第 58 卷 第 16 期/2021 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于集成提升算法的土壤速效氮近红外光谱回归预测

韩亚鲁，李绍稳*，郑文瑞，石胜群，朱先志，金秀

安徽农业大学信息与计算机学院，安徽 合肥 230036

摘要 速效氮含量的预测在土壤养分诊断中具有重要意义，通过特征选择和回归预测算法可有效地提高速效氮光谱检测模型的预测精度。选取了皖南地区的 188 个黄红壤土样本作为对象，利用 7 种预处理方法对光谱数据进行了校正，结合移动窗口法和 5 种智能优化类算法进行特征选择后，再基于多种集成提升(Boosting)算法建立 36 种回归校正模型来分析比较。实验结果表明：基于粒子群优化(PSO)的特征优选算法优选出的 202 个光谱特征主要集中在 600～1000 nm，利用此特征构建出的 Adaptive Boosting(AdaBoost)模型性能最佳，其土壤速效氮预测精度提高到 0.944。所提方法不仅提高了土壤速效氮预测精度，而且在特征区间优选算法上进行了探讨，具有一定的理论价值。

关键词 光谱学；近红外光谱；土壤；速效氮；特征选择；Boosting

中图分类号 S153.6 **文献标志码** A **doi**：10.3788/LOP202158.1630005

Regression Prediction of Soil Available Nitrogen Near-Infrared Spectroscopy Based on Boosting Algorithm

Han Yalu, Li Shaowen*, Zheng Wenrui, Shi Shengqun, Zhu Xianzhi, Jin Xiu

School of Information & Computer, Anhui Agricultural University, Hefei, Anhui 230036, *China*

Abstract The prediction of available nitrogen content has attracted significant attention in soil nutrient diagnosis. The available nitrogen spectrum detection model prediction accuracy can be effectively improved using feature selection and regression prediction algorithms. This study selects 188 yellow-red loam samples in southern Anhui as objects, uses seven preprocessing methods to correct the spectral data. It combines the moving window method and five intelligent optimization algorithms for feature selection. Then, it establishes 36 regression calibration models for analysis and comparison based on different ensemble boosting (Boosting) algorithms. The experimental results show that 202 spectral features selected using the feature optimization algorithm based on particle swarm optimization (PSO) are concentrated in the range of 600–1000 nm. The Adaptive Boosting (AdaBoost) model developed using these features has the best performance, with the prediction accuracy of soil available nitrogen of 0.944. This study improves the prediction accuracy of soil available nitrogen and discusses the optimization algorithm of characteristic interval, which has a certain theoretical value.

Key words spectroscopy; near-infrared spectroscopy; soil; available nitrogen; feature selection; Boosting

OCIS codes 300.6170; 150.1135

1 引 言

速效氮是土壤的重要营养成分，它的含量直接影响到作物的生长发育和最终产量。快速、准确地检测土壤速效氮的含量对测土配方施肥、作物高产稳产都具有重要意义。传统的土壤养分检测方法采用实验室理化检测方法，但该方法存在耗时、费力、低效的缺陷，不能满足精准农业的发展要求。可见

收稿日期：2020-08-28；**修回日期**：2020-09-08；**录用日期**：2020-09-20

基金项目：农业部“948”项目(2015-Z44，2016-X34)、安徽省教育厅课题(KJ2019A0212)

通信作者：*shwli@ahau.edu.cn

近红外(VIS-NIR)光谱分析技术具有快速、无损、高效等特点，被广泛应用于土壤养分定量测定方面[1]。目前常见的近红外光谱仪器波段范围广，但价格较昂贵，而且利用全波段数据建模时，也并非每个波长都能提供有效信息。近红外光谱数据具有数据量多、信噪比低的问题，因此需要对波长进行优选，选择最优光谱特征区间构建光谱快速测量系统，以提高模型稳健性[2]。Zhang 等[3]将蚁群优化算法和互信息结合，选择了与全氮相关的波长点；Xie 等[4]基于遗传算法对土壤养分特征波长进行了选择；纪文君等[5]和 Liu 等[6]分别证明了土壤有机质的光谱响应在 600～800 nm 和 620～810 nm；于雷等[7]也在 600～800 nm 附近提取了最优波长。目前，国内外的研究均聚焦在寻找最优波长点或者土壤有机质的有效波段，但并未对土壤速效氮的最优光谱特征区间进行探讨。

本文的目的是优选出土壤速效氮的近红外光谱特征区间，以提高速效氮含量的预测精度。首先，对光谱数据进行不同的预处理变换，分析其对不同回归模型(Boosting 回归、传统回归)的影响；然后，结合移动窗口(MW)和智能优化类算法，对校正后性能好的光谱数据进行分析，从全波段中选出最优的波长组合变量；最后，权衡不同 Boosting 回归模型的预测精度，优选出土壤速效氮的最优近红外光谱特征区间，成功对土壤速效氮含量进行了预测。所提方法在显著简化模型的同时提高了模型的性能。

2　材料和方法

2.1　土壤样品采集

在皖南的黄山区和石台县完成土壤样本采集工作，该地区主要种植方式为水稻、油菜轮作制。以黄红壤土为研究对象，采样深度为 0～20 cm。首先，通过 5 点对角取样法对土壤进行混合作为一个采样点的样本，去除杂物后装入密封袋保存，每份样本约 1.5 kg，共计采样 188 份；然后，将土壤样本带回实验室进行风干、研磨处理，过 20 目筛后得到实验所需的土壤样本粉末。

2.2　实验数据获取

将每份土样分成 2 份，分别用于土壤光谱数据采集和理化测试。理化测试法采用碱解扩散法测定土壤速效氮含量。使用便携式地物非成像光谱仪(OFS1700，蔚海光学仪器公司)进行光谱数据采集，光谱范围为 200～1700 nm。该仪器内部由两个传感器组合而成，光谱分辨率在 200～900 nm 为 2 nm，在 900～1700 nm 为 5 nm。

室内光谱采集系统如图 1 所示。将处理好的土壤粉末放在直径为 4.5 cm，深为 2.5 cm，内部铺有黑布(防止杂散光干扰)的盛样器皿中，并将土样表面刮平，用光谱仪的反射探头(内置光源)直接接触土壤表面进行光谱采集。随机选取 3 处进行光谱测量，对每个样本测量 10 次反射率光谱后取平均值作为原始光谱数据，测量过程中每 10 份样本进行一次标准白板校正。

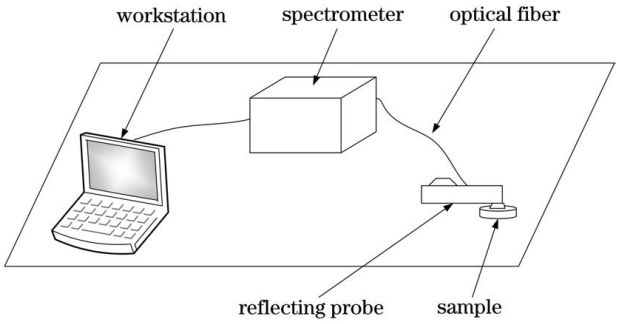

图 1　室内光谱采集系统

Fig. 1　Indoor spectrum acquisition system

2.3　光谱特征选择方法

采用 MW 和 Adaptive Boosting(AdaBoost)相结合的方法从宽谱区中初选出最优信息区间。基本思想是，用 MW 沿全谱区扫描选出多个光谱特征区间[8-9]，通过比较 AdaBoost 模型精度，优选出最佳信息区间。然后基于优选的信息区间，采用 5 种智能优化类算法进行特征选择，这 5 种智能算法分别是随机森林(RF)算法、遗传算法(GA)、模拟退火(SA)算法、粒子群优化(PSO)算法、基于贪心策略的混合遗传算法(GGA)。RF 通过对特征进行重要性度量来实现特征选择[10-11]；GA 是一种模拟自然进化过程的全局搜索算法[12-13]；SA 是一种来源于固体退火过程的随机搜索算法[14-15]；GGA 将贪心策略引入 GA 中，进而进行遗传操作搜索最优解[16]。

PSO 来源于鸟群的群集行为，其中每个粒子都代表问题的每一个可能解[17-18]。所有粒子根据个体极值和全局最优解调整速度和位置，不断迭代得到最优解。速度和位置的数学表达式为[19]

$$\begin{cases} v_{id}^{k} = wv_{id}^{k-1} + c_1 r_1 (p_{\text{best},id} - x_{id}^{k-1}) + c_2 r_2 (g_{\text{best},d} - x_{id}^{k-1}) \\ x_{id}^{k} = x_{id}^{k-1} + v_{id}^{k-1} \end{cases}, \tag{1}$$

式中：w 为惯性因子；k 为当前迭代次数；c 为加速度常数；r 为 0～1 之间均匀分布的随机数；$p_{\mathrm{best},id}$ 为粒子 i 个体最优位置；$g_{\mathrm{best},d}$ 为粒子群体全局最优位置。

2.4　光谱模型构建方法

主要对 5 种建模方法进行分析比较，具体包括最小二乘回归（PLSR）、自适应增强（AdaBoost）回归、梯度提升回归树（GBRT）、极限梯度提升（XGBoost）算法、轻型梯度提升机（LightGBM）算法。

PLSR 可解决光谱数据共线性和冗余问题[20]。AdaBoost、GBRT、XGBoost 以及 LightGBM 都是基于 Boosting 的迭代型算法。Boosting 主要分为两种实现方式，AdaBoost 和 Gradient Boosting，如图 2 所示。其中 AdaBoost 通过改变样本的权重来训练多个弱学习器；GradientBoosting通过改变目标函数来训练多个弱学习器，再基于串行策略将弱学习器组成强学习器。GBRT、XGBoost 和 LightGBM 都是 Gradient Boosting 的具体实现算法。

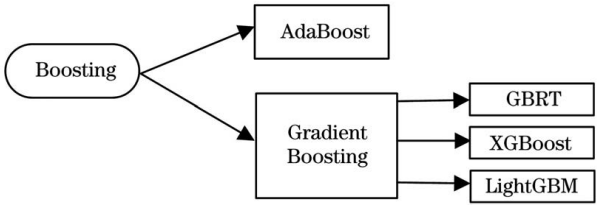

图 2　Boosting 算法分类

Fig. 2　Boosting algorithm classification

Boosting 算法使用决策树作为弱学习器，最小二乘作为损失函数。AdaBoost 算法的思想是对同一个训练集，不断更新权值，训练出不同的弱回归模型，然后弱回归模型集合形成强回归模型。图 3 是 AdaBoost 算法的工作机制[21]。

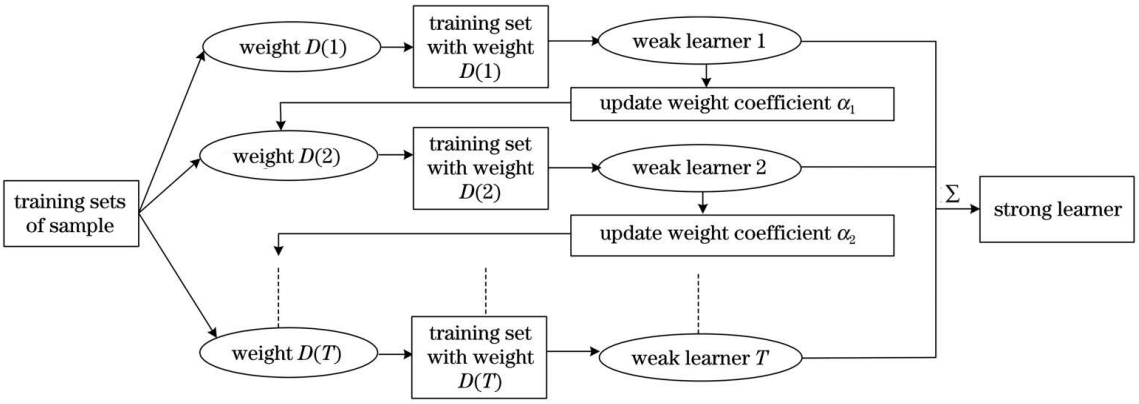

图 3　自适应增强回归原理图

Fig. 3　Schematic of AdaBoost regression

GBRT 是一种迭代的决策树算法，思想是将损失函数的负梯度在当前模型的值作为残差的近似值，拟合一个回归树，并考虑可加性模型 $F(X)$。

$$F(X)=\sum_{m=1}^{M}\gamma_m h_m(X), \tag{2}$$

式中：γ_m 是步长；M 是树的叶子节点数量；$h_m(X)$ 是使损失函数最小化的新添加的树[22-23]。

XGBoost 是基于预排序的决策树算法，核心思想是先通过残差拟合生成多个弱回归模型，然后将多个模型结果累加作为最终预测值[24-25]。其数学表达式为

$$\begin{cases} O_{\mathrm{obj}}=\sum_{i}^{n} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \\ \Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \end{cases}, \tag{3}$$

式中：$l(y_i,\hat{y}_i)$ 是损失函数；$\Omega(f_k)$ 是正则项；T 是叶子结点的个数；w 是叶子权重值；γ 和 λ 均是惩罚系数。

LightGBM 是决策树算法的加强版。在其基础上引入了单边梯度采样（GOSS）、互斥特征绑定（EFB）技术。使用 GOSS 可以减少大量只具有小梯度的数据实例，从而在遍历特征时节省了时间和空间；使用 EFB 可以将许多互斥特征绑定为一个特征，达到了降维的目的[26]。

2.5　评价方法

模型性能指标为决定系数（R^2）、相对分析误差（RPD）和均方根误差（RMSE）。RMSE 用来衡量观测值和真实值之间的偏差。误差越小，R^2 越接近 1，说明模型稳定性越好。根据 Chang 等[27]提出的 RPD 评判等级：$E_{\mathrm{RPD}}\geqslant 2$ 时，模型具有很好的预测效

果，属 A 类模型，可用于定量预测；1.4≤E_{RPD}<2 时，模型有一定的预测效果，属 B 类模型，可用于粗略的预测；E_{RPD}<1.4 时，模型的预测效果较差，属 C 类模型，不能用于定量预测。

2.6　技术路线

所提技术路线如图 4 所示。首先基于不同预处理校正后的光谱数据进行全波段回归建模，对不同模型结果进行分析，得到最优预处理和回归校正方法；然后对最优预处理校正后的数据进行特征选择，将选择的变量输入到优选的回归模型；最后通过比较模型的性能得到最优土壤速效氮定量分析模型和光谱特征区间。

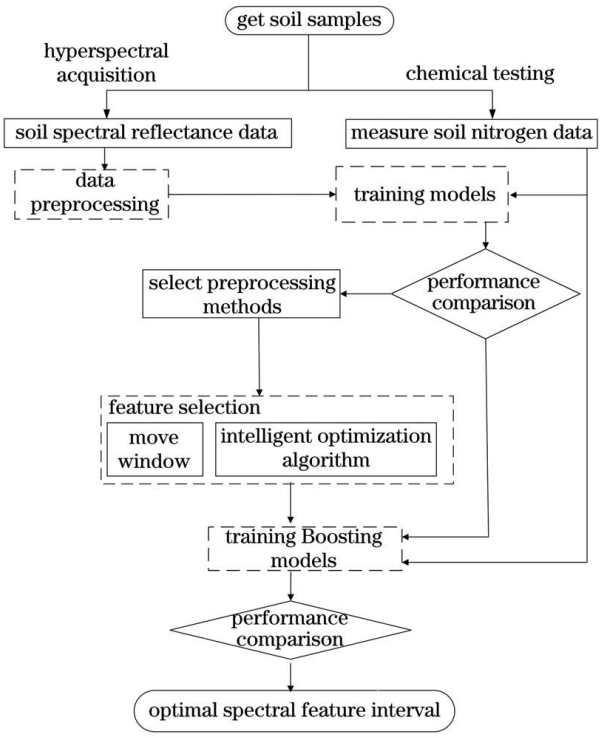

图 4　土壤速效氮的近红外高光谱特征分析技术路线

Fig. 4　Technical route analysis of near-infrared hyperspectral characteristics of available nitrogen in soil

3　结果与分析

3.1　样本分析

采用 Kennard-Stone(KS)方法[28]，将 188 个土壤样本按 7∶3 随机划分为训练集和测试集，训练集共 131 份土壤样本，测试集共 57 份土壤样本。表 1 是训练集、测试集以及全部样本的速效氮含量统计参数，从中可看出，速效氮含量具有明显的梯度差异，且 3 个样本集的均值和标准差较为接近，说明它们具有相似的数据分布结构。因此训练集和测试集均可有效地代表整体数据集的分布特征。

由于首尾波段信噪比较低，故对 350～1655 nm 区域的数据进行研究。图 5 是全部土壤样本的原始和预处理变换后的光谱曲线。采用滤波平滑(SG)、对数变换(LG)、一阶导数变换(FD)、标准正态变换(SNV)及相关组合共 7 种预处理方法对土壤原始光谱数据进行校正处理。从整体趋势来看：所有样本的土壤光谱形态相似，但是不同样本光谱反射率高低不同；反射率曲线具有一定的波动性，尤其在波长 900 nm 左右尤为明显，因为此处是两个传感器的连接处，存在大量噪声，从而产生抖动现象。从反射率变化趋势来看，在 1400 nm 处出现明显吸收谷，根据文献[29]可知该处是水分吸收带，反射率降低。以上分析表明，土壤光谱反射率与土壤速效氮含量之间存在较高的相关性。

表 1　土壤速效氮含量的统计参数

Table 1　Statistical parameters of soil available nitrogen content

Dataset	Number of samples	Max / (mg·kg^{-1})	Min / (mg·kg^{-1})	Median / (mg·kg^{-1})	Mean / (mg·kg^{-1})	Standard / (mg·kg^{-1})
Total set	188	731.584	19.32	132.644	179.083	144.398
Training set	131	731.584	19.32	130.732	179.440	148.217
Testing set	57	687.148	67.62	137.816	178.255	135.211

3.2　基于全波段光谱数据的回归模型分析

利用 PLSR、AdaBoost、GBRT、XGBoost、LightGBM 这 5 种建模方法，基于预处理后的 7 种光谱数据，建立了 35 个土壤速效氮定量分析模型。表 2 为不同预处理方法对回归模型性能的影响，图 6 为不同预处理方法下回归模型测试集的 R^2 和 RPD 变化图。可以看出，Boosting 算法中的 AdaBoost、GBRT 及 XGBoost 算法在 SNV 和 SG+SNV 预处理变换下建立的校正模型性能最优，测试集 R^2 均大于 0.9，RPD 等级均为 A 级，说明模型具有较好的泛化能力和预测精度。因此采用 SNV 和 SG+SNV 预处理后的光谱数据进行特征选择，并结合 AdaBoost、GBRT 及 XGBoost 算法分析建模。

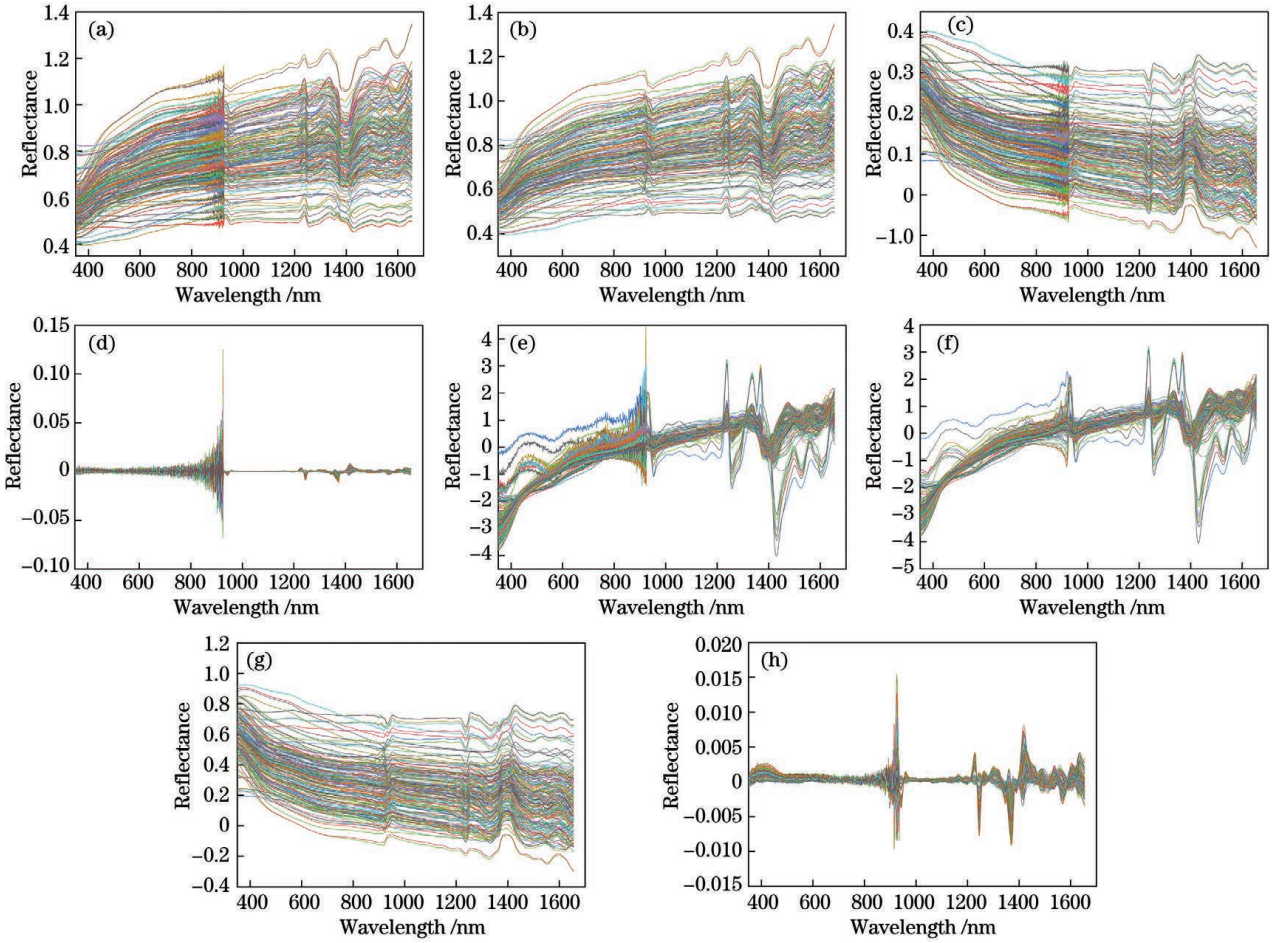

图 5 预处理前后光谱对比。(a)原始光谱;(b) SG;(c) LG;(d) FD;(e) SNV;(f) SG+SNV;(g) SG+LG;(h) SG+FD

Fig. 5 Contrast of spectra before and after preprocess. (a) Original spectra; (b) SG; (c) LG; (d) FD; (e) SNV; (f) SG+SNV; (g) SG+LG; (h) SG+FD

表 2 不同预处理方法对回归模型的影响

Table 2 Influence of different pretreatment methods on regression model

Preprocess method	Regression model	Training set		Testing set		Parameter
		R^2	RPD	R^2	RPD	Number of latent variables
SG	PLSR	0.92	3.62	0.894	3.08	16
LG	PLSR	0.94	4.02	0.898	3.14	11
FD	PLSR	0.95	4.39	0.718	1.88	6
SNV	PLSR	0.96	4.87	0.857	2.64	11
SG+SNV	PLSR	0.94	4.16	0.845	2.54	11
SG+LG	PLSR	0.95	4.25	0.916	3.45	14
SG+FD	PLSR	0.88	2.87	0.738	1.95	6
Preprocess method	Regression model	Training set		Testing set		Parameter
		R^2	RPD	R^2	RPD	Learning rate/number of estimators
SG	GBRT	0.99	28.27	0.610	1.60	0.4/400
LG	GBRT	0.99	28.27	0.508	1.43	0.2/200
FD	GBRT	0.99	28.27	0.668	1.73	0.4/200

表 2(续)

Preprocess method	Regression model	Training set		Testing set		Parameter
		R^2	RPD	R^2	RPD	Learning rate/number of estimators
SNV	GBRT	0.99	34.96	0.915	3.43	0.2/100
SG+SNV	GBRT	0.99	15.13	0.910	3.33	0.4/100
SG+LG	GBRT	0.99	28.27	0.573	1.53	0.4/100
SG+FD	GBRT	0.99	22.27	0.898	3.14	0.1/300
SG	AdaBoost	0.97	5.43	0.644	1.68	0.4/100
LG	AdaBoost	0.95	4.68	0.576	1.54	0.4/200
FD	AdaBoost	0.99	23.74	0.573	1.53	0.1/200
SNV	AdaBoost	0.99	12.14	0.921	3.43	0.2/100
SG+SNV	AdaBoost	0.99	20.07	0.912	3.37	0.1/200
SG+LG	AdaBoost	0.96	4.75	0.319	1.21	0.3/200
SG+FD	AdaBoost	0.99	28.27	0.876	2.84	0.1/200
SG	XGBoost	0.99	28.24	0.745	1.98	0.1/300
LG	XGBoost	0.99	28.24	0.739	1.95	0.1/300
FD	XGBoost	0.99	28.26	0.470	1.15	0.2/100
SNV	XGBoost	0.99	28.21	0.912	3.37	0.4/100
SG+SNV	XGBoost	0.99	25.26	0.908	3.31	0.2/100
SG+LG	XGBoost	0.99	28.25	0.745	1.98	0.1/300
SG+FD	XGBoost	0.99	28.26	0.835	2.46	0.4/100
SG	LightGBM	0.79	2.21		0.81	0.1/400
LG	LightGBM	0.75	1.99		0.69	0.4/400
FD	LightGBM	0.99	19.60	0.521	1.44	0.4/200
SNV	LightGBM	0.99	26.87	0.849	2.57	0.4/100
SG+SNV	LightGBM	0.99	25.44	0.857	2.65	0.4/100
SG+LG	LightGBM	0.79	2.19		0.68	0.3/400
SG+FD	LightGBM	0.99	22.62	0.695	1.81	0.1/200

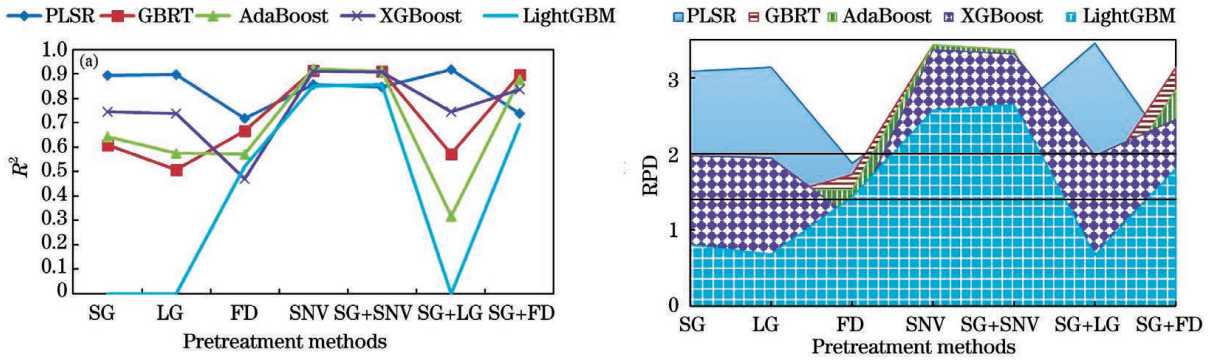

图 6 在不同预处理方法下回归模型测试集的 R^2 和 RPD 值。(a)R^2;(b)RPD

Fig. 6 R^2 and RPD values of regression models with testing obtained by different pretreatment methods. (a) R^2; (b) RPD

3.3 基于特征波段光谱数据的回归模型分析

土壤光谱样本数据会受到水分等其他组分和仪器噪声的影响，此外，速效氮对应的可见近红外光谱信息复杂，区间宽度不等，故先采用移动窗口获得速效氮含量对应的光谱信息区间，再运用更为灵活的智能优化算法进行特征选择。先选取宽度为 400 的窗口对预处理后的光谱数据进行最佳信息区间选择，在 SNV 和 SG+SNV 校正处理的数据下优选出的区间分别是 600～999 nm 和 629～928 nm，测试集的 R^2 分别为 0.923 和 0.917。然后在信息区间中基于特征选择算法搜索更优的波长点变量，迭代 500 次后，RF 分别优选了 200、202 个光谱变量；PSO 分别优选了 202、201 个光谱变量；GA 分别优选了 184、209 个光谱变量；SA 分别优选了 206、206 个光谱变量；GGA 分别优选了 202、200 个光谱变量。

在 SNV、SG+SNV 校正的光谱数据下，基于全波段、特征选择波长点分别建立 AdaBoost、GBRT 及 XGBoost 回归校正模型（共 36 种），并且用未参与建模的 57 个样本对模型性能进行评价。表 3 是在校正的光谱数据下以不同变量建立的定量模型的分析结果。图 7 是以不同变量建立模型的测试集 R^2、RMSE 及 RPD 值变化图。

表 3 基于 SNV、SG+SNV 处理的光谱数据以不同变量建立定量模型分析结果

Table 3 Analysis results of quantitative models with different variables based on the spectral data processed by SNV and SG+SNV

Preprocess method	Algorithm	Wavelength range /nm	Number of variables	Training set		Testing set		Parameter
				R^2	RPD	R^2	RPD	Learning rate/number of estimators
SNV	AdaBoost	350-1655	1305	0.99	12.14	0.921	3.43	0.2/100
SNV	GBRT	350-1655	1305	0.99	34.96	0.915	3.43	0.2/100
SNV	XGBoost	350-1655	1305	0.99	28.21	0.912	3.37	0.4/100
SNV	RF-GBRT	600-999	200	0.99	28.27	0.922	3.57	0.4/300
SNV	PSO-GBRT	602-999	202	0.99	28.27	0.924	3.63	0.2/400
SNV	GA-GBRT	600-999	184	0.99	28.27	0.932	3.83	0.4/300
SNV	SA-GBRT	602-998	206	0.99	28.27	0.941	4.11	0.4/300
SNV	GGA-GBRT	601-999	202	0.99	28.27	0.919	3.52	0.2/400
SNV	RF-AdaBoost	600-999	200	0.99	21.24	0.939	4.06	0.2/100
SNV	PSO-AdaBoost	602-999	202	0.99	18.17	0.944	4.24	0.1/100
SNV	GA- AdaBoost	600-999	184	0.99	12.95	0.940	4.09	0.4/100
SNV	SA-AdaBoost	602-998	206	0.99	24.03	0.937	3.96	0.2/100
SNV	GGA-AdaBoost	601-999	202	0.99	24.21	0.943	4.20	0.4/100
SNV	RF-XGBoost	600-999	200	0.99	28.17	0.929	3.76	0.2/100
SNV	PSO-XGBoost	602-999	202	0.99	28.25	0.821	3.36	0.2/400
SNV	GA-XGBoost	600-999	184	0.99	28.22	0.886	2.96	0.3/100
SNV	SA-XGBoost	602-998	206	0.99	20.21	0.834	2.46	0.1/200
SNV	GGA-XGBoost	601-999	202	0.99	27.96	0.871	2.78	0.1/200
SG+SNV	AdaBoost	350-1655	1305	0.99	20.07	0.912	3.37	0.1/200
SG+SNV	GBRT	350-1655	1305	0.99	15.13	0.910	3.33	0.4/100
SG+SNV	XGBoost	350-1655	1305	0.99	25.26	0.908	3.31	0.2/100
SG+SNV	RF-GBRT	603-999	202	0.99	28.27	0.919	3.53	0.2/200
SG+SNV	PSO-GBRT	607-999	201	0.99	28.27	0.913	3.39	0.1/300
SG+SNV	GA-GBRT	604-999	209	0.99	28.27	0.927	3.69	0.4/100

表 3(续)

Preprocess method	Algorithm	Wavelength range /nm	Number of variables	Training set		Testing set		Parameter
				R^2	RPD	R^2	RPD	Learning rate/number of estimators
SG+SNV	SA-GBRT	607–998	206	0.99	28.27	0.926	3.68	0.4/200
SG+SNV	GGA-GBRT	607–999	200	0.99	28.27	0.900	3.17	0.3/400
SG+SNV	RF-AdaBoost	603–999	202	0.99	12.12	0.919	3.51	0.3/100
SG+SNV	PSO-AdaBoost	607–999	201	0.99	15.41	0.915	3.43	0.4/100
SG+SNV	GA-AdaBoost	604–999	209	0.99	24.16	0.922	3.59	0.2/100
SG+SNV	SA-AdaBoost	607–998	206	0.99	20.24	0.929	3.76	0.1/100
SG+SNV	GGA-AdaBoost	607–999	200	0.99	12.45	0.924	3.64	0.3/300
SG+SNV	RF-XGBoost	600–999	200	0.99	28.18	0.898	3.14	0.4/200
SG+SNV	PSO-XGBoost	602–999	202	0.99	27.45	0.888	2.99	0.4/100
SG+SNV	GA-XGBoost	600–999	184	0.99	24.22	0.885	2.95	0.3/100
SG+SNV	SA-XGBoost	602–998	206	0.99	27.31	0.883	2.92	0.4/100
SG+SNV	GGA-XGBoost	601–999	202	0.99	24.84	0.882	2.91	0.3/100

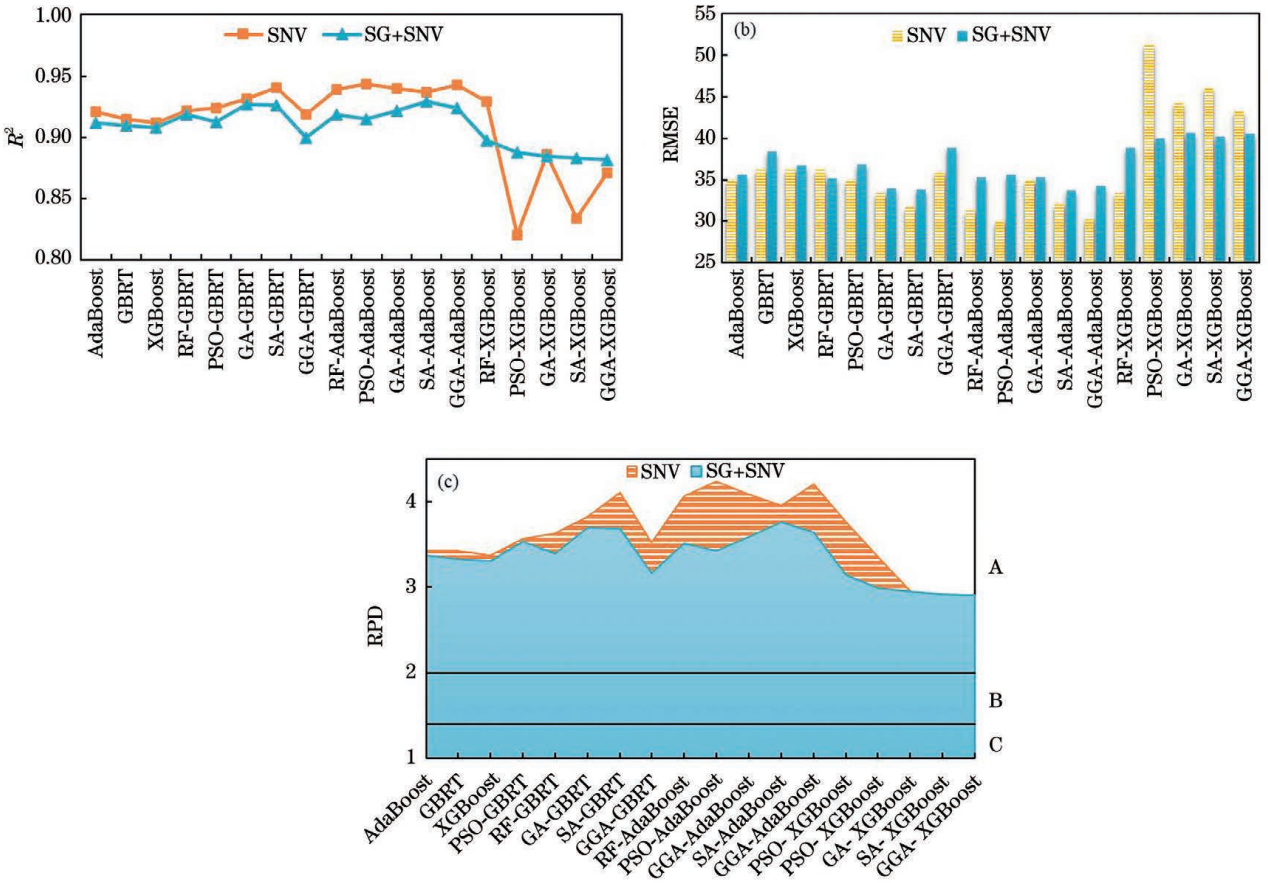

图 7　不同算法的测试集的 R^2、RMSE 和 RPD 值。(a) R^2;(b) RMSE;(c) RPD

Fig. 7　R^2, RMSE, and RPD values of the testing sets of different algorithms. (a) R^2; (b) RMSE; (c) RPD

从表 3 和图 7 可以看出:以特征分析后的变量建立的速效氮定量模型预测精度均为 A 级,在简化模型的同时提高了模型的预测精度;在回归算法方面,AdaBoost 和 GBRT 建模性能要优于 XGBoost;

在预处理方面可以看出，不同预处理方法的建模精度有所差异，相比 SG+SNV，SNV 预处理变换下建立的模型预测精度更高，性能最优。

3.4 最优光谱特征区间分析

测试集的 R^2 均大于 0.94 的模型是 PSO-AdaBoost、GGA-AdaBoost、SA-GBRT、GA-AdaBoost，图 8 为 MW 结合这 4 种特征选择算法优选的特征波长点组合，可以看出，所选波长大多集中在 600～1000 nm，该区域土壤的光谱特性主要是 C—H、O—H、N—H 基团的二级和三级倍频振动吸收引起的。在 SNV 预处理的数据下，基于 PSO 优选的特征变量建立的 AdaBoost 校正模型性能最好，在测试集中实测值和预测值的关系如图 9 所示，该模型预测等级为 A，在减少模型变量的同时提高了模型精度，用于建模的光谱数据点由 1035 个减少到 202 个，预测集的 R^2 提高到了 0.944，说明模型具有较强的泛化性和稳定性。

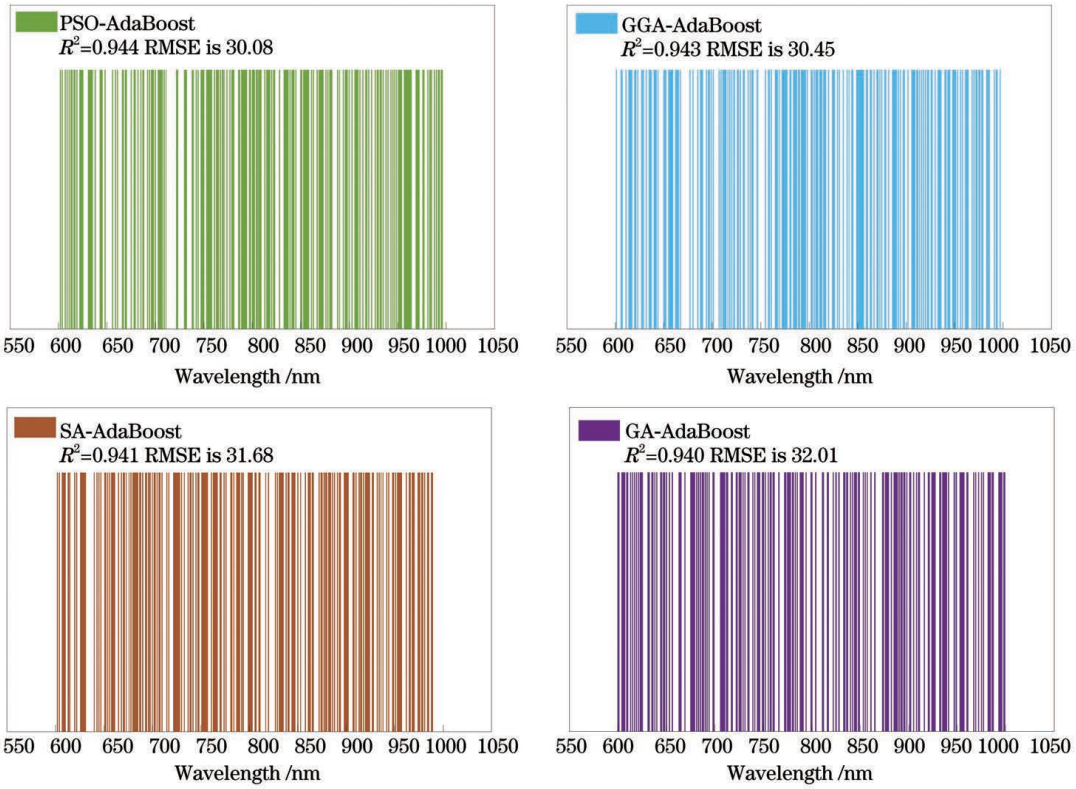

图 8　不同算法选择的最优波长点组合

Fig. 8　Optimal combination of wavelength points selected by different algorithms

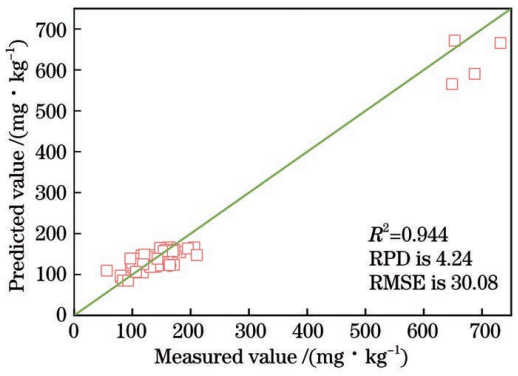

图 9　基于 SNV 的 PSO-AdaBoost 模型预测集中的测量值和预测值

Fig. 9　Measured and predicted values of PSO-AdaBoost model based on SNV in prediction set

4　结　　论

基于 7 种不同预处理方法校正的光谱数据，将移动窗口法和智能优化算法相结合，筛选了土壤速效氮相关的光谱特征点，并基于 Boosting 算法分析建立了不同的土壤速效氮定量分析模型(共 36 个)。实验结果表明：在预测土壤速效氮含量的算法中，Boosting 算法的性能优于传统 PLSR 算法，其中 AdaBoost 对土壤速效氮的拟合效果最好；同时得出并不是特征越多，建模效果越好，对特征进行优化后可有效剔除无用变量，提高模型性能。在 SNV 预处理下建立的 PSO-AdaBoost 校正模型使模型精度提升到了 0.944(波长点分布在 600～1000 nm)，既

显著地简化了模型，又保留了一定的数据冗余，提高了模型的稳健性。纪文君等[5-7]测量并分析了不同地区土壤样品，发现有机质的光谱响应区域均集中在 600～800 nm 波段。速效氮和有机质含量在土壤中呈正相关关系[30-31]，故在光谱分析中，它们的特征波段也具有一定的一致性。又由于氮素在近红外光谱中的直接响应波段主要是有机态的 N—H 键发生能级跃迁的倍频峰和合频峰，且速效氮含量较低，一般仅占总氮的 5%，因此相比有机质，确实需要更多的间接光谱数据信息来反映速效氮的光谱特征。这也证明了实验结果的合理性。

参 考 文 献

[1] Mukherjee S, Laskar S. Vis-NIR-based optical sensor system for estimation of primary nutrients in soil[J]. Journal of Optics, 2019, 48(1): 87-103.

[2] Liu H J, Lü J, Lin M, et al. Application of characteristic wavelength selection method in NIR model of green tea based on genetic algorithms[J]. Journal of Instrumental Analysis, 2007, 26(5): 679-681, 685.
刘辉军, 吕进, 林敏, 等. 基于遗传算法的波长选择方法在绿茶近红外光谱分析模型中的应用[J]. 分析测试学报, 2007, 26(5): 679-681, 685.

[3] Zhang Y, Li M Z, Zheng L H, et al. Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm[J]. Geoderma, 2019, 333: 23-34.

[4] Xie H T, Zhao J S, Wang Q B, et al. Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis[J]. Scientific Reports, 2015, 5: 10930.

[5] Ji W J, Shi Z, Zhou Q, et al. VIS-NIR reflectance spectroscopy of the organic matter in several types of soils[J]. Journal of Infrared and Millimeter Waves, 2012, 31(3): 277-282.
纪文君, 史舟, 周清, 等. 几种不同类型土壤的 VIS-NIR 光谱特性及有机质响应波段[J]. 红外与毫米波学报, 2012, 31(3): 277-282.

[6] Liu H J, Zhang Y Z, Zhang B. Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China[J]. Environmental Monitoring and Assessment, 2008, 154(1/2/3/4): 147-154.

[7] 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.

[8] Shi Z J, Li P F, Lü Y, et al. Region optimization in FT-NIR spectroscopy for determination of MDA in lard with moving window partial least squares[J]. Journal of Chinese Institute of Food Science and Technology, 2014, 14(11): 207-213.
史智佳, 李鹏飞, 吕玉, 等. 移动窗口偏最小二乘法优选猪油丙二醛近红外光谱波段[J]. 中国食品学报, 2014, 14(11): 207-213.

[9] Cheng B, Chen D Z, Wu X H. Near infrared spectral wavelength selection based on moving window-iterative genetic algorithm method [J]. Chinese Journal of Analytical Chemistry, 2006, 34(S1): 123-126.
成飙, 陈德钊, 吴晓华. 基于移动窗口-迭代遗传算法的近红外光谱波长选择方法[J]. 分析化学, 2006, 34(S1): 123-126.

[10] Wang X H, Zheng X L, Han Z Z, et al. Random forests-based hybrid feature selection algorithm for soil potassium content inversion using hyperspectral technology[J]. Spectroscopy and Spectral Analysis, 2018, 38(12): 3883-3889.
王轩慧, 郑西来, 韩仲志, 等. 混合式随机森林的土壤钾含量高光谱反演[J]. 光谱学与光谱分析, 2018, 38(12): 3883-3889.

[11] Kong Q Q, Ding X Q, Gong H L, et al. Research on application of feature selection algorithm based on combination of random forest and game theory in near infrared spectroscopy [J]. Journal of Instrumental Analysis, 2017, 36(10): 1203-1207.
孔清清, 丁香乾, 宫会丽, 等. 基于随机森林结合博弈论的特征选择算法在近红外光谱分类中的应用研究[J]. 分析测试学报, 2017, 36(10): 1203-1207.

[12] 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.
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.

[13] Tu Z H, Ji B P, Meng C Y, et al. Analysis of NIR characteristic wavelengths for apple flesh firmness based on GA and iPLS[J]. Spectroscopy and Spectral Analysis, 2009, 29(10): 2760-2764.
屠振华, 籍保平, 孟超英, 等. 基于遗传算法和间隔偏最小二乘的苹果硬度特征波长分析研究[J]. 光谱学与光谱分析, 2009, 29(10): 2760-2764.

[14] Huang F, Zhang X K, Sun L, et al. Vibration

spectral component analysis based on genetic algorithm and simulated annealing algorithm[J]. Laser & Optoelectronics Progress, 2020, 57(9): 093001.
黄凡, 张旭坤, 孙陆, 等. 基于遗传算法和模拟退火算法的振动光谱成分分析算法[J]. 激光与光电子学进展, 2020, 57(9): 093001.

[15] Dou Y, Sun X R, Liu C L, et al. Near-infrared spectroscopic detection of wheat flour quality using wavelength optimization based on simulated annealing algorithm(SAA)[J]. Food Science, 2016, 37(12): 208-211.
窦颖, 孙晓荣, 刘翠玲, 等. 基于模拟退火算法优化波长的面粉品质检测[J]. 食品科学, 2016, 37(12): 208-211.

[16] Zhao J F, Huang T L, Pang F, et al. Genetic algorithm based on greedy strategy in the 0-1 knapsack problem[C]//2009 Third International Conference on Genetic and Evolutionary Computing, October 14-17, 2009, Guilin, China. New York: IEEE Press, 2009: 105-107.

[17] Bai Q H. Analysis of particle swarm optimization algorithm[J]. Computer and Information Science, 2010, 3(1): 180-184.

[18] Pan L J, Chen R F, Cui R F, 等. Adaptive selection method for analytical lines in laser-induced breakdown spectra[J]. Chinese Journal of Lases, 2020, 47(8): 0811001
潘立剑, 陈蔚芳, 崔榕芳, 等. 激光诱导击穿光谱中分析谱线自适应选择方法[J]. 中国激光, 2020, 47(8): 0811001.

[19] Sakri S B, Abdul Rashid N B, Muhammad Zain Z. Particle swarm optimization feature selection for breast cancer recurrence prediction[J]. IEEE Access, 2018, 6: 29637-29647.

[20] Sampaio P S, Soares A, Castanho A, et al. Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms[J]. Food Chemistry, 2018, 242: 196-204.

[21] Min H, Luo X L. Calibration of soft sensor by using Just-in-time modeling and AdaBoost learning method [J]. Chinese Journal of Chemical Engineering, 2016, 24(8): 1038-1046.

[22] Persson C, Bacher P, Shiga T, et al. Multi-site solar power forecasting using gradient boosted regression trees[J]. Solar Energy, 2017, 150: 423-436.

[23] 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.
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.

[24] 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, 2016, San Francisco, California, USA. New York: ACM, 2016: 785-794.

[25] Tao M Q, Liu J X, Wu Y, et al. Application of XGBoost in gas infrared spectral recognition[J]. Acta Optica Sinica, 2020, 40(7): 0730002.
陶孟琪, 刘家祥, 吴越, 等. XGBoost 在气体红外光谱识别中的应用[J]. 光学学报, 2020, 40(7): 0730002.

[26] Ke G L, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree[C]// Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. New York: Curran Associates, 2017: 3146-3154.

[27] Chang C W, Laird D A, Mausbach M J, et al. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties[J]. Soil Science Society of America Journal, 2001, 65(2): 480-490.

[28] Lee L C, Liong C Y, Jemain A A. Iterative random vs. Kennard-Stone sampling for IR spectrum-based classification task using PLS2-DA[J]. AIP Conference Proceedings, 2018, 1940: 020116.

[29] Ji W, Viscarra Rossel R A, Shi Z. Accounting for the effects of water and the environment on proximally sensed Vis-NIR soil spectra and their calibrations[J]. European Journal of Soil Science, 2015, 66(3): 555-565.

[30] Ma L Y. Effect on the distribution of organic matter content on the available nitrogen in different layer of soil[J]. Chinese Agricultural Science Bulletin, 2010, 26(24): 193-196.
马麟英. 不同土层土壤有机质含量对速效氮分配的影响[J]. 中国农学通报, 2010, 26(24): 193-196.

[31] Zhao Y, Chen W, Li C M, et al. Content of soil organic matter and its relationships with main nutrients on degraded alpine meadow in Eastern Qilian Mountains[J]. Pratacultural Science, 2009, 26(5): 20-25.
赵云, 陈伟, 李春鸣, 等. 东祁连山不同退化程度高寒草甸土壤有机质含量及其与主要养分的关系[J]. 草业科学, 2009, 26(5): 20-25.