

水利水运工程学报

HYDRO-SCIENCE AND ENGINEERING

融合SHAP和TSO-XGBoost模型的水路货运量预测

温泉, 余玉欢, 庄尚德, 牟军敏

Waterway freight volume prediction based on SHAP and TSO-XGBoost model

WEN Quan, YU Yuhuan, ZHUANG Shangde, MOU Junmin

在线阅读 View online: <https://doi.org/10.12170/20240114001>

您可能感兴趣的其他文章

Articles you may be interested in

基于WOA-VMD-XGBoost的混凝土坝变形预测

Deformation prediction of concrete dams using WOA-VMD-XGBoost methodology

水利水运工程学报. 2024(3): 146 <https://doi.org/10.12170/20230425002>

基于XGBoost算法的堆石料南水模型参数反演及应用

NHRI model parameter inversion and application of rockfill based on XGBoost

水利水运工程学报. 2023(3): 111 <https://doi.org/10.12170/20221019002>

水工结构变形预测模型构建与解释

Forecasting and analysis of deformation in hydraulic structures

水利水运工程学报. 2024(2): 125 <https://doi.org/10.12170/20230418001>

人工鱼群算法在河间市水资源优化配置中的应用

Application of artificial fish swarm algorithm in optimal allocation of water resources in Hejian County, China

水利水运工程学报. 2021(3): 74 <https://doi.org/10.12170/20200412001>

基于RBF代理模型和粒子群算法的水交换优化研究

Optimization of water exchange based on RBF surrogate model and particle swarm optimization

水利水运工程学报. 2020(5): 40 <https://doi.org/10.12170/20190724002>

VMD-PSO-LSTM模型的日径流多步预测

Research on multi-step forecast of daily runoff based on VMD-PSO-LSTM model

水利水运工程学报. 2023(4): 81 <https://doi.org/10.12170/20220208001>



扫码进入官网, 阅读更多精彩文章



关注微信公众号, 获得更多资讯信息

DOI:10.12170/20240114001

温泉,余玉欢,庄尚德,等.融合 SHAP 和 TSO-XGBoost 模型的水路货运量预测[J].水利水运工程学报,2024(6): 86-96. (WEN Quan, YU Yuhuan, ZHUANG Shangde, et al. Waterway freight volume prediction based on SHAP and TSO-XGBoost model[J]. Hydro-Science and Engineering, 2024(6): 86-96. (in Chinese))

融合 SHAP 和 TSO-XGBoost 模型的水路货运量预测

温 泉¹,余玉欢²,庄尚德²,牟军敏³

(1. 武汉软件工程职业学院(武汉开放大学) 商学院,湖北 武汉 430079; 2. 长江航道局,湖北 武汉 430023; 3. 武汉理工大学 航运学院,湖北 武汉 430063)

摘要: 水路货运量需求受诸多因素影响,长江干线中游“645”工程实施后,航道通航条件得到了明显改善,为了更好地分析工程实施后货运量变化趋势,提出一种新的水路货运量预测模型。首先,采用二次插值法和 KNN 反距离权重插值法解决高维面板数据中时间粒度不统一与缺失问题,利用层次聚类 and SHAP 值的可解释性综合筛选关键影响因素特征序列,降低预测模型输入数据的维度和规模,引入 Halton 低差异序列和准反射学习策略(QRBL)大幅提升金枪鱼群优化算法(TSO)的寻优效能,增强 TSO 算法对极限梯度提升(XGBoost)模型中决策树数量、决策树的深度、学习速率等决定模型拟合能力的超参组合寻优效果。结果表明,新模型预测精度显著优于对比模型,可更好地适用于多特征影响因素下的水路货运量预测研究。

关键词: 金枪鱼群优化算法(TSO); 极限梯度提升(XGBoost); 水路货运量

中图分类号: U695

文献标志码: A

文章编号: 1009-640X(2024)06-0086-11

长江航道连接重庆、武汉、南京、上海等大型交通枢纽城市及其他众多重要城市,在区域发展总体格局中具有重要战略地位。长江干线“645”工程作为国家“十三五”规划纲要和长江经济带发展规划纲要的重要内容,通过航道整治工程逐步实现武汉至安庆 6.0 m、宜昌至武汉 4.5 m 通航水深,实现万吨海轮常年通达武汉,有效解决长江干线航道“两头深、中间浅”的问题。2021 年 9 月,交通运输部印发《水运“十四五”发展规划》明确了“十四五期,水运需求将总体保持增长态势,呈现高基数、中低速增长的特点”;规划还提出“加快水运大通道扩能升级、大力发展大宗散货水铁联运”等系列重要措施。梳理现有研究成果,相关领域研究者针对水路货运量预测开展了广泛研究。臧文亚等^[1]将灰色 GM 模型与马尔科夫状态转移矩阵相结合对重庆水运货运量进行预测;李文杰等^[2]采用回归分析法、弹性系数法、指数平滑法和投入产出法分别对水路货运量外部影响因素进行了研究,综合预测了涪陵至丰都河段水路货运量变化趋势;温泉^[3]采用四阶段法对长江干线中游货运结构、货运量变化及船型发展趋势进行了综合研究;张浩等^[4]提出基于长江上游区域产业结构测度的三峡枢纽过坝货运量预测模型;戎陆庆^[5]利用灰色 Verhulst 及新陈代谢模型预测了 2015—2017 年广西水路货运量,并采用灰色关联度法分析了影响水路货运量的关键因素。

机器学习模型近年来逐步成为货运量预测领域的热点,陈琛等^[6]提出了基于 FR-NN 构建考虑时空因素的水路货运量预测模型;李红娟等^[7]采用季节分解法对货运量数据进行分解,针对分解后数据特征建立 SAO-LSSVM-MA 预测模型;廖列法等^[8]提出一种改进 BAS 的 Elman 神经网络预测模型对上海港的货运量进行预测;丁涛等^[9]基于区域经济发展、产业结构等因素对过坝货运量的影响,构建 GA-SVM 预测模型对三峡过坝货运量进行了预测;辜勇等^[10]提出了基于网格搜索交叉验证优化的随机森林多式联运货运量预测模型,通过超参寻优进一步提升了模型的预测精度;杨艳等^[11]提出将 ARIMA-LSTM 串联模型按不同权重

收稿日期: 2024-01-14

基金项目: 国家重点研发计划项目(2018YFB1600405);国家自然科学基金资助项目(52271367)

作者简介: 温 泉(1984—),男,山东龙口人,高级工程师,博士,主要从事交通运输规划、机器学习研究。

E-mail: 58071899@qq.com

进行4种组合匹配,分别对某货运公司货运量进行了预测研究。此外,部分研究者^[12-16]将优化算法与长短期记忆模型(LSTM)或门控循环单元模型(GRU)结合应用于铁路货运量预测研究领域。近年来,货运量预测研究主要聚焦于模型选择及模型参数匹配研究,部分研究者将优化算法与常用预测模型相结合进行参数寻优,以提高预测精度,但在高维面板数据特征序列筛选、数据缺失、数据时间粒度不一等与现实息息相关的方面缺乏关注与深入研究。

极限梯度提升(eXtreme Gradient Boosting, XGBoost)作为一种重要的集成学习模型,近年来逐渐受到研究者的关注,并将其应用于不同领域的预测研究。白广栋等^[17]基于旅客选择中转城市数据将 XGBoost 模型应用于空铁中转联运城市推荐研究;吴飞等^[18]将 PCA 用于客观评价指标的降维,采用 SSA 算法对 XGBoost 进行超参寻优,并将 PCA-SSA-XGBoost 预测模型应用于驾驶性预测评估;董东林等^[19]构建 PSO-XGBoost 预测模型并应用于矿井突水水源快速判识研究,其预测精度与泛化能力优于 SVM、RF 等对比模型;王强等^[20]对 10 种气象数据进行降维分析,构建了一种自适应 SA-PSO-XGBoost 预测模型并对南京地区 2016—2017 年气象数据进行预测研究,结果表明采用超参寻优的改进模型优于对比模型预测结果。

为更好分析“645”工程实施后货运量变化趋势,本文选择长江干线湖北段水路货运量作为研究对象,针对高维面板数据中特征序列数据缺失及时间粒度不一致等问题,分别利用二次插值与 KNN 反距离权重插值法对特征数据中时间粒度不一致及缺失问题进行转换与插补。利用 SHAP 值的可解释性与层次聚类法综合研判高维面板数据中的关键特征序列,引入 Halton 低差异序列与准反射学习策略(QRBL)改进金枪鱼优化算法(TSO)的寻优效果,并将改进算法与 XGBoost 预测模型相结合对其决策树数量、决策树的深度、学习速率进行超参寻优,实现超参搜索范围内的最优组合,进一步提升预测模型的预测精度。

1 水路货运量预测模型构建

1.1 极限梯度提升模型(XGBoost)

极限梯度提升(eXtreme Gradient Boosting, XGBoost)^[21]是一种改进梯度提升树(Gradient Boosting Decision Tree, GDBT)算法,在并行计算效率、缺失值处理、预测性能方面优势显著,并在 Kaggle 机器学习竞赛中表现优异而得到广泛使用。XGBoost 的原理是基于梯度提升算法架构,通过多个弱学习器叠加实现树的集成学习,每个弱学习器都是回归树,即第一个回归树拟合样本数据,拟合误差作为第二个回归树的目标继续拟合,以此类推逐步迭代并遵循梯级下降原则,最终预测结果以各回归树预测结果累加形式呈现。

1.2 金枪鱼群优化算法(TSO)

金枪鱼群优化(Tuna swarm optimization, TSO)^[22]主要模拟金枪鱼群体的两种觅食行为,即螺旋觅食和抛物线觅食,本算法具有寻优能力强、收敛速度快且所需调参参数少的特点。因此,拟采用 TSO 算法对 XGBoost 模型进行超参寻优。

TSO 算法种群初始化后,根据随机数 r_1 取值,按算法逻辑判断进入螺旋觅食($r_1 < 0.5$)和抛物线觅食($r_1 \geq 0.5$)两个阶段。

(1)螺旋觅食更新位置

$$X_i(t+1) = \begin{cases} \alpha_1(X_\sigma(t) + \beta|X_\sigma(t) - X_i(t)|) + \alpha_2 X_i(t), & i = 1, r_2 < t/t_{\max} \\ \alpha_1(X_\sigma(t) + \beta|X_\sigma(t) - X_i(t)|) + \alpha_2 X_{i-1}(t), & i = 2, 3, \dots, N_P, r_2 < t/t_{\max} \\ \alpha_1(X_\psi(t) + \beta|X_\psi(t) - X_i(t)|) + \alpha_2 X_i(t), & i = 1, r_2 \geq t/t_{\max} \\ \alpha_1(X_\psi(t) + \beta|X_\psi(t) - X_i(t)|) + \alpha_2 X_{i-1}(t), & i = 2, 3, \dots, N_P, r_2 \geq t/t_{\max} \end{cases} \quad (1)$$

$$\beta = \exp(bl) \cos(2\pi r_3) \quad (2)$$

$$l = \exp\left(3 \cos\left(\left(\left(t_{\max} + \frac{1}{t}\right) - 1\right)\pi\right)\right) \quad (3)$$

(2)抛物线觅食更新位置

$$X_i(t+1) = \begin{cases} X_\psi(t) + r_5(X_\psi(t) - X_i(t)) + T_F \left(\left(1 - \frac{t}{t_{\max}} \right)^{t/t_{\max}} \right)^2 (X_\psi(t) - X_i(t)), & r_4 < 0.5 \\ T_F \left(\left(1 - \frac{t}{t_{\max}} \right)^{t/t_{\max}} \right)^2 X_i(t), & r_4 \geq 0.5 \end{cases} \quad (4)$$

式中: $X_i(t+1)$ 为第 $t+1$ 次迭代的第 i 个个体; $X_i(t)$ 为第 t 次迭代的第 i 个个体; $X_\sigma(t)$ 为第 t 次迭代随机生成的个体参考点; $X_\psi(t)$ 为第 t 次迭代最佳个体; α_1 、 α_2 为控制个体移动趋势的权重系数, 其中 $\alpha_1 = \gamma + (1-\gamma)(t/t_{\max})$, $\alpha_2 = (1-\gamma) - (1-\gamma)(t/t_{\max})$; γ 为常数取值 0.7; t 和 $t_{\max}$ 分别为当前迭代次数与最大迭代次数; r_1 、 r_2 、 r_3 、 r_4 、 r_5 为随机数且取值范围为 $[0, 1]$; T_F 为随机数, 取值 1 或 -1; N_P 为种群数量。

1.3 Halton 低差异序列

优化算法中初始种群粒子的分布范围和均匀程度都会对迭代过程中的收敛速度与搜索效率产生较大影响, 不少研究者采用常规混沌映射、复合混沌映射等方式对优化算法种群初始化进行改进, 在高维特征序列中效果较好。

Halton 低差异序列采用以质数为基的确定性方法构造, 在给定空间内分布更加均匀, 在任意长度的子序列都能均匀地填充整个函数空间, 其特性逐步被应用于优化算法初始化种群领域。设种群规模分别为 100、500、1 000, 维度取 2, 种群粒子取值范围均为 $[0, 1]$, Halton 低差异序列与随机序列产生的初始化种群如图 1 和图 2 所示。其中, Halton 低差异序列较随机序列产生的初始种群粒子分布更均匀, 差异性较小且分布质量更高。因此, 引入 Halton 低差异序列替代 TSO 算法中随机初始化种群粒子。

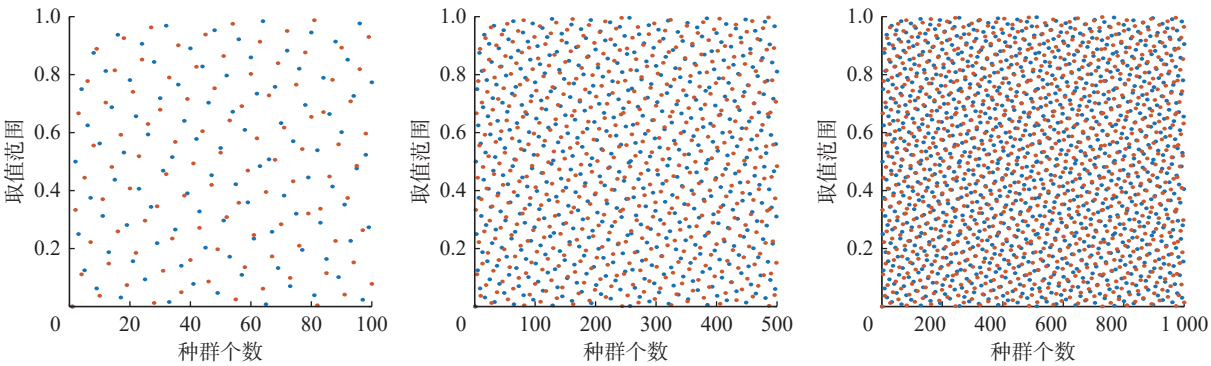


图 1 Halton 序列初始化种群

Fig. 1 Population initialization with Halton sequences

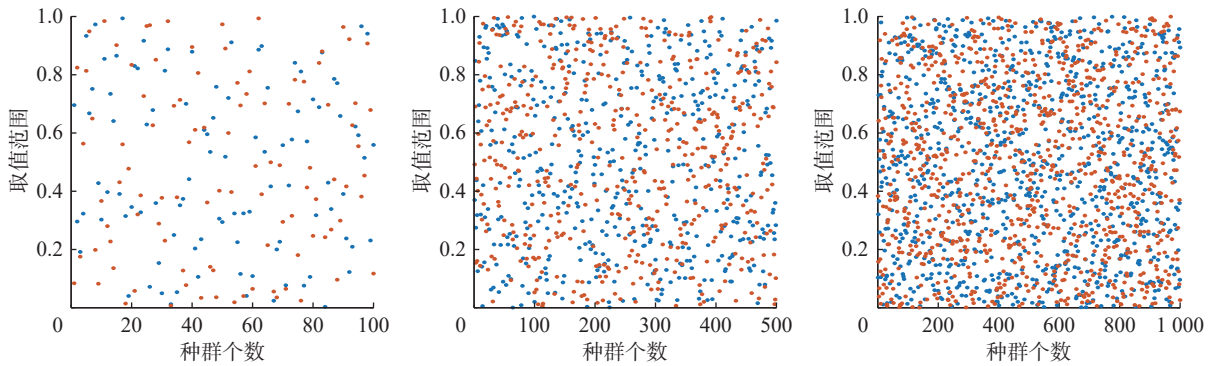


图 2 随机序列初始化种群

Fig. 2 Population initialization with random sequences

1.4 准反射学习 (QRBL) 策略

反向学习策略 (Opposition-based Learning, OBL)^[23] 的核心思想是基于当前最优解产生一个反向解, 增强种群的多样性。Fan 等^[24] 基于反向学习策略提出了一种改进准反射学习策略 (Quasi-reflection-based Learning, QRBL), 对比验证表明其在寻找全局最优解时优于 OBL 策略, 通过引入 QRBL 策略对 TSO 算法迭代解进行二次扰动, 增强了搜索范围内的寻优效率, 其推导公式如式(5)和式(6)所示。

准反射学习策略 (QRBL):

$$M_i = (b_{ui} + b_{li})/2 \quad (5)$$

$$X'_i(t+1) = \begin{cases} X_i(t+1) + (M_i - X_i(t+1))r_6, & X_i(t+1) < M_i \\ X_i(t+1) + (X_i(t+1) - M_i)r_6, & X_i(t+1) \geq M_i \end{cases} \quad (6)$$

式中: $X'_i(t+1)$ 代表第 $t+1$ 次迭代后第 i 个个体的准反射解; M_i 表示种群粒子中第 i 个个体搜索空间取值上下限的中心; b_{ui} 、 b_{li} 分别代表第 i 个个体的搜索空间上、下限; r_6 为 $[0, 1]$ 取值的随机数。

1.5 特征重要性分析方法选择

特征重要性选择是模型筛选关键特征数据的重要一环, 相关领域研究者主要采用皮尔逊相关系数法、斯皮尔曼等级相关系数法、距离相关系数、递归特征消除法、灰色关联度分析法及主成分分析法等筛选关键特征数据。

层次聚类法是目前常用的特征序列降维方法, 其核心思想是将每个样本独立看成一类, 选择类之间的距离条件, 常用选项包括欧氏距离、绝对值距离、切比雪夫距离等, 选定规定条件后将距离最小的一对合并构建一个新类, 并继续计算新类与其他类间的距离, 以此类推逐步合并, 直到所有样本合为一类为止。

SHAP (Shapley Addictive exPlanation) 是基于博弈论中 Shapley Value 的拓展^[25], SHAP 核心思想是计算某一特征加入模型时的边际贡献, 然后考虑该特征在所有特征序列条件下的不同边际贡献, 取其均值进行比选判断大小, 其不仅考虑单变量影响也考虑了各特征序列组合产生的协同效应。因此, 引入 SHAP 的全局解释能力筛选高维面板数据中的关键特征序列, 不仅能降低模型输入数据维度, 更能提升模型的计算效率。

1.6 多策略优化 TSO-XGBoost 预测模型流程

首先针对特征序列进行数据插补与特征重要性分析, 划分训练集与测试集, 引入 Halton 低差异序列提升 TSO 初始解的均匀分布性与覆盖范围, 对 TSO 逻辑搜索寻优后产生的解进行准反射学习策略扰动 (QRBL), 构建新解并比较适应度函数 (见式(7)) 变化, 进一步加强 TSO 算法对 XGBoost 模型超参的全局搜索能力及跳出局部最优的能力, 满足设定的迭代次数后, 将多策略优化 TSO 算法获得的最优超参组合赋予 XGBoost 预测模型, 并对预测结果进行误差分析。

$$F_{\text{fit}} = 0.5 \sqrt{\frac{1}{H} \sum_{h=1}^H (y'_t(h) - y_t(h))^2} + 0.5 \sqrt{\frac{1}{U} \sum_{u=1}^U (y'_t(u) - y_t(u))^2} \quad (7)$$

式中: F_{fit} 代表适应度函数; $y_t(h)$ 代表训练集第 h 个实际值; $y'_t(h)$ 代表训练集第 h 个预测值; $y_t(u)$ 代表测试集第 u 个实际值, $y'_t(u)$ 代表测试集第 u 个预测值; H 为训练集样本数, U 为测试集样本数。

2 实例验证

本文以长江干线中游“645”工程湖北段水路货运量作为研究对象, 数据来源主要依托国家统计局、交通运输部统计数据、湖北省统计局及长江海事局发布的数据统计资料, 数据收集时间为 2015 年 1 月至 2022 年 12 月, 合计 96 个月。

水路货运量影响因素较多, 涵盖国民经济、产业结构、其他运输方式、区域交通基础设施、断面交通流

量等多类指标,相关领域研究者主要采用产业经济分析和系统工程分析的方式确定水路货运量相关影响因素,同时考虑可获得性、适用性、可比性及科学性的选取原则构建影响因素初始特征变量序列^[5]。在梳理已有研究成果的基础上,综合考虑同类研究文献特征序列选取情况、各级政府部门统计数据类型及纳入统计月报的主要货种类型数据,拟分别选取湖北省地区生产总值、进出口总额、居民消费价格指数、铁路货运量、公路货运量、港口货物吞吐量、港口集装箱吞吐量、原煤产量、原油产量、铁矿石产量、商品车产量、水泥产量、钢材产量、武汉长江大桥断面交通流、九江长江大桥断面交通流等 15 个相关领域特征变量。铁路营业里程、长江干线中游航道通航里程等做为区域交通基础设施指标属年度统计数据且变化不大,因此暂不纳入模型特征变量考虑范畴。

2.1 特征序列数据预处理

尽管高维面板数据集来源于各大官方统计机构及主管部门,但部分特征序列数据统计时间粒度存在差异,且个别特征数据存在缺失问题,主要体现在 1、2 月份数据,部分统计单位采用 1、2 月累加合计数据对外发布,针对上述问题需要对特征序列数据进行预处理,为预测模型提供质量可靠的特征数据输入。

2.1.1 地区生产总值数据转化 地区生产总值与产业结构、国民经济、货运流通各环节都存在密切关系,国家统计局、地方统计机构公布的地区生产总值主要以年度、季度数据为主,部分月度公告以累计数据方式公布。为了统一预测模型输入数据的时间粒度,拟将 2015 年 1 月至 2022 年 12 月湖北省地区生产总值季度数据转化为月度数据。参考相关研究领域地区生产总值数据低频转高频方法^[10],拟采用 EViews10 软件对季度数据向月度数据进行转换,同时选择二次插值选项对数据进行填充。湖北省地区生产总值季度数据向月度数据转换结果如表 1 所示(以 2018 年 2、3 季度为例)。

表 1 地区生产总值季度向月度转化

Tab. 1 Conversion of regional GDP from quarterly to monthly

单位: 亿元

时间粒度(季度)		时间粒度(月度)					
2季度	3季度	4月	5月	6月	7月	8月	9月
10 103.57	10 620.01	3 282.82	3 373.38	3 447.35	3 432.11	3 527.37	3 660.51

2.1.2 特征序列数据缺失与处理 上述高维面板数据中,部分月度数据不全,数据缺失对模型训练影响较大,因此需要对缺失数据进行插值。通过 Python 对特征序列数据缺失及现有数据分布情况进行初步分析,由图 3 可知,除地区生产总值、居民消费价格指数、公路货运量、武汉港口集装箱吞吐量、武汉长江大桥断面交通流和九江长江大桥断面交通流数据完整外,其他特征序列数据均存在数据缺失情况。其中,除铁路货运量数据缺失 12.5% 外,其余特征序列数据缺失率均低于 10%,主要缺失月份集中于 1 月和 2 月,需通过插值方式对缺失数据进行插补。

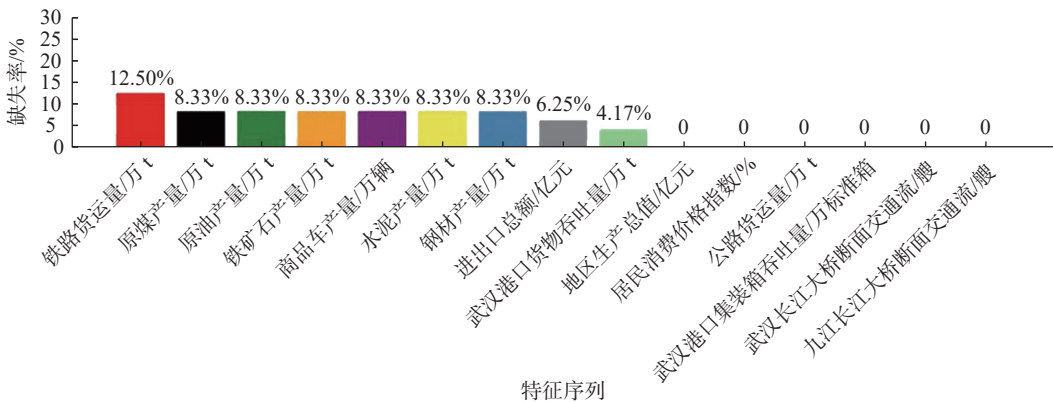


图 3 各特征序列数据缺失率

Fig. 3 Missing rate of each feature sequence data

2.1.3 KNN 反距离权重插值 尽管 XGBoost 这类决策树模型进行仿真计算时,其算法逻辑考虑了数据缺失情景,但模型输入数据的完整性与质量高低对模型计算精度有着重要影响。由前序数据预处理可知,特征序列数据已完成低频数据向高频数据转换,并分析了具体缺失数据的月份及数量,需要采用数据插补方式对其进行填充。目前,常用的数值插补方法主要有线性插值、分段三次埃尔米特插值、三次样条插值、K 最邻近插值(K-Nearest Neighbor, KNN)等。其中, KNN 插值法是目前机器学习领域主流的缺失数据插补方式,部分学者采用反距离权重(Inverse Distance Weighting, IDW)对其进行改进,使用 Python 对特征数据中 8 个含有缺失数据的特征序列进行插补,参数设定插补的相邻样本数为 5,权重函数按点之间的距离倒数表示,距离越近影响越大,反之亦然,计算结果如表 2 所示。

表 2 KNN 反距离权重插值(以湖北铁路货运量缺失值插补为例)
Tab. 2 KNN Inverse distance weighted interpolation (taking the imputation of missing values in Hubei railway freight volume as an example) 单位: 万t

日期	湖北省铁路货运量 (原数据)	湖北省铁路货运量 (KNN反距离权重插值后)	日期	湖北省铁路货运量 (原数据)	湖北省铁路货运量 (KNN反距离权重插值后)
2017-11	590.45	590.45	2018-02		555.10
2017-12	592.56	592.56	2018-03	559.40	559.40
2018-01		555.85	2018-04	561.90	561.90

2.1.4 预测模型特征重要性分析 理论上,模型输入特征序列维度越多模型预测精度愈佳,但维度过大会大大降低模型的计算效率,无法高效、准确地为研究提供合理的数据支持。面对高维面板数据中的 15 个特征序列,首选调用 scipy 库对其进行分析,聚类方式选择离差平方和法,距离采用欧氏距离,当距离为 8 时,原有 15 个特征序列可合并降低为如下特征序列集合(图 4 所示)。

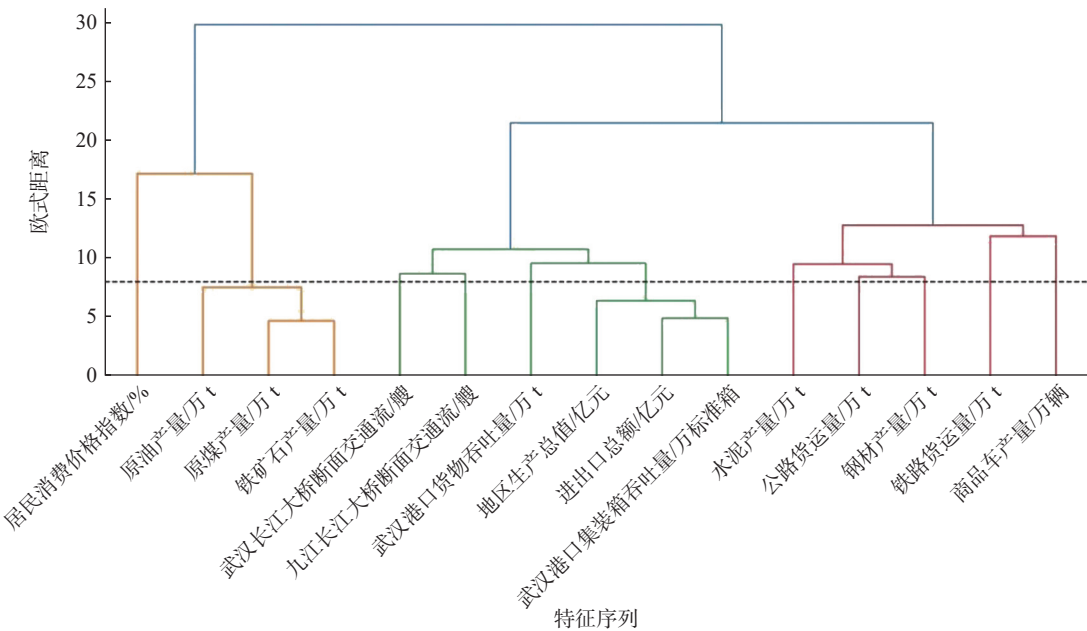


图 4 层次聚类分析图
Fig. 4 Hierarchical clustering analysis

根据层次聚类的结果,初步筛选出居民消费价格指数、地区生产总值等 11 个重要特征,通过计算每一个类别的数据点与所有数据点之间的距离来确定相似性,距离越小,相似度越高,但不能体现特征序列数据在模型训练中的影响效果,而 SHAP 值的全局解释能力则较好展示了特征序列的影响能力与方式。依据图 5 插值后各特征序列 SHAP 值与图 4 层次聚类分析结果,综合考虑两种方式特征序列聚类与排序结果,选取

进出口总额、武汉港口集装箱吞吐量、武汉港口货物吞吐量、地区生产总值、居民消费价格指数、武汉长江大桥断面交通流、原煤产量、铁路货运量、商品车产量、原油产量等 10 类特征序列作为预测模型输入。

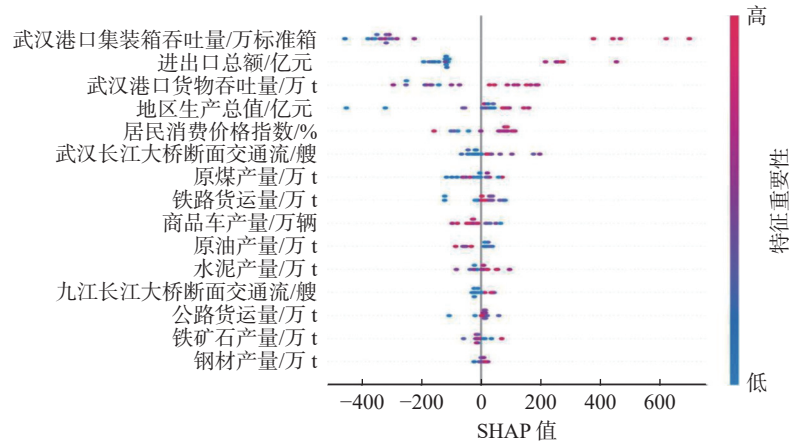


图 5 插值后各特征序列 SHAP 值
Fig. 5 SHAP values of feature sequences after interpolation

2.2 模型参数设定

通过特征序列数据预处理与重要性分析后,构建新的 10 类关键特征序列数据集,按照训练集 90%、测试集 10% 的比例划分(即采用 2015 年 1 月至 2022 年 2 月数据训练模型,2022 年 3 月至 12 月数据测试验证),模型预测结果采用平均绝对值误差(Mean Absolute Error, E_{MAE})、平均绝对百分比误差(Mean Absolute Percentage Error, E_{MAPE})及均方根误差(Mean Squared Error, E_{RMSE})进行评价分析。

为了更好地评估预测效果,选取 BP、LSTM、GRU、LSSVM、XGBoost、TSO-XGBoost、AO-XGBoost、AOA-XGBoost、AVOA-XGBoost、HBA-XGBoost 等热门预测模型及包含超参寻优算法的预测模型进行对比分析。其中 XGBoost 模型决策树数量取值区间 [10, 1 000]、决策树的深度 [1, 100]、学习速率 [0.000 1, 1]。TSO、AO (Aquila Optimizer)、AOA (Arithmetic Optimization Algorithm)、AVOA (African Vultures Optimization Algorithm)、HBA (Honey Badger Algorithm)等优化算法采用默认参数设置,未包含寻优算法的 BP、GRU、LSTM 模型隐含层节点数采用传统经验公式设定,LSSVM 与基础 XGBoost 模型同样选择参数默认。

2.3 数据计算与指标评价

为了比较插值前后及特征序列筛选前后模型预测效果,首先选择 XGBoost 模型(参数默认)与 TSO-XGBoost 模型(超参寻优)对初始 15 个特征序列插值前后(方案 1 和 2)进行比较,再对比方案 2 与 3 的计算结果。由表 3 可知,前述所选对比模型方案 2 预测结果的 E_{MAE} 、 E_{MAPE} 、 E_{RMSE} 均优于方案 1,计算结果表明初始特征序列缺失值插补后对提升模型预测精度有一定帮助。方案 3 与 2 计算结果表明,所选预测模型计算结果并非完全随特征序列数量增加而提升,部分对预测模型训练集有干扰的特征序列经过层次聚类与 SHAP 二者综合筛选剔除后,新特征序列组合(方案 3)对模型预测精度产生了一定程度的提升,特别是 TSO-XGBoost 模型的 E_{MAPE} 仅为 2.96%,其余各项误差评价指标显著优于不含任何超参寻优算法的 XGBoost 模型,进一步说明超参寻优算法在模型预测中的重要作用。

通过缺失值插补及特征序列综合筛选后,将新特征序列组合用于前述包含超参寻优算法的合计 10 种预测模型进行预测误差评估。由表 4、图 6、图 7 可知,多策略优化 TSO-XGBoost 的预测误差最小, E_{MAE} 、 E_{MAPE} 、 E_{RMSE} 分别为 102.233、2.05% 和 145.096,显著优于对比模型。此外,包含超参寻优算法的预测模型误差分析优于不含任何超参寻优算法的预测模型,说明预测模型超参设定对模型预测结果有着较大影响,通过寻优算法在设定搜索范围寻找最佳超参组合,可以有效降低预测模型的 E_{MAE} 、 E_{MAPE} 和 E_{RMSE} 。

表 3 不同方案下特征序列设定与模型选择

Tab. 3 Feature sequence settings and model selection under different schemes

方案设定	特征序列选择	模型选择					
		XGBoost(参数默认)			TSO-XGBoost(超参寻优)		
		E_{MAE}	$E_{MAPE}/\%$	E_{RMSE}	E_{MAE}	$E_{MAPE}/\%$	E_{RMSE}
方案1	初始15个特征序列(插值前)	495.930	9.86	547.391	220.908	4.60	356.899
方案2	初始15个特征序列(插值后)	439.535	8.75	480.357	192.352	4.02	377.056
方案3	10个特征序列(经层次聚类及SHAP筛选)	390.772	7.70	440.283	151.116	2.96	213.010

表 4 各模型预测结果评价指标

Tab. 4 Evaluation metrics for forecasting results of each model

模型类型	平均绝对值误差	平均绝对百分比误差/%	均方根误差	模型类型	平均绝对值误差	平均绝对百分比误差/%	均方根误差
多策略优化TSO-XGBoost	102.233	2.05	145.096	XGBoost	390.772	7.70	440.283
TSO-XGBoost	151.116	2.96	213.010	LSSVM	537.999	10.68	586.099
AOA-XGBoost	188.967	3.93	367.493	LSTM	414.925	8.27	461.438
HBA-XGBoost	182.694	3.79	319.371	GRU	417.339	8.30	497.249
AO-XGBoost	181.533	3.78	319.545	BP	446.253	8.98	544.362
AVOA-XGBoost	187.830	3.92	371.049				

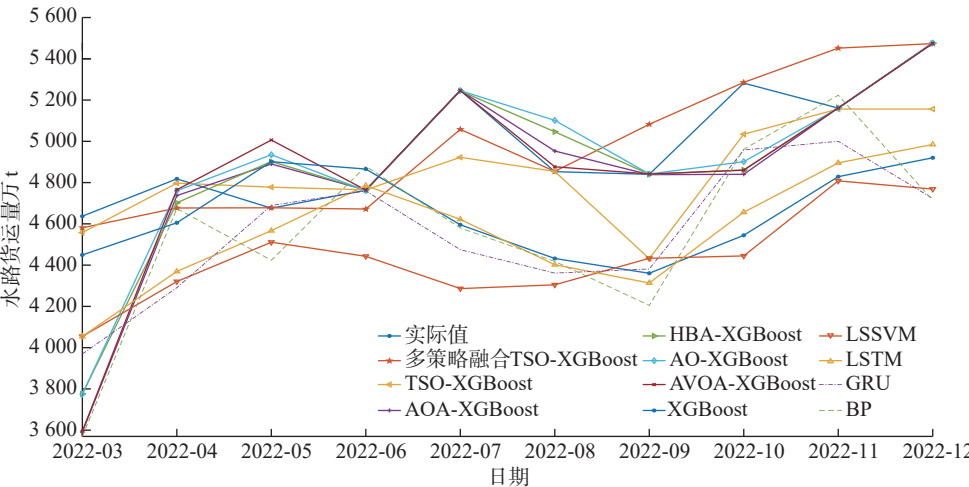


图 6 各对比模型预测值与实际值

Fig. 6 Forecasted versus actual values of comparative models

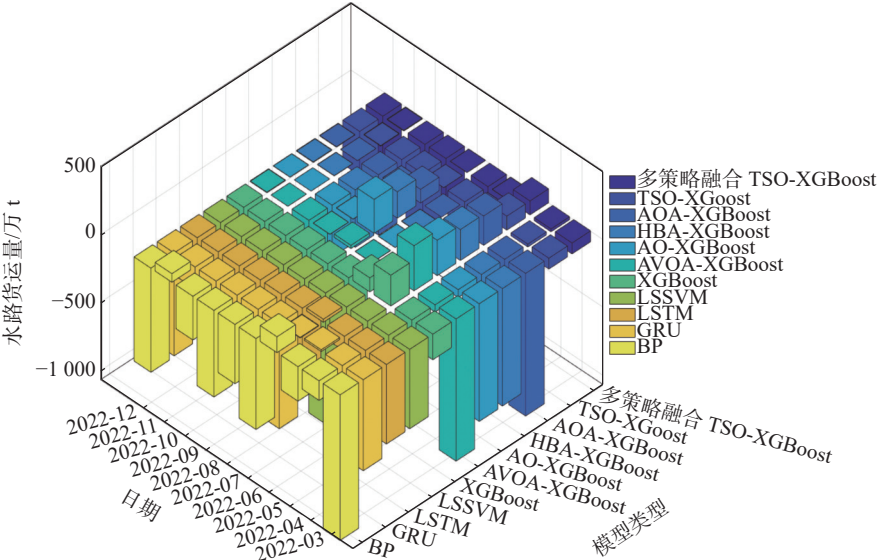


图 7 各对比模型预测值与实际值误差

Fig. 7 Forecasting error of comparative models versus actual values

3 结 语

(1)采用二次插值法和 KNN 反距离权重插值法对特征序列缺失值进行填补,特征序列插值后相对插值前对模型预测精度有一定程度提升。

(2)引入 SHAP 值可解释性与层次聚类的方式对高维面板数据中特征序列进行筛选,可有效降低模型输入计算复杂度,提升模型预测精度。

(3)多策略优化 TSO-XGBoost 模型应用于水路货运量趋势预测研究,其模型预测精度优于对比预测模型,基于其良好的适应性与模型超参寻优性能,可更好地适用于高维面板数据条件下水路货运量长期发展趋势研究。

参 考 文 献:

- [1] 臧文亚,周仲礼,龚灏,等. 基于灰色马尔科夫模型的重庆市水运货运量的预测[J]. 水运工程, 2012(1): 30-33. (ZANG Wenya, ZHOU Zhongli, GONG Hao, et al. Freight volume prediction for Chongqing water transport based on gray Markov[J]. Port & Waterway Engineering, 2012(1): 30-33. (in Chinese))
- [2] 李文杰,贺艺伟,杨胜发,等. 长江上游涪陵至丰都段航道水运量分析预测[J]. 水运工程, 2020(6): 99-105. (LI Wenjie, HE Yiwei, YANG Shengfa, et al. Analysis and prediction on waterway freight from Fuling to Fengdu reach in the Upper Yangtze River[J]. Port & Waterway Engineering, 2020(6): 99-105. (in Chinese))
- [3] 温泉. 长江干线中游运力运量及船型发展研究[R]. 武汉: 长江航道规划设计研究院, 2019: 1-20. (WEN Quan. Research on capacity and ship type development in the middle reaches of the Yangtze River Trunk Line[R]. Wuhan: Changjiang Waterway Institute of Planning and Design, 2019: 1-20. (in Chinese))
- [4] 张浩,肖金龙,温泉,等. 基于长江上游区域产业结构测度的三峡枢纽货运量预测[J]. 水运工程, 2019(11): 58-62. (ZHANG Hao, XIAO Jinlong, WEN Quan, et al. Prediction of Three Gorges ship lock freight volume based on regional industrial structure measurement in upper reaches of the Yangtze River[J]. Port & Waterway Engineering, 2019(11): 58-62. (in Chinese))
- [5] 戎陆庆. 基于灰预测与灰关联的区域水路货运发展研究: 以广西为例[J]. 北京交通大学学报(社会科学版), 2018, 17(2): 109-117. (RONG Luqing. Development of regional waterway freight transportation based on grey prediction and grey relation: taking Guangxi Province as an example[J]. Journal of Beijing Jiaotong University (Social Sciences Edition), 2018, 17(2): 109-117. (in Chinese))
- [6] 陈琛,吴青,高嵩. 基于时空因素的水路短期货运量预测[J]. 上海交通大学学报, 2019, 53(5): 556-562. (CHEN Chen, WU Qing, GAO Song. Short-term shipping freight volume prediction based on temporal-spatial features[J]. Journal of Shanghai Jiao Tong University, 2019, 53(5): 556-562. (in Chinese))
- [7] 李红娟,卢天哲,祝汉灿. 全国货运量预测与货运结构分析[J]. 统计与决策, 2022, 38(18): 179-183. (LI Hongjuan, LU Tianzhe, ZHU Hancan. National freight volume forecast and freight structure analysis[J]. Statistics & Decision, 2022, 38(18): 179-183. (in Chinese))
- [8] 廖列法,欧阳宗英. 改进天牛须搜索优化神经网络的港口货运量预测[J]. 科学技术与工程, 2021, 21(7): 2937-2944. (LIAO Liefu, OUYANG Zongying. Forecast of port freight volume based on improved optimization neural network of beetle antennae search[J]. Science Technology and Engineering, 2021, 21(7): 2937-2944. (in Chinese))
- [9] 丁涛,植链,郭琦,等. 基于 GA-SVM 的三峡大坝过坝货运量预测[J]. 物流技术, 2018, 37(9): 62-65. (DING Tao, ZHI Lian, GUO Qi, et al. Forecasting of over-dam cargo transport volume of Three-Gorges dam based on GA-SVM[J]. Logistics Technology, 2018, 37(9): 62-65. (in Chinese))
- [10] 辜勇,杨泽昭. 基于随机森林的集装箱多式联运货运量预测[J]. 武汉理工大学学报, 2023, 45(1): 35-44. (GU Yong, YANG Zezhao. Prediction of container multimodal transport freight volume based on random forest[J]. Journal of Wuhan University of Technology, 2023, 45(1): 35-44. (in Chinese))

- [11] 杨艳, 黄晴, 龙思, 等. 基于 ARIMA-LSTM 的货运量组合预测方法研究[J]. 交通科学与工程, 2022, 38(2): 102-108. (YANG Yan, HUANG Qing, LONG Si, et al. Research on combined forecasting method of freight volume based on ARIMA-LSTM[J]. Journal of Transport Science and Engineering, 2022, 38(2): 102-108. (in Chinese))
- [12] 徐玉萍, 邓俊翔, 蒋泽华. 基于组合预测模型的铁路货运量预测研究[J]. 铁道科学与工程学报, 2021, 18(1): 243-249. (XU Yuping, DENG Junxiang, JIANG Zehua. Railway freight volume forecasting based on a combined model[J]. Journal of Railway Science and Engineering, 2021, 18(1): 243-249. (in Chinese))
- [13] 周昌野, 李程. 基于 GS-LSTM 模型的铁路货运量预测[J]. 计算机与现代化, 2022(10): 24-28. (ZHOU Changye, LI Cheng. Prediction of railway freight volume based on GS-LSTM model[J]. Computer and Modernization, 2022(10): 24-28. (in Chinese))
- [14] 张仙, 戴家佳, 余奇迪. 基于 SARIMA-PSO-ELM 组合模型的我国铁路货运量预测[J]. 数理统计与管理, 2022, 41(3): 394-401. (ZHANG Xian, DAI Jiajia, YU Qidi. Railway freight volume forecasting based on SARIMA-PSO-ELM combination model[J]. Journal of Applied Statistics and Management, 2022, 41(3): 394-401. (in Chinese))
- [15] 李万, 冯芬玲, 蒋琦玮. 改进粒子群算法优化 LSTM 神经网络的铁路客运量预测[J]. 铁道科学与工程学报, 2018, 15(12): 3274-3280. (LI Wan, FENG Fenling, JIANG Qiwei. Prediction for railway passenger volume based on modified PSO optimized LSTM neural network[J]. Journal of Railway Science and Engineering, 2018, 15(12): 3274-3280. (in Chinese))
- [16] 谭雪, 张小强. 基于 GRU 深度网络的铁路短期货运量预测[J]. 铁道学报, 2020, 42(12): 28-35. (TAN Xue, ZHANG Xiaoqiang. GRU deep neural network based short-term railway freight demand forecasting[J]. Journal of the China Railway Society, 2020, 42(12): 28-35. (in Chinese))
- [17] 白广栋, 翁涅元, 张启蒙, 等. 基于 XGBoost 模型的空铁联运中转城市研究[J]. 铁道运输与经济, 2023, 45(3): 24-31. (BAI Guangdong, WENG Shengyuan, ZHANG Qimeng, et al. Research on transit city of air-rail intermodal transport based on XGBoost model[J]. Railway Transport and Economy, 2023, 45(3): 24-31. (in Chinese))
- [18] 吴飞, 王鹏程, 杨康. 基于 PCA-SSA-XGBoost 的车辆驾驶性评估[J/OL]. 吉林大学学报(工学版). (2023-03-10)[2023-06-26]. <https://doi.org/10.13229/j.cnki.jdxbgxb.20230215>. (WU Fei, WANG Pengcheng, YANG Kang. Drivability evaluation model based on PCA-SSA-XGBoost[J/OL]. Journal of Jilin University (Engineering and Technology Edition). (2023-03-10)[2023-06-26]. <https://doi.org/10.13229/j.cnki.jdxbgxb.20230215>. (in Chinese))
- [19] 董东林, 张陇强, 张恩雨, 等. 基于 PSO-XGBoost 的矿井突水水源快速识别模型[J]. 煤炭科学技术, 2023, 51(7): 72-82. (DONG Donglin, ZHANG Longqiang, ZHANG Enyu, et al. A rapid identification model of mine water inrush based on PSO-XGBoost[J]. Coal Science and Technology, 2023, 51(7): 72-82. (in Chinese))
- [20] 王强, 秦华旺, 齐春帅, 等. 基于自适应 SA-PSO 改进的 XGBoost 气温预测方法[J]. 电子测量技术, 2023, 46(7): 67-72. (WANG Qiang, QIN Huawang, QI Chunshuai, et al. Improved XGBoost temperature prediction method based on SA-PSO[J]. Electronic Measurement Technology, 2023, 46(7): 67-72. (in Chinese))
- [21] CHEN T, 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: ACM, 2016: 785-794.
- [22] XIE L, HAN T, ZHOU H, et al. Tuna swarm optimization: a novel swarm-based metaheuristic algorithm for global optimization[J]. Computational Intelligence and Neuroscience, 2021(9): 1-22.
- [23] TIZHOOSH H R. Opposition-Based Reinforcement Learning[J]. Journal of Advanced Computational Intelligence & Intelligent Informatics, 2006, 10(4): 578-585.
- [24] FAN Q, CHEN Z, XIA Z. A novel quasi-reflected Harris Hawks optimization algorithm for global optimization problems[J]. Soft Computing, 2020, 24(19): 14825-14843.
- [25] LUNDBERG S M, LEE S I. Consistent feature attribution for treeensembles[EB/OL]. (2018-12-12) [2023-06-26]. <https://arxiv.org/abs/1802.03888>.

Waterway freight volume prediction based on SHAP and TSO-XGBoost model

WEN Quan¹, YU Yuhuan², ZHUANG Shangde², MOU Junmin³

(1. Business School, Wuhan Vocational College of Software and Engineering (Wuhan Open University), Wuhan 430079, China; 2. Changjiang Waterway Bureau, Wuhan 430023, China; 3. School of Navigation, Wuhan University of Technology, Wuhan 430063, China)

Abstract: The demand for waterway freight volume is influenced by numerous factors. Following the implementation of the "645" project in the midstream section of the Yangtze River, the navigation conditions have significantly improved. To better analyze the trend changes in freight volume after the project implementation, this study introduces a novel model for forecasting waterway freight volume. Initially, the quadratic interpolation method and the KNN inverse distance weighting interpolation method are employed to address issues of inconsistency in time granularity and missing data in high-dimensional panel data. By utilizing hierarchical clustering and the interpretability of SHAP values, key influence factor feature sequences are comprehensively screened to reduce the dimensions and scale of input data for the forecasting model. The introduction of the Halton low-discrepancy sequences and the Quasi-Reflective Bayesian Learning (QRBL) strategy substantially enhances the optimization efficiency of the Tuna Swarm Optimization (TSO) algorithm, improving the TSO algorithm's optimization effectiveness of hyperparameter combinations, such as the number of decision trees, the depth of decision trees, and learning rate in the eXtreme Gradient Boosting (XGBoost) model that determine the model's fitting ability. The results indicate that the new model significantly outperforms comparative models in forecasting accuracy, demonstrating better applicability for waterway freight volume forecasting under the influence of multiple feature factors.

Key words: Tuna Swarm Optimization (TSO); eXtreme Gradient Boosting (XGBoost); waterway freight volume