文章编号：0253-374X(2025)03-0450-12　　**DOI**：10.11908/j.issn.0253-374x.23264

基于Levy飞行和麻雀搜索算法优化集成学习模型的水质估算

李爱民[1]，康　轩[1]，袁　铮[1]，王海隆[2]，闫翔宇[1]，许有成[2]
(1. 郑州大学 地球科学与技术学院，河南 郑州 450001；2. 郑州大学 水利与交通学院，河南 郑州 450001)

摘要：由于水体的光学复杂性和不同水质参数之间的相互作用，利用集成学习方法估算水质参数具有优势；然而，在建模过程中如何合理选择超参数仍然是一个难题。麻雀搜索算法能够快速搜索集成学习模型的最优参数；而Levy飞行算法可以防止麻雀搜索算法(Sparrow Search Algorithm，SSA)陷入局部最优，并提高模型的准确性和效率。使用Levy飞行算法和麻雀搜索算法对随机森林(Random Forest，RF)、自适应回归(AdaBoost Regression，ABR)和类别提升回归(CatBoost Regression，CBR)3种集成学习模型进行了优化。以郑州东风渠和熊耳河为研究区，基于实测叶绿素a(chlorophyll－a，Chl－a)和总悬浮物(total suspended solids，TSM)数据，构建了LSSA－RF、LSSA－ABR和LSSA－CBR这3种估算模型。实验结果表明：模型经过优化后，各项指标均有不同程度的提高。其中表现最优的是LSSA－CBR模型；CBR模型是在梯度提升框架下进行的建模，对比RF和CBR模型具有更高维度的学习能力。在叶绿素a的估算中，LSSA－CBR估算模型的均方根误差为2.325 μg·L^{-1}，决定系数为0.896；在总悬浮物的估算中，LSSA－CBR模型的均方根误差为1.598 mg·L^{-1}，决定系数为0.882。最后，将精度较好的LSSA－CBR模型应用于卫星Planet影像中，以评估河流叶绿素a和总悬浮物的空间分布情况。研究结果可为环保部门快速了解城市河流水质分布及进行水质评价与管理提供参考。

关键词：叶绿素a；总悬浮物；集成学习模型；Levy飞行—麻雀搜索算法；城市河流
中图分类号：TP751.1；TP79　　**文献标志码**：A

Estimation of Water Quality Parameters Using an Ensemble Learning Model Optimized with Levy Flight and Sparrow Search Algorithms

LI Aimin[1]，*KANG Xuan*[1]，*YUAN Zheng*[1]，*WANG Hailong*[2]，*YAN Xiangyu*[1]，*XU Youcheng*[2]
(1. School of Geo-Science and Technology，Zhengzhou University，Zhengzhou 450001，China；2. School of Water Conservancy and Transportation，Zhengzhou University，Zhengzhou 450001，China)

Abstract：Due to the optical complexity of water bodies and the interactions among various water quality parameters，utilizing ensemble machine learning methods for estimating water quality parameters offers advantages. However，selecting hyperparameters in the modeling process remains challenging. The sparrow search algorithm (SSA) can rapidly search for optimal parameters of ensemble machine learning models，while the Levy flight algorithm prevents SSA from being trapped in local optima，thereby improving the accuracy and efficiency of the model. In this paper，the Levy flight algorithm and SSA were used to optimize three ensemble learning models：random forest (RF)，AdaBoost regression (ABR)，and CatBoost regression (CBR). Taking Zhengzhou Dongfeng Canal and Xiong'er River as the study area，estimation models (LSSA-RF，LSSA-ABR，and LSSA-CBR) were developed based on measured chlorophyll-a and total suspended solids concentrations. The experimental results show that after optimization，various indicators show improvements to varying degrees. Among them，the LSSA-CBR model exhibits the best performance. The CBR model，which is modeled under the gradient boosting framework，demonstrates higher learning capability compared to RF and ABR models. For the estimation of chlorophyll-a，the root mean square error (RMSE) of the LSSA-CBR estimation model is 2.325 μg·L^{-1}，and the coefficient of determination (R^2) is 0.896. For the estimation of total suspended solids，the

收稿日期：2023-08-09
基金项目：河南省自然科学基金面上项目(242300421372)；河南省高等学校重点科研项目(24B170010)
第一作者：李爱民，副教授，主要研究方向为遥感与地理信息技术。E-mail：aiminli@zzu.edu.cn
通信作者：康　轩，硕士生，主要研究方向为水环境遥感监测。E-mail：2465868312@qq.com

论文拓展介绍

RMSE of the LSSA-CBR model is 1.598 mg·L^{-1}, and R^2 is 0.882. Finally, the LSSA-CBR model, demonstrating strong accuracy, was applied to Planet images to evaluate the spatial distribution of chlorophyll-a and total suspended solids in rivers, providing a valuable reference for quickly understanding the distribution of urban river water quality and conducting water quality assessment and management.

Keywords: chlorophyll a; total suspended matter; integrated learning model; Levy flight-sparrow search algorithm; urban river

目前遥感估算水质参数总悬浮物(total suspended solids，TSM)和叶绿素a(chlorophyll-a，Chl-a)的方法很多，传统的统计回归模型主要是多元线性回归等线性方法，但对于成分及影响因素复杂的城市河流水体来说，其光学特征不像大洋水体那样主导因子单一，水质参数与影像数据之间的关系并不严格遵循线性统计规律。伴随着人工智能技术的突飞猛进，许多研究人员开始尝试使用机器学习方法寻找遥感数据与水质参数之间复杂的非线性关系，并使用各种机器学习算法模型实现水质参数的遥感估算[1]。Werther等[2]开发了一种基于Sentinel-3 OLCI和Sentinel-2 MSI数据的贝叶斯神经网络(BNN)，用于估算富营养化湖泊的叶绿素a浓度。机器学习模型能够在一定程度上拟合变量之间的非线性关系，但是模型的性能和稳定性受参数的影响较大，存在参数选取困难的问题，构建稳定可靠的估算模型仍是研究的难点。作为机器学习的主流算法，随机森林[3]、自适应回归[4]、类别提升回归等集成算法近年来逐渐被学者发掘并应用于水质遥感估算。集成学习是一种通过集成多个方法共同决策的机器学习方法，该方法通过集成多个不同模型的估算结果，采用特定规则将这些结果组合，产生更加稳健的估算结果，提高模型的泛化能力和精度[5]。陈点点等[6]采用带交叉验证的网格搜索法分别对CatBoost和随机森林2种机器学习模型进行超参数调优，确定模型最优参数配置，并对比不同模型估算精度，确定最优模型，以少量采样数据估算闽江下游悬浮物浓度，并分析其时空变化特征。Xu等[7]基于GF-6WFV图像和兴凯湖2020年至2021年的少量野外采样数据，研究了3种机器学习模型并集成机器学习算法，证明RF模型精度更高，绘制了2019至2021年兴凯湖Chl-a浓度的时空变化图。但是集成算法超参数的选择直接影响模型的精度和性能，很难通过手动调参找到最优的全局参数，且计算时间较长。一些学者利用遗传算法和粒子群优化算法等群智能算法来优化模型并取得了较好的效果。盛辉等[8]将模拟退火—粒子群算法(SA-PSO)引入到支持向量回归机的参数优化过程中，提出了一种改进SVR(SA-PSO-SVR)的内陆水体化学需氧量(COD)高光谱遥感估算方法。Guo等[9]基于高分二号(GF-2)遥感影像和现场实测悬浮物浓度，以海河一段为研究区，建立偏最小二乘(PLS)算法和粒子群优化(PSO)算法优化反向传播神经网络(BPNN)模型，即PLS-PSO-BPNN模型。麻雀搜索算法(Sparrow Search Algorithm，SSA)是由Xue等[10]借鉴麻雀的群体智慧、觅食和反捕食行为提出的一种新的群体优化算法，可有效缩短计算时间，加强模型的全局搜索能力。Levy飞行则可以避免麻雀搜索算法陷入局部最优，提升模型的精度和效率。

针对集成算法存在参数选取困难、计算时间长等问题，为构建精度高，稳健性好、计算效率高的估算模型，本文以郑州东风渠和熊耳河为研究区，利用Levy飞行对麻雀搜索算法进行改进，利用Levy飞行—麻雀搜索算法(LSSA)来优化随机森林(RF)、自适应回归(ABR)、类别提升回归(CBR)集成的叶绿素a和总悬浮物估算模型，并与传统模型进行精度对比，最后把精度最好的模型应用于Planet影像估算2条河流的叶绿素a和总悬浮物空间分布，旨在探讨利用机器学习模型估算水质中的超参数选取方法，为提高水质估算模型的精度提供参考。

1　研究区域与数据

1.1　研究区域与实测水质数据

以郑州市的东风渠和熊耳河为研究区，如图1所示。结合天气、卫星过境时间等实际情况，于2022年6月7日在东风渠和熊耳河采集水样，采集当天天气状况良好，晴朗无云。按均匀分布原则设置60个采样点，具体位置如图1所示。

采样流程按照《地表水和污水监测技术规范》(HJ/T 91—2002)确定。采样时利用采样器取水面下0.5 m深处水样，采集的水样当日立即送至具有检测资质的检测公司进行检测。叶绿素a的测定使用分光光度法(HJ 897—2017)，总悬浮物的测定使用重量法(GB 11901—1989)，数据如表1所示。

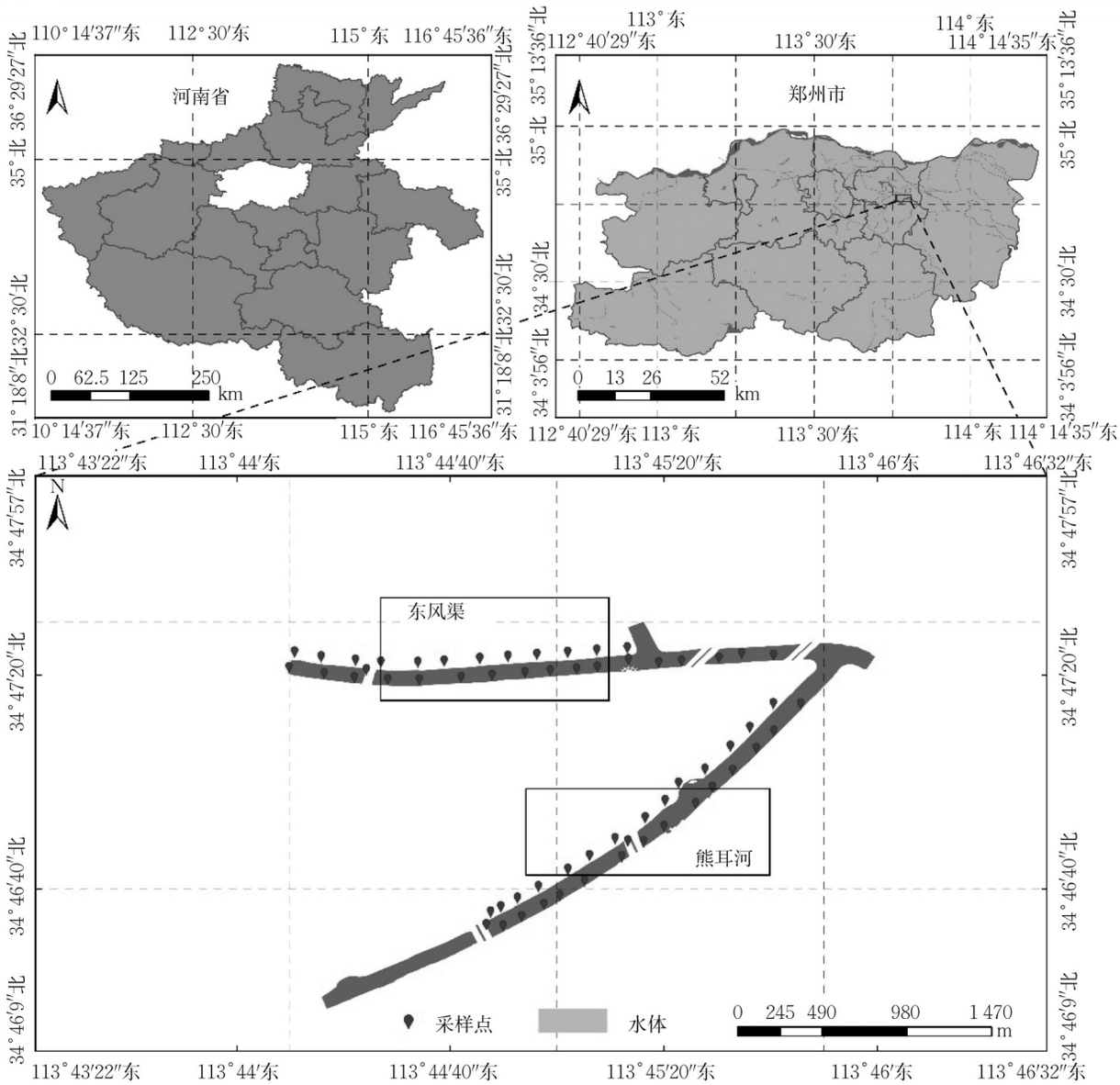

图1 研究区域位置

Fig. 1 Location of the study area

表1 水质数据统计

Tab. 1 Statistics of water quality

水质参数名称	总悬浮物质量浓度/($mg \cdot L^{-1}$)	叶绿素a质量浓度/($\mu g \cdot L^{-1}$)
检测方法	重量法	分光光度法
最小值	15	20
最大值	32	54
均值	21.52	35.38
标准差	4.145	7.774

1.2 遥感数据获取与预处理

使用的Planet卫星影像数据由北京国测星绘信息技术有限公司提供,影像日期为2022年6月7日,与水样采集时间一致,影像完整覆盖研究区。实验获取的影像已进行传感器校正、辐射校正、几何校正及镶嵌拼接处理,因此本文对影像数据的预处理主要有大气校正和裁剪,使用ENVI 5.3软件中的FLAASH工具对影像进行大气校正处理,获取遥感反射率。SuperDove的主要参数和波段信息如表2所示。

基于归一化差异水体指数(NDWI)的方法对熊耳河和东风渠河流进行水体信息的提取。用遥感影像的特定波段进行归一化差值处理,以凸显影像中的水体信息。水体具有正值、非水体具有零或负值,它们分别被增强和抑制。采用绿光波段与近红外波段的比值可以有效抑制植被信息,利用绿光波段和近红外波段之间的运算来构成NDWI,突出影像中的水体[11]。提取的东风渠和熊耳河水体效果图如图

表2　SuperDove波段信息

Tab. 2　Information of SuperDove band

波段	波段名称	波段范围/nm
b1	海岸蓝	431～452
b2	蓝	465～515
b3	绿Ⅰ	513～549
b4	绿Ⅱ	547～583
b5	黄	600～620
b6	红	650～680
b7	红边	697～713
b8	近红外	845～885

2所示。计算式为

$$N_{\mathrm{NDWI}}=\frac{R_{\mathrm{Green}}-R_{\mathrm{NIR}}}{R_{\mathrm{Green}}+R_{\mathrm{NIR}}} \tag{1}$$

式中：R_{Green}为绿光波段反射率；R_{NIR}为近红外波段反射率。

图2　NDWI法提取水体与Planet遥感影像叠加

Fig. 2　NDWI method of extracting water bodies overlaid with Planet remote sensing imagery

2　实验方法和结果

针对叶绿素a和总悬浮物浓度的遥感估算中使用传统机器学习模型存在参数选取困难的问题，基于Levy飞行—麻雀搜索算法(LSSA)对集成算法模型进行优化，构建LSSA-RF、LSSA-ABR和LSSA-CBR这3种模型。麻雀搜索算法在优化问题中的应用已经非常广泛，算法的局部搜索能力极强、收敛速度较快。但麻雀搜索算法的缺点也较为突出，如初始种群分布不均匀、全局搜索能力较弱且跳出局部最优的能力弱，因此该算法具有很大的改进空间。于是引入了Levy飞行策略来对麻雀搜索算法的突出缺点进行了改进。通过在麻雀搜索算法中引入Levy飞行策略可以改善初始种群分布，从而增强种群的多样性，避免过早收敛，增强算法跳出局部最优的能力。研究方法及路线如图3所示。

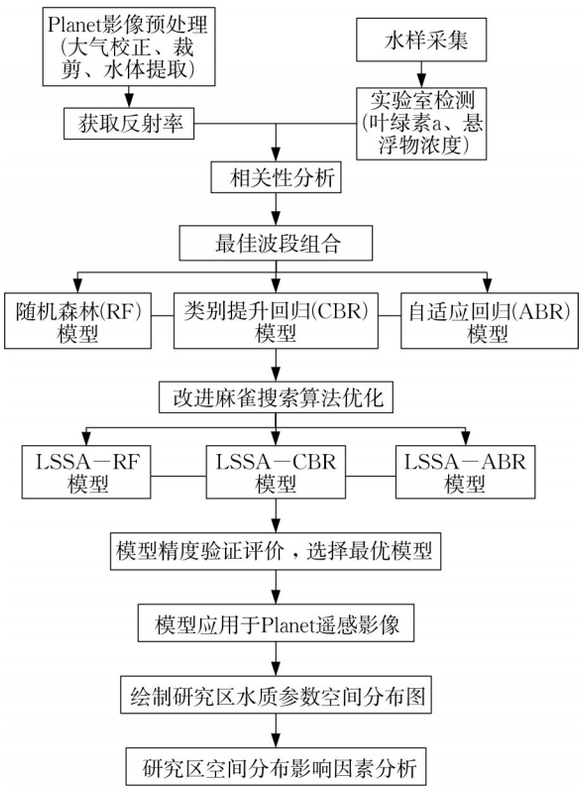

图3　研究方法及技术路线

Fig. 3　Research methodology and technical approach

2.1　敏感波段选择

在建模前先对实测水质参数数据与影像提取的反射率进行Pearson相关性分析，选择敏感波段，相关系数计算式为

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\times\sqrt{\sum(y_i-\bar{y})^2}} \tag{2}$$

式中：x、y为2组变量；x_i、y_i为变量组内第i个数值；$\bar{x}$、$\bar{y}$为2组变量的平均值。

通过计算发现单波段与水质参数浓度值的相关性较低，不适合直接建模。对各种波段组合比较，结果表明部分波段进行组合可以得到高于单波段的相关系数，最终选定参与总悬浮物(TSM)估算建模的波段(R)组合为：R_{b8}/R_{b7}、$(R_{b8}-R_{b7})/(R_{b7}+R_{b8})$、$(R_{b8}-R_{b4})/R_{b4}$、$R_{b3}+R_{b8}$、$R_{b6}+R_{b8}$；参与Chl-a估算建模的波段组合为：$R_{b2}+R_{b7}$、$R_{b6}+R_{b7}$、$R_{b4}\times R_{b7}$、$(R_{b1}+R_{b6})/(R_{b1}+R_{b4})$。各波段组合相关系数具体情况见表3。

2.2　精度评价指标

为了确定最适合于TSM和Chl-a(叶绿素a)估算的模型，使用常用的决定系数(Correlation of Determination，R^2)、均方根误差(Root Mean Square

表3 TSM和Chl-a与波段组合的相关系数

Tab. 3 Correlation coefficients of TSM and Chl-a with band combinations

TSM波段组合	相关系数	Chl-a波段组合	相关系数
R_{b8}/R_{b7}	0.725	$R_{b2}+R_{b7}$	0.704
$(R_{b8}-R_{b7})/(R_{b7}+R_{b8})$	0.821	$R_{b6}+R_{b7}$	0.715
$(R_{b8}-R_{b4})/R_{b4}$	0.791	$R_{b4}\times R_{b7}$	0.737
$R_{b3}+R_{b8}$	0.663	$(R_{b1}+R_{b6})/(R_{b1}+R_{b4})$	0.858
$R_{b6}+R_{b8}$	0.690		

Error，RMSE）和平均相对误差（Mean Relative Error，MRE）这3个指标来评估模型的精度[12]。其中，R^2用于衡量估算值与预测值之间的拟合程度，数值越大表示模型拟合度越高。RMSE对异常值具有高敏感性，能直观地反映估算值和实测值之间的偏差。MRE用于评价各模型估算值与实测值之间的相对偏差。

2.3 集成学习方法

随机森林（Random Forest，RF）是基于bagging框架建立，算法框架如图4所示，通过集合多个决策树来提高模型的预测准确性，模型数据挖掘能力较强，具备准确率高、稳健、参数优化便捷等优点[13-14]。随机是RF算法的一个关键特性，样本特征的随机选取可以有效降低各决策树的相关性，从而进一步提高模型的准确性和稳定性，并避免过度拟合问题。RF模型主要调节的超参数有max_features、min_samples_split、n_estimators。n_estimators为随机森林生成树的个数（即学习器的数量）；max_depth是树的最大深度，即最大复杂度，复杂度一般由高向低的方向调参；min_samples_split为划分内部节点时所需的最小样本数，低于该值的样本不会被划分。

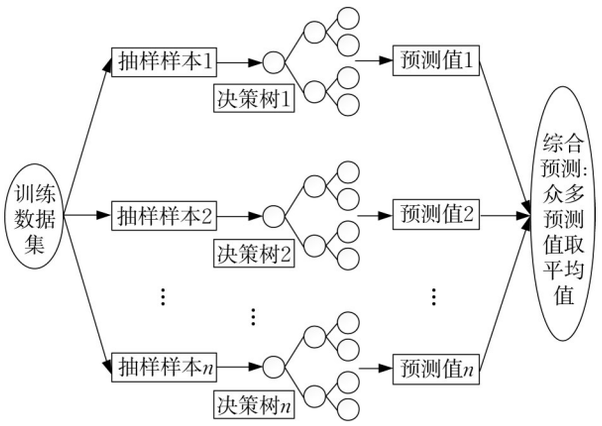

图4 随机森林模型框架

Fig. 4 Framework of random forest model

自适应回归（AdaBoost Regression，ABR）算法是集成学习中boosting类型典型的代表。通过改变回归错误率较大的样本权重来迭代训练一系列弱学习器，从而使下一个学习器更加关注上一轮学习中表现不佳的样本。最后，根据弱学习器的回归错误率对学习器进行加权，并以预测的采样点处的水质参数浓度加权平均值作为最终输出[15]。该算法的原理是从训练好的弱学习器中选出最佳弱学习器，然后通过调整样本权重和弱学习器权重，将最佳弱学习器联合成最终的强学习器。AdaBoost的优点是充分考虑了每个学习者的权重，参数少，在实际应用中不需要调整太多的参数[16]，算法框架如图5。ABR模型主要调节的超参数有n_estimators和learning_rate。n_estimators就是弱学习器的最大迭代次数，或者说最大的弱学习器的个数。learning_rate是权重缩减系数，决定权重的变化量。

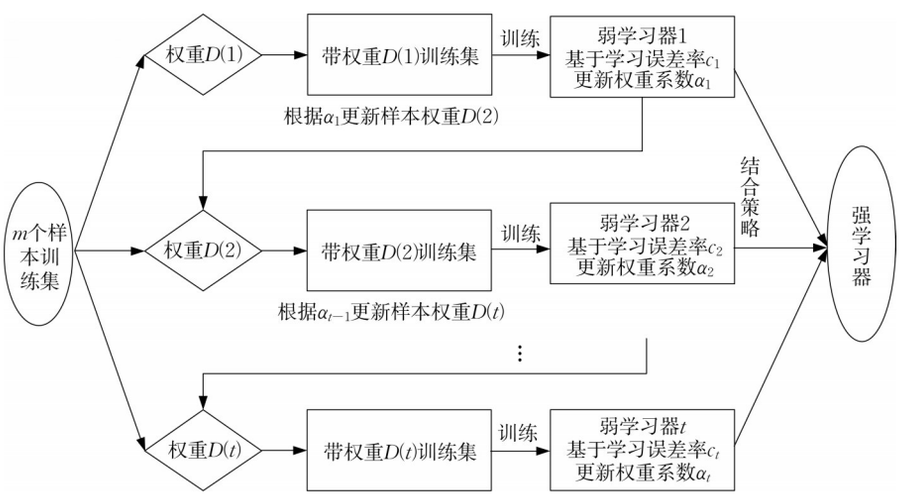

图5 ABR框架

Fig. 5 Framework of ABR

类别提升回归（CatBoost Regression，CBR）是基于Boosting框架的一种算法，由Yandex等[17]在2017年开发。CBR算法基于对称决策树，基学习器实现的参数较少，主要解决的是高效合理地处理类别特征[18]。CatBoost解决了传统Boosting框架存在的预测偏移和梯度偏差的问题，采用排序提升的方法对抗训练集中的噪声点，从而减少过拟合的发生，进而提高算法的准确性和泛化能力。CatBoost还使用了组合类别特征，可以建立特征之间的联系，极大地丰富了特征维度。

2.4　Levy飞行改进麻雀搜索算法优化模型

SSA中的初始种群是随机生成的，随机产生种群不能保证种群的多样性，种群的代表性和质量会受到影响，进而影响算法性能，反向学习策略可以很好地扩大搜索范围来保证种群的多样性，因此使用反向学习策略来初始化种群。反向学习初始化种群的主要过程：先随机生成多个麻雀组成一个种群，然后生成这些个体所对应的反向个体，通过将所有原始个体与反向个体进行比较，去除较差的麻雀个体，把较优的个体留下组成新一代的种群。其中判断“较差”的麻雀个体是通过适应度函数进行量化评估。

Levy飞行步长分布为重尾分布，运动特征为大部分情况下进行随机游走，在随机游走过程中偶发大步跨越，类似飞行[19]。进行寻优计算时，Levy飞行一方面可以在一定区域内小步搜索，另一方面也能在全局中进行大跨越搜索，可以有效保证搜索区域的整体性，避免了原始SSA得到的最优解可能是局部最优值的情况。在麻雀搜索算法模拟实验中，假设整个麻雀种群的个体发生危险且这些麻雀的初始位置是随机产生的，它们的位置更新如下：

$$X_{i,j}^{t+1}=\begin{cases} X_{\text{best}}^{t}+\beta\cdot\left|X_{i,j}^{t}-X_{\text{best}}^{t}\right|, f_i>f_{\text{g}} \\ X_{i,j}^{t}+K\cdot\left(\dfrac{\left|X_{i,j}^{t}-X_{\text{worst}}^{t}\right|}{(f_i-f_{\text{w}})+\varepsilon}\right), f_i=f_{\text{g}} \end{cases} \tag{3}$$

式中：β为步长控制参数，且是一个符合正态分布的随机数；X_{best}^{t}为第t次迭代中全局最优的位置；K为—1～1之间的随机数；f_{g}和f_{w}分别为当前全局最佳和最差适应度值；ε是一个极小常数，用于保证分母始终非零，避免在计算麻雀适应度值时出现除零错误。当$f_i>f_{\text{g}}$时，表明麻雀适应度值较差，在搜寻区域的边缘觅食。当$f_i=f_{\text{g}}$时，表示处于种群中间的麻雀发现危险，需要向其他麻雀靠近。

将Levy飞行策略与预警麻雀的位置更新相结合，对式(3)进行优化，用Levy飞行替代原始SSA预警麻雀随机产生与更新，以此来减小陷入局部最优的可能性，同时也能在搜索区域进行小步长的精细搜索，加强搜索能力。改进式为

$$X_{i,j}^{t+1}=\begin{cases} L(d)X_{\text{best}}^{t}+\beta\cdot\left|X_{i,j}^{t}-L(d)X_{\text{best}}^{t}\right|, f_i>f_{\text{g}} \\ X_{i,j}^{t}+K\cdot\left(\dfrac{\left|X_{i,j}^{t}-X_{\text{worst}}^{t}\right|}{(f_i-f_{\text{w}})+\varepsilon}\right), f_i=f_{\text{g}} \end{cases} \tag{4}$$

$$L(d)=0.01\cdot\frac{r_1\cdot\sigma}{|r_2|} \tag{5}$$

$$\sigma=\left\{\frac{\Gamma(1+\beta)\cdot\sin(\pi\beta/2)}{\Gamma\left(\dfrac{1+\beta}{2}\right)\beta\cdot 2^{(\beta-1)/2}}\right\}^{1/\beta} \tag{6}$$

式中：L为Levy飞行搜索函数；d为维度向量；β为步长控制参数；Γ为伽马函数；r_1和r_2为0～1之间的随机数。

使用选定的反射率波段组合作为输入变量、实测的水质参数浓度数据作为输出数据，把70%的数据作为训练数据、30%的数据作为验证数据。在Python软件的scikit learn开源机器学习库中分别构建RF、ABR和CBR模型。对RF模型调节的参数主要有n_estimators、min_samples_split和max_depth，对ABR模型调节的参数有n_estimators和learning_rate，对CBR模型调节的参数主要有n_estimators和learning_rate和depth，引入LSSA算法对参数进行调节。首先定义麻雀算法，设置麻雀算法种群数量、迭代次数等参数；然后定义一个适应度函数，将适应度函数的值标准化后进行比较，通过将适应度函数的值归一化到一定的范围内，可以更好地比较个体的相对性能，使用[0，1]的范围，用于判断种群个体的优劣。以随机森林为例，将n_estimators、min_samples_split和max_depth的取值作为输入，在设定这些参数值后，RF模型计算得到的R^2作为输出。交叉验证是机器学习中常用的模型构建与验证方法，有助于提高模型的泛化能力，并在一定程度上减少过拟合现象的发生。根据多次实验，将K设置为5，将训练数据分为5组，在5次迭代中，4组用于训练，1组用于测试数据集的模型评估。将训练数据随机分为5份数据，数据间不重复，从中挑选一个子集为测试集，剩余子集用于模型训练，随

后训练模型估算测试集并记录估算偏差,重复上述步骤5次,保证每一个子集都成为过测试集,计算5组数据的平均偏差作为模型精度的估计,来减少模型对数据的敏感性;接下来使用适应度函数和反向学习来初始化麻雀种群,再依次进行发现者位置更新、追随者位置更新、使用Levy飞行改进策略更新意识到危险的麻雀位置,随后进行种群循环。LSSA算法优化过程见图6。经过迭代计算得到输出结果最优的参数设置后,将参数值设置为随机森林模型的最终参数并进行计算,模型输入数据为选定的波段组合,输出为预测的水质参数浓度。优化后各模型的参数最优值如表4所示。

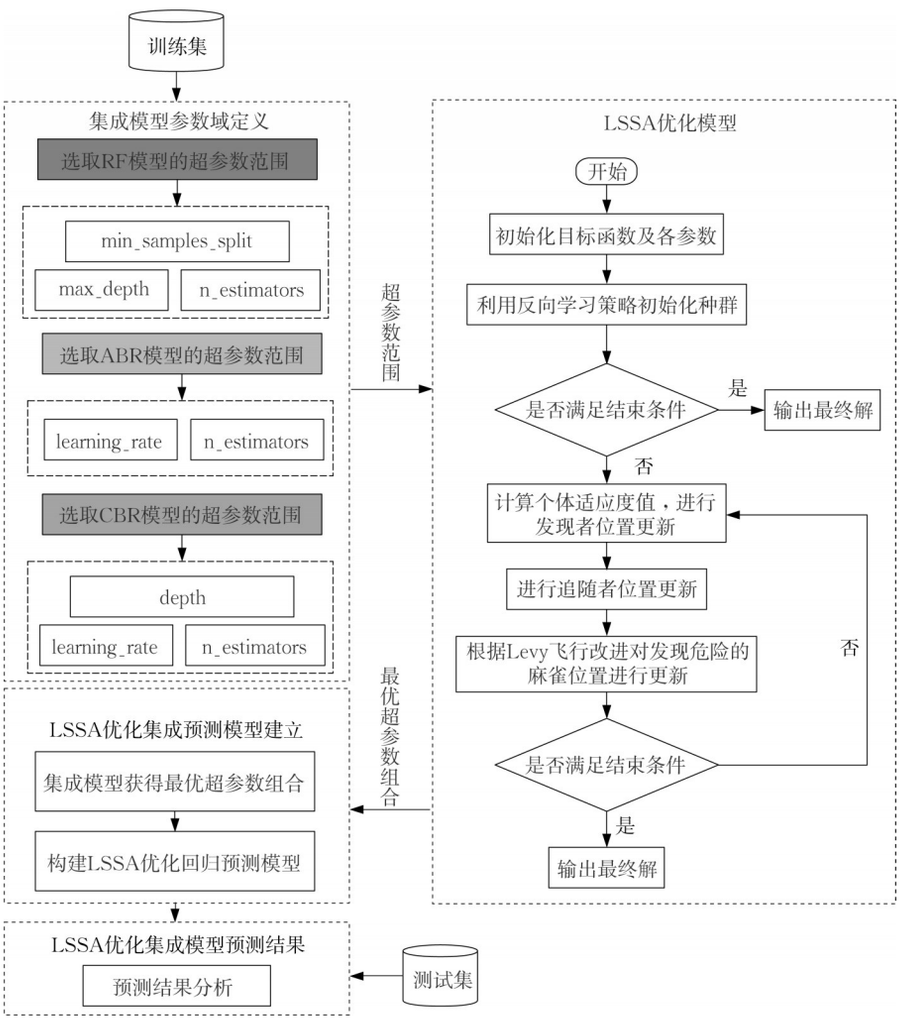

图6 LSSA优化流程

Fig. 6 Optimization process of LSSA

表4 优化后模型的参数

Tab. 4 Parameter profile of optimized model

模型	超参数	反演TSM时模型参数的最优值	反演Chl-a时模型参数的最优值
LSSA-RF	n_estimators	64	80
	min_samples_split	2	2
	max_depth	8	11
LSSA-ABR	n_estimators	90	97
	learning_rate	0.04	0.05
LSSA-CBR	n_estimators	140	214
	learning_rate	0.01	0.01
	depth	4	8

3　实验分析

3.1　模型精度分析

为了更清晰地对比模型的性能，将进行优化后的模型和优化前的模型估算同一水质参数的结果进行对比分析。Chl-a浓度和TSM浓度的估算模型预测结果分别如图7、图8所示。为了更直观地反映各模型的性能差异，将东风渠和熊耳河各模型测试集预测结果进行对比分析，结果如表5和表6所示。

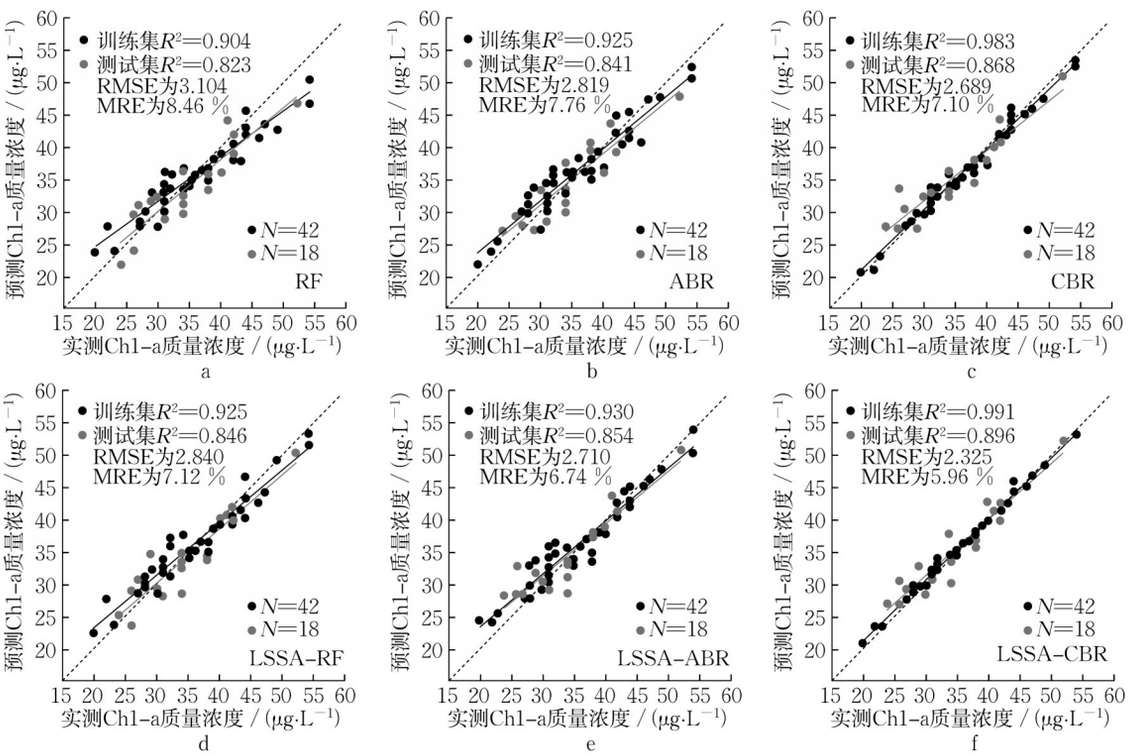

图7　预测Chl-a质量浓度与实测值散点图

Fig. 7　Scatter plot of estimated and measured Chl-a mass concentrations

从Chl-a和TSM浓度的模型预测结果可以看到：在Chl-a浓度的模型预测中，LSSA-CBR模型的RMSE为2.325 μg·L^{-1}，MRE为5.96%，R^2为0.896。在TSM浓度的模型预测中，LSSA-CBR模型的测试集RMSE为1.598 mg·L^{-1}，MRE为6.88%，R^2为0.882。优化后的模型中，LSSA-CBR的模型精度最高，对数据的拟合能力最好。相对于RF和ABR模型，CBR模型是在梯度提升框架下进行建模，具有更高维度的学习能力。使用麻雀搜索算法和Levy飞行算法全局优化后，能够更好地进行参数搜索与优化，找到更优的模型参数组合。3种模型优化后，精度均有不同程度的提升，这也说明了LSSA算法的优越性，可以用于模型的参数优化，提升模型性能。综合来看，LSSA-CBR模型更优，更适用于Chl-a与TSM浓度的反演。

3.2　水质参数估算结果分析

通过比较基于3种模型的训练集和测试集的估算值与实测值的误差，发现LSSA-CBR模型的精度最高，拟合效果最好。将优化得到的LSSA-CBR模型应用于Planet影像，估算研究区总悬浮物浓度空间分布如图9所示，估算叶绿素a浓度空间分布如图10所示，颜色越深表示浓度越大。2种参数估算结果与实测值进行比较，结果如图11所示。

由图9和10可以看出，在东风渠中，西部叶绿素a浓度低于东部叶绿素a浓度，在a处与北面龙湖的交汇口浓度最高，呈现东高西低的趋势；熊耳河除西南部叶绿素a浓度较高外，其他区域浓度相对较低；总悬浮物浓度空间分布与叶绿素a浓度分布情况、区域特征相似。a处为东风渠与北边的龙湖相接的三岔口交汇区，在交汇口北部约100 m处设有橡皮闸。橡皮闸处于关闭状态将河流截断，导致交叉口北部出现一段死水，水体流动性下降，物质沉积，这可能是导致a处水质浓度较高的原因之一。熊耳河中的总悬浮物浓度从中部的b处向东北部有降低的趋

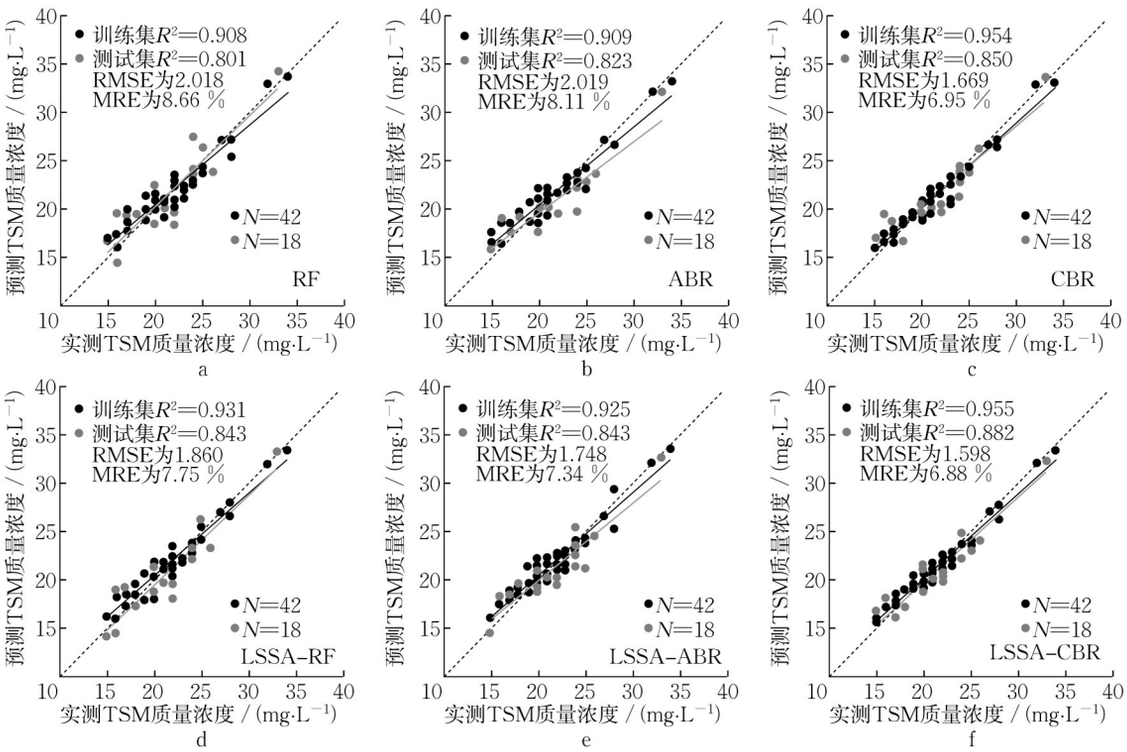

图8 预测TSM质量浓度与实测值散点图

Fig. 8 Scatter plot of estimated and measured TSM mass concentrations

表5 东风渠和熊耳河Chl-a反演精度

Tab. 5 Retrieval accuracy of Chl-a in Dongfengqu and Xiong'er River

模型	R^2	RMSE/(μg·L^{-1})	MRE/%
RF	0.823	3.104	8.46
ABR	0.841	2.819	7.76
CBR	0.868	2.689	7.10
LSSA—RF	0.846	2.840	7.12
LSSA—ABR	0.854	2.710	6.74
LSSA—CBR	0.896	2.325	5.96

表6 东风渠和熊耳河TSM反演精度

Tab. 6 Retrieval accuracy of TSM in Dongfengqu and Xiong'er River

模型	R^2	RMSE/(mg·L^{-1})	MRE/%
RF	0.801	2.018	8.66
ABR	0.823	2.019	8.11
CBR	0.850	1.669	6.95
LSSA—RF	0.843	1.860	7.75
LSSA—ABR	0.843	1.748	7.34
LSSA—CBR	0.882	1.598	6.88

势,现场调查资料记录该处有一弧形水域,水面面积增大,且沿岸设有多片分流小河湾。水面开阔且有小分流可能是此处总悬浮物浓度相对较低的原因。由图11可知,LSSA-CBR模型估算结果和实测数据拟合较好,估算的Chl-a浓度平均值为35.61 μg·L^{-1},标准差为7.356 μg·L^{-1},变异系数为0.206;估算TSM浓度平均值为21.34 mg·L^{-1},标准差为3.746 mg·L^{-1},变异系数为0.175,可知LSSA-CBR模型估算结果接近实际采样点的统计值。

4 讨论

现有的基于机器学习方法估算Chl-a和TSM的研究主要使用单一模型,而机器学习方法的解空间维度通常较高,各方法具有较强的拟合能力,却由于方法各自的缺陷容易陷入局部最优解(即"过拟

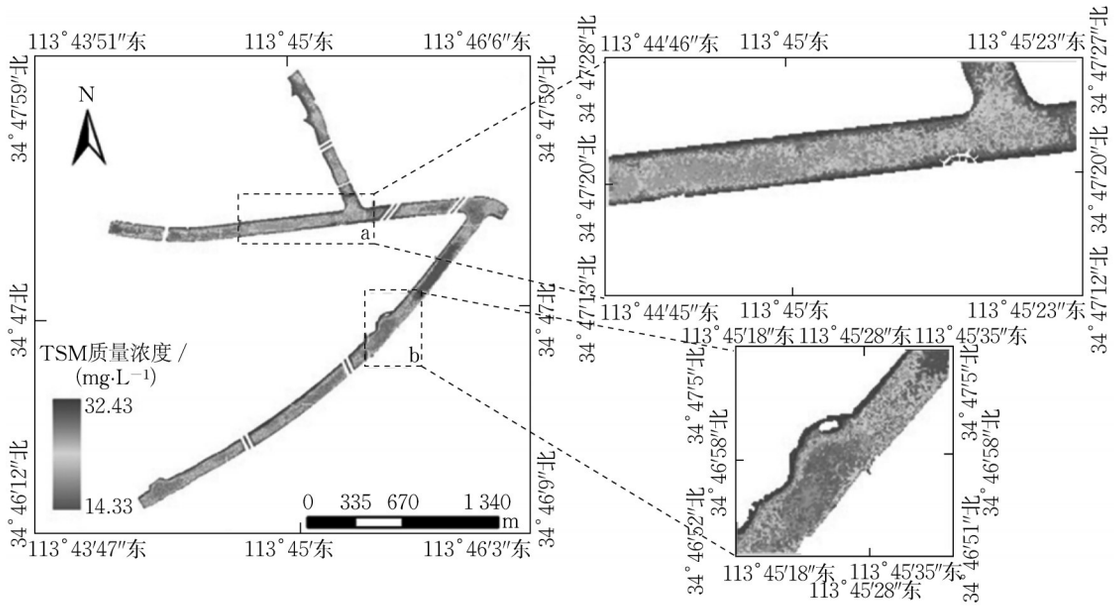

图9　LSSA-CBR模型估算总悬浮物浓度空间分布

Fig. 9　Spatial distribution of TSM concentrations from LSSA-CBR model inversions

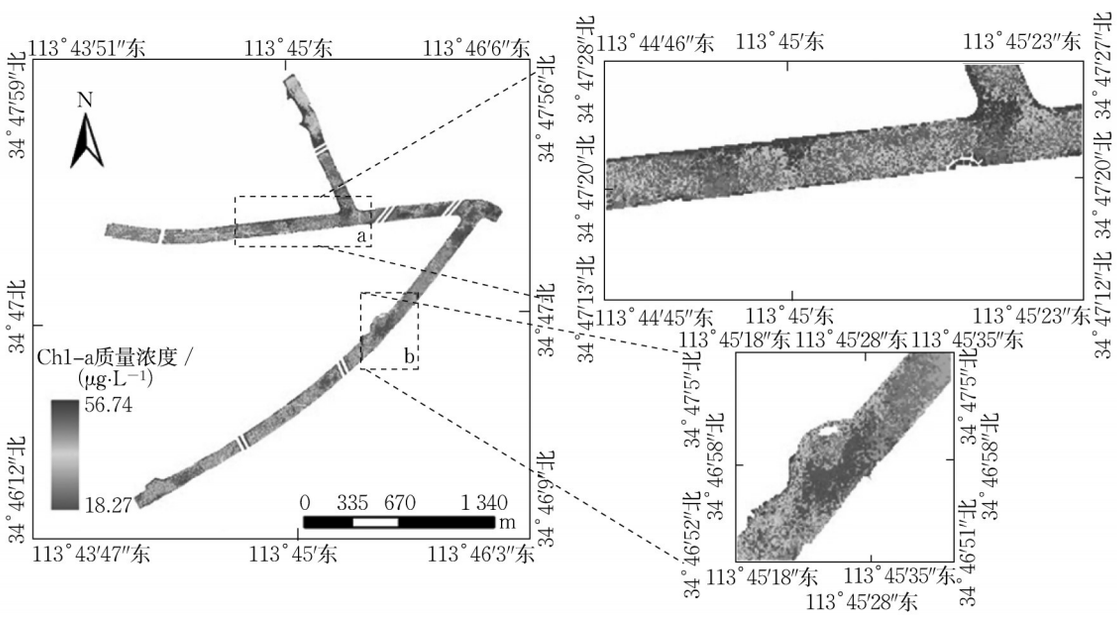

图10　LSSA-CBR模型估算叶绿素a浓度空间分布

Fig. 10　Spatial distribution of Chl-a concentrations from LSSA-CBR model inversions

合”)[21]。集成学习以多个单一机器学习模型作为基模型，采用不同的策略将各个基模型进行组合以实现基模型方法之间的优势互补，最大程度地发挥机器学习方法的潜力[22]，从而提升模型在Chl-a和TSM估算方面的精度和泛化性。麻雀搜索算法在优化问题中的应用已经非常广泛，算法的局部搜索能力极强，收敛速度较快，但麻雀搜索算法的缺点也较为突出，Levy飞行则可以避免麻雀搜索算法陷入局部最优。在调整参数过程中对训练集和测试集处理，使训练集和测试集的R^2分布相近，结合交叉验证通过在不同的数据子集上进行多次训练和测试，更好地评估了模型的泛化性能和过拟合问题[23]。

对比不同反演模型精度可知，LSSA-CBR能够更好地模拟东风渠和熊耳河水体TSM和Chl-a浓度与水体表面遥感反射率的非线性关系。在3种集成算法中，CBR模型反演精度最高，该算法具有使用简单、调节参数较少、准确率极高的特点，最大的特点是可以高效处理类别型特征。除此之外，算法还对

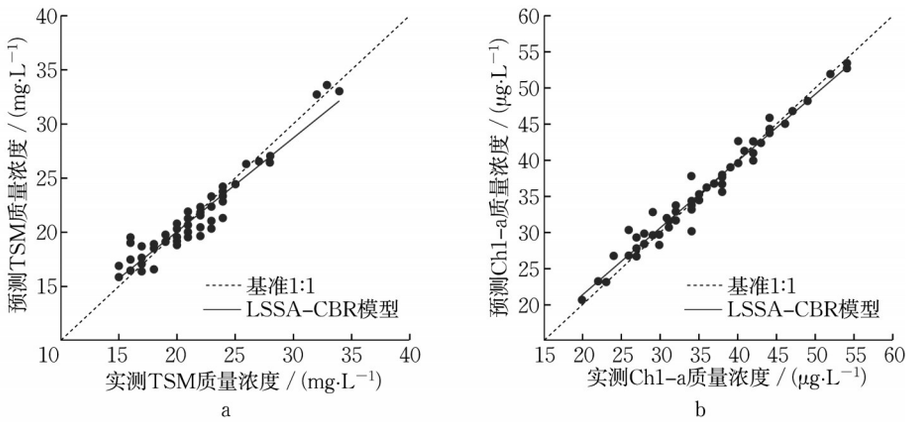

图11 LSSA-CBR模型估算结果散点图

Fig. 11 Scatter of LSSA-CBR model inversion results

GBDT框架的机器学习算法进行了优化,解决了机器学习算法中常见的梯度偏差和预测偏移问题,大幅降低了模型过拟合的发生,提高了算法的泛化能力。使用LSSA飞行算法优化后,精度得到了提升。相比LSSA-RF和LSSA-ABR方法,LSSA-CBR算法具有较强的适应性和抗噪能力,通过学习训练集的特征不断校正、更新样本权重和控制误差来提高TSM和Chl-a浓度遥感反演模型的精度和泛化性能。

5 结论

针对集成算法模型存在参数选取困难、计算时间长等问题,基于水质实测数据和卫星影像数据建立了LSSA优化集成算法模型,通过Planet数据估算了郑州东风渠和熊耳河的叶绿素a及总悬浮物的分布情况,得到以下结论。

(1)引入Levy飞行改进的麻雀搜索算法(LSSA)对RF、ABR、CBR这3个集成算法模型进行优化,构建了LSSA-RF、LSSA-ABR、LSSA-CBR模型。结果显示模型优化后各项指标均有不同程度的提升,其中,LSSA-CBR模型的表现最好,在叶绿素a的估算中,LSSA-CBR模型较优化前的R^2提高了0.028,RMSE降低了0.364 μg·L^{-1}。LSSA-CBR估算模型的RMSE为2.325 μg·L^{-1},R^2为0.896。在总悬浮物的估算中,LSSA-CBR模型较优化前的R^2提高了0.032,RMSE降低了0.071 mg·L^{-1}。LSSA-CBR模型的RMSE为1.598 mg·L^{-1},R^2为0.882。

(2)由估算结果得出,在东风渠中,西部叶绿素a浓度低于东部叶绿素a浓度,在a处与北面龙湖的交汇口浓度最高,呈现东高西低的趋势;熊耳河除西南部叶绿素a浓度较高外,其他区域浓度相对较低;总悬浮物浓度空间分布与叶绿素a浓度分布情况、区域特征相似;Chl-a浓度平均值为35.61 μg·L^{-1},标准差为7.356 μg·L^{-1},变异系数为0.206;估算的TSM浓度平均值为21.34 mg·L^{-1},标准差为3.746 mg·L^{-1},变异系数为0.175,LSSA-CBR模型估算结果接近实际采样点的统计值。

总的来看,Levy飞行改进的麻雀搜索算法优化集成学习模型提升了遥感TSM和Chl-a估算精度和泛化性,在遥感地表监测和信息提取方面表现出很大的潜力。然而,受限于天气和数据采集成本等主客观条件,所采集的实测水质数据量相对较小,所建立的估算模型仅适用于郑州部分水体。后续的研究将重点考虑获取更长时间尺度和更大空间范围的数据,以提升估算模型的适用性。同时,考虑将河流流速、深度和气象因子等信息加入估算模型,以削弱河流特性对估算的影响,进一步提升估算精度。

作者贡献声明:

李爱民:实验方案设计。

康 轩:实验操作。

袁 铮:论文写作和修改。

王海隆:论文写作和修改。

闫翔宇:论文写作和修改。

许有成:论文写作和修改。

参考文献:

[1] KIM Y W, KIM T, SHIN J, *et al*. Validity evaluation of a machine-learning model for chlorophyll a retrieval using Sentinel-2 from inland and coastal waters [J]. Ecological Indicators, 2022,137:108737.

[2] WERTHER M, ODERMATT D, SIMIS S, *et al*. A

Bayesian approach for remote sensing of chlorophyll-a and associated retrieval uncertainty in oligotrophic and mesotrophic lakes[J]. Remote Sensing of Environment, 2022,283:113295.

[3] 李爱民,王海隆,许有成.优化随机森林算法的城市湖泊DOC质量浓度遥感估算[J].郑州大学学报(工学版),2022,43(6):90.
LI Aimin, WANG Hailong, XU Youcheng, *et al*. Remote sensing retrieval of urban lake DOC concentration based on optimized random forest algorithm [J]. Journal of Zhengzhou University(Engineering Science), 2022,43(6):90.

[4] CHEN B, MU X, CHEN P, *et al*. Machine learning-based inversion of water quality parameters intypical reach of the urban river by UAV multispectral data [J]. Ecological Indicators,2021,133:108434.

[5] 嵇晓燕,杨凯,陈亚男,等.基于ARIMA和Prophet的水质预测集成学习模型[J].水资源保护,2022,38(6):111.
JI Xiaoyan, YANG Kai, CHEN Yanan, *et al*. An ensemble learning model for water quality forecast based on ARIMA and Prophet[J]. Water Resources Protection. 2022, 38(6): 111.

[6] 陈点点,陈芸芝,冯险峰,等.基于超参数优化CatBoost算法的河流悬浮物浓度遥感估算[J].地球信息科学学报,2022,24(4):780.
CHEN Diandian, CHEN Yunzhi, FENG Xianfeng, *et al*. Retrieving suspended matter concentration in rivers based on hyperparameter optimized catBoost algorithm [J]. Journal of Geo-information Science, 2022, 24(4): 780.

[7] XU S, LI S, TAO Z, *et al*. Remote sensing of Chlorophyll-a in Xinkai lake using machine learning and GF-6 WFV images [J]. Remote Sensing. 2022, 14(20): 5136.

[8] 盛辉,池海旭,许明明,等.改进SVR的内陆水体COD高光谱遥感估算[J].光谱学与光谱分析,2021,41(11):3565.
SHENG Hui, CHI Haixu, XU Mingming, *et al*. Inland water chemical oxygen demand estimation based on improved SVR for hyperspectral data [J]. Spectroscopy and Spectral Analysis, 2021,41(11):3565.

[9] GUO Q, WU H, JIN H, *et al*. Remote sensing inversion of suspended matter concentration using a neural network model optimized by the partial least squares and particle swarm optimization algorithms[J]. Sustainability 2022, 14: 2221.

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

[11] 王秋燕,陈仁喜,徐佳,等.环境一号卫星影像中水体信息提取方法研究[J].科学技术与工程, 2012, 12(13): 3051.
WANG Qiuyan, CHEN Renxi, XU Jia, *et al*. Research on methods for extracting water body information from HJ—1A/B data[J]. Science Technology and Engineering. 2012, 12(13): 3051.

[12] 李爱民,范猛,秦光铎,等.卷积神经网络模型的遥感估算水质参数COD[J].光谱学与光谱分析,2023,43(2):651.
LI Aimin, FAN Meng, QIN Guangduo, *et al*. Remote sensing inversion of water quality parameter COD of convolutional neural network model [J]. Spectroscopy and spectral analysis. 2023, 43(2): 651.

[13] 杭鑫,曹云,杭蓉蓉,等.基于随机森林算法与高分观测的太湖叶绿素a浓度估算模型[J].气象,2021,47(12):1525.
HANG Xin, CAO Yun, HANG Rongrong, *et al*. Estimation model of Chlorophyll-a concentration in Taihu lake based on random forest algorithm and Gaofen observations [J]. Meteorological Monthly,2021,47(12):1525.

[14] 方馨蕊,温兆飞,陈吉龙,等.随机森林回归模型的悬浮泥沙浓度遥感估算[J].遥感学报,2019,23(4):756.
FANG Xinrui, WEN Zhaofei, CHEN Jilong, *et al*. Remote sensing estimation of suspended sediment concentration based on Random Forest Regression Model [J]. National Remote Sensing Bulletin, 2019,23(4):756.

[15] LIN N, JIANG R Z, LI G J, *et al*. Estimating the heavy metal contents in farmland soil from hyperspectral images based on Stacked AdaBoost ensemble learning [J]. Ecological Indicators, 2022, 143. DOI: doi. org/10.1016/j. ecolind.2022.109330.

[16] BENTEJAC C, CSORGO A, MARTINEZ-MUNOZ G. A comparative analysis of gradient boosting algorithms [J]. Artificial Intelligence Review, 2021,54(3):1937.

[17] PROKHORENKOVA L, GUSEV G, VOROBEV A, *et al*. CatBoost: unbiased boosting with categorical features [M]. Dolgoprudny:[S.n.], 2018.

[18] LI H M, ZHANG G L, ZHONG Q C, *et al*. Prediction of urban forest aboveground carbon using machine learning based on Landsat 8 and Sentinel-2: A case study of Shanghai, China [J]. Remote Sensing, 2023,15(1).

[19] LIU Y H, CAO B Y. A novel ant colony optimization algorithm with Levy flight[J]. IEEE Access, 2020,8:67205.

[20] 张少卿,雷莉萍,宋豪,等.一种基于大气CO_2浓度时空特征的碳排放分区估算方法[J].中国环境科学,2023,43(10):5604.
ZHANG Shaoqing, LEI Liping, SONG Hao, *et al*. A neural network partitioning method for carbon emission estimation based on spatial-temporal clustering of atmospheric CO2 concentration [J]. China Environmental Science, 2023, 43(10): 5604.

[21] 余成,唐毅,潘杨,等.基于无人机遥感和集成学习的苏州市河流悬浮物浓度估算[J].中国环境科学,2023,43(10):5235.
YU Cheng, TANG Yi, PAN Yang, *et al*. Inversion of suspended sediment concentration in rivers of Suzhou based on UAV remote sensing and ensemble learning [J]. China Environmental Science, 2023, 43(10): 5235.

[22] ZHOU Z H. Ensemble methods: foundations and algorithms [M]. Cambridge: CRC press, 2012.

[23] 方韬.基于神经网络的近地面臭氧估算和预测研究[D].上海:上海师范大学,2020.
FANG Tao. Study on estimation and prediction of near-surface ozone based on neural network [D]. Shanghai Normal University,2020.