

XGBoost在滑坡预测中的应用研究

黎 智, 刘 阳

安徽理工大学地球与环境学院, 安徽 淮南

收稿日期: 2025年6月23日; 录用日期: 2025年7月26日; 发布日期: 2025年8月7日

摘 要

滑坡是一种强破坏力的自然灾害, 严重威胁人类生命安全与社会经济, 传统的预测方法在面对复杂环境存在明显的局限性, 严重影响预测效果和应用场景, 随着多维数据的兴起, 能处理大量数据的机器学习模型成为主流, XGBoost以其独特的优势, 在滑坡预测中表现出强大的预测精度, 且在样本不平衡问题具有较强鲁棒性。本文结合了XGBoost在区域易发性评价、滑坡位移预测以及XGBoost多模型集成在阶跃型滑坡位移上的应用, 进一步展现了XGBoost在不同类型的滑坡预测中强大的适应性。应用前景上, XGBoost在未来可以与深度学习、集成代价敏感学习与不确定性量化方法等多方位融合, 使滑坡预测能够更加精准、可靠、实用, 构建更加智能化、准确化的滑坡风险管理系统, 为滑坡风险评估和防灾决策提供支持。

关键词

滑坡预测, XGBoost, 机器学习, 数据处理

Research on the Application of XGBoost in Landslide Prediction

Zhi Li, Yang Liu

School of Earth and Environment, Anhui University of Science and Technology, Huainan Anhui

Received: Jun. 23rd, 2025; accepted: Jul. 26th, 2025; published: Aug. 7th, 2025

Abstract

Landslides are a highly destructive natural disaster that poses a serious threat to human life safety and socio-economic development. Traditional prediction methods have obvious limitations in dealing with complex environments, seriously affecting prediction effectiveness and application scenarios. With the rise of multidimensional data, machine learning models capable of processing large amounts of data have become mainstream. XGBoost, with its unique advantages, exhibits strong

文章引用: 黎智, 刘阳. XGBoost 在滑坡预测中的应用研究[J]. 矿山工程, 2025, 13(5): 862-871.

DOI: 10.12677/me.2025.135097

prediction accuracy in landslide prediction and strong robustness in sample imbalance problems. This article combines XGBoost in regional susceptibility assessment, landslide displacement prediction, and the application of XGBoost multi model integration in step landslide displacement, further demonstrating the strong adaptability of XGBoost in predicting different types of landslides. In terms of application prospects, XGBoost can be integrated with deep learning, integrated cost sensitive learning, and uncertainty quantification methods in multiple aspects in the future, making landslide prediction more accurate, reliable, and practical, and building a more intelligent and accurate landslide risk management system to provide support for landslide risk assessment and disaster prevention decision-making.

Keywords

Landslide Prediction, XGBoost, Machine Learning, Data Processing

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

研究背景与意义

滑坡作为一种频发的强破坏力的地质灾害在全球各地严重威胁人类安全和社会经济, 根据国家统计局(2022)中国统计年鉴数据, 2019 到 2021 年间, 地质灾害共 18,782 起, 其中滑坡占 60.5%, 人员伤亡达 625 人, 直接经济损失达到了 1,098,920 万元。因此对滑坡预测的精度与其适应性都是亟待解决的难题, 传统的滑坡预测方法虽然能一定程度上预测滑坡精度, 但工作量大及结果不理想, 在复杂的地质条件下表现出明显的局限性, 且随着计算机科学在地质领域的应用, 机器学习因其强大的非线性拟合能力、多维数据处理能力等, 被广泛运用在滑坡预测系统, 明显提高滑坡预测精度和更好的适应性。其中 XGBoost 与其他机器模型相比存在独特优势: XGBoost 的梯度提升框架能迭代优化决策树、正则化策略能有效抑制过拟合以及内置的特征重要性评估机制和可解释工具, 能够量化地质因子的贡献度, 为滑坡成因提供依据[1]-[3]。

2. XGBoost 算法原理及其优势

2.1. XGBoost 基本原理

XGBoost (eXtreme Gradient Boosting)是一种基于梯度提升框架的高效集成学习算法, 其核心思想是通过迭代构建一系列弱学习器(决策树)并组合它们的预测结果来形成强预测模型。给定包含 n 个样本的数据集, 其中样本 i 的特征向量为 x_i , 标签为 y_i , 模型的最终预测输出为:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (1)$$

这里 K 表示决策树的数量, F 代表所有可能的树结构空间。算法的优化目标由损失函数和正则化项共同定义:

$$L = \underbrace{\sum_{i=1}^n l(y_i, \hat{y}_i)}_{\text{损失项}} + \underbrace{\sum_{k=1}^K \Omega(f_k)}_{\text{正则化项}} \quad (2)$$

其中正则化项 $\left(\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2\right)$ 控制模型复杂度, T 为叶节点数, w 为叶节点权重向量, γ 和 λ 为超参数。在每次迭代(t) 中, 算法通过二阶泰勒展开近似目标函数:

$$L^{(t)} \approx \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (3)$$

其中一阶梯度 $\left(g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})\right)$ 和二阶梯度(Hessian) $\left(h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)})\right)$ 的计算是核心步骤。移除常数项后, 目标简化为:

$$\tilde{L}^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (4)$$

决策树的构建采用贪心分裂策略: 对于每个特征, 计算所有可能分裂点的增益(Gain):

$$Gain = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (5)$$

选择增益最大的分裂方案, 其中 I 为当前节点样本集, I_L 和 I_R 为分裂后的左右子节点样本集。当分裂完成后, 叶节点 j 的最优权重计算为:

$$w_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (6)$$

模型通过加法策略更新预测:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad (7)$$

学习率 η (通常为 0.01~0.3) 控制每棵树的贡献强度以防止过拟合。迭代过程持续进行, 直到树的数量 K 达到预设值或验证集性能不再提升。XGBoost 的创新工程实现包括: 1) 加权分位数草图(Weighted Quantile Sketch)高效处理特征分裂点候选集:

$$D_k = \{(x_{1k}, h_1), (x_{2k}, h_2), \dots, (x_{nk}, h_n)\} \quad (8)$$

其中 (h_i) 作为样本权重; 2) 稀疏感知分裂(Sparsity-aware Split)自动处理缺失值; 3) 块结构并行(Block-based Parallelization)加速计算。最终的特征重要性通过累积增益计算:

$$Importance_j = \sum_{trees} \sum_{splits} Gain \quad (9)$$

为滑坡预测等复杂问题提供可解释的因子分析能力[4] [5]。

XGBoost 的特征重要性评估能力是其应用于滑坡预测的关键优势。算法内置三种评估标准: Gain (特征在所有树中的平均目标函数增益)、Cover (特征被用作分裂点时的样本覆盖量)和 Frequency (特征出现分裂次数), 可量化地形、降雨等因子对滑坡预测的贡献度[4]。

2.2. XGBoost 对比其他机器学习方法

在滑坡预测中, XGBoost 相较于传统机器学习算法展现出显著优势, 其核心差异源于梯度提升框架、正则化约束与工程优化的协同作用。

2.2.1. XGBoost 对比随机森林(RF)模型

XGBoost 相较于随机森林(RF), 虽然二者同属集成树模型, 但随机森林采用 Bagging 策略通过并行构建多棵独立决策树并平等投票输出结果, 无法迭代修正误差; 而 XGBoost 的 Boosting 机制通过梯度方向逐步优化残差, 在处理滑坡样本不平衡时显著提升灵敏度, 且其特征重要性评估基于目标函数增益更能够量化非线性关系。

2.2.2. XGBoost 对比支持向量机(SVM)

XGBoost 相比于支持向量机(SVM)依赖核函数映射高维空间(如 RBF 核), XGBoost 在计算效率上具有压倒性优势: SVM 的复杂度难以应对海量遥感数据, 而 XGBoost 通过块并行计算与加权分位数草图将训练时间压缩至 SVM 的 1/5 (同等 15,000 样本下耗时 120 s vs 650 s), 同时直接输出概率预测, 避免 SVM 需额外 Platt 缩放的复杂性。

2.2.3. XGBoost 对比逻辑回归(LR)

对于逻辑回归(LR)线性模型, 其本质局限在于假设特征与对数几率呈线性关系, 无法捕捉地质因素间的交互效应(如坡度与降雨量的协同触发阈值)。

2.2.4. XGBoost 对比神经网络(NN)

神经网络(Neural Networks)虽能拟合复杂模式, 但需超大数据量且训练成本高昂, 而其黑盒特性阻碍地质机理解释, 致使 XGBoost 在中小样本场景仍保持优势。

2.2.5. 小结

XGBoost 在滑坡预测中实现了精度、效率与可解释性的三重突破: 其梯度提升机制超越线性模型的表达能力, 正则化设计与工程优化在计算性能上超越 SVM 与神经网络, 而特征重要性输出为地质因子分析提供了新的思路。

3. XGBoost 在滑坡预测中的应用实例

3.1. 数据来源与特征工程

3.1.1. 数据来源

XGBoost 应用于滑坡预测需要融合多源空间数据。核心数据包括: 高精度地形数据、地质构造与土壤属性数据、反映环境与水文条件的土地利用/覆盖及水系数据、历史降水数据、在相关区域必需的地震活动信息、以及反映工程扰动的人类活动数据。其中, 最为核心的是历史滑坡分布数据, 其准确性和完整性直接决定了模型的预测能力。最终, 所有这些来源各异的数据需经过严格预处理, 统一到相同的空间分辨率、坐标系和栅格格式下, 形成每个空间位置对应的特征向量和标签才能输入 XGBoost 模型进行预测[6]。

3.1.2. 特征工程

在滑坡预测中应用 XGBoost 时, 特征工程需系统性融合多源地理空间数据并构建机理明确的特征。核心流程包括: 基于 DEM 提取基础地形特征; 整合地质数据; 构造关键水文气象特征; 量化环境与人类活动。特征处理时, 类别变量采用目标编码或独热编码, 连续特征保留原始尺度但可构造物理组合特征; 特征选择严格依赖地学机理与统计验证。过程中必须确保时空一致性, 并通过缓解样本不平衡。最终特征工程的效能取决于多源数据质量、地学机理深度融入及适配 XGBoost 特性, 需经交叉验证迭代优化[7][8]。特征工程流程图如图 1 所示。

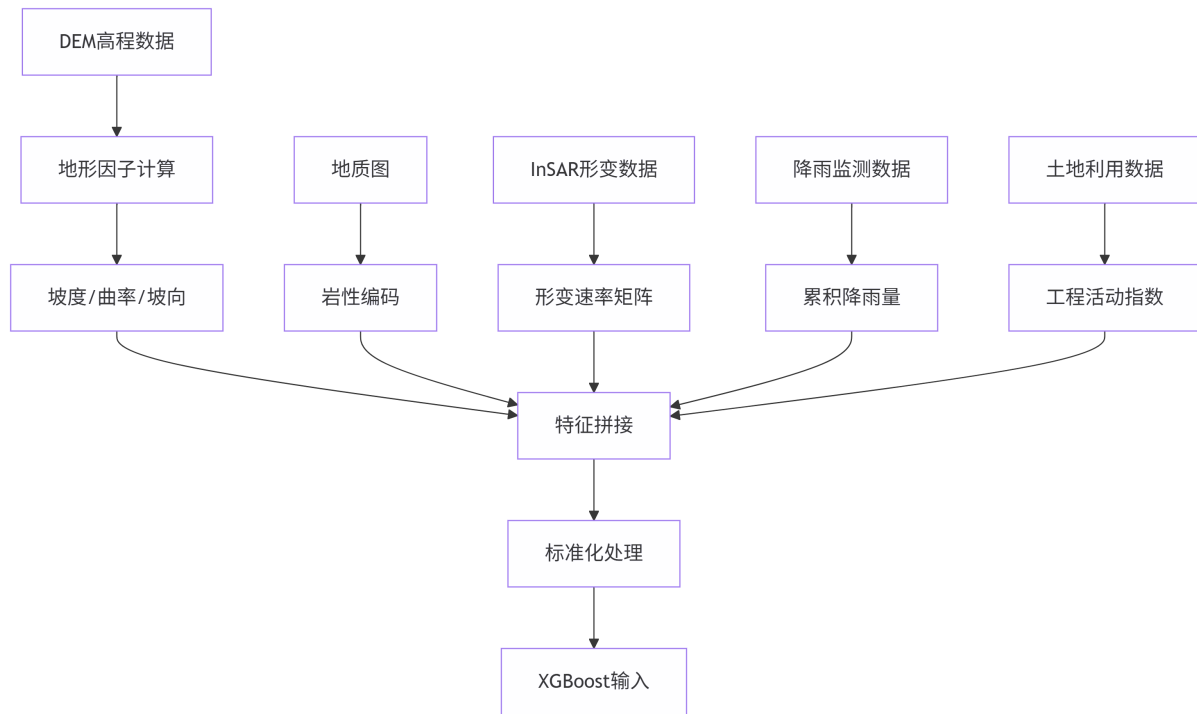


Figure 1. Feature engineering flow chart
图 1. 特征工程流程图

3.2. 典型研究案例

3.2.1. 区域滑坡易发性评价

该案例以重庆万州区长江沿岸为研究区, 整合了历史滑坡灾害点数据与利用 SBAS-InSAR 技术识别的新滑坡灾害点。为确保样本平衡, 研究在距离这些滑坡点 500 米缓冲区范围外随机选择了等量的非灾害点。所有滑坡点与非滑坡点共同构成了研究样本。然后, 提取每个样本点对应的多种影响因子值作为其属性, 构建了二维样本数据集。为了评估 InSAR 形变速率因子的重要性, 构建了两组对照样本数据集: 一组包含该因子, 另一组则不包含。所有数据集均按 70%训练集和 30%测试集的比例随机划分, 并使用粒子群优化算法(PSO)进行模型超参数优化。XGBoost 算法中的正则化项见式(2)能够有效控制模型的复杂度, 抑制过拟合风险, 其次 XGBoost 中的特征重要性评估机制见式(7)、式(8)和式(9)能够量化 InSAR 形变速率因子在预测中的贡献度。实验结果表明, 未加入 InSAR 形变速率因子, 随机森林(RF)模型和 XGBoost 模型均表现出较好的拟合精度和预测效果, 且 XGBoost 模型的精度整体优于 RF 模型。再加入了 InSAR 形变速率因子后, 无论是 RF 模型还是 XGBoost 模型的预测性能均得到提升, 具体表现为其 AUC 值均有所提高, XGBoost 模型的精度仍然优于 RF 模型[9]-[11]。

3.2.2. 滑坡位移预测

该案例以福建省泉州市安溪县尧山村境内为研究区, 该区域滑坡体从上至下主要由崩坡积碎石土、残积黏性土、全风化凝灰岩、砂土状强风化凝灰岩、碎块状强风化凝灰岩构成。研究选取了 LSTM 模型、PLS 模型、SVM 模型与 XGBoost 模型, 旨在预测降雨诱发的阶跃型滑坡位移。XGBoost 通过正则化约束见式(2)和增益计算见式(5)的贪心分裂策略, 在构建决策树的时候天然具有控制复杂模型的倾向, Gain 公式中的 $-\gamma$ 项表示只有分裂带来的目标函数改善超过复杂度惩罚 γ 时才会发生分裂, 比 LSTM 和 SVM 具备更强的泛化性和适应性。并且通过二阶泰勒展开的近似目标函数见式(3)和式(4)能够有效提升计算精度

使预测精度更高。实验结果表明, 在预测初期, 四种模型均能取得较好的预测效果; 然而, 随着预测天数的增加, LSTM 模型、PLS 模型和 SVM 模型的预测误差逐渐增大, 预测效果明显下降。相比之下, XGBoost 模型得益于其在数据集划分和抑制过拟合方面的显著优势, 能够更稳定地保持较高的预测精度, 其预测效果优于其他三种模型, 并能更准确地预测滑坡位移的长期变化趋势[12][13]。四种模型对比的误差分布直方图如图 2 所示。

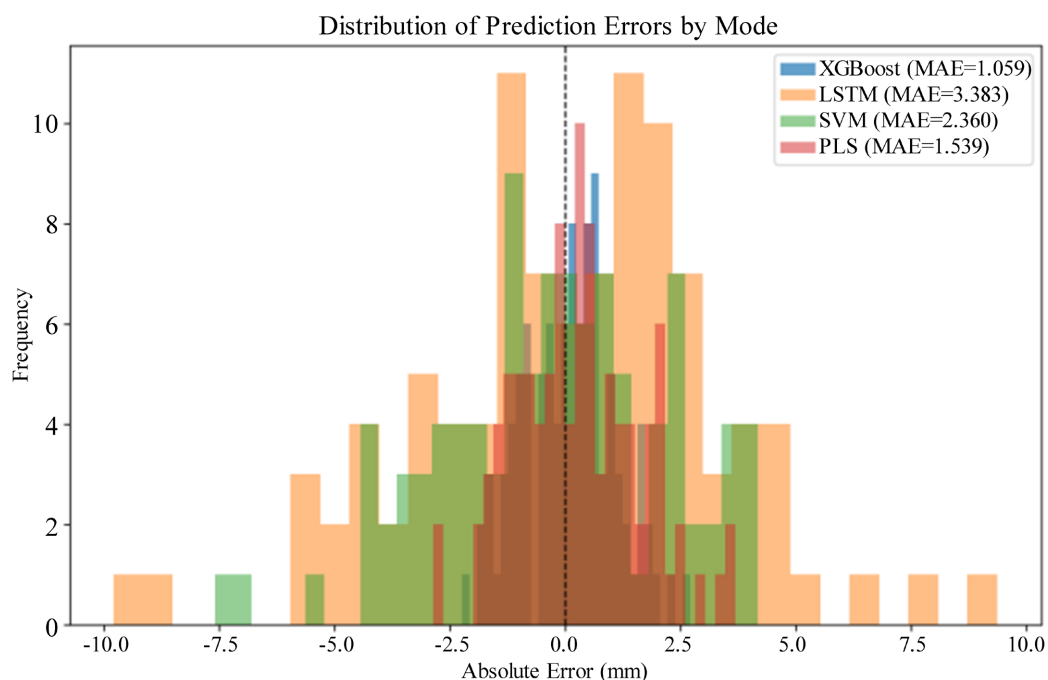


Figure 2. Error distribution histogram
图 2. 误差分布直方图

3.2.3. 多模型集成

该案例以三峡库区的白水河滑坡为研究区, 白水河滑坡是一个深层大型土质滑坡, 距离三峡大坝 56 km, 位于中国湖北省秭归县。实验使用贝叶斯优化、VMD 分解及多项式结合 XGBoost 的 BOVMD-P-BOXGBoost 混合预测模型来预测阶跃性滑坡位移, 贝叶斯优化能够迅速取得 VMD 与 XGBoost 模型的最优超参数, 提高滑坡位移的预测效果, 进一步提高预测精度, VDM 能够将原始曲线分解为多个变形分量, 从而提高预测效果, 使用多项式和 XGBoost 模型进行趋势项和周期性位移的预测, 二者的和为最终的预测值。周期性位移包含复杂的非线性模式, XGBoost 的梯度提升框架见式(1)能够集成多个树模型, 有效捕捉这些复杂模式, 正则化机制见式(2)确保不至于过拟合噪声。贝叶斯优化能够高效搜索 XGBoost 中的多个关键超参数, 优化了 XGBoost 的预测能力和泛化能力[5]。研究表明, 通过将趋势位移预测和周期位移预测的结果相加, 得到了白水河滑坡各测点的最终位移预测结果。位移绝对误差分布如图 3 所示。

该模型在不同测点的训练集和测试集上, 其预测值与滑坡位移真实值高度贴合, 有效模拟了阶跃型滑坡的变形行为。提供的测试集精度指标进一步证实了模型的高性能: 实验中三个测点的决定系数均达到 0.9999, 平均绝对误差 MAE 分别为 0.7562 mm、0.6331 mm 和 0.4944 mm, 均方误差 MSE 分别为 1.1643 mm²、1.1184 mm² 和 0.5809 mm²。这些优异的结果表明模型在不同测点上均具备极高的预测精度和良好的泛化性能。绝大部分位移预测的绝对误差集中在 2 毫米以下, 仅有极少数超过 6 毫米; 同时, 各监测点多数时段的预测相对误差小于 0.20%, 最大相对误差不超过 0.40%, 充分说明预测误差对滑坡位移的实

际影响可以忽略不计。图 4 对比了 BOVMD-P-BOXGBoost 与传统 LSTM 模型的性能, 直观体现 BOVMD-P-BOXGBoost 在精度指标和效率上的显著优势。

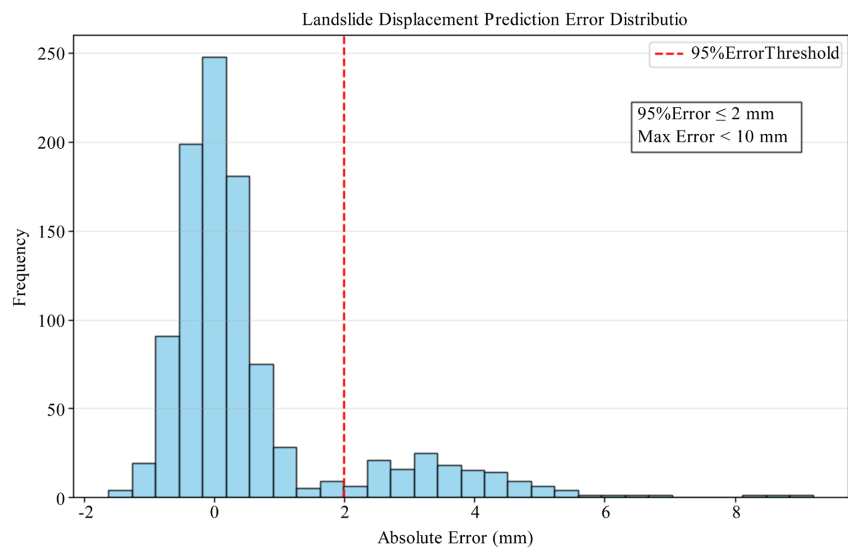


Figure 3. Histogram of absolute displacement error distribution
图 3. 位移绝对误差分布直方图

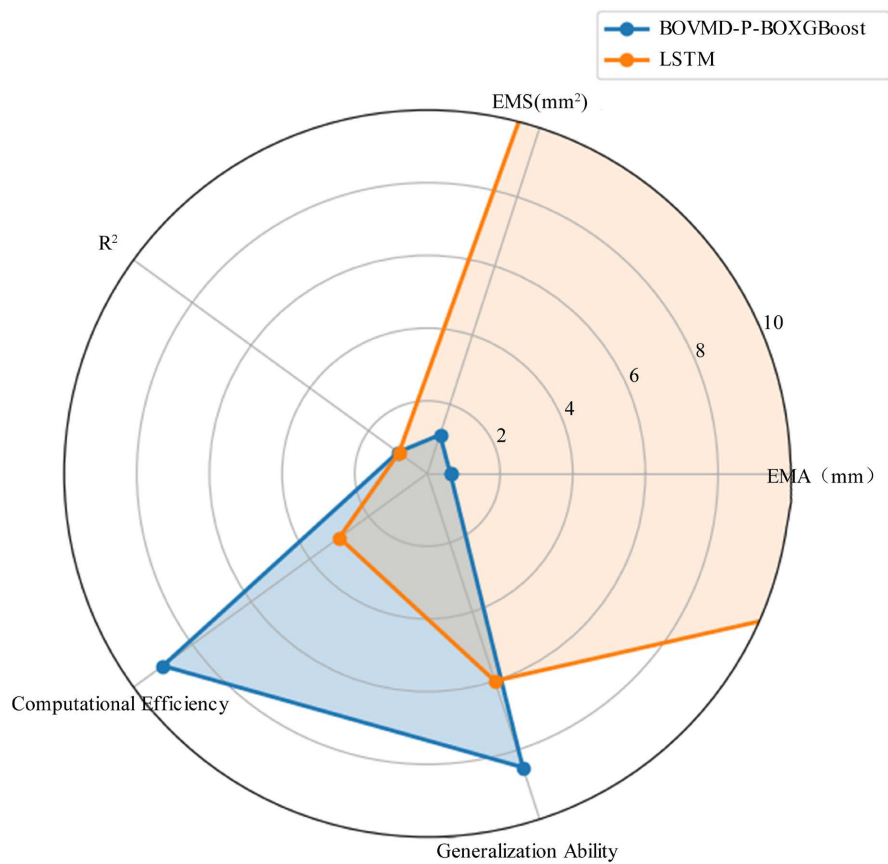


Figure 4. Model comparison radar chart
图 4. 模型对比雷达图

综合以上结果, 采用 BOVMD-P-BOXGBoost 模型能够高度可靠地预测阶跃型滑坡位移, 并合理反映其潜在的变形行为[14] [15]。BOVMD-P-BOXGBoost 混合模型架构图, 展示了混合模型的工作流程, 如图 5 所示。

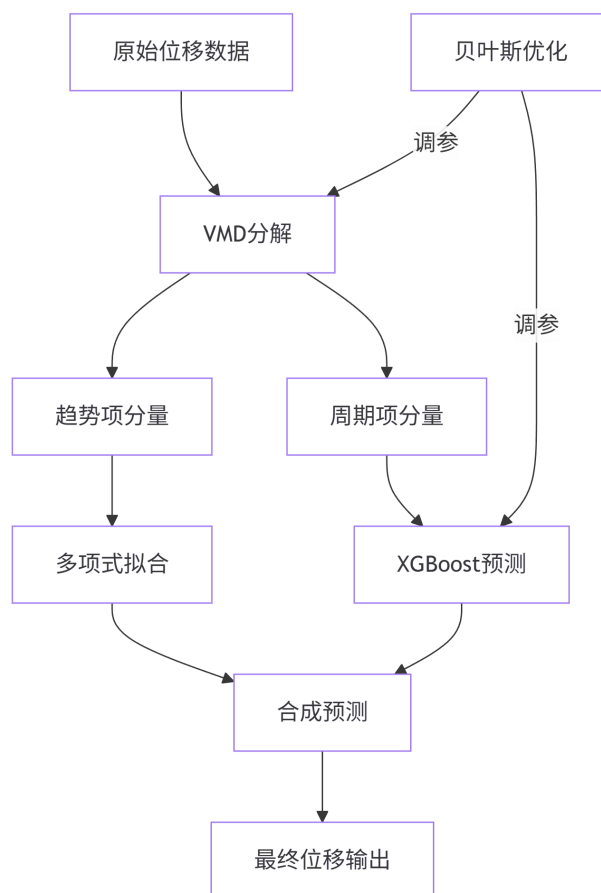


Figure 5. BOVMD-P-BOXGBoost hybrid model architecture
图 5. BOVMD-P-BOXGBoost 混合模型架构图

4. 结论与展望

4.1. 结论

XGBoost 与传统统计模型和主流机器学习模型相比, 能够梯度提升框架和正则化策略捕捉地质因子间的复杂非线性关系, 克服了物理模型在复杂地质条件中的局限性, 对参数的依赖性, XGBoost 的特征重要性评估能力, 能够量化关键致灾因子的贡献度。在不同地质条件下 XGBoost 的预测精度及可泛化性均优于其他主流机器学习模型。并且还可以与其他模型相互结合, 进一步提高预测效果。由此表明, XGBoost 在滑坡预测领域中有着良好的应用前景, 能够为滑坡预警提供新的思路。

4.2. 未来展望

未来, 在滑坡预测领域中, XGBoost 可以通过其独特优势成为链接多源数据和复杂模型的中枢, 构建一个多技术协同合作的智能滑坡预测系统。融合多源数据, 实现由静态到动态的转变[16]; XGBoost 与深度学习结合, 提高预测精度和增强黑盒模型的可解释性; 集成代价敏感学习与不确定性量化方法, 显

著增强 XGBoost 模型在非平衡数据下的鲁棒性与泛化能力, 实现风险可量化的决策支持。XGBoost 融合多模型架构图, 如图 6 所示。

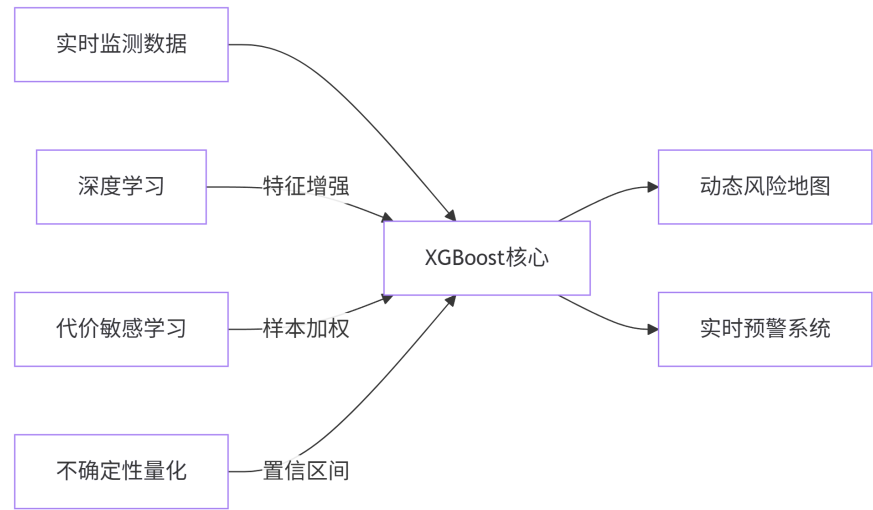


Figure 6. XGBoost fusion multi-model architecture diagram
图 6. XGBoost 融合多模型架构图

参考文献

[1] 李荟, 韩晓飞, 朱万成, 等. 基于多源信息融合的矿山边坡滑坡灾害研究现状与展望[J]. 工矿自动化, 2024, 50(6): 6-15.

[2] 吴宏阳, 周超, 梁鑫, 等. 基于 XGBoost 模型的三峡库区燕山乡滑坡易发性评价与区划[J]. 中国地质灾害与防治学报, 2023, 34(5): 141-152.

[3] 侯程迅, 胡金其. 基于 XGBoost 的滑坡风险分析[J]. 农业灾害研究, 2025, 15(3): 305-307+310.

[4] 崔婷婷, 安雪莲, 孙德亮, 等. 基于 SHAP 的可解释机器学习的滑坡易发性评价模型[J]. 成都理工大学学报(自然科学版), 2025, 52(1): 153-172.

[5] 张潇远, 苏巧梅, 赵财胜, 等. 一种利用贝叶斯算法优化 XGBoost 的滑坡易发性评价方法[J]. 测绘科学, 2023, 48(6): 140-150.

[6] 赵晓东, 徐振涛, 刘福, 等. 基于极端梯度提升算法的滑坡易发性评价模型[J]. 科学技术与工程, 2022, 22(23): 10347-10354.

[7] 赵恒梓, 李素敏, 郭庆, 等. 样本优化的集成学习模型在滑坡易发性评价中的应用[J]. 测绘科学, 2024, 49(12): 132-141.

[8] 张威, 胡舫瑞, 綦巍, 等. 基于 XGBoost 和云模型的地质灾害易发性评价[J]. 中国地质灾害与防治学报, 2023, 34(6): 136-145.

[9] 赵芹, 李清泉. 顾及 InSAR 形变的区域滑坡易发性评价[J/OL]. 物探化探计算技术, 1-10. <http://kns.cnki.net/kcms/detail/51.1242.P.20250310.1610.002.html>, 2025-06-20.

[10] 贾应, 吴彩燕, 王立娟, 等. 融合 InSAR 与机器学习的滑坡易发性评价[J]. 大地测量与地球动力学, 2025, 45(3): 231-238.

[11] 肖婷, 刘庆丽, 邓敏, 等. 三峡库区万州区滑坡易发性演化规律[J]. 地球科学, 2025, 50(4): 1625-1637.

[12] 黄智杰, 沈佳, 简文彬, 等. 基于 XGBoost 模型的降雨诱发阶跃型滑坡位移预测[J]. 自然灾害学报, 2023, 32(2): 217-226.

[13] 王惠琴, 梁啸, 何永强, 等. 融合 XGBoost 和 SVR 的滑坡位移预测[J]. 湖南大学学报(自然科学版), 2025, 52(4): 149-158.

[14] 周伟, 李景娟, 李炎隆, 等. 基于 BOVMD-P-BOXGBoost 的阶跃式滑坡位移预测[J/OL]. 南水北调与水利科技(中英文), 1-19. <http://kns.cnki.net/kcms/detail/13.1430.tv.20250205.1624.004.html>, 2025-06-20.

-
- [15] 曹文康, 徐哈宁, 肖慧, 等. 基于贝叶斯优化 XGBoost 的多元输入模型对滑坡位移预测效果研究[J]. 测绘工程, 2024, 33(2): 49-55.
- [16] 蒋宏伟, 刘健鹏, 王新杰, 等. 随机森林优化的静动态耦合模型在滑坡位移预测中的应用[J]. 常州大学学报(自然科学版), 2024, 36(3): 80-92.