特征变量选择和回归方法相结合的土壤有机质含量估算

李冠稳[1,2**]，高小红[1*]，肖能文[2]，肖云飞[1]

[1]青海师范大学地理科学学院，青海省自然地理与环境过程重点实验室，青海 西宁 810008；

[2]中国环境科学研究院，北京 100012

摘要 针对高光谱数据量大、信息冗余严重的现象，应用稳定竞争性自适应重加权采样（sCARS）、连续投影算法（SPA）、遗传算法（GA）、迭代保留有效信息变量（IRIV）和稳定竞争性自适应重加权采样结合连续投影算法（sCARS-SPA），从全波段光谱数据中筛选特征变量，并利用全波段和特征波段建立偏最小二乘回归（PLSR）、支持向量机（SVM）和随机森林（RF）模型预测土壤有机质含量。结果表明，PLSR 和 SVM 模型结合特征变量选择，不仅提高了模型运算效率，而且模型预测能力较全波段均有一定提高；RF 模型采用特征变量建模，对模型精度的提高不是十分明显，但其构建模型的变量数量却显著减少，大大提高建模效率。RF 模型精度优于 SVM 和 PLSR 模型，IRIV 结合 RF 建立的土壤有机质含量预测模型，变量数仅 63 个，校准集和验证集模型决定系数（R^2）分别为 0.941和 0.96，验证集相对分析误差（R_{PD}）为 4.8。与全波段建模相比，特征变量选择和回归方法相结合，在保证模型精度的同时，可有效提高建模效率。

关键词 光谱学；土壤有机质含量；特征变量选择；回归模型

中图分类号 TP79；S151.9 **文献标识码** A **doi**：10.3788/AOS201939.0930002

Estimation of Soil Organic Matter Content Based on Characteristic Variable Selection and Regression Methods

Li Guanwen[1,2**], Gao Xiaohong[1*], Xiao Nengwen[2], Xiao Yunfei[1]

[1] *Qinghai Provincial Key Laboratory of Physical Geography and Environmental Process,*
School of Geography Sciences,
Qinghai Normal University, Xining, Qinghai 810008, *China*;

[2] *Chinese Research Academy of Environmental Sciences, Beijing* 100012, *China*

Abstract In view of the large amount of soil hyperspectral data and obvious spectral information redundancy, this paper aims to compare prediction abilities of multiple feature variable selection methods for estimating soil organic matter. The stability competitive adaptive reweighted sampling (sCARS), successive projections algorithm (SPA), genetic algorithm (GA), iteratively retained information variables (IRIV), and sCARS-SPA are used to select the characteristic variables from full spectral data. Based on these characteristic bands and full spectral bands, partial least squares regression (PLSR), support vector machine (SVM), and random forest (RF) models are used to predict the soil organic matter content. The results show that the PLSR and SVM models combined with variable selection can not only improve the efficiency of the model, but also improve the model prediction ability over the full bands. The accuracy of RF model constructed with characteristic variables is not obviously improved, but the variable number in the construction model is significantly reduced and the modeling efficiency is greatly improved. Overall, the RF model's accuracy is better than those of the SVM model and the PLSR model. The variable number of the prediction model from the combination of IRIV and RF is only 63, and the coefficients of determination (R^2) from calibration set and validation set are respectively 0.941 and 0.96, and the relative deviation for the validation set R_{PD} is 4.8, showing a very good prediction capacity. Compared to modeling based on the full bands, the combination of characteristic variable selection and regression methods can effectively improve the modeling efficiency while ensuring the accuracy of the model.

收稿日期：2019-03-05；**修回日期**：2019-03-30；**录用日期**：2019-05-05

基金项目：国家自然科学基金(41550003)

* **E-mail**：xiaohonggao226@163.com；** **E-mail**：lgw126522@163.com

Key words spectroscopy; soil organic matter content; characteristic variable selection; regression model
OCIS codes 300.6170; 150.1135; 200.4560; 280.4991

1 引 言

土壤有机质(SOM)含量是衡量土壤质量的一个重要指标,尽管有机质仅占土壤总量的很小一部分,但在促进植物生长发育、改善土壤物理性质等多方面的作用显著[1-2]。高光谱遥感能够快速、大范围、无损地获取土壤信息[3-4],但高光谱包含几百乃至上千个变量,其中有些变量与待测样品信息并无关系,因此对原始光谱进行特征变量或敏感波段选择,研究其是否能够替代全波段,获得较高的预测精度,减少模型工作量及提高模型效率非常有意义。早期如刘焕军等[5]以松嫩平原的黑土、草甸土、黑钙土等耕层土壤有机质含量为研究对象,采用相关分析法得出土壤有机质含量的敏感波段为520 nm。卢艳丽等[6]对东北平原的黑土和潮土光谱采用多元线性回归方法提取特征波段,得出土壤有机质含量的敏感波段为550～830 nm。近年来竞争性自适应重加权(CARS)、稳定竞争性自适应重加权(sCARS)、连续投影算法(SPA)、遗传算法(GA)和迭代保留有效信息变量(IRIV)等变量选择方法被用于土壤有机质估算研究[7]。朱亚星等[8]利用去除有机质实验结合无信息量消除法(UVE)和CARS变量筛选方法得出土壤有机质的敏感波段为561～721 nm与1920～2280 nm。

筛选土壤有机质含量的光谱敏感波段是简化模型和提高模型预测能力的关键。Thielebruhn等[9]对连续小波变换后的可见光-近红外(Vis-NIR)光谱采用CARS进行变量选择,结合偏最小二乘回归(PLSR)模型预测SOM含量,决定系数R^2从0.81提高到0.93。林志丹等[10]对采自涡阳县的130个砂姜黑土土壤样本预测其SOM含量,采用GA提取23个特征变量建立主成分回归模型,决定系数R^2达到0.93。以上研究表明基于变量选择算法挑选特征变量建模,使得模型精度和稳健性提高。

除了光谱变量,预测模型的选择对于Vis-NIR光谱的预测精度也至关重要[11]。Viscarra Rossel等[12]通过对比PLSR模型与多种数据挖掘算法如SVM和RF模型,发现数据挖掘算法精度高于PLSR模型。RF模型不仅有较快的拟合速度,而且对异常值和噪声的敏感度更低,稳健性更好,在模型拟合能力方面优于其他算法[13]。葛翔宇等[14]基于CARS耦合机器学习预测土壤含水量,相较于线性模型,其决定系数R^2从0.617提高到0.918。不同变量选择算法挑选的特征变量并不相同,这会影响模型精度,目前文献中多数为多种变量选择方法结合线性模型,与SVM、RF模型相结合的研究并不多见。因此,需要将多种变量选择方法与多种回归方法相结合对土壤属性进行预测,为设计更高效率、便携性更好的光谱仪器提供理论依据。

本文以青海省湟水流域401个表层土样为研究对象,原始光谱经预处理后,采用sCARS、SPA、GA、IRIV和sCARS-SPA挑选光谱特征波段,并基于全波段和特征波段分别建立PLSR、SVM和RF预测模型,为利用Vis-NIR光谱分析技术快速无损地估测农田土壤有机质含量提供方法支持。

2 数据与方法

2.1 研究区概况

湟水流域位于青海省东北部,为青藏高原一个特殊自然地理单元和生态屏障。地理位置介于36°02′—37°28′N,100°42′—103°04′E之间,流域面积为1.62×10^4 km^2。流域地势整体西北高,东南低,东西长,南北窄,海拔(DEM)在1655～4860 m之间。气候为高原干旱、半干旱大陆性气候。湟水流域土壤类型以栗钙土、黑钙土、灰钙土、山地草甸土、高山草甸土为主,土壤肥沃,为青海省主要粮食生产基地,主要种植春小麦、油菜、马铃薯、青稞、燕麦和玉米等农作物。

2.2 土壤样品采集

分别于2015、2016年10—11月进行土壤采样,此时庄稼已收割完毕,共采集401个湟水流域表层(0～20 cm)土壤样品。考虑到土壤有机质易受到坡度、坡向、土壤母质等自然因素,耕种、施肥灌溉管理措施等人为因素的影响,河谷区域的水浇地及地形平坦区域采用“梅花状”5点采样,坡耕地采用“S”型7点采样,采样点分布见图1。将土样装入密封袋中编号,并于室内自然风干、研磨、过100目筛(150 μm),将过筛后的土壤样品分为两份,分别用于有机质含量测试分析和土壤光谱数据采集。有机质含量采用重铬酸钾-外加热法测定,并应用浓度梯度法确定建模校准集与验证集样本,即将401个土

样有机质含量从高到低排序，按 2∶1比例划分校准集和验证集样本。土壤有机质含量(质量分数)最大值(Max)、最小值(Min)、平均值(Mean)和标准差值(SD)统计见表 1。

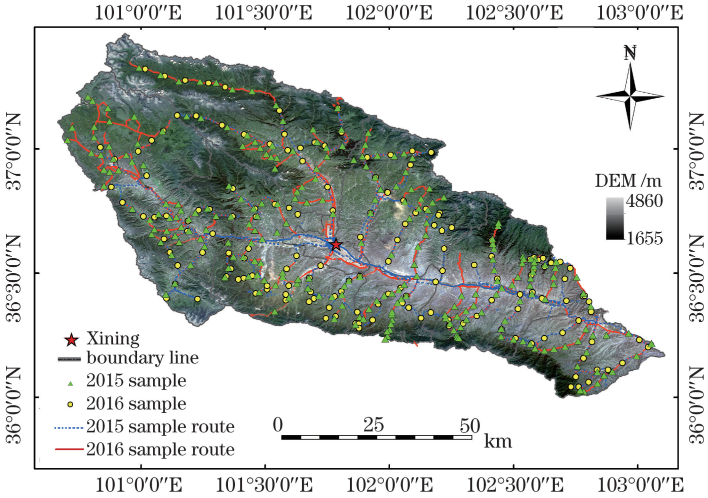

图 1　研究区位置及采样点分布图

Fig. 1　Location of the study area and distribution of soil sampling sites

表 1　校准集和验证集土壤有机质含量统计表

Table 1　Soil organic matter content statistics of calibration set and validation set

Sample set	Samplenumber	Min /(g•kg^{-1})	Max /(g•kg^{-1})	Mean /(g•kg^{-1})	SD
Calibration set	268	4.86	148.74	32.47	23.52
Validation set	133	8.26	133.56	32.16	22.44

2.3　土壤光谱数据采集及预处理

土壤样品的 Vis-NIR 光谱数据采集设备为美国 ASD FieldSpec 4 光谱仪，光谱范围为 350～2500 nm，室内土壤光谱测量参照文献[15]。去除边缘噪声较大的 350～400 nm 和 2401～2500 nm，并参考文献[16]将土壤有机质含量分为高、中、低和非常低 4 类，分别为大于 30 g•kg^{-1}、23～30 g•kg^{-1}、12～22 g•kg^{-1}和小于 12 g•kg^{-1}。土壤样品原始光谱曲线如图 2(a)所示，在 401～1100 nm 波长区域，光谱反射率值随波长变化增加比较明显，且不同等级有机质含量的平均光谱反射率值差异较大；而在 1400～2400 nm 波长区域，光谱反射率值随波长变化增加平缓，且不同等级有机质含量的反射率曲线差别不明显。

光谱采集过程中易受仪器噪声、土壤颗粒分布不均匀及测量随机误差等因素干扰，使所测样品光谱中含有光谱噪声，影响预测模型精度[17-18]，故采用多元散射校正(MSC)、中值滤波(MF)和一阶微分(1st Derivative)对原始光谱依次进行预处理。图 2(b)为经MSC-MF-1st Derivative预处理后的光

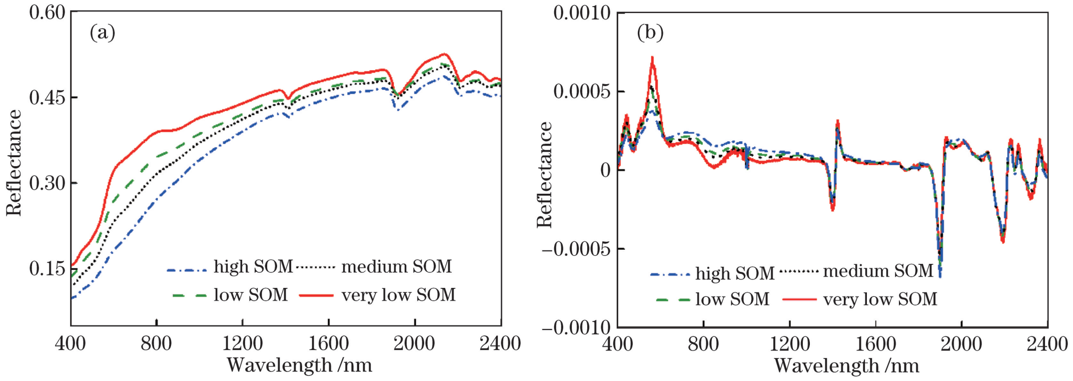

图 2　土壤样品反射率曲线。(a) 原始光谱；(b) MSC-MF-1st Derivative 预处理光谱

Fig. 2　Spectral reflectance curves of soil samples. (a) Raw spectra; (b) spectra after MSC-MF-1st Derivative pre-processing

谱反射曲线。从图中可以看出，光谱预处理后400～900 nm区域不同等级有机质含量光谱曲线之间差异性减小，光谱细节特征更加明显。

2.4 特征变量选择方法

2.4.1 sCARS 变量选择

sCARS以变量的稳定性作为变量选择衡量指标，增强了变量选择的稳定性，并延续CARS变量筛选流程[19]。具体步骤如下：

1）计算每个波长变量的稳定性值 C_i，C_i 定义为

$$C_i = \left| \frac{\bar{b}_i}{s(b_i)} \right| \quad i = 1,2,\cdots,P, \tag{1}$$

式中：C_i 为 M 次蒙特卡罗采样中第 i 个变量的稳定性值，$\bar{b}_i$ 和 $s(b_i)$ 分别为 M 次采样中第 i 个变量的均值和标准偏差，P 为变量数。显然，$\bar{b}_i$ 值越大，$s(b_i)$ 值越小，第 i 个变量的重要性越强；

2）使用强制波长选择和自适应性重加权采样方法（ARS）筛选出一组稳定性较好的变量子集，筛选出的变量数占全波段的比率由指数衰减函数（EDF）计算；

3）循环步骤1）和步骤2），最终得到 K 个变量子集，并基于这些变量建立PLSR模型，对应的均方根误差（R_{MSE}）值最小的变量子集为最后的特征变量，其中 K 为sCARS的循环次数。sCARS通过在MATLAB 2010b中编写代码实现。

2.4.2 SPA 变量选择

SPA的优点是从光谱矩阵中选择最小信息冗余的波长组合，降低波长之间共线性的影响，从而降低模型的复杂度，提高模型的稳定性和准确性[20]。设光谱矩阵为 $\boldsymbol{X}(A\times B)$，$m$ 为提取的特征变量个数，n 为迭代次数。

1）任意选择 $\boldsymbol{X}$ 中的一列 j，把建模集的第 j 列赋值给 x_j，记为变量 $\boldsymbol{x}_{k(0)}$；

2）未被选入的列向量位置集 S 为

$$S = \{j, 1 \leqslant j \leqslant B, j \notin \{k(0),\cdots,k(n\text{-}1)\}\}, \tag{2}$$

分别计算 $\boldsymbol{x}_{k(n\text{-}1)}$ 对未被选中列变量的投影，记为 K_{x_j}，K 为投影算子：

$$K_{x_j} = x_j(x_j^s x_{k(n\text{-}1)}) x_{k(n1)} (x_{k(n\text{-}1)}^s x_{k(n\text{-}1)})^{-1}; \tag{3}$$

3）$k(n\text{-}1) = \arg(\max\|K_{x_j}\|), j \in S, n = n+1,$ (4)

如果 $n < N$，则令 $\boldsymbol{x}_{k(n\text{-}1)}$ 选入列向量，返回2）再计算；

4）采用多元线性回归方法对新选择的变量集 $\{\boldsymbol{x}_{k(0)},\boldsymbol{x}_{k(1)},\cdots,\boldsymbol{x}_{k(m\text{-}1)}\}$ 进行评估，最后得到 m 个最优波长子集 $\{\boldsymbol{x}_{k(0)},\boldsymbol{x}_{k(1)},\cdots,\boldsymbol{x}_{k(m\text{-}1)}\}$。

2.4.3 GA 变量选择

GA通过模拟自然进化过程搜索最优解，具有较高的自适应和全局优化能力[21]。GA具体步骤为：

1）波长编码：对波长进行0/1二进制编码，0表示未选中该波长，1表示选取该波长，0和1随机组合生成一条染色体；

2）选择初始群体：记初始群体个数为 N，每一个染色体长度（波长数）为 q，随机产生 N 个 q 位的0/1随机组合二进制群体；

3）适应度函数：采用PLSR模型中交叉检验均方根误差（R_{MSECV}）为适应度函数；

4）遗传操作：包括选择、交叉和变异3种方式，一般而言，选择轮盘赌法，变异概率为0.01，交叉概率为0.5；

5）终止条件：重复4），若达到设定的最大繁殖代数，则进化过程中得到的具有最大适应度的个体作为最优解输出，计算终止。

2.4.4 IRIV 变量选择

IRIV由随机子集生成、子集模型建立、模型参数分析3个环节构成，是一种基于模型集群分析策略的波长选择算法[22]。相对于一般的波长选择算法，IRIV具有在变量选择时呈现软收缩的特点，因此在保留有效变量方面更为稳妥。需注意的是IRIV虽可以更好地保留变量间的协同效应，但在迭代运算过程中需要建立大量子模型，这使得该算法的计算量较大。

2.4.5 sCARS 结合 SPA 变量选择

sCARS的优点是速度快，最终选出的特征变量的化学意义也比较容易解释，SPA挑选的特征变量冗余度最低，共线性最小。但SPA计算量较大，且选中的波长子集中很可能会纳入一些无关信息甚至是干扰信息，sCARS和SPA的联合使用不仅可以最大程度地降低光谱信息冗余，还可以降低无效波长对SPA计算过程的干扰[23]。

2.5 模型精度评价

采用均方根误差（R_{MSE}）、决定系数（R^2）、验证集标准差（SD）与验证集相对分析误差（R_{PD}）对预测模型进行精度评价。验证集均方根误差（R_{MSEVAL}）越小，R^2、R_{PD} 越大，模型预测效果越好。当 $R_{\mathrm{PD}} \geqslant 2$ 时，表明模型预测性能较好，可对样本进行有效估测；当 $1.4 \leqslant R_{\mathrm{PD}} < 2$ 时，表明模型可对样本进行粗略估测；当 $R_{\mathrm{PD}} < 1.4$ 时，表明模型不能对样本进行估测。

3　结　　果

3.1　特征变量筛选

图 3 为采用 sCARS 挑选特征变量过程。从图 3(a)中可以看出,随着 sCARS 迭代次数的增加,所保留的波长数量逐渐减少,且减少速度由快到慢;图 3(b)中十折交叉验证均方根误差(R_{MSECV})值呈现由大到小再由小到大的变化趋势;图 3(c)为波长变量稳定性轨迹图,图中每条曲线为每个变量的稳定度随迭代次数的变化趋势,将 R_{MSECV} 值最小时的最优变量子集数量用星号标记。sCARS 挑选特征变量过程中具有"粗选"和"精选"两个阶段,当 sCARS 运行次数为 27 次时,R_{MSECV} 值最小,此时共选择 51 个特征变量,仅占总变量数的 2.55%。

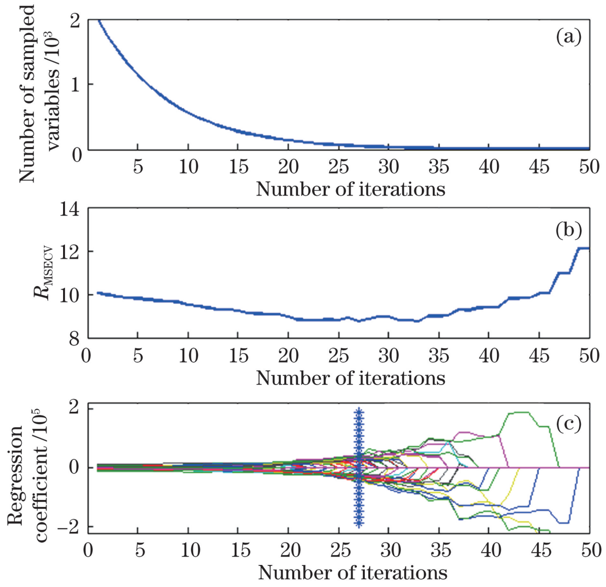

图 3　sCARS 变量筛选流程。(a)变量变化趋势;(b)十折交叉均方根误差;(c)变量回归系数

Fig. 3　Variable selection process by sCARS. (a) Changing trend of variables; (b) 10-fold R_{MSECV} values; (c) regression coefficients of variables

结合光谱数据, SPA 运行时设定特征变量最小波段数为 5,最大波段数为 100,当模型均方根误差最小时对应的变量数为最佳波段数,提取的变量为最佳波段。本研究中利用 SPA 共选择出 5 个最优特征变量,分别为 1361、1758、1909、2049、2213 nm。

图 4 为所有变量被选择的频率,图中 3 条横线表示选择频率阈值,频率阈值越大,选择的变量数越少。横线以上表示保留的特征变量,用于模型构建,横线以下是未被选中变量,不用于建模分析。本研究中采用 GA,从原始光谱中共选取 186 个特征变量,占 Vis-NIR 光谱全部变量的 9.3%。

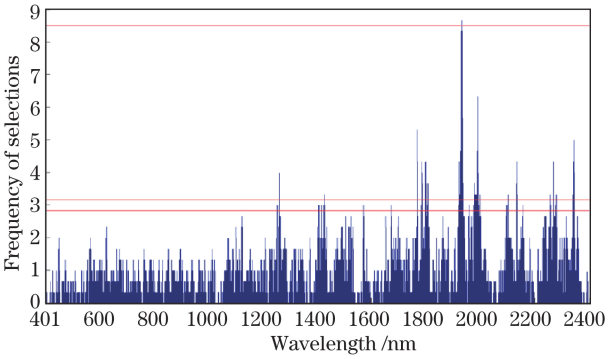

图 4　GA 特征变量筛选过程

Fig. 4　Characteristic variable selection process by GA

IRIV 能够将把单一指标进行波长硬性删除的传统策略转化为较为柔性的波长保留策略,进行波长选择时呈现软收缩的特点,因此在保留有效波长方面更为稳妥。本研究中 IRIV 设定光谱矩阵为 $\boldsymbol{X}$ (500×2000),利用五折交叉验证建立 PLSR 模型选择特征变量,最大主因子数为 10,最终 IRIV 共选择特征变量 63 个,占全部变量的 3.15%。

在 sCARS 处理后,再采用 SPA 挑选特征变量,光谱变量数由 51 个减少到 17 个,如图 5 所示。由图可知随着筛选变量数量的增加,R_{MSE} 值先迅速下降,当变量数为 17 时,R_{MSE} 值趋于稳定状态。图 5(a)中空心小方块表示采用 SPA 得到的最优变量子集个数,图 5(b)为采用 SPA 挑选的最优特征变量点在一条光谱曲线上的分布,分别为 440、444、458、460、493、561、938、976、1905、1916、2025、2253、2305、2314、2318、2390、2398 nm。

图 6 为 5 种变量筛选方法挑选特征变量分布图。由图可知,5 种变量筛选方法挑选的特征变量主要分布在近红外光谱区域。这可能是由于近红外光谱记录的主要是土壤中有机分子基团(O—H、N—H、C—H 及 C=C 等)在不同波长近红外光的倍频与和频吸收信息,而含氢基团是大多数有机物分子结构的基本单元。张娟娟等[24]系统分析了我国东部、中部 5 种不同土壤类型近红外光谱反射率,认为与 SOM 含量相关性较好的波段范围主要集中在 1350～1420 nm、1860～1890 nm、2000～2350 nm 3 个区域。Krishnan 等[25]基于 1130、1350、1398、2210 nm 4 个波段建立了 SOM 估测模型。Bendor 等[26]的研究也表明 1000～2500 nm 的近红外波段区域存在 SOM 相关的吸收波段。

3.2　PLSR 模型

表 2 为全波段和特征波段 PLSR 模型建模结果,其中 R^2_{cal} 和 R^2_{val} 分别为校准集和验证集决定系数,

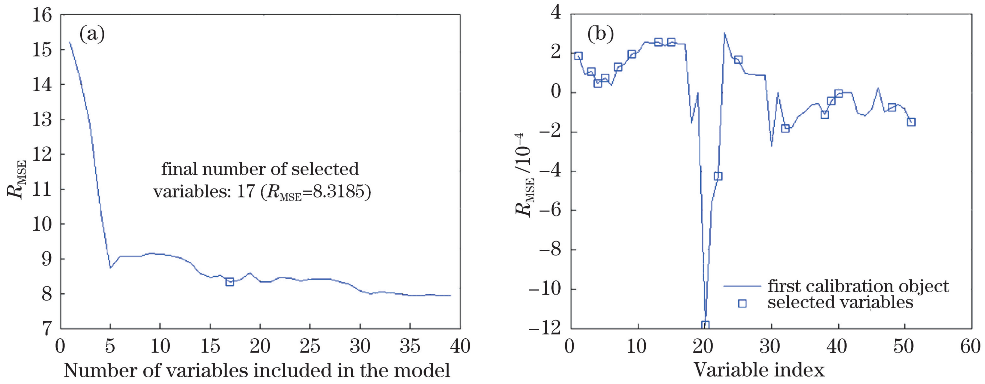

图 5　预处理光谱 sCARS-SPA 特征变量筛选过程。(a)模型变量数;(b)变量指数

Fig. 5　Characteristic variable selection process by sCARS-SPA from the pre-processing spectrum. (a) Number of variables in the model; (b) variable index

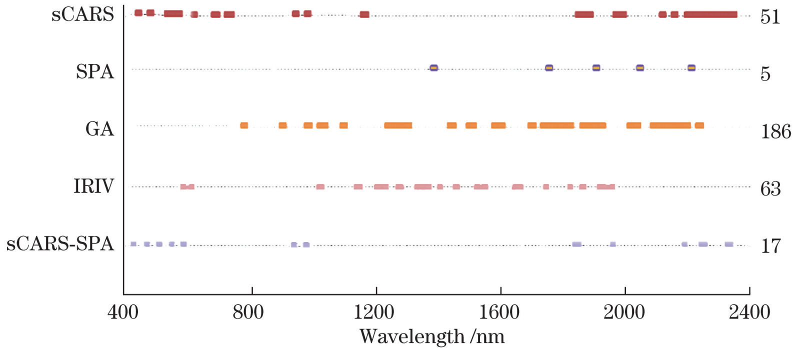

图 6　不同变量筛选方法挑选特征变量分布

Fig. 6　Distribution of characteristic variables with different variable selection methods

R_{MSECAL} 和 R_{MSEVAL} 分别为校准集和验证集均方根误差。由表 2 可知,基于特征波段的 PLSR 模型的预测效果均优于全波段 PLSR 模型,特别是基于 sCARS 挑选的特征变量建模,建模变量数为 51 个,仅占全部变量数的 2.55%,验证集 R^2_{val} 为 0.883,获得的 R_{PD} 为 2.9,模型预测效果最佳。图7为sCARS-PLSR模型校准集和验证集样本实测值和预测值的散点图,正方形代表校准集样本点,圆形代表验证集样本点。从图中可以看出,除验证集 C17、C18、C21、C30、C32、C45、C98 及 C259 点和校准集的 V8、V108 点外,其他样本点均分布在 1∶1直线两侧,这可能是由于这些样本点存在光谱异常或测量值异常。

表 2　不同变量筛选方法的 PLSR 模型精度

Table 2　Accuracies of PLSR model with different variable selection methods

Selection method	Variable number	PC	Calibration set		Validation set		
			R^2_{cal}	R_{MSECAL}	R^2_{val}	R_{MSEVAL}	R_{PD}
Full-spectrum	2000	5	0.842	9.326	0.835	9.069	2.5
sCARS	51	5	0.874	8.327	0.883	7.797	2.9
SPA	5	5	0.850	9.103	0.858	8.525	2.6
GA	186	4	0.842	9.342	0.861	8.415	2.7
IRIV	63	6	0.843	9.300	0.875	8.043	2.8
sCARS-SPA	17	4	0.765	11.391	0.848	8.791	2.6

3.3　SVM 模型

表 3 为全波段和特征波段 SVM 模型建模结果。首先,从表中可知,基于全波段建立的 SVM 模型,验证集 R^2 为 0.74,获得的 R_{PD} 为 1.9,可粗略对样本进行预测;而其校准集 R^2 为 0.91,显著高于验证集 R^2,这可能是由于光谱变量中存在干扰变量或无效变量,导致模型过于复杂,模型出现过拟合。其次,基于 sCARS、SPA、GA、IRIV 和 sCARS-SPA 挑

选特征波长变量后建立预测模型，验证集 R^2 均显著高于全波段模型，且与校准集的 R^2 较为接近，说明所建模型稳定性较好，同时获得的验证 R_{PD} 均大于2，可较好地对样本进行预测，表明建模之前对全波段进行特征变量筛选不仅能够有效提高建模效率，还能提高模型的预测精度。采用 SPA 挑选特征变量建模，建模变量数为 5 个，验证集 R^2 为 0.889，获得的 R_{PD} 为 2.9，模型预测效果最佳。图 8 为 SPA-SVM 模型校准集和验证集样本实测值和预测值的散点图，从图中可以看出，除验证集 C17、C18、C21、C30、C45、C98 及 C259 点和验证集 V8、V108 点外，其他样本点均离 1∶1直线较近。

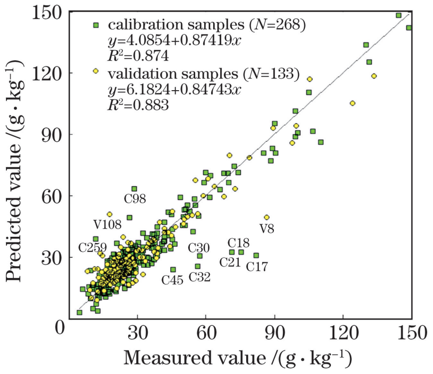

图 7　sCARS-PLSR 模型预测 SOM 含量散点图

Fig. 7　Scatter plot for the measured and predicted value by sCARS-PLSR model

表 3　不同变量筛选方法的 SVM 建模精度

Table 3　Accuracies of SVM model with different variable selection methods

Selection method	Variable number	Optimal parameter		Calibration set		Validation set		
		g (nuclear function)	c (punishment coefficient)	R^2_{cal}	R_{MSECAL}	R^2_{val}	R_{MSEVAL}	R_{PD}
Full-spectrum	2000	0.036	3.031	0.91	7.221	0.74	11.546	1.9
sCARS	51	0.021	3.031	0.881	8.116	0.877	7.918	2.8
SPA	5	0.004	1.741	0.858	8.855	0.889	7.477	3.0
GA	186	0.012	1.741	0.867	8.577	0.871	8.093	2.8
IRIV	63	0.021	3.031	0.869	8.493	0.864	8.307	2.7
sCARS-SPA	17	0.011	2.858	0.877	8.246	0.873	8.052	2.8

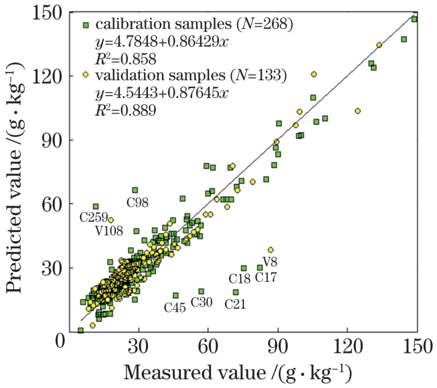

图 8　SPA-SVM 模型预测 SOM 含量散点图

Fig. 8　Scatter plot for the measured and predicted value by SPA-SVM model

3.4　RF 模型

RF 模型是一种分层非参数方法，用于估计独立变量和因变量之间复杂的非线性关系。RF 模型为了提高模型预测精度并避免过拟合，引入两个随机性因素，即森林中单棵树的分类强度和森林中树与树之间的相关强度。这两个随机性因素的引入，使得 RF 模型不容易因变量数远大于建模样本数陷入过拟合，且具有很好的抗噪声能力[27]。RF 模型通过 MATLAB 2010b 编程实现，分类树数目 $n_{\mathrm{tree}}=500$，节点数 m_{try} 为 $\sqrt{p}$，p 为自变量数目。表 4 为全波段和特征波段 RF 建模结果。从表中可知，全波段和特征波段 RF 模型验证集 R^2 值无明显差异，这说明采用特征波段构建 RF 模型对模型预测精度的改善不明显，但其构建模型的变量数却显著减少，可提高建模效率。这可能与 RF 模型具有很好的抗噪声能力，不易因变量数大于建模样本数而陷入过度拟合有关。其中 IRIV-RF 模型预测效果略好，校准集和验证集 R^2 分别为 0.941 和0.96，验证集 R_{PD} 为 4.8，可较好地预测 SOM 含量。图 9 为 IRIV-RF 模型校准集和验证集样本实测值和预测值的散点图。从图中可以看出，IRIV-RF 模型校准集和验证集数据点除 C17、C18、C21、C98、C117、C132、C150 和 V8、V10 外其他数据点均靠近 1∶1直线，且均较 PLSR、SVM 模型数据点更靠近 1∶1直线的两侧。

表 4　不同变量筛选方法 RF 建模精度

Table 4　Accuracies of RF model with different variable selection methods

Selection method	Variable number	Calibration set		Validation set		
		R_{cal}^2	R_{MSECAL}	R_{val}^2	R_{MSEVAL}	R_{PD}
Full-spectrum	2000	0.942	5.817	0.957	4.840	4.6
sCARS	51	0.942	5.781	0.958	4.780	4.7
SPA	5	0.930	6.585	0.954	5.082	4.4
GA	186	0.939	5.894	0.959	4.699	4.8
IRIV	63	0.941	5.927	0.960	4.656	4.8
sCARS-SPA	17	0.940	5.910	0.955	4.971	4.5

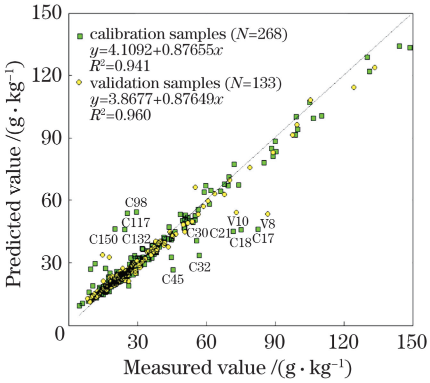

图 9　IRIV-RF 模型预测 SOM 含量散点图

Fig. 9　Scatter plot for the measured and predicted value by IRIV-RF model

4　讨　　论

前文采用 sCARS、SPA、IRIV、GA 和 sCARS-SPA 模型挑选特征变量。基于特征变量的 PLSR 和 SVM 模型精度均高于全波段模型精度，基于特征变量的 RF 模型精度较全波段模型提高不明显，但建模效率大大提升，进一步说明对全波段进行有效变量选择的重要性。

图 10 为不同变量筛选算法的 PLSR、SVM 和 RF 模型建模结果。由图可知，RF 模型的预测效果最佳，优于 SVM 模型和 PLSR 模型。SOM 含量与光谱之间关系较为复杂，PLSR 是一种线性方法，在解决非线性问题时表现不佳，而 SVM 和 RF 可较好地解决独立变量和因变量之间复杂的非线性关系；但 SVM 模型易因较高的频谱噪声引起严重的偏差估计，模型精度降低；RF 模型融合了随机特征选择和 Bagging 算法两大机器学习技术，与传统的分类器算法相比，不但能较好地容忍异常值和噪声，使建立的模型精度较高且具有较好的稳健性，而且能同时处理连续型和离散型数据[28]。如 Viscarra 等[29]采用多种建模方法预测土壤有机碳、黏土含量和 pH 值，结果显示，RF 模型精度要优于 PLSR、MARS、ANN 和 SVM 模型。同样 Douglas 等[30]基于 Vis-NIR 光谱分析技术，通过比较 PLSR 和 RF 模型预测土壤中总石油烃（TPH）含量，发现 RF 模型在处理光谱中非线性因素时，与线性 PLSR 模型相比，提供了更高的预测精度。

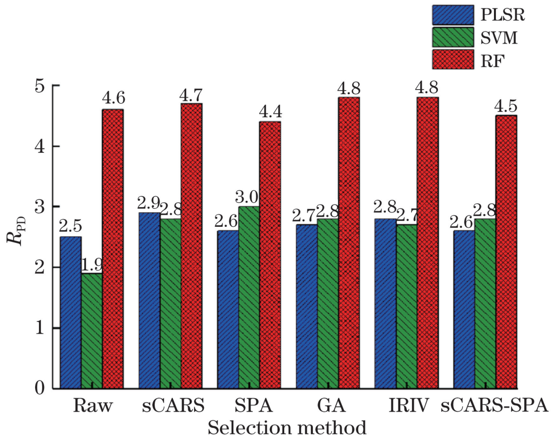

图 10　不同变量筛选方法 PLSR、SVM、RF 模型建模结果

Fig. 10　Results of PLSR, SVM and RF models with different variable selection methods

异常样本的存在对模型的性能产生一定的影响，表 5 为剔除 C17、C18、C21、C30、C45、C98、C259、V8 和 V108 异常点后的模型精度。从表中可知，在剔除这些异常值之后 sCARS-PLSR、sCARS-SVM 和 sCARS-RF 模型精度均有提高，特别是 sCARS-RF 模型验证集 R^2 提高到 0.988，R_{PD} 达到 7.8，预测 SOM 含量性能最佳。因此，在以后的工作中将深入研究异常值剔除对模型精度的影响，如高洪智等[31]基于随机抽取一致性算法（RANSAC）建立有机质含量估算模型，相关系数 r 从 0.891 提高到 0.963。图 11 为剔除异常值前后 sCARS-RF 模型散点图。从图中可知，剔除异常值后，sCARS-RF 模型校准集和验证集数据点均较为均匀地分布在 1∶1直线的两侧。

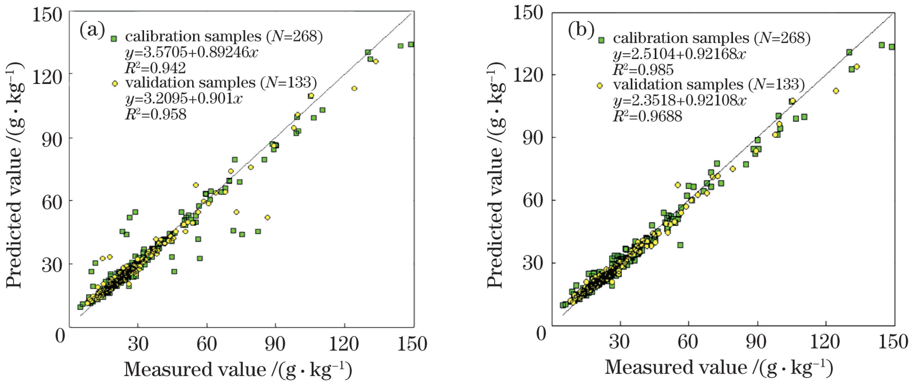

图 11　人工剔除异常值前后 sCARS-RF 模型散点图。(a)剔除异常值前;(b)剔除异常值后

Fig. 11　Scatter plots for the measured and predicted value by sCARS-RF model before and after artificially eliminating outliers. (a) Contain outliers; (b) eliminate outliers

表 5　人工剔除异常值后模型精度

Table 5　Model accuracy after manually eliminating outliers

Model	Calibration set		Validation set		
	R_{cal}^2	R_{MSECAL}	R_{val}^2	R_{MSEVAL}	R_{PD}
sCARS-PLSR	0.943	5.538	0.926	5.987	3.7
sCARS-SVM	0.926	6.092	0.957	4.903	4.6
sCARS-RF	0.985	3.204	0.988	2.865	7.8

5　结　　论

筛选土壤有机质的光谱响应波段是简化模型和提高模型预测能力的关键。本研究以青海省湟水流域表层土样为研究对象,原始光谱经预处理后,采用 sCARS、SPA、GA、IRIV 和 sCARS-SPA 5 种算法挑选光谱特征波段,并引入 PLSR、SVM 和 RF 模型对优选的特征变量构建 SOM 预测模型。PLSR 模型和 SVM 模型精度较全波段均有一定提高,而 RF 模型 R^2 没有明显变化,对模型精度的提高帮助不大,但模型变量数却显著减少,降低了模型复杂度,提高了建模效率。同时 RF 模型精度明显优于 SVM 模型和 PLSR 模型,其在处理多维特征数据和抗噪声能力方面展现出独特的优势。5 种变量选择算法中,基于 IRIV 挑选的特征变量建立的 RF 模型反演 SOM 含量精度和拟合效果最优,较全波段 PLSR 模型,预测集 R^2 从0.835 提高到 0.96,R_{PD}从 2.5 提高到 4.8,实现了对 SOM 含量的精准预测。变量选择算法耦合回归模型,在保证模型精度的同时,大大降低了模型的复杂度,提高了建模效率,为采用光谱分析技术快速无损估测耕地 SOM 含量提供了技术支撑,应用前景广阔。

参　考　文　献

[1] Nan F, Zhu H F, Bi R T. Hyperspectral prediction of soil organic matter content in the reclamation cropland of coal mining areas in the Loess Plateau [J]. Scientia Agricultura Sinica, 2016, 49(11): 2126-2135.
南锋, 朱洪芬, 毕如田. 黄土高原煤矿区复垦农田土壤有机质含量的高光谱预测[J]. 中国农业科学, 2016, 49(11): 2126-2135.

[2] Mishra U, Torn M S, Masanet E, *et al*. Improving regional soil carbon inventories: combining the IPCC carbon inventory method with regression kriging[J]. Geoderma, 2012, 189/190: 288-295.

[3] St Luce M, Ziadi N, Zebarth B J, *et al*. Rapid determination of soil organic matter quality indicators using visible near infrared reflectance spectroscopy [J]. Geoderma, 2014, 232/233/234: 449-458.

[4] Liu Y Q, Chen H Y, Wang R Y, *et al*. Quantitative analysis of soil salt and its main ions based on visible/near infrared spectroscopy in estuary area of Yellow River[J]. Scientia Agricultura Sinica, 2016, 49(10): 1925-1935.
刘亚秋, 陈红艳, 王瑞燕, 等. 基于可见/近红外光谱的黄河口区土壤盐分及其主要离子的定量分析[J]. 中国农业科学, 2016, 49(10): 1925-1935.

[5] Liu H J, Zhang B, Liu D W, *et al*. Study on quantitatively remote sensing typical soils in Songnen plain, northeast China [J]. Journal of Remote Sensing, 2008, 12(4): 647-654.
刘焕军, 张柏, 刘殿伟, 等. 松嫩平原典型土壤高光谱定量遥感研究[J]. 遥感学报, 2008, 12(4): 647-654.

[6] Lu Y L, Bai Y L, Yang L P, *et al*. Prediction and validation of soil organic matter content based on hyperspectrum[J]. Scientia Agricultura Sinica,

2007, 40(9): 1989-1995.
卢艳丽, 白由路, 杨俐苹, 等. 基于高光谱的土壤有机质含量预测模型的建立与评价[J]. 中国农业科学, 2007, 40(9): 1989-1995.

[7] Wang L S, Lu C P, Wang R J, *et al*. Optimization for vis/NIRS prediction model of soil available nitrogen content[J]. Chinese Journal of Luminescence, 2018, 39(7): 1016-1023.
汪六三, 鲁翠萍, 王儒敬, 等. 土壤碱解氮含量可见/近红外光谱预测模型优化[J]. 发光学报, 2018, 39(7): 1016-1023.

[8] Zhu Y X, Yu L, Hong Y S, *et al*. Hyperspectral features and wavelength variables selection methods of soil organic matter[J]. Scientia Agricultura Sinica, 2017, 50(22): 4325-4337.
朱亚星, 于雷, 洪永胜, 等. 土壤有机质高光谱特征与波长变量优选方法[J]. 中国农业科学, 2017, 50(22): 4325-4337.

[9] Vohland M, Ludwig M, Harbich M, *et al*. Using variable selection and wavelets to exploit the full potential of visible-near infrared spectra for predicting soil properties[J]. Journal of Near Infrared Spectroscopy, 2016, 24(3): 255-269.

[10] Lin Z D, Wang Y B, Wang R J, *et al*. Improvements of the vis-NIRS model in the prediction of soil organic matter content using spectral pretreatments, sample selection, and wavelength optimization[J]. Journal of Applied Spectroscopy, 2017, 84(3): 529-534.
林志丹, 汪玉冰, 王儒敬, 等. 波长优选对土壤有机质含量可见光/近红外光谱模型的优化[J]. 发光学报, 2017, 84(3): 529-534.

[11] Nawar S, Buddenbaum H, Hill J, *et al*. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy[J]. Soil and Tillage Research, 2016, 155: 510-522.

[12] Viscarra Rossel R A, Rizzo R, Demattê J A M, *et al*. Spatial modeling of a soil fertility index using visible-near-infrared spectra and terrain attributes[J]. Soil Science Society of America Journal, 2010, 74(4): 1293-1300.

[13] Li M J, Zhang M Y, Cui L J, *et al*. Inversion of Hg content in reed leaf using continuous wavelet transformation and random forest[J]. Chinese Journal of Eco-Agriculture, 2018, 26(11): 1730-1738.
李梦洁, 张曼胤, 崔丽娟, 等. 基于连续小波变换和随机森林的芦苇叶片汞含量反演[J]. 中国生态农业学报, 2018, 26(11): 1730-1738.

[14] Ge X Y, Ding J L, W J Z, *et al*. Estimation of soil moisture content based on competitive adaptive reweighted sampling algorithm combined with machine learning[J]. Acta Optica Sinica, 2018, 38(10): 1030001.
葛翔宇, 丁建丽, 王敬哲, 等. 基于竞争适应重加权采样算法耦合机器学习的土壤含水量估算[J]. 光学学报, 2018, 38(10): 1030001.

[15] Li G W, Gao X H, Yang L Y, *et al*. Estimating soil organic matter contents from different soil particle size using visible and near-infrared reflectance spectrum-a case study of the Huangshui basin[J]. Chinese Journal of Soil Science, 2017, 48(6): 1360-1370.
李冠稳, 高小红, 杨灵玉, 等. 不同粒径土壤有机质含量可见光-近红外光谱估算研究-以湟水流域为例[J]. 土壤通报, 2017, 48(6): 1360-1370.

[16] Conforti M, Castrignanò A, Robustelli G, *et al*. Laboratory-based vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content[J]. Catena, 2015, 124: 60-67.

[17] Chen C, Lu Q P, Peng Z Q. Preprocessing methods of near-infrared spectrum based on NLMS adaptive filtering[J]. Acta Optica Sinica, 2012, 32(5): 0530001.
陈丛, 卢启鹏, 彭忠琦. 基于 NLMS 自适应滤波的近红外光谱去噪处理方法研究[J]. 光学学报, 2012, 32(5): 0530001.

[18] Jiang X Q, Ye Q, Lin Y, *et al*. Inverting study on soil water content based on harmonic analysis and hyperspectral remote sensing[J]. Acta Optica Sinica, 2017, 37(10): 1028001.
姜雪芹, 叶勤, 林怡, 等. 基于谐波分析和高光谱遥感的土壤含水量反演研究[J]. 光学学报, 2017, 37(10): 1028001.

[19] Zhang X Y, Li Q B, Zhang G J. Calibration transfer without standards for spectral analysis based on stable competitive adaptive re-weighted sampling[J]. Spectroscopy and Spectral Analysis, 2014, 34(5): 1429-1433.
张晓羽, 李庆波, 张广军. 基于稳定竞争自适应重加权采样的光谱分析无标模型传递方法[J]. 光谱学与光谱分析, 2014, 34(5): 1429-1433.

[20] Song X Z. Research of three new wavelength selection methods in near infrared spectroscopy quantitative analysis area[D]. Beijing: China Agricultural University, 2017.
宋相中. 近红外光谱定量分析中三种新型波长选择方法研究[D]. 北京: 中国农业大学, 2017.

[21] Chen H Y, Zhao G X, Zhang X H, *et al*. Hyperspectral characteristic and estimation modeling of fluvo-aquic soil alkali hydrolysable nitrogen content based on genetic algorithm in combination with partial

least squares[J]. Chinese Agricultural Science Bulletin, 2015, 31(2): 209-214.

陈红艳, 赵庚星, 张晓辉, 等. 基于遗传算法结合偏最小二乘的潮土碱解氮高光谱特征及含量估测[J]. 中国农学通报, 2015, 31(2): 209-214.

[22] Yun Y H, Wang W T, Tan M L, *et al*. A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration[J]. Analytica Chimica Acta, 2014, 807: 36-43.

[23] Yu L, Hong Y S, Zhou Y, *et al*. Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique[J]. Transactions of the Chinese Society of Agricultural Engineering, 2016, 32(13): 95-102.

于雷, 洪永胜, 周勇, 等. 高光谱估算土壤有机质含量的波长变量筛选方法[J]. 农业工程学报, 2016, 32(13): 95-102.

[24] Zhang J J, Tian Y C, Zhu Y, *et al*. A near-infrared spectral index for estimating soil organic matter content[J]. Chinese Journal of Applied Ecology, 2009, 20(8): 1896-1904.

张娟娟, 田永超, 朱艳, 等. 一种估测土壤有机质含量的近红外光谱参数[J]. 应用生态学报, 2009, 20(8): 1896-1904.

[25] Krishnan P, Alexander J D, Butler B J, *et al*. Reflectance technique for predicting soil organic matter^1[J]. Soil Science Society of America Journal, 1980, 44(6): 1282-1285.

[26] Ben-Dor E, Banin A. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties[J]. Soil Science Society of America Journal, 1995, 59(2): 364-372.

[27] AbdelRahman A M, Pawling J, Ryczko M, *et al*. Targeted metabolomics in cultured cells and tissues by mass spectrometry: method development and validation[J]. Analytica Chimica Acta, 2014, 845: 53-61.

[28] Nawar S, Mouazen A M. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques[J]. Catena, 2017, 151: 118-129.

[29] Rossel R A V, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra[J]. Geoderma, 2010, 158(1/2): 46-54.

[30] Douglas R K, Nawar S, Alamar M C, *et al*. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques[J]. Science of the Total Environment, 2018, 616/617: 147-155.

[31] Gao H Z, Lu Q P, Ding H Q, *et al*. Robust calibration methods of near-infrared spectrum based on random sample consensus algorithm[J]. Acta Optica Sinica, 2013, 33(s2): s230001.

高洪智, 卢启鹏, 丁海泉, 等. 基于随机抽样一致性算法的近红外光谱稳健模型研究[J]. 光学学报, 2013, 33(s2): s230001.