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

葛翔宇[1,2,3**]，丁建丽[1,2,3*]，王敬哲[1,2,3]，王飞[1,2,3]，蔡亮红[1,2,3]，孙慧兰[4]

[1]新疆大学资源与环境科学学院，新疆 乌鲁木齐 830046；

[2]新疆大学绿洲生态教育部重点实验室，新疆 乌鲁木齐 830046；

[3]新疆大学智慧城市与环境建模自治区普通高校重点实验室，新疆 乌鲁木齐 830046；

[4]新疆师范大学地理科学与旅游学院，新疆 乌鲁木齐 830054

摘要 土壤含水量是干旱区地表水-热-溶质耦合运移的关键指标；以干旱区典型样点实测土壤含水量及其室内可见光-近红外光谱数据作为数据集，通过蒙特卡罗交叉验证确定 77 个有效样本；基于竞争适应重加权采样算法筛选出最优光谱变量子集，利用 3 种机器学习方法——BP 神经网络、随机森林回归和极限学习机建立土壤含水量预测模型，进而实现土壤含水量估算模型的优选。结果表明：竞争适应重加权采样算法能有效剔除无关变量，从 2151 个光谱波段中优选出 20 个特征波段，其中 R_{1848} 与土壤含水量的最大相关系数为 0.531；引入偏最小二乘模型和机器学习方法进行对比，分析发现机器学习方法的预测结果比偏最小二乘模型更高；分析比较 BP 神经网络、随机森林回归和极限学习机的建模结果可知：极限学习机模型建模在机器学习方法中的效果最佳，决定系数 $R^2=0.918$，均方根误差 RMSE=0.015，相对分析误差 RPD=3.123，四分位数间隔 RPIQ=3.325；机器学习能显著提升光谱建模反演土壤含水量的精度和稳定性，显示出其在非线性问题中具有很强的透析力和较好的模型稳健性，针对干旱区土壤水分的精准预测和定量估算具有可行性，可为干旱区土壤墒情、精准农业等研究提供科学参考。

关键词 光谱学；土壤含水量估算；机器学习；竞争适应重加权采样算法；极限学习机；随机森林

中图分类号 O436 **文献标识码** A **doi**：10.3788/AOS201838.1030001

Estimation of Soil Moisture Content Based on Competitive Adaptive Reweighted Sampling Algorithm Coupled with Machine Learning

Ge Xiangyu[1,2,3**], Ding Jianli[1,2,3*], Wang Jingzhe[1,2,3], Wang Fei[1,2,3], Cai Lianghong[1,2,3], Sun Huilan[4]

[1] *College of Resource and Environment Sciences, Xinjiang University, Urumqi, Xinjiang* 830046, *China*;

[2] *Key Laboratory of Oasis Ecology, Xinjiang University, Urumqi, Xinjiang* 830046, *China*;

[3] *Key Laboratory of Smart City and Environment Modelling of Higher Education Institute, Xinjiang University, Urumqi, Xinjiang* 830046, *China*;

[4] *School of Geographical Science and Tourism, Xinjiang Normal University, Urumqi, Xinjiang* 830054, *China*

Abstract Soil moisture content is an important indicator that reflects the coupled surface water-heat-solute transport in arid regions. The visible and near-infrared spectroscopy has been widely used for soil moisture content prediction owing to its rapid response. The soil moisture content and corresponding spectral data are obtained in the laboratory; then, the calibration datasets ($n=77$) are selected using Monte Carlo cross-validation algorithm. The competitive adaptive reweighted sampling algorithm is used to optimize spectral variables. Three machine learning algorithms, namely back propagation neural network, random forest regression, and extreme learning machine are used to construct predicting models. The results reveal that competitive adaptive reweighted sampling algorithm can effectively filter and eliminate massive irrelevant variables. Herein, a total of 20 feature bands are divided from all spectral bands, where the band of R_{1848} is the most prominent (the maximum correlation coefficient is 0.531). The performance of models based on machine learning algorithms is superior to those based on partial least squares

收稿日期：2018-04-28；**修回日期**：2018-05-07；**录用日期**：2018-05-12

基金项目：自治区重点实验室专项基金(2016D03001)、国家自然科学基金(41771470，U1303381，41661046)

* **E-mail**：watarid@xju.edu.cn；** **E-mail**：xiangyu_gexj@163.com

regression, with the optimal prediction of the coefficient of determination (R^2), root mean square error of prediction (RMSE), residual predictive deviation (RPD), and ratio of performance to interquartile range (RPIQ). Compared with the predictive effects of all the models, the extreme learning machine-based predicting model is the most effective ($R^2=0.918$, RMSE=0.015, RPD=3.123, and RPIQ=3.325). Compared with common linear models, the machine learning algorithms can effectively improve the precision and stability of the quantitative estimation of soil moisture content. The results provide scientific guidance and baseline data for the accurate monitoring of soil moisture content and precision agriculture in arid regions.

Key words spectroscopy; estimation of soil moisture content; machine learning; competitive adaptive reweighted sampling algorithm; extreme learning machine; random forest

OCIS codes 300.6170; 280.4991; 200.4260

1 引　言

土壤水分在调节陆地-大气间水热传输、能量交换中起着重要作用，深刻影响着气候条件的时空变化[1]。土壤含水量(SMC)是水文、生态、农业等研究领域中最易发生变化的参数之一，对干旱、半干旱区域绿洲农业、绿色生态和水资源管理等具有重要意义[2]。因此，为了满足干旱区农业对 SMC 的精准管理与实时监测，需要采用比传统方法更具有高效性与精准性的新的观测手段。可见光-近红外(Vis-NIR)光谱技术能捕捉到光谱土壤性质中的微小差异，具有高效、易用等特性，在 SMC 研究中得到广泛了应用[3-4]。根据土壤 Vis-NIR 光谱曲线在 1400 nm 与 1900 nm 附近对 SMC 变化的响应规律，利用定量建模手段可以有效地对 SMC 进行估算预测[5]。然而，土壤 Vis-NIR 光谱的噪声增加了挖掘信息的难度，且 SMC 与土壤光谱之间存在非线性、异方差性等复杂关系[6]。因此，优选变量以及利用新的方法提高模型精度、稳定性与泛化性是反演预测 SMC 的热点。

目前，许多研究都是以线性模型进行光谱定量反演的，且模型训练及筛选均基于偏最小二乘(PLSR)[7]、光谱变换[8]等简单方法，在定量表达特征波段和 SMC 的相关性、改善模型稳健性、降低模型过程冗杂度等方面仍待优化。机器学习能综合考量 SMC 光谱响应，并优化复杂过程，提高预测精度。其中的 BP 神经网络(BPNN)、随机森林回归(RFR)和极限学习机(ELM)在解析非线性问题时的表现较好，是建模反演研究中的热点[9-11]。Nawar 等[12]基于 Vis-NIR 土壤光谱，利用神经网络和 RFR 等方法建立了土壤全氮和全碳的预测模型，通过对比分析发现神经网络表现最佳。Liang 等[13]对从 43 个高光谱植被指数中选取的叶面积指数值最佳的植被指数进行分析，并比较人工神经网络和 RFR 模型的反演精度，结果表明 RFR 是其研究中的最佳建模方法，线性回归决定系数 $R^2=0.928$。Khosravi 等[11]为了预测污染土壤中的铅、锌含量，构建了 PLSR、支持向量机和 ELM 共 3 种模型，对土壤进行预测评估，结果表明 ELM 的预测效果最佳，对铅和锌的预测决定系数分别为 0.93 和 0.87，并认为反射光谱结合 ELM 算法是一种快速、精准、易行的方法。可见，BPNN、RFR 和 ELM 这 3 种方法在建模中均取得了良好的建模精度和预测效果，在 Vis-NIR 光谱反演模型中是行之有效的方法，但目前还未见在 SMC 研究中使用这三种方法的文献报道。基于此，本文以干旱区绿洲土壤为研究对象，基于竞争适应重加权采样(CARS)筛选敏感波段，利用 BPNN、RFR 和 ELM 这 3 种机器学习算法构建研究区 SMC 估算模型，以 PLSR 模型为本底进行比较分析，提出干旱区 SMC 最优模拟预测模型，以期提高干旱区 SMC 的 Vis-NIR 光谱预测精度和稳定性。

2 材料与方法

2.1 土壤样本采集

以新疆维吾尔自治区塔里木盆地北缘的渭干河-库车河绿洲(简称渭-库绿洲)为研究区，其坐标为 41°08′～41°55′N、81°06′～83°37′E[14]。根据历史观测点及区域特征，布设采样点数 $n=82$。为了便于验证，利用 GPS 记录采样点位置。各点土壤样品采用五点法混合采集，各采样点的样品分别用铝盒和自封袋取回。在实验室进行处理时，铝盒样品采用室内烘干法得到 SMC，自封袋样品在室内自然风干，研磨后过 2 mm 筛，以测量土壤的 Vis-NIR 光谱数据。

采用美国 ASD FieldSpec3 型光谱仪于暗室中采集土壤的 Vis-NIR 光谱数据，其波长范围为 350～2500 nm，其中 350～1000 nm 波长范围的采样间隔为 1.4 nm，1000～2500 nm 范围的采样间隔为 2 nm，重采样间隔为 1 nm。在暗室中，以 50 W

的卤素灯作为光源，光源与土壤表面的距离为50 cm，卤素灯的天顶角为15°，光谱仪探头与样品之间的距离为10 cm，每次测量前均用白板优化定标。在本实验中，各土样均于4个方向共采集16条光谱曲线，以其算术平均值作为该土样的原始光谱数据。

2.2　数据预处理

将上述土壤光谱原始数据利用Savitzky-Golay(S-G)方法进行平滑处理，处理后的数据作为后续光谱数据集。为了避免异常样本值对优选变量和建模优选结果的影响，利用蒙特卡罗交叉验证(MCCV)对光谱和SMC样本进行验证并剔除异常样本，以降低异常值对研究的掩蔽影响，从而最终确定有效样本[15]。从82个样本点中确定77个有效样本进行后续研究，图1所示为全样本MCCV土壤预测残差的均值-标准差分布。

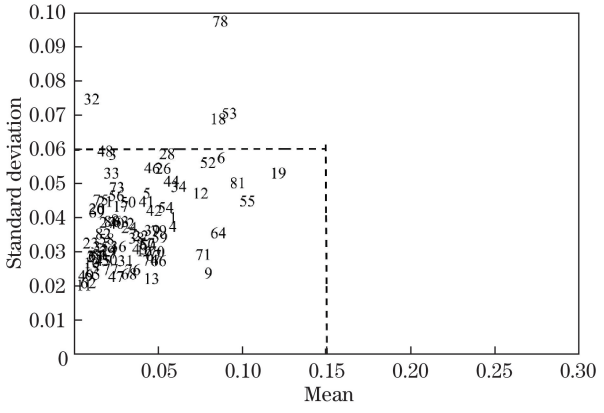

图1　全样本MCCV土壤预测残差的均值-标准差分布

Fig. 1　Mean-standard deviation distribution of soil-residual prediction for full-sample MCCV

2.3　特征波长选择算法

CARS算法为光谱特征波段优选提供了新的方法[16]。CARS算法模仿达尔文进化论中适者生存的原则，根据每个变量的重要程度，以迭代和竞争的方式从n个蒙特卡罗采样运行中顺序选择波长子集。在每次采样运行中，利用指数衰减函数和自适应重加权采样进行执行波长选择和竞争波长选择，根据回归系数选择绝对值大的关键变量，并利用交叉验证选择交叉验证均方根误差(RMSECV)最小的子集。

本研究所用CARS算法及其绘图在MATLAB R2014b软件中实现。

2.4　建模分析方法与精度评价

利用BPNN、RFR和ELM构建SMC预测模型，对比分析3种算法对干旱区SMC建模的效果。选取3种建模方法的原因是综合考虑到ELM与BPNN是本质同类算法，均为基于神经网络的预测模型；RFR与BPNN、ELM是本质异类算法，也是具有代表性的非线性回归预测模型。因此，本研究构建的SMC模型更能体现建模的价值和现实意义。

BPNN是一种多层前馈神经网络，具有自学习能力。本研究中使用的BPNN是由输入层、隐藏层和输出层组成的3层网络[17]。它包括2个阶段：第1个阶段是前馈阶段，外部输入信息在输入节点处向前传播，以在输出单元处计算输出信息信号；第2个阶段是反向阶段，根据输出单元计算和观测信息信号之间的差异，对连接权重进行修改。将实际输出与期望输出进行比较，不满足时进入误差反向阶段，直至输出误差在合理的范围内。参考相关研究[12]将隐藏层设置为10层。

RFR模型是一种基于决策树的集合学习算法。对不平衡的样本而言，该算法能平衡误差，并且实现起来较为简单。决策树代表一组具有分层组织的条件或限制，并且从树根到树叶依次应用[18]。RFR从许多随机抽取的自举样本开始，并在原始训练数据集中进行替换。决策树适用于每个Bootstrap样本，每个决策树的节点是按照一定比例随机抽取的。整个回归问题通过对所有决策树进行平均来得到最终的预测结果。RFR中需要优化2个参数：根据样本实际情况设置的决策树数量(n_{tree}，设置为500)和每个节点的输入变量数(m_{try}，设置为总变量数的1/3)[19]。

ELM是发展于单隐含层前馈神经网络的新型神经网络算法，是为快速训练而设计的单层前馈神经网络算法[20]。训练过程中避免了频繁调整迭代网络的输入权值以及隐元的偏置，并能得到唯一的最优解。其目的是通过使用分配给神经网络隐层的随机权重来降低其训练的复杂性。ELM以学习力迅速、泛化性突出、参数设置便捷等优点来弥补传统神经网络方法中出现的训练时期长、学习率敏感等不足(本文中Hidden nodes为30)。

以上3种方法的建模过程均在MATLAB R2014b软件中实现。

基于Kennard-Stone(K-S)算法进行样本划分，选取62个样本点作为建模集，15个样本点作为验证集，分别建立预测模型。为了量化基于BPNN、RFR和ELM的实测SMC以及预测值建模的效果和性能，选用决定系数R^2、均方根误差(RMSE)、相

对分析误差(RPD)和四分位数间隔(RPIQ)这 4 个参数对模型进行评估。R^2 值越大，表明建模的精度越高。RMSE 表示模型的预测能力，其值的大小与模型精度成反比。RPD 已广泛用于度量评估土壤属性预测模型的准确性；RPD≥2.0 为模型预测土壤特性极佳的指标；1.4≤RPD<2.0 表示模型预测效果可以接受，但有待改进；RPD<1.4 说明模型的可靠性较低，无法对样本实测值进行预测[2]。RPIQ 是四分位数间距与 RMSE 的比值，四分位数间距即样本的第三分位数与第一分位数的差值；RPIQ≥2.2表示模型极佳，1.7≤RPIQ<2.2 表示模型有较均衡的预测能力，RPIQ<1.4 表示模型的可信度低[18]。图 2 所示为实验及模型计算过程流程图。

3　结果与分析

3.1　样品的 SMC

表 1 所示为 SMC 的描述性统计特征。由表 1 可知：建模集与验证集对应的均值分别为 14.06%和 14.83%，变异系数分别为 36.59%和 26.37%；全样本的 SMC 均值为 14.21%，变异系数为 34.58%，为中等变异[2]。全样本的均值和变异系数均介于建模集和验证集之间。

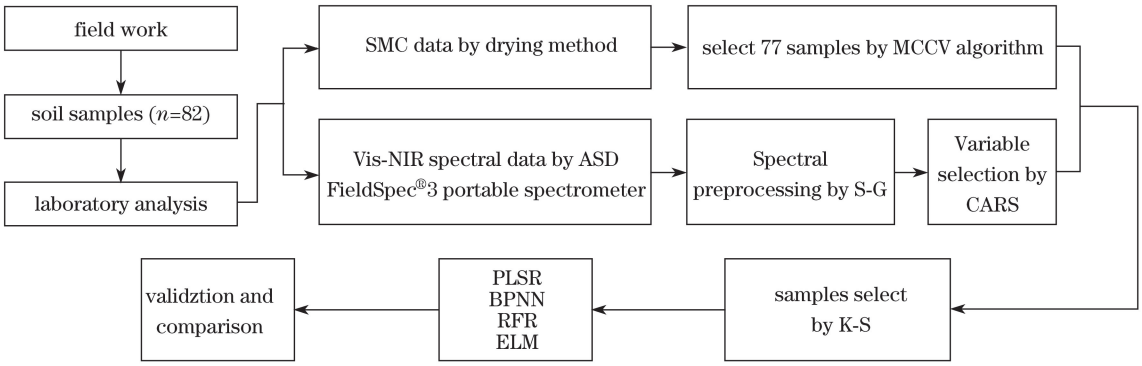

图 2　实验及模型计算过程流程图

Fig. 2　Flow chart of calculation process for experience and model

表 1　土壤样品的 SMC 统计特征

Table 1　Statistical characteristics of SMC of soil samples

Sample type	Number	Maximum	Minimum	Mean	Standard deviation	Coefficient of variation
Whole set	77	0.252	0.021	0.1421	0.049	0.3458
Calibration set	62	0.252	0.021	0.1406	0.051	0.3659
Validation set	15	0.216	0.067	0.1483	0.039	0.2637

3.2　土壤光谱曲线分析和特征波段优选

SMC 与光谱之间具有良好的响应，选取不同含水量的土壤样本的光谱反射率进行分析，如图 3 所示。可知：不同含水量样本的光谱反射率曲线的总体趋势较为一致，但存在一定差异；随着 SMC 增大，土壤光谱反射率与之成反比；在近红外波段的吸收峰深度明显大于在可见光波段的；光谱曲线在可见光波段单调上升，在近红外波段范围内较平缓。总体而言，不同 SMC 的土壤光谱反射率曲线较易被识别。

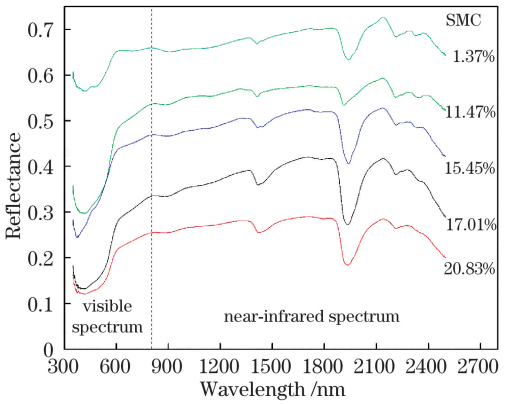

图 3　不同 SMC 土壤的光谱反射率

Fig. 3　Spectral reflectance of soils with different SMCs

采用 CARS 算法来优选特征波段，对放入样本集合的采样次数进行反复迭代，并寻找每次采样的 RMSECV 最小值，将此时采样次数对应的变量视为优选出的变量子集。图 4 所示为 CARS 算法的变量筛选过程。由图 4(a)可知，在指数衰减函数的作用下，变量个数在前 5 次采样过程中有明显递减，说明筛选波段数在迅速减少，此后逐渐平稳。由图 4(b)可知：第 34 次运行结果是 RMSECV 值最

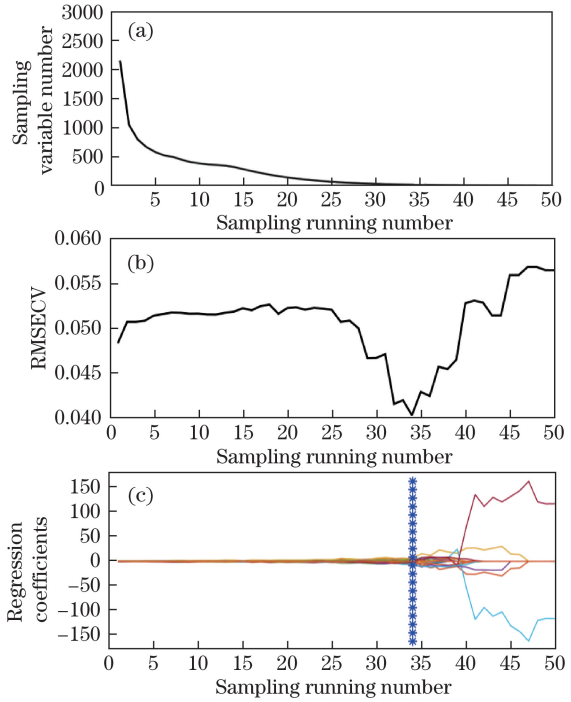

图 4　CARS 算法筛选变量的过程。(a)波长变量个数的变化;(b) RMSECV 的变化;(c) RMSECV 最小时变量回归系数的趋势

Fig. 4　Variable filtering process using CARS. (a) Variation in wavelength variable number; (b) variation in RMSECV; (c) trend of variable regression coefficient when RMSECV is minimum

小,为 0.040;在此之前,曲线基本处于下降的趋势,表明此段过程与 SMC 无关的光谱变量已被筛出并剔除;第 34 次之后,RMSECV 呈陡增趋势,这意味着关于 SMC 的关键光谱变量开始被剔除。由图 4(c)可知,第 34 次采样对应的变量集是 SMC 的最优光谱变量子集,包含 20 个光谱变量:R_{355}、R_{366}、R_{381}、R_{386}、R_{387}、R_{393}、R_{394}、R_{1542}、R_{1549}、R_{1848}、R_{1849}、R_{2312}、R_{2321}、R_{2322}、R_{2323}、R_{2355}、R_{2453}、R_{2454}、R_{2483} 和 R_{2484},如图 5 所示。其中:最大相关波段为 R_{1848},相关系数为 0.531(显著性检验阈值 $P^{**}=\pm0.292$);得到的最优变量子集中位于 1400～2400 nm 范围内的占多数,土壤光谱特征在这范围内以 C═O、C—H、Al—OH、O—H 基团的基频振动以及合频和倍频振动吸收为主要表现形式[3],这也是 Vis-NIR 光谱会在 1400,1900,2200 nm 附近出现特殊吸收峰的主要原因,因此优选出的波段是合理的。

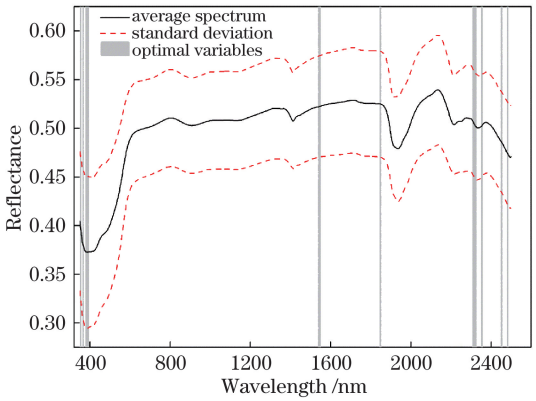

图 5　土壤样本反射率均值及最优光谱波段

Fig. 5　Mean reflectance of soil samples and optimal spectral bands

3.3　SMC 预测模型及验证

将筛选出的最优光谱变量作为 SMC 预测模型所需的预测变量,将经 MCCV 剔除异常值后的 SMC 作为响应变量,分别构建基于特征波长的 BPNN、RFR 和 ELM 的预测 SMC 模型。为了突出 3 种方法的建模效果,引入传统线性回归模型 PLSR 用于对建模方法的结果进行对比。各模型的详细信息如表 2 所示。由表 2 可知,BPNN、RFR 和 ELM 模型的建模效果优于 PLSR 模型的建模效果。由表 4 可以看出:在 4 种方法所构建的 SMC 模型中,PLSR 模型的 RPD 小于 1.4,这说明该方法不适用于本研究区的 SMC 预测;RFR 模型只能对 SMC 进行粗略预测(RPD＜2.0,RPIQ＜2.2);基于 BPNN 和 ELM 这 2 种机器学习方法所构建的模型具有极好的预测能力(RPD＞2.0),且 ELM 模型的预测能力优于 BPNN 模型的预测能力,RPIQ 为 3.325。由此可见,4 种模型按构建模型预测能力由优到劣的

表 2　SMC 预测结果

Table 2　Estimated SMC

Model	Variable number	Calibration set		Prediction set			
		RMSE	R^2	RMSE	R^2	RPD	RPIQ
PLSR	20	0.484	0.478	0.622	0.617	0.522	0.18401
BPNN	20	0.027	0.706	0.024	0.799	2.016	1.90200
RFR	20	0.024	0.872	0.021	0.898	1.647	2.18900
ELM	20	0.016	0.879	0.015	0.918	3.123	3.32500

顺序依次是 ELM、BPNN、RFR、PLSR。对比各模型建模集的 R^2、RMSE 和预测集的 R^2、RMSE 可知：ELM 的预测精度最高，建模集的 R^2 为0.879，RMSE 为 0.016，预测集的 R^2 为 0.918，RMSE 为 0.015；RFR 的预测精度次之，建模集的 R^2 为 0.872，RMSE 为 0.024，预测集的 R^2 为 0.898，RMSE 为 0.021；PLSR 的建模精度最低。由此可见，4 种模型按构建模型精度由高到低的顺序依次是 ELM、RFR、BPNN、PLSR。

综上所述，权衡 4 种建模方法的评价参数，在 SMC 预测模型的建模效果和预测效果中，ELM 模型最为突出。相较于传统的线性回归模型，ELM、RFR 和 BPNN 模型显著提高了预测能力和预测精度。对比 3 种机器学习方法可知，虽然 RFR 模型的预测精度高于 BPNN 模型，但其预测能力不如 BPNN。从 RPIQ 来看，RFR 模型建模效果的可信度比 BPNN 模型更高，而 ELM 在预测能力和模型精度方面均高于 BPNN 和 RFR。图 6 所示为经 CARS 算法筛选后 ELM 建模集的检验结果。

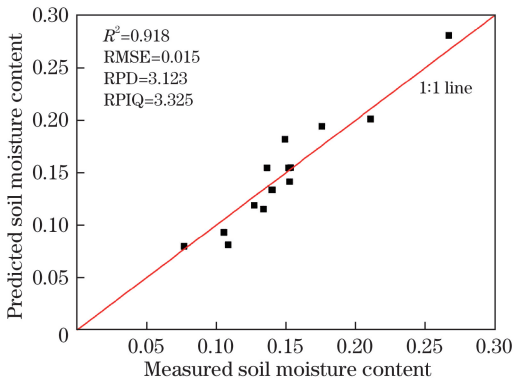

图 6　ELM 模型 SMC 的预测值与实测值

Fig. 6　Predicted and measured SMCs using ELM model

4　讨　　论

本研究的结果表明，CARS 算法可剔除大量冗余波段，优选出有效的特征波段，所得结果可作为 PLSR 和机器学习模型反演中的重要因子。最优变量子集不仅在机理上与 SMC 存在较好的响应，还能在提高预测精度和能力的同时，降低模型训练样本的复杂性，使机器学习模型能在有限样本下反演出理想的结果。

本研究构建模型所用 BPNN、RFR 和 ELM 这 3 种机器学习方法综合考量了本质同类算法和本质异类算法。程术希等[21]通过支持向量机、RFR 和 ELM 这 3 种机器学习方法构建品种鉴别方法，最终得出了 ELM 模型结果最优的结论，这与本研究得到的结果基本一致。在本节中增加 3∶1和 2∶1的建模预测比进行对比分析，讨论各机器学习模型的稳健性，结果如表 3 所示。

由表 3 可知，对于本研究建立的 BPNN、RFR 和 ELM 模型，随建模预测比从 4∶1变化至 2∶1，模型建模集的 R^2、RMSE 和预测集的 R^2、RMSE、RPD、RPIQ 基本保持不变。可见，本研究所用的 3 种机器学习模型的稳健性均较好。其中：RFR 模型在建模预测比为 4∶1时的 RPD 较小，在其他比例时 RFR 模型的 RPD 都比 BPNN 模型的 RPD 大，原因是 RFR 的稳健性相对较差；而 ELM 模型在不同建模预测比时均比 BPNN、RFR 模型表现得更优异，进一步验证了 ELM 模型不仅具有较小的训练误差，预测更精准，而且模型稳健性更好，具有很强的非线性解译建模能力。

表 3　不同建模预测比的 SMC 预测结果

Table 3　Predicted SMC based on different ratios of calibration to prediction

Model	Ratio of calculation to prediction	Variable number	Calibration set		Prediction set			
			RMSE	R^2	RMSE	R^2	RPD	RPIQ
BPNN	62∶15	20	0.027	0.706	0.024	0.799	2.016	1.902
	57∶20	20	0.020	0.842	0.023	0.800	1.826	1.499
	52∶25	20	0.023	0.765	0.024	0.800	1.947	2.010
RFR	62∶15	20	0.024	0.872	0.021	0.898	1.647	2.189
	57∶20	20	0.024	0.863	0.013	0.897	2.217	3.073
	52∶25	20	0.025	0.856	0.014	0.889	2.202	3.041
ELM	62∶15	20	0.016	0.879	0.015	0.918	3.123	3.325
	57∶20	20	0.019	0.869	0.014	0.919	3.102	3.241
	52∶25	20	0.015	0.877	0.016	0.918	2.569	2.958

相较于传统的线性回归模型，本研究使用的机器学习预测模型具有显著的优越性。PLSR 方法已广泛应用于土壤光谱研究[22-23]，尽管能够有效解决自变量间的多重共线性问题，但只能对某些特定的土壤属性与相应 Vis-NIR 光谱之间的潜在线性关系进行模拟。然而土壤性质多呈非标准正态分布，以 PLSR 方法为代表的线性回归可能不适用。相反，如果两者之间存在非线性关系，机器学习方法则通常会得出理想的预测结果。在两类建模方法中，机器学习模型不仅在统计结果上优于 PLSR，在预测能力上也表现出了更好的稳健性和泛化能力。

定量遥感反演的困难在于应用参量不完全是控制遥感信息的主导因子，仅为遥感信息提供弱信号[22,24]。虽然机器学习算法反演的准确性更高，但具有更多参数或超参数，通常需要大规模的复杂训练。理想的算法应该将模拟精度与训练参数、训练时间消耗进行平衡。本研究利用典型样本得到了较高的精度，在下一步研究中，拟从 SMC 和光谱响应机理方面对机器学习的相关参数进行调优，进而更好地诠释 SMC 与 Vis-NIR 光谱的内在联系，并在此基础上，将 Vis-NIR 光谱与现有多光谱遥感系统建立联系，从而为新疆干旱、半干旱地区的 SMC 研究提供科学依据。

5　结　论

以新疆渭-库绿洲为研究对象，利用 82 个土壤样本的 SMC 数据与实验室 Vis-NIR 光谱实测数据，在 CARS 算法的基础上，采用 BPNN、RFR 和 ELM 对优选出的特征波长构建 SMC 预测模型，并引入传统线性模型 PLSR 进行对比。对土壤光谱波长采用 CARS 算法，以达到特征波长的优选效果，并从 2151 个光谱波长中选取 20 个特征变量作为最优光谱变量子集，这 20 个特征变量为：R_{355}、R_{366}、R_{381}、R_{386}、R_{387}、R_{393}、R_{394}、R_{1542}、R_{1549}、R_{1848}、R_{1849}、R_{2312}、R_{2321}、R_{2322}、R_{2323}、R_{2355}、R_{2453}、R_{2454}、R_{2483} 和 R_{2484}。其中特征光谱 R_{1848} 与 SMC 的最大相关系数为 0.531。机器学习在 SMC 预测模型建模过程中缩短了训练时间，大幅提高了模型精度和预测能力，相对于线性模型，预测集的决定系数 R^2 从 0.617 增大到 0.918，RPD 从 0.522 增大到 3.123，RPIQ 从 0.18401增大到 3.325，实现了对 SMC 的精准预测。在 BPNN、RFR 和 ELM 这 3 种机器学习预测模型方法中，按建模精度由高到低的排序结果为 ELM、RFR、BPRR，按模型预测能力由高到低的排序结果为 ELM、BPNN、RFR。ELM 的建模精度和预测能力最好，建模集的 $R^2=0.879$，RMSE＝0.016，预测集的 $R^2=0.918$，RMSE＝0.015，RPD＝3.123，RPIQ＝3.325。说明在基于机器学习的 SMC 预测模型中，ELM 模型的表现最优。

参　考　文　献

[1] Kumar S V, Dirmeyer P A, Peters-Lidard C D, *et al*. Information theoretic evaluation of satellite soil moisture retrievals[J]. Remote Sensing of Environment, 2018, 204: 392-400.

[2] Cai L H, Ding J L. Inversion of soil moisture content based on hyperspectral multi-scale decomposition[J]. Laser & Optoelectronics Progress, 2018, 55(1): 013001.
蔡亮红, 丁建丽. 基于高光谱多尺度分解的土壤含水量反演[J]. 激光与光电子学进展, 2018, 55(1): 013001.

[3] Yu L, Zhu Y X, Hong Y S, *et al*. Determination of soil moisture content by hyperspectral technology with CARS algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2016, 32(22): 138-145.
于雷, 朱亚星, 洪永胜, 等. 高光谱技术结合 CARS 算法预测土壤水分含量[J]. 农业工程学报, 2016, 32(22): 138-145.

[4] Xu C, Zeng W Z, Huang J S, *et al*. Prediction of soil moisture content and soil salt concentration from hyperspectral laboratory and field data[J]. Remote Sensing, 2016, 8(1): 42.

[5] Oltra-Carrió R, Baup F, Fabre S, *et al*. Improvement of soil moisture retrieval from hyperspectral VNIR-SWIR data using clay content information: from laboratory to field experiments[J]. Remote Sensing, 2015, 7(3): 3184-3205.

[6] 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.

[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] Zhang X L, Zhang F, Zhang H W, *et al*. Optimization of soil salt inversion model based on spectral transformation from hyperspectral index[J]. Transactions of the Chinese Society of Agricultural

Engineering, 2018, 34(1): 110-117.
张贤龙, 张飞, 张海威, 等. 基于光谱变换的高光谱指数土壤盐分反演模型优选[J]. 农业工程学报, 2018, 34(1): 110-117.

[9] Diao W Y, Liu G, Hu K L. Estimation of soil water content based on hyperspectral features and the ANN model[J]. Spectroscopy and Spectral Analysis, 2017, 37(3): 841-846.
刁万英, 刘刚, 胡克林. 基于高光谱特征与人工神经网络模型对土壤含水量估算[J]. 光谱学与光谱分析, 2017, 37(3): 841-846.

[10] Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 24-31.

[11] Khosravi V, Ardejani F D, Yousefi S, *et al*. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods[J]. Geoderma, 2018, 318: 29-41.

[12] Nawar S, Mouazen A M. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-NIR spectroscopy measurements of soil total nitrogen and total carbon[J]. Sensors, 2017, 17(10): 2428.

[13] Liang L, Di L P, Zhang L P, *et al*. Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method[J]. Remote Sensing of Environment, 2015, 165: 123-134.

[14] He B Z, Ding J L, Wang F, *et al*. Research on data mining of salinization information based on phenological characters[J]. Acta Ecologica Sinica, 2017, 37(9): 3133-3148.
何宝忠, 丁建丽, 王飞, 等. 基于物候特征的盐渍化信息数据挖掘研究[J]. 生态学报, 2017, 37(9): 3133-3148.

[15] Vohland M, Ludwig M, Thiele-Bruhn S, *et al*. Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection[J]. Geoderma, 2014, 223/224/225: 88-96.

[16] Tan K, Wang H M, Zhang Q Q, *et al*. An improved estimation model for soil heavy metal(loid) concentration retrieval in mining areas using reflectance spectroscopy[J]. Journal of Soils and Sediments, 2018, 18: 2008-2022.

[17] Tang G, Huang Y, Tian K D, *et al*. A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm[J]. Analyst, 2014, 139(19): 4894-4902.

[18] Wijewardane N K, Ge Y F, Morgan C L S. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization[J]. Geoderma, 2016, 267: 92-101.

[19] 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.

[20] Huang G, Huang G B, Song S J, *et al*. Trends in extreme learning machines: a review[J]. Neural Networks, 2015, 61: 32-48.

[21] Cheng S X, Kong W W, Zhang C, *et al*. Variety recognition of Chinese cabbage seeds by hyperspectral imaging combined with machine learning[J]. Spectroscopy and Spectral Analysis, 2014, 34(9): 2519-2522.
程术希, 孔汶汶, 张初, 等. 高光谱与机器学习相结合的大白菜种子品种鉴别研究[J]. 光谱学与光谱分析, 2014, 34(9): 2519-2522.

[22] Wang J Z, Ding J L, Abulimiti A, *et al*. Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS-NIR) spectroscopy, Ebinur Lake Wetland, Northwest China[J]. PeerJ, 2018, 6: e4703.

[23] Yu X, Liu Q, Wang Y B, *et al*. Evaluation of MLSR and PLSR for estimating soil element contents using visible/near-infrared spectroscopy in apple orchards on the Jiaodong Peninsula[J]. Catena, 2016, 137: 340-349.

[24] Li Z, Zhang F, Feng H K, *et al*. Research on the estimation of salt ions of vegetation leaves based on band combination[J]. Acta Optica Sinica, 2017, 37(11): 1128002.
李哲, 张飞, 冯海宽, 等. 基于波段组合的植被叶片盐离子估算研究[J]. 光学学报, 2017, 37(11): 1128002.