DOI:10.13203/j.whugis20160515 文章编号:1671-8860(2018)07-1085-07

基于优化随机森林模型的滑坡易发性评价

刘 坚[1,2] 李树林[1] 陈 涛[1]

1 中国地质大学(武汉)地球物理与空间信息学院,湖北 武汉,430074
2 中国地震局地震研究所地震大地测量重点实验室,湖北 武汉,430071

摘 要:以三峡库区沙镇溪镇-泄滩乡为研究区,探索基于最短描述长度原则的信息增益法对滑坡连续型因子进行离散的效果,计算皮尔森系数去除高相关因子。利用信息量法预测的极低、低易发区随机抽取非滑坡样本点。通过迭代计算袋外误差估计确定较优的随机特征及其数目,将优化后的随机森林对研究区滑坡进行易发性评价,并与逻辑回归等方法进行比较。绘制各算法预测结果的接收灵敏度曲线,其中优化后的随机森林预测结果的曲线下面积较高,达 91.8%,表明优化随机森林模型在滑坡易发性评价中具有较高的预测能力。

关键词:三峡库区;滑坡;离散化;随机森林;优化;易发性评价

中图分类号:P694;P208 **文献标志码**:A

在滑坡风险评估与管理中,滑坡易发性评价方法的探索一直是研究的热点。目前,滑坡易发性评价方法可分为以下几种:①基于经验的定性分析法,通过专家丰富的经验来判断滑坡的易发性,其缺点在于需要丰富的经验知识,主观性强。如刘阳运用经验定性分析模型对延长县滑坡地质灾害进行风险评估[1]。②半定量数学模型,一般有层次分析法、模糊综合评判法等。如许冲等利用层次分析法对汶川地震区滑坡进行易发性分析[2]。③确定性模型,通过斜坡的物理、水文参数计算斜坡稳定性,主要有极限平衡法等。如罗向奎等利用极限平衡法对杨家坝滑坡进行稳定性分析[3]。该方法可靠性高,但需水文、岩土体力学等诸多参数,数据可获取性低使其常局限于单个斜坡的稳定性计算。④定量数学模型,主要有逻辑回归、信息量、支持向量机等。此类方法具有数据可获取性高、预测精度较好等特点,常被用于滑坡易发性评价,但因算法复杂,往往不易解释。如王卫东等将确定性系数与逻辑回归模型运用于贵州省滑坡的危险性评价[4];王佳佳等利用信息量模型对滑坡进行预报预测[5];牛瑞卿等、武雪玲等将支持向量机运用于滑坡的易发性分析[6-7];Pradhan 将模糊逻辑回归模型运用于滑坡易发性评价[8]。

逻辑回归、决策树等定量数学模型多用单个模型进行预测,预测精度往往受限制,且易产生过拟合。为避免此类问题,人们提出了组合多棵决策树的随机森林模型,用于提升预测精度。随机森林可处理高维度、大数据量的数据集,且具有较高的泛化能力,与逻辑回归等传统方法相比具有一定的优势[9]。因此,本文利用随机森林对滑坡易发性进行研究,并从连续型因子离散化和选取样本等角度思考,探索较优的处理方法,通过迭代计算袋外误差估计寻找较优的随机特征以及数目,利用优化后的随机森林对滑坡易发性进行预测。

1 随机森林算法与模型

1.1 连续属性离散化

评价因子中连续型数据的离散化效果对预测结果有一定的影响,但当连续型属性较多且缺少经验时,数据变得不易处理。目前用于滑坡预测的连续型属性离散化并没有统一的方法,多数是根据经验定义、等频率、等宽度、自然断点法等进行处理[9],其离散化效果也常常受研究区限制。随机森林的连续属性离散化算法为基于最小基尼

收稿日期:2017-10-19
项目资助:国家高技术研究发展计划(863 计划)(2012AA121303)。
第一作者:刘坚,博士生,工程师,现从事云计算与地质灾害评估应用研究。linefanliu@163.com
通讯作者:李树林,硕士生。lishulincug@gmail.com

指数的信息增益离散方法，但其随机性使连续型属性的离散结果处于未确定状态，不利于具体滑坡因子的分析。因此，本文采用效果较优的基于最小描述长度原则的信息增益法（entropy based on minimal description length principle，Ent-MDLP）加以解决。具体步骤为：

1）二分递归寻找断点。每次在区间内寻找断点时，有若干候选断点（寻找不同类的相邻点，取它们之间的某点（如中点））。每个候选断点 T 都能将样本集合 S 划分为两个子集，分别计算两个子集的信息熵，然后加权求和，得到关于 T 的分类信息熵 $E(A,T,S)$。取使得分类信息熵最小的断点 T 作为最终选定断点。

2）确定递归停机条件。此处引入最小描述长度原则（minimal description length principle，MDLP），即总体信息量＝描述理论所需信息量＋描述不满足理论的异常所需信息量。停机条件是信息增益 G 应满足：

$$G(A,T,S)=E(S)-E(A,T,S)=E(S)-|S_1|/N\times E(S_1)-|S_2|/N\times E(S_2)>\log_2(N-1)/N+\log_2(3^k-2)-[k\times E(S)-k_1\times E(S_1)-k_2\times E(S_2)] \quad (1)$$

式中，A 为输入变量；T 为断点；S 为样本集合；N 为总样本量；k 为类别数量；$E(S)$ 为样本集 S 的熵；$E(S_1)$、$E(S_2)$为每个子区间内实例集 S_1、S_2 的熵；k_1、k_2为每个子区间的类别数量。式(1)表示增加的信息应大于最小描述长度，其优点是选出的断点为区分类的点，并使分类信息熵最小。

1.2 随机森林模型

随机森林是一种结合装袋法生成多份相互独立的训练集和多棵分类回归树（classification and regression tree，CART）来进行预测的集成学习方法，结果由投票得分最多或取平均决定[10-13]，其主要思想在于多个分类器组合判断的结果优于单个分类器的判断结果。

利用装袋法随机有放回地抽取 n 个（占总样本的2/3）样本作为独立空间训练集，对每个训练集分别建立 CART 树。其中随机选取 m 个因子（$m\leqslant$总因子数量）进行内部节点分支，且不做减枝处理，得到 n 棵独立的随机决策树[10]。综合 n 棵决策树的结果，取投票数最多的类或取其平均值作为结果。每次随机采样中未被抽取的1/3数据称为袋外数据（out of bag，OOB），利用这部分数据来进行内部误差估计，得到每棵树的 OOB 误差，对所有树的 OOB 误差取平均值得到随机森林的 OOB 误差。具体实现过程如图1所示。

OOB 误差是无偏估计，近似于交叉验证得到的误差，且由随机森林的泛化误差界有[12,14]：

$$P^*\leqslant\bar{\rho}(1-s^2)/s^2 \quad (2)$$

式中，P^* 为随机森林的泛化误差；$\bar{\rho}$ 为 CART 树间的相关度平均值；s 为决策树的平均强度。从式(2)可知，要增强随机森林的泛化能力，可减弱决策树间的相关度或增大决策树的强度。对此，通过对 CART 树的特征选择引入随机性，以减弱决策树间的相关度。具体做法为：随机选取 m 个（$m\leqslant$总特征数）特征，按照节点不纯度最小原则从这 m 个特征中选择最优的特征对节点进行分裂，此时 CART 树的强度和相关度受到了 m 的影响[14]。m 过小时，CART 树的强度偏弱；m 过大时，CART 树的强度增加，但 CART 树间的相关度也增加。本文采用迭代法计算不同随机特征数下随机森林的袋外误差，通过寻找最小的袋外误差来确定较优的随机特征数。

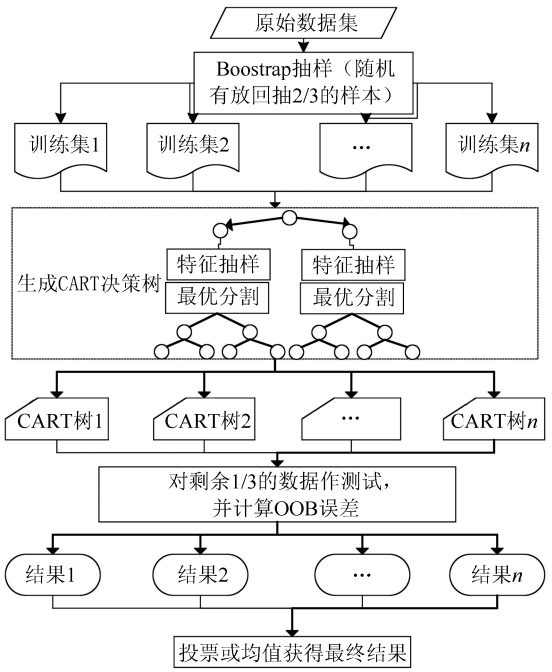

图1 随机森林算法示意图

Fig.1 Diagram of Random Forest Algorithm

2 区域滑坡灾害易发性评价

2.1 研究区概况

研究区位于长江三峡库区内地质环境相近的沙镇溪镇-泄滩乡，具体地理位置见图2。其长约21.6 km，面积约162.2 km^2。研究区地处川东褶皱与鄂西山地结合部，地形主要以高山峡谷为主，山高坡陡，平缓地带稀少，高程范围大致在60～1 150 m内[15]。据滑坡灾害编录资料可知，研究

区内已发生过 68 个滑坡,受 175 m 库水位影响的滑坡有 60 个,不受此库水位影响的滑坡为 8 个。滑坡灾害主要沿长江及其支流青干河两岸展布。该区滑坡灾害的主要诱发因素以降雨和三峡水库水位变动为主,其次为人类工程活动等[16]。

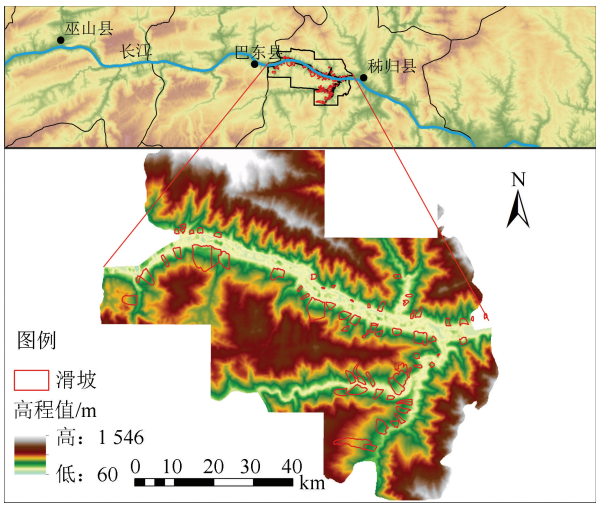

图 2　研究区地理位置及灾害分布图

Fig.2　Location of the Study Area and Distribution Map of Landslide Disaster

2.2　数据源

搜集研究区的遥感影像数据、数字高程模型(digital elevation model, DEM)、地质图和道路等基础地理数据,其分类及特性表见表 1。

表 1　实验数据分类及特性表

Tab.1　Classification of Experimental Data and Characteristics Table

数据类型	空间分辨率	数据用途描述	时间
Sentinel-2A	可见光与全色 10 m,多光谱 20 m	对已有道路等数据校正补充	2016-02-16
Landsat 8	全色 15 m,多光谱 30 m	提取土地利用、NDVI、NDWI 等	2013-09-15
DEM	30 m	提取高程、坡度等地形因子	
地质图	1∶50 000	提取地层岩性、断层等	

利用 ENVI 软件对 Landsat 8 影像提取土地利用、植被指数、地表湿度指数等指标;利用 ENVI 软件对 Sentinel-2A 影像 10 m 可见光波段进行投影配准、裁剪等处理,用于对已有道路等数据的校正与补充;30 m 空间分辨率的 DEM 用于分析该区域的坡度坡向等地形情况;通过 1∶50 000 的地质图获取该区域的地层、构造等地质情况;研究区的滑坡灾害数据来源于三峡库区地质灾害防治工作指挥部提供的三峡库区地质灾害分布数据,用于滑坡特征分析等。

2.3　评价因子

1)地质条件。地质条件属于滑坡灾害的控制因素,往往起着决定性作用。该区出露主要为三叠系和侏罗系等地层,工程岩组以软岩和软硬相间岩为主,西面有少部分硬岩。本文利用因子信息量分析其对滑坡的影响。信息量值越大,对滑坡影响越大;反之,对滑坡影响越小。由图 3(a)可知,硬岩的信息量为负,其值最小,对滑坡发生最不利;软岩和软硬相间两类信息量较大,对滑坡发生有利。通过地层产状、坡度与坡向划分该区的斜坡结构。由图 3(b)可知,伏倾坡、顺倾坡、飘倾坡对滑坡发生有利,逆斜坡、逆向坡对滑坡发生不利。地质构造上,该区位于秭归向斜南翼,断裂主要有仙人桥断裂、马鹿池断裂以及香炉断裂等。利用距断层的欧氏距离来表示断层对滑坡的影响。由图 4(d)可知,断层对滑坡的影响随断层影响距离的增加而呈现出先减后增再减的规律,在 2 318 m 处达到极弱值。

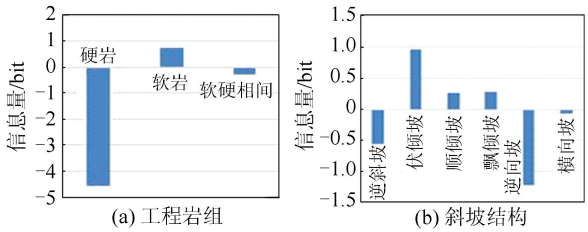

图 3　各因子信息量分布图

Fig.3　Information Distribution of Factors

2)地形地貌。地形地貌控制自然斜坡的临空条件,较大程度决定了滑坡的发育与分布状况。通过 SAGA GIS 软件对 30 m DEM 提取高程、坡度、坡向、凸性等地形地貌因子。由图 4(a)可知,高程的信息量值随着高程的增大而减小,表明高程越大,对滑坡的影响越低。由图 4(e)可知,坡度对滑坡的影响随着坡度的增大而减小。由图 4(f)可知,坡向对滑坡的影响随着坡向的增加而呈现先减后增的趋势,在 283°左右达到最小。由图 4(h)可知,凸性对滑坡的影响随着凸性的增大而减小。

3)水文条件。研究区多为涉水滑坡,强降雨、库水位周期性波动引起的地下水位变化是该区域滑坡的主要诱因。因难以直接获取地下水情况,本文利用 SAGA GIS 从 DEM 中提取了库水位影响、地形湿度指数、径流强度、Melton 崎岖数(一种累计流量的相关指数)等水文因子。通过 Landsat8影像提取地表湿度指数来表示地表湿

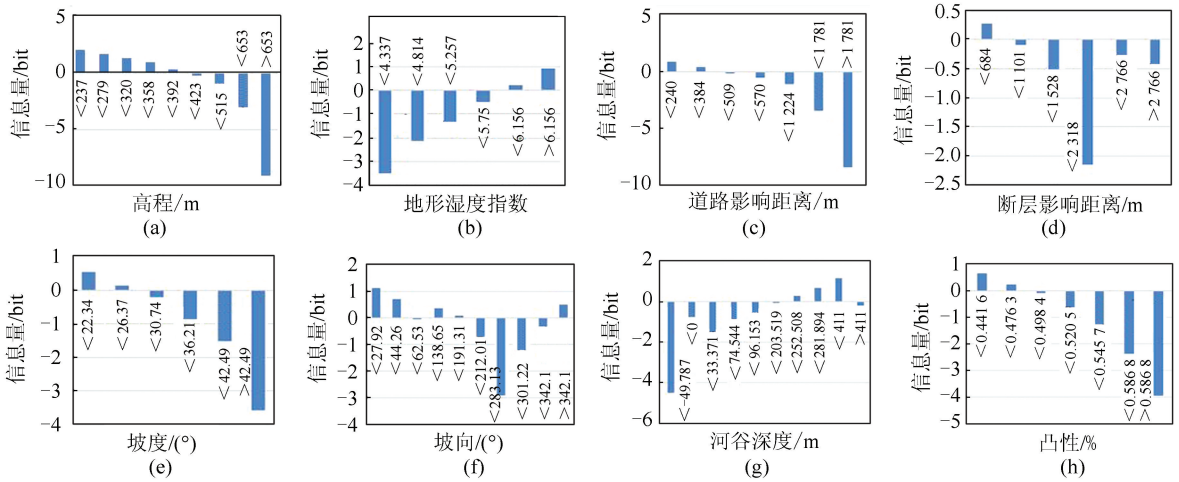

图4 主要因子信息量分布图

Fig.4 Information Distribution of Main Factors

度的情况。由图4(b)可知,随着地形湿度指数的增大,对滑坡的影响逐渐增大。

4)人类工程活动。研究区内受人类工程活动较强的斜坡区域常常是滑坡灾害多发区。利用收集的研究区道路并结合高分影像进行校正,计算距道路的欧氏距离;对经过大气校正后的Landsat 8影像提取归一化植被指数,作为地表植被的覆盖情况;并对Landsat 8影像进行全色融合,利用支持向量机对融合结果进行监督分类,得到该区的土地利用分类情况。由图4(c)可知,随着道路影响距离的增大,对滑坡的影响逐渐减小,这与实际相一致。

5)其他因素。地震通常也是滑坡等地质灾害的诱因。根据中国地震烈度区划图可知,研究区的地震烈度为Ⅵ度,属于地震弱发区,对滑坡的影响较弱,所以暂不考虑地震对该区滑坡的影响。

2.4 基于优化随机森林的滑坡易发性评价

1)连续型因子离散。利用SPSS软件中的最优离散化法(Ent-MDLP)对连续型因子进行离散化,并计算各级的信息量。研究区主要因子的离散效果见图4,具体对滑坡作用见§2.3。

2)因子相关性分析。通常情况下,各因子间存在着一定的相关性,这给模型预测带来信息的冗余。通过在R语言中计算各因子间的皮尔森相关系数,当其绝对值大于0.5,认为具有一定的相关性[15]。据此筛选出道路距离、高程、坡度、坡向等16个因子。

3)评价因子选择。在R语言中利用随机森林模型计算出各因子的不纯度平均减少值,将16个因子进行重要性排序,具体结果见图5。筛选出高程、地形湿度指数、道路影响、断层影响、坡度、坡向、河谷深度、凸性、流域强度、斜坡结构、工程岩组等11个较重要的因子,剔除地表湿度指数、曲率、土地利用类型、Melton崎岖数、归一化植被指数等5个影响较弱的因子。

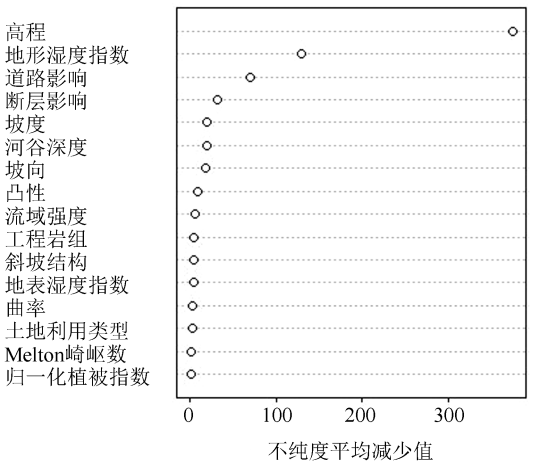

图5 各因子重要性分布图

Fig.5 Importance Distribution of All Factors

4)评价单元与样本选取。以分辨率30 m×30 m作为评价栅格单元大小,共划分180 219个评价栅格单元。在ArcGIS中随机选取80%的滑坡栅格单元作为滑坡训练样本,为避免滑坡的空间自相关性,将滑坡面降采样为90 m×90 m的栅格后再转为点,得到自相关性较弱的滑坡采样点。由于新生滑坡往往发生于暂未发生滑坡的区域,若直接对此类区域进行采样,可能会将潜在滑坡的栅格单元误视为非滑坡样本。为减少此类错误,通过ArcGIS随机生成点工具,对信息量法预测的滑坡极低易发区和低易发区内,随机选取约2倍于滑坡点数目的非滑坡样本点,以减少滑坡与非滑坡数据之间的不平衡性和空间的自相关

性。将滑坡点与非滑坡点合并后,提取各因子相应的数据作为训练数据(滑坡样本点数为 1 000 个,非滑坡样本点数为 2 199 个),剩余的数据则作为测试数据。

5)模型的建立。为寻找出较优的随机特征数,利用 R 语言循环迭代计算不同随机特征数的随机森林(random forest, RF)袋外误差,如图 6 所示。袋外误差越小,对应模型预测的精度越高。由图 6 可知,较优随机特征数为 4 个,且袋外误差并未一直随着随机特征数的增大而减小,当达到一定值时,袋外误差反而增大。此外,确定随机森林的决策树数目为 500 个。

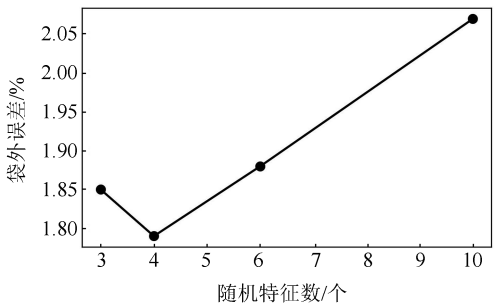

图 6　不同随机特征数下 RF 的 OOB 误差分布图

Fig.6　OOB Error Distribution of Random Forest with Different Numbers of Random Features

2.5　结果与分析

将随机森林、逻辑回归法(Logit)和支持向量机法(support vector machine, SVM)等对滑坡的预测概率作为易发性指数,利用剩余的 20%滑坡测试数据对模型进行检验,计算每种模型预测结果的接收灵敏度曲线(receiver operating characteristic curve, ROC)以及曲线下面积(area under the curve, AUC)(见图 7),进而比较各个模型的预测精度。由图 7 可知,优化后随机森林(optimized random forest, OPRF)预测结果的 AUC 值较高,达 0.918,预测精度比未优化的 RF 有较大提高,同时也高于其他模型的预测精度。

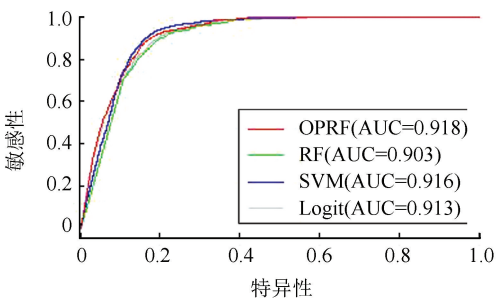

图 7　各种模型预测结果的 ROC 曲线

Fig.7　ROC Curves of Various Models' Prediction Results

对 OPRF 模型计算的滑坡易发性指数结合 Ent-MDLP 法进行分级处理,划分 0～0.060、0.060～0.269、0.269～0.711、0.711～0.960、0.960～1.000 共 5 个级别,分别对应极低易发区、低易发区、中易发区、高易发区、极高易发区 5 个等级。制作优化后随机森林的滑坡易发性分布图。由图 8 可知,滑坡高易发区主要分布于沿长江两岸受水库影响较强且公路或建筑密集的斜坡区域。南面的青干河流域,较典型的滑坡有千将坪滑坡、西陵路滑坡,远离库岸和公路的高山区域滑坡易发性较低。北面泄滩乡已发育的滑坡有庙岭包滑坡、杨坡岭砖厂滑坡等,远离库岸和公路的区域多为滑坡低易发区。长江南岸沿公路区的滑坡易发性明显高于北岸的非公路区。南岸发育的滑坡有树坪滑坡、白水河滑坡、范家坪滑坡等大型深层滑坡,北岸主要以中小型滑坡为主。图 8 表明研究区内滑坡的发育与水库、公路的影响有着较强的相关性。

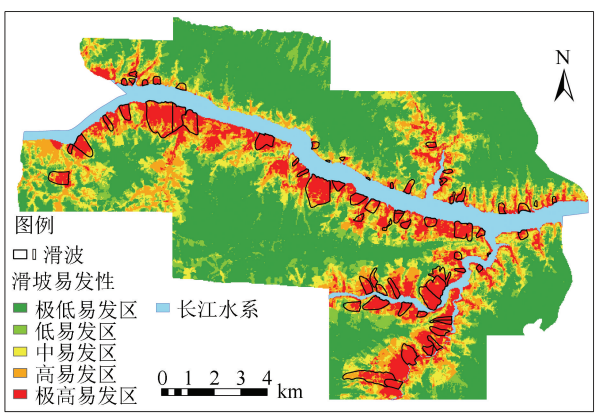

图 8　滑坡易发性分布图

Fig.8　Distribution of Landslide Susceptibility

表 2 为研究区滑坡易发性等级的灾害密度统计情况。由表 2 可知,预测结果中的高易发区和极高易发区占总滑坡面积比例达 95%;滑坡危险性(即滑坡灾害面积密度)从极高易发区到极低易发区呈明显的减小趋势,滑坡极高易发区的灾害面积密度最大,达 0.392 4;极低易发区的灾害面

表 2　危险性分区结果分析表

Tab.2　Analysis Table of Risk Zoning Result

易发区	面积 A/km²	滑坡面积 B/km²	占总滑坡面积比例/%	危险性 B/A
极低易发区	80.818 2	0.014 4	0.13	0.000 2
低易发区	20.517 3	0.050 4	0.47	0.002 5
中易发区	20.303 1	0.456 3	4.28	0.022 5
高易发区	19.963 8	2.065 5	19.36	0.103 5
极高易发区	20.594 7	8.081 1	75.75	0.392 4

积密度最小,几乎为零,这与滑坡实际分布规律相符合。

3 结 语

本文探讨了Ent-MDPL离散法以及随机森林的基本原理,并利用优化后的随机森林对研究区的滑坡易发性进行了评价,得出以下结论:

1)Ent-MDPL方法可较好解决当评价中的连续型因子增多且缺乏足够经验时的离散化问题,离散结果表现出明显的趋势特征,避免了随机森林的随机性给连续型因子分析带来的不便。

2)对于非滑坡区样本选取问题,采用分层抽样的思路,选取信息量模型评价结果中的极低易发区和低易发区进行随机采样,可减少将潜在滑坡点误分为非滑坡点的情况。

3)利用随机森林模型进行因子重要性排序,筛选出高程、地形湿度指数、道路影响等重要因子。本文采用迭代计算不同随机特征数的袋外误差估计来确定其较优参数;通过比较优化后的随机森林与传统模型预测结果的ROC曲线以及AUC值,可知优化后随机森林的预测精度较高。

参 考 文 献

[1] Liu Yang. Extension of the County Landslide Disaster Risk Assessment and Management Research [D]. Xi'an:Chang'an University, 2009(刘阳.延长县滑坡地质灾害风险评估和管理研究[D].西安:长安大学,2009)

[2] Xu Chong, Dai Fuchu, Yao Xin, et al. GIS-Based Landslide Susceptibility Assessment Using Analytical Hierarchy Process in Wenchuan Earthquake Region[J]. *Chinese Journal of Rock Mechanics and Engineering*,2009,28(a02):3 978-3 985 (许冲,戴福初,姚鑫,等. GIS支持下基于层次分析法的汶川地震区滑坡易发性评价[J].岩石力学与工程学报,2009, 28(a02):3 978-3 985)

[3] Luo Xiangkui, Fu Xuhui. Landslide Stability Analysis of Yangjiaba Based Upon Limit Equilibrium Method[J].*Shanxi Architecture*,2009,35(6):108-109(罗向奎,付旭辉.基于极限平衡法的杨家坝滑坡稳定性分析[J].山西建筑,2009,35(6):108-109)

[4] Wang Weidong, Chen Yanping, Zhong Sheng. Landslides Susceptibility Mapped with CF and Logistic Regression Model [J]. *Journal of Central South University (Science and Technology)*, 2009, 40(4): 1 127-1 132 (王卫东, 陈燕平,钟晟.应用CF和Logistic回归模型编制滑坡危险性区划图[J].中南大学学报(自然科学版),2009,40(4): 1 127-1 132)

[5] 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)

[6] Niu Ruiqing, Peng Ling, Ye Runqing, et al. Landslide Susceptibility Assessment Based on Rough Sets and Support Vector Machine[J]. *Journal of Jilin University (Earth Science Edition)*, 2012, 42(2):430-439(牛瑞卿,彭令,叶润青,等. 基于粗糙集的支持向量机滑坡易发性评价[J]. 吉林大学学报(地球科学版), 2012, 42(2): 430-439)

[7] Wu Xueling, Ren Fu, Niu Ruiqing, et al. Landslide Spatial Prediction Based on Slope Units and Support Vector Machines[J]. *Geomatics and Information Science of Wuhan University*, 2013, 38(12):1 499-1 503(武雪玲,任福,牛瑞卿,等. 斜坡单元支持下的滑坡易发性评价支持向量机模型[J].武汉大学学报·信息科学版,2013,38(12):1 499-1 503)

[8] Pradhan B. Manifestation of an Advanced Fuzzy Logic Model Coupled with Geo-information Techniques to Landslide Susceptibility Mapping and Their Comparison with Logistic Regression Modelling[J]. *Environmental and Ecological Statistics*, 2011,18(3): 471-493

[9] Cao Zhengfeng. Study on Optimization of Random Forests Algorithm[D].Beijing:Capital University of Economics and Business,2014(曹正凤.随机森林算法优化研究[D].北京:首都经济贸易大学, 2014)

[10] Breiman L. Random Forests[J]. *Machine Learning*, 2001,45(1): 5-32

[11] Fang Kuangnan, Wu Jianbin, Zhu Jianping, et al. A Review of Technologies on Random Forests[J]. *Statistics & Information Forum*, 2011,26(3):32-38(方匡南,吴见彬,朱建平,等.随机森林方法研究综述[J]. 统计与信息论坛,2011,26(3):32-38)

[12] Li Zhengui. Several Research on Random Forest Improvement [D]. Xiamen: Xiamen University, 2013 (李贞贵.随机森林改进的若干研究[D].厦门:厦门大学,2013)

[13] Dong Shishi, Huang Zhexue. A Brief Theoretical Overview of Random Forests[J].*Journal of Integration Technology*,2013,2(1):1-7 (董师师,黄哲学. 随机森林理论浅析[J].集成技术, 2013,2(1):1-7)

[14] An Zhou. Hard Drive Failure Prediction Based on

Random Forest[D]. Tianjin: Nankai University, 2014(安洲.基于随机森林的硬盘故障预测算法的研究[D].天津:南开大学, 2014)

[15] Peng Ling. Landslide Risk Assessment in the Three Gorges Reservoir[D]. Wuhan: China University of Geosciences, 2013(彭令.三峡库区滑坡灾害风险评估研究[D].武汉:中国地质大学,2013)

[16] Tian Zhengguo, Cheng Wenming, Lu Shuqiang, et al. Control and Triggering Factors Analysis of Landslides and Rockfalls in the Three Gorges Reservoir Area[J]. *Resources Environment & Engineering*, 2013, 27(1):50-55 (田正国,程温鸣,卢书强,等.三峡库区滑坡崩塌发育的控制与诱发因素分析[J].资源环境与工程, 2013, 27(1):50-55)

Landslide Susceptibility Assesment Based on Optimized Random Forest Model

LIU Jian[1,2]　*LI Shulin*[1]　*CHEN Tao*[1]

1　Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China

2　Key Laboratory of Earthquake Geodesy, Institute of Seismology, CEA, Wuhan 430071, China

Abstract: The research area is located in Shazhenxi town and Xietan town of Three Gorges reservoir area in this paper. In order to obtain better results that discrete the continuous factors of landslide, entropy based on minimal description length principle(Ent-MDLP) method is used. To avoid the influence of correlation between factors, we calculate the Pearson correlation coefficient to remove high correlation factor. In order to obtain more accurate non-landslide sample points, the non-landslide sample points are randomly selected from the very low and low susceptible regions predicted by the entropy method. For the optimized random forests model, the optimal random features and its number are determined by iterative calculation of out-of-bag error estimation. Then the optimized random forest is evaluated for the landslide of the study area, and the landslide susceptibility level is divided. The model is compared with the methods of logistic regression, support vector machine and non-optimized random forest. The accuracy of each model is evaluated by plotting the receiver sensitivity curve of each algorithm. The optimized random forest's area is the highest, which the area under the curve is 91.8%. These show that the random forest model is optimized with more high-predictive power in landslide-prone assessment.

Key words: Three Gorges reservoir area; landslide; discretization; random forest; optimization; susceptibility assessment

First author: LIU Jian, PhD candidate, engineer, specializes in cloud computing and geological disaster assessment. E-mail: linefanliu@163.com

Corresponding author: LI Shulin, postgraduate. E-mail: lishulincug@gmail.com

Foundation support: The National High Technology Research and Development Program of China(863 Program), No. 2012AA121303.