DOI: 10.13758/j.cnki.tr.2023.03.025
付平凡, 杨晓静, 苏志诚, 等. 基于集成学习的土壤含水量预测研究——以辽西地区为例. 土壤, 2023, 55(3): 671–681.

基于集成学习的土壤含水量预测研究——以辽西地区为例①

付平凡[1,2]，杨晓静[1,2*]，苏志诚[1,2]，屈艳萍[1,2]，马苗苗[1,2]

(1 中国水利水电科学研究院，北京 100038；2 水利部防洪抗旱减灾工程技术研究中心，北京 100038)

摘 要：准确高效地预测土壤含水量(SMC)对田间水分管理至关重要。本研究利用在辽西地区自建的 3 个站点 2018—2021 年 10 ~ 40 cm 土壤水分自动观测小时数据集，分析研究随机森林(random forest，RF)和梯度提升机(gradient boosting machine，GBM)算法在 SMC 预测方面的适用性，验证不同时间尺度 SMC 的预测结果。同时引入 SHAP(shapley additive explanations)方法表征 5 类(降水、日照时数、平均相对湿度、风速、平均气温)输入变量对 SMC 预测结果的影响，并制定区间划分规则识别变量最大贡献阈值区间。研究结果表明：年尺度下，SMC 预测 GBM 模型和 RF 模型 R^2 分别为 0.982 和 0.888，气温贡献最大，最大贡献区间是 21 ~ 23℃；季尺度下，2 种模型 R^2 分别为 0.935 和 0.863，日照时数贡献最大，最大贡献区间为 2 ~ 4 h。该研究创新应用 SHAP 方法于机器学习输入变量贡献度分析，同时验证了 2 种机器学习算法对 SMC 预测研究的准确性，可为 SMC 相关研究提供参考。

关键词：集成学习；土壤含水量预测；梯度提升机；随机森林；辽宁西部；SHAP 值
中图分类号：S152.7 **文献标志码：**A

Prediction of Soil Moisture Content Based on Ensemble Learning — A Case Study of Western Liaoning Province

FU Pingfan[1,2], YANG Xiaojing[1,2*], SU Zhicheng[1,2], QU Yanping[1,2], MA Miaomiao[1,2]
(1 *China Institute of Water Resources and Hydropower Research*, *Beijing* 100038, *China*; 2 *Research Center of Flood Control*, *Drought Relief*, *and Mitigation Engineering*, *Ministry of Water Resources*, *Beijing* 100038, *China*)

Abstract: Accurate and efficient prediction of soil moisture content (SMC) is vital for field water management. In this study, two types of ensemble learning models (RF and GBM) were used to compare their applicability in SMC prediction based on the automatic hourly SMC data at 10–40 cm during 2018—2021 from three self-built sites in the western Liaoning area, the prediction results were also compared and verified at annual and seasonal scales. The SHAP (Shapley Additive Explanations) method was introduced to quantitatively characterize the effects of five input variables (precipitation, sunshine hour, average relative humidity, wind speed and average temperature) on SMC prediction. Interval division rules were developed to identify the interval of maximum contribution threshold of variables. The results show that R^2 of GBM and RF models are 0.982 and 0.888 respectively on annual scale, temperature is the most important factor with the maximum contribution range of 21–23℃, while R^2 of the two models are 0.935 and 0.863 respectively on seasonal scale, sunshine hour is the most important factor with the maximum contribution range of 2–4 hours. This study innovatively applied SHAP method to analyze the contribution rates of input variables of machine learning, and verified the results of RF and GBM methods in SMC prediction, which can provide reference for related study on SMC.

Key words: Ensemble learning; Soil moisture content forecasting; Gradient boosting machine; Random forest; Western Liaoning; SHAP value

土壤水分是区域水循环、农业灌溉管理和气候变化的特征要素之一，其在水文、气象、农业等学科中也具有重要的作用[1]。土壤含水量(soil moisture content，SMC)是地表植被吸收水分的主要来源，其

①基金项目：江西省“科技+水利”联合计划项目(2022KSG01002)和中国水利水电科学研究院防洪抗旱减灾工程技术研究中心青年创新人才推进项目资助。
* 通讯作者(yxj@iwhr.com)
作者简介：付平凡(1998—)，男，河南信阳人，硕士研究生，主要从事干旱监测研究。E-mail: fupf123456@163.com

对作物的生长发育至关重要[2]。因此，准确预测土壤含水量对作物增产和粮食安全具有重要意义。

目前主要的土壤水分预测方法有经验模型法[3]、土壤水动力学法[4]、时间序列模型法[5]以及机器学习算法[6]等。近年来，随着计算机技术的快速发展，机器学习算法已成为一种重要的预测土壤含水量的手段[7]。集成学习(ensemble learning)是通过构建并结合多个机器学习器来完成任务，具有较强的泛化能力[8]。由于集成学习模型相比传统机器学习模型在性能上表现更为出色，目前集成学习中的随机森林(random forest，RF)和梯度提升机(gradient boosting machine，GBM)模型已在农业干旱监测、骤发性干旱研究等领域有所应用[9-10]。Cai 等[11]结合 GBM 与 RF 模型系统论证了这两种方法预测净生态系统碳交换的有效性；Prodhan 等[12]也将 RF 和 GBM 进行非线性集成，利用 ISI-MP 作物模型定量分析了未来干旱对作物产量的影响。以上研究都表明，RF 和 GBM 模型具有较好的实用性，但此类方法在土壤含水量预测的适用性研究上亟待进一步深入。

由于机器学习模型是黑箱模型，现有的多数研究主要基于评估指标来对模型的结果进行评价，而针对输入变量对预测结果影响的研究还相对较少。为解决这一问题，Lundberg 和 Lee[13]在 2017 年提出了 SHAP(shapley additive explanations)方法，该方法基于合作博弈理论定量化表征每个特征对最终预测值的影响，增加了模型的可解释性。近年来，已有研究利用 SHAP 方法解释机器学习模型，王鑫等[14]融合 LightGBM 模型与 SHAP 方法分析得出了血清胰岛素、葡萄糖浓度和年龄是患者是否患有糖尿病的关键因素；Kannangara 等[15]利用 RF 模型和 SHAP 方法，分析了 9 个输入变量对隧道开挖引起沉降的影响，结果表明土壤类型的影响最大。目前 SHAP 方法已应用于金融欺诈、污水处理、电力系统紧急控制等领域输入变量对预测结果的贡献研究[16-18]，但在土壤水分预测方面上的应用还相对较少。因此，将 SHAP 方法应用于土壤含水量预测研究，可定量识别输入变量对土壤含水量的贡献程度，为缺省输入因子情况下的变量选择提供依据。

本文拟将集成学习中的 RF、GBM 算法应用到土壤含水量预测研究，拓展验证 2 种算法在土壤含水量模拟预测中的适用性；且为解释各变量对于预测模型的影响，引入 SHAP 方法定量评估集成学习模型输入变量的贡献程度，并基于制定的区间划分规则识别特征敏感阈值区间，为解释输入变量对预测值的影响和土壤含水量预测方法的选择提供新的参考。

1 数据与方法

1.1 研究区概况

研究区为辽西地区(119.70° E ~ 122.53° E，40.35° N ~ 42.24° N)，包括阜新、朝阳、葫芦岛和锦州 4 市，属于温带大陆性季风气候，多年平均降水量约为 450 ~ 700 mm，其中夏季降水量约占全年降水量的 2/3。全年四季分明，雨热同期，日照丰富，年均温 7.2 ~ 8.3℃。根据 2021 年辽宁省统计年鉴的结果，辽西地区主要的粮食作物为玉米，占全省粮食作物种植面积的 41.2%。区域内含辽西走廊和辽西北部低山丘陵 2 个区域，地势呈现西北高、东南低的空间分布态势[19]。研究所选墒情站点均位于玉米种植区，其空间分布如图 1 所示。凌海站位于低山丘陵区的凌海市东部，土壤类型为棕壤；孙家湾站位于朝阳市东北部大凌河干流附近，地形为黄土丘陵，土壤类型为褐黄土；叶柏寿站位于朝阳市建平县南部，地形为丘陵坡地，土壤类型为褐黄土。

1.2 数据来源与质量控制

1.2.1 数据来源　研究应用的数据主要包括小时土壤含水量数据集和气象要素数据集。

1)小时土壤含水量数据集来源。2018 年 7 月在辽西地区选址(大凌河出口的凌海站，干流中部附近的孙家湾站，支流的叶柏寿站)并安装 3 套土壤墒情自动监测系统。该系统所使用的土壤水分传感器长期埋设在野外大田的测点中，并基于时域反射原理(time domain reflectometry，TDR)对不同深度土壤进行土壤体积含水量测定。站点的数据时间序列始于 2018 年 7 月，10 ~ 40 cm 深度传感器实时接收间隔为 1 h 的土壤墒情数据。为验证数据的有效性，分季节进行 7 次人工取土实验，利用烘干法将测定的土壤含水量与自动监测站监测结果进行对比，对比结果表明各深度土壤含水量同步监测差值小于 10%。

2)气象要素数据集来源。由于墒情站点的数据序列起始时间为 2018 年 7 月，为匹配对应日期的墒情数据，选择 2018—2021 年气象数据作为模型输入变量，气象数据来源于中国气象数据网(http://data.cma.cn/)。3 个气象站点气象要素包括逐日的降水、日照时数、平均相对湿度、风速、平均气温。

1.2.2 数据质量控制　为降低异常数据对模型预测结果准确性的扰动，从两个方面对数据进行质量控制。

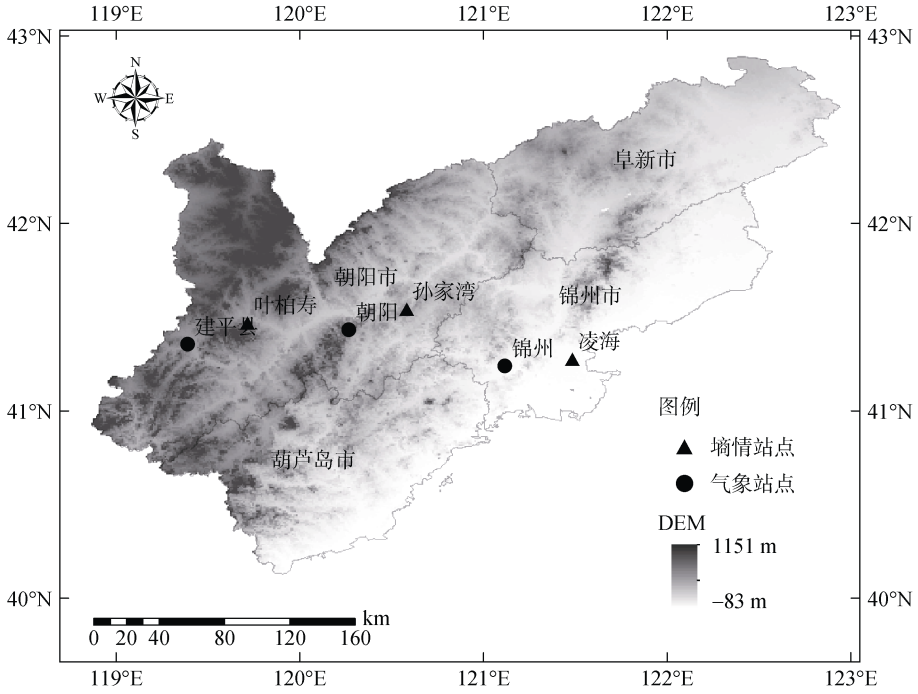

图 1　研究区和 3 个试验站点示意图

Fig. 1　Location of study area and three experimental stations

1)数据有效性控制。为保证数据集的有效性，将墒情站和气象站空值数据剔除后，孙家湾站共有数据 1 099 条，叶柏寿站共有数据 1 202 条，凌海站共有数据 1 177 条。

2)数据量纲控制。为避免不同输入变量之间数量级别和量纲的影响，将输入和输出数据进行归一化处理，计算公式如下：

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

式中：x^*为归一化后的样本集；x 为原始样本集；$x_{\min}$ 为原始样本集最小值，$x_{\max}$ 为原始样本集最大值。将每日 8:00 的土壤墒情数据作为当日值，并按照 8∶2 的分配原则进行训练集和测试集的划分。

1.3　研究方法

1.3.1　集成学习　　集成学习是一种融合多个机器学习模型的集成模型，通过某种融合策略常可获得比单一模型显著优越的泛化性能。集成学习不仅能够实现模型之间的优势互补，还能减少对训练所需数据的依赖程度[20]。常用的融合策略有 3 种：Bagging、Boosting 和 Stacking。本研究采用 Bagging 中的 RF、Boosting 中的梯度提升决策树(GBDT)算法，所使用的 2 种方法的建模过程均在 Python 语言环境下加载 scikit-learn 实现。

梯度提升机(gradient boosting machine，GBM)是由 Friedman[21]提出的一种流行机器学习的集成方法。为了解决回归和分类问题，GBM 通常是以决策树弱模型组合的形式，周期性地构造出一个鲁棒模型。Gradient Boosting 与一般的 Boosting 算法一样，也是一个迭代的过程，Gradient Boosting 每个新的模型是沿着前面模型的残差减少的梯度方向上建立，每次的训练是为了改进上一次的回归结果。为了减少模型的残差(residual)，通常采用牛顿–拉弗森方法(Newton-Raphson method)在残差减少的梯度(Gradient)方向拟合一个新的模型[22]。由 GBM 构建的梯度提升回归模型有 5 个需要优化的参数，分别为学习率(learning_rate)、损失函数(loss)、决策树的数量(n_estimators)、决策树的深度(max_depth)和建立决策树时选择的最大特征数目(max_features)。利用 GridSearchCV 方法[23]进行超参数随机匹配择优，经过调参后，n_estimators=300，max_depth=10，max_features=2，loss='huber'函数，learning_rate=0.1 为最优参数。

随机森林(random forest，RF)算法是一种通过集成大量的决策树来改进分类和回归的方法。Breiman[24]引入的 RF 是一种基于 bootstrap 聚合的决策树集合，通过随机选取广泛应用于回归问题的预测器子集，计算预测变量并基于预测变量的数据分割，得到因变量的均方根误差(RMSE)最佳估计。在 RF 回归中，引入的 RF 算法将自动创建随机决策树群，通过从训练数据集中选择随机变量集，并采用随机有

放回抽样的方法来构建每棵树，最后通过对所有树的均衡化结果来计算观测值的预测值。RF 模型有 3 个需要优化的参数：决策树的数量(n_estimators)、决策树的深度(max_depth)和建立决策树时选择的最大特征数目(max_features)。利用 GridSearchCV 方法进行超参数随机匹配择优，经过调参后，n_estimators=900，max_depth=15，max_features=5 是最优参数。

1.3.2 模型评价指标　选用平均绝对误差(MAE)、决定系数(R^2)[25]和均方根误差(RMSE)3 种指标分别对 GBM、RF 预测模型进行预测效果评估。评价指标计算公式如下：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\left(y_i-\hat{y}_i\right)\right| \tag{2}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{3}$$

$$R^2=1-\frac{\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2}{\sum_{i=1}^{n}\left(\bar{y}_i-y_i\right)^2} \tag{4}$$

式中：$\hat{y}_i$是土壤含水量预测值；y_i是真实值；$\bar{y}_i$是平均值。MAE 是绝对误差的平均值，它能够反映预测值误差的实际情况。RMSE 是含水量估计值与真值之差的平方的期望值，可以评价数据的变化程度。R^2可以消除维数对评价测度的影响。MAE 和 RMSE 越小表明预测结果越好，R^2越大表明预测结果越好。

1.3.3 SHAP 方法　SHAP 方法是一种直观的、合理的解释模型的方法，该方法通过计算每个特征对预测值的贡献来解释特征，所使用的值(SHAP 值)可定量化表征各个特征对预测值的贡献，SHAP 值越大表明该特征对于预测值的贡献越大。SHAP 方法是以合作博弈理论为基础计算 SHAP 值，特征值的 SHAP 值是对所有可能的特征值组合进行加权求和，其公式如下：

$$\phi_j(\mathrm{val})=\sum_{S\subseteq\{1,\cdots,p\}\setminus\{j\}}\frac{|S|!(p-|S|-1)!}{p!}\left(\mathrm{val}\left(S\cup\{j\}\right)-\mathrm{val}(S)\right) \tag{5}$$

式中：S 是模型中使用的特征的子集，表示 j 这些特征不包括在集合 S 中；p 是特征的数量；val(S)是对集合 S 中特征值的预测；ϕ_j表示 val 第 j 个特征的贡献。

2 结果与分析

2.1 年尺度预测结果对比

选取 2018—2021 年土壤 10～40 cm 深度含水量数据进行训练，基于 RF、GBM 算法构建土壤含水量预测模型。对比 2 种模型测试集的预测结果(表 1)发现，10～40 cm 深度预测精度相差较小，R^2差值都在 0.1 以内。GBM 模型预测精度较高，10～40 cm 深度 R^2值均大于 0.94，MAE 和 RMSE 均值均小于 0.006 和 0.026；RF 模型预测精度略差，10 ～ 40 cm 深度 R^2均值范围为 0.881～0.891，MAE 和 RMSE 均值均小于 0.054 和 0.071。

表 1　年尺度下不同站点不同深度土壤含水量 RF 和 GBM 模型预测精度比较

Table 1　Prediction accuracies of soil moisture by RF and GBM models at different depths in different stations on annual scale

站点	土深(cm)	测试集个数	RF			GBM		
			MAE	RMSE	R^2	MAE	RMSE	R^2
孙家湾	10	220	0.048	0.061	0.887	0.004	0.017	0.992
	20	220	0.052	0.066	0.885	0.005	0.028	0.979
	30	220	0.042	0.059	0.869	0.007	0.029	0.968
	40	220	0.074	0.097	0.881	0.006	0.027	0.991
	均值	220	0.054	0.071	0.881	0.006	0.025	0.983
叶柏寿	10	241	0.041	0.054	0.902	0.005	0.021	0.985
	20	241	0.058	0.075	0.882	0.007	0.029	0.982
	30	241	0.061	0.080	0.892	0.008	0.036	0.978
	40	241	0.041	0.055	0.889	0.005	0.018	0.988
	均值	241	0.05	0.066	0.891	0.006	0.026	0.983
凌海	10	236	0.045	0.060	0.886	0.007	0.042	0.945
	20	236	0.032	0.043	0.924	0.004	0.015	0.990
	30	236	0.045	0.064	0.882	0.004	0.017	0.992
	40	236	0.060	0.079	0.871	0.005	0.023	0.989
	均值	236	0.046	0.062	0.891	0.005	0.024	0.979

以孙家湾站为例，2 种模型 10～40 cm 深度测试集土壤含水量样本预测值与实测值基本都在 1∶1 线附近，R^2 值均超过 0.86，GBM 模型的预测值明显更加接近实测值，如图 2 所示。综上所述，对比 2 种模型方法的评价指标，RF 模型和 GBM 模型年尺度下均能精准地预测土壤含水量，但 GBM 模型表现更佳。

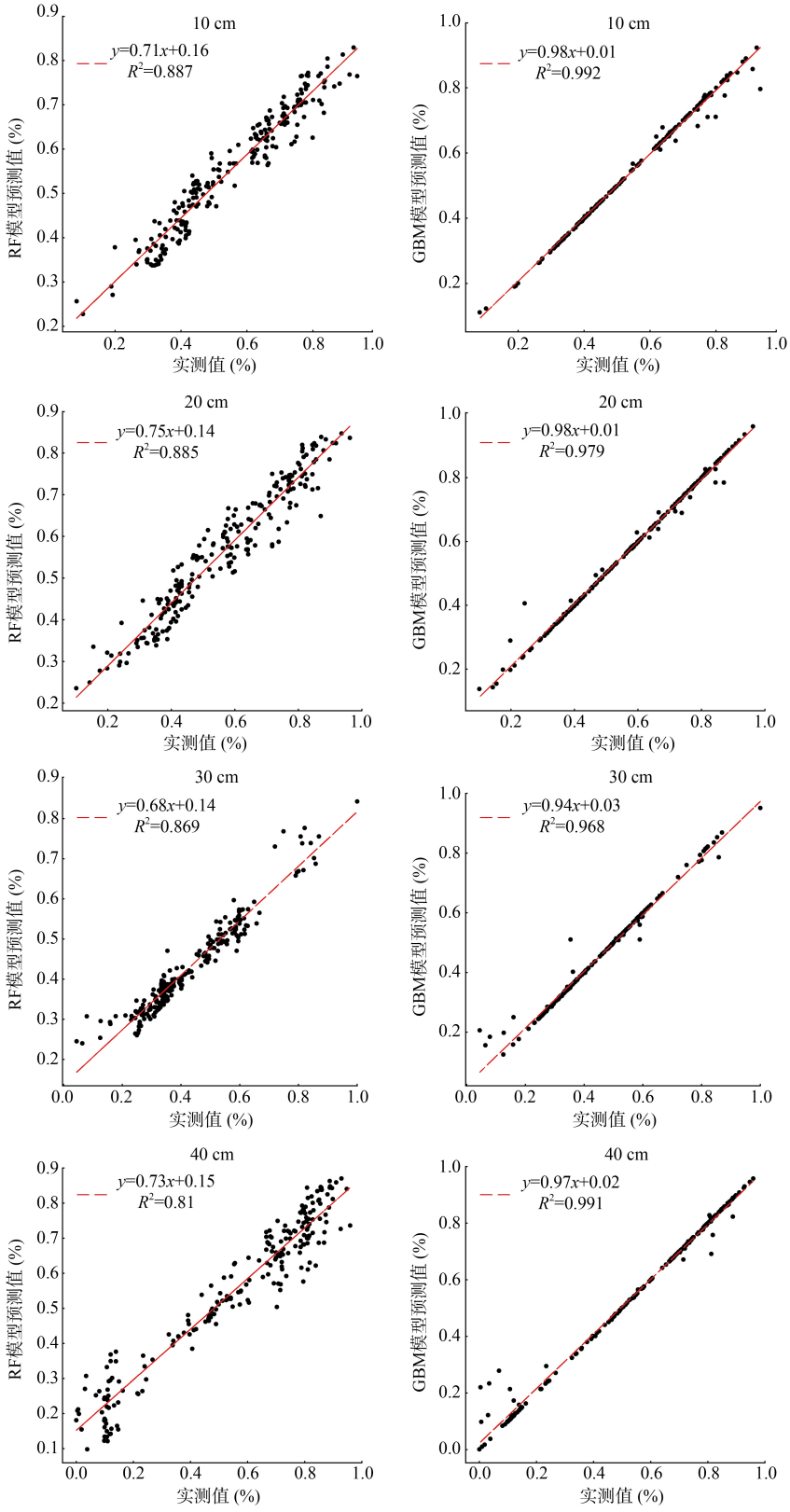

图 2　孙家湾站 10～40 cm 深度土壤含水量预测值与实测值

Fig. 2　Predicted and measured soil moisture values in 10 – 40 cm depth at Sunjiawan Station

2.2 季节尺度预测结果对比

本研究中，季节划分标准为：春季 3—5 月、夏季 6—8 月、秋季 9—11 月、冬季 12 月—次年 2 月。季节尺度的预测结果(表 2)表明，GBM 模型和 RF 模型在不同季节预测土壤含水量均具有较高的精度。GBM 模型 10 ~ 40 cm 深度各季节 R^2 均值范围为 0.931 ~ 0.938，MAE 值均小于 0.026，RMSE 值均小于 0.065；RF 模型 10 ~ 40 cm 深度各季节 R^2 值范围为 0.816 ~ 0.894，MAE 值均小于 0.073，RMSE 值均小于 0.095，精度略低于 GBM 模型。

表 2 不同季节土壤含水量 2 种模型预测精度比较

Table 2 Prediction accuracies of soil moisture by RF and GBM models in different stations in seasonal scale

站点	季节	测试集个数	RF			GBM		
			MAE	RMSE	R^2	MAE	RMSE	R^2
孙家湾	春季	48	0.068	0.084	0.894	0.025	0.052	0.962
	夏季	56	0.058	0.078	0.876	0.021	0.048	0.955
	秋季	57	0.036	0.046	0.889	0.012	0.025	0.966
	冬季	59	0.054	0.076	0.816	0.025	0.064	0.868
	均值	55	0.054	0.071	0.869	0.021	0.047	0.938
叶柏寿	春季	55	0.048	0.067	0.878	0.020	0.045	0.945
	夏季	66	0.068	0.085	0.874	0.014	0.031	0.981
	秋季	65	0.054	0.072	0.845	0.022	0.049	0.924
	冬季	57	0.055	0.074	0.834	0.024	0.062	0.873
	均值	61	0.056	0.075	0.858	0.020	0.047	0.931
凌海	春季	48	0.056	0.076	0.864	0.023	0.045	0.952
	夏季	57	0.072	0.094	0.835	0.024	0.048	0.954
	秋季	72	0.048	0.063	0.862	0.015	0.040	0.923
	冬季	59	0.064	0.084	0.883	0.025	0.064	0.921
	均值	59	0.060	0.079	0.861	0.022	0.049	0.938

对比分析多时间尺度模型预测结果表明：GBM 模型和 RF 模型在年、季尺度下均有较好的预测结果(R^2 均大于 0.816)，GBM 模型的预测精度略高(R^2 均大于 0.868)。年尺度上，2 种模型在 3 个站点不同土层的 R^2 均值皆大于 0.881，RMSE 均值皆小于 0.071，MAE 均值皆小于 0.054；季节尺度上，2 种模型在春季、夏季和秋季的预测结果则更好，春季、夏季和秋季 R^2 均大于 0.835，RMSE 均小于 0.094，MAE 均小于 0.072。

2.3 特征要素贡献度分析

为探究不同时间尺度、不同深度各输入特征要素对预测的土壤含水量的贡献度，将预测结果较好的 GBM 模型与 SHAP 方法进行融合。分别计算年、季尺度下降水、日照时数、平均相对湿度、风速、平均气温这 5 个输入变量的 SHAP 值，并基于 SHAP 值大小判断输入特征对土壤含水量的贡献。

年尺度上，孙家湾站、叶柏寿站和凌海站 10 ~ 40 cm 深度特征要素贡献排序基本一致，均是平均气温贡献最大，降水贡献最小。其中叶柏寿站 10、20 和 40 cm 深度的特征要素贡献排序从高到低分别为平均气温、日照时数、相对湿度、风速和平均气温；30 cm 深度则为平均气温、相对湿度、日照时数、风速和平均气温，如图 3 所示。孙家湾站和凌海站特征要素贡献排序与叶柏寿站一致。

为对比 4 个不同深度、不同季节 5 个输入特征对预测土壤含水量的整体贡献度，利用特征的 SHAP 值之和(整体 SHAP 值)来对比不同深度和季节的结果。

年尺度上，4 个土层深度输入的 5 个气象要素对于预测 10 cm 和 20 cm 土层的土壤含水量贡献较大，且更适用于预测 20 cm 深度的土壤含水量。孙家湾、叶柏寿和凌海站 10 cm 和 20 cm 深度的整体 SHAP 值分别为 7.99、8.07 和 7.98，比 30 cm 和 40 cm 深度分别增加了 10.66%、12.08% 和 1.01%。各站点 20 cm 深度输入变量的整体 SHAP 值分别为 8.48、8.43 和 8.53，比 10 cm 深度分别增加了 12.96%、14.48% 和 14.79%，其中叶柏寿站 SHAP 值如图 3 所示。

由于年尺度 20 cm 土层整体 SHAP 值最高，因此季节尺度上选择 20 cm 土层为代表性土层进行分析。孙家湾站和叶柏寿站贡献最大的特征要素是日照时

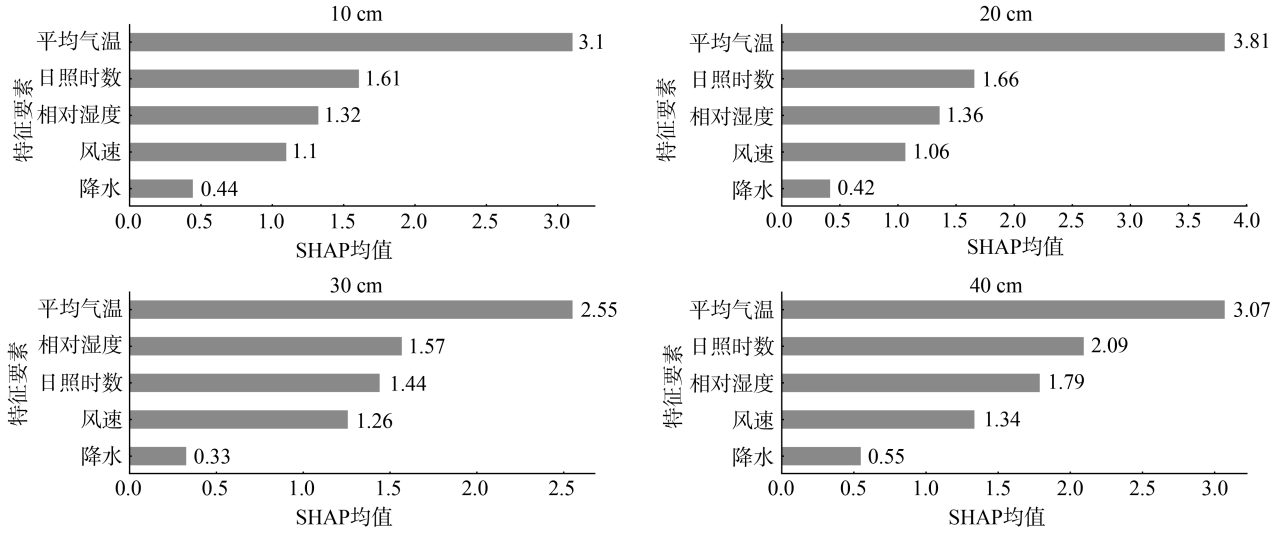

图 3　叶柏寿站 10 ~ 40 cm 土壤预测含水量的特征要素贡献分布

Fig. 3　Characteristic variable contribution to predicted soil moisture in 10–40 cm depth at Yebaishou Station

数，凌海站为平均气温，所有站点对预测结果贡献最低的变量均为降水。夏季整体 SHAP 值高于其余 3 个季节，其中叶柏寿站春季、秋季和冬季整体 SHAP 值相比夏季分别降低 36.5%、18.8% 和 46.2%；凌海站分别降低 47.9%、10.8% 和 5.8%；孙家湾站春季整体 SHAP 值比夏季增加 2.6%，秋季和冬季分别降低 39.6% 和 26.6%，具体结果见表 3。

在年、季尺度上降水贡献均最低，可能有以下两个方面的原因：①辽西地区年降水量区间为 400 ~ 700 mm，且全年 2/3 的降水集中在夏季。孙家湾、叶柏寿和凌海站夏季降水量分别为 319.3、357.2 和 474.43 mm；②无降水日数占比较高。孙家湾、叶柏寿和凌海站年内无降水日数的数据占比分别为 81.1%、79.4% 和 79.5%；尽管降水集中在夏季，但无雨日数仍高于 60%(孙家湾、叶柏寿和凌海站占比分别为 61.8%、62.0% 和 62.3%)。

表 3　20 cm 深度土壤预测含水量不同季节特征要素贡献统计

Table 3　Characteristic variable contribution to predicted soil moisture at 20 cm depth on seasonal scale

站点	季节	SHAP 值					
		降水	日照时数	相对湿度	风速	平均气温	整体 SHAP 值
孙家湾	春季	0.42	2.73	2.54	0.75	1.87	8.30
	夏季	0.81	3.10	1.36	1.03	1.80	8.09
	秋季	0.47	0.83	1.17	0.63	1.80	4.89
	冬季	0.19	1.75	1.72	0.82	1.48	5.94
叶柏寿	春季	0.20	1.64	1.54	0.81	1.38	5.56
	夏季	0.77	2.58	2.52	1.34	1.55	8.76
	秋季	0.82	1.71	1.43	0.89	2.29	7.14
	冬季	0.09	1.02	1.53	0.62	1.45	4.71
凌海	春季	0.17	0.45	0.51	0.39	1.07	2.58
	夏季	0.69	1.16	1.34	0.75	1.46	5.39
	秋季	0.34	0.94	1.17	0.70	1.67	4.81
	冬季	0.11	1.42	1.55	0.75	1.24	5.08

集成学习预测结果的准确性与样本数量和数值变化区间成正比，因此在相同数量样本条件下，较多的无降水日数可使降水贡献小于其他要素。虽然 5 个特征中降水的贡献最低，但对比季节贡献结果可以看出，降水对土壤含水量的贡献度与降水量成正比关系。辽西地区夏季降水最多，贡献度也是四季最高，如图 4 所示。

综上所述，辽西地区降水对土壤含水量贡献较低的主要原因是年内、季节内降水分布不均。已有的研究也表明，降水对于土壤含水量的贡献较小。Clewley

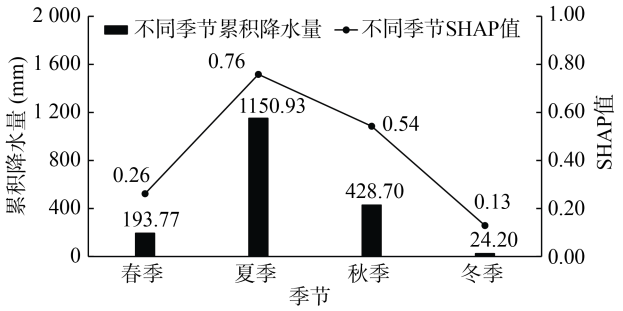

图 4　2018—2021 年不同季节累积降水量和 SHAP 值

Fig. 4　Cumulative precipitation and SHAP values in different seasons from 2018 to 2021

等[26]利用 RF 分析了高程、坡度和降水等对土壤水分的影响，结果表明降水影响最小；Karthikeyan 和 Mishra[27]利用 XGBoost 算法分析了海拔、土壤质地、归一化植被指数(NDVI)和降水对于土壤水分的影响，结果表明降水影响最小。以上研究利用了不同的算法、输入了不同的变量来预测土壤含水量，但结果都表明降水对于土壤水分的影响最小。

为定量识别不同输入特征要素对应的有效阈值区间，制定区间识别划分规则为：①筛选出 SHAP 值大于 0 的点，提取点所在的区间；②将区间等分，分别计算每个区间 SHAP 均值；③比较划分后的区间与原区间 SHAP 均值的大小，最终定量识别不同输入特征对土壤含水量贡献最大的区间。各个特征 SHAP 值大于 0 的区间分别为降水 0 ~ 10 mm、日照时数 0 ~ 8 h、相对湿度 60% ~ 80%、风速 1 ~ 3 m/s、气温 22 ~ 24℃，如图 5 所示。

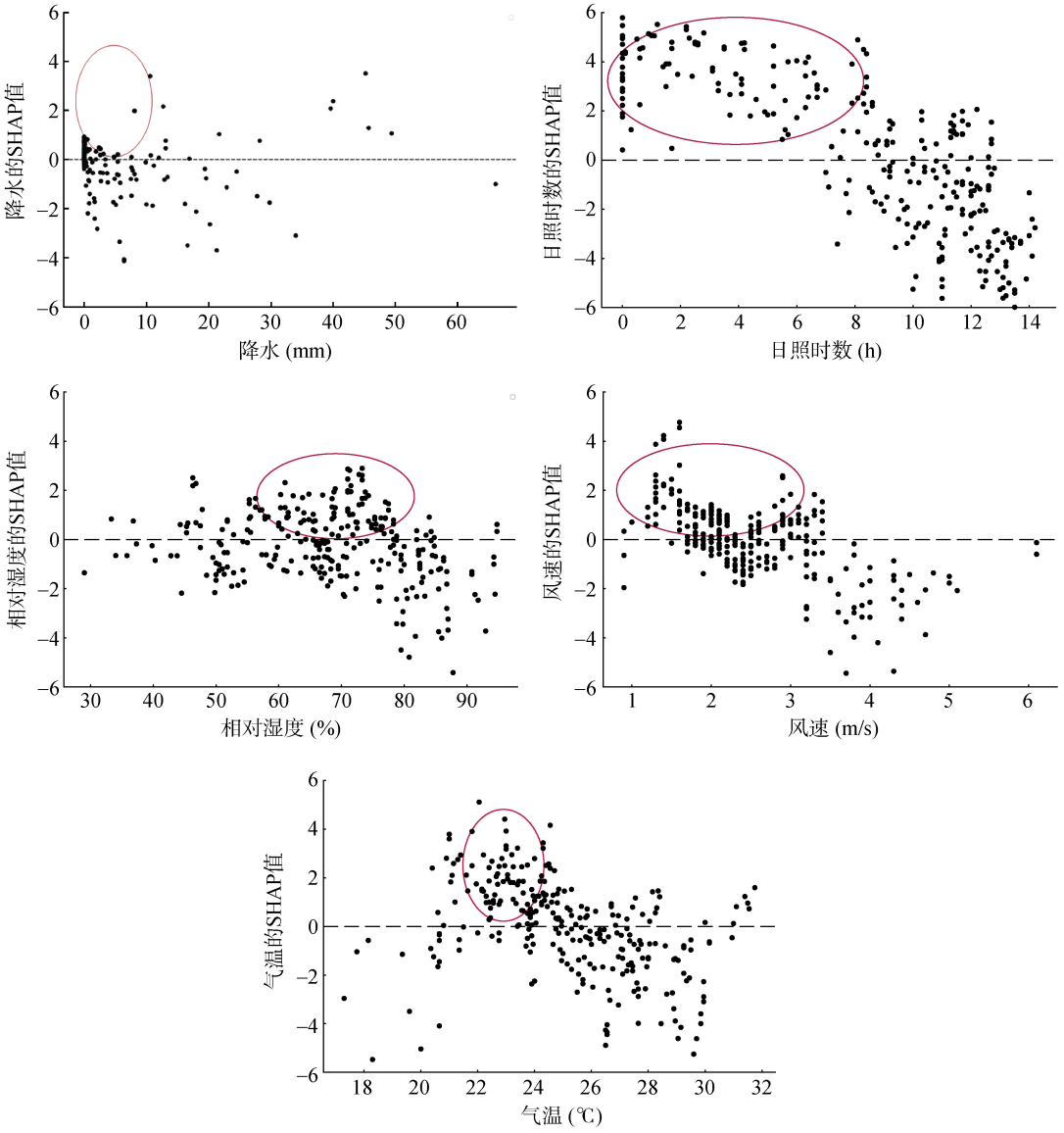

图 5　特征要素贡献依赖图

Fig. 5　Dependence plot of characteristic variable contribution

孙家湾站和叶柏寿站对土壤含水量贡献最大的特征区间一致，都是日照时数贡献最大，最大区间为 2 ~ 4 h；降水贡献最小，贡献最大区间为 0 ~ 5 mm。凌海站却是降水贡献最大，贡献最大的区间为 5 ~ 10 mm；风速贡献最小，贡献最大的区间为 1 ~ 2 m/s。各站点特征区间 SHAP 均值见表 4。

表 4　3 站点各个特征要素区间 SHAP 均值
Table 4　Mean SHAP values at different intervals of characteristic variables in 3 stations

特征	区间	孙家湾	叶柏寿	凌海
降水(mm)	[0,5)	**0.17**	**–0.08**	–0.14
	[5,10]	–1.12	–1.94	**1.58**
	[0,10]	0.07	–0.20	–0.02
日照时数(h)	[0,2)	3.77	5.28	**1.10**
	[2,4]	**3.78**	**5.34**	–1.10
	[0,4]	3.32	3.61	0.34
相对湿度(%)	[60,70)	0.22	–0.16	–1.13
	[70,80]	**0.34**	**1.67**	**0.63**
	[60,80]	0.27	1.03	0.08
风速(m/s)	[1,2)	**0.91**	**2.08**	**0.12**
	[2,3]	0.15	–0.84	–0.27
	[1,3]	0.40	0.02	–0.08
平均气温(℃)	[22,23)	**1.59**	**0.41**	**1.20**
	[23,24]	1.10	–0.23	0.20
	[22,24]	1.30	0.10	0.62

注：表中加粗部分即各个特征要素贡献最大值，所在区间为贡献最大区间。

3　讨论

本研究结果表明，GBM 模型在年、季尺度上的预测精度均高于 RF 模型(R^2 均大于 0.816)。融合 SHAP 方法的 GBM 模型不仅定量计算出了不同土层、不同季节输入变量对土壤含水量的贡献，而且基于区间划分规则识别了特征最大贡献区间。为验证该模型在预测土壤含水量上的优势，从以下两方面进行讨论。

1)与神经网络中最常用的多层感知机(MLP)进行对比验证。目前利用神经网络预测土壤含水量的研究较多[28]，为了验证 GBM 模型和 RF 模型在预测土壤含水量上的优势，采用神经网络中最常用的 MLP 模型进行对比验证。利用 optuna 方法[29]对 MLP 进行 10 次参数择优后，最终确定神经网络隐藏层分别为 70、60 和 20，最优参数分别为 activation='relu'，solver='lbfgs'，max_iter=1400，alpha=0.04。基于参数优选后的结果预测各站点不同深度土壤含水量结果，3 个站点的 MAE 介于 0.065 ~ 0.110，RMSE 介于 0.086 ~ 0.146，R^2 介于 0.423 ~ 0.871，具体结果见表 5。MLP 模型的预测精度明显低于本研究中构建的 2 种土壤含水量预测模型，GBM 模型在 3 个站点的 R^2 均值分别提升了 0.226、0.176 和 0.459；RF 模型在 3 个站点的 R^2 均值分别提升了 0.124、0.084 和 0.371。上述结果表明，本研究使用的集成学习模型相较于 MLP 模型具有显著的优势。

2)与国内外同类研究结果对比。已有的研究也表明，GBM 模型和 RF 模型在土壤含水量预测方面拥有更加良好的表现。Chen[8]等基于 RADARSAT-2 和 Sentinel-2 数据，使用支持向量回归机(SVR)、RF 和梯度提升决策树(GBDT)这 3 种机器学习方法在加拿大安大略省西南部对冬小麦种植区 0 ~ 5 cm 土壤水分进行预测，结果表明，RF 模型结果最优(R^2 为 0.94)，GBMT 模型次之(R^2 为 0.77)，SVR 模型结果最差(R^2 为 3.06)。

表 5　MLP 模型对 10 ~ 40 cm 深度土壤含水量的预测结果
Table 5　Prediction accuracies of soil moisture in 10–40 cm depth by MLP model

站点	土深 (cm)	MLP		
		MAE	RMSE	R^2
孙家湾	10	0.098	0.124	0.690
	20	0.079	0.102	0.759
	30	0.073	0.094	0.709
	40	0.074	0.107	0.871
	均值	0.081	0.107	0.757
叶柏寿	10	0.065	0.086	0.832
	20	0.080	0.108	0.803
	30	0.091	0.122	0.779
	40	0.071	0.097	0.815
	均值	0.077	0.103	0.807
凌海	10	0.108	0.140	0.423
	20	0.096	0.128	0.461
	30	0.096	0.121	0.506
	40	0.110	0.146	0.689
	均值	0.103	0.134	0.520

目前针对特征贡献的研究还相对较少。Clewley 等[26]采集了位于加拿大马尼托巴省南部 SMAP 实验点 2012 年 6—7 月间 13 d 现场数据，利用 RF 算法计算了各输入特征的重要度，结果表明，贡献最大的变量是高程，贡献最小的变量是降水；Cai 等[30]分析了各输入特征与土壤含水量的相关性，结果表明，相对湿度相关性最大，降水相关性最小。本研究利用 SHAP 方法不仅判断出不同时间尺度下最大贡献的特征要素，而且制定了区间划分规则识别输入特征最

大贡献区间，从方法应用范围上进行了提升和改进。

4 结论

本文基于集成学习 Bagging 中的随机森林(RF)、Boosting 中的梯度提升机(GBM)，研究了 2 种算法在辽西地区预测土壤含水量的适用性。在土壤含水量预测的基础上，引入 SHAP 方法定量计算输入特征变量对土壤含水量的贡献，并基于制定的区间划分规则识别特征最大贡献的阈值范围，实现了高精度可解释的土壤水分预测。

1)从模型适用性方面，GBM 模型更适合辽西地区的土壤含水量预测。年、季尺度下，GBM 模型和 RF 模型均适用于辽西地区土壤含水量预测。年尺度下 GBM 模型和 RF 模型 10 ~ 40 cm 深度 R^2 分别为 0.982、0.888；季节尺度下 R^2 分别为 0.935、0.863。

2)对比分析降水、日照时数、平均相对湿度、风速、平均气温 5 个输入要素的贡献度，气温和日照时数贡献较大，其中气温贡献最大范围 21 ~ 23℃；日照时数贡献最大范围为 2 ~ 4 h。年尺度下，气温贡献最大，降水贡献最小；季节尺度下，夏季对于土壤含水量预测的贡献最大，贡献最大的特征要素为日照时数，贡献最小的特征要素为降水。

3)与传统的 MLP 模型结果相比，GBM 模型和 RF 模型 10 ~ 40 cm 深度土壤含水量的预测结果均优于 MLP 模型。孙家湾站、叶柏寿站和凌海站 GBM 模型的 R^2 均值相较于 MLP 模型分别提升了 0.226、0.176 和 0.459，RF 模型较之提升了 0.124、0.084 和 0.371。

4)本研究首次将集成学习算法中的 GBM 模型和 RF 模型应用到辽西地区的土壤含水量预测，验证了 2 种模型在年、季尺度上的有效性。创新引入 SHAP 方法，定量化表征输入特征要素贡献度，并基于制定的区间划分规则计算了区间 SHAP 均值，识别了输入特征最大贡献区间，可为其他地区的土壤含水量预测研究提供新的参考与借鉴。

参考文献：

[1] Zhang D J, Zhou G Q. Estimation of soil moisture from optical and thermal remote sensing: A review[J]. Sensors (Basel, Switzerland), 2016, 16(8): 1308.

[2] 程谅, 焦雄, 邸涵悦, 等. 不同整地措施坡面土壤水分时空分布特征[J]. 土壤学报, 2021, 58(6): 1423–1435.

[3] Hummel J W, Sudduth K A, Hollinger S E. Soil moisture and organic matter prediction of surface and subsurface soils using an NIR soil sensor[J]. Computers and Electronics in Agriculture, 2001, 32(2): 149–165.

[4] 周良臣. 利用土壤水动力学模型预测麦田土壤水分的研究[J]. 节水灌溉, 2007(3): 10–13, 17.

[5] 白冬妹, 郭满才, 郭忠升, 等. 时间序列自回归模型在土壤水分预测中的应用研究[J]. 中国水土保持, 2014(2): 42–45, 69.

[6] 聂红梅, 杨联安, 李新尧, 等. 基于 PCA-SVR 的冬小麦土壤水分预测[J]. 土壤, 2018, 50(4): 812–818.

[7] Padarian J, Minasny B, McBratney A B. Machine learning and soil sciences: A review aided by machine learning tools[J]. SOIL, 2020, 6(1): 35–52.

[8] Chen L, Xing M F, He B B, et al. Estimating soil moisture over winter wheat fields during growing season using machine-learning methods[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 3706–3718.

[9] Zhang L Q, Liu Y, Ren L L, et al. Analysis of flash droughts in China using machine learning[J]. Hydrology and Earth System Sciences, 2022, 26(12): 3241–3261.

[10] Feng P Y, Wang B, Liu L D, et al. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia[J]. Agricultural Systems, 2019, 173: 303–316.

[11] Cai J C, Xu K, Zhu Y H, et al. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest[J]. Applied Energy, 2020, 262: 114566.

[12] Prodhan F A, Zhang J H, Sharma T P P, et al. Projection of future drought and its impact on simulated crop yield over South Asia using ensemble machine learning approach[J]. Science of the Total Environment, 2022, 807: 151029.

[13] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4 - 9, 2017, Long Beach, California, USA. New York: ACM, 2017: 4768–4777.

[14] 王鑫, 廖彬, 李敏, 等. 融合 LightGBM 与 SHAP 的糖尿病预测及其特征分析方法[J]. 小型微型计算机系统, 2022, 43(9): 1877–1885.

[15] Kannangara K K, Zhou W H, Ding Z, et al. Investigation of feature contribution to shield tunneling-induced settlement using Shapley additive explanations method [J]. Journal of Rock Mechanics and Geotechnical Engineering, 2002, 14(4): 1052–1063.

[16] Zhang K, Xu P D, Zhang J. Explainable AI in deep reinforcement learning models: A SHAP method applied in power system emergency control[C]//2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2). October 30 - November 1, 2020, Wuhan, China. IEEE, 2021: 711–716.

[17] Wang D, Thunéll S, Lindberg U, et al. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based

machine learning methods[J]. Journal of Environmental Management, 2022, 301: 113941.
[18] Fukas P, Rebstadt J, Menzel L, et al. Towards explainable artificial intelligence in financial fraud detection: Using shapley additive explanations to explore feature importance[C]//Advanced Information Systems Engineering: 34th International Conference, CAiSE 2022, Leuven, Belgium, June 6-10, 2022, Proceedings. New York: ACM, 2022: 109–126.
[19] 王笑歌. 辽西地区干旱评价及预测研究[D]. 沈阳: 沈阳农业大学, 2019.
[20] 余东行, 张保明, 赵传, 等. 联合卷积神经网络与集成学习的遥感影像场景分类[J]. 遥感学报, 2020, 24(6): 717 场 727.
[21] Friedman J H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189–1232.
[22] 万伦军. 基于梯度提升模型的负相关学习算法的研究与应用[D]. 合肥: 中国科学技术大学, 2014.
[23] Memon N, Patel S B, Patel D P. Comparative analysis of artificial neural network and XGBoost algorithm for PolSAR image classification[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 452–460.
[24] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5–32.
[25] 蔡庆空, 李二俊, 陶亮亮, 等. 基于改进作物散射模型的陕西杨凌区麦田土壤水分反演研究[J]. 土壤, 2020, 52(4): 846–852.
[26] Clewley D, Whitcomb J B, Akbar R, et al. A method for upscaling *in situ* soil moisture measurements to satellite footprint scale using random forests[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(6): 2663–2673.
[27] Karthikeyan L, Mishra A K. Multi-layer high-resolution soil moisture estimation using machine learning over the United States[J]. Remote Sensing of Environment, 2021, 266: 112706.
[28] 范嘉智, 谭诗琪, 罗宇, 等. 长短期记忆神经网络在多时次土壤水分动态预测中的应用[J]. 土壤, 2021, 53(1): 209–216.
[29] Akiba T, Sano S, Yanase T, et al. Optuna: A next-generation hyperparameter optimization framework[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. August 4 - 8, 2019, Anchorage, AK, USA. New York: ACM, 2019: 2623–2631.
[30] Cai Y, Zheng W G, Zhang X, et al. Research on soil moisture prediction model based on deep learning[J]. PLoS One, 2019, 14(4): e0214508.