

基于优化 Logistic 分类算法的山体滑坡研究

刁君华^{1,2}, 冯向萍^{1†}, 马新春²

(1. 新疆农业大学计算机与信息工程学院, 新疆 乌鲁木齐 830052; 2. 新疆电子研究所股份有限公司, 新疆 乌鲁木齐 830013)

摘要:滑坡是一种常见的地质灾害,受到滑坡曲率、坡度等因素的影响。然而这些因素一般是以连续型数据存在,传统的机器学习分类算法无法适应此种数据类型,一般采用人工处理方式将其转换为离散型数据,这种方式不但耗费人力,而且无法准确划分影响因素离散化后的区间范围,产生噪声。针对于此实验提出了一种集成机器学习模型,以 Logistic 算法为分类器,以随机森林算法为连续性数据处理工具,减少人工处理方式产生的噪声影响。模型在滑坡灾害数据集上得以验证,F1 值达到了 90.1%。

关键词:机器学习;人工处理;噪声;Logistic;随机森林

中图分类号:TP181

文献标识码:A

Research on Landslides Based on Optimized Logistic Classification Algorithm

DIAO Junhua^{1,2}, FENG Xiangping^{1†}, MA Xinchun²

(1. School of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi, Xinjiang 830052, China;

2. Xinjiang Electronics Research Institute Co., Ltd., Urumqi, Xinjiang 830013, China)

Abstract: Landslide is a common geological disaster, which is affected by factors such as landslide curvature and slope. However, these factors generally exist in the form of continuous data. Traditional machine learning classification algorithms cannot adapt to this type of data. Manual processing is generally used to convert it into discrete data. This method is not only labor-intensive, but also cannot accurately classify the influencing factors. The discretized interval range generates noise. For this experiment, an integrated machine learning model was proposed, using the Logistic algorithm as the classifier and the random forest algorithm as the continuous data processing tool to reduce the noise impact caused by manual processing. The model was verified on the landslide disaster data set, and the F1 value reached 90.1%.

Key words: machine learning; manual processing; noise; Logistic; random forest

近几年,各个地区滑坡、泥石流、突发性洪灾等事件频频发生。为了确保地质灾害场区人民的生命财产安全、正常生产和安全出行,对地质灾害进行准确的预测,从而减少地质灾害造成的危害,是一项紧迫而艰巨的任务^[1-2]。其中,滑坡是全球范围主要的自然地质灾害之一,通常发生在陡坡或过饱和的土地上,具有突发性和破坏性。因此,预测

滑坡灾害的发生是十分有必要的。

我国不少研究人员通过研究地质构造和地表地形特征,来确定滑坡的潜在危险区域,利用地理信息系统(GIS)^[3]、遥感技术^[4]和激光扫描^[5]等方法,对地形进行高精度测量和分析,揭示潜在滑坡的迹象和特征。需要注意的是,滑坡预测仍然是一个具有挑战性的任务。

收稿日期:2024-03-18

基金项目:新疆维吾尔自治区科技援疆项目计划(2022E02080)

作者简介:刁君华(2000—),女,山东菏泽人,硕士,研究方向:机器学习,深度学习。

†通信联系人,E-mail: 532912349@qq.com

求解滑坡概率的主要方法有 Logistic Regression(LR)^[6]、物理力学模型^[7]、概率密度法^[8]等。其中,以上方法在求解滑坡概率的问题上都存在一定的局限性。物理力学模型通常对初始条件和参数值敏感,如果参数值不准确或初始条件不清楚,很难稳定可靠地预测滑坡概率。概率密度法在滑坡概率估计方面由于模型参数的不确定性和数据的不确定性等,使得在处理这些不确定性方面相对局限。而对于 LR,擅长处理离散特征值,对于包含连续特征值和离散特征值的滑坡灾害数据来说,LR 在预测滑坡概率方面同样存在一定的局限性。因此,本实验在选取 LR 模型进行预测的基础上,使用 Bagging 集成学习算法和 Boosting 集成学习算法来处理数据,经过实验对比随机森林可以更好地解决 LR 在滑坡预测上的局限性。

1 滑坡预测相关理论

1.1 基于 Boosting 的 XGBoost 集成算法

XGBoost(eXtreme Gradient Boosting)算法是一种使用梯度提升算法的集成算法,基于树模型进行预测和建模。通过迭代构建多个弱分类器(决策树),并将这些弱分类器组合成一个强分类器。与传统的梯度提升树算法相比,XGBoost 引入了更多的优化和正则化技术,使得模型具有更高的性能和泛化能力。

给定数据集 D ,具有 n 个样本和 m 个特征,其中 $D \in \{(x_i, y_i) \mid x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$,则该集成学习的预测模型可表示为:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (1)$$

其中, $\hat{y}_i$ 为对样本 x_i 的预测输出值, k 表示该模型中决策树的总数, f_k 表示第 k 个决策树的分类函数, F 是由 k 个子模型组成的模型空间。

XGBoost 的目标函数是由两部分组成的:损失函数和正则化项^[9]。其中损失函数衡量了模型预测值与真实值之间的差异或损失,正则化项则用于控制模型的复杂度,防止过拟合。XGBoost 的目标函数如式(2):

$$\tau^{(t)} = \sum_{i=1}^n L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (2)$$

其中, $L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$ 是损失函数,表示样本 y_i 的真实值与当前模型的预测输出值之间的损失, $\hat{y}_i^{(t-1)}$ 表示第 $(t-1)$ 轮模型的预测输出

值, $f_t(x_i)$ 表示当前轮模型对样本 x_i 的预测输出值, $\Omega(f_t)$ 是正则化部分,用于控制模型的复杂度,防止过拟合。 $\Omega(f_t)$ 的表达式为:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (3)$$

其中, γ 和 λ 是权重系数, T 为树的叶子结点数量, ω 为叶子节点的权重。对式(10)进行二阶泰勒展开可得:

$$\tau^{(t)} = \sum_{i=1}^n [L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} \hat{y}_i f_t^2(x_i)] + \Omega(f_t) \quad (4)$$

其中, g_i 表示损失函数对预测输出值的一阶导数, $\hat{y}_i$ 表示损失函数对预测输出值的二阶导数。

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (5)$$

$$\hat{y}_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)}) \quad (6)$$

1.2 基于 Bagging 的随机森林集成算法

随机森林(Random Forest, RF)是集成学习算法 Bagging 中的一种,随机森林包含很多的决策树分类器,这些决策树可以用于解决分类问题,在构建每棵决策树时,随机森林采用自助采样的方法从训练数据集中有放回的随机抽取样本,在每次决策树的结点划分上,随机森林会从所有特征中选取一个子集进行划分,有助于减少特定特征对模型的影响,增加了模型的多样性,从而提高整体模型的泛化能力和鲁棒性。随机森林生成过程如图 1 所示。

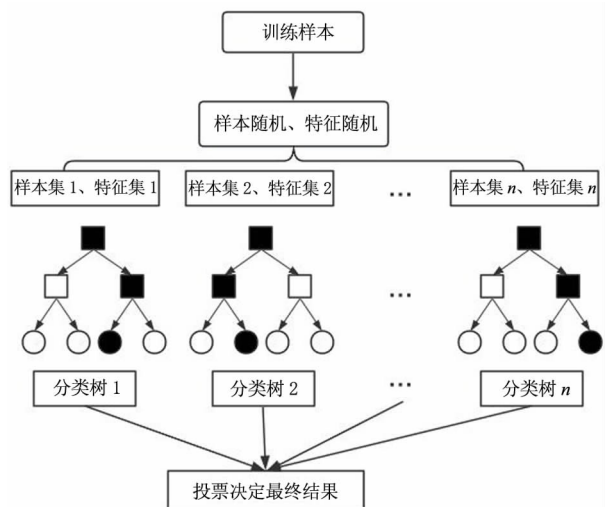


图 1 随机森林生成过程

从训练样本中随机选择一个子集,同时随机选择该子集中的特征属性。将随机选取的子集和特征属性划分为 n 个样本集和特征集,放入建立的 n 个决策树模型中进行训练,然后根据决策树的预测

结果采用投票的方式得到最终预测结果。

1.3 Logistic 回归算法

Logistic(逻辑)回归算法,是一种常用于预测二元或解决分类问题的机器学习算法。它基于线性回归的原理,使用 Sigmoid 函数将线性模型的输出转换为概率,并根据概率进行分类预测。Logistic 回归算法的核心部分是 Sigmoid 函数,如下:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

(7)

其中, $z = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n$, x_i 为数据集的第 i 个特征。逻辑回归模型利用 Sigmoid 函数,将线性模型的预测结果映射到 $[0,1]$ 的范围内,表示样本属于某一类别的概率。逻辑回归函数可表述为:

$$y = \frac{1}{1 + e^{-(w^T x + b)}}$$

(8)

对公式(8)左右分别取对数可得:

$$\ln \frac{y}{1 - y} = w^T x + b$$

(9)

在给定样本 x 的条件下,将样本类别为 1 的概率视为类后验概率。 $P(y = 1 \mid x) = \frac{1}{1 + e^{-z}}$,而 $P(y = 0 \mid x) = \frac{e^{-z}}{1 + e^{-z}}$,因此 $P(y = 1 \mid x) + P(y = 0 \mid x) = 1$,则公式(5)可写为:

$$\ln \frac{P(Y = 1 \mid X)}{P(y = 0 \mid x)} = w^T x + b$$

(10)

设训练数据集为 $Z = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$,其中 $x_i \in R^n, y_i \in (0, 1)$ 用极大似然估计法估计模型参数 w 。设 $P(y = 1 \mid x) = f(x)$, $P(y = 0 \mid x) = 1$,已知似然函数为 $\prod_{i=1}^n [f(x_i)^{y_i} [1 - f(x_i)]^{1-y_i}]$ 。对该似然函数取对数可得:

$$L(w) = \sum_{i=1}^n [y_i \ln f(x_i) + (1 - y_i) \cdot \ln(1 - f(x_i))]$$

(11)

对公式(11)中的 w 求偏导,令 $L(w) = 0$,进而解出 w :

$$\frac{\partial(L(w))}{\partial(w)} = \sum_{i=1}^n (y_i - \frac{e^{w \cdot x_i}}{1 + e^{w \cdot x_i}})$$

(12)

2 实验

2.1 实验设置

本实验采用四川省某地地质灾害监测点的监测数据作为模型数据集,由于大多地质灾害的数据

存在连续型数据和离散型数据。对于 LR 来说,在大多数情况下,直接将连续特征值输入到逻辑回归模型,会使模型的表达能力受限,减缓运算速度,且若存在异常数据,会给模型带来很大干扰,同时,会在本实验数据的应用上存在一定的局限性。因此,在使用 LR 模型进行滑坡预测时,本实验将采用两种常见的 Bagging 中的随机森林集成学习算法和 Boosting 中的 XGBoost 集成学习算法以及人工对样本数据进行离散化处理的三种数据处理方式来提高 LR 在滑坡预测上的准确率。

实验将使用混淆矩阵计算出的精确率(Precision)、召回率(Recall)、F1 值、真负率(TNR)、假正率(FPR)作为评估标准。具体计算公式及混淆矩阵如表 1、表 2 所示。

表 1 模型评价指标定义

评价指标	计算式	
精确率 Precision	$Precision = \frac{T_p}{T_p + F_p}$	(13)
召回率 Recall	$Recall = \frac{T_p}{T_p + F_N}$	(14)
F1 得分 F1 score	$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall}$	(15)
真负率 TNR	$TNR = \frac{T_N}{T_N + F_p}$	(16)
假正率 FPR	$FPR = \frac{F_p}{T_N + F_p}$	(17)

表 2 混淆矩阵

混淆矩阵	预测 0 (没有发生滑坡)	预测 1 (发生滑坡)
真实 0(没有发生滑坡)	T_N	F_p
真实 1(发生滑坡)	F_N	T_p

表 1 中,精确率表示为预测正确的正例数据 T_p 占预测为正例数据 $(T_p + F_p)$ 的比例,召回率表示为预测为正确的正例数据 T_p 占实际为正例数据 $(T_p + F_N)$ 的比例。真负率表示为预测正确的负例数据 T_N 占实际为负例数据 $(T_N + F_p)$ 的比例,假正率则表示为预测错误的负例数据 F_p 占实际为负例数据 $(T_N + F_p)$ 的比例。

表 2 中, T_N 表示为在真实数据中没有发生滑坡,在模型预测时同样也是没有发生滑坡的情况,即模型正确的预测负类别样本的数量。 F_p 表示为在真实数据中没有发生滑坡的情况下,在模型预测时预测为发生滑坡,即模型错误地将负类别样本预测为正类别的数量。 F_N 表示为在真实数据中发生

了滑坡,但预测模型预测为没有发生滑坡,即模型错误地将正类别样本预测为负类别的数量。 T_p 表示为在实际数据下发生了滑坡,在模型预测下同样预测为发生滑坡的情况,即模型正确地预测正类别样本的数量。

2.2 数据离散化处理

本实验将滑坡地质灾害数据中的连续型数据经过算法处理和人工处理两种方式进行离散化。算法处理是使用 XGBoost 和随机森林两种算法进行处理,然后将这两类方式处理后的离散化数据分别输入到 Logistic 模型中进行训练,最终得到预测结果。

2.2.1 人工处理

通过查阅周毅等^[10]、兰剑^[11]对滑坡灾害的影响因素的研究,对实验样本数据中的连续型特征进行

离散化处理。对浅层滑坡的安全系数和岩石的类型这两个特征进行 one-hot 独热编码,将浅层滑坡的安全系数分为 5 类,使其值在{0,1}内选取,且当数据特征为五类中的某一个类别时,值为 1,其余值都为 0。将岩石的类型分为 7 类,分别为 Limestone(石灰岩),Marl(泥灰岩),Evaporties(蒸发岩),Rockfall debris(岩崩碎屑),Moraine on Marls(泥灰岩上的冰碛),Moraine Thick(厚冰碛),Moraine on Evaporites(蒸发岩上的冰碛),使其值在{0,1}内选取,例如,当前特征为石灰岩时,值为 1,其余值全为 0。对其他灾害特征进行阈值的设定和区间的划分来进行离散化处理。以到溪流的距离为例,由于到溪流距离这一属性是连续的,因此,实验将这一特征分为六个分组。具体特征划分如表 3 所示。

表 3 特征指标表

特征	离散处理
strdist(到溪流的距离)	0—100 0,100—200 1,200—300 2,300—400 3,400—500 4,500—600 5,600 以上 6
aspect(坡度)	0—30 0,30—50 1,50—100 2,100—150 3,150—200 4,200 以上 5
basarea(集水区尺寸)	1200—5000 0,5000—10000 1,10000—50000 2,50000—100000 3,100000—500000 4,500000—1000000 5,1000000 以上 6
basin(集水)	1—10 0,10—20 1,20—30 2,30—40 3,40 以上为 4
curvature(坡度曲率)	—11— —8 0,—8— —4 1,—4— —2 2,—2—2 3,2—6 4,6 以上 5
curve_cont(等高线平行坡度曲率)	—13— —9 0,—9— —6 1,—6— —4 2,—4— —2 3,—2 —2 4,2—6 5,6 以上 6
curve_prof(轮廓平行坡度曲率)	—8— —4 0,—4— —2 1,—2—2 2,2—4 3,4—6 4,6 以上 5
curves(坡度引起的高度变化)	0—0.2 0,0.2—0.4 1,0.4—0.62,0.6 以上为 3
drop(谷底以上高度的绝对变化)	1—50 0,50—100 1,100—150 2,15—200 3,200 以上为 4
flowdir(流动的方向)	1—20 0,20—40 1,40—60 2,60—80 3,80—100 4,100 以上为 5
elev(海拔)	528—600 0,600—700 1,700—800 2,800—900 3,900—1000 4,1000—1100 5,1100—1200 6,1200—1300 7,1300 以上 8
cohesion(凝聚)	1—10 0,10—20 1,20 以上为 3
scarpdist(距失效陡坡的距离)	0—50 0,50—100 1,100—150 2,150—200 3,200—250 4,250—300 5,300—350 7,350 以上 8
frictang(摩擦角)	1—18 0,18—26 1,26 以上为 2
slope(坡度梯度)	0—10 0,10—20 1,20—30 2,30—40 3,40—50 4,50—60 5,60 以上为 6
slopeleg(斜坡长度)	43—200 0,200—400 1,400—600 2,600—800 3,800—1000 4,1000—1200 5,1200—1400 6,1400—1800 7,1800—2000 8,2000 以上为 9
specwt(特定的重量)	15—20 0,20—23 1,23 以上为 2

2.2.2 随机森林和 XGBoost 处理

本实验采用随机森林和 XGBoost 两个集成算法处理连续的特征值,由于这两种集成学习算法擅长处理连续的特征值,而 LR 擅长处理离散的特征值。因此,本实验将选用这两种算法对数据进行离散化处理,具体过程图 2 所示。

由图 2 可知,本实验将滑坡地质灾害数据进行

随机划分,其中训练集和测试集划分比例为 4 : 1,之后将训练集和测试集中的离散特征提取出来,命名为 Categorical_Features。再将训练集和测试集中的离散特征剔除,将剔除离散特征后的训练数据集和测试数据集命名为 Continuous_Features 和 Continuous_Test_Features,将 Continuous_Features 输入到随机森林或 XGBoost 中进行训练。

训练好模型后,得到 k 棵树,每棵树上有不同数量的叶子结点,分别记为 $n_1, n_2, n_3, \dots, n_k$,提取每个样本在每棵树的叶子结点的索引,然后将叶子节点的索引转换为稀疏特征,其中,每个样本有 k 个值为 1,其余取值为 0。最后,合并稀疏特征和 categorical_features,将合并后的特征放入 LR 中进行二次训练,最终得到预测结果。

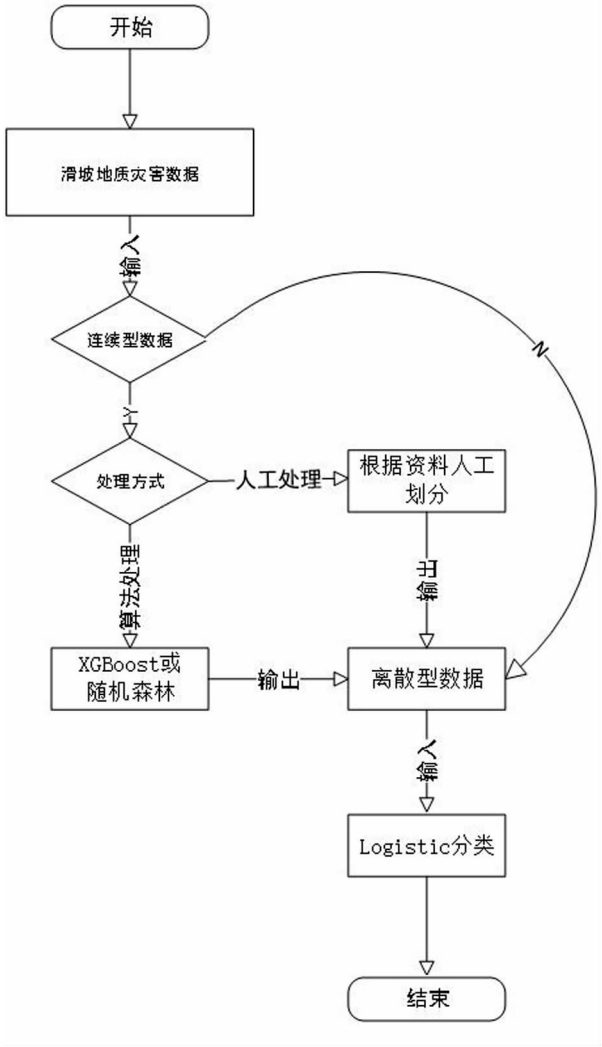


图 2 数据处理过程图

2.3 实验结果

经过以上三种方式对样本数据进行离散化处理,将处理后的离散化数据分别输入到 LR 中进行训练,最终模型的预测结果对比表和预测结果图如表 4 和图 3 所示。

表 4 中显示了经过不同的三种数据处理方式,得到的预测结果的评估指标。从表中可见,使用随机森林处理的数据得到的精确率、召回率、F1 得分较为显著,真负率较高,表示对滑坡预测的判别效果较好。

对于 XGBoost,其在训练数据上更容易过拟

合,特别是在训练数据相对较小的情况下。与此相比,随机森林在某种程度上对过拟合有更好的控制,因为它使用了多个树的投票,有助于降低模型的方差。因此基于 XGBoost 的 Logistic 算法各项指标均劣于基于随机森林的 Logistic 算法。

而人工处理的数据,由于工作人员依靠经验对滑坡影响因素区间的划分具有主观意识,很容易产生噪声,所以最终得到的预测精确率、召回率和 F1 得分显著低于两个集成算法所处理数据的结果,假正率高于两个集成算法。经实验证明,经过集成学习算法处理后的数据,最终的预测结果精度要远高于人工进行离散化处理的结果。

表 4 模型预测结果对比表

模型类型	精确率	召回率	F1 得分	真负率	假正率
XGB-Log	0.858	0.864	0.861	0.922	0.078
RF-Log	0.894	0.907	0.901	0.942	0.058
AI-Log	0.784	0.715	0.748	0.892	0.108

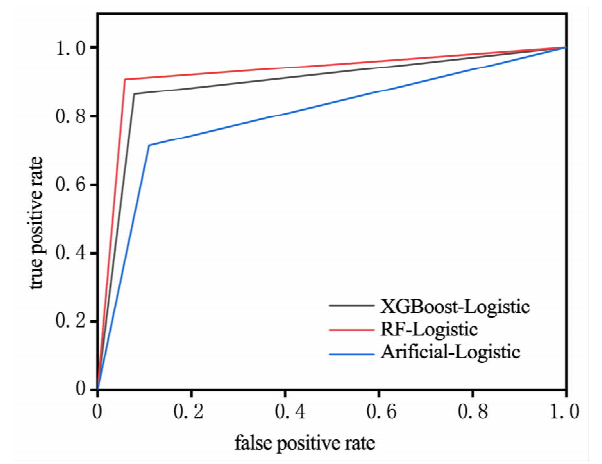


图 3 ROC 曲线

由图 3 可知,横坐标表示为假正率,即预测为正的负样本数与实际负样本数之比。纵坐标表示为真正率,即模型预测为正的样本数占有所有实际正样本数的比例,也可理解为召回率。当 ROC 曲线越靠近左上角时,表示模型的性能越好。因为靠近左上角意味着模型能够在较低的假正率下实现较高的真负率,即模型能够较好地将正样本预测为正样本,并排除大部分负样本。通过观察上述 ROC 曲线图,可以判断经过随机森林处理后的数据,放入 Logistic 模型中预测的性能,与人工处理的数据和随机森林处理的数据相比,更加接近左上角,性能更好。

(下转第 63 页)