激光与光电子学进展

基于海鸥算法优化随机森林的土壤硒含量高光谱反演

谢鹏，王正海*，肖蓓，田雨欣
中山大学地球科学与工程学院，广东 广州 510275

摘要 针对土壤硒含量光谱数据冗余、模型复杂度较高等问题，本研究系统采集含硒土壤若干份，并获取样本硒含量和光谱信息，对原始光谱进行平滑多元散射校正一阶微分(SG-MSC-FD)光谱增强处理，利用稳定性竞争自适应重加权采样(sCARS)等特征提取算法筛选特征波长，建立土壤硒含量的偏最小二乘(PLSR)、支持向量机(SVM)、随机森林(RF)、海鸥优化随机森林(SOA-RF)预测模型，通过对比不同特征筛选下模型的决定系数(R^2)、均方根误差(RMSE)和相对分析误差(RPD)，寻找最佳的组合模型。结果表明：不同特征筛选下的模型精度均有较大提升，其中变量组合集群分析法结合遗传算法(VCPA-GA)精度最高，sCARS算法提取的变量数最少，仅占全波段的0.49%；RF较SVM和PLSR模型有更好的鲁棒性，SOA-RF模型的参数最佳，极大地提升了模型的反演精度。综上，经VCPA-GA特征提取下的SOA-RF模型是最佳的预测模型(R^2=0.92、RMSE为0.08、RPD为2.911)，该模型能够实现对土壤硒含量快速、高效反演。

关键词 土壤硒；高光谱；特征筛选；海鸥优化算法；随机森林
中图分类号 O433.4 **文献标志码** A **DOI**：10.3788/LOP222037

Hyperspectral Inversion of Soil Selenium Content Based on Seagull Algorithm Optimized Random Forest

Xie Peng, Wang Zhenghai*, Xiao Bei, Tian Yuxin
School of Earth Sciences and Engineering, Sun Yat-sen University, Guangzhou 510275, *Guangdong, China*

Abstract The aim of this study is to investigate the problem of redundant soil selenium content spectral data and high model complexity. Several selenium-containing soil samples were collected, and the selenium content and spectral information of the samples were obtained, The raw spectra were preprocessed using Savizkg-Golag multivariate scatter correction first-order differential (SG-MSC-FD), and the feature wavelengths were screened using stability competitive adaptive reweighted sampling (sCARS) and other algorithms to establish the partial least squares regression (PLSR), support vector machine (SVM), random forest (RF), soil selenium-content seagull optimization algorithm (SOA)-RF prediction models. The coefficient of regression (R^2), root mean square error (RMSE) and relative predictive deviation (RPD) values of the models under different feature screenings were compared to determine the best combination model. The results show that the accuracy of the models under different feature filtering is improved. The sCARS algorithm extracts the least number of variables, accounting for only 0.49% of the full band, and the algorithm combined variable combination cluster analysis and genetic algorithm has the highest accuracy. The RF model exhibits better robustness than the SVM and PLSR models, and the inversion accuracy of the models significantly improves with parameter optimization of SOA-RF. In summary, the SOA-RF model with VCPA-GA feature extraction is the best prediction model (R^2=0.92, RMSE is 0.08, RPD is 2.911), and it can achieve rapid and efficient inversion of soil selenium content.

Key words soil Se content; hyperspectrum; feature selection; seagull optimization algorithm; random forest

1 引 言

硒是人体中必不可少的微量元素之一[1-2]。硒以其特殊的“两面性”影响着人的生命安全，当人体内硒含量严重缺乏时，人体会出现克山、大骨节等疾病[3]；当人体内硒含量严重超标时，又会导致硒中毒，特殊的两面性使得硒受到各领域的广泛关注[4]。人体自身无法合成硒，获取硒的途径是食用含硒水果、蔬菜以及其

收稿日期：2022-07-11；**修回日期**：2022-08-10；**录用日期**：2022-08-29；**网络首发日期**：2022-09-09
基金项目：国家自然科学基金(41572316)、广州市科技计划(201804010274)、广东省基础与应用基础研究基金(2020A1515010666)
通信作者：*wzhengh@mail.sysu.edu.cn

他农产品，而土壤是人体中硒的最终来源。因此，研究土壤中的硒含量及分布，对解决人体硒健康至关重要。随着高光谱技术的不断发展，土壤硒含量的预测摆脱了单一的地球化学手段，呈现出快速、无污染、大面积预测的发展前景。近年来，许多学者对土壤硒含量的高光谱预测进行了大量研究。张东辉等[5]通过建立土壤硒含量与光谱反射率的相关性，引进相关系数较大的5个波长点作为光谱参数，建立反演模型，成功实现了土壤硒含量的高光谱预测。

尽管土壤硒含量的高光谱预测前景非常可观，但由于土壤硒含量与光谱反射率的敏感性较低，特征波段的选择，成为了高光谱准确预测土壤硒含量的难点之一。实际上，在可见光和近红外光谱研究中，特征波段的选择是至关重要的。随着预测模型的不断成熟，特征波段的选择受到了大量学者的关注。程介虹等[6]在土壤有机质的高光谱反演中结合迭代保留信息变量（IRIV）和连续投影算法（SPA）实现了数据的高效降维；乔天等[7]基于遗传算法（GA）建立偏最小二乘（PLSR）模型，提升了土壤质地的反演精度。众多的特征选择方法被应用于土壤属性的预测模型中，在降维的同时，特征选择的效率与有用信息保留情况也被加以考虑。因此，选择合适的特征筛选方法可以减少数据冗余，降低模型复杂度，提高模型预测精度[8]。另外，对土壤硒含量的高光谱研究，目前更多利用线性模型，而支持向量机（SVM）、随机森林（RF）等非线性模型在土壤硒含量的反演中利用较少。

因此，本研究通过采集连州地区含硒土壤样本，获取硒含量和光谱反射率数据，在光谱增强和特征选择的基础上，建立线性模型和非线性模型进行土壤硒含量的反演，并对比经过不同特征筛选的各模型精度，以此寻找最佳的特征选择方法和回归模型，为土壤硒含量的大面积预测提供更加高效的途径。

2　材料与方法

基于土壤样本获取光谱信息及硒含量，利用光谱增强手段以及特征波段提取算法，结合PLSR、SVM、RF、海鸥优化随机森林（SOA-RF）预测模型进行土壤硒含量的高光谱反演，所用方法的步骤如图1所示。

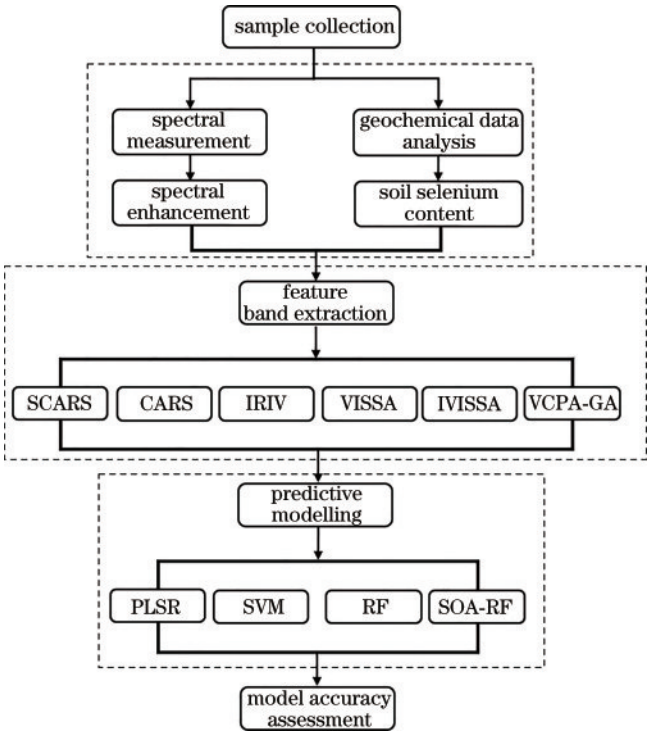

图1　技术路线图

Fig. 1　Technology roadmap

2.1　数据源获取及预处理

本次实验所用的土壤样本采自硒含量较为丰富的连州地区，使用五点采样法共采集50份0～20 cm的土壤样本。利用四酸消解法电感耦合等离子体发射光谱质谱测定土壤硒含量（检测范围为0.006～500 μg·g^{-1}），实验测得土壤硒含量均值为0.72 μg·g^{-1}、最大值为2.38 μg·g^{-1}、最小值为0.20 μg·g^{-1}；使用PSR+3500便携式地物光谱仪（波长范围为350～2500 nm）测量土壤样本的光谱反射率，每个样品测量10次，取平均值作为目标样本的光谱反射率值，部分土壤样品的原始光谱曲线如图2（a）所示。由于测得的土壤样本光谱在350～399 nm、2450～2500 nm波段噪声较

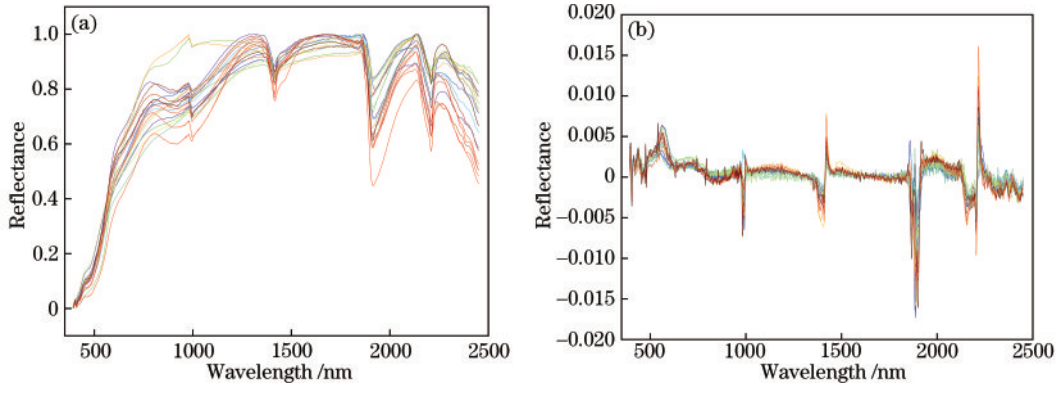

图2　土壤样品光谱曲线。（a）原始光谱；（b）MSC-FD变换光谱

Fig. 2　Spectral curves of soil samples. (a) Original spectra; (b) MSC-FD transformed spectra

大，剔除这2个范围的光谱，另外，考虑到原始光谱反射率与土壤硒含量的相关程度较低，利用卷积平滑结合多元散射校正与一阶微分（SG-MSC-FD）数学变换，对剩余400～2449 nm的波段进行光谱增强处理，经SG-MSC-FD光谱增强处理后的光谱曲线如图2（b）所示，由图2（b）可知，光谱增强处理有效降低了基线漂移的干扰，光谱曲线的特征性明显增强。

2.2 特征波段选择方法

2.2.1 sCARS算法

稳定性竞争自适应重加权采样（sCARS）算法是一种新颖的特征波长选择方法，在近红外光谱选择上具有快速、高效等优势。该算法始终以变量的稳定性为基本准则，变量被选择的概率与该变量的稳定性呈正相关[9]。sCARS算法在获取相应波长变量稳定性值的基础上，采用自适应重加权采样技术（ARS）和指数衰减函数（EDF），筛选出回归系数绝对值大且稳定性高的变量，多次循环迭代此过程。最终通过十折交互检验对每次循环后所得的变量子集进行检验，优选出交互验证均方根误差（RMSECV）最小的变量子集作为特征参数。sCARS算法的步骤如下：

步骤1 开始循环，并计算变量集中各变量的稳定性；

步骤2 利用EDF保留稳定性值较大的变量，通过ARS将保留的较稳定的变量作为新的变量集，再次进行波长筛选；

步骤3 利用EDF与APS不断循环筛选，获得最终变量子集，建立PLS模型，计算相应的RMSECV值；

步骤4 将RMSECV值最小的子集作为最优值输出，绘制迭代过程变化图。

2.3 建模样本集划分

当实验样本中存在异常时，模型精度通常会受到较大的影响，为解决异常样本带来的干扰，本研究通过Origin 2022软件建立箱型图剔除异常值，将剩余的土壤样本进行训练集和预测集的划分。Kennard-Stone（K-S）方法基于光谱变量间的欧氏距离，在特征空间中均匀地选取光谱差异较大的样品，选择的样本数达标即可（初始设置70%为训练集，30%为预测集），该算法可将训练集的一些样本转移到测试集，将测试集的一些样本转移到训练集，以此保证算法运行的随机性、稳定性之间的平衡[10]。本次利用K-S算法划分数据集，划分结果如表1所示。其中，训练样本35个，预测样本14个，由表1可知，训练样本的值完全包含预测样本的数据范围，经过K-S算法获取的数据集能够更好地进行土壤硒含量的预测。

表1 训练集和预测集土壤硒含量描述性统计

Table 1 Descriptive statistics of soil selenium content in training and prediction sets

Sample	Number of sample	Minimum value /(μg·g^{-1})	Maximum value /(μg·g^{-1})	Average /(μg·g^{-1})	Standard deviation
Training set	35	0.200	1.150	0.726	0.256
Prediction set	14	0.295	1.025	0.595	0.213

2.4 预测模型及精度评价

2.4.1 模型原理

PLSR模型包含主成分分析（PCA）、典型相关分析和普通多元线性回归3种方法的优点[11-12]。PLSR模型在土壤养分反演方面，能够克服自变量之间的多重共线性问题，使得预测结果具有较好的稳定性。

SVM是一种将数据通过非线性映射到高纬度空间并展开线性回归的方法[13-14]；该模型在处理小样本、局部最小点等问题上具有较大优势[15]。

RF算法最早来自Leo Breiman和Adele Cutler的理论研究。RF算法能够解决高维特征变量，在处理大数据方面，该模型训练速度较快，对模型参数的兼容能力较强[16]。其基本步骤如下：

步骤1 设训练样本个数为W，特征数目为Z；

步骤2 从W个训练样本中以bootstrap采样方式选择同样本容量的W个样本，对未选择的样本进行预测和误差评估；

步骤3 对每一节点进行随机特征选择，设定特征数目为z，以随机选择的特征为基础，进行最佳分裂方式的选择；

步骤4 在每个训练子集上构建决策树，依据每棵决策树的输出作为最终结果。

海鸥优化算法（SOA）是一种源自海鸥在自然界迁徙、攻击行为的新颖算法[17-18]。该算法主要利用两大搜索步骤进行运算，即全局搜索（迁徙）和局部搜索（攻击）。其中，全局搜索包括避免相互碰撞、确定最佳位置方向以及靠近最佳位置3个条件；局部搜索则是包含海鸥在迁徙过程中攻击速度和角度的不断变化[17]。在RF模型中，生成树的数目（ntree）和树的最大深度值（mdepth）会对模型的精度产生很大影响，故本次利用SOA算法优化RF的参数ntree和mdepth，利用最佳参数值来建立SOA-RF模型，SOA-RF算法具体步骤如下：设置SOA的种群数量为20、最大迭代次数为300，同时设置海鸥迁徙空间等参数值，对比当前迭代次数与最大迭代次数来控制算法的运行，前者不小于后者时模型为最佳状态，SOA-RF模型的运算步骤如图3所示。

2.4.2 模型精度评价

采用决定系数（R^2）、均方根误差（RMSE，f_{RMSE}）和相对分析误差（RPD，f_{RPD}）对预测模型进行精度评价。计算公式为

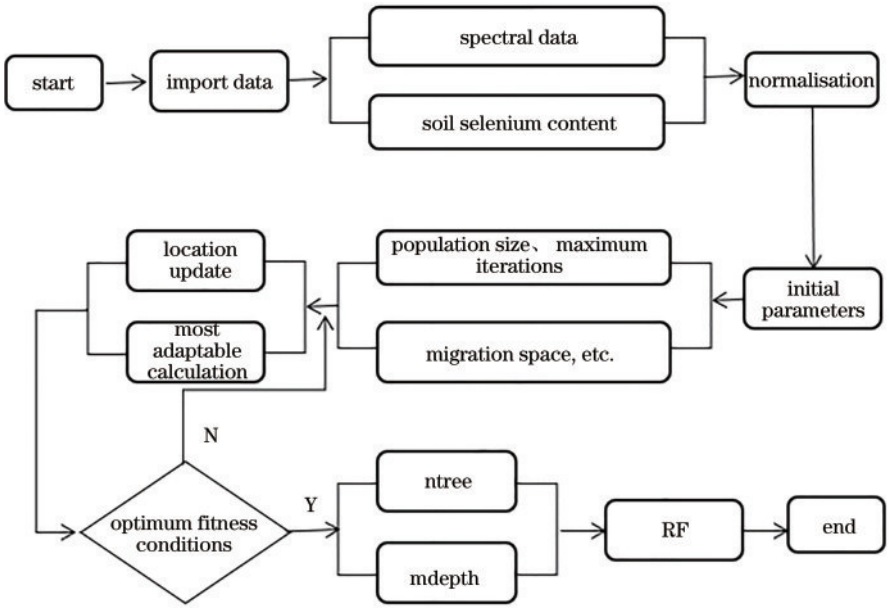

图 3　SOA-RF 流程图

Fig. 3　SOA-RF Flowchart

$$R^2=1-\frac{\sum\left(y-\hat{y}\right)^2}{\sum\left(y-\bar{y}\right)^2},\tag{1}$$

$$f_{\mathrm{RMSE}}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2},\tag{2}$$

$$f_{\mathrm{RPD}}=f_{\mathrm{SD}}/f_{\mathrm{RMSE}},\tag{3}$$

式中：y为实测值；$\hat{y}$模型预测值；$\bar{y}$为样本的平均值；m为样本数。R^2为预测模型对实测值的拟合程度（范围为0～1），R^2越接近于1、RMSE越小，模型的预测效果越好。f_{SD}为分析样本的标准偏差。$f_{\mathrm{RPD}}<1.4$代表模型不可靠；$1.4\leqslant f_{\mathrm{RPD}}\leqslant 2.0$代表模型较为可靠；$f_{\mathrm{RPD}}>2.0$代表模型具备较高可靠性，能够用于反演土壤硒含量。

3　结果与分析

3.1　光谱增强与相关性分析

通过计算原始光谱、MSC-FD变换光谱与土壤硒含量的皮尔逊相关系数，如图4所示。由图4可知，经MSC-FD光谱增强的光谱参数与硒含量的相关性更加明显，其中在1000～1500 nm、2000～2450 nm范围内，相关系数绝对值变化较大，光谱特征点明显增加，为提升特征波段筛选效率，将相关系数绝对值>0.5的波段进行特征选择。

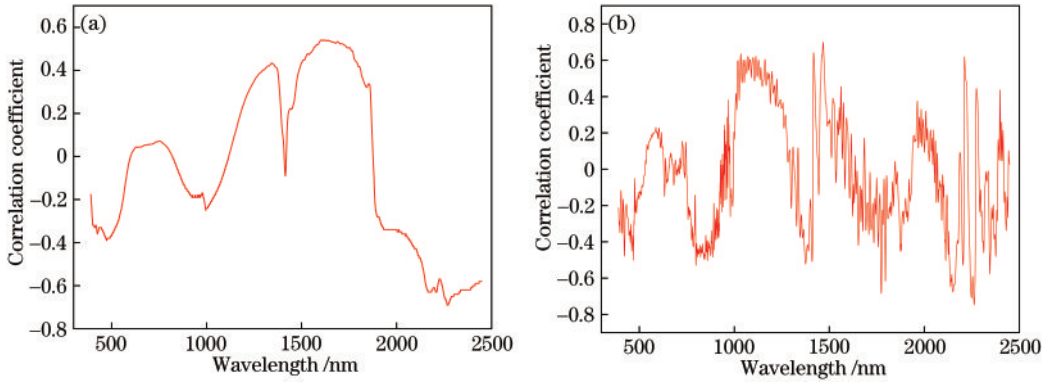

图 4　相关性分析。(a)原始光谱；(b)MSC-FD

Fig. 4　Correlation analysis. (a) Original spectra; (b) MSC-FD

3.2　特征波段选择结果分析

3.2.1　sCARS算法筛选结果

采用sCARS算法对经MSC-FD变化后的光谱进行特征波长筛选，波长筛选过程如图5所示。特征变量筛选过程中所保留的变量数变化趋势如图5(a)所示，由图5(a)可知，随着迭代次数的增加，保留的变量数也逐渐减少，由变化曲线的斜率可以判断出，变量数在经过快速的粗选后进入了精选阶段；在变量筛选过程中，交叉验证均方根误差（RMSECV）值的变化趋势如图5(b)所示，经过40次迭代之后，RMSECV值达到

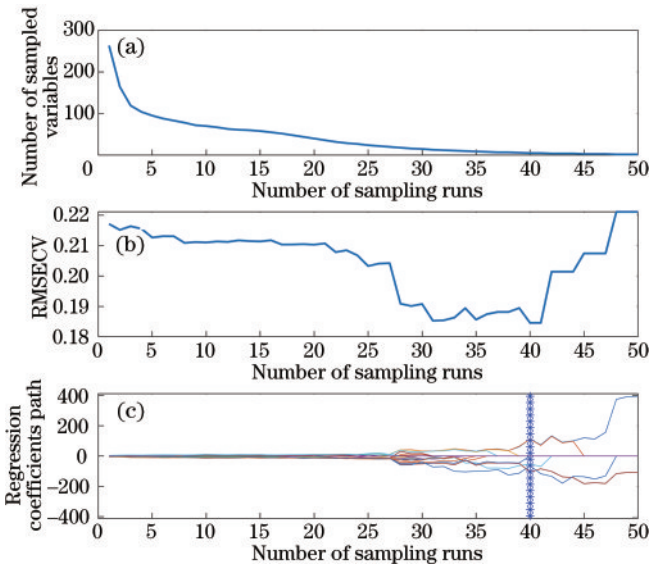

图 5　sCARS 算法特征波段筛选过程

Fig. 5　sCARS algorithm feature band screening process

最小，接着，RMSECV 值又进入了升高阶段，通过 40 次的迭代结果，确定了最佳的特征参数。最终经 sCARS 算法提取了 10 个特征波长，占全波段的 0.49%。

3.2.2　CARS、IRIV、VISSA、IVISSA、VCPA-GA 算法筛选结果

CARS 可将近红外波段中的任意变量作为单一值，通过 ARS 和 PLSR 进行交叉验证，不断更新交叉验证结果，以 RMSECV 值最小的波段组合作为筛选的特征参数[19-20]。本研究利用 CARS 算法共筛选出 14 个特征参数，占全波段的 0.68%。该算法速度快，但 CARS 算法与 sCARS 算法相比，其并不以变量的稳定性为基本准则，而在 sCARS 算法中，变量被选择的概率与该变量的稳定性呈正相关。CARS 算法仅筛选出回归系数最值的最大值的变量，多次循环迭代此过程，CARS 算法与 sCARS 算法不同，sCARS 算法可以筛选出回归系数最值的最大值，且稳定性高的变量，多次循环迭代此过程[21]，故 CARS 算法的稳定性较差。

迭代保留信息变量（IRIV）算法利用加权二进制矩阵采样法产生相关变量组合，基于各变量组合 RMSECV 值的差异性，并结合显著性检验，将所选参数分为强有信息变量、弱有信息变量、无信息变量和干扰变量等 4 种类型。在迭代过程中，不断剔除无信息变量和干扰变量，直至剩余强有信息变量和弱有信息变量，此循环过程停止[6,22]。因此，该算法的运行过程相对于其他算法较为缓慢。经 IRIV 算法，共筛选出特征参数 13 个，占全波段的 0.63%。

变量空间迭代收缩算法（VISSA）是基于模型集群分析策略（MPA）的特征筛选方法。VISSA 在运行过程中，每一步都会缩小空间，并且子模型的预测优于上一个，经过不断优化选择，最终以 RMSECV 值最小时的组合参数作为特征筛选结果[23]。本次通过 VISSA 共筛选出 65 个特征参数，占全波段的 3.17%（筛选过程的 RMSECV 值与 Weight 值变化如图 6 所示）。

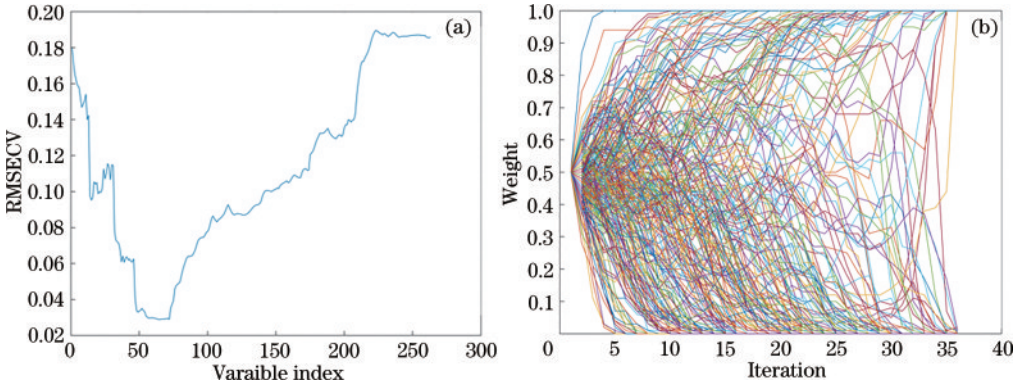

图 6　VISSA 算法特征波段筛选过程。(a) RMSECV 值；(b) Weight 值

Fig. 6　VISSA algorithm feature band screening process. (a) RMSECV values; (b) weight values

区间变量迭代空间收缩法（IVISSA）是邓百川等[24]在 VISSA 的基础上开发出来的一种特征参数选择算法。IVISSA 算法是利用各波段的权重进行整体和局部的优化，在确定各波段的间隔位置和宽度的前提下，采用 RMSECV 作为交叉验证的评价指标，用以筛选含有用信息的波段，通过不断迭代优化，选出与化学元素敏感性最佳的波段作为特征参数。经 IVISSA 算法共筛选出 73 个特征波段，占全波段的 3.56%。

变量组合集群分析法（VCPA）是通过指数衰减函数（EDF）不断缩小和优化参数空间，在数据量较少的变量空间中找出最优的变量子集[25]。GA 是一种自适应的全局概率搜索算法，根据遗传机制和自然选择，通过选择、交叉和变异算法的操作，随着迭代过程的进行，将目标函数值较优的变量保留，将目标函数值较差的变量剔除，最终获得最优的特征参数[7]。此次利用变量组合集群分析法结合遗传算法（VCPA-GA）共筛选出特征变量 42 个，占全波段的 2.08%。

特征筛选算法均在 Matlab R2021a 软件中进行。经过 sCARS、CARS、IRIV、VISSA、IVISSA、VCPA-GA 算法筛选出的特征波段分布情况如图 7 所示。由

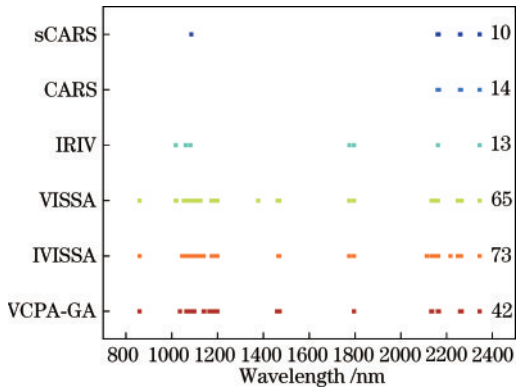

图 7　不同算法提取特征波长分布图

Fig. 7　Distribution of wavelength of features extracted by different algorithms

图 7 可知，所有算法在 2100～2300 nm 范围均提取了部分波长，其中 sCARS 和 CARS 算法更多地保留了 2100～2300 mm 波段的特征点，剔除了其他波段范围的波长点，sCARS 算法获得的变量数最少。

3.3　模型建立与分析

3.3.1　PLSR 模型

将经过 sCARS、CARS、IRIV、VISSA、IVISSA、VCPA-GA 算法筛选出的特征波段和未经筛选的全波段作为自变量，以 49 个土壤样本硒含量作为因变量建立 PLSR 回归模型，通过对比不同特征下模型的 R^2 和 RMSE，以此进行精度评价。模型的预测结果如表 2 所示，由表 2 可知，经特征筛选后的预测模型精度较全波段的模型精度有了很大提高，从验证集的 R^2、RMSE 以及特征筛选的时间进行分析，发现使用 sCARS 算法（R^2=0.49、RMSE 为 0.19）提取特征波段能降低模型的复杂性，也能加快程序运行的速度，预测模型的整体效率有很大的提升。

表 2　不同变量筛选的 PLSR 模型精度评价

Table 2　Evaluation of accuracy of the PLSR model with different variable screenings

Spectral transform	Time /s	Number of variable	Training R^2	Training RMSE	Validation R^2	Validation RMSE
Full spectra	–	2051	0.43	0.21	0.37	0.35
sCARS	＜10.00	10	0.52	0.17	0.49	0.19
CARS	＜10.00	14	0.52	0.19	0.40	0.16
IRIV	145.60	13	0.65	0.15	0.49	0.18
VISSA	51.96	65	0.49	0.19	0.48	0.16
IVISSA	27.11	73	0.50	0.15	0.45	0.20
VCPA-GA	145.53	42	0.54	0.15	0.51	0.19

3.3.2　SVM 模型

建立 sCARS-SVM、CARS-SVM、IRIV-SVM、VISSA-SVM、IVISSA-SVM、VCPA-GA-SVM 以及全波段的 SVM 回归模型，通过对比不同特征筛选下模型的 R^2 和 RMSE，寻找最佳的预测模型。模型的预测结果如表 3 所示，由表 3 可知，经特征筛选后的预测模型精度较全波段的模型精度有很大提高，从验证集的 R^2、RMSE 以及特征筛选的时间进行分析，各个模型的 R^2 从大到小依次为：sCARS-SVM、IVISSA-SVM、CARS-SVM、IRIV-SVM、VCPA-GA-SVM、VISSA-SVM，其中 sCARS-SVM 模型的 R^2 最大（R^2=0.57）；IRIV-SVM 模型的 RMSE 最小，但该模型的运行时间较长，其预测效率大大降低。sCARS 算法可以很好地结合 SVM 模型，有效提升模型的反演精度。

表 3　不同变量筛选的 SVM 模型精度评价

Table 3　Evaluation of accuracy of SVM models with different variable screenings

Spectral transform	Time /s	Number of variable	Training R^2	Training RMSE	Validation R^2	Validation RMSE
Full spectra	–	2051	0.57	0.05	0.34	0.66
sCARS	＜10.00	10	0.59	0.11	0.57	0.14
CARS	＜10.00	14	0.58	0.13	0.55	0.12
IRIV	145.60	13	0.66	0.13	0.55	0.11
VISSA	51.96	65	0.77	0.13	0.53	0.13
IVISSA	27.11	73	0.96	0.01	0.56	0.39
VCPA-GA	145.53	42	0.56	0.11	0.54	0.14

3.3.3　RF 模型

以土壤硒含量为因变量，全波段和经特征提取后的参数为自变量，建立土壤硒含量的 RF 预测模型，不同变量选择下的 RF 模型预测结果如表 4 所示，各模型的预测集 R^2 均大于全波段 RF 模型，具体为 IRIV-RF 模型验证集的 R^2 最高（R^2=0.68），另外，在波段筛选时间上，sCARS-RF 模型运行时间低于 10 s，而 IRIV-RF 模型在同样环境下需耗时 145.60 s。综上，sCARS-RF 预测模型具有极其显著的优势。

3.3.4　SOA-RF 模型

将训练样本和预测样本引入 SOA-RF 模型，进行回归预测，SOA 优化过程的收敛曲线如图 8 所示，当迭代次数为 65 时，模型已收敛。

不同变量选择下的 SOA-RF 模型预测结果如表 5

表 4　不同变量筛选的 RF 模型精度评价

Table 4　Evaluation of accuracy of the RF model with different variable screenings

Spectral transform	Time /s	Number of variable	Training R^2	Training RMSE	Validation R^2	Validation RMSE
Full spectra	–	2051	0.85	0.19	0.47	0.18
sCARS	＜10.00	10	0.82	0.09	0.66	0.13
CARS	＜10.00	14	0.62	0.16	0.49	0.17
IRIV	145.60	13	0.86	0.06	0.68	0.15
VISSA	51.96	65	0.86	0.05	0.65	0.16
IVISSA	27.11	73	0.81	0.06	0.64	0.15
VCPA-GA	145.53	42	0.69	0.11	0.59	0.19

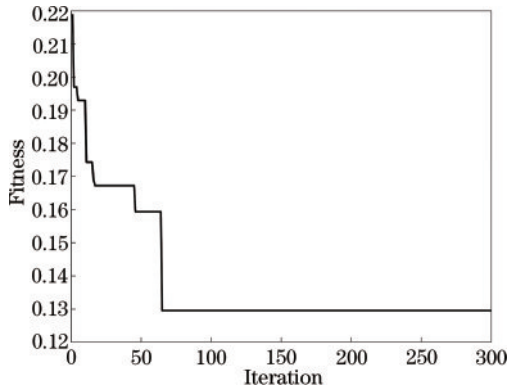

图 8　SOA-RF 算法收敛曲线

Fig. 8　SOA-RF algorithm convergence curve

所示，各模型的预测集 R^2 均大于全波段的 SOA-RF 模型，从大到小依次为：VCPA-GA-SOA-RF、sCARS-SOA-RF、VISSA-SOA-RF、IVISSA-SOA-RF、IRIV-SOA-RF、CARS-SOA-RF，虽然 VCPA-GA-SOA-RF 模型验证集的 R^2（R^2=0.92）大于 sCARS-SOA-RF 模型的 R^2（R^2=0.90），但 sCARS-SOA-RF 模型的 RMSE（RMSE 为 0.06）小于 VCPA-GA-SOA-RF 模型的 RMSE（RMSE 为 0.08），另外，在波段筛选时间上，sCARS-SOA-RF 模型运行时间低于 10 s，而 VCPA-GA-SOA-RF 模型在同样环境下需耗时 145.53 s，极大地影响了模型的预测效率。鉴于 SOA-RF 模型具有较好的预测效果，故进行对比各组合模型的 RPD，如图 9 所示。由图 9 可知，SOA-RF 模型整体效果均较好，所有组合模型的 RPD 均超过了 1.4，可信度较高，其中 VCPA-GA-SOA-RF 模型具有最高的 RPD（RPD 为 2.911）。经过对比模型验证集的 R^2、RMSE 以及波段筛选的时间，最终确定 sCARS 算法，减少了数据冗余，降低了模型复杂度，在模型精度和预测效率上都有较大优势。

表 5　不同变量筛选的 SOA-RF 模型精度评价

Table 5　Evaluation of accuracy of the SOA-RF model with different variable screenings

Spectral transform	Time /s	Number of variable	Training R^2	Training RMSE	Validation R^2	Validation RMSE
Full spectra	–	2051	0.85	0.16	0.75	0.15
sCARS	＜10.00	10	0.93	0.07	0.90	0.06
CARS	＜10.00	14	0.91	0.08	0.80	0.11
IRIV	145.60	13	0.86	0.10	0.81	0.11
VISSA	51.96	65	0.94	0.07	0.84	0.10
IVISSA	27.11	73	0.94	0.06	0.83	0.09
VCPA-GA	145.53	42	0.94	0.07	0.92	0.08

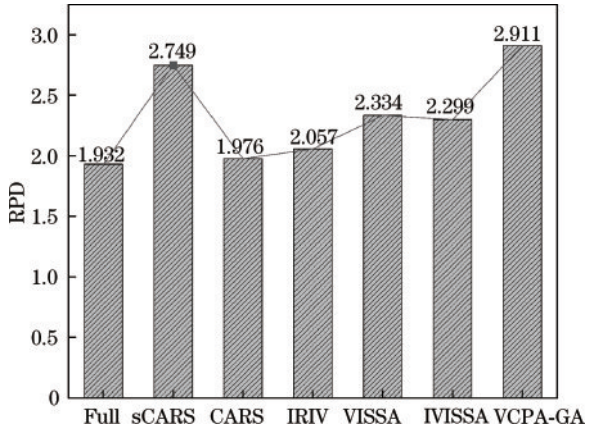

图 9　SOA-RF 模型预测结果（RPD）

Fig. 9　SOA-RF model prediction results(RPD)

4　讨　　论

由于土壤硒在光谱波段上并无代表性的特征波段，需借助特征提取手段来寻找光谱与硒含量之间的化学联系，故进行了不同特征提取下的建模分析，以此增加模型的反演精度。通过对比各模型的反演精度可以发现，sCARS、VCPA-GA 等算法能够很好地与 PLSR、SVM、RF、SOA-RF 模型相结合，用以进行土壤硒含量的高光谱预测，其中，RF、SVM 模型与 PLSR 相比有着更高的精度，说明非线性模型能够更好地体现特征参数的化学意义，土壤硒含量的高光谱反演更适合非线性模型，SOA 成功实现对 RF 模型的参数优化，使模型的精度大幅增加。预测集 VCPA-GA-

SOA-RF 模型的 R^2 最高（R^2=0.92），说明 VCPA-GA 算法提取的特征波长包含了更多的有用信息，sCARS 算法造成了部分有用信息的缺失，其他算法提取的特征参数过多，数据产生较大冗余，使模型精度降低，但 VCPA-GA 算法所耗费时间较长，所选特征参数比 sCARS 算法多，模型的复杂程度较大，故 sCARS-SOA-RF 模型，在土壤硒含量的高光谱反演中，具有极其显著的优势。各模型的最优反演结果如图 10 所示。

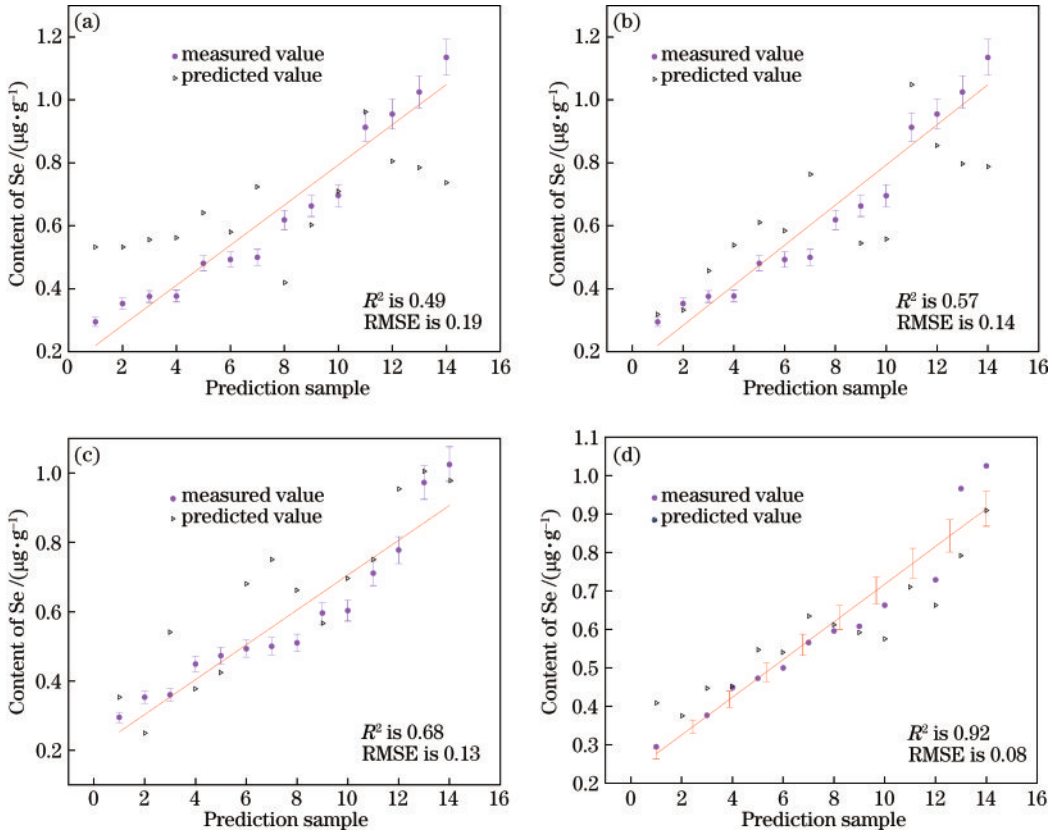

图 10　最优模型实测值与预测值散点图。(a) sCARS-PLSR；(b) IRIV-SVM；(c) sCARS-RF；(d) VCPA-GA-SOA-RF

Fig. 10　Scatter plot of measured and predicted values of the optimal model. (a) sCARS-PLSR; (b) IRIV-SVM; (c) sCARS-RF; (d) VCPA-GA-SOA-RF

5　结　　论

针对土壤硒含量的高光谱反演研究，采用 SG 平滑、多元散射校正一阶微分（MSC-FD）对光谱数据进行预处理，通过 sCARS、CARS、IRIV、VISSA、IVISSA、VCPA-GA 算法筛选特征波长，建立全波段和特征波段的线性模型（PLSR）和非线性模型（SVM、RF），同时利用 SOA 优化 RF 模型的参数 ntree 和 mdepth，经过对比各模型预测精度，寻找土壤硒含量的最佳预测模型，主要得到以下结论：

1）土壤硒含量与原始光谱反射率的相关性较弱，通过 SG-MSC-FD 预处理后，土壤硒含量对光谱反射率的敏感性明显增强，最大相关系数的绝对值从 0.69 提升到了 0.75。

2）特征波段提取对于模型的精度影响较大，sCARS、CARS、IRIV、VISSA、IVISSA、VCPA-GA 算法都不同程度地减少了数据冗余，降低了模型复杂度，提升了模型精度。sCARS 模型与其他模型相比，在运行时间、变量个数、模型精度上都具有更加显著的优势。

3）RF 模型与 SVM 和 PLSR 模型相比，具有更好的鲁棒性，能够更好地体现光谱波段的化学意义，经过 SOA 优化后的 RF 模型，精度有着极大提高。整体来看，VCPA-GA-SOA-RF 模型预测精度最高，sCARS-SOA-RF 模型预测效率最高。经过优化后的 RF 模型可以实现土壤硒含量高效、无损、大面积估测。

4）目前所使用的高光谱数据是基于实验室实测光谱，没有将影像光谱数据结合起来建模分析，故缺乏一定的普适性。在后续研究中，希望能够结合遥感影像与实测光谱数据综合建模；另外，SOA-RF 模型，在不同特征提取下的精度差异明显，也希望能够在近红外与可见光部分分别提取特征波段，进行建模分析，以此尝试提升模型精度。

参　考　文　献

[1] 廖启林，崔晓丹，黄顺生，等. 江苏富硒土壤元素地球化学特征及主要来源[J]. 中国地质，2020，47(6)：1813-1825.
Liao Q L，Cui X D，Huang S S，et al. Element

geochemistry of selenium-enriched soil and its main sources in Jiangsu Province[J]. Geology in China, 2020, 47(6): 1813-1825.

[2] 陈东平, 张金鹏, 聂合飞, 等. 粤北山区连州市土壤硒含量分布特征及影响因素研究[J]. 环境科学学报, 2021, 41(7): 2838-2848.
Chen D P, Zhang J P, Nie H F, et al. Selenium distribution in soils of Lianzhou City, mountain area of northern Guangdong Province and its influencing factors [J]. Acta Scientiae Circumstantiae, 2021, 41(7): 2838-2848.

[3] 迟凤琴, 徐强, 匡恩俊, 等. 黑龙江省土壤硒分布及其影响因素研究[J]. 土壤学报, 2016, 53(5): 1262-1274.
Chi F Q, Xu Q, Kuang E J, et al. Distribution of selenium and its influencing factors in soils of Heilongjiang Province, China[J]. Acta Pedologica Sinica, 2016, 53 (5): 1262-1274.

[4] 王晓丽, 张泽洲, 王张民, 等. 江西宜春市明月山地区土壤和多种作物中硒的含量及形态分布特征[J]. 科学通报, 2022, 67(6): 511-519.
Wang X L, Zhang Z Z, Wang Z M, et al. Content and speciation distribution of selenium in soil and crops in Mingyue Mountain area of Yichun City, Jiangxi Province [J]. Chinese Science Bulletin, 2022, 67(6): 511-519.

[5] 张东辉, 赵英俊, 赵宁博, 等. 一种间接从高光谱数据中提取黑土硒含量的新方法[J]. 光谱学与光谱分析, 2019, 39(7): 2237-2243.
Zhang D H, Zhao Y J, Zhao N B, et al. A new indirect extraction method for selenium content in black soil from hyperspectral data[J]. Spectroscopy and Spectral Analysis, 2019, 39(7): 2237-2243.

[6] 程介虹, 陈争光. 基于迭代保留信息变量和连续投影的近红外光谱波长选择方法[J]. 分析化学, 2021, 49(8): 1402-1409.
Cheng J H, Chen Z G. Wavelength selection method for near infrared spectroscopy based on iteratively retains informative variables and successive projections algorithm [J]. Chinese Journal of Analytical Chemistry, 2021, 49 (8): 1402-1409.

[7] 乔天, 吕成文, 肖文凭, 等. 基于遗传算法的土壤质地高光谱预测模型研究[J]. 土壤通报, 2018, 49(4): 773-778.
Qiao T, Lü C W, Xiao W P, et al. Hyperspectral prediction modeling of soil texture based on genetic algorithm[J]. Chinese Journal of Soil Science, 2018, 49 (4): 773-778.

[8] 张爱武, 董喆, 康孝岩. 基于XGBoost的机载激光雷达与高光谱影像结合的特征选择算法[J]. 中国激光, 2019, 46(4): 0404003.
Zhang A W, Dong Z, Kang X Y. Feature selection algorithms of airborne LiDAR combined with hyperspectral images based on XGBoost[J]. Chinese Journal of Lasers, 2019, 46(4): 0404003.

[9] 刘国海, 夏荣盛, 江辉, 等. 一种基于SCARS策略的近红外特征波长选择方法及其应用[J]. 光谱学与光谱分析, 2014, 34(8): 2094-2097.
Liu G H, Xia R S, Jiang H, et al. A wavelength selection approach of near infrared spectra based on SCARS strategy and its application[J]. Spectroscopy and Spectral Analysis, 2014, 34(8): 2094-2097.

[10] 黄照强, 倪斌. 基于随机变异-Kennard-Stone和偏最小二乘法的土壤重金属镉含量反演: 以雄安新区西南部为例[J]. 地质论评, 2021, 67(5): 1521-1532.
Huang Z Q, Ni B. Retrieval of soil heavy metal Cadmium content based on Random Mutation, Kennard-Stone and partial least squares method: a case study of southwest of Xiong'an New District[J]. Geological Review, 2021, 67(5): 1521-1532.

[11] 金浩哲, 叶浩杰, 偶国富, 等. 基于偏最小二乘法的加氢换热器NH_4Cl结晶温度预测模型[J]. 石油学报(石油加工), 2017, 33(6): 1176-1182.
Jin H Z, Ye H J, Ou G F, et al. Predicting model of ammonium salt crystallization temperature based on partial least squares approach in a hydrogenation heat-exchanger[J]. Acta Petrolei Sinica (Petroleum Processing Section), 2017, 33(6): 1176-1182.

[12] 尚栋, 孙兰香, 齐立峰, 等. 基于循环变量筛选非线性偏最小二乘的LIBS铁矿浆定量分析[J]. 中国激光, 2021, 48(21): 2111001.
Shang D, Sun L X, Qi L F, et al. Quantitative analysis of laser-induced breakdown spectroscopy iron ore slurry based on cyclic variable filtering and nonlinear partial least squares[J]. Chinese Journal of Lasers, 2021, 48 (21): 2111001.

[13] 郭云开, 张思爱, 谢晓峰, 等. 基于GA-SVM的耕地土壤重金属含量高光谱反演方法的研究[J]. 土壤通报, 2021, 52(4): 968-974.
Guo Y K, Zhang S A, Xie X F, et al. The hyperspectral inversion method of heavy metal contents in cultivated soils based on GA-SVM[J]. Chinese Journal of Soil Science, 2021, 52(4): 968-974.

[14] 陈鹏, 齐超, 刘人玮, 等. 基于支持向量机回归的LIBS飞灰含碳量定量分析[J]. 光学学报, 2022, 42(9): 0930003.
Chen P, Qi C, Liu R W, et al. Quantitative analysis of carbon content in fly ash using LIBS based on support vector machine regression[J]. Acta Optica Sinica, 2022, 42(9): 0930003.

[15] 乔守旭, 钟文义, 谭思超, 等. 基于PCA-GA-SVM的竖直下降两相流流型预测[J]. 核动力工程, 2022, 43(3): 85-93.
Qiao S X, Zhong W Y, Tan S C, et al. Prediction of vertical-downward two-phase flow pattern based on PCA-GA-SVM[J]. Nuclear Power Engineering, 2022, 43(3): 85-93.

[16] 李冠稳, 高小红, 肖能文, 等. 基于sCARS-RF算法的高光谱估算土壤有机质含量[J]. 发光学报, 2019, 40(8): 1030-1039.
Li G W, Gao X H, Xiao N W, et al. Estimation soil organic matter contents with hyperspectra based on sCARS and RF algorithms[J]. Chinese Journal of Luminescence, 2019, 40(8): 1030-1039.

[17] Dhiman G, Singh K K, Soni M, et al. MOSOA: a new multi-objective seagull optimization algorithm[J]. Expert

Systems With Applications, 2021, 167: 114150.

[18] Dhiman G, Singh K K, Slowik A, et al. EMoSOA: a new evolutionary multi-objective seagull optimization algorithm for global optimization[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(2): 571-596.

[19] 江晓宇, 李福生, 王清亚, 等. X射线荧光光谱结合CARS变量筛选选择方法用于土壤中铅砷含量的测定[J]. 光谱学与光谱分析, 2022, 42(5): 1535-1540.
Jiang X Y, Li F S, Wang Q Y, et al. Determination of lead and arsenic in soil samples by X fluorescence spectrum combined with CARS variables screening method[J]. Spectroscopy and Spectral Analysis, 2022, 42(5): 1535-1540.

[20] 路皓翔, 张静, 李灵巧, 等. 最小角回归结合竞争性自适应重加权采样的近红外光谱波长选择[J]. 光谱学与光谱分析, 2021, 41(6): 1782-1788.
Lu H X, Zhang J, Li L Q, et al. Least angle regression combined with competitive adaptive re-weighted sampling for NIR spectral wavelength selection[J]. Spectroscopy and Spectral Analysis, 2021, 41(6): 1782-1788.

[21] 李冠稳, 高小红, 肖能文, 等. 特征变量选择和回归方法相结合的土壤有机质含量估算[J]. 光学学报, 2019, 39(9): 0930002.
Li G W, Gao X H, Xiao N W, et al. Estimation of soil organic matter content based on characteristic variable selection and regression methods[J]. Acta Optica Sinica, 2019, 39(9): 0930002.

[22] 纪然仕, 陈晓燕, 刘素珍, 等. 基于高光谱技术和IRIV-FOA-ELM算法的花椒挥发油无损检测[J]. 激光与光电子学进展, 2020, 57(20): 203002.
Ji R S, Chen X Y, Liu S Z, et al. Nondestructive testing of volatile oil of zanthoxylum Bungeanum based on hyperspectral technique and IRIV-FOA-ELM algorithm [J]. Laser & Optoelectronics Progress, 2020, 57(20): 203002.

[23] Deng B C, Yun Y H, Liang Y Z, et al. A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling[J]. The Analyst, 2014, 139(19): 4836-4845.

[24] Deng B C, Yun Y H, Ma P, et al. A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals[J]. The Analyst, 2015, 140(6): 1876-1885.

[25] Yun Y H, Wang W T, Deng B C, et al. Using variable combination population analysis for variable selection in multivariate calibration[J]. Analytica Chimica Acta, 2015, 862: 14-23.