

doi: 10.6046/zrzyyg.2022045

引用格式: 田晨,张金龙,金义蓉,等.一种利用贝叶斯优化的蓝藻遥感分类方法[J].自然资源遥感,2023,35(1):49-56.
(Tian C,Zhang J L,Jin Y R,et al.A remote sensing classification method for cyanobacteria using Bayesian optimization algorithm[J].
Remote Sensing for Natural Resources,2023,35(1):49-56.)

一种利用贝叶斯优化的蓝藻遥感分类方法

田 晨¹,张金龙¹,金义蓉¹,董世元²,王 彬²,张乃祥²

(1. 苏州市水利水务信息调度指挥中心,苏州 215011; 2. 苏州中科蓝迪软件技术有限公司,苏州 215163)

摘要: 利用 Sentinel-2 遥感卫星影像,结合遥感优势以光谱、指数、纹理等 14 种多种特征信息为输入,依托贝叶斯优化算法,设计了一种能自动获取最优超参数组合的 BO-XGBoost 方法,并将其成功应用于 2021 年阳澄湖蓝藻信息提取。结果表明:①通过贝叶斯优化算法获取最优超参数组合,进行训练得到 BO-XGBoost 蓝藻分类模型,其训练结果在测试集和训练集上表现效果良好,准确率高达 96.07%;②将 BO-XGBoost 应用于参与样本集构建的影像,其蓝藻识别结果与人工解译成果对比,2 种方法得到的蓝藻空间分布情况基本一致,交并比最低为 41.31%;③为评价该分类模型在其他时相的适用性,选择其他时相影像数据进行蓝藻提取,BO-XGBoost 与人工解译 2 种方法蓝藻空间分布情况基本一致,交并比最低为 43.85%。

关键词: 贝叶斯优化; BO-XGBoost; 多特征; 蓝藻; Sentinel-2

中图法分类号: TP 79 **文献标志码:** A **文章编号:** 2097-035X(2023)01-0049-08

0 引言

蓝藻大面积暴发,不仅破坏湖泊景观,导致生物量减少,同时也会造成水体自净能力失衡,严重影响湖泊水生态环境^[1]。自“十三五”规划以来,内陆湖泊生态环境得到改善,但是受常年环境问题影响,蓝藻暴发等一系列问题仍然严重威胁水生生态系统。随着遥感技术的发展,越来越多的卫星遥感技术应用于环境治理中^[2-3],利用卫星遥感技术的空间优势可实时快捷地获取蓝藻暴发区域地时空分布特征,掌握蓝藻生长情况,为环境治理、管控提供数据和技术支撑。

目前,国内外众多学者采用多种方法进行遥感影像蓝藻信息提取。其中光谱指数法应用较为广泛,该方法通过蓝藻光谱特性构建能够表征蓝藻特征的光谱指数,主要包括:归一化植被指数(normalized difference vegetation index, NDVI)^[4]、增强型植被指数(enhanced vegetation index, EVI)^[2-3,5]、漂浮藻类指数(floating algae index, FAI)^[5]、替代型漂浮藻类指数(alternative floating algae index, AFAI)^[6]等;同时也有学者进一步结合多个光谱指数构建利

用机器学习分类方法,如:决策树、支持向量机、神经网络、随机森林等蓝藻提取方法^[7-12]。此外还有学者提出了蓝藻提取的非监督分类方法^[13]。综合对比以上几种分类方法,光谱指数法通过专家经验设定固定阈值对蓝藻信息提取,缺乏可拓展性,难以针对长时间序列的遥感影像数据集确定唯一有效阈值进行蓝藻识别;监督分类方法识别精度相比于其他方法精度较高,但在训练过程中需要设定模型超参数,然而对于模型超参数的选择上往往靠过去的经验和大量的人工调试,很难得到最优的网络结构和参数,另一方面训练蓝藻分类模型需要标记大量样本,人工成本较高;非监督分类方法的分类精度则无法保证,需要大量的长时间序列数据进行分析。

XGBoost(eXtreme Gradient Boost)于 2016 年 SIGKDD 大会提出,是一种结合梯度提升和决策树的集成学习方法^[14],能够自动利用 CPU 的多线程进行并行,在算法运行效率和精度上有明显提升,近年来广泛应用于遥感监测领域,如总磷、总氮、叶绿素 a 等水质指标反演^[15],以及遥感地物分类^[16],目前 XGBoost 分类算法在遥感蓝藻监测研究中鲜有报道。机器学习分类模型训练精度受超参数影响较

收稿日期: 2022-02-11; 修订日期: 2022-08-22

基金项目: 苏州市水利水务科技项目“基于光学与 SAR 联合观测的湖泊凤眼莲分布自动化提取研究”(编号: 2021009)资助。

第一作者: 田 晨(1982-),男,硕士,主要从事环境水文学方向研究工作。Email: tc1115@126.com。

通信作者: 董世元(1994-),男,硕士,主要从事遥感信息提取、机器学习、深度学习等研究工作。Email: dongsyrs@163.com。

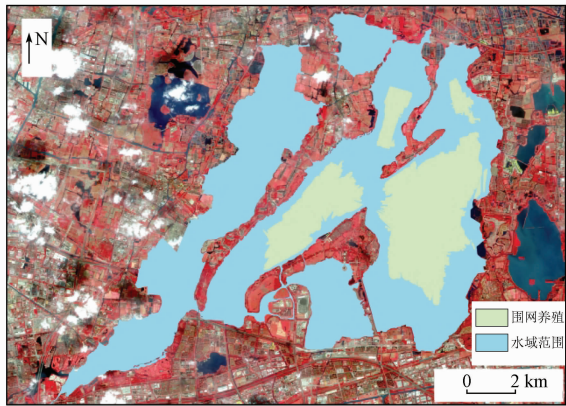
大,较差的超参数组合策略易导致模型过拟合或者欠拟合,贝叶斯优化算法(Bayesian optimization)利用先验知识逼近未知函数的后验分布从而调节最优超参数组合,相较于传统的专家经验设定、网格检索等方法,贝叶斯优化算法更能反映实际问题的本质,并在众多领域中获得应用,如:经济预测、舆情分析等领域^[17]。在以往遥感监测分类算法研究中,机器学习超参数选取一般基于专家经验进行选择,利用贝叶斯优化寻找最优解的方法涉及较少。

光学遥感影像包含光谱特征、指数特征、纹理特征,结合多遥感特征进行蓝藻分类提取可以避免光谱混淆,提升模型精度^[18]。本文以遥感影像包含的光谱、指数、灰度共生矩阵等多特征信息作为输入,利用贝叶斯优化算法寻找XGBoost模型全局最优解对蓝藻识别模型进行训练,生成BO-XGBoost蓝藻识别模型,将该模型应用于2021年阳澄湖蓝藻信息提取,与人工标注结果进行对比验证,并针对BO-XGBoost模型的适用性和优缺点进行评价。本研究的成果可为遥感领域机器学习算法优化问题提供参考,以及分析阳澄湖蓝藻发生规律和进一步湖泊生态环境治理提供技术指导。

1 研究区概况与数据源

1.1 研究区概况

阳澄湖位于江苏省苏州市北部,是典型内陆湖泊,占地约120 km²。阳澄湖由于富营养化引起的蓝藻暴发是水生态安全、生物多样化的重要影响因素,严重威胁苏州市、昆山市饮用水安全^[19]。图1展示了阳澄湖水域范围,其中,阳澄湖包含大面积的大闸蟹围网养殖区域,易与蓝藻发生混淆,经观察发现阳澄湖围网养殖区域常年不变,因此该部分区域不在研究范围内。



(底图为 Sentinel-2 B8(R), B4(G), B3(B) 波段合成影像)

图1 阳澄湖水域范围

Fig.1 Yangcheng Lake water area

1.2 数据获取及处理

阳澄湖湖泊面积较小,及时掌握阳澄湖蓝藻暴发的空间分布形态对卫星遥感影像空间、时间分辨率要求较高。本文使用欧洲航空安全局(European Aviation Safety Agency, ESA)提供的免费数据源 Sentinel-2,其空间分辨率为10 m, S2A 和 S2B 星联合观测时间分辨率可达5 d^[20],能够满足阳澄湖蓝藻暴发监测的需求。Sentinel-2 数据可从 ESA 官方网站(<https://scihub.copernicus.eu/dhus/#/home>)进行下载,下载得到 L1C 数据,该数据已进行过几何纠正和辐射校正,利用官方推荐的 Sen2cor 软件对其进行大气校正,可得到10 m 分辨率的反射率数据。本文使用2021年6—10月份无云或少云的8期遥感数据进行蓝藻监测,表1展示了本文中使用的 Sentinel-2 卫星数据日期。

表1 阳澄湖影像过境时间

Tab.1 Yangcheng Lake image transboundary time			
序号	日期	云层覆盖率/%	载荷类型
1	6月22日	0	S2B 星
2	7月7日	0.85	S2A 星
3	7月12日	0	S2B 星
4	8月26日	0	S2A 星
5	8月31日	0	S2B 星
6	9月30日	2.48	S2B 星
7	10月5日	0	S2A 星
8	10月30日	1.26	S2B 星

2 研究方法

2.1 特征提取

光谱特征是表征地物类型的重要手段之一,本文中选取的光谱特征为 Sentinel-2 影像数据的4个波段。同时选取能够表征蓝藻、水体的光谱指数 NDVI、归一化差异水体指数(normalized difference water index, NDWI),计算公式分别为:

$$NDVI = \frac{NIR - Red}{NIR + Red}, \tag{1}$$

$$NDWI = \frac{Green - NIR}{Green + NIR}, \tag{2}$$

式中 NIR, Red 和 Green 分别为近红外、红光和绿光波段,对应 Sentinel-2 影像 B8, B4 和 B3 波段。

通过光谱特征和光谱指数只能表征单一像元的地物类型,忽略了遥感影像携带的空间纹理特征,本文使用常用的纹理特征提取方法——灰度共生矩阵,通过计算相对位置固定像素对的灰度频率来提取纹理信息^[21]。本文利用 NDVI 作为灰度矩阵,灰度量化级选择值为32,统计像元窗口大小为9像元×

9 像元,提取了 8 个纹理特征,包括: 对比度、相异性、同质性、角二阶矩、方差、均值、最大概率和相关性。

将光谱特征、指数特征、纹理特征共 14 个特征进行融合,多特征图如图 2 所示,并用于 Sentinel-2 的蓝藻分类提取特征集。

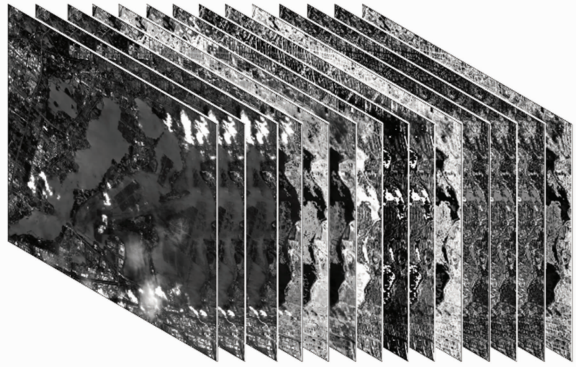


图 2 多特征分布图

Fig. 2 Multi - feature distribution map

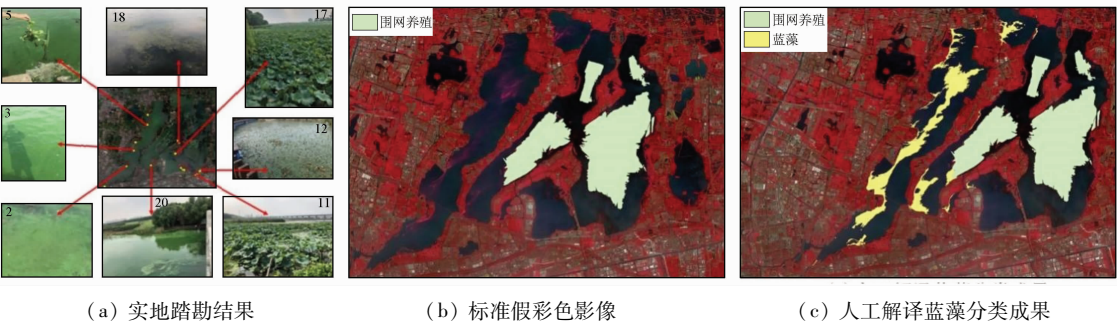


图 3 2021 年 8 月 26 日数据展示

Fig. 3 Data display on August 26, 2021

被、水体、云为背景。剩余 4 期影像数据不参与模型训练过程,以便验证分类模型的泛化能力。

2.3 基于贝叶斯优化的 XGBoost 分类方法

贝叶斯优化是根据目标函数 $f(x)$ 的历史评价结果建立计算先验分布,结合前次迭代得到的观察点 E ,来获取后验分布,进而基于后验信息选择下一个样本点,使目标函数 $f(x)$ 最小^[22]。后验分布的具体计算方法公式为:

$$P(M \mid E) \propto P(E \mid M)P(M) \text{ , } \quad (3)$$

式中: M 为模型目标函数值; E 为观察点表示已观测集合; $P(M)$ 、 $P(M \mid E)$ 分别为 M 的先验和后验概率分布; $P(E \mid M)$ 为似然比概率。 E 和 M 之间满足如下递归关系:

$$E_1 = \{(\theta_1, M_1)\} , D_i = \{D_{i-1}, (\theta_i, M_i)\} \text{ , } \quad (4)$$

式中: M_i 为与 θ_i 对应的目标函数值; D_i 为获取评估点 θ_i 的函数值 M_i ,并将其加入到评估集合中。

另外,贝叶斯优化算法有 2 个核心部分,即概率代理模型和采集函数^[23]。本文选用高斯过程^[24-25]

2.2 样本选取

2021 年 6—10 月期间,选择天气晴朗且有 Sentinel-2 影像数据过境日期对阳澄湖进行实地踏勘。图 3 为 2021 年 8 月 26 日阳澄湖实地踏勘成果和对应日期的 Sentinel-2 影像数据,图 3(a)可以看出点位 2,3,5,20 有藻类聚集,点位 11,12,17,18 有水生植被聚集。观察图 3(a)和图 3(b)蓝藻与其他水生植被在影像呈现的状态,发现蓝藻多以大面积,絮状分布为主,其他水生植被一般有明显的分类边界,以小范围、零散、近陆地分布为主。通过计算 NDVI,并结合实地调研结果,对蓝藻在遥感影像上进行标注,标注结果如图 3(c)。选取表 1 中 4 景遥感影像数据进行样本提取,分别为 7 月 7 日、8 月 26 日、9 月 30 日、10 月 30 日,根据人工标注结果,针对蓝藻、其他植被、水体、云 4 类目标物各随机选取 250 个点位共计 1 000 个样本数据,其中蓝藻为目标,其他植

和 PI 函数^[25]作为概率代理模型和采集函数。

XGBoost 是基于决策树的监督模型,属于集成学习 Boosting 算法的一种,思想是使用多个 CART 分类树作为弱学习器,增加的决策树学习上一轮预测结果的残差,通过不断地增加新树使得损失值最低,最终基于加法模型将多个弱学习器集成为一个强学习器,进而获得一个高精度的机器学习模型。XGBoost 算法需要在训练前设定子模型数量、学习率、树最大深度、惩罚系数等超参数,使用人工检索或网格搜索方法确定最优超参数组合达到最小误差和避免过拟合,会消耗大量计算时间。本文使用贝叶斯优化自动调整 XGBoost 超参数,在有限的检索次数内达到最优值,从而保证得到最优超参数组合进行蓝藻识别和提取。

利用贝叶斯优化算法,对 XGBoost 算法进行 30 次评估,单次评估过程中设置五折交叉验证,并利用准确率指标对目标参数进行评估。在贝叶斯优化过程中,每次观测到的准确率如图 4 所示。在贝叶斯优化评估结束后,得到最优可行点,最优参数取值表

2 所示,在此参数下 XGBoost 模型对蓝藻的识别性
能在验证集上准确度达到 98.71%。

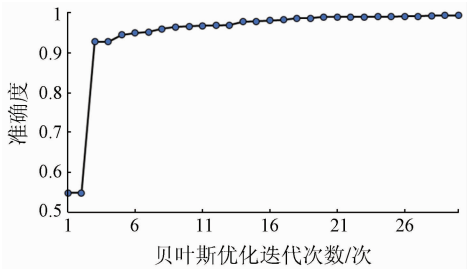


图 4 贝叶斯优化迭代情况

Fig.4 Bayesian optimization iteration case

表 2 XGBoost 最优超参数组合

Tab.2 XGBoost optimal hyperparameter combination

变量	含义	取值范围	步长	最优值
<i>max_depth</i>	子模型最大深度	[5, 15]	1	4
<i>learning_rate</i>	学习率	0.001, 0.01, 0.1, 1, 10	—	0.1
<i>n_estimators</i>	子模型数量	[10, 300]	10	35
<i>subsample</i>	随机采样的比例	(0, 1)	0.1	0.5
<i>colsample_bytree</i>	特征占比	(0, 1]	0.1	0.4
<i>min_child_weight</i>	最小叶节点样本权重和	[1, 10]	1	4

2.4 精度评价

通过混淆矩阵对 BO - XGBoost 蓝藻分类模型
精度进行评估,确定 4 个评价指标,即精确率(Preci-
sion)、召回率(Recall)、F1 指标和 Kappa 系数。另

外,利用交并比(intersection over union, IoU)评价人
工标注成果和识别结果之间的差异大小,以验证分
类模型实际应用效果。精确率描述正确分类的数量
与模型识别出的数量占比;召回率描述正确分类的
数量与真实数量的占比;F1 指标为精确率和召回
率的调和平均数;Kappa 系数描述模型预测结果和
实际分类结果的一致性;IoU 描述正确分类区域与
模型识别结果和真实区域交集的比值。

3 结果与分析

3.1 模型分类结果

采用依托贝叶斯优化生成的最优参数组合,利
用 XGBoost 算法对提取的 1 000 条样本集进行训
练,随机抽取 70% 样本数据作为样本集,30% 样
本数据作为测试集。经训练后生成 BO - XGBoost 模
型,利用精确率、召回率、F1 指标、Kappa 系数对样
本集进行评价,表 3 展示了 BO - XGBoost 分类算
法分别在训练集和测试集上的表现,其中水体、云、其
他植被为背景物,蓝藻为目标物。从表 3 可以看出,
在训练集、测试集上,云类识别精度最高,水体次之,
蓝藻精度较低但 F1 指标仍在 98% 以上。BO - XG-
Boost 在测试集上各项指标精度均高于 96.07%,且测
试集、训练集表现相似相差较小,该模型在样本数据
训练过程中不存在过拟合和欠拟合的问题。

表 3 分类模型精度评价

Tab.3 Classification Model Accuracy Evaluation

评价指标	训练集				测试集			
	蓝藻	水体	云	其他植被	蓝藻	水体	云	其他植被
精确率/%	98.61	98.86	99.88	98.63	96.07	99.32	99.46	98.11
召回率/%	97.82	99.57	99.77	98.97	97.86	98.66	98.93	97.33
F1/%	98.22	99.21	99.82	98.80	96.96	98.99	99.19	97.72
Kappa	0.986 7				0.975 6			

为确定 BO - XGBoost 分类模型的实际应用情
况,利用 7 月 7 日、8 月 26 日、9 月 30 日、10 月 30 日
参与样本集提取的 4 期影像数据进行蓝藻提取。图
5(a) — (d) 展示了 Sentinel - 2 的标准假彩色影像,
图 5(e) — (h) 展示了 BO - XGBoost 的蓝藻分类结
果,其中黄色区域为蓝藻区域,图 5(i) — (l) 展示了
人工判别标注的蓝藻分类结果。对比 BO - XGBoost
和人工解译蓝藻提取成果可发现,人工目视解译成
果的蓝藻空间分布与 BO - XGBoost 识别结果相似,
图中可知 4 期遥感影像数据蓝藻主要暴发区域为阳
澄西湖,2 种方法蓝藻区域表征基本一致。在蓝藻
聚集程度较低的区域,遥感影像上蓝藻与水体边界
不清晰,表征不明显,仅利用真彩色、假彩色、NDVI
等指数进行人工解译的方法,蓝藻提取结果受主观

因素影响较大,尤其蓝藻边界区域的划分。BO -
XGBoost 机器学习分类算法具备一定客观性,同时
该算法从光谱、纹理、光谱指数多个角度对蓝藻区
域进行判别,可有效对蓝藻边界进行区分。

为更好地表征 BO - XGBoost 算法与人工解译
提取的蓝藻结果的区别,选取 8 月 26 日发生大规模
蓝藻聚集区域进行细节对比,如图 6 所示。从图中
可以看出:①BO - XGBoost 相比于人工解译方法,
蓝藻监测区域较大,能够提取出低聚集蓝藻区域;
②BO - XGBoost 算法受岸线混合像元影响,将岸
边植被误提取为蓝藻;③人工解译结果标准不统一,
如图 6(c) 红框范围处蓝藻零星出现,但人为判断
过程中将此处标记为蓝藻。

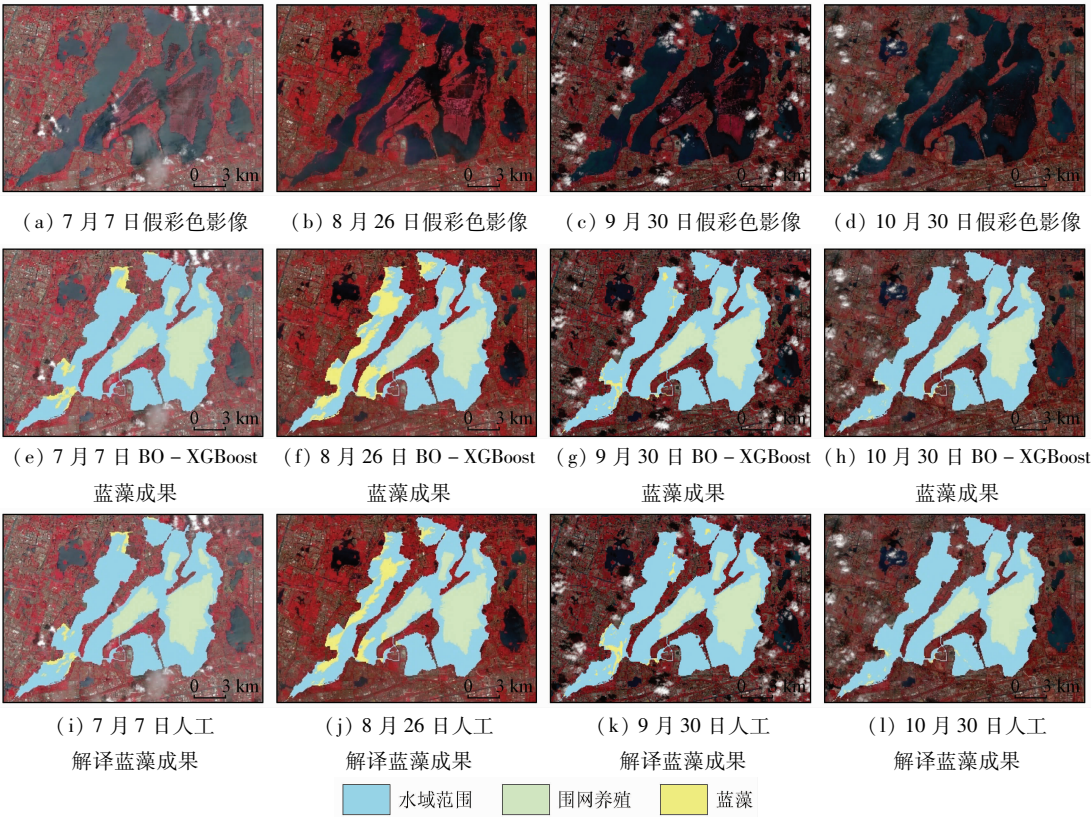


图 5 2021 年参与样本集构建的蓝藻识别成果对比

Fig. 5 Comparison of cyanobacterial identification results participating in the construction of the sample set in 2021

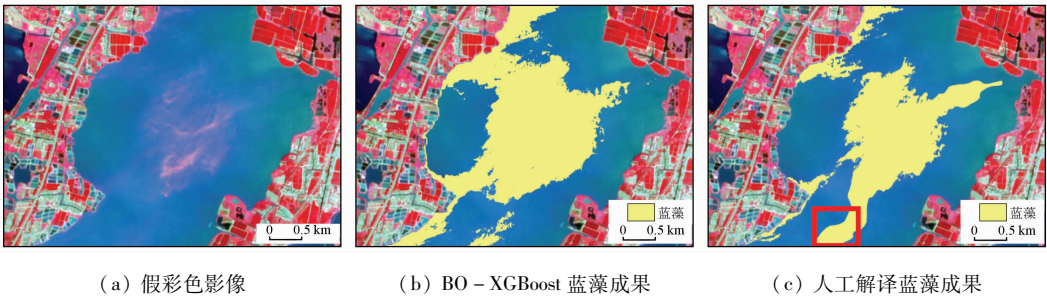


图 6 8 月 26 日蓝藻识别成果细节对比

Fig. 6 Comparison of details of cyanobacteria identification results on August 26

为量化描述 BO - XGBoost 分类模型与人工解译成果之间的蓝藻分类差异程度。表 4 展示了 BO - XGBoost 分类算法和人工解译提取的蓝藻面积,以及 2 种方法得到蓝藻区域的 IoU 计算结果。IoU 指标存在一定差异,2 种方法得到的蓝藻区域 IoU 高于 41.31%,其中 7 月 7 日最高为 55.46%,造成这种差异的主要原因在于:人工解译进行蓝藻提取过程主要利用 NDVI 值和人工目视解译方法进行判别,针对不同区域不同情况下 NDVI 划分阈值,最后再依据人工经验对 NDVI 划分阈值的结果边界进行修订,最终得到蓝藻提取结果,过程中受人为主观影响因素较大。对比图 5 的蓝藻成果细节,IoU 指标差异主要体现在以下 2 点:①算法提取结果受岸边混合像元影响,岸边易被误识别为蓝藻,而人工解

表 4 蓝藻提取面积及 IoU 评价指标
Tab.4 Cyanobacteria extraction area and IoU evaluation index

日期	<i>IoU</i> /%	BO - XGBoost /km ²	人工解译 /km ²
7 月 7 日	55.46	4.289 2	2.743 6
8 月 26 日	41.53	16.521 6	11.235 8
9 月 30 日	41.31	2.305 1	2.897 4
10 月 30 日	43.28	0.896 8	0.812 4

译结果,在人工审核过程中已经去除岸边的细碎斑点;②由于人工解译过程,其结果受人为主观因素影响,其边界判定条件不统一,无法客观表征蓝藻提取结果。

综上所述,基于贝叶斯优化进行 XGBoost 模型超参数选取的方法,提取的超参数组合适用于遥感

蓝藻分类。同时,基于光谱、指数、灰度共生矩阵为多特征输入,其训练结果,应用 Sentinel-2 影像数据时,与人工判读结果区域表征基本一致,蓝藻提取结果和其他地物边界区分明显。

3.2 适用性分析

受大气状况、气象条件、传感器硬件等影响,大部分研究人员利用阈值对蓝藻进行提取时所利用的阈值相差甚远^[26-28]。为验证 BO-XGBoost 分类算法在其他时序遥感影像上的适用性,本文将 6 月 22 日、7 月 12 日、8 月 31 日、10 月 5 日未参与样本提取的 4 期遥感影像数据,分别利用 BO-XGBoost 和人工解译 2 种分类方法进行蓝藻提取。

表 5 展示了 BO-XGBoost 算法和人工解译提取的蓝藻面积和交并比 IoU 指标。IoU 指标在 7 月 12 日最低为 43.85%,10 月 5 日最高为 50.22%。6

月 22 日、7 月 12 日 BO-XGBoost 蓝藻提取面积高于人工解译结果,其余 2 期面积接近。图 7 展示了 BO-XGBoost 算法和人工解译 2 种方法提取的蓝藻空间分布情况,图中黄色区域为蓝藻分布情况,对比图 7(b)和(c)可以发现,2 种方法提取的蓝藻分布情况基本一致,蓝藻暴发区域主要集中于阳澄西湖沿岸附近。

表 5 蓝藻提取面积及 IoU 评价指标
Tab.5 Cyanobacteria extraction area and IoU evaluation index

日期	<i>IoU</i> /%	BO-XGBoost /km ²	人工解译 /km ²
6 月 22 日	44.75	4.073 8	3.167 1
7 月 12 日	43.85	3.465 3	1.671 9
8 月 31 日	44.01	1.486 6	1.642 2
10 月 5 日	50.22	0.585 4	0.678 4

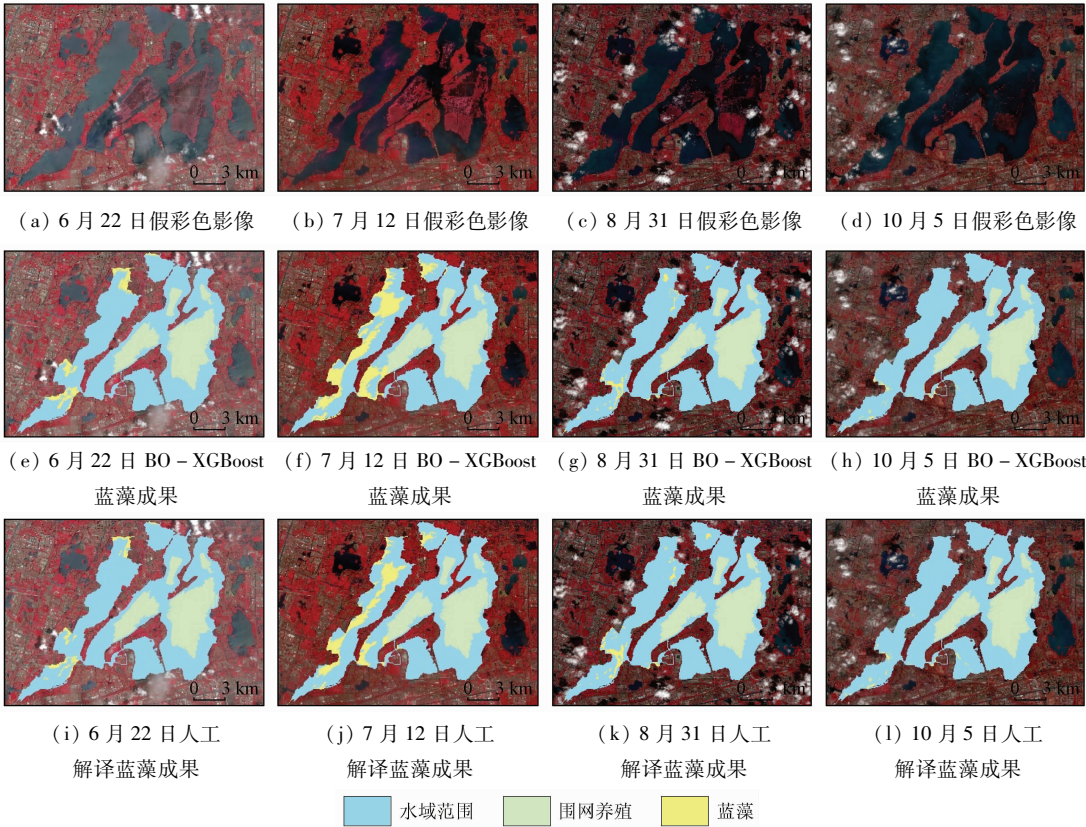


图 7 2021 年未参与样本集构建的蓝藻识别成果对比
Fig. 7 Comparison of cyanobacterial identification results that did not participate in the construction of the sample set in 2021

4 结论

本文以江苏省苏州市阳澄湖为研究区域,利用 Sentinel-2 遥感图像数据以及实地踏勘情况对样本进行标注,以遥感数据特有的光谱、指数、纹理等多特征为分类模型输入,结合贝叶斯优化和 XGBoost

算法构建 BO-XGBoost 蓝藻提取模型。经实验验证该模型算法对 2021 年不同月份蓝藻监测具有很好的适用性,蓝藻分类准确度高达 96%,实际应用中提取的蓝藻与人工提取的蓝藻空间分布结果一致性较高,并经长时序影像数据验证该模型泛化能力强,能够适应不同月份的蓝藻提取和分类应用。

本文提出的 BO-XGBoost 训练模式,可自动确

定 XGBoost 模型最优超参数组合,且该参数组合能够在提升模型精度的同时,避免模型过拟合和欠拟合情况发生。结合遥感影像优势,以光谱、指数、灰度共生矩阵多特征为输入进行模型训练,其蓝藻提别结果与其他地物边界区分明显,蓝藻提取区域与人工解译区域基本一致。表明本文所提方法能够有效对湖泊蓝藻进行自动有效提取。

本文虽然利用贝叶斯优化、XGBoost、多特征进行湖泊蓝藻信息提取,同时验证了在相同区域其他时相遥感影像数据蓝藻信息提取的有效性,但仍存在不足之处在于,该蓝藻识别模型是否可直接应用于其他湖泊进行蓝藻信息提取有待验证;输入多特征进行模型训练时,光谱、指数、灰度共生矩阵包含的 14 类特征哪一类特征更能表征蓝藻分布情况,并未进行验证。后续工作将针对其他湖泊区域进行实验以验证模型的泛化能力,另外基于基尼指数、卡方检验等方法对特征重要性进行评价。

参考文献 (References) :

[1] 胡智华,林妙丽,李 港,等. 城市闸控河流浮游植物群落结构特征及影响因素[J]. 环境科学学报,2021,41(9):3631 – 3640.

Hu Z H,Lin M L,Li G,et al. Phytoplankton community characteristics and influencing factors of the urban gate – controlled rivers [J]. Acta Scientiae Circumstantiae,2021,41(9):3631 – 3640.

[2] 岩 腊,龙 笛,白亮亮,等. 基于多源信息的水资源立体监测研究综述[J]. 遥感学报,2020,24(7):787 – 803.

Yan L,Long D,Bai L L,et al. A review on water resources stereoscopic monitoring systems based on multisource data[J]. National Remote Sensing Bulletin,2020,24(7):787 – 803.

[3] 谢 勇,李凯云,李家国,等. Sentinel – 2A 影像珠海市污染水体识别[J]. 遥感信息,2021,36(3):1 – 9.

Xie Y,Li K Y,Li J G,et al. Water pollution identification in Zhuhai City based on Sentinel – 2A imagery[J]. Remote Sensing Information,2021,36(3):1 – 9.

[4] 邱亚会,卢剑波. 浒苔遥感监测研究进展[J]. 生态学报,2015,35(15):4977 – 4985.

Qiu Y H,Lu J B. Advances in the monitoring of Enteromorpha prolifera using remote sensing[J]. Acta Ecologica Sinica,2015,35(15):4977 – 4985.

[5] Xing Q G,Hu C M. Mapping macroalgal blooms in the Yellow Sea and East China Sea using HJ – 1 and Landsat data: Application of a virtual baseline reflectance height technique[J]. Remote Sensing of Environment,2016,178:113 – 126.

[6] 陈春波,周宝同,田永中,等. 环境一号卫星多光谱数据在巢湖动态监测中的应用[J]. 中国环境监测,2014,30(1):200 – 204.

Chen C B,Zhou B T,Tian Y Z,et al. Application of environmental satellite HJ1A/1B – CCD data for Cyanobacteria dynamic monitoring in Chaohu Lake [J]. Environmental Monitoring of China,2014,30(1):200 – 204.

[7] Hu C. A novel ocean color index to detect floating algae in the global oceans [J]. Remote Sensing of Environment,2009,113(10):2118 – 2129.

[8] Fang C,Song K,Shang Y,et al. Remote sensing of harmful algal blooms variability for Lake Hulun using adjusted FAI (AFAI) Algorithm[J]. Journal of Environmental Informatics,2018:54 – 68.

[9] Otukey J R,Blaschke T. Land cover change assessment using decision trees,support vector machines and maximum likelihood classification algorithms[J]. International Journal of Applied Earth Observation and Geoinformation,2010,12:27 – 31.

[10] 李旭青,刘世盟,李 龙,等. 基于 RF 算法优选多时相特征的冬小麦空间分布自动解译[J]. 农业机械学报,2019,50(6):218 – 225.

Li X Q,Liu S M,Li L,et al. Automatic interpretation of spatial distribution of winter wheat based on random forest algorithm to optimize multi – temporal features[J]. Transactions of the Chinese Society for Agricultural Machinery,2019,50(6):218 – 225.

[11] 李苓苓,潘耀忠,张锦水,等. 支持向量机与分类后验概率空间变化向量分析法相结合的冬小麦种植面积测量方法[J]. 农业工程学报,2010,26(9):210 – 217.

Li L L,Pan Y Z,Zhang J S,et al. Method of winter wheat planting area estimation based on support vector machine and post – classification changed vector analysis[J]. Transactions of the Chinese Society of Agricultural Engineering,2010,26(9):210 – 217.

[12] 赵 莲,张锦水,胡潭潭,等. 变端元混合像元分解冬小麦种植面积测量方法[J]. 国土资源遥感,2011,23(1):66 – 72. doi:10.6046/gtzyyg.2011.01.13.

Zhao L,Zhang J S,Hu T G,et al. The application of the dynamic endmember linear spectral unmixing model to winter wheat area estimation[J]. Remote Sensing for Land and Resources,2011,23(1):66 – 72. doi:10.6046/gtzyyg.2011.01.13.

[13] 王怡人,王胜强,喻 樾,等. 一种提取南黄海浒苔的自适应阈值遥感算法[J]. 遥感信息,2021,36(2):120 – 129.

Wang Y R,Wang S Q,Yu Y,et al. An adaptive threshold algorithm for detecting ulva prolifera in southern Yellow Sea by remote sensing[J]. Remote Sensing Information,2021,36(2):120 – 129.

[14] Chen T W,Guestrin C. XGBoost: A scalable tree boosting system [C]//22nd ACM SIGKDD International Conference. ACM,2016.

[15] 卢雪梅,苏 华. 基于 OLCI 数据的福建近海悬浮物浓度遥感反演[J]. 环境科学学报,2020,40(8):2819 – 2827.

Lu X M,Su H. Retrieving total suspended matter concentration in Fujian coastal waters using OLCI data[J]. Acta Scientiae Circumstantiae,2020,40(8):2819 – 2827.

[16] 黄 宇,潘 励. 基于显著性图像与纹理特征的遥感影像云检测[J]. 测绘地理信息,2021,46(2):16 – 19.

Huang Y,Pan L. Cloud detection of remote sensing images based on significance maps and texture features[J]. Journal of Geomatics,2021,46(2):16 – 19.

[17] 徐 逸,甄佳宁,蒋侠朋,等. 无人机遥感与 XGBoost 的红树林物种分类[J]. 遥感学报,2021,25(3):737 – 752.

Xu Y,Zhen J N,Jiang X P,et al. Mangrove species classification with UAV – based remote sensing data and XGBoost[J]. National Remote Sensing Bulletin,2021,25(3):737 – 752.

[18] 李 鹏,冯存前,许旭光,等. 一种利用贝叶斯优化的弹道目标微动分类网络[J]. 西安电子科技大学学报,2021,48(5):139 –

148.

Li P, Feng C Q, Xu X G, et al. Ballistic target fretting classification network based on Bayesian optimization[J]. Journal of Xidian University, 2021, 48(5): 139 – 148.

[19] 孙淑文, 夏 霆, 杨金艳, 等. 苏州湖泊水源地浮游植物群落特征及营养状况分析[J]. 中国农村水利水电, 2020(8): 51 – 55.

Sun S W, Xia T, Yang J Y, et al. An analysis of the phytoplankton community characteristics and nutrition status in lake – type drinking water sources in Suzhou[J]. China Rural Water and Hydropower, 2020(8): 51 – 55.

[20] 刘瑞清, 李加林, 孙 超, 等. 基于 Sentinel – 2 遥感时间序列植被物候特征的盐城滨海湿地植被分类[J]. 地理学报, 2021, 76(7): 1680 – 1692.

Liu R Q, Li J L, Sun C, et al. Classification of Yancheng coastal wetland vegetation based on vegetation phenological characteristics derived from Sentinel – 2 time – series[J]. Acta Geographica Sinica, 2021, 76(7): 1680 – 1692.

[21] Guo X J, Zhang C C, Luo W R, et al. Urban impervious surface extraction based on multi – features and random forest[J]. IEEE Access, 2020, 8: 226609 – 226623.

[22] Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for hyper – parameter optimization[J]. Advances in Neural Information Processing Systems, 2011, 24(24): 2546 – 2554.

[23] Tanaka R, Iwata H. Bayesian optimization for genomic selection: A method for discovering the best genotype among a large number of candidates[J]. Theoretical and Applied Genetics, 2017, 131(1): 1 – 13.

[24] Seeger M. Gaussian processes for machine learning[J]. International Journal of Neural Systems, 2004, 14(2): 69 – 106.

[25] Colopy G W, Roberts S J, Clifton D A. Bayesian optimization of personalized models for patient vital – sign monitoring[J]. IEEE Journal of Biomedical and Health Informatics, 2018, 22(2): 301.

[26] 陈 云, 戴锦芳. 基于遥感数据的太湖蓝藻水华信息识别方法[J]. 湖泊科学, 2008, 20(2): 179 – 183.

Chen Y, Dai J F. Extraction methods of cyanobacteria bloom in Lake Taihu based on RS data[J]. Journal of Lake Science, 2008, 20(2): 179 – 183.

[27] 张东彦, 尹 勋, 余 宝, 等. 多源卫星遥感数据监测巢湖蓝藻水华爆发研究[J]. 红外与激光工程, 2019, 48(7): 303 – 314.

Zhang D Y, Yin X, She B, et al. Using multi – source satellite imagery data to monitor cyanobacterial blooms of Chaohu Lake[J]. Infrared and Laser Engineering, 2019, 48(7): 303 – 314.

[28] 刘海秋, 任恒奎, 牛鑫鑫, 等. 基于 Sentinel – 2 遥感影像的巢湖蓝藻水华提取方法研究[J]. 生态环境学报, 2021, 30(1): 146 – 155.

Liu H Q, Ren H K, Niu X X, et al. Extraction of cyanobacteria bloom in Chaohu Lake based on Sentinel – 2 remote sensing images[J]. Ecology and Environmental Sciences, 2021, 30(1): 146 – 155.

A remote sensing classification method for cyanobacteria using Bayesian optimization algorithm

TIAN Chen¹, ZHANG Jinlong¹, JIN Yirong¹, DONG Shiyuan², WANG Bin², ZHANG Naixiang²
(1. Suzhou Water Conservancy and Water Information Dispatching Command Center, Suzhou 215011, China; 2. Suzhou Land Think Software Technology Company Limited of Chinese Academy of Science, Suzhou 215163, China)

Abstract: With 14 types of multi – feature information, such as spectrum, index, and texture, of remote sensing images from satellite Sentinel – 2 as input and using the Bayesian optimization algorithm, this study designed the BO – XGBoost method used to automatically obtain the optimal hyperparameter combination. This method was successfully applied to the information extraction of cyanobacteria in Yangcheng Lake in 2021. The results show that: ① The optimal hyperparameter combination was obtained using the Bayesian optimization algorithm, and then the BO – XGBoost cyanobacteria classification model was established through obtaining. The training results performed well on the test and training sets, with an accuracy rate of up to 96.07%; ② The BO – XGBoost method was applied to the images used in the sample set. The comparison between the cyanobacteria identification results and the manual interpretation results shows that the two methods yielded roughly the same spatial distribution of cyanobacteria, with a lowest intersection over union (IoU) of 41.31%; ③ To evaluate the applicability of the BO – XGBoost method in other periods, images of other periods were selected for the information extraction of cyanobacteria. As a result, both BO – XGBoost and manual interpretation also yielded roughly the same spatial distribution of cyanobacteria, with a lowest IoU of 43.85%.

Keywords: Bayesian optimization; BO – XGBoost; multi – feature; cyanobacteria; Sentinel – 2

(责任编辑: 张 仙)