基于数据预处理的岩爆等级预测模型及精度优化*

王宇航[1],周宗红[1],李国才[2],刘剑[1]

(1.昆明理工大学 国土资源工程学院, 云南 昆明 650093;
2.新平鲁电矿业有限公司, 云南 玉溪市 653100)

摘要:岩爆预测精度对岩体工程灾害预测具有重要的现实意义,精确有效的数据预处理是后续预测工作的基础。通过收集国内外471组岩爆案例建立岩爆数据库,选取围岩最大切向应力、抗压强度、抗拉强度和弹性能量指数作为特征指标,并结合10种机器学习算法构建预测模型。为消除样本中离群值对预测模型的干扰,将离群值清洗范围缩小至单一等级内,根据岩爆烈度等级逐级检测并处理离群值。提出自适应过采样(ADASYN)改善数据分布,在保留少数类样本数据特征的情况下对原始少数类数据进行样本合成,解决各岩爆等级样本不平衡问题。引入遗传算法(GA)对高稳定性模型参数寻优,并结合混淆矩阵和多个评价指标对模型深度评估。研究表明:ADASYN方法将模型综合准确率提升11.58%,并选出最优性能GA-XGBoost模型,预测准确率和加权平均$F1$值均达到93%;将模型应用于锦屏二级水电站、三山岛金矿和马路坪矿,预测结果与现场情况有较好的一致性,可为今后岩爆预测提供新方法。

关键词:岩爆预测;离群值;数据预处理;机器学习;模型评估
中图分类号:TD311 **文献标识码**:A
文章编号:1005-2763(2024)11-0101-09

0 引言

岩爆是一种典型的岩体塑性破坏现象,表现为硬脆性岩石在开挖状态下导致弹性应变能急剧释放,发生动力失稳[1]。随着地下工程不断向深部推进,高地应力地区岩爆现象频繁发生,严重危害人员、设备和国家财产安全[2]。因此,找到精准预测岩爆的方法,并做出有效防护尤为重要。

国内外专家在岩爆预测领域做了大量研究,从不同角度提出岩爆预测的方法,但是由于岩爆机理十分复杂,仅靠单一判据获得的预测结果准确率并不高。近年来,通过实例样本数据的综合预测方法发展迅速[3],将机器学习算法与岩石力学交叉融合,从实际出发全面解决预测岩爆问题。孙臣生[4]以非线性科学理论为指导,建立了采用9个指标判据的BP神经网络改进预测模型,结合工程实例对模型进行验证。田睿等[5]通过建立了RF-AHP-云模型、IGSO-SVM和DA-DNN3种岩爆预测模型,结合工程实例进行评估分析,减少了人为因素对广义神经网络的影响。为解决岩爆预测中存在的大数据不平衡问题,汤志立等[6]考虑多因素的岩爆预测模型,研究了5种过采样方法及5种客观赋权法对模型性能的影响。李明亮等[7]用T-分布邻域嵌入(T-SNE)降维方法,对数据进行降维可视化,最后,对建立的6种岩爆预测模型进行分析评估。刘晓悦等[8]利用天牛须搜索算法(Beetle Antennace Search Algorithm, BAS)算法解决支持向量机(Support Vector Machine, SVM)中的重要参数C与gamma择优问题,并引用AdaBoost集成学习算法对BAS-SVM弱学习器进行强化训练,解决了单一分类器不稳定问题。

综合上述研究成果,利用机器学习算法构建岩爆预测模型时,需要数据预处理来提升模型精度,但模型的可靠性要基于大量工程实例数据,数据过少会导致模型过拟合。针对目前岩爆案例整合不足的问题,本文收集国内外岩爆案例共471组,建立数据库,并结合10种算法构建预测模型。当前研究易忽略样本中的离群值以及数据结构不均衡的问题,本文提出依据岩爆烈度等级逐级进行离群值处理,缩小离群值检测范围,做到消除单一等级范围内离群

* **收稿日期**:2023-12-22
基金项目:国家自然科学基金资助项目(52264019,51864023)
作者简介:王宇航(1997—),男,辽宁凌海人,硕士研究生,主要从事采矿与岩石力学等方面的研究工作。E-mail:814145045@qq.com
通信作者:周宗红(1967—),男,安徽宿州人,教授,博士生导师,主要从事采矿与岩石力学教学与研究等方面的工作。E-mail:zhou20051001@163.com

值的干扰;提出自适应过采样(ADASYN)改善数据结构,并引入 SMOTE 过采样、SMOTETomek 综合采样进行对比分析;用遗传算法(GA)对预测模型进行参数寻优,再次提高预测准确率。通过多种评价指标评估,选出最优性能模型,为岩爆预测提供新方法。

1 原理分析

1.1 XGBoost 算法基本原理

XGBoost(eXtreme Gradient Boosting)是一种集成学习方法,通过组合多个弱学习器来构建一个强大的预测模型[9]。在训练过程中,XGBoost 首先初始化一个弱学习器,然后通过梯度下降的方式迭代优化每个决策树。在每一次迭代中,XGBoost 计算当前模型的梯度和二阶导数,然后使用这些信息来构造一个新的决策树,该决策树能够减少模型的损失函数。通过重复这个过程,XGBoost 逐步改善模型的性能,直到达到预定的迭代次数或损失函数收敛。最后,在预测阶段,XGBoost 将每个样本输入到训练好的多个决策树中,并根据决策树的预测结果进行投票或加权平均,得到最终的分类结果。

XGBoost 模型如下:

$$\hat{y}_i = \sum_{t=1}^{n} f_t(x_i) \tag{1}$$

式中:n 为树的数目;f_t 为第 t 个基模型;$\hat{y}_i$ 为预测值;x_i 为输入的第 i 个数据。

XGBoost 的目标函数可以表示如下:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

$$\Omega(f_k) = \gamma T + \lambda \frac{1}{2} \sum_{j=1}^{T} \omega_j^2 \tag{3}$$

式中:$\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ 为训练损失,l 指损失函数;$\sum_{k=1}^{K} \Omega(f_k)$ 为正则化损失;y_i 为真实值;T 为叶子节点个数;ω_j 为第 j 个叶子节点权重;γ 为控制叶子节点的个数;λ 保证叶子节点的权重不至于太大。为了方便计算,运用泰勒公式进行二阶展开,目标函数的近似表示:

$$Obj = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{4}$$

$$\begin{cases} g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \\ h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \end{cases} \tag{5}$$

式中:g_i 为一阶导数;h_i 为二阶导数。

将正则化项代入上式,并进一步简化(将各个叶子节点中样本合并)得到如下:

$$Obj = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T = \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2} (H_j + \lambda) \omega_j^2 \right] + \gamma T \tag{6}$$

式中:$G_j = \sum_{i \in I_j} g_i$;$H_j = \sum_{i \in I_j} h_i$;不难发现,这个函数是关于叶子节点权重$\omega_j$ 的二次函数,其最值点 w_j^* 和最值 Obj 分别为:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{7}$$

$$Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{8}$$

1.2 遗传算法原理分析

遗传算法(Genetic Algorithm,GA)是一种通过模拟自然进化过程搜索最优解的方法[10]。它根据问题的目标函数构造一个适值函数(Fitness Function),对一个由多个解(每个解对应一个染色体)构成的种群进行评估、遗传运算、选择,经多代繁殖,获得适应值最好的个体作为问题的最优解。遗传算法的具体操作步骤如下:

(1) 生成初始化种群;

(2) 计算种群中每个个体的适应度值;

(3) 判断是否满足迭代停止条件,如满足,则输出当前最优结果;否则,转到步骤(4);

(4) 种群更新操作,即对种群进行复制、交叉及变异等操作,产生出新一代种群转到步骤(2)。

2 样本库及数据分析

2.1 岩爆数据库的建立

岩爆预测模型的输入选择尤为重要,诱发岩爆主要有内、外两种因素[11]。在高地应力环境下,开挖硐室会使应力集中和应力重新分布,此为岩爆发生的外部环境因素变化;围岩最大切向应力反映了岩爆外部因素,即地应力特征。围岩自身的力学属性为内部因素,往往硬岩和脆性岩石更易出现岩爆;岩石的单轴抗压强度、抗拉强度和弹性能量指数,即代表岩石累计弹性能量的能力。根

据王元汉等[12]的研究成果,在充分考虑内、外因素对岩爆的影响的条件下,本文选取围岩最大切向应力(Maximum Tangential Stress, MTS)、抗压强度(Uniaxial Compressive Strength, UCS)、抗拉强度(Utimate Tensile Strength, UTS)和弹性能量指数(Elastic Energy Indes, EEI)作为预测模型的特征指标。

基于案例分析的岩爆预测方法中岩爆案例的数量和质量决定模型的可靠性,在有些研究中所参考的案例较少,建立的预测模型泛化性较差[13]。因此,本文收集国内、外岩爆工程实例共 471 组作为样本数据库。现有岩爆评价体系中,通常将岩爆烈度等级分为无岩爆(Ⅰ)、轻微岩爆(Ⅱ)、中等岩爆(Ⅲ)、强烈岩爆(Ⅳ)。在所建立的岩爆数据库中,岩爆烈度等级分布见表 1。

表 1 岩爆烈度等级分布情况

Table 1 Distribution of rockburst intensity grade

岩爆烈度等级	占比/%
无岩爆	14.5
轻微岩爆	27.4
中等岩爆	38.6
强烈岩爆	19.5

2.2 数据预处理

为使预测模型有更高的准确率,首先要对原始数据做预处理,为了更直观地描绘样本数据分布情况,绘制 4 个特征的高斯函数分布曲线,以及不同特征中分 4 个岩爆等级的箱线图,见图 1。

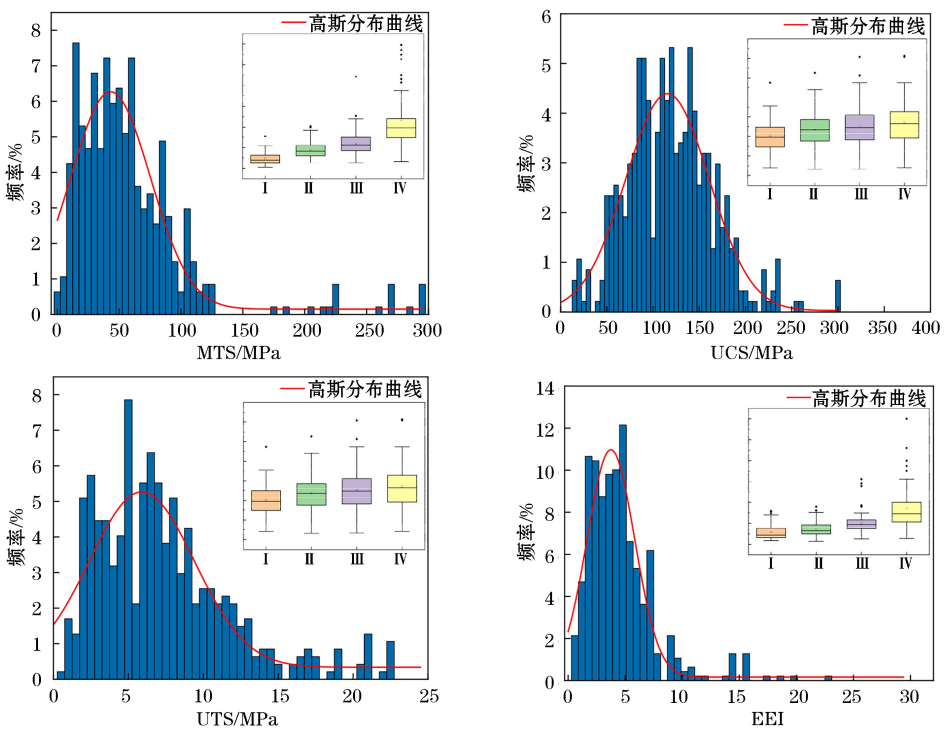


图 1 原始数据分布曲线及箱线图

Fig.1 Raw data distribution curves and box plots

图 1 中可见数量级的巨大差异和样本明显左偏现象;图中红色曲线为样本分布曲线,箱线图中实心菱形点代表样本数据极端情况,箱中的水平实线表示中位数,空心方形点表示均值,箱的边界处上、下水平线分别代表第 3 个和第 1 个四分位点(颜色标识见电子版)。由于不同特征指标的收集能力不同,就会存在一些数据误差,如某些个别样本的值与大多数其他测量值不同,将这种情况称之为异常数据或离群值[3]。为消除离群值表现出的不合理性和特殊性,提高智能模型精度,通常会对异常值直接剔除,或者对特征指标整体作异常值处理。本文提出将样本中的特征按照Ⅰ、Ⅱ、Ⅲ和Ⅳ级岩爆逐级检测离群值。通过图 3 中箱线图处理后,计算 1.5 倍四分位差的值,凡超出上须和下须的值即判定为离群值;为消除离群值干扰,本文用上、下须对应的值进行替换。离群值替换数据见表 2。

表 2 分等级离群值替换数据

Table 2 Replace data of hierarchical outliers

特征指标	岩爆等级	异常值	替换值
MTS/MPa	Ⅰ	77.69,77.69	57.95
	Ⅱ	99.09,102.38	95.20
	Ⅲ	127.13,221	122.76
	Ⅳ	225,225,225,225,274.3,274.3,274.3,274.3,297.8,297.8,297.8,297.8,207,215,285.8,263.6	190.79
UCS/MPa	Ⅰ	237.1	199.03
	Ⅱ	263	225.10
	Ⅲ	304,256.50	251.03
	Ⅳ	306.58,304.2,304.21	262.17
UTS/MPa	Ⅰ	10.6,9.52,11.2,11.5,10.2,11.96,11.96,11.96,11.96,17.66	9.45
	Ⅱ	22.6,16.72,22.6,22.4	15.33
	Ⅲ	21,21	19.42
	Ⅳ	22.6,22.6,22.6	21.08
EEI	Ⅰ	7.4,7.8,7.9	7.25
	Ⅱ	8.1,8,9	7.85
	Ⅲ	15.6,14.5,15.6,14.5,9.3,13.9,9.1,9.1,9,9	8.86
	Ⅳ	30,20,23,17.6,18.7,30	17.05

本文所选的 4 组岩爆特征指标,均为数值型数据,为消除数量级和量纲的影响,对数据库样本进行典型归一化处理。

3 构建岩爆预测模型

由于近年来机器学习被广泛应用在工程领域[14],为解决岩爆预测问题,本文引用经典机器学习算法共 10 种,分别是 SVM,KNN,MLP,RF,XGBoost,GBDT,LDA,NB,AdaBoost 和 DT,构建岩爆烈度等级预测模型。为防止模型过拟合,本文将预处理后的岩爆样本数据按 8∶2 比例随机分割训练集和测试集。模型预测步骤见图 2。

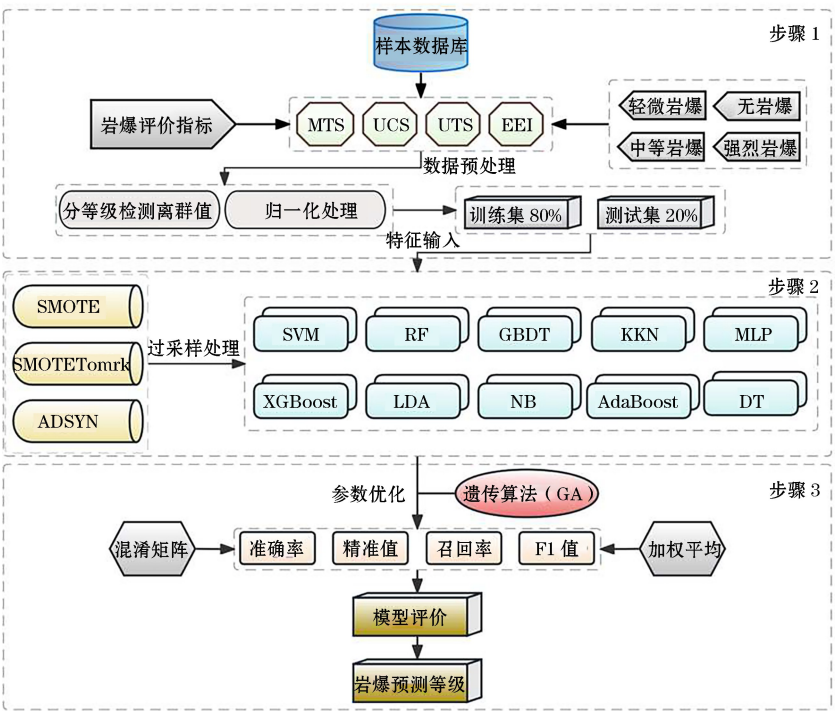


图 2 岩爆预测模型

Fig.2 Rockburst prediction model

3.1 多种采样方法

在机器学习中,每个样本对优化目标的贡献一般是相同的,若多数类比少数类大得多,就会导致分类边界更倾向多数类[15]。本文数据库中有Ⅰ级岩爆样本68组,Ⅱ级岩爆样本129组,Ⅲ级岩爆样本182组,Ⅳ级岩爆样本92组。其中,Ⅲ级岩爆样本约是Ⅰ级岩爆样本的2.7倍,反映出原始数据的不均衡性。用欠采样方法处理数据会丢失含有重要信息的多数类样本,随机过采样是采用简单复制来增加样本,易导致模型过拟合。为使岩爆样本在进行重采样后保持原有数据结构,本文提出自适应过采样(ADASYN)处理样本,增加对少数类样本的训练,使数据分布达到平衡;并结合SMOTE过采样、SMOTETomek综合采样对比分析。处理后各岩爆等级样本数目如图3所示。

由图3可知,重采样处理后不同岩爆等级间的样本比例已发生变化,由于每种采样方法核心思想不同,采样后各等级样本数目并非完全一致;ADASYN和SMOTETomek处理后各等级样本比例接近1∶1∶1∶1;SMOTE处理后各等级样本均为182组。

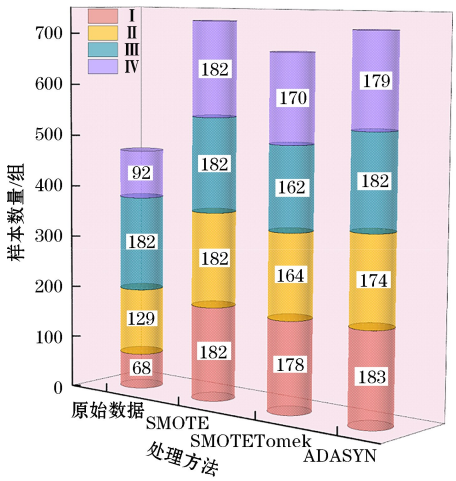


图3 不同处理方法下4个岩爆等级样本数量

Fig.3 The number of samples of 4 rockburst grades under different treatment methods

3.2 预测结果分析

将未采样样本和3种不同采样方法处理后的样本,分别输入基于10种算法的岩爆预测模型进行训练评估。测试集准确率见表3。

表3 不同采样方法下10种机器学习算法模型预测准确率

Table 3 Prediction accuracy of 10 machine learning algorithms under different sampling methods %

采样方法	不同机器学习算法预测准确率										综合准确率
	SVM	RF	GBDT	KNN	MLP	XGBoost	LDA	NB	Adaboost	DT	
未过采样	54.74	62.11	64.21	60.20	55.79	61.05	54.74	56.84	45.26	53.68	56.86
SMOTE	58.22	79.45	75.34	70.55	56.16	78.77	57.53	60.96	57.53	74.65	66.92
SMOTETomek	52.59	79.25	78.52	70.37	55.56	81.48	52.59	48.89	59.26	72.59	65.11
ADASYN	60.42	80.55	79.86	77.48	56.94	81.94	53.47	54.86	55.56	83.33	68.44
均值	56.49	75.34	74.48	69.65	56.11	75.81	54.58	55.39	54.40	71.06	—

由表3可知,样本数据未过采样时,10种算法模型的综合准确率为56.86%,经过SMOTE过采样后,模型综合准确率为66.92%,较未过采样时提升10.06个百分点;经过SMOTETomek综合采样后,模型综合准确率为65.11%,较未过采样时提升8.25个百分点;ADASYN过采样后模型综合准确率最高,为68.44%,较未采样时提升11.58个百分点,其中RF(80.55%)、XGBoost(81.94%)和DT(83.33%)3种模型准确率超过80%。

10种算法模型中以RF和XGBoost稳定性最佳,二者在4种采样状态下的准确率均值都超过了75%;经ADASYN方法处理后,RF模型准确率由62.11%提高到80.55%,增加了18.44个百分点;XGBoost模型准确率由61.05%提高到81.94%,增加了20.89个百分点。因此,ADASYN采样方法对模型性能提升最大,对解决样本数据不均衡问题有明显效果。

3.3 模型参数优化

为提高岩爆预测模型精度,本文引用遗传算法(GA)对RF模型和XGBoost模型的框架参数和学习器参数进行优化。利用遗传算法的反复交叉和重新个体评估操作,找到全局最优解。两种算法主要参数优化情况见表4。

依据上部分的结论,在ADASYN处理样本的基础上,对RF和XGBoost模型泛化能力进行深度评估有重要意义。本文引入准确率(Accuracy,A)、精确率(Precision,P_1)、召回率(Recall,R)、$F1$值和加权平均(Weighted Average)从各岩爆等级到整体评估模型的性能。为了更好理解评价指标,表5总结了分类模型的预测结果的场景。在混

淆矩阵中,样本可分为真正例(TP)、假正例(FP)、真反例(TN)和假反例(FN)。具体公式如下:

$$A=\frac{TP+TN}{TP+FP+FP+FN} \tag{9}$$

$$P_1=\frac{TP}{TP+FP} \tag{10}$$

$$R=\frac{TP}{TP+FN} \tag{11}$$

$$F1=\frac{2\times P_1\times R}{P_1+R} \tag{12}$$

表 4 RF 和 XGBoost 参数优化

Table 4 RF and XGBoost parameter optimization

算法组合	参数名称	参数范围	最优值
GA-RF	n_estimators	[100,600]	533
	min_samples_split	[2,50]	3
	min_samples_leaf	[2,50]	2
	max_leaf_nodes	[100,100]	770
	max_features	auto	auto
	max_depth	None	None
GA-XGBoost	learning_rate	[0.01,0.1]	0.1
	gamma	[0,1]	0.3
	reg_alpha	[0,1]	0.1
	max_depth	[3,50]	20
	min_child_weight	[1,10]	3
	subsample	[0.5,1]	1
	colsample_bytree	[0.5,1]	0.8

表 5 混淆矩阵

Table 5 Confusion matrix

实际等级	预测等级	
	正例(T)	反例(F)
正例(P)	真正例(TP)	假反例(FN)
反例(N)	假正例(FP)	真反例(TN)

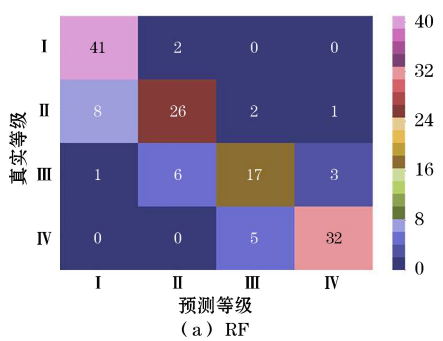
将初始模型与 GA 优化后模型对比分析,测试集结果以 4 种岩爆等级展开,见表 6;并绘制相应的混淆矩阵图,见图 4。

通过混淆矩阵计算出不同岩爆等级预测准确率,由表 6 可知,经过遗传算法进行参数优化后,GA-RF 预测准确率达到 90%,较未优化时提高 9 个百分点;GA-XGBoost 预测准确率达到 93%,较未优化时提高 11 个百分点。分等级观察图 4 和表 6 发现,RF 和 XGBoost 对Ⅱ级岩爆和Ⅲ级岩爆的误判率较高;整体上观察到二者对岩爆预测更倾向于低级别,但是并未出现将Ⅰ级岩爆误判为Ⅳ级,或者将Ⅳ级岩爆误判为Ⅰ级的情况,属于可接受误差范围。经遗传算法优化后,预测倾向性偏低情况缓解;GA-RF 对Ⅳ级岩爆预测能力提升较为突出,$F1$ 值达到 95%;加权平均 $F1$ 值为 90%。GA-XGBoost 模型对Ⅰ级岩爆预测准确率达到 100%,$F1$ 值达到 98%,可以精准判断Ⅰ级岩爆;对Ⅱ级岩爆和Ⅳ级岩爆预测准确率均超过 90%,$F1$ 值分别为 92%和 93%,加权平均 $F1$ 值达到 93%,已是本文预测模型中的最大值。

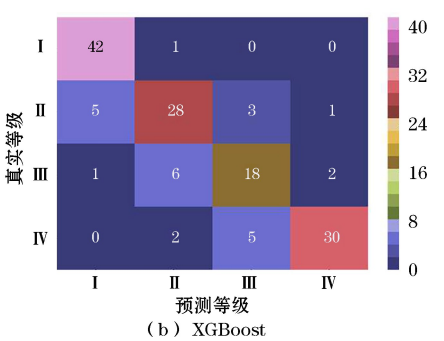
表 6 优化前后 4 等级预测结果对比

Table 6 Comparison of the prediction results of the four grades before and after optimization

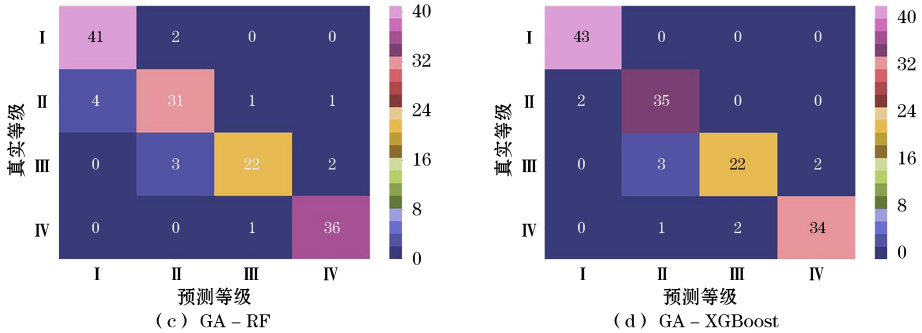
算法	岩爆等级	精确率	召回率	$F1$ 值	样本数量/组
RF	Ⅰ	0.82	0.95	0.88	43
(准确率=0.81)	Ⅱ	0.76	0.7	0.73	37
	Ⅲ	0.71	0.63	0.67	27
	Ⅳ	0.89	0.86	0.88	37
	加权平均值	0.80	0.81	0.80	144
XGBoost	Ⅰ	0.88	0.98	0.92	43
(准确率=0.82)	Ⅱ	0.76	0.76	0.76	37
	Ⅲ	0.69	0.67	0.68	27
	Ⅳ	0.91	0.81	0.86	37
	加权平均值	0.82	0.82	0.82	144
GA-RF	Ⅰ	0.91	0.95	0.93	43
(准确率=0.9)	Ⅱ	0.86	0.84	0.85	37
	Ⅲ	0.92	0.81	0.86	27
	Ⅳ	0.92	0.97	0.95	37
	加权平均值	0.90	0.90	0.90	144
GA-XGBoost	Ⅰ	0.96	1.00	0.98	43
(准确率=0.93)	Ⅱ	0.91	0.95	0.92	37
	Ⅲ	0.92	0.81	0.86	27
	Ⅳ	0.84	0.92	0.93	37
	加权平均值	0.93	0.93	0.93	144



(a) RF



(b) XGBoost



(c) GA-RF (d) GA-XGBoost

图 4 混淆矩阵测试集

Fig.4 Confusion matrix test set

综合来看,测试集准确率上升,说明经遗传算法(GA)优化后,在一定程度上避免了过拟合,提高了模型的泛化能力,优化效果显著。在实际发生岩爆工程中,不仅要考虑模型的准确率,还要承担预测错误的风险;整体比较后,认为 GA-XGBoost 模型更加可靠稳定,综合预测能力更强。

4 工程实例验证

为了进一步验证本文构建 GA-XGBoost 预测模型的准确性和适应性,选取锦屏二级水电站岩爆实例 9 组、三山岛金矿岩爆实例 11 组和马路坪矿岩爆实例 7 组,共计 27 组应用于本文预测模型;3 项工程地质情况详见文献[19]至文献[22]。预测结果如图 5 所示。

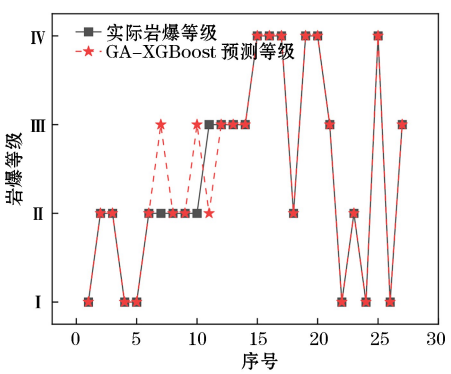


图 5 岩爆预测等级

Fig.5 Rockburst prediction grade

由图 5 可知,27 组岩爆实例中精准预测 24 组,预测模型将第 7 组、第 10 组和第 11 组判别为邻近等级,属于可接受误差范围。预测模型准确率为 88.89%,由此可见经遗传算法(GA)优化的 XGBoost 模型,可以实现对目标工程可能发生的岩爆等级预测,为地下工程中岩爆安全防护提供参考,以减轻岩爆危害。

5 结论

(1) 本文通过收集 471 组岩爆工程实例建立数据库,解决此前由于样本数量和质量较差而影响预测模型泛化能力的问题;并提出根据岩爆烈度等级逐级进行离群值处理,将离群值检测范围缩小至单一等级内,有效防止极端数据干扰预测模型精度。

(2) 引入 ADASYN 方法改善数据分布,并结合 10 种机器学习算法构建岩爆预测模型。同未处理样本、SMOTE 过采样和 SMOTETomek 综合采样处理,对比分析得到 ADASYN 方法对模型性能提升最高;有效解决样本不均衡导致的预测结果倾向多数类问题,既保存住原始样本的有效信息,又不会发生过拟合。通过综合表现优选出预测稳定性较高的 RF 算法模型和 XGBoost 算法模型。

(3) 用遗传算法(GA)对影响预测模型准确率的框架参数及学习器参数进行择优,提高收敛速度,并在反复交叉操作中得到全局最优值;经 GA 优化后的 XGBoost 模型的加权平均 $F1$ 值达到 93%,大幅提升了预测精度。通过将本文构建模型应用于锦屏二级水电站、三山岛金矿和马路坪矿,预测结果与现场情况有较好的一致性,验证了本文预测模型的可靠性。

参考文献(References):

[1] 乔木,周宗红,李岳峰,等.基于主客观赋权-物元可拓模型优选岩爆倾向性预测方法[J].有色金属工程,2022,12(8):119-130.
QIAO Mu, ZHOU Zonghong, LI Yuefeng, et al. Optimal of rockburst tendency prediction method based on subjective and objective weighting-matter-element extension model[J]. Nonferrous Metals Engineering,2022,12(8):119-130.

[2] 刘剑,周宗红,刘军,等.基于主成分分析和改进 Bayes 判别的岩爆等级预测[J].采矿与岩层控制工程学报,2022,4(5):16-26.
LIU Jian, ZHOU Zonghong, LIU Jun, et al. Prediction of rockburst grade based on principal component analysis and improved Bayesian discrimination analysis[J]. Journal of Mining And strata Control Engineering,2022,4(5):16-26.

[3] 贾义鹏.岩爆预测方法与理论模型研究[D].杭州:浙江大学,2014.
JIA Yipeng. Study on prediction method and theoretical model of rockburst[D].Hangzhou:Zhejiang University,2014.

[4] 孙臣生.基于改进 MATLAB-BP 神经网络算法的隧道岩爆预测模型[J].重庆交通大学学报(自然科学版),2019,38(10):41-49.
SUN Chensheng. A prediction model of rock burst in tunnel based on improved MATLAB-BP neural network[J].Journal of Chongqing Jiaotong University(Natural Science),2019,38(10):41-49.

[5] 田睿,孟海东,陈世江,等.基于机器学习的 3 种岩爆烈度分级预测模型对比研究[J].黄金科学技术,2020,28(6):920-929.
TIAN Rui, MENG Haidong, CHEN Shijiang, et al. Comparative study on three rockburst prediction models of intensity classification based on machine learning[J]. Gold Science and Technology,2020,28(6):920-929.

[6] 汤志立,王雪,徐千军.基于过采样和客观赋权法的岩爆预测[J].清华大学学报(自然科学版),2021,61(6):543-555.
TANG Zhili, WANG Xue, XU Qianjun. Rockburst prediction based on oversampling and objective weighting method[J]. Journal of Tsinghua University (Science and Technology),2021,61(6):543-555.

[7] 李明亮,李克钢,秦庆词,等.岩爆烈度等级预测的机器学习算法模型探讨及选择[J].岩石力学与工程学报,2021,40(增刊 1):2806-2816.
LI Mingliang, LI Kegang, QIN Qingci, et al. Discussion and selection of machine learning algorithm model for rockburst intensity grade prediction[J].Chinese Journal of Rock Mechanics and Engineering,2021,40(Sup.1):2806-2816.

[8] 刘晓悦,季红瑜.基于 AdaBoost-BAS-SVM 模型的岩爆预测研究[J].金属矿山,2021(10):28-34.
LIU Xiaoyue, JI Hongyu. Research on rockburst prediction based on Ada Boost-BAS-SVM model[J].Metal Mine,2021(10):28-34.

[9] 陈振宇,刘金波,李晨,等.基于 LSTM 与 XGBoost 组合模型的超短期电力负荷预测[J].电网技术,2020,44(2):614-620.
CHEN Zhenyu, LIU Jinbo, LI Chen, et al. Ultra short-term power forecasting based on combined LSTM-XGBoost model[J].Power System Technology,2020,44(2):614-620.

[10] 李少波,宋启松,李志昂,等.遗传算法在机器人路径规划中的研究综述[J].科学技术与工程,2020,20(2):423-431.
LI Shaobo, SONG Qisong, LI Zhi'ang, et al. Review of genetic algorithm in robot path planning [J]. Science Technology and Engineering,2020,20(2):423-431.

[11] 谢学斌,李德玄,孔令燕,等.基于 CRITIC-XGB 算法的岩爆倾向等级预测模型[J].岩石力学与工程学报,2020,39(10):1975-1982.
XIE Xuebin, LI Dexuan, KONG Lingyan, et al. Rockburst model based on CRITIC-XGB algorithm[J].Chinese Journal of Rock Mechanics and Engineering,2020,39(10):1975-1982.

[12] 王元汉,李卧东,李启光,等.岩爆预测的模糊数学综合评判方法[J].岩石力学与工程学报,1998(5):15-23.
WANG Yuanhan, LI Wodong, LI Qiguang, et al. Fuzzy mathematical comprehensive evaluation method for rockburst prediction[J].Chinese Journal of Rock Mechanics and Engineering,1998(5):15-23.

[13] 汤志立,徐千军.基于 9 种机器学习算法的岩爆预测研究[J].岩石力学与工程学报,2020,39(4):773-781.
TANG Zhili, XU Qianjun. Rockburst predictionbased on nine machine learning algorithms[J]. Chinese Journal of Rock Mechanics and Engineering,2020,39(4):773-781.

[14] 刘剑,周宗红.基于修正散点图矩阵与随机森林的岩爆等级预测[J].有色金属工程,2022,12(3):120-128.
LIU Jian, ZHOU Zonghong. Rockburst grade prediction based on modified scatter graph matrix and random forest [J].Nonferrous Metals Engineering,2022,12(3):120-128.

[15] 李艳霞,柴毅,胡友强,等.不平衡数据分类方法综述[J].控制与决策,2019,34(4):673-688.
LI Yanxia, CHAI Yi, HU Youqiang, et al. Review of imbalanced data classification methods [J]. Control and Decision,2019,34(4):673-688.

[16] 王乐,韩萌,李小娟,等.不平衡数据集分类方法综述[J].计算机工程与应用,2021,57(22):42-52.
WANG Le, HAN Meng, LI Xiaojuan, et al. Review of classification methods for unbalanced data sets[J]. Computer Engineering and Applications,2021,57(22):42-52.

[17] 魏志强,张浩,陈龙.一种采用 SmoteTomek 和 LightGBM 算法的 Web 异常检测模型[J].小型微型计算机系统,2020,41(3):587-592.
WEI Zhiqiang, ZHANG Hao, CHEN Long. Web anomaly detection model using SmoteTomek and LightGBM algorithms[J].Journal of Chinese Computer Systems,2020,41(3):587-592.

[18] 崔少泽,赵森尧,王延章.基于 ADASYN-IFA-Stacking 的再入院患者风险预测方法[J].系统工程理论与实践,2021,41(3):744-758.
CUI Shaoze, ZHAO Senyao, WANG Yanzhang. Risk prediction method for readmission patients based on ADASYN-IFA-Stacking[J].Systems Engineering-Theory & Practice,2021,41(3):744-758.

[19] ZHANG Q C, ZHOU H, FENG T X. An index for estimating the stability of brittle surrounding rock mass: FAI and its engineering application[J]. Rock Mechanics and Rock Engineering,2011,44(4):401-414.

[20] LIU Huanbin, ZHAO Guoyan, XIAO Peng, et al. Ensemble tree model for long-term rockburst prediction in incomplete datasets[J]. Minerals,2023,13(1):103.

[21] LI Diyuan, LIU Zida, XIAO Peng, et al. Intelligent rockburst prediction model with sample category balance using feedforward neural network and Bayesian optimization [J]. Underground Space,2022,7(5):833-846.

[22] 杨金林,李夕兵,周子龙,等.基于粗糙集理论的岩爆预测模糊综合评价[J].金属矿山,2010(6):26-29.
YANG Jinlin, LI Xibing, ZHOU Zilong, et al. A fuzzy assessment method of rock-burst prediction based on rough set theory[J].Metal Mine,2010(6):26-29.

Prediction Model and Accuracy Optimization of Rockburst Grade Based on Data Preprocessing

WANG Yuhang[1], *ZHOU Zonghong*[1], *LI Guocai*[2], *LIU Jian*[1]

(1.College of Land and Resources Engineering, Kunming University of Science and Technology, Kunming, Yunnan 650093, China;
2.Xinping Ludian Mining Co., Ltd., Yuxi, Yunnan 653100, China)

Abstract: The accuracy of rockburst prediction has an important practical significance for the prediction of rock mass engineering disasters. Accurate and effective data preprocessing is the basis of subsequent prediction work. The rockburst database was established by collecting 471 groups of rockburst cases at home and abroad. The maximum tangential stress, compressive strength, tensile strength and elastic energy indexes of surrounding rocks were selected as the characteristic indexes, and the prediction model was constructed by combining 10 machine learning algorithms. In order to eliminate the interference of outliers in the samples to the prediction model, the outlier cleaning range was reduced to a single level, and the outliers were detected and processed step by step according to the rockburst intensity level. An adaptive oversampling (ADASYN) was proposed to improve the data distribution, and the sample synthesis of the original minority class data was carried out under the condition of retaining the characteristics of the minority class sample data, so as to solve the problem of sample imbalance of each rock burst grade. The genetic algorithm (GA) was introduced to optimize the parameters of the high stability model, and the model was deeply evaluated by combining the confusion matrix and multiple evaluation indexes. The research shows that the ADASYN method improves the comprehensive accuracy of the model by 11.58%, and GA-XGBoost model has been selected as the optimal performance. The prediction accuracy and weighted average F1 value reach 93%. The model was applied to the Jinping Ⅱ Hydropower Station, Sanshandao Gold Mine, and Maluping Mine, and the predicted results showed good consistency with the on-site conditions, providing a new method for predicting rock bursts in the future.

Key words: Rockburst prediction, Outlier, Data preprocessing, Machine learning, Model evaluation