基于梯度提升树的土壤速效磷高光谱回归预测方法

金秀，朱先志，李绍稳*，王文才，齐海军
安徽农业大学信息与计算机学院，安徽 合肥 230036

摘要 在前期研究基础上，利用皖北地区砂姜黑土的 193 个土壤样本的可见近红外高光谱(350～1700 nm)数据，结合非线性和线性的核函数，对 9 种算法进行模型的首次优化；再利用随机森林、提升树和梯度提升树三种集成学习算法进行模型组合和二次优化。通过模型比较，优选并组合了 Sigmoid 函数的偏最小二乘、线性的支持向量回归、径向基的支持向量回归和 Sigmoid 函数的支持向量回归 4 个单模型，集成算法优化后发现，梯度提升树算法的预测结果最优。与单模型的预测结果相比，梯度提升树模型组合的决定系数为 0.86，提高了 17.8%，相对分析误差系数为 2.55，从 B 等级提升到 A，不仅在准确率上有显著提高，且组合模型过拟合更低，泛化性好。因此，梯度提升树的集成学习可结合多种模型优势，通过高光谱的模型集成来提升土壤速效磷的预测结果精确度。

关键词 成像系统；土壤速效磷；高光谱；回归算法；集成学习
中图分类号 S153.6 **文献标识码** A **doi**：10.3788/LOP56.131102

Predicting Soil Available Phosphorus by Hyperspectral Regression Method Based on Gradient Boosting Decision Tree

Jin Xiu, Zhu Xianzhi, Li Shaowen*, Wang Wencai, Qi Haijun
School of Information & Computer, Anhui Agricultural University, Hefei, Anhui 230036, *China*

Abstract Based on the previous studies, visible near-infrared hyperspectral (350-1700 nm) data of 193 samples from sandy ginger black soil in northern Anhui province are firstly used to optimize the nine models by combing the nonlinear and linear kernel functions. Then, model combination and secondary optimization are performed via three integrated learning algorithms based on the random forest, boosting tree, and gradient boosting decision tree (GBDT). Four single models, including partial least squares of Sigmoid function, linear support vector regression, radial basis support vector regression, and support vector regression of Sigmoid function, are selected and combined by model comparison. After optimization of the integrated algorithms, it is found that the prediction results of the GBDT algorithm are optimal. The determination coefficient of the GBDT algorithm is 0.86, which is 17.8% higher than that of the single model, and the relative analysis error coefficient is 2.55, which is significantly improved from grade B to A. The GBDT algorithm not only improves the accuracy, but also has low overfitting degree and good generalization performance. Therefore, the GBDT algorithm can be combined with the advantages of multiple models and improve the accuracy of the prediction results of soil available phosphorus through hyperspectral model integration.

Key words imaging systems; soil available phosphorus; hyperspectrum; regression algorithms; integrated learning
OCIS codes 110.4234; 50.4600

1 引 言

在土壤速效养分的定量检测方法中，近红外高光谱与传统实验室理化测试方法相比，具有无损、快速等优点。高光谱数据分辨率高、范围宽，其可见近红外区范围内可表达多种成分信息，从而可以建立

收稿日期：2019-01-04；**修回日期**：2019-01-21；**录用日期**：2019-01-25
基金项目：农业部引进国际先进农业科学技术计划(948 计划)项目(2015-Z44，2016-X34)、国家重点研发计划(2018YFF021350601)

* **E-mail**：shwli@ahau.edu.cn

回归模型来预测土壤成分，具有广阔的应用前景[1]。在国内外研究中，土壤速效磷的回归模型建立都取得了较高的性能[2-5]，研究中大部分使用偏最小二乘法(PLS)、支持向量机(SVM)、神经网络等算法来进行回归模型构建。每个建模算法都有各自的特点，其中 PLS 的应用最为广泛，因为其可解决高光谱数据共线性和冗余等问题[6-8]。在速效磷高光谱预测中，2016 年 Sarathjith 等[9]在 350～2500 nm 波段范围，利用离散小波和支持向量回归构建模型来预测速效磷，其速效磷的相对分析误差为 2.27。同年，张佳佳等[10]针对南方丘陵稻田土进行土壤全磷和有效磷的分析，在 350～170 nm 的光谱数据基础上利用多项式回归模型发现其相对分析误差系数分别为 1.43 和 1.54。2018 年，齐海军等[11]在 400～1000 nm 光谱范围内进行了土壤速效磷的回归建模，模型在相对分析误差上提高到 2.29，但是其结果与早期 Sarathjith 等建立的模型相同，且测试集上相对分析误差的提高程度较少，同时还分析了成像光谱和非成像光谱在土壤速效磷的预测差异性[12]，为本研究进一步提供了优化的实验方案。

本文针对土壤速效养分的高光谱(350～1700 nm)数据进行单模型优化和多模型组合结果评价、分析和比较。由于高光谱回归建模算法较多[13-18]，不同模型各具优缺点，因此构建并优化了 9 个单模型，同时利用了堆叠(Stacking)算法进行模型组合，对比了随机森林、提升树和梯度提升树(GBDT)的集成学习算法，最终获得土壤速效磷的最佳预测结果。

2 基本原理

2.1 回归算法

偏最小二乘法回归(PLSR)、支持向量回归(SVR)和岭回归均属于适用性较广的化学计量学建模方法，尤其 PLSR 和 SVR 被广泛用于光谱分析中[19]；SVR 是经典的监督类型回归方法之一，但其缺点为在数据量过大时建模速度较慢[20]；岭回归也是一种专用于共线性数据分析的有偏估计回归方法，通过正则算法放弃最小二乘法的无偏性，以损失信息、降低精度获得回归模型[21]。土壤的速效磷光谱特征可能同时具有线性和非线性关系，因此针对 3 种回归算法都建立了非线性模型，通过线性和非线性模型对比来研究光谱特征规律。本文主要使用 2 个非线性核函数，第 1 个为径向基函数[22]，其表示 1 个取值仅依赖于离原点距离的实值函数，是回归算法中常用的非线性核函数，公式如下

$$K(x,x')=\exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right),\tag{1}$$

式中：$\|x-x'\|_2^2$ 为两个特征向量(x,x')之间的平方欧几里得距离；σ 是自由参数。

第 2 个为 Sigmoid 函数[23]，也被称为 S 型生长曲线，其公式如下

$$S(x)=\frac{1}{1+\exp(-x)},\tag{2}$$

式中：x 为自变量。本文使用网格搜索方法进行了高维参数调优[24]。

2.2 集成算法

集成学习算法是通过构建并组合多个模型来完成学习任务，其中模型集成提升效果主要有两种条件：1)单模型性能提升；2)模型之间的差异性增大[25-27]。本文模型组合中使用 Stacking 结合策略，通过比较模型组合后的差异性，来获得最优模型组合。Stacking 方法相对于其他方法更稳定，过拟合风险更低[28]，其思想如图 1 所示。图中 R_i 为第 i 种回归模型，总共有 n 种不同的模型，N_i 为第 i 种回归模型参数的新特征数据，总共生成 n 个特征。

集成学习的初级算法使用了线性和非线性 PLS、SVR 和岭回归，次级建模算法使用了随机森林算法、提升树和梯度提升树[29]。因此，采用 Stacking 方法对单模型生成新建模集，初级学习算法的输出作为样例输入特征，而初始样本的标记仍作为样例标记，通过次级学习算法进行再次建模预测。次级训练集是由初级学习算法产生的，因此，在 Stacking 算法中为避免过拟合，使用了交叉验证的方式产生次级学习算法的训练样本，如在图 1 中通过 K 折交叉验证方法产生次级算法的新数据集。在次级建模算法中，先使用随机森林算法。随机森林[30-31]是在决策树基础上进行的延伸，能解决决策树泛化性弱的问题，通过内部包含若干个独立的子决策树模型进行汇总得到最终预测结果，其中随机森林中需要优化的最重要的参数有决策树数量和决策树最大深度。

提升树与随机森林都是以回归树为基本学习器的提升方法，但是随机森林与前者不同的是每一步都是独立抽样的，而提升树采用加法模型与向前分布算法，利用前一轮模型中对每个样本预测的偏差作为新的建模集，整体的构建实际上就是不断地拟合前一轮模型预测误差的过程。提升树模型算法公式如下

$$f_M(x)=\sum_{m=1}^{M}T(x;\Theta_m),\tag{3}$$

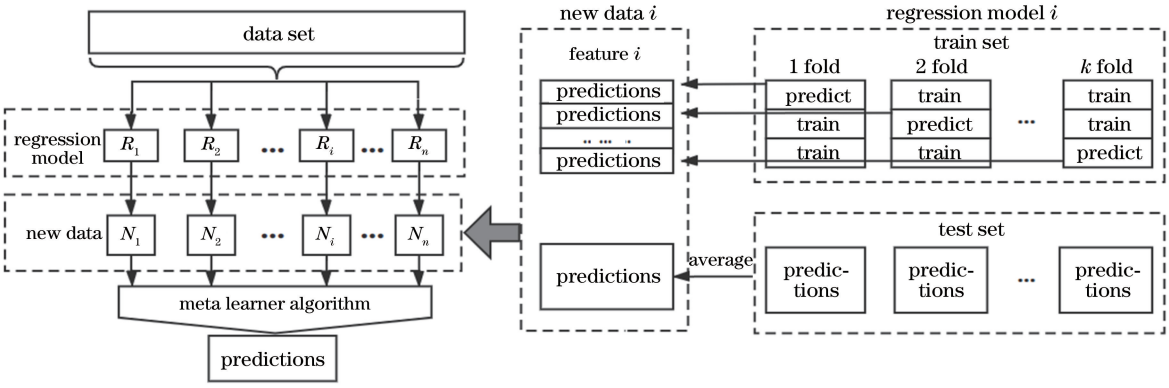

图 1 Stacking 方法

Fig. 1 Stacking method

式中：$T(x;\Theta_m)$表示决策树；Θ_m 为决策树的参数；M 为树的个数，m 为序号。

梯度提升树算法与提升树类似，但与随机森林相比对异常值更加敏感，后者通过减小模型的方差提高性能，而前者是通过减少模型偏差来提高性能。从性能上比较，梯度提升数相对于随机森林具有更好的泛化性能，主要因为梯度提升数不再使用残差作为训练数据，而是利用最速下降法的近似方法，用损失函数求梯度后进行计算，其中损失函数的负梯度如下所示

$$-\left\{\frac{\partial L\left[y_n,f(x_n)\right]}{\partial f(x_n)}\right\}_{f(x)=f_{m-1}(x)}, \tag{4}$$

(4)式的值为第 m 轮的第 n 个样本的损失函数的负梯度，将其作为提升树算法中的残差近似值来拟合回归模型；x_n、y_n 为训练数据集和标签值；L 为损失函数；$f(x)$为模型的预测结果。

2.3 评价方法

评价方法主要由决定系数 R^2、方均根误差 f_{RMSE}和相对分析误差 f_{RPD}来进行比较分析[32]。当 $f_{\mathrm{RPD}}>2.0$ 时模型精确度为 A 类，为良好预测能力；当 $1.4\leqslant f_{\mathrm{RPD}}\leqslant 2.0$ 时模型精确度为 B 类，为中等预测能力；当 $f_{\mathrm{RPD}}<1.4$ 时模型精确度为 C 类，为较差预测能力[33]。R^2 和 f_{RPD}值越大，f_{RMSE}值越小，模型的性能越好。因此，本文重点参考预测结果的 R^2 和 f_{RPD}值来评价模型方法。

3 实 验

野外土壤样品为安徽省皖北地区采集的表层深度为 0～20 cm 的土层，其类型为砂浆黑土，共采集 193 个土壤样本。将所采土壤样本放室内自然风干，经研磨后，人工捡出土壤中的石块、秸秆等干扰物，进行过筛处理，将样品分为两份，一份用于实验室理化检测，一份用于光谱分析。土壤实验室理化检测在安徽农业大学资源与环境学院土壤学实验室支持下完成，土壤速效磷的理化检测使用碳酸氢钠浸提-钼锑抗分光光度法[34]。土壤高光谱数据采集使用的是地物光谱仪(OFS1700，Ocean Optics，美国)，光谱范围为 350～1700 nm，光谱分辨率在 900 nm和 1700 nm 处分别为 2 nm 和 5 nm。在室内条件下，将处理后的土壤放入培养皿中并将土样表面刮平，用地物光谱仪的反射探头(前端密封橡胶圈和内置光源的结合，利于随时随地创造人工暗室条件，有效避免杂散光的影响)直接接触土壤表面进行量测，如图 2 所示。每个样品量测 10 次反射率光谱后取算术平均值，量测过程中每 10 份样本进行 1 次标准白板校正。

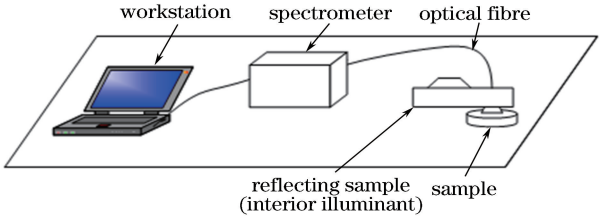

图 2 室内高光谱采集系统

Fig. 2 Indoor hyperspectral acquisition system

土壤样品总数为 193，土壤样本的划分使用 Kennard-Stone 算法[35]，将其分为大约 3∶1的建模集和测试集，其中建模集有 144 个样本，测试集有 49 个样本。土壤样品速效磷含量的统计参数如表 1 所示，其中 Max 为样本中速效磷含量最大值，Min 为样本中速效磷含量最小值，Average 为平均值，Standard deviation 为标准差。从表中可知，建模样本与总样本的数据分布有相似的数据范围和分布情况，但测试样本与总样本具有较小偏差，测试集的预测结果可以充分比较不同算法间的准确性和稳定性，因此本文针对测试集来评价土壤速效磷模型的回归预测性能。

表 1　土壤速效磷含量的统计参数

Table 1　Statistical parameters of soil available phosphorus content

Type	Sample	Max /(mg·kg^{-1})	Min /(mg·kg^{-1})	Average /(mg·kg^{-1})	Standard deviation /(mg·kg^{-1})
Total	193	34.96	0.03	10.56	9.36
Training	144	34.96	0.03	10.94	9.49
Testing	49	32.24	0.60	9.01	8.99

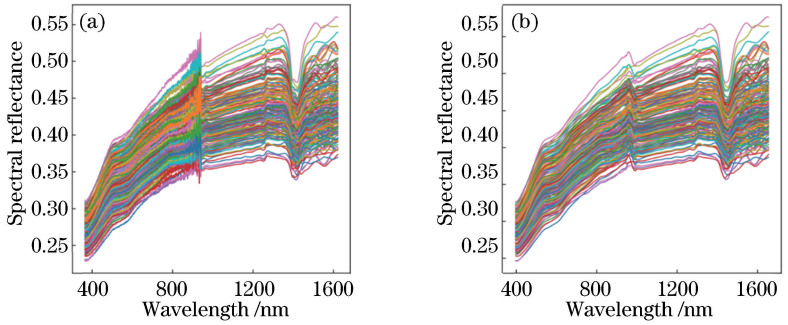

图 3　土壤高光谱反射率。(a)原始光谱;(b)平滑后光谱

Fig. 3　Hyperspectral reflectance of soil. (a) Original spectra; (b) smoothing spectra

由于光谱范围的两端信噪比较低,因此剔除了两端少部分光谱数据,选取有效波段区域为 380～1610 nm,土壤的原始光谱反射率如图 3(a)所示,其中高频噪声较为明显。对光谱反射率进行 Savitaky-Golay 卷积平滑[36]消除噪声后的数据如图 3(b)所示。

从原始光谱反射率可以看出整体反射率偏低,这是由于砂姜黑土颜色偏深的影响。光谱在 350～500 nm 和 570～850 nm 呈现明显上升趋势,在 830 nm左右光谱出现明显波动,可能与土壤中的有机质对光谱的吸收有关[37];在 500～570 nm 呈现几乎零增长趋势,可能与土壤中的铁氧化物对光谱的吸收有关[38]。在 950～1657 nm 光谱整体增长趋势较缓,在 1250 nm 处出现明显波动,这是由于土壤中的铁氧化物对光谱的吸收减弱[39];在 1450 nm 处出现明显的吸收谷,有相关文献[40]报道,该处是水分的吸收带,故反射率降低。以上分析表明,土壤的光谱反射率与土壤速效磷含量之间存在较好的相关性。

4　分析与讨论

模型训练中 PLS 算法参数较为复杂,通过比较不同的潜在变量(LV)来选择最优模型,因此必须使用不同数量 LV 的 f_{RMSE} 趋势进行比较和分析。从图 4 可知,线性的偏最小二乘(Linear-PLS)、径向基的偏最小二乘(RBF-PLS)和 Sigmoid 函数的偏最小二乘(Sigmoid-PLS)3 个模型的 f_{RMSE} 变化趋势基本一致。当 LV 个数增加时,线性和非线性 PLS 的 f_{RMSE} 值都呈现了先降低再升高的状态;当 LV 个数为 11 时,Linear-PLS 方均根误差达到最小值;当 LV 个数为 11 时,RBF-PLS 方均根误差达到最小;当 LV 个数为 10 时,Sigmoid-PLS 方均根误差达到最小。从 LV 趋势图中可知,线性和非线性的 PLS 模型其参数具有较小差别,模型相似性较高。两个非线性模型 RBF-PLS 和 Sigmoid-PLS 的 f_{RMSE} 相对于线性模型有了显著的降低,尤其在最低点上线性模型的表现较差。因此,从 PLS 算法中可知,非线性的 PLS 对光谱特征的拟合程度更高,更能构建出有效的回归模型。

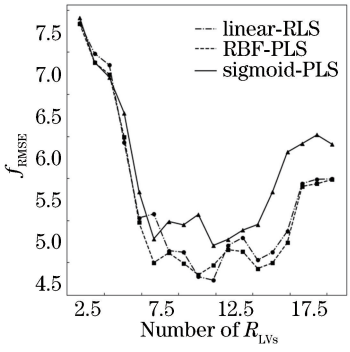

图 4　线性和非线性 PLS 中不同 LV 对应的方均根误差值

Fig. 4　f_{RMSE} values of different LV numbers in linear and nonlinear PLS

比较本文训练模型,Linear-PLS、Linear-SVR 和 Linear-Ridge 为线性模型,RBF-PLS、RBF-SVR、RBF-Ridge、Sigmoid-PLS、Sigmoid-SVR、Sigmoid -

Ridge 为非线性模型，每个模型都通过网格式调参，得到的最优模型参数如表 2 所示。在单模型调参中，R_{LVs} 为 PLS 模型潜在变量，其值越大，模型的拟合能力越强。R_{gamma} 为 RBF，sigmoid 核函数的参数，其值越大，高斯分布越窄，模型复杂度越大，拟合能力越强。R_{cofe0} 为 sigmoid 内核和多项式内核的截距参数，其可以用来控制数据的缩放。C 为 SVM 模型的惩罚因子，模型的正则化强度越大，模型的复杂度越高，拟合能力越强。R_{Alpha} 为 Ridge 中的参数，其值越大，正则化的强度越大。由表 2 的预测精度分类可知，所有模型的精确度为 B 类，属于中等水平，其中 RBF-PLS 和 Sigmoid-PLS 的 R^2 最高，为最优单模型。由表 2 可知，PLS 在建模效果上优于 SVR 和 Ridge。从建模集的结果分析可知，线性和非线性模型之间差异性不大，从测试集的结果分析可知，除 Ridge 模型外其他非线性模型在 f_{RPD} 和 R^2 上都明显优于线性模型。在 PLS 算法中，非线性 R^2 最高为 0.73，线性 $R^2=0.80$；在 SVR 算法中，非线性 R^2 最高为 0.72，线性 $R^2=0.69$；在 Ridge 中，非线性 R^2 最高为 0.7，线性 $R^2=0.69$。比较可知，测试集中非线性结果明显高于线性，其中 RBF 核函数更加适合光谱曲线拟合。

由以上分析可知，土壤速效磷的回归模型在预测精度分类等级上相同；但其光谱特征更趋向非线性关系，因此非线性模型与线性模型具有差异性，同时非线性模型更优于线性模型。下一步研究将遍历模型所有的组合方式，优选出单模型，然后比较 3 种次级集成学习算法的测试集结果。

表 2　最优单模型的测试结果

Table 2　Testing results of optimal single model

Modeling method	Training set		Testing set		Prediction level (testing set)	Parameter
	f_{RPD}	R^2	f_{RPD}	R^2		
PLS	1.66	0.73	1.65	0.68	B	$R_{\mathrm{LVs}}=11$
RBF-PLS	1.58	0.71	1.79	0.73	B	$R_{\mathrm{LVs}}=11$, $R_{\mathrm{gamma}}=0.016$
Sigmoid-PLS	1.55	0.70	1.75	0.73	B	$R_{\mathrm{LVs}}=10$, $R_{\mathrm{gamma}}=0.00085$, R_{cofe0} cofe0$=4.5$
SVR	1.60	0.74	1.53	0.69	B	$C=10000$
RBF-SVR	1.70	0.76	1.66	0.72	B	$C=2000000$, $R_{\mathrm{gamma}}=0.0028$
Sigmoid-SVR	1.59	0.73	1.55	0.70	B	$C=10^{11}$, $R_{\mathrm{gamma}}=0.000001$, $R_{\mathrm{cofe0}}=0$
Ridge	1.60	0.74	1.50	0.69	B	$R_{\mathrm{Alpha}}=0.001$
RBF-Ridge	1.55	0.74	1.50	0.70	B	$R_{\mathrm{Alpha}}=0.00006$, $R_{\mathrm{gamma}}=0.01$
Sigmoid-Ridge	1.52	0.72	1.50	0.69	B	$R_{\mathrm{Alpha}}=4\times10^{-7}$, $R_{\mathrm{gamma}}=0.0005$, $R_{\mathrm{cofe0}}=0.9$

土壤速效磷的各预测模型间具有一定的差异性，通过模型组合可提高预测精度。在 Stacking 方法基础上，研究比较了 9 种模型的 502 种组合结果。遍历分析可得 4 个单模型组合为最优，分别为 Sigmoid-PLS、Linear-SVR、RBF-SVR 和 Sigmoid-SVR。最优单模型中既有线性模型 SVR，也有非线性模型 Sigmoid-PLS、RBF-SVR 和 Sigmoid-SVR，因此线性和非线性的组合是可以提高结果准确率的。同时，Ridge 算法在组合中全部被剔除，由此可见，其算法在高光谱建模中具有一定的缺陷。

先比较随机森林、提升树和梯度提升树三种集成学习算法，其中，图 5 所示为梯度提升树参数调优结果，R_{loss} 在算法中主要有 4 种值：方均差损失函数 F_{ls}，如图 5(a)所示；方均差和最小偏差集合损失函数 F_{huber}，如图 5(b)所示；分位数损失函数 F_{quantile}，如图 5(c)所示；最小绝对偏差损失函数 F_{lad}，如图 5(d)所示。算法参数优化中，$R_{\mathrm{n\_estimators}}$ 为最大迭代次数，其值越大拟合能力越强，但也易造成过拟合，因此每种 loss 值条件下最优 $R_{\mathrm{n\_estimators}}$ 值也有所不同。$R_{\mathrm{learning\_rate}}$ 为权重缩减系数，$R_{\mathrm{max\_depth}}$ 为决策树最大深度。从图 5 每个子图的 f_{RPD} 值波动区间可知，R_{loss} 为“lad”，$R_{\mathrm{n\_estimators}}=310$ 的梯度提升模型组合比其他更优，因此在图 5(d)中 $R_{\mathrm{max\_depth}}=4$ 时，$R_{\mathrm{learning\_rate}}$ 为 0.29 和 0.41 时，f_{RPD} 值为两个最高点。因此，比较图 5(d)两个最高点的结果可以发现，当 $R_{\mathrm{learning\_rate}}=0.41$ 时，建模集的 $f_{\mathrm{RPD}}=3.7$，测试集的 $f_{\mathrm{RPD}}=2.61$；$R_{\mathrm{learning\_rate}}=0.29$ 时，建模集的 $f_{\mathrm{RPD}}=2.56$，测试集的 $f_{\mathrm{RPD}}=2.55$，虽然后者测试集的 f_{RPD} 低 0.01，但其泛化性远优于前者。最终，梯度提升树模型最优结果的参数设置如下：$R_{\mathrm{learning\_rate}}=0.29$，$R_{\mathrm{max\_depth}}=4$，$R_{\mathrm{loss}}$ 为“lad”，$R_{\mathrm{n\_estimators}}=310$。

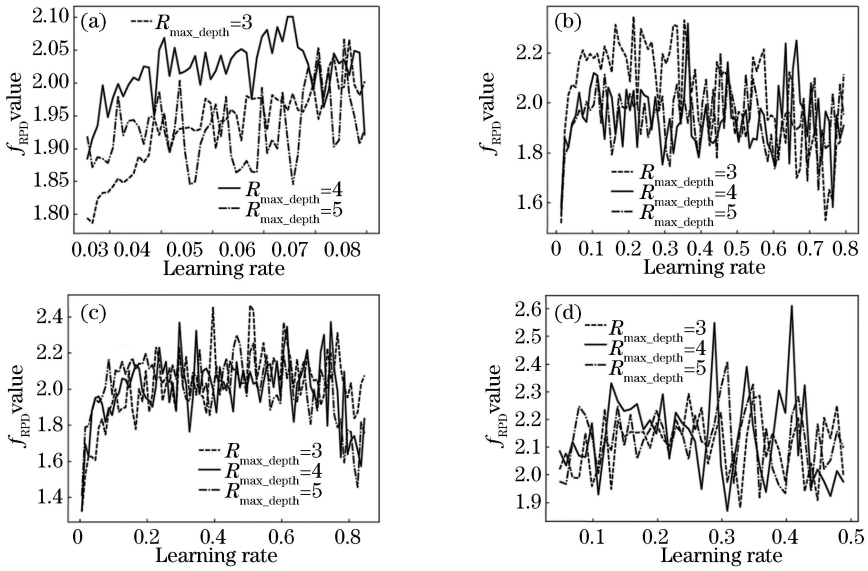

图 5 梯度提升树模型参数优化。(a) $R_{loss}=F_{ls}$, $R_{n\_estimators}=100$;(b) $R_{loss}=F_{huber}$, $R_{n\_estimators}=200$;(c) $R_{loss}=F_{quantile}$, $R_{n\_estimators}=200$;(d) $R_{loss}=F_{lad}$, $R_{n\_estimators}=310$

Fig. 5 Parameter optimization of GBDT model. (a) $R_{loss}=F_{ls}$, $R_{n\_estimators}=100$; (b) $R_{loss}=F_{huber}$, $R_{n\_estimators}=200$; (c) $R_{loss}=F_{quantile}$, $R_{n\_estimators}=200$; (d) $R_{loss}=F_{lad}$, $R_{n\_estimators}=310$

由表 3 和图 6 可知,模型组合后预测集的 f_{RPD} 值均超过了 2,从预测精度分类上为 A 类,因此多模型组合的结果远比单模型要优秀。从建模集和测试集的评价结果比较可知,提升树算法建模集的 R^2 达到了 0.9,但是由于其测试集的 R^2 仅为 0.82,因此提升树算法的模型组合产生了较高的过拟合。梯度提升树的建模集的 $R^2=0.88$, $f_{RPD}=2.56$,测试集 $R^2=0.86$, $f_{RPD}=2.55$,在算法中均为最高且其泛化程度最好。因此,由 3 种算法结果比较可知,梯度提升树的多模型组合结果不仅优于单模型,而且相对随机森林和提升树都具有一定程度的优化,因此在精确度和泛化性上为最优方案。

表 3 多种模型组合结果

Table 3 Results of multi-model combination

Ensemble method	Training set		Testing set		Prediction level (testing set)	Parameter
	f_{RPD}	R^2	f_{RPD}	R^2		
Random forest	2.10	0.84	2.08	0.84	A	$R_{n\_estimators}$, $R_{max\_depth}=5$
Boosting tree	2.86	0.90	2.12	0.82	A	$R_{n\_estimators}=300$, $R_{learning\_rate}=0.01$, $R_{max\_depth}=5$, $R_{loss}=F_{linear}$
GBDT	2.56	0.88	2.55	0.86	A	$R_{n\_estimators}=310$, $R_{learning\_rate}=0.29$, $R_{max\_depth}=4$, $R_{loss}=F_{lad}$

5 结 论

土壤速效磷的高光谱预测过程中,由于非同类的土壤化学成分具有较大的差异性,导致其光谱特征具有复杂度高、提取困难且不具有共性等问题,而光谱全波段进行单独回归算法的建模也易出现准确率低、泛化能力差等缺陷。针对光谱全波段,使用了 PLS、SVR 和岭回归算法[5-10],和随机森林、提升树和梯度提升树集成算法,以提高土壤速效磷的高光谱预测精度为目的,进行模型构建、优化、组合和比较。

本次实验使用光谱范围为 350~1700 nm 的非成像方式,相对于 400~1000 nm 的成像方式,在土壤速效磷上的相关性有显著提高[12]。从单模型分析中发现,PLS 模型在预测效果上最优,其次为 SVR 模型,最差为 Ridge 模型。而在线性算法预测效果中,3 个算法基本相同,但是在非线性的算法中,PLS 明显优于其他两种,尤其在 Sigmoid 的函数上 PLS 算法测试结果的 R^2 为 0.73,远高于 SVR 的 0.7和 0.69,因此 PLS 在单模型中泛化性强且预测效果较优。模型算法中线性和非线性核函数对数据预测效果具有一定的影响,其中高斯分布的非线性

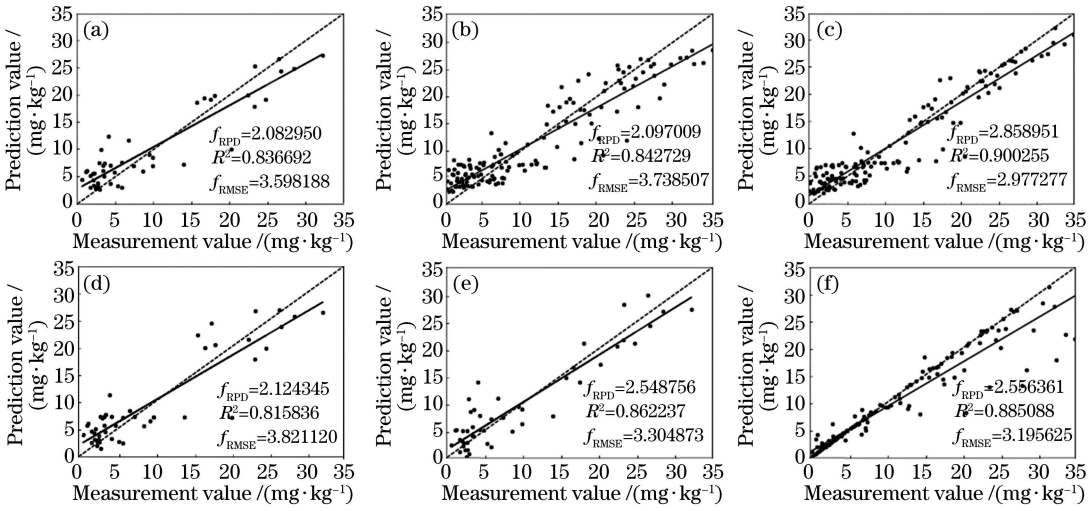

图 6 不同模型集成算法的结果。(a)基于建模集的随机森林结果；(b)基于测试集的随机森林结果；(c)基于建模集的提升树结果；(d)基于测试集的提升树结果；(e)基于建模集的梯度提升树结果；(f)基于测试集的梯度提升树结果

Fig. 6 Results of different model integration algorithms. (a) Results of random forest based on modeling set; (b) results of random forest based on testing set; (c) results of boosting tree based on modeling set; (d) results of boosting tree based on testing set; (e) results of GBDT based on modeling set; (f) results of GBDT based on testing set

核函数相对于其他核函数而言具有更好的拟合能力，因此该特性反映出了土壤高光谱特征具有高复杂性、提取困难的特点。本文着重讨论了多模型组合，从预测的精度分类上 3 个算法都属于 A 等级，但梯度提升树 R^2 为0.86，其次随机森林 R^2 为0.84，最差的提升数 R^2 为 0.82。其中从预测效果上分析梯度提升树算法最优，其预测效果相对于最优单模型 f_{RPD} 的 1.79 提高约 30%，R^2 的 0.73 提高 17.8%，并且与文献[11]的 f_{RPD} 值相比提高了 10%，精确度有较高的提升。

研究发现，集成算法中的梯度提升树是在提升树算法基础上利用最速下降法进行拟合的，在算法复杂度上优于随机森林和提升树，在测试集的评价中，其准确率和泛化性高于单模型和其他集成算法。多模型组合在土壤速效磷高光谱预测中具有两个优势：1)针对特征较为复杂的土壤光谱，不同模型的组合弥补了算法在回归预测时的缺陷，可大幅度提升其精确度；2)对不同类型的土壤、不同范围的光谱区域，通过多模型组合的集成算法可以抑制单模型的过拟合性，使得模型具有较高的泛化程度。因此，基于梯度提升树的多模型组合方法不仅有利于提高土壤高光谱的回归预测精度，而且有利于提升其在实际应用中的可行性。

参 考 文 献

[1] 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.

[2] Wu Q, Yang Y H, Xu Z L, *et al*. Applying local neural network and visible/near-infrared spectroscopy to estimating available nitrogen, phosphorus and potassium in soil[J]. Spectroscopy and Spectral Analysis, 2014, 34(8): 2102-2105.
吴茜, 杨宇虹, 徐照丽, 等. 应用局部神经网络和可见/近红外光谱法估测土壤有效氮磷钾[J]. 光谱学与光谱分析, 2014, 34(8): 2102-2105.

[3] Li X Y, Fan P P, Hou G L, *et al*. Rapid detection of soil nutrients based on visible and near infrared spectroscopy [J]. Spectroscopy and Spectral Analysis, 2017, 37(11): 3562-3566.
李雪莹, 范萍萍, 侯广利, 等. 可见-近红外光谱的土壤养分快速检测[J]. 光谱学与光谱分析, 2017, 37(11): 3562-3566.

[4] Shao Y N, He Y. Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy[J]. Soil Research, 2011, 49(2): 166-172.

[5] Jia S Y, Yang X L, Li G, *et al*. Quantitatively determination of available phosphorus and available potassium in soil by near infrared spectroscopy combining with recursive partial least squares[J]. Spectroscopy and Spectral Analysis, 2015, 35(9): 2516-2520.
贾生尧, 杨祥龙, 李光, 等. 近红外光谱技术结合递

归偏最小二乘算法对土壤速效磷与速效钾含量测定研究[J]. 光谱学与光谱分析, 2015, 35(9): 2516-2520.

[6] Gatius F, Miralbés C, David C, *et al*. Comparison of CCA and PLS to explore and model NIR data[J]. Chemometrics and Intelligent Laboratory Systems, 2017, 164: 76-82.

[7] Kawamura K, Tsujimoto Y, Rabenarivo M, *et al*. Vis-NIR spectroscopy and PLS regression with waveband selection for estimating the total C and N of paddy soils in Madagascar[J]. Remote Sensing, 2017, 9(10): 1081.

[8] Genisheva Z, Quintelas C, Mesquita D P, *et al*. New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR)[J]. Food Chemistry, 2018, 246: 172-178.

[9] Sarathjith M C, Das B S, Wani S P, *et al*. Comparison of data mining approaches for estimating soil nutrient contents using diffuse reflectance spectroscopy[J]. Current Science, 2016, 110(6): 1031-1037.

[10] Zhang J J, Guo X, Zhao X M. Hyperspectral characteristics and inversion models of total phosphorus and available phosphorus in paddy fields in southern hilly China[J]. Jiangsu Agricultural Sciences, 2016, 44(7): 522-525.
张佳佳, 郭熙, 赵小敏. 南方丘陵稻田土壤全磷、有效磷高光谱特征与反演模型[J]. 江苏农业科学, 2016, 44(7): 522-525.

[11] Qi H J, Li S W, Arnon K, *et al*. Prediction method of soil available phosphorus using hyperspectral data based on PLS-BPNN[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(2): 166-172.
齐海军, 李绍稳, Karnieli Arnon, 等. 基于 PLS-BPNN 算法的土壤速效磷高光谱回归预测方法[J]. 农业机械学报, 2018, 49(2): 166-172.

[12] Wang W C, Li S W, Qi H J, *et al*. The difference analysis of soil available phosphors content imaging and non-imaging spectra prediction[J]. Jiangsu Journal of Agricultural Sciences, 2018, 34(4): 811-817.
王文才, 李绍稳, 齐海军, 等. 土壤速效磷含量成像和非成像光谱预测差异性分析[J]. 江苏农业学报, 2018, 34(4): 811-817.

[13] Fu Z L. A universal ensemble learning algorithm[J]. Journal of Computer Research and Development, 2013, 50(4): 861-872.
付忠良. 通用集成学习算法的构造[J]. 计算机研究与发展, 2013, 50(4): 861-872.

[14] Kaneko H, Funatsu K. Applicability domain based on ensemble learning in classification and regression analyses[J]. Journal of Chemical Information and Modeling, 2014, 54(9): 2469-2482.

[15] Okujeni A, van der Linden S, Suess S, *et al*. Ensemble learning from synthetically mixed training data for quantifying urban land cover with support vector regression[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(4): 1640-1650.

[16] Mesquita D P P, Gomes J P P, Souza Junior A H. Ensemble of efficient minimal learning machines for classification and regression[J]. Neural Processing Letters, 2017, 46(3): 751-766.

[17] Zheng M D, Xiong H G, Qiao J F, *et al*. Remote sensing inversion of soil organic matter based on broad band and narrow band comprehensive spectral index[J]. Laser & Optoelectronics Progress, 2018, 55(7): 072801.
郑曼迪, 熊黑钢, 乔娟峰, 等. 基于宽波段与窄波段综合光谱指数的土壤有机质遥感反演[J]. 激光与光电子学进展, 2018, 55(7): 072801.

[18] Ying L N, Zhou W D. Comparative analysis of multiple chemometrics methods in application of laser-induced breakdown spectroscopy for quantitative analysis of soil elements[J]. Acta Optica Sinica, 2018, 38(12): 1214002.
应璐娜, 周卫东. 对比分析多种化学计量学方法在激光诱导击穿光谱土壤元素定量分析中的应用[J]. 光学学报, 2018, 38(12): 1214002.

[19] 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.

[20] Zou T T, Wang Y, Song H L. Near infrared spectroscopy combined with support vector regression applied for rapid and nondestructive detection of adulterate goat milk powder[J]. Journal of Chinese Institute of Food Science and Technology, 2017, 17(8): 261-267.
邹婷婷, 王莹, 宋焕禄. 牛乳清粉掺伪羊乳粉的近红外光谱法快速无损检测[J]. 中国食品学报, 2017, 17(8): 261-267.

[21] Ni W D, Nørgaard L, Mørup M. Non-linear calibration models for near infrared spectroscopy[J]. Analytica Chimica Acta, 2014, 813: 1-14.

[22] Nie P C, Wu D, Yang Y, *et al*. Fast determination of boiling time of yardlong bean using visible and near infrared spectroscopy and chemometrics[J]. Journal of Food Engineering, 2012, 109(1): 155-161.

[23] Ting J A, D′Souza A, Vijayakumar S, *et al*. Efficient learning and feature selection in high-dimensional regression[J]. Neural Computation, 2010, 22(4): 831-886.

[24] Kalika D, Morton K D, Collins L M, *et al*. Hyperbolic and PLSDA filter algorithms to detect buried threats in GPR data[J]. Proceedings of SPIE, 2014, 9072: 90720U.

[25] Jain A, Smarra F, Mangharam R. Data predictive control using regression trees and ensemble learning[C]//2017 IEEE 56th Annual Conference on Decision and Control (CDC), December 12-15, 2017, Melbourne, VIC, Australia. New York: IEEE, 2017: 4446-4451.

[26] Kaneko H. Automatic outlier sample detection based on regression analysis and repeated ensemble learning[J]. Chemometrics and Intelligent Laboratory Systems, 2018, 177: 74-82.

[27] Kabir A, Ruiz C, Alvarez S A, *et al*. Regression, classification and ensemble machine learning approaches to forecasting clinical outcomes in ischemic stroke[M]//Peixoto N, Silveira M, Ali H, *et al*. Biomedical engineering systems and technologies. Cham: Springer, 2018, 881: 376-402.

[28] Alazzam I, Alsmadi I, Akour M. Software fault proneness prediction: a comparative study between bagging, boosting, and stacking ensemble and base learner methods[J]. International Journal of Data Analysis Techniques and Strategies, 2017, 9(1): 1-16.

[29] Li S F, Jia M Z, Dong D M. Fast measurement of sugar in fruits using near infrared spectroscopy combined with random forest algorithm[J]. Spectroscopy and Spectral Analysis, 2018, 38(6): 1766-1771.
李盛芳, 贾敏智, 董大明. 随机森林算法的水果糖分近红外光谱测量[J]. 光谱学与光谱分析, 2018, 38(6): 1766-1771.

[30] Ge X Y, Ding J L, Wang J Z, *et al*. Estimation of soil moisture content based on competitive adaptive reweighted sampling algorithm coupled 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.

[31] Kong Q Q, Ding X Q, Gong H L. Application of improved random forest pruning algorithm in tobacco origin identification of near infrared spectrum[J]. Laser & Optoelectronics Progress, 2018, 55(1): 013006.
孔清清, 丁香乾, 宫会丽. 改进的修剪随机森林算法在烟叶近红外光谱产地识别中的应用研究[J]. 激光与光电子学进展, 2018, 55(1): 013006.

[32] Gao Y, Cui L J, Lei B, *et al*. Estimating soil organic carbon content with visible-near-infrared (Vis-NIR) spectroscopy[J]. Applied Spectroscopy, 2014, 68(7): 712-722.

[33] 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.

[34] Shi Y F, Chang S P. A study of determining the available phosphorus in high orgnaic soils by means of $NaHCO_3$ extraction, ammonium molybdate-tartaric emetic-ascrbic acid colorimetry[J]. Journal of Gansu Agricultural University, 1984, 19(2): 108-111.
石应福, 常淑平. 对碳酸氢钠浸提—钼锑抗比色法测定高含量有机质土壤有效磷的改进试验[J]. 甘肃农大学报, 1984, 19(2): 108-111.

[35] Claeys D D, Verstraelen T, Pauwels E, *et al*. Conformational sampling of macrocyclic alkenes using a Kennard-Stone-based algorithm[J]. The Journal of Physical Chemistry A, 2010, 114(25): 6879-6887.

[36] Liu G S, Guo H S, Pan T, *et al*. Vis-NIR spectroscopic pattern recognition combined with SG smoothing applied to breed screening of transgenic sugarcane[J]. Spectroscopy and Spectral Analysis, 2014, 34(10): 2701-2706.
刘桂松, 郭昊淞, 潘涛, 等. Vis-NIR 光谱模式识别结合 SG 平滑用于转基因甘蔗育种筛查[J]. 光谱学与光谱分析, 2014, 34(10): 2701-2706.

[37] Bayer A, Bachmann M, Müller A, *et al*. A comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African ecosystem[J]. Applied and Environmental Soil Science, 2012, 2012: 971252.

[38] 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.

[39] Peng J, Zhang Y Z, Zhou Q, *et al*. The progress on the relationship physics-chemistry properties with spectrum characteristic of the soil [J]. Chinese Journal of Soil Science, 2009, 40(5): 1204-1208.
彭杰, 张杨珠, 周清, 等. 土壤理化特性与土壤光谱特征关系的研究进展[J]. 土壤通报, 2009, 40(5): 1204-1208.

[40] 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.