doi: 10.12097/gbc.2023.07.020

增量学习在滑坡易发性评价中的应用
——以甘肃省天水市为例

严天笑[1,2]，张建通[3*]，朱月琴[2*]，刘浩然[1,2]，朱浩濛[4]
YAN Tianxiao[1,2], ZHANG Jiantong[3*], ZHU Yueqin[2*], LIU Haoran[1,2], ZHU Haomeng[4]

1. 防灾科技学院, 河北 廊坊 065201;
2. 应急管理部国家自然灾害防治研究院, 北京 100085;
3. 交信北斗科技有限公司, 北京 100011;
4. 浙江省地质院, 浙江 杭州 310000
1. *Institute of Disaster Prevention*, *Langfang* 065201, *Heibei*, *China;*
2. *National Institute of Natural Hazards*, *Ministry of Emergency Management of China*, *Beijing* 100085, *China;*
3. *Jiaoxin Beidou Technology Company*, *Beijing* 100011, *China;*
4. *Zhejiang Institute of Geosciences*, *Hangzhou* 310000, *Zhejiang*, *China*

摘要: 为了提升机器学习模型在滑坡易发性评价任务中的泛化能力，以甘肃天水市为例，采用基于 LightGBM 的增量学习模型，并利用 Autogluon 自动机器学习框架实现模型的超参数优化和堆叠，以及使用 SHAP 可解释框架进行特征选择和数据异常分析，构建了适用于滑坡易发性评价的增量学习模型。通过在天水市不同区域采集的滑坡灾害数据进行模型验证，结果表明，基于增量学习的滑坡易发性评价模型能够有效地识别和预测滑坡易发区域，根据新数据集自适应调整模型，并且提高模型的性能。

关键词: 滑坡易发性；机器学习；增量学习；特征选择；可解释性
中图分类号: P642.22　　**文献标志码:** A　　**文章编号:** 1671−2552(2024)04−0630−11

Yan T X, Zhang J T, Zhu Y Q, Liu H R, Zhu H M. Application of incremental learning in landslide susceptibility assessment: A case study of Tianshui, Gansu Province. *Geological Bulletin of China*, 2024, 43(4): 630−640

Abstract: To enhance the generalization ability of machine learning models in the assessment of landslide susceptibility, this paper takes the city of Tianshui as an example and employs an incremental learning model based on LightGBM. By utilizing the Autogluon automated machine learning framework, the model's hyperparameter optimization and mdoel stacking are implemented. Additionally, the SHAP explainable framework is used for feature selection and data anomaly analysis. By using the above methods we construct an incremental learning model suitable for landslide susceptibility assessment. Model validation using landslide disaster data collected from various regions in Tianshui city demonstrates that the incremental learning model for landslide susceptibility can effectively identify and predict landslide-prone areas. It adapts to new datasets by self-adjusting the model and improves model performance.

Key words: suspectibility of landslide; machine learning; incremental learning; feature selection; interpretability

收稿日期: 2023−07−20; **修订日期:** 2023−11−15
资助项目: 应急管理部国家自然灾害防治研究院基本科研业务专项(编号: ZDJ2022-45)、国家自然科学基金项目《大数据环境下的滑坡危险性评估模型构建方法研究》(批准号:41872253) 和河北省大学生创新创业训练计划项目《InSAR 与深度学习技术相结合的白格地区滑坡形变监测与识别》(编号: S202211775007)
作者简介: 严天笑(1998−), 男, 在读硕士生, 从事灾害信息处理研究。E−mail: yurenzi@126.com
*** 通信作者:** 张建通(1977−), 男, 高级工程师, 从事交通安全、卫星导航系统等研究与应用推广。E−mail: zhangjiantong@cttic.cn
朱月琴 (1975−) , 女, 博士, 高级工程师, 从事地质大数据、地图综合与可视化研究工作。E−mail: yueqinzhu@163.com

滑坡作为常见的突发性地质灾害, 严重威胁人民生命财产安全, 对滑坡灾害进行监测具有重要意义。甘肃省天水市黄土灾害频发, 本文以天水市历史滑坡灾害数据作为研究对象构建易发性评价模型, 滑坡易发性评价方法包括基于公式的统计方法和机器学习方法。机器学习方法具有更快的数据处理速度, 并可以更准确地描述滑坡因子与易发性之间的非线性关系, 被广泛应用于滑坡易发性预测和地质灾害风险评价中。

国内很早就将机器学习技术应用于滑坡易发性预测和地质灾害风险评价(严武文,2010), 并据此制定了防灾减灾策略。为了提高评价结果的准确性, 研究者们往往从数据优化和模型改进 2 个方面优化机器学习模型: 滑坡主控因子和负样本的选取是影响滑坡易发性分析结果的关键因素, 通过确定最优评估因子组合可以训练得到最适合研究区域的机器学习模型(黄发明等, 2022; 刘纪平等, 2022)。超参数选取和混合模型的应用解决了单个模型在准确性和训练效率方面的局限性(武雪玲等, 2016; 邓念东等, 2022)。针对天水市的滑坡易发性研究工作也已取得大量研究成果,包括使用多元线性回归模型(邵葆蓉等,2020)和 BP 神经网络(邵葆蓉等,2020)建立评价模型。这些研究主要集中在比较和应用不同的模型在该研究区的效果, 本文则针对易发性评价模型泛用能力差的特点研究增量学习在其中的应用。

大量研究证明, 机器学习方法在滑坡易发性评价中得到了卓有成效的应用。近几年的集成模型对数据的拟合能力达到了较高的水平(Merghadi et al., 2020), 模型精度很难有进一步提升。传统机器学习算法仍存在一些缺点, 包括难以计算大量数据、模型泛化能力差等, 而地质学具有很强的时空异质性, 针对这一问题, 增量学习方法可能成为一种很好的解决方式。目前增量学习更多用在监测预警、实时检测、决策支持等需要不断融合新数据的场景, 例如 Wu et al.(2019)在基于水文知识的贝叶斯网络模型基础上, 提出了一种增量学习方案, 旨在通过更新网络参数来提高模型的自我改进和自适应能力, 该方案针对小型河流的洪水预测问题进行研究, 并取得了显著的成果; Wu et al.(2013)则基于支持率向量机模型提出了一种在先验知识较少的情况下监测预警岩爆灾害的增量学习模型; Huang et al.(2022)结合贝叶斯网络和增量学习, 构建了一种新的用于滑坡易发性评价的机器学习模型, 该模型随着新数据的增加预测精度逐渐提高(Huang et al., 2022)。与传统机器学习模型相比, 结合增量学习的贝叶斯网络模型可以更有效地利用历史滑坡数据, 在新的滑坡数据加入时进行学习, 从而更好地预测未来的滑坡。

本文围绕如何快速准确地对跨区域滑坡数据进行易发性评价这一问题, 研究增量学习方法在滑坡易发性评价中的应用。

1　增量学习方法及领域应用现状

1.1　增量学习理论方法

增量学习(Incremental Learning)也可称作终生学习或持续学习, 是一种适用于动态数据集的机器学习方法。相较于传统的机器学习方法, 增量学习在保留已有训练结果的基础上, 对新增数据进行针对性地再学习, 从而不断迭代更新模型。这意味着支持增量学习的模型具有较好的灵活性和泛化能力, 对数据变化适应能力强, 从而适用于解决许多实际问题。

目前, 增量学习模型在许多领域已经获得了成功应用, 在电网短路故障位置自动识别和机械故障诊断等领域, 基于增量学习的算法有效识别了风险位置并拥有较高的识别精度(李世其等, 2006; 王洪林等, 2022)。研究结果表明, 增量学习能够有效克服实际应用中传统方法数据需求量大、更新困难等多种局限, 进一步提高模型的准确性和时效性。

增量学习方法可以应用于支持向量机和决策树等多种机器学习算法, 其在不同算法中的实现思路有所差异, 在支持向量机中, 可以通过增量学习对超平面参数进行逐步更新, 并采用矩阵运算和核函数技巧减少计算量, 从而实现对划分超平面的快速适应(李挺等, 2018); 在决策树模型中, 增量学习往往是通过对叶节点属性和划分准则的更新来实现的。

1.2　基于 LightGBM 的增量学习模型

相比于其他基于决策树算法的梯度提升模型, LightGBM 更注重于训练速度、内存消耗和准确率方面的优化, 其采用基于梯度的单边采样和独立特征合并技术分别处理大量数据实例和特征, 使得其内存消耗和计算速度显著高于如 XGBoost 的其他集成学习模型。此外 LightGBM 采用 Leaf-wise(叶子生长)策略, 每次选择当前所有叶子中分裂增益最大的

叶子进行分裂，从而在迭代中快速建立决策树。相较于 level-wise 生长策略，叶子生长策略能够更好地降低误差，但同时也容易过拟合，长出较深的决策树（Ke et al., 2017）。

作为一种兼具高效率与高准确率的集成学习算法，LigtGBM 已经在滑坡问题上获得了显著的成果，特别是在滑坡灾害易发性与危险性评价方面，表现出优异的性能（赵泽园等, 2020; 张博等, 2023）。结果表明，相比于 XGBOOST 等其他的大规模集成学习模型，LightGBM 在计算速度上更快，资源消耗上更低，而准确度上则更高。因此，它极其适用于滑坡相关问题的研究工作。

LightGBM 提供了 2 种增量学习机制，第一种是在已经训练好的模型基础上，在新数据集上训练新的树，并将新旧两批次训练树的结果叠加作为最终的输出；另一种是基于现有的树结构对新数据进行更新，直接在已建立的树的叶子上添加新的叶子节点，不需要对原有的树进行修改，快速更新模型。

2 基于增量学习的滑坡易发性评价模型构建

2.1 基于增量学习的滑坡易发性评价模型架构

增量学习可应用于样本数据集较少和数据记录十分庞大的情况。在滑坡易发性评价的应用场景中，当发现模型预测性能下降时，往往需要更新原始训练集重新建模，这种做法在工作量、时间和资源上都存在浪费的问题（庄维嘉等, 2022）。相较而言，增量学习策略可以仅针对新增数据进行持续更新，避免重复训练整个数据集，降低计算复杂度，提高模型泛化能力，达到平衡模型性能和效率的效果。

本次研究旨在探究增量学习技术在滑坡易发性评价领域的应用，深入研究增量学习技术在滑坡易发性评价过程中的具体实现方法和效果，实验流程包括：①基于斜坡单元的滑坡因子数据集制作与处理；②使用研究区 1 数据，基于自动机器学习框架的滑坡预测模型的构建，具体来说是其内置的超参数寻优功能训练 LightGBM 模型；③在基本 LightGBM 模型的基础上使用增量学习功能训练研究区 2 数据，并结合自动机器学习框架的模型堆叠功能构建与优化模型；④对比增量学习前后的模型验证精度。

在具体环节中，由于 LightGBM 是一种基于决策树的梯度提升框架，为了构建具有较高预测准确率的模型，必须找到合适的参数组合，包括学习率 (learning_rate)、叶子节点数 (num_leaves)、特征子抽样比例 (feature_fraction)、最大树深度 (max_depth) 等，这些参数对模型的预测准确率有很大的影响。本次研究使用抽样方法、AutoGluon 自动机器学习框架、SHAP 可解释框架优化数据集、改进机器学习模型提高模型预测的精度。

2.2 数据抽样

在进行滑坡易发性分析构建数据集时，需要包括正负样本数据。其中，正样本数据一般是历史滑坡清单，由专业人员进行现场勘察得到，具有较高的可靠性。而负样本数据则需要进行样本筛选，本文选择使用集成学习模型 Adaboost 作为先验模型，选择概率低于阈值的样本作为负样本数据，并使用下采样算法平衡数据集，最后采用聚类抽样去除噪声，减少计算量，具体流程见图 1。这样可以确保数据集中同时包含了滑坡和非滑坡数据样本，有助于提高模型训练和预测的准确性、可靠性。

2.3 自动机器学习框架

以往基于机器学习的滑坡易发性制图预测精度高度依赖于数据处理、模型选取和超参数组合的选择，其中数据处理、调参等工作费时费力且重复度高。自动机器学习框架的出现大大简化了这些流程，增加了研究的可重复性和准确性，已有研究者开始将自动机器学习框架应用于滑坡易发性研究，并利用其进行全球尺度的滑坡易发性制图（王毅等, 2022）。结果表明，基于自动机器学习的滑坡易发性模型性能更稳定、精度更优越。如图 2 所示，本文采用亚马逊开源的 AutoGluon 自动机器学习框架辅助进行如下任务。

超参数寻优：Autogluon 采用基于贝叶斯优化的策略进行参数调优，在预定义的参数空间内随机选择一组参数作为初始参数，以该组参数训练模型并根据预设的评估指标对预测性能进行评估。之后使

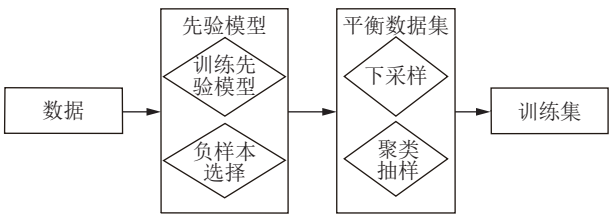

图 1 样本优化流程

Fig. 1 Sample optimization process

用高斯过程模型预测不同参数设置下的预测性能，并据此调整参数组合进行模型训练，直到找到最佳参数组合以获得最佳的预测性能。使用这种优化策略，Autogluon 可以快速高效地调整模型参数，并找到最佳的预测结果。

交叉验证: Autogluon 能够自动化选择最佳的交叉验证方法和参数，并将其融入模型选择和超参数搜索的优化过程中，以提高模型的性能和泛化能力。

Stacking 集成: Autogluon 提供了自动化的 Stacking 集成学习方法，利用交叉验证技术同时训练多个不同类型的基学习器(组成集成学习模型的各个机器学习模型)，并在组合之后进行自适应的超参数优化，以增强整个集成模型的泛化能力和准确性(Erickson et al., 2020)。

总体来说，Autogluon 可以帮助用户减少模型和数据处理方面的工作负担，更加高效地开发和部署机器学习模型，从而提高模型性能和准确度。本文计划使用 AutoGluon 框架进行数据处理、增量学习模型的参数寻优和模型堆叠以提高模型精度。

2.4　SHAP 可解释模型

除线性回归、逻辑回归、决策树等精度较低的机器学习模型外，集成学习等复杂机器学习方法常被称作是“黑箱模型”，这些模型普遍缺乏可解释性，为了解决这一问题，可解释性机器学习方法被提出。

SHAP 具有完善的理论基础，它使用博弈论中的 Shapley 值，通过为特征分配一个重要性值(SHAP 值)来解释特定的预测。假设第 i 个样本为 X_i，共有 n 个特征，模型对第 i 个样本的预测值为 Y_i，以均值作为整个可解释模型的基线，记为 y_{base}，则该样本的 Shapely 值计算公式为:

$$Y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{in})$$

如图 3 所示，SHAP 通过计算每个特征对于该样本预测结果的 Shapley 值，分析模型中每个特征的贡献和特征之间的交互作用；通过计算每个数据点的相对贡献，检测相对于其他点具有不同特征影响的数据点(Lundberg et al., 2017)。综上所述，SHAP 可以用来进行特征分析和异常数据检测。

3　研究区概况和数据准备

3.1　研究区概况

天水市位于甘肃省东南部，地处陇西黄土高原过渡地带，地势西北高、东南低，市区平均海拔 1100 m，呈条带状分布，是典型的河谷型城市。地质构造复杂，黄土结构疏松、分布广泛，岩层以软弱的泥岩和黄土混杂堆积物为主。受地质条件影响，天水市地质环境脆弱，滑坡、泥石流、崩塌灾害频发，其中滑坡分布十分广泛，威胁人民生命财产安全，严重制约城市建设发展(齐娜等, 2022)。研究区概况及采集到

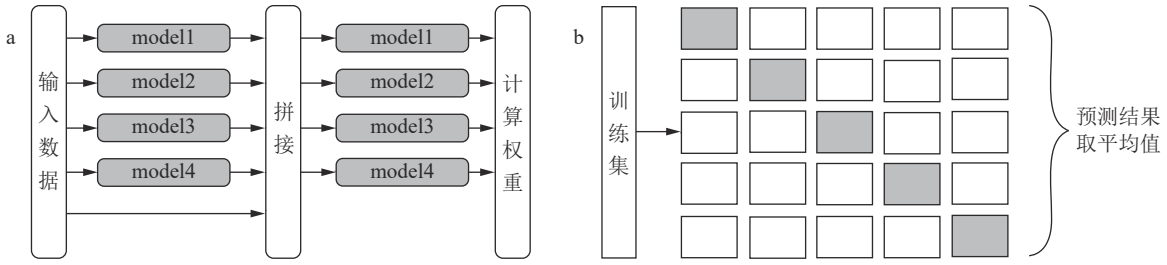

图 2　AutoGluon 模型多层堆叠(a)和 K 折交叉验证(b)

Fig. 2　AutoGluon model(a)and K-fold cross-validation(b)

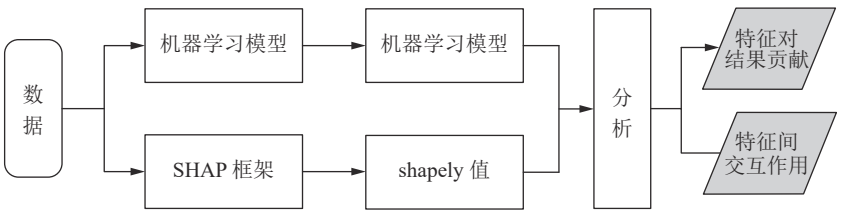

图 3　SHAP 可解释模型分析流程

Fig. 3　SHAP interpretable model analysis process

的历史滑坡灾害空间分布情况见图 4，可以看到历史滑坡灾害数据主要采集于两片区域。

根据《天水市人民政府关于全市地质灾害防治工作情况的报告——2021 年 6 月 29 日在市七届人大常委会第四十次会议上》，截至 2021 年，全市排查出地质灾害隐患点 2005 处，其中滑坡 1184 处，达到地质灾害总数的 59%。近些年滑坡灾害仍时有发生，2020 年，强降雨导致天水市麦积区公路线塌方，交通中断；2021 年，天水娘娘坝持续降雨导致滑坡复发，摧毁坡下两间房屋。

前人针对天水市滑坡灾害的易发性进行了深入探讨，并取得大量成果（邵葆蓉等,2020；康孟羽等,2022）。然而，这些研究主要集中在评估和比较不同的模型在该研究区的效果，同时根据该地区特有的地质条件，采用各种不同的因子组合进行研究。本文根据天水市历史滑坡灾害点数据的地理分布特点，探讨增量学习模型在处理不同区域数据时的适用性。

3.2 数据来源

本文使用实地调查数据、卫星影像数据和 GIS 工具构建滑坡数据集，涉及到的研究数据包括：①滑坡数据库来源于中国地质科学院地质力学研究所的历史滑坡灾害点数据，数据以 shp 矢量文件格式存储。②数字高程影像（DEM）数据来源于中国地质科学院地质力学研究所收集的 25m*25m DEM 数据，数据以 tif 栅格影像文件格式存储。

归一化植被指数（NDVI，反映植被覆盖量的变化）、归一化水体指数（NDWI，反映水体分布信息）和归一化建筑指数（NDBI，反映建筑用地信息）通过对 Landsat 8 卫星影像进行解译而得出。这些指数利用不同波段的反射率数据，经过辐射定标、大气校正计、归一化处理，提供了关于地表植被、水体和建筑物分布情况的重要信息。

道路、土壤、岩石、降水等灾害因子数据来源于中国科学院地理科学与资源环境科学与数据中心公开数据集。

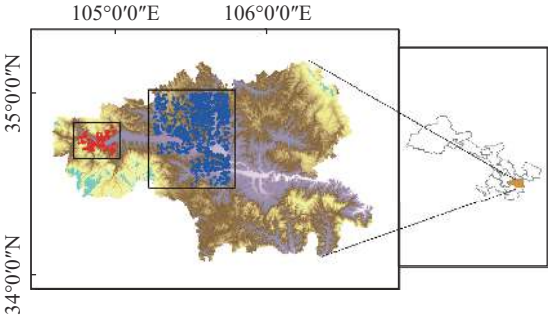

图 4 天水市滑坡灾害分布图

Fig. 4 Landslide distribution map of Tianshui City

3.3 斜坡单元划分

滑坡灾害区域评价常用的评价单元有栅格单元和斜坡单元，斜坡单元是由山谷线和山脊线围成的区域，为滑坡发生的基本单元，在描述边坡力学机制、岩性、环境边界等方面都优于栅格单元。传统基于水文分析法的斜坡单元划分方法容易需要人工设置阈值，存在大量破损面和冗余分界。本文在提取坡向和山体阴影图的基础上，采用 MSS（多尺度分割）划分斜坡单元，其划分情况如图 5 所示，可以看到基于多尺度分割算法划分的斜坡单元边缘平滑，破碎面少。

4 基于增量学习的天水市滑坡易发性评价分析

4.1 评价因子选择

滑坡易发性是多种孕灾因子耦合的结果，用于滑坡易发性评价的特征往往因地理位置的不同而存在差异。为了进行综合分析，本文在总结归纳已有专家学者研究成果的基础上，基于研究区滑坡规模和分布特点、专家咨询和野外调查成果及其他研究区的研究成果，从地质构造、水文环境、人为活动、土地植被等多个方面选取了 21 个特征作为评价因子（图 6）。

在地质构造方面，选择高程、坡度、坡向、曲率、地形粗糙度、地形起伏度、地形湿度、岩性、断层距离作为滑坡的控制因素，这些因子反映了地表形态、稳定性、岩性等地质特征。在水文环境方面，选择降

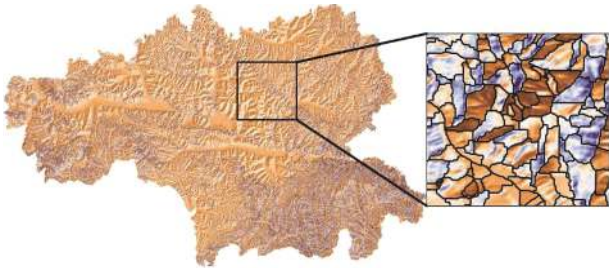

图 5 基于多尺度图像分割的斜坡单元划分

Fig. 5 Slope unit division based on multi-scale image segmentation

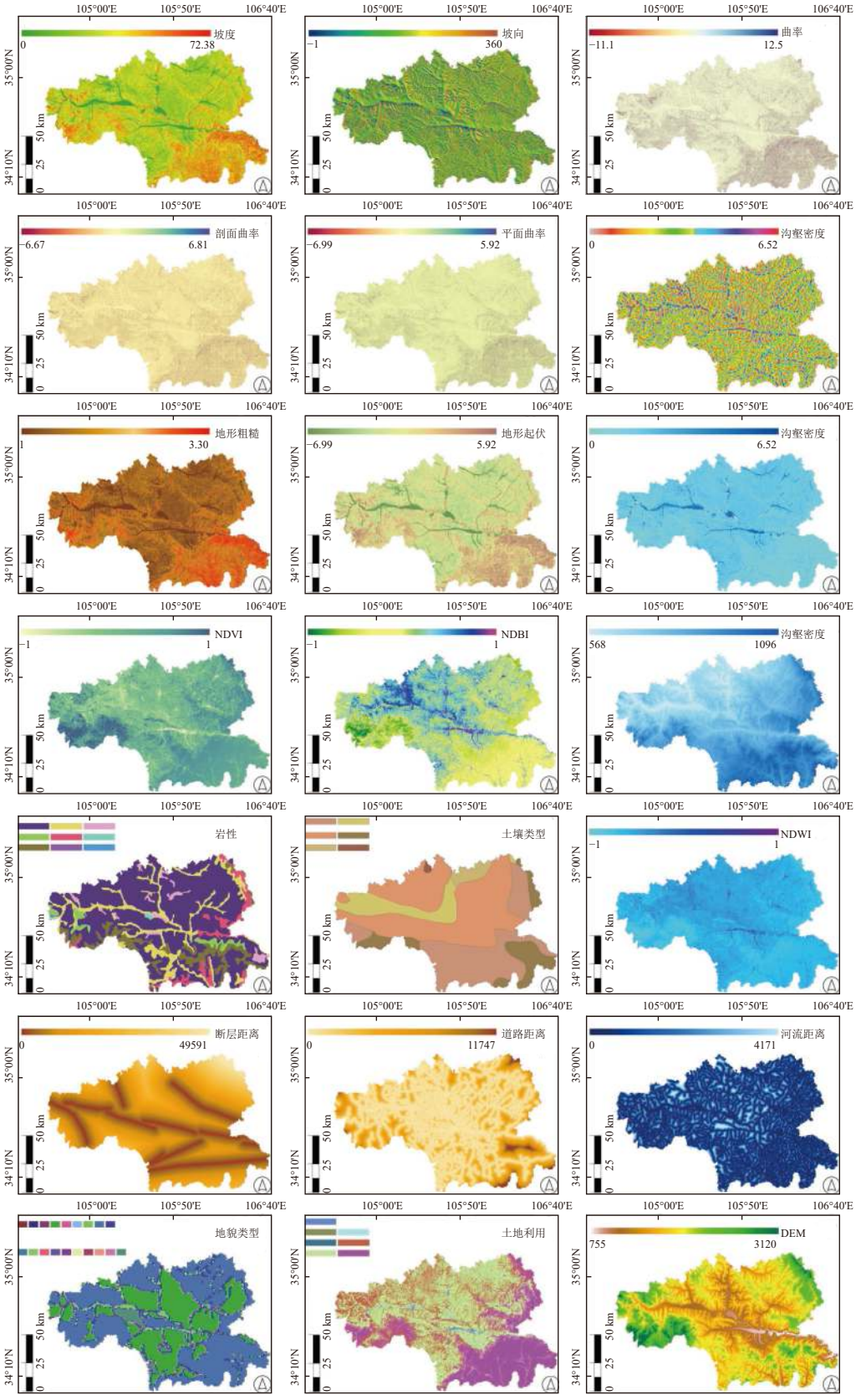

图 6　滑坡影响因子图

Fig. 6　Landslide impact factor diagrams

水量、NDVI、NDWI、距河流距离作为滑坡控制因素，其中降雨是滑坡发生的重要驱动因素，NDVI 反映了地表植被情况，NDWI 反映了水体分布情况，距河流距离反映了滑坡与河流之间的距离，这些都可能增加水分的聚集和土壤侵蚀的风险，从而增加滑坡的潜在危险性。

在人类活动方面，选择距道路距离、土地利用、NDBI 作为滑坡控制因素，其中距道路距离反映了滑坡与道路的距离，土地利用反映了土地的功能和利用方式，NDBI 则反映了建筑分布情况，这些人类活动都可能导致土壤的扰动，影响滑坡的形成和发展。

在土地植被方面，选择沟壑密度、地貌类型、土壤类型作为滑坡控制因素，其中沟壑的存在可能会增加水流的聚集和土壤侵蚀的风险，不同的土壤、地貌类型具有不同的水分保持能力和土壤稳定性，这些都可能影响滑坡的发育发展。

使用 ArcGIS 等软件进行数据处理、提取，制作了研究区评价因子数据集，汇总至斜坡单元，得到带属性数据的研究区表格数据，部分样本信息如表 1 所示。

4.2 样本优化与特征处理

基于 MSS 斜坡单元划分方法对天水市研究区进行斜坡单元划分，共得到 157985 个斜坡单元，机器学习模型对数据质量有较高的依赖性，为了方便机器学习模型训练，需要进行样本优化和特征选择。

在特征选择方面，集成模型相较于单个模型可以从不同方面捕获数据的特征，通常在集成模型中进行特征选择是不必要的，但为了去除数据集中的冗余特征，降低模型性能开销，为了剔除特征，采用 SHAP 算法基于边缘贡献度对各特征进行 SHAP 计算。对数据进行特征聚类分析和特征贡献度排序后，各个特征的重要性得分如图 7 所示，可以看到 NDBI、NDWI、NDVI 之间，坡度、地形湿度、地形粗糙度、地形起伏度之间，曲率、剖面曲率、平面曲率之间具有较高的关联度，结合贡献度排序和特征关联性，选择去除平面曲率、剖面曲率、NDVI、NDWI、地形粗糙度、坡度、地形起伏 7 个特征。

4.3 基于 LightGBM 的滑坡易发性评价

抽样后的数据集根据地理位置分布划分为研究区 1 和研究区 2，具体情况见图 1，将研究区 1 数据用于训练 LightGBM 模型，设置 learning_rate、n_estimators、max_depth、max_features、num_leaves、feature_fraction、min_data_in_leaf 为寻优参数并设置搜索范围，使用 Autogluon 参数寻优后最佳模型参数如表 2 所示。

经过优化后的模型与其他几种机器学习模型性能对比结果如图 8 所示，其中 8−a 图为 ROC 曲线对比图，横坐标 FPR 表示模型正确识别正例占总正例的比例，纵坐标 TPR 表示错误识别正例站总正例的比值。图 8−b 则为 3 个指标（召回率、精确率、F1 值）对比图，其中召回率反映正确预测正例占实际正例的比值，精确率反映正确预测正例占预测正例的比值，F1 则为召回率和精确率的调和平均值。

可以看到，经过 AutoGluon 参数调优后的 LightGBM 模型有较大程度的精度提升，计算其混淆矩阵，召回率为 0.85，F1 值为 0.81。然而将该模型应

表 1 滑坡样本信息

Table 1 Landslide sample information

序号	DEM	NDBI	NDVI	NDWI	剖面曲率	土地利用	土壤	地形湿度	地形粗糙	地形起伏	是否滑坡
1	1882.51	−0.44	0.81	−0.71	−0.31	2	8	5.14	1.14	50.57	0
2	1837.37	−0.44	0.81	−0.72	−0.16	2	8	4.76	1.17	63.51	0
3	1786.96	−0.45	0.82	−0.72	0.18	2	8	4.80	1.21	77.71	0
4	1826.43	−0.44	0.80	−0.71	−0.05	2	8	4.73	1.19	75.81	0
5	1820.29	−0.44	0.80	−0.71	−0.17	2	8	4.90	1.19	70.70	0
6	1767.20	−0.45	0.82	−0.73	0.17	2	8	4.59	1.16	68.66	0
1027	1627.71	−0.23	0.60	−0.56	−0.68	1	1	5.39	1.08	54.24	1
1028	1545.51	−0.34	0.73	−0.66	0.10	1	1	5.26	1.08	49.08	1
1029	1549.01	−0.29	0.63	−0.59	0.16	1	1	6.11	1.10	53.96	1
1030	1585.05	−0.18	0.46	−0.46	0.01	1	1	4.83	1.11	55.04	1
1031	1540.65	−0.25	0.56	−0.54	−0.08	1	1	5.46	1.07	44.02	1

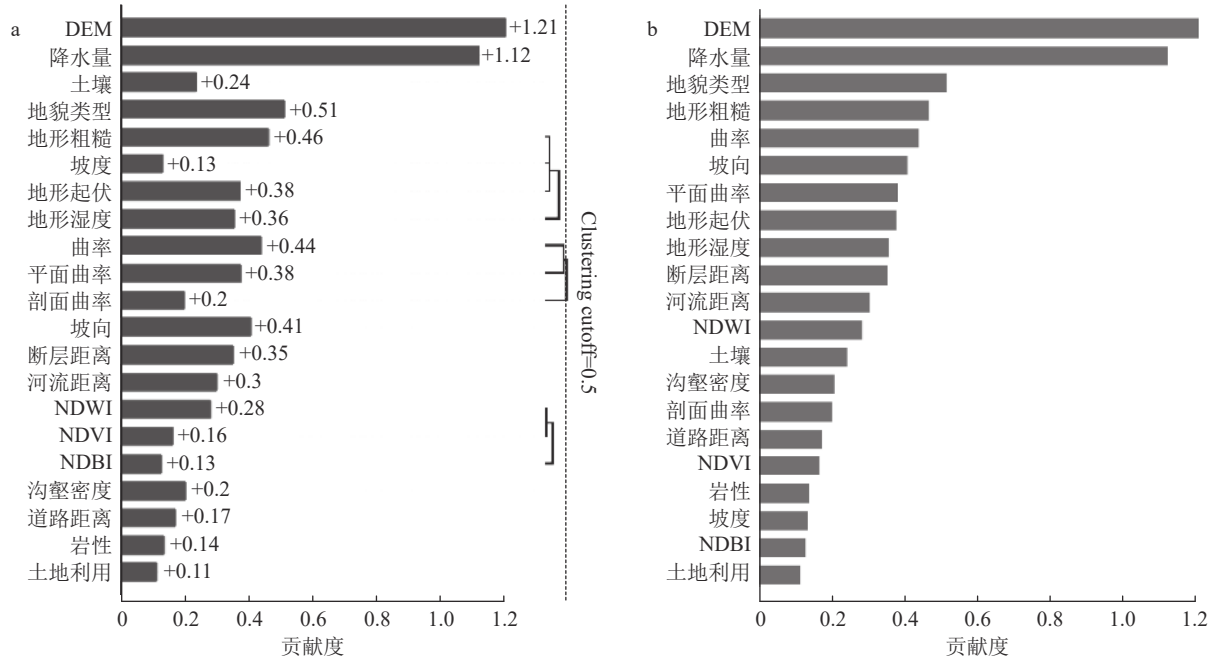

图 7　SHAP 特征聚类图(a)和 SHAP 特征排序图(b)

Fig. 7　SHAP feature clustering plot (a) and SHAP feature ranking plot (b)

表 2　最优参数

Table 2　Optimal parameters

参数名称	learning_rate	n_estimators	max_depth	max_features	num_leaves	feature_fraction	min_data_in_leaf	num_boost_round
最优参数	0.1615	448	8	0.3049	77	0.4447	6	92

用于研究区 2 后, 模型精度发生下降, AUC 值仅为 0.79329。使用 SHAP 进行异常数据检测, 其中蓝线表示对预测的负向贡献, 红线表示特征对预测的正向贡献, 可以发现模型泛化能力较差, 对新区域数据预测能力弱, 通过 ArcGIS 分析发现, 新区域的土壤类型、地貌类型均有新数据出现(图 9)。

4.4　基于增量学习改进滑坡易发性计算分析

为了提高模型泛化能力, 使其对新数据的预测能力得到提升, 使用基于 LightGBM 的增量学习模块对新数据进行增量学习, 同时采用 K 折交叉验证和模型堆叠, 防止数据过拟合和充分利用数据, 提高预测模型准确率。在本次研究中, 为了使预测结果更偏重于初始模型的输出, 选择在原始模型基础上继续扩张叶节点的增量学习策略(图 10)。

经过增量学习后, 模型对新区域的预测性能得到了大幅提高, 同时增量学习的目的是为了保证新得到的分类器对新增训练样本集和原训练样本集均能实现较好的识别, 需要调低增量学习时的学习率。在多次尝试后, 经过比较原区域与新区域数据的预测结果, 选择将学习率设置为 0.01, 并设置指标降低时中断训练, 根据训练完成后的模型绘制研究区 1 和研究区 2 预测结果的 ROC 曲线(图 11)。

表 3 对研究区 1 训练前后的 F1 值和召回率进行了对比, 可以发现学习新数据后对原数据集预测能力也有所上升, 但值得注意的是, 在召回率方面, 新模型的性能有所下降。

4.5　滑坡易发性区划

将经过增量学习的机器学习模型应用于天水市的滑坡易发性评价, 将模型预测的滑坡发生概率作为滑坡易发性指数, 按照自然断点法将其分为 5 个易发等级: 低、中低、中、中高、高。其中, 滑坡易发性指数在 0 ~ 0.3 之间的地区划分为低易发等级; 在 0.3 ~ 0.5 之间的地区划分为中易发等级; 在 0.5 ~ 0.8 之间的地区划分为中高易发等级; 在 0.8 ~ 1 之间的地区划分为高易发等级。通过这种划分方法, 可以得到滑坡易发性评价图(图 12)。

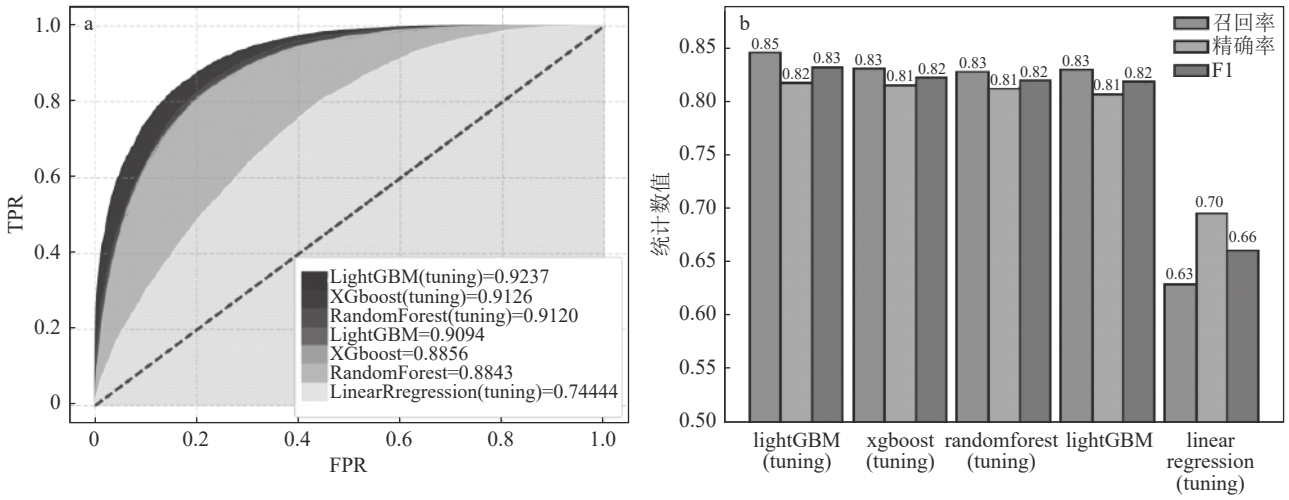

图 8 经过 AutoGluon 优化的 LightGBM 与其他模型性能对比

Fig. 8 Performance comparison of autogluon and other models

a—P-R 曲线图, 反映准确率和召回率之间的关系, 曲线下方面积越大越好; b—各模型之间召回率、精确率和 F1 值对比图

使用 ArcGIS 对易发性结果和滑坡历史数据进行空间连接分析, 结果显示, 发生过滑坡的斜坡单元有 98% 落入中高易发区间和高易发区。该图可以清晰地展示不同地区的滑坡易发性情况, 并较准确地判断滑坡是否发生, 有助于灾害预测和防范措施的制定。

5 结束语

本文以天水市为例, 研究了基于 LightGBM 的增量学习模型在滑坡易发性评价中的应用。结果表明, 基于增量学习的评价模型展现出了良好的性能和应用前景。首先, 该模型在基本保持对原数据预测能力的基础上, 显著提升了对新数据的预测性能。其次, 通过引入新数据集, 该方法能够迅速更新和优化模型, 保持模型的高性能表现。对增量学习前后的模型性能进行比较, 发现其总体上性能有所提升, 意味着该模型能够及时捕捉新数据中的潜在特征和规律, 从而更准确地识别和预测滑坡易发区域。

滑坡易发性评价工作中, 由于数据的时空异质性, 数据集经常需要更新。传统的机器学习模型需

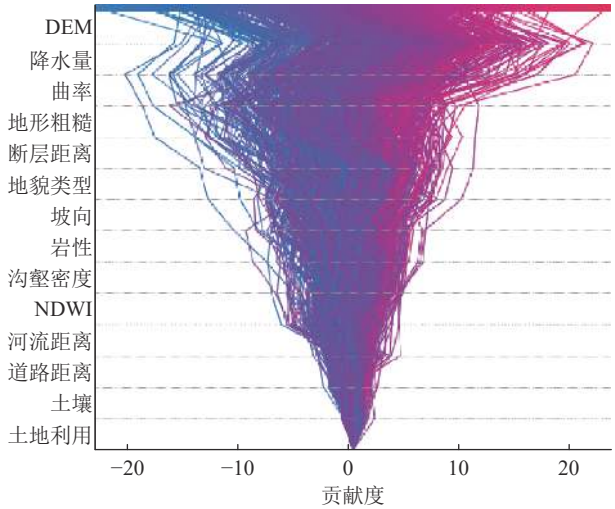

图 9 LightGBM 模型的决策图分析

Fig. 9 Decision Graph analysis of LightGBM model

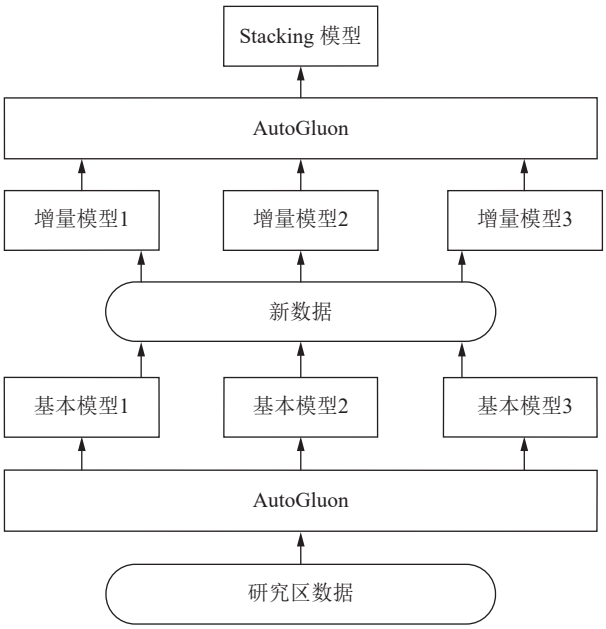

图 10 基于 LightGBM 的增量学习流程

Fig. 10 Incremental learning process based on LightGBM

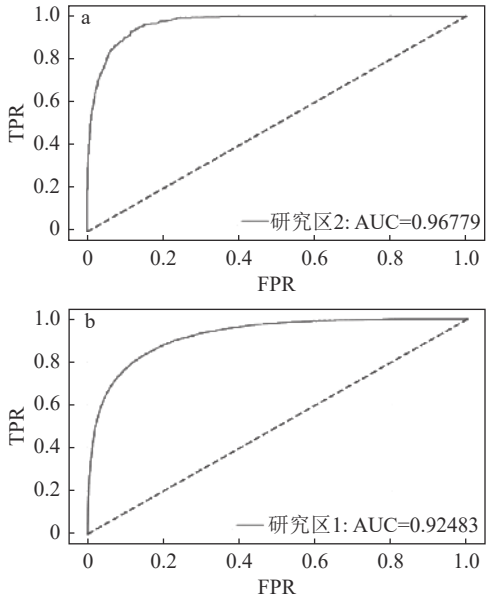

图 11　增量学习后预测效果

Fig. 11　Predictive effect after incremental learning

要重新训练和调整, 这样的过程耗费了大量时间和计算资源。与之相反, 基于增量学习的模型通过引入新数据能够快速自我更新和优化模型, 在保证效率的同时也能够保持模型的高性能。因此, 该模型

表 3　训练前后召回率和 F1 值

Table 3　Recall rate and F1 value before and after training

模型指标	召回率	精确率	F1值
增量训练前	0.846	0.807	0.826
增量训练后	0.842	0.820	0.831

在实践中具有重要的意义和广泛的应用价值。

需要注意的是, 该方法仍存在不可避免的缺陷。例如, LightGBM 的增量学习策略会导致新增的树偏向于新数据, 从而降低对原数据的预测性能。为了解决这个问题, 可以考虑调整学习率来平衡新旧数据的影响。此外, 随着树的增加, 模型体积也会增大, 这需要在实际应用中进行权衡和优化。因此, 在合适条件下, 全量学习仍然是机器学习的最优解。

为进一步推动相关领域的发展, 未来的研究可以从以下几个方面进行改进: 首先, 需要探索更加高效的方法来改进增量学习模型, 以解决模型容易过度学习的问题, 从而提高模型的稳定性和鲁棒性。其次, 需要引入更多相关数据, 并探索训练大模型预测滑坡易发性的方法, 这将为今后的模型训练提供更可靠的基础。

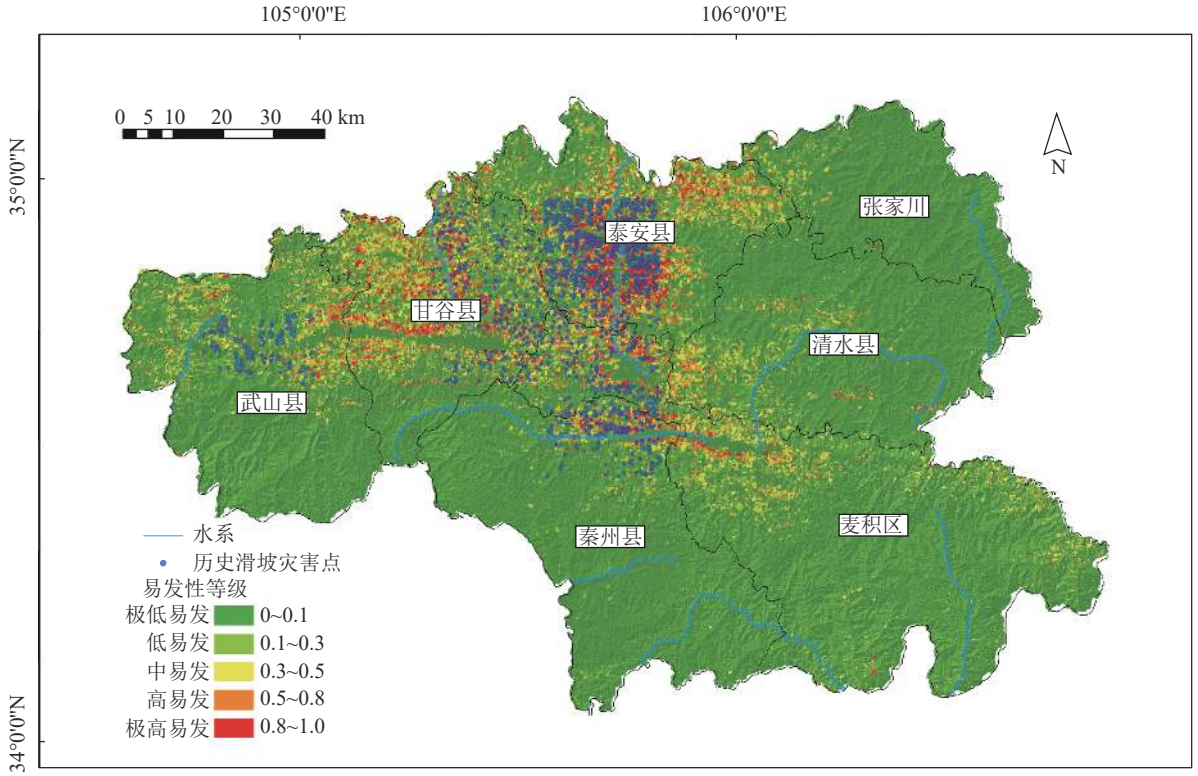

图 12　天水市滑坡易发性评价图

Fig. 12　Landslide susceptibility assessment map of Tianshui City

参考文献

Erickson N, Mueller J, Shirkov A, et al. 2020. Autogluon−tabular: Robust and accurate automl for structured data[J]. arXiv, preprint arXiv: 2003.06505.

Huang F, Ye Z, Zhou X, et al. 2022. Landslide susceptibility prediction using an incremental learning Bayesian Network model considering the continuously updated landslide inventories[J]. Bulletin of Engineering Geology and the Environment, 81(6): 250.

Ke G, Meng Q, Finley T, et al. 2017. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems, 2017: 3149−3157.

Lundberg S M, Lee S I. 2017. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems, 2017: 30.

Merghadi A, Yunus A P, Dou J, et al. 2020. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance[J]. Earth−Science Reviews, 207: 103225.

Wu C, Jia R, Qiu T, et al. 2013. Rock burst monitoring and early warning based on incremental learning method with SVM[J]. Research Journal of Information Technology, 5(4): 121−124.

Wu Y, Xu W, Yu Q, et al. 2019. Hierarchical Bayesian network based incremental model for flood prediction[C]//MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25. Springer International Publishing: 556−566.

邓念东, 李宇新, 崔阳阳, 等. 2022. 基于机器学习混合模型的滑坡易发性评价[J]. 科学技术与工程, 22(14): 5539−5547.

黄发明, 胡松雁, 闫学涯, 等. 2022. 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, 41(2): 79−90.

康孟羽, 朱月琴, 陈晨, 等. 2022. 基于多元非线性回归和 BP 神经网络的滑坡滑动距离预测模型研究[J]. 地质通报, 41(12): 2281−2289.

李世其, 段学燕, 刘燕. 2006. 一种决策树增量学习算法在故障诊断中的应用[J]. 华中科技大学学报: 自然科学版, 34(4): 79−81.

李挺, 洪镇南, 刘智勇, 等. 2018. 基于增量单类支持向量机的工业控制系统入侵检测[J]. 信息与控制, 47(06): 756−761.

刘纪平, 梁恩婕, 徐胜华, 等. 2022. 顾及样本优化选择的多核支持向量机滑坡灾害易发性分析评价[J]. 测绘学报, 51(10): 2034−2045.

齐娜, 胡良柏. 2022. 天水盆地黄土滑坡特征与分布规律分析[J]. 甘肃科技, 38(22): 28−33.

邵葆蓉, 孙即超, 朱月琴, 等. 2020. 基于多元回归的黄土滑坡滑动距离预测模型探讨——以甘肃天水地区为例[J]. 地质通报, 39(12): 1993−2003.

王毅, 陈曦, 唐贵希, 等. 2022. 基于自动机器学习的全球尺度滑坡灾害易发性预测[J]. 资源环境与工程, 36(5): 604−613.

王洪林, 董春林, 董俊, 等. 2022. 基于支持向量机增量学习算法的高压电网短路故障位置自动识别[J]. 电气自动化, 44(4): 34−36.

武雪玲, 沈少青, 牛瑞卿. 2016. GIS 支持下应用 PSO-SVM 模型预测滑坡易发性[J]. 武汉大学学报 (信息科学版), 41(5): 665−671.

严武文. 2010, 基于粗集——神经网络的区域滑坡灾害易发性预测研究[D]. 中国地质大学硕士学位论文.

张博, 向旭, 贾俊龙, 等. 2023. 基于 LightGBM 的天然气管道周围滑坡灾害预测方法[J]. 吉林大学学报 (理学版), 61(2): 338−346.

赵泽园, 罗菲. 2020. 基于 LightGBM 模型的区域滑坡危险性评价研究[J]. 内蒙古煤炭经济, 5: 48−49.

庄维嘉, 谭文安, 林瑞钦, 等. 2022. GA-LightGBM 模型及其在车辆保险需求预测中应用[J]. 上海第二工业大学学报, 39(4): 339−346.