文章编号：0559-9350(2026)02-0232-13

基于细菌觅食优化Stacking集成学习的钻孔效率预测模型研究

关　涛[1]，缴春祺[1]，俞　澎[1]，余　佳[1]，郭振邦[1]，程正飞[2]
（1. 天津大学 水利工程智能建设与运维全国重点实验室，天津　300072；
2. 水电水利规划设计总院有限公司，北京　100120）

摘要：钻孔效率是堆石坝料场开挖施工进度仿真的关键参数，其预测准确性直接关系到仿真模型的可靠性。针对现有数学方法预测效率较低、单一学习器难以满足仿真精度要求、集成学习研究中超参数调整方法局部搜索精细度不足的问题，本文提出了一种基于细菌觅食优化Stacking集成学习的钻孔效率预测模型。首先，以某堆石坝现场采集的钻孔效率数据为目标变量，以其影响因素（如钻孔深度、岩石性质、高程等）为特征变量构建数据集；其次，采用XGBoost、LightGBM和MLP三种异质基学习器并行训练，并引入细菌觅食优化算法模拟趋化和繁殖行为，通过R^2曲线实时追踪，迭代优化各基学习器的超参数，确保输出稳定的“元特征”；最后，将各基学习器的预测结果输入支持向量回归（SVR）元学习器，通过整合多模型的互补信息，在抑制偏差与方差的同时获得集成预测结果。实验结果表明，经细菌觅食优化后，各基学习器的R^2均可达到0.93以上，*PCC*值均超过0.97，集成模型在整个样本数据集上的学习曲线也平滑稳定，残差分析显示预测值与真实值的残差序列在零均值线附近均匀分布，最终结果的*PCC*值接近0.98，可以满足施工过程仿真需求。

关键词：钻孔效率；施工仿真；Stacking集成学习；XGBoost；LightGBM；MLP；支持向量机
中图分类号：TV641.4　　**文献标识码：**A　　**doi：**10.3724/j.slxb.20250402

1　研究背景

堆石坝施工仿真技术通过进度模拟解决资源分配等问题，其可靠性依赖于仿真参数精度。钻孔效率作为关键参数，预测准确性直接影响模型效果。作为影响爆破作业及大坝填筑进度分析准确性的重要参数，如何根据已知条件分析得到钻孔效率是目前施工进度仿真分析的一项重要课题。然而，现有的研究中存在数学方法[1-4]预测效率较低、单一学习器难以满足仿真精度要求、集成学习研究中运用的超参数（hyperparameter）调整方法局部搜索精细度不足等问题。由于可全面考虑复杂非线性关系，机器学习技术可以为钻孔效率预测提供重要手段，使预测过程更加简单高效。

目前，已有诸多学者采用包括机器学习在内的方法进行了岩体钻进效率预测研究[5-7]。Zhou等[8]提出了一种基于已识别钻井条件的在线混合预测模型，解决了实际钻井数据机械钻速建模困难的问题；Abbas等[9]考虑了多个影响因素的影响，采用人工神经网络模型（Artificial Neural Network，ANN）预测钻机钻速与斜井钻进时间；Jia等[10]提出遗传算法优化的反向传播神经网络（Genetic Algorithm-Back Propagation Neural Network，GA-BP）模型有效解决了传统BP神经网络在钻孔预测方面的局限性；王晓玲等[11]基于Q学习改进猎人猎物优化算法优化了超级学习器集成学习模型，并结合沙普利加性解释

收稿日期：2025-07-17；网络首发日期：2026-01-20
网络首发地址：https://link.cnki.net/urlid/11.1882.TV.20260119.1710.002
基金项目：高寒条件下考虑鲁棒性的高拱坝施工过程智能仿真（52379131）；国家自然科学基金原创探索计划项目（52350417）；国家自然科学基金联合基金重点项目（U24B20111）；国家自然科学基金项目（52279137）；中国电力建设股份有限公司科技项目（DJ-ZDXM-2020-50）
作者简介：关涛（1989—），博士，副教授，主要从事水利水电工程施工智能化研究。E-mail：tao.guan@tju.edu.cn
通信作者：余佳（1992—），博士，副教授，主要从事水利工程智能建设研究。E-mail：yujia@tju.edu.cn

(SHAP)理论，在钻孔预测精度大幅度提高的同时实现了土石方开挖钻孔效率的高精度预测与影响因素可解释分析。由于钻孔效率受岩体特性、钻具参数与操作条件等多重因素交织影响，单一学习器在预测时往往存在泛化能力有限、参数敏感性高等问题，难以满足高精度施工仿真的要求。为此，集成学习策略被引入以进一步提升预测精度并克服单一模型的局限性[12-14]。在多种集成学习方法中[15]，Stacking(堆叠)通过将多个基学习器的预测输出作为新的特征输入至元学习器，实现多源预测信息的融合与重用。该策略不仅保留了原始数据的信息结构，还能有效降低模型偏差，从而提高整体预测的稳健性与泛化能力。例如，Ibrahim等[16]提出了集成优化与缩放的混合集成学习模型(Hybrid Ensemble Learning with Integrated Optimization and Scaling，HELIOS-Stack)，通过堆叠集成随机森林回归、XGBoost与多层感知机等基础模型，实现了岩石节理粗糙度的高精度预测；Lv等[17]针对小样本数据提出了基于改进布谷鸟搜索算法优化的Stacking集成学习(Improved Cuckoo Search Optimized Stacking-based Ensemble Learning，ICSO-Stacking)的钻进效率预测模型，通过集成四种异质机器学习算法，提高了预测精度；Tian等[18]借助Stacking方法全面考虑了钢筋混凝土剪力墙抗剪承载力的影响因素，实现了构件抗剪承载力的高精度预测；邓怡徽等[19]通过ABC优化Stacking集成策略的多分类器组合，实现了虚拟钻孔地层属性的高精度预测。

上述研究成果实现了机器学习在工程领域的应用，然而机器学习模型在构建的过程中通常需要处理高维及非线性的优化问题。此类问题的目标函数往往存在大量局部极小值、梯度消失、爆炸现象以及参数空间的高度冗余性等问题，使得传统解析方法难以直接求解全局最优解。同时，现有集成学习研究中超参数优化算法存在局部搜索细度不足的问题，因此，引入高效、鲁棒、全面的优化算法成为解决此类问题的必然选择。细菌觅食优化算法(Bacterial Foraging Optimization，BFO)作为一种广泛应用的群体无导数优化方法，在对基学习器进行超参数调优时具有显著优势。算法通过模拟细菌趋化、繁殖与迁徙等行为，兼顾局部精搜与全局扰动，能在多峰、无导数或含噪目标上有效跳出局部最优并便于并行化。相较于粒子群优化与差分进化等算法，BFO在局部微调与全局跳出方面机制更明确，更适合对超参数进行全局优化。在利用BFO算法对模型超参数进行优化并解决工程实际问题的各种研究中，王佳俊等[20]针对堆石坝压实质量非线性回归预测问题，使用基于BFO的自适应混沌改进策略(Adaptive chaotic Bacteria Foraging Algorithm，AC-BFA)，优化核方法与模糊逻辑模型的超参数，显著提升了模型拟合精度与鲁棒性；Alrashed等[21]利用BFO算法对模型参考自适应系统中的PI控制器参数进行动态优化，解决了永磁同步电机参数与速度估计中的非线性回归问题；Anand等[22]利用BFO对熔融沉积建模中PETG材料的工艺参数进行了优化，实现了材料拉伸强度和弯曲强度的显著提升。可见，BFO算法通过趋化过程的动态调整与消除扩散机制能在高维参数空间中保持全局探索与局部精细搜索的平衡，尤其适合多超参数场景下的复杂寻优。

综上所述，本研究利用Stacking集成学习方法，同时基于BFO对超参数进行调节建立预测模型，从而可以获得更准确的钻进效率预测数据，以实现施工仿真参数的高精度建模。

2　研究框架

本文所提出的基于细菌觅食优化Stacking集成学习的钻孔效率预测模型框架如图1所示。该框架旨在通过一个系统化、智能化的建模流程，解决堆石坝钻孔作业中多因素耦合的非线性预测难题。首先基于我国西南某大型堆石坝工程实际采集的施工数据采集样本，明确以“钻孔效率”作为模型预测的目标变量(Y)，并系统地遴选了钻孔深度、岩石性质、高程等10个核心参数作为特征变量(X_1，X_2，…，X_n)，确保了模型输入与工程机理的紧密结合。其次，框架构建一个异质基学习器集合，选取了在理论基础上具有显著差异性和互补性的三种算法：极端梯度提升机(Extreme Gradient Boosting，XGBoost)侧重于精度与正则化、轻量梯度提升机(Light Gradient Boosting Machine，LightGBM)侧重效率与特征处理以及侧重非线性映射的多层感知机(Multilayer Perceptron，MLP)模型。图中展示了各算法的关键原

理，其输出结果的多样性是后续集成成功的关键前提。再次，本框架的创新亮点在于引入了BFO算法对基学习器进行独立超参数优化。与传统繁琐的手动调参相比，BFO模拟细菌群体的趋化、繁殖等行为，在参数空间中进行自适应地全局寻优。优化过程以5折交叉验证的R^2作为适应度函数，其收敛过程由R^2曲线实时监控与验证。此步骤确保了每个基学习器在输入至下一环节前，均已达到其独立的最优性能状态，输出稳定而精确的初级预测结果。然后，经过优化的基学习器预测结果作为新的训练特征，输入至元学习器。本研究选用支持向量回归(Support Vector Regression，SVR)，原因在于其能够找到最优决策边界以最大化间隔，能最有效地加权组合基学习器的预测结果，实现对单一模型偏差与方差的协同抑制与互补，从而生成最终统一的预测模型。最后，框架通过一个多维度、可视化的模型效果评价体系形成闭环验证。通过残差图分析误差分布、学习曲线诊断模型泛化能力、以及预测值与真实值对比(PCC)图直观展示一致性，从多角度充分论证了所构建模型的有效性、可靠性及优越性。

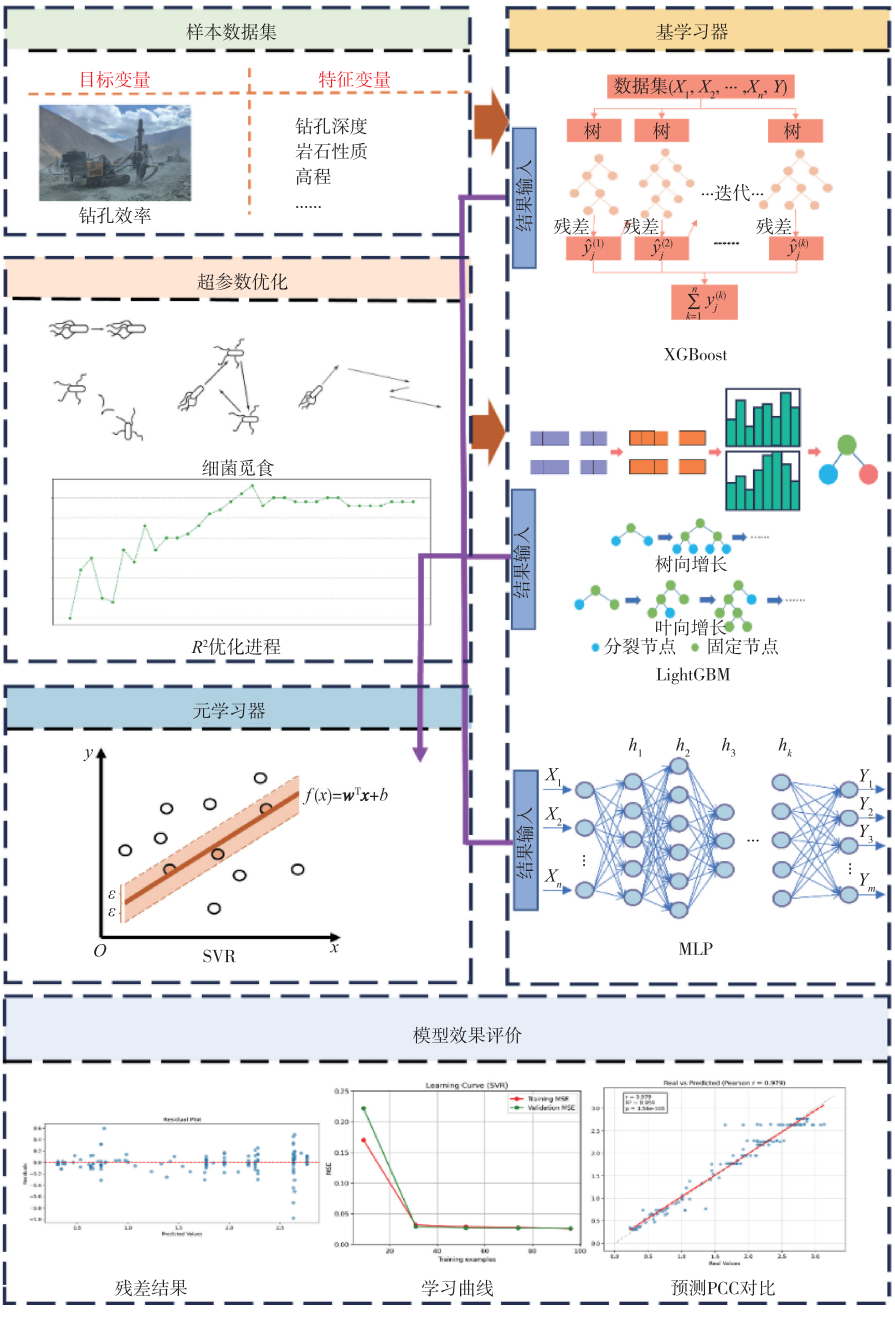


图1　钻孔效率集成学习预测模型框架图

Fig. 1　Framework of the ensemble learning prediction model for drilling efficiency

综上所述，该流程图清晰地阐述了一个从数据准备到模型验证的完整、先进的建模逻辑。其核心价值在于通过BFO优化提升个体性能，再通过Stacking集成发挥协同效应，最终为复杂钻孔施工场景的预测问题提供了一个精准、稳健且自动化的解决方案。

3 钻孔效率预测集成学习模型

在Stacking框架下，基学习器与元学习器通过分层协作共同提升模型的泛化能力。本研究将基学习器作为第一层模型，其预测结果输出为包含三个特征向量和一个目标向量的中间样本集。元学习器作为第二层模型对中间样本集进行学习训练并输出最终结果。由于XGBoost的正则化树结构可解析特征交互、LightGBM的直方图算法可优化高维稀疏建模、多层感知机(MLP)的深度非线性可映射捕捉全局复杂模式所形成的异质架构互补优势，本研究选用这三种异质学习器模型进行融合作为基学习器，通过树模型局部精准拟合与神经网络全域映射的协同机制，实现特征空间多尺度覆盖，并利用模型误差的异源分布特性，有效增强集成系统对钻孔效率及其影响因素复杂关系的泛化鲁棒性，而后以SVR模型作为元学习器进行模型搭建。

3.1 基学习器

3.1.1 XGBoost XGBoost的核心机制是通过前向分布算法（Additive Training）构建多棵提升树，将多个弱学习器集成，从而获得高性能模型。其核心思想是在每一轮迭代中，通过添加新的决策树逐步优化钻孔效率预测结果，同时严格约束模型复杂度以防止过拟合。XGBoost设f_t为第t次迭代时决策树的模型函数，每棵决策树被分解为结构函数q和叶节点权重w两个部分，q将样本依据工程特征(如高程、钻机型号、钻杆数量等)映射到叶节点，w则表示该叶节点对应的钻孔效率预测值。为了控制模型复杂度，XGBoost引入正则化项：

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \tag{1}$$

式中：$\Omega(f_t)$为正则化项；T为叶节点总数；γ和λ为超参数，分别为惩罚叶节点数量和权重数值；w_j为第j片叶对应的权重。为了最小化目标函数，采用加法训练的方法，将第t步的目标函数表示为：

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i,\ \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2}$$

式中：$Obj^{(t)}$为目标函数；n为训练集中的样本数；y_i为真实值；$\hat{y}_i^{(t-1)}$为第i个样本在迭代$t-1$时的预测结果；f_t为本轮新增的决策树模型。其中的损失函数$l()$衡量预测值与真实值的偏差，正则化项约束新增树的结构复杂度。

3.1.2 LightGBM LightGBM的创新在于引入直方图算法、梯度单边采样(Gradient-based One-Side Sampling，GOSS)和叶向增长(Leaf-wise)策略，优化决策树构建过程，在提升训练效率的同时控制模型复杂度。该模型通过将连续钻孔影响特征离散为区间，基于统计信息高效确定最优分裂点，显著降低特征遍历的计算开销。GOSS方法则通过对样本按梯度绝对值排序，以$a\%$大梯度样本的采样比例，$b\%$为小梯度样本的采样比例，保留前$a\%$高梯度样本，并在其余样本中随机采样$b\%$，在压缩数据量的同时保留关键信息。为修正样本不均带来的估计偏差，对低梯度样本的梯度进行缩放处理：

$$g_i' = \begin{cases} g_i & i \in S_h \\ \dfrac{1-a\%}{b\%} g_i & i \in S_r \end{cases} \tag{3}$$

式中：g_i为样本i的梯度值；S_h为高梯度样本集合；S_r为随机采样样本集合。该机制在保证模型梯度估计精度的同时，降低了训练样本规模，从而显著提升训练效率。

同时模型采用叶向增长策略，每次选择叶子中增益最大的节点分裂，以更少叶子节点实现高精度，同时通过最大深度限制和正则化项控制复杂度，防止过拟合。其目标函数融合损失函数与正则项：

$$Obj = \sum_{i=1}^{n} l(y_i,\ \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{4}$$

式中：$l(y_i,\ \hat{y}_i)$衡量钻孔效率样本i的预测值$\hat{y}_i$与真实值y_i的偏差；正则项$\Omega(f_t)$(见式(2))约束第t棵树的复杂度。通过迭代优化每棵树，逐步提升模型整体钻孔效率预测性能，同时抑制过拟合。

3.1.3 MLP MLP是一种神经网络架构，由输入层、一个或多个隐藏层以及输出层组成。钻孔效率样本数据从输入层进入，经过隐藏层中神经元的加权求和与非线性激活函数处理后，传递到输出层得到最终结果。以单隐层结构为例，其前向传播过程为：

$$\begin{cases} \boldsymbol{h} = f(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) \\ \boldsymbol{y} = \boldsymbol{W}_2 \boldsymbol{h} + \boldsymbol{b}_2 \end{cases} \tag{5}$$

式中：$\boldsymbol{h}$为隐藏层数；$\boldsymbol{x}$为输入向量；$\boldsymbol{W}_1$，$\boldsymbol{W}_2$和$\boldsymbol{b}_1$和$\boldsymbol{b}_2$为权重与偏置；$f(\ )$为激活函数；$\boldsymbol{y}$为输出预测。

模型根据具体任务设定相应的损失函数衡量预测结果与真实标签之间的差异，训练过程中，模型通过反向传播算法计算梯度，并结合梯度下降法更新参数，从而最小化损失实现学习。

3.2 元学习器 SVR是通过构造ε不敏感间隔带实现鲁棒回归建模，该模型在高维特征空间中寻找一个最优超平面，使得预测值与真实值的偏差在容忍区间$\pm\varepsilon$内的同时最大化间隔宽度。SVR对给定训练集通过非线性映射将样本投影到高维空间，构造回归函数$f(x) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + \boldsymbol{b}$，其中$\boldsymbol{w}$为权重，$\boldsymbol{b}$为偏置。为平衡模型复杂度与预测误差，目标函数定义为：

$$\min_{\boldsymbol{w},b,\xi,\xi^*} \left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \right) \tag{6}$$

式中：$\frac{1}{2}\|\boldsymbol{w}\|^2$为正则化项控制超平面复杂度；$C\sum_{i=1}^{m}(\xi_i + \xi_i^*)$为惩罚项，用于对超出$\varepsilon$间隔带的样本(即松弛变量不为0的情况）进行惩罚，$C$为惩罚系数，$\xi_i$和$\xi_i^*$为松弛变量，其约束条件为：

$$\begin{cases} y_i - f(x_i) \leqslant \varepsilon + \xi_i \\ f(x_i) - y_i \leqslant \varepsilon + \xi_i^* \\ \xi_i,\ \xi_i^* \geqslant 0 \end{cases} \tag{7}$$

式中ε定义间隔带宽度(预测误差小于等于ε时不惩罚)。模型通过引入核函数隐式计算高维内积，避免了显式映射求解后仅依赖支持向量(位于间隔带外的样本)确定模型，使计算效率与噪声免疫力显著优于传统回归方法。

在Stacking框架中，元学习器性能取决于基学习器的预测质量与多样性。基学习器从高程、钻机型号、钻杆数量等工程参数提取互补特征，准确且差异化的输出有助于元学习器整合优势、提升整体精度；反之，偏差或高方差会降低输入信噪比、削弱集成效果。为此，本研究用BFO算法调优基学习器超参数，以提高个体预测能力并为元学习器提供高信噪比输入，改善Stacking模型性能。

4 细菌觅食优化算法(BFO)

Passino等[23]借鉴人体内大肠杆菌的觅食行为机理，提出了一种新型智能优化算法——细菌觅食优化算法(BFO)。该算法通过模拟细菌群体在复杂环境中的趋化、繁殖与迁徙行为，实现全局搜索与最优解逼近。在本研究中，BFO算法用于优化钻孔效率预测模型中基学习器的超参数结构。BFO算法通过模拟细菌群体在参数空间中的动态搜索过程，提高模型对实际钻孔影响因素(如钻孔高程、钻机型号、钻杆数量等)的拟合与泛化能力，其步骤如下：

(1)初始化参数。细菌趋向性操作的次数为N_c，细菌趋向性操作中在一个方向上前进的最大步数为N_s，细菌进行繁殖操作的次数为N_{re}，粒子进行迁徙操作的次数为N_{ed}，粒子的迁徙概率为p_{ed}，细菌种群规模的大小为S(为了繁殖取一半的需要，此值取偶数)；

(2)初始化细菌位置，计算细菌的初始适应度值J，本文采用R^2作为优化算法适应度值。采用式

(8)初始细菌的位置，即待优化的初始超参数组合：

$$X = x_{\min} + rand() \times (x_{\max} - x_{\min}) \tag{8}$$

式中：X为初始化的细菌位置，代表了一个待优化的超参数组合；$x_{\min}$为搜索空间的下界(最小值)，即该超参数允许取值的最小边界；$x_{\max}$为搜索空间的上界(最大值)，即该超参数允许取值的最大边界；$rand()$为随机因子，指在[0，1]区间内均匀分布的随机数，它决定了初始位置在上下界之间的比例位置；$x_{\min}$-$x_{\max}$为搜索区间的长度，表示搜索范围的跨度。以X为超参数组合得出的预测结果所计算出的R^2值即为初始适应度值J。

(3)设$P(i, j, k, l)$为第i个细菌在第l次迁徙、第k次敏殖，第j次趋向循环时的位置(即钻孔效率预测中基学习器的一个超参数组合)。其中，迁徙操作循环：$l\in[1, N_{ed}]$；繁殖操作循环：$k\in[1, N_{re}]$；趋向操作循环；$j\in[1, N_c]$。

(4)执行细菌趋向性操作。

翻转：按式(9)更新细菌位置(超参数组合)

$$P(i, j+1, k, l) = P(i, j, k, l) + C(i) \times \frac{\boldsymbol{\Delta}(i)}{\sqrt{\boldsymbol{\Delta}^{\mathrm{T}}(i) \times \boldsymbol{\Delta}(i)}} \tag{9}$$

式中：$\boldsymbol{\Delta}(i)$为细菌翻转时在随机方向上的一个单位向量；$C(i)$为细菌在解空间中向前游动一步的步长。

游动：游动过程中，若细菌翻转后的位置适应度优于当前位置，即超参数组合更新后R^2更高，则沿该方向持续游动，直至适应度不再提升或达到最大游动步数N_s；若翻转后适应度未改善则停止游动。

(5)执行细菌繁殖操作。对细菌适应度值进行排序，淘汰掉适应度较差的一半($S/2$)细菌，剩余适应度较好的细菌进行繁殖，每个细菌分裂成两个完全相同的细菌，保持种群规模S不变。

(6)执行迁徙操作。细菌完成繁殖操作后，每个细菌以概率p_{ed}被重新随机分布到寻优空间中。在解空间中按照公式初始化；

(7)循环结束。条件判断，满足则结束，输出结果，不满足则从步骤(3)循环。

在本研究中，基学习器XGBoost和LightGBM作为基于GBDT树模型的改进，BFO算法对两个模型参数的优化思路相似：通过树数量与学习率的动态平衡防止因岩石性质等参数的突变导致的预测震荡，对单个树深度的限制则防止模型过于复杂出现的过拟合现象。同时，利用动态调整钻孔样本训练集权重分配的子采样策略使模型更好地适应多源特征变量下的不同特点的数据分布，提升模型在复杂数据下的适应性与泛化能力。区别于树模型，针对MLP的隐藏层数量与神经元分布，BFO通过层数与神经元的平衡避免梯度消失，递减约束增强特征抽象并抑制过拟合，结合自适应学习率与决定系数优化目标提升训练稳定性，动态整数步长衰减和驱散重置策略加速网络结构优化，使MLP以精简结构提升钻孔数据回归任务的R^2分数并减少资源消耗。

5 实验结果与分析

5.1 样本数据集 为验证本文搭建的集成学习预测模型，选用我国西南某大型堆石坝石料场的钻孔效率相关资料，并对收集到的数据预处理，将量化后的钻孔效率影响因素作为特征变量，综合目标变量钻孔效率，最终得到包含152行与12列的样本数据集。样本数据集的80%采用5折交叉验证的方式，每次将其中的4/5用作训练集，1/5用作验证集，而样本数据集的剩下20%用作测试集，训练后对全部训练样本进行预测得到预测数据。

本研究以单孔作业持续时间为钻孔效率，影响因素如表1所示。高程的变化会改变含氧量从而影响设备功率与人员状态。钻孔深度增大会直接增加单孔工作量并延长作业时间。不适宜施工的温度以及天气(降雨、低温)会限制操作、增加维护与停歇，从而延长工时。钻杆数量与人员配备数量影响装卸效率——更多辅助人员可缩短作业时间；加钻杆数量是为持续钻进而接长钻柱的操作次数，卸钻杆数量是为起钻作业而卸下钻柱的操作次数，二者随孔深增加而增加，因此增卸钻杆数量会增减操作时

间。岩石类型不同对施工环境有一定影响，岩性越坚硬或层理越复杂，钻削阻力与磨损增大，进尺速率下降。钻机型号不同会在故障率与功率输出上存在差异，工作时段亦会影响效率。此外，上述影响施工因素的交互效应也可能显著影响最终效率，故本研究样本以钻孔效率为目标，选取高程、钻孔深度、温度、天气、人员配备、加钻杆情况、卸钻杆情况、岩石类型、钻机型号以及包括工作时间段的其他信息等10个特征变量进行建模，以量化考虑多种因素的影响。

表1　钻孔效率影响因素

Table 1　Factors influencing drilling efficiency

序号	因素	特征描述
1	高程	工作面高度
2	深度	每个钻孔的深度
3	温度	钻进期间的大气温度
4	天气	分类变量，(0，0)(0，1)(1，0)(1，1)分别代表云、雨、雪、晴四种不同的天气类型
5	人员配备	每个钻孔分配的协助操作机器的工人数量
6	加钻杆情况	每个钻机增加钻杆的数量和时间
7	卸钻杆情况	每个钻机卸载钻杆的数量和时间
8	岩石类型	0代表某工程中板岩，1代表某工程中砂岩
9	钻机型号	钻机类型：(0，0)(0，1)(1，0)(1，1)代表不同的四种类型
10	其他信息	钻进作业的年、月、日和小时等详细信息

5.2　超参数优化　由于XGBoost、LightGBM和MLP三类模型基于各自结构特性，通过BFO算法进行协同参数调优与搜索，在精度、泛化及效率上实现针对性优化。

以XGBoost模型为例，BFO算法对模型七个参数进行动态调整，由于算力限制，本研究在算法运行过程中设置早停机制，在连续15轮迭代后适应度没有显著差异的情况下得出最终超参数调整结果。表2显示BFO算法在XGBoost模型训练过程中根据R^2进行的超参数调整，根据与表2类似的基学习器优化数据得出图2基学习器的优化进程曲线。如图2(b)所示，前期XGBoost的R^2值在0.9475～0.9495间波动，反映参数空间的探索调整，随后进入快速提升期，表明算法捕获到关键优化方向，经过对模型超参数的多次组合调整后，在早停机制的影响下适应度保持在0.95范围时没有了显著的变化，此时找到模型最佳的超参数组合。

表2　XGBoost的参数优化过程

Table 2　XGBoost parameter optimization process

优化次数	估计值	最大深度	学习率	子样本比例	特征采样比例	L1正则化系数	L2正则化系数	R^2
1	294	9	0.2764	0.814	0.372	0.924	1.246	0.9475
2	194	9	0.2057	0.756	0.300	0.731	0.966	0.9487
3	188	6	0.2578	0.752	0.300	0.671	1.186	0.9490
……				……				
31	248	3	0.2506	0.669	0.310	0.784	1.174	0.9504
32	248	3	0.2506	0.669	0.312	0.784	1.174	0.9504
33	242	3	0.2506	0.669	0.305	0.784	1.174	0.9504

5.3　基学习器结果　基学习器的钻孔效率预测结果与高程、钻杆配置等多维度影响因素直接关联，并构成元学习器的输入向量。基学习器的训练结果表明，BFO算法显著提升了三个基学习器的预测精度，使其输出结果达到了Stacking模型对基学习器的高标准要求，为构建高质量的Stacking模型奠定了基础。

图 3 以 XGBoost 为例表明基学习器对真实值数据的预测效果。通过柱状图对比展示了模型在多个样本上的预测值与真实值。柱状图统计了模型对样本的高估与低估情况，分别标注了高估样本数量及其平均偏差和低估样本数量及其平均偏差。从图中可以看出，高估或低估的样本数量均较少，其平均偏差也保持在较小范围，总体误差可控。

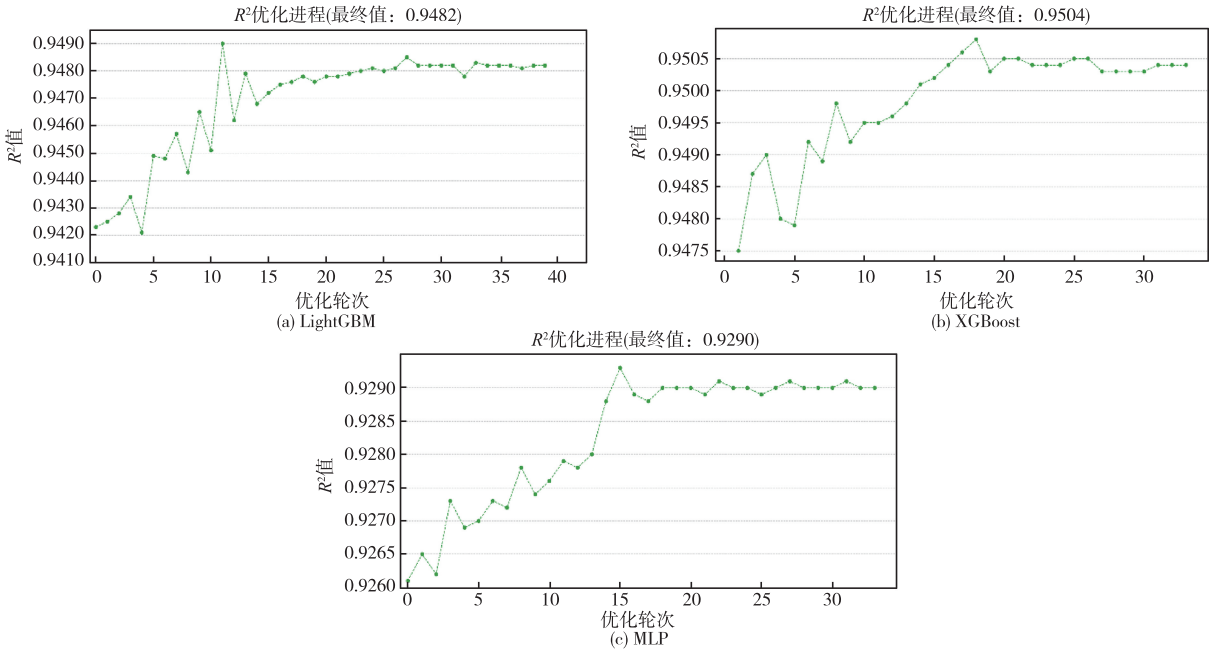


图 2　BFO 算法优化进程

Fig. 2　BFO algorithm optimization process

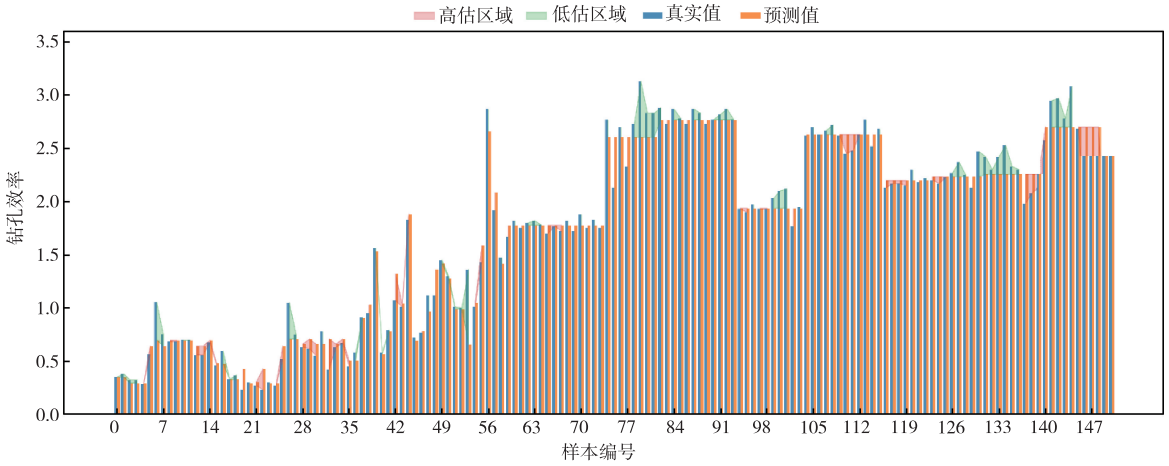


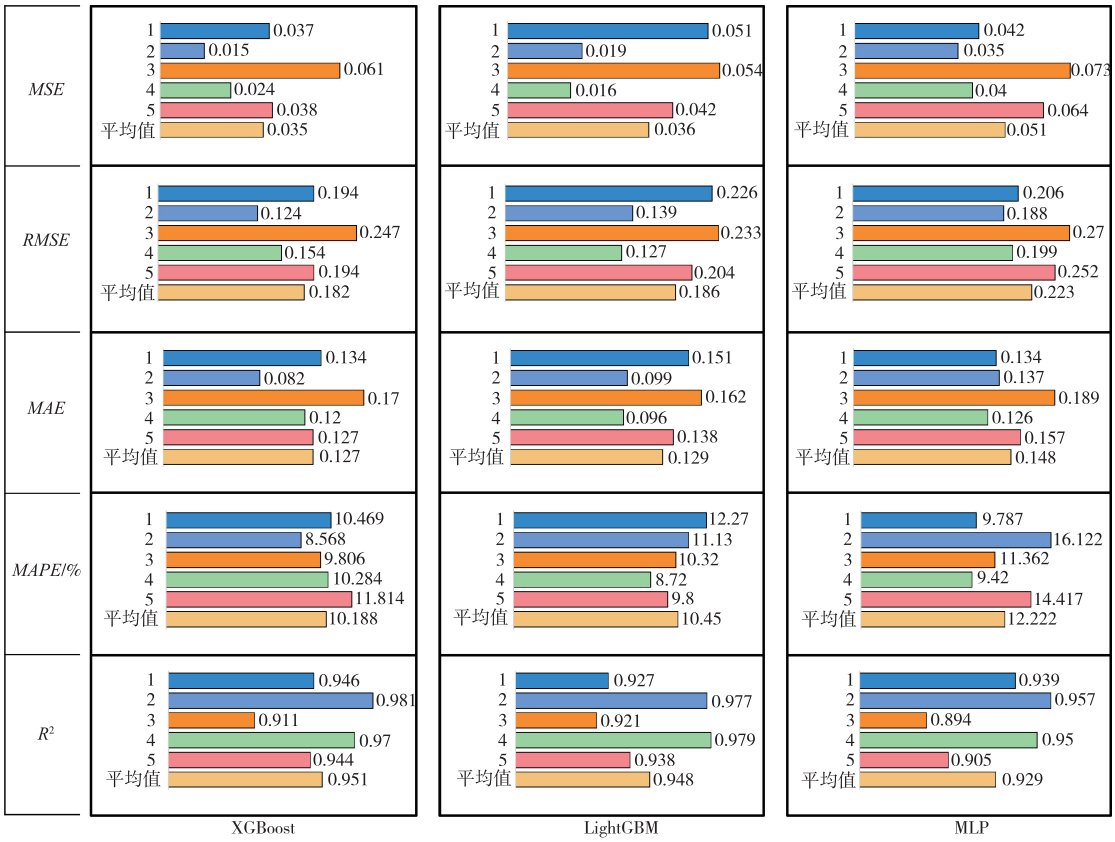
图 3　XGBoost 预测对比图

Fig. 3　XGBoost prediction comparison chart

钻孔效率与其影响因素的非线性关系以及样本数据集中影响因素多种维度的复杂性使得模型的效能体现在多个方面，所以本研究选用包括均方误差(*MSE*)、均方根误差(*RMSE*)、平均绝对误差(*MAE*)、平均绝对百分误差(*MAPE*)、决定系数(R^2)共 5 个评价指标对模型的训练情况进行评价，为确保模型泛化能力。同时，本文还引进皮尔逊相关系数(Pearson Correlation Coefficient，PCC)作为实测值与预测值之间关系的评价指标，更为直观的表现出预测值与实际值之间的差距。*PCC* 计算公式如下所示：

$$PCC = \frac{\sum_{i=1}^{n}(\hat{y}_i - E(\hat{y}_i))(y_i - E(y_i))}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - E(\hat{y}_i))^2} \times \sqrt{\sum_{i=1}^{n}(y_i - E(y_i))^2}} \tag{10}$$

式中：$\hat{y}_i$为预测值；$E(\hat{y}_i)$为预测值的平均值；y_i为真实值；$E(y_i)$为真实值的平均值。*PCC*在-1到1之间，1表示完全正相关，-1表示完全负相关，0表示没有线性关系。此指标更能够直接反应模型预测的效果，从而更加直观的找出最优参数。



注：1—5指基学习器第1—5折。

图4　基学习器五折交叉验证结果

Fig. 4　Five-fold cross-validation results of the base learners

如图4所示，XGBoost在五折交叉验证中表现稳定：*MSE*均值为0.035、*RMSE*均值为0.182，波动小；相较于未优化的*MSE*为0.170、*RMSE*为0.410，整体误差大幅下降，模型对异常值的处理能力增强了60%以上；*MAE*均值为0.127、*MAPE*均值为10.188%验证了误差控制可靠性，R^2均值为0.951表明模型可捕捉约95%的钻孔影响因素样本数据变异。

LightGBM评价所得*MSE*、*RMSE*、*MAE*、*MAPE*和R^2的均值分别为0.036、0.180、0.129、10.450%和0.948；而未优化时分别为0.067、0.260、0.200、16.240%和0.900。尽管个别折次有波动，但整体钻孔效率预测泛化能力强，误差类指标较未优化模型降低约30%以上，R^2显著提升，预测精度与拟合能力均得到实质优化。

MLP的*MSE*、*RMSE*、*MAE*、*MAPE*和R^2分别为0.051、0.223、0.148、12.220%和0.929，略逊于树模型，但R^2仍较高。部分折次接近树模型优化后的水平，说明MLP在捕捉钻进参数与钻孔效率间非线性关系上具有一定潜力，能为复杂工况下钻孔效率预测提供与树模型互补的差异化建模思路。

根据基学习器预测值与真实值的对比分析，结合五折交叉验证的评估结果(图4)，可判断所获得的模

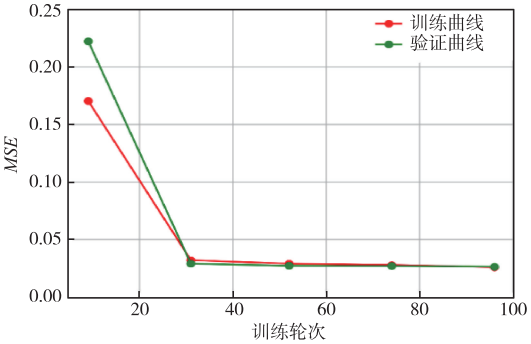


图5　元学习器学习曲线

Fig. 5　Meta-learner learning curve

型输出已满足 Stacking 策略第一层的要求。在此基础上，将三种基学习器的预测结果作为新的特征变量，构建元学习器的训练样本集。下文将围绕元学习器的训练过程及其在测试集上的预测性能进行深入探讨。

5.4　元学习器结果　SVR 通过核函数处理基学习器输出与最终预测结果之间的复杂非线性关系，使集成模型能更准确地融合多个基学习器的预测结果，从而提升整体性能。图 5 以 *MSE* 为评价指标，绘制出模型训练与验证过程中 *MSE* 随样本数据的增加而变化的情况，直观呈现了 SVR 在处理基学习器输出与最终预测结果非线性关系中的实际优化效果。

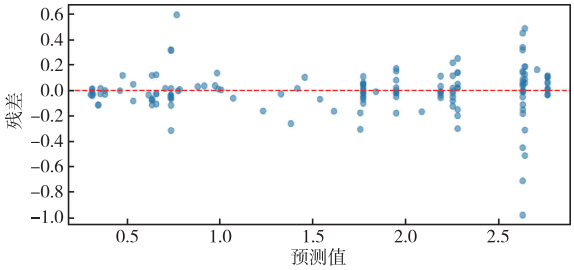


图 6　预测结果残差图

Fig. 6　Prediction results residual plot

图 5 中显示，随着训练样本量增加，训练曲线和验证曲线均呈现先快速下降后平稳收敛的趋势。当样本量超过 60 后，两曲线稳定在 0.05 ~ 0.10 区间且基本重合，说明元学习器在充分数据支持下达到稳定状态，泛化间隙极小，既未出现过拟合又具备较强噪声鲁棒性。

通过残差图比较最终结果与钻孔效率样本真实值之间的关系，如图 6 所示，预测结果整体表现良好，在预测值的主体区间内，蓝色散点围绕红色零线（$y = 0$）随机分布，正负残差对称且纵向波动范围稳定，表明模型对大部分数据的预测符合同方差性和无系统性偏差的回归假设，核心规律捕捉有效。散点未呈现 U 型、喇叭型等非线性或异方差趋势，说明 SVR 核函数选择合理，预测效果良好。

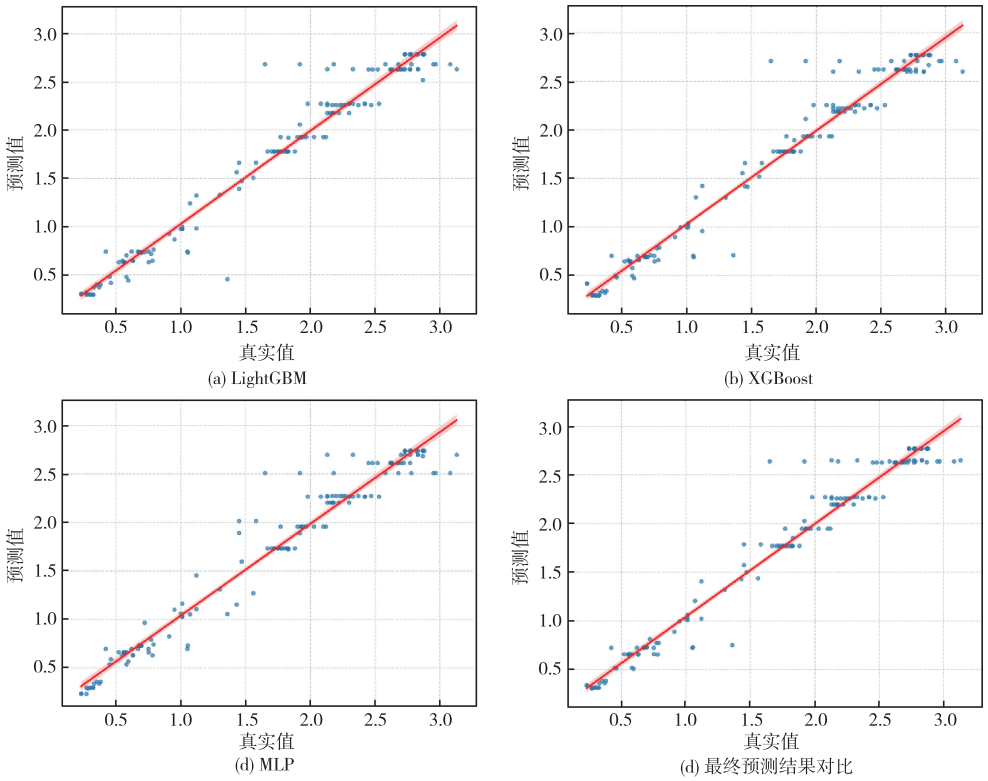


图 7　*PCC* 结果对比图

Fig. 7　*PCC* results comparison chart

通过五折交叉验证和图 7 所示 *PCC* 分析结果对比可见，基学习器在钻孔效率样本数据集上的训练效果良好，其预测值与真实值之间的 *PCC* 达到了 0.977，说明预测性能处于较高水平。采用 Stacking 集成策略后，元学习器的预测结果 *PCC* 值进一步提高至 0.979，相应的 R^2 值达到 0.959。

对比Lv等[17]提出的ICSO-Stacking集成学习模型，*PCC*所示的预测结果提高6.1%，R^2提高10.36%。

上述结果表明，优化后的集成学习模型具有较强的学习能力和高精度的预测性能，并且对复杂钻孔效率数据中的异常值表现出较强的鲁棒性。优化算法相较于手动调整参数，能够更好地避免人为选择参数时的盲目性，进一步提升预测准确度，增强模型稳定性。集成后的模型拥有强泛化能力，能够很好地预测钻孔效率，从而为施工进度仿真提供重要的参数依据。

6 结论

针对现有数学方法预测效率较低、单一学习器难以满足仿真精度要求、集成学习研究中超参数调整方法局部搜索精细度不足的问题，本文建立了基于BFO优化Stacking集成学习的钻孔效率预测模型，实现了数据中多层次特征的协同挖掘与偏差-方差的全局优化。本研究成果如下：

(1)利用Stacking集成学习方法将XGBoost、LightGBM、MLP三种异质学习器作为基学习器，对量化后的钻孔效率及其影响因素样本数据集进行预测并将结果作为输入变量训练元学习器SVR得到最终的集成学习模型。

(2)为克服基学习器复杂的调参组合问题，采用BFO对三个基学习器进行参数调整，并在钻孔样本训练过程中，利用五折交叉验证的R^2为适应度找出了在该过程中的最优参数组合。

(3)针对某大型堆石坝建设中的钻孔样本数据进行模型训练并对模型结果进行分析，对比现今相似算法模型，实验结果有了显著提升。*PCC*指数逼近0.98表明模型能够对复杂环境下的钻孔效率进行准确的预测，其结果达到仿真模型需求。

随着我国水利水电工程建设向高寒高海拔地区的逐步推进，以及智能建造理念的持续深化，水利工程施工进度仿真技术的研发已成为水利水电工程领域的重要研究课题。本研究提出的模型为水利工程开挖仿真参数的预测提供了新的模型训练思路与技术方法。

参 考 文 献：

[1] 肖浩汉，曹瑞琅，王玉杰，等. 随钻监测数据预处理方法研究[J]. 水利学报，2024，55(11)：1379-1390. (XIAO H H，CAO R L，WANG Y J，et al. Research on preprocessing methods for minitoring drilling data[J]. Journal of Hydraulic Engineering，2024，55(11)：1379-1390. (in Chinese))

[2] 魏迎奇，陈树铭，蔡红，等. 复杂地质单元的三维智能建模方法[J]. 水利学报，2014，45(S2)：19-25. (WEI Y Q，CHEN S M，CAI H，et al. Research on intelligent modeling method of complex geological unit[J]. Journal of Hydraulic Engineering，2014，45(S2)：19-25. (in Chinese))

[3] ABBASPOUR H，DREBENSTEDT C，BADRODDIN M，et al. Optimized design of drilling and blasting operations in open pit mines under technical and economic uncertainties by system dynamic modelling[J]. International Journal of Mining Science and Technology，2018，28(6)：839-848.

[4] BANI M A，ABBAS A K，ALSABA M，et al. Improving drilling performance through optimizing controllable drilling parameters[J]. Journal of Petroleum Exploration and Production，2021，11(3)：1223-1232.

[5] DARBOR M，FARAMARZI L，SHARIFZADEH M. Performance assessment of rotary drilling using non-linear multiple regression analysis and multilayer perceptron neural network[J]. Bulletin of Engineering Geology and the Environment，2019，78(3)：1501-1513.

[6] KOOPIALIPOOR M，TOOTOONCHI H，JAHED A D，et al. Application of deep neural networks in predicting the penetration rate of tunnel boring machines[J]. Bulletin of Engineering Geology and the Environment，2019，78(8)：6347-6360.

[7] SHI X，LIU G，GONG X L，et al. An efficient approach for real-time prediction of rate of penetration in offshore drilling[J]. Mathematical Problems in Engineering，2016，2016(1)：3575380.

[8] ZHOU Y，CHEN X，ZHAO H B，et al. A novel rate of penetration prediction model with identified condition for

the complex geological drilling process[J]. Journal of Process Control, 2021, 100: 30-40.
[9] ABBAS A K, RUSHDI S, ALSABA M, et al. Drilling rate of penetration prediction of high-angled wells using artificial neural networks[J]. Journal of Energy Resources Technology, 2019, 141(11): 112904.
[10] JIA C D, ZHANG J Y, KONG X D, et al. Prediction of drilling efficiency for rotary drilling rig based on an improved back propagation neural network algorithm[J]. Machines, 2024, 12(7): 438.
[11] 王晓玲, 胡亦宁, 张君, 等. 土石方开挖钻孔效率预测可解释超级学习器集成学习模型[J]. 天津大学学报: 自然科学与工程技术版, 2025, 58(4): 331-342. (WANG X L, HU Y N, ZHANG J, et al. Interpretable super learner ensemble learning model for earth-rock excavation drilling efficiency prediction[J]. Journal of Tianjin University:Science and Technology, 2025, 58(4): 331-342. (in Chinese))
[12] 张君, 王金国, 余佳, 等. 多模态感知驱动下高堆石坝施工仿真参数集成深度学习模型[J]. 水利学报, 2022, 53(9): 1049-1063, 1072. (ZHANG J, WANG J G, YU J, et al. Multimodal perception-driven high rockfill dam construction simulation input modeling using an ensemble deep learning model[J]. Journal of Hydraulic Engineering, 2022, 53(9): 1049-1063, 1072. (in Chinese))
[13] 韦林, 段凯, 刘效东, 等. 基于多源数据与多模型集成的流域人为蒸散发变异评估[J]. 水利学报, 2022, 53(4): 433-444. (WEI L, DUAN K, LIU X D, et al. Assessing human-induced evapotranspiration change based on multi-source data and Bayesian model averaging at the basin scale[J]. Journal of Hydraulic Engineering, 2022, 53(4): 433-444. (in Chinese))
[14] 余红玲, 王晓玲, 任炳昱, 等. 土石坝渗流性态分析的IAO-XGBoost集成学习模型与预测结果解释[J]. 水利学报, 2023, 54(10): 1195-1209. (YU H L, WANG X L, REN B Y, et al. IAO-XGBoost ensemble learning model for seepage behavior analysis of earth-rock dam and interpretation of prediction results[J]. Journal of Hydraulic Engineering, 2023, 54(10): 1195-1209. (in Chinese))
[15] MENDES-MOREIRA J, SOARES C, JORGE A M, et al. Ensemble approaches for regression: A survey[J]. ACM Computing Surveys(csur), 2012, 45(1): 1-40.
[16] UMAR I H, LIN H, LIU H W, et al. HELIOS-stack: A novel hybrid ensemble learning approach for precise joint roughness coefficient prediction in rock discontinuity analysis[J]. Materials, 2025, 18(8): 1807.
[17] LV F, YU J, ZHANG J, et al. A novel stacking-based ensemble learning model for drilling efficiency prediction in earth-rock excavation[J]. Journal of Zhejiang University: Science A, 2022, 23(12): 1027-1046.
[18] TIAN S M, NI X Y, WANG Y. Seismic shear strength prediction of reinforced concrete shear walls by stacking multiple machine learning models[J]. Applied Sciences, 2025, 15(5): 2268.
[19] 邓怡徽, 邹艳红, 李延申. 混合ABC-Stacking机器学习的钻孔数据地层三维隐式建模方法[J]. 成都理工大学学报: 自然科学版, 2025, 52(5): 1020-1034. (DENG Y H, ZOU Y H, LI Y S. A three-dimensional implicit formation modeling method for drilling data with a mixed artificial bee colony stacking machine learning strategy[J]. Journal of Chengdu University of Technology: Science & Technology Edition, 2025, 52(5): 1020-1034. (in Chinese))
[20] 王佳俊, 钟登华, 关涛, 等. 基于KM和AC-BFA模糊逻辑的堆石坝压实质量实时评价[J]. 水力发电学报, 2019, 38(3): 165-178. (WANG J J, ZHONG D H, GUAN T, et al. Real-time evaluation of compaction quality of earth-rock dams using fuzzy logic based on KM and AC-BFA[J]. Journal of Hydroelectric Engineering, 2019, 38(3): 165-178. (in Chinese))
[21] ALRASHED M M, ELNAGGAR M F, FLAH A, et al. Online intelligent parameter and speed estimation of permanent magnet synchronous motors using bacterial foraging optimization[J]. Science and Technology for Energy Transition, 2025, 80: 33.
[22] ANAND S, DHANKHAR G, SATYARTHI M K. Optimizing mechanical properties of PETG fabricated parts via PSO-BFO hybrid algorithm in fused deposition modelling[J]. International Journal on Interactive Design and Manufacturing, 2025, 19(8): 5875-5891.
[23] PASSINO K M. Biomimicry of bacterial foraging for distributed optimization and control[J]. IEEE Control Systems Magazine, 2002, 22(3): 52-67.

Research on a borehole drilling efficiency prediction model based on bacterial foraging optimization stacking ensemble learning

GUAN Tao[1], JIAO Chunqi[1], YU Peng[1], YU Jia[1], GUO Zhenbang[1], CHEN Zhengfei[2]
(1. *State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, Tianjin* 300072, *China*;
2. *China Renewable Energy Engineering Institute, Beijing* 100120, *China*)

Abstract: Borehole drilling efficiency is a key parameter in construction progress simulation of rock-fill dam quarry excavation, and its prediction accuracy directly affects the reliability of the simulation model. To address the issue of low prediction accuracy efficiency in existing mathematical methods, the difficulty of single learners meeting simulation accuracy requirements, and insufficient local search refinement in hyperparameter tuning within ensemble learning research, this paper proposes a borehole drilling efficiency prediction model based on Bacterial Foraging Optimization for Stacking ensemble learning. First, a dataset was constructed using on-site drilling efficiency data from a rock-fill dam as the target variable and its influencing factors (e.g., borehole depth, rock properties, altitude) as feature variables. Second, three heterogeneous base learners (XGBoost, LightGBM, and MLP) were trained in parallel, and the Bacterial Foraging Optimization algorithm—simulating chemotaxis and reproduction—was introduced to iteratively optimize each base learner's hyperparameters by tracking the R^2 curve in real time, ensuring stable "meta-features" output. Finally, the base learners' predictions were input to a Support Vector Regression (SVR) meta-learner; by integrating complementary information from multiple models, the ensemble prediction was obtained while suppressing bias and variance. Experimental results show that after Bacterial Foraging Optimization, each base learner's R^2 can exceed 0.93 and *PCC* are all above 0.97; the ensemble model's learning curve over the full sample set is smooth and stable, residual analysis indicates residuals are evenly distributed around the zero-mean line, and the final *PCC* approaches 0.98, meeting the requirements of construction process simulation.
Keywords: borehole drilling efficiency; construction simulation; Stacking ensemble learning; XGBoost; LightGBM; MLP; support vector machine

(责任编辑：王　婧)

(上接第 231 页)

A comparative study on numerical simulation and prototype observation of thermal effluents from a nuclear power plant

CHEN Xiaoli, ZHANG Dongyang, YUAN Jue, DUAN Yafei
(*China Institute of Water Resources and Hydropower Research, Beijing* 100038, *China*)

Abstract: Prototype observation of thermal discharge from operating power plants is the primary means adopted in China for verifying the accuracy of thermal-discharge numerical models. However, the temperature fields obtained from such prototype observations incorporate complex background water temperatures, while the models usually assume a uniform background temperature. Since the impact of baseline values on verification accuracy is unclear, there is an urgent need to carry out relevant research. This paper takes a certain nuclear power plant site as an example. The water temperature variation processes in the project sea area through a real-time surface heat flux simulation approach based on spatial and temporal varied metrological and hydrological conditions. The numerical model is then validated using diverse in-situ observation data sources, including satellite remote sensing, airborne remote sensing, and surface water measurements. The calculated temperature results are in good agreement with the measurements. Then the impact of the background water temperature value on the extraction of the temperature rise field is analyzed, as well as the sensitivity of numerical simulation results to the values of the diffusion coefficient, heat dissipation coefficient, and wind field. The results show the background temperature distribution has obvious effect on the observed temperature rise determination. Using a uniform background value will result in an opposite seasonal trend in the temperature rise field—overestimation in summer and underestimation in winter. The marine wind field has the greatest impact on the pattern and covered area of the temperature rise field. The diffusion coefficient and heat dissipation coefficient show different in their sensitivity to the calculated temperature rise result, as the diffusion coefficient has a greater impact on the 4 ℃ temperature rise zone, while the heat dissipation coefficient has a greater impact on the 1 ℃ temperature rise zone. This study provides an important reference for enhancing the accuracy of thermal discharge modeling and promoting thermal discharge impact assessment.
Keywords: thermal discharge; background water temperature; prototype observation; numerical simulation; wind field

(责任编辑：鲁　婧)