引文格式:吴宏阳,周超,梁鑫,等.基于样本优化策略的滑坡易发性评价[J].武汉大学学报(信息科学版),2024,49(8):1492-1502.DOI:10.13203/j.whugis20220527

Citation: WU Hongyang, ZHOU Chao, LIANG Xin, et al. Evaluation of Landslide Susceptibility Based on Sample Optimization Strategy[J]. Geomatics and Information Science of Wuhan University, 2024, 49(8):1492-1502.DOI:10.13203/j.whugis20220527

基于样本优化策略的滑坡易发性评价

吴宏阳[1] 周 超[1,2] 梁 鑫[3] 王 悦[3] 袁鹏程[1] 吴立星[3]

1 中国地质大学(武汉)地理与信息工程学院,湖北 武汉,430074

2 三峡库区地质灾害野外监测与预警示范中心,重庆,404199

3 中国地质大学(武汉)工程学院,湖北 武汉,430074

摘 要:准确的易发性评价结果能够对滑坡带来的危险进行精准防控。样本优化是滑坡易发性评价的重要方法,可有效解决不平衡样本产生的决策边界偏移问题,提升滑坡易发性评价精度。以中国重庆市万州区东南区域为例,选取地层、土地利用、高程等10个影响因子构建滑坡易发性评价指标体系,应用频率比方法定量分析滑坡与指标之间的关系,在此基础上分别利用深度神经网络模型(deep neural networks, DNN)、过采样-深度神经网络模型(synthetic minority oversampling technique-DNN, SMOTE-DNN)、混合采样-深度神经网络耦合模型(one-class support vector machine-SMOTE-DNN, OS-DNN)、混合采样-深度神经网络-*K*均值聚类耦合模型(OS-DNN-*K*-means)进行滑坡易发性评价。结果表明,距道路距离、土地利用、地层是研究区滑坡发育的主要控制因子。精度评价结果发现OS-DNN-*K*-means(95.61%)和OS-DNN(91.16%)相较于模型SMOTE-DNN(87.97%)和DNN(81.40%)更能有效提高滑坡预测精度。通过混合采样和半监督分类进行样本优化能够有效解决研究区样本不平衡问题,为滑坡灾害空间预测提供新技术支撑。

关键词:滑坡;易发性建模;深度神经网络;混合采样;*K*均值聚类;样本优化策略

中图分类号:P642.4;P237 **文献标识码**:A **收稿日期**:2022-12-24

DOI:10.13203/j.whugis20220527 **文章编号**:1671-8860(2024)08-1492-11

Evaluation of Landslide Susceptibility Based on Sample Optimization Strategy

WU Hongyang[1] *ZHOU Chao*[1,2] *LIANG Xin*[3] *WANG Yue*[3] *YUAN Pengcheng*[1] *WU Lixing*[3]

1 School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China

2 Research Center of Geohazard Monitoring and Warning in the Three Gorges Reservoir, Chongqing 404199, China

3 Faculty of Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China

Abstract: **Objectives**: Accurate susceptibility evaluation results can accurately prevent and control the dangers caused by landslides. Sample optimization is an important method for landslide susceptibility evaluation, which can effectively solve the problem of decision boundary offset generated by unbalanced samples and improve the accuracy of landslide susceptibility evaluation. **Methods**: Taking the southeast area of Wanzhou District of Chongqing, China as an example, ten influencing factors such as strata, land use and elevation were selected to construct a landslide susceptibility evaluation index system, and the relationship between landslide and the indices was quantitatively analyzed by frequency ratio method, and on this basis, deep neural network model (DNN), synthetic minority oversampling technique-DNN model (SMOTE-DNN), one-class support vector machine-DNN coupling model (OS-DNN), and OS-DNN-*K*-means clustering coupling model (OS-DNN-*K*-means) were used to evaluate landslide susceptibility. **Results**: The results show that the distance from the road, land use and strata are the main control factors for landslide development in the study area. The accuracy evaluation results show that OS-DNN-*K*-means

基金项目:国家自然科学基金(42371094,41907253)。

第一作者:吴宏阳,硕士,主要从事地质灾害风险评价与系统开发。wuhongyangpower@163.com

通讯作者:周超,博士,副教授。zhouchao@cug.edu.cn

(95.61%) and OS-DNN (91.16%) could improve the landslide prediction accuracy more effectively compared with SMOTE-DNN (87.97%) and DNN (81.40%). **Conclusions**: Sample optimization through mixed sampling and semi-supervised classification can effectively solve the problem of sample imbalance in the study area, and provide new technical support for spatial prediction of landslide disasters.

Key words: landslides; landslide susceptibility mapping; deep neural networks; mixed sampling; *K*-means clustering; sample optimization strategy

中国地形地貌、地质条件复杂多样,地质灾害频发,其中滑坡发生次数最多,造成人员伤亡和经济损失最为严重[1-3]。为减少滑坡带来的危害,中国从20世纪90年代起,推进防灾减灾工程,自此每年因为滑坡灾害造成的人员伤亡和经济损失大幅度减少,这也导致中国滑坡灾害发生形式发生了很大的改变,呈现出点多、面广、小型微型化的特点[4],滑坡的高隐蔽性、高分散性导致样本出现不平衡情况。开展精细化滑坡易发性评价,识别潜在滑坡隐患,对于区域国土空间规划有重要作用。

随着计算机算力的增加,以GIS为基础,结合机器学习的方法在滑坡易发性评价领域中被广泛应用,如逻辑回归模型、人工神经网络模型、随机森林模型等[5-8],相较于传统的知识驱动模型,其结果具有更强的量化能力,然而这些模型被视为只有零个或者一个隐藏层的浅层学习方法,存在大量缺陷,如限制训练时间、难收敛、局部最优等[9]。相比之下,深度学习以其卓越的性能逐渐被运用到滑坡易发性评价建模中。深度学习模型能通过设置多层隐藏层、多种激活函数和优化函数来发现指标之间深层次的联系,为处理非线性数据提供了稳定的性能[10]。尽管现有研究表明深度学习在不同采样策略下相较于机器学习具有更好的预测性能[11],但准确预测的前提是大量的样本支持[12]。而在实际研究中,滑坡与非滑坡的样本数量往往是不均衡的,这种数据被称为不平衡数据[13]。在滑坡易发性评价中,数据量较少的滑坡具有更高的统计价值,不平衡的样本导致模型很难从数据量少的滑坡中获取特征,而是过多地关注非滑坡信息,使得模型存在偏差和过拟合问题[14]。目前针对样本不平衡问题主要有两种解决方法:一是对建模样本比例进行重构,主要分为减少多数类数量和增加少数类数量,即通过欠采样和过采样使样本中不同类别信息达到平衡[15];二是随机抽取样本,通过随机采样实现样本信息平衡[16]。

欠采样和过采样是常用的两种样本比例重构方法。其中欠采样通过减少不平衡数据中多数类样本数量来平衡数据数量,但这种方法容易删除代表性数据,进而丢失重要信息,导致模型学习不到完整分类特征[17];过采样通过增加不平衡数据中少数类样本来提高预测精度,但容易产生数据冗余,出现模型过拟合问题[18];随机采样能够有效地解决样本信息不足和过拟合问题,但采样具有随机性,结果不稳定[19]。针对上述问题,本文提出一种混合采样的方法来纯化和增添样本,并利用半监督方式从非滑坡样本中增添高置信度滑坡样本来优化滑坡易发性建模时样本不平衡问题。

以中国重庆市万州区东南区域为研究区,选取地层、土地利用、高程等十个指标因子构建滑坡易发性评价指标体系,应用频率比方法判别各指标对滑坡发育是影响关系,在此基础上分别利用深度神经网络(deep neural networks, DNN)、过采样-深度神经网络模型(synthetic minority oversampling technique-DNN, SMOTE-DNN)、混合采样-深度神经网络耦合模型(one-class support vector machine-SMOTE-DNN, OS-DNN)和混合采样-深度神经网络-*K*均值聚类耦合模型(OS-DNN-*K*-means)进行滑坡易发性评价,并利用受试者工作特征曲线(receiver operating characteristic curve, ROC)和精度评价模型验证混合采样算法和半监督分类算法。

1　方法原理

传统滑坡易发性评价建模将研究区已发生滑坡作为滑坡样本,滑坡外样本作为非滑坡样本进行易发性训练样本建模,但将低质量的非滑坡样本等同于高质量的滑坡样本会给模型建立引入噪声,降低分类的准确性,并且导致模型建模时出现数据不平衡情况。为进一步纯化现有非滑坡样本和补充新滑坡样本,本文提出了一种混合采样的新方法。如图1所示,首先利用单分类支持向量机(one-class support vector machine, OCSVM)构建超平面对非滑坡数据进行初步分类,在保留非滑坡样本重要信息前提下,从大量

非滑坡数据中筛选出纯化程度更高的非滑坡样本,实现非滑坡数据的欠采样过程;其次为平衡滑坡与非滑坡数据比例,对现有滑坡数据进行SMOTE处理,通过对现有数据进行分析生成新滑坡样本,扩充滑坡样本数据集,实现滑坡数据的过采样过程。通过以上两种方法实现样本数据的优化,即混合采样算法。为了说明混合采样的优势,将其与随机采样、SMOTE采样的精度进行对比,并在混合采样基础上耦合K-means构建半监督模型进一步提升预测精度。

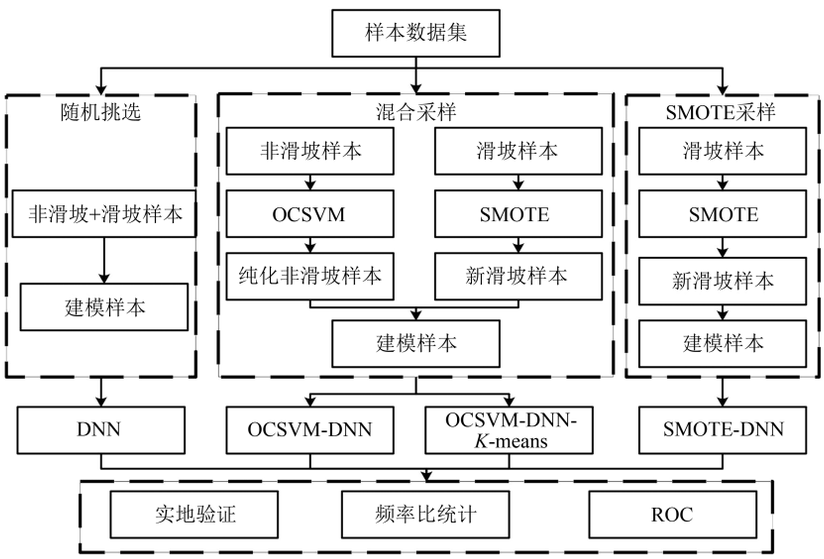

图1 技术路线流程图

Fig. 1 Flowchart of the Proposed Method

1.1 混合采样算法

1.1.1 OCSVM欠拟合采样

支持向量机(support vector machine, SVM)是一种按监督学习方式对数据进行二分类的广义线性分类器,与传统SVM不同,OCSVM是一种非监督分类算法,常常被用作异常点监测。它通过构建原点和训练数据之间的超平面,判断测试数据与训练数据之间的相似性,如果测试数据和训练数据相似,则将其归类为相似样本,记为1,否则记为-1(图2)[20]。

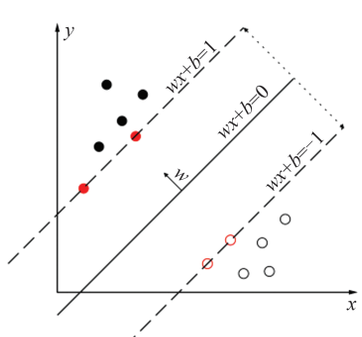

图2 OCSVM模型

Fig. 2 OCSVM Model

以斜率w、自变量x和常数项b构造分离超平面$wx+b=0$,对于任意线性可分的数据集来说,这样的超平面有无数个,但是几何间隔最大的分离超平面是唯一的。对于给定的数据集合T和超平面$wx+b=0$,数据集合中样本点(x_i,y_i)与超平面的几何间隔γ_i为:

$$\gamma_i=y_i\left(\frac{w}{\|w\|}\times x_i+\frac{b}{\|w\|}\right) \tag{1}$$

超平面在所有几何间隔中最小值为γ,即支持向量到超平面距离,因此可以将求解超平面距离问题变为求解约束最优化问题:

$$y_i\left(\frac{w}{\|w\|}\times x_i+\frac{b}{\|w\|}\right)^3\gamma \tag{2}$$

通过拉格朗日乘子法得到最终计算函数$f(x)$为:

$$f(x)=\mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i^*K(x_i,x_j)-b\right) \tag{3}$$

式中,α_i^*为拉格朗日乘子法下不等式约束;$K(x_i,x_j)$为核函数。如果将点(x_i,x_j)代入$f(x)$得到的结果大于0,则预测为1,否则预测为-1,以此实现对非滑坡数据的分离和纯化。

1.1.2 SMOTE

SMOTE技术是基于随机过采样算法的一种改进方案。随机过采样采取简单复制样本的策略来增加少数类样本数量,此方法容易产生模型过拟合的问题,导致模型泛化性不强,而SMOTE

算法的基本思想是对少数类样本进行分析并在少数类样本基础上人工合成新样本添加到数据集中[21]。

算法流程为:(1)选择滑坡中的样本x,以欧氏距离为标准计算它到少数类样本集中其他所有样本的距离,得到其k近邻;(2)从其k近邻中随机选择若干个样本,假设选择的近邻为x_n,对于每一个随机选出的近邻x_n,分别与原样本x按如下公式:

$$x_{new}=x+\text{rand}(0,1)\times|x_n-x| \tag{4}$$

构建新的滑坡样本[22],并将此样本添加到建模数据中用于训练。式(4)中,rand(0,1)表示随机产生一个大于等于0及小于1的均匀分布实数。

1.2　DNN

深度神经网络模型是由多个简单的神经元组成,每个神经元可以作为一个单独的感知机模型。一个完整的深度神经网络模型具有一个输入层、多个隐藏层和一个输出层(图3)。在神经网络结构中,层与层之间是全连接的,即第i层的任意一个神经元一定与第$i+1$层的任意一个神经元相连。DNN训练过程主要可以分为信号正向传递过程和误差反向传递过程。从输入层开始接收滑坡信息,识别神经元之间的连接权,依据权重对现有滑坡数据进行处理后,传递到下一个神经元进行分析,每个神经元的状态只会影响下一层神经元的状态。在输出层判断数据实际值和期望值的误差,误差过大则会转向反向传递,从输出层开始向隐藏层传递,调整隐藏层中神经元的权重,直到误差小于设置的阈值或训练次数达到设置上限[23]。DNN通过调整网络结构中权重使得目标函数尽可能拟合标签,一般使用梯度下降法作为权重调整算法。

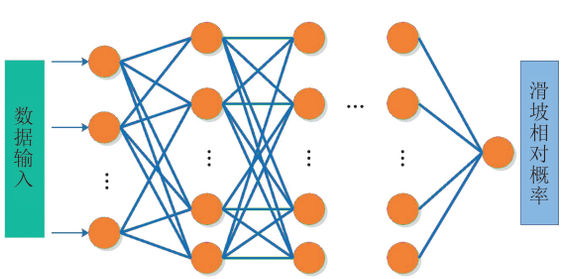

图3　DNN模型

Fig. 3　DNN Model

1.3　DNN-*K*-means

K-means算法是一种无监督聚类分析算法,通过选择离种子点最近均值的方法来对数据进行聚集。主要流程为:在数据中设置K个聚类中心,然后计算每个聚类对象到聚类中心的距离,将聚类对象分配给最近的聚类中心,聚类中心及分配的聚类对象作为一个聚类。每进行一次分配,聚类中心会根据现有的对象重新计算位置,不断迭代,直到不再进行分配或者误差平方和局部最小[24]。

DNN-*K*-means是一种半监督深度学习分类算法,其以深度神经网络模型的结果来协助分类和聚类,通过多次迭代提升模型预测精度。主要过程为:首先将已知滑坡设为已标记样本,剩余点设为未标记样本,随机抽取一定数量已标记样本和相同数量未标记样本预训练DNN,然后将全区剩余数据输入模型进行预测。通过*K*-means聚类算法对结果进行聚类,其中与已标记样本特征高度一致的样本定义为高置信度样本,得到对应伪标签,并更新已标记样本和未标记样本。重复上述训练过程,直到模型损失函数小于预设值或模型达到迭代次数(图4)。

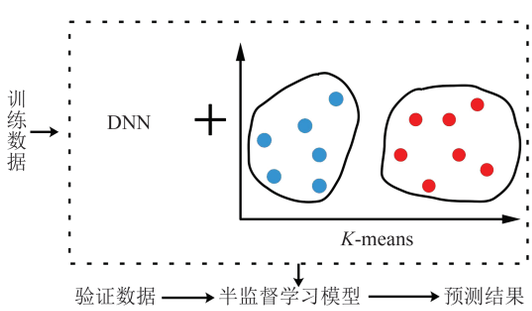

图4　DNN-*K*-Means模型

Fig. 4　DNN-*K*-Means Model

2　研究区概况与滑坡易发性评价建模

2.1　区域概况

研究区位于重庆市万州区东南部,其范围为108°31′E~108°43′E,30°34′N~31°42′N,面积约180.54 km²(图5)。研究区位于四川盆地东北部,属构造剥蚀中浅切割丘陵和峡谷地貌,区域四周丘陵起伏、中部地势相对平缓,整体呈漏勺状,绝对高程为251~1 245 m。研究区处于川东褶皱束方斗山背斜和齐岳山背斜之间,北跨龙驹坝背斜与赶场向斜,南靠马头场向斜。区域北部的背斜弧形转折部应力较为集中,加之受茨竹垭正断层和自生桥逆断层作用,内侧处于轴部的嘉陵江组地层被破坏[25]。区内地层岩性复杂,主要出露地层为侏罗系新田沟组、沙溪庙组、自流井组、珍珠冲组和三叠系巴东组、嘉陵江组。区内山间冲沟密布,与多条小溪构成区内地表径流网络。每年

5月—9月是研究区降雨集中时间段,在暴雨冲刷下,进一步加剧斜坡失稳,诱发滑坡灾害。自2018年,为落实研究区脱贫攻坚计划,区内人类工程活动愈加剧烈,对斜坡稳定性造成了强烈的破坏。主要人类工程活动包括房屋扩建和道路修建(如G318、X556及乡村公路的陆路交通网络体系)。

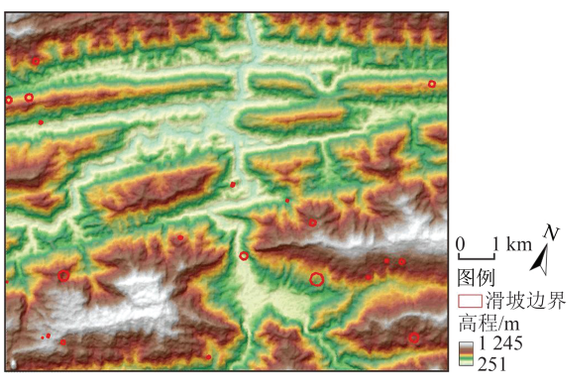

图5 研究区位置及滑坡分布

Fig. 5 Distribution of Landslides

2.2 滑坡数据概况

根据万州区实地调研结果建立滑坡灾害编录数据库。数据库显示区内共发育滑坡灾害21处,均由长期降雨或暴雨导致,其中小型滑坡13处,中型滑坡7处,大型滑坡1处。以30 m×30 m大小栅格划分研究区,共得到200 601个栅格,其中滑坡栅格629个,非滑坡栅格199 972个,两者比例近似0.31%。因此研究区滑坡与非滑坡数量存在不平衡情况,且研究区滑坡大多分布在丘陵起伏的山丘旁,地理空间分布特征较为单一。

2.3 评价指标体系建立

通过调研前人研究基础[26-28],结合研究区滑坡发育情况及数据可获取情况,并考虑到研究区无断层分布且距离长江过远,因此选择坡度、高程、斜坡结构、土地利用、地层、距道路距离、坡向、斜坡形态、地形湿度指数(topographic wetness index, TWI)、植被归一化指数(normalized difference vegetation index, NDVI)10个指标因子建立研究区评价指标体系,其中坡度、高程、坡向、斜坡形态、TWI来自于地形数据,地层来自于1∶5万地质图,斜坡结构来自于倾向倾角,NDVI由Landsat影像计算得到,各数据来源情况如表1所示。利用ArcGIS中波段集统计工具检验各指标因子之间的相关性,结果显示各指标因子之间的相关系数均小于0.36(表2),各指标因子之间呈弱相关或不相关,因此这10个指标可以直接用于滑坡易发性建模分析。

表1 实验数据来源

Tab. 1 Data Information of This Study

类型	比例尺或精度	来源
Landsat影像	30 m	地理空间数据云
地形数据	30 m	地理空间数据云
地质图	1∶50 000	万州区自然资源管理局提供
土地利用数据	30 m	EULUC-China数据集
倾角	30 m	野外实测插值
倾向	30 m	野外实测插值

表2 指标相关性

Tab. 2 Correlation of Indicator Factors

指标	土地利用	距道路距离	斜坡结构	NDVI	地层	斜坡形态	TWI	高程	坡向	坡度
土地利用	1	−0.02	0.01	0.14	−0.05	−0.03	−0.08	0.07	0.01	0.18
距道路距离	−0.02	1	0	−0.04	0.13	0	0	0.1	0.07	−0.06
斜坡结构	0.01	0	1	−0.16	0.05	0	−0.01	−0.01	0.36	0
NDVI	0.14	−0.04	−0.16	1	−0.12	−0.04	−0.09	0.03	−0.15	0.15
地层	−0.05	0.13	0.05	−0.12	1	−0.01	0	0.36	0.02	0
斜坡形态	−0.03	0	0	−0.04	−0.01	1	0.28	−0.08	0	−0.01
TWI	−0.08	0	−0.01	−0.09	−0	0.28	1	−0.13	−0.01	−0.09
高程	0.07	0.1	−0.01	0.03	0.36	−0.08	−0.13	1	0	0.09
坡向	0.01	0.07	0.36	−0.15	0.02	0	−0.01	0	1	0.03
坡度	0.18	−0.06	0	0.15	0	−0.01	−0.09	0.09	0.03	1

为进一步探究指标对于滑坡发育影响作用,参考前人对万州区的研究[29-30],首先对各指标因子以较小的间隔进行重分类,利用频率比方法计算各指标频率比值,在曲线值突变点进行二次分级[31]。

1) 坡度。坡度是滑坡产生的重要因素,坡度太小,无法为滑坡滑动产生足够动力,坡度太大,不利于岩土体进行堆积。将研究区坡度分为4级(见表3和图6(a))。研究区中滑坡基本集中于9°~24°区域,其频率比值为1.23,此区间容易堆积崩坡积物,又能为滑坡滑动产生足够动力。

表 3　各因素状态频率比表

Tab. 3　The Weighted Information Values of Each Factor State

指标	分级	频率比	指标	分级	频率比
坡度	[0°, 9°)	0.79	NDVI	[0, 0.15)	0.00
	[9°, 24°)	1.23		[0.15, 0.225)	0.65
	[24°, 36°)	0.92		[0.225, 0.375)	1.04
	[36°, 75°]	0.44		[0.375, 1]	1.00
坡向	[0°, 90°)	−0.69	地层	沙溪庙、新田沟	0.71
	[90°, 198°)	0.41		自流井、珍珠冲	2.43
	[198°, 252°)	0.60		巴东、嘉陵江	0.04
	[252°, 360°]	−0.59		须家河、雷口坡	2.32
高程	[251, 500) m	0.36	斜坡结构	顺向飘倾坡	1.34
	[500, 800) m	1.56		顺斜坡	1.45
	[800, 1 000) m	0.80		横向坡	1.06
	[1 000, 1 250] m	0.00		逆斜坡、逆向坡	0.55
TWI	[1, 5)	1.01	距道路距离	[0, 300) m	0.15
	[5, 8)	0.80		[300, 900) m	0.14
	[8, 10)	1.14		[900, 1 150) m	1.56
	[10, 15)	0.99		[1 150, 20 000] m	1.26
	[15, 28]	0.34		内向凹形坡(V/V)、内向凸形坡(V/X)	1.34
土地利用	建筑用地	2.63	斜坡形态	内向直线坡(V/GE)	1.45
	林地	0.79		外向凹形坡(X/V)、外向凸形坡(X/X)、外向直线坡(X/GE)	1.06
	裸地、农业用地	0.33		直线凹形坡(GR/V)、直线凸形坡(GR/X)、直线形直坡(GR/GE)	0.55

2)高程。不同高程下,人类工程活动和植被覆盖情况存在差异,进而导致斜坡稳定性在不同高程下存在区别。研究区高程为 251～1 245 m,可将研究区高程分为 4 级(见表 3 和图 6(b))。研究区中滑坡基本集中于 500～800 m 区域,频率比值为 1.56,此区间存在大量居民点和道路网。

3)斜坡结构。将斜坡坡度坡向与岩层倾角倾向之间关系进行判断,当岩层倾斜方向和斜坡面倾向方向一致时,易产生滑坡。将研究区斜坡结构分为 4 级(见表 3 和图 6(c))。研究区顺斜坡最易发育滑坡,其频率比值为 1.45。

4)土地利用。土地利用类型从侧面表征了人类工程活动强度,不同地表覆盖类型的斜坡稳定性不同。将研究区土地利用分为 3 级(见表 3 和图 6(d))。研究区中滑坡集中发育在人类工程活动较为丰富的建筑用地区域,频率比值为 2.63。

5)坡向。不同坡向上居民点和植被分布存在较大差异,导致局部地区水热比和气候分布存在差异,进而影响土壤含水量和斜坡稳定性。将研究区坡向分为 4 级(见表 3 和图 6(e))。研究区中滑坡基本集中于 198°～252°区域,频率比值为 0.6°,此区间居民点大量分布,存在较多人类工程活动。

6)斜坡形态。斜坡形态直观地展示了地表在空间上的几何特征,通常用地形的曲率特征来表示,其中剖面曲率控制物体在地表流动的速度,平面曲率控制物体流动的方向,二者共同决定一个物体运动的条件。研究区斜坡形态分为 4 级(见表 3 和图 6(f))。研究区中滑坡基本集中分布于内向直线坡,频率比值为 1.45。

7)地层。地层是滑坡发育的物质基础,其所独有的物理特性、渗透性等对滑坡发育起到基础控制作用。将研究区地层分为四类(表 3 和图 6(g))。自流井和珍珠冲组为泥岩夹砂岩或泥灰岩薄层,层内空隙较大,渗水性较好,在降雨情况下容易断裂或失滑,滑坡发育最为明显,频率比值为 2.43。

8)距道路距离。公路、铁路等工程的修建会造成坡脚开挖、坡体失稳,进而导致滑坡的发生。将研究区距道路距离分为 4 级(表 3 和图 6(h))。研究区大量道路的修建给边坡稳定性带来了巨大破坏,导致滑坡主要发生在距道路距离为 900～1 150 m 区域内,频率比值为 1.56。

9)TWI。TWI 反映了地形的富水程度,可以根据汇流累积量在时空上的变化,反映径流在地形中的流动情况,因此 TWI 也是影响滑坡的重要因素。将研究区 TWI 分为 5 级(表 3 和图 6(i))。指研究区中滑坡基本集中在 TWI 为 8～10 区域,频率比值为 1.14。

10)NDVI。植被覆盖度直接影响到植被对于地下水的补给作用,植被过少,暴雨和持续降雨会导致地下水大量渗入地表,使得斜坡抗剪强度减小。将研究区NDVI分为4级(表3和图6(j))。研究区中滑坡基本集中于NDVI值为[0.225,0.375)的区域,频率比值为1.04。

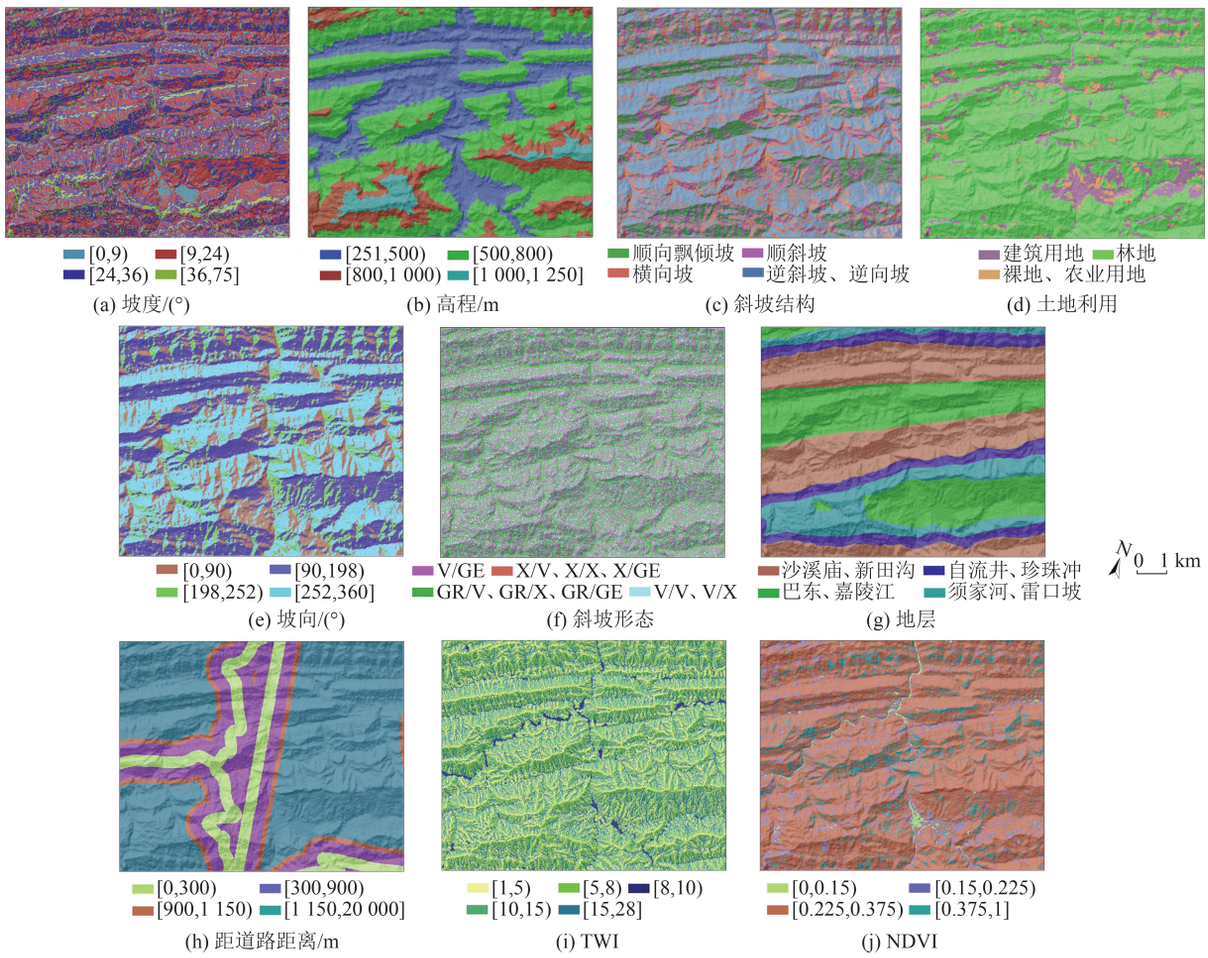

图6 研究区易发性评价指标图

Fig. 6 Landslide Susceptibility Factors of the Study Area

2.4 建模样本优化

目前在进行滑坡易发性评价时,大多是选择部分滑坡数据作为正样本,从滑坡外选择相同数量的点数据作为负样本进行训练,但这种方法筛选出的负样本不纯,很难保证非滑坡中不存在潜在滑坡,因此需要对非滑坡数据提纯。本研究首先利用OCSVM得到高度纯化后的非滑坡数据(图7),然后随机选择70%的滑坡数据,利用合成SMOTE进行样本的扩充,并选择相同数量的纯化非滑坡用于建模训练。

2.5 评价模型建立

DNN模型的预测过程中,主要影响参数包括隐藏层层数、隐藏层对应节点数量、激活函数等。目前常用的激活函数有Sigmoid、ReLU和Tanh,其中Sigmoid会出现梯度消失情况,且运行效率慢,ReLU遇到负数时则会完全失效,而Tanh收敛速度快,迭代次数少,并且能够解决部分Sigmoid不以0为中心进行输出的问题,因此本文设置Tanh为模型激活函数。考虑到部分指标因子差异较小且对滑坡影响因素不同,因此选择交叉熵作为损失函数来扩大数据差异性,加速模型收敛。此外,参考前人对深度学习模型的调参经验[32-33],结合研究区数据进行多次实验(图8),最终确定DNN网络结构由一层输入层、三层隐藏层、一层输出层组成,每个隐藏层中包含10个节点,DNN训练次数为500,学习率为0.01,此外为减少内存使用量,提高计算效率,采用自适应矩优化算法,它是一种结合自适应梯度算法和均方根传播优点的替代优化算法,能够完美替代随机梯度下降。为实现OS-DNN-*K*-means在聚类时得到较高的簇内相似度和较低的簇间相似度,在试算法下设置*K*-means聚类数为5。

2.6 结果对比与分析

对指标数据进行归一化后,利用训练数据建

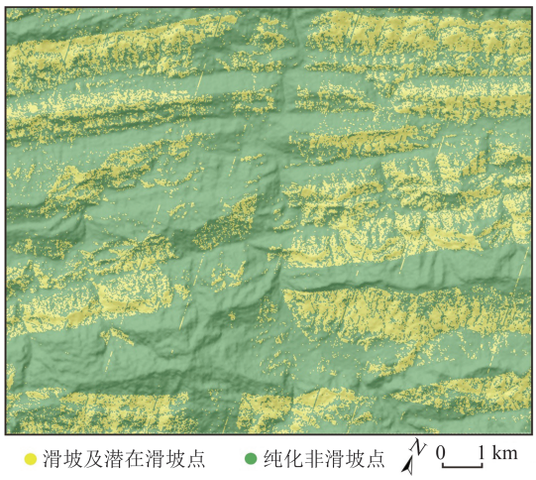

图 7　纯化滑坡分布

Fig. 7　Purification Landslide Distribution

模,并检验模型精度,得到全区易发性评价结果,将其大致分为低易发性(85%)、中易发性(5%)、高易发性(5%)、极高易发性(5%)4个区间。整体来看,OS-DNN-*K*-means(图 9(d))的预测结果相较于 OS-DNN(图 9(c))、SMOTE-DNN(图 9(b))和 DNN(图 9(a))的更符合实际滑坡分布,且碎块状点较少、连续性更强;局部上,OS-DNN-*K*-means 预测结果能够更好地贴近实际滑坡边界。

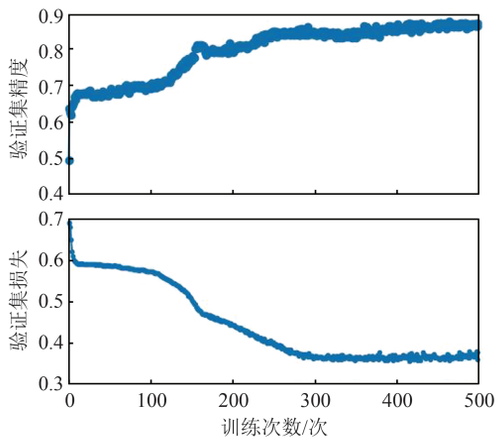

图 8　参数与预测精度关系曲线

Fig. 8　Parameters and Prediction Accuracy Relationship Curves

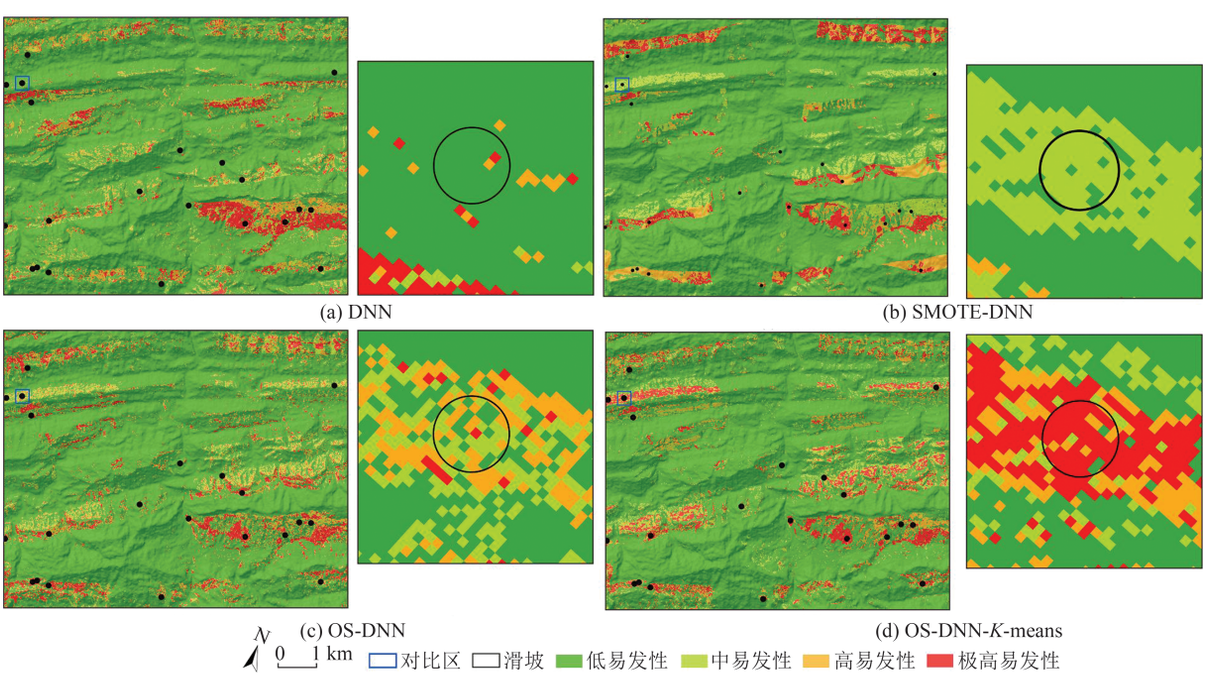

图 9　滑坡易发性分级图

Fig. 9　Classification Maps of Landslide Susceptibility

为进一步评价易发性结果,分别统计在各易发性等级中滑坡数量占比情况,结果如表 4 所示。DNN、SMOTE-DNN、OS-DNN、OS-DNN-*K*-means 在极高易发区中滑坡比率分别为 5.52、7.29、9.41 和 13.06。OS-DNN-*K*-means 能较好地预测滑坡分布情况。通过 ROC 和曲线下面积(area under curve, AUC)[34](图 10),可以看出 OS-DNN-*K*-means 模型预测精度为 95.61%,优于 OS-DNN(92.14%)、SMOTE-DNN(87.97%)和 DNN(81.40%)。

SMOTE-DNN 方法通过衍生高置信度滑坡样本用于建模训练,相较于随机挑选样本进行建模训练的 DNN 方法,精度可提高 6.57%,OS-DNN 利用混合采样的方法挑选高质量非滑坡样本和衍生高置信度滑坡样本用于滑坡易发性评价建模,其建模结果与 SMOTE-DNN 方法相比,精度可提高 4.17%;OS-DNN-*K*-means 在 OS-DNN 基础上利用半监督方法从未标记样本中挑选滑坡高置信度滑坡样本进行建模学习,可以进

表4 各易发性等级中滑坡数量占比

Tab. 4 Proportion of the Number of Landslides in Each Susceptibility Grade

模型	易发性等级	发生滑坡栅格数b	栅格总数c	占总滑坡比例d/%	占总栅格比例e/%	滑坡比率(d/e)
DNN	低	290	166 265	0.46	0.83	0.56
	中	61	11 666	0.10	0.06	1.67
	高	95	12 106	0.15	0.06	2.50
	极高	183	10 564	0.29	0.05	5.52
SMOTE-DNN	低	163	166 440	0.26	0.83	0.31
	中	100	11 458	0.16	0.06	2.78
	高	118	11 855	0.19	0.06	3.17
	极高	248	10 848	0.39	0.05	7.29
OS-DNN	低	85	165 954	0.14	0.83	0.16
	中	72	12 048	0.11	0.06	1.91
	高	138	11 279	0.22	0.06	3.90
	极高	334	11 320	0.53	0.06	9.41
OS-DNN-*K*-means	低	25	165 465	0.04	0.82	0.05
	中	37	12 340	0.06	0.06	0.96
	高	93	11 225	0.15	0.06	2.64
	极高	474	11 571	0.75	0.06	13.06

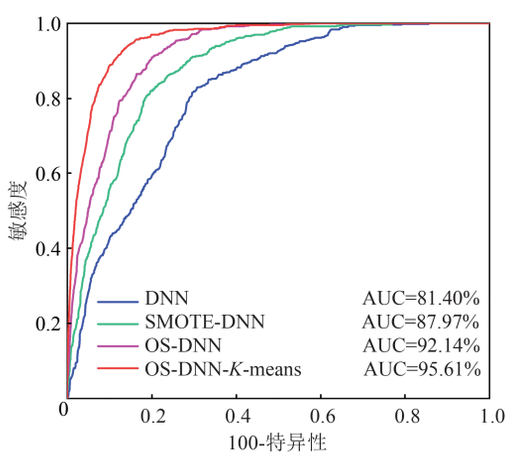

图10 模型ROC曲线图

Fig. 10 ROC Curves of the Model

一步增加模型对区域滑坡发育的认知,其建模结果与OS-DNN相比,精度可提高3.47%。

3 结 语

1)从空间异质性来看,不同的指标因子对滑坡发育的影响作用不同,通过构建易发性评价指标体系,并应用频率比模型来分析滑坡发育与指标之间的关系。在本研究区中,距道路距离(900~1 150 m)、土地利用(建筑用地)和地层(自流井和珍珠冲)是滑坡空间发育的主要控制因素。

2)进行滑坡易发性建模时,非滑坡质量低和滑坡数量少往往导致建模样本特征不充分,使得训练模型精度较低。混合采样方法可以实现非滑坡样本的纯化和新滑坡样本的生成,在此基础上进行滑坡易发性评价可以有效解决样本特征不足带来的影响,提高易发性评价预测精度;不仅能够减少低质量非滑坡数据干扰,还能通过增大滑坡样本数量实现样本平衡,更能避免随机采样的不确定性,达到提升预测精度的目的。

3)半监督学习可以有效减少在有限标记样本下模型的过度拟合,增强模型的泛化能力。此外,通过筛选和拓展高质量训练样本使得模型能够更充分地捕捉滑坡分布的复杂性,并有效提取未标记样本中包含的滑坡信息,实现用较少的标签样本达到较高的预测精度,能为数据不平衡地区进行滑坡易发性评价提供参考。

参 考 文 献

[1] Zhou C, Cao Y, Yin K, et al. Landslide Characterization Applying Sentinel-1 Images and InSAR Technique: The Muyubao Landslide in the Three Gorges Reservoir Area, China[J]. *Remote Sensing*, 2020, 12: 3385.

[2] Zhou Chao, Yin Kunlong, Cao Ying, et al. Characteristic Comparison of Seepage-Driven and Buoyancy-Driven Landslides in Three Gorges Reservoir Area, China[J]. *Engineering Geology*, 2022, 301: 106590.

[3] Liang Xin, Yin Kunlong, Chen Lixia, et al. Flow-solid Coupling Characteristics and Stability Analysis of Ganjingzi Landslide in the Wu Gorge Under Reservoir Water Level Fluctuation and Rainfall[J]. *The Chinese Journal of Geological Hazard and Control*,

2019, 30(1): 30-40. (梁鑫, 殷坤龙, 陈丽霞, 等. 库水位波动及降雨作用下巫峡干井子滑坡流-固耦合特征及稳定性分析[J]. 中国地质灾害与防治学报, 2019, 30(1): 30-40.)

[4] Qiu Haijun. Study on the Regional Landslide Characteristic Analysis and Hazard Assessment: A Case Study of Ningqiang County [D]. Xi'an: Northwest University, 2012. (邱海军. 区域滑坡崩塌地质灾害特征分析及其易发性和危险性评价研究: 以宁强县为例[D]. 西安: 西北大学, 2012.)

[5] Bragagnolo L, Silva R, Grzybowski J. Artificial Neural Network Ensembles Applied to the Mapping of Landslide Susceptibility[J]. *Catena*, 2020, 184: 104240.

[6] Bai S B, Wang J, Lü G N, et al. GIS-based Logistic Regression for Landslide Susceptibility Mapping of the Zhongxian Segment in the Three Gorges Area, China[J]. *Geomorphology*, 2010, 115(1): 23-31.

[7] Long J J, Liu Y, Li C D, et al. A Novel Model for Regional Susceptibility Mapping of Rainfall-Reservoir Induced Landslides in Jurassic Slide-prone Strata of Western Hubei Province, Three Gorges Reservoir Area [J]. *Stochastic Environmental Research and Risk Assessment*, 2021, 35(7): 1403-1426.

[8] Chen Fei, Cai Chao, Li Xiaoshuang, et al. Evaluation of Landslide Susceptibility Based on Information Quantity and Neural Network Model [J]. *Chinese Journal of Rock Mechanics and Engineering*, 2020, 39(S1): 2859-2870. (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.)

[9] Ghorbanzadeh O, Shahabi H, Crivellari A, et al. Landslide Detection Using Deep Learning and Object-based Image Analysis [J]. *Landslides*, 2022, 19(4): 929-939.

[10] Shahabi H, Rahimzad M, Tavakkoli Piralilou S, et al. Unsupervised Deep Learning for Landslide Detection from Multispectral Sentinel-2 Imagery [J]. *Remote sensing*, 2021, 13(22): 4698.

[11] Dou J, Yunus A P, Merghadi A, et al. Different Sampling Strategies for Predicting Landslide Susceptibilities Are Deemed less Consequential with Deep Learning [J]. *The Science of the Total Environment*, 2020, 720: 137320.

[12] Yao J, Qin S, Qiao S, et al. Assessment of Landslide Susceptibility Combining Deep Learning with Semi-Supervised Learning in Jiaohe County, Jilin Province, China[J]. *Applied Sciences*, 2020, 10(16): 5640.

[13] Chawla N V, Japkowicz N, Kotcz A. Editorial: Special Issue on Learning from Imbalanced Data Sets [J]. *ACM SIGKDD Explorations Newsletter: Special Issue on Learning from Imbalanced Datasets*, 2007, 6(1): 1-6.

[14] Wang Bowen, Wang Jingsheng, Wu Enzhong. SMOTENC-XGBoost Driver Traffic Safety Assessment Model for Unbalanced Dataset [J]. *Science Technology and Engineering*, 2023, 23(2): 831-837. (王博文, 王景升, 吴恩重. 面向不平衡数据集的SMOTENC-XGBoost驾驶人交通安全评估模型[J]. 科学技术与工程, 2023, 23(2): 831-837.)

[15] Liu X Y, Wu J X, Zhou Z H. Exploratory Undersampling for Class-Imbalance Learning [J]. *IEEE Transactions on Systems, Man, and Cybernetics Part B, Cybernetics: A Publication of the IEEE Systems, Man, and Cybernetics Society*, 2009, 39(2): 539-550.

[16] Tsai C F, Lin W C, Hu Y H, et al. Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection [J]. *Inf Sci*, 2019, 477: 47-54.

[17] Huang Faming, Chen Bin, Mao Daxiong, et al. Landslide Susceptibility Prediction Modeling and Interpretability Based on Self-screening Deep Learning Model [J]. *Earth Science*, 2023, 48(5): 1696-1710. (黄发明, 陈彬, 毛达雄, 等. 基于自筛选深度学习的滑坡易发性预测建模及其可解释性[J]. 地球科学, 2023, 48(5): 1696-1710.)

[18] Liu Xiwen, Duan Longzhen, Duan Wenying. Fuzzy C-means Clustering Based Undersampling in Clusters [J]. *Journal of Nanchang University (Natural Science)*, 2021, 45(5): 437-444. (刘稀文, 段隆振, 段文影. 基于FCM的簇内欠采样算法[J]. 南昌大学学报(理科版), 2021, 45(5): 437-444.)

[19] Seiffert C, Khoshgoftaar T M, Van Hulse J, et al. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance [J]. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, 2010, 40(1): 185-197.

[20] Sheng Ming, Chen Lingshan, Wang Junjie, et al. Abnormity Detection Method for In-vehicle CAN Bus Based on One-class SVM [J]. *Automobile Technology*, 2020(5): 21-25. (盛铭, 陈凌珊, 汪俊杰, 等. 基于单分类支持向量机的CAN总线异常检测方法[J]. 汽车技术, 2020(5): 21-25.)

[21] Wang Y M, Wu X L, Chen Z J, et al. Optimizing the Predictive Ability of Machine Learning Methods for Landslide Susceptibility Mapping Using SMOTE for Lishui City in Zhejiang Province, China[J]. *International Journal of Environmental Research and Public Health*, 2019, 16(3): 368.

[22] Li Xu, Chen Jiadui, Wu Yongming, et al. Classification Strategy of Imbalanced Data in Manufacturing

Process Based on Improved SMOTE[J/]. *Computer Engineering and Application*, 2022, 58(16): 284-291.(黎旭,陈家兑,吴永明,等.基于改进SMOTE的制造过程不平衡数据分类策略[J].计算机工程与应用,2022,58(16):284-291)

[23] Jiang Wandong,Xi Jiangbo,Li Zhenhong,et al. Landslide Detection and Segmentation Using Mask R-CNN with Simulated Hard Samples[J]. *Geomatics and Information Science of Wuhan University*, 2023, 48(12): 1931-1942.(姜万冬,席江波,李振洪,等.模拟困难样本的Mask R-CNN滑坡分割识别[J].武汉大学学报(信息科学版),2023,48(12):1931-1942.)

[24] Liao Yiqian, Yue Xianchang, Wu Xiongbin, et al. Amplitude Calibration of High-Frequency Radar Arrays Based on AIS and Canopy+*K*-means Algorithm[J]. *Modern Radar*,2023,45(9):9-15.(廖一迁,岳显昌,吴雄斌,等.基于AIS和Canopy+*K*-means算法的高频雷达阵列幅相校准[J].现代雷达,2023,45(9):9-15.)

[25] Xiao T, Segoni S, Liang X, et al. Generating Soil Thickness Maps by Means of Geomorphological-empirical Approach and Random Forest Algorithm in Wanzhou County, Three Gorges Reservoir[J]. *Geoscience Frontiers*, 2023,14(2),101514.

[26] Guo Zizheng, Yin Kunlong, Fu Sheng, et al. Evaluation of Landslide Susceptibility Based on GIS and WOE-BP Model [J]. *Earth Science*, 2019, 44(12): 4299-4312.(郭子正,殷坤龙,付圣,等.基于GIS与WOE-BP模型的滑坡易发性评价[J].地球科学,2019,44(12):4299-4312.)

[27] Wu Xueling, Yang Jingyu, Niu Ruiqing. A Landslide Susceptibility Assessment Method Using SMOTE and Convolutional Neural Network [J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(8): 1223-1232.(武雪玲,杨经宇,牛瑞卿.一种结合SMOTE和卷积神经网络的滑坡易发性评价方法[J].武汉大学学报(信息科学版),2020,45(8):1223-1232.)

[28] Wang Jiajia, Yin Kunlong, Xiao Lili. Landslide Susceptibility Assessment Based on GIS and Weighted Information Value: A Case Study of Wanzhou District, Three Gorges Reservoir[J]. *Chinese Journal of Rock Mechanics and Engineering*, 2014, 33(4): 797-808.(王佳佳,殷坤龙,肖莉丽.基于GIS和信息量的滑坡灾害易发性评价:以三峡库区万州区为例[J].岩石力学与工程学报,2014,33(4):797-808.)

[29] Liu Yuanbo, Niu Ruiqing, Yu Xianyu, et al. Application of the Rotation Forest Model in Landslide Susceptibility Assessment[J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(6): 959-964.(刘渊博,牛瑞卿,于宪煜,等.旋转森林模型在滑坡易发性评价中的应用研究[J].武汉大学学报(信息科学版),2018,43(6):959-964.)

[30] Guo Zizheng, Yin Kunlong, Huang Faming, et al. Evaluation of Landslide Susceptibility Based on Landslide Classification and Weighted Frequency Ratio Model[J]. *Chinese Journal of Rock Mechanics and Engineering*, 2019, 38(2): 287-300.(郭子正,殷坤龙,黄发明,等.基于滑坡分类和加权频率比模型的滑坡易发性评价[J].岩石力学与工程学报,2019,38(2):287-300.)

[31] Zhou Chao,Yin Kunlong,Xiang Zhangbo,et al. Quantitative Evaluation of the Landslide Susceptibility in Chun'an County Based on GIS[J]. *Safety and Environmental Engineering*,2015,22(1):45-50.(周超,殷坤龙,向章波,等.基于GIS的淳安县滑坡易发性定量评价[J].安全与环境工程,2015,22(1):45-50.)

[32] Wang Zan, Yan Ming, Liu Shuang, et al. Survey on Testing of Deep Neural Networks[J]. *Journal of Software*, 2020, 31(5): 1255-1275.(王赞,闫明,刘爽,等.深度神经网络测试研究综述[J].软件学报,2020,31(5):1255-1275.)

[33] Zhu Juntao, Yao Guangle, Zhang Gexiang, et al. Survey of few Shot Learning of Deep Neural Network[J]. *Computer Engineering and Applications*, 2021, 57(7): 22-33.(祝钧桃,姚光乐,张葛祥,等.深度神经网络的小样本学习综述[J].计算机工程与应用,2021,57(7):22-33.)

[34] Zhou C, Yin K L, Cao Y, et al. Landslide Susceptibility Modeling Applying Machine Learning Methods: A Case Study from Longju in the Three Gorges Reservoir Area, China [J]. *Computers & Geosciences*, 2018, 112: 23-37.