

基于XGBoost和SHAP的海滩波浪爬高预测研究

张建^{1,2}, 丁佩^{1,2}, 刘楷操^{1,2}, 路川藤³

(1. 珠海市规划设计研究院, 广东 珠海 519000; 2. 广东省滨海地区防灾减灾工程技术研究中心, 广东 珠海 519000; 3. 南京水利科学研究院, 江苏 南京 210029)

摘 要: 海滩波浪爬高预测是海岸侵蚀防护和防灾减灾的关键技术支撑。针对现有经验公式在精确度、泛化性等方面的不足, 将极限梯度提升模型XGBoost引入到波浪爬高预测中, 利用1 400多个来自实验室和现场观测的海滩波浪爬高数据, 通过贝叶斯优化进行超参数调整, 建立基于XGBoost的海滩波浪爬高预测模型。此外, 还将可解释机器学习框架SHAP与XGBoost模型结合, 以挖掘波浪爬高预测结果的关键特征。评估结果表明: XGBoost模型的决定系数为0.957, 均方根误差为0.384 m, 显著优于其他经验公式, 整体预测可靠稳定; SHAP分析也表明XGBoost模型的预测趋势符合真实走向, 且Iribarren数在海滩波浪爬高预测中起着关键作用。

关键词: 机器学习; 波浪爬高; 极限梯度提升模型; 贝叶斯优化; 可解释机器学习框架

中图分类号: P731.33 **文献标识码:** A **文章编号:** 1003-0239(2025)02-0001-08

0 引言

波浪爬高即波浪在斜坡破碎后在斜坡上引起的水位振荡最大高程, 一直都是沿海工程师和管理人员极感兴趣的研究领域^[1]。在极端风暴事件中, 波浪爬高已成为沿海洪水的重要因素, 在极端条件下波浪会越过沙丘顶部形成越浪, 威胁近岸居民的生命财产安全。波浪爬高在沉积物运输中同样也发挥关键作用, 为海滩侵蚀防护、防灾减灾、海岸工程设计提供重要依据^[2-3]。

通常情况下, 不规则波的爬高统计采用一段时间内超过2%的波浪爬高($R_{2\%}$)值来表示。在近岸破碎带发生非线性过程以及水下沙坝连续变化的情况下, 波浪传播难以预测, 因此, 除物理实验外, 研究人员通常采用包含海滩剖面坡度、深水有效波高和谱峰周期等参数的经验公式(如常用的Stockdon公式^[4])来评估波浪爬高。然而, 经验公式由于其非线性过程无法转化为简单的预测因子导致存在误差^[5], 因此预测结果存在相当大的不确定

性。一些学者利用具有物理意义的数值模型模拟波浪爬高, 其中Xbeach数值模型有着较广泛的应用^[6]。在当今计算资源更易获取的时代, 利用机器学习与深度学习提高波浪爬高预测能力、减少预测的不确定性已成为一种较为先进的方法。Abolfathi等^[7]利用M5决策树算法提高了预测波浪爬高精度; Tarwidi等^[8]采用了400多个基于实验室观测的波浪爬高数据作为训练集, 构建了极限梯度提升(eXtreme Gradient Boosting, XGBoost)模型, 用于预测斜坡海滩的波浪爬高, 这个研究展现了XGBoost模型在波浪爬升预测中的良好应用; Beuzen等^[9]将数据驱动的高斯过程方法与数值模型结合使用, 开发出更为复杂和准确的集合预测方法。

本研究旨在利用XGBoost方法开发一种基于机器学习的方法, 准确预测各种海岸坡度和波幅下的波浪爬高, 通过对大型数据库进行训练, 利用贝叶斯优化进行XGBoost超参数调优, 将优化后的模型与典型公式的波浪爬高预测进行对比, 以验证XGBoost模型的预测能力。此外, 将可解释机器学习

收稿日期: 2024-02-04。

基金项目: 水利部重大科技项目(SKS-2022087)。

作者简介: 张建(1983-), 男, 高级工程师, 硕士, 主要从事城市防洪(潮)、河道整治等设计及研究工作。E-mail: 121939233@qq.com

习框架 (SHapley Additive exPlanation, SHAP) 与模型结合,挖掘波浪爬高预测结果的关键特征,增强算法的可解释性。

1 数据来源

数据来自波浪爬高数据集(网址: <https://coastalhub.science/data>)^[5],它整合了多个公开发表的野外工作成果,包含了实验室与现场观测数据,尤其是来自多个砾石海滩与砂质海滩的现场观测记录。通过数据清洗,删除无波浪爬高的数据,最终选择 1 432 组数据进行应用。表 1 为数据集描述,其中离岸显著波高(H_s)和波谱峰周期(T_p)分别为 0.02~7.17 m 和 0.81~23.68 s,坡度($\tan\beta$)为 0.01~0.29,泥沙中值粒径(D_{50})为 0~50 mm。

表 1 数据集描述

Tab.1 Description of the database

| 数据描述 | $R_{2\%}/\text{m}$ | H_s/m | T_p/s | $\tan\beta$ | D_{50}/mm |
|-------|--------------------|----------------|----------------|-------------|--------------------|
| 平均值 | 2.27 | 1.85 | 9.27 | 0.12 | 9.45 |
| 标准差 | 1.78 | 1.32 | 3.56 | 0.06 | 17.36 |
| 最小 | 0.03 | 0.02 | 0.81 | 0.01 | 0 |
| 50% 值 | 1.85 | 1.83 | 10.00 | 0.11 | 0.90 |
| 最大 | 12.67 | 7.17 | 23.68 | 0.29 | 50.00 |

2 方法

2.1 模型基本原理

2.1.1 XGBoost 模型

梯度提升决策树 (Gradient Boosting Decision Tree, GBDT) 是一种集成学习方法。它通过结合多个弱学习器,基于梯度下降思想,在每一轮通过最小化损失函数的梯度方向来纠正上一轮模型的误差,逐步提升整体性能。XGBoost^[10]是在 GBDT 的基础上改进而来,通过构建多个决策树模型来实现梯度提升,每个树模型都是一个弱学习器,通过累加多个树的输出得到最终预测结果。相较于 GBDT, XGBoost 的算法主要在损失函数上进行了改进,通过二阶泰勒展开来提高求解效率,同时引入了正则化项来控制模型的复杂度,从而防止了过

拟合,使得 XGBoost 有更好的计算效率与泛化能力,因此在机器学习竞赛与科学研究及应用中有着广泛应用。XGBoost 算法的目标函数公式 Obj 由自身的训练损失函数 $\ell(y_i, \hat{y}_i)$ 和正则化惩罚项 $\Omega(f_k)$ 相加而成。公式为:

$$\text{Obj} = \sum_{i=1}^n \ell(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

式中: y_i 表示 i 真实值, \hat{y}_i 表示第 i 个预测值; f_k 表示模型的第 k 个树。

2.1.2 贝叶斯优化

贝叶斯优化 (Bayesian Optimization, BO) 是一种用于优化黑盒函数的迭代算法,其中目标函数形式未知。贝叶斯优化通过建立高斯过程模型来估计函数值的分布,在参数空间中智能地选择采样点,观察对应的函数值,并通过不断更新先验分布来逼近最优解。该算法通过平衡探索和利用,可以动态调整采样策略,最终在有限预算内找到全局最小值或最大值,因此被广泛应用于超参数的调整优化中。

2.1.3 SHAP 模型

机器学习在回归预测方面具有很大的潜力,其黑盒模型的特性也构成了学习的障碍之一。在实际工作中,为评价机器学习的黑盒模型中各变量对模型预测的贡献,通常采用 Shapley 值分析各特征的影响程度与方向。Shapley 值源于博弈论,主要用于量化每个特征对模型预测的贡献。SHAP 框架也基于此原理,通过计算每个特征加入模型时的 Shapley 值,来反映每个特征的贡献程度与正负性^[11]。SHAP 将 Shapley 值的表现形式转化为可加的特征归因方法。计算公式为:

$$y_j = y_{\text{base}} + f(x_{j1}) + f(x_{j2}) + f(x_{j3}) + \cdots f(x_{jk}) \quad (2)$$

式中: y_j 为模型预测值; $f(x_{jk})$ 表示第 j 个样本中第 k 个特征对最终预测的贡献 Shapley 值。

2.2 特征工程

Hunt^[12]研究了倾斜的不透水结构中的波浪驱动爬高,并提出了波浪爬升与 Iribarren 数之间的直接经验关系:

$$\frac{R_{2\%}}{H_s} = K\xi_0 \quad (3)$$

式中: K 为常数; ξ_0 是 Iribarren 数, 用于表示海滩的地貌动力学特征与近海波浪参数的结合参数, 是一个广泛应用在波浪爬高预测中的无量纲参数^[13-15]。Iribarren 数定义如下:

$$\xi_0 = \frac{\tan\beta}{\sqrt{H_s/L_p}} \quad (4)$$

$$L_p = \frac{gT_p^2}{2\pi} \quad (5)$$

波陡 (H_0/L_0) 是波浪爬升的一个重要因素^[16], 底部摩擦力也同样不可忽视^[17]。Teng 等^[18]在不同坡度的光滑和粗糙的平面海滩上进行了波浪爬高实验, 实验结果表明, 河床粗糙度显著降低了缓坡的最大爬高, 而在陡坡上则可忽略不计, 研究证明了海滩坡度和底部摩擦力在波浪爬高过程中的重要性。

综上, 出于建模目的, 将波浪爬高转为无量纲化, 建立以下函数:

$$\frac{R_{2\%}}{H_0} = f(\xi_0, \tan\beta, H_s/L_p, D_{50}/H_s) \quad (6)$$

2.3 模型训练与调优

XGBoost 模型需经过训练与调参, 才能得到更好的精度和泛化能力, 主要调参流程见图 1, 超参数组合见表 2。在整个过程中, 通过调整多个超参数来提高模型精度和防止过度拟合。 \max_depth 用于控制树的深度, 防止模型由于太专注于训练数据而影响了泛化能力。学习率 ($learning_rate$) 也是关键参数, 用于防止过拟合, 但需注意过小的学习率可能导致收敛速度较慢, 需要更多的树来构建集成, 通常设置学习率为 0.01~0.30。为了避免单个树过于复杂导致过拟合, 使用正则化参数对树的复杂性进行惩罚, 即 λ 和 α , 其中 λ 用于控制叶节点中参数的 L2 正则化项的权重, 而 α 用于控制 L1 正则化项的权重, 通过最小化训练数据的误差来严格执行两个参数的正则化, 有助于提高模型的泛化性。除了上述超参数外, 其他关键参数还包括 $subsample$ (子样本比例)、 \min_child_weight (最小叶子节点样本权重和)、 $n_estimators$ (迭代次数)、 γ (最小划分损失减少值) 以及 $colsample_bytree$ (每棵树的特征采样比例)。通过精心调整这些参

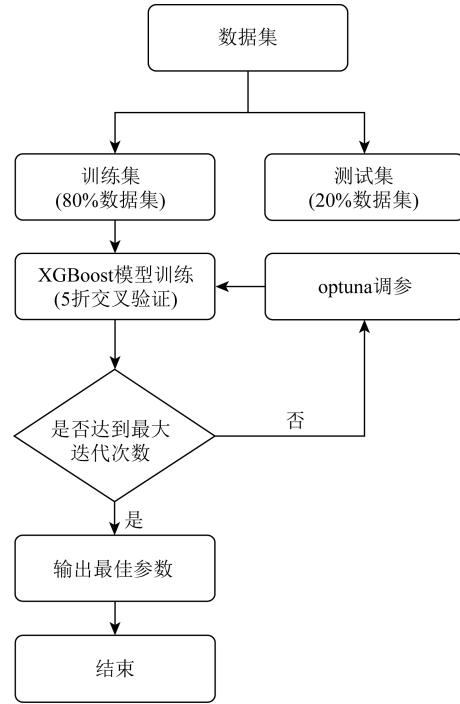


图 1 XGBoost 模型调参流程图

Fig.1 XGBoost model tuning flowchart

表 2 XGBoost 超参数组合

Tab.2 XGBoost hyperparameter combinations

| 名称 | 说明 | 范围 |
|------------------|-------------|--------------|
| lambda | L2 正则化参数 | (1e-8, 1.0) |
| alpha | L1 正则化参数 | (1e-8, 1.0) |
| max_depth | 树的最大深度 | (3, 25) |
| gamma | 最小划分损失减少值 | (1e-8, 1.0) |
| subsample | 子样本比例 | (0.5, 1.0) |
| colsample_bytree | 每棵树的特征采样比例 | (0.5, 1.0) |
| min_child_weight | 最小叶子节点样本权重和 | (1, 10) |
| n_estimators | 迭代次数 | (50, 500) |
| learning_rate | 学习率 | (0.01, 0.30) |

数, 可以更有效地优化 XGBoost 模型, 提高其在训练和测试数据上的性能。

XGBoost 模型的超参数调优过程采用了贝叶斯优化方法。它是一种自动调整超参数的方法, 与传统的网格搜索或随机搜索相比, 贝叶斯优化更为智能和高效。贝叶斯优化的核心思想是通过构建概率模型和代理模型 (通常是高斯过程), 来近似表

现目标函数和参数之间的关系,有助于在参数空间中智能地选择接近最优的参数组合,避免对整个参数空间的不必要遍历。Optuna 是 Python 常用的贝叶斯优化库^[19], 本文将用于 XGBoost 模型关键参数和取值范围的优化。为提升模型泛化能力,使用 80% 的数据进行随机训练,保留 20% 用于测试。此外,我们还引入了 K 折交叉验证($K=5$)进行训练。 K 折交叉验证是将数据分成 5 个子集,在 5 次训练中,每次使用 4 个子集进行训练,剩余 1 个子集用于验证。这样的循环训练重复 5 次,可确保模型在不同子集上

的性能评估,提高模型对不同数据分布的适应性。

2.4 经验公式

本节介绍几种常用的波浪爬高公式,它们具有一定的精确性,得到了广泛应用。Holman^[13] 将 Iribarren 数应用到天然海滩中,得出式(7),虽然该公式提出较早,但仍有较高的精确性。公式为:

$$R_{2\%} = 0.83 \tan \beta \sqrt{H_s L_p} + 0.2 H_s \quad (7)$$

Stockdon 等^[4]结合多次测量,提出了目前最广

$$R_{2\%} = 0.043 (H_s L_p)^{0.5} \quad (\xi_0 < 0.3)$$

$$R_{2\%} = 1.1 \left[0.35 \tan \beta (H_s L_p)^{0.5} + \frac{H_s L_p (0.563 \tan \beta^2 + 0.004)^{0.5}}{2} \right] \quad (0.3 < \xi_0 < 3.5) \quad (8)$$

泛接受的波浪爬高的公式:

Vousdoukas 等^[14]利用葡萄牙一个大潮汐海滩上的观测数据,提出了 $R_{2\%}$ 的计算方法:

$$R_{2\%} = 0.53 \tan \beta \sqrt{H_s L_p} + 0.58 \tan \beta H_s + 0.45 \quad (9)$$

Atkinson 等^[15]根据 6 个现场模型和 1 个大规模实验室 $R_{2\%}$ 数据衍生模型的最佳拟合,开发了一个新的 $R_{2\%}$ 公式,同样也使用 Hunt^[12] 公式进行建模。新公式为:

$$R_{2\%} = 0.92 \tan \beta \sqrt{H_s L_p} + 0.16 H_s \quad (10)$$

Power 等^[20]通过基因表达式编程进行显示公式编程,提取出一个包含波陡、无量纲糙率和坡度的公式,精度较为良好,其收集使用的数据库也是本文数据库的主要来源。该公式较为繁长,受篇幅限制,本文不予展示,具体可见文献[20]。

2.5 评价指标

利用均方根误差 (Root Mean Square Error, RMSE) 和 决定系数 (R^2) 两个参数,将各经验公式或 XGBoost 的预测结果与实测结果进行评估。计算公式分别为:

$$\text{均方根误差: } E_{\text{RMS}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

$$\text{决定系数: } R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (12)$$

式中: n 是样本数量; y_i 是观察值, \hat{y}_i 是模型的预测

值, \bar{y} 为均值。一般而言, RMSE 用于测量模型预测误差的标准偏差, RMSE 越低, 表示模型预测越精确; R^2 表示模型能够解释目标变量方差的比例, R^2 越接近 1, 表示模型越能够解释目标变量的变异性, 越接近 0, 表示模型的解释能力越弱。

3 结果及讨论

3.1 模型比较

模型的应用能力主要体现在其处理未经训练的数据集的性能表现上。通过图 2 可以明显观察到, XGBoost 模型在测试集中展现出卓越的应用能力。具体而言, 模型在测试中的预测值与实测值的 RMSE 为 0.389 m, R^2 为 0.957 (见图 2a)。XGBoost 模型的出色表现不仅在测试集上得到验证, 而且在整个数据集中, 模型的 RMSE 和 R^2 值为 0.225 m 和 0.984 (见图 2b), 显示了其在广泛数据范围内的稳健性。通过 D_{50} 小于 2 mm (砂质海滩) 和大于 2 mm (砾石海滩) 的区分, 可以进一步观察到, 在两种海滩场景下, XGBoost 模型同样表现出色, 巩固了其在不同应用场景中的鲁棒性。

从表 3 可以明显观察到, 在 $D_{50} < 2$ mm 的数据集中, Stockdon、Holman、Vousdoukas 以及 Atkinson 公式都展现出良好的性能, 其 R^2 值维持在 0.6 以上, 相比之下, Power 公式的性能相对较差。然而, 在 $D_{50} > 2$ mm 的数据集中, 这主要是由于 Power 公式的 R^2 值显著提升, 而其他方程的适用性却相对较

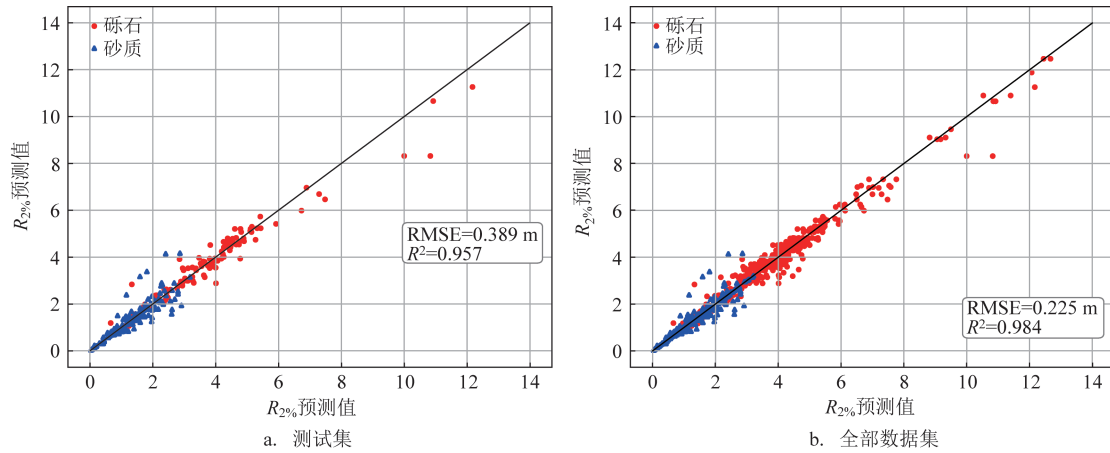


图2 模型计算值与实测值比较

Fig.2 Comparison between model calculated and measured values

表3 模型与其他经验公式比较

Tab.3 Comparison with other empirical formulas

| 模型/经验公式 | $D_{50} < 2$ mm 数据集 | | $D_{50} > 2$ mm 数据集 | | 数据集 | |
|-------------------------|---------------------|-------|---------------------|--------|--------|-------|
| | RMSE/m | R^2 | RMSE/m | R^2 | RMSE/m | R^2 |
| Holman公式(式7) | 0.430 | 0.691 | 1.436 | 0.356 | 1.035 | 0.660 |
| Stockdon公式(式8) | 0.433 | 0.687 | 1.668 | 0.130 | 1.189 | 0.551 |
| Vousdoukas公式(式9) | 0.451 | 0.660 | 1.873 | -0.097 | 1.329 | 0.440 |
| Atkinson公式(式10) | 0.439 | 0.678 | 1.464 | 0.330 | 1.056 | 0.646 |
| Power公式 ^[20] | 0.701 | 0.179 | 0.811 | 0.795 | 0.754 | 0.819 |
| XGBoost | 0.164 | 0.955 | 0.279 | 0.976 | 0.225 | 0.984 |

低,这主要是由于Power公式特别适用于砾石海滩的情境,而该数据集并不完全平衡。各经验公式的侧重均有所不同,但XGBoost模型在这两种数据集中都表现出较高的应用性和精度。

在这项研究中,采用了不同文献的公开数据集,这些数据的测量方式和情境不一,数据来源于实验或现场,因此并非完全可靠,这导致数据库可靠性降低,复杂性增加。数据质量是模型验证的重要因素,影响公式/模型的稳健性。另外,数据的分布也会带来不确定性,通过表1可以看出,该数据集并不均衡,数据主要集中在 D_{50} 较小的区域,而多数经验公式采用的数据也来自这一区域。理想做法是如果模型在某些数据稀疏的区域表现不佳,需要考虑收集更多的相关数据来扩展训练集。这有助于确保模型能够对未见数据点做出准确预测,而不仅仅在已知数据点附近表现良好。应用XGBoost

模型可以进行准确预测,预测误差很小,这可能是由于防止过拟合的超参数(例如 subsample, min_child_weight 和 lambda 等)发挥了很大作用。这些超参数创建了一个更通用的模型,它对数据集中特定数据点的依赖性较小。此模型在未见数据上表现同样出色,而不仅仅是在训练集上。此外,本文还在XGBoost模型训练中采用了交叉验证,这为模型性能提供了更稳健和可靠的评估,减少了过拟合的风险。

3.2 模型预测结果解释

本文利用SHAP框架对XGBoost模型的特征重要性进行分析。一个特征的Shapley值越高,预测值相对受影响程度越高。通过Shapley值的平均绝对值可以得出每个特征的相对重要性。从图3可以看出, ξ_0 在模型中最重要,次之为 D_{50}/H_s 、 $\tan\beta$ 、

H_s/L_p 。在SHAP摘要图中(见图4),如果基于x轴的Shapley值为负,表示它具有负向贡献,如果为正,则表示具有正向影响,值越大贡献越明显。此外,红色代表大的特征值,蓝色代表小的特征值。通过图4可以发现,随着 ξ_0 、 $\tan\beta$ 值的增大,模型的Shapley值增大,两特征值分别与Shapley值成正相关关系; H_s/L_p 与模型的Shapley值呈负相关关系;随着相对粗糙度(D_{50}/H_s)的增大,Shapley值减小,负向作用明显,然而该趋势为非线性变化,其一方面可能来自于 D_{50} 取值的不确定性,另一方面则可能因为 D_{50}/H_s 与其他特征交互作用导致其对模型的贡献呈现非线性变化,如在大坡度情况下相对粗糙度对爬高的贡献可以忽略^[18]。

图5为单个/两个特征的依赖作用,横轴表示特征值,左侧纵轴为Shapley值,散点代表每个样本,颜色越红代表所交互的特征在该样本上的值越大。从单个特征来看, ξ_0 增大,模型的Shapley值增大

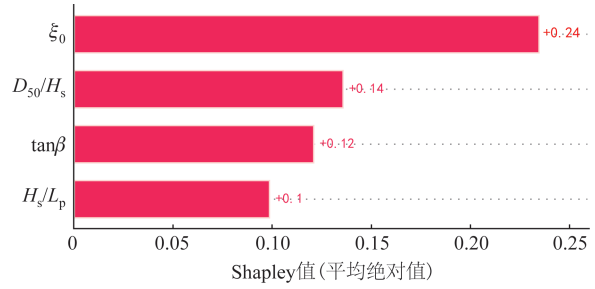


图3 SHAP的特征重要性图

Fig.3 Importance of SHAP's features

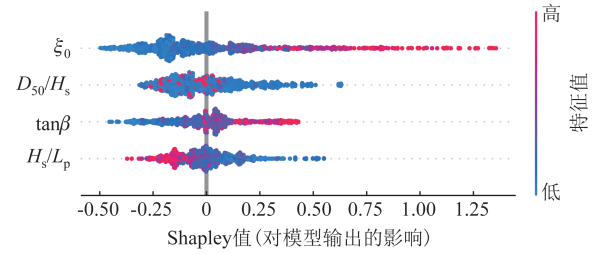


图4 SHAP摘要图

Fig.4 SHAP summary plot

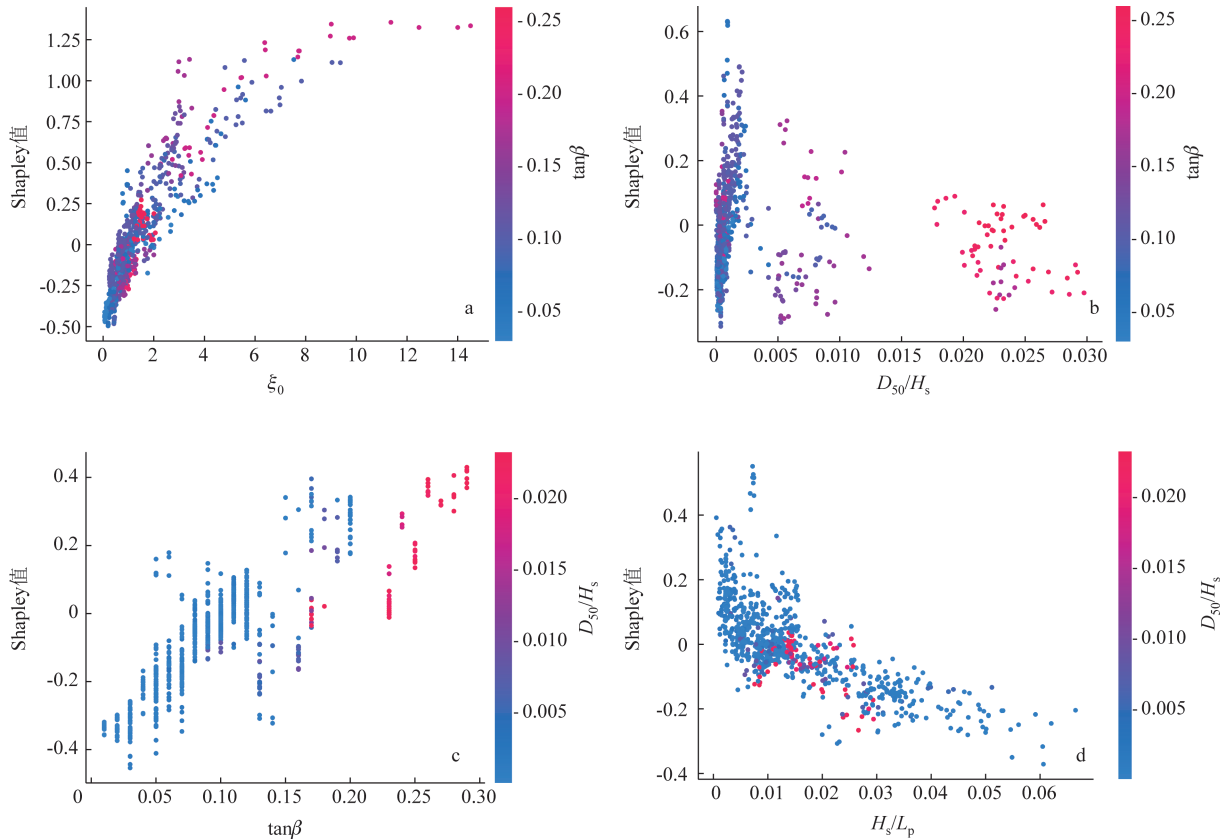


图5 部分特征依赖图

Fig.5 Partial Dependence Plot

(见图5a),对模型的正向影响越大。 D_{50}/H_s 增加,Shapley值趋于降低(见图5b),到达一定程度后,Shapley值的变化很小,这一观察结果与物理实践一致,即当粗糙度增加时,波浪爬高减小,当坡度到达一定程度时,粗糙度的影响程度较小。另外,在 D_{50}/H_s 为0附近观察到较大的垂直变化,即小的相对粗糙度对模型的贡献是不稳定的,这种现象可能与泥沙 D_{50} 采样的不确定性、泥沙不均匀性等相关。 $\tan\beta$ 与Shapley值成正相关关系(见图5c), $\tan\beta$ 值增大,其正向影响增大,较小的坡度为波浪破碎提供了充足的空间与时间,从而耗散更多能量,故 $\tan\beta$ 与波浪爬高成正比。 H_s/L_p 与模型的Shapley值呈负相关关系(见图5d),波陡常用作波浪破碎指标,波陡越大,波浪越容易破碎,能量耗散越大,波浪爬高衰减,故波陡与模型的Shapley值呈现负相关。

在图5中,选取与该特征交互作用最明显的一个其他特征作为颜色轴,从两个特征的交互影响作用来看, $\xi_0/H_s/L_p$ 与其他特征的交互并不明显。当 $D_{50}/H_s > 0.015$ 时,在较大的坡度($\tan\beta$)交互影响下,随着 D_{50}/H_s 的增大,Shapley值基本保持稳定(见图5b);然而在 D_{50}/H_s 较大的情况下,随着 $\tan\beta$ 的增大,Shapley值依然有正向增加的趋势(见图5c)。由此可以看出,在较大的坡度下, $\tan\beta$ 的影响程度较 D_{50}/H_s 更加显著, D_{50}/H_s 对模型的贡献较小。

4 结论

机器学习依靠其在处理大规模和复杂数据方面的稳健性,近年来已成为波浪模型开发的重要方法。本文采用基于极限梯度提升(XGBoost)的机器学习方法,用于预测海滩上的波浪爬高。利用1400多个来自实验室和现场观测的海滩波浪爬高数据作为训练集来构建XGBoost模型。通过贝叶斯优化进行超参数调整,得到一个优化的XGBoost模型,同时采用SHAP框架分析XGBoost的特征。结论如下:

①在海滩波浪爬高预测中,XGBoost模型在测试集的相关系数(R^2)为0.957,均方根误差(RMSE)为0.384 m,在整个数据集中,模型的RMSE和 R^2 值为0.225 m和0.984,模型表现出较好的精度与泛化能力。

②经验公式通常局限于特定范围的坡度与粒径内,而XGBoost模型适用于更宽广的海滩坡度、波浪条件和粗糙系数范围。在预测波浪爬高方面,优化后的XGBoost方法是现有经验公式和传统数值模型的可行替代方案。

③将可解释机器学习框架(SHAP)与XGBoost模型结合,挖掘波浪爬高预测结果的关键特征,增强算法的可解释性,从而提高了波浪爬高预测结果的可信度。Shapley值表明,该模型中不同特征的重要性依次为 ξ_0 、 D_{50}/H_s 、 $\tan\beta$ 、 H_s/L_p ,其中 ξ_0 、 $\tan\beta$ 值增大,对模型的正向贡献也增大,而 H_s/L_p 越大,对模型的负向贡献也越大;相对粗糙度 D_{50}/H_s 增大,对模型的负向贡献也越大,但通过分析特征交互作用后发现,在较大的 $\tan\beta$ 的情况下,相对粗糙度对模型的贡献不明显。

参考文献:

- [1] 尹航. 视频图像海滩动力地貌监测与信息提取方法研究[D]. 厦门: 自然资源部第三海洋研究所, 2022.
YIN H. Beach dynamic geomorphology monitoring and information extraction based on video imagery[D]. Xiamen: Third Institute of Oceanography, MNR, 2022.
- [2] 王广生, 童林龙, 罗梦岩, 等. 贝宁海滩上波浪传播演变特性研究[J]. 河海大学学报(自然科学版), 2023, 51(6): 123-129.
WANG G S, TONG L L, LUO M Y, et al. Study on wave propagation and evolution characteristics over a beach in Benin[J]. Journal of Hohai University (Natural Sciences), 2023, 51(6): 123-129.
- [3] 邱星, 董玉祥. 海岸沙丘对风暴潮响应研究进展与展望[J]. 地球科学进展, 2022, 37(8): 811-821.
QIU X, DONG Y X. Research progress and prospect of the response of coastal dunes to storm surge[J]. Advances in Earth Science, 2022, 37(8): 811-821.
- [4] STOCKDON H F, HOLMAN R A, HOWD P A, et al. Empirical parameterization of setup, swash, and runup[J]. Coastal Engineering, 2006, 53(7): 573-588.
- [5] DA SILVA P G, COCO G, GARNIER R, et al. On the prediction of runup, setup and swash on beaches[J]. Earth-Science Reviews, 2020, 204: 103148.
- [6] LERMA A N, PEDREROS R, ROBINET A, et al. Simulating wave setup and runup during storm conditions on a complex barred beach [J]. Coastal Engineering, 2017, 123: 29-41.
- [7] ABOLFATHI S, YEGANEH-BAKHTIARY A, HAMZE-ZIABARI S M, et al. Wave runup prediction using M5' model tree algorithm [J]. Ocean Engineering, 2016, 112: 76-81.

- [8] TARWIDI D, PUDJAPRASETYA S R, ADYTIA D, et al. An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach[J]. *MethodsX*, 2023, 10: 102119.
- [9] BEUZEN T, GOLDSTEIN E B, SPLINTER K D. Ensemble models from machine learning: an example of wave runup and coastal dune erosion[J]. *Natural Hazards and Earth System Sciences*, 2019, 19(10): 2295-2309.
- [10] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: Association for Computing Machinery, 2016: 785-794.
- [11] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017: 4768-4777.
- [12] HUNT JR I A. Design of seawalls and breakwaters[J]. *Journal of the Waterways and Harbors Division*, 1959, 85(3): 123-152.
- [13] HOLMAN R A. Extreme value statistics for wave run-up on a natural beach[J]. *Coastal Engineering*, 1986, 9(6): 527-544.
- [14] VOUSDOKAS M I, WZIATEK D, ALMEIDA L P. Coastal vulnerability assessment based on video wave run-up observations at a mesotidal, steep-sloped beach[J]. *Ocean Dynamics*, 2012, 62(1): 123-137.
- [15] ATKINSON A L, POWER H E, MOURA T, et al. Assessment of runup predictions by empirical models on non-truncated beaches on the south-east Australian coast[J]. *Coastal Engineering*, 2017, 119: 15-31.
- [16] DIWEDAR A I. Investigating the effect of wave parameters on wave runup[J]. *Alexandria Engineering Journal*, 2016, 55(1): 627-633.
- [17] WU D H, LIU H J. Effects of the bed roughness and beach slope on the non-breaking solitary wave runup height[J]. *Coastal Engineering*, 2022, 174: 104122.
- [18] TENG M H, FENG K L, LIAO T I. Experimental study on long wave run-up on plane beaches[C]//*The Tenth International Offshore and Polar Engineering Conference*. Seattle: OnePetro, 2000.
- [19] AKIBA T, SANO S, YANASE T, et al. Optuna: A next-generation hyperparameter optimization framework[C]//*Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage: Association for Computing Machinery, 2019: 2623-2631.
- [20] POWER H E, GHARABAGHI B, BONAKDARI H, et al. Prediction of wave runup on beaches using Gene-Expression Programming and empirical relationships[J]. *Coastal Engineering*, 2019, 144: 47-61.

A study on beach wave run-up prediction based on XGBoost and SHAP

ZHANG Jian^{1,2}, DING Pei^{1,2}, LIU Kaicao^{1,2}, LU Chuanteng³

(1. Zhuhai Institute of Urban Planning & Design, Zhuhai 519000, China; 2. Guangdong Coastal Area Disaster Prevention and Mitigation Engineering Technology Research Center, Zhuhai 519000, China; 3. Nanjing Hydraulic Research Institute, Nanjing 210029, China)

Abstract: Beach wave run-up prediction is a key technical support for coastal erosion protection, disaster prevention and mitigation. In view of the shortcomings of the existing empirical formulas in terms of accuracy and generalization, the XGBoost model is introduced into wave run-up prediction, and more than 1 400 laboratory and field observations of beach wave run-up are used as a dataset, and hyperparameter tuning is carried out by using Bayesian optimization, which in turn establishes an XGBoost-based wave run-up prediction model. The XGBoost model is used to predict beach wave height, and SHAP, an interpretable machine learning framework, is combined with the XGBoost model to explore the key features of the wave height prediction results. The evaluation results show that the R-squared of the XGBoost model is 0.957, and the root-mean-square error is 0.384 m, which is significantly better than other empirical formulas, and the overall prediction is reliable and stable, meanwhile SHAP shows that the XGBoost model predicted trend is in line with the true value direction and Iribarren number plays a key role in beach wave run-up prediction.

Key words: machine learning; wave run-up; XGBoost; bayesian optimization; SHAP