激光与光电子学进展 先进成像

基于GA-XGBoost模型的GF-5卫星影像土壤重金属含量反演研究

柏晗[1]，杨耘[1,2*]，崔琴芳[3]，贾鹏[4]，王丽霞[1]

[1]长安大学地质工程与测绘学院，陕西 西安 710054；

[2]自然资源部退化及未利用土地整治工程重点实验室，陕西 西安 710016；

[3]苍穹数码技术股份有限公司，陕西 西安 710001；

[4]长庆工程设计有限公司，陕西 西安 710018

摘要 随着高光谱成像技术的发展，利用国产高光谱影像进行大范围土壤参数反演成为了可能，但其反演精度仍有待提高。因此，以陕西大西沟矿区为例，以GF-5高光谱卫星影像以及实测的土壤样本数据为数据源，提出了一种基于遗传算法特征选择的XGBoost土壤铜元素反演模型(GA-XGBoost)。首先，对预处理后的影像数据进行连续统去除等光谱变换，并利用蒙特卡罗交叉验证法(MCCV)剔除异常土壤样本；最后，分别建立基于相关系数与遗传算法特征选择的XGBoost重金属含量反演模型。实验结果表明，相同光谱变换条件下，与基于相关系数特征选择的XGBoost模型相比，所提GA-XGBoost模型性能均有明显改善，其中基于连续统去除变换的GA-XGBoost模型反演效果最优，均方根误差为4.85 mg·kg^{-1}，拟合优度达0.84，相对预测误差值为2.0。利用该模型进行研究区土壤Cu含量空间分布反演结果表明，该区域开采区周边及道路两侧受到Cu的污染较严重，这一规律与实地调查结果一致。

关键词 光谱学；遥感；高光谱；土壤重金属；极端梯度提升算法；遗传算法

中图分类号 TP79 **文献标志码** A **DOI**：10.3788/LOP202259.1230001

Retrieval of Heavy Metal Content in Soil Using GF-5 Satellite Images Based on GA-XGBoost Model

Bai Han[1], Yang Yun[1,2*], Cui Qinfang[3], Jia Peng[4], Wang Lixia[1]

[1]*College of Geology Engineering and Surveying, Chang'an University, Xi'an* 710054, *Shaanxi, China*;

[2]*Key laboratory of Degraded and Unused Land Consolidation Engineering, Ministry of Natural Resources, Xi'an* 710016, *Shaanx , China*;

[3]*Technologies Co., Ltd., Xi'an* 710001, *Shaanxi, China*;

[4]*Changqing Engineering Design Co., Ltd., Xi'an* 710018, *Shaanxi, China*

Abstract The rapid development of hyperspectral imaging technology has increased the use of domestic hyperspectral images for the inversion of soil parameters in a wide range. However, the accuracy needs to be improved. Therefore, by considering the Daxigou mining area in Shaanxi Province and taking GF-5 hyperspectral satellite images and measured soil samples as data sources, we proposed an XGBoost inversion model based on

收稿日期：2021-09-08；**修回日期**：2021-09-14；**录用日期**：2021-09-23

基金项目：陕西省自然科学基金(2022JM-163)、国家重点研发计划(2018YFC1504805)、中央高校基本科研业务费(300102269205、300102269201)

通信作者：*yangyunbox@163.com

genetic algorithm feature selection (GA-XGBoost). First, the preprocessed image data were transformed by continuum removal and logarithm of spectral reciprocal. Then, the Monte Carlo cross-validation method was used to remove abnormal soil samples. Finally, The XGBoost heavy metal content inversion models based on correlation coefficient and genetic algorithm feature selection were established respectively. The results show that the performance of the proposed GA-XGBoost model significantly improved compared with the XGBoost model based on correlation coefficient feature selection under the same spectral transformation. Furthermore, the GA-XGBoost model based on continuum removal transformation has the best inversion accuracy, with a root mean square error of 4.85 mg·kg^{-1}, goodness fit of 0.84, and relative prediction error of 2.0. The inversion results of the spatial distribution of soil Cu content in the study area using the model show that the surrounding of the mining area and both sides of the road are seriously polluted by Cu, which is consistent with the field survey results.

Key words spectroscopy; remote sensing; hyperspectral; soil heavy metals; extreme gradient boosting; genetic algorithm

1 引 言

矿区周边土壤重金属污染严重,危及人类健康。如何高效地获取矿区土壤重金属污染信息,为污染防治、生态环境保护以及人类健康提供重要的技术支持,是我国当前面临的关键科学问题之一。地面高光谱技术可以获取土壤近似连续的光谱信息,可以对土壤成分进行快速、准确的估算,如土壤有机质[1]、有机碳[2]、重金属[3]等,但无法实现大规模监测[4]。

高光谱遥感影像可实现周期性观测,覆盖范围大,光谱分辨率高,使得直接利用高光谱影像获取土壤重金属含量的空间分布成为了可能[5]。其原理与基于地物光谱测量的土壤重金属反演相同,即利用土壤样本重金属含量数据与遥感影像光谱数据建立模型,最终将反演模型应用到影像中,实现重金属含量的空间反演[6]。国产高分五号卫星(GF-5)搭载的可见光-短波红外高光谱相机(AHSI)是目前性能指标最先进的载荷之一[7],其光谱分辨率较Hyperion影像、环境一号(HJ-1)高光谱影像有很大提升,在诸多领域得到了应用,如土壤有机质估算[8]等,但在土壤重金属反演中应用较少。

以往对土壤成分的反演研究中,为了实现光谱快速降维,多采用基于相关系数的特征选择方法(CFS),即利用光谱数据与土壤成分之间的相关性分析初步筛选特征[8-9],或将选择的特征作为最终输入[10-11]建立模型,该方法简单、快速、解释性强,然而这种方法筛选条件有较强的主观性,同时由于土壤重金属与光谱之间往往不是简单的线性关系[4,6],基于线性假设的相关性分析很可能忽略掉土壤中微弱的重金属光谱信息,通常无法获得最优精度。随着机器学习的发展,越来越多的特征选择方法被应用到土壤成分的反演中,其中遗传算法(GA)是一种模拟生物种群进化方式的优化算法,以最大化目标函数(通常为精度)为目标,在特征空间中进行启发式搜索,可以快速获取近似最优解,实现特征选择,已广泛用于高光谱估算土壤成分含量的研究中[12-13]。

在重金属反演模型方面,以往研究大多使用基于线性假设的反演模型,如偏最小二乘回归[14]、多元逐步回归[15]等,但这些线性模型反演精度有待提高。近年来,越来越多的机器学习模型得到了应用,如神经网络[16]、支持向量机(SVM)[17]模型等,这些模型可以克服线性模型的局限性,并能取得较高精度,但是这些模型无法对结果给出合理解释。极端梯度提升(XGBoost)[18]是一种以决策树为基学习器的集成学习模型,可以处理非线性问题,并且训练效率高,通过特征的重要性排序,可以对模型进行很好的解释,在土地覆盖分类[19]、土壤含水量估算[20]等领域取得了不错的效果。

针对以上问题,本文以陕西大西沟矿区为例,利用GF-5影像光谱数据及实测土壤重金属含量数据建立了基于遗传算法特征选择的XGBoost重金属含量反演模型(GA-XGBoost),有效挖掘了重金属微弱信息,提高了反演精度,并对模型进行合理解释,为快速实现矿区周边环境监测提供了技术支持。

2 研究数据与预处理

2.1 研究区概况

大西沟矿区位于陕西省商洛市柞水县小岭镇。矿区及周围富含铜、铅、砷等多种重金属元素。该区域高差大,地形复杂,主要土地利用类型为采矿区、工矿设施、居民区等。采矿活动造成了土壤重

金属污染，需要定期对该矿区进行土壤重金属污染调查与监测。

2.2　土壤样本数据采集、分析

考虑到研究区地形起伏较大，部分区域无法到达，沿道路及坡中部可到达的位置布设了43处采集点。采样时间为2019年10月，采样时遵循多点取样、等量混合的原则，记录了采样点区域中心的地理坐标、土地利用类型等信息。将收集的43个样本在实验室内依据规范[21]对Cu元素含量进行检测。

2.3　GF-5影像数据及处理

GF-5卫星是2018年发射的国产高光谱卫星，搭载的AHSI能获取精细的地物光谱信息，具有广泛的应用前景，其与国内外其他常用高光谱卫星参数对比如表1所示。本实验组从自然资源部国土卫星遥感应用中心获取了接近采样时间的GF-5影像，时间为2019年11月15日。

对原始GF-5影像进行了预处理，剔除了受水汽影响严重的波段、成像质量差的波段以及短、波红外与可见光重合波段，并去除了部分波段的条纹噪声、坏线，处理后剩余285个波段。对预处理后的影像进行辐射校正处理，以去除大气等环境的影响，获取地面准确反射率；利用Landsat 8陆地成像仪（Landsat 8 OLI）同期影像数据、该区域先进星载热发射和反射辐射仪全球数字高程模型数据进行正射校正，与基准影像之间误差在一个像元以内；由于地形起伏较大，阴影区域与向阳区域同一地物光谱差异较大，需进行地形校正以消除其影响，本实验组尝试了常用的地形校正模型：C、SCS+C、VECA、Teillet，经目视比较Teillet模型阴影消除效果较好，故最终采用Teillet模型校正影像。

表1　部分高光谱卫星参数比较

Table 1　Comparison of hyperspectral satellite parameters

Satellite and its sensor	Band	Wavelength /nm	Spectral resolution /nm	Spatial resolution /m
GF-5(AHSI)	330	450–2500	5(VNIR),10(SWIR)	30
HJ-1-A(HSI)	105	450–1050	2–9	100
EO-1(Hyperion)	220	400–2500	10	30

2.4　异常土壤样本去除

土壤样本中可能存在异常样本，即光谱值或重金属含量值存在异常，不利于模型的拟合，因此需要对异常样本进行识别与去除。蒙特卡罗交叉验证（MCCV）方法[22]通过随机划分数据集建立大量的偏最小二乘回归模型，并根据模型预测残差的均值-方差图寻找异常样本。具体方法如下：每次取70%作为训练集，30%作为测试集，建立偏最小二乘回归模型，循环2500次，统计各个样本值在测试集中的预测残差，生成预测残差的均值-标准差图，如图1所示。

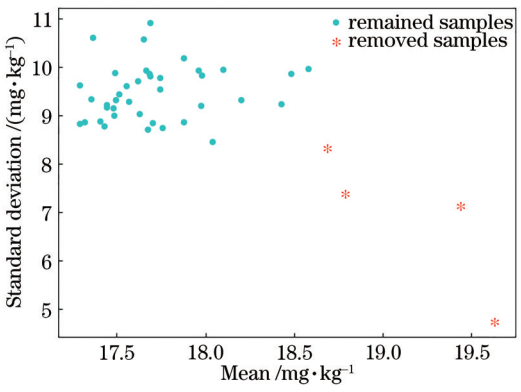

图1　MCCV法均值-标准差分布图

Fig. 1　Mean-standard deviation distribution of MCCV method

图1中，孤立点（高均值、高标准差）可能为异常点。为了最大限度保留样本，本实验组仅去除了明显异常值点。剔除异常样本后，将剩余39个有效样本的Cu元素含量的统计值与该区域的国家土壤元素统计数据[23]进行对比，结果如表2所示。

表2　样本Cu元素含量实测值与区域背景值的统计比较

Table 2　Comparison between measured Cu content and regional background value　　unit: mg·kg^{-1}

Item	Minimum	Median	Maximum	Mean
Background	6.8	19.5	43.6	21.4
Sample	24.0	46.0	75.0	48.2

从表1可以看出，剔除异常样本后，Cu元素含量在[24.0 mg·kg^{-1}, 75.0 mg·kg^{-1}]之间，最小值、最大值及平均值等各项统计指标均超过该区域的国家土壤元素统计数据。

3　研究方法

3.1　GF-5影像光谱变换及其特征分析

由于土壤中重金属含量低，在原始影像中表现出的光谱信号微弱，而光谱变换可以突出部分特征波段[4,6]。本实验组选取了连续统去除（CR）、光谱倒数对数变换来进行光谱增强。CR变换将反射率归

一到对应光谱背景中，可以有效突出特征波段[24]；光谱倒数的对数变换可以减少因光照变化引起的乘性因素影响[25]，样本点原始光谱（R）及两种光谱变换后的光谱曲线（CR和log R^{-1}）如图2所示。

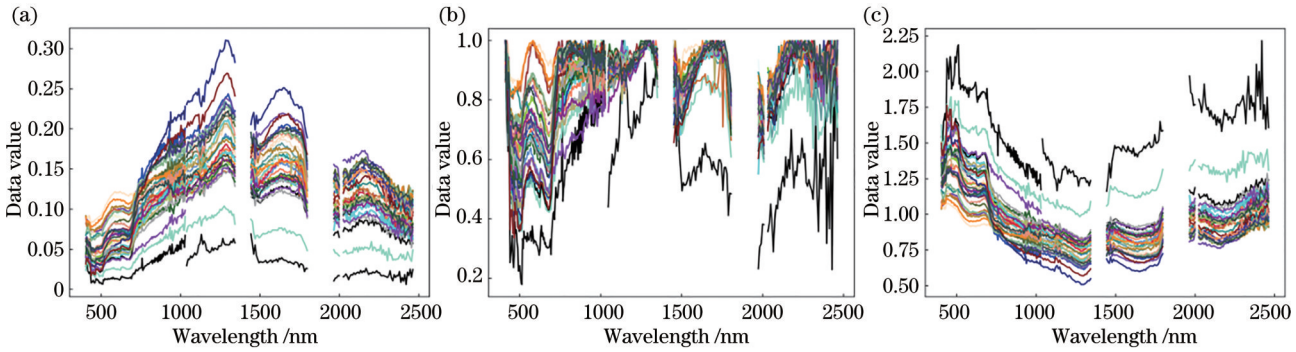

图2　不同光谱变换的光谱曲线。(a)原始光谱；(b)连续统去除光谱；(c)倒数对数光谱

Fig. 2　Spectral curves of different kinds of spectral transformations. (a) R; (b) CR; (c) log R^{-1}

从图3可以看出：与原始光谱曲线相比，CR变换后的光谱曲线中吸收光谱特征得到突出；而经倒数对数变换后的光谱曲线与原始光谱变化趋势基本相反，光谱值区间得到了放大。为进一步选取Cu元素的特征波段，对Cu含量与影像光谱值进行了Pearson相关性分析。Pearson相关系数可以描述重金属含量与光谱之间的线性相关性，其计算公式为

$$r_i=\frac{\mathrm{cov}(X_i,Y)}{\sqrt{\mathrm{var}(X_i)\,\mathrm{var}(Y)}},\tag{1}$$

式中：cov(X_i, Y)表示第i个波段光谱值与重金属含量之间的协方差；var(X_i)为第i个波段光谱值的方差；var(Y)为重金属含量的方差。原始光谱及其变换后光谱值与重金属含量之间的相关性如图3所示。

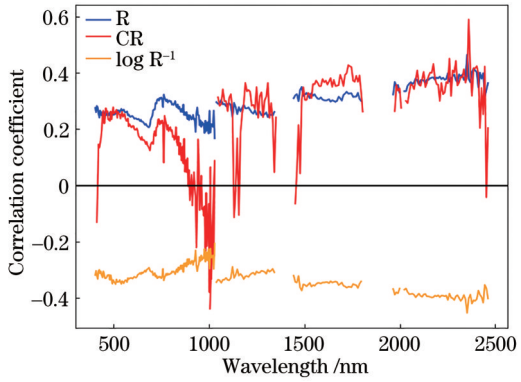

图3　不同光谱变换与重金属含量的相关性分析

Fig. 3　Correlation analysis of different spectral transformations and heavy metal content

对图3的光谱值与Cu含量的相关系数统计分析如表3所示，其中显著性波段为相关系数t检验中0.05水平显著的波段，即$p<0.05$。

综合分析图3和表3，可以看出：3种光谱与重金属含量的相关系数均在2344 nm处达到最大，CR变换相关系数最值最大，达0.590；光谱倒数的对数与原始光谱值呈现负相关，CR变换后的相关系数正负均有；各光谱变换后与重金属含量之间的显著性波段均增加，但经倒数对数变换后的相关系数最值有所下降。

表3　原始光谱及其两次变换与Cu含量的相关系数统计

Table 3　Statistics of correlation coefficient between original spectrum and its two transformations with Cu content

Spectrum transform	Maximum absolute correlation coefficient	Band	Number of significant bands
R	0.464**	R_{2344}	77
CR	0.590**	R_{2344}	95
log R^{-1}	0.453**	R_{2344}	186

Notes: R_{2344} represents the band at 2344 nm, ** represents significant value at the $p<0.01$ level.

3.2　基于遗传算法特征选择的XGBoost反演模型

XGBoost是基于梯度提升树（GBDT）改进的，它们同属于Boosting算法。GDBT将目标函数泰勒展开到一阶，容易遗漏目标函数部分信息；而XGBoost将目标函数泰勒展开到二阶，有助于提升模型效果，同时XGBoost通过特征预排序实现了特征维度上的并行计算，大大提高了运行效率[18]。XGBoost算法流程如下。

在每一次迭代过程中，XGBoost需要最小化的目标函数训练模型的表达式为

$$f_{\mathrm{obj}}^{t}=L(\theta)+\Omega(f_t),\tag{2}$$

$$\begin{cases}L(\theta)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\\ \Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^{\,2}\end{cases},\tag{3}$$

式中：θ为模型参数，$L(\theta)$为损失函数，通常为均方误差；$\Omega(f_t)$为正则化项，用于防止模型出现过拟合；y_i，$\hat{y}_i$分别为第i个样本的真值与预测值；γ为加入新叶子节点的惩罚代价；T为叶子节点的个数；λ为正则化参数；w_j为第j个叶子节点的权重。每一次迭代加入新的树来拟合上一棵树的预测值与真值的残差，即

$$\hat{y}_i^{(0)}=0,\hat{y}_i^{(1)}=f_1(x_i)=\hat{y}_i^{(0)}+f_1(x_i),\cdots,\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i),\tag{4}$$

式中：$\hat{y}_i^{(t)}$为第t个树模型的预测值；第t次迭代加入的函数$f_t(x_i)$需使对应的目标函数f_{obj}^t最小。算法达到最大迭代次数时停止运行，得到最终结果。

由于样本数量较少，光谱特征较多，XGBoost模型容易出现过拟合现象，因此本实验组提出了基于遗传算法特征选择的XGBoost反演模型，在建立XGBoost模型过程中引入遗传算法进行特征选择，然后利用所选最优特征子集构建了XGBoost模型，用于对土壤重金属含量的反演。具体步骤如下：

1）在特征空间（所有可能的特征组合）中随机产生多个特征子集形成特征子集集合（含30个特征子集），其中特征子集称为个体，特征子集集合称为种群；对个体进行二进制编码，以方便后续产生新个体，并利用XGBoost模型预测值的均方根误差（RMSE）对每个个体进行评价。

2）选择RMSE低的个体直接进入下一次迭代，同时利用两点交叉（交叉概率为0.6）交换两个个体的部分特征，利用突变（突变概率为0.01）改变单个个体的部分特征，以避免算法陷入局部最优解，至此完成一次迭代，产生新的种群。

3）算法运行至最大迭代次数（150）后，输出种群中最优的个体。由于遗传算法具有随机性，需多次执行取其最优结果。

4）将最优的个体，即最优特征子集作为输入，建立XGBoost模型。

4　实验及分析

4.1　数据集划分与模型评价指标

为客观评价模型性能，需要合理划分数据集。本实验组采用浓度梯度法对数据集进行划分：按照样本的重金属含量大小排序，每隔几个样本取出1个作为测试集。取80%作为训练集，20%作为测试集，结果如表4所示。

从表4可以看出：训练集、测试集、总集各个统计指标均相近，数据集划分较为合理。采用RMSE作为辅助评价标准，相对预测误差（RPD）与拟合优度（R^2）作为主要模型评价标准，其中RMSE反映观测值与实测值之间的偏差，其值越小，表明模型反演精度越高；RPD反映模型的预测能力；R^2反映模型的拟合效果，其取值范围为[0,1]，R^2越高表明模型拟合效果越好。上述指标计算公式分别为

$$\begin{cases}E_{\text{RMSE}}=\sqrt{\dfrac{\sum_{i=1}^{N}(\hat{y}_i-y_i)^2}{N}}\\ D_{\text{RPD}}=\dfrac{\sqrt{\sum_{i=1}^{N}(y_i-\overline{\hat{y}_i})^2}}{(N-1)^2\cdot E_{\text{RMSE}}},\\ R^2=\dfrac{\sum_{i=1}^{N}(y_i-\overline{\hat{y}_i})^2}{\sum_{i=1}^{N}(\hat{y}_i-\overline{\hat{y}_i})^2}\end{cases}\tag{5}$$

式中：$\overline{\hat{y}_i}$为所有预测值的平均值；N为样本总数。

表4　训练集和测试集样本统计特征

Table 4　Statistical characteristics of training set and testing set

Sample set	Sample size	Maximum /(mg·kg^{-1})	Minimum /(mg·kg^{-1})	Mean /(mg·kg^{-1})	Standard deviation /(mg·kg^{-1})
Total set	39	75.0	24.0	48.2	13.38
Training set	31	75.0	26.0	49.1	13.5
Testing set	8	64.0	24.0	45.0	13.1

依据以下准则来评价模型的性能[6]：当$R^2\geqslant0.9$，$D_{\text{RPD}}\geqslant3.0$时，视为优秀模型；当$0.81\leqslant R^2\leqslant0.9$，$2.5\leqslant D_{\text{RPD}}\leqslant3.0$时，视为良好模型；当$0.65\leqslant R^2\leqslant0.81$，$2.0\leqslant D_{\text{RPD}}\leqslant2.5$时，视为近似模型；当$0.5\leqslant R^2\leqslant0.65$，$1.5\leqslant D_{\text{RPD}}\leqslant2.0$时，该模型具有一定反演能力；当$R^2\leqslant0.5$，$D_{\text{RPD}}\leqslant1.5$时，该模型不能反演出正确的结果。

4.2　结果评价与分析

4.2.1　土壤重金属反演结果评价与对比分析

为了验证GA-XGBoost模型的有效性，以R^2、

RMSE及RPD 3个指标对模型性能进行评价，并与基于相关系数的特征选择（CFS）建立的XGBoost模型（CFS-XGB）进行了比较，其中CFS-XGB所输入的特征为各个光谱（R、log R^{-1}、CR）的显著性波段。各模型3个性能指标对比如表5所示。

表5　基于遗传算法与相关系数特征选择下XGBoost模型性能对比

Table 5　Precision comparison of XGBoost model based on GA and CFS feature selection methods

Selection method	Model	Training set			Testing set			Number of bands
		R^2	RMSE	RPD	R^2	RMSE	RPD	
CFS	R-CFS-XGB	0.95	3.05	4.1	0.51	8.57	1.1	77
	CR-CFS-XGB	0.92	3.77	3.0	0.61	7.59	1.6	95
	log R^{-1}-CFS-XGB	0.94	3.12	3.9	0.53	8.36	1.3	186
GA	R-GA-XGB	0.95	3.07	3.8	0.61	7.60	1.5	139
	CR-GA-XGB	0.94	3.38	3.3	0.84	4.85	2.0	137
	log R^{-1}-GA-XGB	0.91	3.86	2.7	0.65	7.48	1.7	110

Notes: R-CFS-XGB represents XGBoost model that using R as the input feature after CFS feature selection, and R-GA-XGB represents XGBoost model that using R as the input feature after GA feature selection, etc.

由表5可以看出：相同光谱类型条件下，GA-XGBoost模型性能比CFS-XGBoost均有明显改善，表明该模型具有一定反演能力；R-XGBoost，log R^{-1}-XGBoost模型测试集的R^2值提高了0.1左右，RMSE值下降了约1 mg·kg^{-1}，RPD值提高了约0.3；CR变换下的CR-GA-XGBoost模型效果最好，测试集R^2值达到0.84，RPD值为2，达到近似模型标准；对于CFS特征选择方式下的3种XGBoost模型，CR-CFS-XGBoost模型性能最优，其测试集上的R^2=0.61，D_{RPD}=1.6，表明该模型具有一定反演能力，但未达到近似模型标准，其他模型均不能作为反演模型。

两种特征选择下的模型预测值与真值散点图如图4、5所示。从图中可以看出，CFS-XGBoost在训练集表现较好，散点分布在$y=x$附近，但测试集散点大多偏离$y=x$线，说明模型存在严重过拟合，

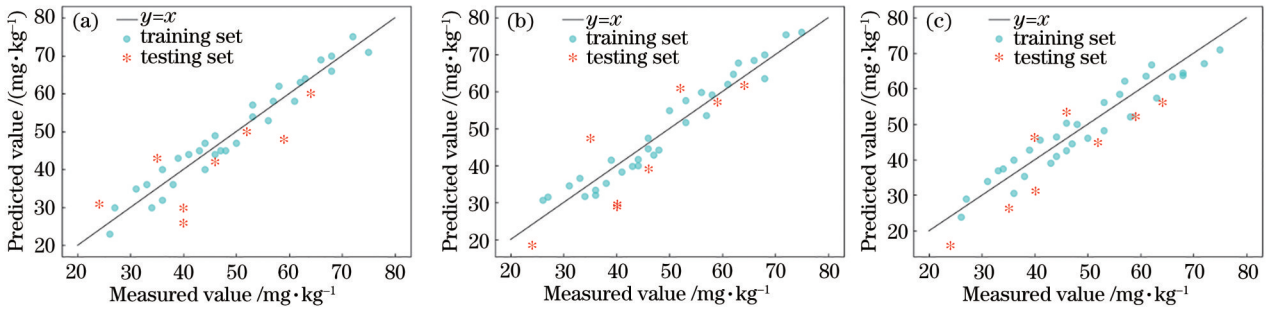

图4　CFS-XGBoost模型预测结果散点图。(a)原始光谱；(b)倒数对数光谱；(c)连续统去除光谱

Fig. 4　Scatter diagrams of prediction results of CFS-XGBoost. (a) R; (b) log R^{-1}; (c) CR

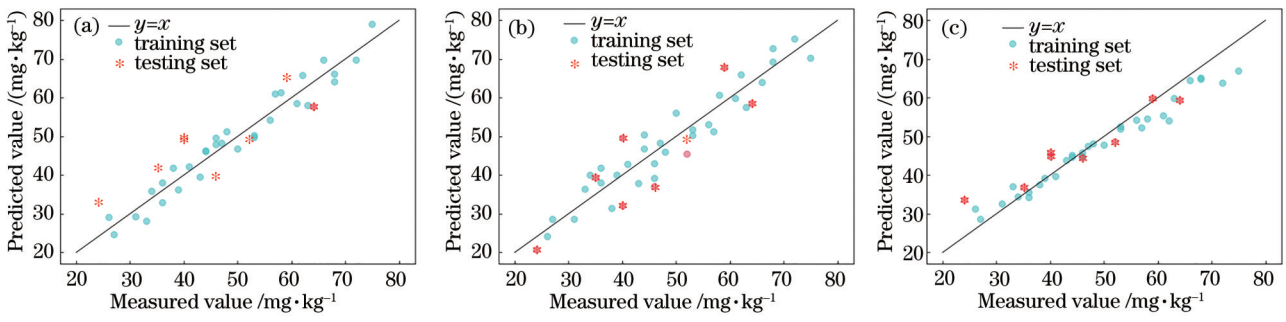

图5　GA-XGBoost模型预测结果散点图。(a)原始光谱；(b)倒数对数光谱；(c)连续统去除光谱

Fig. 5　Scatter diagram of prediction results of GA-XGBoost. (a) R; (b) log R^{-1}; (c) CR

而GA-XGBoost模型测试集散点更接近$y=x$线，模型过拟合现象减弱。

4.2.2　波段选择结果对比及特征波段分析

为进一步分析所提模型在光谱特征选择方面的有效性，对CR-GA-XGBoost选取的137个特征与CR-CFS-XGBoost选取的95个特征进行了对比，结果如图6所示，其中光谱曲线为CR样本平均光谱曲线，GA、CFS、GA&CFS分别表示GA特征选择独有的波段、相关系数特征选择独有的波段以及两种方法共有的波段。

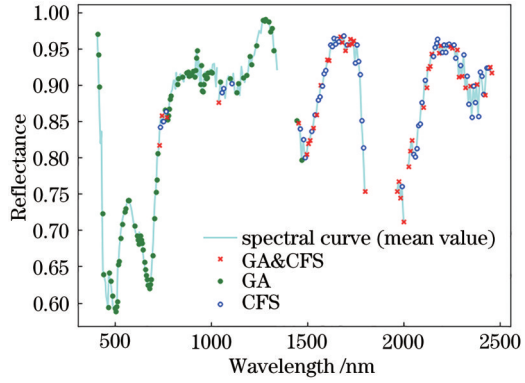

图6　基于相关系数的特征选择与GA特征选择结果

Fig. 6　Results of correlation coefficient feature selection and GA feature selection

从图6可以看出：利用相关系数选择的波段分布在短波红外（1038～2464 nm），而在可见光-近红外区间几乎没有波段入选；而GA选择的波段在各个波段均有分布，且大部分集中在光谱曲线的吸收或反射峰。GA所选波段中显著性波段共38个，仅占比28%，说明土壤中微弱信号（非显著性波段）可以有效提高预测精度。土壤光谱与其重金属含量之间关系复杂，仅仅依靠相关系数进行特征选择虽然可以减小数据冗余，但容易忽略土壤中的微弱信号，导致反演结果不佳。

XGBoost模型可利用特征在分裂时的作用进行特征重要性排序，对所选特征的相对重要性进行量化，从而进一步分析，提高模型解释性。以在所有树中某特征被用来分裂节点的次数（weight）为依据，利用CR-GA-XGBoost模型对特征进行重要性排序，weight较高的前20个特征及其重要性如图7所示。

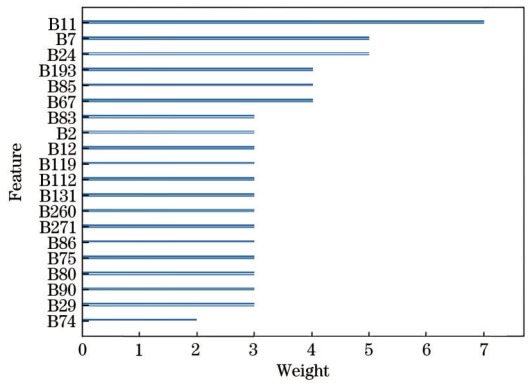

图7　XGBoost对特征的评分

Fig. 7　Feature importance scores given by XGBoost

由图7可知，前6个特征相对重要，weight达到了4以上。第3.1节基于相关性分析得到的相关性最大的波段B271（2344 nm）仅为并列第七，说明通过相关系数特征选择的波段并非最优特征波段。前20个特征中，B7（428 nm）、B24（501 nm）、B2（407 nm）、B12（450 nm）与贺军亮等[26]的研究相类似，这些波段附近存在氧化铁的微弱吸收峰；B131（959 nm）、B260（2252 nm）、B271（2344 nm）与涂宇龙等[27]的研究一致，而2200 nm附近是有机质的特征波段。重金属元素通常吸附于有机质、铁氧化物中[4-6,27]，本实验所采样本大部分属于旱作地褐土土壤，有机质、铁氧化物较多，表明XGBoost的特征排序使得利用GF-5影像进行Cu含量反演的机理得到了较好的解释。

4.2.3　土壤重金属空间分布规律分析

选取第4.2.1节中最优模型（CR-GA-XGBoost）对研究区重金属含量空间分布进行反演：将遥感影像光谱数据输入到模型中，得到重金属含量的空间分布图，每个像元的值为其重金属含量值，为了方便对结果进行分析，对重金属含量值进行了分类，结果如图8所示。

为了进一步分析重金属含量的空间分布特征，对反演的研究区Cu元素含量值进行了统计分析，结果如表6所示。

综合分析表6、图8，可以看出：全区大部分区域（99.93%）超出国家土壤统计数据Cu元素含量的算术平均值（21.4 mg·kg^{-1}），均值达到

表6　铜金属含量估计结果统计

Table 6　Statistics of copper content estimation results

Content of copper /(mg·kg^{-1})	19.79–21.40	21.40–38.76	38.76–45.02	45.02–51.83	51.83–66.75
Percentage /%	0.07	53.06	21.07	17.86	7.94

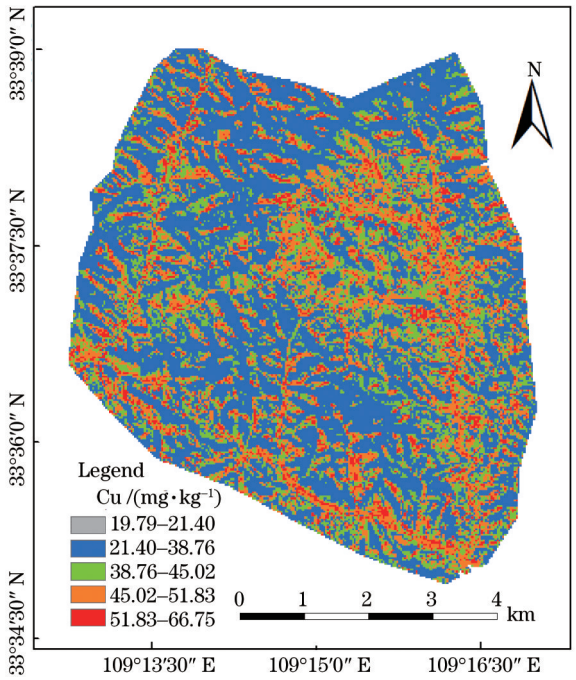

图 8　Cu元素含量空间分布图

Fig. 8　Spatial distribution of Cu content

38.45 mg·kg^{-1}，表明该区域土壤受到一定的Cu污染。结合该区域道路分布图以及矿区位置图可知，受污染相对较为严重的区域（大于51.83 mg·kg^{-1}）分布在道路两旁以及矿区周围。经实地调查，矿区源源不断出入矿车，且运输途中粉尘污染严重，因此这可能是矿石碎屑在运输过程中从车辆中掉落以及扬尘导致的。

5　结　　论

以陕西大西沟矿区为研究区域，利用GF-5影像光谱及其变换光谱建立了基于遗传算法特征选择的XGBoost重金属含量反演模型，实验结果表明：基于遗传算法的特征选择可以很好地识别出重金属微弱信号，并有效改善模型过拟合问题，建立的最优模型CR-GA-XGBoost提高了研究区Cu元素含量的反演精度，达到了近似模型的标准；依据XGBoost特征重要性得到土壤Cu含量特征波段为407，428，450，501，959，2252，2344 nm，与前人研究结果相似，且与有机质、铁氧化物光谱特征波段一致。利用所提模型反演的研究区重金属含量空间分布可以得出如下规律：研究区受到了一定的污染，且污染研究区域分布于矿区、道路周围。研究也为国产GF-5卫星影像在土壤重金属含量反演等定量遥感应用中提供了一种可借鉴的技术方法。但本实验组的研究还不完善，由于高光谱影像空间分辨率的限制，影像中常出现混合像元，反演精度难以提高，混合像元分解技术可以有效解决该问题。未来将引入数字高程模型等其他辅助数据，挖掘更多有效的影响因子，进一步提升土壤重金属反演的精度和智能化程度。

参　考　文　献

[1] 马国林，丁建丽，张子鹏. 基于土壤协变量与VIS-NIR光谱估算土壤有机质含量的研究[J]. 激光与光电子学进展，2020，57(19)：192801.
Ma G L，Ding J L，Zhang Z P. Soil organic matter content estimation based on soil covariate and VIS-NIR spectroscopy[J]. Laser & Optoelectronics Progress，2020，57(19)：192801.

[2] 赵启东，葛翔宇，丁建丽，等. 结合分数阶微分技术与机器学习算法的土壤有机碳含量光谱估测[J]. 激光与光电子学进展，2020，57(15)：153001.
Zhao Q D，Ge X Y，Ding J L，et al. Combination of fractional order differential and machine learning algorithm for spectral estimation of soil organic carbon content[J]. Laser & Optoelectronics Progress，2020，57(15)：153001.

[3] 张霞，王一博，孙伟超，等. 基于铁氧化物特征光谱和改进遗传算法反演土壤Pb含量[J]. 农业工程学报，2020，36(16)：103-109.
Zhang X，Wang Y B，Sun W C，et al. Inversion of Pb content in soil based on iron oxide characteristic spectrum and improved genetic algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering，2020，36(16)：103-109.

[4] Wang F H，Gao J，Zha Y. Hyperspectral sensing of heavy metals in soil and vegetation：feasibility and challenges[J]. ISPRS Journal of Photogrammetry and Remote Sensing，2018，136：73-84.

[5] 刘彦平，罗晴，程和发. 高光谱遥感技术在土壤重金属含量测定领域的应用与发展[J]. 农业环境科学学报，2020，39(12)：2699-2709.
Liu Y P，Luo Q，Cheng H F. Application and development of hyperspectral remote sensing technology to determine the heavy metal content in soil[J]. Journal of Agro-Environment Science，2020，39(12)：2699-2709.

[6] Shi T Z，Chen Y Y，Liu Y L，et al. Visible and near-infrared reflectance spectroscopy：an alternative for monitoring soil contamination by heavy metals[J]. Journal of Hazardous Materials，2014，265：166-176.

[7] 刘银年. 高光谱成像遥感载荷技术的现状与发展[J]. 遥感学报, 2021, 25(1): 439-459.
Liu Y N. Development of hyperspectral imaging remote sensing technology[J]. National Remote Sensing Bulletin, 2021, 25(1): 439-459.

[8] Meng X T, Bao Y L, Ye Q, et al. Soil organic matter prediction model with satellite hyperspectral image based on optimized denoising method[J]. Remote Sensing, 2021, 13(12): 2273.

[9] Meng X T, Bao Y L, Liu J G, et al. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data[J]. International Journal of Applied Earth Observation and Geoinformation, 2020, 89: 102111.

[10] 林楠, 刘海琪, 杨佳佳, 等. BA-Adaboost模型的黑土区土壤养分含量高光谱估测[J]. 光谱学与光谱分析, 2020, 40(12): 3825-3831.
Lin N, Liu H Q, Yang J J, et al. Hyperspectral estimation of soil nutrient content in the black soil region based on BA-Adaboost[J]. Spectroscopy and Spectral Analysis, 2020, 40(12): 3825-3831.

[11] Wei L F, Yuan Z R, Zhong Y F, et al. An improved gradient boosting regression tree estimation model for soil heavy metal (arsenic) pollution monitoring using hyperspectral remote sensing[J]. Applied Sciences, 2019, 9(9): 1943.

[12] Jiang G, Zhou S G, Cui S C, et al. Exploring the potential of HySpex hyperspectral imagery for extraction of copper content[J]. Sensors, 2020, 20(21): 6325.

[13] 王轩慧, 陈建毅, 郑西来, 等. 基于SGA-RF算法的农业土壤镉浓度反演研究[J]. 农业机械学报, 2018, 49(10): 261-269.
Wang X H, Chen J Y, Zheng X L, et al. Inversion of cadmium content in agriculture soil based on SGA-RF algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(10): 261-269.

[14] 杨灵玉, 高小红, 张威, 等. 基于Hyperion影像植被光谱的土壤重金属含量空间分布反演: 以青海省玉树县为例[J]. 应用生态学报, 2016, 27(6): 1775-1784.
Yang L Y, Gao X H, Zhang W, et al. Estimating heavy metal concentrations in topsoil from vegetation reflectance spectra of Hyperion images: a case study of Yushu County, Qinghai, China[J]. Chinese Journal of Applied Ecology, 2016, 27(6): 1775-1784.

[15] 屈永华, 焦思红, 刘素红, 等. 从高光谱卫星数据中提取植被覆盖区铜污染信息[J]. 光谱学与光谱分析, 2015, 35(11): 3176-3181.
Qu Y H, Jiao S H, Liu S H, et al. Retrieval of copper pollution information from hyperspectral satellite data in a vegetation cover mining area[J]. Spectroscopy and Spectral Analysis, 2015, 35(11): 3176-3181.

[16] Liu P, Liu Z H, Hu Y M, et al. Integrating a hybrid back propagation neural network and particle swarm optimization for estimating soil heavy metal contents using hyperspectral data[J]. Sustainability, 2019, 11(2): 419.

[17] 袁自然, 魏立飞, 张杨熙, 等. 优化CARS结合PSO-SVM算法农田土壤重金属砷含量高光谱反演分析[J]. 光谱学与光谱分析, 2020, 40(2): 567-573.
Yuan Z R, Wei L F, Zhang Y X, et al. Hyperspectral inversion and analysis of heavy metal arsenic content in farmland soil based on optimizing CARS combined with PSO-SVM algorithm[J]. Spectroscopy and Spectral Analysis, 2020, 40(2): 567-573.

[18] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, San Francisco, California, USA. New York: ACM, 2016: 785-794.

[19] Abdi A M. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data[J]. GIScience & Remote Sensing, 2020, 57(1): 1-20.

[20] 田美玲, 葛翔宇, 丁建丽, 等. 耦合机器学习和机载高光谱数据的土壤含水量估算[J]. 激光与光电子学进展, 2020, 57(9): 093002.
Tian M L, Ge X Y, Ding J L, et al. Coupled machine learning and unmanned aerial vehicle based hyperspectral data for soil moisture content estimation [J]. Laser & Optoelectronics Progress, 2020, 57(9): 093002.

[21] 中华人民共和国生态环境部. 土壤和沉积物 铜、锌、铅、镍、铬的测定 火焰原子吸收分光光度法: HJ 491—2019[S]. 北京: 中国环境科学出版社, 2019.
Ministry of Ecology and Environment of the People's Republic of China. Soil and sediment determination of copper, zinc, lead, nickel and chromium: flame atomic absorption spectrophotometry: HJ 491—2019 [S]. Beijing: China Environmental Science Press, 2019.

[22] 刘智超, 蔡文生, 邵学广. 蒙特卡罗交叉验证用于近红外光谱奇异样本的识别[J]. 中国科学(B辑: 化学), 2008, 38(4): 316-323.
Liu Z C, Cai W S, Shao X G. Identify singular samples in near infrared spectroscopy based on

Monte Carlo Cross Validation[J]. Science in China (Series B: Chemistry), 2008, 38(4): 316-323.

[23] 国家环境保护局. 中国土壤元素背景值[M]. 北京: 中国环境科学出版社, 1990.
National Environment Protection Agency. Backgrounds value of soil elements in China[M]. Beijing: China Environment Science Press, 1990.

[24] 王维, 沈润平, 吉曹翔. 基于高光谱的土壤重金属铜的反演研究[J]. 遥感技术与应用, 2011, 26(3): 348-354.
Wang W, Shen R P, Ji C X. Study on heavy metal Cu based on hyperspectral remote sensing[J]. Remote Sensing Technology and Application, 2011, 26(3): 348-354.

[25] 徐良骥, 李青青, 朱小美, 等. 煤矸石充填复垦重构土壤重金属含量高光谱反演[J]. 光谱学与光谱分析, 2017, 37(12): 3839-3844.
Xu L J, Li Q Q, Zhu X M, et al. Hyperspectral inversion of heavy metal content in coal gangue filling reclamation land[J]. Spectroscopy and Spectral Analysis, 2017, 37(12): 3839-3844.

[26] 贺军亮, 崔军丽, 张淑媛, 等. 基于偏最小二乘的土壤重金属铜含量高光谱估算[J]. 遥感技术与应用, 2019, 34(5): 998-1004.
He J L, Cui J L, Zhang S Y, et al. Hyperspectral estimation of heavy metal Cu content in soil based on partial least square method[J]. Remote Sensing Technology and Application, 2019, 34(5): 998-1004.

[27] 涂宇龙, 邹滨, 姜晓璐, 等. 矿区土壤Cu含量高光谱反演建模[J]. 光谱学与光谱分析, 2018, 38(2): 575-581.
Tu Y L, Zou B, Jiang X L, et al. Hyperspectral remote sensing based modeling of Cu content in mining soil[J]. Spectroscopy and Spectral Analysis, 2018, 38(2): 575-581.