胡原野, 王收军, 陈松贵, 等. 基于随机森林的复坡堤越浪量预测研究[J]. 海洋学报, 2021, 43(10): 106–114, doi:10.12284/hyxb2021133
Hu Yuanye, Wang Shoujun, Chen Songgui, et al. Overtopping prediction for composite slope breakwater based on random forest method[J]. Haiyang Xuebao, 2021, 43(10): 106–114, doi:10.12284/hyxb2021133

基于随机森林的复坡堤越浪量预测研究

胡原野[1,2]，王收军[1]，陈松贵[2*]，柳叶[2]，王家伟[1,2]，田昀艳[1,2]

(1. 天津理工大学 机电工程国家级实验教学示范中心, 天津 300384; 2. 交通运输部天津水运工程科学研究院 港口水工建筑技术国家工程实验室, 天津 300456)

摘要: 针对复坡堤越浪量的计算问题, 提出了采用随机森林算法预测越浪量的方法。首先, 通过对欧洲 CLASH 数据集进行筛选, 挑选出符合复坡堤越浪量预测的数据; 其次, 对数据做无量纲化处理, 建立以随机森林为基础的复坡堤越浪量预测模型, 并通过网格搜索 (GridSearchCV) 方法对模型进行调参以改善模型的性能; 最后, 利用决定系数R^2来评估模型的精度, 并将随机森林模型与集成神经网络模型做了预测能力的对比, 同时还给出了随机森林模型各个特征参数对预测精度的重要性。结果显示, 随机森林模型的决定系数为 92.7%, 集成神经网络模型的决定系数为 87.7%, 表明随机森林模型对越浪量具有更强的学习和预测能力。通过对特征重要性的分析, 墙顶高程对模型预测精度的影响最大, 堤顶高程次之, 堤脚宽度影响最小。

关键词: 随机森林; 越浪量; 复坡堤; 决定系数; 特征重要性; 预测

中图分类号: U661　　**文献标志码:** A　　**文章编号:** 0253−4193(2021)10−0106−09

1　引言

防波堤作为港口建设中一种重要的水工建筑, 对保护堤后建筑起着重要的作用。越浪量是指波浪越过防波堤的水量, 通常用单位宽度上每秒水体越过防波堤的水量来度量。越浪量是防波堤设计的重要指标, 对堤后结构物和堤面的安全有直接的影响。复坡堤是最为常见的防波堤类型之一, 相比于单坡堤, 其结构更为复杂, 越浪量的计算更为困难; 且目前国内尚无规范可循。本文提出了一种有效精确的复坡堤越浪量估算方法, 对防波堤设计及提高防波堤安全性具有重要的意义。

国内外学者在越浪量方面的研究都做了很多的工作。对越浪量估算方法的研究大体分为 3 类: 经验公式法、数值模拟法和机器学习法。经验公式法主要是通过建立实验模型, 对实验数据进行分析, 然后总结出经验公式; 数值模拟法利用计算机建立研究模型, 结合有限元, 通过数值计算的方法实现对问题的研究; 机器学习法将训练样本数据植入到计算机中, 通过机器学习算法来模拟人类学习的过程, 以此来实现对新样本的预测。王红等[1]通过物理实验分析了单坡堤上不规则波越浪量的相关因子, 并由建筑物形态和波浪特征来确定越浪量, 其成果被《港口与航道水文规范》[2]采纳; 范红霞[3]通过搜集多种防波堤类型资料, 建立水槽物理实验, 分析了各影响因素对越浪量的影响, 并给出了一种计算越浪量的方法。陈国平等[4]通过物理实验分析了不规则波作用下的越浪量, 并发现影响越浪量和波浪爬高的因素基本相同, 从而提出了不规则波作用下越浪量计算公式。陈松贵等[5]和 Liu 等[6]通过水槽实验分别研究了规则波和

收稿日期: 2020−06−27; **修订日期:** 2021−04−09。

基金项目: 国家自然科学基金(52001149, 52039005, 51861165102); 中国科协青年人才托举工程(2018QNRC001); 中央级公益性科研院所基本科研业务费(TKS20200204, TKS20210102, TKS20210110); 天津市科技计划项目 (17PTYPHZ00080)。

作者简介: 胡原野(1995—), 男, 河南省许昌市人, 主要从事波浪与结构物相互作用方向研究。E-mail: 1140799473@qq.com

*** 通信作者:** 陈松贵(1987—), 男, 天津市人, 副研究员, 主要从事波浪理论及波浪与结构物相互作用研究。E-mail: chensg05@163.com

不规则波作用下岛礁陡变地形上直立堤越浪规律, 给出了平均越浪量的计算公式。Owen [7] 在不考虑斜坡粗糙度的情况下, 通过一系列实验推导出越浪量计算公式。van de Meer 等[8] 对斜坡堤越浪量做了大量的工作, 综合考虑了防波堤参数和波浪参数的影响, 提出了斜坡堤上平均越浪量公式, 该公式被欧洲大多数国家使用。美国《海岸工程手册》中采用的就是 Ward 和 Ahrens[9] 通过实验计算的越浪量公式。舒叶华等[10] 通过对复式结构海堤越浪量进行研究, 比较了国内外常见的复式海堤的越浪量计算方法。Oliveira 等[11] 基于粒子有限元法(PFEM)建立了数值波浪水槽模型, 模拟了不可渗透海堤的越浪过程, 给出了一种求解越浪量的工具。关大玮[12] 应用 FLOW-3D 建立了可模拟规则波和不规则波浪的三维数值水槽, 并模拟了复坡堤上越浪过程, 将结果与实验数据对比, 吻合较好。董志等[13] 采用数值模拟的方法, 利用 RANS 方程和 VOF 法建立数值波浪水槽, 针对复式海堤分别进行了规则波和不规则波越浪的数值模拟。而 van Gent 等[14] 采用了人工神经网络的方法对越浪量做了预测, 并给出越浪量在不同置信区间的值。Formentin 等[15] 在前人的基础上增加了模型的输入参数, 对模型做了进一步的完善。刘诗学等[16] 采用人工神经网络方法对单坡式防波堤越浪量做了估算。Liu 等[17] 通过使用深水波参数作为输入开发了一种反向传播的人工神经网络模型, 来预测珊瑚礁上不透水的垂直海堤的越浪量。传统的经验公式法通常需要消耗大量的人力、物力资源, 且公式的推导过程较为繁琐; 数值模拟法通常需要为了达到相应的精度, 而需要非常大的计算量, 对计算机性能要求较高; 神经网络方法在经济效益方面具有一定优势, 但是仍存在一些不足之处, 比如全局参数搜索比较困难, 对奇异样本敏感, 容易陷入局部最优。

随机森林是近几年来兴起的一种基于统计学的人工智能算法。它基于决策树结构组成的强学习器, 是一种集成学习算法, 该算法对异常数据有较高的容忍性, 且能够直接处理高维度样本[18]。目前, 该方法极少应用于越浪量预测方面。本文提出利用随机森林算法预测越浪量, 为越浪量的计算提供了一种新的方法。

2　数据获取与处理

2.1　CLASH 数据集介绍

“CLASH”是欧盟启动的一个项目计划, 它搜集了各国有关越浪量的实验数据, 组成了较为丰富的越浪量数据集。该数据集有 1 万多条数据, 包含了多种防波堤类型, 每条数据都包含波要素参数、越浪量和防波堤结构参数。此外, 数据集包含有关实验可靠性和结构复杂性的一些信息, RF 表示实验可靠性, 取值在 1～4 之间, RF 值越小说明实验可信度越高, 相反则说明实验可靠性越低; CF 表示断面的复杂度, 取值在 1～4 之间, CF 值越大表示断面越复杂, 反之亦然。

2.2　数据处理

本文主要研究复坡堤越浪量, 根据复坡堤的结构特点, 选取以下参数: 堤前有效波高$H_{m0,t}$、堤前谱周期$T_{m-1,0,t}$、波浪入射角β、堤前水深h、坡度m、堤脚浸没水深h_t、堤脚宽度B_t、护面块体粗糙度γ_f、平台以下结构与水平面正切值$\cot\alpha_d$、平台宽度B、平台上水深h_b、波浪爬高范围内的平均坡度(包含平台)$\cot\alpha_{incl}$、护面块体的平均粒径D、堤顶高程A_c、胸墙顶高程R_c、肩台宽度G_c。结构示意图如图 1 所示。

数据处理是指对数据集进行筛选、整理, 删除错误、无效和有缺失值的数据。经过一系列处理, 将原

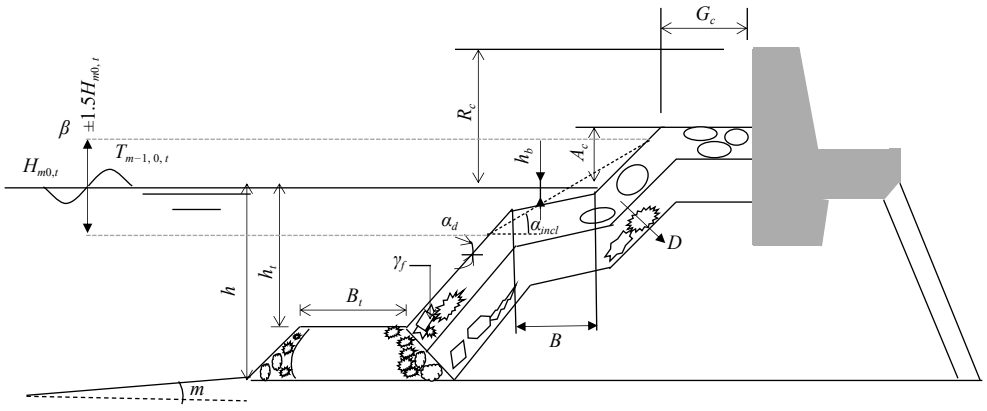

图 1　复坡堤参数示意图

Fig. 1　Schematic diagram of composite slope breakwater parameters

始数据变为可供模型直接使用的数据。其方法如下：

（1）删除标签为 Non-core data 的数据；

（2）删除$q < 10^{-6}$ m³/(s·m)的数据；

（3）删除有缺失值的数据；

（4）删除$CF = 4$和$RF = 4$的数据行。

经过对数据的处理，用于模型使用的数据量为 2 462 条。

2.3 无量纲化

由于越浪量数据集是在特定的实验条件下测量的，会存在不同组次数据的比尺不同，为了消除实验模型比尺和数据量纲之间的差异，需要对数据进行无量纲化。对于每条数据，根据$L_{m-1,0} = gT_{m-1,0}^2/(2\pi)$求出波长，然后按以下方法进行无量纲化：

（1）计算出$H_{m0,t}/L_{m-1,0,t}$；

（2）计算出$h/L_{m-1,0,t}$；

（3）水平方向参数除以波长；

（4）竖直方向参数除以波高；

（5）角度和地貌参数保持不变；

（6）越浪量采用 EurOtop 手册[18]中方法进行无量纲化，并对其进行归一化

$$q_{\mathrm{AD}} = \frac{q}{\sqrt{gH_{m0,t}^3}}, \tag{1}$$

$$q^* = \frac{\log_{10}(q_{\mathrm{AD}}) - \min\left\{\log_{10}(q_{\mathrm{AD}})\right\}}{\max\left\{\log_{10}(q_{\mathrm{AD}})\right\} - \min\left\{\log_{10}(q_{\mathrm{AD}})\right\}}, \tag{2}$$

式中，q为越浪量(m³/(s·m))；$H_{m0,t}$为堤前有效波高(m)；g为重力常量，取9.8 m/s²；q_{AD}为无量纲化后的越浪量(m³/(s·m))；$q*$为归一化后的越浪量(m³/(s·m))。

3 随机森林算法

3.1 随机森林原理

随机森林是一种基于决策树模型的集成学习算法，通过对样本数据随机抽样组成多个不同的决策树，再把决策树计算结果通过某种组合策略来获得随机森林的预测结果。随机森林可以看作是决策树的整合择优。因此，随机森林通常比单纯的决策树模型具有更好的拟合能力，且随机森林在分类问题和回归问题上都具有较好的效果。本文建立的越浪量预测模型就是随机森林在回归问题上的体现。

3.1.1 决策树

决策树是随机森林的基本组成单元，也是一种机器学习算法，它的建立过程基于树形结构，主要由内部节点、树枝和叶节点组成。如图 2 所示，最上面的是根节点，严格来说，根节点也属于内部节点。树的建立过程就是节点分化的过程，每一次节点划分都会

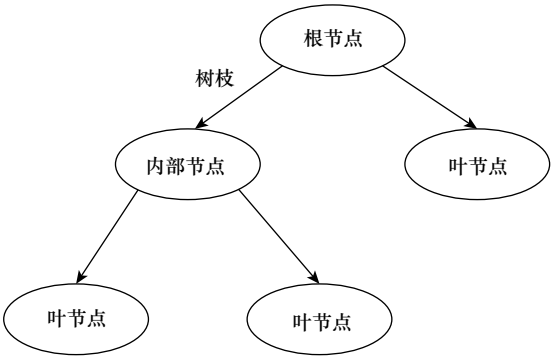

图 2 决策树基本结构示意图

Fig. 2 Schematic diagram of the basic structure of the decision tree

得到对应的输出，即经过分化多了一个节点。经过有限次的条件划分结束后，最终每个单元的输出也就确定了，即叶节点。一般来说，随着模型复杂程度的提高，决策树也随之长得很大。

决策树理论的核心就是如何最优地确定切分点。随着决策树的逐渐长大，样本划分的也越来越细，也就是各节点的样本纯度也会越高(即越来越趋于同一类)。每次逐步划分当前所有特征中的所有取值，然后基于平方误差最小化准则选择最优的切分点。比如切分点为训练集中第j个特征变量$x^{(j)}$，且$x^{(j)}$的值为s，定义区域$R_1(j,s) = \{x|x^{(j)} \leqslant s\}$和区域$R_2(j,s) = \{x|x^{(j)} > s\}$，然后确定$j$和$s$，使得平方误差最小，即求解下式[19]

$$\min_{j,s}\left[\sum_{x_i \in R_1(j,s)}(y_i - \bar{y}_1)^2 + \sum_{x_i \in R_2(j,s)}(y_i - \bar{y}_2)^2\right], \tag{3}$$

式中，y_i为输出变量；$\bar{y}_1$为在区域R_1上y_i的均值；$\bar{y}_2$为在区域R_2上y_i的均值。

在确定出最优的(j,s)后，该节点就会划分为两个子节点，然后对每个子节点重复上述过程，直到满足条件停止。

3.1.2 随机森林算法结构

随机森林是由一系列决策树组成的一种强学习器，根据 Bagging 集成方法来提高算法的精度。具体步骤如下：

（1）从越浪量样本集有放回地随机抽取n个训练集，原始样本集中会有约 36.8% 的样本未被抽到，把该部分数据称为袋外数据(OOB)。

（2）利用抽取的n个训练集组成n棵决策树，在分裂过程，其中在每一个内部节点从M个特征中随机选择m个特征进行分裂($M \geqslant m$)。这样通过特征的随机性增加了各决策树之间的差异性。

(3)经过训练, 每一颗决策树都会对样本做出回归预测, 分别得到n个预测结果q_1, q_2, q_3, …, q_n。

(4)采用平均法的方式, 将n棵决策树的输出结果综合平均, 最后得到预测结果q, 即: $q=\frac{1}{n}\sum_{i=1}^{n}q_i$。因此, 基于随机森林的复坡堤越浪量预测模型结构如图 3 所示。

3.2　模型的建立

本文利用 Python 提供的 Numpy 和 Pandas 库对数据进行处理, 使数据转化为可直接供模型使用的数据类型。利用 Sklearn 建立越浪量预测模型, 把经过处理后的数据输入到建立的模型中。其中, 将数据集随机地划分为两部分: 90% 作为训练集供模型学习, 10% 作为测试集用来评估模型的性能。表 1 为处理后的无量纲数据的分布特征。

3.3　模型调参

模型参数的调节对模型性能有非常重要的影响, 本文主要对影响随机森林精度较大的 3 个参数做优化, 分别为决策树的数量(n_estimators)、决策树的最大深度(max_depth)和随机选择的最大特征数(max_features)。综合考虑模型精度和运行时间成本, 给上述 3 个参数选取多个适当的值, 取值范围如表 2。

其中, n_estimators 的取值步长为 10; max_depth 的取值步长为 5; max_features 有两种取值: auto 表示取所有的特征, sqrt 表示取特征数的平方根。

本文利用 Sklearn 库中网格搜索(GridSearchCV)方法对 3 个参数进行调优, 该方法只需把设置好范围的需要调优的参数输入到此算法中, 它就会遍历整个范围获得多种参数组合, 这样就能方便快捷得到最优的结果。经过网格搜索计算, 得到的最优参数取值 n_estimators=150, max_depth=20, max_features='sqrt'。

4　结果分析

为了评估随机森林算法对越浪量的预测精度, 通过比较预测值和真实值来直观判断。同时, 通过R^2(决定系数)来定量计算模型的预测精度, R^2的值越接近 1, 则预测值和真实值越接近, 表明模型越好; 反之, R^2越接近 0, 则表明模型越差。决定系数[20]计算公式为

$$R^2=1-\frac{\sum_{i=1}^{N}\left(q_{rf_i}-\bar{q}\right)^2}{\sum_{i=1}^{N}\left(q_i-\bar{q}\right)^2},\tag{4}$$

式中, N为样本数; q_{rf_i}为预测值(第i个样本); $\bar{q}$为真实值的平均值; q_i为真实值(第i个样本)。

4.1　预测结果分析

将划分好的数据集输入到建立好的随机森林模型中, 分别得到训练集和测试集的预测结果, 如图 4 和图 5。训练集预测结果表示模型对样本数据的学

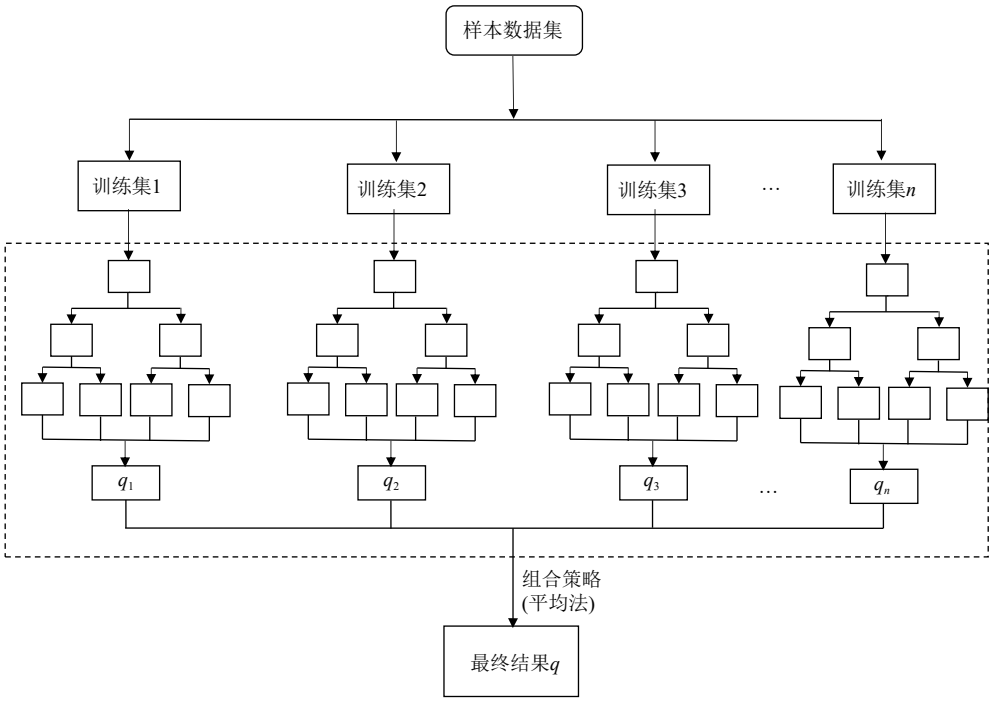

图 3　基于随机森林的复坡堤越浪量预测模型结构图

Fig. 3　Structure diagram of overtopping prediction model of composite slope breakwater based on random forest

表 1 无量纲化后输入参数分布特征

Table 1 Distribution characteristics of input parameters after dimensionless

特征参数	平均值	最大值	最小值	标准差
$H_{m0,t}/L_{m-1,0,t}$	0.033	0.087	0.005	0.012
β	0.716	80.000	0.000	4.820
$h/L_{m-1,0,t}$	0.135	0.666	0.010	0.102
$h_t/H_{m0,t}$	3.467	22.566	0.429	2.403
$B_t/L_{m-1,0,t}$	0.017	0.396	0.000	0.050
$h_b/H_{m0,t}$	0.173	7.826	–2.652	1.014
$B/L_{m-1,0,t}$	0.093	0.973	0.000	0.109
$A_c/H_{m0,t}$	1.155	4.216	–5.242	0.581
$R_c/H_{m0,t}$	1.246	6.032	0.000	0.531
$G_c/L_{m-1,0,t}$	0.023	0.257	0.000	0.039
m	417.755	1 050.000	10.000	455.659
$\cot\alpha_d$	1.672	7.000	0.000	1.331
$\cot\alpha_{incl}$	2.584	11.299	–1.331	2.096
γ_f	0.790	1.000	0.380	0.269
$D/H_{m0,t}$	0.118	0.807	0.000	0.152
目标参数				
q*	0.001 668	0.165	0.000 001	0.010 84

表 2 重要参数取值范围

Table 2 Value range of important parameters

重要参数	取值范围
n_estimators	10～200
max_depth	10～50
max_features	auto, sqrt

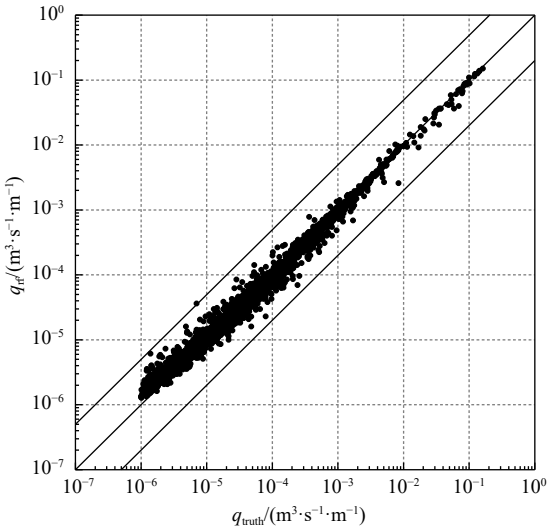

图 4 训练集预测结果(随机森林)

Fig. 4 Prediction result of training set (random forest)

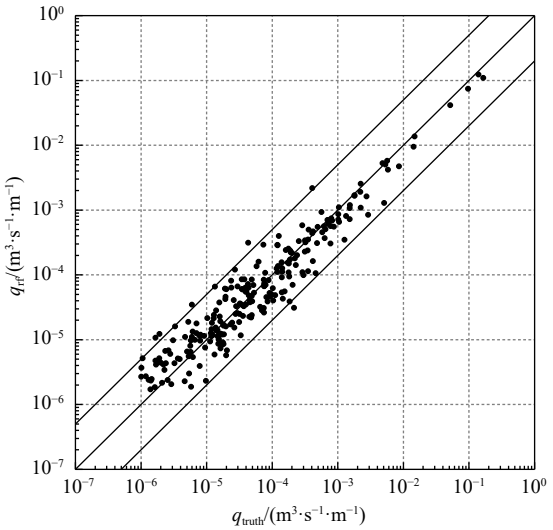

图 5 测试集预测结果(随机森林)

Fig. 5 Prediction result of testing set (random forest)

习能力,测试集的预测结果表示模型对测试数据的泛化能力,即对新样本的适应能力。

从图 4 和图 5 可以看出,训练集预测结果基本都在 5 倍误差区间内(两侧实线之间),且决定系数 R^2 = 98.8%,表明该模型具有很好的学习能力;测试集与训练集是完全不同的数据集,这部分并没有参与模型的训练,其结果依然能够基本上落在 5 倍误差区间内,且决定系数R^2 = 92.7%,预测结果很可靠,表明模型对新样本也具有很强的适应能力。

4.2 与集成神经网络算法的对比

为了进一步验证随机森林算法预测复坡堤越浪量的精度,与集成神经网络算法预测结果做了对比。采用与随机森林算法相同的数据集,建立了基于集成神经网络算法的越浪量预测模型。神经网络模型包含 3 层:输入层、隐含层和输出层。由于目前没有足够的理论确定神经元个数,常采用逐步试验法选择结果较好的,神经元个数最小的组,以免过拟合。因此最终确定输入层神经元数为 15,隐含层神经元数为 25,输出层神经元数为 1。激活函数选择选取双曲正切函数(tanh),输入参数采用 max-min 归一化。构建 100 个网络模型,然后将这 100 个模型的输出结果通过平均的方法得到最终集成神经网络的输出结果。集成神经网络模型的预测结果如图 6 和图 7 所示,训练集大部分落在 5 倍误差范围内,也有比较多的点落在了 5 倍范围之外,决定系数R^2 = 91.9%;而测试集的

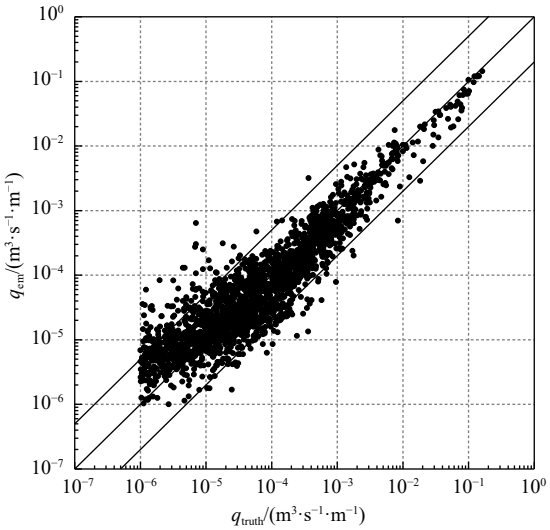

图 6　训练集预测结果比较(集成神经网络)

Fig. 6　Prediction result of training set (ensemble neural network)

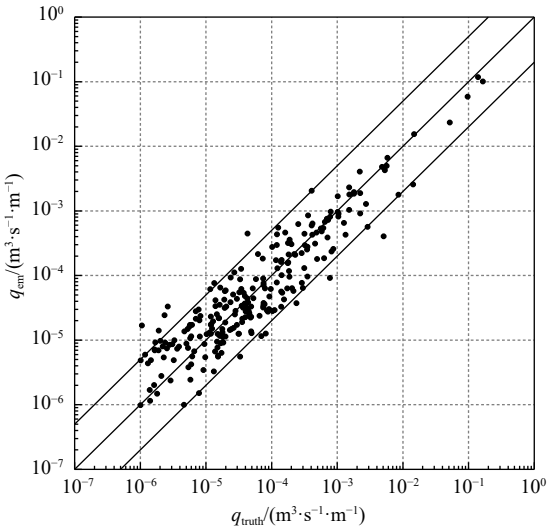

图 7　测试集预测结果比较(集成神经网络)

Fig. 7　Prediction result of testing set (ensemble neural network)

结果基本都落在 5 倍误差范围内, 决定系数R^2 = 87.7%。对比两种算法训练集的预测结果(图 4 和图 6)发现, 随机森林算法的结果明显比集成神经网络更集中在 45°理想线附近(中间的实线), 且决定系数也高于集成神经网络, 这说明随机森林比集成神经网络具有更强的学习能力; 对比两种算法的测试集结果(图 5 和图 7), 直观上不容易看出明显的差别, 但根据两者的决定系数可知, 随机森林依然要高于集成神经网络, 说明随机森林算法的泛化能力也更好。综上所述, 不管是训练集还是测试集, 本文建立的随机森林模型的准确度都要优于集成神经网络。这是由于在构建随机森林模型时, 每棵决策树的训练集是通过 Bagging 集成方法抽样, 且决策树分裂时采用随机选择, 这使得随机森林中的决策树多样性增加, 从而更好地发挥了集成思想的作用。

4.3　特征参数对预测精度的影响

分析特征参数的重要性就是探究哪些特征对模型的影响大, 哪些特征对模型的影响小, 这样有助于更好的做特征筛选, 即对于影响特别小的特征, 对模型来说或许会被认为是噪点, 可以选择丢弃。一方面, 可以提高模型的精度; 另一方面, 利于减小模型的计算量, 从而提高效率。

通过随机森林模型对越浪量预测的同时, 模型可以评估所有输入特征对预测结果的重要性。其原理是: 在建模过程中随机森林会挑选出某一个特征对其加入噪声, 然后观测对计算结果的影响, 最后比较各特征之间的影响大小。一般用袋外数据误差评价。方法是: 对于一颗决策树, 计算 OOB 的误差e_1, 对于特征参数X_i, 置换 OOB 中的第X_i列, 保持其他列不变, 再次计算袋外误差e_2, 用e_1-e_2表示特征参数X_i的重要性。最后把所有决策树计算得到的e_1-e_2取平均, 即特征参数X_i对随机森林模型的重要性。袋外误差e的计算公式为

$$e=\frac{1}{N}\sqrt{\sum_{i=1}^{N}(\hat{q}_i-q_i)^2}, \tag{5}$$

式中, $\hat{q}_i$、q_i分别为第i个样本的预测值和真实值; N为对应的样本数。因此, 特征参数X_i的重要性评分[21]为

$$\mathrm{Score}_{X_i}=\frac{\sum_{t=1}^{n}(e_{2_t}^{X_i}-e_{1_t}^{X_i})}{n}, \tag{6}$$

式中, n是随机森林中树的个数; $e_{1_t}^{X_i}$表示特征参数X_i在置换之前的第t棵树的袋外误差; $e_{2_t}^{X_i}$表示特征参数X_i在置换后的第t棵树的袋外误差。如果对某个特征加入噪声, 随机森林的袋外准确率大幅减小了, 说明该特征的重要程度较高。通过随机森林算法对特征计算重要性评分, 然后对其进行归一化后, 就得到特征重要性。由于有些特征参数本身具有无量纲特性, 且这类数据之间的差异较大, 这里不予考虑。我们只讨论经过无量纲化后的特征对预测结果的影响, 如图 8。

由图 8 可知, 重要性最高的特征参数为墙顶高程R_c, 其次是堤顶高程A_c和平台上水深h_b, 再者是平台宽度B和波陡$H_{m0,t}/L_{m-1,0,t}$, 而堤前水深h、堤脚浸没水深h_t和肩台宽度G_c的重要性相当, 护面块体平均粒径D和堤脚宽度B_t对预测结果的影响最小。不难理解, 墙顶能够有效的阻挡波浪越过堤顶, 随着墙顶高程的

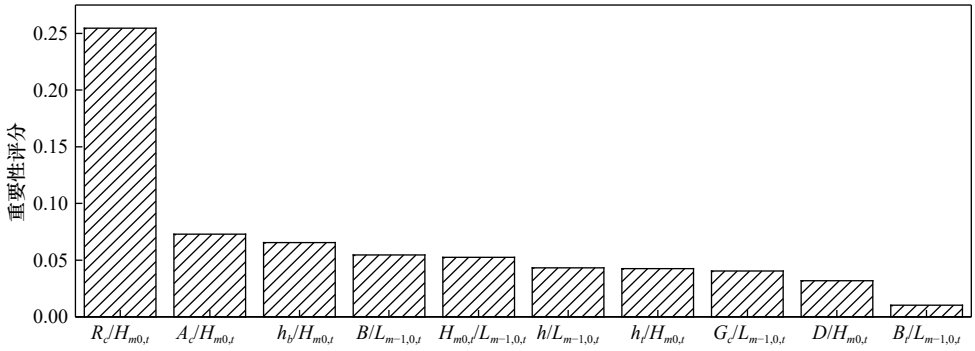

图 8 模型特征参数重要性评价

Fig. 8 Importance evaluation of model characteristic parameters

增加，波浪在挡浪墙处的破碎更加剧烈，大部分水体将被挡在海浪侧。需要消耗更多的能量波浪才能越过堤顶；倘若波浪超过堤顶，就会有比较大的可能发生越浪，堤顶高程的增加对减少越浪具有重要的意义；平台可以削减海侧方向来的波浪，而且设置在静水位附近时的削弱效果最好[22]，平台宽度一定程度上会影响波浪的爬高；波浪在堤脚附近时，由于浅水变形使得波陡变大最终发生破碎，导致波浪损失能量，因此会对越浪量造成一些影响；护面块体粒径大小主要是以渗透率和孔隙率的形式影响越浪量，对越浪量的影响不大；而堤脚宽度对越浪的影响非常小。

分析各特征对预测精度影响，从模型角度讲，可以对模型做特征选择，丢弃对预测精度影响小的特征，保留影响大的特征，来进一步提高模型的精度；从工程角度讲，了解影响越浪量大小的因素，有利于控制越浪量，为设计防波堤提供参考。

5 结论

本文以欧洲 CLASH 项目作为数据支撑，利用 Python 构建了基于随机森林算法的复坡堤越浪量预测模型并通过调参使模型得以优化，从而提高了模型的准确率。为了验证本文提出的越浪量预测模型的准确度，将本文建立的随机森林预测模型与集成神经网络模型的预测精度进行对比，结果显示，随机森林的预测精度要优于集成神经网络。此外，随机森林算法还给出了特征参数对模型预测精度的影响大小，为进一步对特征参数做筛选提供依据。通过本文的研究，实现了将随机森林算法应用于越浪量预测领域，为计算复坡堤越浪量提供了一种新的方法，对设计防波堤和提高防波堤安全性具有较大的实际应用价值。

参考文献：

[1] 王红, 周家宝, 章家昌. 单坡堤上不规则波越浪量的估算[J]. 水利水运科学研究, 1996(1): 58−63.
Wang Hong, Zhou Jiabao, Zhang Jiachang. Estimating of irregular wave overtopping quantities on single sloping[J]. Journal of Nanjing Hydraulic Research Institute, 1996(1): 58−63.

[2] 中华人民共和国交通运输部. JTS 145−2015, 港口与航道水文规范[S]. 北京: 人民交通出版社, 2015.
Ministry of Transport of China. JTS 145−2015, Code of Hydrology for Harbour and Waterway[S]. Beijing: China Communications Press, 2015.

[3] 范红霞. 斜坡式海堤越浪量及越浪流试验研究[D]. 南京: 河海大学, 2006.
Fan Hongxia. Expermental study on the overtopping discharge and overtopping flow of the sloped[D]. Nanjing: Hohai University, 2006.

[4] 陈国平, 周益人, 严士常. 不规则波作用下海堤越浪量试验研究[J]. 水运工程, 2010(3): 1−6.
Chen Guoping, Zhou Yiren, Yan Shichang. Test study on wave overtopping under irregular wave action[J]. Port & Waterway Engineering, 2010(3): 1−6.

[5] 陈松贵, 王泽明, 张弛, 等. 珊瑚礁地形上直立式防浪堤越浪大水槽实验[J]. 科学通报, 2019, 64(28/29): 3049−3058.
Chen Songgui, Wang Zeming, Zhang Chi, et al. Experiment on wave overtopping of a vertical seawall on coral reefs in large wave flume[J]. Chinese Science Bulletin, 2019, 64(28/29): 3049−3058.

[6] Liu Ye, Li Shaowu, Chen Songgui, et al. Random wave overtopping of vertical seawalls on coral reefs[J]. Applied Ocean Research, 2020, 100: 102166.

[7] Owen M W. Design of seawalls allowing for wave overtopping[R]. Wallingford: Hydraulics Research Station,1980.

[8] Allsop N W H, Bruce T, de Rouck J, et al. EurOtop II manual on wave overtopping of sea defences and related structures[R]. Deflt: Tech-

nical Adcisory Committee on Flood Defence, 2016.
[9] Ward D L, Ahrens J P. Overtopping rates for seawalls[R]. Coastal Engineering Research Center Vicksburg MS, 1992.
[10] 舒叶华, 徐宇航, 谢先坤. 复式海堤结构越浪量计算方法比较[J]. 水运工程, 2019(5): 27−31.
Shu Yehua, Xu Yuhang, Xie Xiankun. Comparison of calculation methods for overtopping discharge of composite seawall structure[J]. Port & Waterway Engineering, 2019(5): 27−31.
[11] Oliveira T C A, Sánchez-Arcilla A, Gironella X. Simulation of wave overtopping of maritime structures in a numerical wave flume[J]. Journal of Applied Mathematics, 2012, 2012: 246146.
[12] 关大玮. 规则波与不规则波的海堤越浪数值模拟[D]. 广州: 华南理工大学, 2016.
Guan Dawei. Numerical simulation of regular and irregular wave overtopping against seawalls[D]. Guangzhou: South China University of Technology, 2016.
[13] 董志, 关大玮, 苗青, 等. 复式海堤上规则波和不规则波越浪数值模拟研究[J]. 中国农村水利水电, 2020(3): 112−118.
Dong Zhi, Guan Dawei, Miao Qing, et al. Numerical simulation of regular and irregular waves overtopping on composite section seadike[J]. China Rural Water and Hydropower, 2020(3): 112−118.
[14] van Gent M R A, van den Boogaard H F P, Pozueta B, et al. Neural network modelling of wave overtopping at coastal structures[J]. Coastal Engineering, 2007, 54(8): 586−593.
[15] Formentin S M, Zanuttigh B, van der Meer J W. A neural network tool for predicting wave reflection, overtopping and transmission[J]. Coastal Engineering Journal, 2017, 59(1): 1750006.
[16] 刘诗学, 王收军, 陈松贵, 等. 基于人工智能的单坡式防波堤越浪量评估方法研究与应用[J]. 水道港口, 2019, 40(5): 541−546, 587.
Liu Shixue, Wang Shoujun, Chen Songgui, et al. Research and application of artificial intelligence based method for overtopping assessment of straight slopes[J]. Journal of Waterway and Harbor, 2019, 40(5): 541−546, 587.
[17] Liu Ye, Li Shaowu, Zhao Xin, et al. Artificial neural network prediction of overtopping rate for impermeable vertical seawalls on coral reefs[J]. Journal of Waterway, Port, Coastal, and Ocean Engineering, 2020, 146(4): 04020015.
[18] 孙明喆, 毕瑶家, 孙驰. 改进随机森林算法综述[J]. 现代信息科技, 2019, 3(20): 28−30.
Sun Mingzhe, Bi Yaojia, Sun Chi. A survey of improved random forest algorithms[J]. Modern Informationn Technology, 2019, 3(20): 28−30.
[19] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
Li Hang. Statistical Learning Methods[M]. Beijing: Tsinghua University Press, 2012.
[20] 贾怀勤. 应用统计[M]. 北京: 外经济贸易大学出版社, 1998.
Jia Huaiqin. Applied Statistics[M]. Beijing: University of International Business and Economics Press, 1998.
[21] 林子聪, 任向宁, 朱阿兴, 等. 基于随机森林算法的耕地质量定级指标体系研究[J]. 华南农业大学学报, 2020, 41(4): 38−48.
Lin Zicong, Ren Xiangning, Zhu Axing, et al. Research on the index system of cultivated land quality grading based on random forest algorithm[J]. Journal of South China Agricultural University, 2020, 41(4): 38−48.
[22] 陈国平, 余广明, 章家昌. 平台高程与宽度对不规则波爬高的影响[J]. 海洋工程, 1992, 10(4): 59−67.
Chen Guoping, Yu Guangming, Zhang Jiachang. The effect of berm width and elevation on irregular wave run-up[J]. The Ocean Engineering, 1992, 10(4): 59−67.

Overtopping prediction for composite slope breakwater based on random forest method

Hu Yuanye [1,2], Wang Shoujun [1], Chen Songgui [2], Liu Ye [2], Wang Jiawei [1,2], Tian Yunyan [1,2]

(1. *National Demonstration Center for Experimental Mechanical and Electrical Engineering Education*, *Tianjin University of Technology*, *Tianjin* 300384, *China*; 2. *National Engineering Laboratory for Port Hydraulic Construction Technology*, *Tianjin Research Institute for Transport Engineering*, *Tianjin* 300456, *China*)

Abstract: Aiming at the problem of calculating overtopping of the composite slope breakwater, a prediction model of the overtopping for the composite slope based on the random forest method is proposed. Firstly, by filtering the European CLASH data set, the data consistent with the prediction of overtopping of the composite slope breakwater are selected. Secondly, after dimensionless processing of the data, overtopping prediction model is established based on random forest method, and improved by adjusting the model parameters according to GridSearchCV. Fi-

nally, the coefficient of determination R^2 is used to evaluate the accuracy of the model, and the prediction ability of the model is compared with the ensemble neural network model. The effect of each feature parameter of the random forest model on the prediction accuracy is assessed. The results show that the coefficient of determination of the random forest model is 92.7%, and the coefficient of determination of the ensemble neural network model is 87.7%, indicating the random forest model has a stronger prediction ability for predicting overtopping. Wall height with respect to static water level has the greatest influence on the prediction accuracy of the model, the height of the top of the embankment is the second, and the width of the foot of the embankment least.

Key words: random forest; overtopping; composite slope breakwater; coefficient of determination; feature importance; prediction