doi: 10.12097/gbc.2023.11.008

基于优化随机森林模型的降雨群发滑坡易发性评价
——以西秦岭极端降雨事件为例

刘帅[1,2,3]，王涛[1,2,3]*，曹佳文[4]*，刘甲美[1,2,3]，张帅[1,2,3]，辛鹏[1,2,3]
LIU Shuai[1,2,3], WANG Tao[1,2,3]*, CAO Jiawen[4]*, LIU Jiamei[1,2,3], ZHANG Shuai[1,2,3], XIN Peng[1,2,3]

1. 中国地质科学院地质力学研究所, 北京 100081;
2. 自然资源部活动构造与地质安全重点实验室, 北京 100081;
3. 自然资源部陕西宝鸡地质灾害野外科学观测研究站, 陕西 宝鸡 721001;
4. 中国地质调查局, 北京 100037
1. *Institute of Geomechanics*, *Chinese Academy of Geological Sciences* , *Beijing* 100081, *China;*
2. *Key Laboratory of Active Tectonics and Geological Safety*, *Ministry of Natural Resources*, *Beijing* 100081, *China;*
3. *Observation and Research Station of Geological Disaster in Baoji*, *Shaanxi Province*, *Ministry of Natural Resources*, *Baoji* 721001, *Shaanxi*, *China;*
4. *China Geological Survey*, *Beijing* 100037, *China*

摘要: 随机森林模型（RF）是在滑坡易发性评价中广泛应用的机器学习模型之一。针对制约随机森林模型评价应用质量的难点问题，以西秦岭山区娘娘坝镇极端降雨诱发的 2 万余处群发滑坡为例，从滑坡−非滑坡样本筛选方法、影响因子选取、联结方法应用和超参数优化 4 个方面开展了模型优化及与常规模型评价的对比研究。通过区域滑坡易发性评价和有效性比较可知，2 种情形评价均取得理想结果，优化随机森林评价结果 AUC（精度曲线下的面积）可达 0.877，对比常规评价结果更优，表明该优化方法可以明显提升随机森林模型在区域降雨滑坡评价中的效果和学习效率，可为气候变化背景下极端降雨群发滑坡灾害易发性评估提供参考。

关键词: 滑坡；易发性；随机森林；极端降雨；西秦岭
中图分类号: P642.22; TP181 **文献标志码:** A **文章编号:** 1671−2552(2024)06−0958−13

Liu S, Wang T, Cao J W, Liu J M, Zhang S, Xin P. Susceptibility assessment of precipitation-induced mass landslides based on optimal random forest model: Taking the extreme precipitation event in western Qinling mountains as an example. *Geological Bulletin of China*, 2024, 43(6): 958−970

Abstract: Random forest model (RF) is one of the widely used machine learning models for landslide susceptibility assessment. Aiming at the difficult problems that restrict the application quality of random forest model assessment, taking more than 20000 extreme rainfall landslides induced by extreme rainfall in Niangniangba Town, western Qinling Mountains as an example, the model optimization and comparison with conventional model evaluation were carried out mainly from four aspects: landslide−non−landslide

收稿日期: 2023−11−06; **修订日期:** 2023−12−10
资助项目: 中国地质调查局项目《全国地质灾害风险区划技术方法研究》(编号: DD20221738)、甘肃省地质环境监测院《典型乡镇地质灾害风险定量评估关键技术研究》和自然资源部科技人才工程资助项目《气候变化、构造活跃及人类活动加剧背景下地质灾害风险减轻策略研究》(编号: 121106000000180039-2207)
作者简介: 刘帅(1997−), 男, 在读博士生, 从事地质灾害危险性评价研究。E−mail: 1048485867@qq.com
* **通信作者:** 王涛(1982−), 男, 研究员, 从事地质灾害致灾机理与风险评估防控研究。E−mail: wangtaoig@cags.ac.cn
曹佳文(1979−), 男, 高级工程师, 从事地质灾害、城市地质及重大工程地质安全风险等管理工作。E−mail: cjiawen@mail.cgs.gov.cn

sample screening method, influence factor selection, coupling method application and hyper−parameter optimization. Based on the above optimization, the regional landslide susceptibility evaluation and effectiveness comparison of typical towns−Niangniangba Town are carried out. The evaluation of both situations has achieved ideal results. The optimized random forest evaluation result AUC can reach 0.877, which is better than the conventional assessment results. It shows that the optimization method can obviously improve the assessment effect and learning efficiency of random forest model in regional rainfall landslide, and can provide reference for the risk assessment of extreme rainfall landslide hazard under the background of climate change.

Key words: landslide; susceptibility; random forest; extreme rainfall; West Qinling

滑坡灾害分布范围广、发生频率高、灾害损失程度严重，极大地威胁着人类生命财产安全(许强等, 2004; Khan et al., 2021)。滑坡易发性评价是滑坡风险识别和预测评价的重要基础(Havenith et al., 2015)。尤其是在气候变化背景下，近年来极端降雨群发灾害频发多发，例如 2021 年 7 月 20 日河南郑州特大暴雨灾害、2023 年 7 月 29 日北京门头沟特大暴雨灾害等，均引发大规模山体滑坡等地质灾害，造成严重的人员伤亡和财产损失。因此，面向事件的评价研究对重大灾难事件复盘和未来预测具有重要意义(Alvioli et al., 2018; 殷跃平等, 2024)。

数据驱动模型因其鲁棒性、准确性最佳，在滑坡易发性评价中被广泛应用，先后经历了数理统计、机器学习 2 个阶段(Ado et al., 2022)。信息量等数理统计模型通常依赖于大量先验知识，且运算程度较低，未能充分挖掘利用精度较高的数据，存在一定的局限性。机器学习模型，在海量数据积累和算力提升背景下，因存在数据量较大的运算优势并能获得较高的预测精度，且不需要过多先验知识而被广泛应用于滑坡易发性评价预测中，其成功率和有效性明显高于传统数理统计，应用愈发广泛，主要包括支持向量机 (Yao et al. 2008)、随机森林 (吴润泽等 2021; Deng et al., 2022)、深度学习模型(Bui et al., 2020)等机器学习模型。在现有机器学习模型中，随机森林使用广泛、效果更好(Sun et al., 2020; Huang et al., 2022)，在众多研究中对比支持向量机等其他机器学习模型能得到较高精度的模型预测结果 (Guilherme et al., 2019)。然而，常规的机器学习模型目前存在许多问题及疑点，不能较好地实现非滑坡样本的选择 (Yao et al., 2022)，降低了机器学习自身的“学习”效率；模型的可解释性未得到足够的深化(Jie et al., 2020)。传统的因子离散性分级方式在模型中的应用并未发挥机器学习的运算优势，一定程度上导致部分信息丢失从而降低模型性能(郭飞等, 2022)。超参数保留默认参数或以多次重复试验的方式未能客观体现适用于不同地质环境条件下超参数设置的合理性(周晓亭, 2022)，导致模型精确率降低。因此，对于机器学习模型在滑坡易发性评价中的优化研究十分必要。

针对上述机器学习模型存在的问题及疑点，本文选取西秦岭山区娘娘坝镇地区为例，开展模型优化和评价验证研究，探索基于滑坡的面状矢量数据，建立缓冲区，确定非滑坡选取区域，改进非滑坡样本筛选方法；利用 Pearson 相关性分析与随机森林基尼系数进行因子筛选择优，拟解决现存信息冗余问题；从影响因子定量表征及归一化处理方面，探索传统离散性分级存在的信息丢失问题解决方法；基于遗传算法对随机森林超参数自适应寻优，探寻进一步提升模型自身属性的有效方法，将系统性的优化随机森林模型应用于降雨群发滑坡易发性评价中，并选择常规随机森林模型评价结果进行对比，以期优化随机森林模型，并提升其在滑坡易发性评价中的效果。

1　西秦岭极端事件及数据准备

1.1　研究区概况

研究区为西秦岭山区 2013 年天水市“7·25”群发暴雨地质灾害事件发生区，综合考虑人员密集和群发灾害集中分布的 206 km^2 范围地段，区内发生大量浅表层流滑灾害，孕灾条件具有代表意义。研究区娘娘坝镇位于秦州区南部，地处西秦岭南麓，气候湿润，雨量充足，地形切割强烈，沟谷发育，地形坡度大，最高海拔 2169 m，最低海拔 687 m(图 1)。

1.2　数据准备

1.2.1　滑坡编录

滑坡编录数据是进行滑坡易发性评价的基础资料，编录数据的完整性与准确性对滑坡易发性评价具有重要意义。2013 年 6 月 19 日—7 月 26 日，尤其是 7 月 25 日天水市出现大范围强降雨天气，造成了大面积、群发性地质灾害，持续强降雨造成该镇

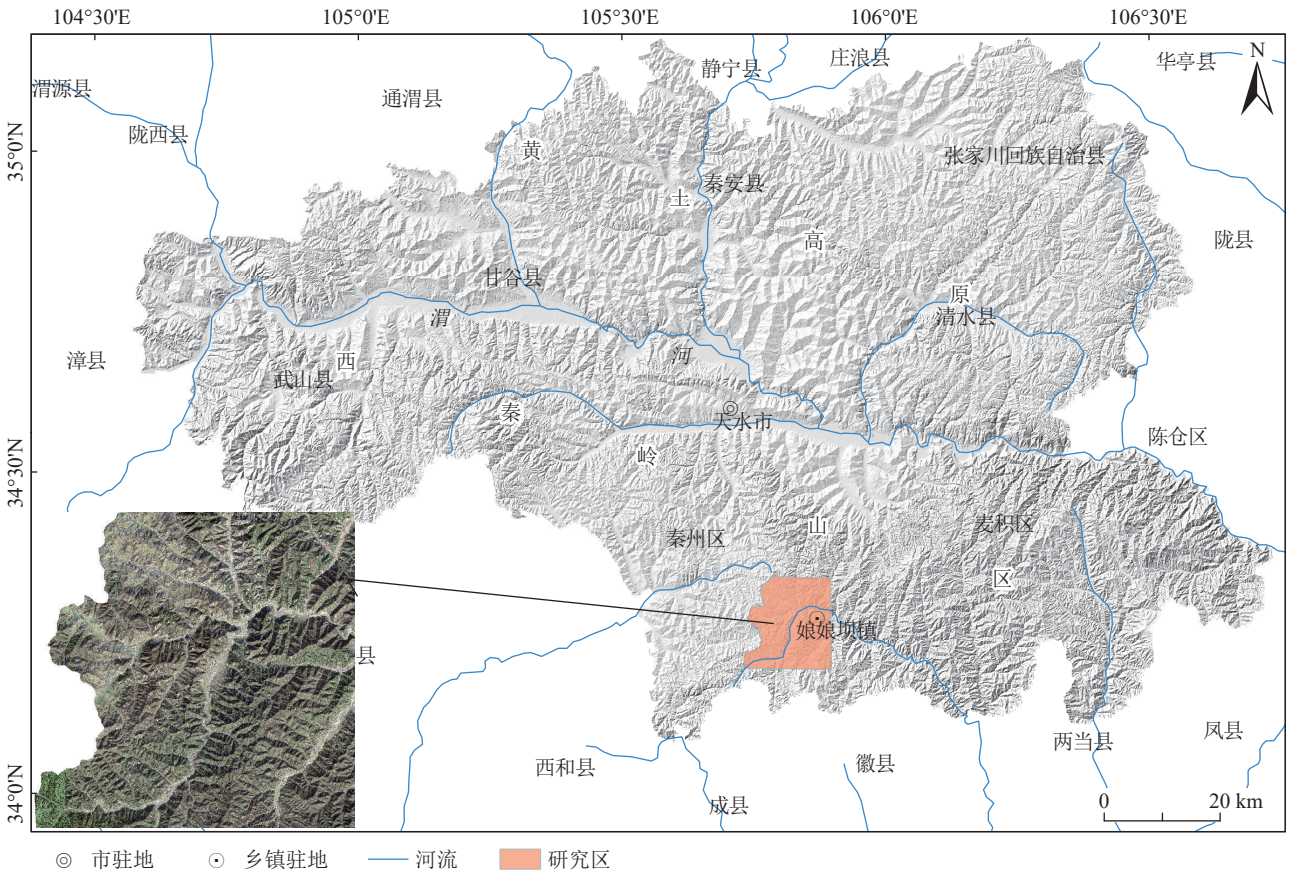

图 1 研究区地理位置图

Fig. 1 Geographical map of the study area

3439 户居民受灾，42 名群众死亡(郭富赟等, 2015; 黄森, 2021)。利用分辨率为 0.5 m 的灾后 Pleiades 遥感影像，对研究区“7·25”群发滑坡灾害通过人工解译、机器学习分类等方式进行遥感解译(图 2)，共获取滑坡 20362 处，总滑坡面积 8.25 km^2，占研究区总面积的 4.1%，规模以小型为主，类型主要为新近堆积黄土流滑和残坡积碎石土滑坡。

1.2.2 影响因子

降雨滑坡的形成主要受降雨量、地形地貌、地层岩性等因素的综合影响(唐辉明, 2018; Ahmad et al., 2023)。鉴于各地孕灾条件的差异，各类因素对滑坡发育的贡献程度不同。其中，地层岩性为地质灾害发育提供了物质基础 (Abbas et al., 2023)，地形地貌控制着地质灾害的空间边界条件 (Achu et al., 2023)，河流指示坡脚侵蚀及坡体的水文地质特征，增加斜坡的不稳定性 (Li et al., 2023)，地质构造既控制地形地貌，又控制岩体结构及其组合特征，对地质灾害的发育起综合控制影响作用 (刘帅等, 2023)，降雨则是滑坡的有效孕灾条件(María et al., 2023)。根据研究区域滑坡特征及相关研究成果(Reichenbach et al., 2018)，本文初步选取前 3 日有效降雨、高程、坡度等 17 个影响因子作为滑坡易发性评价输入变量，进行单因素分析。详细数据及来源见表 1。

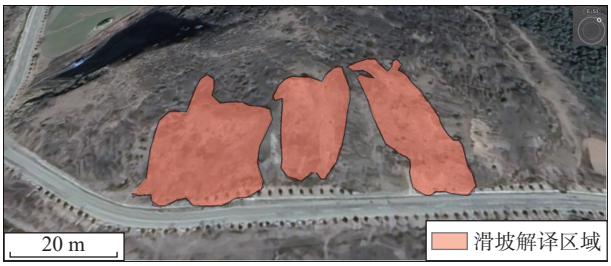

图 2 典型滑坡遥感解译结果

Fig. 2 Remote sensing interpretation results of typical landslide

2 随机森林模型优化方法

随机森林(Redom Forest, 简称 RF)是一种用于分类和回归的监督学习算法，在数据样本上建立决策树，然后从每个样本中得到预测结果，最后通过投

表 1　影响因子数据来源统计结果

Table 1　Statistical results of influence factor data sources

分类	序号	要素	致灾指示意义、数据说明及处理方法
地形地貌	1	坡度	指示坡体属性，ALOS DEM(12.5 m分辨率)
	2	坡向	
	3	高程	
	4	平面曲率	
	5	剖面曲率	
	6	起伏度	
地层岩性	7	地层岩性分区	指示斜坡岩土体的物理力学强度特征, 基于1∶20万地质图(公开版)
河流	8	距一级河流距离	指示坡脚侵蚀及坡体的水文地质特征, 使用1∶25万基础地理信息数据
	9	距二级河流距离	
道路	10	距一级道路距离	指示人类工程活动对斜坡破坏的影响, 使用1∶25万基础地理信息数据
	11	距二级道路距离	
前期有效降雨	12	前3日有效降雨	指示降雨有效入渗对滑坡发育的影响, 通过气象台站降雨数据插值获取, 1 km栅格
地质构造	13	距主要断层距离	指示地质构造对地质灾害的形成发展的影响, 基于1∶20万地质图(公开版)
环境地质特征	14	土地利用类型	指示影响斜坡发育的环境地质特征, 基于Landsat8遥感影响数据, 30 m
	15	修正归一化水体指数(MNDWI)	
	16	归一化建筑指数(NDBI)	
	17	归一化植被指数(NDVI)	

票的方式选择最优解。它是一种比单一决策树更好的集成方法, 通过对结果进行平均减少过拟合, 从而提高预测性能(Youssef et al., 2016)。具体运算流程见图 3。

为提升模型预测精度, 本文在样本筛选策略、影响因子选取、联接方法应用、超参数寻优 4 个方面进行全面优化。为保证易发性评价结果的成功率与预测率, 本文涉及的优化 RF 与常规 RF 模型均采用相同优化后输入样本数据集, 最终通过 ROC 曲线与数据统计分析进行评价结果对比分析。具体技术流程见图 4。

2.1　样本筛选方法

建立缓冲区的方法可有效避免非滑坡样本选取落入滑坡区域, 确保选取的非滑坡为真正的“非滑坡”。本文基于光学遥感解译的滑坡面状矢量数据, 为防止选取的非滑坡落入滑坡边界区域, 在 GIS 平台中以滑坡面为基准建立 50 m 缓冲区, 确定非滑坡选取区域, 以随机抽样的方式, 选取等量非滑坡数据, 共同组成滑坡−非滑坡建模样本(图 5)。该方法克服了非滑坡样本错选问题, 相比以往以点状滑坡数据筛选非滑坡样本有更高的准确性。

2.2　随机森林 *Gini* 系数

随机森林算法在选择属性时采用 *Gini* 系数, 表征各影响因子的特征重要性, 即对滑坡发生的贡献程度, 并采用二分递归分割技术生成结构简洁的二叉树作为影响因子选取依据, 避免了因子选取的盲目性和主观性。给定影响因子数据集 D, 计算 D 中每个因子的 *Gini* 系数(Wen et al.,2017)。

数据集 *Gini* 系数度量:

$$Gini(D) = 1 - \sum_{i=1}^{K} p_i^2 \qquad (1)$$

式中: K 表示影响因子类别个数; p_i 表示数据集中第 i 类因子所占比例。

计算因子 A 的 *Gini* 系数时, 根据 A 因子的取值, 将数据集分割为 $D1$,$D2$ 两个子集。计算每个取值对应的 *Gini* 系数。经加权后得到 A 因子的 *Gini* 系数。

$$Gini(D,A) = \frac{|D1|}{|D|}Gini(D1) + \frac{|D2|}{|D|}Gini(D2) \qquad (2)$$

2.3　Pearson 相关性分析

Pearson 相关性分析适用于服从正态分布的两定量型变量, 若两变量通过绘制散点图后发现存在线性趋势, 可以通过计算 Pearson 相关系数描述两变量

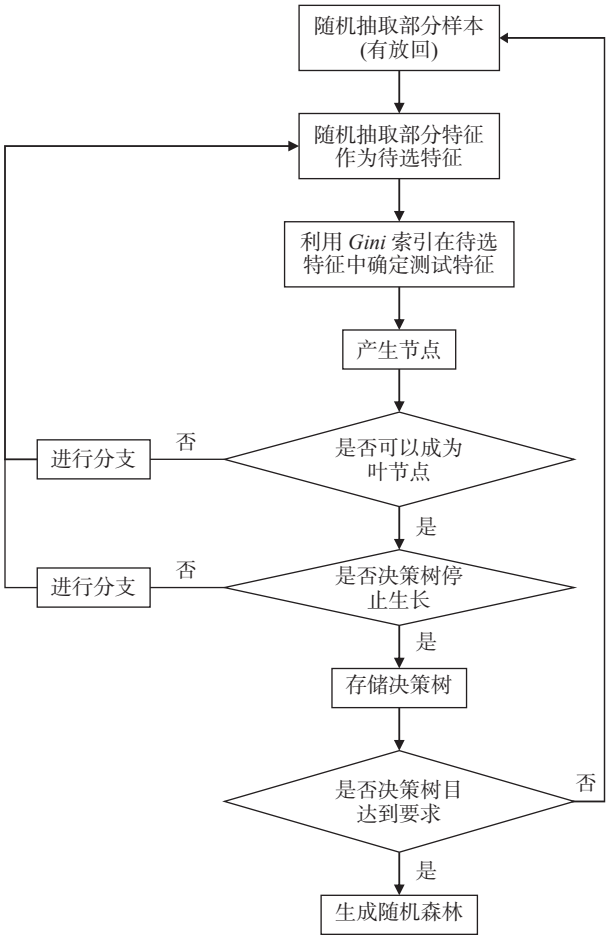

图 3 随机森林算法流程图

Fig. 3 Flow chart of RF algorithm

的线性相关性(Danika et al., 2019),并据此解决各影响因子自相关问题,避免了参数选取时有关因素重复考虑带来的信息冗余。

2 个影响因子之间的 Pearson 相关系数定义为 2 个因子之间的协方差和标准差的商:

$$\rho X,Y = \frac{cov(X.Y)}{\sigma X \sigma Y} = \frac{E[(X-EX)(Y-EY)]}{\sigma X \sigma Y} \quad (3)$$

式中:ρ 为总体相关系数;X,Y 为比较的 2 个因子;cov 为协方差;E 为期望值;σ 为标准差。估算样本的协方差和标准差,可得到 Pearson 相关系数:

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \quad (4)$$

式中:r 为 Pearson 相关系数;X_i, Y_i 为比较的 2 个因子;$\overline{X}$, $\overline{Y}$ 分别为 2 个因子的平均值。

2.4 影响因子量化提取

本文的量化提取是基于微分理念,避免传统连续型因子分级离散化,提取各连续型因子的实际数值应用于数据样本统计分析,充分体现各因子数据的真实性;对于离散型因子则采用频率比的方式定量表征,降低其离散性。将数量化后的各影响因子作归一化处理,消除不同因子间的量纲影响,使数据处于同一数量级,保证数据处理的方便和程序运行时的收敛加快。

其中,频率比(Tareq et al., 2011)计算公式如下:

$$FR_{ij} = \frac{N_{ij}/N}{S_{ij}/S} \quad (5)$$

式中:i 为影响因子;j 为因子下的分级;FR_{ij} 为第 i 个因子,第 j 个等级的频率比;N_{ij} 为第 i 个因子,第 j 个等级的滑坡数;N 为研究区的滑坡总数;S_{ij} 为第 i 个因子,第 j 个等级的栅格数;S 为研究区的总栅格数。

归一化(Du et al., 2017)计算公式如下:

$$Y = \frac{X-X_{min}}{X_{max}-X_{min}} \quad (6)$$

式中:Y 为求取的归一化值;X 为被归一化的某一个值;X_{min} 为数据集中的最小值;X_{max} 为数据集中的最大值。

2.5 遗传算法超参数优化

遗传算法是自适应启发式搜索算法,属于进化算法的一部分,是基于自然选择和遗传学的思想,利用随机搜索提供的历史数据,可以指导搜索到解决方案空间中性能更好的区域,通常用于为优化问题和搜索问题生成高质量的解决方案(Ke et al., 2012),避免了以往随机森林模型超参数设置的主观调试。具体算法流程见图 6。

3 滑坡易发性评价

3.1 输入样本建立

3.1.1 滑坡-非滑坡样本

以滑坡面为基础,建立 50 m 缓冲区从而反向确定非滑坡选取区域,以生成随机点的方式在非滑坡区域随机生成等量的非滑坡点与滑坡数据,共同组成 40724 个输入滑坡-非滑坡样本数据,参与统计计算与模型训练。

3.1.2 影响因子量化处理

初步选取的 17 个因子统一连续性量化及归一化处理,降低离散性的同时,也使所有影响因子归于同一数据维度,保证了机器学习的学习效率。各影响因子量化结果见表 2。

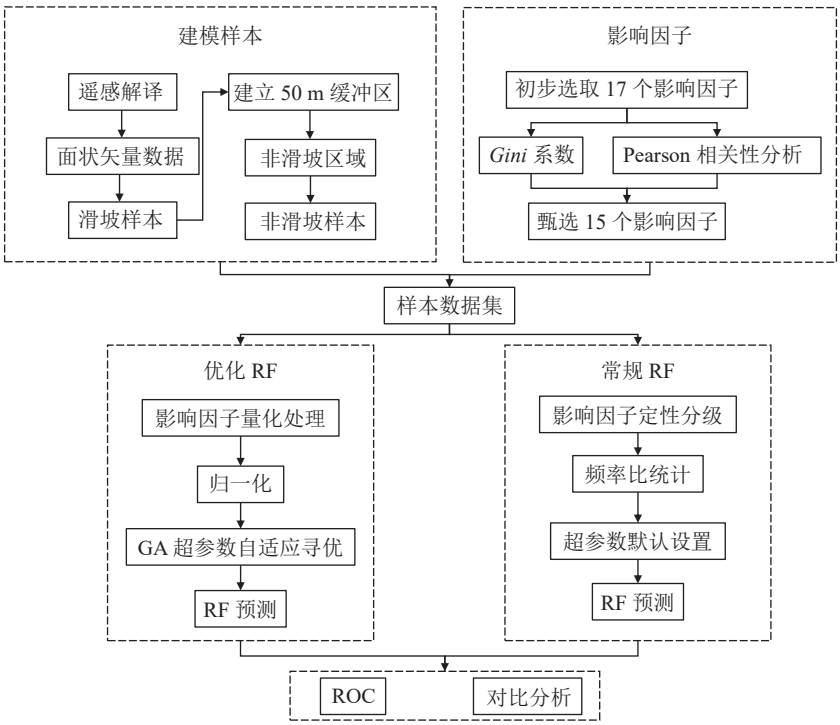

图 4　滑坡易发性评价技术流程图

Fig. 4　Technical flow chart of landslide susceptibility assessment

表 2　影响因子量化结果

Table 2　Quantitative results of impact factors

类型	序号	分类	要素	实际取值范围	归一化取值
连续性	1	地形地貌	坡度	0~81°	0 ~ 1
	2		坡向	0~360°	
	3		高程	1471~2030 m	
	4		平面曲率	−284~320	
	5		剖面曲率	−373~297	
	6		起伏度	0−171 m	
	7	河流	距一级河流距离	0~5594 m	
	8		距二级河流距离	0~2396 m	
	9	道路	距一级道路距离	0~4315 m	
	10		距二级道路距离	0~3819 m	
	11	前期有效降雨	前3日有效降雨	9−159 mm	
	12	地质构造	距主要断层距离	0~8047 m	
	13	环境地质特征	修正归一化水体指数(MNDWI)	−0.40~0.27	
	14		归一化建筑指数(NDBI)	−0.37~0.14	
	15		归一化植被指数(NDVI)	−0.04~0.70	
离散型	16	环境地质特征	土地利用类型	0~2.07(频率比)	
	17	地层岩性	地层岩性分区	0.09~1.50(频率比)	

3.1.3 影响因子甄选

对于初选的 17 个影响因子，结合 Pearson 相关性系数与随机森林 *Gini* 系数，最终剔除 2 个影响因子。从图 7 可以看出，由于归一化建筑指数(NDBI)与归一化植被指数(NDVI)的 Pearson 相关性系数较大，为−0.854，需剔除其一。同时，基于随机森林 *Gini* 系数对各因子进行滑坡贡献程度的特征重要性分析(图 8)，显示 NDBI 的特征重要性较低，因此剔除。另外，影响因子特征重要性分析结果得出，剖面曲率的特征重要性为 0，即对于滑坡发育剖面曲率未作出贡献，因此也剔除。最终甄选出高程、坡度、坡向、起伏度、平面曲率、地层岩性、距断层距离、距一级河流距离、距二级河流距离、距一级道路距离、距二级道路距离、土地利用类型、归一化植被指数(NDVI)、修正归一化水体指数(MNDWI)、前

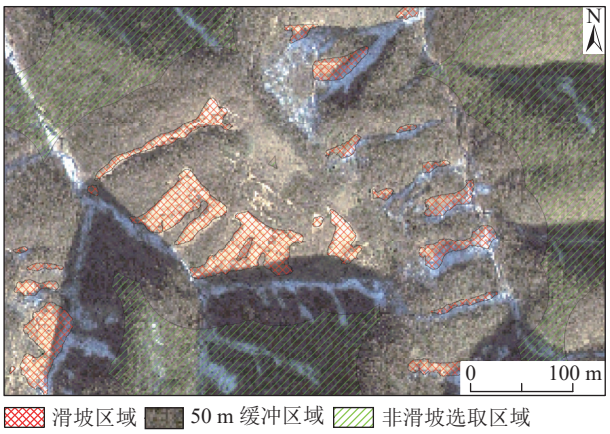

图 5 非滑坡选取区域示意图

Fig. 5 Non-landslide selection area diagram

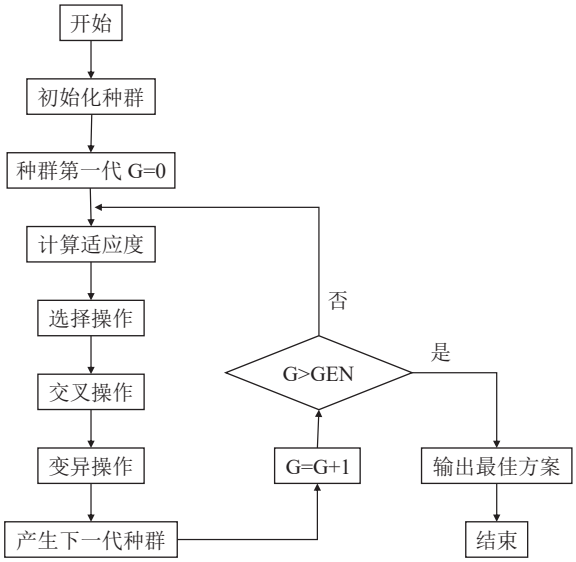

图 6 遗传算法流程图

Fig. 6 Flow chart of genetic algorithm

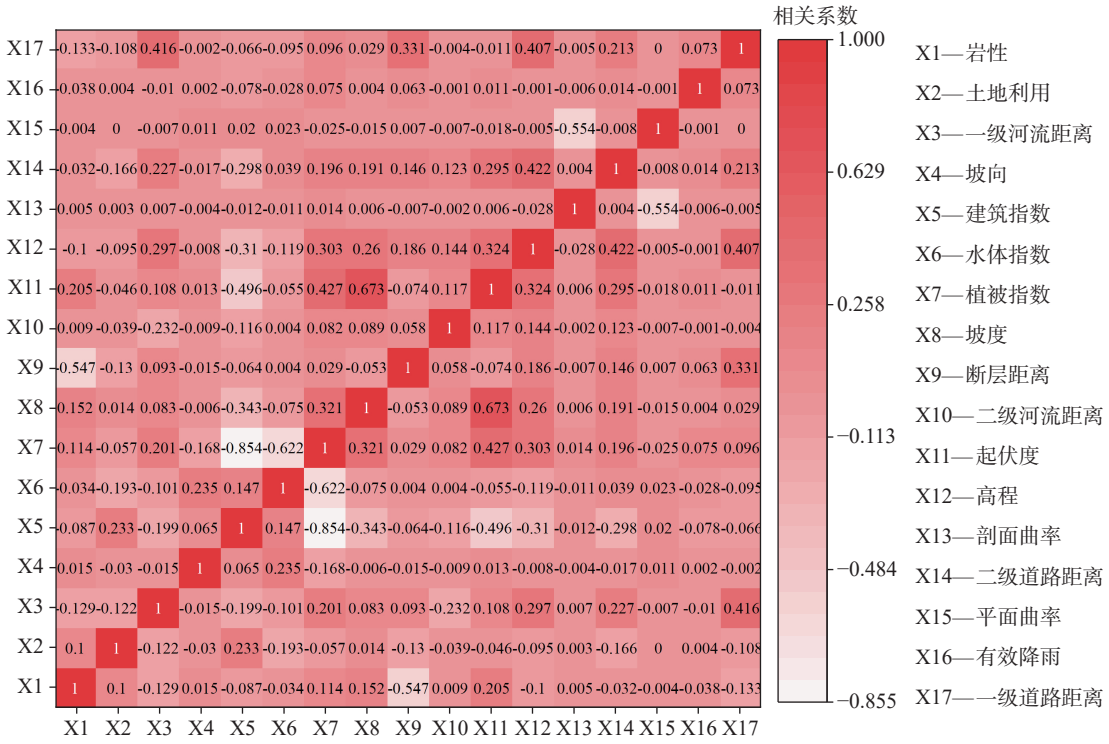

图 7 Pearson 相关系数热力图

Fig. 7 Pearson correlation coefficient heat map

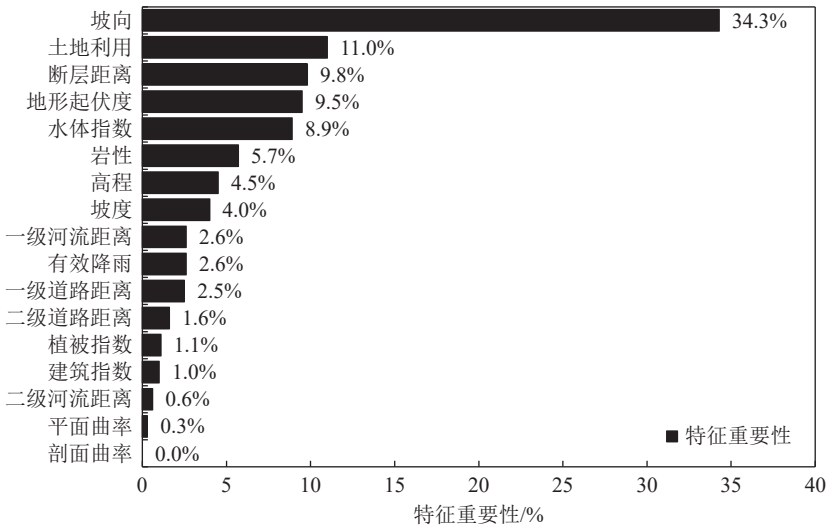

图 8　影响因子特征重要性

Fig. 8　Importance of impact factor features

3 日有效降雨 15 个影响因子参与研究区域易发性评价(图 9)。

3.2　随机森林模型训练结果

基于遗传算法的模型超参数自适应寻优前后，训练时长有明显变化，对比默认参数设置，寻优训练时长达到近 54 h，除决策树数量减少外，其余寻优参数的数量均有增加(表 3)。

断层距离在优化后由占比第五上升到占比第三的位置，比重由 8.6% 上升至 9.6%，其余基于 *Gini* 系数的影响因子特征重要性在优化前后各影响因子对滑坡贡献程度无明显变化(图 10)。

对比优化前后，随机森林模型训练精确率有明显提升(表 4)，F1 为召回率与精确率的调和平均，是模型训练结果的重要参考值。常规与优化随机森林训练结果均有较高的准确性，但优化后的训练结果提升至 0.827，具有更高的准确性。

表 3　优化前后超参数对比

Table 3　Comparison of hyperparameters before and after optimization

参数名	优化参数值	默认参数值
训练用时	53 h 42 min 58 s	2 s
决策树数量	50	100
内部节点分裂的最小样本数	50	2
叶子节点的最小样本数	48	1
树的最大深度	12	10
叶子节点的最大数量	101	50

3.3　滑坡易发性评价结果

将训练好的模型用于预测娘娘坝镇区域滑坡易发性，分别得到优化前后的滑坡易发性图(图 11)。此外，根据自然间断点法将研究区域划分为高、中、低、非 4 个易发性等级，并计算其相应面积与分区滑坡面积(表 5)。对比常规 RF 预测结果，优化 RF 预测结果高、中、低、非 4 个易发等级所占面积逐级递增；优化 RF 预测结果高、中易发区滑坡分布相对集中，相比常规 RF 预测结果，呈现更明显的规律性与合理性。在局部尺度上，以娘娘坝镇西南部的滑坡集中分布区为例，可以看出优化 RF 的易发性分区结果与实际蓝色滑坡区较一致，且优于常规 RF 的评价结果(图 12)。

3.4　ROC 曲线

AUC 表示 ROC 曲线下的面积，主要用于衡量模型的泛化性能(Isidro et al., 2019)。从 ROC 曲线(图 13)可以看出，模型易发性预测结果均有较高的准确性，优化 RF 在 AUC 的值为 0.877，优于常规 RF 预测结果。

本文优化与常规随机森林评价结果均有很高的准确率，因二者均建立在相同的滑坡-非滑坡输入样本的基础上，保证了随机森林评价结果的精度。常规随机森林评价结果已达到较高的预测率，而优化随机森林模型是在影响因子定量表征与超参数设置上进一步优化和完善的，结果精度有进一步提升，亦可证明本文优化方法的可行性和实际效果。

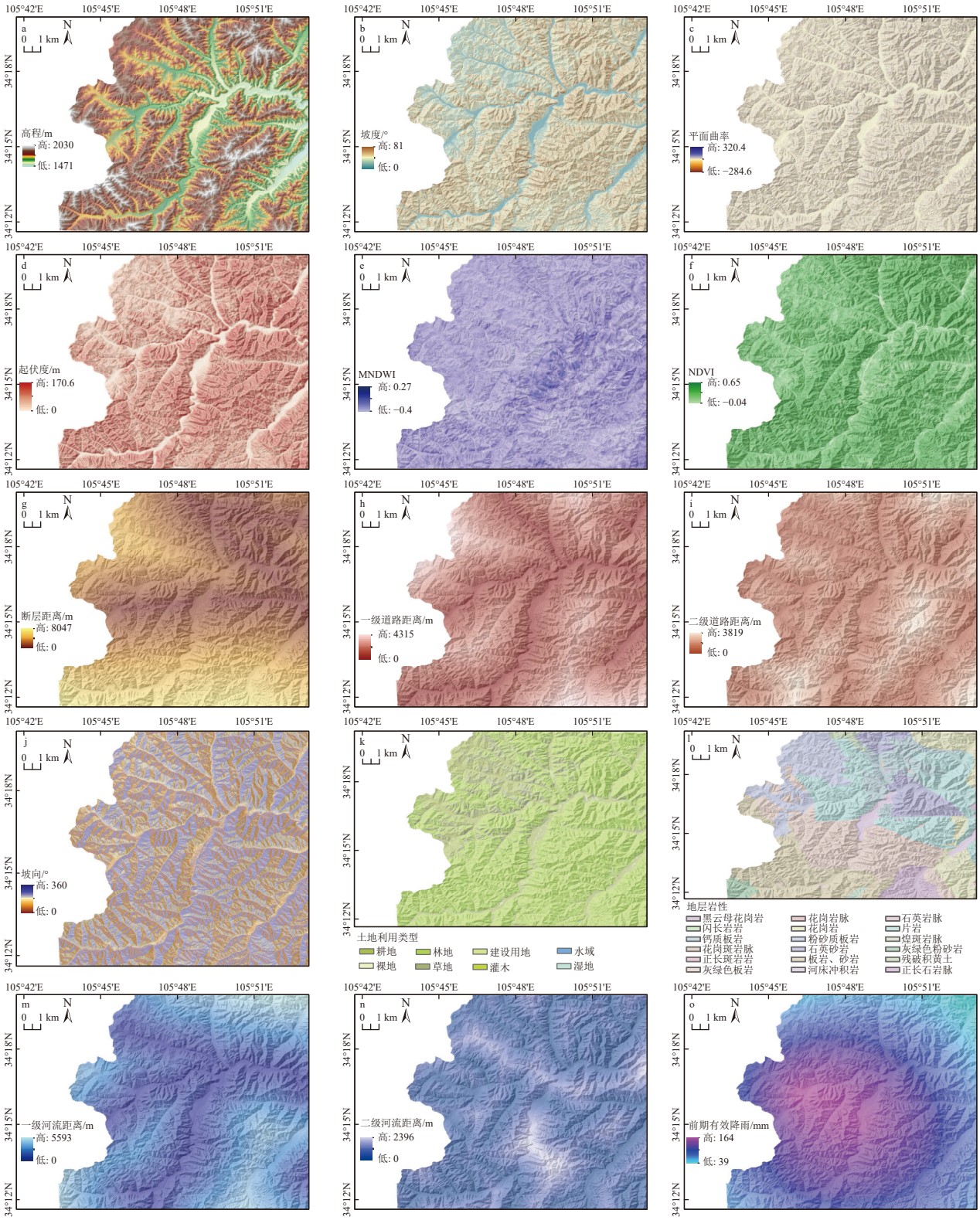

图 9 甄选影响因子分布图

Fig. 9 Selection of impact factor distribution map

a—高程; b—坡度; c—平面曲率; d—起伏度; e—MNDWI; f—NDVI; g—距断层距离; h—距一级道路距离; i—距二级道路距离; j—坡向; k—土地利用类型; l—岩性; m—距一级河流距离; n—距二级河流距离; o—前 3 日有效降雨

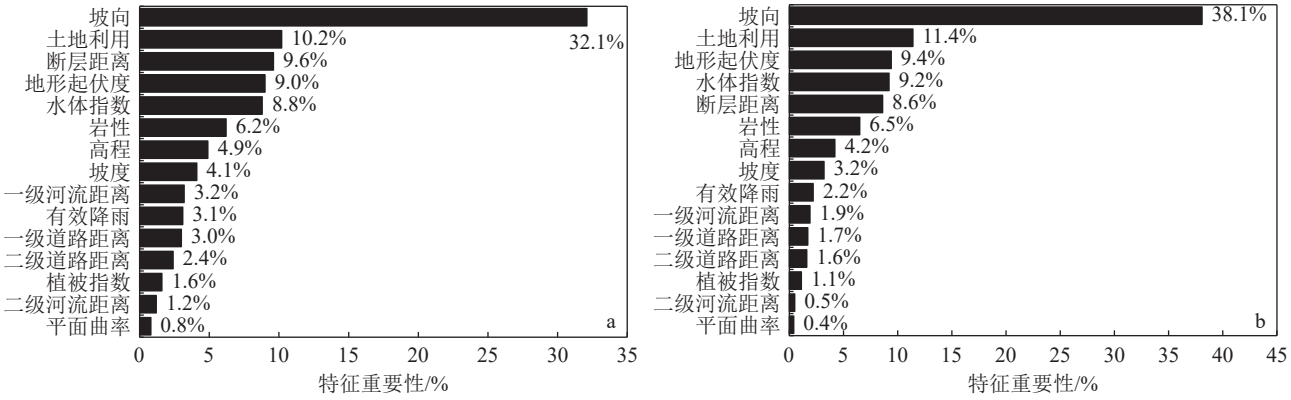

图 10　优化前后影响因子特征重要性对比

Fig. 10　Comparison of the importance of impact factor features before and after optimization

a—优化 RF; b—常规 RF

表 4　优化前后训练集结果对比

Table 4　Comparison of training set results before and after optimization

训练集	准确率	召回率	精确率	F1
优化RF	0.827	0.827	0.827	0.827
常规RF	0.794	0.794	0.794	0.794

4　讨　论

针对西秦岭天水地区, 前人运用线性回归、加权信息量等传统数据驱动模型开展区域滑坡易发性评价(孟晓捷等, 2022), 评价结果均取得较好的准确性, AUC 值为 0.75。在此基础上, Ma et al. (2023)基于 TRIGRS 模型分析了天水地区浅层黄土滑坡的稳定性, 确定了滑坡降雨阈值, 得出坡度条件对阈值影响显著; 李霞等(2023)结合多源数据, 共选取 13 个影响因子, 运用传统证据权法对天水地区进行滑坡易发性评价, 结果 AUC 值为 0.847, 评价结果较好。借鉴前人研究, 本文运用优化后机器学习模型, 结合因子筛选方法共选取 15 个影响因子参与滑坡易发性评价, 结果精度达到 0.877, 对比前人研究, 评价结果的准确性进一步提升。

随机森林等机器学习模型预测结果的优劣很大程度取决于输入学习样本的质量, 可以通过不同途径实现对样本质量的改善。本文非滑坡样本的选取是在保证样本科学完整性的滑坡面状数据的基础上作缓冲, 反向建立非滑坡随机选取区域; 有些研究则以滑坡点状数据为基础, 通过聚类分析(黄发明等, 2018)、基于划分滑坡占比网格的非均匀采样(Yang et al., 2023)等技术方法筛选出较准确的非滑坡样本数据集。由此可知, 从滑坡样本与非滑坡筛选方法 2 个角度优化均能提升模型精度。

机器学习具备强大的计算能力, 可实现基于微分理念的影响因子实际数值运算。传统数理统计模型中经验性的因子分级方式易造成关键控灾因子区间信息异化或丢失, 本文针对离散型因子, 采用频率比的形式对其定量表征, 实现了机器学习数据处理的方便和程序运行时的收敛加快。

本文基于遗传算法的随机森林模型, 超参数自适应寻优结果精度对比超参数默认设置的训练结果有所提升, 但对于遗传算法自身超参数也未作过多调试与研究, 导致训练结果精度虽有提升但效果不明显。

在海量数据积累与算力提升背景下, 继续深化机器学习在地质灾害评价中的应用具有实质性前景(郭飞等, 2023)。针对本文在天水地区评价结果、数据、方法等方面研究的优势与局限, 笔者认为: 在滑坡-非滑坡样本数据选取方面, 在构建滑坡面状矢量样本基础上, 运用准确率更高的非滑坡样本筛选方法不失为有前景的研究方向; 在联接方法应用上, 探索不同离散型因子的定量表征方式, 降低数据的离散性, 使其尽可能契合机器学习运算机理, 是影响因子处理应用的关键; 在模型优化上, 适用于随机森林等机器学习模型的超参数自适应寻优算法在不牺牲模型精度的前提下, 自身参数调试优化是下一步超参数寻优的研究方向。

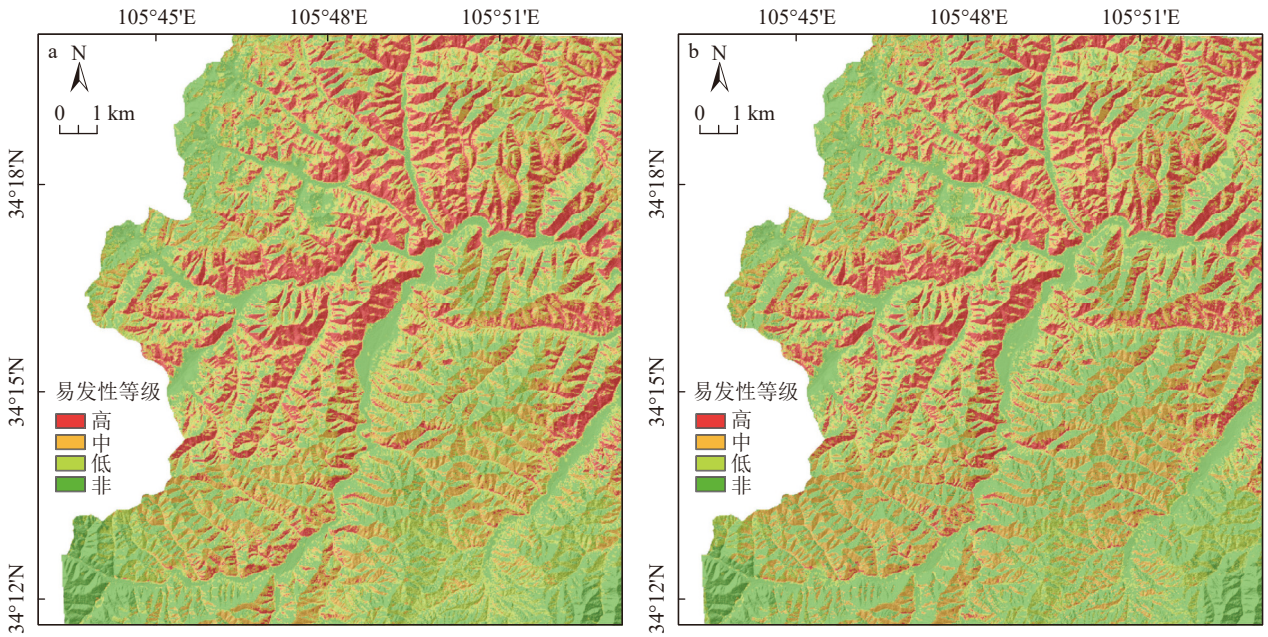

图 11 滑坡易发性评价结果对比图

Fig. 11 Comparison diagram of landslide susceptibility assessment results

a—优化 RF; b—常规 RF

表 5 滑坡易发性评价分区统计结果

Table 5 Landslide susceptibility assessment zoning statistical results

预测模型	易发性等级							
	高		中		低		非	
	分区面积/%	滑坡面积/%	分区面积/%	滑坡面积/%	分区面积/%	滑坡面积/%	分区面积/%	滑坡面积/%
优化RF	18.67	14.04	24.63	3.73	26.32	1.29	30.37	0.27
常规RF	17.47	13.56	24.37	2.98	20.91	1.62	37.25	0.39

5 结 论

本文以西秦岭山区天水娘娘坝镇地区为例，从滑坡–非滑坡样本筛选、影响因子选取、联接方法应用、超参数优化 4 个方面对随机森林模型在滑坡易发性评价的应用进行优化和对比研究，主要取得以下结论。

（1）非滑坡样本选取上，基于降雨事件的 20362 处降雨滑坡面状矢量数据样本的完整编目库，以建立 50 m 缓冲区反向选取的方式建立非滑坡选取区域，有效避免了所选取非滑坡点落入滑坡边界内造成信息误差，减少了将潜在滑坡误分为非滑坡的情况。

（2）影响因子选取上，通过 Pearson 相关性分析与基于随机森林 *Gini* 系数的滑坡发育贡献程度相结合的方法，最终甄选了前期有效降雨、坡度等 15 个影响因子。连续型因子直接提取其实际数值，离散型因子通过频率比进行定量表征，最后所有量化因子均作归一化处理，消除不同因子间的量纲影响，使数据处于同一数量级，提升了机器学习模型样本训练效率。

（3）超参数优化上，结合遗传算法对随机森林超参数进行自适应寻优，在耗时近 54 h 的模型训练后，得出决策树数量、树的最大深度等超参数的寻优结果。通过对比优化随机森林与常规随机森林模型预测结果的 ROC 曲线及 AUC 值，AUC 精度值达到 0.877，可知优化随机森林模型预测精度更高，有较高的成功率与预测率，可为山地丘陵区降雨滑坡机器学习易发性评价提供参考。

致谢：应急管理部国家自然灾害防治研究院的

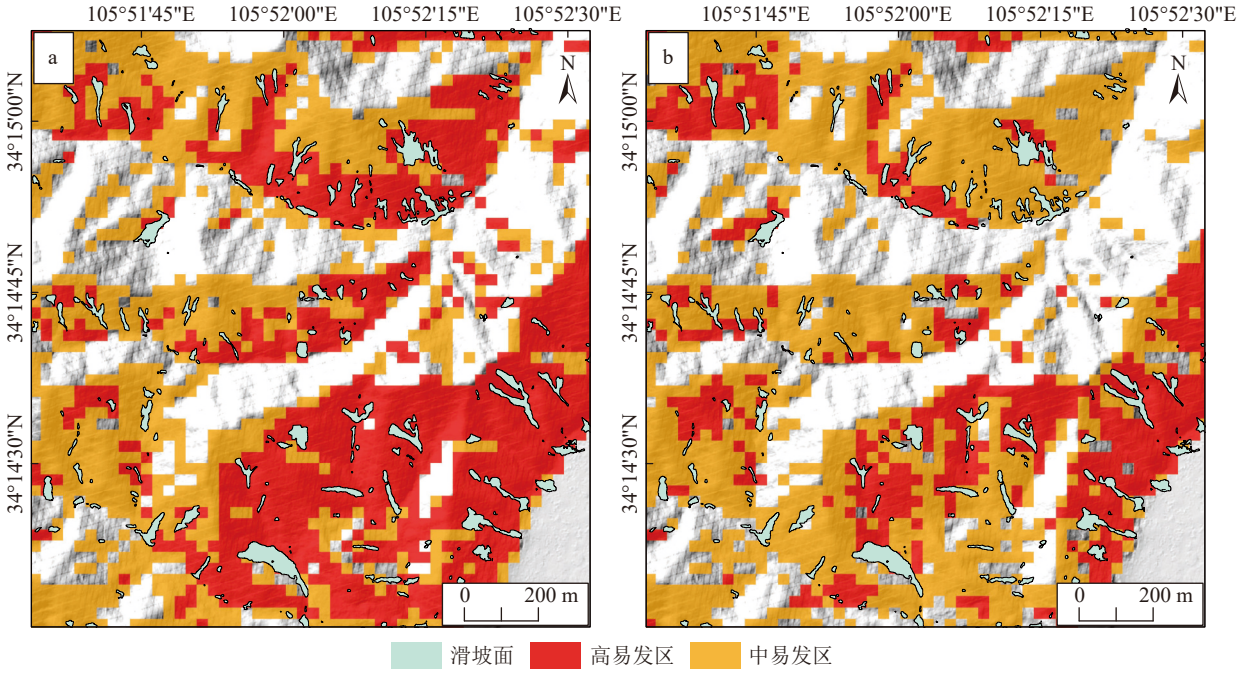

图 12　滑坡集中区易发性评价结果对比图

Fig. 12　Comparison chart of susceptibility assessment results of landslide concentration area

a—优化 RF; b—常规 RF

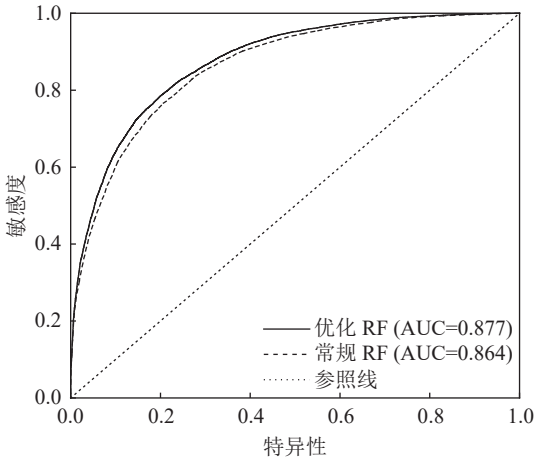

图 13　滑坡易发性评价结果 ROC 曲线

Fig. 13　ROC curve of landslide susceptibility assessment results

许冲研究员提供了部分滑坡数据；甘肃省地质环境监测院张永军、宋晓玲教授级高工提供了“7·25”群发灾害事件有关指导，谨表谢意。

参考文献

Abbas F, Zhang F, Ismail M, et al. 2023. Optimizing machine learning algorithms for landslide susceptibility mapping along the Karakoram Highway, Gilgit Baltistan, Pakistan: A comparative study of baseline, bayesian, and metaheuristic hyperparameter optimization techniques[J]. Sensors (Basel, Switzerland), 23(15): 6843.

Achu A L, Aju C D, Di N M, et al. 2023. Machine−learning based landslide susceptibility modelling with emphasis on uncertainty analysis[J]. Geoscience Frontiers, 14(6): 101657.

Ado M, Amitab K, Maji A K, et al. 2022. Landslide susceptibility mapping using machine learning: A literature survey[J]. Remote Sensing, 14(13): 3029.

Ahmad A, Farida M, Juita N, et al. 2023. Soil micromorphology for modeling spatial on landslide susceptibility mapping: A case study in Kelara subwatershed, Jeneponto Regency of South Sulawesi, Indonesia[J]. Natural Hazards, 118(2): 1445−1462.

Alvioli M M, Melillo F, Guzzetti M, et al. 2018. Implications of climate change on landslide hazard in Central Italy[J]. Sci. Total Environ., 630: 1528−1543.

Bui D, Tsangartos P, Nguyen T, et al. 2020. Comparing the prediction performance of a deep learning neural network model with conventional machine leaning models in landside susceptibility assessment[J]. Catena, 188: 104426.

Danika S, Edward M C, Sylvia H M, et al. 2019. A Mixed−Methods Evaluation of a gender affirmative education program for families of trans young people[J]. Journal of GLBT Family Studies, 16(1): 18−31.

Deng H, Wu X T, Zhang W J, et al. 2022. Slope−unit scale landslide susceptibility mapping based on the random forest model in deep valley areas[J]. Remote Sensing, 14(17): 4245.

Du G L, Zhang Y S, Iqbal J, et al. 2017. Landslide susceptibility mapping using an integrated model of information value method and logistic

regression in the Bailongjiang watershed, Gansu Province, China[J]. Journal of Mountain Science, 14(2): 249−268.

Guilherme G O, Luis F C, Laurindo A G, et al. 2019. Random forest and artificial neural networks in landslide susceptibility modeling: a case study of the Fão River Basin, Southern Brazil[J]. Natural Hazards, 99(2): 1049−1073.

Havenith H B, Torgoev A, Schlöge R, et al. 2015. Tienshan geohazards database: Landslide susceptibility analysis[J]. Geomorphology, 249: 2856.

Huang F M, Chen J W, Liu W P, et al. 2022. Regional rainfall−induced landslide hazard warning based on landslide susceptibility mapping and a critical rainfall threshold[J]. Geomorphology, 408: 108236.

Isidro C, Miguel A C, Francisco G, et al. 2019. A ROC analysis−based classification method for landslide susceptibility maps[J]. Landslides, 16(2): 265−282.

Jie D, Ali P Y, Abdelaziz M, et al. 2020. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning[J]. Science of the Total Environment, 720: 137320.

Ke W, Chang M W, Fang Qi, et al. 2012. Search of the most dangerous in slide slope stability analysis based on genetic algorithm[J]. Applied Mechanics and Materials, 1799: 166−169.

Khan R, Yousaf S, Haseeb, A, et al. 2021. Exploring a Design of landslide monitoring system[J]. Complexity, 2021: 1−13.

Li R, Huang S Y, Dou H Q. 2023. Dynamic risk assessment of landslide hazard for large−scale photovoltaic power plants under extreme rainfall conditions[J]. Water, 15(15): 2832.

Ma S Y, Shao X Y, Xu C, et al. 2023. Physically−based rainfall−induced landslide thresholds for the Tianshui area of Loess Plateau, China by TRIGRS model[J]. Catena, 233: 107499.

María C H, Laura P C, Iván L H, et al. 2023. Landslide susceptibility analysis on the vicinity of Bogotá−Villavicencio road (Eastern Cordillera of the Colombian Andes)[J]. Remote Sensing, 15(15): 3870.

Reichenbach P M, Rossi B D, Malamud M M, et al. 2018. A review of statistically−based landslide susceptibility models[J]. Earth−Science Reviews 180: 60−91.

Sun D L, Wen H J, Wang D Z, et al. 2020. A Random Forest Model of landslide susceptibility mapping based on Hyperparameter Optimization using bayes algorithm[J]. Geomorphology, 362: 107201.

Tareq H M, Juhari M A, Abdul G R, et al. 2011. Landslide susceptibility assessment using frequency ratio model applied to an area along the E−W highway (Gerik−Jeli)[J]. American Journal of Environmental Sciences, 7(1): 47−56.

Wen F, Xin S W, Yan B C, et al. 2017. Landslide susceptibility assessment using the certainty factor and analytic hierarchy process[J]. Journal of Mountain Science, 14(5): 906−925.

Yang H, Shi P, Quincey D. et al. 2023. A heterogeneous sampling strategy to model earthquake−triggered landslides[J]. Int. J. Disaster Risk Sci., 14(4): 636−648.

Yao J, Qin S, Qiao S, et al. 2022. Application of a two−step sampling strategy based on deep neural network for landslide susceptibility mapping[J]. Bulletin of Engineering Geology and the Environment, 81: 1−20.

Yao X, Tham L, Dai F, et al. 2008. Landslide susceptibility mapping based on support vector machine: A case study on natural slopes of Hong Kong, China[J]. Geomorphology, 101: 572−582.

Youssef A M, Pourg H, Hamid R, et al. 2016. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia[J]. Landslides, 13(5): 839.

郭飞, 赖鹏, 陈洋, 等. 2022. 不同环境因子联接方法对崩岗易发性评价的影响[J]. 水土保持通报, (425): 123−130.

郭飞, 赖鹏, 黄发明, 等. 2023. 基于知识图谱的滑坡易发性评价文献综述及研究进展[J/OL]. 地球科学, 1−33. [2024−05−14]. http://kns.cnki.net/kcms/detail/42.1874.P.20230713.1234.002.html.

郭富赟, 孟兴民, 黎志恒, 等. 2015. 天水市“7·25”群发性地质灾害特征及成因[J]. 山地学报, 33(1): 100−107.

黄发明, 殷坤龙, 蒋水华, 等. 2018. 基于聚类分析和支持向量机的滑坡易发性评价[J]. 岩石力学与工程学报, 37(1): 156−167.

黄森. 2021. 天水市“7·25”群发性降雨滑坡灾害预警模型研究[D]. 西北大学硕士学位论文.

李霞, 宿星, 张满银, 等. 2023. 基于证据权法与多源数据的陇中生态脆弱区滑坡敏感性评价——以天水市为例[J]. 冰川冻土, 45(1): 67−79.

刘帅, 朱杰勇, 杨得虎, 等. 2023. 基于斜坡单元与随机森林模型的元阳县崩滑地质灾害易发性评价[J]. 中国地质灾害与防治学报, 34(4): 144−150.

孟晓捷, 张新社, 曾庆铭, 等. 2022. 基于加权信息量法的黄土滑坡易发性评价——以 1∶5 万天水市麦积幅为例[J]. 西北地质, 55(2): 249−259.

唐辉明, 鲁莎. 2018. 三峡库区黄土坡滑坡滑带空间分布特征研究[J]. 工程地质学报, 26(1): 129−136.

吴润泽, 胡旭东, 梅红波, 等. 2021. 基于随机森林的滑坡空间易发性评价: 以三峡库区湖北段为例[J]. 地球科学, 46: 321−330.

许强, 黄润秋, 李秀珍. 2004. 滑坡时间预测预报研究进展[J]. 地球科学进展, (3): 478−483.

殷跃平, 高少华. 2024. 高位远程地质灾害研究: 回顾与展望[J]. 中国地质灾害与防治学报, 35(1): 1−18.

周晓亭, 黄发明, 吴伟成, 等. 2022. 基于耦合信息量法选择负样本的区域滑坡易发性预测[J]. 工程科学与技术, 54(3): 25−35.