基于信息量和机器学习的新疆托克逊县地质灾害易发性评价*

李凤霞[1]，刘桂萍[1†]，杜光辉[1]，魏 震[1]，杨焱青[2]
（1. 新疆大学 地质与矿业工程学院，新疆 乌鲁木齐 830017；2. 新疆大学 计算机科学与技术学院，新疆 乌鲁木齐 830017）

摘　要：地质灾害易发性评价对防灾减灾工作至关重要，有效的评价方法与评价模型在地质灾害易发性评估中发挥着重要作用. 本文以ArcGIS为平台，采用信息量法和机器学习方法，构建了IV-RF、IV-XGBoost、IV-CatBoost、IV-KNN四种耦合模型，对托克逊县地质灾害易发性进行评价，并利用SHAP值深入剖析最高AUC值的耦合模型，明确各影响因子对预测结果的贡献. 结果表明：在四种模型中，IV-CatBoost模型具有较高精度，其中距道路距离、距水系距离和地形起伏度是最重要的三个影响因子. 研究区地质灾害极高易发区、高易发区主要位于阿拉沟山、鱼儿沟、甘沟等沟谷沿线，阿拉沟河和乌斯通沟中下游地区，以及吐-和高速公路（G3012）甘沟段、S301沿线.

关键词：耦合模型；机器学习；信息量；易发性评价；地质灾害

DOI：10.13568/j.cnki.651094.651316. 2024.12.26.0001

中图分类号：P642　**文献标识码：**A　**文章编号：**2096-7675(2025)04-0469-016

引文格式：李凤霞，刘桂萍，杜光辉，魏震，杨焱青. 基于信息量和机器学习的新疆托克逊县地质灾害易发性评价[J]. 新疆大学学报(自然科学版中英文)，2025，42(4)：469-484.

英文引文格式：LI Fengxia，LIU Guiping，DU Guanghui，WEI Zhen，YANG Yanqing. Evaluation of geological disaster susceptibility based on information quantity and machine learning in Toksun County，Xinjiang[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English)，2025，42(4)：469-484.

Evaluation of Geological Disaster Susceptibility Based on Information Quantity and Machine Learning in Toksun County, Xinjiang

LI Fengxia[1], LIU Guiping[1], DU Guanghui[1], WEI Zhen[1], YANG Yanqing[2]
（*1. School of Geology and Mining Engineering, Xinjiang University, Urumqi Xinjiang 830017, China;*
2. School of Computer Science and Technology, Xinjiang University, Urumqi Xinjiang 830017, China）

Abstract： The evaluation of geological disaster susceptibility is crucial for disaster prevention and mitigation. It is very important to select effective evaluation methods and models for the assessment of geological disaster susceptibility. By integrating the information quantity method and machine learning to construct four coupling models, namely IV-RF, IV-XGBoost, IV-CatBoost and IV-KNN based on ArcGIS, this paper constructs coupling models for geological disaster susceptibility evaluation in Toksun County. Using SHAP values to deeply analyze the coupled model with the highest AUC value, clarify the contribution of each influencing factor to the prediction results. The results show that the IV-CatBoost model has higher accuracy among the four models. In the IV-CatBoost model, the top three factors in terms of importance are the distance from the road, the distance from the water system and the topographic relief. The extremely high and high prone areas of geological disasters are predominantly situated along the valleys of Alagou mountain, Yuergou and Gangou, the middle and lower reaches of Alagou river and Wusitonggou, the Gangou section of Turpan-Hotan expressway (G3012), and along S301 road.

Key words： coupling model; machine learning; information quantity; susceptibility evaluation; geological disaster

* **收稿日期：**2024-12-26

基金项目：新疆维吾尔自治区吐鲁番市托克逊县2023年自然灾害防治体系建设项目“新疆托克逊县地质灾害风险调查评价”（GYZB-TLFDZFZ2023-01）；新疆大学2024年国家级大学生创新训练计划项目“基于机器学习的锆石成因分类研究”（202410755020）.

作者简介：李凤霞（1999—），女，硕士生，从事地质灾害评价与防治的研究，E-mail：lfx0725a@163.com.

† **通讯作者：**刘桂萍（1982—），女，博士，副教授，主要从事构造地质学的研究，E-mail：ping0991@163.com.

0 引 言

托克逊县位于吐鲁番盆地西部，县域内道路建设、矿产资源开发、渠道修建等人类工程活动对地质环境造成了不同程度的改变及破坏，导致崩塌、泥石流等地质灾害频发，严重威胁当地居民生命财产安全. 托克逊县山区较多，很多道路沿坡脚傍山而建，其中吐-和高速公路G3012是一条连接南北疆的高速公路. G3012甘沟路段穿越了中天山东段觉罗塔格山，是吐-和高速的咽喉之地，长约70 km，道路建设时的相关活动导致区域内崩塌等灾害发育广泛. 2021年7月11日，在持续强降雨条件影响下，托克逊县辖区S301线130 km、120 km等多处发生山体滑坡、泥石流等自然灾害，大量泥石被冲至路面导致道路堵塞，造成了巨大交通安全隐患[1]；2024年6月1日，吐-和高速G3012甘沟路段72 km+500 m处发生山体塌方，导致交通中断、部分车辆滞留，严重威胁过往车乘人员的安全[2]. 因此，开展地质灾害易发性评价对当地的地质灾害防治工作及经济社会可持续发展具有重要现实意义.

地质灾害易发性评价的方法有定性和定量两类，主要考虑地形地貌、地质条件、气象水文等影响因素，评价目标为地质灾害发生的可能性和潜在位置[3−4]. 采用定性方法进行地质灾害易发性评价主要依赖于专家经验，具有一定的主观因素[5]. 而信息量、确定性系数、逻辑回归等以数理统计为主的定量方法，极大程度上克服了主观因素的影响，广泛应用于地质灾害易发性评价[6−8]. 但基于专家经验的定性方法和传统数理统计模型无法准确解释不同类型评价因子间的非线性关系，预测能力不足[9−10]. 随着人工智能技术的快速发展，采用机器学习模型进行地质灾害易发性评价的研究越来越多[11−15]，但独立的机器学习模型极易造成数据的过度拟合[16]. 近年来，许多学者不再采用单一的评价方法，而是尝试构建耦合模型进行地质灾害易发性评价，以提高模型评价的准确性[17−19]，其中数理统计与机器学习耦合模型的应用较为广泛[20−21]. 陈飞等[22]通过构建信息量与神经网络相结合的评价模型，对江西省赣州市上犹县开展滑坡易发性评价，结果表明耦合模型评价精度较高，比单一信息量模型的评价精度高5.1%. Zhao等[23]采用分形理论、信息量与随机森林混合的模型对研究区进行滑坡易发性评价，并与信息量、反向传播神经网络和模糊逻辑评价模型对比，结果表明耦合模型的评价精度较高. He等[24]构建了基于信息量、频率比模型与逻辑回归、支持向量机和随机森林的耦合模型，对云南省昭通市威信县进行滑坡易发性评价，结果表明耦合模型的精度普遍高于单一模型.

托克逊县气候条件与灾害发生时机与以往研究有较大不同，为克服定性评价的主观认识不足以及检验人工智能方法的适用性，本文采用信息量与机器学习的耦合模型首次对托克逊县开展地质灾害易发性评价研究. 主要通过ROC曲线（Receiver Operating Characteristic Curves）对模型进行比较与验证，确定适用于研究区地质灾害易发性评价的模型，为该地区开展地质灾害防治工作提供参考.

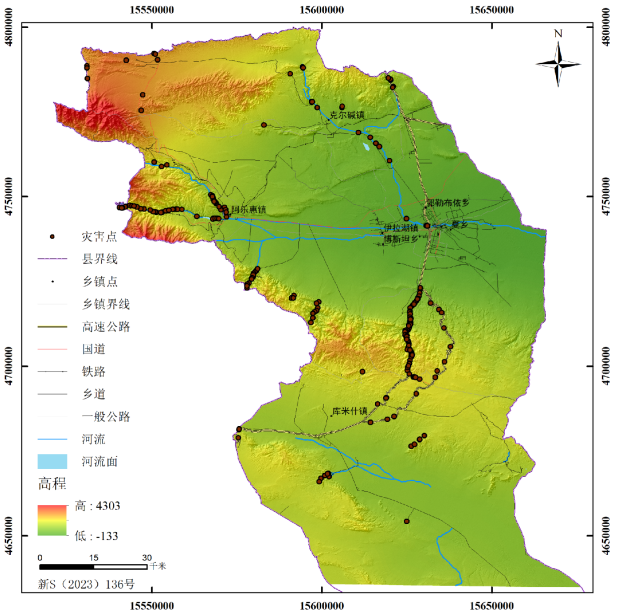

图 1 研究区概况

1　研究区概况

托克逊县位于中国新疆维吾尔自治区中东部，天山南麓，吐鲁番盆地西部，地理坐标位于东经87°23′～89°15′、北纬41°19′～43°19′之间；地质构造上位于哈萨克斯坦板块的准噶尔微板块、伊犁微板块以及塔里木古陆板块的对接带部位及其附近；托克逊县三面环山，整体呈南北走向的长条形，地形地貌为三山夹两洼，主要由侵蚀剥蚀作用形成的中山地貌单元构成，地势西北高、东南低. 山区海拔高程约为+1 000～+4 303 m，绿洲海拔高程约为-200～+200 m，东部最低海拔高程约为-133 m，研究区概况如图1所示.

研究区内地层出露较为完整，包括古生界志留系、泥盆系、石炭系、二叠系、中生界三叠系、侏罗系和新生界古近系、新近系、第四系. 其中泥盆系、石炭系和第四系的地层分布最广. 泥盆系岩性主要为砂岩、凝灰岩、凝灰砂岩夹碧玉岩、凝灰质粉砂岩. 石炭系岩性主要为灰岩夹砂岩、粉砂岩、砾岩等. 二叠系岩性主要为凝灰岩、安山岩、钙质砂岩、砾岩和泥质粉砂岩. 第四系在山前平原地带、河谷地区广泛分布，岩性主要为砂卵砾石、砂砾石、碎屑砂土、细砂和中细砂等. 研究区内侵入岩主要为华力西中期侵入岩，以中粒花岗岩和黑云母花岗岩为主.

2　数据和方法

2.1　研究区数据

本文基础数据来源于新疆维吾尔自治区“托克逊县地质灾害风险调查评价”项目（表1）. 经统计，研究区内地质灾害点包括崩塌139处、泥石流77处、滑坡1处. 崩塌灾害点主要分布在托克逊县西部山区的阿拉沟及山前沟口处和南部山区的甘沟及山前沟口处；泥石流灾害主要分布于西部山区的阿拉沟沟谷流域以及南部山区甘沟流域和山前沟口地带. 托克逊县滑坡仅发育1处，位于阿乐惠镇. 研究区所使用栅格数据的像元大小统一重采样至15 m×15 m.

表 1　基础数据来源

数据名称	数据来源	数据格式
历史灾害点	托克逊县自然资源局	矢量
高程	地理空间数据云	栅格
地层岩性	1∶20万和1∶5万区域地质图（托克逊县自然资源局）	矢量
断层	1∶20万和1∶5万区域地质图（托克逊县自然资源局）	矢量
河流	Open Street Map	矢量
年均降雨量	国家气象科学数据中心	栅格
归一化植被指数NDVI	国家科技基础条件平台-国家生态科学数据中心	栅格
道路	Open Street Map	矢量

2.2　研究方法

2.2.1　信息量

信息量法是地质灾害易发性评价领域中一种比较成熟且应用十分广泛的模型[25]. 信息量模型反映了在一定地质环境条件下，多种影响因素组合对地质灾害易发性程度的贡献，信息量值的大小代表地质灾害发生的可能性，信息量值越高，地质灾害发生的概率越大，易发性程度越高. 某影响因素在特定分级状态下的信息量计算方法为

$$I(x_i,X)=\ln\frac{N_i/N}{S_i/S}, \tag{1}$$

式中：x_i为影响因素X对应的第i类分级区间；N_i为在x_i区间内发育的地质灾害数量；N为研究区地质灾害发育的总数量；S_i为在x_i区间内所对应的栅格单元数量；S为研究区栅格单元总数量.

将不同影响因素分级状态下对应的信息量值进行叠加，得到总的信息量值[26]，即

$$I_{总}=\sum_{i=1}^{n}I(x_i,X), \tag{2}$$

式中：$I_{总}$为某影响因素X内地质灾害发生的总信息量；n为影响因素分级数量.

2.2.2 随机森林

随机森林（Random Forest, RF）是一种基于决策树的集成学习算法，通过自助采样法从原始训练数据集有放回地抽取多个子数据集，利用这些子数据集构建多棵决策树. 当对新的数据进行预测时，将新数据输入到每一棵决策树中得到预测结果，通常采用投票的方式（多数表决）确定最终的分类结果. 随机森林因其准确性高、鲁棒性好和可以并行化计算的优点，被广泛应用于地质灾害易发性评价中.

2.2.3 XGBoost

XGBoost（eXtreme Gradient Boosting）是一种高效的梯度提升树算法，以决策树作为基学习器，通过不断添加树来拟合前序模型的残差. 构建树时，采用贪心算法寻找最优分裂点，考虑了一阶和二阶导数信息，使损失函数下降更快. XGBoost算法通过正则化项控制模型复杂度，从而防止过拟合. 此外，XGBoost算法对缺失数据有独特处理机制，能自动学习最优的分裂方[27]. 训练过程中，支持并行计算特征分裂的增益，提高了训练效率，有助于更有效地利用数据特征，生成高精度的预测模型[28].

2.2.4 CatBoost

CatBoost（Categorical Boosting）是一种基于梯度提升决策树的机器学习算法，通过迭代构建对称树结构的决策树来最小化损失函[29]. CatBoost算法凭借在处理类别型特征方面的高效率和高稳定性，以及在降低梯度偏差和预测偏移方面的创新技术，成为一个强大的机器学习工具，适用于多种复杂的分类预测任务.

2.2.5 KNN

KNN（K-Nearest Neighbors）是一种基于实例的学习算法，依据邻近性原则计算最近邻点，并从这些最近邻的类别中选择投票数最高的类来决定新数据点的类别. KNN算法简单易懂，不需要训练模型参数，就可以适应不同的数据分布. 无论是线性还是非线性的数据分布，只要有充足的数据，KNN算法就可以进行较好的分类或预测[30].

2.2.6 信息量耦合机器学习模型

信息量耦合机器学习模型通过信息量法计算不同影响因子各区间IV值，利用ArcGIS软件提取样本点处各影响因子的IV值. 将IV值作为机器学习模型的输入变量，构建托克逊县地质灾害易发性评价的耦合模型，具体评价流程如图2所示.

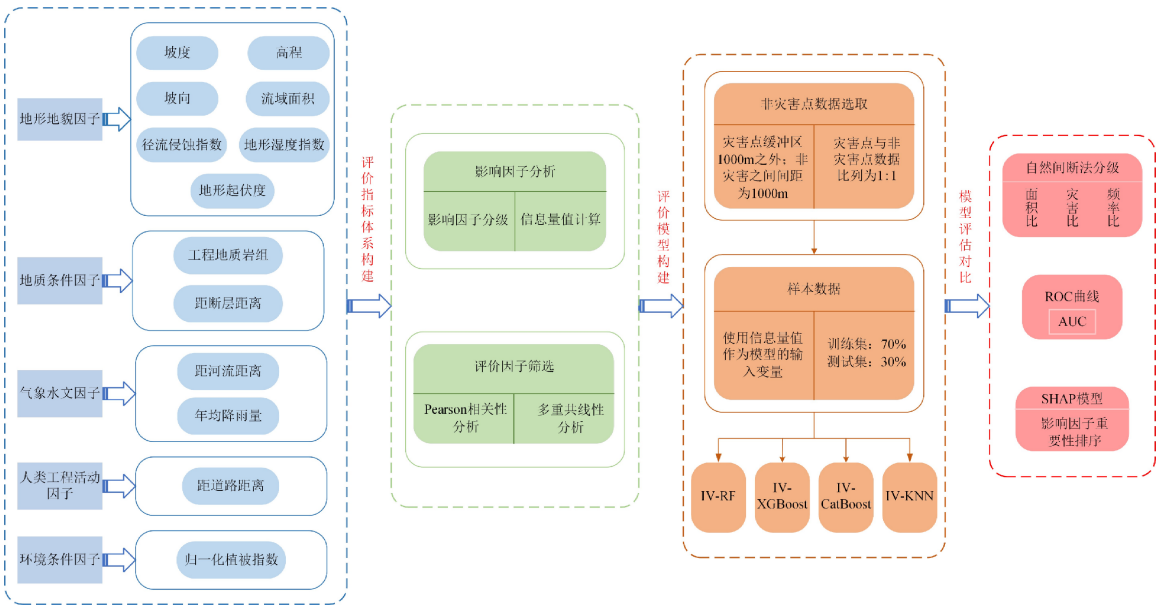

图 2 耦合模型的地质灾害易发性评价流程

3 地质灾害影响因子分析

3.1 因子分级与信息量值

利用ArcGIS中重分类工具，按自然间断法对研究区影响因子进行分级（图3），统计灾害点分布与各影响因子分级关系，并通过式（1）计算各影响因子分级的信息量值（表2）.

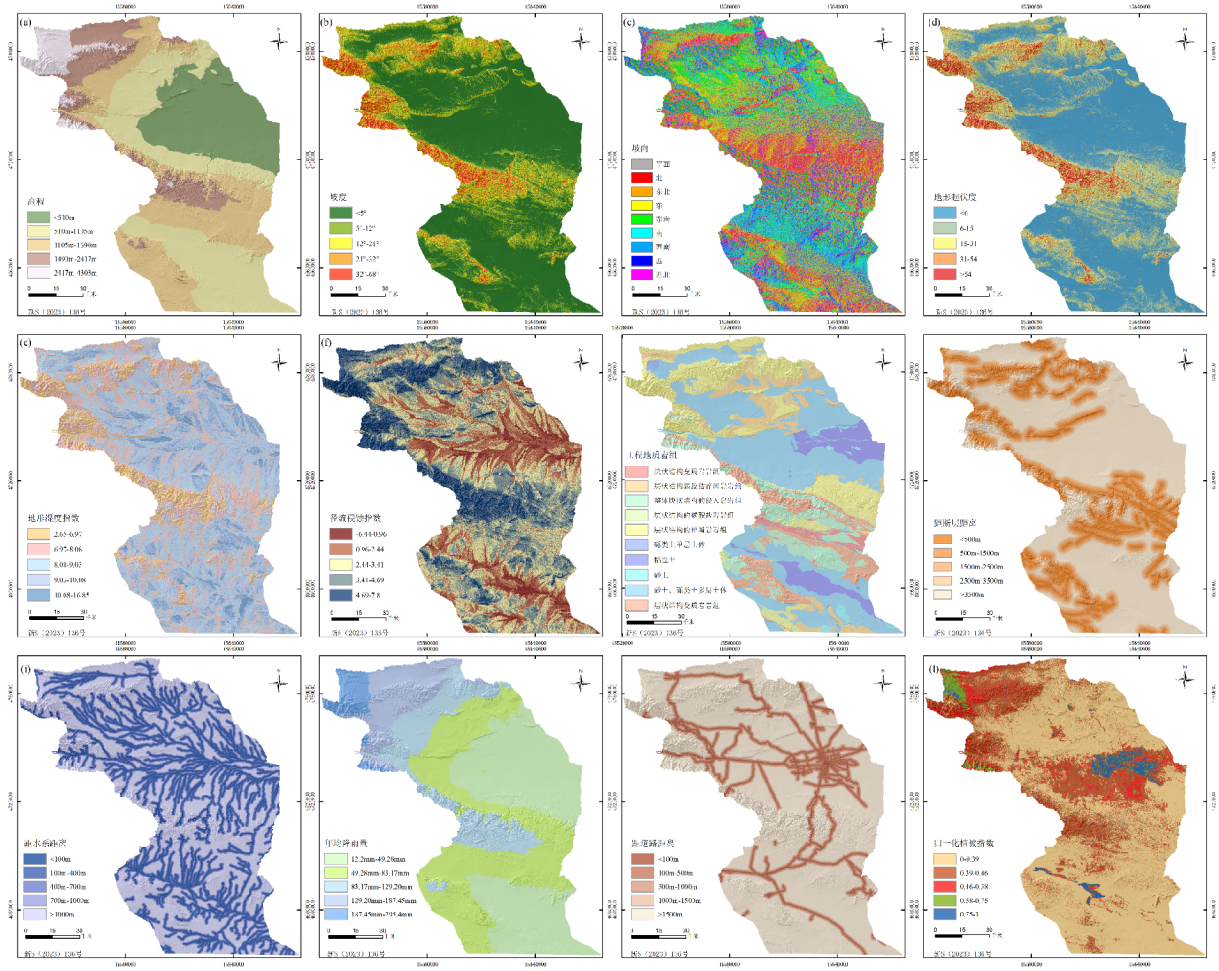

图 3　地质灾害影响因子分级

表 2　影响因子信息量值

评价因子	评价因子分级	N_i/N	S_i/S	信息量值
高程/m	<510	0.078 3	0.221 2	-1.038 1
	[510,1 105)	0.442 4	0.318 6	0.328 3
	[1 105,1 690)	0.405 5	0.282 7	0.360 7
	[1 690,2 417)	0.041 5	0.123 1	-1.088 1
	[2 417,4 303]	0.032 3	0.054 3	-0.521 2
坡度/°	<4.77	0.124 4	0.622 6	-1.610 2
	[4.77,11.67)	0.373 3	0.189 2	0.679 6
	[11.67,20.95)	0.267 3	0.091 4	1.073 0
	[20.95,32.09)	0.138 2	0.060 0	0.834 4
	[32.09,67.63]	0.096 8	0.036 8	0.966 3
坡向	东	0.124 4	0.139 6	-0.114 8
	东北	0.179 7	0.162 2	0.102 7
	东南	0.059 9	0.152 5	-0.934 3
	北	0.147 5	0.113 4	0.262 7
	南	0.069 1	0.145 2	-0.742 3
	平面	0.000 0	0.030 8	0.000 0
	西	0.152 1	0.068 9	0.792 4
	西北	0.156 7	0.076 2	0.720 5
	西南	0.110 6	0.111 3	-0.006 0

续表 2

评价因子	评价因子分级	N_i/N	S_i/S	信息量值
地形起伏度	<6	0.069 1	0.534 0	-2.044 5
	$[6,15)$	0.350 2	0.245 3	0.356 1
	$[15,31)$	0.336 4	0.120 7	1.025 0
	$[31,54)$	0.156 7	0.069 9	0.807 7
	≥ 54	0.087 6	0.030 1	1.066 5
地形湿度指数TWI	$[2.65,6.97)$	0.465 4	0.101 3	1.525 1
	$[6.97,8.08)$	0.308 8	0.240 2	0.250 9
	$[8.08,9.03)$	0.175 1	0.297 8	-0.530 9
	$[9.03,10.08)$	0.050 7	0.264 9	-1.653 6
	$[10.08,16.85]$	0.000 0	0.095 8	0.000 0
径流侵蚀指数SPI	[-6.44,0.96)	0.004 6	0.079 3	-2.845 4
	$[0.96,2.24)$	0.078 3	0.204 6	-0.959 8
	$[2.24,3.41)$	0.198 2	0.281 2	-0.350 0
	$[3.41,4.69)$	0.483 9	0.265 7	0.599 5
	$[4.69,7.80]$	0.235 0	0.169 3	0.328 3
工程地质岩组	层状结构的碎屑岩岩组	0.290 3	0.200 2	0.371 6
	整体块状结构的侵入岩岩组	0.290 3	0.098 9	1.076 4
	层状结构弱胶结碎屑岩岩组	0.055 3	0.091 7	-0.505 5
	黏性土	0.000 0	0.079 4	0.000 0
	砂土、砾类土多层土体	0.198 2	0.390 9	-0.679 4
	层状结构变质岩岩组	0.156 7	0.104 9	0.401 2
	砂土	0.000 0	0.008 6	0.000 0
	层状结构的碳酸盐岩岩组	0.004 6	0.006 3	-0.312 2
	块状结构变质岩岩组	0.004 6	0.016 6	-1.283 3
	砾类土单层土体	0.000 0	0.002 4	0.000 0
距断层距离/m	<500	0.161 3	0.093 2	0.548 1
	$[500,1\ 500)$	0.331 8	0.160 8	0.724 5
	$[1\ 500,2\ 500)$	0.207 4	0.131 1	0.458 7
	$[2\ 500,3\ 500)$	0.142 9	0.106 4	0.294 7
	$\geq 3\ 500$	0.156 7	0.508 5	-1.177 3
距水系距离/m	<100	0.635 9	0.074 5	2.143 8
	$[100,400)$	0.230 4	0.194 9	0.167 5
	$[400,700)$	0.018 4	0.153 6	-2.120 1
	$[700,1\ 000)$	0.036 9	0.125 8	-1.227 1
	$\geq 1\ 000$	0.078 3	0.451 2	-1.750 9
年均降雨量/mm	$[12.20,49.28)$	0.110 6	0.359 5	-1.178 8
	$[49.28,83.17)$	0.460 8	0.331 1	0.330 7
	$[83.17,129.20)$	0.327 2	0.158 0	0.728 0
	$[129.20,187.45)$	0.064 5	0.111 8	-0.549 7
	$[187.45,295.40]$	0.036 9	0.039 7	-0.073 1
距道路距离/m	<100	0.769 6	0.026 6	3.366 1
	$[100,500)$	0.041 5	0.085 3	-0.721 4
	$[500,1\ 000)$	0.055 3	0.086 9	-0.451 7
	$[1\ 000,1\ 500)$	0.013 8	0.074 2	-1.680 5
	$\geq 1\ 500$	0.119 8	0.727 0	-1.803 0

续表 2

评价因子	评价因子分级	N_i/N	S_i/S	信息量值
归一化植被指数NDVI	[0,0.39)	0.539 2	0.631 9	-0.158 7
	[0.39,0.46)	0.258 1	0.235 5	0.091 4
	[0.46,0.58)	0.119 8	0.074 0	0.482 6
	[0.58,0.75)	0.059 9	0.032 6	0.608 9
	[0.75,1]	0.023 0	0.026 0	-0.122 2

3.2　影响因子分析

3.2.1　地形地貌因子

1）高程是反映地貌特征的一个重要指标，不同高程地区的植被类型及覆盖率、降雨情况、人类工程活动等因素存在差异. 研究区内灾害点主要集中分布在高程为510~1 105 m和1 105~1 690 m两个区间内（图4(a)），占灾害点总数的84.79%. 510~1 690 m高程区间内，信息量值和灾害点密度较高，因为该区间内人类工程活动频繁，岩石风化强烈，从而导致地质灾害发育较多.

2）坡度直接关系到地表物质的稳定性和坡体的承受力，与土壤稳定性、水文条件、植被覆盖率等多种因素相关[31]. 研究区内灾害点在坡度为5°~12°区间分布最多（图4(b)），占灾害点总数的37.33%. 坡度为12°~21°区间内，灾害点密度和信息量值最大，因为该区间内人工修路开挖坡脚及过度采掘切坡等工程活动使边坡稳定性变差，地质灾害发生的概率增大.

3）坡向决定了地表接收阳光的角度和强度，进而影响地表的温度、水分蒸发、植被生长等环境条件. 研究区内灾害点在东北向分布最多（图4(c)），占灾害点总数的17.97%. 西向的灾害点密度和信息量值最大，灾害点密度为0.035个/平方千米，信息量值为0.792，该向引发地质灾害概率较高.

4）地形起伏度是最大高程与最小高程的差值，可以表示局部小区域的地形起伏情况. 研究区内灾害点在地形起伏度6~15区间分布最多（图4(d)），占灾害点总数的35.02%. 信息量值和灾害点密度在地形起伏度大于等于54区间达到最大值，表明地形起伏度大的区域发生地质灾害的可能性较大.

5）地形湿度指数（TWI）是衡量地形对水流累积和土壤湿度影响的指标，量化了地形对水文过程的控制作用. 研究区内灾害点在TWI值2.65~6.97区间分布最多（图4(e)），占灾害点总数的46.54%，该区间内灾害点密度和信息量值均为最大.

6）径流侵蚀指数（SPI）揭示了地形对水流侵蚀能力的控制作用，SPI值高的区域，水流侵蚀能力强，反之则较弱. 研究区内SPI值为3.41~4.69区间灾害点分布最多（图4(f)），占灾害点总数的48.39%，该区间内灾害点密度和信息量值均为最大.

3.2.2　地质条件因子

1）工程地质岩组. 岩土体是地质灾害发生的物质基础，其性质和状态对地质灾害的发生和发展起着决定性作用. 研究区的灾害点主要集中分布在层状结构的碎屑岩岩组和整体块状结构的侵入岩岩组区域内（图4(g)），占灾害点总数的58.06%. 整体块状结构的侵入岩岩组区域内信息量值和灾害点密度均为最大，该岩组岩性以花岗岩、闪长岩为主，岩石裸露，长期风化作用逐渐破坏岩石的结构，导致岩石强度降低，容易发生破碎.

2）距断层距离. 断层附近地质结构复杂，岩石的破碎程度高，裂缝和节理发育，坡体完整性差，崩塌等地质灾害发育的可能性较高. 研究区内灾害点在距断层距离为500~1 500 m区间分布最多（图4(h)），占灾害点总数的33.18%，信息量值在该区间最大.

3.2.3　气象水文因子

1）距水系距离. 河流、湖泊等水系是地质灾害发生和发展的重要影响因素，河流持续的冲刷和侵蚀作用导致岸坡稳定性降低，发生崩塌、滑坡等自然地质灾害风险较高. 研究区内灾害点在距水系距离小于100 m区间分布最多（图4(i)），占灾害点总数的63.6%，该区间内灾害点密度和信息量值均为最大. 随着距离水系距离的增加，灾害点数量明显下降，灾害点密度逐渐降低.

2）年均降雨量. 持续的降雨、暴雨，可能使地下岩土体的结构发生变化. 比如软化、裂隙扩大等，降低岩土体边坡的稳定性，从而更容易发生地质灾害. 本文降雨量数据采用2018—2022年的年均降雨量. 研究区内灾害

点在年均降雨量为49.28~83.17 mm区间分布最多（图4(j)），占灾害点总数的46.08%. 年均降雨量为83.17~129.20 mm区间灾害点密度和信息量值最大.

3.2.4 人类工程活动因子

研究区内对地质灾害影响最明显的人类工程活动因子为距道路距离. 道路建设时对山地进行切割开挖，破坏原有的地质结构，使得边坡内部的应力状态发生改变，边坡稳定性降低，遇地震或暴雨时易发生地质灾害. 研究区内灾害点在距道路距离小于100 m区间分布最多（图4(k)），占灾害点总数的76.95%，信息量值和灾害点密度在此区间均为最大.

3.2.5 环境条件因子

归一化植被指数（NDVI）是一种通过比较近红外和红光波段的反射率来反映植被覆盖和生长状况的参数指标. 植被的根系能够牢固地固定土壤和岩石，提高地表的稳定性，降低泥石流等地质灾害的风险. 研究区内灾害点在NDVI为0~0.39区间分布最多（图4(l)），占灾害点总数的53.92%，在NDVI为0.58~0.75区间灾害点密度和信息量值最大.

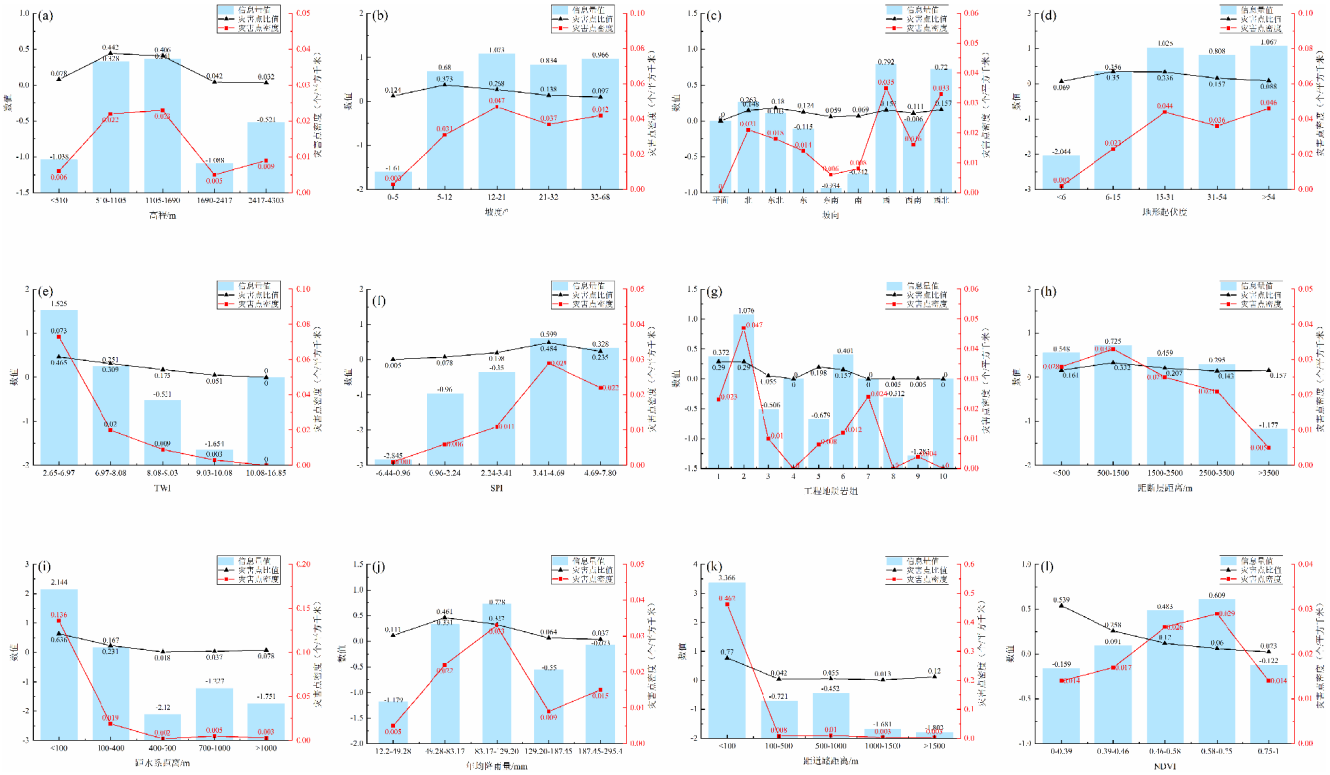

图 4 地质灾害影响因素分级信息量值与灾害点分布关系

注：(g) 的横坐标中，1为层状结构的碎屑岩岩组；2为整体块状结构的侵入岩岩组；3为层状结构弱胶结碎屑岩岩组；4为黏性土；5为砂土、砾类土多层土体；6为层状结构变质岩岩组；7为砂土；8为层状结构的碳酸盐岩岩组；9为块状结构变质岩岩组；10为砾类土单层土体

3.3 影响因子相关性与多重共线性分析

地质灾害影响因素多且关系复杂，不同影响因子之间可能存在相互作用和联系，不完全独立. 若将初步选取的12个影响因子直接输入到模型中，可能导致模型的不稳定和预测结果的不准确. 因此，本文采用皮尔逊相关系数法和多重共线性分析对初步选取的12个因子进行分析与筛选.

3.3.1 相关性分析

皮尔逊相关系数表示两个变量之间的相关程度，取值范围为[-1, 1]. 相关系数绝对值越大表示变量之间的相关性越显著，判断标准如表3所示[32].

由图5可知，坡度和地形起伏度因子的相关系数为0.865，两个因子之间存在高度相关性. 然而仅依靠皮尔逊相关系数无法确定剔除相关性较高的因子，故需结合因子多重共线性分析结果对影响因子进行筛选.

表 3　相关系数绝对值判断标准

相关系数绝对值取值范围	相关程度
[0,0.3]	极弱或无相关
(0.3,0.5]	弱相关
(0.5,0.8]	中等程度相关
(0.8,1]	高度相关

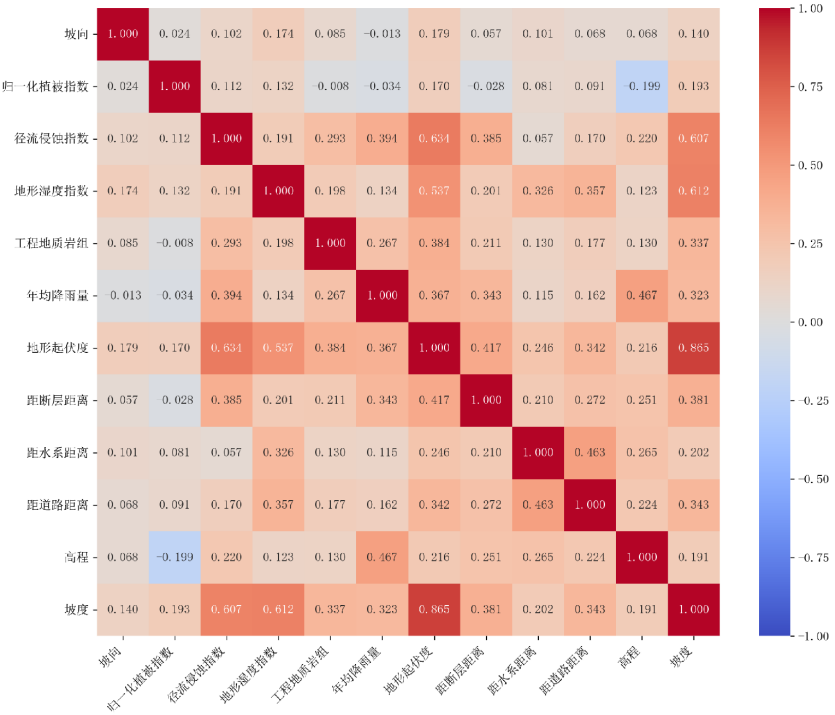

图 5　皮尔逊相关系数矩阵

3.3.2　多重共线性分析

影响因子之间存在多重共线性会导致模型预测性能下降. 方差膨胀因子（VIF）可以表示两种或两种以上变量之间的相互依赖关系[33]. VIF值越高，因子间的多重共线性程度越严重. 通常认为当VIF>5时，因子之间的多重共线性将对模型性能造成较大影响[34]. 由表4可知，12个影响因子的VIF值都不超过5，故影响因子间不存在高度多重共线性问题. 结合皮尔逊相关系数分析结果，剔除VIF值较高的坡度因子.

表 4　多重共线性分析结果

影响因子	VIF
高程	1.466
坡向	1.062
工程地质岩组	1.214
NDVI	1.138
TWI	1.907
SPI	2.031
距断层距离	1.357
距道路距离	1.436
距水系距离	1.429
地形起伏度	4.689
坡度	4.954
年均降雨量	1.550

4 地质灾害易发性评价

4.1 样本数据构建

以研究区217处灾害点为正样本，标签记为1. 非灾害点选取区域为灾害点缓冲区1 000 m之外，且非灾害点之间间距为1 000 m. 使用ArcGIS中创建随机点工具按照1∶1的比例随机提取217个非灾害点作为负样本，标签记为0. 提取样本点处各评价因子的信息量值组成总的样本数据，将其按照7∶3的比例划分为训练数据集和测试数据集.

4.2 易发性分区

将训练好的IV-RF、IV-XGBoost、IV-CatBoost和IV-KNN耦合模型分别用于预测研究区中每个栅格单元发生地质灾害的概率. 使用ArcGIS软件中的重分类工具，按自然间断法将预测结果划分为四个等级，得到耦合模型的地质灾害易发性评价区划（图6）.

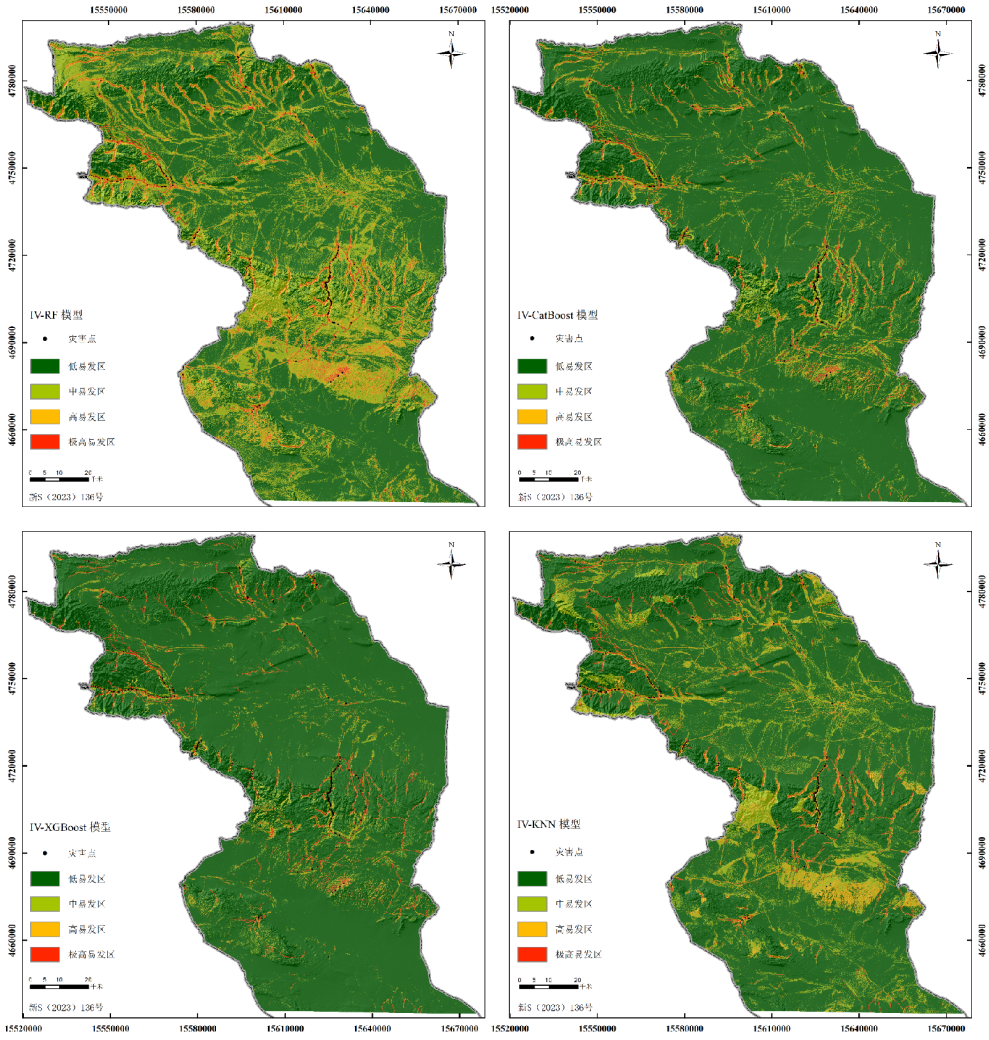

图 6 耦合模型的地质灾害易发性评价区划

四种耦合模型的易发性评价结果显示，极高易发区和高易发区主要集中分布在西北部中山区阿拉沟、鱼儿沟、北东部中山区G30线后沟段和中部觉罗塔格山G3012线甘沟段. 整体而言，四种模型的易发性分区结果较为相似，极高易发区和高易发区位于侵蚀、剥蚀作用强烈的中山区及山前地带. 区域内地形陡峭，沟谷切割较深，裸露基岩风化强烈，为构造上升强烈地段，沟谷松散堆积物丰富，公路沿线易形成崩塌、泥石流灾害. 由表5可知，四种模型中IV-XGBoost的极高易发区面积最小，IV-RF的极高易发区面积最大；IV-XGBoost和IV-KNN的低易发区灾害点数量相同，IV-KNN的极高易发区内灾害点数量最少；研究区灾害点大多分布在高易发区和极高易发区，极高易发区的灾害点数量最多，随着易发性等级的提升，各模型的频率比呈逐渐上升趋势.

表 5　耦合模型的易发性分区结果统计

评价模型	易发性分区	面积/km^2	面积占比/%	灾害点数量/个	灾害点占比/%	频率比
IV-RF	低	10 084.70	74.08	1	0.46	0.006
	中	2 262.95	16.62	2	0.92	0.055
	高	943.14	6.93	2	0.92	0.133
	极高	323.11	2.37	212	97.70	41.224
IV-CatBoost	低	11 643.05	85.52	2	0.92	0.011
	中	1 336.21	9.82	1	0.46	0.047
	高	321.89	2.36	3	1.38	0.585
	极高	312.75	2.30	211	97.24	42.278
IV-XGBoost	低	12 425.62	91.27	3	1.38	0.015
	中	655.69	4.82	1	0.46	0.095
	高	220.26	1.62	3	1.38	0.852
	极高	312.33	2.29	210	96.78	42.262
IV-KNN	低	10 530.12	77.35	3	1.38	0.018
	中	2 060.50	15.14	5	2.30	0.152
	高	707.04	5.19	15	6.91	1.331
	极高	316.24	2.32	194	89.41	38.539

注：频率比为灾害点占比与面积占比的比值

4.3　模型精度评价及验证

受试者工作特征曲线（ROC）是被广泛用于评估分类模型性能的一种图形化工具，其X轴表示假正例率、Y轴表示真正例率[35−36]．AUC值表示ROC曲线右下方区域的面积，其值与1越接近表明模型的预测性能越好．由图7可知，IV-KNN模型的AUC值最低，模型预测结果较差，IV-CatBoost模型的AUC值最高，模型预测性能表现最好．因此，本节基于IV-CatBoost模型得到的地质灾害易发性评价结果进行模型验证．

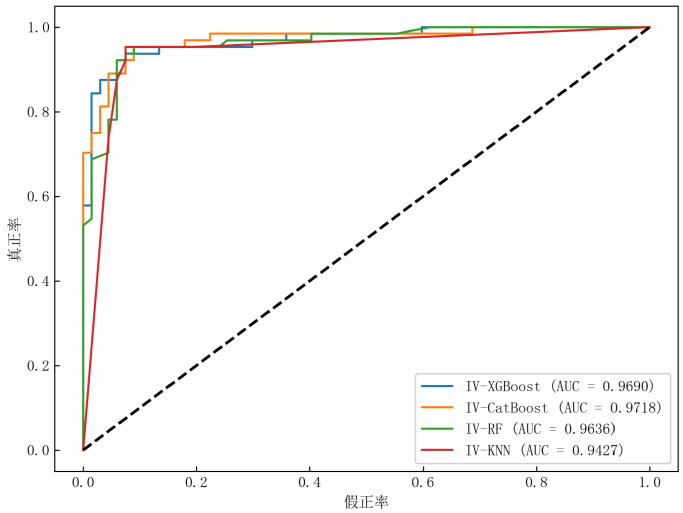

图 7　信息量耦合模型的ROC曲线

将IV-CatBoost模型的易发性评价区划结果与野外调查（图9(a)中A、B为野外验证区域）得到的灾害点进行验证（图8）．野外调查中，阿拉沟托克逊段的阿乐惠镇上游是泥石流发育地段，分布泥石流灾害点18处，称为阿拉沟泥石流群（图8(a∼c)，图9(b)）．典型案列为N53泥石流（图8(c)），据调查该处曾因暴雨引发山洪（泥石流），造成2人伤亡、1人失联．吐-和高速公路（G3012）甘沟段道路沿坡脚傍山而建，坡上岩体节理裂隙较发育，人工切坡修路破坏了斜坡的整体稳定性，发育了多处岩质崩塌（图8(d∼f)，图9(c)），其形成机制主要为地形切割造成的高陡边坡在卸荷过程中经历应力重分布，导致边坡卸荷区域产生拉张裂缝．这些裂缝与原有的构造裂隙和层理结构面等其他裂隙组合，逐渐形成潜在的危岩体．在震动、降水及风化等外界因素的影响下，危岩体会突然脱离母体发生崩塌，对坡脚路过的行人及车辆造成威胁．

图 8 野外灾害点图集

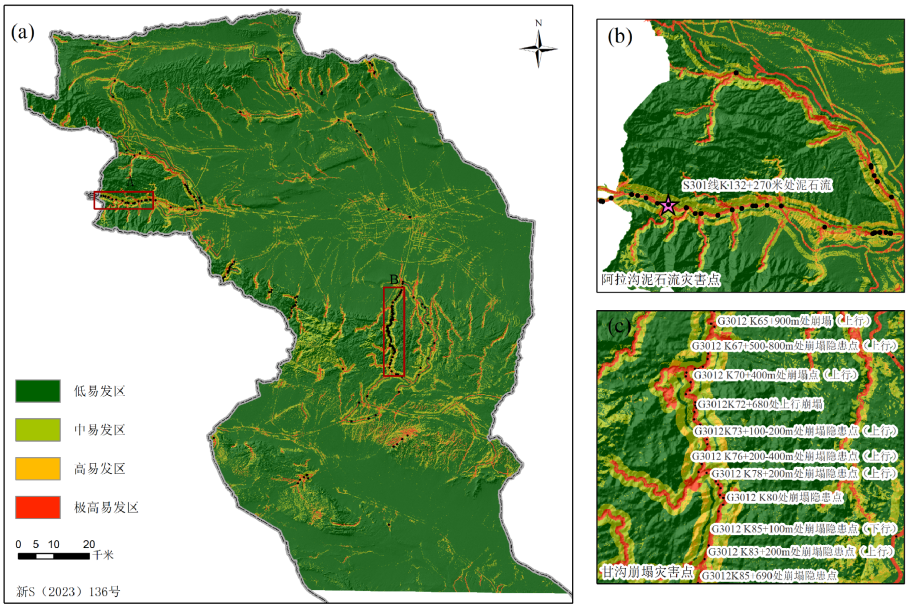

图 9 IV-CatBoost模型野外验证区域

4.4 最优模型的SHAP值解释

SHAP方法由Lundberg和Lee在2017年提出，是一个强大的模型解释工具，旨在提高机器学习模型的可解释性[37−38]. 本文通过SHAP方法中的汇总图对预测性能最优的IV-CatBoost模型进行深入分析.

研究区地质灾害汇总如图10所示，每个点为一个样本，横轴表示SHAP值大小，纵轴表示影响因子. 点的颜色从蓝到红，表示影响因子的值从小到大. 点位越靠近右侧，表示影响因子SHAP值越大，从而对地质灾害的正向贡献度越大. 由影响因子SHAP值随因子值的变化过程可知，随着影响因子值逐渐增大，其SHAP值逐渐变大，对地质灾害的正向贡献度也越大，表明信息量值高的区域易发生地质灾害.

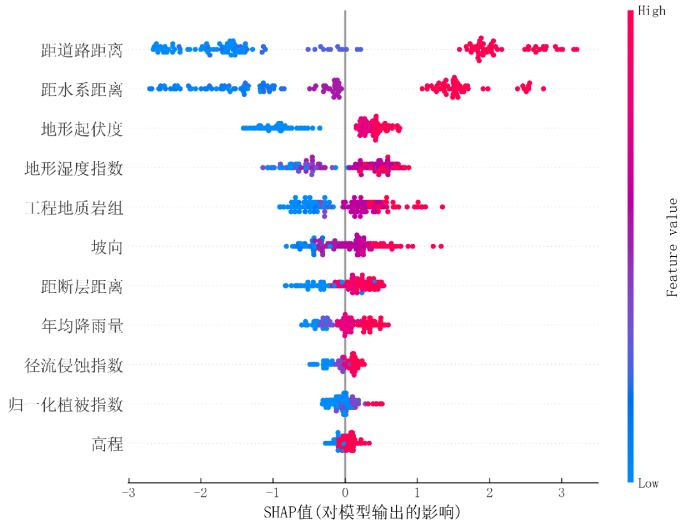

图 10　研究区地质灾害重要影响因子的SHAP值贡献度分布

对图10中每个影响因子所有样本的SHAP值求平均值，得到地质灾害影响因子重要性排序图[39]. 由图11可知，对研究区地质灾害影响最大的三个因子为距道路距离、距水系距离和地形起伏度. 距道路距离信息量值最大的为小于100 m区间，此区间人类工程活动频繁，道路修建、采矿活动等导致道路两侧岩体遭到破坏，地质环境较为脆弱. 公路沿线的人工切坡在路基一侧形成高陡的人工边坡，坡体稳定性较低，极易形成崩塌等地质灾害. 山区地形狭窄，崩塌灾害的发生及开挖道路堆积的松散物又为泥石流灾害的发生提供了物源条件，在雨季极易引发泥石流灾害，威胁过往车辆及行人的安全. 距水系距离信息量值最大的为小于100 m区间，该区间河流的侵蚀作用会削弱河岸的稳定性，导致河岸坍塌等灾害. 此外，修建水库、河道整治等人类活动在该区间较为集中，这些活动会破坏自然河岸结构，从而增加地质灾害的风险. 中部平原白杨河沿河两岸区域地表被厚度较厚的第四系沉积物覆盖，其河岸常年受上游支流洪水的冲蚀，易发生崩塌地质灾害. 地形起伏度信息量值最大的为大于等于54区间，该区间地貌类型为侵蚀剥蚀作用形成的中山区，植被覆盖率低，土壤易被侵蚀. 区间内地表高度差异显著，坡度较陡，岩石和土壤整体稳定性较低，使得崩塌和泥石流等灾害的发育概率较高.

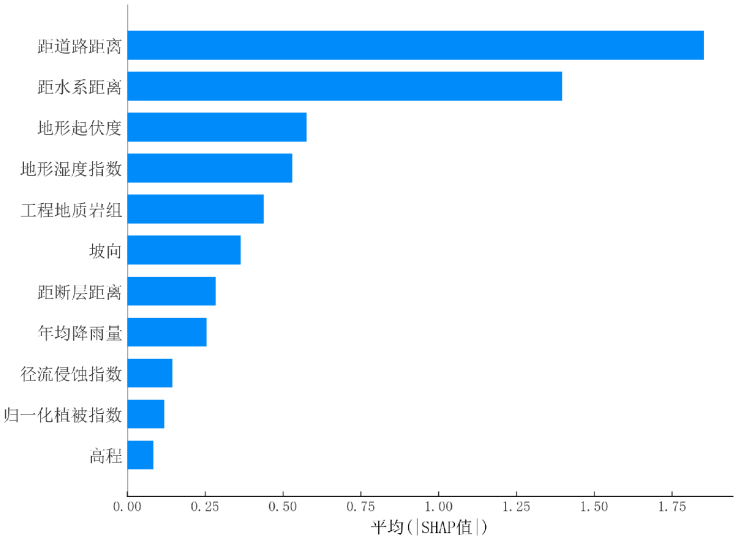

图 11　研究区地质灾害影响因子重要性排序图

5　结 论

1）通过分析研究区地质灾害点分布与影响因子分级信息量值的关系，得出灾害点在以下区间内信息量值最大：高程为1 105~1 690 m、坡度为12°~21°、坡向为西方向、地形起伏度为大于等于54、TWI为2.65~6.97、SPI为

3.41～4.69、工程地质岩组为整体块状结构的侵入岩岩组、距断层距离为500～1 500 m、距水系距离小于100 m、年均降雨量为83.17～129.20 mm、距道路距离小于100 m、NDVI为0.58～0.75.

2）根据ROC曲线对IV-RF、IV-XGBoost、IV-CatBoost和IV-KNN模型进行精度检验，四种耦合模型的AUC值都大于0.9，均可作为托克逊县地质灾害易发性评价模型，其中IV-CatBoost模型性能表现最佳. 研究区地质灾害极高易发区、高易发区主要集中在阿乐惠镇、博斯坦乡西南部、夏乡西南部和克尔碱镇北东部.

3）通过SHAP模型对IV-CatBoost模型进行深入分析，距道路距离、距水系距离和地形起伏度是控制研究区地质灾害分布的前三个重要因子，人类工程活动因子对研究区地质灾害的发生影响显著.

本文收集的评价因子有限，虽经过比较分析，但评价模型仍可能存在一定的偶然性. 在今后的工作中，可考虑进一步增加评价因子以及优化评价模型.

参考文献：

[1] 李昊轩. 暴雨致山体泥石流 交警抢险排洪保畅通[EB/OL]. (2021-07-12) [2024-10-22]. http://xj.cnr.cn/xjfw_1/dzxw/20210712/t20210712_525532812.shtml.
LI H X. Debris flow caused by rainstorm, traffic police rescue and flood discharge ensure smooth operation[EB/OL]. (2021-07-12) [2024-10-22]. http://xj.cnr.cn/xjfw_1/dzxw/20210712/t20210712_525532812.shtml. (in Chinese)

[2] 丁骁. 吐和高速G3012线甘沟路段发生山体塌方 造成交通中断[EB/OL]. (2024-06-02) [2024-10-22]. https://news.cnr.cn/native/gd/20240602/t20240602_526726888.shtml.
DING X. A landslide occurred on the Gangou section of the Turpan-Hotan expressway G3012, causing traffic interruption[EB/OL]. (2024-06-02) [2024-10-22]. https://news.cnr.cn/native/gd/20240602/t20240602_526726888.shtml. (in Chinese)

[3] HUANG F M, ZHANG J, ZHOU C B, et al. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction[J]. Landslides, 2020, 17: 217-229.

[4] SEGONI S, PAPPAFICO G, LUTI T, et al. Landslide susceptibility assessment in complex geological settings: Sensitivity to geological information and insights on its parameterization[J]. Landslides, 2020, 17: 2443-2453.

[5] 李志，陈宁生，侯儒宁，等. 基于机器学习的伊犁河谷黄土区泥石流易发性评估[J]. 中国地质灾害与防治学报，2024，35(3)：129-140.
LI Z, CHEN N S, HOU R N, et al. Susceptibility assessment of debris flow disaster based on machine learning models in the loess area along Yili Valley[J]. The Chinese Journal of Geological Hazard and Control, 2024, 35(3): 129-140. (in Chinese)

[6] 胡涛，樊鑫，王硕，等. 基于逻辑回归模型和3S技术的思南县滑坡易发性评价[J]. 地质科技通报，2020，39(2)：113-121.
HU T, FAN X, WANG S, et al. Landslide susceptibility evaluation of Sinan County using logistics regression model and 3S technology[J]. Bulletin of Geological Science and Technology, 2020, 39(2): 113-121. (in Chinese)

[7] 魏雨. 西南山区高速公路沿线地质灾害易发性评价：以攀枝花至盐源高速为例[D]. 成都：成都理工大学，2020.
WEI Y. Geological disaster susceptibility assessment along expressway in southwest mountain area: A case study of Panzhihua-Yanyuan expressway[D]. Chengdu: Chengdu University of Technology, 2020. (in Chinese)

[8] 温鑫，范宣梅，陈兰，等. 基于信息量模型的地质灾害易发性评价：以川东南古蔺县为例[J]. 地质科技通报，2022，41(2)：290-299.
WEN X, FAN X M, CHEN L, et al. Susceptibility assessment of geological disasters based on an information value model: A case of Gulin County in southeast Sichuan[J]. Bulletin of Geological Science and Technology, 2022, 41(2): 290-299. (in Chinese)

[9] 詹华炜. 基于机器学习的崩塌滑坡易发性评价方法研究：以广州市白云区为例[D]. 广州：广州大学，2023.
ZHAN H W. Study on evaluation method of landslide susceptibility based on machine learning: A case study of Baiyun district, Guangzhou[D]. Guangzhou: Guangzhou University, 2023. (in Chinese)

[10] 黄军朋，张紫昭. 基于机器学习的高寒山区矿山地质灾害易发性研究[J]. 中国矿业大学学报，2024，53(5)：960-976.
HUANG J P, ZHANG Z Z. Study on the mining geological hazard susceptibility assessment in alpine areas using machine learning models[J]. Journal of China University of Mining & Technology, 2024, 53(5): 960-976. (in Chinese)

[11] LING X, ZHU Y Q, MING D P, et al. Feature engineering of geohazard susceptibility analysis based on the random forest algorithm: Taking Tianshui City, Gansu Province, as an example[J]. Remote Sensing, 2022, 14: 5658.

[12] 田乃满，兰恒星，伍宇明，等. 人工神经网络和决策树模型在滑坡易发性分析中的性能对比[J]. 地球信息科学学报，2020，22(12)：2304-2316.

TIAN N M，LAN H X，WU Y M，et al. Performance comparison of BP artificial neural network and CART decision tree model in landslide susceptibility prediction[J]. Journal of Geo-Information Science，2020，22(12)：2304-2316. (in Chinese)

[13] 王毅，方志策，牛瑞卿，等. 基于深度学习的滑坡灾害易发性分析[J]. 地球信息科学学报，2021，23(12)：2244-2260.
WANG Y，FANG Z C，NIU R Q，et al. Landslide susceptibility analysis based on deep learning[J]. Journal of Geo-Information Science，2021，23(12)：2244-2260. (in Chinese)

[14] USTA Z，AKINCI H，AKIN A T. Comparison of tree-based ensemble learning algorithms for landslide susceptibility mapping in Murgul (Artvin)，Turkey[J]. Earth Science Informatics，2024，17：1459-1481.

[15] YAN M Q，YANG J R，NI X Y，et al. Urban waterlogging susceptibility assessment based on hybrid ensemble machine learning models：A case study in the metropolitan area in Beijing，China[J]. Journal of Hydrology，2024，630：130695.

[16] 孔嘉旭，庄建琦，彭建兵，等. 基于信息量和卷积神经网络的黄土高原滑坡易发性评价[J]. 地球科学，2023，48(5)：1711-1729.
KONG J X，ZHUANG J Q，PENG J B，et al. Evaluation of landslide susceptibility in Chinese Loess Plateau based on Ⅳ-RF and Ⅳ-CNN coupling models[J]. Earth Science，2023，48(5)：1711-1729. (in Chinese)

[17] 饶姗姗，冷小鹏. 基于RSIV-RF模型的凉山州泥石流易发性评价[J]. 地质科技通报，2024，43(1)：275-287.
RAO S S，LENG X P. Debris flow susceptibility evaluation of Liangshan Prefecture based on the RSIV-RF model[J]. Bulletin of Geological Science and Technology，2024，43(1)：275-287. (in Chinese)

[18] 张钟远，邓明国，徐世光，等. 镇康县滑坡易发性评价模型对比研究[J]. 岩石力学与工程学报，2022，41(1)：157-171.
ZHANG Z Y，DENG M G，XU S G，et al. Comparison of landslide susceptibility assessment models in Zhenkang County，Yunnan Province，China[J]. Chinese Journal of Rock Mechanics and Engineering，2022，41(1)：157-171. (in Chinese)

[19] 谢维安，谷士飞，向星多，等. 基于信息量与多模型耦合的碎屑岩区滑坡易发性分区评价[J]. 自然灾害学报，2023，32(1)：236-244.
XIE W A，GU S F，XIANG X D，et al. Zoning evaluation of landslide susceptibility in clastic rock areas based on information and multi-model coupling[J]. Journal of Natural Disasters，2023，32(1)：236-244. (in Chinese)

[20] 黄立鑫，郝君明，李旺平，等. 基于RBF神经网络-信息量耦合模型的滑坡易发性评价：以甘肃岷县为例[J]. 中国地质灾害与防治学报，2021，32(6)：116-126.
HUANG L X，HAO J M，LI W P，et al. Landslide susceptibility assessment by the coupling method of RBF neural network and information value：A case study in Min Xian，Gansu Province[J]. The Chinese Journal of Geological Hazard and Control，2021，32(6)：116-126. (in Chinese)

[21] 李凯新. 信息量耦合机器学习模型的西山煤田滑坡易发性评价[D]. 太原：太原理工大学，2023.
LI K X. Evaluation of landslide susceptibility of Xishan coalfield by information value coupled with machine learning models[D]. Taiyuan：Taiyuan University of Technology，2023. (in Chinese)

[22] 陈飞，蔡超，李小双，等. 基于信息量与神经网络模型的滑坡易发性评价[J]. 岩石力学与工程学报，2020，39(S1)：2859-2870.
CHEN F，CAI C，LI X S，et al. Evaluation of landslide susceptibility based on information volume and neural network model[J]. Chinese Journal of Rock Mechanics and Engineering，2020，39(S1)：2859-2870. (in Chinese)

[23] ZHAO B B，GE Y F，CHEN H Z. Landslide susceptibility assessment for a transmission line in Gansu Province，China by using a hybrid approach of fractal theory，information value，and random forest models[J]. Environmental Earth Sciences，2021，80：441.

[24] HE W C，CHEN G P，ZHAO J S，et al. Landslide susceptibility evaluation of machine learning based on information volume and frequency ratio：A case study of Weixin County，China[J]. Sensors，2023，23：2549.

[25] 王雪林. 四川省宜宾市叙州区滑坡灾害易发性评价[D]. 绵阳：西南科技大学，2023.
WANG X L. Susceptibility evaluation of landslide disaster in Xuzhou district of Yibin City，Sichuan Province[D]. Mianyang：Southwest University of Science and Technology，2023. (in Chinese)

[26] 孟凡奇，高峰，林波，等. 基于AHP和信息量模型的地质灾害易发性评价：以鲁东片区为例[J]. 灾害学，2023，38(3)：111-117.
MENG F Q，GAO F，LIN B，et al. Evaluation of geological disaster susceptibility in eastern Shandong based on AHP and information model[J]. Journal of Catastrophology，2023，38(3)：111-117. (in Chinese)

[27] ABDO H G，RICHI S M. Application of machine learning in the assessment of landslide susceptibility：A case study of mountainous eastern Mediterranean region，Syria[J]. Journal of King Saud University - Science，2024，36：103174.

[28] WANG Z，HONG T Z，PIETTE M A. Building thermal load prediction through shallow machine learning and deep learning[J]. Applied Energy，2020，263：114683.

[29] PHAM K，KIM D，LE C V，et al. Dual tree-boosting framework for estimating warning levels of rainfall-induced land-

slides[J]. Landslides, 2022, 19: 2249-2262.

[30] JARI A, KHADDARI A, HAJAJ S, et al. Landslide susceptibility mapping using multi-criteria decision-making(MCDM), statistical, and machine learning models in the Aube department, France[J]. Earth, 2023, 4: 698-713.

[31] 杨勇. 基于集成学习算法的衡阳市滑坡易发性研究[D]. 荆州: 长江大学, 2022.
YANG Y. Assessment of landslide susceptibility based on ensemble learning algorithm in Hengyang City[D]. Jingzhou: Yangtze University, 2022. (in Chinese)

[32] 王晨阳. 基于机器学习的滑坡地质灾害预测模型研究[D]. 西安: 西安科技大学, 2022.
WANG C Y. Research on landslide geological disaster prediction model based on machine learning[D]. Xi'an: Xi'an University of Science and Technology, 2022. (in Chinese)

[33] 杨灿. 基于机器学习的滑坡灾害易发性评价: 以安化县为例[D]. 长沙: 中南大学, 2022.
YANG C. Machine learning-based landslide susceptibility assessment: A case study in Anhua County[D]. Changsha: Central South University, 2022. (in Chinese)

[34] 贺倩, 汪明, 刘凯. 基于Logistic回归和MCMC方法评价地震滑坡敏感性[J]. 水土保持研究, 2022, 29(3): 396-403+410.
HE Q, WANG M, LIU K. Assessement on earthquake-triggered landslide susceptibility based on Logistic regression and MCMC method[J]. Research of Soil and Water Conservation, 2022, 29(3): 396-403+410. (in Chinese)

[35] SWETS J A. Measuring the accuracy of diagnostic systems[J]. Science, 1988, 240(4857): 1285-1293.

[36] 胡旭东. 基于集成学习的地质灾害易发性评价研究: 以云南省泸水县为例[D]. 武汉: 中国地质大学（武汉）, 2019.
HU X D. Study on the geo-hazards susceptibility assessment based on a novel ensemble learning framework: Application to the Lushui County, Yunnan Province[D]. Wuhan: China University of Geosciences(Wuhan), 2019. (in Chinese)

[37] LUNDBERG S M, ERION G, CHEN H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nature Machine Intelligence, 2020, 2(1): 56-67.

[38] YANG S L, TAN J Y, LUO D Y, et al. Sample size effects on landslide susceptibility models: A comparative study of heuristic, statistical, machine learning, deep learning and ensemble learning models with SHAP analysis[J]. Computers & Geosciences, 2024, 193: 105723.

[39] 周新植. 滑坡易发性机器学习优化模型及可解释性研究[D]. 重庆: 重庆大学, 2022.
ZHOU X Z. Study on machine learning optimization model and interpretability of landslide susceptibility[D]. Chongqing: Chongqing University, 2022. (in Chinese)

责任编辑: 张自强