DOI：10. 13471/j. cnki. acta. snus. 2022D065

基于PBLC算法的滑坡空间易发性分析*

黄伟钧，李佳豪，刘子越，胡晓梅，黄华兵，李文楷

中山大学地理科学与规划学院，广东 广州 510006

摘　要：滑坡空间易发性统计模型的构建需要正样本（滑坡点）和负样本（非滑坡点）两类数据，但历史观测数据仅记录了正样本，而负样本的选取容易受到正样本污染，因为没有滑坡记录的地方也可能在过去或未来发生滑坡，从而导致模型的预测精度与稳定性受到影响。针对此问题，将前期提出的半监督学习算法PBLC（positive and background learning with constraints）应用于滑坡空间易发性分析，探讨其解决负样本污染问题的有效性。本文以粤东地区为研究区，选择高程、坡度、坡向、剖面曲率、距离道路最短距离、距离断层线最短距离、距水系最短距离、年平均降雨量、归一化植被指数和地理坐标共11个影响因子作为环境变量。结果表明，与传统的人工神经网络模型相比，基于PBLC算法的预测概率取值范围更为合理，预测结果更加稳定，且预测精度随背景样本数量增加而提高；粤东地区的滑坡灾害高易发区集中于北部和西南区域，坡度和高程是影响该地区滑坡易发性的主要因子。结果表明，半监督学习算法PBLC可以有效解决滑坡统计建模过程负样本污染的问题，提高模型预测精度。

关键词：滑坡易发性；带约束的正样本–背景学习；人工神经网络；未标记数据；粤东地区

中图分类号：P237　　**文献标志码：**A　　**文章编号：**2097－0137（2023）04－0054－11

Spatial susceptibility analysis of landslide based on PBLC algorithm

HUANG Weijun，*LI Jiahao*，*LIU Ziyue*，*HU Xiaomei*，*HUANG Huabing*，*LI Wenkai*

School of Geography and Planning，*Sun Yat-sen University*，*Guangzhou 510006*，*China*

Abstract：Statistical modeling of landslide susceptibility requires both positive（landslide）and negative（non-landslide）samples，but historical records of landslides only contain information on positive data. Selecting negative samples from areas without historical landslide records is problematic because landslides could have occurred without being observed or will occur in the future. This problem is referred to as case-control sampling with contaminated controls，which will affect the predictive accuracy and robustness of statistical models. To address this problem，we propose applying a semi-supervised learning algorithm PBLC（positive and background learning with constraints）and investigate its effectiveness in landslide susceptibility modeling. Taking Eastern Guangdong Province as the study area，we select 11 environmental variables，including elevation，slope，aspect，profile curvature，the shortest distance from roads，the shortest distance from fault lines，the shortest distance from rivers，mean annual precipitation，normalized difference vegetation index，and spatial coordinates，to investigate the effectiveness of the PBLC algorithm. Experimental results show that traditional artificial neural network underestimates the probabilities of landslide occurrences，and the degree of underestimation is affected by the number of negative samples. By contrast，the predicted probabilities of landslide occur-

* **收稿日期：**2022－09－02　　**录用日期：**2022－11－08　　**网络首发日期：**2023－03－31
基金项目：广东省基础与应用基础研究基金（2020A1515010764）
作者简介：黄伟钧（1998年生），男；**研究方向：**遥感技术与应用、统计模拟；E-mail：huangwj53@mail2.sysu.edu.cn
通信作者：李文楷（1982年生），男；**研究方向：**遥感技术与应用、统计模拟；E-mail：liwenk3@mail.sysu.edu.cn

rences by PBLC are more accurate and robust. The predicted landslide susceptibility map by PBLC shows that the areas with high susceptibility class are concentrated in the northern and southwestern regions in Eastern Guangdong Province, and slope and elevation are two of the most important factors that affect landslide susceptibility in the study area. We conclude that the semisupervised learning algorithm PBLC is effective in addressing the case-control sampling with contaminated controls in landslide susceptibility modeling.

Key words: landslide susceptibility; positive and background learning with constraints; artificial neural network; unlabeled data; Eastern Guangdong Province

我国拥有复杂多样的地质地貌环境，是世界上受地质灾害影响最大的国家之一（Brabb，1991）。开展滑坡空间易发性分析对防灾减灾政策制定与资源配置具有重要的指导意义，可有效保障生活在潜在滑坡区人民的生命财产安全。目前滑坡空间易发性分析方法主要分为两类。一类是利用岩土工程学方法，基于滑坡过程物理机理进行易发性评估，但此类方法需要测量复杂的参数，监测难度大、成本高，难以进行大范围分析与应用（Schiliro et al.，2016）。另一类是基于滑坡历史纪录和影响因子，构建统计模型进行易发性评价分析（Reichenbach et al.，2018）。与物理机理模型相比，统计模型的可行性更好，因此统计分析和机器学习模型被广泛应用于滑坡空间易发性分析。刘艺梁等（2010）运用改进的多层前馈人工神经网络（ANN，artificial neural network）模型和Logistic回归模型，基于坡度、坡向等9项评价指标，对三峡库区滑坡空间易发性进行分析。Yilmaz（2010）采用支持向量机（SVM，support vector machine）、Logistic回归和ANN等方法，利用高程、坡度、坡向等影响因子对滑坡易发性进行分区。Hu et al.（2019）利用SVM、ANN和随机森林（RF，random forest）3种机器学习模型，对九寨沟灾区开展滑坡风险分析。Goyes-Peñafiel et al.（2021）使用Logistic回归和证据权重的方法对哥伦比亚波帕扬地区的滑坡灾害易发性进行分析。刘坚等（2018）使用信息量法随机抽取不发生滑坡的点作为负样本数据，利用RF模型对三峡库区进行滑坡空间易发性分析。王毅等（2021）对江西省上饶市铅山县的滑坡易发性进行分析，结果表明卷积神经网络模型比传统的Logistic回归模型精度更高，且异质集成模型能大幅度提高预测精度。

构建滑坡空间易发性统计模型需要正样本（滑坡点）和负样本（非滑坡点）两类数据。正样本即为历史滑坡记录，负样本往往从无滑坡记录的区域随机选取，这种采样方式难以保证负样本不会受到正样本的污染，因为无滑坡记录的区域未来有可能会发生滑坡，或者已经发生过滑坡但没有被记录下来。如果负样本中掺杂了部分正样本，这些正样本在模型训练中被错误地当作负样本会导致估算的易发性概率值偏低。此外，负样本的数量往往是主观确定（例如与正样本数量相近），这会导致训练集中正样本的占比与该区域滑坡灾害先验概率不符，进而影响后验概率估算的准确性。

机器学习中的正样本-未标记学习（PUL，positive-unlabeled learning）算法属于半监督学习中的一个分支，可以解决一类分类问题中负样本缺失的问题（Castelli et al.，1996；Lee et al.，2003；Liu et al.，2003）。Elkan和Noto（2008）提出的PUL算法使用正样本和未标记数据训练模型对概率预测进行校正，但该算法要求正样本和未标记数据来自简单随机采样。Li et al.（2011）提出的正样本-背景学习（PBL，positive and background learning）算法则适合于正样本和未标记数据来自于分层随机采样。然而，PUL与PBL算法均需要对概率预测进行后续校正处理，容易出现概率高估现象。近期，Li et al.（2020）提出了新的带约束的正样本-背景学习（PBLC， positive and background learning with constraints）算法，可以使用正样本和背景数据对模型训练并直接得到校正的概率值，适用于分层随机采样情景，并在遥感影像一类分类中体现出较好的性能。

滑坡易发性分析属于一类分类问题，适合使用半监督学习来解决负样本污染问题。但是，目前在滑坡易发性分析中相对较缺乏半监督学习方法的应用与验证。历史滑坡位置点属于有标记的正样本，而无滑坡记录的位置点属于未标记数据而非单纯的负样本，且这两类数据来自分层随机采样。因此，本文采用针对分层采样的半监督学习算法PBLC，探讨其在滑坡易发性分析中的有效

性，并与经典的ANN方法进行对比。

1 PBLC算法原理

在一类分类问题中，用正样本($y=1$)和负样本($y=0$)训练一个分类器得到的概率预测结果为特定环境条件下属于正类别的条件概率，即 $\mathrm{P}(y=1\mid x)$，其中 x 表示环境影响因子。当使用有标记的正样本和无标记样本训练模型时，得到的概率预测结果为 $\mathrm{P}(s=1\mid x,\mu=1)$，其中 $s=1$ 表示有标记样本，$s=0$ 表示无标记样本，$\mu=1$ 表示分层随机采样。有标记的样本即正样本，故 $s=1$ 时 $y=1$，但无标记样本可能是正样本或负样本，故 $s=0$ 时 y 信息未知。采用正样本-未标记数据训练得到的概率模型 $\mathrm{P}(s=1\mid x,\mu=1)$ 与期望得到的概率模型 $\mathrm{P}(y=1\mid x)$ 存在如下关系(Li，2011)：

$$\mathrm{P}(s=1\mid x,\mu=1)=\frac{\mathrm{P}(y=1\mid x)}{\mathrm{P}(y=1\mid x)+(1-c)/c},\quad(1)$$

其中 c 代表训练集中正样本被标记出来的比例，即 $c=\mathrm{P}(s=1\mid y=1)$。假设训练集中有标记的正样本数量为 n_1，未标记样本数量 n_0，研究区内正类别的先验概率为 $\mathrm{P}(y=1)$，则

$$c=n_1/[n_1+n_0\times\mathrm{P}(y=1)].\quad(2)$$

若能明确 c 的值，可以先用正样本和未标记样本训练得到模型 $\mathrm{P}(s=1\mid x,\mu=1)$，再根据公式(1)推导得到目标概率模型 $\mathrm{P}(y=1\mid x)$. 然而，先验概率 $\mathrm{P}(y=1)$ 往往是未知的，因此 c 也未知。

令 $\mathrm{P}(y=1\mid x)=f(x,\omega)$，$\mathrm{P}(s=1\mid x,\mu=1)=g(x,\beta)$，其中 f 和 g 为某种形式的数学函数，ω 和 β 是对应模型的待估参数。由于训练集中包含了标记和未标记样本，可以通过最小化交叉熵损失函数估算模型参数 β：

$$L(\beta)=-\sum_{i=1}^{n}\left\{s_i\log\left[g(x_i,\beta)\right]+(1-s_i)\log\left[1-g(x_i,\beta)\right]\right\},\quad(3)$$

其中 n 为样本 (x_i,s_i) 的总数量，即 $i=1,2,\cdots,n$. 根据式(1)，以上损失函数可以改写为

$$L(\omega,c)=-\sum_{i=1}^{n}\left\{s_i\log\left[\frac{f(x_i,\omega)}{f(x_i,\omega)+(1-c)/c}\right]+(1-s_i)\log\left[1-\frac{f(x_i,\omega)}{f(x_i,\omega)+(1-c)/c}\right]\right\},\quad(4)$$

由于缺乏先验概率信息，通过最小化公式(4)直接求解模型参数 ω 变得更加困难(Hastie et al.，2013；Ward et al.，2009)。为了解决这个问题，可以将后验概率最大值 $P_{\max}$ 作为正则项添加到式(4)中，得到

$$L(\omega,c)=-\sum_{i=1}^{n}\left\{s_i\log\left[\frac{f(x_i,\omega)}{f(x_i,\omega)+(1-c)/c}\right]+(1-s_i)\log\left[1-\frac{f(x_i,\omega)}{f(x_i,\omega)+(1-c)/c}\right]\right\}+\lambda\left|\max\left[f(x,\omega)\right]-P_{\max}\right|^2,\quad(5)$$

其中 λ 是正则化参数，而后验概率最大值 $P_{\max}$ 可以设置为1. c 将作为参数在训练期间与 ω 一起优化。

因此，在缺乏可靠负样本情况下，可以基于正样本和未标记数据，使用公式(5)的损失函数来训练一个分类器，直接得到目标概率模型 $\mathrm{P}(y=1\mid x)$ 的参数 ω. 该算法称为带约束的正样本–背景学习(PBLC)算法(Li et al., 2020)。

2 研究区与实验设计

2.1 研究区

本文以粤东地区为研究区，该地区包括汕头、梅州、汕尾、潮州和揭阳5个市，总面积超过3万 km^2，属东南亚季风区，雨热同季，是全国光、热和水资源较丰富的地区。受海洋性东南季风影响，粤东地区雨量充沛，其中海丰、陆丰、揭西等雨量高区是省暴雨高区之一(曾广建，2022)。粤东地区地势呈现西北高，东南低的格局，西部、北部多山地，尤其以梅州市山地最多，中部较多丘陵，沿海多为冲积平原。粤东地区滑坡灾害频发，严重威胁人们的正常生活，其中梅州是广东省内滑坡最高发的地区，2005—2009年间共发生滑坡102场次(林泽雨等，2019)。

2.2 研究数据及预处理

研究区内滑坡灾害点数据来自于中国科学院资源环境科学与数据中心。经对原始滑坡灾害数据进行筛选、去除冗余信息，共统计出有记载的949个滑坡灾害点，其中556个位于梅州，204个位于潮州，126个位于汕尾，55个位于揭阳，8个位于汕头，其空间分布如图1所示。由图1可以看出，粤东地区滑坡点多集中在梅州。梅州山脉较多，一方面山体中天然形成的边坡数量较多，另一方面由于人类工程建设不断扩深，例如建设房屋道路等工程建设需要对山体进行削坡，削坡处理不当会加剧滑坡发生的可能性(郭子正等，2019；Costache，2019；Sifa et al.，2020)。

参考滑坡预测研究中的常用影响因子，本文选择了高程(DEM，digital elevation model)、坡度、坡向、

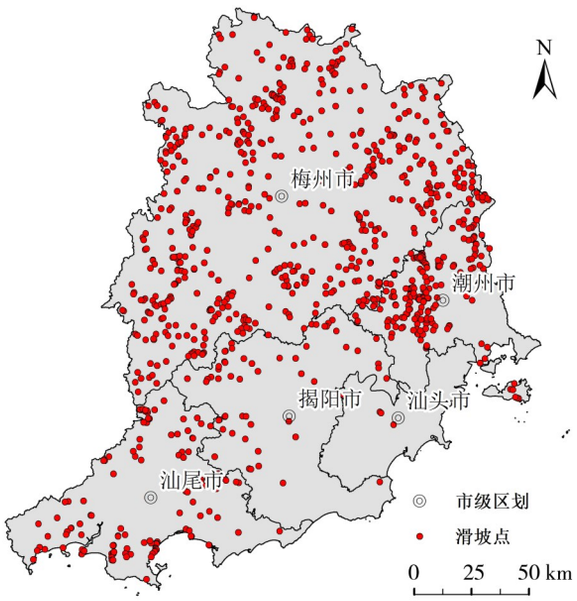

图1　粤东地区滑坡空间分布图

Fig. 1　Spatial distribution map of landslides in Eastern Guangdong Province

剖面曲率、距离道路最短距离、距离断层线最短距离、距水系最短距离、年平均降雨量、归一化植被指数(NDVI，normalized difference vegetation index)作为滑坡空间易发性预测的环境变量(李松林等，2020；Batar et al.，2021；Lucchese et al.，2021)。此外，考虑到滑坡灾害的发生具有一定的区域聚集性，我们把地理坐标(经纬度)也作为影响因子。其中DEM数据收集自ASTER GDEM。历年滑坡数据、断层线数据、坡度、坡向、降雨量数据均来自中国科学院资源环境科学与数据中心。降雨量数据为近5 a年平均降雨量。NDVI由欧空局Sentinel-2光学影像计算得出。水系、道路数据均来自于全国基础地理数据库，具体的数据格式与空间分辨率如表1所示。

由于收集到的环境因子有不同的数据格式、投影平面和坐标系统等，需要进行数据预处理操作。主要包括以下几个部分：数据镶嵌、统一坐标系、统一数据格式与分辨率、数据裁剪和归一化处理。本文使用的DEM、坡度、坡向、剖面曲率、NDVI、降雨量数据为栅格数据，而水系、道路、断层线为矢量数据。为方便后续研究，本文将所有环境因子数据统一为栅格数据。对于水系、道路、断层数据，本文使用欧氏距离工具计算到河流最小距离、到道路最短距离、到断层最短距离并输出保存为栅格数据图层。为了统一数据源空间尺度，本文统一使用30 m作为栅格化的格网大小，对需要做欧氏距离计算的图层数据在环境设置以DEM数据(30 m ×30 m)的栅格大小作为基准进行栅格捕捉设置。对于分辨率低的降雨量数据，本文采用重采样中的双线性内插法生成30 m ×30 m分辨率数据，将所有环境因子图层空间尺度统一。最后，将所有环境因子进行最小值-最大值归一化，将数据线性转换到［0，1］的区间，消除不同量纲的影响。经过统计分析，以上11个影响因子的Spearman相关系数均小于0.8，且方差膨胀因子VIF(variance inflation factor)值均小5，因此所有因子都被用于模型的训练和预测。

2.3　实验设计

ANN是滑坡易发性分析的经典方法，具有非线性拟合能力，可以估算后验概率(Lucchese et al.，2021；Richard et al.，1991；Yilmaz，2010)。因此，本文选取ANN为分类器构建滑坡空间易发性模型，并对耦合PBLC算法前后两种模型的预测结果进行对

表1　数据来源与数据类型

Table 1　Data source and data type

数据名称	数据来源	数据类型	分辨率
滑坡灾害	中国科学院资源环境科学与数据中心	矢量	
DEM	ASTER GDEM	栅格	30 m ×30 m
坡度	中国科学院资源环境科学与数据中心	栅格	30 m ×30 m
坡向	中国科学院资源环境科学与数据中心	栅格	30 m ×30 m
剖面曲率	由DEM数据计算得到	栅格	30 m ×30 m
降雨量	中国科学院资源环境科学与数据中心	栅格	1 km ×1 km
NDVI	欧空局Sentinel-2光学影像	栅格	30 m ×30 m
河流水系数据	全国基础地理数据库	矢量	
道路网络	全国基础地理数据库	矢量	
断层线数据	中国科学院资源环境科学与数据中心	矢量	

比，即 ANN 与 ANN-PBLC。本次实验将粤东地区收集到的 949 个历史滑坡灾害点作为有标记的正样本，并从研究区随机选取一定数量无标记的背景数据，组成样本数据集，其中 70% 作为训练集，30% 作为测试集。无标记背景样本数量的确定具有主观性，为了探讨其对模型的影响，本实验固定正样本的数量，并按照正样本与背景样本比例分别为 1∶1、1∶2、1∶3、1∶4、1∶5 选取对应数量的背景样本进行对照实验。为消除偶然性影响，本实验将每种情景下的随机采样过程进行 10 次重复实验。根据实验数据调整学习率，迭代次数、λ 正则化等参数。

为了保证不同实验条件下的模型评价结果具有可比性，所有模型的评价需要使用统一的独立测试数据，故测试数据中正样本和背景数据的比例固定为 1∶5。使用接受者操作特征曲线线下面积（AUC， area under the receiver operating characteristic curve）和 F_{pb} 作为模型的评价指标，同时还统计研究区预测结果的概率分布并绘制频率直方图。其中，AUC 是对模型的概率预测结果进行评价，其计算需要正样本和负样本（Davis et al.，2006）。由于缺乏真实可靠的负样本，用无标记背景数据替代负样本是文献中常用的方法（Jiménez-Valverde，2012；Peterson et al.，2008）。根据相关研究，由于背景数据是受到正样本污染的伪负样本，计算得到的 AUC 值会偏低，因此 AUC 的绝对值缺乏意义，但相对值仍然可以对模型的性能进行排序（Li et al.，2021；Lobo et al.，2008；Sofaer et al.，2019）。F 值是信息提取中常用的精度指标，可以对模型的二值预测结果进行评价，但也需要正样本和负样本（Sokolova et al.，2006；van Rijsbergen，1979）。F_{pb} 是缺乏负样本情况下对 F 值的替代指标，其绝对值也缺乏意义，但相对值可以对模型二值预测结果进行排序（Li et al.，2013）。

$$F_{pb} = \frac{2 \times TP}{TP + FN + FP}, \tag{6}$$

其中 TP 为正确预测为滑坡的数量，FN 为被错误预测为非滑坡的数量，FP 为未标记数据被预测为滑坡的数量。计算 F_{pb} 时，需要将概率值转换为二元值（即滑坡与非滑坡）。如果模型的概率预测值 P(y=1|x) 比较合理，可以选择 0.5 作为阈值进行二值预测；反之，如果模型的概率预测不合理，选择 0.5 为阈值将会产生较低的预测精度。因此，我们统一使用 0.5 作为阈值将概率预测二值化并计算 F_{pb} 指标，以此间接评价模型概率预测的合理性。

3 结果与讨论

3.1 模型评价

不同正样本-背景数据比例下 10 次重复实验的模型性能对比如表 2，其中 P_{min}，P_{ave} 和 P_{max} 分别为预测概率值的最小值、平均值和最大值。可以看出，在不同样本比例下，模型 ANN 与 ANN-PBLC 的 AUC 值比较接近，且随着样本比例从 1∶1~1∶5，模型的 AUC 值呈现上升趋势。例如，当样本比例为 1∶1 时，ANN 与 ANN-PBLC 的 AUC 值分别为 0.695 4 和 0.706 9；当样本比例为 1∶5 时，两个模型对应的 AUC 值分别为 0.729 9 和 0.732 3。与 AUC 指标不一样，两个模型的 F_{pb} 值存在明显的差异：

表 2 ANN 和 ANN-PBLC 两个模型在不同正样本-背景数据比例下的性能对比

Table 2 Comparison of performances of ANN and ANN-PBLC under different ratio values between the numbers of positive and background samples

样本比例	数值	ANN					ANN-PBLC				
		P_{min}	P_{ave}	P_{max}	AUC	F_{pb}	P_{min}	P_{ave}	P_{max}	AUC	F_{pb}
1∶1	平均值	0.121 5	0.473 4	0.755 0	0.695 4	0.457 4	0.018 0	0.501 0	0.971 2	0.706 9	0.470 0
	标准差	0.044 6	0.013 4	0.035 6	0.019 5	0.024 1	0.010 7	0.036 3	0.014 8	0.014 2	0.027 4
1∶2	平均值	0.059 0	0.321 5	0.635 4	0.715 4	0.255 9	0.012 6	0.464 6	0.970 8	0.717 0	0.483 5
	标准差	0.013 4	0.011 8	0.042 2	0.016 0	0.090 5	0.010 4	0.032 3	0.006 7	0.018 3	0.026 2
1∶3	平均值	0.024 9	0.240 6	0.576 8	0.721 5	0.045 3	0.005 6	0.431 1	0.971 0	0.721 5	0.489 5
	标准差	0.015 2	0.004 7	0.045 0	0.018 5	0.028 5	0.002 9	0.034 4	0.009 9	0.014 0	0.027 4
1∶4	平均值	0.019 5	0.193 5	0.541 0	0.724 6	0.013 4	0.005 7	0.429 7	0.978 0	0.728 4	0.494 4
	标准差	0.011 5	0.006 7	0.048 9	0.015 9	0.011 8	0.003 1	0.049 3	0.009 3	0.014 9	0.033 4
1∶5	平均值	0.011 9	0.166 4	0.497 3	0.729 9	0.004 1	0.004 3	0.441 7	0.982 6	0.732 3	0.500 2
	标准差	0.005 1	0.007 4	0.037 1	0.014 7	0.005 7	0.002 5	0.046 8	0.007 9	0.013 2	0.026 5

在同样的样本比例下，ANN-PBLC 的 F_{pb} 值高于 ANN；随着样本比例从 1∶1~1∶5，ANN-PBLC 的 F_{pb} 值增加，而 ANN 的 F_{pb} 值则迅速下降。当样本比例为 1∶1 时，ANN 与 ANN-PBLC 的 F_{pb} 值分别为 0.457 4 和 0.470 0；当样本比例为 1∶5 时，两个模型对应的 F_{pb} 值分别为 0.004 1 和 0.500 2。此外，ANN 模型预测概率的最小值变化不大，但平均值和最大值随着样本比例从 1∶1~1∶5 逐步变小。相反，ANN-PBLC 模型预测概率的最小值、平均值和最大值在不同样本比例下均比较稳定。

图 2 为两个模型在不同样本比例下预测概率的频率直方图。理论上，研究区内滑坡易发性概率值应该覆盖［0，1］区间，且不同样本量下预测的概率分布应该接近。由图 2 可以看到，ANN-PBLC 模型概率取值范围分布较为均匀，在概率值为 1 的附近也有分布，且随着背景样本的增加，ANN-PBLC 模型预测概率的分布直方图比较稳定，符合实际情况。相反，ANN 模型预测概率的直方图存在明显的左偏现象，在概率值为 1 的附近缺乏分布，且随着背景样本的增加概率左偏问题越来越严重。根据图 3，两个模型的概率预测值空间格局接近，但 ANN 的预测概率值较低。当样本比例为 1∶1 时，大部分的滑坡灾害点均落于两个模型概率值较大的地方；当样本比例为 1∶5 时，ANN-PBLC 的概率预测图变化不大，但 ANN 的预测概率值明显变低，大多数的滑坡灾害点落于 ANN 预测概率值较低的地方。

综上可知，ANN 模型只有在样本比例为 1∶1 时预测结果较为合理，但其概率预测值与 ANN-PBLC 相比仍然偏低。粤东地区是一个滑坡灾害频繁的区域，根据图 3 的预测结果来看，除了已知的历史滑坡灾害点之外，研究区内尚有很多高风险区域，因此从无滑坡记录的区域里面选择负样本时很容易将高风险地方当作负样本，这是造成 ANN 模型概率预测值偏低的一个主要原因。此外，主观设定样本比例是造成模型概率预测偏差的另一个原因。根据机器学习的原理（Hastie et al.，2013；Ward et al.，2009），只有当训练集中正样本的占比符合研究区的先验概率 $P(y=1)$ 时模型才可以正确地估算后验概率 $P(y=1|x)$，但 ANN 的分层采样方式造成正样本的占比往往与研究区的先验概率信息不符。相反，PBLC 是一种针对分层采样方式的半监督学习算法，未标记的背景数据并不是单纯地被当作负样本，而是作为正样本和负样本的混合，在模型训练中根据贝叶斯原理进行了概率校正，在理论上可以准确地估算 $P(y=1|x)$（Li et al.，2020）。相应地，本次实验中 ANN-PBLC 模型的概率预测结果均比 ANN 更加合理，且在不同样本比例下其预测结果比较一致。因此，ANN-PBLC 可以使用大量的无标记背景数据来提高模型预测精度且同时保持合理的概率分布。

本次实验中 ANN 和 ANN-PBLC 在样本量最大时的 AUC 值均在 0.73 左右。AUC 衡量的是模型的排序能力，与模型预测概率的相对值有关而与绝对值无关（Lobo et al.，2008）。根据公式（1），ANN 的预测结果 $P(s=1\mid x,\ \mu=1)$ 与 ANN-PBLC 的预测结果 $P(y=1\mid x)$ 存在单调递增关系，两个模型的排序能力相接近，因此 AUC 值也接近。需要注意的是，测试数据中的负样本也同样受到部分正样本的污染，因此计算出来的 AUC 值要比真实值偏低（Jiménez-Valverde，2012；Li et al.，2021）。如果采用 Li et al.（2021）提出的校正方法，则本次实验的 AUC 值在 0.84 左右。田春山等（2016）使用 Logistic 模型对广东省地质灾害易发性评价中的 AUC 值为 0.78，与本次实验结果较为接近。王毅等（2021）使用了深度学习方法对江西省上饶市铅山县的滑坡易发性进行分析，其 AUC 值为 0.83，而集成模型的 AUC 值可达 0.91。PBLC 是一种模型训练方法，可以与多种分类器耦合，而本次研究仅以 ANN 为例。在今后研究中，需要进一步探讨该算法与多种分类器（尤其是深度学习和集成模型）的耦合效果。

3.2　粤东地区滑坡空间易发性

根据以上分析，模型 ANN-PBLC 在正样本−背景数据比例为 1∶5 的情况下精度最高，因此选择该模型对粤东地区滑坡空间易发性制图分析，结果如图 4。按照预测概率值 P 将研究区划分为 4 个等级：极低易发区（$P\leqslant 0.25$）、较低易发区（$0.25\leqslant P<0.5$）、较高易发区（$0.5\leqslant P<0.75$）、极高易发区（$P\geqslant 0.75$）。此外，选择 0.5 的概率阈值将预测结果进行二值化，将研究区划分为滑坡发生和不发生两种类别。由图 4 可见，粤东地区存在较多的滑坡灾害高易发区，主要集中于北半部，西南区域也有部分高易发区存在，而低易发区则主要集中于东南区域地势平坦的地方。根据表 3 的滑坡易发性等级统计结果可见，随着滑坡风险等级的增加，滑坡的频率比也逐步增大，由极低易发区的 0.149 3 增加到极高易发区的 3.067 3，说明该区域

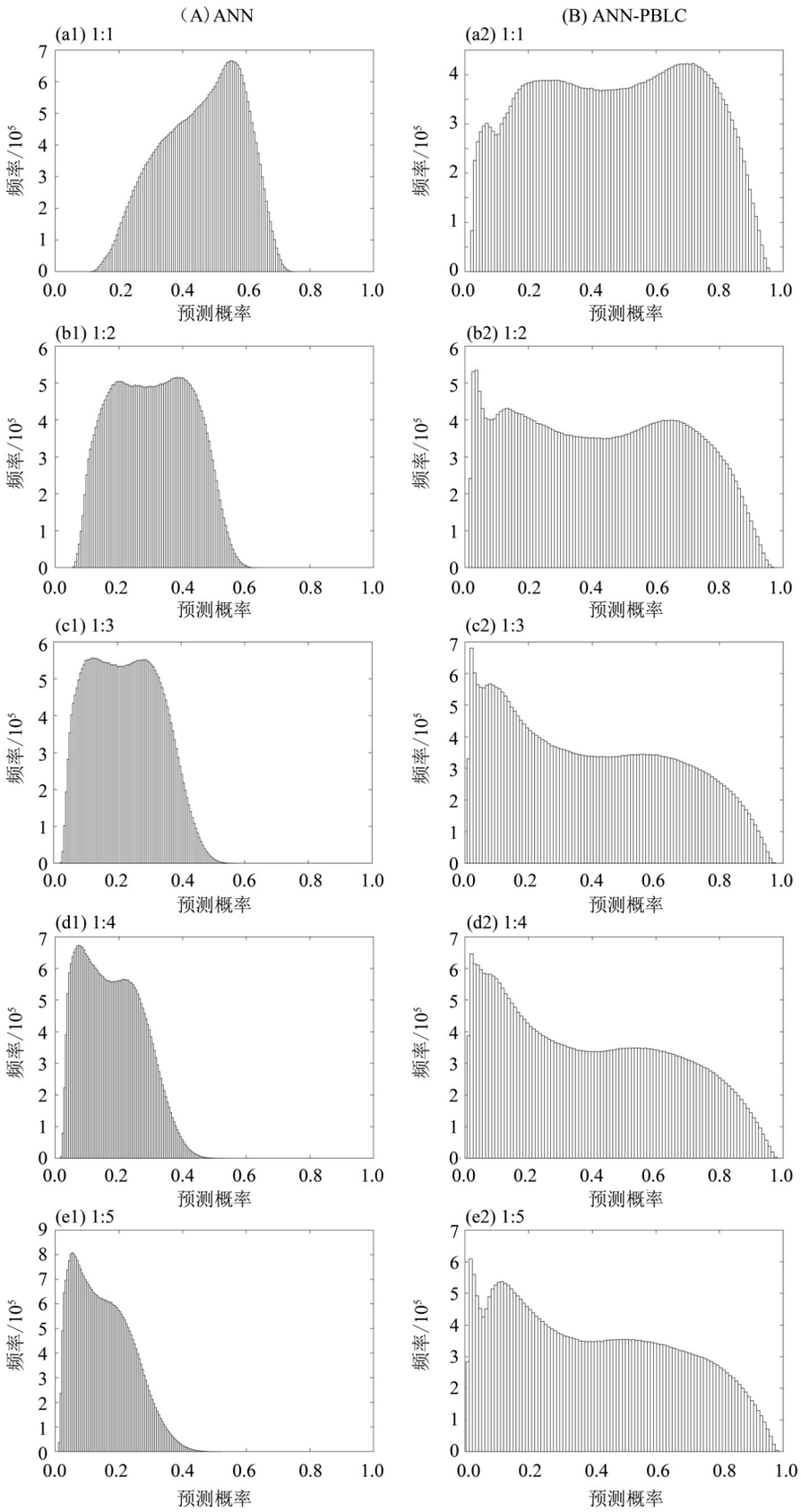

图 2　ANN 和 ANN-PBLC 在不同正样本−背景数据比例下预测概率分布直方图

Fig. 2　The histograms of predicted probabilities by ANN and ANN-PBLC with different ratio values between the numbers of positive and background samples: (a1, a2) 1:1, (b1,b2) 1:2, (c1,c2) 1:3, (d1,d2) 1:4, (e1,e2) 1:5

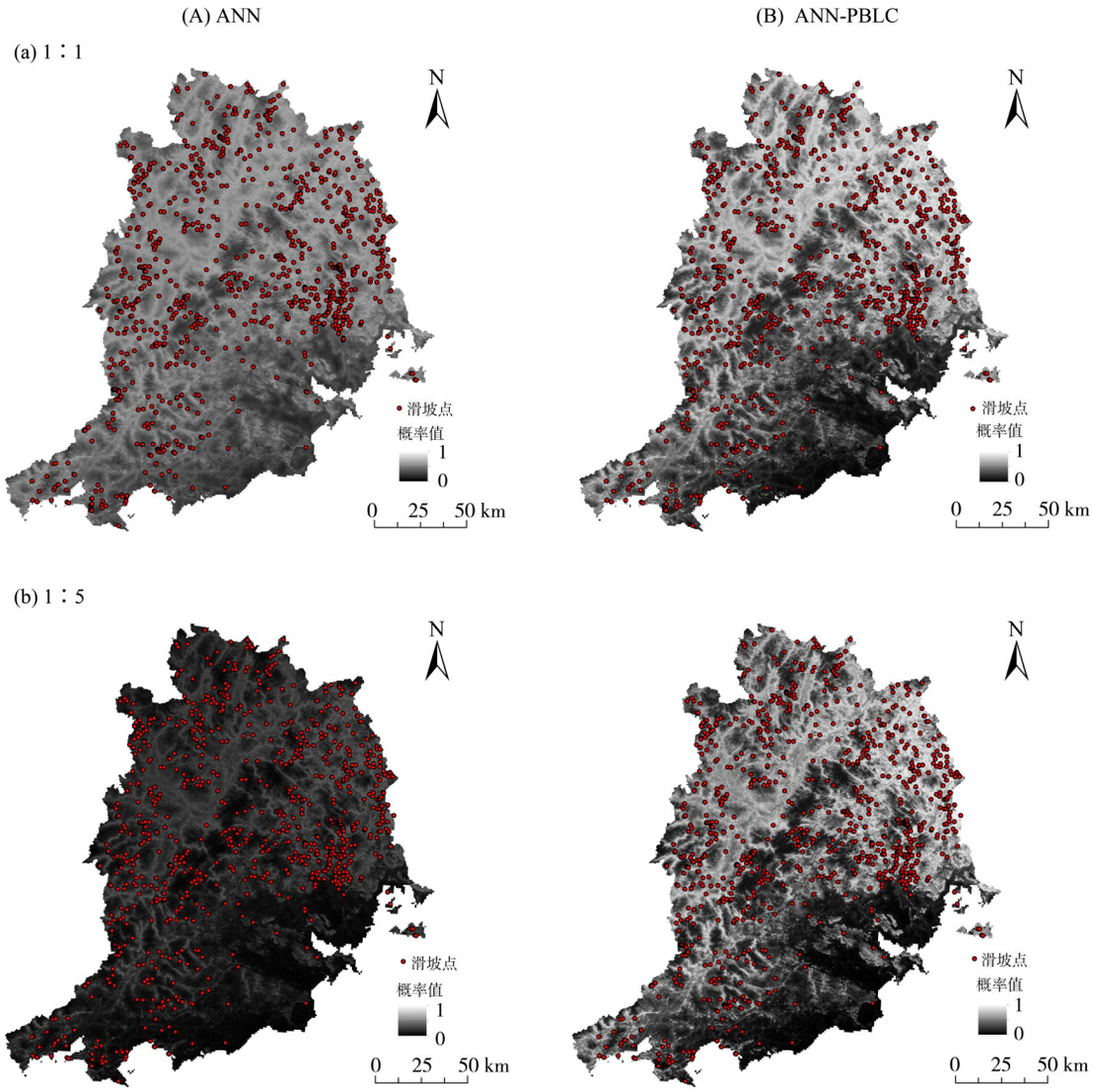

图3　ANN和ANN-PBLC在不同正样本-背景数据比例下概率预测图

Fig. 3　The predicted probability maps by ANN and ANN-PBLC with different ratio values between the numbers of positive and background samples

表3　粤东地区滑坡易发性等级统计

Table 3　The statistics of landslide susceptibility levels in Eastern Guangdong Province

易发性级别	滑坡数量/次	分级栅格数/个	滑坡比例	面积比例	频率比
极低	50	12 032 052	0. 052 7	0. 352 8	0. 149 3
较低	174	9 173 630	0. 183 4	0. 269 0	0. 681 7
较高	346	8 459 899	0. 364 6	0. 248 0	1. 469 9
极高	379	4 440 629	0. 399 4	0. 130 2	3. 067 3

滑坡易发性等级划分具有合理性(陈飞等，2020；田春山等，2016)。从图4的二值图来看，研究区内的历史滑坡灾害点大多数位于模型预测的滑坡发生区域，同时也有大量的预测发生区域中尚未有滑坡灾害记录，这些地方在今后的滑坡灾害防控工作中需要重点关注。表4统计了每个城市中不同滑坡易发性等级的比例，其中梅州和潮州两个城市中的高易发区比例较高(约50%)，而汕尾、揭阳和汕头3个城市的高风险区域比例较低。

本研究选择了11个环境因子来预测滑坡易发性。在模型预测时将某一个环境因子变成随机值，可以分析该因子对模型预测精度的影响，即置换

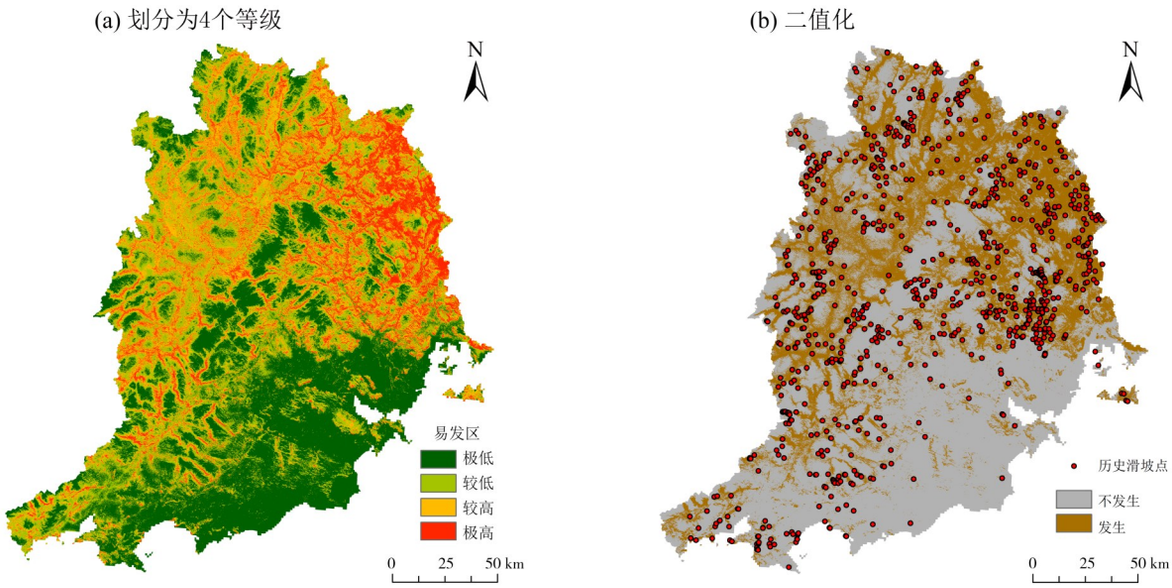

图4 基于ANN-PBLC的粤东地区滑坡易发性预测图

Fig. 4 The predicted landslide susceptibility maps by ANN-PBLC

表4 不同城市滑坡易发性等级的比例

Table 4 The percentages of landslide susceptibility levels in different cities %

易发性级别	梅州	潮州	汕尾	揭阳	汕头
极低	17. 40	26. 10	51. 00	64. 53	77. 38
较低	30. 36	21. 78	29. 64	22. 08	13. 22
较高	34. 41	27. 10	14. 67	10. 12	7. 51
极高	17. 84	25. 02	4. 69	3. 28	1. 88

重要性(Permutation importance)(Breiman，2001)。以AUC为评价指标，通过对以上ANN-PBLC模型进行置换重要性分析，各环境因子的重要性排序如图5所示。其中，坡度和高程的重要性最高，两者的贡献均在20%左右。地理坐标(经纬度)在本次实验中也体现了较大的贡献，尤其是纬度，两者的贡献合计达20%左右。根据图1，研究区内大多数的历史灾害记录集中于北半部和西南区域，具有明显的空间聚集性，这也是地理坐标有较大贡献的原因。由于缺乏数据，本次研究未把其他潜在的环境因子(例如岩性、地质年代、土壤等)纳入分析，这也是本研究的不足之处。

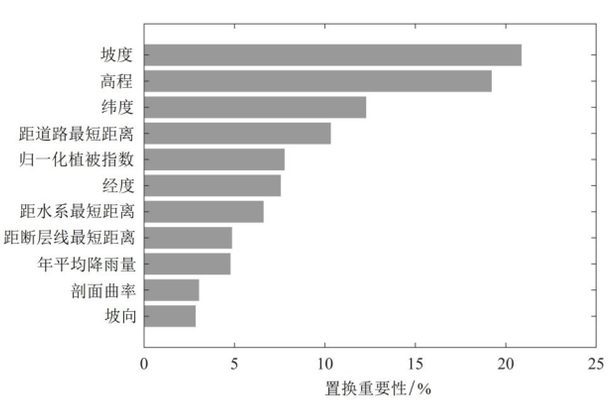

图5 环境因子重要性

Fig. 5 The importance of environmental factors

4 结 论

本文选取高程、坡度、坡向、剖面曲率、距离道路最短距离、距离断层线最短距离、距水系最短距离、年平均降雨量、NDVI和地理坐标共11个影响因子作为滑坡空间易发性预测的环境变量，通过半监督学习算法PBLC耦合ANN模型，对粤东地区的滑坡空间易发性进行了分析评估。实验结果表明：ANN-PBLC模型的概率预测结果比ANN模型更加合理和稳定，且ANN-PBLC模型的AUC和F_{pb}值随背景样本增加而提升，说明该模型可以通过增加未标记背景样本帮助模型训练，提高预测精度。根据ANN-PBLC模型的预测结果，粤东地区滑坡空间易发性预测结果表明该地区滑

坡灾害高易发区集中于北半部和西南区域，其中梅州和潮州两个城市的高易发区比例较高，而汕尾、揭阳和汕头3个城市的高风险区域比例较低；坡度和高程是影响该区域滑坡易发性的主要因素，且滑坡易发性具有明显的空间聚集特征。

在真实的地理环境中，滑坡的负样本数据容易受到正样本的污染，从而导致模型结果出现偏差。本研究表明半监督学习算法PBLC可以有效解决滑坡统计建模过程负样本污染的问题，提高模型预测精度，克服了传统ANN方法预测概率偏低且取值范围随背景样本增加而缩小的问题，对滑坡灾害的风险区划与防灾减灾策略制定具有指导意义。在今后研究中，将结合更多的案例对该方法进行评估与验证，并进一步探讨该算法与其他分类器(包括深度学习模型)的耦合效果。

参考文献：

陈飞，蔡超，李小双，等，2020. 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，39(S1)：2859-2870.

郭子正，殷坤龙，付圣，等，2019. 基于GIS与WOE-BP模型的滑坡易发性评价[J]. 地球科学，44(12)：4299-4312.

李松林，许强，汤明高，等，2020. 三峡库区滑坡空间发育规律及其关键影响因子[J]. 地球科学，45(1)：341-354.

林泽雨，刘爱华，2019. 广东地区滑坡灾害分布特征与预警措施分析[J].人民长江，50(S1)：90-92.

刘坚，李树林，陈涛，2018. 基于优化随机森林模型的滑坡易发性评价[J]. 武汉大学学报(信息科学版)，43(7)：1085-1091.

刘艺梁，殷坤龙，刘斌，2010. 逻辑回归和人工神经网络模型在滑坡灾害空间预测中的应用[J]. 水文地质工程地质，37(5)：92-96.

田春山，刘希林，汪佳，2016. 基于CF和Logistic回归模型的广东省地质灾害易发性评价[J]. 水文地质工程地质，43(6)：154-161+170.

王毅，方志策，牛瑞卿，等，2021. 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，23(12)：2244-2260.

曾广建，2022. 粤东地区降水量空间插值方法研究[J]. 广东水利水电，(4)：27-32.

BATAR A K，WATANABE T，2021. Landslide susceptibility mapping and assessment using geospatial platforms and weights of evidence（WoE）method in the Indian Himalayan region：Recent developments，gaps，and future directions[J]. ISPRS Int J Geo Inf，10(3)：114-141.

BRABB E E，1991. The world landslide problem[J]. Episodes，14(1)：52-61.

BREIMAN L，2001. Random forests[J]. Mach Lang，45(1)：5-32.

CASTELLI V，COVER T M，1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter[J].IEEE Trans Inf Theory，42(6): 2102-2117.

COSTACHE R，2019. Flash-flood potential index mapping using weights of evidence，decision trees models and their novel hybrid ensemble[J]. Stoch Environ Res Risk Assess，33(7)：1375-1402.

DAVIS J，GOADRICH M，2006. The relationship between precision-recall and ROC curves[C]//Proceedings of the 23rd International Conference on Machine learning. USA：233-240.

ELKAN C，NOTO K，2008. Learning classifiers from only positive and unlabeled data[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. USA：213-220.

GOYES-PEÑAFIEL P，HERNANDEZ-ROJAS A，2021. Landslide susceptibility index based on the integration of logistic regression and weights of evidence：A case study in Popayan，Colombia[J]. Eng Geol，280：105958-105966.

HASTIE T，FITHIAN W，2013. Inference from presence-only data；the ongoing controversy[J]. Ecography，36(8)：864-867.

HU Q，ZHOU Y，WANG S，et al，2019. Improving the accuracy of landslide detection in “off-site” area by machine learning model portability comparison：a case study of Jiuzhaigou earthquake，China[J]. Remote Sens，11(21)：2530-2550.

JIMÉNEZ-VALVERDE A，2012. Insights into the area under the receiver operating characteristic curve（AUC）as a discrimination measure in species distribution modelling[J]. Glob Ecol Biogeogr，21(4)：498-507.

LEE W S，LIU B，2003. Learning with positive and unlabeled examples using weighted logistic regression[C]//Proceedings of the Twentieth International Conference（ICML 2003）.USA：448-455.

LI W，GUO Q，ELKAN C，2011. Can we model the probability of presence of species without absence data?[J]. Ecography，

34(6)：1096−1105.

LI W，GUO Q，ELKAN C，2020. One−class remote sensing classification from positive and unlabeled background data［J］. IEEE J Sel Top Appl Earth Obs Remote Sens，14：730−746.

LI W，GUO Q，2013. How to assess the prediction accuracy of species presence−absence models without absence data?［J］. Ecography，36(7)：788−799.

LI W，GUO Q，2021. Plotting receiver operating characteristic and precision−recall curves from presence and background data［J］. Ecol Evol，11(15)：10192−10206.

LIU B，DAI Y，LI X，et al，2003. Building text classifiers using positive and unlabeled examples［C］//Third IEEE International Conference on Data Mining. USA：179−186.

LOBO J M, JIMÉNEZ−VALVERDE A, REAL R, 2008. AUC: A misleading measure of the performance of predictive distribution models[J]. Global Ecol and Biogeography, 17 (2): 145−151.

LUCCHESE L V，DE OLIVEIRA G G，PEDROLLO O C，2021. Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using artificial neural networks［J］.CATENA，198：105067−105077.

PETERSON A T，PAPEŞ M，SOBERÓN J，2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling［J］. Ecol Model，213（1）：63−72.

REICHENBACH P，ROSSI M，MALAMUD B D，et al，2018.A review of statistically−based landslide susceptibility models［J］. Earth Sci Rev，180：60−91.

RICHARD M D，LIPPMANN R P，1991. Neural network classifiers estimate Bayesian a posteriori probabilities［J］. Neural Comput，3(4)：461−483.

SCHILIRÒ L，MONTRASIO L，MUGNOZZA G S，2016. Prediction of shallow landslide occurrence：Validation of a physically−based approach through a real case study［J］. Sci Total Environ，569/570：134−144.

SIFA S F，MAHMUD T，TARIN M A，et al，2020. Event−based landslide susceptibility mapping using weights of evidence（WoE）and modified frequency ratio（MFR）model：A case study of Rangamati district in Bangladesh［J］. Geolo Ecol Landsc，4(3)：222−235.

SOFAER H R，HOETING J A，JARNEVICH C S，2019. The area under the precision−recall curve as a performance metric for rare binary events［J］. Methods Ecol Evol，10（4）：565−577.

SOKOLOVA M, JAPKOWICZ N, SZPAKOWICZ S, 2006. Beyond accuracy, F−score and ROC: A family of discriminant measures for performance evaluation[M]//Lecture Notes in Computer Science. Berlin, Heidelberg: Springer: 1015−1021.

Van RIJSBERGEN C J，1979. Information retrieval［M］.2nd ed. London：Butterworths.

WARD G，HASTIE T，BARRY S，et al，2009. Presence-only data and the EM algorithm［J］. Biometrics，65（2）：554−563.

YILMAZ I，2010. Comparison of landslide susceptibility mapping methodologies for Koyulhisar，Turkey：Conditional probability，logistic regression，artificial neural networks，and support vector machine［J］. Environ Earth Sci，61（4）：821−836.

（责任编辑　秦社彩）