露天煤矿边坡稳定性的随机森林预测模型

温廷新,张波

辽宁工程技术大学系统工程研究所,葫芦岛 125105

摘要　边坡工程是露天煤矿中的重点工程,边坡的稳定性关系着煤矿的安全生产。边坡稳定性预测是边坡防治工作的前提,针对煤矿边坡工程稳定性预测的复杂性,为了快速、有效地判别煤矿边坡稳定性,利用随机森林算法建立煤矿边坡稳定性预测模型。通过选取与煤矿边坡工程密切相关的岩石重度、黏聚力、内摩擦角、边坡角、边坡高度、孔隙水压力6个指标作为边坡稳定性的影响因素,即为随机森林预测模型的输入,边坡稳定性状态作为随机森林预测模型的输出,通过随机森林算法建立边坡稳定性影响因素与边坡稳定状态之间的非线性关系。利用煤矿实测30组边坡稳定性数据作为随机森林预测模型的训练数据集,进行模型的学习训练;另用12组边坡稳定性数据作为预测模型的测试数据,通过训练好的边坡稳定性预测模型进行测试;为了验证随机森林预测模型的准确率,同时与SVM和BP神经网络的测试数据进行比较。结果说明,选取煤矿边坡稳定性的6个指标建立的随机森林预测模型,人工控制参数较少、结构简单、容易实现,且具有较高的准确度,边坡稳定状态预测结果与煤矿边坡工程实际状态相吻合,能有效预测边坡稳定性状态,指导煤矿边坡防治工作的开展。

关键词　随机森林;CART;边坡稳定性;露天煤矿

中图分类号　TP306.1　　**文献标志码**　A　　**doi**　10.3981/j.issn.1000-7857.2014.h1.018

Prediction Model for Open-pit Coal Mine Slope Stability Based on Random Forest

WEN Tingxin, ZHANG Bo

System Engineering Institute, Liaoning Technological University, Huludao 125105, China

Abstract　Slope engineering is a key project in open-pit coal mines. The stability of the slope is closely related to safety production of coal mines. Slope stability prediction is a prerequisite in slope control, faced with complexities. To quickly and effectively determine the coal mine slope stability, this paper establishes a prediction model using the random forest algorithm. Six factors influencing the slop stability were selected as input of the prediction model, including the gravity density of rocks, cohesive force, internal friction angle, slope angle, slope height and pore water pressure, and slope stability status was selected as output of the prediction model. The random forest algorithm was used to establish the nonlinear relationship between slope stability factors and stability status. The 30 sets of measured data were used as training data set to learn and train the random forest slope stability prediction model. In addition, 12 groups of data as slope stability test data were used to test the trained prediction models. In the meantime, the accuracy of the random forest prediction models was tested by comparing them with the SVM and BP neural network prediction models. The results show that the random forest prediction model based on the selected six factors has less manual control parameters, simple structure and high accuracy. The predictive results coincide with the actual state of the slope project, indicating that the prediction model is able to

收稿日期:2013-10-25;修回日期:2013-11-18

基金项目:辽宁省突发事件应急管理多元化IS体系设计项目(LT2010048);山东省突发事件多元应急信息系统研究与构建项目(ZR2010FL012);校企合作基金项目(SCDY2012018)

作者简介:温廷新,副教授,研究方向为数据挖掘和知识管理,电子信箱:wen_tx@163.com

引用格式:温廷新, 张波. 露天煤矿边坡稳定性的随机森林预测模型[J]. 科技导报, 2014, 32(4/5): 105-109.

predict the slope stability effectively and provide guidance to coal mine slope prevention work.

Keywords random forest; CART; slope stability; open-pit coal mine

随着社会经济的发展,边坡工程的种类越来越多,边坡的高度也越来越高。边坡失稳不仅会造成严重的经济损失,而且可能危及到边坡周边人民的生命财产安全。例如,意大利北部瓦依昂水库1963年10月9日发生的灾难性顺层滑坡,滑坡体积达2亿4000万 m^3,使大约2600人丧生,造成了巨大的财产损失。而在各类边坡工程中,露天煤矿边坡工程的边坡高度几乎都超过百米,有些甚至高达几百米。边坡不稳定,即使一个小小的滑坡,也可能严重阻碍煤矿开采工作,造成直接或间接的经济损失。滑坡已成为同地震和火山并列的全球性三大地质灾害之一[1,2]。边坡稳定性是煤矿工程效益分析的重点之一,关系着工程建设的成败,是矿山工程安全的根本保障[3]。因此,露天煤矿边坡稳定性预测分析是当前矿山工程关注的热点,构建快速有效的边坡预测分析模型对煤矿边坡防治工作具有重大意义。

边坡稳定性预测分析主要有确定性和非确定性两种方法。确定性方法主要有极限平衡分析法[4]、数值方法[5]等,此类方法在对边坡稳定性进行分析时,将影响因素大量简化,导致理论结果与实际相差甚远。非确定性方法出现了一些智能算法,如程纬华等[6]将BP神经网络(BPNN)运用于露天矿边坡稳定性预测中。但是人工神经网络为了保证预测的准确率,需要大量测试数据,而且BP神经网络的学习是基于梯度下降的,算法存在过学习、易陷入局部最小值点等缺陷,使得网络的预测精度不高。乔金丽等[7]用遗传算法进行边坡稳定性的预测,但是遗传算法收敛速度慢,并且算法容易早熟、操作比较复杂、泛化能力较弱。马海兴等[8]将最小二乘支持向量机(LSSVM)用于边坡稳定性研究中,LSSVM对于小样本非线性数据有很强的泛化能力,但是LSSVM预测性能的好坏依赖于惩罚参数C和核函数参数σ的选取。以上研究都取得了较好的效果,但各方法也存在不足,为了保证边坡稳定性预测结果与实际情况相符,露天煤矿边坡稳定性预测的准确率需要进一步提高。

随机森林算法(RF)是Breiman[9]在2001年提出的。随机森林算法是基于决策树的分类器集成算法,它的基本单元是决策树,是决策树的集成。随机森林算法融合了Bagging和随机特征选取两大机器学习技术,因此拥有比以往算法更多的优势[10]。随机森林算法具有需要调整的参数较少、分类速度快、不必担心过度拟合、能有效处理高维大样本数据、能估计哪个特征在分类中更具有重要性以及具有较强的抗噪音能力等特点,它避免了决策树中出现的过拟合问题。本文用随机森林算法建立煤矿边坡稳定性预测模型,并运用该模型对煤矿边坡稳定性数据进行训练和预测,预测结果表明随机森林预测模型具有较高的准确度。

1 随机森林算法

1.1 分类与回归树

分类与回归树(classification and regression tree,CART)是一种非常有趣并且十分有效的非参数分类和回归方法,由分类树和回归树两部分组成[11]。它采用一种二分递归分割的技术,将当前的样本集分为两个子样本集,使得生成的决策树的每个非叶子节点都有两个分支。CART通过构建二叉树达到预测目的,生成的决策树是结构简单的二叉树,因此,CART是简单决策树算法。

以下是CART算法描述:其中T代表当前样本集,当前候选属性集用T_attributelist表示。

1)创建根节点N;

2)为N分配类别;

3)如果T都属于同一类别或者T中只剩下一个样本则返回N为叶节点,为其分配属性;

4)遍历T_attributelist中每个属性,执行该属性上的一个划分,计算此划分的基尼系数,用G表示;

5)N的测试属性test_attribute为T_attribute中最小基尼系数G的属性;

6)划分T得到T_1、T_2两个子集;

7)对子集T_1和T_2,分别重复步骤1)~6)。

CART算法考虑到每个节点都有成为叶节点的可能,对每个叶节点都分配类别。CART算法用基尼系数作为属性划分的标准,G的计算公式为

$$G(D)=1-\sum_{i=1}^{m}p_i^2 \tag{1}$$

式中,p_i为样本集D中元素属于某类的概率。对于元素的二元分裂则由下式计算基尼系数G

$$G'(D)=\frac{|D_1|}{|D|}G(D_1)+\frac{|D_2|}{|D|}G(D_2) \tag{2}$$

式(2)将D划分为D_1和D_2两个子集,这两个子集按照式(1)计算,D进行二分裂之后的基尼系数值$G'(D)$是由两个子集的值按(2)式得到,对于单列属性的二元分裂要选取$G'(D)$最小的一个作为该属性列上的一个合理划分。

1.2 随机森林算法

随机森林算法是一种集成学习算法,常见的集成学习算法还包括装袋算法和提升算法。集成学习是一种机器学习范式,是国际机器学习界研究的热点,集成学习改善了单一方法的不足。随机森林是以CART为基本分类器,它包含多个由Bagging集成学习技术训练得到的决策树,当输入待分类的样本时,最终的分类结果由单个决策树的输出结果投票决定[12,13]。

随机森林是一个树型分类器 $h(\boldsymbol{x},\boldsymbol{\theta}_k,k=1,2,\cdots,n)$ 的集合。其中,元分类器 $h(\boldsymbol{x},\boldsymbol{\theta}_k)$ 是用CART算法构建的没有剪枝的分类回归树;$\boldsymbol{x}$ 为输入向量;$\boldsymbol{\theta}_k$ 为独立同分布的随机向量,决定了单棵树的生长过程;森林的输出采用简单多数投票法(针对分类)或单棵树输出结果的简单平均(针对回归)得到。随机森林是通过自助法(boot-strap)重采样技术,不断生成训练样本和测试样本,由训练样本生成多个分类树组成随机森林,测试数据的分类结果按分类树投票多少形成的分数而定,其步骤如下:

1) 从原始训练数据集 $S=\{(x_i,y_i)\}(i=1,\cdots,n)$ 中boot_strap抽样生成 k 个训练样本集,每个样本集是每棵分类树的全部训练数据。

2) 每个训练样本集生长成为一棵不剪枝叶的分类树 h_i。在树的每个节点处从 M 个特征中随机挑选 m 个特征($m\leqslant M$),在每个节点上从 m 个特征中依据基尼系数选取最优特征进行分支生长。这棵分类树进行充分生长,使每个节点的不纯度达到最小,不进行通常的剪枝操作。

3) 根据生成的多个树分类器对新的测试数据 $x_t,t=1,\cdots,p$ (t 为测试数据个数)进行预测,分类结果按每个树分类器的投票多少而定,即分类公式为

$$f(x_t)=\text{majority vote}\{h_i(x_i)\}_{i=1}^{n\_tree} \tag{3}$$

式中,用 majority vote 表示多数投票,n_tree 为随机森林中树的个数。在训练过程中每次抽样生成自助训练样本集,原始训练数据集中不在自助样本中的剩余数据被称为袋外数据(out-of-bag,OOB),OOB数据被用来预测分类的正确率,每次的预测结果进行汇总得到错误率的OOB估计。

随机森林的边缘函数[14]

$$r(\boldsymbol{x},\boldsymbol{y})=p\boldsymbol{\theta}[h(\boldsymbol{x},\boldsymbol{\theta})=\boldsymbol{y}]-\max_{j\neq y}p\boldsymbol{\theta}[h(\boldsymbol{x},\boldsymbol{\theta})=j] \tag{4}$$

分类器 $h(\boldsymbol{x},\boldsymbol{\theta})$ 的强度

$$s=E_{x,y}r(\boldsymbol{x},\boldsymbol{y}) \tag{5}$$

式中,$E_{x,y}$ 为输入变量和输出变量计算出的均值误差,假设 $s\geqslant 0$,根据切比雪夫不等式,由式(4)、(5)可以得到

$$PE^*\leqslant \text{var}(r)/s^2 \tag{6}$$

式中,PE^* 为随机森林的泛化误差,var(r) 根据文献[12]的方法推导获得,最终可以得到随机森林的泛化误差上界为

$$PE^*\leqslant \rho(1-s^2)/s^2 \tag{7}$$

式(7)中,ρ 为相关系数的均值,s 为分类器的强度。随机森林通过在每个节点处随机选择特征进行分支,最小化各棵分类树之间的相关性,提高了分类精确度。

随机森林形成过程如图1所示。

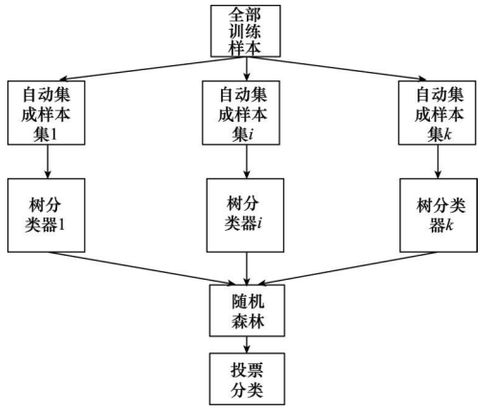


图1　随机森林

Fig. 1　Random forest

2　实验预测

2.1　边坡稳定性影响因素

露天煤矿的边坡是开挖煤矿形成的,是一个开放的自然系统,其稳定性受多种因素的综合影响,归纳起来有岩石强度、水文条件、边坡角、边坡高度、岩石结构、孔隙水压、地震及人类工程活动等。大量实际边坡工程及研究表明,对边坡稳定性起决定作用的因素有:岩石重度、黏聚力、内摩擦角、边坡角、边坡高度、孔隙水压力[15]。本文选取上述6个指标作为边坡稳定性的主要影响因素。根据文献[2],边坡的类别可划为两类:一是边坡稳定,二是边坡破坏。

2.2　实验结果

用于本次实验训练和测试的数据来源于文献[2]的39组数据,边坡类别是定性数据,作为预测模型的输出,将用2表示边坡稳定,1表示边坡破坏,用前30组数据作为训练样本集,利用随机森林算法建立边坡稳定性预测模型,通过样本数据训练确定岩石重度、黏聚力、内摩擦角、边坡角、边坡高度、孔隙水压力6个指标与煤矿边坡两个类别之间的非线性关系。随机森林有两个重要参数,一是树节点预选的变量个数 m_try,二是随机森林中树的个数 n_tree。本文在Matlab 7.0平台上做随机森林编程训练,设置变量个数 $m\_try=5$,随机森林中树的个数 $n\_tree=400$,影响边坡稳定性的6个判别指标作为预测模型的输入,边坡稳定性的两个类别作为预测模型的输出,通过训练数据集进行预测模型的训练,确定影响因素与类别之间的非线性关系,然后对未知的边坡稳定性数据进行测试。在样本集的训练过程中,图2是袋外数据OOB的错误率变化,图3为输入样本集中6个特征重要性的均值分别在准确率和基尼系数中的变化。

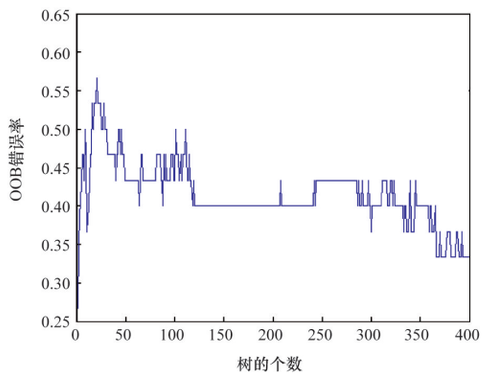
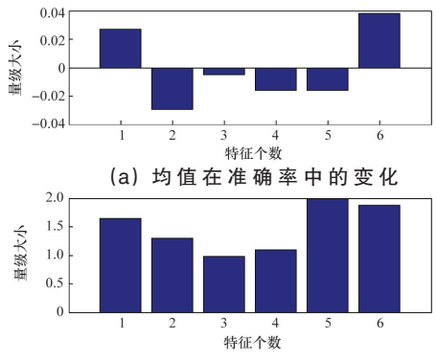


图2　OOB错误率

Fig. 2　OOB error rate



(a) 均值在准确率中的变化

(b) 均值在基尼指标中的变化

图3 特征重要性

Fig. 3 Significance of the features

随机森林训练好之后,用表1中12组数据进行测试,为了验证其预测分类的准确率,同时用支持向量机(SVM)和BP神经网络进行训练测试,随机森林(RF)测试类别、SVM测试类别和BP神经网络(BPNN)测试类别与真实类别的对比结果如表2所示(带*的数据表示预测值与真实值不符)。从表2可看出,RF测试的类别与实际类别相符合,SVM和BP神经网络预测的结果与边坡实际情况存在出入,RF测试正确判断率要高于SVM和BP神经网络测试的正确率。图4为随机森林在Matlab平台上的测试结果,2为边坡稳定状态,1为边坡破坏状态。根据图中预测结果与真实结果的对比显示,该预测模型错误判断率为0,由此可以看出,基于随机森林建立的边坡稳定性预测模型准确度更高。

表1 测试样本数据

Table 1 Test sample data

序号	岩石重度/(kN·m^{-3})	黏聚力/kPa	内摩擦角/(°)	边坡角/(°)	边坡高度/m	孔隙水压力/kPa	实际类别
1	18.84	15.32	30	25	10.67	0.38	稳定
2	22.40	100	45	45	15	0.25	稳定
3	24	0	40	33	8	0.30	稳定
4	27	10	39	40	470	0.25	稳定
5	25	46	35	47	443	0.25	稳定
6	20	20	36	45	50	0.25	破坏
7	19.63	11.97	20	22	21.19	0.40	破坏
8	21.82	8.62	32	28	12.8	0.49	破坏
9	25	55	36	45	299	0.25	稳定
10	27.3	10	39	40	480	0.25	稳定
11	25	46	35	46	393	0.25	稳定
12	25	48	40	49	330	0.25	稳定

表2 预测模型测试结果

Table 2 Test results of the prediction model

序号	真实类别	RF测试类别	SVM测试类别	BPNN测试类别
1	稳定	稳定	稳定	稳定
2	稳定	稳定	稳定	稳定
3	稳定	稳定	稳定	稳定
4	稳定	稳定	稳定	破坏*
5	稳定	稳定	稳定	稳定
6	破坏	破坏	破坏	破坏
7	破坏	破坏	破坏	破坏
8	破坏	破坏	稳定*	破坏
9	稳定	稳定	稳定	稳定
10	稳定	稳定	稳定	稳定
11	稳定	稳定	稳定	破坏*
12	稳定	稳定	稳定	稳定

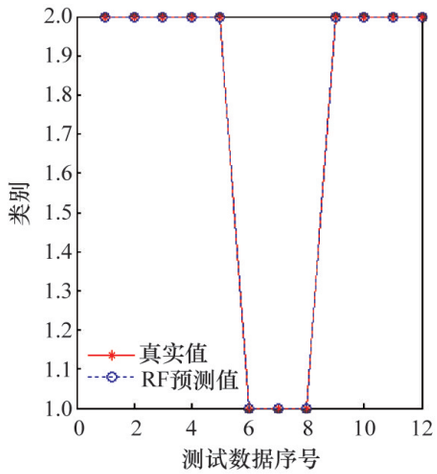


图4 RF预测结果

Fig. 4 RF prediction results

3　结论

露天煤矿边坡不稳定给煤矿安全生产带来了隐患,对煤矿安全生产起着制约作用,其边坡稳定性预测一直受到诸多因素的影响。本文尝试着将随机森林理论运用到煤矿边坡稳定性预测问题中,在借鉴国内外文献的基础上,综合选取了岩石重度、黏聚力、内摩擦角、边坡角、边坡高度、孔隙水压力6个指标作为煤矿边坡稳定的影响因素。运用随机森林集成算法,构建了基于随机森林的露天煤矿边坡稳定性预测分析模型,通过该预测方法对煤矿边坡稳定性进行预测,通过预测的结果对比,表明该随机森林预测模型具有较高的正确预测率,为煤矿边坡防治工作提供指导,同时,随机森林算法结构简单易实现,也为煤矿边坡稳定性预测提供了一种新的思路。在后续的工作中,需要进一步考虑非均值边坡的预测以及软弱层因素对于边坡的影响,收集更加丰富的数据,从定性和定量两个方面分析提取更具代表性的边坡影响因素,提高边坡预测模型的正确判别能力。

参考文献(References)

[1] Sharma R K, Mehta B S, Jamwal C S. Cut slope stability evaluation of NH-21 along Nalayan-Gambhrola section, Bilaspur district, Himachal Pradesh, India[J]. Nat Hazards, 2013, 66(6): 249-270.

[2] 张豪, 罗亦泳. 基于人工免疫算法的边坡稳定性预测模型[J]. 煤炭学报, 2012, 37(6): 911-917.
Zhang Hao, Luo Yiyong. Prediction model for slope stability based on artificial immune algorithm[J]. Journal of China Coal Society, 2012, 37(6): 911-917.

[3] 何方维, 朱明, 刘文生, 等. BP网络在露天矿边坡角优化中的应用[J]. 金属矿山, 2011(1): 35-38.
He Fangwei, Zhu Ming, Liu Wensheng, et al. Application of BP artificial neural network in optimization of open-pit slope angle[J]. Metal Mine, 2011(1): 35-38.

[4] 张均锋, 丁烨. 边坡稳定性分析的三维极限平衡法及应用[J]. 岩石力学与工程学报, 2005, 24(3): 365-370.
Zhang Junfeng, Ding Ye. Generalized 3D limit-equilibrium method for slope stability analysis and its application[J]. Chinese Journal of Rock Mechanics and Engineering, 2005, 24(3): 365-370.

[5] Chen Z, Wang X, Haberfield C, et al. A three-dimensional slope stability analysis method using the upper bound theorem Part I theory and methods[J]. International Journal of Rock Mechanics & Mining Sciences, 2001, 38(3): 369-378.

[6] 程纬华, 乔登攀, 张磊, 等. BP神经网络在露天矿边坡稳定性分析中的应用[J]. 矿冶, 2012, 21(2): 10-15.
Cheng Weihua, Qiao Dengpan, Zhang Lei, et al. Application of BP networks in the stability analysis of slopes in the open-pit mine[J]. Mining & Metallurgy, 2012, 21(2): 10-15.

[7] 乔金丽, 刘波, 李艳艳, 等. 基于遗传规划的边坡稳定安全系数预测[J]. 煤炭学报, 2010, 35(9): 1466-1469.
Qiao Jinli, Liu Bo, Li Yanyan, et al. The prediction of the safety factor of the slope stability based on genetic programming[J]. Journal of China Coal Society, 2010, 35(9): 1466-1469.

[8] 马海兴, 张刚. 基于LS-SVM的边坡稳定性预测研究[J]. 宁夏大学学报: 自然科学版, 2012, 33(3): 250-253.
Ma Haixing, Zhang Gang. Study on slope stability prediction based on LS-SVM[J]. Journal of Ningxia University: Natural Science Edition, 2012, 33(3): 250-253.

[9] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[10] 马昕, 郭静, 孙啸. 蛋白质中RNA-结合残基预测的随机森林模型[J]. 东南大学学报: 自然科学版, 2012, 42(1): 50-54.
Ma Xin, Guo Jing, Sun Xiao. Prediction of RNA-binding residues in proteins using random forest[J]. Journal of Southwest University: Natural Science Edition, 2012, 42(1): 50-54.

[11] 赵小欢, 夏靖波, 李明辉. 基于随机森林算法的网络流量分类方法[J]. 中国电子科学研究院学报, 2013, 8(2): 184-190.
Zhao Xiaohuan, Xia Jingbo, Li Minghui. Research on classification of network traffic based on random forests algorithm[J]. Journal of China Academy of Electronics and Information Technology, 2013, 8(2): 184-190.

[12] 董师师, 黄哲学. 随机森林理论浅析[J]. 集成技术, 2013, 2(1): 1-7.
Dong Shishi, Huang Zhexue. A brief theoretical overview of random forests[J]. Journal of Integration Technology, 2013, 2(1): 1-7.

[13] 杨帆, 林琛, 周绮凤, 等. 基于随机森林的潜在k近邻算法及其在基因表达数据分类中的应用[J]. 系统工程理论与实践, 2012, 32(4): 815-825.
Yang Fan, Lin Chen, Zhou Qifeng, et al. Random forest based potential k nearest neighbor classifier and its application in gene expression data[J]. Systems Engineering Theory & Practice, 2012, 32(4): 815-825.

[14] 庄进发, 罗键, 彭彦卿, 等. 基于改进随机森林的故障诊断方法研究[J]. 计算机集成制造系统, 2009, 15(4): 777-785.
Zhuang Jinfa, Luo Jian, Peng Yanqing, et al. Fault diagnosis method based on modified random forests[J]. Computer Integrated Manufacturing Systems, 2009, 15(4): 777-785.

[15] Gorog P, Torok A. Slope stability assessment of weathered clay by using field data and computer modeling: a case study from Budapest [J]. Natural Hazards and Earth System Sciences, 2007, 7(3): 417-422.

(责任编辑　侯澄芝,马宇红)