

随机森林算法在覆岩两带预测中的应用

李向东, 袁同佟

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28日; 录用日期: 2025年7月28日; 发布日期: 2025年8月7日

摘 要

为了应对煤矿开采过程中“垮落带”和“导水裂隙带”(两带)高度预测难的问题, 运用了基于随机森林算法建立两带高度预测模型, 在综合考虑各种复杂的地质构造、岩层组合、采厚、采深等因素的影响下, 利用随机森林算法本身的双重随机性(Bagging自助采样以及特征随机选择)和抗噪声性能好、能自动选择重要特征的优点, 建立了两带高度的预测模型, 用决定系数(R^2)来评价模型的优劣, 结果表明: 此模型的平均预测 R^2 为0.8704, “强解释能力”($0.8 \leq R^2 < 0.9$), 可以为工程实际提供可靠的依据, 相较于线性回归、KNN等传统方法来说, 随机森林算法有较强的鲁棒性和准确性。

关键词

垮落带, 导水裂隙带, 随机森林算法, 预测模型

Application of Random Forest Algorithm in the Prediction of Overburden and Two Zones

Xiangdong Li, Tongtong Yuan

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

Received: Jun. 28th, 2025; accepted: Jul. 28th, 2025; published: Aug. 7th, 2025

Abstract

In order to deal with the problem of difficulty in predicting the height of the “collapse zone” and the “water conduction fracture zone” (the two zones) in the process of coal mining, the height prediction model of the two zones was established based on the random forest algorithm, and the prediction model of the height of the two belts was established by taking into account the influence of various complex geological structures, rock formations, mining thickness, mining depth and other

文章引用: 李向东, 袁同佟. 随机森林算法在覆岩两带预测中的应用[J]. 矿山工程, 2025, 13(5): 855-861.

DOI: 10.12677/me.2025.135096

factors, and using the advantages of the random forest algorithm itself (Bagging self-sampling and random selection of features) and the advantages of good noise resistance and automatic selection of important features. The results show that the average prediction R^2 of the model is 0.8704, and the “strong explanatory ability” ($0.8 \leq R^2 < 0.9$) can provide a reliable basis for engineering practice, and the random forest algorithm has strong robustness and accuracy compared with traditional methods such as linear regression and KNN.

Keywords

Collapse Zone, Zones of Water-Conducting Fractures, Random Forest Algorithm, Predictive Models

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

研究意义与背景

由于近年来煤矿开采深度不断加深、强度不断增大,浅部资源面临枯竭的压力,为了满足生产需求,转向了更为困难的煤层进行开采,“两带”即“垮落带”和“导水裂隙带”的问题越发突出,带来的危害也越来越大。据国家统计局中国统计年鉴的数据可知:从2015到2022年间,顶板事故(垮落带)发生的次数占当年煤矿事故总数的比例基本都达到了60%~70%,死亡人数占比则为40%~50%;透水事故(导水裂隙带)发生的次数占全年煤矿事故总数的比例为5%~10%,死亡人数占比为15%~25%。由此可知,预测“两带”高度问题对于煤矿安全生产及矿井涌水量保障是极其重要的,同时也是核心技术问题,在矿山生命周期的所有阶段都发挥着重要作用。利用计算机技术发展的契机,机器学习技术应运而生,并已经广泛应用于各领域,其中机器学习领域的随机森林算法也被引用到此预测建模中,在预测时比传统的贝叶斯模型[1]更适用且效果更好。

2. 随机森林算法的原理及优势

2.1. 机器学习的原理

通过从数据中自动发现规律来实现在未知数据上的预测或决策,使得机器学习算法能够使计算机系统实现从数据到预测或决策的自动化,把大量的数据给到机器学习算法,获取其中存在的内在联系,形成需要的模型,本文使用随机森林对覆岩“两带”高度进行预测及分析,针对裂隙带和垮落带分别做出研究。在本研究中,涵盖了数据采集、特征选择和模型训练等流程,用随机森林算法处理地质问题综合效果最好,可较好地总结出地质特点与规律,在覆岩两带高度预测方面建立模型后可以得到较好的预测结果[2]。

2.2. 随机森林的原理

随机森林作为机器学习中主要的分类器之一,它是由许多的独立同分布的决策树构成决策树主要研究样本的规律。随机森林具有双重随机性,每棵决策树使用自助采样法从原始训练集随机抽取样本(有放回)(样本随机,基于Bagging策略)分裂节点时,仅从随机子集中选择最优分裂特征,打破特征间相关性(特征随机)[3]。

随机森林算法流程主要如图 1 流程图所示:

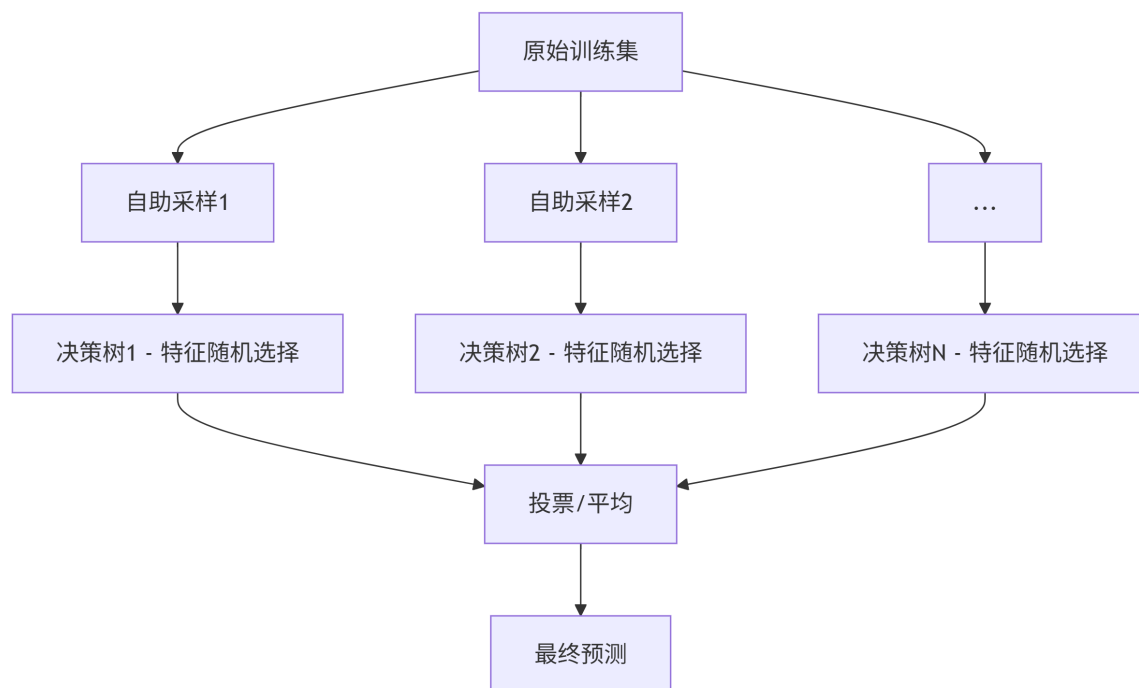


Figure 1. Flow chart
图 1. 流程图

相比于其他算法,随机森林在高维特征方面拥有很大的优势,在这之上会自行选取有效而重要的地质变量(比如岩性厚度、倾角、抗压强度)排除掉部分无关的特征,将重点放在有效的信息上[4];同时应用到覆岩两带的高度预测之上能充分发挥出地层条件的复杂性及数据特征多样性,并可以从中发现数据间的非线性联系,而且还可以结合更多复杂的地质参数之间的非线性耦合作用的关系,这样就使得随机森林算法的有效性得到很大的提高,并且不会因为某几个特征导致预测结果不准确的情况发生。

2.3. 随机森林的优势

2.3.1. 非线性建模能力优秀

垮落带与导水裂隙带的发育高度受控于多方面的因素,如:地质构造、岩性组合、煤层倾角、采厚、采深以及采煤方法等等。线性回归假定变量之间存在线性关系,在实际工作中两者的关系是错综复杂的,很难用直线来加以表达,且不一定严格地呈线性关系。而随机森林算法可以发现更加复杂的非线性关系及变量间的互作关系,而不需要像线性模型一样去做过于复杂的特征工程(比如多项式变换)。

2.3.2. 对噪声和异常值的鲁棒性

地质勘探与工程测量的采集数据不可避免地存在一些噪声、误差以及异常值。而随机森林采用 Bagging (自助采样)和特征随机子集选择法构建出多棵决策树,并分别通过计算其平均值(回归)或投票(分类)来得出最后的结果,利用了集成学习的思想使某个偶然的数据样本或者某个孤立的数据点不会对最终的预测结果产生过多影响,进一步保证了模型的稳定性。

2.3.3. 自动特征选择和处理高维特征

预测垮落带与导水裂隙带要考虑到很多方面的东西:地质、采矿、水文等。而随机森林是建立每棵

决策树的过程中都是按照特删选择来选取属性进行分裂的,这就相当于其本身就是一个隐式的正则化,所以当出现较多的属性的时候,随机森林可以有效的防止过拟合的发生;除此之外,随机森林还能够提供一些特征重要性的判别标准(如根据基尼不纯度降低或袋外错误来进行),并能直接告诉研究人员采厚、关键层强度和岩性组合指数等多少个变量的影响最大。

2.3.4. 较好的泛化能力与较低的过拟合风险(通过 Bagging)

Bagging 是随机森林算法防止过拟合的重要手段,它会从原始样本中有放回地进行多次采样得到不同训练子集,为每一颗树提供不同的训练集,最后聚合各个树的结果,这样可以降低模型方差,增强对未知数据的泛化能力,可以适当调节每棵树最大深度、叶子节点最小样本数量等参数对模型复杂度进行调整。

2.3.5. 能够处理混合数据类型

预测模型含有连续型变量(例如采深、采高、岩层强度)与类别型变量(例如岩性类别、顶板管理方式、开采方式),随机森林算法可以不需要做独热编码处理即可直接应用到这两种类型的特征变量上,也不会造成维度上升的情况以及稀疏性的难题。

2.4. 对比与总结

针对包含地压矿井灾害等复杂的地质采矿因素所引起的非线性问题(如垮落带、导水裂隙带)进行预测,运用基于随机森林的处理方法取得了较好的结果,随机森林算法具有很好的综合优势。由于线性回归存在的线性假设过于僵化,不能准确刻画出实际的关系;KNN 容易受到高维的影响,效率较低且无解释性,而随机森林能够利用树木结构进行建模,且易于使用 Bagging 集成、随机特征选取等方式进行建模,模型的鲁棒性强于其他两类算法,因此该算法的稳定性更好。与集成树的顶级算法 XGBoost 和 CatBoost 相比,随机森林算法虽然没有达到极限精度(尤其是在经过 XGBoost 或 CatBoost 调参之后),但在一些方面还是拥有一定优势[5]:① 更强大的抗异常值能力;② 更有意义、更容易理解的特征重要性输出(这有助于工程决策的解释原因);③ 更简单、更鲁棒的调优方式(参数较少且简单);④ 直接支持混合数据类型的无缝训练,不需要两者的任何预处理方式(连续/类别)。另外,随机森林不需要对数据做特殊的预处理,直接提供给 Bagging 机制就能够保证更好的泛化性能。因此,在充分兼顾模型鲁棒性、可解释性、易用性及稳定性这一要求的工程实现过程中,通常来说,RF 是较为适宜且高效的方案,而 XGBoost/CatBoost 则可以在有精度提升需求的情况下(但是其自身调优的成本很高),用作一个良好的备选方案,但是对于线性回归以及 KNN 的使用而言,则需要有所警惕[6]。

3. 预测模型建立

上一节我们讲了随机森林的优势,在这一节里,我们将基于随机森林算法对覆岩两带的高度进行预测,并使用 R^2 值来评价所建立的模型,再将该模型用于分别预测垮落带和导水裂隙带的覆岩。并且我们利用 Windows11 系统以及 Matlab2024b 实现了该过程。Matlab 是由美国 mathworks 公司推出的高性能数值计算和科学计算的平台软件,具有矩阵运算、数据可视化、算法开发和多领域的仿真的优点,主要应用于工程学、科学和数学领域。凭借 Matlab 强大的数值计算和交互式的编程功能作为编写代码的平台。

3.1. 主要流程

(1) 数据准备:准备好 csv 或者 excel 格式的数据表格,并将表内数据信息填入,其中数据包括覆岩岩性、采高、煤层倾角、工作面斜长、采深、开采方法以及垮落带高度、导水裂隙带高度等。再将这些数据进行预处理后,设置随机森林超参数,即设置特征子集大小、最大深度、树的数量等等。

(2) 特征工程: 构造复合特征, 如岩性脆塑性系数。处理缺失值, 将类别变量变法(分层 = 1, 炮采 = 2, 综采 = 3, 综放 = 4 等等)。

(3) 构建决策树: 通过自助抽样法有放回地抽取子集, 从子集中随机选择特征, 循环生成多棵回归决策树。

(4) 训练模型和预测输出

通过所有决策树完成后再将它们合成为随机森林模型, 训练完成后将待预测样本输入, 然后依次经过每一个树, 最后取得每棵树的预测值的均值即为该点覆岩两带高度预测值。总流程如图 2 总流程图所示:

总流程如图 2 总流程图所示:

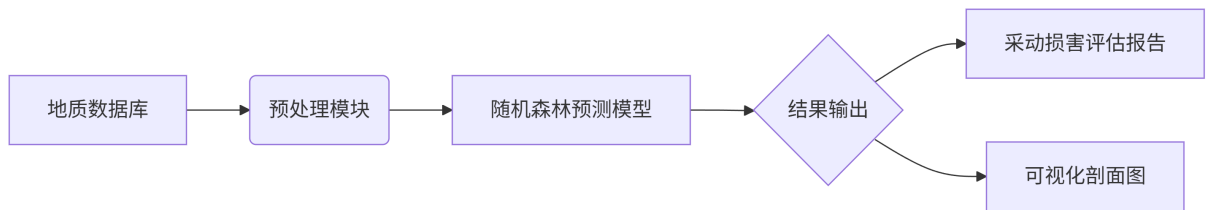


Figure 2. Total flow chart
图 2. 总流程图

3.2. 模型评价

3.2.1. 决定系数(R^2)

决定系数(R^2)是回归模型的基本评价标准, 用来反映模型对因变量变异的解释程度。其中数学公式表示为式(1) [7]:

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

其中 SS_{res} 表示残差平方和(预测误差的平方和), 反映模型未解释的变异。 SS_{tot} 表示总平方和(数据自身的方差), 反映目标变量的总变异。 $\bar{y}$ 表示目标变量的均值(“最朴素模型”的预测值)。

3.2.2. 性能判断标准

决定系数(R^2)的判断标准如表 1 所示:

Table 1. Performance judging criteria
表 1. 性能判断标准

R^2 范围	模型解释能力	可靠性分析
$R^2 \geq 0.9$	极强	优秀, 可直接用于决策
$0.8 \leq R^2 < 0.9$	强	良好, 推荐部署
$0.6 \leq R^2 < 0.8$	中等	需优化特征/算法
$R^2 < 0.6$	弱	不可用, 重新设计模型

注: 地质系统受多因素非线性耦合(如采深、岩性、构造应力), 若 $R^2 > 0.9$, 需警惕过拟合(如模型过度依赖某个矿区局部特征)。

3.2.3. 评价结果

由图3可知随机森林算法应用于两带高度预测后, 垮落带的决定系数 R^2 为 0.8694, 裂隙带的决定系数 R^2 为 0.8714, 平均决定系数 R^2 为 0.8704, 根据表1的数据来看, 该算法建立的模型可靠性良好, 推荐部署。

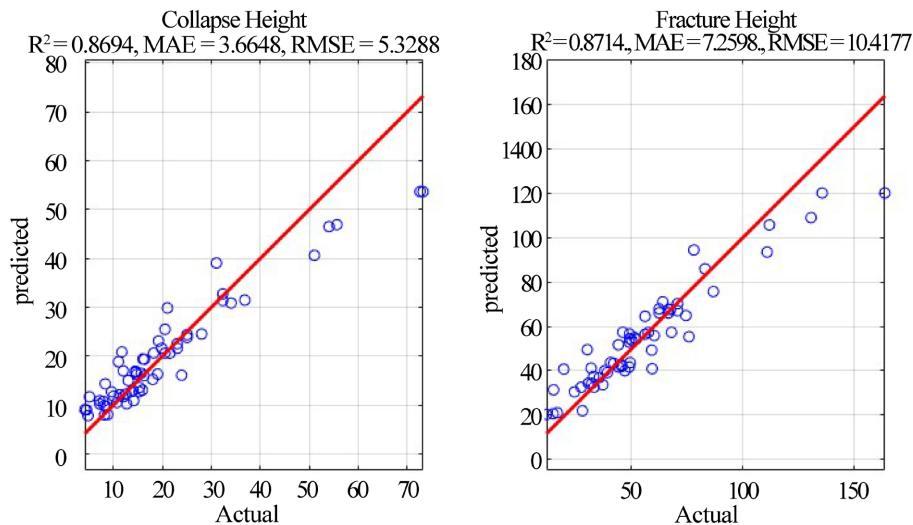


Figure 3. Evaluation results
图 3. 评价结果

4. 结论与展望

4.1. 结论

两带高度预测是防范顶板事故(垮落带相关, 占煤矿事故 60%~70%)、透水事故(导水裂隙带相关, 占 5%~10%)的关键技术手段, 有助于生产安全、生态安全。本文从随机森林算法切入, 一方面分析其有较好的非线性建模能力, 可以精准挖掘出地质参数之间的耦合关系; 特征分析能够智能挖掘关键因素并剔除冗余, 抗噪能力强; 基于 Bagging 机制 + 特征随机性的优势下具备良好的泛化能力; 另一方面, 说明本文所搭建模型评估结果完全达到预期效果, 并优于传统线性方法, 满足工程实现需求。

4.2. 展望

单一的随机森林算法就已经具有较为强大的预测能力, 基于此, 探索随机森林算法与其他算法的集成模式(如: 随机森林算法集成 XGBoost、CatBoost 等), 利用随机森林算法的优势以及 Boosting 类算法高精度的优势, 提升预测性能; 引入时序数据分析方法(如 LSTM)实现开采过程中的两带高度动态监测和预警, 进一步满足矿井巷道顶板底板的有效预警需要; 在试验中要注意针对特定矿区中避免模型过拟合(注意 $R^2 > 0.9$ 的问题); 此外继续开发嵌入式系统, 融合传感器实时数据流到已有模型, 将模型落地于实际智能矿山监测平台应用中。

参考文献

- [1] 孔春芳, 田倩, 刘健, 等. 基于集成学习模型与贝叶斯优化算法的成矿预测[J/OL]. 地学前缘: 1-18. <https://doi.org/10.13745/j.esf.sf.2025.4.66>, 2025-06-28.
- [2] 令狐曦. 机器学习模型测评技术研究[学位论文]. 北京: 北京邮电大学, 2019.

-
- [3] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014, 44(1): 137-141.
 - [4] 刘少泽, 崔美娟, 付晓伟, 等. 顾及缓冲区范围与负样本优化的随机森林地质灾害易发性评价[J]. 科学技术与工程, 2025, 25(15): 6220-6229.
 - [5] 李占山, 刘兆赓. 基于 XGBoost 的特征选择算法[J]. 通信学报, 2019, 40(10): 101-108.
 - [6] 潘自辉, 肖正利, 黄光体, 等. 机载激光雷达数据与机器学习算法的森林蓄积量估测模型构建精度评价——基于 KNN、XGBoost 与 RF 模型反演算法[J]. 湖北林业科技, 2025, 54(2): 34-44, 50.
 - [7] 胥雪炎, 李补喜. 不同被解释变量选择对决定系数 R^2 的影响研究[J]. 太原科技大学学报, 2007(5): 363-365.