DOI: 10.20103/j.stxb.202403150536

王可月,王轶夫,陈馨,郑峻鹏,李杰,孙玉军.基于集成学习算法和 Optuna 调优的江西省森林碳储量遥感估测.生态学报,2025,45(2):685-700.
Wang K Y, Wang Y F, Chen X, Zheng J P, Li J, Sun Y J.Remote sensing estimation of forest carbon storage in Jiangxi Province based on ensemble learning algorithm and Optuna tuning.Acta Ecologica Sinica,2025,45(2):685-700.

基于集成学习算法和 Optuna 调优的江西省森林碳储量遥感估测

王可月[1],王轶夫[1,*],陈 馨[1],郑峻鹏[2],李 杰[3],孙玉军[1]

1 北京林业大学森林资源和环境管理国家林业和草原重点实验室, 北京 100083

2 北京市十三陵林场管理处, 北京 102200

3 北京市园林绿化大数据中心, 北京 101118

摘要:了解森林碳储量对于完整、准确地量化碳排放及气候变化背景下的环境监测至关重要,借助遥感数据源是估算区域尺度碳储量的有效方法。以江西省为研究区,基于第七次国家森林资源连续清查样地数据与 Landsat-5 TM 遥感数据,通过 GEE 平台对影像进行处理,将递归特征消除(RFE)、Boruta 两种特征选择方法与支持向量机(SVR),包括随机森林(RF)、极端梯度提升(XGBoost)和堆叠集成(Stacking)在内的三种集成学习算法相结合,分析不同模型的估测精度。此外,运用 Optuna 超参数优化框架来确定各模型的超参数。根据最优估测模型来反演江西省森林碳储量并绘制空间分布图,选用地理探测器对碳储量的空间分布格局进行驱动力分析。结果表明:(1)根据特征重要性排名,RFE 筛选出 30 个变量,Boruta 筛选出 11 个变量,合适的特征子集与回归算法相结合能显著提升估测的准确性。(2)基于 Optuna 对各模型的超参数进行迭代调优,发现不同特征子集与机器学习算法相结合,超参数取值和重要性在模型中差异较大。其中 RFE 筛选的最优特征子集与 Stacking 算法结合进行回归拟合时获得了最好的估测效果(R^2 = 0.527,RMSE = 15.85Mg/hm^2,MAE = 12.31Mg/hm^2),该模型有效利用训练数据,结合多种算法的优点以减少偏差,显著改善森林碳密度高值低估和低值高估的问题。(3)最优估测模型反演得到江西省 2006 年的森林碳密度平均值为 33.356Mg/hm^2(2.585—88.943Mg/hm^2),森林碳储量总量为 321.507Tg。(4)自然环境因子中海拔和坡度是影响碳储量空间分布格局的主要驱动因子;所有因子在交互作用下呈非线性增强和双因子增强,其空间分布格局是自然因素和人为因素协同作用的结果。

关键词:森林碳储量遥感估测;集成学习算法;Optuna 超参数调优;堆叠集成算法;碳密度;地理探测器

Remote sensing estimation of forest carbon storage in Jiangxi Province based on ensemble learning algorithm and Optuna tuning

WANG Keyue[1], WANG Yifu[1,*], CHEN Xin[1], ZHENG Junpeng[2], LI Jie[3], SUN Yujun[1]

1 *State Forest and Grassland Administration Key Laboratory of Resources and Environmental Management*, *Beijing Forest University*, *Beijing* 100083, *China*

2 *Beijing Ming Tombs Forest Center*, *Beijing* 102200, *China*

3 *Beijing Landscaping Big Data Center*, *Beijing* 101118, *China*

Abstract: Understanding forest carbon storage has been crucial for the complete and accurate quantification of carbon emissions and environmental monitoring in the context of climate change. The use of remote sensing data sources has proven to be an effective method for estimating carbon reserves at the regional scale. Taked Jiangxi Province as the research area, using the seventh National Forest Continuous Inventory (NFCI) data along with Landsat-5 TM remote sensing data. Image

基金项目:江西省林业局科技创新专项([202133]);中央高校基本科研业务费专项资金(BFUKF202404,PTYX202407)

收稿日期:2024-03-15; **网络出版日期**:2024-09-23

* 通讯作者 Corresponding author.E-mail: wyfbing@163.com

processing steps were performed on the Google Earth Engine (GEE) platform. Forest carbon storage was estimated by employing two feature selection methods(recursive feature elimination (RFE) and Boruta), combined with support vector regression (SVR) and three ensemble learning algorithms, including random forest (RF), extreme gradient boosting (XGBoost), and stacking integration (Stacking), to comparatively analyze the estimation accuracies of different models in detail. In addition, the Optuna hyperparameter optimization framework was used to determine the hyperparameters of each model. Inverted the forest carbon storage in Jiangxi Province based on the optimal estimation model and drew a spatial distribution map. The driving forces of carbon stocks' spatial distribution patterns were analyzed using geographic detectors. The results show that: (1) According to the feature importance ranking, RFE screened out 30 variables and Boruta screened out 11 variables. The combination of appropriate feature subsets and regression algorithms can significantly improve the accuracy of estimation. (2) Optuna-based iteratively adjusted the hyperparameters of each model. It was found that when different subsets of features were combined with machine learning algorithms, there was a significant disparity in the importance and values of hyperparameters within the model. The optimal feature subset was screened using RFE, which achieved the best predictive performance when utilized in regression simulations combined with Stacking models (R^2 = 0.527, RMSE = 15.85Mg/hm^2, MAE = 12.31Mg/hm^2). The model effectively utilized the training data and combined the advantages of multiple algorithms to reduce bias, which significantly improved the problems of underestimation of high carbon density values and overestimation of low carbon density values. (3) The optimal estimation model was inverted to obtain the average forest carbon density in Jiangxi Province in 2006 was 33.356Mg/hm^2(2.585—88.943Mg/hm^2), and the total forest carbon stock was 321.507Tg. (4) Among the natural environment factors, elevation and slope were the main driving factors influencing the spatial distribution pattern of carbon stocks. All factors showed nonlinear enhancement and two-factor enhancement under interaction. The spatial distribution pattern of carbon storage was the result of the synergistic effect of natural and anthropogenic factors.

Key Words: forest carbon storage remote sensing estimation; ensemble learning algorithm; Optuna hyperparameter tuning; Stacking; carbon density; Geodectetor

森林是陆地生态系统中最重要的“碳库”,森林植被碳储量占陆地植被碳储量的 77%,森林土壤碳储量约占全球土壤碳储量的 39%,其作为碳源和碳汇,在应对全球气候变化中发挥着至关重要的作用[1]。森林碳储量是森林生态系统最基本的定量特征,能反映森林物质循环、能量流动及植物与环境之间的复杂关系[2—3]。由 150 多个国家于 1992 年签署的《联合国气候变化框架公约》,要求缔约方定期提交有关温室气体排放的报告;《巴黎协定》中特别强调了通过森林可持续管理来增加碳汇,自此国际组织与各国林业部门更加重视森林生态系统的固碳能力以及人类活动导致的碳储量变化,这是量化森林在气候变化中的作用和制定可持续发展政策的必要步骤[4—5]。如何准确快速地估测大尺度森林碳储量以及探究其空间分布格局的驱动力对落实“双碳”目标具有重要意义。

森林碳储量的估测主要通过传统的实地测量和基于遥感数据源反演。在林分尺度上,利用实地测量估测的碳储量更为准确,但在大尺度区域内则较为困难,成本高且耗时费力,对生态环境破坏性大[6]。相比较而言,遥感数据广泛的覆盖性和优越的重复性,可以有效监测区域尺度森林碳储量及其时空变化格局。国内外学者借助时序性的遥感数据对不同尺度碳储量的估测展开了大量的研究[7—8]。研究发现,合成孔径雷达(Synthetic Aperture Radar,SAR)和机载激光雷达(Light Detection and Ranging,LiDAR)对森林冠层具有一定的穿透能力,可以获取植被的垂直结构,克服光学传感器易受云雾饱和的影响,但 SAR 的信号容易受到复杂地形和冠层结构的影响[9—10],LiDAR 其高昂成本和时空上重复获取的局限性阻碍了大尺度碳储量的估测。光学遥感数据虽然对茂密森林的穿透能力不足,估测精度较低,但因其长时间序列、空间覆盖范围广等特点,仍然是大区域尺度估测森林碳储量的主要数据源之一。在中等分辨率的光学遥感影像中,Landsat 系列卫星提

供超过 40 年的对地观测数据[11],具有多波段的光谱信息和广泛的时空覆盖范围,且开放数据可以免费获取和使用。目前,大多数估算模型受地区、遥感数据源等限制,模型可移植性差,因此对于森林碳储量的估测仍需进行深层次的优化[12]。

中等分辨率影像强调变量的选择和建模方法的稳健性,特征选择是提高模型性能的关键,适宜的方法可以最大限度地减少数据负载,解决非正态、共线性等问题,并提高预测的准确性[13]。遥感数据可以获取多光谱反射率、植被指数、纹理特征、地形特征等潜在变量,例如,Luo 等[14]基于 Landsat-8 OLI 影像数据,使用递归特征消除(Recursive Feature Elimination,RFE)、基于随机森林的特征选择(Variable Selection Using Random Forests,VSURF)、LASSO(Least Absolute Shrinkage and Selection Operator)三种特征选择方法,发现 RFE 保留了特征变量丰富的信息,对提升森林生物量估测精度有显著作用;王平[15]分别使用 Boruta 改进的支持向量机(Support Vector Regression,SVR)、LASSO、随机森林(Random Forest,RF),结果显示改进的 Boruta-SVR 估测准确性更高。然而,没有研究表明某种方法能筛选出对所有模型都有良好优化效果的变量组合,因此,需要进一步根据实际应用确定最佳输入特征集、输入特征数量和最佳拟合效果间的平衡关系。

森林碳储量遥感估测模型主要包括参数回归模型和非参数机器学习模型。参数回归模型在统计分析基础上,以多元线性回归和主成分回归等方法,构建样地实测数据与遥感变量之间的回归关系来估测碳储量。然而,参数回归模型无法完全捕捉遥感数据中变量与碳储量之间的复杂关系,此外,变量间可能存在的特征共线性导致估测精度较低。非参数模型不需要服从特定分布的样本,可以忽略变量之间的共线性,特征变量的利用率和模型的准确性高。作为非参数模型的代表,神经网络[16]、RF[17]、SVR[18]等机器学习模型被广泛用于碳储量估测,但训练过程中预测值和实测值的一致性仍然较差,对于碳密度高值点和低值点的预测能力较低。集成学习是机器学习的一个分支,通过不同的策略结合多个弱学习器来提高模型的整体性能,解决回归和分类问题[19]。集成学习通常指 Bagging、Boosting 和 Stacking 技术,以在弱模型中引入高可变性[20]。Bagging 和 Boosting 的弱模型为单一建模算法,Bagging 中弱学习器存在强依赖关系,为“同质并行集成”,RF 就是最经典的 Bagging 算法之一,通过从训练集中生成随机样本并将单一学习器用不同样本拟合来创建多样性,将若干个弱学习器集成为强学习器,降低方差以提高估测的准确性;Boosting 遵循迭代过程,为“顺序集成学习”,通过调整前一个弱学习器中训练集的权重以减小最终模型的偏差,包括极端梯度提升算法(Extreme Gradient Boosting,XGBoost)。堆叠集成(Stacking)算法集成不同的建模算法,每个模型都经过独立训练,多个基模型和元模型的结合产生“异构并行集成”,对于增强模型训练效果起重要作用[21—22],可以有效纠正基学习器对训练数据产生的偏差,提高估测的精度。尽管大多数算法已用于碳储量的遥感估测,但是整个江西省地形条件复杂,森林的异质性较高,如何利用遥感技术准确估测森林碳储量仍是一个挑战。Stacking 算法通常比单个经过训练的模型性能更优,它已被成功用于监督学习[23],所以该算法是否可以改进影像易饱和、大气干扰性大等因素造成的碳密度低值高估和高值低估,并提高模型估测的准确性? 这些问题还需要经过实证分析。

在机器学习领域,超参数的调整对于提高模型性能至关重要。不同于训练过程中产生的参数,超参数在训练前设置,其取值会显著影响模型的性能。目前常用的调优方法有随机搜索、网格搜索和贝叶斯优化。研究发现,在处理高维度和非均匀分布的数据时,虽然随机搜索比网格搜索更高效[24],却存在平均效果差的缺点。贝叶斯参数优化则避免了冗余操作,兼顾速度和效果两个目标[25]。但在实际工作中,对于不熟悉集成学习工作原理的人来说,如何根据需求平衡模型性能与调参成本之间的关系,搜寻超参数的最优组合却是一个难点。本研究引入 Optuna 超参数优化框架,作为一种新兴工具,它支持多种优化策略,包括贝叶斯优化、遗传算法等多种高效算法,同时与可视化工具相结合,使得用户能够轻松实现多功能框架的配置[26—27],对超参数的重要性进行分析,将其与 Stacking 算法相融合,多核处理器并行搜索来实现快速准确地估测森林碳储量。

简而言之,不同的特征变量组合、估测模型和超参数都会影响森林碳储量估测的准确性。本文以江西省为研究区,通过多种特征选择方法筛选特征子集,基于四种机器学习方法和 Optuna 超参数优化框架来构建森林碳储量估测模型,并对模型的性能进行评估,旨在利用最优估测模型对江西省森林碳储量进行计算,快速且

高精度地绘制江西省森林碳储量空间分布图,并通过地理探测器对碳储量空间分布格局的驱动因素进行了分析,为江西省可持续发展政策的制定提供科学依据。

1 研究区概况

研究区为江西省全境(图 1),地处中国东南部,长江中下游南岸,面积约 1669.46 万 hm^2,属亚热带季风气候,年均温为 16.3—19.5℃,年平均降水量为 1341—1943mm。所处地形复杂多样,北部为赣江平原,中部为丘陵盆地,南部为赣南丘陵和武夷山脉。该地是亚热带植物区系的起源中心之一,植被类型丰富,树种多样,主要包括杉木(*Cunninghamia lanceolata*)、马尾松(*Pinus massonian*)、湿地松(*Pinus elliottii*)等针叶树种,以及枫香(*Liquidambar formosana*)、木荷(*Schima superba*)、樟(*Camphora officinarum*)等阔叶树种。

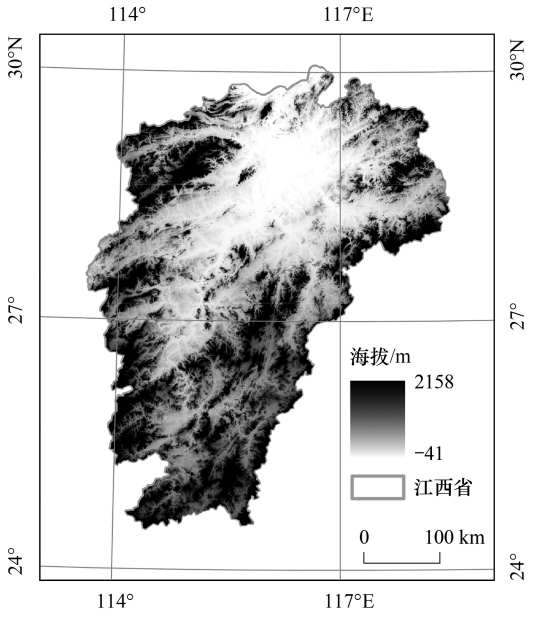

图 1 研究区示意图

Fig.1 Research area schematic diagram

2 研究数据和方法

本研究的方法主要包括三个步骤:特征提取与选择、碳储量估测模型的构建与评估、绘制森林碳储量空间分布图。通过将实地调查数据与遥感数据进行预处理,对遥感特征变量进行提取并利用两种方法筛选得到特征子集。基于四种机器学习算法,以特征子集为自变量,碳储量为因变量回归建模,对模型性能进行评估,选择最优估测模型反演得到江西省森林碳储量空间分布格局,采用地理探测器研究不同因子对碳储量空间分布的影响。

2.1 数据收集和预处理

2.1.1 实测数据收集和处理

实测数据为 2006 年江西省第七次国家森林资源连续清查固定样地数据,样地沿投影坐标系横纵轴在 8km×8km 的网格交点上布设,面积为 0.0667hm^2,形状为方形,起测直径为 5cm,调查并记录样地坐标、优势树种以及样木的胸径、树高、立木类型、检尺类型等信息。选取立木类型为有林地,检尺类型为保留木、进界木、漏测木、胸径错测木和树种错测木的树木进行碳储量的估测,其中错测木需对照两期数据订正。将样地数据与遥感影像数据叠加,剔除影像上被云覆盖的样地,最终保留了 974 个样地作为研究样本。通过对数据进行分析,优势树种分为杉木、马尾松、湿地松、栎类、枫香、木荷、其它硬阔和其它软阔八类。

采用单木生物量模型和含碳系数计算单株林木的碳储量,不同树种(组)的参数见表 1,合计得到样地森林碳密度(Mg/hm^2),其值变化范围为 1.584—168.039Mg/hm^2,平均值为 34.455Mg/hm^2。

2.1.2 卫星影像采集与预处理

基于 Google earth engine(GEE)平台获取 2006 年 8—10 月的 Landsat-5 TM Collection 2 Level-2A 级表观反射率数据,该数据经过了大气预处理和几何精校正,质量和一致性较高。利用 CFMASK(The C Function of Mask)云处理算法和 median 函数进行去云和影像镶嵌,生成研究区的年度合成影像(图 2)。

2.2 特征变量提取

光谱变量、地形因子为碳储量的估测提供基本信息,植被指数、纹理特征是与森林参数、森林生物量密切相关的因子,具有估测森林碳储量的潜力。本研究共选取 164 个特征变量来构建森林碳储量估测模型,包括 6 个光谱变量,87 个植被指数,68 个纹理特征因子和 3 个地形特征因子(表 2)。本研究从预处理后的

Landsat-5 TM 数据中提取原始波段值,并计算植被指数;纹理特征在 GEE 平台运用灰度共生矩阵导出,这一过程中先用加权线性组合将其转换为灰度级影像,再计算四种窗口大小(3×3、5×5、7×7、9×9)的纹理特征值[34—35];地形因子从 NASA(National Aeronautics and Space Administration)提供的空间分辨率为 30m 的 SRTM DEM 数据(https://cmr.earthdata.nasa.gov)中获取。

表 1　江西省主要树种单木生物量模型和含碳系数

Table 1　Biomass models and carbon coefficients of major tree species in Jiangxi Province

树种(组) Species (Group)	模型 Model	含碳系数 Carbon coefficient	参考文献 References
杉木 *Cunninghamia lanceolata*	$M_A = 0.032718D^{2.11093}H^{0.60212}$ $M_B = 0.008199D^{2.62298}H^{-0.00956}$	0.5201	[28]
马尾松 *Pinus massoniana*	$M_A = 0.078D^{2.115}H^{0.433}$ $M_B = 0.008828D^{2.73828}H^{-0.080255}$	0.4596	[29]
湿地松 *Pinus elliottii*	$M_A = 0.083890D^{2.44091}$ $M_B = 0.043570D^{2.22877}$	0.4936	[30]
栎类 *Quercus* L.	$M_A = 0.21360D^{2.30416}$ $M_B = 0.110595D^{2.05730}$	0.5004	[31]
枫香 *Liquidambar formosana*	$M_A = 0.10615D^{2.46650}$ $M_B = 0.09552D^{2.14190}$	0.4668	[32]
木荷 *Schima superba*	$M_A = 0.17685D^{2.26314}$ $M_B = 0.064079D^{2.19784}$	0.4706	[33]
其他硬阔 Other hardwood broad-leaved tree	$M_S = 0.044(D^2H)\ 0.9169$; $M_P = 0.023(D^2H)\ 0.7115$ $M_R = 0.0104(D^2H)\ 0.9994$; $M_L = 0.0188(D^2H)\ 0.8024$ $M_A = M_S + M_P + M_R + M_L$; $M_B = 0.0197(D^2H)\ 0.8963$	0.4834	[29]
其它软阔 Other softwood broad-leaved tree	$M_A = 0.0495502\ (D^2H)^{0.952453}$; $M_B = \frac{M_A}{3.85}$	0.4956	[29]

M_A:地上生物量 Aboveground biomass;M_B:地下生物量 Underground biomass;M_S:树干生物量 Trunk biomass;M_P:树皮生物量 Bark biomass;M_R:树枝生物量 Branch biomass;M_L:树叶生物量 Leaf biomass;D:胸高直径 Diameter at breast height;H:树高 Height

2.3　特征选择的方法

从多光谱数据中生成的特征容易产生多重共线性,设置相关性阈值(相关性>0.98)去除多重共线性的影响,将剩余变量用于特征选择。为避免模型训练速度过慢,许多研究会对特征进行筛选。递归特征消除(RFE)被认为是嵌入式选择方法,通过对基模型反复训练逐一消除特征,依据消除顺序对特征进行打分,筛选出最优特征子集。Boruta 为包裹式方法,通常在外部使用模型来评估特征重要性,考虑了决策树的平均损失精度的波动值,将初始特征与影子特征相结合,评估初始特征和对应影子特征的重要性,初始特征显著优于对应的影子特征,则标注为重要特征。通过 RFE 和 Boruta 两种特征选择方法来评估不同变量的特征重要性,筛选出最优特征子集[25]。

2.4　森林碳储量估测的机器学习算法

选择随机森林、支持向量机、极端梯度提升、堆叠集成四种机器学习算法对森林碳储量进行估测。

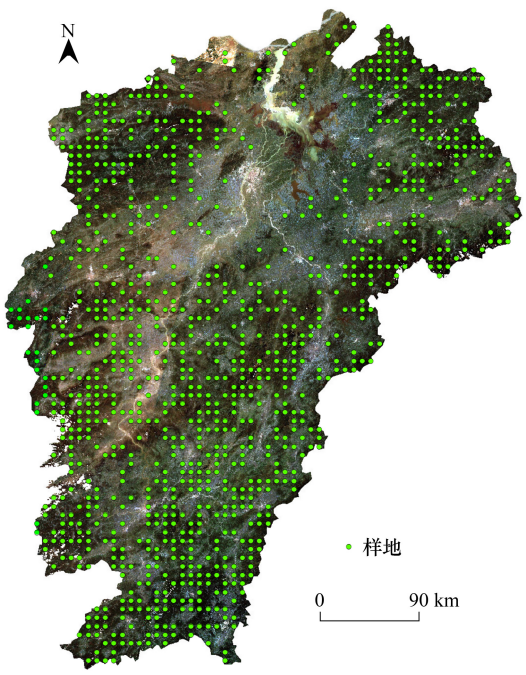

图 2　江西省 2006 年 Landsat-5 TM 合成影像

Fig.2　Landsat-5 TM composite image of Jiangxi Province in 2006

表 2 本研究中提取的特征变量

Table 2 Feature variables were extracted in the study

变量类型 Variable type	变量个数 Variable number	变量名称 Variable name	描述 Description
光谱变量 Spectral variables	6	$B1$,$B2$,$B3$,$B4$,$B5$,$B7$	Landsat-5 表观反射率:$B1$ 为 Blue;$B2$ 为 Green;$B3$ 为 Red;$B4$ 为 NIR;$B5$ 为 SWIR1;$B7$ 为 SWIR2
植被指数 Vegetation indexes	87	NDVI	$(B4 - B3)/(B4 + B3)$
		RVI	$B4/B3$
		DVI	$B4 - B3$
		SAVI	$(1 + L)\ (B4 - B3)/(B4 + B3 + L)$
		MSAVI	$(2 \times B4 + 1 - [(2 \times B4 + 1)^2 - 8 \times (B4 - B3)]^{1/2})/2$
		EVI	$2.5 \times (B4 - B3)/(B4 + 6 \times B3 - 7.5 \times B1 + 1.5)$
		GNDVI	$(B4 - B2)/(B4 + B2)$
		NLI	$(B4^2 - B3)/(B4^2 + B3)$
		OSAVI	$(B4 - B3)/(B4 + B3 + 0.16)$
		IIVI	$(B4 - B5)/(B4 + B5)$
		mNDVI	$(B4 - B3)/(B4 + B3 - 2 \times B1)$
		TNDVI	$[(B4 - B3)/((B4 + B3) + 0.5)]^{1/2}$
		二波段比值植被指数(B_{i_j})	B_i/B_j ,i、j 为 1、2、3、4、5、7,且 $i \neq j$。
		三波段比值植被指数(B_{i_jk})	$B_i/(B_j + B_k)$,i、j、k 为 1、2、3、4、5、7,且 $i \neq j \neq k$。
纹理特征 Texture features	68	ASM_w、CON_w、CORR_w、VAR_w、IDM_w、SAVG_w、SVAR_w、SENT_w、ENT_w、DENT_w、DISS_w、DVAR_w、SHADE_w、PROM_w、INE_w、IMCOR1_w、IMCOR2_w	基于 GEE 平台将 $B2$、$B3$ 和 $B4$ 进行波段合成,利用 ee.Image.glemTexture 函数进行计算。XXX_w 表示在 Landsat-5 TM 中分别使用窗口大小为 $w \times w$ 获取的纹理特征因子,w 取值可为 3、5、7、9。
地形特征 Terrain features	3	海拔(Elevation) 坡度(Slope) 坡向(Aspect)	利用 ee.Alorithms.Terrain 进行计算

NDVI:归一化植被指数 Normalized difference vegetation index;RVI:简单比值指数 Ratio vegetation index;DVI:差值植被指数 Difference vegetation index;SAVI:土壤调整植被指数 Soil adjusted vegetation index;MSAVI:修正型土壤调节植被指数 Modified soil adjusted vegetation index;EVI:增强型植被指数 Enhanced vegetation index;GNDVI:归一化绿度植被指数 Green normalized vegetation index;NLI:非线性植被指数 Nonlinear vegetation index;OSAVI:优化土壤调整植被指数 Optimized soil-adjusted vegetation index;IIVI:红外植被指数 Infrared vegetation index;mNDVI:改进归一化植被指数 Modified normalized difference vegetation index;TNDVI:转换归一化植被指数 Transformed normalized difference vegetation index;ASM:角二阶矩 Angular second moment;CON:对比度 Contrast;CORR:相关性 Correlation;VAR:方差 Variance;IDM:逆差异矩 Inverse difference moment;SAVG:总平均值 Sum average;SVAR:总方差 Sum variance;SENT:总熵 Sum entropy;ENT:熵 Entropy;DENT:差异熵 Difference entropy;DISS:差异 Dissimilarity;DVAR:差异方差 Difference variance;SHADE:阴影度 Cluster shade;PROM:集聚度 Cluster prominence;INE:惯性 Inertia ;IMCOR1:信息相关性 1 Information measure of corr.1;IMCOR2:信息相关性 2 Information measure of corr.2;L 为通过冠层时红光和近红外消光差异,取值为 0.5;植被指数中二波段比值植被指数 15 个,三波段比值植被指数 60 个

2.4.1 随机森林

随机森林(RF)作为集成学习中 Bagging 最经典的算法之一,通过创建多个决策树预测来得到最终的结果。它在每个决策树中对目标变量进行独立预测,最终输出所有预测的平均值,有助于生成去自相关的决策树降低方差以提高预测的准确性,能更敏锐地处理高维数据[36]。

2.4.2 支持向量机

支持向量机(SVR)是 Vapnik 于 1982 年提出的一种用于回归的训练算法[37]。根据输入特征的数量来创建一个多维空间,每个特征代表这个空间中的一个坐标轴,目标是用指定的核函数生成一个包含最多样本点的超平面,平面周围设置一个“间隔”,使得间隔内样本点的预测值与真实值差异最小。

2.4.3 极端梯度提升

极端梯度提升(XGBoost)是基于梯度提升框架下实现的监督集成算法[38]。XGBoost 根据前面学习器的

学习误差率更新训练样本的权重,在每次迭代过程中评估学习器的残差,增加错误样本被包含在下一棵决策树中的权重,通过逐步减小残差的方式来提高模型估测的准确性。

2.4.4　堆叠集成

堆叠集成(Stacking)作为经典的集成算法之一,与 RF 和 XGBoost 集成同质学习器不同,它将多样化的基学习器集成以改进模型评估的准确性,元模型的高泛化能力又能修正训练过程的偏差[39]。有研究表明,将 3—4 个基学习器堆叠使用效果最佳,简单的线性模型用作元模型可以对模型做平滑解释[23,40]。本研究选择 RF、SVR、XGBoost 作为基学习器,将 MLR(Multiple Linear Regression)、Ridge、Lasso 作为备选的元学习器。

2.5　超参数调优

Optuna 是一个超参数自动优化框架,通过迭代调用和评估不同参数值的目标函数来获取最优解[41],整体是一种基于改进贝叶斯的试错算法,终止预测可能性较小的参数区间,让更多的算力用于可能性更高的区域以提高搜索效率。本研究将 Optuan 超参数优化框架与四种机器学习模型相结合,设置 Optuna 的目标函数、优化方向和迭代次数,来确定每个估测模型的最优超参数组合。

2.6　模型评估

首先,为了提高模型的收敛速度,考虑到不同类型特征表现出的取值范围,利用 Min-Max 函数将各特征变量的取值归一化到 0—1 区间[36]。其次,将所有样本随机分成训练集(80%)和验证集(20%),训练模型时利用五折交叉验证来优化参数。最后,使用验证集计算决定系数(Coefficient of Determination,R^2)、均方根误差(Root Mean Squared Error,RMSE)、平均绝对误差(Mean Absolute Error,MAE)、均方误差(Mean Squared Error,MSE)等指标来评估模型的估测精度。

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

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

式中,y_i为实测的样地生物量,$\bar{y}_i$为实测的样地生物量均值,$\hat{y}_i$为预测的样地生物量,i 为方程中的数据个数,n 为样本数量。

2.7　森林植被碳储量空间分布制图

本研究通过 Globeland 30[42]中 2000 年和 2010 年数据将江西省十年间的土地利用类型变化分为森林、非森林、森林干扰和恢复这三类。结合 MCD12Q1.061(https://lpdaac.usgs.gov)中 2006 年数据,判定森林发生干扰和恢复的区域在 2006 年是否为森林,将这一类中判定为森林的区域与 10 年间土地利用类型一直为森林的区域进行镶嵌,生成空间分辨率为 30m 的 2006 年江西省森林分布图。然后,选取最优碳储量估测模型将 Landsat-5 TM 的 2006 年合成影像进行反演,得到江西省森林碳储量空间分布图。利用地理探测器以温度、降水、人口、海拔、坡度和坡向为研究变量,探究不同因子对森林碳储量分布格局的驱动影响,其中温度和降水存储于 GEE 平台的气候数据集("ECMWF/ERA5/MONTHLY"),人口数据由中国科学院资源环境科学数据中心下载,其核心思想为若 X 对 Y 有重要影响时,两者在空间分布上应该具有相似性,并且用 q 值来度量变量对碳储量空间分布的解释力大小,q 值越大表示解释力越强[43]。

3 结果与分析

3.1 变量选择结果

为了消除变量间多重共线性的影响,剔除了自相关性高的特征因子,剩余 99 个特征因子用于进行特征筛选,其中包括 6 个光谱变量、48 个植被指数、42 个纹理特征因子和 3 个地形特征因子。

在 RFE 特征选择过程中,我们选择随机森林作为基模型,在训练过程中利用评价测试分数进行模型性能评估,依据最小的 MSE 来确定特征子集的个数。随着输入变量数量的增多,MSE 在 0—30 的区间呈现持续下降的趋势,MSE 在 30—99 区间呈平稳趋势(图 3)。因此,确定选择的变量数量为 30,这样既保留了原始数据集的大部分信息,也减少了模型运行负载。图 4 是 RFE 方法选出的最优特征子集中重要性前 30 的特征变量排序,这些特征包括 4 个光谱变量、14 个植被指数、10 个纹理特征和 2 个地形特征。

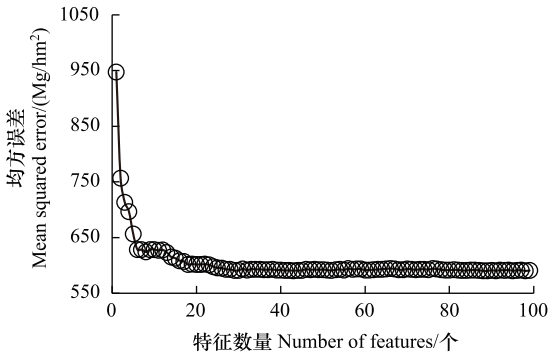

图 3 RFE 特征选择的准确性随特征数量的变化趋势

Fig.3 The accuracy of RFE feature selection changes with the number of features

RFE:递归特征消除 Recursive feature elimination

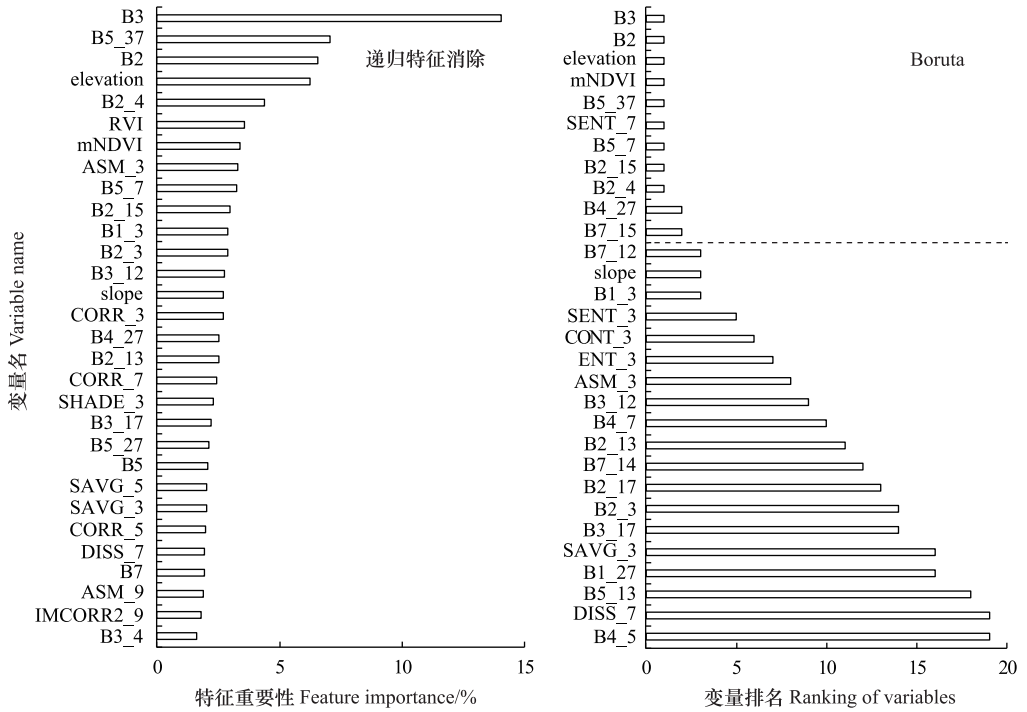

图 4 RFE 和 Boruta 方法确定的重要性排序前 30 的特征变量

Fig.4 The top 30 feature variables ranked by importance determined using RFE and Boruta methods

*B*1:蓝波段 Blue;*B*2:绿波段 Green;*B*3:红波段 Red;B4:近红外波段 Near infrared;*B*5:短波红外 1 波段 Shortwave infrared 1;*B*7:短波红外 2 波段 Shortwave infrared 2;RVI:简单比值指数 Ratio vegetation index;mNDVI:改进归一化植被指数 Modified normalized difference vegetation index;B_{i_j}和B_{i_jk}分别为二波段比值植被指数和三波段比值植被指数,i、j、k 可取 1、2、3、4、5、7;其中 XXX_w 表示在 $w\times w$ 窗口大小下的纹理特征值,ASM:角二阶矩 Angular second moment;CORR:相关性 Correlation;SAVG:总平均值 Sum average; DISS:差异 Dissimilarity; SENT:总熵 Sum entropy;ENT:熵 Entropy;SHADE:阴影度 Cluster shade;IMCORR2:信息相关性 2 Information measure of corr.2;elevation:海拔;slope:坡度

Boruta 算法通过给每个特征一个等级来表示在特征选择过程中的相对重要性,等级为 1 和 2 的变量分别

对应显著特征和不确定特征,其他等级为不显著特征。本研究中除显著特征外,也保留不确定特征,对后续预测精度效果改进有帮助。最后通过 Boruta 筛选出 11 个特征变量(图 4),分别为 *B*2,*B*3,B2_4,B5_7,mNDVI,B2_15,B4_27,B5_37,B7_15,SENT_7,elevation。

3.2　模型性能的比较

3.2.1　模型超参数的确定

基于 Optuna 超参数优化框架,设置迭代次数为 300,得到不同最优特征子集与四种机器学习组合的超参数调优结果(表 3)。Optuna 提供了 Web 仪表盘用于可视化和分析研究,由图 5 发现使用不同的特征选择方法,超参数在模型中的重要性表现略有差异。在 RF 模型中,叶节点最小样本数(min_samples_leaf)超参数重要性最高,约为 40%,该参数在模型的拟合中具有重要影响;核函数(kernel)对于 SVR 的超参数贡献度最高,说明了核函数的选取对于 SVR 模型至关重要;XGBoost 模型中学习率(learning_rate)具有很高的超参数重要性,为 69%和 88%,学习率较高的模型可以快速达到拟合最佳点,提升模型的性能;Stacking 模型中,基模型的超参数在不同的特征子集中差异较大,元模型选择了 Lasso 算法,只需要调节正则化强度(alpha)将模型达到最优性能。

表 3　模型超参数优化的结果

Table 3　Results of model hyperparameter optimization

模型 Model	最优特征子集 Optimal feature subset	
	递归特征消除 Recursive feature elimination	Boruta
随机森林 Random forest	'n_estimators':87,'max_depth':467, 'max_features':5,'min_samples_split':6, 'min_samples_leaf':4,'random_state':41	'n_estimators':1056,'max_depth':463, 'max_features':1,'min_samples_split':10, 'min_samples_leaf':7,'random_state':42
支持向量机 Support vector regression	'kernel': 'poly','C':4.69,'gamma':0.84	'kernel': 'linear','C':29.99,'gamma':7.66e-05
极端梯度提升 Extreme gradient boosting	'n_estimators':3000,'max_depth':3, 'learning_rate':0.0015,'subsample':0.7, 'colsample_bytree':0.23,'reg_alpha':0.32, 'reg_lambda':0.0002	'n_estimators':464,'max_depth':21, 'learning_rate':0.0065,'subsample':0.1, 'colsample_bytree':0.28, 'reg_alpha':1.2e-06,'reg_lambda':9.36e-05
堆叠集成 Stacking	'rf_n_estimators':148,'rf_max_depth':279, 'max_features':9,'min_samples_split':8, 'min_samples_leaf':8,'random_state':43, 'kernel': 'linear','C':13.35,'gamma':7.89e-05, 'xgb_n_estimators':2947,'xgb_max_depth':1, 'learning_rate':0.0067,'subsample':0.5, 'colsample_bytree':0.89,'reg_alpha':0.01, 'reg_lambda':62.62, 'Meta-model': Lasso,'alpha':6.1654	'rf_n_estimators':156,'rf_max_depth':470, 'max_features':1,'min_samples_split':10, 'min_samples_leaf':9,'random_state':43, 'kernel': 'rbf','C':13.34,'gamma':1.146, 'xgb_n_estimators':1776,'xgb_max_depth':2, 'learning_rate':0.0037,'subsample':0.9, 'colsample_bytree':0.11,'reg_alpha':6.96, 'reg_lambda':7.33e-05, 'Meta-model': Lasso,'alpha':0.0059

n_estimators:学习器数量;max_depth:树的最大深度;max_features:最大特征数;min_samples_split:节点最小样本数;min_samples_leaf:叶节点最小样本数;random_state:种子数;kernel:核函数;C:惩罚参数;gamma:核函数系数;learning_rate:学习率; subsample:样本抽样比例;colsample_bytree:决策树中抽样比例;reg_alpha:L1 正则化项系数;reg_lambda:L2 正则化项系数;rf_max_depth 和 rf_n_estimators 为 Stacking 算法中基模型为 RF 的超参数,xgb_max_depth 和 xgb_n_estimators 为基模型为 XGBoost 的超参数;Meta-model:元模型;alpha:正则化强度

3.2.2　最优特征子集中特征变量的重要性评估

本研究统一选取置换特征重要性(Permutation Feature Importance,PFI)来评估通过 RFE 和 Boruta 两种方法选出的最优特征子集中每个特征变量在不同机器学习算法模型中的贡献度。PFI 基本思想是通过对输入特征进行随机顺序置换,来评估它们对模型性能的影响,PFI 值越高,说明模型对该特征变量的依赖性高,即该特征变量对于模型性能越重要。

在 RFE 最优特征子集中,发现“B4_27”在 RF、Stacking 中具有最高的特征重要性,分别为 16%和 12%;SVR 模型中,纹理特征“SAVG_3”特征重要性为 15%;在 XGBoost 模型中,比值植被指数“B2_13”对碳储量估测模型的预测能力有积极贡献;“B2_13”对四个模型都有较大的贡献(图 6)。

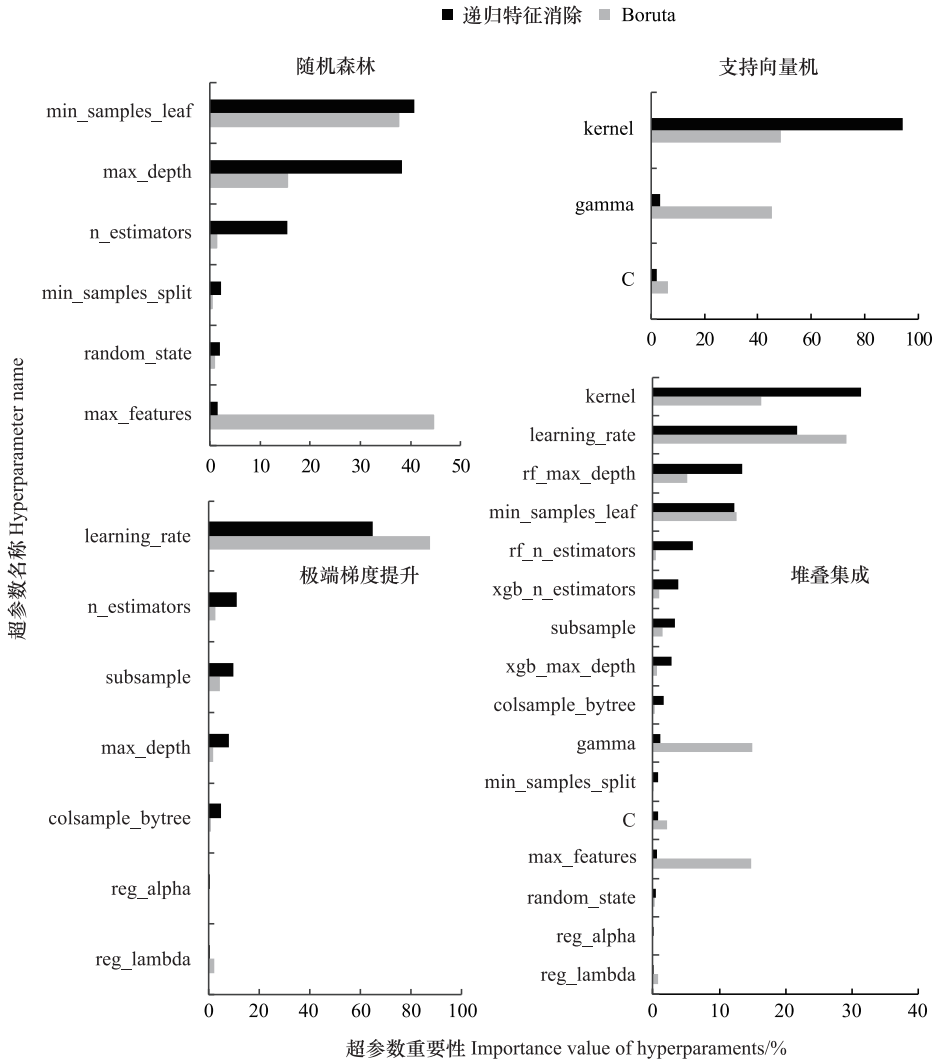

图 5 模型中各超参数的重要性

Fig.5 The importance of each hyperparameter in the model

min_samples_leaf:叶节点最小样本数;max_depth:树的最大深度;n_estimators:学习器数量;min_samples_split:节点最小样本数;random_state:种子数;max_features:最大特征数;kernel:核函数;gamma:核函数系数;C:惩罚参数;learning_rate:学习率; subsample:样本抽样比例;colsample_bytree:决策树中抽样比例;reg_alpha:L1 正则化项系数;reg_lambda:L2 正则化项系数;rf_max_depth 和 rf_n_estimators 为 Stacking 算法中基模型为 RF 的超参数,xgb_max_depth 和 xgb_n_estimators 为基模型为 XGBoost 的超参数

Boruta 筛选的 11 个特征中,“mNDVI”在 RF、XGBoost 和 Stacking 具有最高的特征重要性,分别为 27%、23%和 21%。而“B4_27”对 SVR 模型具有最高的特征重要性,为 39%(图 7)。

这也就说明,最优特征子集和机器学习算法的不同都会导致各个特征变量在模型中的重要性产生差异。部分特征变量重要性为负值,说明这些变量在训练过程中对拟合模型不重要,甚至可能降低模型性能,例如,RF 模型中的“B5_37”等。

3.2.3 不同机器学习算法的比较分析

在两种特征选择方法下,对比分析了 4 种机器学习算法构建的森林碳储量估测模型。结果表明,对于 RF、XGBoost 和 Stacking 算法,利用 RFE 筛选的最优特征子集建模时,模型性能显著提高,R^2 分别提升了 0.019、0.009 和 0.033,斜率更加趋近于 1;而对于 SVR 算法,使用 Boruta 筛选的最优特征子集作为自变量时,模型性能提升更为显著,R^2 提升了 0.056,斜率增加 0.0721。在特征子集相同的情况下,基于单一学习器集成

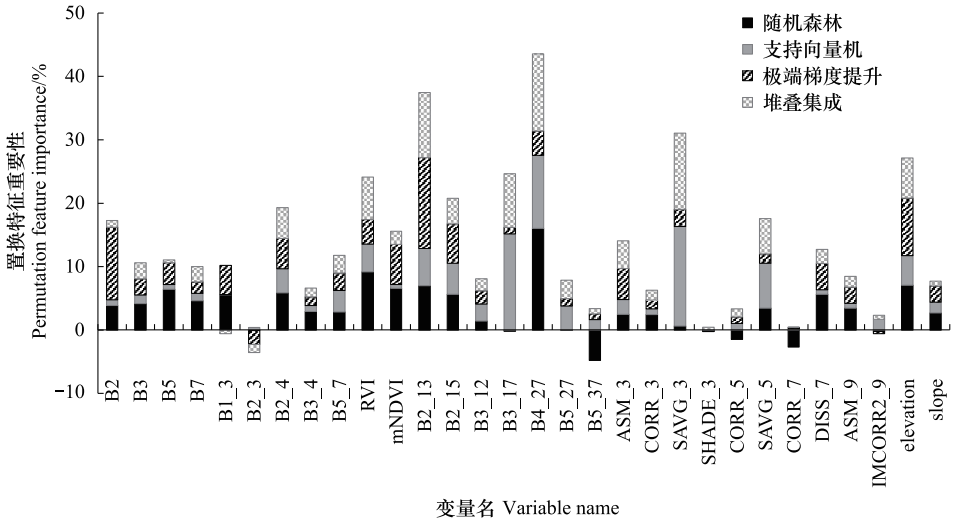

图 6　RFE 最优特征子集的置换特征重要性

Fig.6　Permutation feature importance of RFE optimal feature subset

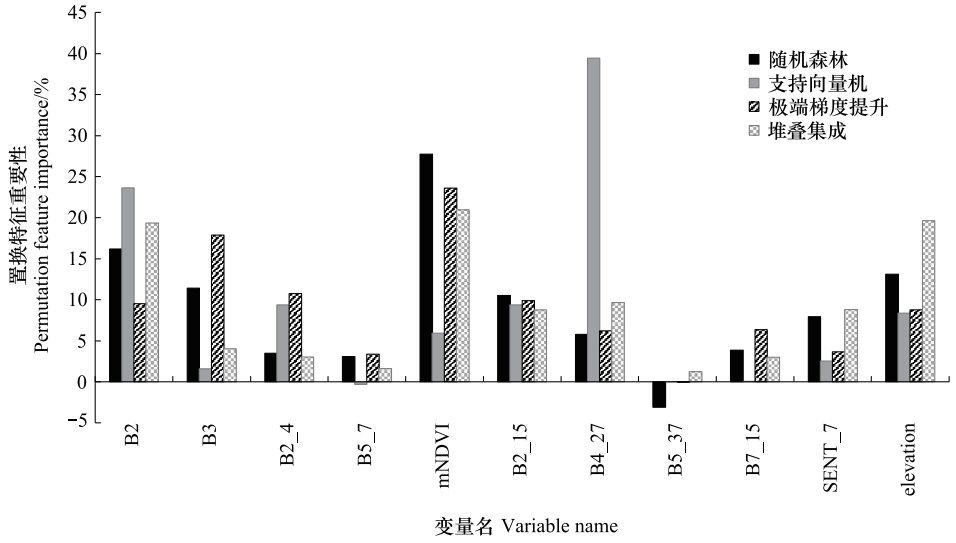

图 7　Boruta 最优特征子集的置换特征重要性

Fig.7　Permutation feature importance of Boruta optimal feature subset

的 XGBoost 模型预测能力强于 RF 和 SVR,而以 RF、SVR、XGBoost 作为基模型的 Stacking 集成学习模型相较于单一学习器,斜率为 0.4772—0.5049,R^2 为 0.520—0.527,RMSE 为 15.85—15.98Mg/hm^2,MAE 为 12.31—12.42Mg/hm^2(图 8),能更好得捕捉到数据间的变化趋势,预测值与实测值的一致性较好,模型的拟合优度更佳。

因此,选择合适的特征子集可以降低不同机器学习算法的预测偏差并提升拟合效果,例如,SVR 算法与 Boruta 选择的特征相结合;另一方面,集成学习在训练过程中能够充分提取变量信息,而 Stacking 算法结合多个学习器的优点,以生成高性能、准确且方差较小的模型以改进模型的预测能力,斜率为 0.5049 表明所获得估测结果的散点图更接近 $y=x$ 这条直线,在一定程度上改进了光学遥感中存在的低值高估和高值低估问题,能够提供更加准确的森林碳储量估测结果。

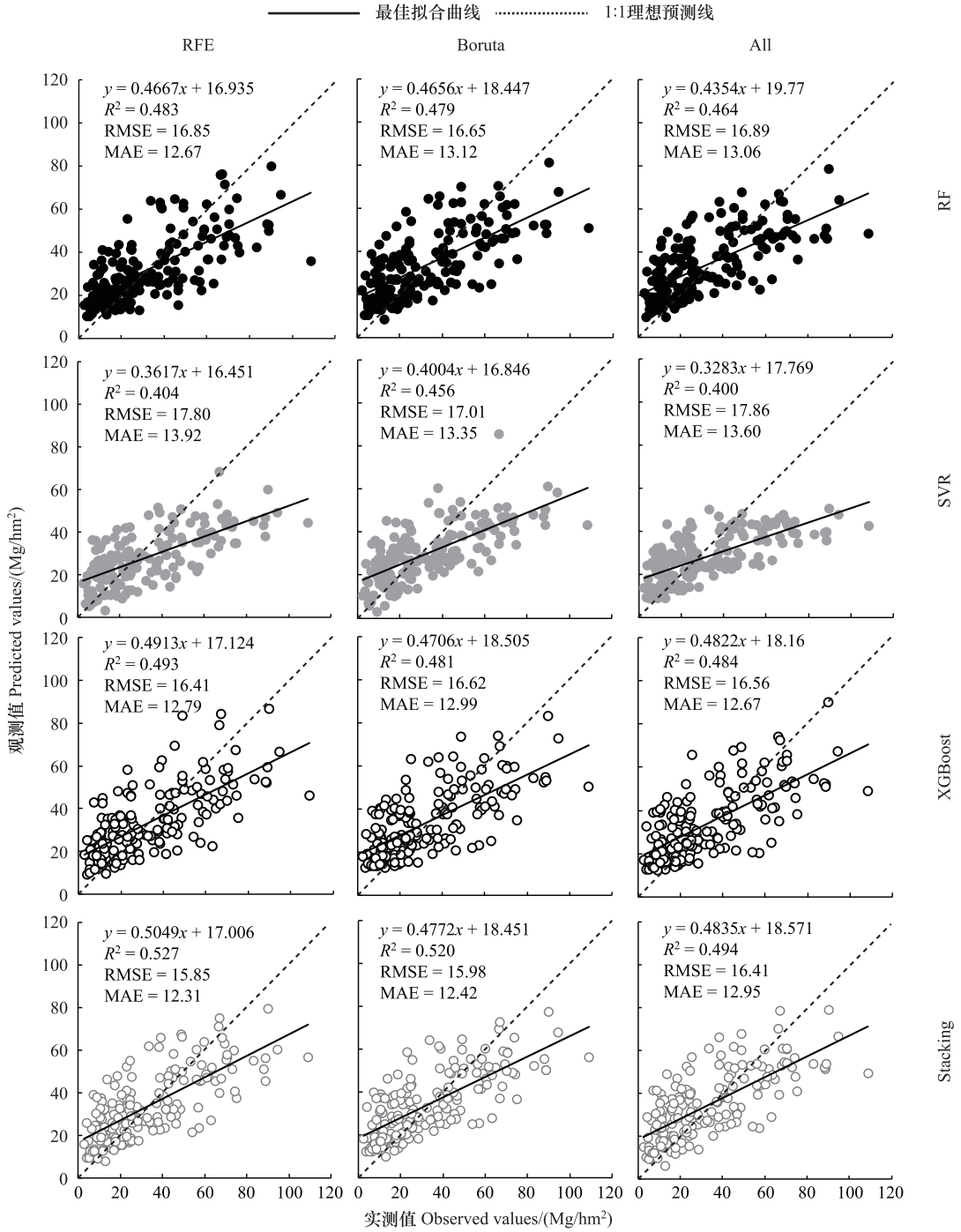

图 8 使用 RF、SVR、XGBoost 和 Stacking 算法绘制森林碳储量预测值与观察值的散点图

Fig.8 Using RF, SVR, XGBoost and Stacking algorithms to draw a scatter plot of forest carbon storage predicted values and observed values

RMSE:均方根误差 Root mean squared erroe;MAE:平均绝对误差 MAE Mean Absolute Error;All 为未经特征选择的特征集

3.3 江西省森林碳储量空间分布

通过以上分析,本研究采用集成学习 Stacking 算法,利用 RFE 筛选的最优特征子集为自变量,构建江西省森林碳储量估测模型,并模拟研究区森林碳储量空间分布。结果表明(图 9),江西省森林碳储量总量为 321.507Tg,平均碳密度为 33.356Mg/hm^2(2.585—88.943Mg/hm^2)。江西省森林碳储量空间分布与江西省自然地理有着重要联系。江西省东部为武夷山脉,西部是罗霄山脉,南部有大庾岭和九连山,三面环山,北部为鄱

阳湖湖滨平原,中部丘陵平原交错分布,全省呈一个整体向北处鄱阳湖开口的地貌。森林碳密度较高的区域大多分布在东北、西北和南部等山地较多的地区,中部、北部平原以及南部的章江、贡水流域等地区因其水热条件适宜,田地面积分布广泛,森林碳密度相对较低。

探究森林碳储量空间分布格局与各因子之间的驱动关系,结果(表 4)表明,该地区温度、降水、人口、海拔、坡度等因子 p 值均为 0,与碳储量分布有显著性关系。其中,海拔和坡度的 q 值最高,分别为 0.3142 和 0.1766,是影响碳储量空间分布格局的主要驱动因子。此外,双因子交互的解释力均高于单因子,其中人口、海拔、坡度是交互解释力增强的关联驱动因子,降水与温度、人口、海拔、坡向以及坡向与温度、海拔、坡度、人口间交互均呈非线性增强作用,其他因子间交互均呈双因子增强。

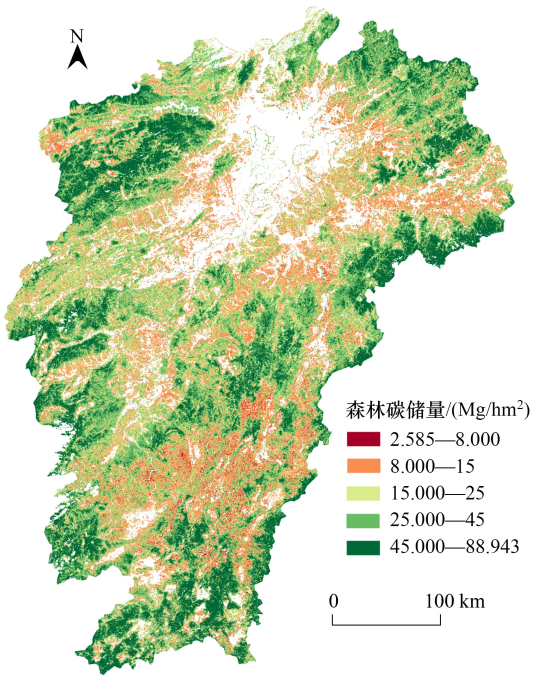

图 9　江西省森林碳储量空间分布图

Fig.9　Spatial distribution map of forest carbon storage in Jiangxi Province

4　讨论

4.1　估测模型结果评价

4.1.1　特征变量的评价

对于部分机器学习算法,更多的特征变量能提供充足的信息,但也可能因数据冗余而导致模型预测性能下降。本研究通过 RFE 和 Boruta 两种方法筛选得到的最优特征子集,与四种机器学习算法组合时发现,RF、XGBoost、Stacking 与 RFE 特征子集结合时,对于碳密度高值和低值区域的预测更为准确,模型拟合效果得到增强;SVR 则与 Boruta 特征子集结合时模型性能有更大提升,由此可见,合适的变量组合可以提高森林碳储量估测的精度。这与 Luo 等[14]的研究结果相一致,他发现不同特征选择方法与回归算法相结合的预测效果优于使用单一机器学习算法同时进行特征筛选和回归的模型。从变量的重要性来看,比值植被指数和 mNDVI 的特征重要性较高,然而,本研究中纹理特征因子并没有表现出较高的重要性,与 Li 等[44]的研究中影像纹理与生物量之间存在高度相关性的结果存在差异,这可能是由于不同树种之间的图像纹理规律差异较大,在估测大尺度碳储量时,植被类型和结构差异较大影响了纹理特征的可用性,而年度影像合成过程中,光照和植物间的遮盖又加剧了这种影响;另外,也有研究认为在结构较简单的森林中,植被指数变量相关性强于纹理 特征[45]。此外,本研究从提取的变量中获取冠层的平面信息,不涉及垂直结构,而在植被覆盖率较高的

表 4　不同因子解释力的大小

Table 4　The explanatory of the different factors

探测因子 Detection factor	q 值						P
	温度 Temperature	降水 Precipitation	人口 Population	海拔 Elevation	坡度 Slope	坡向 Aspect	
温度 Temperature	0.0389						0.000
降水 Precipitation	0.0736	0.0136					0.000
人口 Population	0.0911	0.0896	0.0661				0.000
海拔 Elevation	0.3279	0.3288	0.3313	0.3142			0.000
坡度 Slope	0.0911	0.1858	0.2186	0.3506	0.1766		0.000
坡向 Aspect	0.0454	0.0203	0.0720	0.3182	0.1809	0.0017	0.1738

q 值表示该变量对碳储量的空间分异解释力的大小;P 值为在 0.05 水平下进行显著性检验的结果

森林中,饱和现象导致的估测偏差可能是本研究预测精度偏低的原因之一。激光雷达技术提取的特征变量可以获取地表到冠层的垂直结构信息,缓解饱和度并加深对森林三维结构的理解[46—47],因此对于大尺度的碳储量估测,可以考虑将光学遥感影像与 SAR 影像相结合构建兼具高时空分辨率的长时间序列影像,但需要解决影像融合和复杂地形、大气散射对信号衰减的问题。

4.1.2 估测模型的评价

本研究中 XGBoost 作为单个学习器的预测效果比 RF 和 SVR 好,它能够基于先前的决策树生成新的树纠正误差,还改进了对于正则化学习目标的处理,但是该算法容易过拟合;RF 与 XGBoost 相比,预测性能存在轻微的差异,但是 RF 在训练过程中运行时间过长,预测潜力相对较低。值得关注的是,作为集成算法,Stacking 将前面三种算法作为基模型进行堆叠,与其他模型进行对比,在碳密度较高的林分中,Stacking 算法估测结果更准确,但是不足之处在于,虽然它整合了所有基学习器的预测结果,通过捕捉不同模型优势,降低随机偏差来提高预测的准确性,但是并没有完全解决高值低估的问题, 在之前的研究中该问题也常出现,可能是样本选取不合适,使得模型无法在训练集外进行推断结果[48]。此外,三种集成学习算法对于碳密度低值点的预测效果不佳,可能是影像信息受到林下植被和土壤等的影响,并且集成学习在综合多个模型的结果进行平滑输出时,结果会偏向数据的整体趋势,从而忽略了低值点的特殊性。

4.1.3 估测结果的不确定性

本研究估测的 2006 年江西省森林碳密度平均值为 33.356Mg/hm^2。与一些同类研究的估测结果相近,例如,李海奎等[49]利用第七次全国森林资源清查数据,通过生物量经验方程及含碳系数求得江西省乔木林碳密度为 30.010Mg/hm^2。Wang 等[50]基于高精度曲面建模方法将遥感数据和清查数据进行融合,模拟估算江西省森林碳密度为 30.600Mg/hm^2。然而,廖凯涛等[51]利用 GLAS 的激光雷达数据结合 TM 光学影像中获取的植被指数,采用回归模型估算江西省森林覆盖区的平均生物量为 141.213Mg/hm^2,与本研究估算结果差异较大,究其原因,一方面前者研究中未建立不同树种的生物量模型,实测数据获取的时间不一致,研究区气候条件变化较大,树种差异性和生长潜力差异等均会影响估测的准确性;另一方面,本研究可能存在样地坐标与遥感影像精准匹配的问题,导致研究结果存在一定不确定性[52]。

可用于森林碳储量遥感建模的特征因子众多,如本研究中采用了光谱变量、植被指数、纹理特征和地形特征,除此之外还有一些潜在变量,例如,气候因子、冠层高度等。众所周知,不同预测变量的重要性会随着研究区的环境差异有所不同[53]。江西省森林大多处在山地丘陵,气温、降水量、土壤性质等变化较大,植被物候的复杂性也导致植被形态各异,通过 RFE 方法筛选出的特征变量中,“B4_27”和“B2_13”对这种环境变化的解释能力较强。中等空间分辨率的 Landsat-5 TM 数据的表观反射率受地面植被、土壤和云层影响较大,当碳储量高到一定程度时,容易出现影像信息的饱和,导致林分出现高值低估问题,为此有研究指出可以将更高空间分辨率的光学遥感数据与激光雷达数据融合[54],在模型中引入冠层高度变量,同时从混合像元分解、数据清洗等方向来改进估测精度[55]。此外,在利用 Landsat-5 TM 数据估测复杂林分碳储量时,光谱变量对林分结构不敏感,特别是对成熟林的估测效果较差[56],因此,可以引入树种组成和林分密度等变量,按照森林类型和结构对模型进行分层拟合[57],增加模型精度的同时提高模型鲁棒性。

4.2 江西省森林碳储量分布格局驱动因子分析

地理探测器结果表明,海拔和坡度是江西省森林碳储量空间分异的主要驱动因子,温度、降水、人口等驱动因子对碳储量空间分布产生一定的作用,但整体上解释性较低。交互因子的探测结果表明,不同因子间的结合对碳储量分布格局呈非线性增强和双因子增强,说明在研究过程中考虑多因子的协同作用的必要性。江西省森林碳储量分布一方面主要受到单个自然因素的驱动,其中海拔和坡度等地形因子通过改变研究区的小气候来调节植物的生理生态过程,进而影响森林碳储量的空间分布格局。另一方面,自然因素与人为因素间交互作用,在研究区中高海拔地区,人口的减少以及天然林资源保护工程等影响,在一定程度上减少了人类活动对于森林的干扰,因此森林碳储量处于一个较高的水平;降水与其他因子交互呈非线性增强作用表明森林

生长对于水分、温度、光照等条件的依赖性,森林生长越好其碳储量越高。

5　结论

(1)特征选择对于碳储量的预测效果有一定的影响,除 SVR 之外,各机器学习算法与 RFE 筛选的特征子集相结合而构建的森林碳储量估测模型效果更优。

(2)Optuna 超参数优化框架融合机器学习算法,将超参数的重要性可视化,不同特征子集的超参数取值有差异,在 Stacking 集成学习模型中,核函数和学习率的调优对模型影响显著。

(3)集成学习算法在森林碳储量估测模型构建中具有巨大的潜力,其中以 RFE 为特征选择方法,以 RF、SVR、XGBoost 为基模型的 Stacking 集成学习模型估测精度最高,R^2为 0.527,RMSE 为 15.85Mg/hm^2,MAE 为 12.31Mg/hm^2,用此模型估测得到江西省森林碳密度平均值为 33.356Mg/hm^2,森林碳储量总量为 321.507Tg。

(4)森林碳储量的分布受到多重因素共同作用的影响。海拔、坡度是主要驱动因子,其次是人口、温度和降水。各因子在交互作用下驱动力增强,自然因素与人为因素对碳储量的分布格局产生了协同作用。

参考文献(References):

[1] 杨元合, 石岳, 孙文娟, 常锦峰, 朱剑霄, 陈蕾伊, 王欣, 郭焱培, 张宏图, 于凌飞, 赵淑清, 徐亢, 朱江玲, 沈海花, 王媛媛, 彭云峰, 赵霞, 王襄平, 胡会峰, 陈世苹, 黄玫, 温学发, 王少鹏, 朱彪, 牛书丽, 唐志尧, 刘玲莉, 方精云. 中国及全球陆地生态系统碳源汇特征及其对碳中和的贡献. 中国科学: 生命科学, 2022, 52(4): 534-574.

[2] Pan Y D, Birdsey R A, Fang J Y, Houghton R, Kauppi P E, Kurz W A, Phillips O L, Shvidenko A, Lewis S L, Canadell J G, Ciais P, Jackson R B, Pacala S W, McGuire A D, Piao S L, Rautiainen A, Sitch S, Hayes D. A large and persistent carbon sink in the world's forests. Science, 2011, 333(6045): 988-993.

[3] Kauppi P E, Mielikäinen K, Kuusela K. Biomass and carbon budget of European forests, 1971 to 1990. Science, 1992, 256(5053): 70-74.

[4] Brown S. Measuring carbon in forests: current status and future challenges. Environmental Pollution, 2002, 116(3): 363-372.

[5] Zhao X, Ma X W, Chen B Y, Shang Y P, Song M L. Challenges toward carbon neutrality in China: strategies and countermeasures. Resources, Conservation and Recycling, 2022, 176: 105959.

[6] Li Y C, Li M Y, Li C, Liu Z Z. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Scientific Reports, 2020, 10(1): 9952.

[7] 曹霖. 基于 Sentinel-2 影像的延庆区森林蓄积量遥感估测研究[D]. 北京: 北京林业大学, 2019.

[8] Zheng Z J, Schmid B, Zeng Y, Schuman M C, Zhao D, Schaepman M E, Morsdorf F. Remotely sensed functional diversity and its association with productivity in a subtropical forest. Remote Sensing of Environment, 2023, 290: 113530.

[9] Santoro M, Beaudoin A, Beer C, Cartus O, Fransson J E S, Hall R J, Pathe C, Schmullius C, Schepaschenko D, Shvidenko A, Thurner M, Wegmüller U. Forest growing stock volume of the Northern Hemisphere: Spatially explicit estimates for 2010 derived from Envisat ASAR. Remote Sensing of Environment, 2015, 168: 316-334.

[10] Zhang H B, Zhu J J, Wang C C, Lin H, Long J P, Zhao L, Fu H Q, Liu Z W. Forest growing stock volume estimation in subtropical mountain areas using PALSAR-2 L-band PolSAR data. Forests, 2019, 10(3): 276.

[11] Maciel D A, Pahlevan N, Barbosa C C F, Martins V S, Smith B, O'Shea R E, Balasubramanian S V, Saranathan A M, Novo E M L M. Towards global long-term water transparency products from the Landsat archive. Remote Sensing of Environment, 2023, 299: 113889.

[12] 郝晴, 黄昌. 森林地上生物量遥感估算研究综述. 植物生态学报, 2023, 47(10): 1356-1374.

[13] Rasel S M M, Chang H C, Ralph T J, Saintilan N, Diti I J. Application of feature selection methods and machine learning algorithms for saltmarsh biomass estimation using Worldview-2 imagery. Geocarto International, 2021, 36(10): 1075-1099.

[14] Luo M, Wang Y F, Xie Y H, Zhou L, Qiao J J, Qiu S Y, Sun Y J. Combination of feature selection and CatBoost for prediction: the first application to the estimation of aboveground biomass. Forests, 2021, 12(2): 216.

[15] 王平. 基于改进 Boruta 算法的森林地上生物量遥感估测研究[D]. 长沙: 中南林业科技大学, 2022.

[16] Foody G M, Boyd D S, Cutler M E J. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. Remote Sensing of Environment, 2003, 85(4): 463-474.

[17] 曹军, 张加龙, 肖庆琳, 王飞平, 韩雪莲, 黄屹杰. 基于随机森林和蒙特卡洛的高山松地上碳储量估测及不确定性分析. 林业科学研究, 2023, 36(5): 131-139.

[18] Gleason C J, Im J. Forest biomass estimation from airborne LiDAR data using machine learning approaches. Remote Sensing of Environment, 2012, 125: 80-91.

[19] Temesgen A, Petri P, Tino J, James M, Jann e H. Towards tree-based systems disturbance monitoring of tropical mosaic landscape using a time series ensemble learning approach. Remote Sensing of Environment, 2023, 299:113876.

[20] Mienye I D, Sun Y X. A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access, 2022, 10: 99129-99149.

[21] Li X Y, Zhang M, Long J P, Lin H. A novel method for estimating spatial distribution of forest above-ground biomass based on multispectral fusion

data and ensemble learning algorithm. Remote Sensing, 2021, 13(19): 3910.
[22] Cho D, Yoo C, Im J, Lee Y, Lee J. Improvement of spatial interpolation accuracy of daily maximum air temperature in urban areas using a stacking ensemble technique. GIScience & Remote Sensing, 2020, 57(5): 633-649.
[23] Breiman L. Stacked regressions. Machine Learning, 1996, 24(1): 49-64.
[24] Yu T C. Optuna Hyperparameters Optimization for Discriminating Life-Threatening Ventricular Arrhythmia Problem[D]. New York: State University of New York at Buffalo, 2023.
[25] Lin L, Zhang J, Zhang N, Shi J C, Chen C. Optimized LightGBM power fingerprint identification based on entropy features. Entropy, 2022, 24(11): 1558.
[26] Sipper M. High per parameter: a large-scale study of hyperparameter tuning for machine learning algorithms. Algorithms, 2022, 15(9): 315.
[27] Chintakindi S, Alsamhan A, Abidi M H, Kumar M P. Annealing of monel 400 alloy using principal component analysis, hyper-parameter optimization, machine learning techniques, and multi-objective particle swarm optimization. International Journal of Computational Intelligence Systems, 2022, 15(1): 18.
[28] 国家林业局. LY/T 2264—2014 立木生物量模型及碳计量参数——杉木. 北京: 中国标准出版社, 2014.
[29] 赵菡. 江西省主要树种不同立地等级的地上生物量与不确定性估计[D]. 北京: 中国林业科学研究院, 2017.
[30] 国家林业局. LY/T 2261—2014 立木生物量模型及碳计量参数——湿地松. 北京: 中国标准出版社, 2014.
[31] 国家林业局. LY/T 2658—2016 立木生物量模型及碳计量参数——栎树. 北京: 中国标准出版社, 2016.
[32] 国家林业局. LY/T 2661—2016 立木生物量模型及碳计量参数——枫香. 北京: 中国标准出版社, 2016.
[33] 国家林业局. LY/T 2660—2016 立木生物量模型及碳计量参数——木荷. 北京: 中国标准出版社, 2016.
[34] Haralick R M, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, 1973, SMC-3(6): 610-621.
[35] Conners R W, Trivedi M M, Harlow C A. Segmentation of a high-resolution urban scene using texture operators. Computer Vision, Graphics, and Image Processing, 1984, 25(3): 273-310.
[36] Sarkar S, Sagan V, Bhadra S, Rhodes K, Pokharel M, Fritschi F B. Soybean seed composition prediction from standing crops using PlanetScope satellite imagery and machine learning. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 204: 257-274.
[37] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273-297.
[38] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. San Francisco California USA: Knowledge Discovery and Data Mining. 2016.
[39] Zhang Y Z, Ma J, Liang S L, Li X S, Liu J D. A stacking ensemble algorithm for improving the biases of forest aboveground biomass estimations from multiple remotely sensed datasets. GIScience & Remote Sensing, 2022, 59(1): 234-249.
[40] Zhou Z H, Wu J X, Tang W. Ensembling neural networks: many could be better than all. Artificial Intelligence, 2002, 137(1/2): 239-263.
[41] Srinivas P, Katarya R. hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. Biomedical Signal Processing and Control, 2022, 73: 103456.
[42] Chen J, Chen J, Liao A P, Cao X, Chen L J, Chen X H, He C Y, Han G, Peng S, Lu M, Zhang W W, Tong X H, Mills J. Global land cover mapping at 30m resolution: a POK-based operational approach. ISPRS Journal of Photogrammetry and Remote Sensing, 2015, 103: 7-27.
[43] 王劲峰, 徐成东. 地理探测器:原理与展望. 地理学报, 2017, 72(1): 116-134.
[44] Li D, Gu X F, Pang Y, Chen B W, Liu L X. Estimation of forest aboveground biomass and leaf area index based on digital aerial photograph data in Northeast China. Forests, 2018, 9(5): 275.
[45] Lu D S, Chen Q, Wang G X, Moran E, Batistella M, Zhang M Z, Vaglio Laurin G, Saah D. Aboveground forest biomass estimation with landsat and LiDAR data and uncertainty analysis of the estimates. International Journal of Forestry Research, 2012, 2012(1): 436537.
[46] Walter J D C, Edwards J, McDonald G, Kuchel H. Estimating biomass and canopy height with LiDAR for field crop breeding. Frontiers in Plant Science, 2019, 10: 1145.
[47] Tian L, Qu Y H, Qi J B. Estimation of forest LAI using discrete airborne LiDAR: a review. Remote Sensing, 2021, 13(12): 2408.
[48] 唐君, 陆应诚, 焦俊男, 刘建强, 胡连波, 丁静, 邢前国, 王福涛, 宋庆君, 陈艳拢, 田礼乔, 王心源, 刘锦超. 中国近海绿潮生物量的卫星光学遥感估算. 遥感学报, 2023, 27(11): 2484-2498.
[49] 李海奎, 雷渊才, 曾伟生. 基于森林清查资料的中国森林植被碳储量. 林业科学, 2011, 47(7): 7-12.
[50] Wang Y F, Yue T X, Lei Y C, Du Z P, Zhao M W. Uncertainty of forest biomass carbon patterns simulation on provincial scale: a case study in Jiangxi Province, China. Journal of Geographical Sciences, 2016, 26(5): 568-584.
[51] 廖凯涛, 齐述华, 王成, 王点. 结合 GLAS 和 TM 卫星数据的江西省森林高度和生物量制图. 遥感技术与应用, 2018, 33(4): 713-720.
[52] 郑嘉豪, 孙超, 林昀, 李璐, 刘永超. 基于 Landsat 像元级时间序列的海岸带盐沼植被分类. 遥感学报, 2023, 27(6): 1400-1413.
[53] Stelmaszczuk-Górska M, Rodriguez-Veiga P, Ackermann N, Thiel C, Balzter H, Schmullius C. Non-parametric retrieval of aboveground biomass in Siberian boreal forests with ALOS PALSAR interferometric coherence and backscatter intensity. Journal of Imaging, 2015, 2(1): 1.
[54] 谭雨欣, 田义超, 黄卓梅, 张强, 陶进, 刘虹秀, 杨永伟, 张亚丽, 林俊良, 邓静雯. 北部湾茅尾海无瓣海桑红树林地上生物量反演——基于 XGBoost 机器学习算法. 生态学报, 2023, 43(11): 4674-4688.
[55] 张磊. 基于 Landsat 时间序列影像的区域不透水面提取研究[D]. 武汉: 武汉大学, 2017.
[56] Lu D S. Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon. International Journal of Remote Sensing, 2005, 26(12): 2509-2525.
[57] Su Y J, Guo Q H, Xue B L, Hu T Y, Alvarez O, Tao S L, Fang J Y. Spatial distribution of forest aboveground biomass in China: estimation through combination of spaceborne lidar, optical imagery, and forest inventory data. Remote Sensing of Environment, 2016, 173: 187-199.