

文章编号:1671-251X(2021)03-0060-06

DOI:10.13272/j.issn.1671-251x.2020100038

基于 EMD 特征提取与随机森林的煤矸识别方法

窦希杰^{1,2}, 王世博^{1,2}, 刘后广¹, 陈钱有^{1,2}, 邹文才^{1,2}, 卢召栋^{1,2}

(1. 中国矿业大学 机电工程学院, 江苏 徐州 221116;

2. 中国矿业大学 矿山智能采掘装备协同创新中心, 江苏 徐州 221116)



扫码移动阅读

摘要:基于振动信号辨识是实现综放开采煤矸识别的有效手段,现有方法在识别准确性和有效性方面有待进一步研究。提出了一种基于经验模态分解(EMD)特征提取与随机森林(RF)的煤矸识别方法。采用加速度传感器及数据采集仪采集了某综放工作面煤和矸石冲击液压支架尾梁产生的振动信号,分别对 2 种信号进行 EMD,得到一系列本征模态函数(IMF);根据 EMD 结果选取有效 IMF,分别提取 IMF 能量、峭度、矩阵奇异值及对应熵作为特征向量,采用各特征向量独立训练 RF 模型,根据各 RF 模型对测试样本的识别结果筛选特征向量,并建立特征数据集;采用特征数据集训练 RF 模型,采用训练好的 RF 模型实现煤矸识别。测试结果表明:该方法对 200 组煤矸测试样本的识别准确率达 96.5%,且当 RF 模型中决策树数量设置为 100 或 150 时识别准确率最高,对测试样本进行特征提取与识别的耗时不超过 0.2 s,满足综放工作面煤矸识别准确性和实时性要求。

关键词:煤炭开采;综放开采;放顶煤;煤矸识别;液压支架尾梁;振动信号;经验模态分解;随机森林
中图分类号:TD821 文献标志码:A

Coal and gangue identification method based on EMD feature extraction and random forest

DOU Xijie^{1,2}, WANG Shibo^{1,2}, LIU Houguang¹, CHEN Qianyou^{1,2},
ZOU Wencai^{1,2}, LU Zhaodong^{1,2}

(1. School of Mechanical and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China; 2. Intelligent Mining Equipment Collaborative Innovation Center, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Identification based on vibration signals is an effective method to realize coal and gangue identification in fully mechanized mining. The existing method needs to be further studied in terms of identification accuracy and effectiveness. A coal and gangue identification method based on empirical mode decomposition(EMD) feature extraction and random forest(RF) is proposed in this study. The acceleration sensor and data acquisition instrument are used to collect the vibration signals generated by the impact of coal and gangue on the tail beam of the hydraulic support in a fully mechanized working face. Then the two signals are processed by EMD respectively so as to obtain a series of intrinsic mode functions(IMF). The effective IMFs are selected according to the EMD results, and the IMF energy, kurtosis, matrix singular values and corresponding entropy are extracted as feature vectors. Each feature vector is used to train the RF model independently. The feature vectors are filtered according to the identification results of each RF model on the test samples, and the feature data set is established. The feature data set is used to train the RF model, and the trained RF model is applied to realize the coal and gangue identification. The test

收稿日期:2020-10-22;修回日期:2021-03-10;责任编辑:李明。

基金项目:国家重点研发计划资助项目(2018YFC0604503);江苏高校优势学科建设工程资助项目(PAPD)。

作者简介:窦希杰(1994-),男,山东滨州人,硕士研究生,研究方向为煤矸识别,E-mail:douxijie@cumt.edu.cn。通信作者:王世博(1979-),男,河北新乡人,教授,博士,研究方向为煤矿智能化开采,E-mail:wangshb@cumt.edu.cn。

引用格式:窦希杰,王世博,刘后广,等.基于 EMD 特征提取与随机森林的煤矸识别方法[J].工矿自动化,2021,47(3):60-65.

DOU Xijie, WANG Shibo, LIU Houguang, et al. Coal and gangue identification method based on EMD feature extraction and random forest[J]. Industry and Mine Automation, 2021, 47(3): 60-65.

results show that the identification accuracy of the method reaches 96.5% for 200 sets of coal and gangue test samples, and the highest identification accuracy is achieved when the number of decision trees in the RF model is set to 100 or 150. Furthermore, the time consumed for feature extraction and identification of test samples is less than 0.2 s, which meets the requirements of accuracy and real time of coal and gangue identification in fully mechanized working face.

Key words: coal mining; fully mechanized mining; top coal caving; coal and gangue identification; tail beam of hydraulic support; vibration signal; empirical mode decomposition; random forest

0 引言

放顶煤过程的智能化是制约智能化综放开采的主要技术瓶颈^[1-2]。在放顶煤过程中,根据煤层赋存条件变化进行煤矸精准识别,并根据识别结果实时自动调整放煤口启闭,是实现自动化放煤的关键^[3],不仅能够降低混矸率,提高煤炭质量,还能使放煤工人远离综放工作面,减少恶劣环境对工人健康的影响。

对放顶煤过程中产生的振动信号进行辨识是实现煤矸识别的有效手段之一。近年来,学者们针对该方法进行了大量研究。文献[4]分析了煤和矸石冲击产生的振动信号频谱特征,得出了2种振动信号频率不同的结论,但未根据频谱特征进一步研究煤矸识别技术。文献[5]分析了放顶煤过程中液压支架后尾梁及刮板输送机处采集的振动信号,认为液压支架后尾梁更适合作为煤矸冲击振动信号测点,进一步分析了不同工况下振动信号的时域特征,得出了方差、偏度与峭度指标对工况变化敏感的结论,但放顶煤过程的复杂性使得尾梁振动信号具有非平稳特性,时域特征不能准确地表征振动信号。文献[6]采用小波分析方法对煤和矸石冲击产生的振动信号进行特征提取,设计了神经网络模型实现煤矸识别,但建模时使用的训练样本和测试样本较少,模型有效性有待进一步验证。

本文提出了一种基于经验模态分解(Empirical Mode Decomposition, EMD)特征提取和随机森林(Random Forest, RF)的煤矸识别方法。该方法对综放现场采集的大量煤和矸石冲击液压支架尾梁产生的振动样本信号进行EMD,在分解产生的本征模态函数(Intrinsic Mode Function, IMF)上进行特征提取与筛选,形成最优化的特征数据集训练RF模型,提高了煤矸识别的准确性;与BP神经网络、支持向量机等机器学习算法相比,采用的RF算法可直接对特征数据集进行分类,无需进一步处理,保证了煤矸识别效率。

1 基于EMD的特征提取

1.1 EMD基本原理

EMD是N. E. Huang等^[7-8]在对瞬时频率概

念研究基础上提出的一种自适应分解的信号处理方法,在机械故障诊断^[9]、模态参数识别^[10]等工程领域得到了广泛应用。EMD可将信号中不同时间尺度的波动逐级分解出来,产生一系列IMF。各个IMF需满足2个条件^[11]:①数据序列中极值点和过零点最多相差1个。②任一时刻由信号的局部极大值与局部极小值定义的包络平均值为0。原始信号 S 经EMD后可表示为

$$S = \sum_{j=1}^n c_j + r \quad (1)$$

式中: n 为分解得到的IMF个数; c_j 为第 j 个IMF; r 为残差信号,代表信号的平均趋势。

1.2 特征提取

为了获取表征原始信号的特征向量,对各样本信号进行EMD,根据分解结果选取有效IMF,进一步提取IMF能量、峭度、矩阵奇异值及对应的熵作为特征向量,并对各特征向量的提取效率及有效性进行比较,完成特征筛选,建立特征数据集。

IMF能量为

$$E_j = \sum_{i=1}^N |c_j(i)|^2 \quad j = 1, 2, \dots, m \quad (2)$$

式中: N 为样本包含的数据点数; m 为有效IMF个数; $c_j(i)$ 为第 j 个有效IMF的第 i 个数据点。

IMF峭度为

$$K_j = \frac{E(c_j(i) - \mu_j)}{\sigma_j^4} \quad (3)$$

式中: $E(\cdot)$ 为期望函数; μ_j 为第 j 个有效IMF的均值; σ_j 为第 j 个有效IMF的标准差。

根据奇异值分解定义^[12],可对各有效IMF组成的 $m \times N$ 矩阵 C 进行奇异值分解:

$$C = UQV^H \quad (4)$$

式中: U, V 分别为 m 阶、 N 阶酉矩阵; $Q = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$, $\Sigma = \text{diag}(\omega_1, \omega_2, \dots, \omega_r)$, r 为矩阵 C 的秩, $\omega_1, \omega_2, \dots, \omega_r$ 为矩阵 C 的奇异值, $\omega_1 \geq \omega_2 \geq \dots \geq \omega_r > 0$ 。

信息熵是系统不确定性的定量评价指标^[13-15],可表征信号分布状态的不确定性和复杂程度。对于包含 m 个元素的离散随机变量 $X = \{x_1, x_2, \dots, x_m\}$,令第 j 个元素出现的概率为 p_j , $\sum_{j=1}^m p_j = 1$,则

X 的信息熵定义为

$$H(X) = \sum_{j=1}^m p_j \log_2 \frac{1}{p_j} \quad (5)$$

依据式(5),可求得 IMF 能量熵、奇异值熵及峭度熵。

2 RF 原理

RF 的实质是一个包含多棵决策树的组合分类器,在一定程度上克服了单分类器的局限性^[16-17],通过集成提高了分类器的稳定性。RF 采用 Bootstrap 方法进行重采样^[18],产生多个训练集;利用每个训练集生成对应的决策树,在构建决策树时采用随机分裂属性集方法。使用训练好的 RF 模型可直接对特征数据集进行分类,简化了计算环节,减少了特征提取耗时。

2.1 Bootstrap 重采样

设集合 T 中有 k 个样本,若每次有放回地从该集合中抽取 1 个样本,抽取 k 次形成的新集合 T^* 中不包含第 $t(t=1,2,\dots,k)$ 个样本的概率为

$$p_t = \left(1 - \frac{1}{k}\right)^k \quad (6)$$

当 k 趋于无穷时,有

$$\lim_{k \rightarrow \infty} p_t = \lim_{k \rightarrow \infty} \left(1 - \frac{1}{k}\right)^k = \exp(-1) \approx 0.368 \quad (7)$$

由式(7)可知,虽然 T^* 与 T 中样本数均为 k ,但 T^* 中可能包含了重复样本,且 T^* 中约包含 T 中 63.2% 的样本。

2.2 RF 算法流程

RF 算法流程如图 1 所示。

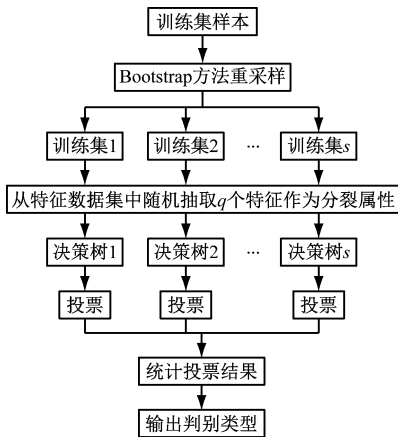


图 1 RF 算法流程

Fig. 1 RF algorithm flow

采用 Bootstrap 方法对特征数据集进行重采样,产生 s 个训练集。之后利用训练集生成对应的决策树,在每个非叶子节点选择分裂属性之前,从特征数据集中随机抽取 q 个特征作为当前节点的分裂属性。使每棵树完整生长,不进行剪枝操作,最终所有决策树构成一个 RF。当有测试集样本输入时,

RF 中每棵决策树都会输出一个结果,采用投票方法将 s 棵决策树中输出最多的类别作为测试集样本的类别输出。

3 煤矸识别流程

基于 EMD 特征提取和 RF 的煤矸识别流程如图 2 所示。首先,对放顶煤过程中采集的振动信号进行等长度截取预处理,得到一系列放煤和放矸石振动样本信号;其次,对各样本信号进行 EMD,根据分解结果选取有效 IMF,提取 IMF 能量、矩阵奇异值、峭度及对应的熵作为特征向量;再次,使用各特征向量独立训练 RF 模型,并将测试集样本数据输入训练好的 RF 模型测试特征向量的有效性,根据识别结果完成特征向量筛选,建立特征数据集;最后,使用特征数据集训练 RF 模型,通过训练好的 RF 模型实现煤矸识别。



图 2 煤矸识别流程

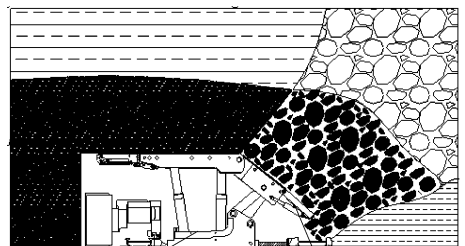
Fig. 2 Coal and gangue identification flow

4 煤矸识别试验

4.1 数据采集

煤矸识别试验使用的放煤和放矸石振动信号来源于同煤大唐塔山煤矿有限公司 8222 综放工作面。工作面全长 230.5 m,平均煤层厚度为 14.36 m。夹矸 6~17 层,夹矸单层厚度为 0.05~0.82 m。各可采煤层的物理性质相似,呈碎块状、块状、条带状结构,弱玻璃光泽,水平层理,煤层坚固性系数为 2.7~3.7。夹矸多为灰褐色高岭岩、灰白色高岭质泥岩、灰黑色炭质泥岩,坚固性系数为 4.0~4.5。

采用 1A946E 型 IEPE 压电式加速度传感器及 DH5925N 型便携式数据采集仪对放顶煤过程中顶煤和矸石冲击 ZF17000/27.5/42D 型液压支架尾梁产生的振动信号进行采集与记录。仪器设置位置如图 3 所示。采用螺纹安装方式将加速度传感器安装在液压支架尾梁背面,并通过信号线将加速度传感器与数据采集仪连接,如图 4 所示。数据采集仪布置在液压支架两立柱之间,如图 5 所示。



加速度传感器安装位置 数据采集仪安装位置

图 3 振动采集仪器布置

Fig. 3 Arrangement of vibration acquisition instruments

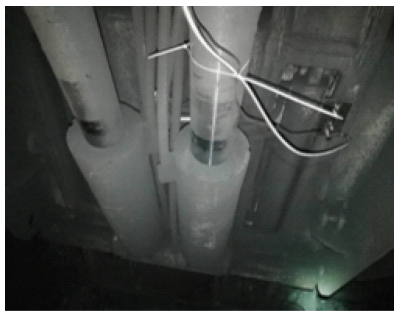


图 4 加速度传感器安装现场

Fig. 4 Installation site of acceleration sensor

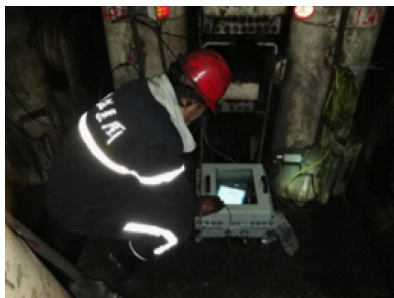
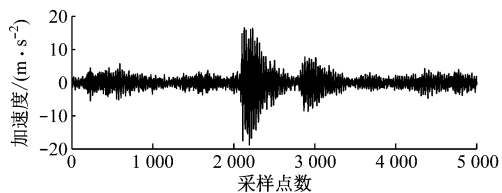


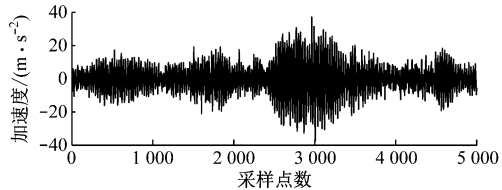
图 5 数据采集仪安装现场

Fig. 5 Installation site of data acquisition

试验开始时,将数据采集仪的采样频率设置为 25.6 kHz,采集并存储放顶煤过程中煤和矸石冲击液压支架产生的振动信号,并根据放煤工人的提示记录放煤和放矸石 2 种工况的开始及结束时刻。放煤和放矸石振动信号时域波形如图 6 所示。



(a) 放煤振动信号



(b) 放矸石振动信号

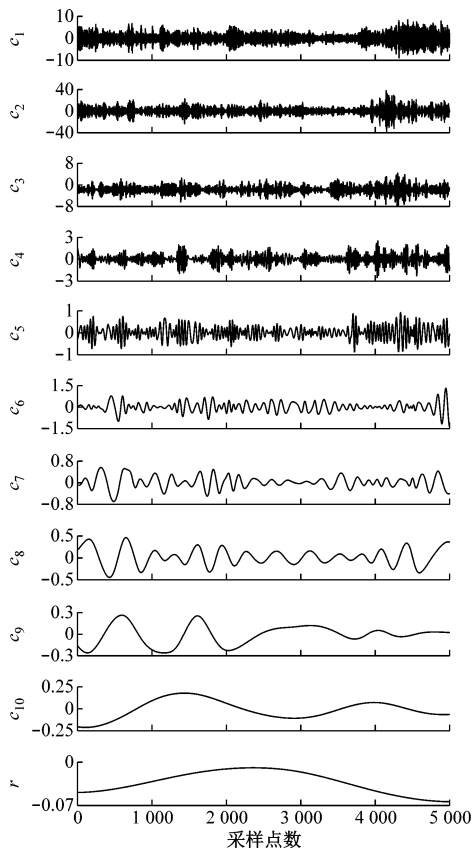
图 6 振动信号时域波形

Fig. 6 Temporal waveforms of vibration signals

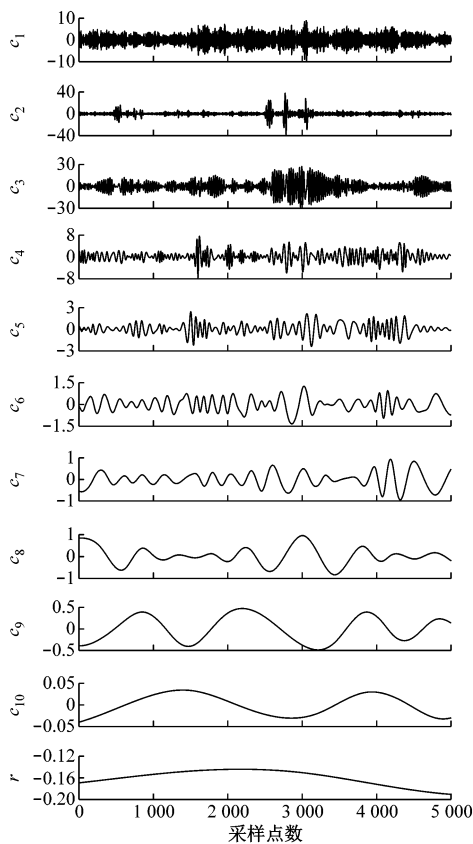
4.2 振动信号特征提取及特征数据集构建

对放顶煤过程中采集的放煤和放矸石 2 种工况下的振动信号进行 EMD,结果如图 7 所示。可看出 2 种振动信号经 EMD 后各得到 10 个 IMF 及 1 个残差分量,且对应的 IMF 幅值和波形具有明确区别。另外放煤和放矸石振动信号的能量均集中在前 8 个 IMF 中,因此将前 8 个 IMF 作为有效 IMF,提取 IMF 能量、峭度、矩阵奇异值及对应的熵作为特征向量。

对采集的原始振动信号进行预处理,以 5 000



(a) 放煤振动信号



(b) 放矸石振动信号

图 7 振动信号 EMD 结果

Fig. 7 EMD results of vibration signals

点作为 1 个样本长度,共得到放煤和放矸石振动信号各 1 100 个样本数据。从原始样本中随机抽取放

煤和放矸石振动信号各 1 000 个样本作为训练集、100 个样本作为测试集,将放煤状态标签设置为 1、放矸石状态标签设置为 0,进行 RF 模型训练及测试。

为验证所提取特征的有效性,使用各特征向量分别训练 RF 模型。设置 RF 所含决策树数量 $s = 500$,分裂属性集中属性个数 $q = M/2$ (M 为特征总数)。为保证识别结果的可靠性,每次均使用随机产生的训练集及测试集进行建模及识别,并将 10 次识别的准确率平均值作为最终结果。采用不同特征向量进行 RF 建模及识别的结果见表 1。

表 1 采用不同特征向量的煤矸识别结果

Table 1 Coal and gangue identification results by use of different feature vectors

特征	识别准确率/%		
	放煤	放矸石	平均值
IMF 能量	91.6	97.0	94.30
IMF 能量+IMF 能量熵	94.1	97.2	95.65
IMF 峭度	63.7	86.8	75.25
IMF 峭度+IMF 峭度熵	64.6	86.8	75.70
IMF 矩阵奇异值	88.3	85.4	86.85
IMF 矩阵奇异值+IMF 奇异值熵	90.0	84.7	87.35

由表 1 可知,使用 IMF 能量训练得到的 RF 模型对测试集样本的识别准确率最高,达 94.30%;采用 IMF 峭度训练得到的 RF 模型识别准确率最低,仅为 75.25%;各特征向量加入对应的熵特征再进行训练得到的 RF 模型识别准确率均小幅提升。将 IMF 峭度特征剔除,使用 IMF 能量(8 维)、IMF 能量熵(1 维)、IMF 矩阵奇异值(8 维)、IMF 奇异值熵(1 维)共 18 维数据建立特征数据集,由特征数据集训练得到的 RF 模型对放煤和放矸石各 100 组振动样本的识别准确率达 96.5%。

4.3 RF 模型性能评估

RF 模型决策树数量对其泛化性能有一定影响。为获得 RF 模型的最优决策树数量,对包含不同数量决策树的 RF 模型进行评估。具体方法:设置步长为 50,使决策树数量在 50~1 000 范围内变动;对于每个确定的决策树数量,均使用 4.2 节中得到的特征数据集建立 100 个 RF 模型,取其识别结果的平均值作为当前 RF 模型的识别准确率。决策树数量不同的 RF 模型识别准确率如图 8 所示。可看出当决策树数量设置为 100 或 150 时,RF 模型识别效果最优,此时对测试集样本的识别准确率达 97%。

考虑到综放工作面煤矸识别的实时性需求,对特征数据集中各特征向量的提取耗时及 RF 模型对测试集样本的识别耗时进行统计。具体方法:从测

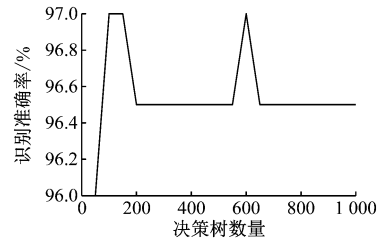


图 8 决策树数量对 RF 模型识别性能的影响

Fig. 8 Influence of decision tree number on RF identification performance

试集样本中随机抽取 10 组数据,统计 10 组数据特征提取的平均耗时;将 10 组数据特征向量输入 RF 模型,计算其平均识别耗时。统计结果见表 2。

表 2 特征提取及识别耗时统计

Table 2 Time consuming statistics of feature extraction and identification

特征	耗时/s		
	特征提取	识别	合计
IMF 能量+IMF 能量熵	0.104 3	0.037 8	0.142 1
IMF 矩阵奇异值+IMF 奇异值熵	0.110 5	0.037 6	0.148 1
IMF 能量+IMF 能量熵+IMF 矩阵奇异值+IMF 奇异值熵	0.132 0	0.050 7	0.182 7

由表 2 可知,使用 IMF 能量及其熵组成的特征向量训练 RF 模型时,模型对测试集样本进行特征提取与识别的耗时为 0.142 1s,使用 IMF 矩阵奇异值及其熵组成的特征向量训练时耗时为 0.148 1s,使用组合特征向量训练时耗时为 0.182 7 s。

5 结论

(1) 在综放现场采集了大量的煤、矸石冲击液压支架尾梁产生的振动信号,对振动信号进行 EMD,采用 IMF 能量、峭度、矩阵奇异值及对应的熵对振动信号进行定量描述,并采用不同的特征向量训练 RF 模型,根据 RF 模型对未知样本的识别能力筛选特征,由此建立特征数据集。根据特征数据集建立 RF 模型,对 200 组测试集样本进行识别,准确率达 96.5%。

(2) 从泛化能力和识别效率 2 个方面研究了 RF 模型的性能,结果表明当决策树数量设置为 100 或 150 时,RF 模型的识别准确率最高,达 97%,同时对测试集样本进行特征提取与识别的耗时不超过 0.2 s。研究结果验证了本文方法可快速、准确地实现煤矸识别。

参考文献 (References):

[1] 于斌,徐刚,黄志增,等. 特厚煤层智能化综放开采理论与关键技术架构[J]. 煤炭学报, 2019, 44(1):

- 42-53.
YU Bin, XU Gang, HUANG Zhizeng, et al. Theory and its key technology framework of intelligentized fully-mechanized caving mining in extremely thick coal seam[J]. Journal of China Coal Society, 2019, 44(1):42-53.
- [2] 张宁波, 鲁岩, 刘长友, 等. 综放开采煤矸自动识别基础研究[J]. 采矿与安全工程学报, 2014, 31(4): 532-536.
ZHANG Ningbo, LU Yan, LIU Changyou, et al. Basic study on automatic detection of coal and gangue in the fully mechanized top coal caving mining[J]. Journal of Mining & Safety Engineering, 2014, 31(4):532-536.
- [3] 张宁波, 刘长友, 陈现辉, 等. 综放煤矸低水平自然射线的涨落规律及测量识别分析[J]. 煤炭学报, 2015, 40(5):988-993.
ZHANG Ningbo, LIU Changyou, CHEN Xianhui, et al. Measurement analysis on the fluctuation characteristics of low level natural radiation from gangue[J]. Journal of China Coal Society, 2015, 40(5):988-993.
- [4] 张守祥, 张艳丽, 王永强, 等. 综采工作面煤矸频谱特征[J]. 煤炭学报, 2007, 32(9):971-974.
ZHANG Shouxian, ZHANG Yanli, WANG Yongqiang, et al. Caved coal and rock spectrum on longwall face[J]. Journal of China Coal Society, 2007, 32(9):971-974.
- [5] 薛光辉, 赵新赢, 柳二猛, 等. 基于振动信号时域特征的综放工作面煤岩识别[J]. 煤炭科学技术, 2015, 43(12):92-97.
XUE Guanghui, ZHAO Xinying, LIU Ermeng, et al. Time-domain characteristic extraction of coal and rock vibration signal in fully-mechanized top coal caving face[J]. Coal Science and Technology, 2015, 43(12): 92-97.
- [6] 李一鸣, 符世琛, 李瑞, 等. 垮落煤岩性状识别研究[J]. 工矿自动化, 2017, 43(2):24-28.
LI Yiming, FU Shichen, LI Rui, et al. Reseach on identification of caving coal and rock traits [J]. Industry and Mine Automation, 2017, 43(2):24-28.
- [7] HUANG N E, SHEN Z, LONG S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceeding of the Royal Society London A, 1998, 454(1971):903-995.
- [8] HUANG N E. A new method for nonlinear and nonstationary time series analysis: empirical mode decomposition and Hilbert spectral analysis [J]. Proceedings of the International Society for Optical Engineering, 2000, 4056:197-209.
- [9] 林选, 田慕琴. 基于小波包和 EMD 的异步电动机轴承故障诊断方法[J]. 工矿自动化, 2010, 36(6): 49-53.
LIN Xuan, TIAN Muqin. Fault diagnosis method of bearing of asynchronous motor based on wavelet-packet and EMD[J]. Industry and Mine Automation, 2010, 36(6):49-53.
- [10] 李康强, 冯志鹏. 基于 EMD 和能量算子的模态参数识别在行星齿轮箱中的应用[J]. 振动与冲击, 2018, 37(8):1-8.
LI Kangqiang, FENG Zhipeng. Modal parameter identification based on empirical mode decomposition and energy operator for planetary gearboxes [J]. Journal of Vibration and Shock, 2018, 37(8):1-8.
- [11] 付晓波. 经验模态分解法理论研究与应[D]. 太原: 太原理工大学, 2013.
FU Xiaobo. Research on empirical mode decomposition theory and its application [D]. Taiyuan: Taiyuan University of Technology, 2013.
- [12] 方保镛, 周继东, 李医民. 矩阵论[M]. 2 版. 北京: 清华大学出版社, 2013.
FANG Baorong, ZHOU Jidong, LI Yimin. Matrix theory[M]. 2nd Edition. Beijing: Tsinghua University Press, 2013.
- [13] 李梅. 信息论基础与应用[M]. 北京: 电子工业出版社, 2016.
LI Mei. Foundation and application of information theory[M]. Beijing: Publishing House of Electronics Industry, 2016.
- [14] HARTE J, NEWMAN E A. Maximum information entropy: a foundation for ecological theory[J]. Trends in Ecology & Evolution, 2014, 29(7):384-389.
- [15] 赵志宏, 杨绍普. 一种基于样本熵的轴承故障诊断方法[J]. 振动与冲击, 2012, 31(6):136-140.
ZHAO Zhihong, YANG Shaopu. Sample entropy-based roller bearing fault diagnosis method [J]. Journal of Vibration and Shock, 2012, 31(6): 136-140.
- [16] LIAW A, WIENER M. Classification and regression by random forest[J]. R News, 2002, 2(3):18-22.
- [17] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3):32-38.
FANG Kuangnan, WU Jianbin, ZHU Jianping, et al. A review of technologies on random forests [J]. Statistics & Information Forum, 2011, 26(3):32-38.
- [18] 曹正凤. 随机森林算法优化研究[D]. 北京: 首都经济贸易大学, 2014.
CAO Zhengfeng. Study on optimization of random forests algorithm [D]. Beijing: Capital University of Economics and Business, 2014.