基于随机森林回归的船舶特涂维修的日能耗预测

甘瑞平[1]，任新民[2]，姜军[2]，李鹏[3]，周小兵[1]
1. 云南大学信息学院，云南 昆明 650504；
2. 友联船厂（蛇口）有限公司，广东 深圳 518067；
3. 深圳市中科银狐机器人有限公司，广东 深圳 518216

摘要
特殊涂装（简称特涂）维修是修船工作的核心内容，能耗的预测是船舶智能能效优化中的一项重要任务。使用随机森林回归（RFR）模型对船舶特涂维修日能耗进行分析，去除异常值、随机化和标准化数据集，然后使用RFR模型对船舶日能耗历史数据进行训练拟和，利用带交叉验证的网格搜索优化RFR模型，使用优化后的RFR模型对船舶特涂维修日能耗数据进行分析，并与其他模型进行对比实验。结果表明，优化后的RFR模型预测效果优于多种其他模型，R^2值达93.25%，均方误差明显更低。

关键词
能耗预测；随机森林回归；LOF算法；船舶特涂

中图分类号：TP31 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024018

Prediction of daily energy consumption for ship special coating maintenance based on stochastic forest regression

GAN Ruiping[1], REN Xinmin[2], JIANG Jun[2], LI Peng[3], ZHOU Xiaobing[1]
1. School of Information Science & Engineering, Yunnan University, Kunming 650504, China
2. You Lian Dockyards (Shekou) Co., Ltd., Shenzhen 518067, China
3. Info Robot Co., Ltd., Shenzhen 518216, China

Abstract
Predicting energy consumption is an important task in the intelligent energy efficiency optimization of ship maintenance, with special coating (spec coat) being the core aspect. In this experiment, the random forest regression (RFR) model was employed to analyze the daily energy consumption of ship maintenance for special coating. The dataset was preprocessed by removing outliers, randomizing and standardizing the data. Subsequently, the RFR model was trained and fitted using historical data of daily energy consumption in ship maintenance. The RFR model was optimized using grid search with

cross-validation, and analysis of daily energy consumption data for ship special coating maintenance using optimized RFR model. Comparative experiments were conducted with other models. The results revealed that the optimized RFR model outperformed several other models, achieving an R-squared value of 93.25% and significantly lower mean squared error (MSE).

Key words
energy consumption prediction, random forest regression, LOF algorithm, ship special coating

0 引言

进入21世纪以来，国际经济联系日益紧密，国际贸易的发展超过了物质生产的发展，达到了前所未有的高度，已经成为衔接世界经济的重要纽带。随着物流行业的不断发展，越来越多的企业开始将物流作为其策略重点，物流产业链越来越复杂，需求越来越多元化，运输行业也随之不断升级。同时，跨境贸易的不断增加也推动了运输行业的发展。而船舶运输是运输行业最重要的一环，世界上通过船舶运输的货物约占总运输货物的80%，已然成为运输业的标杆。运载化学品和石油的船舶由于装载货物的特殊性，需要定期使用特别的涂层进行维修和保养（即船舶特涂业务）。因此，在修船厂对船舶进行维修不仅需要大量的材料与人员支撑，同时还需要庞大的能源消耗，包括维修设备的电力消耗等。为了提高修船厂的盈利空间，需要对船舶进行科学的维护和管理，控制和优化船舶维修和保养过程中的设备能源消耗。因此，开发船舶能效优化技术具有重要意义[1-2]，随着大数据和机器学习等技术的不断发展，以及“E航海”战略的施行，面向船舶的智能能效管理技术的研究与应用[3]将是大势所趋，通过记录并解读船舶修造过程中的数据，将其用于对船舶能耗影响要素的归纳分析，进而建立能耗预测模型，可以对船舶能耗进行详细的描绘和预测[4]。

1 相关研究

近年来，船舶特涂维修的发展研究主要集中于优化特涂维修流程。如绿色涂料技术，研发出许多无铅、低挥发性有机化合物（VOC）的环保涂料[5]，以减少对环境的影响；涂装工艺的改进，合理安排作业计划，减少涂装设备的启停次数，降低能源消耗；引入更高效的涂装设备和工具，如喷涂机和辊涂机，可以减少涂料的浪费和能源消耗。目前关于船舶特涂维修过程中能耗分析的研究较少，即如何统筹兼顾能耗影响的各个环节与因素，如杨永刚[6]提出了一种基于精益工具的能耗管理框架，根据车间生产流程和管理重点实施了精益能耗管理，通过应用精益能耗管理方法，分段涂装车间能够显著减少能源消耗。尽管上述方法在船舶特涂维修领域有一定的优化效果，但因为缺乏综合性的能耗分析，不能做到上述设备、材料以及管理的最优匹配，从而导致成本升高，限制了改进能源效率的潜力。因此，为了进一步提高改进能源效率的潜力，基于特涂维修能耗分析的结果，需要提出优化策略和建议，以减少船舶特种涂装维修过程中的能源消耗。现在的工艺已经使用许多方法来改善能源的消耗，研究人员综合了机器学习、深度学习等人工智能算法来预测涂装和船舶能源的消耗，这其中涉及涂装生产能耗预测、船舶油耗预测、船舶设备能耗预测以及船舶航行能耗预测。陆应康[7]提出了一种基于

贝叶斯优化的LSTM-CNN预测方法，利用LSTM预测能耗数据特征，并结合CNN进行能耗特征重构，将两种神经网络组合的模型作为汽车涂装车间的能耗预测模型。Bocchetti等[8]利用实际船舶监测数据建立了一种多元线性回归模型，通过该模型能够有效地预测船舶的能耗。Yan等[9]基于从实际船舶收集的能源效率数据创建了神经网络模型，从而能够评估和预测船舶能效水平。Pagoropoulos等[10]提出一种采用支持向量机的方法用于实现船舶能效评估，并经过结果剖析验证了该方法的有效性。BAL BEŞIKÇI等[11]以油船为对象，利用正午报告数据建立了基于人工神经网络的能效预测模型与决策系统。Wickramanayake等[12]基于机器学习的船舶能耗预测方法进行了系统的剖析，并比较了随机森林、梯度增强和神经网络方法在多变量时间序列的舰队能耗预测问题上的有效性。研究结果表明，采用随机森林技术可以得到更准确的预测结果。YANG等[13]提出可以采用基于遗传算法的灰箱模型来进行船舶能耗预测，这种方法是有效的。此外，孙双休等[14]提出了一种最小二乘支持向量机模型，用于分析和预测船舶集中空调系统的能耗。还有一种方法是采用k-means聚类算法对大量船舶航行数据进行聚类分析，孙峰等[15]使用这种方法得到了船舶主机在不同转速下负荷和油耗率的变化规律。Wang等[16]针对主机油耗受多种因素影响的问题，提出了一种基于绝对收缩选择算子（least absolute shrinkage and selection operator）的能耗回归模型，用于对船舶主机油耗进行预测和分析。同时，Yuan等[17]将人工神经网络和高斯过程应用于船舶能耗评价，并进行了实验研究，实验结果表明，通过速度优化可以有效地降低船舶的能耗。Leifson等[18]使用人工神经网络完成模型内部的参数确定工作，并在考虑风浪对船舶油耗影响的同时，加入附着物这一被人们广泛忽略的影响因素，使建立的白箱模型具有更高的适用性。叶睿等[19]使用一艘丹麦籍客滚轮的运营数据，基于人工神经网络建立了船舶油耗预测模型。船舶特种涂装维修能耗往往涉及多个输入特征及多个影响因素之间的高度复杂和非线性关系，而RFR模型可以有效处理多维特征、提供特征重要性排序和非线性关系，并能够捕捉输入特征之间的相互作用，这对于理解船舶特种涂装维修能耗的关键因素非常有帮助，通过分析特征重要性来识别对能耗影响最大的因素，从而指导优化策略和决策。同时，RFR模型对异常值和噪声具有较好的鲁棒性。本文在总结前人对船舶能耗的研究基础上，提出了一种基于随机森林回归模型的方法来预测船舶特涂维修日能耗的方法。

在特涂维修的能耗分析中，使用RFR算法会面临一些困难和挑战。能耗分析涉及的数据可能受到噪声、缺失值或异常值的影响，这些因素会对RFR模型的性能产生负面影响，因此在数据预处理阶段需要采用最优的方法处理这些问题。在特征选择上，特涂维修涉及的特征可能很多，需要仔细筛选具有预测能耗能力的特征，以提高模型的准确性。RFR模型在处理高维数据和复杂关系方面表现出色，能够处理大量特征变量和非线性关系。它能够捕捉特涂维修中各种因素对能耗的复杂影响，从而提供准确的能耗预测。同时，RFR模型的鲁棒性较高，这使其在特涂维修的应用中更加可靠。

2 数据预处理与特征选择

本文以友联船厂（蛇口）有限公司（以下简称友联船厂）的10艘进行特涂作业的

货轮作为数据源，包括萨法、托玛琳、坦桑石、丹娜、古姆达、西姆斯、黎明之光、海洋石油116、新道恩、雷姆，采集了从2021年9月19日到2022年11月28日的船舶特涂作业的相关数据，见**表1**，收集并整理成可用于船舶特涂维修能耗预测的数据。

采用船舶特涂业务相关变量作为能耗预测的影响因素，包括当日维修的舱室数据量（n_carbin）、当日维修的面积（area）、各类工序权重之和（press）、船舶类型（type）、各类特涂设备数（鼓风机数（Ebm）、吸砂机数（Essm）、除湿机数（Ed）、其他设备（Eo，包括高压冲水机、抽风机、热风机等使用频率较低、能耗较少的设备）和设备总数（Ea）。

2.1 数据预处理

2.1.1 局部异常因子算法（LOF算法）

首先对数据进行预处理，目前的特涂能耗主要是基于传感器采集的数据。由于船厂特涂部门实行的策略是船坞每进入一艘船就对其进行维修，所以多数时期的数据是对同步维修进度的采集与统计，数据会出现局部异常值和一定的顺序性的情况，而局部异常值会降低模型预测的准确率，因此需要对异常值进行检测与处理。本文采用局部异常因子算法（LOF算法）来检测并删除数据的异常值。数据的箱线图如**图1**所示。

局部异常因子算法是一种基于密度的经典异常值检测算法，被广泛应用于工业异常检测等领域。该算法通过计算每个数

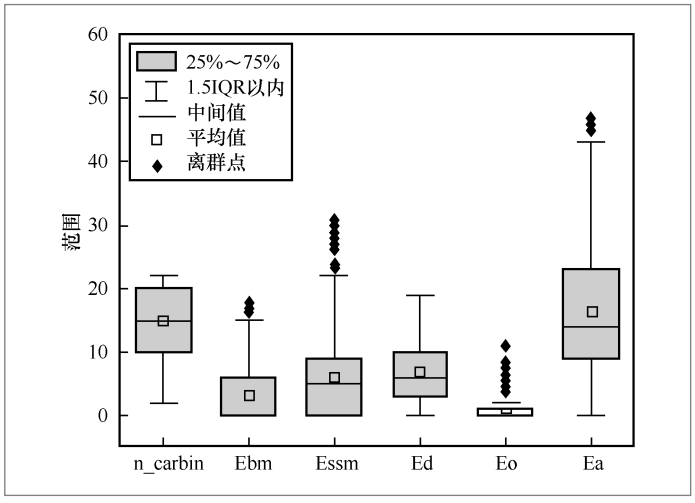


图1　数据的箱线图

表1　RFR模型分析中使用的船舶特涂信息

影响因素	舱室数量/个	施工面积/m^2	工序/道	设备总数/台	总能耗/（kW·h）
萨法	22	27 600	10	40	1 200 162.8
托玛琳	18	24 801	10	52	753 397.85
坦桑石	8	11 091	10	25	255 251.77
丹娜	22	27 454	10	53	828 235.33
古姆达	15	10 753	10	37	492 620.4
西姆斯	22	27 454	10	54	1 081 057.03
黎明之光	14	10 753	10	43	648 553.79
海洋石油116	16	177 573	8	96	1 861 483.63
新道恩	14	8 851	10	30	265 683.9406
雷姆	22	27 454	10	39	908 837.6865

据点周围的密度来确定异常点。具体来说，算法会计算每个数据点的局部密度，并将其与邻居的密度进行比较。如果该点的密度远低于其邻居，则该点被认为是异常点。由于该算法对样本分布非常敏感，因此能够有效地检测出异常点。

局部异常因子的构造主要涉及5个关键定义：样本的第k距离、样本的k距离邻域、样本的可达距离、样本的局部可达密度以及样本的局部异常因子的计算。这5个定义在构建局部异常因子时扮演着重要角色，并相互交织在一起，为异常检测提供了基本框架。其主要构造步骤如下。

样本的第k距离用于衡量样本与其邻域中第k个最近样本之间的距离，其定义为 $\mathrm{dist}_k(O)=\mathrm{dist}(O,U)$ ，$\mathrm{dist}_k(O)$ 为样本O的第k距离，表示该样本邻域内距离最近的第k个点U与O之间的距离为 $\mathrm{dist}(O,U)$ 。

样本的k距离邻域 $\mathrm{Neighbor}_k(O)$ 定义为与样本的距离小于等于 $\mathrm{dist}_k(O)$ 的都是样本O的k距离邻域。每一个样本至少会存在 k 个k距离邻域，因为与样本O的距离等于 $\mathrm{dist}_k(O)$ 的样本O可能存在数个。

样本的可达距离是指从样本到另一个样本的最短路径上的最大距离：

$$\begin{aligned}&\mathrm{reachable}-\mathrm{dist}_k(U,O)=\\&\max\left\{\mathrm{dist}_k(O),\mathrm{dist}(U,O)\right\}\end{aligned}\qquad(1)$$

样本点U到样本点O的可达距离为O的第k距离与两样本间距离的最大值。当样本点U靠近样本点O时，可达距离为较小值 $\mathrm{dist}_k(O)$ ；当两者之间距离较大时，可达距离为较大值 $\mathrm{dist}(U,O)$ 。因此，可达距离可以用来描述样本点周围的密度情况，即样本点周围距离较近的区域密度较大，相反则密度较小。

根据前文的3个定义，样本的局部可达密度则表示样本周围邻域内样本的平均可达距离的倒数：

$$\rho(U)=\frac{\left|\mathrm{Neighbor}_k(U)\right|}{\sum_{O\in\mathrm{Neighbor}_k(U)}\mathrm{reachable}-\mathrm{dist}_k(U,O)},\qquad(2)$$

可以观察发现当样本U与其邻域内的元素都很接近时，式(2)的分母值接近较小值 $\mathrm{dist}_k(O)$ ，此时 $\rho(U)$ 值较大，表示该样本处在一个密度较大的区域中。

样本的局部异常因子通过结合样本的局部可达密度和其邻域内其他样本的局部可达密度，用于度量样本相对于其邻域的异常程度：

$$\mathrm{LOF}_k(U)=\frac{\sum_{O\in\mathrm{Neighbor}_k(U)}\frac{\rho(O)}{\rho(U)}}{\left|\mathrm{Neighbor}_k(U)\right|}\qquad(3)$$

式（3）体现了样本U邻域内的样本和其密度比值的平均值。当$\mathrm{LOF}_k(U)$的值在1附近时，说明该样本与周围样本处在密度相似的区域当中；当$\mathrm{LOF}_k(U)$的值远小于1时，说明该样本处于一个高密度区域之中；当$\mathrm{LOF}_k(U)$的值远大于1时，说明该样本处于一个密度低的区域，有可能是一个异常样本。

这些定义共同构成了局部异常因子的基本构建要素，并在异常检测领域发挥着重要的作用。

2.1.2 数据标准化

因为数据在采集时有时间上的先后顺序，在模型学习时会被当作一种特征学习，从而导致过拟合，为了避免这种情况的发生，本文对数据进行随机化处理以消除这种效应，使模型能够学习到更多的信息，从而提高模型的准确性和稳定性。同时，在对数

据进行处理的过程中发现，数据的取值范围较大且数据分布不均匀，这些对模型的拟和和评估效果都有影响，尤其是对模型的评估指标影响更明显。因此，为了使评估指标具有直观的意义，通常需要对数据进行标准化处理。数据密度曲线如**图2**所示。

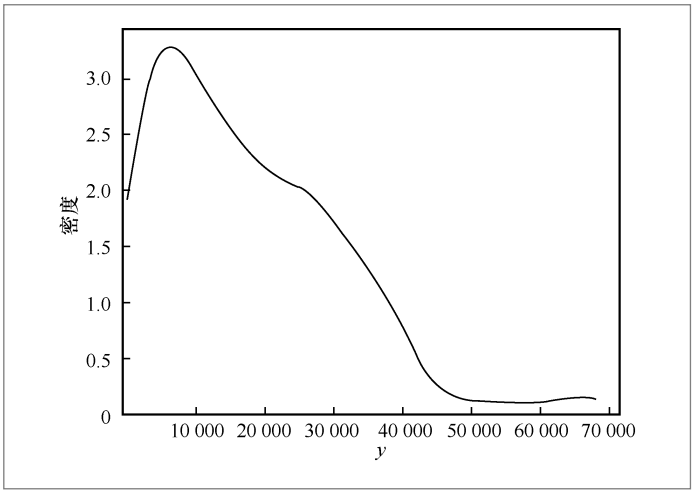


图 2　数据密度曲线

本文采用Z-score标准化对数据做标准化处理，它基于数据的均值和标准差，将原始数据转换为具有均值为0、标准差为1的正态分布。具体而言，对于给定的数据集，Z-score的标准化过程如下。

计算数据的均值（mean）和标准差（standard deviation即std）：

$$\text{mean} = \left(\frac{1}{n}\right)\sum X \qquad (4)$$

$$\text{std} = \sqrt{\left(\left(\frac{1}{n}\right)\sum(X - \text{mean})^2\right)} \qquad (5)$$

其中，n是数据集的样本数量，X是数据集中的每个样本。

对每个数据样本进行标准化转换：

$$\text{X\_scaled} = (X - \text{mean}) / \text{std} \qquad (6)$$

其中，X_scaled是标准化后的值，X是原始值，mean是数据的均值，std是数据的标准差。通过Z-score进行标准化处理后，数据更具有可比性和可解释性。

2.2 特征工程

船舶特涂维修能耗影响因素是通过在友联船厂了解船舶特涂业务流程及能耗预测相关领域的研究经验选择的。船舶的尺寸是一个重要的因素，较大的船舶拥有更多的舱室和更大的维修面积，需要更多的涂料和施工时间，因此会有更高的能耗。船舶特涂维修过程包含10个工序，分别是搭架、预打砂吸砂、结构处理、冲洗化清、干燥、主打砂吸砂、第一度油漆、第二度油漆、底部完工修补和拆架。每个工序都对应着固定的权重值。在设备方面，鼓风机在搭架、预打砂吸砂、结构处理、冲洗化清、干燥和主打砂吸砂工序中经常被使用。吸砂机仅用于预打砂吸砂和主打砂吸砂这两个工序。除湿机主要用于冲洗化清、干燥、第一度油漆和第二度油漆。高压冲水机虽然用于预打砂吸砂和主打砂吸砂，但使用频率较低。抽风机和热风机的使用工序不固定且频率较低。因此，船舶每日维修的能耗影响因素主要有当日维修的舱室数量（n_carbin）、当日维修的面积（area）、各类工序权重之和（press）、船舶类型（type）、鼓风机数（Ebm）、吸砂机数（Essm）、除湿机数（Ed）、其他设备数（Eo）和设备总数（Ea）。为了提高所选模型的拟和度，本文采用经典的皮尔逊算法来进行特征筛选，保证模型不会因为特征的筛选而欠拟合，通过实验对比，去掉与日能耗（y）呈负相关的特征。根据**图3**所示的特征相关性的值，选择舱室数据量（n_carbin）、当日维修的面积（area）、各类工序权重之和（press）、船舶类型（type）、各类特涂设备数（吸砂机数（Essm）、除

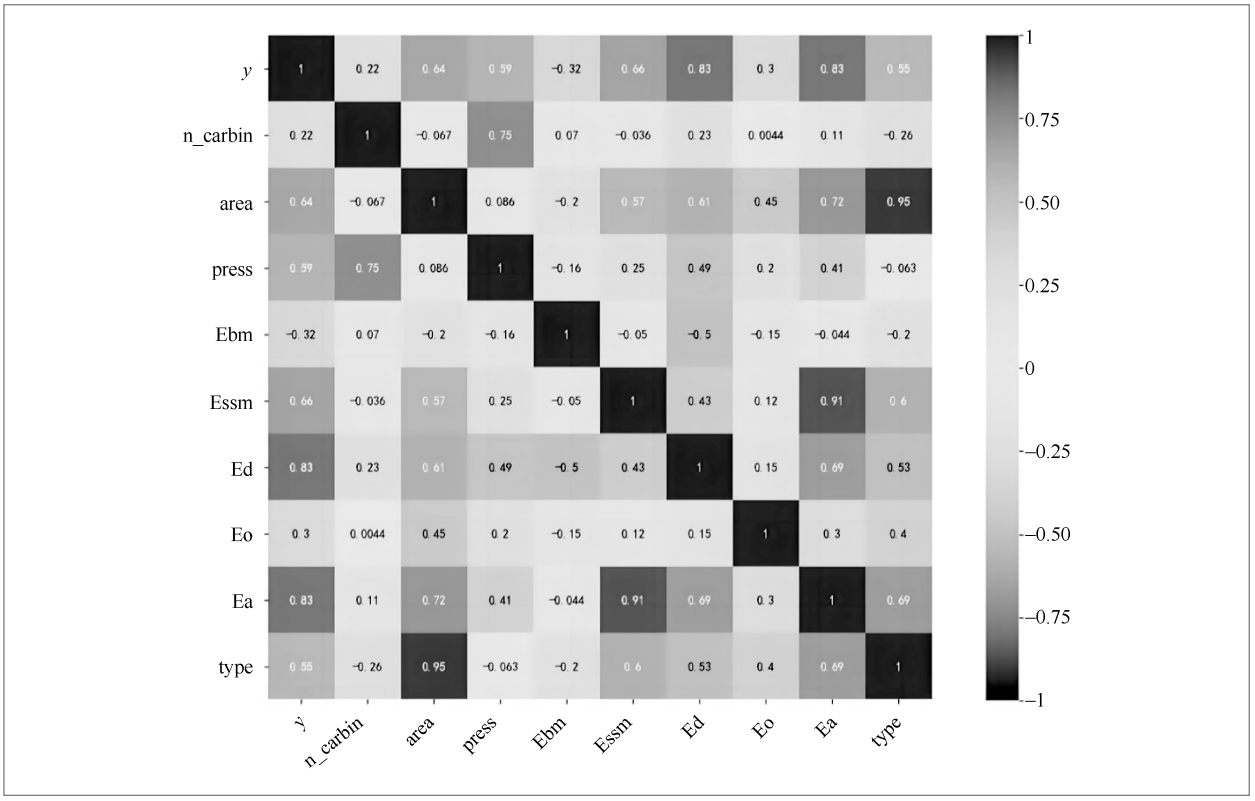


图 3 各变量之间的热力相关图

湿机数（Ed）、其他设备数（Eo，包括高压冲水机、抽风机、热风机等使用频率较低，能耗较少的设备）和设备总数（Ea）作为特征。

3 系统模型设计

3.1 随机森林回归

随机森林回归模型是一种基于自助法（bootstrap）重复抽样技术的模型，它由多棵决策树组成。设随机森林模型的输入向量为X，输出向量为Y，通过随机的行列向量形成组合模型$\{h(X,\vartheta_k),k=1,L,p\}$，即决策树能够将数值型变量转化为预测因子，从而实现随机森林模型的构建。该模型是非线性的，并且可以处理多元变量。每棵决策树$\{h(X,\vartheta_k)\}$的值除将它们组合在一起并取均值以获得最终的预测值外，组合模型还应满足每个模型基于独立的训练集形成随机森林的条件。随机抽取的向量Y与数值型预测向量$h(X)$的推广误差均方为$E_{X,Y}(Y-h(X))$。

随机森林回归有以下特征。

（1）在进行回归时，当随机森林中树的数量无穷大时，式（5）将在所有位置上成立。

$$E_{X,Y}(Y-av_k h(X,\vartheta_k))^2 \rightarrow E_{X,Y}(Y-E_{\vartheta}h(X,\vartheta))^2 \quad (7)$$

因此，随机森林回归模型的计算式为：$E_{\vartheta}h(X,\vartheta)$，使用模型评估时，$k$可以设定为无限大：

$$Y=av_k h(X,\vartheta_k) \quad (8)$$

在此情况下，对树的误差进行具体分析。假设整个随机森林的泛化误差表示为PE^*，则单个回归树的平均泛化误差可以用来表示为：

$$PE^*(tree)=E_\vartheta E_{X,Y}(Y-h(X,\vartheta))^2 \quad (9)$$

通过计算每棵树的平均泛化误差，可以了解单个回归树的预测性能。这个平均泛化误差反映了单个树的预测能力，而随机森林的优势在于通过组合多个树的预测结果来降低整体的泛化误差。因此，通过分析树的误差，可以评估和优化随机森林模型的性能。

（2）如果对于所有的ϑ，都有$E(Y)=E_x h(X,\vartheta)$，则：

$$PE^*(forest)\leqslant\bar{\rho}PE^*(tree) \quad (10)$$

其中，$\bar{\rho}$为$Y-h(X,\vartheta)$与$Y-h(X,\vartheta^*)$余项之间的加权相关系数，ϑ与ϑ^*两者之间彼此是相互独立的。这表明随机森林模型的泛化误差相比于单独的分类树的泛化误差降低了$\bar{\rho}$倍，且引入随机化变量ϑ、ϑ^*可以降低$\bar{\rho}$。因此，随机森林模型在降低泛化误差方面表现出更强的能力。

随机森林模型采用随机样本选择和基于多个决策树的投票机制，因此其具有强大的抗干扰能力。这种抗干扰能力使随机森林模型在船舶特涂作业日能耗预测模型中具有很高的适用性。此外，随机森林模型还具有无过度拟合现象等优点，这些优点进一步加强了其在船舶特涂作业日能耗预测模型中的应用前景。

3.2 带交叉验证的网格搜索

为了使RFR模型的预测效果达到最优，需要对模型中的一些超参数进行设置。然而，这些超参数的选择往往依赖于研究者的经验，而不是基于科学严谨性。这意味着模型的性能可能受到超参数选择的影响，而这些选择缺乏科学依据。因此，为了提高模型的科学严谨性和预测效果，本文采用带交叉验证的网格搜索方法（GridSearchCV）。该方法通过穷举所有可能的参数组合来寻找最优的超参数组合，同时还包含了交叉验证，可以更加准确地评估模型的性能。在使用该方法时，只需要指定参数的取值范围，就可以保证找到精度最高的参数。由于本文所用的数据集较小，因此选择了GridSearchCV来寻找最优超参数组合。

3.3 模型结构

特涂维修设备的用电量具有复杂的不确定性。为了能够准确地预测船舶特涂维修单船的用电量，本文对数据集用LOF算法去除数据中的异常值并标准化，然后对数据进行随机化。使用皮尔逊系数筛选特征，在保证模型预测精度的同时减少模型搭建的复杂度。同时，GridSearchCV优化算法可以有效地解决RFR模型由于不合理的参数设置而导致的效果不佳的问题，它可以科学地处理全局优化问题。基于LOF算法和GridSearchCV算法与RFR模型结合，提出了一种船舶特涂维修日能耗预测组合模型，其模型结构如**图4**所示。

3.4 评估指标

均方误差（mean squared error，MSE）是一种常用的衡量模型预测结果与真实值之间差异的方法。MSE是一个非负的值，它的值越小，表示模型的预测结果与真实值之间的差异越小。

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-y_{predi})^2 \quad (11)$$

平均绝对误差（mean absolute er-

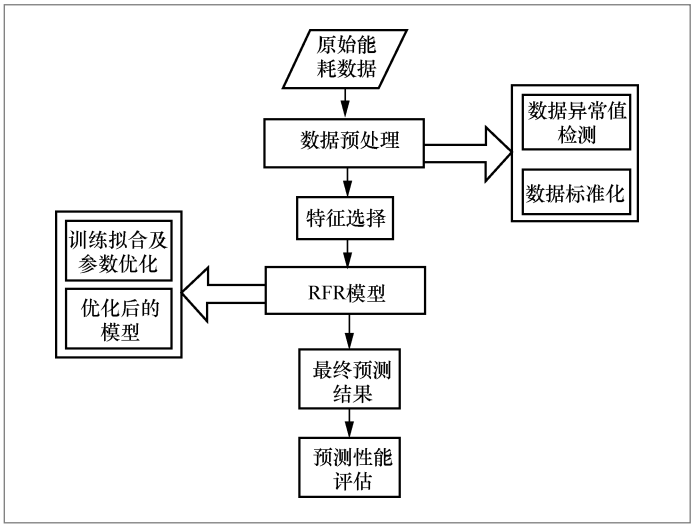


图 4 船舶特涂维修日能耗预测组合模型结构图

ror，MAE）是预测值与实际值之间的绝对差异的平均值。与MSE不同，MAE不考虑误差的平方，而是使用绝对值。MAE的值越小越好。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-y_{\mathrm{predi}}\right| \tag{12}$$

R平方（R-squared，R^2）是一种用于评估模型拟合优度的标准化指标，便于不同模型之间的比较，它表示模型解释了因变量变异性的比例。R^2的取值范围为0到1，越接近1表示模型的预测效果越好。

$$R^2=1-\left[\left(\sum_{i=1}^{n}(y_i-y_{\mathrm{predi}})^2\right)/\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)\right] \tag{13}$$

解释方差（explained variance，EV）是一种用于评估模型预测能力的指标，它表示模型对因变量的变异性解释了多少，取值范围在0和1之间，越接近1表示模型的预测效果越好。

$$\mathrm{EV}=1-\frac{\mathrm{Var}\left\{y_i-y_{\mathrm{predi}}\right\}}{\mathrm{Var}\left\{y_i\right\}} \tag{14}$$

其中，y_i、y_{predi}分别表示目标变量的实际值与预测值，n表示预测样本总数，$\bar{y}$为测试数据集中目标变量的平均值。

4 实验及结果

4.1 实验描述

在进行船舶特涂维修的日能耗分析实验时，采用随机森林回归（RFR）作为一种模型，并将其与其他常用的回归模型进行比较，以扩大实验的对比范围。其他回归模型包括传统的多元线性回归（LR）、多项式回归（PR）、支持向量机回归（SVR）和极限梯度提升（XGBR）模型。首先对数据集用LOF算法去除数据中的异常值，然后对数据进行标准化和随机化，使用皮尔逊系数筛选特征，在保证模型预测精度的同时降低模型搭建的复杂度。通过Grid-SearchCV优化算法将这5个模型的参数都调整至最优值，最后采用五折交叉验证的方法输出均方误差、平均绝对误差、解释方差和R^2，对所有模型进行比较分析。

4.2 模型参数设置

经过优化后，各个模型的最优超参数设置见**表2**，未涉及的模型参数一律采用默认值。

4.3 实验结果

本实验的目标是通过对比不同能耗预测模型的预测效果，验证RFR模型在预测能耗方面是否具有优越性。为此，本实验采用了五折交叉验证的方法对所有模型进行评估。使用五折交叉验证能够更准确地评估模型的性能，并对其在不同数据集上的泛化能力进行全面的评估，以确保结

果具有统计意义。与传统的多元线性回归（LR）、多项式回归（PR）和支持向量机回归（SVR）等模型相比，RFR模型具有更强的非线性建模能力和更好的抗噪声能力。而与近年来广受关注的极限梯度提升（XGBR）模型相比，RFR模型在模型训练的速度和模型解释性方面更具有优势。各个模型预测效果见**表3**，本实验所用的模型RFR的最优结果为：MSE仅有0.067，R^2为93.25%，EV为93.29%，MAE为0.181。

从**表3**可以看出LR、SVR和PR的MSE均为0.13~0.155，远高于RFR的0.067，而XGBR的MSE为0.080，高出RFR的MSE值16%左右；并且RFR的R^2与EV的值为93%左右，而LR、SVR和PR的R^2与EV值均为84%~87%，比RFR的R^2与EV的值低了6%~9%，可见RFR表现优秀；另外LR、SVR和PR的MAE值分别为0.283、0.267和0.238，而RFR的MAE值为0.181，LR、SVR和PR的MAE值高出RFR的MAE31%以上，XGBR的MAE为0.192，略高于RFR的值。以上结果表明随机森林回归在船舶特涂维修的日能耗分析上具有更好的性能。这意味着RFR模型能够更准确地预测船舶特涂维修的能耗情况，并提供更可靠的结果。

在实验过程中，虽然使用了五折交叉验证后的结果，但因为数据划分过程的随机性，每一次的迭代都会产生不同的结果，为体现本实验模型优异的泛化性能，本文先对比了5次迭代后不同模型的MSE结果，如**图5**所示。

从**图5**可以看出，RFR模型的均方误差每次迭代结果均小于其他模型，并且其值大部分在0.08以下，这说明RFR模型具有很强的预测能力，能够在预测能耗时保持较小的误差。然后对比了5次迭代后不同模型的R^2、EV和MAE结果，如**图6**所示，从图中可以看出，RFR模型每次迭代的结果均高于其他模型，且其R^2和EV的值均在0.9以上，而MAE值均在0.2以下，这说明RFR模型具有很好的预测能力，能够在预测能耗时保持较小的误差；并且在**图6**的3个子图中，RFR模型的评估指标值在每一次迭代后，波动都非常小，足见RFR模型的稳定性。综上所述，随机森林回归模型所有评价指标均为最优，充分证明了其在能耗预测方面的优越性和在能耗

表 2　各模型最优超参数

RFR		XGBR		SVR	
参数名	参数值	参数名	参数值	参数名	参数值
n_estimators	200	n_estimators	191	C	500
max_depth	14	max_depth	9	kernel	linear
max_features	5	max_features	5	cache_size	500
random_state	50	random_state	100		
		eta	0.3		

表 3　模型评估指标对比

模型	MSE	R^2	EV	MAE
LR	0.137	85.95%	86.05%	0.283
PR	0.136	86.45%	86.98%	0.238
RFR	0.067	93.25%	93.29%	0.181
SVR	0.151	84.62%	84.71%	0.267
XGBR	0.080	91.94%	92.11%	0.192

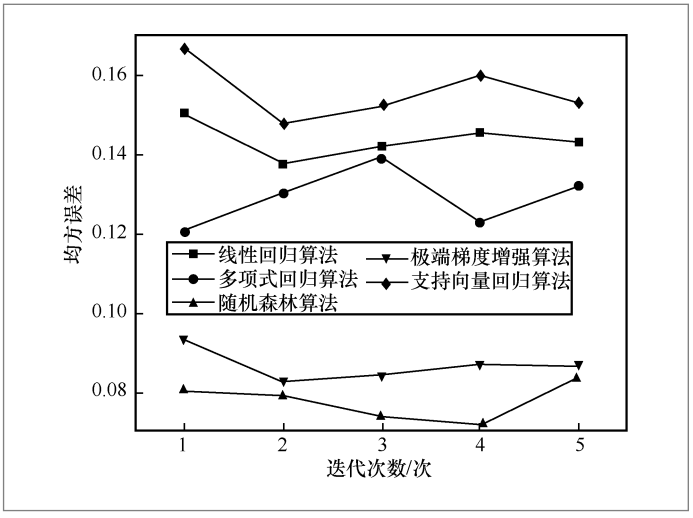


图 5　各模型均方误差的精度对比

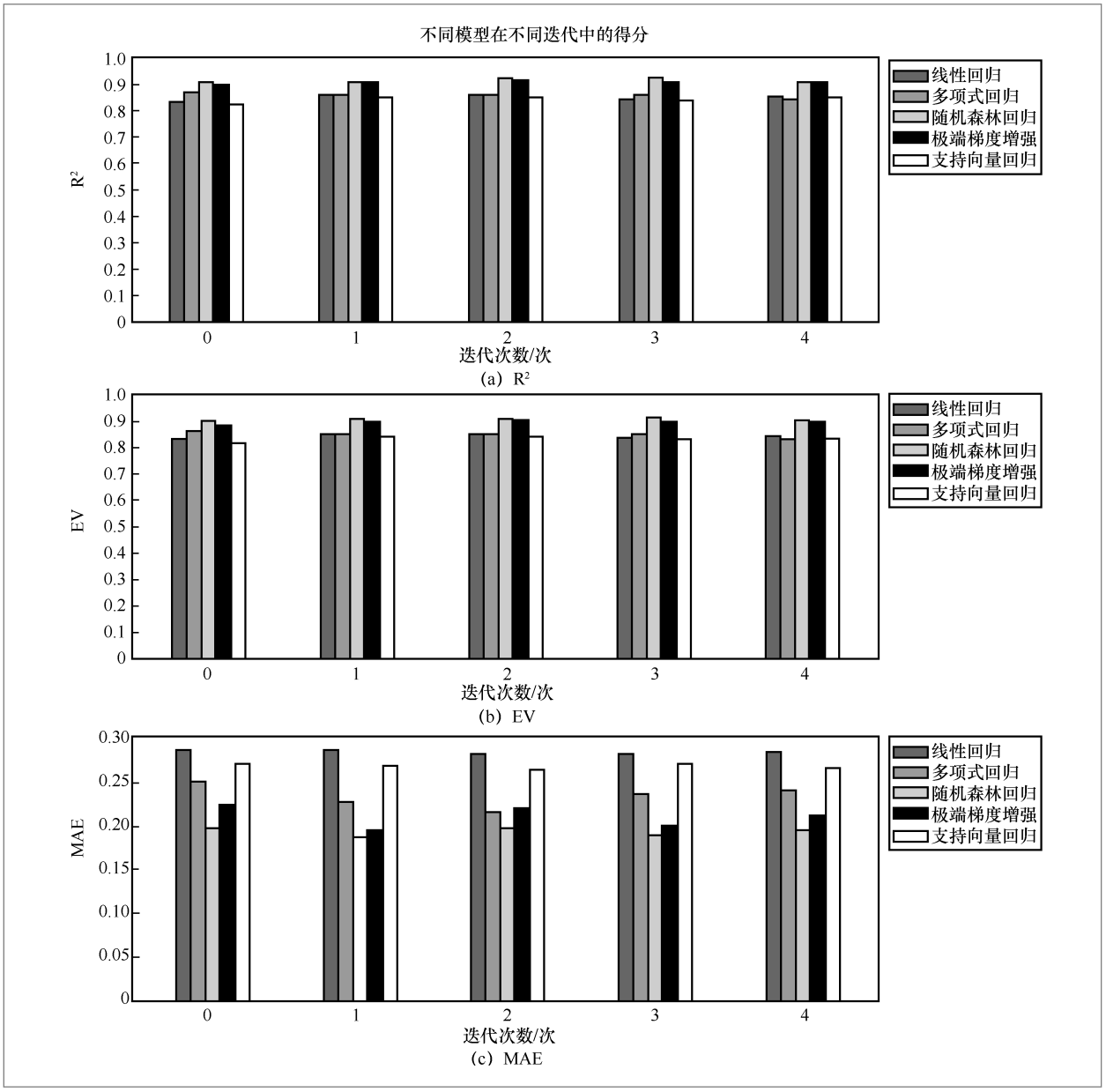


图 6 各模型的精度对比

预测研究方向具有广泛的适用性。

此外，为了增加RFR模型预测结果的可解释性，本文采用基于博弈论的SHAP[20]方法与RFR模型相结合，计算特征变量的贡献度，在综合考虑算法速度和准确率的情况下，本文设定决策树个数K=200，特征总数M=8。通过运行程序得到8个特征的贡献度，并根据这些贡献度绘制了特征分析图，如**图7**所示。**图7**中的横轴表示对船舶特涂工序能耗的正负影响，其中正值表示对能耗有正向影响，负值表示对能耗有负向影响。每个点代表了每个特征在每个样本上的SHAP值。通过观察特征分析图和特征的SHAP值，可以得出以下结论：排名第一的变量SHAP值大于0的点少于小于0的点，说明除湿机数（Ed）

对船舶特涂工序能耗的影响是负向的，即在满足施工完成量的要求上，设备使用数越多，其能耗反而会变少；排名第二的变量是设备总数（Ea），其对能耗的影响既有正向影响，也有负向影响，因此在整个数据集中设备总数（Ea）的SHAP值正负分布均匀；排名第三的变量是各类工序权重之和（press），其SHAP值大于0的点少于小于0的点，说明船舶在相比于前一日的特涂工序上，其每日的特涂工序都有变化；排名最后的变量船舶类型（type）的SHAP值接近0，这是由于特涂作业中不同类型的船舶在同时期的涂装工序的差异性不大。

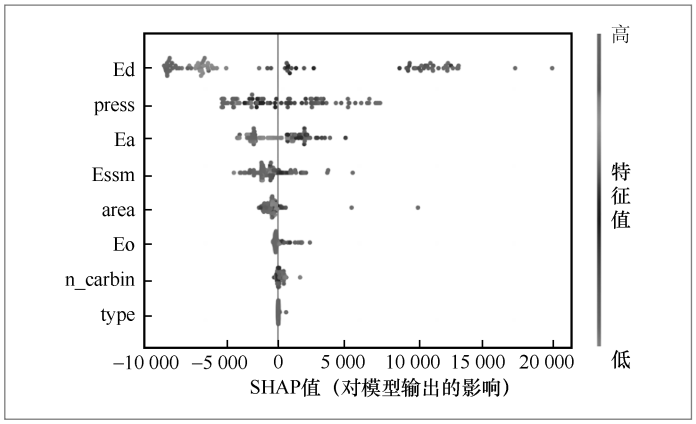


图 7　SHAP 特征分析

5 结束语

本文提出了一种基于随机森林回归的船舶特涂维修的日能耗预测方法。该方法通过多源传感器采集船舶特涂维修能耗及其影响因素数据，然后对其进行预处理、特征选择和模型匹配化寻优等步骤。在预处理阶段，采用局部异常因子算法检测并删除异常值，并对数据进行随机化以及标准化处理。接着进行特征选择，去除冗余特征。然后使用GridSearchCV对随机森林回归模型进行匹配化寻优，使模型与当前输入数据适配性最好。最后将处理好的数据输入优化好的随机森林回归模型中，对能耗数据进行预测。

为了验证该方法的性能，本文进行了实验并与其他方法进行比较。实验结果表明，相比其他模型，随机森林回归模型具有更高的预测精度和鲁棒性，以及更稳定的预测性能。其中，使用的3种评估指标都取得了最优结果（MSE仅有0.067，R^2为93.25%，EV为93.29%）。这表明该方法能够更准确地预测船舶特涂维修的日能耗，具有很高的应用价值。本文提出的模型可以为利用机器学习探索能耗预测提供一个有益的研究思路，在工业生产消耗的能源研究中有优异的效果。同时，采用基于博弈论的SHAP方法计算变量的特征贡献度，分析了其与能耗的关系。结果表明，除湿机数（Ed）对模型预测的贡献度最大，船舶类型（type）对模型预测的贡献度最小。

尽管本文的模型在船舶特涂维修能耗预测中取得了优秀的表现，但是还有很大的提升空间。例如，可以进一步挖掘船舶特涂维修能耗的影响因素，并探索更有效的数据预处理和特征选择方法，以提高预测精度和稳定性。此外，随着后续工作的进行，采集的实验数据量将会增加，可以通过引入更多的特征和优化算法来提高预测性能。除船舶特涂维修能耗预测之外，该技术还可以在其他领域得到应用。例如，在建筑物、工厂等设施的维修能耗预测中，该技术可以提高设施的维护效率和降低维护成本。另外，该技术还可以应用于其他能耗预测场景，例如交通运输、船舶运营行业等。

参考文献:

[1] WANG K, YAN X P, YUAN Y P, et al.

Real-time optimization of ship energy efficiency based on the prediction technology of working condition[J]. Transportation Research Part D: Transport and Environment, 2016, 46: 81-93.

[2] 严新平, 刘佳仑, 范爱龙, 等. 智能船舶技术发展与趋势简述[J]. 船舶工程, 2020, 42(3): 15-20.
YAN X P, LIU J L, FAN A L, et al. Brief introduction to the development and trend of intelligent ship technology[J]. Ship Engineering, 2020, 42(3): 15-20.

[3] 王凯, 胡唯唯, 黄连忠, 等. 船舶智能能效优化关键技术研究现状与展望[J]. 中国舰船研究, 2021, 16(1): 181-192, 199.
WANG K, HU W W, HUANG L Z, et al. Research progress and prospects of ship intelligent energy efficiency optimization key technologies[J]. Chinese Journal of Ship Research, 2021, 16(1): 181-192, 199.

[4] CHAAL M. Ship operational performance modelling for voyage optimization through fuel consumption minimization[D]. Malmo: World Maritime University, 2018.

[5] 盛同范, 叶峰, 王晨昊, 等. 船舶行业室内涂装VOCs治理和节能方案的探讨[J]. 船舶物资与市场, 2021, 29(6): 53-54.
SHENG T F, YE F, WANG C H, et al. Discussion on VOCs treatment and energy saving scheme of indoor painting in shipbuilding industry[J]. Marine Equipment/Materials & Marketing, 2021, 29(6): 53-54.

[6] 杨永刚. 分段涂装车间精益能耗管理应用研究[D]. 上海: 上海交通大学, 2016.
YANG Y G. Study on the lean energy management in the Painting & Blasting workshop[D].Shanghai: Shanghai Jiao Tong University, 2016.

[7] 陆应康. 涂装车间能耗预测方法研究与系统实现[D]. 武汉: 武汉理工大学, 2020.
LU Y K. Research on energy consumption prediction method of painting workshop and its system realization[D]. Wuhan: Wuhan University of Technology, 2020.

[8] BOCCHETTI D, LEPORE A, PALUMBO B, et al. A statistical approach to ship fuel consumption monitoring[J]. Journal of Ship Research, 2015, 59(3): 162-171.

[9] YAN X P, SUN X, YIN Q Z. Multiparameter sensitivity analysis of operational energy efficiency for inland river ships based on backpropagation neural network method[J]. Marine Technology Society Journal, 2015, 49(1): 148-153.

[10] PAGOROPOULOS A, MØLLER A H, MCALOONE T C. Applying multi-class support vector machines for performance assessment of shipping operations: the case of tanker vessels[J]. Ocean Engineering, 2017, 140: 1-6.

[11] BAL BEŞIKÇI E, ARSLAN O, TURAN O, et al. An artificial neural network based decision support system for energy efficient ship operations[J]. Computers & Operations Research, 2016, 66: 393-401.

[12] WICKRAMANAYAKE S, DILUM BANDARA H M N. Fuel consumption prediction of fleet vehicles using Machine Learning: a comparative study[C]// Proceedings of the 2016 Moratuwa Engineering Research Conference. Piscataway: IEEE Press, 2016: 90-95.

[13] YANG L Q, CHEN G, RYTTER N G M, et al. A genetic algorithm-based grey-box model for ship fuel consumption prediction towards sustainable shipping[J]. Annals of Operations Research, 2019.

[14] 孙双林, 杨倩. 基于最小二乘支持向量机的船舶集中空调系统能耗预测[J]. 舰船科学技术, 2020, 42(2): 184-186.
SUN S L, YANG Q. Energy consumption prediction of ship central air conditioning system based on least square support vector machine[J]. Ship Science and Technology, 2020, 42(2): 184-186.

[15] 孙峰, 黄连忠, 刘伊凡, 等. 一种应用数据挖掘技术评估柴油机性能的方法[J]. 大连海事大学学报, 2017, 43(3): 83-88.
SUN F, HUANG L Z, LIU Y F, et al.

A method of evaluating diesel engine performance by using data mining technology[J]. Journal of Dalian Maritime University, 2017, 43(3): 83–88.

[16] WANG S Z, JI B X, ZHAO J S, et al. Predicting ship fuel consumption based on LASSO regression[J]. Transportation Research Part D: Transport and Environment, 2018, 65: 817–824.

[17] YUAN J, WEI S M. Comparison of using artificial neural network and Gaussian process in ship energy consumption evaluation[J]. DEStech Transactions on Environment, Energy and Earth Sciences, 2019.

[18] LEIFSSON L Þ, SÆVARSDÓTTIR H, SIGURÐSSON S Þ, et al. Grey-box modeling of an ocean vessel for operational optimization[J]. Simulation Modelling Practice and Theory, 2008, 16(8): 923–932.

[19] 叶睿, 许劲松. 基于人工神经网络的船舶油耗模型[J]. 船舶工程, 2016, 38(3): 85–88.
YE R, XU J S. Vessel fuel consumption model based on neural network[J]. Ship Engineering, 2016, 38(3): 85–88.

[20] 丁珍妮, 陈华友, 朱家明. 三角模糊数组合预测模型及其Shapley值近似解法[J]. 统计与决策, 2019, 35(24): 68–72.
DING Z N, CHEN H Y, ZHU J M. Combined forecasting model of triangular fuzzy number and its approximate solution method of Shapley value[J]. Statistics & Decision, 2019, 35(24): 68–72.

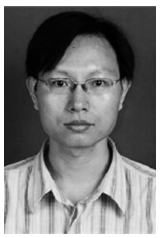
作者简介

甘瑞平（1994– ），男，云南大学信息学院硕士生，主要研究方向为工业大数据建模、自然语言处理。

任新民（1966– ），男，友联船厂（蛇口）有限公司高级工程师，主要研究方向为物联网、大数据。

姜军（1976– ），男，友联船厂（蛇口）有限公司高级工程师，主要研究方向为物联网、大数据。

李鹏(1984-),男,就职于深圳市中科银狐机器人有限公司,主要研究方向为数据采集与分析、智能控制。



周小兵(1975-),男,博士,云南大学信息学院教授,主要研究方向为自然语言处理、计算机应用。

收稿日期: 2023-09-27

基金项目: 深圳大学稳定保障计划项目(No. 20200829114939001);深圳信息职业技术学院校级创新科研团队项目(No. TD2020E001);珠江三角洲水资源配置工程科研项目(No. CD88-QT01-2022-0068)

Foundation Item: Shenzhen University Stability Support Plan (No. 20200829114939001), Project of Shenzhen Institute of Information Technology School-level Innovative, Scientific Research Team (No. TD2020E001), The Pearl River Delta Water Resources Allocation Engineering Scientific Research Project (No. CD88-QT01-2022-0068)