基于 XGBoost 与可见-近红外光谱的煤矸识别方法

李 瑞[1]，李 博[1*]，王学文[1]，刘 涛[1]，李廉洁[1,2]，樊书祥[2]

1. 太原理工大学机械与运载工程学院，山西 太原 030024
2. 北京农业智能装备技术研究中心，北京 100097

摘 要 煤矸智能识别是实现综放开采智能化亟待研发的新技术；可见-近红外光谱技术具有环保、实时等优势，满足煤矸智能分选的要求。为解决基于可见-近红外光谱的煤矸识别问题，引入在数据科学竞赛中表现出色的极端梯度提升树(XGBoost)算法。搭建可见-近红外光谱实验平台采集来自山西西铭、陕西神木、内蒙古巴隆图煤矿的块状煤与矸石样品在370～1 049 nm波段的反射光谱；利用黑白校正、始末波段去除、SG卷积平滑和标准正态变量变换(SNV)对采集的原始光谱进行预处理，以减少光照不均、噪声以及光程差的影响。依据三个煤矿煤与矸石样品反射光谱的差异划分实验组和测试组，实验组差异微小，用于对比不同模型的性能，挑选最佳算法；测试组差异较明显，用于测试最佳算法在其他煤矿下的表现，检验算法对不同煤矿的适用性。在实验组的实验中，首先基于XGBoost算法建立煤与矸石分类模型，并引入常用的机器学习分类算法k近邻法(KNN)、随机森林(RF)、支持向量机(SVM)做对比，结果表明XGBoost的表现最佳，十折交叉验证的平均准确度(ACC_{10})、分类准确度(ACC)与AUC值分别达到0.957 2，0.970 5与0.971 6，体现出较强的稳定性与分类能力。其次为降低数据维度减少模型运算量，使用递归特征选择(RFE)、连续投影算法(SPA)与竞争性自适应重加权算法(CARS)分别进行特征波长的选择并与上述四种分类算法结合构建简化分类模型，经测试RFE与XGBoost组合的简化模型表现最佳，ACC_{10}，ACC与AUC值分别为0.965 7，0.980 3与0.980 3且数据维度降至9，在降低数据维度的同时提高了模型的稳定性与分类能力。在测试组的实验中，基于优选出的XGBoost与RFE-XGB算法建立的模型，同样可以实现对其他矿区煤与矸石稳定精确地识别，且简化模型表现更好，与实验组结果一致。

关键词 XGBoost；可见-近红外光谱；煤矸石分选；黑色背景；无损检测

中图分类号：TD94 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)09-2947-09

引 言

煤炭在我国的能源结构中占据主体地位；综放开采是开采大储量特厚煤层的主要方法，而煤矸智能识别是实现综放开采智能化亟待研发的新技术[1]。目前的识别方法有射线法[2]与图像识别法[3]，但分别存在放射源安全隐患与易受粉尘、光照等环境条件影响的不足。可见-近红外光谱技术具有实时、无污染、高信噪比、仪器成本低等诸多优势，满足煤矸分选的要求。基于可见-近红外光谱的煤的品质测定[4]、煤种分类[5]、煤系岩石识别[6]等领域的研究为煤矸识别提供了理论基础。在此基础上，学者们提出了多种基于可见-近红外光谱的煤与矸石的分类方法。杨恩等[7]建立了GRB-KPCA-SVM模型实现了对烟煤与碳质页岩(块状与粉末)的识别；Mao[8]等利用IAM-ELM算法实现了室外环境下煤与矸石的分类；Xiao[9]等建立了ICELM-LRF模型实现了对不同煤种以及煤与矸石的分类；Hu[10]等基于LBP算法对样本多光谱图像进行特征提取，并采用GS-SVM分类器实现对煤与矸石的分类。在目前的煤矸识别研究中，样品大多仅采自一个煤矿，针对该煤矿样品设计出的算法对其他矿区的适用性有待检验；部分研究虽使用了不同煤矿的煤与矸石样品，但样品预处理，如清洁、研磨等，脱离了实际应用背景。此外大多研究采集样本光谱时使用的背景与实际背景也存在较大差异。

Chen[11]等提出的极端梯度提升树(XGBoost)算法，是一种Boosting算法，引入正则化项且支持并行运算，具有快

收稿日期：2021-10-19，**修订日期**：2022-04-04

基金项目：国家自然科学基金项目(51804207，51875386)，山西省“1331”工程项目资助

作者简介：李 瑞，1998年生，太原理工大学机械与运载工程学院硕士研究生 e-mail：lirui0063@link.tyut.edu.cn

*通讯作者 e-mail：libo@tyut.edu.cn

速、准确、可解释等特点，模型表现较 BP 神经网络与 SVM 等有更高的准确率[12]，广泛应用在医学、遥感、电气、电子商务、故障诊断等领域。但在可见-近红外光谱分析领域应用较少，尤其是煤矸识别方向还未见研究报道。

本工作进行了以下研究：(1)基于 XGBoost 算法建立了块状煤与矸石黑色背景下的可见-近红外光谱分类模型，并与常用机器学习分类模型进行比较；(2)基于多种特征选择算法筛选的特征波长建立 XGBoost 以及对比算法的简化模型，并选出最优简化模型；(3)在相同实验条件下更换其他煤矿的煤与矸石作为样品对 XGBoost 算法以及选出的最优特征选择与分类的组合简化算法进行煤矿适用性的检验。

1 XGBoost 算法

极端梯度提升算法(XGBoost)，为梯度提升决策树(GBDT)的一种实现方法，以分类与回归树(CART)为子模型，通过梯度提升实现多个 CART 子模型的集成学习[13-14]。XGBoost 对目标函数进行二阶泰勒展开，一阶导数与二阶导数同时参与运算，提升了模型优化过程中的收敛速度。算法的步骤如下

目标函数

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

正则项

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

式(1)和式(2)中，y_i 为第 i 个样本的真实值，$\hat{y}_i$ 为第 i 个样本的预测值，f_k 为第 k 棵决策树，$l(\cdot)$为损失函数，$\Omega(\cdot)$为正则项，T 为当前决策树下叶子节点个数，w_j 为第 j 个叶子节点的权重，γ 和 λ 分别为叶子数目与叶子权重的正则化系数。

根据贪心算法，每次建立新的弱学习器时都以使目标函数降低最大为目标，第 k 棵树的目标函数为

$$\mathrm{Obj}^{(k)}=\sum_{i=1}^{N}l(y_i,\hat{y}_i^{k-1}+f_k(x_i))+\Omega(f_k)+C \tag{3}$$

式(3)中，$f_k(x_i)$为第 i 个样本在第 k 棵树下的输出分数，$C=\sum_{k=1}^{K-1}\Omega(f_k)$ 即前 $K-1$ 棵树的正则项和，k 轮迭代时为常数项。

对 l 在 $\hat{y}_i^{k-1}$ 处进行二阶泰勒展开，展开后的一阶导数与二阶导数以 g_i 与 h_i 表示。第 j 个叶子节点的样本集合为 I_j，定义 $G_j=\sum_{i\in I_j}g_i$，$H_j=\sum_{i\in I_j}h_i$，并去掉常数项后的目标函数为

$$\mathrm{Obj}^{(k)}=\sum_{j=1}^{T}\Big(\sum_{i\in I_j}g_iw_j+\frac{1}{2}\sum_{i\in I_j}h_iw_j^2\Big)+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{4}$$

$$\mathrm{Obj}^{(k)}=\sum_{j=1}^{T}\Big(G_jw_j+\frac{1}{2}(H_j+\lambda)w_j^2\Big)+\gamma T \tag{5}$$

化简后的目标函数对 w_j 求一阶导并令导数为 0，求得 w_j 的最优解为

$$w_j^*=-\frac{G_j}{H_j+\lambda} \tag{6}$$

目标函数的最优解为

$$\mathrm{Obj}^{(s)}=-\frac{1}{2}\sum_{j=1}^{T}\Big(\frac{G_j^2}{H_j+\lambda}\Big)+\gamma T \tag{7}$$

XGBoost 分裂节点时遍历所有特征，以带来最大分裂增益 Gain 的点作为分裂节点，定义某特征在整个模型构建中作为分裂结点的次数为 weight，该特征作为分裂节点的平均增益为 gain。

$$\mathrm{Gain}=\frac{1}{2}\Big[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\Big]-\gamma \tag{8}$$

$$\mathrm{gain}=\frac{\sum\mathrm{Gain}}{\mathrm{weight}} \tag{9}$$

2 实验部分

2.1 样本与光谱数据采集

在我国煤炭的主要产地山西省、陕西省与内蒙古自治区，分别采集部分煤与矸石样品。如表 1 所示，依据产地对样品进行编号与信息统计。如图 1 所示，不同产地的样品形状、颜色存在较大差异。

表 1 样品信息

Table 1 Samples information

产地及煤矿	样品类别及编号	外观特征	样品数量	采集光谱数量
山西西铭(Ⅰ)	煤(Ⅰ.1)	黑色，有光泽	93	169
	岩(Ⅰ.2)	黑色，无光泽	91	169
陕西神木(Ⅱ)	煤(Ⅱ.1)	黑色，有光泽	20	39
	岩(Ⅱ.2)	黑褐色，无光泽	18	33
内蒙古巴隆图(Ⅲ)	煤(Ⅲ.1)	黑褐色，无光泽	26	49
	岩(Ⅲ.2)	灰白色，无光泽	23	43

图 2 为搭建的可见-近红外光谱采集系统。使用 150 W 卤素灯(JCR15V150WBAU，USHIO，USA)作为光源，置于光纤探头一侧，距背景平面垂直距离 500 mm，入射角度调整至与垂直面夹角为 30°。光谱检测光纤(1SMA1S-SI0.6-1500S，Nanjing Shen Lue Technology Co.，Ltd，China)的反射光纤探头由干板夹(GCM-1311M，Daheng Optics，China)夹持，可通过支座(GCM-030304M，Daheng Optics，China)与支杆夹(GCM-55，Daheng Optics，China)调节探头位置与角度，经调试置于检测样本平面正上方 200 mm 处。可见-近红外光谱仪(USB2000＋，Ocean Insight，USA)与检测光纤连接，使用光谱仪配套软件(OceanView2.0.7，Ocean Insight，USA)采集反射光谱，采集的波长范围为 370～1 049 nm，波长数为 2 048 个。采集平台以黑色纸板作为采集背景，原始块状样品作为采集样品，模拟选煤环境，并以黑布与外界隔开，避免外界杂散光影响。

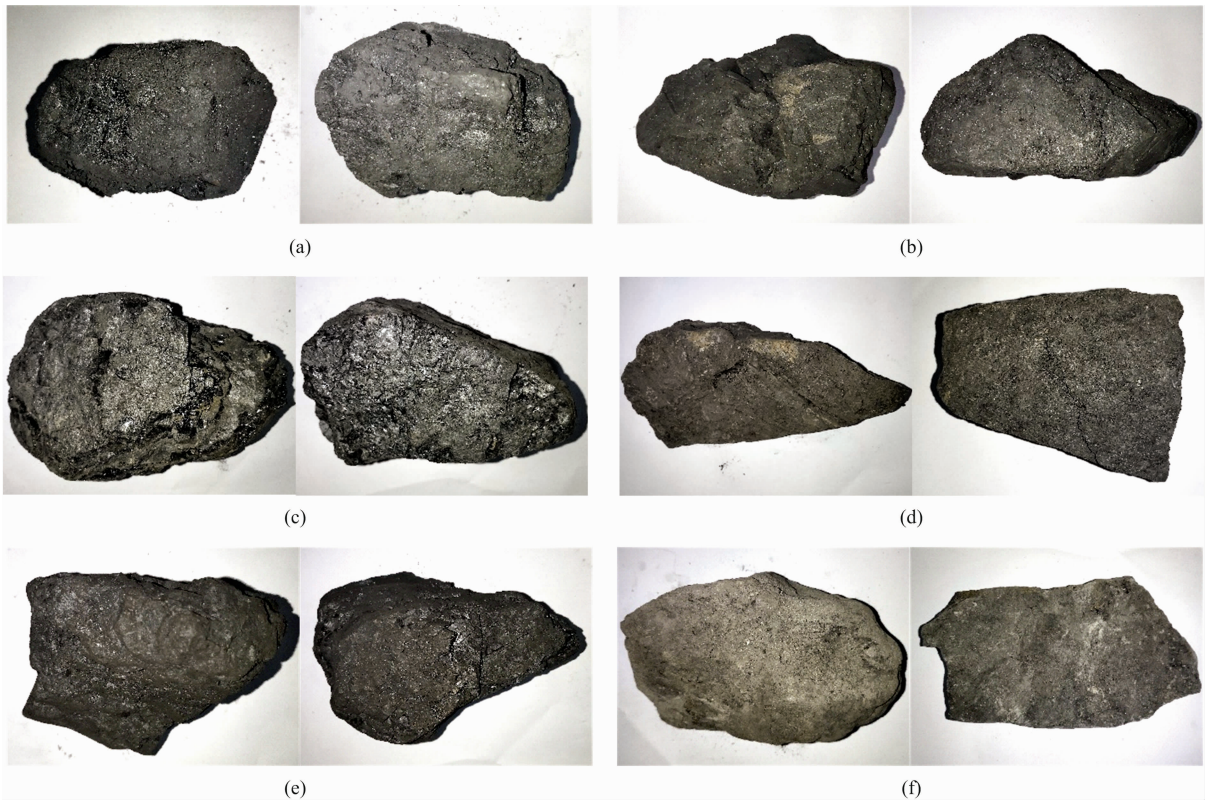

图 1　不同煤矿的煤(a，c，e)与矸石(b，d，f)样品

(a)，(b)：西铭煤矿；(c)，(d)：神木煤矿；(e)，(f)：巴隆图煤矿

Fig. 1　Coal samples (a，c，e) and gangue samples (b，d，f) from different coal mines

(a)，(b)：Ximing coal mine；(c)，(d)：Shenmu coal mine；(e)，(f)：Balongtu coal mine

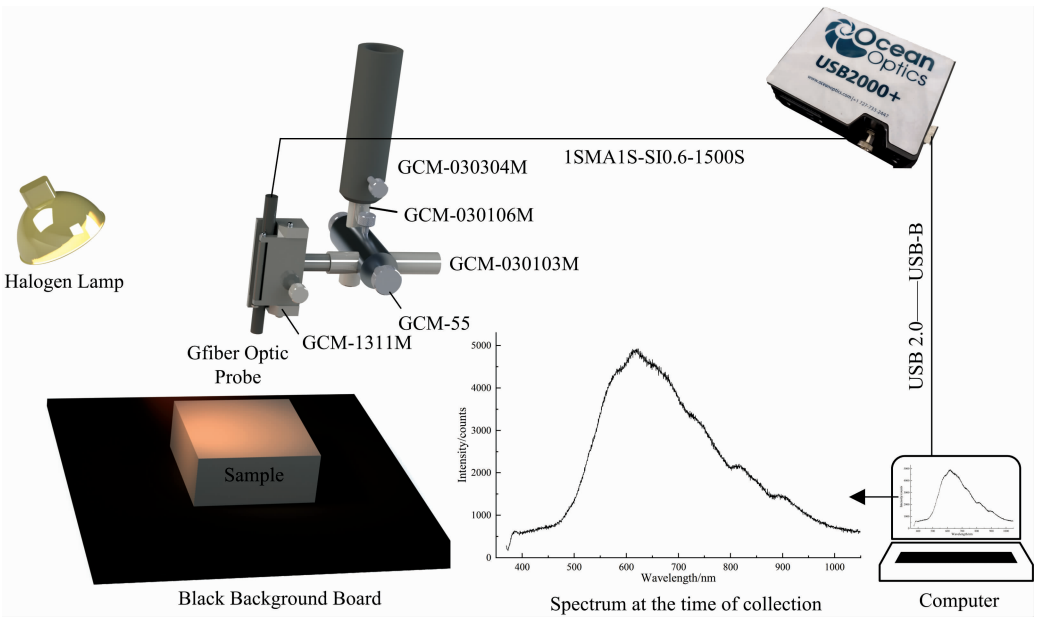

图 2　可见-近红外光谱采集系统

Fig. 2　Visible and near-infrared spectroscopy collection system

采集数据前，将光谱仪和光源预热 15～20 min 待光谱曲线稳定后再进行采集。为避免光谱饱和，采集时积分时间设置为 9 ms，同时利用平均值采样法以连续三次采集取平均的光谱数据作为最终数据，以提高数据的准确性。在室内相

同环境下采集样品的反射光强 L，每个样品通过翻转采集面采集 1～2 条光谱，并在每次连续采集前利用聚四氟乙烯为材料的标准白板采集白参考 L_{white}，关闭光源盖住光纤探头采集暗参考 L_{black}。

2.2 光谱预处理

原始光谱数据因受仪器、采集环境等影响含有噪声与冗余信息，增加数据维度的同时，掩盖了关键信息。因此进行光谱预处理，去除噪声及冗余信息，提取特征信息。根据实验原数据的特点，使用以下方法进行预处理。

(1)黑白校正与波段挑选

为减小光照不均与暗噪声的影响，参照式(10)对原始样品数据进行黑白校正以求得相对反射率 R，并以去除含有大量随机噪声的始末波段(369.51～517.69，859.37～1 049.07 nm)后的反射光谱波段(518.00～857.05 nm)作为后续预处理数据，特征维度由 2 048 降至 1 000。

$$R = \frac{L - L_{\text{black}}}{L_{\text{white}} - L_{\text{black}}} \tag{10}$$

(2)Savitzky-Golay(SG)卷积平滑

使用 Savitzky-Golay(SG)卷积平滑法对反射光谱进行去噪，选择一次多项式拟合，窗口设为 29，即式(11)中 $w=14$。

$$R_{k,\ \text{smooth}} = \frac{1}{\sum_{i=-w}^{+w} h_i} \sum_{i=-w}^{+w} R_{k+i} h_i \tag{11}$$

式(4)中，k 为当前波长，h_i 为平滑系数，$2w+1$ 为窗口内的波长数。

(3)标准正态变量变换(SNV)

由于煤矸样品高度在 40～95 mm 将引起光程变化，使用标准正态变量变换(SNV)来消除煤矸样品高度变化对反射光谱的影响。实现公式为式(13)，使用前将反射光谱单位按式(12)进行转换。

$$A = \lg(1/R) \tag{12}$$

$$R_{\text{SNV}} = \frac{A - \overline{A}}{\sqrt{\dfrac{\sum_{k=1}^{m} (A_k - \overline{A})^2}{m-1}}} \tag{13}$$

式(13)中，m 为波长数，本文中为 1 000，$\overline{A} = \dfrac{\sum_{i=1}^{m} A_i}{m}$。

在得到预处理后的光谱曲线后，根据同一煤矿煤与矸石反射光谱差异的显著性，将样品划分为实验组与测试组。实验组煤与矸石的反射光谱差异微小，用于比较全波段与特征波长光谱下不同模型的表现，选出最佳的算法；测试组煤与矸石的反射光谱差异较实验组明显，用于对选出的最佳算法进行矿区适应性检验。

2.3 XGBoost 建模与模型评价指标

使用 python 3.8.3 语言以及 XGBoost，Scikit-Learn 和 Numpy 等开源工具包，在 Jupyter notebook 编译器中进行建模。训练集与测试集以 7∶3 的比例随机划分，通过学习曲线与网格搜索对模型超参数进行寻优，同时通过训练集十折交叉验证平均分类准确度 ACC_{10} 对模型超参数进行调整以避免过拟合，以期提高测试集的分类准确度 ACC。为更好的评价模型性能，加入 AUC (area under the curve)值为评价指标，通过上述三个指标可以综合体现模型的稳定性与分类能力。

为更好地评价 XGBoost 模型对可见光-近红外波段煤矸石分类的性能，引入 k 近邻法(KNN)、随机森林(RF)、支持向量机(SVM)三种分类算法做对比。

2.4 特征选择算法与简化模型寻优

光谱数据虽经剔除噪声波段预处理，降低了维度，但仍为含有冗余信息的多维数据。因此利用特征选择对光谱数据进行再次降维，以降低模型计算量，缩短模型运行时间，为进一步在线识别研究奠定基础。选用递归特征选择算法(RFE)、连续投影算法(SPA)和竞争性自适应重加权算法(CARS)进行特征选择，并基于特征选择后的波长数据建立简化分类模型，通过评估模型寻找出最优的特征波长选择与机器学习分类算法的组合。

(1)递归特征选择(RFE)

递归特征选择是一种基于模型特征重要性的特征选择算法，在每轮迭代时依据特征重要性排名消除若干末尾特征，并将包含保留特征的数据集作为下一轮的训练样本，直至将特征降低到一定维度。XGBoost 模型可以通过特征重要性度量指标对特征重要性进行排序，本工作以式(9)中 gian 的归一化权重 score 为度量指标作为特征重要性排序依据，并通过迭代模型的五折交叉验证均方根误差(RMSECV)进行最佳特征维度的选择。

(2)连续投影算法(SPA)

连续投影算法的选择原理为，从选入某一波长数据开始，循环计算新选入的特征波长在未选入特征的投影，将投影向量最大的波长选出，直至选出的波长数量达到设定值。为得到最佳的起始波长与选出波长数量，利用多元线性回归分析(MLR)对划分后的数据集进行建模，以测试集均方根误差(RMSE)为模型评价指标，选出最佳特征波长。

(3)竞争性自适应重加权算法(CARS)

竞争性自适应重加权算法在每次循环中通过蒙特卡洛采样法从训练集中选出一定比例样本建立偏最小二乘(PLS)回归模型。以指数衰减的方式去除回归系数较小的特征，并在达到蒙特卡洛采样次数时停止，以交叉验证均方根误差最小的特征集为最佳特征波长集。设定采样次数为 50，采样比例为 0.8。

2.5 最佳算法的矿区适用性检验

不同产地的煤与矸石物理与化学性质存在差异，光谱曲线也存在差异，选出的最佳算法在单一煤矿样品分类中的表现存在局限性，因此有必要对实验组选出的最佳算法进行煤矿适用性检验。通过使用选出的最优算法建立测试组全波段以及特征波长光谱的分类模型，测试模型的性能来检验算法的煤矿适用性。

3 结果与讨论

3.1 反射光谱的分组

原始光谱在经黑白校正、波段挑选、SG 卷积平滑与 SNV 预处理后，降低了环境条件的影响，并去除了冗余信

息，将特征维度由 2 048 降至 1 000。图 3 为预处理后的三个煤矿的光谱图像，反映出了以下信息：

Ⅰ煤矿煤与矸石的反射光谱曲线整体较相似，均在 619 nm 附近出现吸收峰，750 nm 附近出现吸收谷，但在 750～857 nm 波段煤与矸石曲线斜率差异明显。

Ⅱ煤矿煤与矸石反射光谱曲线差异明显，煤的光谱曲线在 625 和 820 nm 附近出现吸收峰；矸石的光谱曲线在 680 nm 附近产生吸收峰，在 730 nm 附近出现吸收谷，另外有部分样品在 825 nm 附近出现吸收峰。

Ⅲ煤矿煤与矸石反射光谱曲线差异同样明显，煤的光谱曲线在 680 和 820 nm 附近出现吸收峰；矸石的光谱曲线在 680 nm 附近出现吸收峰，736 nm 附近出现吸收谷。

依据三个煤矿煤与矸石样品反射光谱曲线差异的显著性，将差异微小的Ⅰ煤矿样品作为实验组，用于对比不同模型的性能，挑选最佳算法；将差异较明显的Ⅱ与Ⅲ煤矿样品作为测试组，用于测试选出的最佳算法的模型性能，检验算法对不同煤矿的适用性。

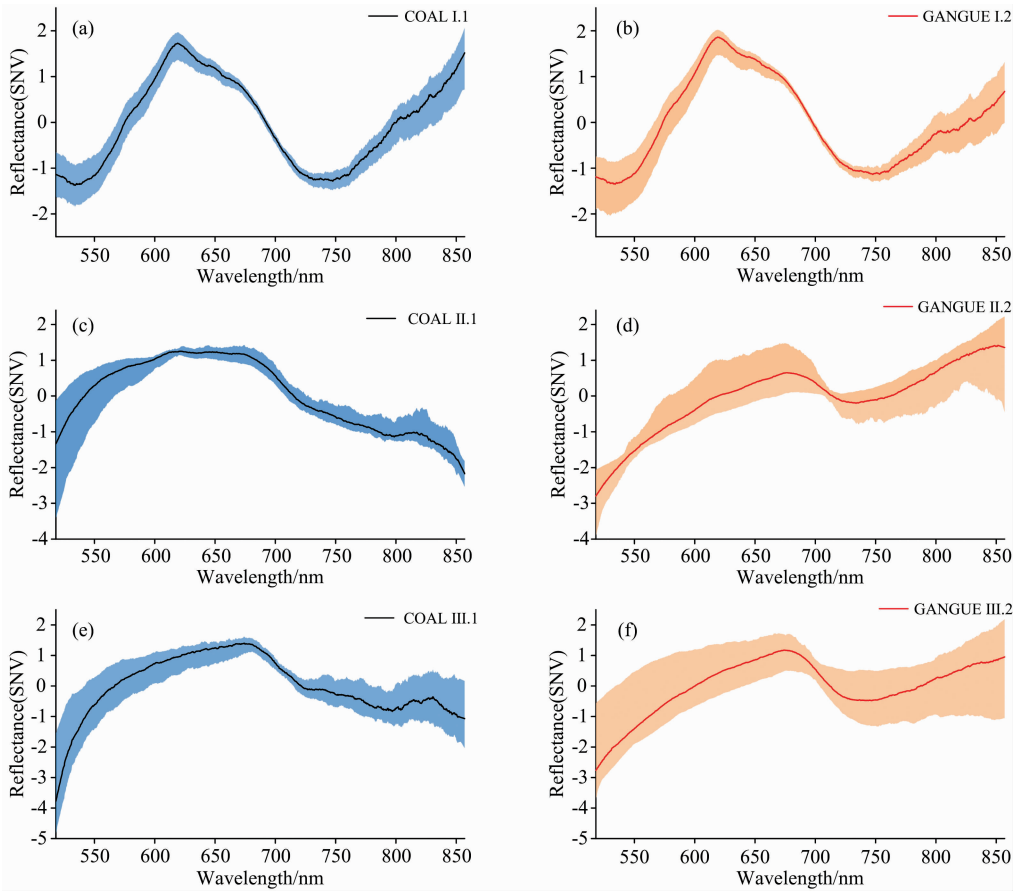

图 3　Ⅰ：(a)(b)，Ⅱ：(c)(d)，Ⅲ：(e)(f)煤矿煤与矸石预处理后的光谱

Fig. 3　Spectra of coal and gangue in Ⅰ：(a)(b)，Ⅱ：(c)(d)，Ⅲ：(e)(f) mines after pretreatment

3.2　基于全波段光谱的煤与矸石 XGBoost 分类模型

表 2 为基于实验组全波段光谱的煤与矸石 XGBoost 分类模型以及 k 近邻法(KNN)、随机森林(RF)、支持向量机(SVM)三种对比算法的模型分类表现，XGBoost 表现最佳，ACC_{10}，ACC，AUC 分别达到 0.957 2，0.970 5 和 0.971 6，较其他模型有更强的分类能力，可以稳定且准确地完成煤与矸石的分类。

3.3　基于特征波长的煤与矸石分类模型

为降低光谱数据的维度，缩短模型训练与预测的时间，分别使用 RFE，SPA 和 CARS 算法对预处理后的光谱数据进行特征选择。

RFE 的筛选过程如图 4 所示，特征数量为 9 时，RMSECV 达到最小，挑选出的特征波长为 673.15，689.55，692.27，698.72，699.06，699.74，700.41，704.14 和 707.18 nm。

表 2　基于全波段光谱的不同分类模型对比

Table 2　Comparison of different classification models based on the full-band spectra

Sample origin	Number of variables	Model	ACC_{10}	ACC	AUC
Ⅰ	1 000	KNN	0.948 5	0.941 1	0.941 5
		RF	0.953 0	0.960 7	0.961 1
		SVM	0.944 3	0.960 7	0.961 1
		XGBoost	0.957 2	0.970 5	0.971 6

SPA 的筛选过程如图 5 所示，最佳波长数为 5，即：518.00，588.17，671.78，690.56 和 718.30 nm。

CARS 的筛选过程如图 6 所示，在第 23 次采样时，RMSECV 值最小，此时的变量子集为最佳特征波长集，含有 61 个波长。

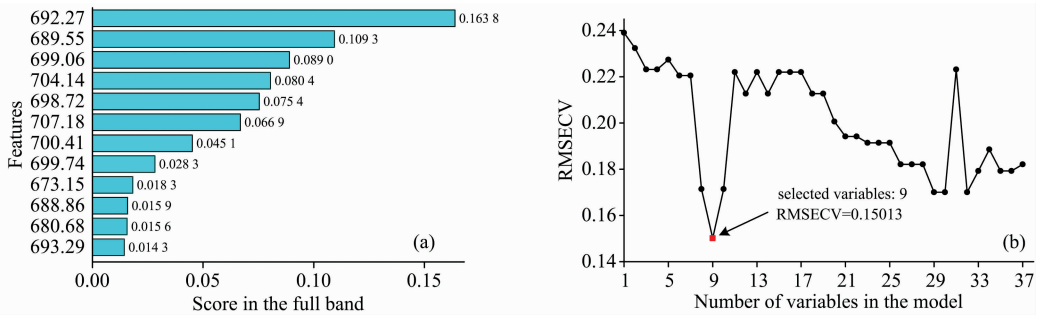

图 4 RFE 波长筛选过程

Fig. 4 The process of variable selection by RFE

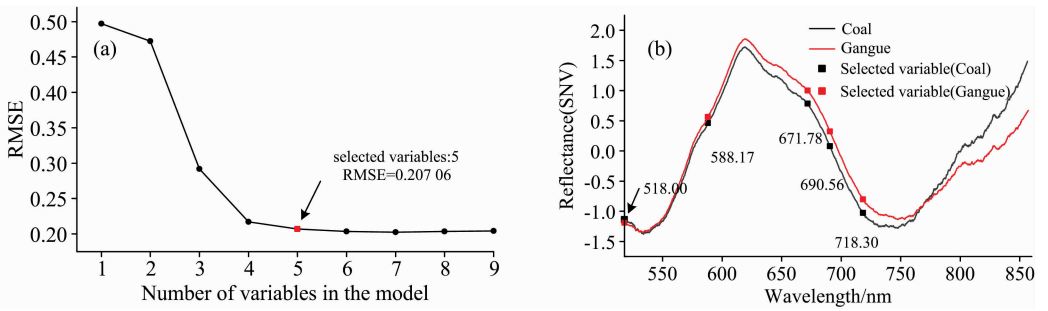

图 5 SPA 波长筛选过程

Fig. 5 The process of variable selection by SPA

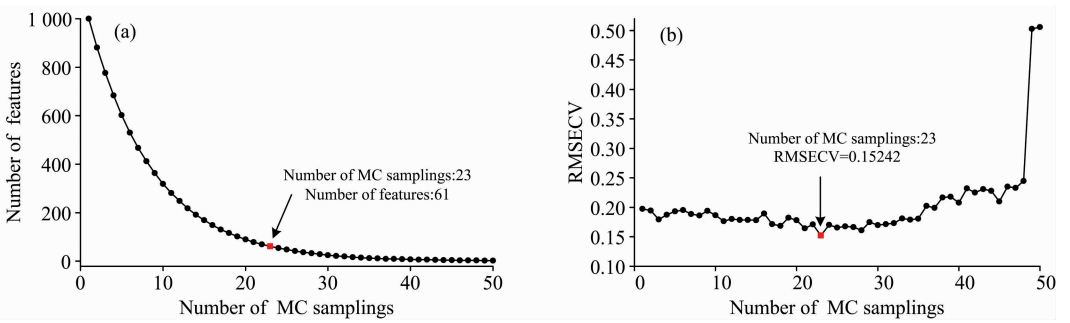

图 6 CARS 波长筛选过程

Fig. 6 The process of variable selection by CARS

将上述三种方法选出的特征波长与四种分类算法两两组合建立基于特征波长的煤与矸石分类模型，经测试模型表现如表 3 所示。

首先对比特征波长与全波段光谱数据建立的模型分类表现，基于特征波长的分类模型大多提高了模型的稳定性，并且与测试集下的分类准确度相近，并有部分模型准确度与 AUC 值有所提升。可见上述三种特征选择方法不仅有效地剔除了数据中的冗余信息，而且保留了关键信息，在降低数据复杂度的同时，有利于模型更稳定地分类。

分别对比四种模型用相同特征选择算法的模型分类表现可知，对 RFE 算法，XGBoost 模型较其他三种算法优势明显；对 SPA 算法，XGBoost，RF 和 SVM 分类表现较 KNN 更好；对 CARS 算法，SVM 分类效果最优，XGBoost 次之。单从模型表现来看，表现最好的是 CARS-SVM，其次为仅 ACC_{10} 上略差的 RFE-XGB。但由于 CARS 选出的特征波长数量为 61，相比 RFE 选出的 9 维特征，增加了采集、预处理以及模型训练时的数据复杂度与处理时间，对于实际应用中要求快速判别的煤与矸石的在线分选，RFE-XGB 更合适。因此以 RFE-XGB 作为最佳的基于特征波长光谱的分类算法。

表 3　基于特征波长光谱的不同分类模型预测结果

Table 3　The prediction results of different classification models based on characteristic wavelengths

Sample origin	Variable selection methods	Number of variables	Model	ACC_{10}	ACC	AUC
Ⅰ	RFE	9	KNN	0.948 3	0.950 9	0.950 9
			RF	0.948 3	0.950 9	0.952 0
			SVM	0.956 8	0.960 7	0.952 0
			XGBoost	0.965 7	0.980 3	0.980 3
	SPA	5	KNN	0.953 0	0.950 9	0.951 2
			RF	0.953 0	0.960 7	0.960 7
			SVM	0.965 7	0.960 7	0.961 1
			XGBoost	0.957 2	0.960 7	0.960 7
	CARS	61	KNN	0.957 0	0.950 9	0.951 2
			RF	0.948 9	0.950 9	0.950 9
			SVM	0.978 6	0.980 3	0.980 3
			XGBoost	0.961 5	0.960 7	0.960 7

3.4　最佳算法的矿区适用性检验

为验证在实验组——西铭煤矿煤与矸石样品全波段和特征波长光谱下选出的表现最好的 XGBoost 和 RFE-XGB 算法是否适用于其他矿区，分别建立基于测试组——神木与巴隆图煤矿煤与矸石样品反射光谱数据的 XGBoost 和 RFE-XGB 模型。

RFE 的波长数与 RMSECV 如图 7，神木与巴隆图煤矿的最佳特征波长数分别为 3 个与 7 个，挑选出的特征波长分别为 528.12，534.25，843.46 nm 与 706.84，707.18，707.51，709.54，711.90，839.98，851.70 nm。

模型表现如表 4 所示，对于神木与巴隆图煤矿基于全波段数据建立的 XGBoost 模型表现同样很好，准确率分别达到 1 和 0.964 2，AUC 值达到 1 和 0.9687，在测试集中表现出较佳的分类能力，且 ACC_{10} 分别达到 0.955 0 与 0.940 4，模型也具有较好的稳定性。基于两矿区样品建立 RFE-XGB 模型较全波段 XGBoost 模型均显著提高了模型的稳定性，并且巴隆图煤矿模型的准确率与 AUC 值有了明显提升，分类能力更强。综上所述，XGBoost 与 RFE-XGB 算法适用于其他矿区的煤与矸石样品，建立的模型与西铭煤矿样品的模型表现一致，都具有较强的分类能力与稳定性，并且 RFE-XGB 模型表现更好。

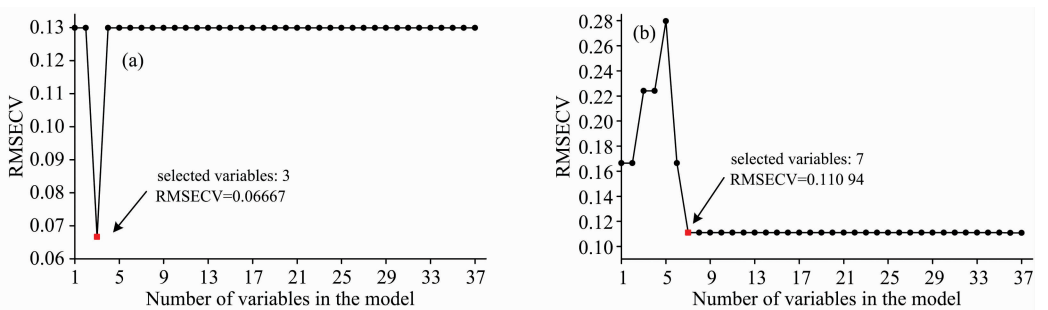

图 7　测试组的 RFE 波长筛选过程

Fig. 7　The variables selection process of the test group by RFE

表 4　测试组的模型预测结果

Table 4　The model prediction results of the test group

Model	Sample origin	Variable selection methods	Number of variables	ACC_{10}	ACC	AUC
XGBoost	Ⅱ	RFE	1 000	0.955 0	1.000 0	1.000 0
			3	0.975 0	1.000 0	1.000 0
	Ⅲ	RFE	1 000	0.940 4	0.964 2	0.968 7
			7	0.954 7	1.000 0	1.000 0

4 结 论

针对同一煤矿黑色背景下不同高度的块状煤与矸石的分类问题，结合预处理方法——黑白校正、始末波段去除、SG卷积平滑与标准正态变量变换，建立了基于XGBoost算法的煤与矸石可见-近红外光谱分类模型，得到以下结论：

(1)基于全波段光谱数据的XGBoost模型较KNN，RF和SVM三种常用的分类模型表现更好，在实验组下ACC_{10}，ACC，AUC分别为0.957 2，0.970 5和0.971 6，具有更强的稳定性与分类能力，可以实现对煤与矸石稳定而准确地分类。

(2)基于特征波长光谱的模型较全波段，在显著降低特征维度的同时，提升了模型的性能，尤其是模型的稳定性。RFE-XGB模型较其他特征选择方法与分类模型的组合表现最优，在实验组下的特征维度、ACC_{10}，ACC，AUC分别为9，0.965 7，0.980 3，0.980 3，不仅显著降低了数据复杂度，而且提升了模型性能，更适合于实际煤矸分选短时高效的要求。

(3)基于神木与巴隆图煤矿煤与矸石建立的XGBoost模型与RFE-XGB模型均与西铭煤矿表现一致，XGBoost模型实现了对煤与矸石较精确的分类，RFE-XGB模型可实现特征降维并提升模型的稳定性。XGBoost与RFE-XGB算法对于不同煤矿的煤与矸石分类具有良好的适用性，为基于可见-近红外光谱进行煤和矸石分类提供了参考方法。

References

[1] YU Bin, XU Gang, HUANG Zhi-zeng, et al(于 斌，徐 刚，黄志增，等). Journal of China Coal Society(煤炭学报)，2019，44(1)：42.

[2] Zhang N, Liu C. Scientific Reports，2018，8(1)：190.

[3] LI Bo, WANG Xue-wen, GAO Xin-yu, et al(李 博，王学文，高新宇，等). China Powder Science and Technology(中国粉体技术)，2021，27(4)：77.

[4] Wang S H, Zhao Y, Hu R, et al. Chinese Journal of Analytical Chemistry，2019，47 (4)：E19034.

[5] SONG Liang, LIU Shan-jun, MAO Ya-chun, et al(宋 亮，刘善军，毛亚纯，等). Journal of Northeastern University·Natural Science(东北大学学报·自然科学版)，2017，38(10)：1473.

[6] YANG En, WANG Shi-bo, GE Shi-rong, et al(杨 恩，王世博，葛世荣，等). Industry and Mine Automation(工矿自动化)，2017，45(3)：46.

[7] Yang E, Ge S, Wang S, et al. Journal of Spectroscopy，2018，2018：2754908.

[8] Mao Y, Le B T, Xiao D, et al. Optics and Laser Technology，2019，114：10.

[9] Xiao D, Li H, Sun X. ACS Omega，2020，5(40)：25772.

[10] Hu F, Zhou M, Yan P, et al. IEEE Access，2019，7：169697.

[11] Chen T, Guestrin C. XGBoost：A Scalable Tree Boosting System. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2016，785.

[12] HUANG Qing, XIE He-liang(黄 卿，谢合亮). Mathematics in Practice and Theory(数学的实践与认识)，2018，48(8)：297.

[13] TAO Meng-qi, LIU Jia-xiang, WU Yue, et al(陶孟琪，刘家祥，吴 越，等). Acta Optica Sinica(光学学报)，2020，40 (7)：0730002.

[14] JI Hui-jie, NI Feng, LIU Jiang, et al(冀慧杰，倪 枫，刘 姜，等). Computer Technology and Development(计算机技术与发展)，2021，31(5)：21.

A Classification Method of Coal and Gangue Based on XGBoost and Visible-Near Infrared Spectroscopy

LI Rui[1], LI Bo[1*], WANG Xue-wen[1], LIU Tao[1], LI Lian-jie[1,2], FAN Shu-xiang[2]

1. College of Mechanical and Vehicle Engineering, Taiyuan University of Technology, Taiyuan 030024, China
2. Beijing Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China

Abstract Intelligent recognition of coal and gangue is a new technology that needs to be developed urgently to realize the intelligentization of fully mechanized caving mining. Visible-near infrared spectroscopy technology has many advantages such as environmental friendly and real-time, which meets the requirements of intelligent separation of coal and gangue. The Extreme Gradient Boosting Tree (XGBoost) algorithm which performed well in data science competitions, was introducedto achieve the recognition of coal and gangue based on visible-near infrared spectroscopy. Firstly, a visible-near infrared spectroscopy experimental platform was built to collect the reflectance spectra of lump coal and gangue samples from Shanxi Ximing, Shaanxi Shenmu, and Inner Mongolia Balongtu coal mines in the range of 370～1 049 nm. The collected original spectra were preprocessed through black and white correction, method of removing the start and end bands, Savitzky-Golay (SG) smoothing and standard normal variable transformation (SNV) to reduce the effects of uneven illumination, noise and optical path difference. Secondly, the experimental group and test group were divided according to the difference of reflection spectrum of samples from different mines. The experimental group had a minor difference, which was used to compare the performance of different models and select the best algorithm; the difference of test groups was obvious, which was used to test the performance of the best algorithm in other coal mines and verified the applicability of the algorithm to different coal mines. In the experiment of the experimental group, the coal and gangue classification model was established based on the XGBoost algorithm, and the commonly used machine learning classification algorithms k-nearest neighbor method (KNN), random forest (RF), support vector machine (SVM), which were introduced for comparison. The results showed XGBoost performed best. The average accuracy of 10-fold cross-validation (ACC_{10}), classification accuracy (ACC), and AUC values respectively reached 0.957 2, 0.970 5, and 0.971 6, showing strong stability and classification capabilities. Then in order to reduce the data dimension and calculations, recursive feature elimination (RFE), successive projections algorithm (SPA) and competitive adaptive reweighted sampling (CARS) were used to select the characteristic wavelength and combined with the above four classification algorithms to construct a simplified classification model, respectively. The simplified model of the combination of RFE and XGBoost(RFE-XGB) performed best in the test. The ACC_{10}, ACC, AUC was 0.965 7, 0.980 3, 0.980 3, respectively, and the data dimension reduced to 9. Simplified model improved the stability and classification ability of the model while reducing the data dimension. In the experiment of the test group, the model based on XGBoost and RFE-XGB algorithms can also achieve stable and accurate recognition of coal and gangue in other coal mines, and the simplified model performed better, which was consistent with the results of the experimental group.

Keywords XGBoost; Visible and near-infrared; Coal and gangue separation; Black background; Nondestructive detection

(Received Oct. 19, 2021; accepted Apr. 4, 2022)

* Corresponding author