

DOI:10.16031/j.cnki.issn.1003-8035.2019.05.04

“阶跃式”滑坡突变预测与核心因子提取的 平衡集成树模型

何少其^{1,2}, 刘元雪^{1,2}, 梁叶³, 刘娜², 赵久彬¹

(1. 陆军勤务学院岩土力学与地质环境保护重庆市重点实验室, 重庆 401311;

2. 重庆市地质矿产测试中心, 重庆 400042; 3. 重庆长江勘测设计院有限公司, 重庆 401147)

摘要:“阶跃式”滑坡在复杂多变的地质环境作用下呈现突变与稳定交替、不平衡的演化形态, 根据此特点提出了边坡突变与稳定分类的平衡集成树模型, 建立高维地质环境影响因子与致变之间的关联, 并应用于三峡库区 26 个具有“阶跃”特征的滑坡。考虑到直接使用分类器难以捕捉有效信息, 模型利用了合成边界少数类过采样原理适当提升突变样本比率, 再分别组合随机森林和梯度提升树进行优化和训练, 并对测试集和预测集进行评估校核。同数据平衡前后的不同模型进行对比实验, 平衡集成树模型能够有效提高突变预测的整体精度, 并量化得出了所有地质环境因子的特征重要性指标, 最终应用于样本外的王爷爷滑坡结果表明, 模型取得了较高的预报水平。该方法能够实现突变的有效预测, 并发现核心影响因子及其数据分布特征, 为研究“阶跃式”滑坡机理和预警提供了新的思路。

关键词:“阶跃式”滑坡; 平衡集成树; 三峡库区; 突变; SMOTE; 随机森林; 梯度提升树

中图分类号: P642.22

文献标识码: A

文章编号: 1003-8035(2019)05-0027-10

Balanced decision tree ensemble model for catastrophe prediction and key factors' extraction of step-like landslides

HE Shaoqi^{1,2}, LIU Yuanxue^{1,2}, LIANG Ye³, LIU Na², ZHAO Jiubin¹

(1. Army Logistics University of PLA, Chongqing Key Laboratory of Geomechanics and Geoenvironment Protection, Chongqing 401311, China; 2. Chongqing Testing Center of Geology and Mineral Resources, Chongqing 400042, China; 3. Chongqing Yangtze Survey Design Institute Co., Ltd, Chongqing 401147, China)

Abstract: Step-like landslides show abrupt change, stable alternation and unbalanced evolution under the complex and changeable geological environment. Based on this feature, a balanced decision tree ensemble model of slope catastrophe and stable classification is proposed to establish a correlation between high-dimensional geological environmental impact factors and mutations, and it is applied to 26 typical step-like landslides in the Three Gorges Reservoir Area. Considering that it is difficult to capture effective information by directly using the classifier, the model uses the synthetic borderline minority over-sampling principle to appropriately increase the mutation sample ratio, and then combines the random forest and the gradient boosted decision tree for optimization and training, and evaluates and verifies the test set and prediction set. Compared

收稿日期: 2019-01-12; 修订日期: 2019-04-27

基金项目: 国家自然科学基金项目(41877219); 重庆市自然科学基金项目(cstc2019jcyj-msxmBX0585); 重庆市规划和自然资源局科技计划项目(KJ-2018016); 陆军勤务学院研究生创新项目(LY180510)

第一作者: 何少其(1990-), 男, 江西婺源人, 硕士研究生, 主要从事滑坡监测大数据预警方面的研究工作。E-mail: hesqlq@outlook.com

通讯作者: 刘元雪(1969-), 男, 重庆人, 博士, 教授, 博士生导师, 主要从事岩土体本构关系与地下工程稳定性的教学与科研工作。E-mail: lyuanxue@vip.sina.com

with different models before and after data balance, the balanced decision tree ensemble model can effectively improve the overall accuracy of the catastrophe prediction, also the feature importance index of all geological environmental factors is obtained. Finally, the results applied to the Wangyemiao Landslide show that the optimization model has achieved a higher level of forecasting. This method can achieve effective catastrophe prediction, and finds the core influence factors and their data distribution characteristics, which provides the ideas for studying the step-like landslide mechanism and early warning.

Keywords: step-like landslide; balanced decision tree ensemble; Three Gorges Reservoir Area; catastrophe; SMOTE; random forest; gradient boosted decision tree

0 引言

“阶跃式”滑坡是三峡库区的典型滑坡类型^[1],具有分布广泛、危害性大、成因复杂、难以预测等特点^[2]。大量研究表明,库区降雨和水位升降是引起此类滑坡失稳的主要诱因^[3-4],受此影响滑体的位移时间曲线呈现突变与稳定交替的演化形态,而突变又与稳定性下降直接相关^[5]。因此,利用海量的位移和环境数据进行分析、挖掘是研究和预防此类滑坡灾害的有效途径^[6],其中提升突变的预测精度和确定核心相关因子是关键问题。

近年来,“阶跃式”滑坡研究分析多以位移预测为主,应用了时间序列、支持向量机、神经网络等算法及其组合。彭令等^[7]、MIAO F S 等^[8]分别利用时间序列将位移分解,并通过多项式函数和支持向量机耦合模型分别对趋势项和季节项进行拟合,并有效应用于总位移预测;李骅锦等^[9]提出了 ARMA 与 LASSO-ELM 的联合模型,并选取 Copula 函数对诱因数据与位移预测数据的联合分布进行描述并提取诱因阈值和位移风险值;杨背背等^[10]通过多项式拟合趋势项位移、LSTM 模型拟合周期项位移,获得了预测精度较高的叠加动态模型。这类方法应用于滑坡位移跟踪时能够达到较高精准度,但多适用于单滑坡体预报,难以获取共性规律。与此类方法不同的是,文章提出了考虑高维地质环境影响因子作用下区域内“阶跃式”滑坡突变与稳定的分类优化模型,目前类似的思想和方法还不多见。

从数据层面来看,“阶跃式”滑坡的突变和稳定数据在数量上存在严重的不平衡问题,势必引起分类器性能下降。针对此类问题,CHAWLA N V 等^[11]提出了合成少数类过采样技术(Synthetic Minority Over-sampling Technique, SMOTE),HAN H 等^[12]通过增强分类边界附近少数类过采样得到改进的 borderline-SMOTE 算法。从分类器层面来看,单分类器如 k 近邻(KNN)、支持向量机(SVM)、逻辑回归(LR)、朴素贝

叶斯(NB)、决策树(CART)等精度受限,且易造成模型过拟合。而随机森林(Random Forest, RF)^[13]和梯度提升树(Gradient Boosted Decision Tree, GBDT)^[14]这两种集成树算法利用了算法集成的思想,具有泛化性能优、精度高、计算速度快、多分类预测等优势。

基于以上考虑,本研究设计的平衡集成树,就是通过 borderline-SMOTE 中的 1 型(简称 borsmote1)合成控制适当比例的边界突变少数类以平衡数据源,对随机森林和梯度提升树进行优化;设计分层十折交叉验证和网格搜索的组合,利用剪枝提升泛化性能,最终使集成树提升至较高预测精度的平衡状态;结合海量“阶跃式”滑坡数据集,对多源异构数据进行归一化处理并用于训练与测试,结果评估利用了监督学习中的受试者工作特征(Receiver Operating Characteristics, ROC)标准^[15],并定义综合精度平均值指标,模型得出的预测性能整体优于单集成树以及平衡前后的不同分类器;以树节点纯度计算对 26 个地质环境因子进行了排名,相比于常用的相关性系数,能够有效提取出核心影响因子。通过平衡集成树算法,使“阶跃式”滑坡的突变预测达到了较高的预报水准,确定出的核心相关因子可作为后续机理研究参考分析,为区域内滑坡理论分析与防治提供新的思想与方法。

1 平衡集成树算法

平衡集成树模型融合了 borsmote1 与集成树的算法优势,并结合实际的“阶跃式”滑坡数据进行应用。它的主要思想是利用训练集突变样本合成边界少数类,使数据达到测试分类性能最佳的平衡状态,从而对 RF 和 GBDT 进行改进和优化,并建立该状态下的最终模型。

1.1 数据平衡规则

传统的过采样和欠采样方法或造成模型的过拟合、或以牺牲大量宝贵数据为代价,均难以满足数据激增的发展趋势。因此,本文采用突变边界样本合成的方法平衡数据层。具体规则为:

(1)“阶跃式”滑坡数据全集为 T , 突变样本集合为 $P = \{p_1, p_2, \dots, p_{pn}\}$, 稳定样本集合为 $N = \{n_1, n_2, \dots, n_{nn}\}$, 其中 pn 和 nn 分别为突变样本和稳定样本的数量。在 T 范围内寻找 P 中所有样本 $p_i (i = 1, 2, \dots, pn)$ 的 m 最近邻, m' 是 m 最近邻中稳定样本数。

(2) 当 $m' = m$, 表明 p_i 被多数类样本完全包裹, 此时被判别为噪声点; 当 $m' \in [m/2, m)$, 表明 p_i 的 m 最近邻中稳定样本多于突变样本, 突变极大可能被错分为稳定, 此时被判别为危险点; 当 $m' \in [0, m/2)$, 表明 p_i 的 m 最近邻中突变样本多于稳定样本, 此时被判别为安全点。

(3) 在广义欧式空间中, 危险点即是 P 的边界样本, 记为 p^* 。

(4) 通过不同比率调节对 p^* 随机选取的最近邻中 k 个突变样本 p_j , 按照公式(1)合成属于突变类的新样本 p_{new} , 所有新样本加入 P 组成新的合成突变样本集合 P_{new} 。

$$p_{new} = p^* + rand \times (p_j - p^*) \quad (1)$$

式中, $rand$ ——在区间 $(0, 1)$ 内的随机数, $j = 1, 2, \dots, k$ 。

1.2 集成树参数优化

集成树参数优化的目标是改善集成树的普适性与鲁棒性, 有效保证其对新数据的预测精度。对于 RF 和 GBDT 两种集成树而言, 已被证明子树的数量越多, 对随机状态的适应过程越稳定, 但是训练和预测时长相应增加。通过实际测试与权衡, 本研究均使用 300 棵子树用于集成树构建。由于 GBDT 采用了 Boosting^[16] 的策略, 利用损失函数 (Loss Function) 在梯度下降方向持续迭代提升模型准确率, 本文通过设置学习率为 0.1 来调节减少残差的速率。同时, 为了模型达到最优泛化的目的, 根据机器学习经验法则拆分出总数据集的 25% 作为测试集, 测试集不参与模型的训练过程, 只在最后用于评估训练模型对于新数据的泛化表现。训练过程中, 采取分层十折交叉验证和网格搜索手段来综合选择划分特征子集、控制树的深度等进行参数调优, 对子树进行预剪枝操作。选取对高维多噪音数据敏感的基尼系数 (Gini Impurity) 作为纯度计算标准, 节点不纯度值越低, 模型拟合越好。

1.3 计算方法

平衡集成树模型应用于“阶跃式”滑坡突变预测与特征提取的计算方法为:

(1) 将多源、海量、异构数据归一化为可挖掘的数据信息, 构建原始数据集, 并拆分为训练集和测试集;

(2) 训练集通过不同比率参数合成突变边界少数类样本, 得到平衡化数据集, 供 RF、GBDT 训练模型;

(3) 以测试集综合精度平均值为评估指标, 选定最佳测试精度下的参数优化组合, 建立平衡集成树模型;

(4) 预报区域内“阶跃式”滑坡突变, 根据模型特征重要性提取核心相关因子 (图 1)。

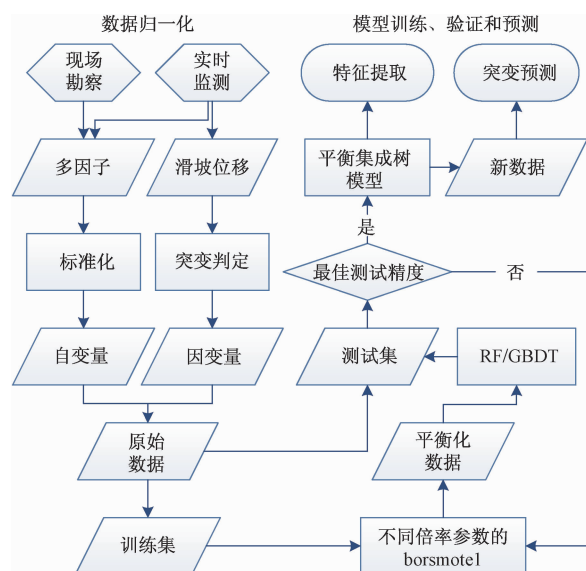


图 1 平衡集成树模型流程

Fig. 1 Flow chat of the balanced decision tree ensemble model

1.4 评估指标

结合滑坡预测的实际情况, 采用 ROC 标准用于预测结果评估, 具体指标包括: 召回率 (True Positive Rate, TPR), 为正类样本中被正确预测的占比; f_1 测度 (f_1 -score, f_1), 为召回率和准确率 (被预测为正类样本中实为正类的占比) 的调和平均; AUC 值 (Area Under Roc Curve, AUC), 等价于分类器对正类中随机样本评分高于反类中随机样本的概率。

定义有关概念如下: 稳定多数类作为正类计算时, 召回率为 TPR^0 , f_1 测度为 f_1^0 ; 同理, 突变少数类作为正类计算时, 召回率为 TPR^1 , f_1 测度为 f_1^1 ; $\bar{f}_1 = (f_1^1 + f_1^0)/2$, 表示稳定样本、突变样本的综合精度平均值, 与传统计算精度 (Accuracy)、单类别的 f_1 测度相比, 综合精度平均值指标可以比较全面地评价不平衡数据下模型对突变与稳定样本预测的整体精度。所有评估指标均默认表示测试和预测精度。

2 研究区滑坡概况及数据集

2.1 研究区滑坡地质环境

本文选取的三峡库区内 26 个“阶跃式”岸坡

(图2),均为堆积层-基岩组合形式,以侏罗系地层的砂岩、泥岩为主,地层岩性具有易风化、遇水软化等特点,剪出口相近,前缘位于175 m库水位以下。库区水位为145~175 m,涨落高差达30 m,相应的库岸地灾水位变化也基本如此。该区属亚热带季风性湿润气

候,全年的气候特征为湿度大、降雨充沛,年平均降雨量达到996.7~1309.9 mm。夏季长而热、多暴雨,秋多阴雨,冬暖、春旱。丰富的水资源为“阶跃式”滑坡灾害提供了外部条件,使得岩土体强度降低,坡体自重变化,产生动水压力和浮托力,使得坡体稳定性急剧下降。

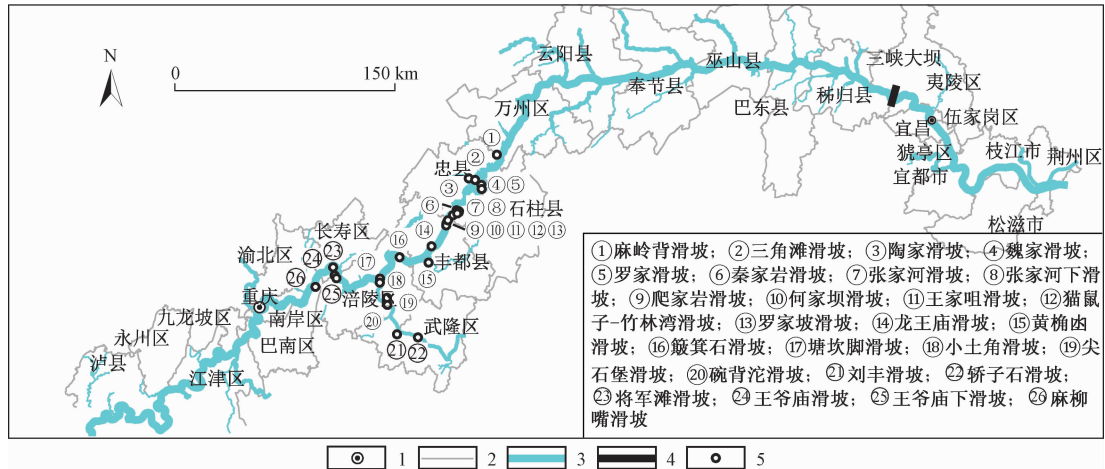


图2 研究区滑坡分布

Fig. 2 Distribution of landslides in the study area

1—行政中心;2—区县界;3—库区;4—三峡大坝;5—滑坡。

2.2 数据归一化处理

滑坡的地质环境因子分为内部和外部^[17]。内部因子数据包括地形地貌、地层岩性和地质构造^[18]等,基本趋于稳定,由现场勘查获取;外部因子数据包括降雨和库水^[19]等,它随时间活跃变化,由监测点实时监控。因子选取尽量体现真实数据的差异性,但由于调研数据难以包含所有地质情况,部分因子及分级类别不够齐全。汇编数据以月为周期,并考虑了降雨和库水的滞后效应^[20-21],最终选取可获取的、较为完整的26个地质环境因子(表1)。为了验证模型可行性、实用性,随机抽取王爷庙滑坡作为预测集校核模型,另25个滑坡构成因子计算、机器学习及测试的总数据集。

表1 26个滑坡地质环境因子

Table 1 26 Geological environment factors of landslides

信息源	因子类型
地质内部数据	地形地貌(坡度、滑体厚度、剖面形态、坡高、坡长、坡宽、前缘高程、坡向、距175 m水位线水平距离)
	地层岩性(基岩地层岩性、堆积层岩性)
	地质构造(岩层产状关系)
	地下水(地下水类型、地表水补给)
地质外部数据	降雨影响(当月降雨量、前1个月降雨量、前2个月降雨量、单日最大降雨量、降雨超过0.1 mm天数)
	库水影响(月平均库水位、月最低库水位、月最高库水位、当月水位变化量、前1个月水位变化量、前2个月水位变化量)
	力学机制(滑坡类型)

数据的归一化是算法实现的前提。由于数据平衡化是基于广义欧式距离实现的,针对离散型因子,先采用各因子相应的二级因子突变频率比^[22]进行数据连续化,作为衡量因子的量化指标:

$$D_{i,j} = \frac{n_{i,j}}{m_{i,j}} \quad (2)$$

式中, $n_{i,j}$ ——各类二级因子突变样本;

$m_{i,j}$ ——该二级因子集合,其中*i*为环境因子序号,*j*为所在因子分类的二级因子序号。

再将所有连续型在内的因子离差标准化,映射至[0,1]区间消除量纲影响,作为自变量供模型使用,离散型因子分级与连续化结果如表2至表9所示。

表2 滑坡类型因子分级与连续化

Table 2 Classification and continuous of landslide type

分级序号	二级因子	突变频率比	标准化值
1	牵引式	0.077 1	1.000 0
2	混合式	0.057 2	0.179 0
3	推移式	0.052 8	0.000 0

表3 剖面形态因子分级与连续化

Table 3 Classification and continuous of profile morphology

分级序号	二级因子	突变频率比	标准化值
1	直线形	0.068 4	1.000 0
2	折线形	0.057 4	0.194 7
3	圆弧形	0.054 7	0.000 0

表 4 基岩地层岩性因子分级与连续化

Table 4 Classification and continuous of bedrock lithological

分级序号	二级因子	突变频率比	标准化值
1	S ₁₁	0.111 1	1.000 0
2	J _{1-2zl}	0.107 8	0.951 4
3	T ₂₁	0.098 8	0.816 3
4	J _{2s}	0.071 0	0.403 1
5	T _{1j}	0.065 5	0.321 5
6	J _{1z}	0.062 5	0.276 5
7	J _{3p}	0.053 1	0.136 7
8	J _{3sn}	0.043 9	0.000 0

表 5 堆积层岩性因子分级与连续化

Table 5 Classification and continuous of accumulative formation lithological

分级序号	二级因子	突变频率比	标准化值
1	Q_4^{el+dl}	0.102 0	1.000 0
2	Q_4^{del} 、 Q_4^{el+dl} 、 Q_4^{al+pl}	0.075 9	0.527 8
3	Q_4^{el+dl} 、 Q_4^{al+pl}	0.072 8	0.471 7
4	Q_4^{del}	0.065 0	0.332 3
5	Q_4^{del} 、 Q_4^{el+dl}	0.046 6	0.000 0

表 6 岩层产状关系因子分级与连续化

Table 6 Classification and continuous of rock stratum occurrence

分级序号	二级因子	突变频率比	标准化值
1	横向	0.103 7	1.000 0
2	斜向	0.068 1	0.338 9
3	顺向	0.063 3	0.250 7
4	逆向	0.049 8	0.000 0

表 7 坡向因子分级与连续化

Table 7 Classification and continuous of aspect

分级序号	二级因子	突变频率比	标准化值
1	阴坡	0.113 0	1.000 0
2	半阴坡	0.066 3	0.249 7
3	半阳坡	0.050 7	0.000 0

表 8 地下水类型因子分级与连续化

Table 8 Classification and continuous of groundwater type

分级序号	二级因子	突变频率比	标准化值
1	孔隙-裂隙水	0.064 1	1.000 0
2	裂隙-溶隙水	0.061 7	0.000 0

表 9 地下水补给因子分级与连续化

Table 9 Classification and continuous of groundwater recharge

分级序号	二级因子	突变频率比	标准化值
1	大气降水、地表水	0.092 6	1.000 0
2	大气降水	0.061 2	0.000 0

滑坡地表水平位移监测数据作为突变依据,结合“阶跃式”滑坡曲线连续突变的特征,以统计学中离群值检验的四分位上界^[23]对各监测点月位移量进行划

分,用作数据集的因变量,其中突变记作 1、稳定记作 0。将突变频次与降雨量、库水位比较,发现三者具有较强的相关性。图 3 为 2006 年 6 月至 2018 年 8 月期间,研究区“阶跃式”滑坡总突变频次与月平均降雨量、月平均库水位的关系。

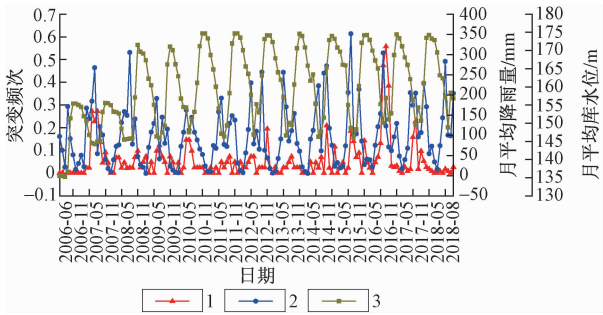


图 3 “阶跃式”滑坡的总突变频次与月平均降雨量、月平均库水位的相关性曲线

Fig. 3 The correlation curve between the catastrophe frequency of step-like landslides and monthly average rainfall, monthly average reservoir water level

1—突变频次;2—月平均降雨量;3—月平均库水位。

2.3 “阶跃式”滑坡数据集

建立所有变量总和的“阶跃式”滑坡数据库,总数据集按 3:1 划分为训练集和测试集,其中训练集根据不同过采样比率参数合成新的样本集用于集成树训练使用,测试集不参与模型的训练过程用于评估训练模型对于新数据的泛化表现。如表 10 所示,数据集中样本数目分布严重不平衡,直接用于传统分类算法必将导致训练结果更倾向于稳定类,对突变类学习效率降低。因此,在数据层面对目标变量进行平衡优化是提升突变信息学习效率的突破点。

表 10 “阶跃式”滑坡数据集

Table 10 The data set of step-like landslide

数据集	样本数/条	突变样本数/条	稳定样本数/条	样本比例/%
训练集	6 156	396	5 760	6.88
测试集	2 052	130	1 922	6.76
总数据集	8 208	526	7 682	6.85

3 模型应用与分析

3.1 模型测试结果比较

模型测试结果比较分为三步:(1)采用初始训练集分别训练传统单分类算法(KNN、SVM、LR、NB、CART)和集成树算法(RF、GBDT),并进行测试结果评估;(2)采用比率参数为 1 的平衡化训练集对上述方法进行训练并测试评估;(3)通过比率参数搜索构造

两种最佳平衡状态下的平衡集成树(Borsmote1-RF 和 Borsmote1-GBDT)模型。

(1)传统单分类算法与集成树算法的测试结果比较

实验对所选模型分别进行了泛化调参,用没有参与训练的测试集进行结果评估(表 11)。

表 11 初始训练集对应的不同算法测试结果
Table 11 Different algorithm test performances
corresponding to the raw training set

类别	算法	TPR ¹ /%	f_1^1 /%	$\bar{f}_1$ /%	AUC/%
单分类算法	KNN	29.23	36.02	66.28	82.02
	SVM	47.69	50.00	73.39	82.76
	LR	0.77	1.49	49.11	79.13
	NB	52.31	30.36	60.91	77.17
	CART	48.46	45.49	70.77	72.06
集成树算法	RF	47.69	52.77	74.95	88.38
	GBDT	47.69	51.67	74.33	89.62

结果显示集成树模型应用于“阶跃式”滑坡数据集的测试精度指标普遍高于单分类模型。其中 NB 的 TPR¹ 值虽然较高,但是其它指标处于较低水平,表明在突变预测过程中,将大量的稳定样本错误预报为突变。两种集成树模型的 $\bar{f}_1$ 和 AUC 指标均分别达到 74% 和 88% 之上。

(2)比率参数为 1 的平衡优化算法比较

比例参数表示过采样后突变类数量与稳定类之比,采用比率参数为 1 对测试集进行平衡操作,此时突变类与稳定类数量相等,以此分别对(1)中的所有方法进行优化。训练过程同样采取了泛化措施,优化后的训练模型用于测试集预测结果评估(表 12)。

表 12 比率参数为 1 的平衡优化算法测试结果
Table 12 Balanced optimization algorithm test performances
at a ratio of one

类别	算法	TPR ¹ /%	f_1^1 /%	$\bar{f}_1$ /%	AUC/%
单分类优化算法	Borsmote1-KNN	63.08	40.30	66.86	82.68
	Borsmote1-SVM	59.82	52.14	74.47	85.45
	Borsmote1-LR	70.77	24.50	53.80	79.30
	Borsmote1-NB	73.85	22.15	50.65	76.32
	Borsmote1-CART	56.15	48.03	71.93	75.36
集成树优化算法	Borsmote1-RF	66.15	57.14	76.88	90.23
	Borsmote1-GBDT	55.39	54.96	75.95	89.48

经平衡化后,模型测试精度整体提高,AUC 值平均增加了 1.10%,只有少数指标出现了下降,证明数据的平衡化对突变预测性能提升有效。LR 和 NB 优化模型 TPR¹ 值显著提升,然而会造成过量的平稳被错分为突变,这点由 f_1^1 和 $\bar{f}_1$ 偏低可知。两种集成树优

化后的 TPR¹ 值分别大幅提升了 18.46% 和 7.70%, Borsmote1-RF 的 AUC 值达到了 90.23%。因此,考虑采用不同比率值进一步优化,最大限度提高突变预测水准。

(3)构造最佳平衡状态下的平衡集成树模型

合成突变类过程中,比率参数使合成突变类数线性增加,决定了数据平衡的程度,直接影响分类器性能。之前的实验已经证明比率值为 1 时平衡优化的集成树算法精度最高,通过调节比率值可以进一步挖掘训练模型的最佳平衡状态。每种组合的平衡优化集成树模型同样经过泛化训练获得,整个平衡过程的结果如图 4 所示。

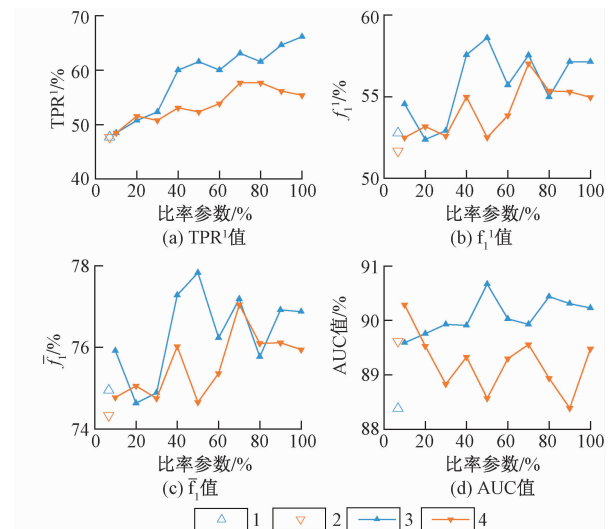


图 4 TPR¹ f_1^1 $\bar{f}_1$ 和 AUC 预测精度随测试集比率参数变化曲线图

Fig. 4 The values of TPR¹ f_1^1 $\bar{f}_1$ and AUC vary with ratio parameters

1—RF; 2—GBDT; 3—Borsmote1-RF; 4—Borsmote1-GBDT。

图 4 说明了数据的平衡化过程对平衡集成树性能的影响。TPR¹ 值随比率参数增大基本呈上升趋势; f_1^1 、 $\bar{f}_1$ 和 AUC 值整体上先快速上升然后缓慢下降,中间有较大波动,三者趋势比较相似。Borsmote1-RF 模型在比率为 0.5 时,测试精度最高,而 Borsmote1-GBDT 模型当比率为 0.7 时效果最佳,可以认为这两点为两种模型的最佳平衡状态。当比率继续增大时,模型测试结果精度下降,这是因为合成少数类样本并非真实样本,过多亦会造成数据集质量降低,平衡化程度并非越平均越好。

选取最佳平衡状态的优化模型作为最终模型,整体预测的正确率再次提升,只有 Borsmote1-RF 模型的

TPR¹ 较比率为 1 时少量下降。最终模型的 $\bar{f}_1$ 值相比于 RF 和 GBDT 分别增加 2.88% 和 2.72% ;AUC 值中 Borsmote1-RF 增加 2.29% ,Borsmote1-GBDT 则保持在原有水平;两种模型在 TPR¹ 指标上相应增加了 13.85% 和 9.70% ,这对突变的预报尤为关键,且对其余指标提升做出了贡献(表 13)。

表 13 最佳平衡状态下的平衡集成树测试结果
Table13 Test results of balanced decision tree ensemble
under optimal equilibrium

类别	算法	TPR ¹ /%	f_1^1 /%	$\bar{f}_1$ /%	AUC /%
平衡	Borsmote1-RF(0.5 比率)	61.54	58.61	77.83	90.67
集成树	Borsmote1-GBDT(0.7 比率)	57.39	57.03	77.05	89.56

3.2 核心因子提取

因子与突变的关联是非线性和离散的,而采用相关性系数分析相关性前提是线性情况,因此采用以往常用方法提取因子特征具有局限性。平衡集成树模型对于“阶跃式”滑坡突变预测的精度得到了有效提升,表明模型在数据平衡过程中经过边界增强、噪音消减更易于捕捉到了突变的核心相关因子,可以认为基于平衡集成树信息增益率给出的因子特征重要性比集成树更具有参考与实用价值。按照 borsmote1-RF(0.5 比率)模型计算的 26 个地质环境因子特征重要性进行降序排列,图 5 得出集成树改进前后该研究区的因子重要性。

由图 5 可知,外部因子较内部因子影响更为显著,其中降雨、库水位和内部因子三者的比重和约为 27.85%、

52.43% 和 21.00% ,该比例受因素数量影响,但可作为定性参考分析。可见,库水位为三峡库区“阶跃式”滑坡的主导因素,且水位下降导致失稳的作用强于水位上涨;降雨为主要因素,其中前 1 个月降雨量和单日最大降雨量影响较大。

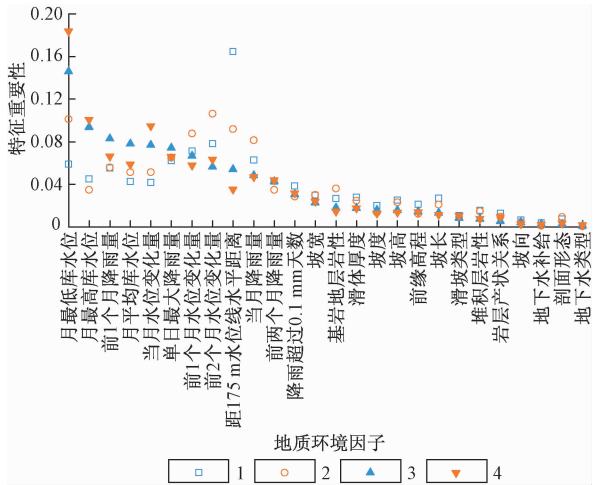


图 5 模型因子重要性对比

Fig.5 Factor importance of models

1—RF;2—GBDT;3—Borsmote1-RF(0.5 比率);4—Borsmote1-GBDT(0.7 比率)。

图 6 为 borsmote1(0.5 比率)情况下,主要因子所对应的平衡化突变类、稳定类和初始突变类三者的分布特征。经过平衡化的数据集降低了树节点划分突变的不纯度值,且各因子的分布特征均能够得到较好的可解释性。如图 6(a)、图 6(b)、图 6(d)和图 6(e)所

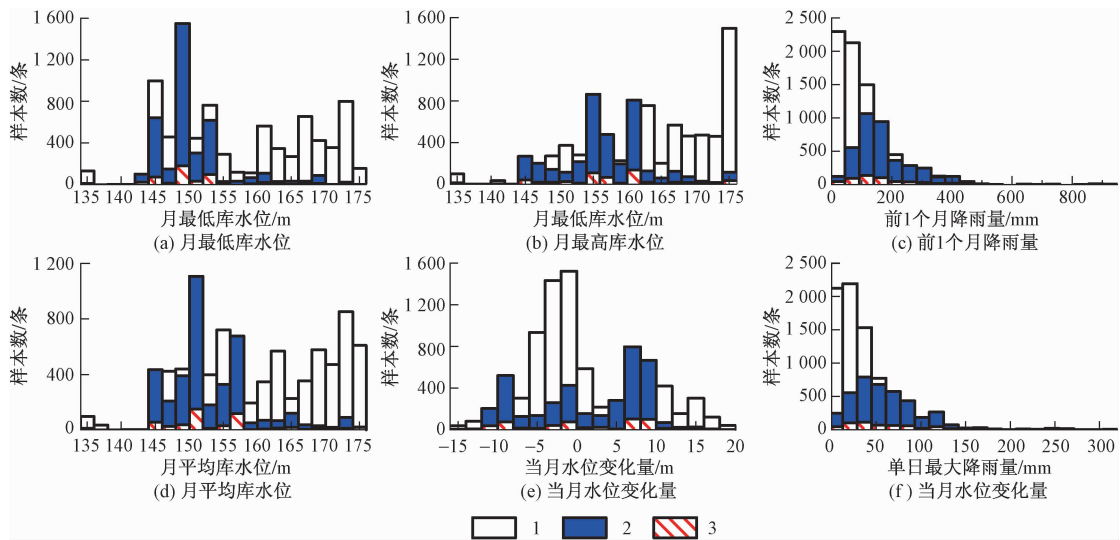


图 6 主要因子对应的稳定类、平衡化突变类和初始突变类三者的分布特征

Fig.6 Distribution characteristics of balanced mutants, stable classes and initial mutations corresponding to major factors

1—稳定类;2—平衡化突变类;3—初始突变类。

示,突变集中发生在月最低库水位[144 m,154 m]、月最高库水位[153 m,162 m]和月平均库水位[144 m,158 m]段,当月水位变化量则多发于[-12 m,-8 m]和[4 m,10 m]两个区间段(突变发生的[-4 m,0 m]区段是稳定类的集中区段,因此突变频率相对较低),表明库水位的骤升和骤降均是诱发“阶跃式”滑坡失稳的重要因素;如图6(c)、图6(d)所示,单日最大降雨量为40 mm和前1个月降雨量为125 mm左右,是“阶跃式”滑坡突变启动的一个转折点,且均与突变呈正相关。

3.3 王爷庙滑坡案例分析

王爷庙滑坡地处重庆市涪陵区石和乡长江右岸,平面呈长舌状,微地貌呈缓坡,地形总体为西高东低;平均长650 m,宽150 m,标高140~288 m,坡向85°,坡角约20°。经现场勘查,滑体物质为第四系冲洪积层(Q_4^{al+pl})及残坡积层(Q_4^{el+dl})的粉砂土和粉质黏土夹碎(块)石,厚5~15 m;下伏基岩以侏罗纪中下统自流井组($J_{1-2}zl$)泥岩为主。滑坡区位于丰盛场背斜东翼,产状 $155^\circ \angle 27^\circ$,为大角度相交的斜向斜坡,裂隙较发育。该滑坡属新生型的中型浅层牵引式土层滑坡。

历年来,滑坡受雨季降水(图7)和库水波动(图8)影响,滑体沿基岩接触面向下滑动,前缘经江水浸泡、冲刷和淘蚀而失去支撑,变形加剧,处于欠稳定状态。其中中部WYM05号监测点累计地表水平位移达5 681.6 mm,该点自2007年雨季开始变形,时空曲线为“阶跃式”发展,历经多个突变时段(主要为2007年5~8月,2008年9月,2009年2月、5~11月,2010年4~6月,2011年5月,2012年5~7月,2013年6月,2014年5~12月,2015年6~7月,2016年4~10月),位移方向 $81^\circ \sim 85^\circ$ 同主滑方向基本一致。

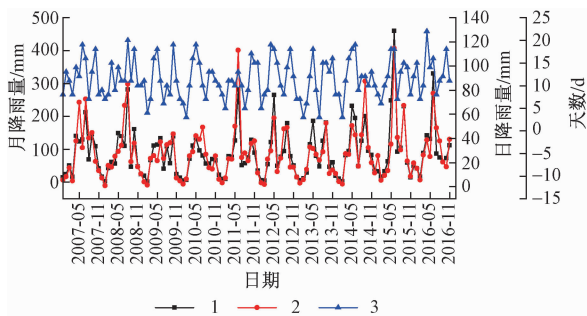


图7 王爷庙滑坡降雨监测曲线

Fig.7 Monitoring curves for precipitation of Wangyemiao landslide

1—当月降雨量;2—单日最大降雨量;3—降雨超过0.1 mm天数。

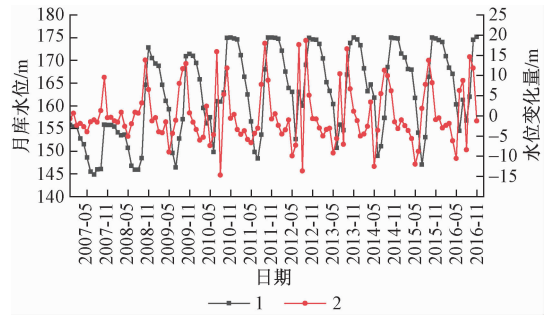


图8 王爷庙滑坡库水监测曲线

Fig.8 Monitoring curves for reservoir water level of Wangyemiao landslide

1—月平均库水位;2—当月水位变化量。

经平衡集成树的 Borsmote1-RF(0.5 比率)和 Borsmote1-GBDT(0.7 比率)模型学习计算,预测出的突变时刻与位移监测曲线相吻合, f_1 值分别为80.88%和78.75%,图9为平衡集成树模型预测结果。

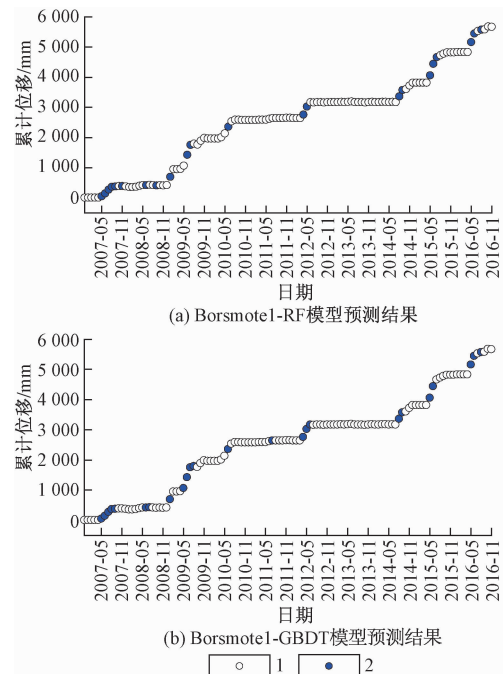


图9 WYM05号监测点累计位移曲线和预测结果

Fig.9 Accumulated displacement curves and prediction results of monitoring point WYM05

1—预测为稳定类;2—预测为突变类。

4 结论和展望

本文提出了平衡集成树模型对“阶跃式”滑坡突变预测和特征提取进行了研究分析,得出如下结论:

(1)模型考虑了滑坡监测数据处理中的不平衡问

题,通过合成边界突变类,使模型能够有效获取“阶跃式”滑坡突变与复杂地质环境因子间的关联。与传统算法相比较,该算法实现了海量监测数据下的有效泛化和准确分析,解决了传统突变预测方法效率低的问题;

(2)该方法能够较好反映研究区内不同时段不同条件下因子对“阶跃式”滑坡失稳的影响作用。本文探讨了以月为周期的预测分析,并进行因子特征重要性比较,研究方法及所得规律可为相关工程和研究提供重要的参考价值;

(3)通过测试集和预测集结果评价,模型性能达到了实际预报要求,验证了平衡集成树模型应用于“阶跃式”滑坡风险评估与机理研究的可行性。同时,处理和解决滑坡灾害数据分析中的数据不平衡问题,是实现智能滑坡灾害分析与大数据预警的必然趋势。

参考文献:

- [1] 许强,汤明高,徐开祥,等. 滑坡时空演化规律及预警预报研究[J]. 岩石力学与工程学报,2008,27(6):1104-1112.
XU Qiang, TANG Minggao, XU Kaixiang, et al. Research on space time evolution laws and early warning prediction of landslides[J]. Chinese Journal of Rock Mechanics and Engineering, 2008, 27(6): 1104-1112.
- [2] 易庆林,张明玉,文凯,等. 三峡库区白水河滑坡变形特征及影响因素的阶段分析[J]. 三峡大学学报(自然科学版),2017,39(1):38-42.
YI Qinglin, ZHANG Mingyu, WEN Kai, et al. Periodic analysis of deformation characteristics and influential factors of Baishuihe Landslide in Three Gorges Reservoir area[J]. Journal of China Three Gorges University(Natural Sciences), 2017, 39(1): 38-42.
- [3] 黄达,顾东明,陈智强,等. 三峡库区塔坪 H2 古滑坡台阶状复活变形的库水-降雨耦合作用机制[J]. 岩土工程学报,2017,39(12):2203-2211.
HUANG Da, GU Dongming, CHEN Zhiqiang, et al. Hybrid effects of rainfall and reservoir level fluctuation on old Taping H2 landslide in Wushan County in Three Gorges Reservoir area[J]. Chinese Journal of Geotechnical Engineering, 2017, 39(12): 2203-2211.
- [4] 张振华,钱明明,位伟. 基于改进破坏接近度的千江坪岸坡失稳机制分析[J]. 岩石力学与工程学报,2018,37(6):1371-1384.

- ZHANG Zhenhua, QIAN Mingming, WEI Wei. Analysis of failure mechanism of Qianjiangping slope based on improved failure approach index[J]. Chinese Journal of Rock Mechanics and Engineering, 2018, 37(6): 1371-1384.
- [5] 陈卫兵,郑颖人,冯夏庭,等. 考虑岩土体流变特性的强度折减法研究[J]. 岩土力学,2008,29(1):101-105.
CHEN Weibing, ZHENG Yingren, FENG Xiating, et al. Study on strength reduction technique considering rheological property of rock and soil medium[J]. Rock and Soil Mechanics, 2008, 29(1): 101-105.
- [6] 赵久彬,刘元雪,宋林波,等. 大数据关键技术在滑坡监测预警系统中的应用[J]. 重庆理工大学学报(自然科学),2018,32(2):182-190.
ZHAO Jiubin, LIU Yuanxue, SONG Linbo, et al. Application of big data processing and monitoring technologies in landslide monitoring and warning system[J]. Journal of Chongqing University of Technology(Natural Science), 2018, 32(2): 182-190.
- [7] 彭令,牛瑞卿,吴婷. 时间序列分析于支持向量机的滑坡位移预测[J]. 浙江大学学报(工学版),2013,47(9):1672-1679.
PENG Ling, NIU Ruiqing, WU Ting. Time series analysis and support vector machine for landslide displacement prediction[J]. Journal of Zhejiang University(Engineering Science), 2013, 47(9): 1672-1679.
- [8] MIAO F S, WU Y P, XIE Y H, et al. Prediction of landslide displacement with step like behavior based on multialgorithm optimization and a support vector regression model[J]. Landslides, 2017, 15(3): 475-488.
- [9] 李骅锦,许强,何雨森,等. “阶跃式”滑坡位移预测及阈值分析的 ARMA-(LASSO-ELM)-Copula 模型[J]. 岩石力学与工程学报. 2017,36(增刊2):4075-4084.
LI Huajin, XU Qiang, HE Yusen, et al. An ARMA-(LASSO-ELM)-Copula framework for landslide displacement prediction and threshold computing of the displacement of step-like landslides[J]. Chinese Journal of Rock Mechanics and Engineering, 2017, 36(S2): 4075-4084.
- [10] 杨背背,殷坤龙,杜娟. 基于时间序列与长短期记忆网络的滑坡位移动态预测模型[J]. 岩石力学与工程学报,2018,37(10):2334-2343.

- YANG Beibei, YIN Kunlong, DU Juan. A model for predicting landslide displacement based on time series and long and short term memory neural network[J]. Chinese Journal of Rock Mechanics and Engineering, 2018, 37(10): 2334 – 2343.
- [11] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321 – 357.
- [12] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [J]. Advances in Intelligent Computing, 2005, 3644(1): 878 – 887.
- [13] 林威伟, 钟登华, 胡炜, 等. 基于随机森林算法的土石坝压实质量动态评价研究[J]. 水利学报, 2018, 49(8): 945 – 955.
- LIN Weiwei, ZHONG Denghua, HU Wei, et al. Study on dynamic evaluation of compaction quality of earth rock dam based on Random Forest [J]. Journal of Hydraulic Engineering, 2018, 49(8): 945 – 955.
- [14] ZHANG C S, LIU C C, ZHANG X L, et al. An up-to-date comparison of state-of-the-art classification algorithms [J]. Expert Systems with Applications, 2017, 82: 128 – 150.
- [15] POWERS D M W. Evaluation: from Precision, Recall and F-Factor to ROC, informedness, markedness & correlation [J]. Journal of Machine Learning Technologies, 2011, 2(1): 37 – 63.
- [16] FRIEDMAN JEROME H. Stochastic gradient boosting [J]. Computational Statistics & Data Analysis, 2002, 38(4): 367 – 378.
- [17] 乔建平. 长江三峡库区蓄水后滑坡危险性预测研究——以重庆市万州区库岸为例[M]. 北京: 科学出版社, 2012.
- QIAO Jianping. Study on the risk prediction of landslide after impoundment in the Three Gorges reservoir area of the Yangtze River: taking the bank of Wanzhou District in Chongqing as an example [M]. Beijing: Science Press, 2012.
- [18] 殷跃平. 三峡库区边坡结构及失稳模式研究[J]. 工程地质学报, 2005, 13(2): 145 – 154.
- YIN Yueping. Human-cutting slope structure and failure pattern at the Three Gorges Reservoir [J]. Journal of Engineering Geology, 2005, 13(2): 145 – 154.
- [19] 乔素云, 马云彪, 尹思杰, 等. 基于岩土参数变异特性的三峡库区白水河滑坡失稳概率分析[J]. 中国地质灾害与防治学报, 2018, 29(5): 38 – 43.
- QIAO Suyun, MA Yunbiao, YIN Sijie, et al. Probability analysis of Baishuihe landslide in the Three Gorges Reservoir area based on variation characteristics of geotechnical parameters [J]. The Chinese Journal of Geological Hazard and Control, 2018, 29(5): 38 – 43.
- [20] 简文星, 许强, 童龙云. 三峡库区黄土坡滑坡降雨入渗模型研究[J]. 岩土力学, 2013, 34(12): 3527 – 3533.
- JIAN Wenxing, XU Qiang, TONG Longyun. Rainfall infiltration model of Huangtupo landslide in Three Gorges Reservoir area [J]. Rock and Soil Mechanics, 2013, 34(12): 3527 – 3533.
- [21] 郑颖人, 唐晓松. 库水作用下的边(滑)坡稳定性分析[J]. 岩土工程学报, 2007, 29(8): 1115 – 1121.
- ZHENG Yingren, TANG Xiaosong. Stability analysis of slopes under drawdown condition of reservoirs [J]. Chinese Journal of Geotechnical Engineering, 2007, 29(8): 1115 – 1121.
- [22] 黄发明, 殷坤龙, 蒋水华, 等. 基于聚类分析和支持向量机的滑坡易发性评价[J]. 岩石力学与工程学报, 2018, 37(1): 156 – 167.
- HUANG Faming, YIN Kunlong, JIANG Shuihua, et al. Landslide susceptibility assessment based on clustering analysis and support vector machine [J]. Chinese Journal of Rock Mechanics and Engineering, 2018, 37(1): 156 – 167.
- [23] MOORE D S, MCCABE G P, CRAIG B A. Introduction to the practice of statistics [M]. New York: W. H. Freeman and Company, 2012.