文章编号:1671-4598(2025)07-0105-09 DOI:10.16526/j.cnki.11-4762/tp.2025.07.013 中图分类号:TP642.2;TP391.4;TP183 文献标识码:A

地下工程场景下基于改进随机森林的滑坡预测方法

许文学，陈金磊，陈宗清，柳倩男，李 玉
（中国人民解放军 93204 部队，北京 100068）

摘要：为解决传统滑坡预测方法依赖专家经验、难以适应复杂地下工程场景的问题，研究提出一种改进随机森林方法以提升预测精度与泛化能力；通过专业设备采集地下工程滑坡关键参数，经对不完备数据处理构建高质量数据集；对比分析 k 近邻、支持向量机、决策树等分类器的全局与局部分类性能，确定随机森林为最优基础分类器；创新性引入多决策树相关性度量方法，以特征空间内积计算量化决策树间冗余性，通过最优阈值筛选构建改进随机森林；实验表明，改进随机森林的分类精度达 93.05%，其 Precision、Recall 和 F_1-score 指标在不同标签数据上均保持最高稳定性，验证了改进神经网络在整体与局部分类性能上的双重优势；实际工程应用验证了该方法在多维复杂场景下的有效性，为地质灾害智能化预测提供了可靠解决方案。

关键词：滑坡预测；随机森林；相关性度量；数据清洗；地下工程

Landslide Prediction Method Based on Improved Random Forest in Underground Engineering Scenarios

XU Wenxue, CHEN Jinlei, CHEN Zongqing, LIU Qiannan, LI Yu
(Unit 95899, PLA, Beijing 100068, China)

Abstract: To address the limitations of traditional landslide prediction methods that rely on expert experience and struggle to adapt to complex underground engineering scenarios, an improved random forest method is proposed to enhance prediction accuracy and generalization capability. Key parameters of landslides in underground engineering environments are collected using specialized equipment, and a high-quality dataset is generated through incomplete data processing. Compared with the global and local classification performance of classifiers such as k-nearest neighbors, support vector machines, and decision trees, random forest is identified as an optimal base classifier. An innovative multi-decision tree correlation measurement method is introduced, which quantifies redundancy between trees through inner product calculations in feature space and constructs an improved random forest by optimal threshold filtering. Experimental results show that, the improved random forest achieves a classification accuracy of 93.05%, with the Precision, Recall, and F_1-score maintaining the highest stability on different labeled data, validating its dual advantages in both global and local classification performance. Practical engineering applications verify its effectiveness in multidimensional complex scenarios, providing a reliable solution for intelligent geological hazard prediction.

Keywords: landslide prediction; random forest; correlation measurement; data cleaning; underground engineering

0 引言

地下工程建设在现代城市化及基础设施发展中起着至关重要的作用，尤其是在隧道、地下空间、地铁等项目的建设过程中，地质条件和自然灾害成为影响工程安全和施工进度的主要因素之一[1]。滑坡作为一种典型的地质灾害，对地下工程的安全性构成了严重威胁。滑坡的发生不仅会导致巨大的经济损失，还可能危及生命安全，因此，滑坡的预测和预防在地下工程建设中显得尤为重要。有效的滑坡预测可以为工程建设提供及时的预

收稿日期:2025-02-25； **修回日期**:2025-04-22。

作者简介:许文学(1980-),男,硕士,工程师。

引用格式:许文学,陈金磊,陈宗清,等.地下工程场景下基于改进随机森林的滑坡预测方法[J].计算机测量与控制,2025,33(7):105-113.

警，帮助工程人员采取适当的防护措施，从而减少灾害损失，确保地下工程的安全和稳定[2]。

传统滑坡预测方法主要基于地质调查、经验规则、统计分析等手段，依靠专家知识和大量的现场数据来推测滑坡发生的概率[3]。早期的滑坡预测方法一般使用地质调查数据，结合坡度、地质类型、土壤水分、降水量等地质和气象因素，通过建立风险评估模型进行滑坡预测。例如，文献［4］提出了一种基于均值的低阶自回归张量完成（MLATC）的滑坡位移时间序列预测模型。文献［5］使用了数据驱动滑坡易发性评估方法和基于物理的滑坡稳定性评估方法进行水库滑坡预测。文献［6］开发了基于 GIS 的陕南土石接触带滑坡地质灾害预测方法，建立滑坡地质灾害信息值预测模型。文献［7］提出了降雨条件下水库滑坡的 ILF-FFT 预测预警模型，基于速度和加速度时程函数的极限曲率，实现了滑坡演化的四阶段定量划分方法。

然而，传统滑坡预测方法存在一定的局限性。首先，这些方法通常依赖于专家经验和直觉，缺乏系统的科学依据和数据支持，容易导致预测结果的主观性和偏差。其次，传统滑坡预测方法多依赖于地质背景的局部数据，难以有效处理地下工程环境中复杂的多维度、多变量问题。更重要的是，传统滑坡预测方法的预测精度往往受到数据质量的限制，对于大规模滑坡预测来说，传统滑坡预测方法的适应性和泛化能力较弱，无法满足实际工程应用的高要求。因此，亟需开发新的滑坡预测方法，以提高预测精度、适用性和可靠性。

随着人工智能技术的发展，基于机器学习的智能滑坡预测方法逐渐成为研究的热点。其原理是通过分析历史滑坡事件与相关环境因素（如地形、地质、气象等）的关系，构建模型以预测新的滑坡发生风险。该过程首先通过收集和处理大量的数据，提取出关键特征，然后利用分类器（如支持向量机、决策树、随机森林等）对数据进行训练，学习滑坡发生的规律和模式。训练好的分类器可以在面对新的地质条件时，通过输入相应的特征数据来预测滑坡的可能性。机器学习能够处理复杂的非线性关系，挖掘出潜在的模式和关联，提供比传统滑坡预测方法更为准确和动态的滑坡预测[8]。许多学者开展相关研究，如，文献［9］提出了一种新的模糊深度学习（FuDL）模型，用于近实时地震引发的滑坡空间预测。文献［10］提出了一种基于混沌高斯变异麻雀搜索算法优化 BP 神经网络（CG-SSA-BP）的滑坡位移预测方法。文献［11］使用机器学习算法建立了类似类型滑坡变形、混合类型滑坡变形和单个滑坡变形的预测模型。尤其的，随机森林作为一种集成学习方法，由于其在处理高维数据时的优势，也在滑坡预测中得到了广泛应用，特别是在处理非线性问题时表现出了较强的预测能力。如，文献［12］利用贝叶斯超参数优化随机森林的超参数，然后选择最优超参数进行滑坡易发性映射。文献［13］使用随机森林来预测台湾中部曾文河流域（TRW）的滑坡区域。

尽管智能滑坡预测方法相较于传统滑坡预测方法具有较高的准确性和适应性，但现有研究仍存在一些问题。首先，现有的机器学习分类器大多在实验室环境下进行验证，使用的数据集往往为公开数据集，缺乏对地下工程场景的具体适应性，导致这些分类器在实际应用中的效果较差。例如，许多研究使用的是地质学研究中的公开数据集，这些数据集的环境条件与实际地下工程中所面临的复杂情况存在较大差异。因此，用于滑坡预测的分类器需要结合实际工程环境中的复杂数据，才能更好地满足工程应用需求。其次，现有研究大多专注于选择最先进的分类器，或者在现有分类器基础上进行优化，尤其是在参数调优上进行大量探索，以寻找最佳组合。然而，这些方法忽视了分类器的理论深度和优化空间，往往缺乏从构建原理上对分类器进行优化。例如，某些研究通过深度学习方法或增强学习对现有模型进行微调，虽然在某些数据集上取得了较好的结果，但由于缺乏系统性的优化，导致了分类器的泛化能力和稳定性较差。因此，分类器选择与优化应同时关注理论基础与实际数据的相互配合，而非单纯依赖参数调优。最后，许多智能滑坡预测方法在方案设计上较为简单，通常选择某一分类器进行实验，但未能进行充分的对比分析，未给出最优分类器的选择依据。虽然许多研究对不同分类器进行了初步尝试，但通常没有进行详细的性能对比，也未明确给出不同算法在特定工程环境下的适用性。例如，文献［14］使用改进随机森林对山体滑坡进行预测，但未提供详细的性能对比分析，也没有充分考虑数据集特征与算法适配性之间的关系。因此，应当加强对不同分类器的系统性对比分析，从而为选择最适合的分类器提供更加科学的依据，以确保智能滑坡预测方法在实际工程应用中的有效性和可操作性。

因此，针对现有研究存在的不足，本文提出了一种基于改进随机森林的滑坡预测方法，旨在解决现有传统与智能滑坡预测方法在地下工程场景下应用的局限性。首先，在真实的地下工程场景中，采用专用设备对滑坡发生的关键参数进行现场采集，并利用采集到的原始数据构建初步的数据集。随后，通过对数据集进行数据清洗，包括缺失值和异常值处理，确保了数据的高质量和可靠性，为后续建模奠定了坚实的基础。其次，在数据集上训练了多种最新的或经典的适用数据分类的分类器和神经网络，并对它们在数据集上的整体分类性能以及在不同标签的数据上的局部分类性能进行了详细的对比分析。通过综合评估，最终选择了随机森林作为最优分

类器，并详细阐述了选择的科学依据和其在特定工程环境中的适应性。最后，基于对随机森林构建原理的深入理解，提出了一种基于多决策树相关性度量的改进随机森林方法，并与经典随机森林及其他改进版本进行了充分的对比研究。结果表明，改进后的随机森林在滑坡预测中的表现更加优越，进一步提升了预测精度和稳定性，有效解决了传统和智能滑坡预测方法在地下工程场景中应用的局限性。本文有效弥补了现有滑坡预测方法中存在的不足，不仅为地下工程场景下滑坡预测提供了一种新的解决方案，也为滑坡预测领域的理论与实践发展提供了重要的参考和借鉴。

本文的主要贡献如下：(1) 采集真实地下工程应用场景下的数据并进行数据清洗，构建高质量与高可靠数据集。(2) 综合评估不同分类器和神经网络的整体和局部分类性能，提供最优分类器的选择依据。(3) 提出了基于多决策树相关性度量的改进随机森林，提升了滑坡预测的精度与稳定性。

1　经典随机森林简述

随机森林是基于集成学习的组合分类器，由多个 CART（以二叉树形式构建的决策树）构成。它在抽样阶段采用 Bagging 建立多个样本子集[15]，并分别在每个样本子集上训练一个 CART，得到多个 CART。根据合适的组合方式，将多个 CART 组合来构建随机森林。当使用随机森林解决分类问题时，多个 CART 并行的对同一分类对象进行预测并给出分类结果。通过多数表决规则对多个 CART 给出的分类结果进行投票，得到最终分类结果[16]。随机森林构建流程如下。

步骤一：参考 Bagging 在含有 N 个样本的样本集上进行 n 次随机抽样，由抽取的 n 个样本来构建样本子集。重复 k 次上述过程，得到 k 个样本子集，每个样本子集包含 n 个样本。

步骤二：分别在 k 个样本子集上训练一个 CART，得到 k 个 CART。其中，CART 中的每个节点都是从包含 M 个特征的特征集中选取 m 个特征进行构建的。每个样本子集都有一个对应的特征子集。在构建每个 CART 时，是选择基尼指数最小的特征来分裂节点，计算公式如下。CART 中的其他节点都采取相同的决策标准进行构建，直到该节点的所有样本都属于同一类别或者已经到达树的最大深度：

$$\mathrm{Gini}(p) = \sum_{i=1}^{L} p_i(1-p_i) \tag{1}$$

式中，L 表示样本子集中标签的个数，i 表示某个标签，p_i 表示样本的标签是 i 的概率。

步骤三：在上述两个步骤的基础上，得到多个样本子集和特征子集，分别用于训练多个 CART，进而构建随机森林。通常，用于构建随机森林的多个 CART 不进行剪枝。

步骤四：对分类对象进行预测时，多个 CART 独立、并行的进行预测并给出分类结果。通过对多个分类结果进行多数表决，得到一个综合决策结果作为随机森林的分类结果。

2　基于改进随机森林的滑坡预测方法

2.1　基于多决策树相关性度量的改进随机森林

如 1 节中所述，在构建随机森林的过程中，样本子集和特征子集都是随机选择的。这可能会导致训练的某些 CART 的分类性能较差，而这些 CART 同样对随机森林分类性能的贡献较小。因此，本文提出一种基于多决策树相关性度量的改进随机森林方法。具体的，对于每个 CART，预留 5 个样本子集来测试它的分类性能，统计它在 5 个样本子集上取得的平均分类精度。在此基础上，根据各 CART 取得的平均分类精度，对多个 CART 进行降序排列。

此外，在 CART 的构建过程中，采用的有放回的随机抽样方法，这可能会出现两个 CART 之间相似度较大的情况，即，两者之间存在较强的相关性。这可能会导致出现冗余分类结果的问题。因此，精简相关性较强的 CART，可以提高随机森林的计算效率，但 CART 之间的相关性也并不是越弱越好。一方面，如果 CART 之间的相关性太弱，会导致构建随机森林的 CART 的数量不足，降低随机森林的分类性能和稳定性。另一方面，CART 之间的相关性太弱意味着通过 Bagging 抽取的样本子集之间的交集较少，会导致抽取的样本子集不能覆盖原始样本集中的所有样本。这会导致构建的随机森林不能拥有全局的掌控能力，即其分类性能并不完备。

针对上述问题，在本文提出的改进随机森林方法中，使用向量内积法对 CART 的相关性进行度量。并且，以随机森林取得的分类精度为评价指标，寻找合适的内积阈值。如果 CART 之间的内积数值大于内积阈值，则它们被判断为彼此之间相关性较强。这样，一对 CART 中取得平均分类精度较低的那个需要被删除。

在创建随机森林的过程中，本文在预设数量的基础上多训练一定数量的 CART。然后，综合考虑 CART 取得的平均分类精度与彼此之间的相关性，删除那些相关性较强且分类性能较差的 CART，直至剩余 CART 的数量达到预设数量为止。这样，被保留的 CART 是较为优秀的，由此构建的随机森林能够取得更好的分类性能和计算效率。本文提出的改进随机森林方法的具体实现步骤如下，其设计流程如图 1 所示。

步骤一：通过不放回的随机抽样方式从原始样本集中分别选出 5 个样本子集，作为评估每个 CART 分类

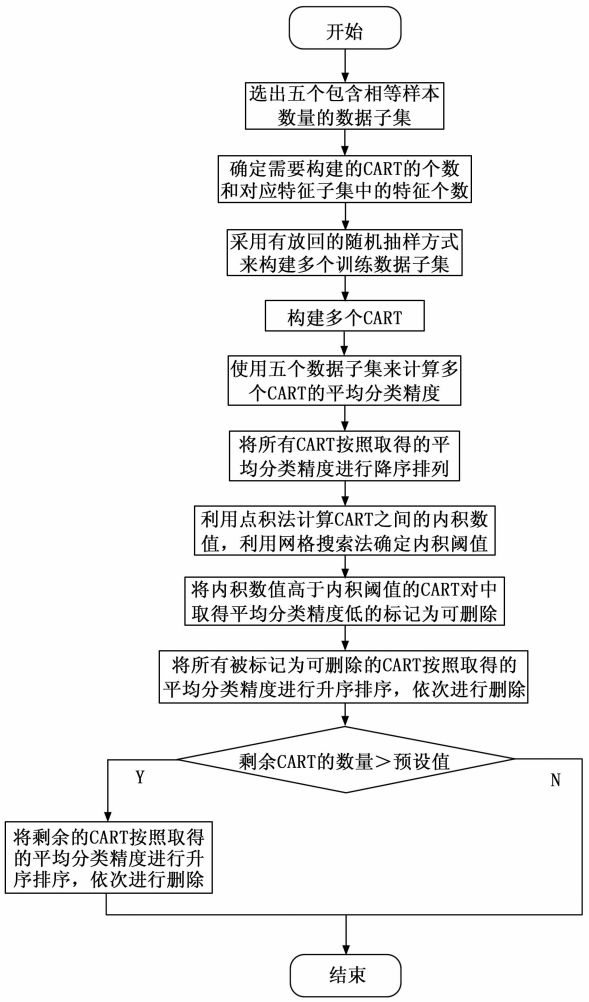

图 1 改进随机森林方法的设计流程

性能的测试集。需要说明的是，5 个样本子集包含的样本数量是相同的，但是具体包含的样本是不同的，且这些样本的标签是已知的。

步骤二：确定需要训练的 CART 的个数 N 和对应特征子集中的特征个数。利用 Bagging 在原始样本集的剩余样本上进行 $(1+m)\times N$ 次有放回的随机抽样，构建 $(1+m)\times N$ 个训练集。在此基础上，训练 $(1+m)\times N$ 个 CART。其中，$m\times N$ 个 CART 是在预设数量的基础上多训练一定数量的 CART。

步骤三：应用每个 CART 在 5 个样本子集上进行预测，将取得的分类精度表示为 a_i^j。其中，$i=1,2,\cdots,(1+m)\times N$，表示第 i 个 CART，$j=1,2,3,4,5$，表示第 j 个样本子集。

步骤四：计算第 i 个 CART 取得的平均分类精度，公式为：$\bar{a}_i=\dfrac{a_i^1+a_i^2+a_i^3+a_i^4+a_i^5}{5}$。

步骤五：将 CART 按照取得的平均分类精度进行降序排列。

步骤六：采用公式（2）所示的向量点积法计算并保存 CART 之间的内积数值。以随机森林取得的分类精度为评价指标，结合网格搜索法寻找使随机森林取得最高分类精度的内积阈值，称为最优内积阈值 t。这样，对于那些内积数值小于内积阈值的一对 CART，不进行处理。对于那些内积数值大于内积阈值的一对 CART，将其中取得平均分类精度低的 CART 标记为可删除：

$$\mathrm{Sim}(D_i,D_j)=\mathrm{acos}(W_i\cdot W_j) \tag{2}$$

式中，D_i 和 D_j 分别表示两个 CART，W_i 表示 CARTi 对应的特征子集，W_j 表示 CARTj 对应的特征子集，acos（·）表示反余弦函数。

$W_i\cdot W_j$ 的详细计算公式如下：

$$W_i\cdot W_j=\frac{\sum_{m=1}^{p}\sum q_{n=1}I(W_{im}=W_{jn})(W_{im}\cdot W_{jn})\cdot 180^\circ}{\sum_{m=1}^{p}\sum_{n=1}^{q}(W_{im}\cdot W_{jn})} \tag{3}$$

式中，W_{im} 表示特征子集 W_i 中的第 m 个特征，W_{jn} 表示特征子集 W_j 中的第 n 个特征；I 是指示函数，只有当 $W_{im}=W_{jn}$ 时，$I(W_{im}=W_{jn})=1$，否则 $I(W_{im}=W_{jn})=0$；$W_{im}=W_{jn}$ 表示不同的特征子集中选择了相同的特征。

步骤七：将所有被标记为可删除的 CART 按照取得的平均分类精度进行升序排列，并依次进行删除，直到剩余的 CART 的数量为 N。即，剩余的 CART 的数量达到预设数量。需要说明的是，如果被标记为可删除的 CART 都被删除后，剩余的那些未被标记的 CART 的数量还大于 N，则根据步骤五给出的 CART 的排序进行依次删除，直到剩余的 CART 的数量为 N。

步骤八：利用保留的 N 个 CART 来构建随机森林。结合多数表决规则，确定改进随机森林的决策结果。

2.2 滑坡预测方法

在 2.1 节改进随机森林方法的基础上，本文进一步提出滑坡预测方法，具体实现步骤如下，其设计流程如图 2 所示。

步骤一：在斜坡上布置专用的参数采集终端，实时采集多个地质参数特征并添加标签，构建原始数据，建立原始数据集和原始特征集。

步骤二：对原始数据集进行异常值与缺失值检测，视为不完备数据，根据同一规则进行不完备数据处理，得到完备数据集。

步骤三：在完备数据集和原始特征集的基础上，参考 2.1 节所提基于多决策树相关性度量的改进随机森林的设计流程，确定最优内积阈值 t，构建用于滑坡预测的改进随机森林。

步骤四：对于未知状态下参数采集终端采集的多个地质参数特征，进行不完备数据检测，构建待测数据，

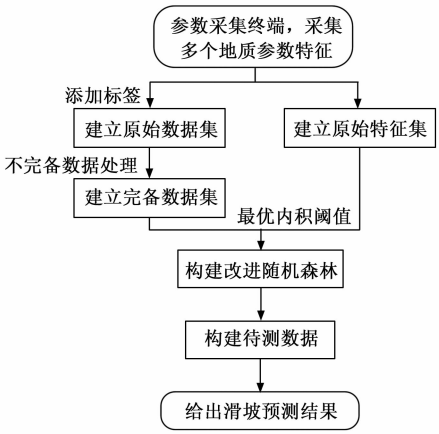

图 2　滑坡预测方法的设计流程

由改进随机森林给出预测结果。

3　实验与分析

3.1　原始数据集

前期，作者们开发了一套参数采集终端放置于地下工程场景下多个斜坡上，参数采集终端连接的集成传感器模块插入斜坡来采集地质参数，包括：倾斜角度、地表位移、水平位移、垂直位移、水压力、土壤形变、土壤应力、土壤湿度、土壤温度、剪切强度和电导率，共 11 个地质参数，称为 11 个特征。这样，参数采集终端在运行过程中，每次采集会得到 11 个地质参数的数值、形成一条数据，多次采集最终会得到多条数据。根据地下工程场景下滑坡预测的实际需求，本文将这些数据分为 4 个类别，包括：正常、小幅水土流失、水土流失和滑坡，量化为“0”“1”“2”和“3”4 种标签，并为这些数据添加对应的标签。

以中国南方某省份的地下工程为例，在 2022 年 1 月至 10 月期间，使用参数采集终端采集了大量数据，并根据实际情况为数据添加了上述标签，构建了一个原始数据集，其详细描述如表 1 所示。值得说明的是，在构建原始数据集之前，对于采集的原始大量数据，人为的进行了数据平衡处理，确保 4 种标签的数据的数量大致相等。具体的，以标签为“3”的数据的数量为参考数量，分别在其他 3 种标签的数据中筛选近似相等数量的数据。经过统计，原始数据集共包含 30 389 条数据。

表 1　原始数据集的详细描述

标签	数据个数	备注
0	7 591	正常
1	7 609	小幅水土流失
2	7 613	水土流失
3	7 576	滑坡

3.2　完备数据集

参考正态分布的 3σ 原则，本文对原始数据集进行异常值检测。具体的，以 11 个特征为基础单位，每次对属于同一特征的 30 389 个数值进行检测，将超过 3 倍标准差的数值视为异常值。这样，那些存在异常值的数据被称为异常数据。同样的，本文对原始数据集进行缺失值检测。具体的，本文将那些数值为“NULL”及数值为“?”等特殊符号的数值视为缺失值。这样，那些存在缺失值的数据被称为缺失数据。在此基础上，本文将异常数据和缺失数据统一称为不完备数据。值得说明的是，不完备数据中可能存在一些既包含异常值也包含缺失值的数据。

经过统计，标签为“0”的数据中存在 43 条不完备数据，标签为“1”的数据中存在 51 条不完备数据，标签为“2”的数据中存在 45 条不完备数据，标签为“3”的数据中存在 56 条不完备数据，共 195 条不完备数据。文献［17］的研究表明，对于一个数据集，如果异常数据的占比不超过 2%，或者缺失数据的占比不超过 3%，则可以直接丢弃这些异常数据或缺失数据，因为它们对数据集的完备性及多样性不构成威胁。

经过计算，标签为“0”的不完备数据在标签为“0”的数据中的占比约为 0.57%，标签为“1”的不完备数据在标签为“1”的数据中的占比约为 0.67，标签为“2”的不完备数据在标签为“2”的数据中的占比约为 0.59，标签为“3”的不完备数据在标签为“3”的数据中的占比约为 0.74%，所有不完备数据在数据集中的占比约为 0.64%。可以看出，不管是在各个标签的数据上，还是在整个数据集上，不完备数据的占比都没有超过 1%，更不用说其中异常数据或缺失数据的占比。因此，本文直接去除原始数据集中的不完备数据，由此得到地下工程场景下的高质量的完备数据集，其详细描述如表 2 所示。

表 2　完备数据集的详细描述

标签	数据个数	备注
0	7 548	正常
1	7 558	小幅水土流失
2	7 568	水土流失
3	7 520	滑坡

3.3　最优分类器

机器学习中的分类器通常被分为 3 种类型，基于数学理论的分类器、基于树模型的分类器和深度学习模型。前者包括 k 近邻[18]（kNN）和支持向量机[19]（SVM）。中者包括决策树[20]和随机森林[21]。后者包括多层感知机[22]（MLP）和一维卷积神经网络[23]（1D-CNN）。

对于表 2 所示的完备数据集，本文按照 7：3 的比例将其划分为训练集和测试集，在训练集上分别训练上述 6 个分类器，获取它们取得最高分类精度时的参数配置，如表 3 所示，并在测试集上再次测试它们的分类性能，统计它们取得的分类精度，以及它们在 4 种标签的数据上取得的 Precision、Recall 和 F_1-score。表 4 给出了 6 个分类器取得的分类精度。表 5 至表 7 给出了 6 个分类器取得的 Precision、Recall 和 F_1-score。

表 3 6 个分类器的最优参数配置

分类器	最优参数配置
kNN	n_neighbors = 5，weights = 'distance'，metric_params = 'minkowski'
SVM	kernel = 'rbf'，C = 0.1，gamma = 0.02
决策树	criterion = 'gini'
随机森林	n_estimators = 120，criterion = 'gini'，max_features = 5
MLP	Sigmoid 激活函数，学习率为 0.08，L_1 正则化
1D－CNN	卷积核为 1×3，步长为 1

表 4 6 个分类器取得的分类精度

分类器	分类精度/%
kNN	85.15
SVM	88.03
决策树	86.41
随机森林	89.84
MLP	78.92
1D－CNN	87.77

表 5 6 个分类器取得的 Precision

分类器	4 种标签的数据			
	0	1	2	3
kNN	0.88	0.87	0.85	0.82
SVM	0.9	0.89	0.89	0.86
决策树	0.88	0.88	0.86	0.84
随机森林	**0.91**	**0.9**	**0.9**	**0.89**
MLP	0.81	0.79	0.79	0.77
1D－CNN	0.9	0.89	0.87	0.85

表 6 6 个分类器取得的 Recall

分类器	4 种标签的数据			
	0	1	2	3
kNN	0.87	0.87	0.85	0.83
SVM	0.89	0.89	0.88	0.86
决策树	0.87	0.88	0.85	0.85
随机森林	**0.9**	**0.89**	**0.9**	**0.89**
MLP	0.82	0.8	0.78	0.78
1D－CNN	0.9	0.88	0.88	0.86

表 7 6 个分类器取得的 F_1-score

分类器	4 种标签的数据			
	0	1	2	3
kNN	0.87	0.87	0.85	0.82
SVM	0.89	0.89	0.88	0.86
决策树	0.87	0.88	0.85	0.84
随机森林	**0.9**	**0.89**	**0.9**	**0.9**
MLP	0.81	0.79	0.78	0.77
1D－CNN	0.9	0.88	0.87	0.85

从表 4 至表 7 中可以看出，随机森林取得最高的分类精度，同样在各标签上取得的最高的 Precision、Recall 和 F_1-score。这表明，随机森林在 6 种分类器中表现突出，侧面说明基于树模型的分类器用于本文滑坡预测的优越性。实际上，作者们前期的探索同样证明随机森林表现最好。这有效证明了本文以随机森林为基础分类器，开展改进随机森林的必要性与正确性。

3.4 改进随机森林

参考第 2.1 节中提出的基于多决策树相关性度量的改进随机森林方法，本文分别对默认参数配置的随机森林和 3.3 节中经过参数寻优的随机森林进行改进，得到两个改进随机森林，分别称为默认改进随机森林和优化改进随机森林。如 2.1 节中所述，在此过程中，确定改进随机森林的内积阈值至关重要。本文将默认改进随机森林和优化改进随机森林的内积阈值分别设置为 21 和 25，对此进行说明。

本文在训练集上训练多个 CART，计算 CART 之间的内积数值。在此基础上，根据历史经验，本文将内积阈值 t 的取值范围设置为 5 至 29。这样，借助网格搜索法，本文将内积阈值 t 的搜索范围设置为 5 至 29。当内积阈值 t 顺序取 5 至 29 之间的任一个值时，本文将高于当前内积阈值 t 的一对 CART 中取得平均分类精度低的删除，用剩余的 CART 构建随机森林，计算其在测试集上取得的分类精度。由此，本文顺序构建 25 个随机森林，并得到 25 个分类精度。此时，通过比较哪个随机森林取得最高的分类精度，其对应的内积阈值 t 被视为最优的内积阈值 t。图 3 给出了内积阈值 t 与随机森林取得的分类精度之间的关系。

从图 3 中可以看出，当内积阈值为 21 时，默认改进随机森林取得的分类精度最高。当内积阈值为 25 时，优化改进随机森林取得的分类精度最高。此后，当内积阈值继续增加时，两个改进随机森林取得的分类精度均略有下降后保持稳定。因此，本文将默认改进随机森林的内积阈值设置为 21，将优化改进随机森林的内积阈值设置为 25。

在此基础上，本文统计两个改进随机森林取得的分类精度。其中，默认改进随机森林和优化改进随机森林

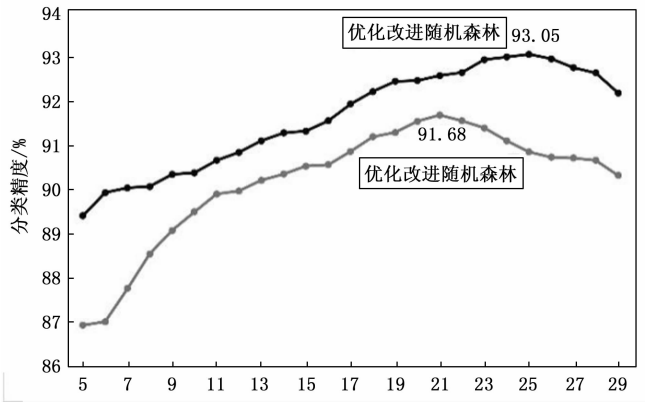

图 3　内积阈值的取值与取得的分类精度之间的关系

取得的分类精度分别为 91.68%和 93.05%。相较于 3.3 节中经过参数寻优的随机森林取得的 89.84%的分类精度，分别提升了 1.84%和 3.21%。这初步说明了本文提出的改进随机森林方法的可行性、实用性与优越性。其中，默认改进随机森林相较于经过参数寻优的随机森林取得的分类精度优势充分说明，仅对分类器的参数进行微调并不能提升其泛化能力和稳定性，但从构建原理上开展系统优化能够有效解决这一问题。这充分说明了开展本研究的必要性。优化改进随机森林取得最显著的分类精度充分说明，同时开展参数微调和系统优化能够同步提升原分类器的分类性能，是值得考虑的集成优化方案。此外，表 8 至表 10 给出了 3 个改进随机森林（包含经过参数寻优的随机森林）在 4 种标签的数据上取得的 Precision、Recall 和 F_1-score。

表 8　3 个改进随机森林取得的 Precision

改进随机森林	4 种标签的数据			
	0	1	2	3
默认改进随机森林	0.93	0.93	0.92	0.91
优化改进随机森林	**0.95**	**0.94**	**0.94**	**0.92**
参数寻优随机森林	0.91	0.9	0.9	0.89

表 9　3 个改进随机森林取得的 Recall

改进随机森林	4 种标签的数据			
	0	1	2	3
默认改进随机森林	0.93	0.92	0.92	0.91
优化改进随机森林	**0.94**	**0.94**	**0.93**	**0.93**
参数寻优随机森林	0.9	0.89	0.9	0.89

表 10　3 个改进随机森林取得的 F_1-score

改进随机森林	4 种标签的数据			
	0	1	2	3
默认改进随机森林	0.93	0.92	0.92	0.91
优化改进随机森林	**0.94**	**0.94**	**0.93**	**0.92**
参数寻优随机森林	0.9	0.89	0.9	0.9

从表 8 至表 10 中可以看出，优化改进随机森林在 4 种标签的数据上取得最高的 Precision、Recall 和 F_1-score，较默认改进随机森林和参数寻优随机森林优势明显。这证明优化改进随机森林在数据集上取得的整体分类性能最显著。同时，默认改进随机森林取得的 Precision、Recall 和 F_1-score 均高于参数寻优随机森林取得的。这有效说明了本文提出的改进随机森林方法的可行性与优越性。并且，优化改进随机森林在 4 种标签的数据上取得的 Precision、Recall 和 F_1-score 同样最均衡和最稳定，彼此之间差异最小。这证明其在各标签的数据上取得的局部分类性能最显著且最稳定。同时，默认改进随机森林取得的 Precision、Recall 和 F_1-score 的差异较参数寻优随机森林取得的 Precision、Recall 和 F_1-score 的差异小。这有效说明了本文提出的改进随机森林方法的稳定性。综合来说，本文提出的改进随机森林方法能够同时提升随机森林的整体分类性能与局部分类性能，明显优于现有参数寻优方法。

3.5　综合对比

在本小节，本文寻找最新文献中用于滑坡预测的改进随机森林，包括贝叶斯超参数优化随机森林（BH-RF）[12]，贝叶斯优化随机森林组合卡尔曼滤波器（BORF-KF）[24]，以及级联林随机森林框架（CF-RF）[25]。本文在训练集上训练上述 3 个改进随机森林，并在测试集上测试它们的分类性能，统计它们取得的分类精度，以及它们在 4 种标签的数据上取得的 Precision、Recall 和 F_1-score，用于与 3.4 节中的优化改进随机森林进行对比。表 11 给出了 4 个改进随机森林取得的分类精度。表 12 至表 14 给出了 4 个改进随机森林取得的 Precision、Recall 和 F_1-score。

表 11　4 个改进随机森林取得的分类精度

改进随机森林	分类精度/%
BH-RF	90.36
BORF-KF	90.28
CF-RF	89.92
优化改进随机森林	93.05

从表 11 中可以看出，依然是本文构建的优化改进随机森林取得最高的分类精度为 93.05%，较 BH-RF、BORF-KF 和 CF-RF 取得的 90.36%、90.28%和 89.92%的分类精度，分别高出 2.69%、2.77%和 3.13%，优势明显。值得说明的是，3.4 节中构建的默认改进随机森林取得的分类精度为 91.68%，也高于 BH-RF、BORF-KF 和 CF-RF 取得的分类精度。这充分说明本文提出的改进随机森林方法的优越性，在数据集上取得了最好的整体分类性能。

实际上，BH-RF 也是一种参数寻优的改进随机森林方法，BORF-KF 和 CF-RF 都是将随机森林与其他分

类器组合的改进随机森林方法。如引言中所述，这些方法专注于在现有分类器基础上进行优化，尤其是在参数调优上进行大量探索，或使用增强学习对现有模型进行微调，都未能从随机森林的构建原理出发来开展系统性优化研究。这充分证明了开展本研究的正确性和必要性。

此外，相较于 3.3 节中 kNN、SVM、决策树、MLP 和 1D-CNN 取得的分类精度，优化改进随机森林、默认改进随机森林、BH-RF、BORF-KF 和 CF-RF 取得的分类精度具有明显优势，前者分类性能最优的支持向量机取得的 88.03%的分类精度明显低于后者分类性能最差的 CF-RF 取得的 89.92%的分类精度。这充分展示了随机森林这一基于集成学习的经典的分类器用于数据分类的优越性，也再次证明了在 3.3 节中选择随机森林为基础分类器的正确性。

表 12 4 个改进随机森林取得的 Precision

改进随机森林	4 种标签的数据			
	0	1	2	3
BH-RF	0.92	0.9	0.9	0.88
BORF-KF	0.91	0.9	0.89	0.88
CF-RF	0.92	0.91	0.89	0.88
优化改进随机森林	0.95	0.94	0.94	0.92

表 13 4 个改进随机森林取得的 Recall

改进随机森林	4 种标签的数据			
	0	1	2	3
BH-RF	0.91	0.9	0.9	0.89
BORF-KF	0.91	0.9	0.89	0.89
CF-RF	0.91	0.9	0.9	0.89
优化改进随机森林	0.94	0.94	0.93	0.93

表 14 4 个改进随机森林取得的 F_1-score

改进随机森林	4 种标签的数据			
	0	1	2	3
BH-RF	0.91	0.9	0.9	0.88
BORF-KF	0.91	0.9	0.89	0.88
CF-RF	0.91	0.9	0.89	0.88
优化改进随机森林	0.94	0.94	0.93	0.92

从表 12 至表 14 中可以看出，本文构建的优化改进随机森林取得最高的 Precision、Recall 和 F_1-score，明显高于 BH-RF、BORF-KF 和 CF-RF 取得的 Precision、Recall 和 F_1-score。并且，优化改进随机森林在 4 种标签的数据上取得的 Precision、Recall 和 F_1-score 同样最均衡和最稳定，彼此之间差异最小。这再次说明了本文构建的优化改进随机森林在各标签的数据上取得最显著且最稳定的局部分类性能，展示其相较于 BH-RF、BORF-KF 和 CF-RF 的局部分类性能优势。综合来说，相较于现有文献提出的用于滑坡预测的改进随机森林，本文构建的优化改进随机森林的整体分类性能与局部分类性能均更优。本文提出的改进随机森林方法的可行性与正确性被证明，其可行性、实用性、稳定性与优越性得到充分证明。

3.6 性能评估指标

在开展本研究的过程中，需要使用性能评估指标对分类器的分类性能进行评估。本文使用的性能评估指标主要包括分类精度、Precision、Recall 和 F_1-score[26]。

数据集为 $D=\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，其中 y_i 是数据 x_i 的真实标签，$R(x_i)$ 是分类器 R 给出的预测标签。分类精度表示为标签预测正确的数据的数量占总数据数量的比例，计算公式如下：

$$\mathrm{acc}(R;D)=\frac{1}{n}\sum_{i=1}^{n} I[R(x_i)=y_i] \tag{4}$$

式中，I 是指示函数，当 $R(x_i)=y_i$ 时，$I[R(x_i)=y_i]=1$。

此外，Precision、Recall 和 F_1-score 能够评估分类器在数据集中某个标签的数据上的局部分类性能。其中，Precision 表示预测为正类的数据中有多少比例的数据是真正的正类。Recall 表示数据中有多少比例的正类数据被正确预测。F_1-score 是两者的有机结合。表 15 给出了二分类结果混淆矩阵[27]。*Precision*、*Recall* 和 F_1-score 的计算公式如下：

$$Precision=\frac{TP}{TP+FP} \tag{5}$$

$$Recall=\frac{TP}{TP+FN} \tag{6}$$

$$F_1\text{-score}=2\cdot\frac{P\cdot R}{P+R} \tag{7}$$

表 15 二分类结果混淆矩阵

真实类别	预测类别	
	正类	反类
正类	TP	FN
反类	FP	TF

4 结束语

本研究针对地下工程滑坡预测中传统方法泛化能力不足、现有机器学习分类器稳定性差的问题，提出了一种基于相关性度量的改进随机森林方法。通过融合真实场景数据采集、多分类器性能对比与决策树冗余性优化，构建了兼顾全局与局部分类性能的预测模型。实验表明，改进随机森林在精度与稳定性上均显著超越传统方法及现有改进方法（如 BH-RF、BORF-KF 等），其分类指标均衡性为复杂场景下的滑坡风险分级提供了可靠依据。实际案例中，该方法成功应用于中国南方某省份的军用地下工程滑坡预警，准确识别潜在风险区域，

验证了其工程实用价值。

未来研究可从三方面进一步拓展：其一，探索多源异构数据（如地质雷达、微震监测）的融合机制，以增强模型对隐蔽性滑坡特征的捕捉能力；其二，结合在线学习与增量训练优化模型动态适应性，满足地下工程实时预测需求；其三，将改进框架扩展至泥石流、塌陷等其他地质灾害预测领域，推动智能化防灾体系的跨场景应用。此外，如何降低高精度传感器的部署成本、提升算法在边缘计算设备上的运行效率，也将是工程落地的重要研究方向。

参考文献：

[1] 李守雷，梁为群，陈晓斌，等. 城市地下空间安全监测与预警指标研究 [J]. 地质与勘探，2024，60 (1)：95－104.

[2] 唐辉明. 重大滑坡预测预报研究进展与展望 [J]. 地质科技通报，2022，41 (6)：1－13.

[3] 蒋　树，王义锋，刘　科，等. 滑坡灾害空间预测方法研究综述 [J]. 人民长江，2017，48 (21)：67－73.

[4] WANG C H，ZHAO Y J. Time series prediction model of landslide displacement using mean-based low-rank autoregressive tensor completion [J]. Applied Sciences-Basel，2023，13 (8)：5214.

[5] XUE Z H，FENG W K，YI X Y，et al. Integrating data-driven and physically based landslide susceptibility methods using matrix models to predict reservoir landslides [J]. Advances in Space Research，2024，73 (3)：1702－1720.

[6] NIU H T，SHAO S J，GAO J Q，JING H. Research on GIS-based information value model for landslide geological hazards prediction in soil-rock contact zone in southern Shaanxi [J]. Physics and Chemistry of the Earth，2024，133：103515.

[7] WANG J W，LIU Y L，ZHANG G C，et al. Reservoir landslide displacement prediction under rainfall based on the ILF-FFT method [J]. Bulletin of Engineering Geology and the Environmen，2023，82 (5)：179.

[8] 黄发明，胡松雁，闫学涯，等. 基于机器学习的滑坡易发性预测建模及其主控因子识别 [J]. 地质科技通报，2022，41 (2)：79－90.

[9] ZHANG A M，WANG X M，PEDRYCZ W，et al. Near real-time spatial prediction of earthquake-triggered landslides based on global inventories from 2008 to 2022 [J]. Soil Dynamics and Earthquake Engineering，2024，185：108890.

[10] FANG L，YUE J P，XING Y. Research on landslide displacement prediction based on DES-CGSSA-BP model [J]. Processes，2023，11 (5)：1559.

[11] MA J，YANG Q，ZHANG M Z，et al. Data-driven deformation prediction of accumulation landslides in the middle Qinling-Bashan mountains area [J]. Water，2024，16 (3)：464.

[12] WANG S B，ZHUANG J Q，ZHENG J，et al. Application of bayesian hyperparameter optimized random forest and XGBoost model for landslide susceptibility mapping [J]. Frontiers in Earth Science，2021，9：712240.

[13] CHENG Y S，YU T T，SON N T. Random forests for landslide prediction in Tsengwen river watershed，central Taiwan [J]. Remote Sensing，2021，13 (2)：199.

[14] WU X Y，SONG Y B，CHEN W，et al. Analysis of geological hazard susceptibility of landslides in muli county based on random forest algorithm [J]. Sustainability，2023，15 (5)：4328.

[15] 邬礼扬，曾韬睿，刘谢攀，等. 基于集成学习建模的滑坡易发性评价 [J]. 地球科学，2024，49 (10)：3841－3854.

[16] 方匡南，吴见彬，朱建平，等. 随机森林方法研究综述 [J]. 统计与信息论坛，2011，26 (3)：32－38.

[17] SUN Z G，GAO M M，JIANG A P，et al. Incomplete data processing method based on the measurement of missing rate and abnormal degree：Take the loose particle localization data set as an example [J]. Expert Systems with Applications，2023，216：119411.

[18] 孙志刚，王国涛，高萌萌，等. 基于 kNN 优化算法的密封电子设备多余物定位技术 [J]. 电子测量与仪器学报，2021，35 (3)：94－104.

[19] 孙志刚，王国涛，高萌萌，等. 参数优化支持向量机的密封电子设备多余物定位方法研究 [J]. 电子测量与仪器学报，2021，35 (8)：162－174.

[20] 栾丽华，吉根林. 决策树分类技术研究 [J]. 计算机工程，2004，(9)：94－96.

[21] 吕红燕，冯　倩. 随机森林算法研究综述 [J]. 河北省科学院学报，2019，36 (3)：37－41.

[22] 张　婧，周怡欣，胡　涵，等. 基于知识采纳模型和多层感知机神经网络的评论有用性识别研究 [J]. 中国管理科学，2022，30 (4)：264－274.

[23] 曲建岭，余　路，袁　涛，等. 基于一维卷积神经网络的滚动轴承自适应故障诊断算法 [J]. 仪器仪表学报，2018，39 (7)：134－143.

[24] ZHANG N F，ZHANG W，LIAO K，et al. Deformation prediction of reservoir landslides based on a Bayesian optimized random forest-combined Kalman filter [J]. 2022，81：197.

[25] CHEN S J，PAN Y T，LU C D，et al. Landslide spatial prediction based on cascade forest and stacking ensemble learning algorithm [J]. International Journal of Systems Science，2024，Early Access.

[26] 李旭然，丁晓红. 机器学习的五大类别及其主要算法综述 [J]. 软件导刊，2019，18 (7)：4－9.

[27] 杨剑锋，乔佩蕊，李永梅，等. 机器学习分类问题及算法研究综述 [J]. 统计与决策，2019，35 (6)：36－40.