J}$ —— W_i 、 J_i 的平均值。

r 取值范围为-1.0~1.0,越接近 1.0,正相关越显著,越接近-1.0,负相关性越显著。构造统计量 t 检验相关性是否显著,计算公式如下:

$$t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \tag{5}$$

统计量 t 遵从自由度 $\nu=n-2$ 的 t 分布。给定显著性水平 α ,查 t 分布表,若 $t>t_\alpha$,认为相关系数是显著的。

2 结果与分析

2.1 相关性分析

为了确定所选气象、水文、水质因子对太湖藻华面积模拟的有效性,提高藻华面积模型的准确性,采用 Pearson 相关系数去除相关性不显著的变量。分别计算当天、提前 1 d 到提前 30 d 共计 31 种时间尺度条件下,全太湖、贡湖、南部沿岸区、中西北湖区藻华面积,以及气象、水文、水质数据之间的 Pearson 相关系数,发现当天、提前 1 d、提前 2 d 时间尺度相关系数明显高于其他时段。Pearson 相关系数较显著的因子如表 1 所示,全太湖藻华面积与水温、风速、降水等气象因子相关性较高,其中,与当天无锡 2:00、8:00 风速负相关性较显著,分别达到-0.17、-0.23。不同时间尺度下的贡湖藻华面积与水温、浊度、总氮、氨氮正相关性均较高,相关系数均可达 0.15 以上。南部沿岸区藻华面积与水温呈显著正相关性,相关系数均可达 0.19 以上,与当日东山站 2:00、8:00 风速呈显著负相关性。中西北湖区藻华面积与当日宜兴站 2:00、8:00 风速的负相关性较显著。

2.2 模型分析结果

将训练集和验证集分别代入训练好的全太湖、贡湖、南部沿岸区、中西北湖区藻华面积模拟模型中,并对输出值进行归一化处理。模拟值与观测值之间的 N_{SE} 与 P_B 见表 2。对于不同时间尺度(当天、提前 1 d、提前 2 d)下的全太湖藻华面积回归模型训练集与验证集而言,XGBoost 回归模型 N_{SE} 最高、 P_B 最低,其次是 LSTM 回归模型、SVM 回归模型;验证期的 XGBoost 回归模型,当天的模型 1 与提前 1 d 的模型 2,相对优于提前 2 d 的模型 3。

对于不同时间尺度下贡湖、南部沿岸区、中西北湖区的藻华面积回归模型 N_{SE} ,XGBoost 模型训练集与验证集 N_{SE} 普遍较高,训练集 LSTM 模型的 N_{SE} 高于 SVM 模型,验证集 SVM 模型的 N_{SE} 高于 LSTM 模型;验证期的贡湖 SVM 模型、藻华面积模型 1、南部沿岸区 XGBoost 藻华面积模型 1、中西北湖区 XGBoost 藻华面积模型 2

的 N_{SE} 高于相应湖区的其他时间尺度模型,与全太湖藻华面积回归模型表现也基本一致。

表1 藻华面积与气象、水文、水质因子之间的 Pearson 相关系数

Table 1 Pearson relative coefficient between cyanobacterial bloom area and factors including meteorological, hydrological and water quality factors

因变量	自变量	Pearson 相关系数		
		模型 1	模型 2	模型 3
全太湖藻华面积	贡湖风浪站水温	0.13 *	0.12 *	0.11 *
	贡湖风浪站电导率	0.13 *	0.14 *	0.05
	无锡气象站 2:00 风速	-0.17 **	0.01	0.01
	无锡气象站 8:00 风速	-0.23 **	0.01	-0.13 *
	太湖区降水量	0.06	0.13 *	0.01
贡湖藻华面积	贡湖水温	0.20 **	0.18 **	0.18 **
	贡湖 pH	0.08	0.11 *	0.12 *
	贡湖浊度	0.17 **	0.24 **	0.19 **
	贡湖总氮	0.20 **	0.18 **	0.19 **
	贡湖氨氮	0.15 **	0.17 **	0.15 **
	无锡气象站 8:00 风速	-0.11 *	0.01	-0.06
南部沿岸区藻华面积	南部代表站水温	0.19 **	0.20 **	0.21 **
	东山气象站 2:00 风速	-0.16 **	0.01	0.01
	东山气象站 8:00 风速	-0.21 **	0.06	0.04
中西北湖区藻华面积	西部代表站水温	0.13 *	0.12 *	0.10
	宜兴气象站 2:00 风速	-0.16 **	0.01	-0.07
	宜兴气象站 8:00 风速	-0.17 **	0.01	-0.11 *

注:模型 1、模型 2、模型 3 分别代表用当天、提前 1 d、提前 2 d 的气象预报因子预报当天的藻华面积,以下同。* 表示在 $\alpha=0.05$ 水平上显著相关,** 表示在 $\alpha=0.01$ 水平上显著相关。

表2 3 种机器学习回归模型评价结果

Table 2 Evaluation results of three machine-learning regression methods

湖区	模型	训练集 N_{SE}			验证集 N_{SE}			训练集 P_B			验证集 P_B		
		模型 1	模型 2	模型 3	模型 1	模型 2	模型 3	模型 1	模型 2	模型 3	模型 1	模型 2	模型 3
全太湖	SVM	0.39	0.17	0.34	0.09	0.02	0.18	0.22	0.40	0.27	0.25	0.47	0.40
	LSTM	0.68	0.69	0.79	0.08	0.08	0.24	0.14	0.11	0.14	0.18	-0.11	0.18
	XGBoost	0.83	0.83	0.85	0.37	0.23	0.16	-0.01	-0.01	0.01	0.12	0.04	-0.11
贡湖	SVM	0.29	0.34	0.49	0.25	0.03	0.16	0.33	0.33	0.17	-0.21	-0.22	-0.46
	LSTM	0.74	0.74	0.76	0.01	0.01	0.01	0.02	0.01	0.04	-0.95	-0.66	-0.97
	XGBoost	0.86	0.86	0.86	0.18	0.07	0.16	-0.04	-0.05	-0.03	-0.83	-0.45	-0.32
南部沿岸区	SVM	0.29	0.15	0.51	0.03	0.10	0.04	0.35	0.30	0.17	0.06	0.41	0.29
	LSTM	0.96	0.96	0.98	0.01	0.01	0.01	0.01	0.01	0.01	-0.88	-0.98	-0.88
	XGBoost	0.98	0.98	0.99	0.43	0.06	0.11	-0.01	-0.01	-0.01	-0.45	0.02	-0.17
中西北湖区	SVM	0.44	0.49	0.48	0.23	0.16	0.09	0.27	0.22	0.22	0.11	0.13	0.11
	LSTM	0.69	0.74	0.82	0.01	0.01	0.01	0.16	0.14	0.10	0.10	0.03	0.20
	XGBoost	0.87	0.88	0.88	0.22	0.24	0.14	-0.03	-0.01	-0.01	-0.21	-0.22	-0.19

《太湖蓝藻水华评价方法(试行)》将蓝藻水华面积分为 4 类: $<240\text{ km}^2$ 、 $[240,600)\text{ km}^2$ 、 $[600,1\,000)\text{ km}^2$ 、 $>1\,000\text{ km}^2$,相应面积水域发生水华分别称为零星湖区、局部湖区、区域、大范围水华暴发^[47]。以全太湖藻华面积过程线拟合效果为例,分析 SVM、LSTM、XGBoost 回归模型模拟结果,见图 2~4。模拟结果表明:不同时间尺度下 SVM、LSTM 回归模型训练集的模拟值偏低,XGBoost 回归模型训练集模拟值与实测值基本接近;尽管 SVM、XGBoost 回归模型验证集模拟值较实测值偏小,但有效模拟了零星湖区、局部湖区水华的发展趋势;LSTM 回归模型验证集的模拟值与实测值差异较大,不合理跳动较多,这也验证了神经网络模型存在过学习、泛化能力差的固有缺陷。

从上述分析可知,回归模型在验证期模拟效果不太好,考虑到藻华面积预测与防控业务中主要关注藻华面积与相应湖区面积比例的范围,对精确数值关注较少。为进一步改进藻华面积模型,采用分类算法输出离散的类别值。将藻华面积与相应湖区面积比例按照 $<10\%$ 、 $[10\%,25\%)$ 、 $[25\%,50\%)$ 、 $>50\%$ 分成 4 类,分析不同时间尺度下 SVM、LSTM、XGBoost 分类模型的正确率,见表 3。表 3 结果表明:全太湖各模型的训练集

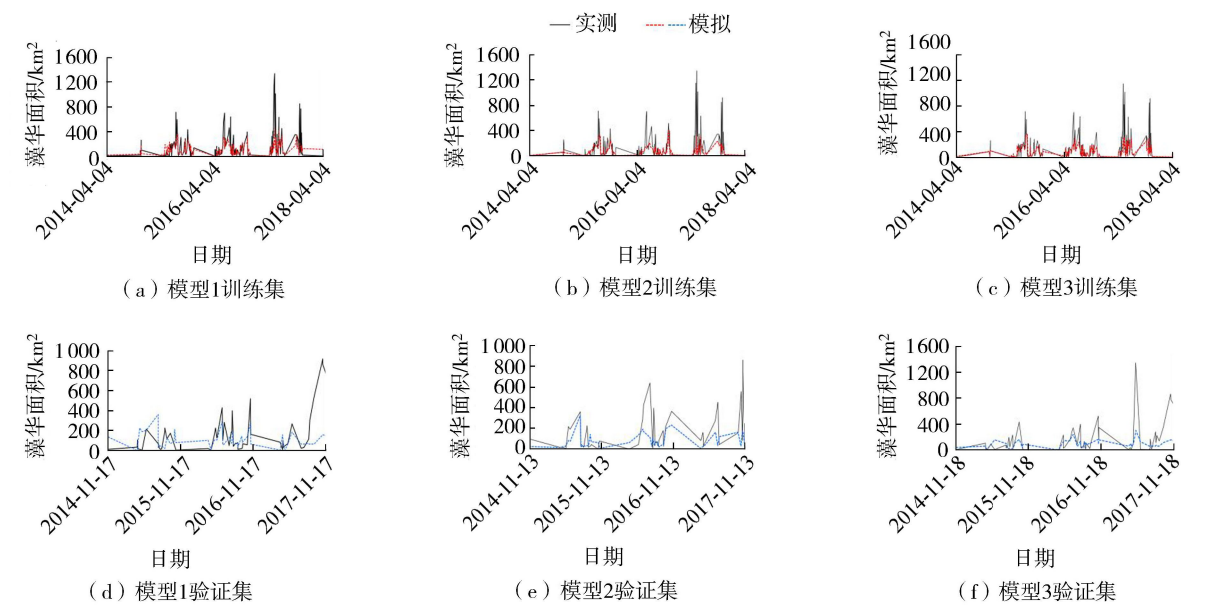


图 2 基 SVM 于回归模型的全太湖藻华面积训练集与验证集模拟

Fig.2 Training set and verification set simulations of cyanobacterial bloom area with SVM regression model in Taihu Lake

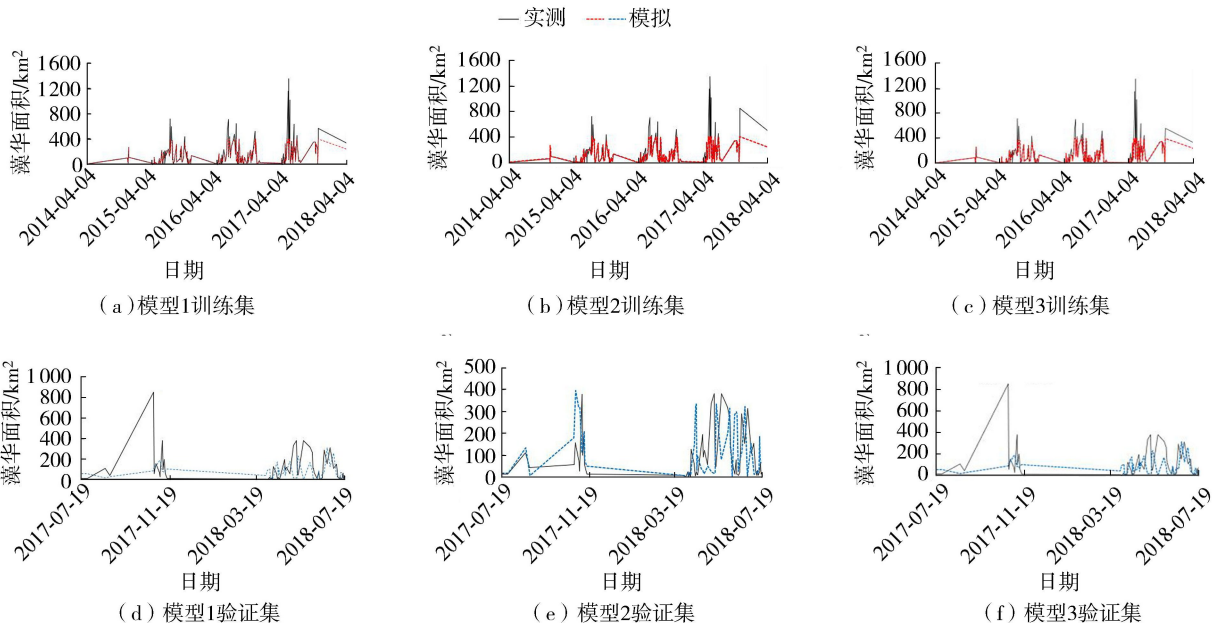


图 3 基于 LSTM 回归模型的全太湖藻华面积训练集与验证集模拟

Fig.3 Training set and verification set simulations of cyanobacterial bloom area with LSTM regression model in Taihu Lake

正确率达到 0.97 以上,验证集 SVM、XGBoost 分类模型正确率达到 0.80 以上,LSTM 分类模型正确率略低,为 0.66 ~ 0.86,误差较大,见图 5。贡湖各模型的训练集正确率达到 0.89 以上,验证集正确率达到 0.82 以上;南部沿岸区各模型的训练集正确率达到 0.97 以上,验证集正确率达到 0.80 以上;中西北湖区训练集正确率达到 0.83 以上,验证集 XGBoost 模型正确率达到 0.80 以上,好于 LSTM 分类模型、SVM 分类模型。总体而言,贡湖藻华面积 SVM、LSTM、XGBoost 分类模型验证集正确率普遍优于其他湖区,可能与贡湖风浪站相对于其他环湖站点位置更加偏向于贡湖内部,气象、水文、水质数据代表性更强有关。

从上述研究可知,SVM、LSTM 模型在藻华面积较低时,模拟与实测结果较接近;当藻华面积超过600 km²时,模拟与实测结果偏差较大。可能原因有 3 个:一是太湖藻华面积高值相对较少,且水华从暴发到消退较快,在运用机器学习方法训练模型时,如果目标值数量不均等、数值差异大,SVM、LSTM 模型一般偏向于数量多、数值低的藻华面积的样本。二是LSTM较适合大数据的机器学习,在数据量足够多的情况下效果较

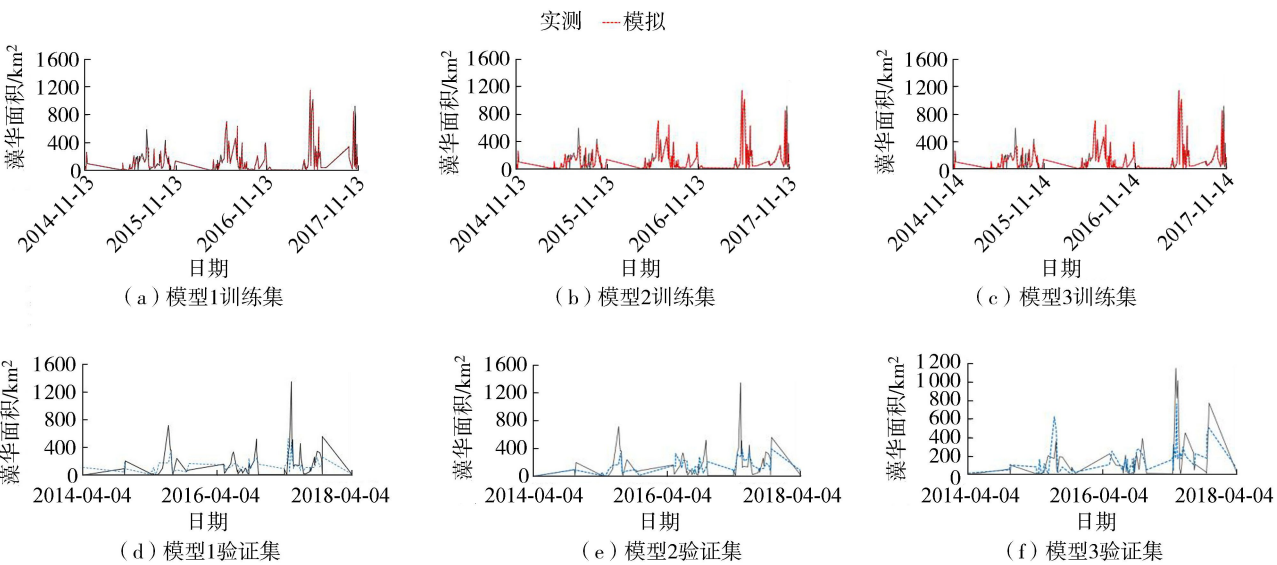


图4 基于XGBoost回归模型的全太湖藻华面积训练集与验证集模拟

Fig.4 Training set and verification set simulations of cyanobacterial bloom area with XGBoost regression model in Taihu Lake

表3 3种机器学习分类模型评价结果

Table 3 Evaluation results of three machine-learning classification methods

湖区	模型	训练集正确率			验证集正确率		
		模型1	模型2	模型3	模型1	模型2	模型3
全太湖	SVM	0.98	0.98	0.97	0.82	0.80	0.80
	LSTM	0.99	0.98	0.98	0.86	0.66	0.75
	XGBoost	0.98	0.98	0.97	0.82	0.80	0.82
贡湖	SVM	0.89	0.90	0.89	0.92	0.90	0.85
	LSTM	0.99	0.99	0.99	0.95	0.82	0.84
	XGBoost	0.99	0.99	0.99	0.89	0.89	0.89
南部沿岸区	SVM	0.99	0.97	0.98	0.80	0.83	0.81
	LSTM	0.99	0.99	0.99	0.84	0.80	0.80
	XGBoost	0.99	0.99	0.99	0.84	0.85	0.83
中西北湖区	SVM	0.83	0.83	0.83	0.63	0.55	0.58
	LSTM	0.98	0.98	0.99	0.74	0.82	0.75
	XGBoost	0.95	0.95	0.95	0.84	0.80	0.80

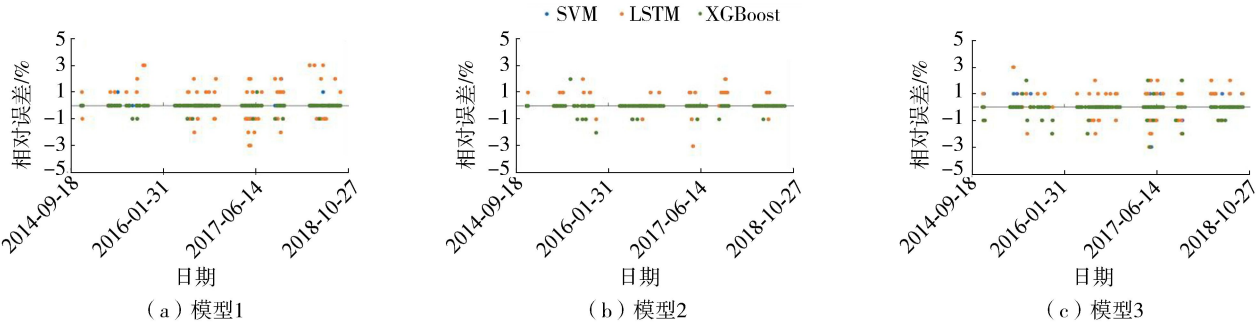


图5 不同时间尺度下SVM、LSTM、XGBoost分类模型的藻华面积验证集模拟

Fig.5 Verification of cyanobacterial bloom areas with SVM, LSTM and XGBoost classification models in Taihu Lake

优,而本研究的数据量可能远远达不到大数据的要求,LSTM 无法很好地保证模型的泛化能力,存在陷入局部最优以及收敛时间长等问题^[48],导致模拟效果不佳。三是太湖风浪作用显著,适度的间歇性风浪扰动可以促进氮磷等营养盐释放,刺激蓝藻细胞生理特征改变,对太湖藻华迁移聚集影响较大,而模型只考虑了风场,流场数据的缺乏导致水动力输移考虑可能不够。总体而言,由于 XGBoost 模型在损失函数寻优过程中加

入了正则项来控制过拟合现象,模拟藻华面积与实测相比,精度较高、稳健性较好。

3 结 论

- a. XGBoost 全太湖与分区的回归模型训练集与验证集模拟效果均较好,LSTM 回归模型训练集较好,但验证集较差,SVM 回归模型介于 XGBoost 与 LSTM 回归模型之间;SVM、XGBoost 回归模型的验证集模拟值偏小,但有效模拟了藻华的发展趋势;选取当天、提前 1 d 的气象水文水质因子作为太湖及分区藻华面积模型的输入,模拟结果较好。
- b. 在全太湖、中西北湖区藻华面积模拟中,XGBoost 分类模型正确率较高;在贡湖、南部沿岸区藻华面积模拟中,SVM、LSTM、XGBoost 分类模型正确率均较高。
- c. SVM、LSTM 模型在藻华面积较低时模拟与实测结果较接近,当藻华面积较高时模拟与实测偏差较大,可能与机器学习方法训练模型一般偏向于数量多、数值低的藻华面积的样本有关;XGBoost 模型精度较高、稳健性较好,适合在太湖藻华面积预测中推广。

参考文献:

[1] 杨柳燕,杨欣妍,任丽曼,等. 太湖蓝藻水华暴发机制与控制对策[J]. 湖泊科学,2019, 31(1):18-27. (YANG Liuyan, YANG Xinyan, REN Liman, et al. Mechanism and control strategy of cyanobacterial bloom in Lake Taihu[J]. Journal of Lake Sciences, 2019, 31(1): 18-27. (in Chinese))

[2] 杭鑫,李心怡,谢小萍. 基于通径分析法的太湖蓝藻水华定量气象评估模型[J]. 湖泊科学, 2019, 31(2): 345-354. (HANG Xin, LI Xinyi, XIE Xiaoping. The quantitative meteorological evaluation model of cyanobacterial bloom in Lake Taihu based on path analysis[J]. Journal of Lake Sciences, 2019, 31(2): 345-354. (in Chinese))

[3] THOMAS M K, LITCHMAN E. Effects of temperature and nitrogen availability on the growth of invasive and native cyanobacteria[J]. Hydrobiologia, 2016,763(1): 357-369.

[4] WU Tingfeng, QIN Boqiang, BROOKES J D, et al. The influence of changes in wind pattern on the areal extension of surface cyanobacterial blooms in a large shallow lake in China[J]. Science of the Total Environment, 2015, 518/519(5): 24-30.

[5] 丁文浩,秦伯强,吴挺峰,等. 夏季风下的太湖风场-流场野外观测研究[J]. 河海大学学报(自然科学版), 2020, 48(2): 102-108. (DING Wenhao, QIN Boqiang, WU Tingfeng. et al. Study on Taihu Lake' s wind field and flow field under summer monsoon[J]. Journal of Hohai University (Natural Sciences), 2020, 48(2):102-108. (in Chinese))

[6] 王华,陈华鑫,徐兆安,等. 2010—2017 年太湖总磷浓度变化趋势分析及成因探讨[J]. 湖泊科学, 2019, 31(4): 919-929. (WANG Hua, CHEN Huaxin, XU Zhaoan, et al. Variation trend of total phosphorus and its controlling factors in Lake Taihu, 2010—2017[J]. Journal of Lake Sciences, 2019, 31(4): 919-929. (in Chinese))

[7] 甘南琴,魏念,宋立荣. 微囊藻毒素生物学功能研究进展[J]. 湖泊科学, 2017, 29(1):1-8. (GAN Nanqin, WEI Nian, SONG Lirong. Recent progress in research of the biological function of microcystins[J]. Journal of Lake Sciences, 2017, 29(1):1-8. (in Chinese))

[8] 戴秀丽,钱佩琪,叶凉,等. 太湖水体氮、磷浓度演变趋势(1985—2015 年)[J]. 湖泊科学, 2016, 28(5): 935-943. (DAI Xiuli, QIAN Peiqi, YELiang. et al. Changes in nitrogen and phosphorus concentrations in Lake Taihu, 1985—2015[J]. Journal of Lake Sciences, 2016, 28(5): 935-943. (in Chinese))

[9] 孔繁翔,高光. 大型浅水富营养化湖泊中蓝藻水华形成机理的思考[J]. 生态学报, 2005, 25(3): 589-595. (KONG Fanxiang, GAO Guang. Hypothesis on cyanobacteria bloom-forming mechanism in large shallow eutrophic lakes[J]. Acta Ecologica Sinica, 2005, 25(3): 589-595. (in Chinese))

[10] 朱伟,陈怀民,王若辰,等. 2017 年太湖水华面积偏大的原因分析[J]. 湖泊科学, 2019, 31(3): 621-632. (ZHU Wei, CHEN Huaimin, WANG Ruochen, et al. Analysis on the reasons for the large bloom area of Lake Taihu in 2017[J]. Journal of Lake Sciences, 2019, 31(3): 621-632. (in Chinese))

[11] 朱广伟,秦伯强,张运林,等. 2005—2017 年北部太湖水体叶绿素 a 和营养盐变化及影响因素[J]. 湖泊科学, 2018, 30(2): 279-295. (ZHU Guangwei, QIN Boqiang, ZHANG Yunlin, et al. Variation and driving factors of nutrients and chlorophyll-a concentrations in northern region of Lake Taihu, China, 2005—2017[J]. Journal of Lake Sciences, 2018, 30(2): 279-295. (in Chinese))

[12] WU Tingfeng, QIN Boqiang, ZHU Guangwei, et al. Dynamic of cyanobacteria bloom formation during short-term hydrodynamic fluctuation in a large shallow, eutrophic, and wind-exposed lake Taihu, China[J]. Environmental Science and Pollution Research, 2013, 20(12):8546-8556.

- [13] 张民,阳振,史晓丽.太湖蓝藻水华的扩张与驱动因素[J].湖泊科学,2019,31(2):336-344. (ZHANG Min, YANG Zhen, SHI Xiaoli. Expansion and drivers of cyanobacterial blooms in Lake Taihu[J]. Journal of Lake Sciences, 2019, 31(2):336-344. (in Chinese))
- [14] XU Hai, PAERL H, QIN Boqiang, et al. Nitrogen and phosphorus inputs control phytoplankton growth in eutrophic Lake Taihu, China[J]. Limnology and Oceanography, 2010, 55(1):420-432.
- [15] PAERL H, HAVENS K E, XU Hai, et al. Mitigating eutrophication and toxic cyanobacterial blooms in large lakes: the evolution of a dual nutrient (N and P) reduction paradigm[J]. Hydrobiologia, 2019(3):1-17.
- [16] PAERL H, PAUL V J. Climate change: links to global expansion of harmful cyanobacteria[J]. Water Research, 2012, 46(5): 1349-1363.
- [17] 孔繁翔,马荣华,高俊峰,等.太湖蓝藻水华的预防、预测和预警的理论与实践[J].湖泊科学,2009,21(3):314-328. (KONG Fanxiang, MA Ronghua, GAO Junfeng, et al. The theory and practice of prevention, forecast and warning on cyanobacteria bloom in Lake Taihu[J]. Journal of Lake Sciences, 2009, 21(3):314-328. (in Chinese))
- [18] 杭鑫,罗晓春,谢小萍.太湖蓝藻水华形成的适宜气象指标[J].应用气象与科学试验,2019,47(1):171-178. (HANG Xin, LUO Xiaochun, XIE Xiaoping. Suitable meteorological indicators for formation of cyanobacteria blooms in Taihu Lake[J]. Meteorological Science and Technology, 2019, 47(1):171-178. (in Chinese))
- [19] 秦伯强,杨桂军,马健荣,等.太湖蓝藻水华"暴发"的动态特征及其机制[J].科学通报,2016,61(7):759-770. (QIN Boqiang, YANG Guijun, MA Jianrong, et al. Dynamics of variability and mechanism of harmful cyanobacteria bloom in Lake Taihu, China[J]. Chinese Science Bulletin, 2016, 61(7):759-770. (in Chinese))
- [20] 王长友,于洋,孙运坤,等.基于ELCOM-CAEDYM模型的太湖蓝藻水华早期预测探讨[J].中国环境科学,2013,33(3):491-502. (WANG Changyou, YU Yang, SUN Yunkun, et al. The discussion of the early forecasting of cyanobacteria bloom in the Lake Taihu based on ELCOM-CAEDYM model[J]. China Environmental Science, 2013, 33(3):491-502. (in Chinese))
- [21] 仝玉华,周洪亮,黄浙丰,等.一种自优化RBF神经网络的叶绿素a浓度时序预测模型[J].生态学报,2011,31(22):6788-6795. (TONGYuhua, ZHOU Hongliang, HUANG Zhefeng, et al. Time series prediction of the concentration of chlorophyll-a based on RBF neural network with parameters self-optimizing[J]. Acta Ecologica Sinica, 2011, 31(22):6788-6795. (in Chinese))
- [22] 王万良,张兆娟,高楠,等.基于人工智能技术的大数据分析方法研究进展[J].计算机集成制造系统,2019,25(3):529-547. (WANG Wanliang, ZHANG Zhaojuan, GAO Nan, et al. Progress of big data analytics methods based on artificial intelligence technology[J]. Computer Integrated Manufacturing Systems, 2019, 25(3):529-547. (in Chinese))
- [23] CRAMER S, KAMPOURIDIS M, FREITAS A A, et al. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives[J]. Expert System with Applications, 2017,85:169-181.
- [24] BISHOP C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.
- [25] MURPHY K P. Machine learning: a probabilistic perspective[M]. London: MIT Press, 2012.
- [26] 蒋志.机器学习方法在雷达定量测量降水及临近预报中的应用研究[D].北京:中国气象科学研究院,2012.
- [27] 苏辉东,贾仰文,倪广恒,等.机器学习在径流预测中的应用研究[J].中国农村水利水电,2018(6):40-43. (SU Huidong, JIA Yangwen, NI Guangheng, et al. The application of machine learning in runoff prediction[J]. China Rural Water and Hydropower, 2018(6):40-43. (in Chinese))
- [28] 刘朔孺,杨敏,张方辉,等.基于支持向量机分类的嘉陵江草街水库甲藻水华预警[J].湖泊科学,2015,27(1):38-43. (LIU Suru, YANG Min, ZHANG Fanghui, et al. Research on early warning of dinoflagellate bloom in Caojie Reservoir base on support vector machine classification[J]. Journal of Lake Sciences, 2015, 27(1):38-43. (in Chinese))
- [29] 张艳会,徐兆安,陈求稳,等.基于BP人工神经网络和模糊理论的太湖蓝藻水华发生风险评价[J].长江流域资源与环境,2011,20(9):1092-1097. (ZHANG Yanhui, XU Zhaoan, CHEN Qiuwen, et al. Integration of bp artificial neural network and fuzzy theory on evaluating cyanobacteria-dominant bloom occurrence risk of Lake Taihu[J]. Resources and Environment in the Yangtza Basin, 2011, 20(9):1092-1097. (in Chinese))
- [30] 于家斌,尚方方,王小艺,等.基于遗传算法改进的一阶滞后滤波和长短期记忆网络的蓝藻水华预测方法[J].计算机应用,2018,38(7):2119-2123. (YU Jiabin, SHANG Fangfang, WANG Xiaoyi, et al. Cyanobacterial bloom forecast method based on genetic algorithm-first order lag filter and long short-term memory network[J]. Journal of Computer Applications, 2018, 38(7):2119-2123. (in Chinese))
- [31] 罗晓春,杭鑫,曹云,等.太湖富营养化条件下影响蓝藻水华的主导气象因子[J].湖泊科学,2019,31(5):1248-1258. (LUO Xiaochun, HANG Xin, CAO Yun, et al. Dominant meteorological factors affecting cyanobacterial blooms under eutrophication in Lake Taihu[J]. Journal of Lake Sciences, 2019, 31(5):1248-1258. (in Chinese))
- [32] 易娟,徐枫,高怡,等.2007年以来环太湖22条主要河流水质变化及其对太湖的影响[J].湖泊科学,2016,28(6):

- 1167-1174. (YI Juan, XU Feng, GAO Yi, et al. Variations of water quality of the major 22 inflow rivers since 2007 and impacts on Lake Taihu[J]. Journal of Lake Sciences, 2016, 28(6): 1167-1174. (in Chinese))
- [33] 张恒敢,顾克军,张斯梅. 用经验正交函数提取太湖 MODIS-EVI 时空分布特征[J]. 生态学杂志, 2018, 37(12): 3802-3808. (ZHANG Henggan, GU Kejun, ZHANG Simei. Extracting temporal and spatial distribution features of Lake Taihu from MODIS-EVI data by empirical orthogonal function analysis[J]. Chinese Journal of Ecology, 2018, 37(12): 3802-3808. (in Chinese))
- [34] 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 Chengcheng, CHEN Qiuwen, XU Qiang, 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. (in Chinese))
- [35] 刘朔孺,杨敏,张方辉,等. 基于支持向量机分类的嘉陵江草街水库甲藻水华预警[J]. 湖泊科学, 2015, 27(1): 38-43. (LIU Suru, YANG Min, ZHANG Fanghui, et al. Research on early warning of dinoflagellate bloom in Caojie Reservoir base on support vector machine classification[J]. Journal of Lake Sciences, 2015, 27(1): 38-43. (in Chinese))
- [36] 张成成,沈爱春,张晓晴,等. 应用支持向量机评价太湖富营养化状态[J]. 生态学报, 2013, 33(23): 7563-7569. (ZHANG Chengcheng, SHEN Aichun, ZHANG Xiaoqing, et al. Application of support vector machine to evaluate the eutrophication status of Taihu Lake[J]. Acta Ecologica Sinica, 2013, 33(23): 7563-7569. (in Chinese))
- [37] 陈曦,仇蕾,黄泽元. 最小二乘支持向量机在太湖流域水质评价中的应用[J]. 水生态学杂志, 2013, 34(6): 16-21. (CHEN Xi, QIU Lei, HUANG Zeyuan. Application of Least Square-Support Vector Machine in water quality assessment of Taihu Lake Basin[J]. Journal of Hydroecology, 2013, 34(6): 16-21. (in Chinese))
- [38] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.
- [39] 秦文虎,陈溪莹. 基于长短时记忆网络的水质预测模型研究[J]. 安全与环境学报, 2020, 20(1): 328-334. (QIN Wenhui, CHEN Xiying. Water quality forecast and prediction model based on long-and short-term memory network[J]. Journal of Safety and Environment, 2020, 20(1): 328-334. (in Chinese))
- [40] 郭燕,赖锡军. 基于长短时记忆神经网络的鄱阳湖水位预测[J]. 湖泊科学, 2020, 32(3): 865-876. (GUO Yan, LAI Xijun. Water level prediction of Lake Poyang based on long short-term memory neural network[J]. Journal of Lake Sciences, 2020, 32(3): 865-876. (in Chinese))
- [41] 徐高扬,刘姚. LSTM 网络在台风路径预测中的应用[J]. 计算机与现代化, 2019, 9(5): 64-68. (XU Gaoyang, LIU Yao. Application of LSTM in typhoon path prediction[J]. Computer and Modernization, 2019, 9(5): 64-68. (in Chinese))
- [42] 刘亚新,樊启祥,尚毅梓,等. 基于 LSTM 神经网络的水电站短期水位预测方法[J]. 水利水电科技进展, 2019, 39(2): 56-60. (LIU Yaxin, FAN Qixiang, SHANG Yizi, et al. Short-term water level prediction method for hydropower station based on LSTM neural network[J]. Advances in Science and Technology of Water Resources, 2019, 39(2): 56-60. (in Chinese))
- [43] 张红斌,邱蝶蝶,邹任重,等. 基于极端梯度提升树算法的图像属性标注[J]. 山东大学学报(工学版), 2019, 49(2): 8-16. (ZHANG Hongbin, QIU Diedie, ZU Renzhong, et al. Image attribute annotation based on extreme gradient boosting algorithm[J]. Journal of Shandong University(Engineering Science), 2019, 49(2): 8-16. (in Chinese))
- [44] 张永梅,陈惠妮,张奕. 基于 XGBoost 的雾霾预测方法[J]. 计算机工程与设计, 2019, 40(12): 3631-3638. (ZHANG Yongmei, CHEN Huini, ZHANG Yi. Image attribute annotation based on extreme gradient boosting algorithm[J]. Journal of Shandong University(Engineering Science), 2019, 40(12): 3631-3638. (in Chinese))
- [45] 孟杭,黄细霞,涂修建. 基于时间序列和 Xgboost 的钢卷仓储吞吐量预测[J]. 计算机应用, 2019, 39(增刊2): 24-28. (MENG Hang, HUANG Xixia, TU Xiujian. Prediction of steel coil storage throughput based on time series and Xgboost[J]. Journal of Computer Applications, 2019, 39(Sup2): 24-28. (in Chinese))
- [46] 黄宇玲,吴达胜,方陆明. 基于逐步回归的 XGboost 方法的森林蓄积量估测[J]. 中南林业科技大学学报, 2020, 40(6): 72-80. (HUANG Yuling, WU Dasheng, FANG Luming. Forest stock volume estimation based on XGboost method of stepwise regression[J]. Journal of Central South University of Forestry & Technology, 2020, 40(6): 72-80. (in Chinese))
- [47] 水利部太湖流域管理局. 2018 年太湖健康状况报告[R]. 北京:水利部太湖流域管理局, 2019.
- [48] 裴洪平,罗妮娜,蒋勇. 利用 BP 神经网络方法预测西湖叶绿素 a 的浓度[J]. 生态学报, 2004, 24(2): 246-251. (PEI Hongping, LUO Nina, JIANG Yong. Applications of back propagation neural network for predicting the concentration of chlorophyll-a in West Lake[J]. Acta Ecologica Sinica, 2004, 24(2): 246-251. (in Chinese))