doi:10.3963/j.issn.1001-487X.2023.02.014

基于特征工程的XGBoost爆破块度预测研究*

夏淑媛,董永峰,王利琴
(河北工业大学 人工智能与数据科学学院,天津 300401)

摘 要: 露天矿山台阶爆破后矿岩的平均块度是衡量爆破质量的重要指标。早期研究主要依靠经验公式总结、岩体力学模型计算等方法,这些方法存在准确率不够、主观性强等缺点。近期,机器学习算法应用于块度预测,但基本通过专家经验选用固定的特征来进行预测且预测稳定性不足,泛化能力差。针对以上缺点,提出一种基于特征工程的极端梯度提升树(XGBoost)爆破块度预测模型。以太原袁家村铁矿为研究区,采集近半年的爆破数据作为原始数据,综合考虑影响平均块度的各方面因素。首先使用随机森林(RF)的袋外估计和互信息(MI)两种方法分别进行特征选择,其次将不同方法选择的特征子集集成并利用特征之间的互信息进行去冗余,最后以MSE的值为评价指标选出最优特征子集表征爆破,完成基于数据驱动的特征选择。更进一步,在最优特征子集上采用XGBoost算法进行块度预测,通过均方误差(MSE)、平均绝对误差(MAE)两个指标构成模型的评价体系将文章所提方法与其他传统机器学习算法进行对比。对比结果表明:文章提出方法比传统机器学习算法的预测准确率更高,可以为爆破的管理与控制提供科学指导。

关键词: 随机森林; 互信息; Xgboost模型; 平均块度
中图分类号: TD235.3 **文献标识码:** A **文章编号:** 1001-487X(2023)02-0097-05

Study on Blasting Lumpiness by XGBoost Model based on Feature Engineering

XIA Shu-yuan, DONG Yong-feng, WANG Li-qin
(School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China)

Abstract: The average lumpiness of ore rock is an important index to measure the blasting quality. The early research mainly relies on empirical formula summary, rock mechanics model calculation, which have shortcomings such as insufficient accuracy and strong subjectivity. Recently, , machine learning algorithm is applied for prediction, but still have problems such as empirical feature selection, insufficient model prediction stability, and poor generalization ability for the prediction of blasting material fragmentation. Aiming at above shortcomings, an extreme Gradient Boosting (xgboost) blasting fragmentation prediction model based on Feature Engineering is proposed. Taking Yuanjiacun Iron Mine in Taiyuan as the research area, engineering data are collected, Random Forest(RF) and Mutual Information (MI) are used for feature selection respectively, and the two feature subsets are integrated to obtain the best feature subset based on the value of MSE. XGBoost is used to predict the block size on the optimal feature subset, and the evaluation system is composed of two indexes: Mean Square Error (MSE) and Mean Absolute Error(MAE). The proposed method is compared with other traditional machine learning algorithms, and the results show that it is better than others. Furthermore, it can provide scientific guidance for the management and control of blasting.

收稿日期: 2023-01-04
作者简介: 夏淑媛(1986-),女,实验师、硕士研究生学历,主要从事数据挖掘、机器学习、物联网等方面的教学和科研工作,(E-mail)447463736@qq.com。
通讯作者: 董永峰(1976-),男,教授、博士研究生学历,主要从事大数据,知识图谱,机器学习等方面的教学和科研工作,(E-mail)dongyf@hebut.edu.cn。
基金项目: 河北省高等学校科学技术研究项目(ZD2022082);河北省高等教育教学改革研究与实践项目(2020GJJG027)

Key words: random forest; mutual information; XGBoost-model; average lumpiness

爆破是矿石开采中最重要的环节,评价爆破效果最重要的参数之一就是块度[1,2],爆破块度不仅影响爆破综合成本,还影响装载、运输等后续环节的效率[3,4],因此实现爆破设计参数的优化,对爆破块度进行预测和控制是爆破施工的重要目标。爆破早期研究主要依靠现场爆破试验、经验公式总结、岩体力学模型计算等方法[5],其中以 Cunningham 提出的 KUZ-RAM 模型为代表[6],能够较好地预测爆破块度。随着计算机科学与人工智能技术的创新与发展,针对传统爆破料堆块度预测方法的不足,机器学习(Machine Learning)方法正逐渐应用于爆破块度预测问题,并逐渐从经验层次向自动化、智能化层次发展[2,5,7-9]。美国的 Kulatilake 和土耳其 T Hudaverdi 等提出采用人工神经元网络(ANN)方法预测岩石爆破中碎石的平均块度[7],通过多种算法模型对比,证明了神经网络模型的可行性;史秀志等得出 SVM、LS-SVR 方法预测岩石爆破块度优于 Kuz-Ram 公式法[10,11];T Hudaverdi 等基于多元回归分析(MVRA)方法[12],提出了考虑岩石节理特性、炸药性质及钻孔参数的爆破块度预测模型,但拟合精度仍有待提高。王仁超等随机森林回归方法在模型预测性能上优于 BPNN、SVR 模型[13];叶海旺等提出 LOO-XGBoost 模型主要针对爆破领域的小样本问题[14],预测性能好于同条件下的 SVR、BPNN、RF 以及 10 折交叉验证下的 XGBoost 模型。但上述所有研究采用的均是 T Hudaverdi 构建的数据库中的 91 个爆破数据,选取爆破特征时直接采用爆破工程研究者提出的比率形式,这就导致数据来源单一,输入特征单一且过于依靠专家经验的问题凸显。针对爆破特征选择过程存在主观性及所采用传统机器学习算法性能不佳等问题,提出一种基于特征工程的 XGBoost 的爆破块度预测方法,真正地实现使用工程数据来选择影响爆破效果的特征,并通过构建模型对这些特征进行分析预测,辅助相关爆破专业人员进行爆破相关参数配置。首先整理,筛选,剔除明显异常工况数据,接着使用随机森林以及互信息以 MSE 为评价指标分别进行特征选择,形成特征子集 M 和 N,更进一步,依据最小冗余原则将 M 和 N 进行集成,形成最优特征子集 U。最后,在 U 上使用不同模型计算,通过比较不同模型的 MSE 和 MAE 指标值,验证本文所提方法的可行性和优异性。

1 特征选择与极限梯度提升算法原理

1.1 基于随机森林(RF)特征选择

对每一棵决策树,选择相应的袋外数据(Out-Of-Bag,OOB)计算袋外数据误差 errOOB1(Out-Of-Bag1 Error,errOOB1);随机对袋外数据 OOB 所有样本的特征 X 加入噪声干扰,再次计算袋外数据误差,记为 errOOB2;计算所有决策树的测试平均误差,以平均精度下降率(Mean Decrease in Accuracy,MDA)作为指标进行特征重要性计算[15],MDA 公式如下

$$MDA = \sum_{i=1}^{n} (errOOB_{i1} - errOOB_{i2})/N \quad (1)$$

如果加入随机噪声后,袋外数据准确率大幅度下降,说明这个特征对于样本的预测结果有很大影响,进而说明重要程度比较高。依此方法对爆破相关特征进行重要性排序,进行特征选择;以 XGBoost 算法为基准模型依次放入 topK 个特征验证不同特征子集的 MSE 值,选取 MSE 值最小的 topK 个特征为特征子集 M。

1.2 基于互信息(MI)特征选择

互信息的公式为

$$I(X,Y) = \int_X\int_Y P(X,Y)\log\frac{P(X,Y)}{P(X)P(Y)} = H(Y)\text{-}H(Y \mid X) \quad (2)$$

式中,$I(X,Y)$ 表示由 X 引入而使 Y 的不确定度减小的量。$I(X,Y)$ 越大,表明两个变量相关性越大,$I(X,Y)$ 取 0 时,代表 X 与 Y 独立。

基于最大相关最小冗余准则,计算特征与平均块度的互信息,删除互信息为 0 的特征;计算留下特征之间的互信息,若两两互信息值大,则保留一个特征,最终形成特征子集 N。

1.3 XGBoost

XGBoost 是由 K 个基模型组成的一个加法模型,设我们第 t 次迭代训练的树模型

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (3)$$

XGBoost 的目标函数定义如下

$$Obj = \sum_{i=1}^{n} l(y_i,\hat{y}_i) + \sum_{i=1}^{t} l(y_i,\hat{y}_i)\Omega(f_t) \quad (4)$$

式中,$\sum_{i=1}^{t}\Omega(f_i)$ 是将全部 t 棵树的复杂度进行求和。根据公式(3)(4),以第 t 步的模型为例,目标函数可以写成

$$Obj^{(t)} = \sum_{i=1}^{n} l[y_i,\hat{y}_i^{(t-1)} + f_t(x_i)] + \Omega(f_t) + constant \quad (5)$$

将公式(5)泰勒展开,第 t 步时,$l[y_i,\hat{y}_i^{(t-1)}]$ 是常数,去掉全部常数项,目标函数为

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{6}$$

式中,$g_i = \frac{\partial l[y_i, \hat{y}_i^{(t-1)}]}{\partial t \hat{y}_i^{(t-1)}}$,$h_i = \frac{\partial^2 l[y_i, \hat{y}_i^{(t-1)}]}{\partial [\hat{y}_i^{(t-1)}]}$。

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

式(7)定义了一棵树,其中 T 为叶子结点的复杂度,γ 为惩罚系数,w_j 为叶子结点的权重。将式(7)代入式(6),整理最终目标函数为

$$Obj^{(t)} \approx \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2\right] + \gamma T \tag{8}$$

$G_j = \sum_{i \in I_j} g_j$ 叶子结点 j 所包含样本的一阶偏导数累加之和,$H_j = \sum_{i \in I_j} h_j$ 叶子结点 j 所包含样本的二阶偏导数累加之和。

2　块度预测模型建立

2.1　数据预处理

采集现场爆破数据,获取爆破设计参数与实例如表 1 所示。

表 1　爆破设计参数示例

Table 1　Example of blasting design parameters

序号	W	B/m	D/mm	L/m	m	H	e	Wd/m	S/m	T/m
1	0.91	4.17	140	10.6	0.790	15.1	0.47	8.73	5.29	4.5

将中孔爆破块状图复制到现状图中,测出首排孔抵抗线 W,底盘抵抗线 Wd,孔距 S,最小抵抗线 B,间距系数 $m(B/S)$;当排距 B 小于 6 m 时,孔径 D 取 140 ms,大于 6 m 时,D 取 310 ms;在深孔药量计算表中,不计超钻部分的装药长度 L 是用实际的孔深减去超深减去堵塞长度;H 是台阶高度,T 堵塞长度;e 一般取值为 0.3 ~ 0.6 之间,在实地场景中经过统计取值为 0.47。岩石信息参数与实例如表 2 所示。

表 2　岩石信息参数示例

Table 2　Examples of rock information parameters

编号	E	A	容重(kg · m^{-3})	抗压强度/MPa	抗拉强度/MPa	泊松比	内聚力/MPA	内摩擦角/°
1	80.64	10	3301.0	139.05	16.09	0.142	23.65	53.43

岩石数据除了 A,都是由矿山工程地质及岩体力学试验计算得出。A 取值大小与岩石节理,裂隙发育程度有关,中硬岩 $A=7$,节理发育岩 $A=10$,节理不发育坚硬岩 $A=13$。如 βu(辉绿岩)、AFQ(含铁石英岩)取 10,AC(绢云片岩),AS(As 白云片岩)取 7。

炸药信息参数与实例如表 3 所示。

表 3　炸药信息参数示例

Table 3　Example of explosive information parameters

编号	pf	Q/kg	EZ
1	0.64	210.00	100

炸药单耗 pf、单孔装药量 Q 从爆破通知单里直接取出,采用的是乳化炸药,炸药相对重要威力 EZ 取 100。

引入 Hudaverdi 开发的数据库中的比率形式来表征数据,具体实例如表 4 所示。

表 4　比率形式参数实例

Table 4　Examples of ratio form parameters

编号	S/B	H/B	B/D	T/B
1	1.2686	3.6211	29.7857	1.0791

以爆破平均块度 $X50$ 为输出,其值是由 split 图像分析软件获得。一次爆炸后,将一颗篮球放在爆堆上,从多个方向抓取爆堆图像分别进行计算,再将计算值求和取平均作为此次爆破平均块度值。

2.2　建模流程

删除异常数据;基于随机森林以及互信息方法进行特征选择获得特征子集 M 和 N;集成 M 与 N 并经计算特征之间的互信息值,去冗余获得最优特征子集 U;将经过特征选择后的数据划分为训练样本和测试样本,为模型的训练及测试备好数据集;利用训练样本对 Xgboost 模型进行训练,得出该训练样本对应的爆破块度预测模型;利用测试样本对于训练好的的模型进行检验。将 SVR、GBDT、线性回归三种传统机器学习算法与 XGBOOST 算法模型进行横向对比,检验模型优越性。建模流程如图 1 所示。

3　实例分析

3.1　特征选择

依据太钢袁家村铁矿 2020.9 ~ 2021.2 连续 200 多次爆破记录数据,经过整理,筛选,剔除明显异常工况数据,留取约 150 条有效数据进行训练,其

中每条数据包含 23 个特征。

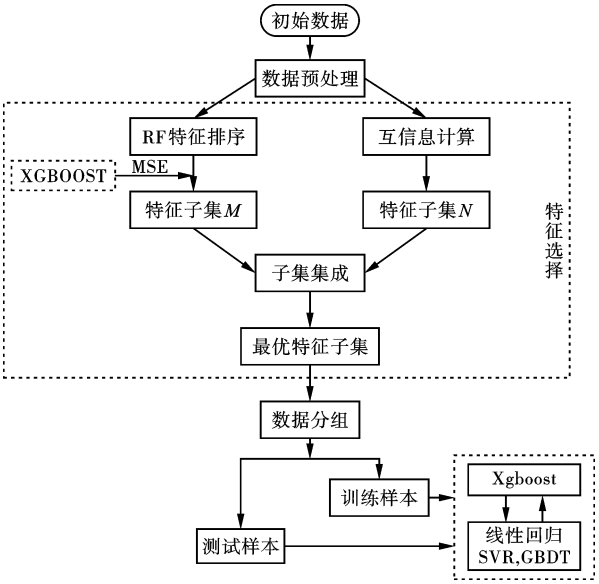

图 1 爆破块度预测模型构建流程

Fig. 1 Construction process of blasting fragmentation prediction model

使用随机森林袋外估计计算，为了减少算法的随机性，对特征进行 5 次计算取平均值进行排序。如表 5 所示。

对表中特征进行特征重要度排序，结果如图 2 所示。

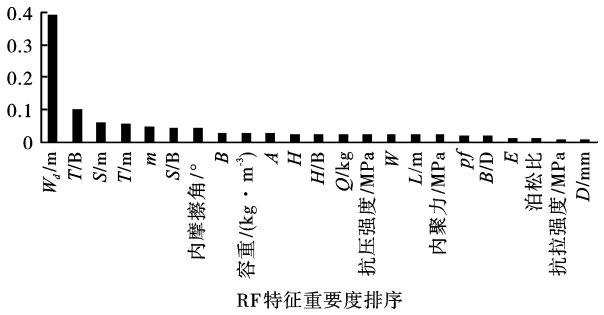

图 2 随机森林特征重要度排序

Fig. 2 Ranking of random forest feature importance

选取特征重要度为 topK 的特征作为特征子集，以 XGBoost 算法为基准模型运行验证不同特征子集的 MSE 值结果如图 3 所示。

表 5 *RF* 特征重要性统计表

Table 5 Statistics of *RF* feature importance

特征	实验序号					平均值
	1	2	3	4	5	
W	0.0262	0.011	0.0324	0.0174	0.0157	0.0205
B	0.0200	0.0231	0.0430	0.0237	0.0167	0.0253
D	0.0003	0.0006	0.0004	0.0003	0.0004	0.0004
L	0.0346	0.0108	0.0258	0.0235	0.0080	0.0205
m	0.0582	0.0370	0.0591	0.0536	0.0108	0.0437
H	0.0207	0.0207	0.0212	0.0258	0.0199	0.0217
Wd	0.2644	0.3667	0.1988	0.4001	0.7247	0.3910
S	0.0313	0.0714	0.0649	0.1078	0.0197	0.0590
T	0.0098	0.0396	0.1975	0.0155	0.0066	0.0538
A	0.0372	0.0140	0.0268	0.0261	0.0206	0.0249
E	0.0088	0.0076	0.0082	0.0125	0.0036	0.0082
pf	0.0221	0.0072	0.0129	0.0119	0.0229	0.0154
Q	0.0169	0.0226	0.0331	0.0263	0.0062	0.0210
容重	0.0296	0.0346	0.0362	0.0192	0.0053	0.0250
抗压强度	0.0222	0.0340	0.0214	0.0229	0.0035	0.0208
抗拉强度	0.0079	0.0029	0.0072	0.0061	0.0026	0.0053
泊松比	0.0084	0.0076	0.0071	0.0095	0.0029	0.0071
内聚力	0.0216	0.0354	0.0197	0.0208	0.0032	0.0202
内摩擦角	0.0553	0.0300	0.0648	0.0313	0.0203	0.0403
S/B	0.0753	0.0249	0.0512	0.0390	0.0167	0.0414
H/B	0.0244	0.0255	0.0216	0.016	0.0181	0.0212
B/D	0.0195	0.0084	0.0186	0.0137	0.0139	0.0148
T/B	0.1852	0.1642	0.0279	0.0763	0.0378	0.0983

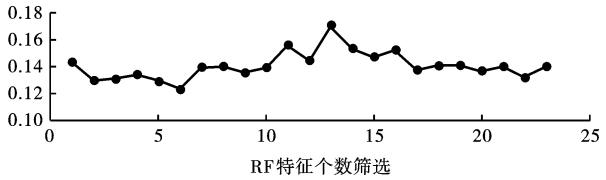

图 3　特征个数筛选

Fig. 3　Selecting the number of features

图 3 中显示,当选取前 6 个特征时,模型的 *MSE* 值最低。确定特征子集 $M\{Wd(m),T/B,S(m),T(m),m,S/B\}$。

计算特征与平均块度的互信息值,去除 $I=0$ 的特征,剩下为 W,m,Wd,T,E,容重,抗压强度,抗拉强度,泊松比,内聚力,S/B。计算特征之间的互信息值并归一化,I(抗拉强度,抗压强度) = 0.8645;I(抗拉强度,内聚力) = 0.9643;I(泊松比,容重) = 0.6674;$I(m,S/B)$ = 0.9943,结合 *RF* 对于特征重要度的排序留下一个特征,最终获得特征子集 $N\{W,m,Wd,T,E$,容重,抗拉强度$\}$。

集合特征子集 M 与 N。再次计算特征之间的互信息值去冗余。最终确定 $U\{Wd(m),T/B,S(m),T(m),m,W,E$,容重,抗拉强度$\}$。集成两种特征选择方法,有效地克服了 *RF* 不能遍历所有特征组合以及互信息选取特征泛化能力差的缺点。

以 XGboost 算法为基准模型,对特征选择的效果进行验证,分别选取 RF、互信息以及集成后的特征进行横向的对比试验,对比结果如表 6 所示。

表 6　特征选择方案对比

Table 6　Comparison of feature selection schemes

方法	RF	互信息	特征集成
MSE 值	0.1237	0.1299	0.0810

由表可知,使用特征集成方法,不仅留取了对 *X*50 重要的特征,而且还提高模型训练的准确度,保证了选取特征的稳定性和全面性。

3.2　验证预测模型

利用筛选出来的特征,分别选取线性回归(Linear Regression),支持向量回归(SVR),梯度提升树回归(GBDT)和 Xgboost 四种模型进行实验分析,采用均方误差(Mean Square Error, MSE)、平均绝对误差(Mean Absolute Error,MAE)作为实验结果的评价指标,MSE 和 MAE 的计算公式如下

$$MSE = \frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\ test - \hat{y}_{test}^{(i)}]^2$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y^{(i)}\ test - \hat{y}_{test}^{(i)}|$$

实验结果如表 7 所示。

表 7　模型预测结果对比

Table 7　Comparison of model prediction results

模型应用算法	MSE	MAE
Linear Regression	0.0939	0.1691
XGBoost	0.0588	0.1680
SVR	0.0673	0.1744
GBDT	0.0990	0.1880

通过比较可以看出 XGBoost 和 SVR 的性能相似,但明显好于 LR 以及 GBDT。比较 MAE 平均绝对误差,可发现 XGBoost 以及 LR 的性能要略好于其他两种算法,综上可知 XGBoost 在预测性能要好于其他三种模型。

4　结论

利用工程数据,通过比较不同特征选择方法的 MSE 值,证明集成后的特征更能表征爆破,通过分析选取的特征,发现影响爆破效果的因素包含岩体、炸药以及爆破设计参数各方面信息,完成了依靠数据驱动的特征选择;除此之外,利用已选特征组合将 XGBoost 和其他预测模型进行对比,证明 XGBoost 模型较传统的机器学习预测准确率上有较大的提升,进一步论证了基于特征工程的 XGBoost 爆破块度预测模型能为矿石爆破施工提供指导,为爆破施工智能化管理与控制提供可能。

参考文献(References)

[1]　李　想.爆破工程安全评估系统化管理研究及软件开发[D].贵阳:贵州大学,2015.

[1]　LI Xiang. The research about systematic management of blasting engineering's safety assessment and software development[D]. Guiyang: Guizhou University, 2015. (in Chinese)

[2]　史秀志,郭　霆,尚雪义,等.基于 PCA-BP 神经网络的岩石爆破平均粒径预测[J].爆破,2016,33(2):55-61.

[2]　SHI Xiu-zhi, GUO Ting, SHANG Xue-yi, et al. Prediction of mean particle size of rock blast based on combination of PCA and BP neural networks[J]. Blasting, 2016, 33(2): 55-61. (in Chinese)

[3]　王仁超,李世鹏,徐跃明,等.抽水蓄能电站开挖施工仿真研究[J].水力发电学报,2018,37(3):18-27.

[3]　WANG Ren-chao, LI Shi-peng, XU Yue-ming, et al. Study on excavation and construction simulation of pumped storage power station]J]. Journal of Hydroelectric Engineering, 2018, 37(3): 18-27. (in Chinese)

(下转第 131 页)