

BA-Adaboost 模型的黑土区土壤养分含量高光谱估测

林楠^{1,2}, 刘海琪³, 杨佳佳⁴, 吴梦红¹, 刘翰霖¹

1. 吉林建筑大学测绘与勘查工程学院, 吉林 长春 130118
2. 吉林大学地球科学学院, 吉林 长春 130026
3. 东北大学资源与土木工程学院, 辽宁 沈阳 110819
4. 中国地质调查局沈阳地质调查中心, 辽宁 沈阳 110034

摘要 黑土中的有机质、磷和钾等养分元素在作物生长过程中起着至关重要的作用, 研究黑土养分元素的分布特征, 开展元素含量的定量计算, 对黑土地的科学管理和环境保护具有重要意义。基于黑龙江省讷河市 80 份黑土样品和高光谱实测数据, 分析了光谱反射率、反射率一阶微分、反射率倒数对数、反射率倒数对数一阶微分与土壤有机质、磷元素和钾元素含量的相关性, 并利用相关系数法提取敏感波段。针对机器学习模型中参数值优化选择问题, 引入蝙蝠算法(BA)并与 Adaboost 模型相结合, 利用 BA 对 Adaboost 模型中的最大迭代次数 n 和弱学习器权重缩减系数 v 两个核心参数进行寻优计算, 选择 CART 决策树为模型的弱回归学习器, 决定系数作为参数优化的目标函数值, 构建 BA-Adaboost 土壤养分含量高光谱预测模型, 定量估测土壤有机质、磷元素和钾元素含量, 结果表明: BA-Adaboost 组合模型可以快速搜索全局最优参数, 经 BA 优化后的 Adaboost 模型精度和可靠性显著提高, 3 种元素中, 土壤有机质估测精度最高, 决定系数和均方根误差分别为 0.864 和 $0.152 \text{ g} \cdot \text{kg}^{-1}$, 对比优化前模型预测精度提高了 14.2% 和 25.4%, 说明构建的 BA-Adaboost 模型在土壤元素含量高光谱估测中具有一定的应用前景, 是一种高效的估测方法。

关键词 黑土区; 土壤养分含量; 高光谱估测; 蝙蝠算法; Adaboost 模型

中图分类号: TP79 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2020)12-3825-07

引言

土壤养分含量快速估测对作物生长监测和黑土科学管理具有重要的意义。高光谱数据可以通过数千个紧密排列的波长记录土壤信息。众多研究表明, 土壤高光谱数据与土壤含水率^[1]、养分含量^[2-3]、重金属含量^[4]等土壤指标存在密切关系, 利用高光谱技术进行土壤元素含量估测已取得了丰硕的研究成果。

目前, 利用机器学习算法建立土壤高光谱特征和元素含量之间的反演模型是土壤元素含量估测的主要方法, 其中支持向量机^[5]、极限学习机^[6]、随机森林^[7]等模型都得到了广泛的应用。机器学习算法通常需要预先设定一些模型参数值, 由于预定义的参数值很可能不包含全局最优的参数值, 导致机器学习模型达不到最佳效果。为了克服机器学习模型在寻找最佳模型参数时所存在的问题, 遗传算法、粒子群算法等经典优化算法被用来优化机器学习模型的内部参数。然

而粒子群等经典优化算法对初始参数设置比较敏感, 寻优过程中容易陷入局部最优解, 导致算法后期的收敛速度变慢。蝙蝠算法是一种新兴的群体智能方法, 在寻优过程中, 蝙蝠算法模仿蝙蝠声波脉冲响度、频率的自适应调节过程, 实现全局寻优和局部寻优过程的自由切换, 从而使算法的全局搜索能力和局部搜索能力得到平衡, 在模型参数寻优计算方面表现优异^[8]。本工作对黑土土壤有机质、磷元素和钾元素的光谱特征进行详细分析, 选择元素最佳光谱变换形式及相关性较高的敏感波段, 将蝙蝠算法和 Adaboost 机器学习模型组合构建土壤养分含量估测模型, 利用蝙蝠算法求解 Adaboost 模型建模中的关键参数, 并比较模型参数优化前后的估测精度, 为黑土养分含量高光谱估测提供一种高效的新方法。

1 实验部分

1.1 土壤样品采集

土壤样品采集区位于黑龙江省讷河市, 该区地处松辽平

收稿日期: 2020-05-21, 修订日期: 2020-09-06

基金项目: 国家自然科学基金项目(41702357), 吉林省教育厅科学技术项目(JJKH20180608KJ)资助

作者简介: 林楠, 1984年生, 吉林建筑大学测绘与勘查工程学院副教授 e-mail: linnanzc@126.com

原的北端,大小兴安岭南缘,是黑土分布的典型区域。区内耕地面积广阔,农产品种类丰富,是甜菜和马铃薯的盛产地。2019年4月在该区内挑选典型黑土地块,设置5公里大小的网格作为采样单元,采样点设计结合第二次全国土地调查成果合理分布,所布设的采样点要能够代表采样单元的土

壤性质,试验区内共采集土壤样本80个(图1),采集样品过程中确保采样距离公路至少150m,以每个采样点位置为中心,在其周围5m×5m范围内进行样品采集,每个采样点位置共采集5份黑土样本,取样深度为表土层的15cm以内,将样本充分混合后装入采样袋。

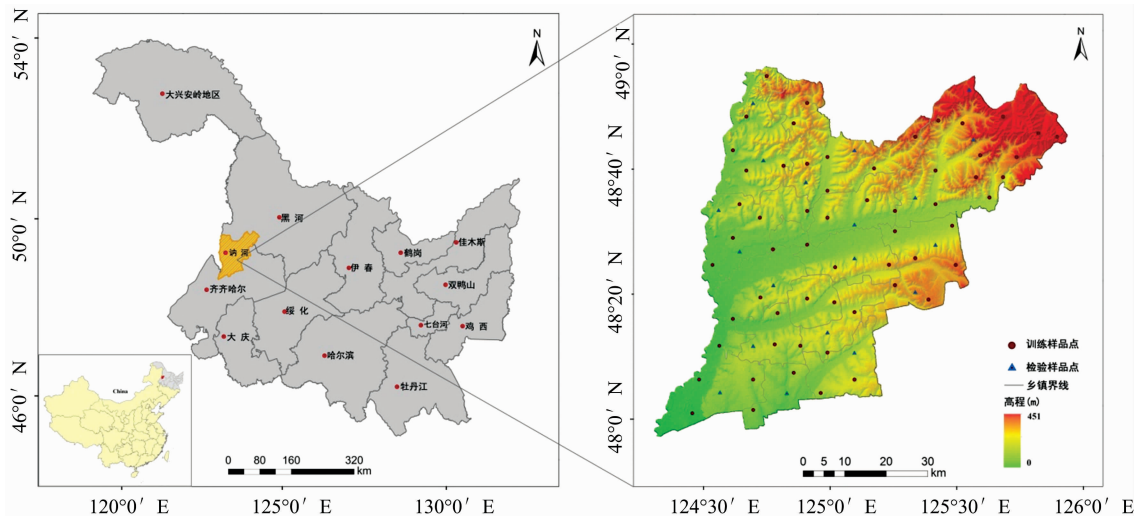


图1 研究区和样品采集点地理位置

Fig. 1 Location of study area and sampling collection points

1.2 元素含量测量及样本划分

将样品中的秸秆、砂砾等杂物剔除,风干后研磨过筛,使得土壤的粒径小于0.25mm;将样品分成两份,一份用于测定土壤元素含量,另一份用于室内高光谱测量。结合多目标样品分析质量要求等技术规范,参照《土壤农业化学分析方法》,选用重铬酸钾容量法测定土壤有机质含量,选用X射线荧光光谱法进行土壤磷元素和钾元素含量测量^[9]。对元素含量的测量结果进行统计分析(表1),将80份土壤样品按元素含量由低到高分为20组,每组随机抽取1个样品放入验证集,共20个样品作为验证集,其余60个样品为训练集。

表1 土壤样本养分元素含量基本信息

Table 1 Statistics of nutrient contents in soil samples

养分元素	样本数	最小值/ (g·kg ⁻¹)	最大值/ (g·kg ⁻¹)	均值/ (g·kg ⁻¹)	标准差/ (g·kg ⁻¹)
有机质	80	2.138	5.517	3.364	0.535
磷	80	0.452	0.988	0.682	0.087
钾	80	1.875	2.240	2.036	0.062

1.3 光谱采集及预处理

在暗室中进行土壤光谱测量,测量仪选用ASD Filed-Spec4地物波谱仪,为提高光谱测量数据的精度,取5次光谱测量的均值作为土壤样品的反射光谱数据。由于光谱数据在350~399nm以及2451~2500nm范围的噪声较大,信噪比低,对土壤元素与反射率关系的分析会造成干扰,所以将400nm前及2450nm后的数据剔除。光谱仪采样间隔为1nm,即在400~2450nm范围内得到2051个波段,由于光谱分辨率较高,波段数较多,相邻波段之间可能存在信息

重叠,更容易受到噪声的影响,因此,将光谱数据进行重采样处理,设置采样间隔为10nm。在去噪和重采样处理的基础上,将原始光谱反射率进行一阶微分、倒数对数等特征变换,不同的变换形式可以帮助准确、快速地找到峰谷,并通过峰谷确定相应的波长,从而确定敏感波段。

1.4 模型概述

蝙蝠算法(Bat algorithm, BA)是一种启发式搜索算法,它模拟蝙蝠使用声纳探测猎物和躲避障碍物,通过模拟蝙蝠飞行寻找猎物的过程来模拟优化搜索过程,在计算过程中利用求解问题的适应度值来选取蝙蝠的位置,利用优胜劣汰的进化过程来模拟较优可行解代替较差可行解的迭代搜索过程^[10]。基于BA的基本原理,在算法各项参数被初始化后,从d维搜索空间中的一个随机位置 z_i 开始启发式搜索。以固定的频率、不同的波长和音强搜索猎物,搜索过程中,蝙蝠根据接近猎物的距离自动调整波长的大小。经全局搜索后更新每只蝙蝠的飞行速度和空间位置,并计算目标函数的适应度值,速度和空间位置更新公式如式(1)^[11]

$$\begin{aligned}
 f_i &= f_{\min} + (f_{\max} - f_{\min})\beta \\
 v_i^{t+1} &= v_i^t + (z_i^t - z^*)f_i \\
 z_i^{t+1} &= z_i^t + v_i^{t+1}
 \end{aligned} \quad (1)$$

其中, v_i^t 和 v_i^{t+1} 分别表示蝙蝠个体*i*在*t*和*t+1*时刻的飞行速度; z_i^t 和 z_i^{t+1} 分别表示蝙蝠个体*i*在*t*和*t+1*时刻的位置; z^* 代表全局最优位置。频率 f_i 为蝙蝠个体*i*在搜索时的脉冲频率,其中, β 为介于[0,1]间的随机数, (f_{\min}, f_{\max}) 为脉冲频率范围。每次迭代过后,音强和频度会根据脉冲响度衰减系数和脉冲频度增加系数进行更新计算。

Adaboost算法基本思想是针对同一训练集训练能力一

般的弱回归器,通过叠加方法构建一个训练能力很强的强回归器。其算法本身是通过改变样本分布权重来实现的,根据每次训练集中每个样本的预测精度以及上次的总体预测精度来计算每个弱回归器的权重,同时更新每个样本的分布权重,最后将每次训练得到的回归器结果加权求和,作为强回归器最后的输出结果^[12]。建模过程中,弱学习器的最大迭代次数 n 和弱学习器的权重缩减系数 v 是两个重要的参数,迭代次数 n 设置过小会导致模型拟合不充分, n 设置太大则会导致模型拟合过度,而较小的权重缩减系数 v 意味着需要更多的弱学习器的迭代次数,所以通常这两个参数要一起进行优化调整^[13]。

1.5 模型精度检验

为了评价模型预测能力及模型的稳定性,选择决定系数 (R^2) 和均方根误差 (RMSE) 对模型的建模效果进行评价

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - Y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - Y_i)^2} \quad (2)$$

其中 n 为样本量, y_i 为实测值, Y_i 为预测值, \bar{y} 为实测值的平均值。预测模型计算的 RMSE 越低, R^2 越接近 1, 预测模型的准确性和稳定性越高。

2 结果与讨论

2.1 光谱特征变换

经处理后的室内土壤样品光谱曲线如图 2(a) 所示,从图中可以看出,采集的黑土样品原始反射率在 $0 \sim 0.9$ 之间,各样品光谱曲线波动形状相似,在可见光波段,随着波长的增加,反射率逐渐增加,直到 $1\ 200\ \text{nm}$ 处反射率趋于稳定。近红外区土壤光谱反射率总体高于可见光区,两个明显的波谷分布在 $1\ 400$ 和 $1\ 900\ \text{nm}$ 附近,主要是土壤中残留的水分和空气中水蒸气的吸收造成的,而在 $2\ 200\ \text{nm}$ 处有轻微的凹陷,是受土壤中存在的粘土矿物的影响。图 2(b)–(d) 分别为原始反射率经一阶微分 (R')、倒数对数 ($\lg 1/R$)、倒数对数一阶微分 $[(\lg 1/R)']$ 变换后的光谱曲线,从变化结果可以看出,一阶微分变换可以对原始光谱变化起到放大作用,经变换后反射率在 $1\ 400$, $1\ 900$ 和 $2\ 200\ \text{nm}$ 处的波动更大。

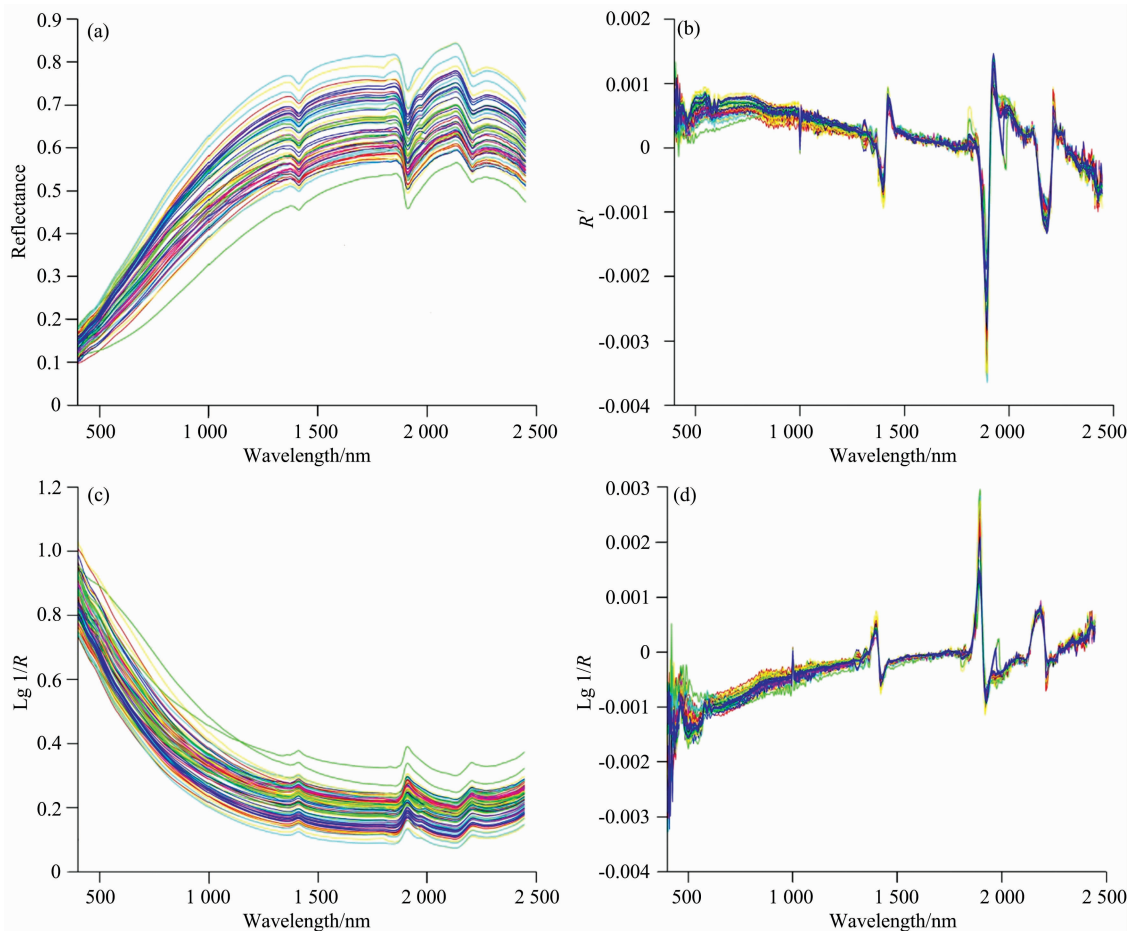


图 2 反射率及其变换的土壤光谱曲线

Fig. 2 Soil spectral reflectance curves and its transformations

2.2 相关性分析及特征波段选择

分别计算了土壤有机质、磷、钾含量与土壤反射率的相关系数,并绘制相关系数曲线图(图 3)。从图 3(a)原始光谱的相关系数可以看出,土壤有机质和磷元素含量与光谱反射率呈负相关,而钾元素含量则相反。与原始光谱反射率相

比,变换后的光谱数据与土壤有机质、磷、钾含量的相关性更高,其中,一阶微分变换形式与土壤有机质、磷、钾含量的相关系数表现为正、负交叉,波峰和波谷较多,而且经一阶微分变换后,各元素的最高相关系数显著提高。

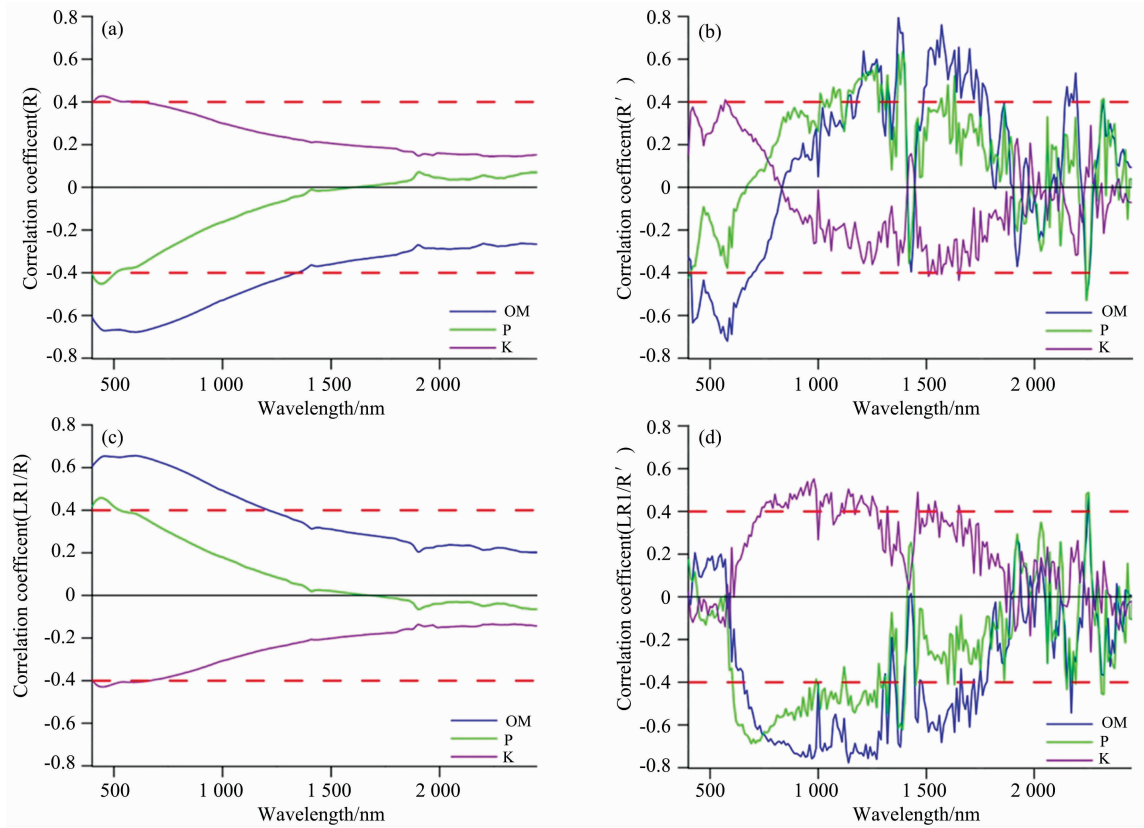


图 3 土壤元素含量与光谱相关系数曲线

Fig. 3 Correlation coefficient curves of soil element content and spectral reflectance

选取相关系数大于 0.4 的敏感波段作为预测模型的样本输入数据(表 2)。由统计结果可以看出,经过不同的光谱特征变换,土壤元素含量与光谱反射率的相关系数有所提高,其中,土壤有机质对应的最佳变换形式为一阶微分,与光谱反射率数据的相关系数最高为 0.796,波长范围在 1 370 mm

附近;土壤磷元素对应的最佳变换形式为倒数对数的一阶微分,相关系数最高为-0.688,波长范围在 690 mm 附近;土壤钾元素对应的最佳变换形式为倒数对数一阶微分,相关系数最高为 0.553,波长范围在 980 mm 附近。

表 2 最大相关系数和敏感波段

Table 2 Maximum correlation coefficients and sensitive bands

元素		光谱变化形式			
		R	R'	$\lg 1/R$	$(\lg 1/R)'$
有机质	最大值	-0.678	0.796	0.656	-0.778
	对应波段	600	1 370	600	1 140
	敏感波段数量	95	85	82	101
磷元素	最大值	-0.452	0.637	0.458	-0.688
	对应波段	440	1 390	440	690
	敏感波段数量	11	38	13	82
钾元素	最大值	0.428	-0.436	-0.429	0.553
	对应波段	450	1 650	700~710	980
	敏感波段数量	23	5	26	51

2.3 基于 BA-AdaBoost 模型的预测

将选取 60 个训练样本最佳光谱变换形式的敏感波段和对应元素含量值作为 Adaboost 模型的建模数据, 3 种元素的建模波段数分别为 85 个、82 个和 51 个, 选择 CART 决策树为 Adaboost 模型的弱回归学习器, 采用蝙蝠搜索算法对 Adaboost 模型的最大迭代次数 n 和弱学习器权重缩减系数 v 进行寻优。

基于 BA 的基本原理, 首先需对模型各项参数进行初始化, BA 初始化参数较多, 但除了迭代次数外其他参数敏感性不强, 可选择默认缺省参数进行初始化。BA 算法的搜索空间是由最大迭代次数 n 和弱学习器权重缩减系数 v 为坐标轴组成的二维空间, 迭代搜索过程从搜索空间内的 L 个随机位置开始, 在每次迭代过程中, 利用每只蝙蝠在空间中位置

的二维坐标 (n, v) 作为 Adaboost 模型的初始化参数, 然后基于 Adaboost 模型对样本数据训练建模, 并计算预测结果。根据计算结果选取 R^2 值最大时对应的位置作为蝙蝠当前最优位置, 利用式(1)更新每只蝙蝠的空间位置。设置种群大小 $L=20$, 脉冲频率范围 $f_{\min}=0, f_{\max}=1$ 、脉冲音强范围 $A_{\min}=0, A_{\max}=1$, 脉冲响度衰减系数 $\alpha=0.9$, 脉冲频度增加系数 $\gamma=0.9$, 作为缺省参数, 并计算了不同迭代次数对应的 R^2 值(图 4)。从图中可以看出, 随着迭代次数 T 的增加, R^2 值逐渐增大, 3 种元素中, 钾元素收敛速度最快, 当迭代数达到 10 次的时候, R^2 值达到了最大, 有机质和磷元素分别是当迭代数达到 12 次和 16 次达到最大。建模估测时, 选取 R^2 最大时对应的 n 和 v 的值作为 Adaboost 模型的建模参数。

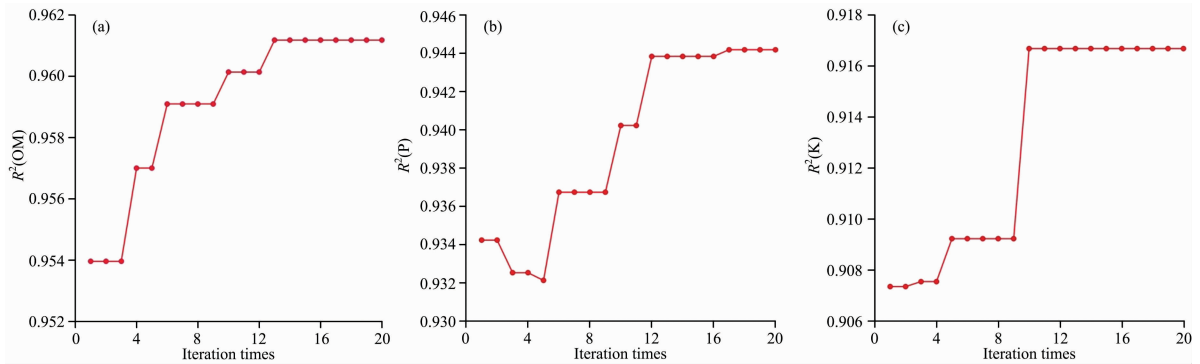


图 4 BA-Adaboost 模型 R^2 变化曲线

Fig. 4 R^2 variation curves of BA-Adaboost

2.4 模型精度分析

为了分析 BA-Adaboost 模型的预测精度, 分别计算了 BA 优化前后模型训练集和验证集的 R^2 和 RMSE(表 3)。从表中可以看出无论是训练集还是验证集有机质的高光谱预测

精度均高于磷元素和钾元素, 经过 BA 寻优后, Adaboost 模型的精度有了较大的提升, 从验证集结果来看, 磷元素的 R^2 变化最为显著, 由 0.695 提高至 0.808; 有机质的 RMSEP 变化最大, 由 $0.204 \text{ g} \cdot \text{kg}^{-1}$ 下降至 $0.152 \text{ g} \cdot \text{kg}^{-1}$ 。

表 3 模型预测精度对比分析

Table 3 Comparison of estimation accuracies

元素	光谱变化形式	预测模型	训练集		验证集	
			R^2_t	RMSEC	R^2_v	RMSEP
有机质	一阶微分	Adaboost	0.925	0.140	0.756	0.204
		BA-Adaboost	0.961	0.101	0.864	0.152
磷元素	倒数一阶微分	Adaboost	0.909	0.025	0.695	0.039
		BA-Adaboost	0.944	0.019	0.808	0.031
钾元素	倒数一阶微分	Adaboost	0.877	0.019	0.651	0.029
		BA-Adaboost	0.917	0.016	0.761	0.024

利用模型的预测值和实测值绘制预测结果拟合图, 进一步对比分析 3 种元素模型拟合效果(图 5)。从图中可以看出, 与 Adaboost 相比, 优化后的 BA-Adaboost 模型的预测值更紧密地分布在 1 : 1 左右, 数据拟合能力和稳定性优于 Adaboost 模型。同时可以看出训练集的拟合效果比验证集更优异, 验证集样本点分布相对分散, 说明机器学习模型对训练样本数据的可靠性要求较高。

3 结 论

以黑龙江省讷河市 80 个黑土样品为研究对象, 对黑土土壤有机质、磷元素和钾元素的光谱特征进行分析, 构建 BA-Adaboost 模型对元素含量进行估测, 主要结论如下:

(1) 光谱数据的特征变换处理可以增强光谱特征, 一阶

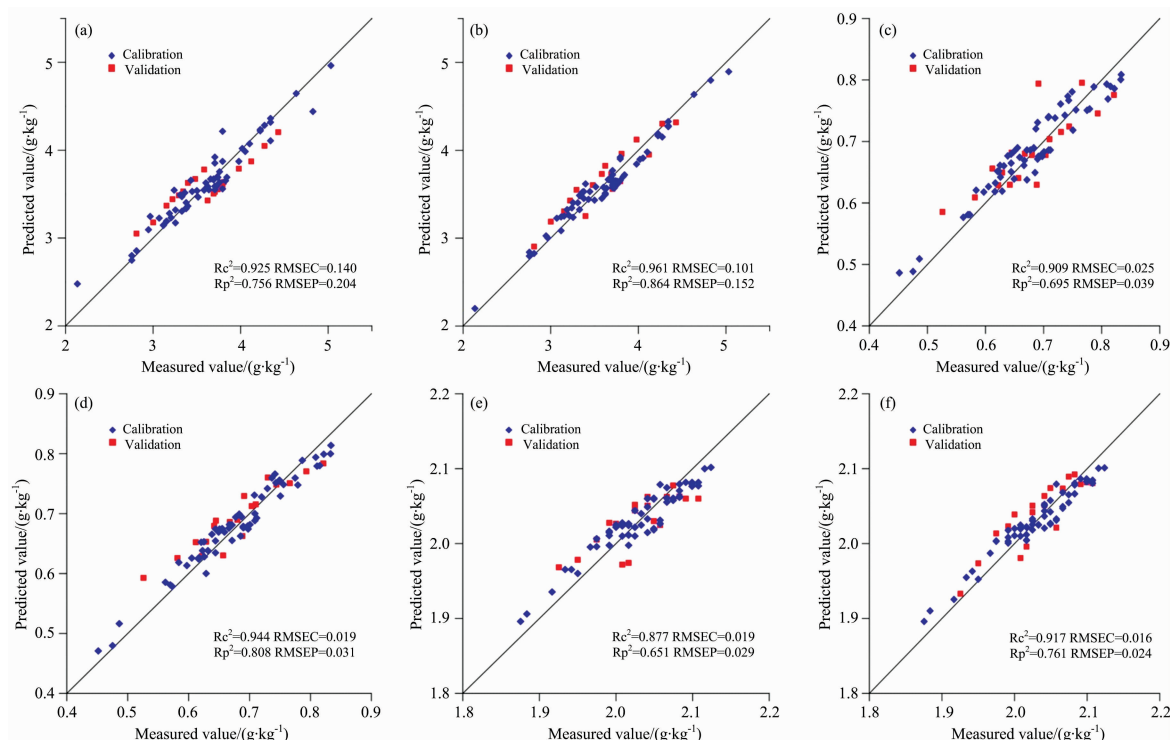


图 5 元素含量实测值和预测值

(a): OM 含量 Adaboost 模型; (b): OM 含量 BA-Adaboost 模型; (c): P 含量 Adaboost 模型;
(d): P 含量 BA-Adaboost 模型; (e): K 含量 Adaboost 模型; (f): K 含量 BA-Adaboost 模型

Fig. 5 Measured and predicted values of element content

(a): Adaboost of OM content; (b): BA-Adaboost of OM content; (c): Adaboost of P content;
(d): BA-Adaboost of P content; (e): Adaboost of K content; (f): BA-Adaboost of K content

微分变化有效地突出了光谱曲线的波峰和波谷, 提高了光谱反射率与元素含量的相关系数, 有机质、磷元素、钾元素分别在 1 370, 690 和 980 mm 处相关系数达到最大值。

(2) 蝙蝠算法在参数寻优过程中能够动态控制全局寻优和局部寻优过程之间的自由切换, 利用蝙蝠算法对 Adaboost 模型参数进行优化计算, 避免了估测模型陷入局部极小值, 与传统参数寻优方法相比, 该方法可更快更准确地找到全局最优参数。

(3) 将 BA 与 Adaboost 模型相结合, 构建 BA-Adaboost 土壤含量估测模型, 该组合模型只需设置搜索空间, 然后自动搜索模型最优参数值; 对比 BA 算法优化前后的估测精度可以看出, 优化后 BA-Adaboost 模型 R^2 增大、RMSE 变小, 估测精度明显提高, 说明 BA-Adaboost 模型在土壤元素含量高光谱估测中具有一定的适用性, 扩展了机器学习模型在土壤成分估测研究中的应用。

References

- [1] WU Long-guo, WANG Song-lei, HE Jian-guo(吴龙国, 王松磊, 何建国). Spectroscopy and Spectral Analysis(光谱学光谱分析), 2018, 38(8): 2563.
- [2] LIU Xiu-ying, SHI Zhao-yong, CHANG Qing-rui, et al(刘秀英, 石兆勇, 常庆瑞, 等). Acta Pedologica Sinica(土壤学报), 2018, 55(2): 325.
- [3] NIE Zhe, LI Xiu-fen, LÜ Jia-xin, et al(聂哲, 李秀芬, 吕家欣, 等). Chinese Journal of Soil Science(土壤通报), 2019, 303(6): 1285.
- [4] Kemper T, Sommer S. Environmental Science & Technology, 2002, 36(12): 2742.
- [5] TAN Kun, ZHANG Qian-qian, CAO Qian, et al(谭琨, 张倩倩, 曹茜, 等). Earth Science-Journal of China University of Geosciences(地球科学—中国地质大学学报), 2015, 40(8): 1339.
- [6] ZHOU Peng, YANG Wei, LI Min-zan, et al(周鹏, 杨玮, 李民赞, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2017, 48(S1): 271.
- [7] Tan K, Ma W, Wu F, et al. Environmental Monitoring and Assessment, 2019, 191(7): 1.
- [8] Chen Y, Wu W, Zhao Q. Minerals, 2019, 9(5): 317.

- [9] LU Ru-kun(鲁如坤). Methods of Soil Agricultural Chemical Analysis(土壤农业化学分析方法). Beijing: China Agriculture and Science Press(北京: 中国农业科技出版社), 2000. 638.
- [10] WU Chen-wen, LIU Xiao-guang, WEI Li-xin(吴辰文, 刘小光, 魏立鑫). Journal of Northwest University • Natural Science Edition(西北大学学报 • 自然科学版), 2019, 49(4): 597.
- [11] MENG Kai-lu, YUE Ke-qiang, SHANG Jun-na(孟凯露, 岳克强, 尚俊娜). Computer Engineering(计算机工程), 2019, 45(4): 189.
- [12] XU Ting, ZHANG Xiang, ZHANG Ya-kun, et al(徐婷, 张香, 张亚坤, 等). Journal of Safety and Environment(安全与环境学报), 2019, 19(4): 1273.
- [13] YANG Xiao, WANG Zhi-zhang, ZHOU Zi-yong, et al(杨笑, 王志章, 周子勇, 等). Acta Petrolei Sinica(石油学报), 2019, 40(4): 457.

Hyperspectral Estimation of Soil Nutrient Content in the Black Soil Region Based on BA-Adaboost

LIN Nan^{1,2}, LIU Hai-qi³, YANG Jia-jia⁴, WU Meng-hong¹, LIU Han-lin¹

1. College of Surveying and Prospecting Engineering, Jilin Jianzhu University, Changchun 130118, China

2. College of Earth Science, Jilin University, Changchun 130026, China

3. College of Resources and Civil Engineering, Northeastern University, Shenyang 110819, China

4. Shenyang Geological Survey Center, China Geological Survey, Shenyang 110034, China

Abstract Organic matter (OM), phosphorus (P) and potassium (K) in black soil play a crucial role during crop growth. Studying the distribution of nutrient elements in black soil and carrying out quantitative calculation of element content is of great significance to the scientific management of black soil and environmental protection. Based on 80 black soil samples collected from Nehe city, Heilongjiang province and hyperspectral measured data, the correlation between 4 forms of spectral reflectance including original, first differential, absorbance transformation, and first differential of absorbance and soil OM, P, K contents were calculated, and the correlation coefficient method was used to extract the sensitive bands. For the optimization of parameter values in the machine learning model, the bat algorithm (BA) was introduced and combined with Adaboost model. Using BA to optimize the two important parameters of Adaboost, namely, maximum iterations n and weak learner weight reduction factor v , selecting CART decision tree and determination coefficient as the weak regression learner of the model and objective function value of the parameter optimization respectively, the BA-Adaboost model was constructed for estimating soil OM, P, K contents. The results showed that the BA-Adaboost model could quickly search the optimal global parameters, and the accuracy and reliability of the Adaboost were improved significantly after BA optimization, where the prediction accuracy of OM was the highest among the three nutrient elements. And determination coefficient and root mean square error were 0.864 and $0.152 \text{ g} \cdot \text{kg}^{-1}$ respectively, which were improved by 14.2% and 25.4% compared with before optimization. Therefore, the constructed BA-Adaboost model has potential in the hyperspectral estimation of soil element content and is an efficient estimation method.

Keywords Black soil region; Soil nutrient content; Hyperspectral estimation; Bat algorithm; Adaboost model

(Received May 21, 2020; accepted Sep. 6, 2020)