

【材料科学 / Materials Science】

基于改进 Stacking 集成学习的深层油井管腐蚀预测

黄晗¹, 陈长风², 贾小兰², 张玉洁¹, 石丽伟³, 王立群¹

1) 中国石油大学(北京)理学院, 北京 102249; 2) 中国石油大学(北京)新能源与材料学院, 北京 102249; 3) 中国政法大学科学技术教学部, 北京 102249

摘要: 为提升深层复杂环境下油井管平均腐蚀与点蚀速率的预测精度, 并优化传统 Stacking 集成学习未充分考虑基学习器异质性的问题, 提出了一种基于决定系数 R^2 的改进 Stacking 集成学习算法. 该算法集成了 XGBoost(extreme gradient boosting)模型、随机森林(random forest, RF)模型、支持向量回归(support vector regression, SVR)模型和梯度提升决策树(gradient boosting decision tree, GBDT)模型 4 种机器学习算法作为基学习器, 并基于决定系数 R^2 为基学习器的输出结果进行权重赋值, 作为元学习器的输入数据集. 实验结果显示, 与传统 Stacking 集成方法相比, 改进后的模型在平均腐蚀速率预测上, 平均绝对误差和均方误差分别降低了 25.9% 和 9.7%, 决定系数提高了 2.3%; 在点蚀速率预测上, 平均绝对误差和均方误差分别降低了 11.6% 和 2.0%, 决定系数提高了 2.7%, 证明了本算法的有效性. 研究成果可为深层油井管腐蚀防控与安全运维提供支撑.

关键词: 腐蚀科学与防护; Stacking 集成学习; 深层油井管材腐蚀; 机器学习; XGBoost; 随机森林; 支持向量回归; 梯度提升决策树

中图分类号: TE983; TP181

文献标志码: A

DOI: 10.3724/SP.J.1249.2026.01007

Corrosion prediction of deep oil well tubing based on improved Stacking ensemble learning approach

HUANG Han¹, CHEN Changfeng², JIA Xiaolan², ZHANG Yujie¹,
SHI Liwei³, and WANG Liqun¹

1) College of Science, China University of Petroleum (Beijing), Beijing 102249, P. R. China

2) College of New Energy and Materials, China University of Petroleum (Beijing), Beijing 102249, P. R. China

3) Department of Science and Technology, China University of Political Science and Law, Beijing 102249, P. R. China

Abstract: To enhance the prediction accuracy for both average corrosion rate and pitting corrosion rate of oil well tubing in deep complex environments, and to address the issue of insufficient consideration of base learner heterogeneity in traditional Stacking ensemble learning, an improved Stacking ensemble learning algorithm based on the coefficient of determination (R^2) is proposed. This algorithm integrates four machine learning models as base learners: extreme gradient boosting (XGBoost), random forest (RF), support vector regression (SVR), and gradient boosting decision tree (GBDT). The outputs of these base learners are weighted according to their respective R^2 , and the weighted combination forms the input dataset for the meta-learner. Experimental results demonstrate that,

Received: 2025-03-12; **Revised:** 2025-11-08; **Accepted:** 2025-11-18; **Online (CNKI):** 2025-12-30

Foundation: National Natural Science Foundation of China (12171482, U23A20301); Project of State Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum-Beijing (PRE/DX-2504)

Corresponding author: Professor WANG Liqun (wliqunhmlly@cup.edu.cn)

Citation: HUANG Han, CHEN Changfeng, JIA Xiaolan, et al. Corrosion prediction of deep oil well tubing based on improved Stacking ensemble learning approach [J]. Journal of Shenzhen University Science and Engineering, 2026, 43(1): 7-16. (in Chinese)

Open access: This article is licensed under the CC BY 4.0 License (<https://creativecommons.org/licenses/by/4.0/>).



<http://journal.szu.edu.cn>

compared with the traditional Stacking ensemble method, the improved model achieves a 25.9% reduction in mean absolute error (MAE) and a 9.7% reduction in mean squared error (MSE) for average corrosion rate prediction, alongside a 2.3% increase in the R^2 . For pitting corrosion rate prediction, it yields reductions of 11.6% for MAE and 2.0% for MSE, respectively, with a 2.7% increase for R^2 . These results validate the effectiveness of the proposed algorithm, and the research findings provide valuable support for corrosion prevention, control and safe operational maintenance of deep oil well tubing.

Key words: corrosion science and protection; Stacking ensemble learning; corrosion of deep oil well tubing; machine learning; XGBoost; random forest; support vector regression; gradient boosting decision tree

随着全球能源需求攀升, 油气勘探开发向深层、超深层推进. 中国深层油气资源储量丰富、开发潜力巨大, 是重要增储上产领域^[1-2]. 但高深度环境下地层温度升高会大幅加剧油井管的腐蚀问题, 给安全高效开采带来严峻挑战^[3-4], 精确预测油井管腐蚀, 结合理论分析制定维修周期, 可有效减少腐蚀故障, 保障管道稳定运行, 规避经济损失^[5-6].

目前, 管道腐蚀预测模型已形成经验模型、半经验模型、机理模型及机器学习算法模型4大类^[7]. 经验模型如 Equinor ASA 与 Saga 石油公司的 Norsok M506 模型, 是石油工业材料选择与腐蚀裕量评估标准之一, 可预测 20 ~ 150 °C、pH 值为 3.5 ~ 6.5 的金属腐蚀^[8-9]; 半经验模型以 De. Waard 模型为代表, 其中, De. Waard 95 模型应用最广, 适用于温度低于 80 °C 的金属腐蚀预测^[10-12]. 但经验与半经验模型难以应对实际场景中腐蚀过程的高度非线性特征, 及温度、压力和介质成分等复杂因素的交织影响, 预测存在较大不确定性与误差^[13].

近年来, 基于机器学习的管道腐蚀预测模型逐步发展, 能从数据中学习规律, 捕捉多因素交互作用与综合效应^[14]. 如 REN 等^[15]用 BP 神经网络预测天然气长输管道内腐蚀速率; LIAO 等^[16]通过构建人工神经网络模型并结合粒子群等算法训练实现了湿式集气管道内腐蚀速率预测; ZHANG 等^[17]将支持向量机与不等区间灰色预测法结合, 建立了天然气管道腐蚀预测模型; OSSAI^[18]基于前馈子空间聚类神经网络建立管道腐蚀缺陷深度时变模型. 但此类模型在深层油田管道腐蚀预测中的应用仍较少. 为构建适用于深层油田的管道腐蚀预测机器学习模型, 本研究提出一种基于决定系数 R^2 的改进 Stacking 集成学习算法, 用于深层油田的管道平均腐蚀和点蚀速率的预测. 该方法集成 XGBoost (extreme gradient boosting)、随机森林 (random forest, RF)、支持向量回归 (support vector regression, SVR) 和梯

度提升决策树 (gradient boosting decision tree, GBDT) 4 种模型, 并依据决定系数 R^2 为各基学习器分配集成权重, 以提高模型的预测性能.

1 Stacking 集成学习

1.1 基学习器

1.1.1 XGBoost

XGBoost^[19]是在梯度提升框架下实现的机器学习算法, 对于给定的数据样本 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. 其中, n 为数据样本总数量. 训练出 k 个决策树模型 $f_i(x)$, $i = 1, 2, \dots, k$, 分别对数据样本进行映射, 将每个输入样本分配到不同的叶子节点, 每个叶子节点对应的分数即为在此叶子节点的数据样本在本棵决策树的预测值, 样本的预测结果就是其对每棵树的预测值之和. XGBoost 模型定义为

$$\hat{y}_i = \sum_{j=1}^k f_j(x_i) \quad (1)$$

其中, $f_j(x_i)$ 表示样本 x_i 在第 j 棵树的预测值; $\hat{y}_i$ 为第 i 个预测值.

在训练过程中, 设定 XGBoost 模型的目标函数为

$$G = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{i=1}^k \Omega(f_i) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \Omega(f_i) \quad (2)$$

$$\Omega(f_i) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (3)$$

其中, $L(y_i, \hat{y}_i)$ 为损失函数, 表示第 i 个样本的预测值和真实值之间的误差, y_i 为第 i 个观测值; $\Omega(f_i)$ 为模型的复杂度, 包含控制树结构的 γT 和正则项 $\frac{1}{2} \lambda \sum_{j=1}^T w_j^2$, 分别限制叶子节点数和叶子节点的权重分数; T 为叶子节点的个数; w_j 为第 j 个叶子节点的权重; γ 和 λ 为控制惩罚程度的超参数. 目标

函数中 $\hat{y}_i^t = \hat{y}_i^{(t-1)} + \eta f(x_i)$, 其中, t 为迭代次数; η 为学习率.

1.1.2 随机森林模型

随机森林^[20-22]是一种经典的 Bagging 算法, 它是通过构建多个决策树并将它们的预测结果通过共同投票或加权平均进行汇总, 得到最终的预测结果. 决策树模型的构建为: 对于给定的数据样本 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, 建立决策树模型, 对于任意的划分特征, 其对应的划分点将数据样本划分成两组数据集, 使两组数据集各自的均方差以及均方差之和最小时, 所对应的特征和特征值划分点即为所求, 其表达式为

$$\min_{j,s} \left[\min_{x_i \in D_1(j,s)} \sum (y_i - c_1)^2 + \min_{x_i \in D_2(j,s)} \sum (y_i - c_2)^2 \right] \quad (4)$$

其中, D_1 和 D_2 为划分出的两个数据集; c_1 和 c_2 分别为对应两个数据集中的样本输出均值.

对于确定的划分 (j, s) , 其输出值为

$$D_1(j, s) = \{x | x^{(j)} \leq s\}, D_2(j, s) = \{x | x^{(j)} > s\},$$

$$\hat{c}_m = \frac{1}{n_m} \sum_{x_i \in D_m(j,s)} y_i, x \in D_m, m = 1, 2 \quad (5)$$

其中, $x^{(j)}$ 为输入样本 x 的第 j 个特征值; n_m 为数据集 D_m 中的样本数量.

输入空间划分为 P 个区域 $R_1, R_2, \dots, R_P$, 则有

$$f(x) = \sum_{i=1}^P \hat{c}_i I \quad (6)$$

其中, I 为指示函数, 若 $x \in R_i$ 则 $I = 1$, 否则为 0.

随机森林的建立步骤如下: 采用 bootstrap 采样方法, 随机且有放回地抽取出 n 个训练集的子样本集 $\{x_1, x_2, \dots, x_n\}$, 对每个子样本集建立对应的决策树, 并且对于数据集的所有特征项, 从中随机抽取 d 个特征作为决策树的分裂特征集, 重复上述步骤后构建出多个决策树, 通过集成处理得到最终的随机森林模型.

1.1.3 SVR 模型

支持向量回归模型^[23]属于支持向量机 (support vector machine, SVM) 的一种扩展, 对于给定的数据样本 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, 通过非线性函数 $\varphi(x_i)$ 将 x_i 映射到 n 维特征空间, 构造出回归函数:

$$\hat{y}_i = w^T \varphi(x_i) + b \quad (7)$$

其中, w 和 b 分别是目标函数和偏置项的权重系数.

为了找出 w 和 b , 优化函数为

$$\min_{w,b,\xi,\hat{\xi}} \frac{1}{2} w^T w + C \sum_{i=1}^N (\hat{\xi}_i + \xi_i) \quad (8)$$

满足:

$$\begin{cases} y_i - (w^T \varphi(x_i) + b) \leq \varepsilon + \xi_i \\ (w^T \varphi(x_i) + b) - y_i \leq \varepsilon + \hat{\xi}_i \\ \hat{\xi}_i, \xi_i \geq 0, i = 1, 2, \dots, n \end{cases} \quad (9)$$

其中, ε 为不敏感损失函数的容忍阈值; 引入 ξ_i 和 $\hat{\xi}_i$ 作为松弛变量, 分别用于度量第 i 个样本的真实值 y_i 与预测上边界 $(\hat{y}_i + \varepsilon)$ 与下边界 $(\hat{y}_i - \varepsilon)$ 的距离; C 为惩罚因子.

为了解决模型回归拟合问题, 引入不敏感损失函数 L_ε ,

$$L_\varepsilon(y_i - \hat{y}_i) = \begin{cases} 0, & |y_i - \hat{y}_i| \leq \varepsilon \\ |y_i - \hat{y}_i| - \varepsilon, & \text{其他} \end{cases} \quad (10)$$

引入拉格朗日函数, 则最优解问题转化为对偶问题, 其表达式为

$$\max_{\alpha} \left[-\frac{1}{2} \sum_{i,j=1}^N (\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j) k(x_i, x_j) - \varepsilon \sum_{i=1}^N (\alpha_i + \hat{\alpha}_i) + \sum_{i=1}^N y_i (\alpha_i - \hat{\alpha}_i) \right] \quad (11)$$

其中, α_i 和 $\hat{\alpha}_i$ 为拉格朗日乘子, 均为非负变量, 用于对偶问题求解.

应用 $k(x_i, x_j) = \varphi^T(x_i) \varphi(x_j)$ 核函数能够将非线性数据映射到高维特征空间, 则回归函数变为

$$\hat{y}_i = w^T \varphi(x_i) + b = \sum_{i=1}^N (\alpha_i - \hat{\alpha}_i) k(x_i, x) + b \quad (12)$$

模型核函数 $k(x_i, x)$ 分别采用了 RBF 核函数和 Poly 核函数.

RBF 核函数表达式为

$$k(x_i, x) = \exp(-\gamma \|x_i - x\|^2) \quad (13)$$

其中, γ 为 RBF 核函数的带宽超参数; $\|x_i - x\|^2$ 为样本 x_i 与待预测样本 x 的欧几里得距离.

Poly 核函数表达式为

$$k(x_i, x) = [(x_i x) + 1]^d \quad (14)$$

其中, d 为 Poly 核函数的多项式阶数.

1.1.4 GBDT 模型

梯度提升决策树模型^[24-25]结合了 Boosting 和决策树的思想, 其基本原理是迭代地训练决策树, 每次训练都基于之前训练的结果来进行优化. 在训练过程中, GBDT 使用加法模型和函数优化方法, 并

基于梯度下降的思想来逐步减少预测误差. 对于给定的数据样本 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, 初始化弱学习器为

$$f_0(x) = \arg \min_c \sum_{i=1}^n L(y_i, c) \quad (15)$$

其中, L 为损失函数; c 为损失函数最小的常数值.

建立 M 棵回归树, 对于任意样本 (x_i, y_i) , 计算对应的第 m 棵树的负梯度, 其表达式为

$$r_{i,m} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x) = f_{m-1}(x)} \quad (16)$$

其中, $r_{i,m}$ 为残差值; $f_{m-1}(x)$ 为第 $m-1$ 棵回归树的预测值.

将 $r_{i,m}$ 作为样本新的真实值, 即使用 $(x_i, r_{i,m})$ 作为后续回归树的训练数据, 产生第 m 棵回归树, 对应的叶子节点区域为 $R_{j,m}$, $j = 1, 2, \dots, T$. 回归模型的最佳拟合值为

$$c_{j,m} = \arg \min_r \sum_{x_i \in R_{j,m}} L(y_i, f_{m-1}(x_i) + r) \quad (17)$$

其中, r 为第 m 棵树第 j 个叶子节点的拟合值候选变量.

对强学习器 $f_m(x_i)$ 进行更新,

$$f_m(x_i) = f_{m-1}(x_i) + \sum_{j=1}^T c_{j,m} I \quad (18)$$

最终, 得到强学习器 $f(x_i)$ 即为预测模型,

$$f(x_i) = f_M(x_i) = f_0(x_i) + \sum_{m=1}^M \sum_{j=1}^T c_{j,m} I \quad (19)$$

1.2 Stacking 集成学习

Stacking 集成学习算法框架如图 1 所示. 它采用集成策略, 将多个基学习器的输出结果通过元学习器进行泛化整合, 有效融合各基学习器的优势, 相比于单个基学习器具有更好的预测性能. 其中, 为了避免模型复杂度过高导致的过拟合问题, 常用的做法是在训练基学习器时采用 K 折交叉验证.

2 基于决定系数 R^2 的改进 Stacking 集成学习

有研究表明, Stacking 集成学习算法的输出结果在很大程度上受到元学习器所使用的训练数据集特征属性的影响^[26]. 训练数据集中包含的错误信息越少, 最终的预测准确性就越高^[27]. 在生成元学习

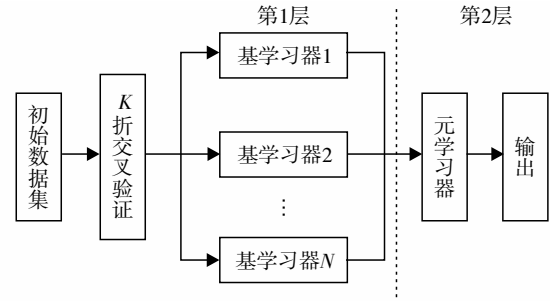


图1 Stacking集成学习算法框架

Fig. 1 Framework of stacking-based ensemble learning algorithm.

器的训练数据集时, 传统Stacking集成学习算法通常是将多个基学习器的输出进行平均处理. 然而, 由于基学习器的性能存在差异, 这些差异可能会影响集成学习的最终效果, 预测准确的基学习器可以有效地提升元学习器的学习效果, 而如果基学习器的预测错误, 则可能使元学习器的性能达不到最优, 从而忽略了某些具有高预测能力的基学习器的优势, 最终影响整个算法的预测精度.

为了解决传统Stacking集成学习算法存在的问题, 在构建元学习器的输入样本集合时, 可以根据特定标准对特征矩阵的权重进行合理分配, 为了使预测精度更高的基学习器对整个模型的效果有较大的影响, 以不同基学习器的预测值和真实值之间的误差为基础来确定权重. 鉴于本研究目的在于实现腐蚀速率的预测, 应尽量使模型的样本预测值更接近真实值, 因此采用决定系数 R^2 作为评判基学习器的指标, 具体步骤如下:

1) 对基学习器采用 K 折交叉验证, 在每一折都进行训练得到一个预测模型, 其决定系数 r_n 为

$$r_n = 1 - \frac{\sum_{x_i \in F_n} (y_i - \hat{y}_i)^2}{\sum_{x_i \in F_n} (y_i - \bar{y})^2} \quad (20)$$

其中, r_n 为模型在第 n 折的决定系数; F_n 为 K 折划分出的第 n 折的数据集; $\bar{y}$ 为实际观测值的均值.

2) 每个基学习器的总体决定系数为

$$R_i = \frac{\sum_{n=1}^K r_n}{K} \quad (21)$$

3) 通过基学习器的总体决定系数, 计算每个基学习器的权重为

$$W_i = \frac{R_i}{\sum_{j=1}^N R_j} \quad (22)$$

其中, N 为模型的数量。

4) 结合 K 折交叉验证中基学习器的输出结果及其对应权重, 构建精度加权后的训练数据集 $\{W_1Y_1, W_2Y_2, \dots, W_bY_b\}$ 。其中, Y_b 为第 b 个基学习器的输出; W_b 为第 b 个基学习器的权重系数。

5) 将精度加权以后的数据集输入到元学习器进行训练, 使用训练好的元学习器对测试数据集进行预测, 得到最终的输出结果。经过精度加权后的模型结构如图2所示。

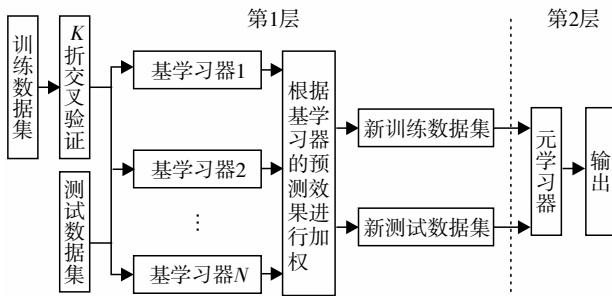


图2 改进的Stacking集成学习模型结构

Fig. 2 Structure diagram of the improved stacking-based ensemble learning model.

3 数据处理及模型优化

3.1 数据集介绍

本研究数据来自深层油井管材腐蚀的室内模拟实验, 包括环境因素、离子浓度、材质和腐蚀情况。环境因素为温度、含水率、 H_2S 分压和 CO_2 分压, 离子浓度为 Cl^- 浓度、 $Mg^{2+}+Ca^{2+}$ 浓度和 SO_4^{2-} 浓度, 材质为P110钢和P110S钢, 腐蚀情况为平均腐蚀速率和点蚀速率。部分样本数据请扫描论文末页右下角二维码查看补充材料表S1。

3.2 数据处理

3.2.1 数据筛选

当环境含水率(质量分数)不足100%时, 所记录的腐蚀情况仅限于平均腐蚀速率而未涵盖点蚀速率。因此, 对2种腐蚀数据进行筛选, 最终获得1594组平均腐蚀速率数据和178组点蚀速率数据。

3.2.2 数据标准化

由于数据集的特性, 为防止因为输入特征数值之间的数量级差距过大造成误差过大, 采用标准化对输入数据进行处理, 以此保证各个油管参数的取值处于合理的分布范围:

$$x_i^* = \frac{x_i - \mu_i}{\sigma_i} \quad (23)$$

其中, x_i^* 为标准化处理后的变量数值; μ_i 为数据序列均值; σ_i 为数据序列标准差。

3.2.3 数据特征工程

由于机器学习算法大多只能处理数值类型的数据, 对于特征类型的数据无法直接处理。基于此, 针对实验数据集中的材质信息的特征变量进行独热编码处理, 将其转化为数值类型变量。由于实验数据中有2种类型的管道材质, 因此设置材质参数为2, 分别在对应的材质参数取值为1, 另一种材质参数取值为0, 处理后的数据请扫描论文末页右下角二维码查看补充材料表S2。

3.2.4 数据扩充

由于通过实验得到的点蚀速率数据容易出现数据分布不均, 数据量少的情况, 为了缓解过拟合问题并平衡类别分布, 使用添加随机扰动项的方法对训练数据集进行数据扩充, 生成随机扰动项为

$$x^* = x + a_x, \quad a_x \sim U(-n_x, n_x) \quad (24)$$

其中, x^* 和 x 为随机扰动添加后和添加前的变量数值; a_x 为变量 x 对应的随机扰动数值; n_x 为变量 x 对应的随机扰动范围。本实验中, n_x 的取值依据各特征变量的实验测量精度和固有波动范围确定, 温度扰动的 n_x 设定为其原始值的2%, 分压和离子浓度的 n_x 设定为其原始值的1%, 含水率不进行扰动。

为确保扩充数据的物理合理性, 在生成后对所有数据进行范围检查和截断处理, 强制要求所有特征变量(如离子浓度、分压和含水率等)处于物理上可行的正数范围内, 且含水率均为100%。此操作仅在训练集上进行, 测试集保持不变以保障模型评估的公正性。为验证数据扩充策略的有效性并评估过拟合风险, 本研究系统测试了不同扩充倍数对模型性能的影响, 结果如图3所示。当扩充倍数为5时, 继续增加数据量对性能提升有限, 因此选择将143组点蚀训练集增至858组。该策略通过在特征空间的局部邻域内创造合理变体, 有效增强了模型的泛化能力, 而未引入原始物理空间下的非物理值问题。

3.3 评价指标

油井管材腐蚀预测本质上是一个回归预测问题, 使用平均绝对误差(mean absolute error, MAE) η_{MAE} 、均方误差(mean squared error, MSE) η_{MSE} 以及决定系数 R^2 三个指标来评价预测结果的准确性, 计

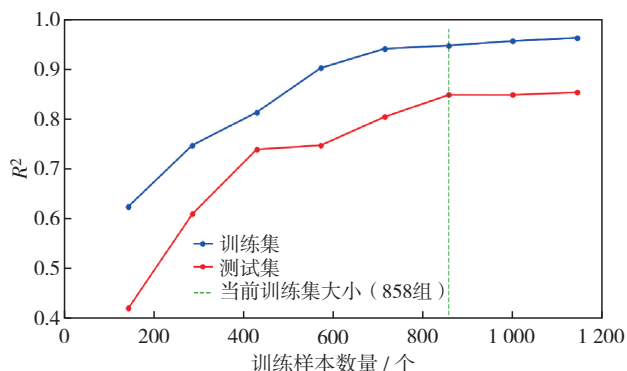


图3 点蚀速率学习曲线

Fig. 3 Learning curves (R^2) for pitting corrosion rate prediction with increasing training sample size: training set (blue line) and testing set (red line). The green dashed line indicates the current training set size (858 samples).

算公式分别为

$$\eta_{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (25)$$

$$\eta_{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (26)$$

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (27)$$

η_{MAE} 和 η_{MSE} 越小, 说明预测结果与真实值之间的偏差越小, 模型的拟合能力越强; R^2 衡量的是全部特征变量对目标变量变异的解释程度, 可以直观反映模型回归拟合效果, 取值范围为[0, 1]. R^2 值越接近 1, 模型拟合效果越好.

3.4 模型优化

为了使模型性能达到最优, 使用网格搜索和 5 折交叉验证对基学习器的超参数进行参数调优, 在给定的参数范围内进行穷举式搜索, 将各个超参数的取值进行排列组合, 使得所有可能组合的结果生成“网格”, 并对所有超参数组合进行模型训练, 寻找出性能最优的参数组合, 流程如图 4.

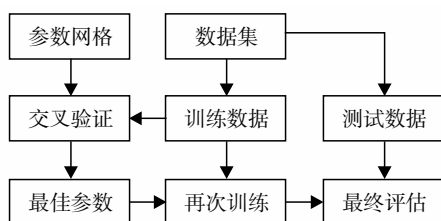


图4 超参数调优流程

Fig. 4 Flowchart of hyperparameter tuning process.

4 实验验证

4.1 传统 Stacking 集成模型的构造

4.1.1 Stacking 学习器的选择

本研究充分考虑常用的回归预测算法及其特点, 选择 XGBoost 回归模型、随机森林回归模型、SVR 回归模型以及 GBDT 回归模型参与构建基学习器的堆叠模型. 其中, XGBoost 具有并行计算和正则化能力, 能够处理非线性关系并防止过拟合, 但参数调优较复杂; 随机森林通过集成提高预测精度, 具有较好抗噪能力和泛化性能, 但可能产生过拟合; SVR 通过在高维空间中寻找最优超平面来进行回归预测, 适用于处理非线性及高维数据问题, 但在应对庞大数据集时性能会受到一定限制; GBDT 通过迭代拟合残差来优化预测结果, 能够捕捉到数据中的复杂关系, 但计算复杂度相对较高. 在选择元学习器时, 通常期望其能降低 Stacking 集成模型的复杂度, 减少内存消耗并提升运算速度. 因此, 本研究采用训练效率高、拟合性能优越的线性回归模型作为元学习器, 以便对基学习器的预测结果进行更深入的优化处理.

4.1.2 Stacking 基学习器的优化

将各基学习器在原始数据集上进行单独预测并对结果进行分析. 针对 XGBoost 回归模型、随机森林回归模型、SVR 回归模型以及 GBDT 回归模型进行超参数调优. 4 种回归模型的超参数取值如表 1 所示. 首先, 对每个需要调优的超参数设置取值范围, 通过 5 折交叉验证将训练集分成 5 组, 每组的样本不重复. 每次使用 4 组样本作为训练集来训练模型, 剩下 1 组样本作为测试集来计算模型的损失函数. 最后, 将 5 次训练得到的损失函数的平均值作为模型在数据集上最终的损失函数值, 结合 5 折交叉验证获得最优的模型.

将原始数据分为训练样本和测试样本, 对每个样本在最佳参数下的预测结果进行分析. 使用传统 Stacking 集成模型, 采用 10 折交叉验证, 将 4 种基学习器进行集成, 使用相同的训练集和测试集进行观测. 5 种回归模型的预测性能如表 2 和表 3 所示. 从表 2 和表 3 可见, 随机森林回归模型对于平均腐蚀速率的预测误差最小, 表现较好, 其 R^2 值达到了 0.897, XGBoost 回归模型对于点蚀速率的预测误差最小, 表现最好, 其 R^2 值达到了 0.819.

表1 四种基学习器预测性能的最佳参数
Table 1 Optimal parameters for predicting performance of four base learners

模型	超参数集	
	平均腐蚀预测模型	点蚀预测模型
XGBoost	学习率 = 0.1	学习率 = 0.1
	决策树个数 = 100	决策树个数 = 300
	最大树深 = 6	最大树深 = 10
随机森林	决策树个数 = 100	决策树个数 = 71
	最大树深 = 20	最大树深 = 15
	评价标准 = MSE	评价标准 = MAE
SVR	epsilon = 0.1	epsilon = 0.5
	核函数 = RBF	核函数 = Poly
GBDT	学习率 = 0.1	学习率 = 0.1
	决策树个数 = 100	决策树个数 = 100
	最大树深 = 3	最大树深 = 4

表2 五种回归模型的平均腐蚀速率预测性能
Table 2 Prediction performances of the average corrosion rate by five regression models

模型	$\eta_{\text{MSE}}/(\text{mm} \cdot \text{a}^{-1})$	$\eta_{\text{MAE}}/(\text{mm} \cdot \text{a}^{-1})$	R^2
XGBoost	0.042	0.082	0.877
随机森林	0.035	0.079	0.897
SVR	0.181	0.155	0.472
GBDT	0.080	0.143	0.768
传统 Stacking	0.027	0.072	0.921

表3 五种回归模型的点蚀速率预测性能
Table 3 Prediction performances of the pitting corrosion rate by five regression models

模型	$\eta_{\text{MSE}}/(\text{mm} \cdot \text{a}^{-1})$	$\eta_{\text{MAE}}/(\text{mm} \cdot \text{a}^{-1})$	R^2
XGBoost	0.045	0.149	0.819
随机森林	0.059	0.181	0.761
SVR	0.141	0.317	0.433
GBDT	0.098	0.246	0.605
传统 Stacking	0.043	0.147	0.824

将基学习器中表现最好的回归模型与集成了4种基学习的传统 Stacking 回归模型预测效果进行对比, 可以看出传统 Stacking 回归模型都具有更好的表现效果, 其中, 对于平均腐蚀速率的预测, 传统 Stacking 回归模型 MSE 下降了 22.9%, MAE 下降了 8.9%, R^2 提高了 2.7%; 对于点蚀速率的预测, 传统 Stacking 回归模型 MSE 下降了 4.4%, MAE

下降了 1.3%, R^2 提高了 0.61%.

4.2 基于 R^2 的改进 Stacking 集成模型的验证

使用本研究提出的基于 R^2 的改进 Stacking 集成模型, 对训练集和测试集进行模型训练和预测. 基学习器采用 10 折交叉验证, 得到各基学习器的总体决定系数 R^2 及权重如表 4 和表 5 所示.

表4 平均腐蚀速率预测的基学习器决定系数 R^2 及权重
Table 4 R^2 and weights of base learners for the average corrosion rate prediction

指标	XGBoost	随机森林	SVR	GBDT
R^2	0.700	0.782	0.361	0.510
权重	0.297	0.332	0.153	0.217

表5 点蚀速率预测的基学习器决定系数 R^2 及权重
Table 5 R^2 and weights of base learners for the pitting corrosion rate prediction

指标	XGBoost	随机森林	SVR	GBDT
R^2	0.642	0.646	0.451	0.523
权重	0.284	0.285	0.199	0.231

依据计算得到的权重, 将精度加权以后的数据集输入到元学习器进行训练, 使用训练好的元学习器对测试数据集进行预测, 得到改进 Stacking 集成模型的预测结果, 其预测性能如表 6 所示, 改进前后的 Stacking 集成模型对测试集的预测结果对比如图 5 所示. 为了更直观地展示各种回归模型的预测性能对比, 将各模型的决定系数 R^2 进行了汇总, 对比结果如图 6 所示. 由图 6 可知, 对比传统的 Stacking 集成模型, 改进后的 Stacking 集成模型对于平均腐蚀速率的预测 MSE 下降了 25.9%, MAE 下降了 9.7%, R^2 为 0.942, 提高了 2.3%, 而对于点蚀速率的预测, 改进后的 Stacking 集成模型的 MSE 下降了 11.6%, MAE 下降了 2.0%, R^2 为 0.846, 提高了 2.7%. 可以看出, 传统 Stacking 集成模型的预测性能要优于单个的基学习, 进一步地, 本研究所提出的基于决定系数 R^2 的改进

表6 改进 Stacking 集成模型的腐蚀速率预测性能
Table 6 Prediction performance of the corrosion rate by improved Stacking ensemble model

预测目标	$\eta_{\text{MSE}}/(\text{mm} \cdot \text{a}^{-1})$	$\eta_{\text{MAE}}/(\text{mm} \cdot \text{a}^{-1})$	R^2
平均腐蚀速率	0.020	0.065	0.942
点蚀速率	0.038	0.144	0.846

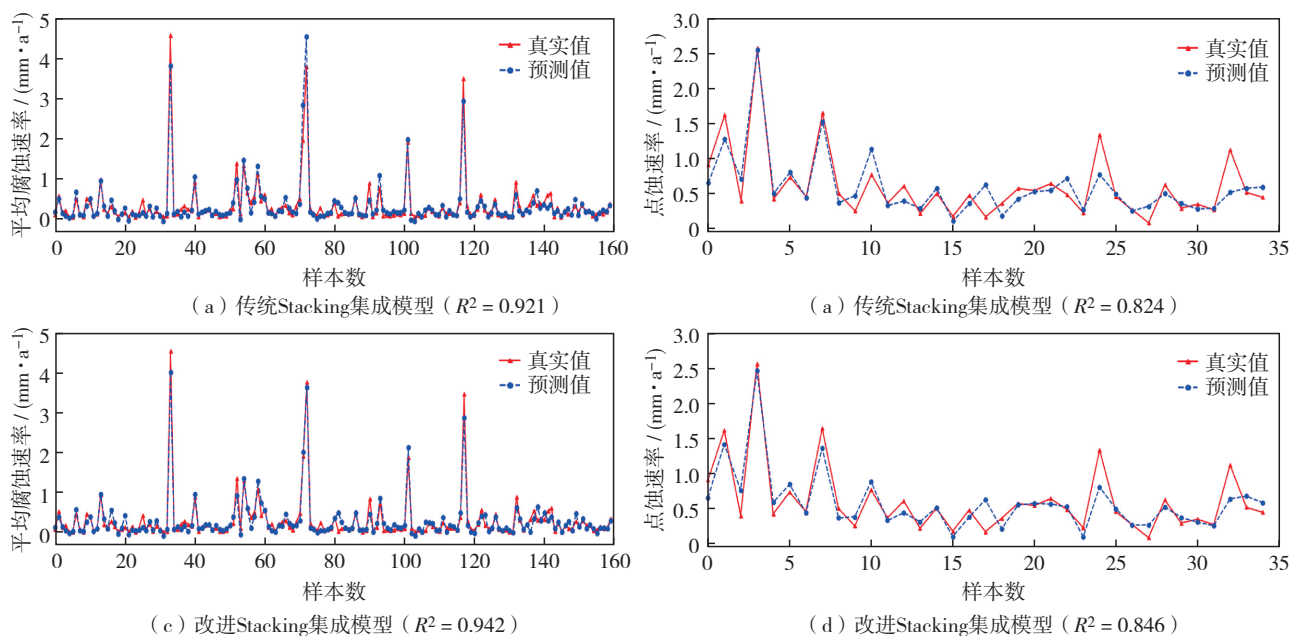


图5 传统Stacking集成模型和改进Stacking集成模型对平均腐蚀速率及点蚀速率的预测结果

Fig. 5 Prediction results for (a) average corrosion rate, (b) pitting corrosion rate using the traditional Stacking ensemble model, (c) average corrosion rate and (d) pitting corrosion rate using the improved Stacking ensemble model, with the red triangles represent actual values and the blue circles represent predictive values.

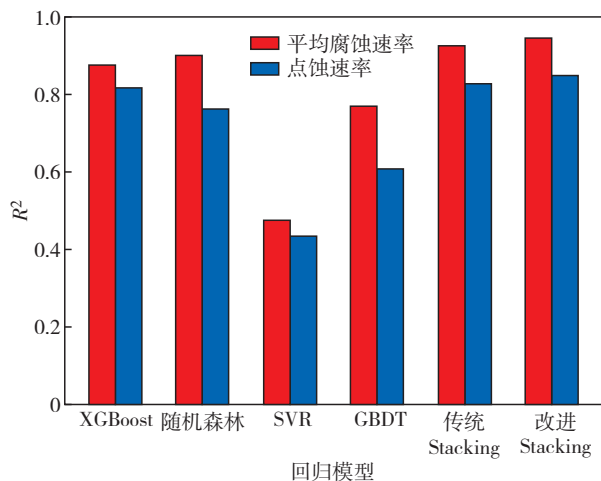


图6 六种回归模型的腐蚀速率预测决定系数 R^2 对比

Fig. 6 Comparison of the R -squared (R^2) for average corrosion rate predictions (red column) and pitting corrosion rate predictions (blue column) by six regression models.

Stacking集成模型又在所有评价指标上均展现出了相较于传统Stacking集成模型的优越性。

5 结 论

基于深层油田管材腐蚀的室内模拟实验数据,通过数据特征处理和数据扩充等工作,得到了适用于机器学习模型的油井管材腐蚀数据集。对于预测

模型的构造,首先使用了传统Stacking集成模型,其中,基学习器使用了XGBoost回归模型、随机森林回归模型、SVR回归模型以及GBDT回归模型,元学习器使用线性回归模型。通过实验对比可知,传统Stacking集成模型的预测效果要优于各个基学习器回归模型。针对传统Stacking集成学习方法在第1层学习中未充分考虑基学习器间差异性所带来的问题,提出一种基于决定系数 R^2 的改进Stacking集成回归模型。该模型是在基学习器进行 K 折交叉验证训练时,计算出每一折模型的决定系数 R^2 ,从而获得每个基学习器的总体决定系数,并以此得到每个基学习器的权重,结合基学习器的输出结果及其对应权重,构建精度加权后的训练数据集作为元学习器的输入,使用测试数据集对训练好的元学习器进行预测,得到最终的预测结果。通过实验可知,相比较于传统Stacking集成回归模型,改进后的Stacking集成回归模型应用在油井管材腐蚀的预测中具有更好的预测效果。在平均腐蚀速率的预测上,平均绝对误差和均方误差分别降低了25.9%和9.7%,决定系数提高了2.3%;在点蚀速率的预测上,平均绝对误差和均方误差分别降低了11.6%和2.0%,决定系数提高了2.7%,结果表明,本研究算法具有良好的预测性能。

基金项目: 国家自然科学基金资助项目(12171482, U23A20301); 中国石油大学油气资源与工程全国重点实验室资助项目(PRE/DX-2504).

作者简介: 黄晗(2023216113@student.cup.edu.cn), 中国石油大学(北京)硕士研究生. 研究方向: 大数据分析.

引文: 黄晗, 陈长风, 贾小兰, 等. 基于改进Stacking集成学习的深层油井管腐蚀预测[J]. 深圳大学学报理工版, 2026, 43(1): 7-16.

参考文献 / References:

- [1] 王兆明, 温志新, 贺正军, 等. 全球近10年油气勘探新进展特点与启示[J]. 中国石油勘探, 2022, 27(2): 27-37.
WANG Zhaoming, WEN Zhixin, HE Zhengjun, et al. Characteristics and enlightenment of new progress in global oil and gas exploration in recent ten years [J]. China Petroleum Exploration, 2022, 27(2): 27-37. (in Chinese)
- [2] 王震, 王思允. 全球深水油气勘探开发进展与研究综述[J]. 天然气与石油, 2024, 42(6): 1-10.
WANG Zhen, WANG Siyun. Comprehensive review of advances and research in global deepwater exploration and development [J]. Natural Gas and Oil, 2024, 42(6): 1-10. (in Chinese)
- [3] 郭旭升, 胡宗全, 李双建, 等. 深层—超深层天然气勘探研究进展与展望[J]. 石油科学通报, 2023, 8(4): 461-474.
GUO Xusheng, HU Zongquan, LI Shuangjian, et al. Progress and prospect of natural gas exploration and research in deep and ultra-deep strata [J]. Petroleum Science Bulletin, 2023, 8(4): 461-474. (in Chinese)
- [4] 宋成立. 多因素作用下油气集输管道的腐蚀行为及预测模型[D]. 西安: 西北大学, 2022.
SONG Chengli. Corrosion behavior and prediction model of oil and gas gathering pipelines under multi-factors [D]. Xi'an: Northwest University, 2022. (in Chinese)
- [5] 卢主飞, 贺毅龙, 刘树岭. 油管腐蚀分析与防护技术研究[J]. 中国设备工程, 2024(17): 167-169.
LU Zhufei, HE Yilong, LIU Shuling. Research on corrosion analysis and protection technology of oil pipes [J]. China Plant Engineering, 2024 (17): 167-169. (in Chinese)
- [6] 王志华, 洪家骏, 张宏奇. 机理与数据融合驱动的原油集输管道腐蚀失效预测[J]. 管道保护, 2025, 2(1): 2-13.
WANG Zhihua, HONG Jiajun, ZHANG Hongqi. Corrosion failure prediction of crude oil gathering and transportation pipelines driven by mechanism and data [J]. Pipeline Protection, 2025, 2(1): 2-13. (in Chinese)
- [7] XU Lei, WANG Yunfu, MO Lin, et al. The research progress and prospect of data mining methods on corrosion prediction of oil and gas pipelines [J]. Engineering Failure Analysis, 2023, 144: 106951.
- [8] OLSEN S. CO₂ corrosion prediction by use of the NORSOK M-506 model - guidelines and limitations [C]// Proceedings of World Corrosion Congress 2003. Richardson, USA: OnePetro, 2003: NACE-03623.
- [9] 郑度奎, 李昊燃, 程远鹏, 等. 油气管道的CO₂腐蚀预测模型和预测方法[J]. 腐蚀与防护, 2020, 41(3): 48-53, 58.
ZHENG Dukui, LI Haoran, CHENG Yuanpeng, et al. CO₂ corrosion prediction model and prediction method for oil and gas pipeline [J]. Corrosion & Protection, 2020, 41(3): 48-53, 58. (in Chinese)
- [10] DE WAARD C, MILLIAMS D E. Carbonic acid corrosion of steel [J]. Corrosion, 1975, 31(5): 177-181.
- [11] 李霞, 王丹丹, 夏明磊, 等. 腐蚀模型与海底管道内检测数据的对比研究[J]. 清洗世界, 2022, 38(9): 34-36.
LI Xia, WANG Dandan, XIA Minglei, et al. Comparative study of corrosion model and internal inspection data of submarine pipeline [J]. Cleaning World, 2022, 38(9): 34-36. (in Chinese)
- [12] 李磊, 朱恩雄, 丰劲松, 等. 某集气管道苛刻服役工况下CO₂腐蚀分布规律仿真模拟[J]. 腐蚀与防护, 2025, 46(1): 105-110.
LI Lei, ZHU Enxiong, FENG Jinsong, et al. Simulation of CO₂ corrosion distribution in a gas gathering pipeline under harsh service conditions [J]. Corrosion & Protection, 2025, 46(1): 105-110. (in Chinese)
- [13] XUE Hao, GAO Qiuying, ZHAO Yang, et al. A life prediction model for P110S steel in deep-well environments with H₂S/CO₂ coexistence based on multi-factor chemometric drive [J]. Corrosion Science, 2024, 240: 112475.
- [14] 张璐瑶. 基于退化建模的海底原油管道腐蚀风险评估方法研究[D]. 西安: 西安建筑科技大学, 2022.
ZHANG Luyao. Research on corrosion risk assessment method of subsea crude oil pipeline based on degradation modeling [D]. Xi'an: Xi'an University of Architecture and Technology, 2022. (in Chinese)
- [15] REN Chiyuan, QIAO Wei, TIAN Xin. Natural gas pipeline corrosion rate prediction model based on BP

- neural network [C]// Fuzzy Engineering and Operations Research. Berlin: Springer Berlin Heidelberg, 2012: 449-455.
- [16] LIAO Kexi, YAO Quanke, WU Xia, et al. A numerical corrosion rate prediction method for direct assessment of wet gas gathering pipelines internal corrosion [J]. *Energies*, 2012, 5(10): 3892-3907.
- [17] ZHANG Weidong, SHEN Bin, AI Yibo, et al. Gas pipeline corrosion prediction based on modified support vector machine and unequal interval model [J]. *Applied Mechanics and Materials*, 2013, 373-375: 1987-1994.
- [18] OSSAI C I. Corrosion defect modelling of aged pipelines with a feed-forward multi-layer neural network for leak and burst failure estimation [J]. *Engineering Failure Analysis*, 2020, 110: 104397.
- [19] CHEN Tianqi, GUESTRIN C. XGBoost: a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: Association for Computing Machinery, 2016: 785-794.
- [20] BREIMAN L. Random forests [J]. *Machine Learning*, 2001, 45(1): 5-32.
- [21] MOSES A, CHEN Ding, WAN Peng, et al. Prediction of electrochemical corrosion behavior of magnesium alloy using machine learning methods [J]. *Materials Today Communications*, 2023, 37: 107285.
- [22] HUANG Sheng, WANG Li, LI Zaoyuan, et al. Machine learning-based prediction model for CO₂-induced corrosion on oil well cement under high-pressure and high-temperature condition [J]. *Construction and Building Materials*, 2024, 414: 134999.
- [23] NASERZADEH Z, NOHEGAR A. Development of HGAPSO-SVR corrosion prediction approach for offshore oil and gas pipelines [J]. *Journal of Loss Prevention in the Process Industries*, 2023, 84: 105092.
- [24] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. *The Annals of Statistics*, 2001, 29(5): 1189-1232.
- [25] DONG Zhenzhen, ZHANG Min, LI Weirong, et al. Development of a predictive model for carbon dioxide corrosion rate and severity based on machine learning algorithms [J]. *Materials*, 2024, 17(16): 4046.
- [26] DEMİR N, DALKILIÇ G. Modified stacking ensemble approach to detect network intrusion [J]. *Turkish Journal of Electrical Engineering and Computer Sciences*, 2018, 26(1): 418-433.
- [27] 邓威, 郭钊秀, 李勇, 等. 基于特征选择和Stacking集成学习的配电网网损预测[J]. *电力系统保护与控制*, 2020, 48(15): 108-115.
- DENG Wei, GUO Yixiu, LI Yong, et al. Power losses prediction based on feature selection and Stacking integrated learning [J]. *Power System Protection and Control*, 2020, 48(15): 108-115. (in Chinese)
- 【中文责编：坪梓；英文责编：远鹏】



补充材料