

DOI:10.3876/j.issn.1000-1980.2020.04.010

基于 Stacking 集成机器学习的波浪预报

沈晖华^{1,2}, 时 健^{1,2}, 徐佳丽², 朱士鹏³, 郑金海^{1,2}

(1. 海岸灾害及防护教育部重点实验室(河海大学), 江苏 南京 210098;
2. 河海大学港口海岸与近海工程学院, 江苏 南京 210098; 3. 南京航空航天大学自动化学院, 江苏 南京 211106)

摘要: 采用多层感知器模型、随机森林模型为第一层子模型, 极端树模型为第二层元模型, 建立基于 Stacking 集成机器学习的波浪预报算法, 并引入邻域平均法抑制在拐点处产生的数值震荡。以长江口外海 2016 年 1—9 月的风速和中国近海波高数据为数据源, 利用机器学习风速与有效波高之间的关系, 将 2016 年 10—11 月的风速、波高数据用于预报结果的对比分析, 预报前 45 d R^2 拟合优度达到 0.97 以上, 平均误差最大值为 0.08 m, 平均相对误差最大值为 0.05, 预报结果与波浪谱模型结果趋势一致, 准确度较高; 预报结果后 15 d 误差增长较快, 这与训练集数据中寒潮浪占比较少有关。

关键词: 集成机器学习; Stacking 算法; 波浪预报方法; 长江口外海海域; 中国近海波浪数据
中图分类号: P731.33 **文献标志码:** A **文章编号:** 1000-1980(2020)04-0354-05

Wave forecasting algorithm with stacking ensemble machine learning method

SHEN Huihua^{1,2}, SHI Jian^{1,2}, XU Jiali², ZHU Shipeng³, ZHENG Jinhai^{1,2}

(1. Key Laboratory of Coastal Disaster and Protection(Hohai University),
Ministry of Education, Nanjing 210098, China;
2. College of Harbor, Coastal and Offshore Engineering, Hohai University, Nanjing 210098, China;
3. College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: A wave forecasting algorithm is established based on the stacking ensemble machine learning method. The algorithm includes two sublayers. The multi-layer perceptron model and the random forest model are used in the first sublayer, and the extreme randomized tree model is used in the second sublayer. To suppress the numerical oscillation of the predicted results, the neighborhood averaging method is introduced into the algorithm. To establish relations between the wind speed and the significant wave height, the wind speed and China wave (CWAVE) data from January to September 2016 in the offshore area of the Yangtze River Estuary are used. By using the wind data from October to November 2016, the corresponding significant wave height is predicted by the wave forecasting algorithm. The results show that the values of R^2 between the predicted wave height and CWAVE data are larger than 0.97 in the 45 days from 1 October, the mean error is less than 0.08 m, and the mean relative error is less than 0.05. The trends of the variations of the significant wave height can be accurately predicted by the algorithm. However, the error increases in the last 15 day, which is due to lack of cold wave data in the training set.

Key words: ensemble machine learning; Stacking algorithm; wave forecasting method; sea area outside the Yangtze River Estuary; China offshore wave data

基金项目: 国家杰出青年科学基金(51425901); 国家自然科学基金(41930538); 中国博士后科学基金(2018M632220)
作者简介: 沈晖华(1995—), 男, 硕士研究生, 主要从事波浪数值模拟预报研究。E-mail: huihs@hhu.edu.cn
通信作者: 郑金海, 教授。E-mail: jhzheng@hhu.edu.cn
引用本文: 沈晖华, 时健, 徐佳丽, 等. 基于 Stacking 集成机器学习的波浪预报[J]. 河海大学学报(自然科学版), 2020, 48(4): 354-358.
SHEN Huihua, SHI Jian, XU Jiali, et al. Wave forecasting algorithm with stacking ensemble machine learning method[J]. Journal of Hohai University(Natural Sciences), 2020, 48(4): 354-358.

波浪是海工结构物、海上运输及作业工具运转时需要考虑的重要安全因素。开展准确高效的波浪预报,不仅有助于海上生产活动的开展,也有助于增强海工结构物、海上运输工具以及人员应对灾害的能力,具有重要的科研价值和社会经济意义。

伴随机器学习的发展,目前应用于波浪预报的机器学习模型主要分为 2 类,一类是神经网络模型^[1-5],该模型模仿人类大脑对于多数据、多特征的处理方式,以人工神经元为单位处理、分析数据,具备解决非线性问题的能力,适用于随机性较强的波浪研究,但神经网络模型具有原理可解释性不强、调参耗时耗力的缺点;另一类是决策树模型^[6-7],与神经网络模型相比,决策树模型的优势在于其建模规则清晰可见、便于理解、训练时间更短,缺点是对异常值过于敏感、泛化能力差。针对波浪预报,现有研究成果大都关注于单一模型的应用,而机器学习的发展趋势是同时使用多个算法,以期规避单一模型自身的缺陷,获得比单一模型更好的拟合表现和更小的误差。然而,目前未见该类集成机器学习算法应用于波浪预报研究的报道。

集成机器学习是 20 世纪 90 年代以来机器学习的主流方法^[8],该方法旨在利用多个学习器(模型)来解决同一问题,获得比单个学习器更优异的表现^[9]。目前主流的集成机器学习方法有:Boosting^[10]、Bagging^[11]和 Stacking^[12-13]。不同集成机器学习方法的工作原理和理论基础存在差异,如 Stacking 使用不同的个体学习器去构建算法,而 Boosting 和 Bagging 都是基于同种学习器的集成机器学习方法。本研究数据集变量因素多,数据量大,可能多个映射关系,单一模型难以穷举。因此选择 Stacking 方法建立波浪预报模型,可以集成多个不同的方法,建立多个映射关系,以充分学习波浪数据集。

1 算法构建

如图 1 所示,训练样本 $D\{(x_i^{(t)}, y_i^{(t)})\}$ 包括与有效波高有关的特征数据(x_i)和有效波高(y_i),上标(t)表示模型当前层数。第一层模型学习训练集之后,会输出各个体学习器所预报的有效波高。此时直接用第一层模型的训练集训练第二层模型或会产生过拟合。过拟合是指模型由于高度适应当前训练集而无法对同类数据集进行拟合的情况。为了避免过拟合风险,需要生成新的训练集 D' 。在新的训练集中,第一层模型的输出结果 $h_t^{(j)}(x_i)$ 被当作新一层模型的样本输入,其中 t 表示第一层模型中个体学习器的编号, j 表示样本编号。有效波高样本 y_i 保持不变,形成新的训练集 D' , D' 用于训练第二层模型。以上 2 层机器学习模型组成了基于 Stacking 集成机器学习的波浪预报算法。

模型架构如图 1 所示,采用多层感知器^[14]、随机森林^[15] 2 种拟合度较好的学习器为第一层集成机器学习算法,针对波浪数据进行充分学习。由于不同的模型在假设空间和原理上有所差异,第一层算法可以认为是从被扩大的假设空间中寻求更好假设的过程,又因为第一层算法采用了复杂的非线性变化提取数据特征,所以容易产生过拟合情况。第二层集成机器学习算法采用极端树模型^[16],可以降低在第一层算法中累计的过拟合风险,并从第一层扩大的假设空间中寻找更好的近似,从而获得比单一模型更好的表现,提升预报的准确率和稳定性。以上 2 层集成机器学习算法构成基于 Stacking 集成机器学习的波浪预报算法。

2 研究对象

2.1 样本数据

机器学习需要大量的数据用于模型训练,但高分辨率长时间序列的波浪数据难以获取,因此本文以 Shi 等^[17]所建立的中国近海波浪数据库(China wave database, 以下简称 CWAVE)作为数据源。CWAVE 利用 CFSR (climate forecast system reanalysis) 提供的水面以上 10 m 逐时风场数据驱动波浪模型 TOMAWAC,追算了中国近海 1979-01-01—2017-12-31 为期 39 年的有效波高、平均周期等波浪要素。本研究采用 CWAVE 2016 年 1—9 月波高、风速数据作为训练样本,2016 年 10—11 月风速数据用于波浪预报。

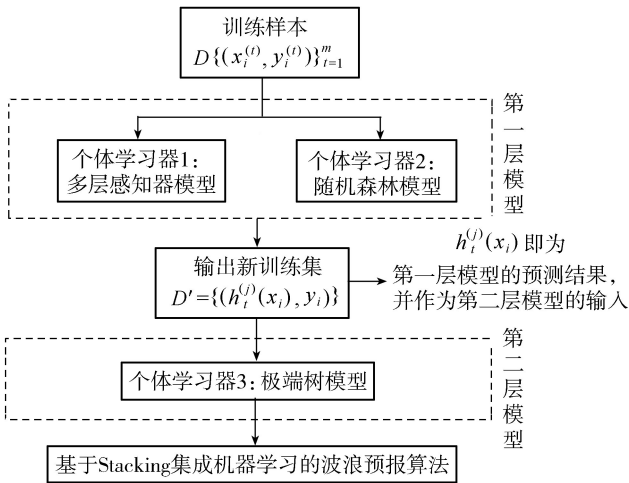


图 1 Stacking 集成机器学习波浪预报算法模型架构

Fig.1 Modeling framework of wave forecasting algorithm based on the stacking ensemble machine learning method

2.2 研究区域

研究区域如图 2(a) 所示,为 $1^{\circ} \times 1^{\circ}$ 的矩形区域,位于 $123.5^{\circ}\text{E} \sim 124.5^{\circ}\text{E}$ 、 $30.5^{\circ}\text{N} \sim 31.5^{\circ}\text{N}$ 。该区域包含东海浮标 (123.5°E , 31.2°N), CWAVE 数据库 2016 年有效波高计算结果经过东海浮标验证,数据质量较高,误差较小^[17],因此选用该区域 CWAVE 有效波高结果作为波浪预报算法的训练集和验证集。

本研究采用矩形网格,将研究区域划分为 31×31 的网格(图 2(b)),网格间距与 CWAVE 数据库中模型计算网格尺寸一致。用于评估算法模拟性能的网格点位于四角、边缘以及内部(图 2 黑色点),可用于检验波浪预报算法在不同区域的精度。

2.3 数值震荡的处理

由于 Stacking 集成机器学习算法下的随机森林模型、极端树模型均以决策树作为个体学习器,而决策树认为局部极值属于异常值,因此容易在波高值拐点处出现数值震荡(图 3(a))。

为减少算法数值震荡,引入邻域平均法^[18],定义 K 参数,其数学表达式如下:

$$K(i) = \frac{|4y_i - y_{i-1} - y_{i-2} - y_{i+1} - y_{i+2}|}{4}$$

(1)

式中: K ——邻域参数,为 4 倍的 i 时刻预报数据与 i 前 2 个时刻以及后 2 个时刻的数据所得绝对值的平均值; y_i ——算法预报数据,不同下标表示 i 前、后不同时刻的预报数据。

K 参数用于刻画 i 时刻与前、后时刻数据的情况,根据该参数可以设置触发抑制震荡的数学条件。本研究所设触发条件为 K 大于 0.2。当触发条件时,将 i 时刻的数据改为前、后 2 个时刻预报数据的平均值。

图 3(b) 为将 K 参数加入波浪预报算法后的有效波高计算结果。可以看出,在波高值拐点所产生的数值震荡大为减少,且加入 K 参数前计算结果均方根误差为 0.229,加入后均方根误差下降为 0.144,说明该方法可有效抑制有效波高预报的数值震荡,提升算法的预报精度。

2.4 结果验证

利用 2016 年 1—9 月数据对波浪预报算法进行训练,选取 3 个网格点的 10 月有效波高预报结果如图 4 所示。可以看出有效波高

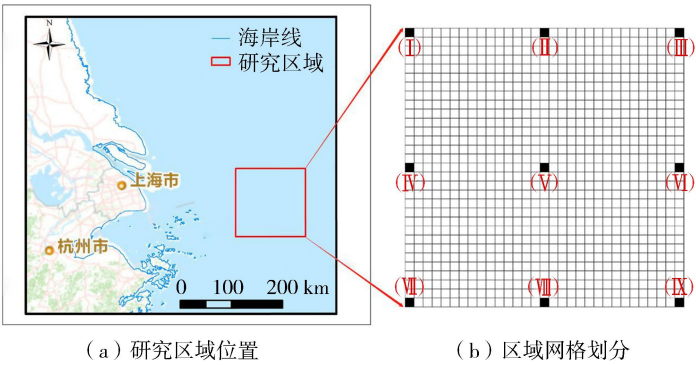


图 2 研究区域位置及其网格划分示意图
Fig.2 Schematic diagram of study area location and its grid division

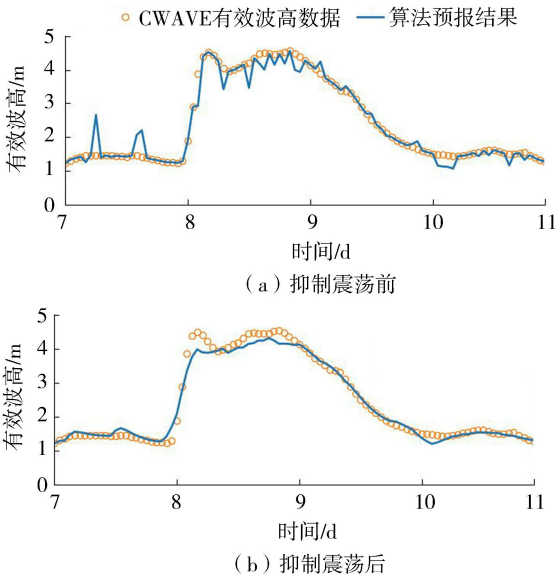


图 3 引入邻域平均法前、后预报结果与 CWAVE 数据对比

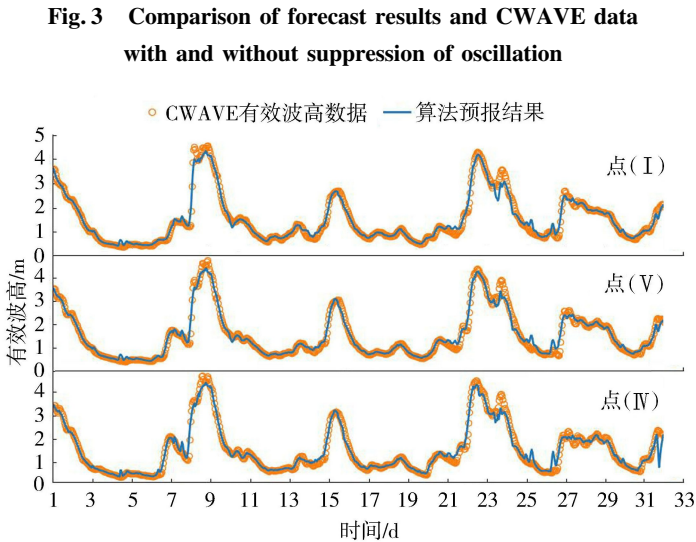


图 4 有效波高预报结果与 CWAVE 数据对比
Fig.4 Comparison of predicted significant wave height and CWAVE data

预报结果与 CWAVE 数据趋势一致,量值上也吻合较好。

3 预报结果分析

3.1 误差统计分析

为定量分析波浪预报结果与 CWAVE 数据之间的误差,引入平均误差 δ_{ME} 和平均相对误差 δ_{MRE} 进行预报结果精度分析(图 5)。

图 5(a)中时间表示自 10 月 1 日起的预测天数。统计结果表明,随着预测时间的增加,平均误差呈增加趋势,但在 1~45 d 内增长速率较低,最大平均误差为 0.08 m;45~60 d 误差增长较快,最大平均误差出现在第 54 天,为 0.17 m。图 5(b)为平均相对误差随预测时间变化曲线,曲线变化趋势与平均误差类似,1~45 d 预报结果精度较高,平均相对误差最大值为 0.05,46~60 d 预报结果平均相对误差值快速增加。

3.2 相关性分析

引入 R^2 分析预报结果与 CWAVE 数据的相关关系, R^2 计算公式如下:

$$R^2 = 1 - \frac{\sum (\hat{y} - y)^2}{\sum (y - \bar{y})^2}$$

(2)

式中: R^2 ——拟合优度,表示 CWAVE 有效波高数据能够被风速数据解释的比例,取值范围为 $0 \leq R^2 \leq 1$, R^2 越接近 1,说明算法所建立的风速与有效波高的相关关系越强; $\hat{y}$ ——算法预报结果; y ——CWAVE 有效波高; $\bar{y}$ ——CWAVE 有效波高平均值。

图 6 列出了每隔 15 d 的 R^2 和预报结果与 CWAVE 数据散点图,1~45 d 预报结果准确度较高, R^2 为 0.97 以上,说明波浪预报算法能够准确学习风速与有效波高的相关关系,精确预报研究区域的有效波高值。1~45 d 预报结果与 CWAVE 数据误差较小,多数数据点落在 45° 对角线上,算法对 3 m 以上有效波高预测精度也较高,数据点都落在 10% 误差线以内。

图 6(d) 为 46~60 d 预报结果与 CWAVE 数据散点图。可以看出,算法对小于 1 m 波浪预报值偏小,1~2 m 波浪预报值偏大, R^2 值仅为 0.41,说明波浪预报算法不能准确预测这一时段的有效波高变化。如图 5 所示,预报结果误差第 50 天(11 月 19 日)加速增长,这段时间研究区域受强寒潮天气影响^[19],风向和波向为西北方向,波浪以当地生成的风浪为主。而算法训练集数据为 1—9 月风速和有效波高数据,寒潮浪占比较少,训练集数据不足导致算法难以准确预测寒潮浪的波高变化。

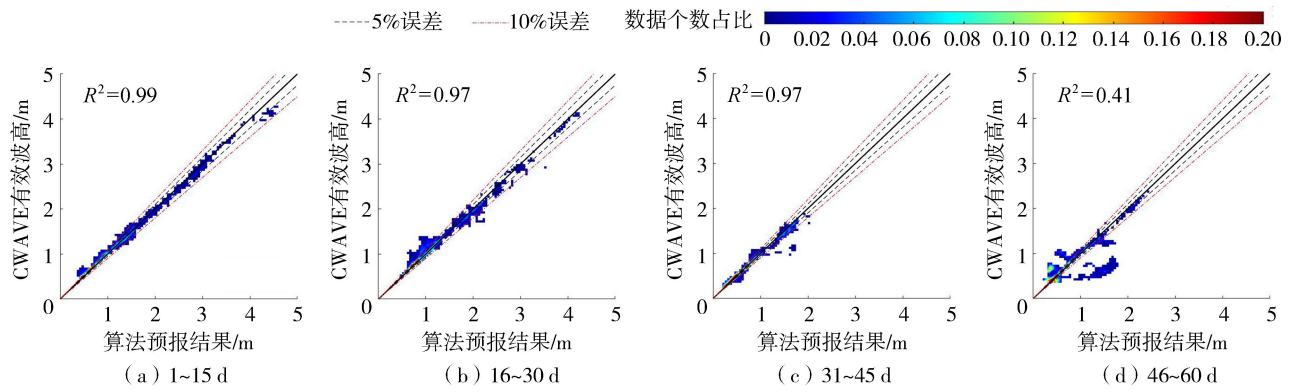


图 6 预报结果与 CWAVE 数据散点图

Fig. 6 Scatter chart of the predicted results and CWAVE data

综上所述,通过集成机器学习算法建立风速与有效波高的映射关系,可实现高精度的有效波高预报。但由于本研究训练样本时间长度仅为 9 个月,未包含足够的寒潮浪训练数据,下一步可通过完善训练数据集,建立更加完整的风速与有效波高的映射关系,从而提高算法波浪预报精度。

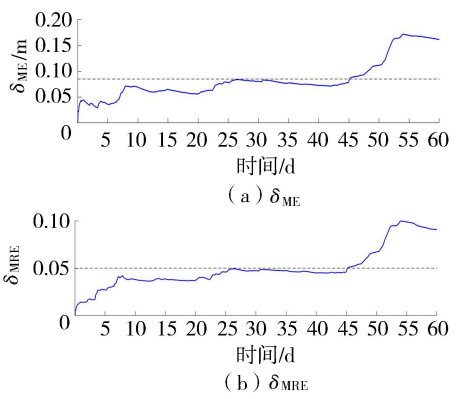


图 5 δ_{ME} 和 δ_{MRE} 时间序列

Fig. 5 Time series of δ_{ME} and δ_{MRE}

4 结 语

采用中国近海波浪数据库(CWAVE)所提供的波高数据和 CFSR 再分析风场数据,以多层感知器、随机森林、极端树作为个体学习器,建立了基于 Stacking 集成机器学习的波浪预报算法。

以长江口外海 $1^{\circ}\times 1^{\circ}$ 的区域为对象,利用 2016 年 1—9 月 CWAVE 有效波高数据和 CFSR 风速数据作为训练样本,以风速为输入参数预报了 2016 年 10—11 月研究区域的有效波高。结果显示波浪预报算法建立了精确的风速与有效波高相关关系,可以准确预测研究区域的波高变化。45 d 内平均误差最大值为 0.08 m,相对平均误差最大值为 0.05。相关性分析发现 45 d 预报结果与 CWAVE 有效波高数据的 R^2 拟合优度达到了 0.97,说明预报结果与波浪谱模型结果趋势一致,准确度较高。

参考文献:

[1] DEO M C, JHA A, CHAPHEKAR A S, et al. Neural networks for wave forecasting[J]. Ocean Engineering, 2001, 28(7): 889-898.

[2] MAKARYNSKY O, MAKARYNSKA D. Wave prediction and data supplementation with artificial neural networks[J]. Journal of Coastal Research, 2007, 23(4): 951-960.

[3] TSAI Chingpiao, LIN Chang, SHEN Jian. Neural network for wave forecasting among multi-stations[J]. Ocean Engineering, 2002, 29(13): 1683-1695.

[4] MAKARYNSKY O, PIRES-SILVA A A, MAKARYNSKA D, et al. Artificial neural networks in wave predictions at the west coast of Portugal[J]. Computers & Geosciences, 2005, 31(4): 415-424.

[5] JAIN P, DEO M C. Neural networks in ocean engineering[J]. Ships and Offshore Structures, 2006, 1(1): 25-35.

[6] MAHJOABI J, ETEMAD-SHAHIDI A. An alternative approach for the prediction of significant wave heights based on classification and regression trees[J]. Applied Ocean Research, 2008, 30(3): 172-177.

[7] ETEMAD-SHAHIDI A, MAHJOABI J. Comparison between M5' model tree and neural networks for prediction of significant wave height in Lake Superior[J]. Ocean Engineering, 2009, 36(15): 1175-1181.

[8] ZHOU Zhihua. Ensemble methods: foundations and algorithms[M]. Florida: CRC Press, 2012.

[9] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.

[10] SCHAPIRE R E. The strength of weak learnability[J]. Machine Learning, 1990, 5(2): 197-227.

[11] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[12] BREIMAN L. Stacked regressions[J]. Machine Learning, 1996, 24(1): 49-64.

[13] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.

[14] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators [J]. Neural Networks, 1989, 2(5): 359-366.

[15] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[16] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

[17] SHI Jian, ZHENG Jinhai, ZHANG Chi, et al. A 39-year high resolution wave hindcast for the Chinese coast: model validation and wave climate analysis[J]. Ocean Engineering, 2019, 183: 224-235.

[18] GONZALEZ R C, WOODS R E. Digital image processing[M]. London: Pearson, 2018.

[19] 余洋. 2016 年 11 月一次强寒潮南下天气特征分析[J]. 气象科技进展, 2018, 8(4): 22-28. (YU Yang. Analysis of the strong cold wave moving southwards in November 2016[J]. Advances in Meteorological Science and Technology, 2018, 8(4): 22-28. (in Chinese))

(收稿日期:2019-07-05 编辑:高建群)