doi:10.6041/j.issn.1000-1298.2024.08.035

基于中红外光谱特征增强和集成学习的土壤有机碳含量估算模型研究

唐澳华[1,2]　杨贵军[1,2]　杨　悦[2,3]　陈伟男[1,2]　徐新刚[2]　徐　波[2]　高美玲[1]　张　静[1]

(1.长安大学地质工程与测绘学院,西安 710054;
2.农业农村部农业遥感机理与定量遥感重点实验室,北京 100097;
3.长安大学土地工程学院,西安 710054)

摘要:中红外光谱数据在实现土壤有机碳含量的准确、低成本快速预测方面具有巨大潜力。为提高光谱数据估算模型的普适性,本研究利用光谱特征增强策略,并基于 Stacking 算法结合多种机器学习方法构建了一种高鲁棒性的土壤有机碳含量估算模型。采用多种光谱特征增强方法及其组合对土壤中红外光谱进行特征增强,筛选最佳策略;通过应用 Stacking 算法结合多种机器学习方法构建集成模型,以提高模型泛化能力;将集成模型估算性能与偏最小二乘回归模型(PLSR)、梯度提升树(GBT)和一维卷积神经网络(1D-CNN)模型进行比较分析。研究结果表明,最佳光谱特征增强策略可以显著提高土壤光谱数据与土壤有机碳含量的相关性,最佳 Pearson 相关系数达到 -0.82;相较于 PLSR、GBT 和 1D-CNN 等模型,集成模型在各光谱数据下均表现出较高的估算精度,特别是在一阶导变换结合多元散射校正的光谱特征增强策略下,集成模型展现出优良的估算性能(决定系数 R^2 = 0.92,均方根误差为 1.18 g/kg,相对分析误差为 3.52)。本研究方法能够快速、准确地估算土壤有机碳含量,可为现代农业管理提供科学依据。

关键词:土壤有机碳;估算模型;集成学习;光谱特征增强;中红外光谱

中图分类号:S151.9　　**文献标识码**:A　　**文章编号**:1000-1298(2024)08-0382-09　　**OSID**:

Estimation Model of Soil Organic Carbon Content Based on Mid-infrared Spectral Characteristics Enhancement and Ensemble Learning

TANG Aohua[1,2]　YANG Guijun[1,2]　YANG Yue[2,3]　CHEN Weinan[1,2]　XU Xin'gang[2]　XU Bo[2]
GAO Meiling[1]　ZHANG Jing[1]

(1. *College of Geological Engineering and Geomatics*, *Chang'an University*, *Xi'an* 710054, *China*
2. *Key Laboratory of Quantitative Remote Sensing in Agriculture*, *Ministry of Agriculture and Rural Affairs*, *Beijing* 100097, *China*
3. *School of Land Engineering*, *Chang'an University*, *Xi'an* 710054, *China*)

Abstract: Mid-infrared spectral data holds immense potential for accurate, cost-effective, and rapid prediction of soil organic carbon (SOC) content. To enhance the universality of spectroscopic data estimation models, a spectroscopic feature enhancement strategy was employed and combined multiple machine learning methods by using the Stacking algorithm to construct a robust model for estimating SOC content. Various spectroscopic feature enhancement methods and their combinations were applied to enhance the features of mid-infrared soil spectra and select the optimal strategies. The Stacking algorithm was used in conjunction with multiple machine learning methods to build an ensemble model, aiming to improve the model's generalization ability. The estimation performance of the ensemble model was compared with that of partial least squares regression (PLSR), gradient boosting trees (GBT), and 1-dimensional convolutional neural network (1D-CNN) models. The results demonstrated that the optimal spectral characteristics enhancement strategy can significantly improve the correlation between soil spectra and soil organic carbon content, and the optimal Pearson correlation coefficient reached -0.82.

收稿日期:2023-11-23　修回日期:2024-01-23

基金项目:云南省科技计划项目(202202AE090013)、黑龙江省"揭榜挂帅"科技攻关项目(2021ZXJ05A0502)和重庆市技术创新与应用发展专项(cstc2021jscx-gksbX0064)

作者简介:唐澳华(1999—),男,硕士生,主要从事农业遥感研究,E-mail:2022126042@chd.edu.cn

通信作者:杨贵军(1975—),男,研究员,主要从事农业定量遥感机理模型及应用研究,E-mail:guijun.yang@163.com

Compared with PLSR, GBT, and 1D－CNN models, the ensemble model exhibited higher estimation accuracy and robustness across various spectral datasets. In particular, under the spectral characteristic enhancement strategy of first derivative combined with multivariate scatter correction, the ensemble model demonstrated excellent estimation performance (R^2 = 0.92, RMSE was 1.18 g/kg, RPD was 3.52). The proposed method enabled timely and accurate estimation of SOC, which can provide a scientific basis for modern agricultural management.

Key words: soil organic carbon; estimation model; ensemble learning; spectral characteristics enhancement; mid-infrared spectroscopy

0　引言

土壤有机碳(Soil organic carbon,SOC)是评价土壤质量的关键指标,在农业生产管理中扮演不可或缺的角色,同时也是全球碳循环过程中的重要组成部分[1-2]。然而,传统的土壤有机碳测量方法,如化学分析和燃烧法等,存在耗时、成本高以及不适用于大面积土壤样品测定等缺点,无法满足现代农业的生产需求。因此,寻求一种快速、准确且低成本的SOC 含量测定方法已成为土壤科学的紧迫任务。

近年来,土壤光谱学在 SOC 含量预测方面取得了显著的进展,被认为是传统实验室分析的有效替代方案,该方法具有样品制备简单、测量周期短、成本较低等诸多优势[3-4]。可见-近红外(Visible-to-near-infrared,V－NIR;波长 400 ~2 500 nm)和中红外(Mid-infrared,MIR;波数为 400 ~4 000 cm^{-1}或波长为 2 500 ~25 000 nm)反射光谱依赖于不同官能团与电磁辐射之间的相互作用,已被证实能够用于定量估算 SOC 含量。MIR 范围内土壤中含碳基团和矿物质展现出强烈的基本分子振动,而 V－NIR 范围则显现一些低能电子跃迁和含氢基团的倍频和合频峰。这种分子振动的特征使得 MIR 范围内的能量吸收相较于 V－NIR 中的分子振动更为强烈和明显[5]。此外,国内外研究表明,相较于 V－NIR 光谱等数据,中红外光谱数据可以更精确地估算 SOC 含量[6-8]。例如,HONG 等[9]对 V－NIR 和 MIR 融合进行研究,结果表明 MIR 光谱估算精度优于 V－NIR 光谱;NG 等[7]使用卷积神经网络(Convolutional neural network,CNN)研究了 V－NIR 和 MIR 光谱对土壤养分的估算性能,结果显示 MIR 光谱的估算精度优于 V－NIR。综上所述,合理利用 MIR 光谱中土壤碳基团和矿物质基本振动的特征信息,在一定程度上可以较好地估算 SOC 含量。

然而,原始光谱数据受到土壤粒径、测量条件以及测量仪器等因素的影响而存在噪声[10]。因此,采用适当的光谱特征增强方法消除光谱噪声至关重要。目前,常用的光谱特征增强方法包括一阶导变换(First derivative,FD)、倒数之对数变换(lg(1/R),Log)、标准正态变化(Standard normal variation,SNV)和多元散射校正(Multivariate scatter correction,MSC)等。这些方法均能在一定程度上提高原始光谱数据的质量,例如 FD 能突出原始光谱的特征,有效消除测量背景等因素对光谱的影响;Log 变换能提高光谱波段的差异性,降低光照条件和地形差异对光谱的干扰;MSC 和 SNV 能消除土壤粒径等因素引起的散射影响[11-13]。此外,不同建模方法在估算 SOC 含量时精度存在差异。目前,主要建模方法可分为线性模型和非线性模型。偏最小二乘回归(Partial least squares regression,PLSR)作为一种经典的线性回归模型,在光谱估算 SOC 含量方面得到了广泛应用[14]。然而,由于土壤成分的复杂性,土壤光谱数据与 SOC 含量之间并非简单的线性关系,因此在估算过程中,线性模型存在一定局限性。近年来,非线性模型,如支持向量机回归(Support vector machine regression,SVR)、随机森林(Random forest, RF)、Cubist、高斯过程回归(Gaussian process regression,GPR)等,在估算 SOC 含量方面应用越来越广泛。大量研究表明,相较于线性模型,非线性模型在一定程度上具有更高的估算精度[15-17]。张智韬等[18]利用分数阶微分结合支持向量机和 RF 来监测土壤有机质含量,结果显示最佳模型组合有效提高了模型精度。因此,通过将有效的光谱特征增强策略与非线性估算模型结合,能有效消除测量条件等因素引起的光谱噪声,并构建 SOC 含量估算模型,为准确估算 SOC 含量提供可行且有效的手段。

鉴于土壤成分的多样性和复杂性,导致不同土壤样本之间存在显著的光谱特征差异,这些差异会对光谱特征增强方法和模型构建造成影响。本文着重构建一个快速有效的估算模型,以提高对 SOC 含量的估算精度。研究表明,利用集成学习算法将多种机器学习进行结合可以有效提高模型精度和泛化能力[19]。基于此背景,本研究选用国际土壤参考和信息中心(International Soil Reference and Information Center,ISRIC)提供的中红外土壤光谱数据集作为研究对象,采用不同的光谱特征增强策略,

并利用堆叠泛化(Stacking generalisation,SG)集成学习框架结合 RF、SVR、极限梯度提升树(eXtreme gradient boosting,XGBoost)、多层感知机(Multilayer perceptron,MLP)模型,构建 SG 集成模型。最后,通过与 PLSR、梯度提升树(Gradient boosting trees,GBT)以及一维卷积神经网络(1-dimensional convolutional neural network,1D-CNN)模型进行比较,以评估集成模型估算 SOC 含量的性能,为 SOC 含量的精准无损估算以及指导农业生产提供技术支撑和参考。

1 材料与方法

1.1 数据来源

选用国际土壤参考和信息中心提供的公开土壤光谱数据库作为研究对象,该数据库由世界农林业中心的土壤和植物光谱诊断实验室测量,涵盖来自亚洲、欧洲、北美洲等全球 50 多个国家的 4 000 多个土壤样本[20]。为确保数据准确性和一致性,该数据集中的土壤样品经过严格的标准化处理和测量,土壤样品经过风干、碾磨和过筛(<2 mm)处理后,进行化学分析和光谱测量[21]。SOC 含量采用重铬酸钾氧化法进行测定,土壤中红外光谱数据使用傅里叶变换中红外光谱仪(Tensor 27 型)在实验室条件下测量获得。中红外光谱范围为 602 ~ 7 496 cm^{-1},分辨率为 4 cm^{-1},采样间隔为 2 cm^{-1}。

1.2 数据处理与样本划分

本研究剔除 ISRIC 数据库中 SOC 含量数据缺失的样本,最终选取 3 723 个样本作为研究对象,样本分布如图 1 所示。此外,删除 MIR 光谱数据中噪声较大且信息量较低的 4 000 ~ 7 496 cm^{-1} 光谱范围,仅保留信息更为丰富的 602 ~ 4 000 cm^{-1} 光谱范围[15]。

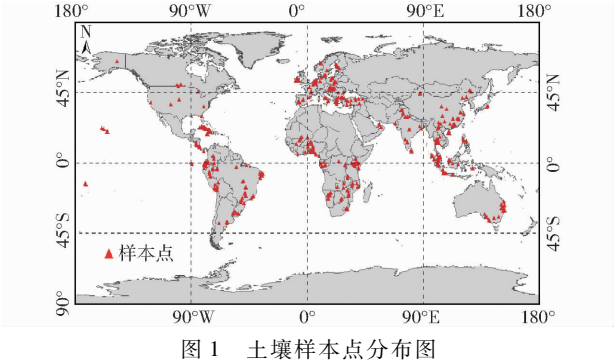

图 1 土壤样本点分布图

Fig. 1 Distribution map of soil sample points

样本集采用分层随机抽样的方法划分训练集和测试集。具体而言,根据总数据集中 SOC 含量的直方图分布进行样本分层,每个直方图层级表示不同范围的 SOC 含量,共分为 20 个层级。在此基础上,采用随机抽样的方法从每个直方图层级中随机选取训练集与测试集样本。其中,训练集包含 2 972 个样本,占总数的 80%;测试集包含 751 个样本,占总数的 20%。分层随机抽样确保了总数据集、训练集和测试集在 SOC 含量分布方面的相似性,从而使测试集数据能够更好地代表全部样本,用于全面评估模型的准确性。

1.3 光谱特征增强方法

光谱特征增强方法在提高光谱数据质量方面发挥重要作用,能有效削弱或消除测量环境和其他因素引起的噪声[12]。本研究采用 FD、SNV、MSC 及其组合对原始光谱进行处理。其中,FD 通过计算光谱数据中的微分数值,能有效减弱或消除基线漂移和噪声干扰,从而突出光谱曲线特征;SNV 能减少或消除光谱数据中的偏移和缩放差异,使经过处理后的光谱数据符合标准正态分布;MSC 则有助于消除测量过程中土壤粒径等因素导致的散射影响,突出样品的光谱特征[22]。此外,将 FD、SNV 和 MSC 进行组合处理,能在一定程度上提高数据信噪比。因此,本研究共采用 6 种光谱特征增强策略,分别为 FD、SNV、MSC、FD + SNV、FD + MSC 和 FD + SNV + MSC。

1.4 建模方法

选用 SVR、RF、XGBoost 和 MLP 为基础模型,使用 XGBoost 作为元模型构建 SG 模型,并与 PLSR、GBT 和 1D-CNN 比较模型精度,以评估 SG 模型估算性能。

1.4.1 线性回归模型

PLSR 是一种经典的线性回归模型,可同时实现回归建模、简化数据结构和分析变量间相关性,在光谱表征土壤组分研究中得到广泛应用[23]。本研究通过网格搜索对 PLSR 中最重要的主成分参数进行优化。

1.4.2 机器学习模型

本研究涉及的机器学习模型有 SVR、RF、GBT、XGBoost 和 MLP。其中,SVR 利用核函数将样本映射到高维空间进行回归分析,以获得最佳估算结果[24];RF[25]、GBT[26] 和 XGBoost[27] 模型基于不同的结合方法将多个弱学习器进行组合,以提高模型预测精度;MLP 是一种多层前馈神经网络,具有出色的非线性拟合和泛化能力[28]。

1.4.3 深度学习模型

1D-CNN 是一种专门用于处理一维数据序列的深度学习模型,通过卷积操作捕捉输入数据中的局部特征和模式[29]。本研究采用包含 3 层卷积网络的 1D-CNN 模型估算 SOC 含量。

1.4.4　集成模型

Stacking 算法是一种常见的集成学习方法，包括高层模型和低层模型。低层模型是基础学习器（选定的各基础模型），高层模型是 SG 模型中的元学习器。SG 模型利用基础学习器的输出作为输入来训练元模型进行估算[30]。本研究采用 5 折交叉验证对基础学习器进行训练和估算，然后利用基础学习器的估算结果训练元模型估算 SOC 含量[31]。SG 模型中的元学习器有助于过滤基础学习器中的估算偏差，提高集成学习估算精度和泛化能力。

上述模型参数优化过程均采用网格搜索结合 5 折交叉验证的方法，在 Python 3.9 版本中完成。SG 集成模型框架如图 2 所示。

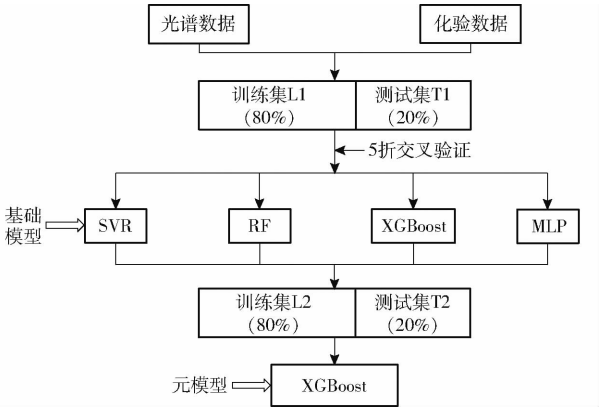

图 2　SG 集成模型框架

Fig. 2　Framework of SG integrated model

1.5　技术路线

本研究技术路线如图 3 所示。

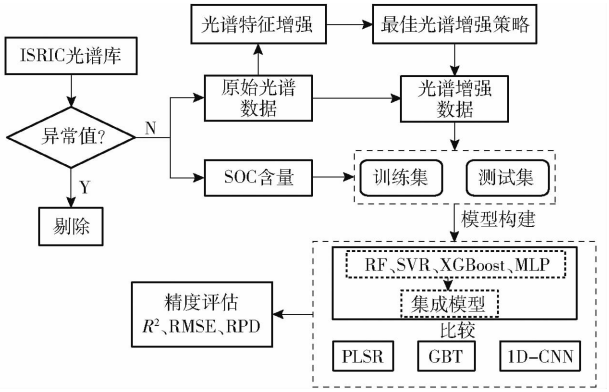

图 3　技术路线图

Fig. 3　Technology roadmap

1.6　精度评价

以测试集决定系数（Determination coefficient，R^2）、均方根误差（Root mean square error，RMSE）和相对分析误差（Relative percent deviation，RPD）作为评估模型精度的指标。其中，R^2 越接近 1，RMSE 越小，RPD 越大则说明模型性能越优。通常，RPD 可分为 3 个等级：RPD 大于 2 表示模型预测性能很好，RPD 为 1.4 ~ 2 表示预测性能一般，RPD 为 0 ~ 1.4 表示预测性能较差[32]。

2　结果与分析

2.1　SOC 含量统计结果

本研究中，SOC 含量范围为 0.01 ~ 60 g/kg，其平均值和标准偏差分别为 1.20 g/kg 和 2.66 g/kg（表 1）。然而，SOC 含量偏度和变异系数分别为 10.57 和 221.67%，说明数据集中 SOC 含量呈现明显右偏。这表明 SOC 含量变异性较高，数据分布不均匀，即不同样本之间的 SOC 含量存在显著差异。训练集和测试集具有相近的平均值，这是分层随机抽样得到的结果。然而，测试集数据标准差和变异系数明显高于训练集，这是因为数据集中高 SOC 含量样本分布不均且数量较少所导致，其中 SOC 含量大于 10 g/kg 的样本数量远少于低 SOC 含量的样本。

表 1　土壤有机碳含量统计

Tab. 1　Statistics of soil organic carbon content

数据集	均值/（g·kg^{-1}）	标准差/（g·kg^{-1}）	最小值/（g·kg^{-1}）	最大值/（g·kg^{-1}）	偏度	变异系数/%
总数据集	1.20	2.66	0.01	60.0	10.57	221.67
训练集	1.13	2.12	0.01	45.8	8.05	187.61
测试集	1.45	4.13	0.01	60.0	9.47	285.82

2.2　光谱特征波数相关性分析

图 4 为采取不同光谱特征增强策略处理后，光谱数据与 SOC 含量的相关性。从图 4 可以看出，光谱特征增强方法可以有效提高土壤光谱数据与 SOC 含量之间的相关性，但各种策略之间存在明显差异。具体而言，使用 FD + MSC、FD + SNV + MSC、FD + SNV 等光谱增强策略处理后，光谱数据与 SOC 含量之间的相关性显著（$P < 0.01$），最大相关系数分别为 −0.82、−0.82 和 0.71。然而，使用 SNV、MSC、SNV + MSC 等策略处理后，光谱数据与 SOC 含量之间的相关性一般，最大相关系数绝对值分别为 0.55、0.49 和 0.56。此外，以 Pearson 相关性系数绝对值大于 0.4（$P < 0.01$）作为依据筛选特征，各种特征增强策略后的光谱数据中满足该条件的波数数量由多到少依次为 FD + MSC（287）、FD + SNV + MSC（268）、SNV + MSC（185）、SNV（162）、MSC（157）、FD + SNV（134）、FD（63）、原始光谱（0）。

综上所述，采用 FD + MSC 和 FD + SNV + MSC 光谱特征增强策略可以显著提高中红外光谱数据与 SOC 含量之间的相关性。因此，选择这两种光谱增强策略进行进一步分析。

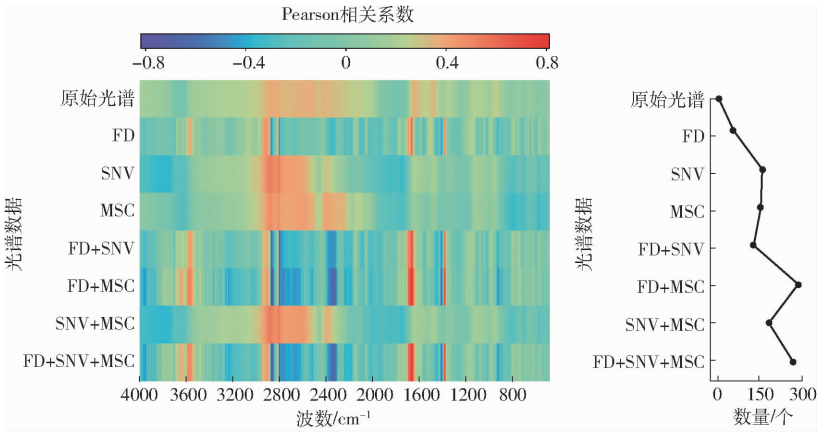

图 4 有机碳含量与光谱特征相关系数

Fig. 4 Correlation coefficient of organic carbon content and spectral feature

2.3 基础模型估算结果

采用原始光谱以及经过 FD + MSC 和 FD + SNV + MSC 特征增强处理后的光谱数据构建 SOC 含量估算模型，图 5 为各基础模型在测试集上的估算精度。在原始光谱中，SVR 与 MLP 模型估算精度较低，而 RF 和 XGboost 模型表现相近且较优。尤其是 XGboost 模型估算精度最佳，R^2、RMSE、RPD 分别为 0.74、2.09 g/kg 和 1.97。经过 FD + MSC 特征增强处理后，SVR 模型表现较差，而 RF、XGboost 和 MLP 模型表现出较好的估算性能，R^2 均高于 0.80。其中，MLP 模型精度最高，R^2 为 0.90，RMSE 和 RPD 分别为 1.26 g/kg 和 3.27。在经过 FD + SNV + MSC 特征增强后，4 个基础模型都展现出较高的估算精度，其 R^2 均高于 0.80。尤其是 MLP 模型表现最佳，其 R^2 达到 0.91，RMSE 为 1.26 g/kg。

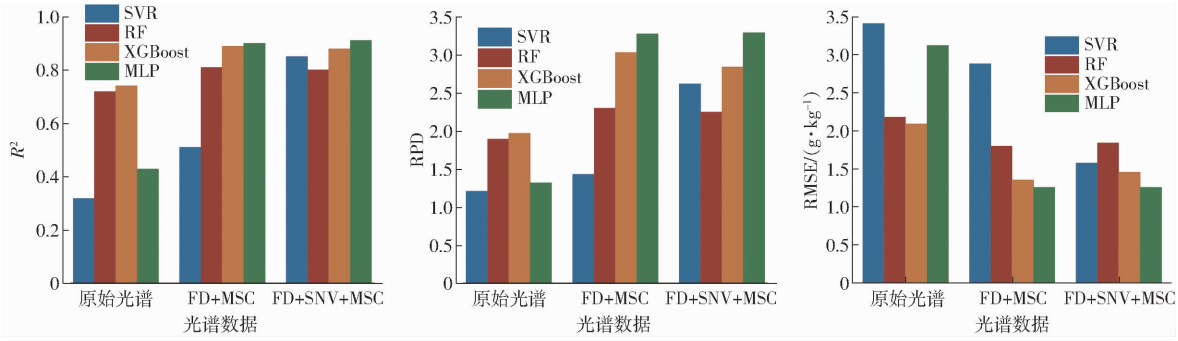

图 5 基础模型估算精度对比

Fig. 5 Comparison of base model estimation accuracy

此外，XGBoost 模型在不同光谱数据中都表现出稳定且优异的 SOC 含量估算效果。总体而言，XGboost 模型在 3 种光谱数据中的 R^2、RMSE 和 RPD 分别平均为 0.84、1.64 g/kg 和 2.62。

2.4 集成模型估算精度

将各基础模型输出的 SOC 含量估算值作为输入，以 XGboost 模型为元模型构建 SG 模型，各模型估算精度见表 2。相较于各基础模型，SG 模型在 3 种不同光谱数据下表现出最佳估算精度，RPD 均大于 2，说明模型具有较强鲁棒性。特别是在经过 FD + MSC 特征增强处理后，SG 模型精度达到最佳，R^2 为 0.92，RMSE 和 RPD 分别为 1.18 g/kg 和 3.52。并且由图 6 可见，SG 模型误差主要分布在 −0.5 ~ 0.5 g/kg 范围内，这表明 SG 模型估算性能较优。然而，在 FD + MSC 以及 FD + SNV + MSC 特征增强处理后，相较于 MLP 模型，SG 模型估算性能提升较小。此外，当 SOC 含量接近或大于 30 g/kg 时，SG 模型存在饱和效应，估算误差较大。

通过对比 PLSR、GBT 和 1D − CNN 模型可知，在原始光谱中，PLSR 模型估算精度较低，而 GBT 和 1D − CNN 模型估算性能良好；经过 FD + MSC 以及 FD + SNV + MSC 光谱增强后，3 种模型都呈现良好的估算性能，其 R^2 均大于 0.85、RPD 大于 2.5，RMSE 小于 1.6 g/kg。总体而言，SG 模型在原始光谱和处理后的光谱数据中表现均优于 PLSR、GBT 和 1D − CNN 模型，展示出较高的鲁棒性。4 种模型总体性能从高到低依次为 SG 模型、1D − CNN、GBT、PLSR。因此，这也证实集成学习作为一种强大的机器学习方法，相较于单一机器学习模型，能有效提高模型预测性能。

表 2　SG 模型与其他模型估算精度对比

Tab. 2　Comparison of estimation accuracy between SG model and other models

光谱数据	SG 模型			PLSR			GBT			1D－CNN		
	R^2	RMSE/(g·kg^{-1})	RPD	R^2	RMSE/(g·kg^{-1})	RPD	R^2	RMSE/(g·kg^{-1})	RPD	R^2	RMSE/(g·kg^{-1})	RPD
原始光谱	0.82	1.76	2.35	0.66	2.41	1.71	0.74	2.09	1.97	0.78	1.93	2.14
FD + MSC	0.92	1.18	3.52	0.86	1.56	2.65	0.86	1.56	2.64	0.87	1.49	2.76
FD + SNV + MSC	0.91	1.23	3.37	0.86	1.56	2.65	0.86	1.52	2.70	0.88	1.43	2.88

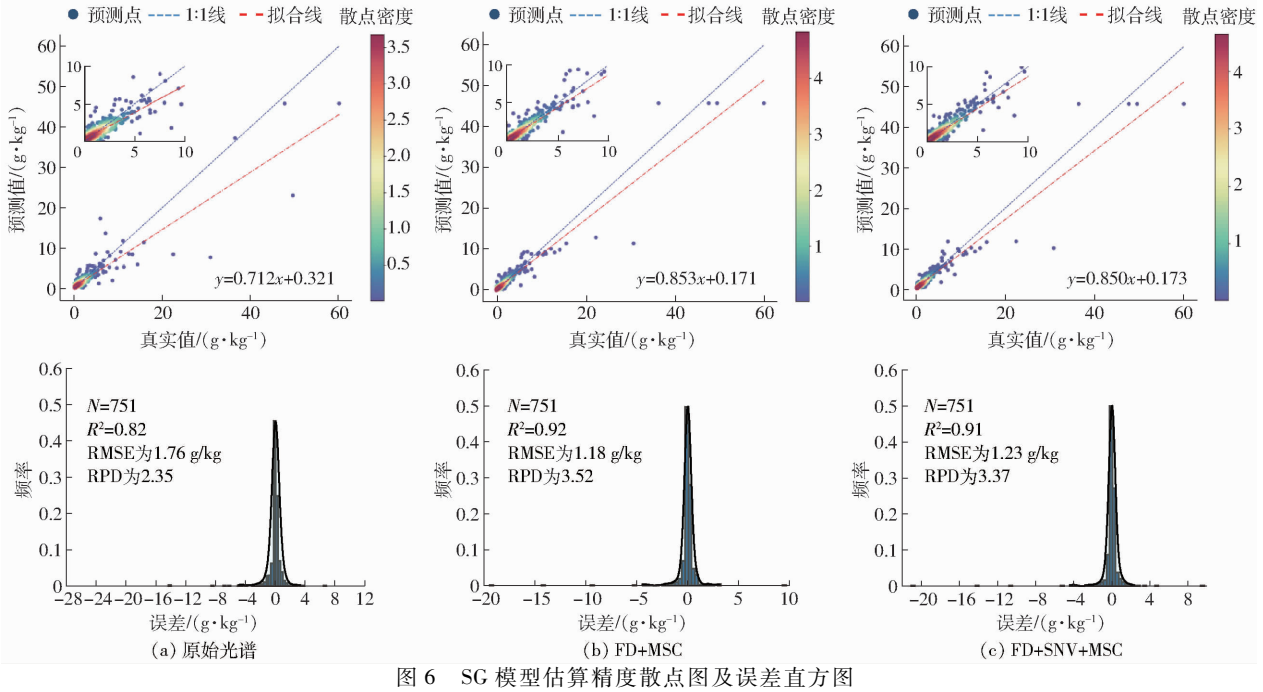

图 6　SG 模型估算精度散点图及误差直方图

Fig. 6　Scatter plots and error histograms of SG model estimation accuracy

3　讨论

3.1　土壤有机碳含量分布对建模精度的影响

本研究以 ISRIC 提供的全球公开土壤光谱数据库为研究对象，该数据库包含超过 4 000 个土壤光谱数据以及相应的土壤属性信息，为土壤组分研究提供了重要的数据来源。然而，统计结果（图 7）显示，该数据库中的 SOC 含量呈现显著的不均衡分布。具体而言，SOC 含量小于 0.5 g/kg 的样本数约占总数据集的 50%，而 SOC 含量大于 10 g/kg 的土壤样本仅占不到 1%。这样的分布导致 SOC 含量的描述性统计呈现出明显的偏斜和异常的变异系数（表 1）。与此同时，训练集和测试集中高 SOC 含量的样本相对较少，这使得各个模型在学习过程中更加偏向于低 SOC 含量部分，而对高 SOC 含量样本的学习效果较差。这也导致本研究中各模型在估算高 SOC 含量样本时产生较大的偏差，从而导致 SG 模型在高 SOC 含量部分出现饱和现象（图 6）。因此，未来的研究中应适当增加高 SOC 含量的样本数据，使得数据结构更加合理，从而提高模型估算性能。

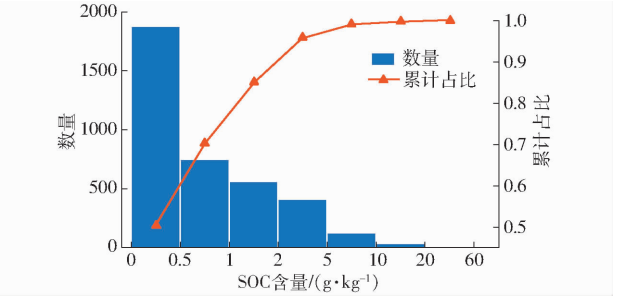

图 7　ISRIC 土壤光谱数据结构

Fig. 7　Structure of ISRIC soil spectral data

3.2　光谱特征增强策略对估算结果的影响

光谱采集过程中测量条件会给土壤光谱引入噪声，从而对最终模型估算结果造成误差。目前，许多研究致力于削弱或消除测量条件产生的光谱噪声，如一阶导数变换、小波变换等。本研究中，采用目前应用最广泛的 3 种光谱增强方法（FD、SNV 和 MSC）进行研究，这些方法在已有文献中均展示出优异性能[12,17,33]。因此，本研究探究 FD、SNV、MSC 及其组合对模型性能的影响，结果显示 FD + MSC 以及 FD + SNV + MSC 的组合策略效果较佳。这是由于原始光谱经过 FD 处理后能有效消除测量背景引起

的误差,从而增加了土壤成分信息在光谱数据中的信息占比;随后采取 SNV 和 MSC 可以进一步消除土壤粒径等造成的散射影响,提高光谱数据的信噪比,进而提升了估算模型精度。

此外,本研究通过 Pearson 相关性分析初步探讨 FD、SNV、MSC 3 种方法不同组合顺序对光谱数据的影响。如图 8(FSM 表示特征增强组合顺序为 FD + SNV + MSC,其他类推)所示,不同组合顺序处理后,光谱数据与 SOC 含量的 Pearson 相关性存在较大的差异。SFM 和 FSM 组合策略处理后土壤光谱数据与 SOC 含量呈现较高的相关性,而 SMF 和 MSF 组合策略效果较差。这可能是因为不同组合策略在突显光谱特征方面存在差异。对于 SFM 和 FSM 组合,可能更有效地强调了与 SOC 含量相关的特定频谱范围。而 SMF 和 MSF 组合可能突出了与 SOC 含量不相关的特征,导致引入了更多噪声或削弱了相应的特征信息。此外,在不同顺序组合中,后续处理方法可能将前期处理中的特征波段视为噪声而削弱,从而影响了光谱数据与 SOC 含量之间的相关性。因此,未来的研究中有必要深入探讨不同光谱增强方法组合对光谱数据的影响,以期为土壤光谱学找到一种普适性和高效性的光谱增强方法。

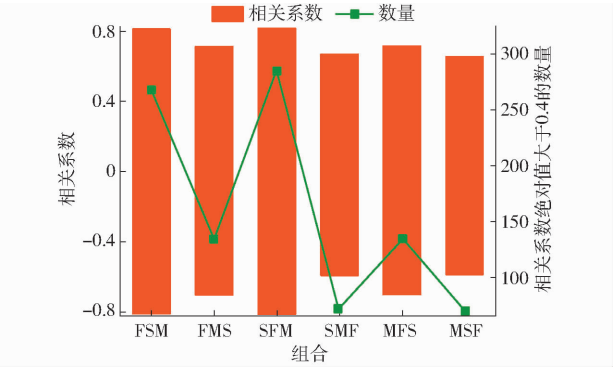

图 8 不同光谱特征增强组合顺序与有机碳含量的相关性分析

Fig. 8 Correlation analysis of combination order of different spectral feature enhancements with organic carbon content

3.3 集成模型估算 SOC 含量的优势、不足与展望

随着计算机技术的发展,学者们已经利用多种估算模型结合光谱数据来估算 SOC 含量[34]。然而,迄今为止尚未发现普适性较强的估算模型,这表明传统的单一模型可能存在一定局限性。集成学习方法作为一种新兴的机器学习方法,通过融合多个预测模型,综合各模型优势,以提高模型的预测精度和鲁棒性[35]。本研究基于 Stacking 算法结合多种机器学习模型估算 SOC 含量。结果显示,相比于单一估算模型,SG 模型在整体上表现出更高的估算精度和优越的鲁棒性,在不同光谱数据下都呈现较优的估算结果,这与先前研究结果相一致。例如,张秀全等[36]采用 V - NIR 光谱结合堆叠泛化集成模型获得了优于其他算法的模型精度;NAWAR 等[37]利用 MIR 光谱结合该集成框架也取得了较优的模型性能。

然而,在特定的光谱条件下,集成学习的准确性相对于最佳基础模型并未表现出显著的提升。在本研究中,这一现象可能是由于数据集分布不均匀,导致各基础模型在高 SOC 含量样本方面存在较大的偏差。这进一步影响了 SG 模型在高 SOC 含量部分的估算效果,在 SOC 含量大于 30 g/kg 部分出现了饱和现象。此外,目前集成学习研究中通常选择特定的线性模型或非线性模型进行融合,但并未明确说明选择这些模型的依据。

因此,集成学习在使用土壤光谱估算 SOC 含量方面的优势仍需进一步探索和挖掘。未来的研究方向可以着重从以下方面进行:①研究数据分布不均匀的影响。进一步深入研究不均匀数据分布对集成学习准确性的影响,以找到针对这一问题的解决方案。②探索其他集成学习方法。采用不同的集成学习方法,并从各个机器学习模型的本质出发,深入挖掘不同模型之间的多样性。这可以包括集成更多的基础模型,使用不同的融合策略或者尝试集成学习和深度学习方法的结合,以提高估算模型的适应性和鲁棒性。③多源数据融合。探索基于集成学习的 V - NIR、MIR、激光诱导击穿光谱等多源数据融合的 SOC 含量探测方法,以提高估算的全面性和可靠性。通过综合多种光谱数据源,可以更全面地捕捉土壤光谱信息,提高估算的全面性和可靠性。通过这些研究,有望更全面、准确地理解集成学习在土壤光谱估算 SOC 含量中的潜力,为未来的农田管理和土壤质量监测提供更可靠的技术支持。这将有助于推动土壤光谱学的发展,并提高其在实际应用中的准确性和实用性。

4 结论

(1)原始光谱数据与 SOC 含量的相关性较差,但采用光谱特征增强策略能有效提高光谱数据的特征,特别是经过 FD + MSC 和 FD + SNV + MSC 组合策略处理后,Pearson 相关系数绝对值达到 0.82。

(2)在 4 个基础模型中,XGboost 模型表现出较好的稳定性和估算精度;而在相应特征增强后的光谱数据下,MLP 模型表现出最佳估算精度,其 R^2 和 RMSE 分别达到 0.91 和 1.26 g/kg。

(3)相较于其他单一模型,SG 集成模型在各光谱数据中表现出更优异的鲁棒性和准确的估算结果。在 FD + MSC 和 FD + SNV + MSC 光谱特征增强策略条件下,集成学习模型 R^2 均不小于 0.91,RPD 不小于 3.37,能够准确地估算 SOC 含量。

参 考 文 献

[1] 刘焕军, 赵春江, 王纪华, 等. 黑土典型区土壤有机质遥感反演[J]. 农业工程学报, 2011, 27(8): 211 - 215.
LIU Huanjun, ZHAO Chunjiang, WANG Jihua, et al. Soil organic matter predicting with remote sensing image in typical blacksoil area of Northeast China[J]. Transactions of the CSAE, 2011, 27(8): 211 - 215. (in Chinese)

[2] MINASNY B, MALONE B P, MCBRATNEY A B, et al. Soil carbon 4 per mille[J]. Geoderma, 2017, 292: 59 - 86.

[3] WIJEWARDANE N K, GE Y, WILLS S, et al. Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library[J]. Soil Science Society of America Journal, 2018, 82(3): 722 - 731.

[4] 尼加提·卡斯木, 茹克亚·萨吾提, 师庆东, 等. 基于优化光谱指数的土壤有机质含量估算[J]. 农业机械学报, 2018, 49(11): 155 - 163.
NIJAT Kasim, RUKEYA Sawut, SHI Qingdong, et al. Estimation of soil organic matter content based on optimized spectral index[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(11): 155 - 163. (in Chinese)

[5] 吴裕, 申广荣, 刘璐, 等. 土壤有机质高光谱特征及其反演研究[J]. 上海交通大学学报(农业科学版), 2019, 37(4): 37 - 44.
WU Yu, SHEN Guangrong, LIU Lu, et al. Hypespectral characteristics of soil organic matter and inversion methods[J]. Journal of Shanghai Jiaotong University(Agricutural Science), 2019, 37(4): 37 - 44. (in Chinese)

[6] HAGHI R K, PÉREZ-FERNÁNDEZ E, ROBERTSON A H J. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: a comparison of near infrared and mid-infrared spectroscopy [J]. Geoderma, 2021, 396: 115071.

[7] NG W, MINASNY B, MONTAZEROLGHAEM M, et al. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra[J]. Geoderma, 2019, 352: 251 - 267.

[8] GOYDARAGH M G, TAGHIZADEH-MEHRJARDI R, JAFARZADEH A A, et al. Using environmental variables and Fourier transform infrared spectroscopy to predict soil organic carbon[J]. CATENA, 2021, 202: 105280.

[9] HONG Y, MUNNAF M A, GUERRERO A, et al. Fusion of visible-to-near-infrared and mid-infrared spectroscopy to estimate soil organic carbon[J]. Soil and Tillage Research, 2022, 217: 105284.

[10] 勾宇轩, 赵云泽, 李勇, 等. 基于 CWT - sCARS 的东北旱作农田土壤有机质高光谱反演[J]. 农业机械学报, 2022, 53(3): 331 - 337.
GOU Yuxuan, ZHAO Yunze, LI Yong, et al. Soil organic matter content in dryland farmland in Northeast China with hyperspectral reflectance based on CWT - sCARS[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(3): 331 - 337. (in Chinese)

[11] SANTOS U J D, DEMATTÊ J A D M, MENEZES R S C, et al. Predicting carbon and nitrogen by visible near-infrared (Vis - NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil[J]. Geoderma Regional, 2020, 23: e00333.

[12] 第五鹏瑶, 卞希慧, 王姿方, 等. 光谱预处理方法选择研究[J]. 光谱学与光谱分析, 2019, 39(9): 2800 - 2806.
DIWU Pengyao, BIAN Xihui, WANG Zifang, et al. Study on the selection of spectral preprocessing methods[J]. Spectroscopy and Spectral Analysis, 2019, 39(9): 2800 - 2806. (in Chinese)

[13] 张俊华, 尚天浩, 陈睿华, 等. 基于光谱 FOD 与优化指数的银川平原土壤有机质含量反演[J]. 农业机械学报, 2022, 53(11): 379 - 387.
ZHANG Junhua, SHANG Tianhao, CHEN Ruihua, et al. Inversion of soil organic matter content in Yinchuan Plain using field spectral fractional-order derivatives combined with spectral optimization index[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(11): 379 - 387. (in Chinese)

[14] 于雷, 洪永胜, 耿雷, 等. 基于偏最小二乘回归的土壤有机质含量高光谱估算[J]. 农业工程学报, 2015, 31(14): 103 - 109.
YU Lei, HONG Yongsheng, GENG Lei, et al. Hyperspectral estimation of soil organic matter content based on partial least squares regression[J]. Transactions of the CSAE, 2015, 31(14): 103 - 109. (in Chinese)

[15] 孟鑫鑫, 于雷, 周勇, 等. 基于可见近红外和中红外近地面光谱数据融合的土壤有机碳含量反演[J]. 土壤通报, 2022, 53(2): 301 - 307.
MENG Xinxin, YU Lei, ZHOU Yong, et al. Predicing organic carbon using data fusion of visible near-infrared and middle infrared spectra by proximal soil sensing[J]. Chinese Journal of Soil Science, 2022, 53(2): 301 - 307. (in Chinese)

[16] DEISS L, MARGENOT A J, CULMAN S W, et al. Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy[J]. Geoderma, 2020, 365: 114227.

[17] 叶勤, 姜雪芹, 李西灿, 等. 基于高光谱数据的土壤有机质含量反演模型比较[J]. 农业机械学报, 2017, 48(3): 164 - 172.

YE Qin, JIANG Xueqin, LI Xican, et al. Comparison on inversion model of soil organic matter content based on hyperspectral data[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(3): 164 - 172. (in Chinese)

[18] 张智韬, 劳聪聪, 王海峰, 等. 基于 FOD 和 SVMDA - RF 的土壤有机质含量高光谱预测[J]. 农业机械学报, 2020, 51(1): 156 - 167.
ZHANG Zhitao, LAO Congcong, WANG Haifeng, et al. Estimation of desert soil organic matter through hyperspectra based on fractional-order derivatives and SVMDA - RF[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(1): 156 - 167. (in Chinese)

[19] LIANG M, CHANG T, AN B, et al. A stacking ensemble learning framework for genomic prediction[J]. Frontiers in Genetics, 2021, 12: 600040.

[20] SAFAIE M, HOSSEINPOUR-ZARNAQ M, OMID M, et al. Using deep neural networks for evaluation of soil quality based on Vis - NIR spectroscopy[J]. Earth Science Informatics, 2024, 17(1): 271 - 281.

[21] TERHOEVEN-URSELMANS T, VAGEN T G, SPAARGAREN O, et al. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library[J]. Soil Science Society of America Journal, 2010, 74(5): 1792 - 1799.

[22] 谭洋. 高光谱土壤有机质含量反演精度影响因素分析及模型优化[D]. 长春: 吉林大学, 2022.
TAN Yang. Analysis on influence facters and model optimization of soil organic matter prediction accuracy based on Vis - NIR spectra[D]. Changchun: Jilin University, 2022. (in Chinese)

[23] GELADI P, KOWALSKI B R. Partial least-squares regression: a tutorial[J]. Analytica Chimica Acta, 1986, 185: 1 - 17.

[24] DRUCKER H, BURGES C J C, KAUFMAN L, et al. Support vector regression machines[C]// Advances in Neural Information Processing Systems, 1996.

[25] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5 - 32.

[26] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189 - 1232.

[27] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]// Knowledge Discovery and Data Mining. ACM, 2016.

[28] 张驰, 郭媛, 黎明. 人工神经网络模型发展及应用综述[J]. 计算机工程与应用, 2021, 57(11): 57 - 69.
ZHANG Chi, GUO Yuan, LI Ming. Review of development and application of artificial neural network models[J]. Computer Engineering and Applications, 2021, 57(11): 57 - 69. (in Chinese)

[29] KIRANYAZ S, AVCI O, ABDELJABER O, et al. 1D convolutional neural networks and applications: a survey[J]. Mechanical Systems and Signal Processing, 2021, 151: 107398.

[30] WANG S Q, YANG J, CHOU K C. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition[J]. Journal of Theoretical Biology, 2006, 242(4): 941 - 946.

[31] 费帅鹏, 禹小龙, 兰铭, 等. 基于高光谱遥感和集成学习方法的冬小麦产量估测研究[J]. 中国农业科学, 2021, 54(16): 3417 - 3427.
FEI Shuaipeng, YU Xiaolong, LAN Ming, et al. Research on winter wheat yield estimation based on hyperspectral remote sensing and ensemble learning method[J]. Scientia Agricultura Sinica, 2021, 54(16): 3417 - 3427. (in Chinese)

[32] VISCARRA ROSSEL R A, MCGLYNN R N, MCBRATNEY A B. Determining the composition of mineral-organic mixes using UV - Vis - NIR diffuse reflectance spectroscopy[J]. Geoderma, 2006, 137(1): 70 - 82.

[33] 谭洋, 姜琦刚, 刘骅欣, 等. 基于多尺度 SNV - CWT 特征的黑土有机质、水分、总铁及 pH 值估测[J]. 光谱学与光谱分析, 2021, 41(11): 3424 - 3430.
TAN Yang, JIANG Qigang, LIU Huaxin, et al. Estimation of organic matter, moisture, total iron and pH from back soil based on multi scales SNV - CWT transformation[J]. Spectroscopy and Spectral Analysis, 2021, 41(11): 3424 - 3430. (in Chinese)

[34] VISCARRA R R A, HICKS W S. Soil organic carbon and its fractions estimated by visible-near infrared transfer functions: Vis - NIR estimates of organic carbon and its fractions[J]. European Journal of Soil Science, 2015, 66(3): 438 - 450.

[35] 袁培森, 杨承林, 宋玉红, 等. 基于 Stacking 集成学习的水稻表型组学实体分类研究[J]. 农业机械学报, 2019, 50(11): 144 - 152.
YUAN Peisen, YANG Chenglin, SONG Yuhong, et al. Classification of rice phenomics entities based on Stacking ensemble learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11): 144 - 152. (in Chinese)

[36] 张秀全, 李志伟, 郑德聪, 等. 基于近红外-可见光高光谱的堆叠泛化模型褐土有机质预测[J]. 光谱学与光谱分析, 2023, 43(3): 903 - 910.
ZHANG Xiuquan, LI Zhiwei, ZHENG Decong, et al. Vis - NIR hyperspectral prediction of soil organic matter based on stacking generalization model[J]. Spectroscopy and Spectral Analysis, 2023, 43(3): 903 - 910. (in Chinese)

[37] NAWAR S, MOUAZEN A M. Combining mid infrared spectroscopy with stacked generalisation machine learning for prediction of key soil properties[J]. European Journal of Soil Science, 2022, 73(6): e13323.