

基于机器学习算法的棉田土壤钾、钠离子量预测

唐茂淞^{1,2}, 张楠^{1,2}, 李国辉^{1,2}, 赵泽艺^{1,2}, 李明发³, 王兴鹏^{1,2,4*}

(1.塔里木大学 水利与建筑工程学院, 新疆 阿拉尔 843300; 2.塔里木大学 现代农业工程重点实验室, 新疆 阿拉尔 843300; 3.新疆生产建设兵团第一师水文水资源管理中心, 新疆 阿拉尔 843300; 4.农业农村部西北绿洲节水农业重点实验室, 新疆 石河子 832000)

摘要:【目的】比较4种机器学习方法对南疆棉田土壤 K^+ 、 Na^+ 量的预测结果, 确定一种预测准确度较高的机器学习模型作为可供参考的方法。【方法】采用支持向量回归(SVR)、随机森林回归(RFR)、K-最近邻回归(KNNR)和梯度提升回归树(GBRT)4种机器学习算法, 2020年棉田土壤 K^+ 、 Na^+ 量数据训练模型, 2021年实测数据进行测试验证。使用平均绝对误差(MAE)、均方根误差(RMSE)和决定系数(R^2)对模型预测结果进行评估。【结果】4种模型(SVR、RFR、KNNR和GBRT)对测试样本 K^+ 量预测的MAE分别为0.100、0.169、0.169 g/kg和0.167 g/kg; RMSE分别为0.119、0.218、0.218 g/kg和0.223 g/kg; R^2 分别为0.687、0.437、0.430和0.395。对测试样本 Na^+ 量预测的MAE分别为0.841、2.841、2.826 g/kg和2.856 g/kg; RMSE分别为1.154、3.658、3.630 g/kg和3.650 g/kg; R^2 分别为0.838、0.299、0.219和0.200。将测试样本 K^+ 、 Na^+ 量分别按4个土层深度(0~10、10~20、20~30 cm和30~40 cm)进行预测时, SVR模型的误差值最小, 其对 K^+ 量按照4个深度预测的MAE分别为0.122、0.114、0.056 g/kg和0.106 g/kg, RMSE分别为0.135、0.135、0.069 g/kg和0.126 g/kg; 对 Na^+ 量预测的MAE分别为0.540、0.619、0.835 g/kg和1.371 g/kg, RMSE分别为0.636、0.748、1.198 g/kg和1.710 g/kg。【结论】SVR模型预测 K^+ 、 Na^+ 量的精度最高, 可推荐作为南疆棉田土壤钾、钠离子量预测的一种方法。

关键词: 南疆棉田; 土壤盐分离子; 机器学习; 回归预测模型

中图分类号: TP181

文献标志码: A

doi: 10.13522/j.cnki.ggps.2022405

OSID: 

唐茂淞, 张楠, 李国辉, 等. 基于机器学习算法的棉田土壤钾、钠离子量预测[J]. 灌溉排水学报, 2023, 42(9): 32-39.

TANG Maosong, ZHANG Nan, LI Guohui, et al. Predicting Soil K^+ and Na^+ Contents in Cotton Field Using Machine Learning Algorithm[J]. Journal of Irrigation and Drainage, 2023, 42(9): 32-39.

0 引言

【研究意义】降水稀少、蒸发强烈、气候干旱及土壤母质含盐量高是造成南疆土壤盐渍化严重的内在因素^[1], 而不合理的灌溉、施肥及过度开发则进一步加重了这一过程^[2]。土壤中钾(K^+)、钠(Na^+)等离子量过高, 将会阻滞作物对养分的吸收、抑制生长进而导致作物产量降低^[3-6]。【研究进展】目前, 使用机器学习算法对土壤理化指标进行预测已成为研究的热点^[7-9]。研究表明, 支持向量回归(Support Vector Regression, SVR)模型对土壤体质量和孔隙度的预测能够使 R^2 分别达到0.867和0.743^[10], 对土壤含水率的预测效果也较为理想^[11]。而基于灰狼优化算法的SVR校正模型对土壤镉量有

着更高的预测精度^[12]。利用随机森林回归(Random Forest Regression, RFR)模型对土壤有机质的空间分布^[13]、土壤pH值预测也能够取得较好结果^[14]; 当采用遗传算法和贝叶斯优化算法分别对RFR进行改进后, 在对土壤含盐量的反演过程中能够达到较佳结果^[15]。借助K-最近邻回归(K-Nearest Neighbor Regression, KNNR)对土壤有机质的空间分布进行预测^[16], 对土壤水热变化趋势的模拟精度较高^[17]。利用梯度提升回归树(Gradient Boosting Regression Tree, GBRT)建立的土壤电导率的估算模型表现出较高的预测能力^[18], GBRT也在对土壤水分^[19]和土壤镍量^[20]的高光谱反演模拟中具有优势。【切入点】南疆地区土壤次生盐渍化程度十分严重, 离子毒害导致了棉花生产力下降趋势明显。为此, 如何对上述土壤盐分离子量进行精确的估算, 将会对预判土壤盐渍化程度及提出适宜的防治措施具有重要意义。

【拟解决的关键问题】本研究基于不同的盐分处理, 利用SVR、RFR、KNNR和GBRT机器学习预测模型对棉花不同生育阶段及不同土层深度的土壤 K^+ 、 Na^+ 量进行预测, 以期对南疆棉田土壤盐分离子量预

收稿日期: 2022-07-20 修回日期: 2023-05-11 网络出版日期: 2023-09-13
基金项目: “十四五”国家重点研发计划项目(2022YFD1900505); 兵团重大科技项目(2021AA003); 塔里木大学研究生科研创新项目(TDGRI202143)

作者简介: 唐茂淞(1997-), 男, 硕士研究生, 主要从事灌溉排水理论与节水灌溉研究。E-mail: tms765951540@gmail.com

通信作者: 王兴鹏(1978-), 男, 教授, 博士生导师, 博士, 主要从事旱区水资源高效利用及水环境保护研究。E-mail: 13999068354@163.com

©《灌溉排水学报》编辑部, 开放获取 CC BY-NC-ND 协议

测提供一种新的方法。

1 材料与方法

1.1 试验区概况

试验区位于新疆生产建设兵团第一师水利灌溉试验站（40.32°N，81.17°E，海拔 1 014 m），属暖温带极端大陆性干旱荒漠气候，冬季寒冷，夏季炎热，昼夜温差大，降水量小，蒸发量大。试验站多年平均气温 11.3 °C，年平均降水量 46 mm，蒸发量 1 877~2 559 mm，日照时间 2 950 h，无霜期 207 d，地下水埋深 3.5~5.0 m，平均含盐量 1.7 g/kg。试验站土壤质地为沙壤土，0~100 cm 土壤平均体积质量为 1.56 g/cm³，pH 值为 8.6。

1.2 试验设计

2020 年棉花播种和收获日期分别为 4 月 26 日和 10 月 27 日，棉田设置了 1 mS/cm（T1）、2 mS/cm（T2）和 3 mS/cm（T3）3 个盐分梯度。2021 年棉花播种和收获日期分别为 4 月 22 日和 11 月 8 日，棉田设置了 2 mS/cm（T4）、5 mS/cm（T5）和 8 mS/cm（T6）3 个盐分梯度。每个处理 3 次重复，棉花施肥、除草、打药、打顶等田间管理措施均与当地农田一致。

1.3 样品采集与测定

分别在棉花苗期、蕾期、花铃前期和花铃后期取土测定土壤盐分离子量，取样深度分别为 0~10、10~20、20~30 cm 和 30~40 cm。土样自然风干后，以土水质量比为 1:5 制备浸提液，采用火焰光度计法测定土壤 K⁺、Na⁺量。

1.4 模型训练与验证

本文采用 SVR、RFR、KNNR 和 GBRT 这 4 种机器学习构建回归预测模型。训练样本为 2020 年实测的 K⁺、Na⁺量，2 种离子量各有 144 个训练数据（包括 3 次重复值），测试样本为 2021 年实测的 K⁺、Na⁺量，2 种离子量各有 48 个测试数据（包括 3 次重复均值）。模型分别以 K⁺、Na⁺量为因变量，自变量为初始盐分梯度、土层深度和棉花生育期，分类变量量化标准如表 1 所示，通过 R 语言程序进行标准化消除自变量之间的量纲差异。使用 MAE、RMSE 和 R² 对模拟结果进行评估，其中，MAE 和 RMSE 越小，模型预测值越接近实测值，R² 越接近于 1，模型拟合性能越好。

表 1 分类变量量化标准

Table 1 Quantitative criteria for categorical variables

变量名称	量化标准
土层深度	0~10 cm-1、10~20 cm-2、20~30 cm-3、30~40 cm-4
棉花生育期	苗期-1、蕾期-2、花铃前期-3、花铃后期-4

1.5 机器学习模型

1.5.1 支持向量回归（SVR）

SVR 模型是一种广义线性模型，通过核函数进行线性回归处理^[21]，SVR 作为一种最大似然方法，基于结构风险最小化原则可以克服过度拟合的问题^[22]。SVR 表达式如式（1）所示：

$$f(x) = \omega x + b = \sum_{n=1}^l (a_n - a_n^*) K(x_n, x) + b, \quad (1)$$

式中： $f(x)$ 为回归函数； ω 和 b 为超平面的系数； $x \in R^n$ ； a_n 和 a_n^* 为样本支持向量； $K(x_n, x)$ 为非线性核函数。

SVR 的建模预测流程为：首先将低维特征空间映射到高维空间，再通过线性回归实现低维特征空间数据的预测。SVR 作为一种监督学习算法，使用对称损失函数进行训练，具有出色的泛化能力和高预测精度。

1.5.2 随机森林回归（RFR）

RFR 模型是一种典型的机器学习算法，由分类与回归树组成，以强大的非线性拟合能力避免了出现过拟合现象^[23]，结合加权平均原则对决策树的监测结果进行计算后得到最终监测结果^[24]。RFR 表达式如式（2）所示。

$$\hat{y}(x) = \sum_{i=1}^n \omega_i(x) Y_i, \quad (2)$$

式中： $\hat{y}$ 为最终监测值； n 为因变量个数； $\omega_i(x)$ 为每个因变量观测值的权重； Y_i 为因变量的观测值。

RFR 的建模预测流程为：对每棵树建立一个独立的决策树回归模型，最终预测结果为所有树的平均值。

1.5.3 K-最近邻回归（KNNR）

KNNR 模型是一种非参数模型，基于距离度量找出训练集中与其最靠近的 k 个训练样本，KNNR 的欧式距离表达式如式（3）所示。

$$d = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}, \quad (3)$$

式中： d 为测试样本和指定训练样本之间的欧氏距离； k 为最近邻样本数量； x_i 为第 1 个点的第 n 维坐标， y_i 为第 2 个点的第 n 维坐标。

KNNR 的建模预测流程为：首先确定距离计算方法，确定 k 值大小，再从训练集中找到 k 个与测试样本距离最接近的样本，最后使用 k 个近邻的均值作为测试样本的预测值^[25]。KNNR 算法的缺点比较明显，其对近邻数 k 的取值比较敏感，若 k 值过小容易引发过拟合，若 k 值过大可能会增大近似误差，且 KNNR 计算样本需要较长时间，特别当数据量很大时，可能会导致内存溢出。

1.5.4 梯度提升回归树（GBRT）

GBRT 模型是一种函数空间优化算法，能适应

复杂的非线性关系, GBRT 在每一次迭代后产生一个精度不高的弱学习器, 每个学习器之间不独立, 最终将弱学习器集成可以实现较高的精度^[26]。GBRT 表达式如式 (4) 所示。

$$F_m(x) = \sum_{i=1}^n \theta_i f_i(x), \quad (4)$$

式中: $F_m(x)$ 为最终的模型; n 为弱学习器的数量; θ_i 为减少过拟合的系数; $f_i(x)$ 为弱学习器。

GBRT 的建模预测流程为: 首先使用一个弱学习器输出预测结果, 再使用第二个学习器去学习特征到残差的映射, 将 2 个学习器的输出结果相加得到最终预测结果, 经过多次迭代后, 可得到最终预测结果。GBRT 算法本质上是一个基于树的模型, 它集成了由 CART 算法生成的几个弱学习器。GBRT 方法属于集成学习的增强类别, 对于增强学习算法, 得分高的弱学习器将获得更高的权重。

1.6 数据处理

首先采用 Microsoft Excel 2019 录入和整理试验数据, 然后分别采用 R 语言 tidyverse 程序包、e1071 程序包、randomForest 程序包、caret 程序包、h2o 程序包和 ggplot2 程序包进行数据预处理、建立 SVR 模型、RFR 模型、KNNR 模型、GBRT 模型和作图。

2 结果与分析

2.1 不同盐分处理的棉田土壤钾、钠离子量变化

不同盐分处理下棉花全生育期内 0~40 cm 土层 K^+ 、 Na^+ 量均值变化如图 1 所示。通过 2 a 的试验发现, K^+ 量随着棉花生育期呈先增加后减少的趋势, 蕾期和花铃前期的 K^+ 量要高于苗期和花铃后期。 Na^+ 量在 2020 年的试验中随着棉花生育期呈下降的趋势, 在花铃后期达到最小, 而在 2021 年的试验中, Na^+ 量变化较为稳定, 各生育阶段变化值较小, 不同盐分处理对 K^+ 、 Na^+ 量影响明显。

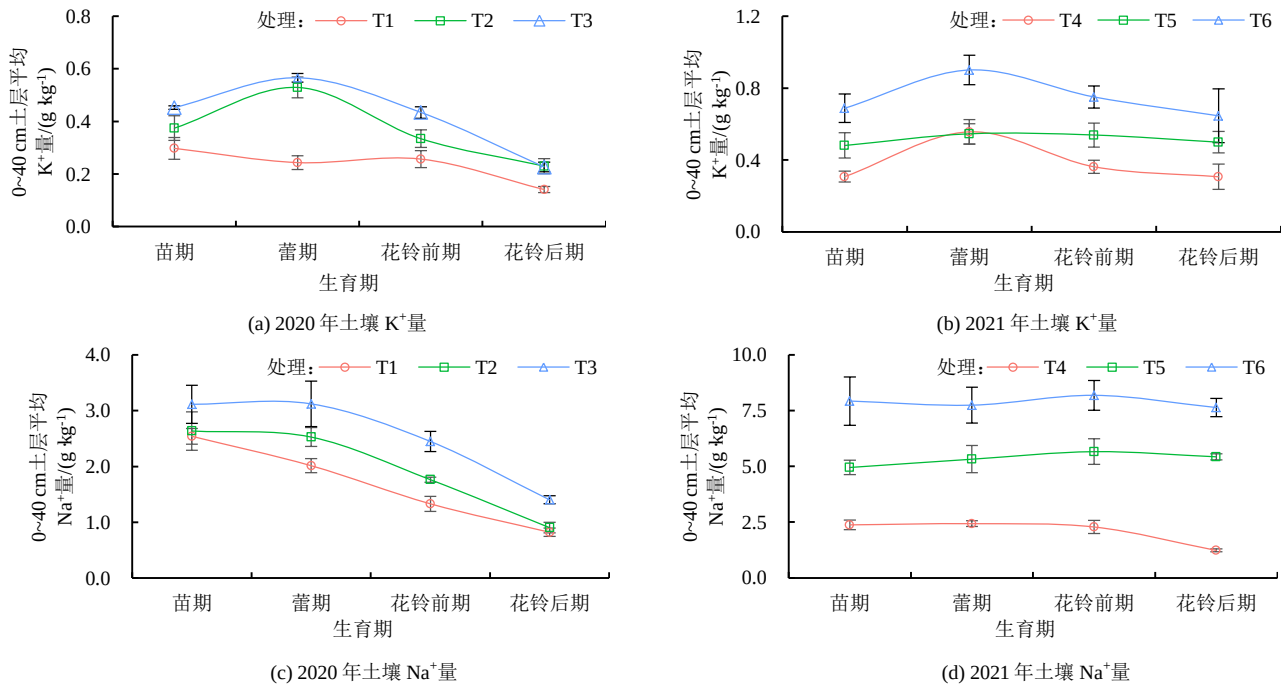


图 1 全生育期 K^+ 、 Na^+ 量变化

Fig.1 Changes of K^+ and Na^+ contents in the whole growth period

2.2 土壤盐分钾、钠离子量的统计特征

以 2020、2021 年棉田土壤 K^+ 、 Na^+ 量实测数据为总体样本。以 2020 年 K^+ 、 Na^+ 量为训练样本, 2021 年 K^+ 、 Na^+ 量为测试样本, 不同离子量统计特征如图 2 所示。图中 n 为样本数量, $mean$ 为平均值, sd 为标准差, vc 为变异系数。由图 2 可知, K^+ 量样本集的变异系数范围为 38.72%~50.82%, Na^+ 量样本集的变异系数范围为 48.8%~73.47%。由于变异系数均在 10%~100% 的区间内, 样本集均属于中等变异。从箱线图可以看出, K^+ 量总体样本和训练样本各出现 1 个异常值, Na^+ 量总体样本和训练样本分别出现

16 个和 3 个异常值, 但由于样本集中的异常值均为极个别情况下对土壤离子量的实测值, 故无需对伪异常进行处理。

2.3 SVR 模型的建立与验证

使用 R 程序对训练样本训练模型时, 需修改 svm() 函数的参数, 参数 x 和 y 分别为训练样本的自变量和因变量, 参数 $type$ 选择 “eps-regression” 建立回归预测, 参数 $kernel$ 选择 “radial” 使用高斯核。参数 $gamma$ 是进行非线性预测的超参数, 超参数 $cost$ 是惩罚因子, 由于使用了高斯核则需要对 $gamma$ 和 $cost$ 进行同步优化, 使建立的模型精度达

到最佳。使用试错法结合 bootstrap 采样进行超参数搜索,通过 tune.controls()函数设置使用“bootstrap”采样的方法进行参数搜索,得到超参数 γ 和 $cost$ 较优取值如表 2 所示。

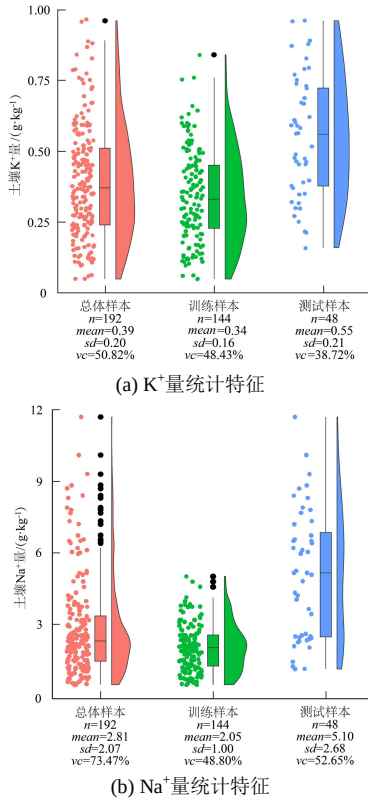


图 2 不同离子量的统计特征

Fig.2 Statistical characteristics of different ion contents

表 2 SVR 模型的超参数调优

Table 2 Super parameter optimization of SVR model

离子	超参数设置	
	γ	$cost$
K^+	0.001	7
Na^+	0.005	1 600

SVR 模型对训练样本和测试样本的预测结果如表 3 所示。在训练样本中,SVR 对 Na^+ 量取得的决定系数最大, R^2 为 0.900。在测试样本中,SVR 对 Na^+ 量取得的决定系数最大, R^2 为 0.838。综合来看,SVR 对 Na^+ 量的预测效果较好。

表 3 SVR 模型的预测结果

Table 3 Prediction results of SVR model

离子	训练样本			测试样本		
	$MAE/(g \cdot kg^{-1})$	$RMSE/(g \cdot kg^{-1})$	R^2	$MAE/(g \cdot kg^{-1})$	$RMSE/(g \cdot kg^{-1})$	R^2
K^+	0.073	0.097	0.701	0.100	0.119	0.687
Na^+	0.234	0.317	0.900	0.841	1.154	0.838

2.4 RFR 模型的建立与验证

本研究使用 R 程序对数值型训练样本训练模型时, randomForest()函数会根据输入变量建立回归预测模型,参数 x 和 y 分别为训练样本的自变量和因变量,参数 $data$ 为训练样本数据,超参数 $ntree$ 表示生长树的数量。使用试错法结合 tuneRF()函数寻

找较优的超参数,随着 ntreeTry 参数取值递增, OBB Error 模型误差先迅速降低,随后略微增加,在误差最小处确定超参数 $ntree$ 的值,得到超参数 $ntree$ 取值如表 4 所示。

表 4 RFR 模型的超参数调优

Table 4 Super parameter optimization of RFR model

离子	超参数设置
	$ntree$
K^+	5
Na^+	6

RFR 模型对训练样本和测试样本的预测结果如表 5 所示。在训练样本中,RFR 对 Na^+ 量取得的决定系数最大, R^2 为 0.838。在测试样本中,RFR 对 K^+ 量取得的决定系数最大, R^2 为 0.437,但是对 Na^+ 量取得的决定系数较小, R^2 为 0.299。综合来看,RFR 对 K^+ 量的预测效果较好。

表 5 RFR 模型的预测结果

Table 5 Prediction results of RFR model

离子	训练样本			测试样本		
	$MAE/(g \cdot kg^{-1})$	$RMSE/(g \cdot kg^{-1})$	R^2	$MAE/(g \cdot kg^{-1})$	$RMSE/(g \cdot kg^{-1})$	R^2
K^+	0.061	0.077	0.837	0.169	0.218	0.437
Na^+	0.381	0.478	0.838	2.841	3.658	0.299

2.5 KNNR 模型的建立与验证

本研究使用 R 程序对训练样本训练模型时,需修改 knnreg()函数的参数,参数 x 和 y 分别为训练样本的自变量和因变量,参数 $data$ 为训练样本数据,超参数 k 的值表明近邻的个数。随着近邻的个数由 0 开始增加,模型的预测误差会快速减小,在近邻数取得某个值后误差会缓慢增大,由此可确定影响模型精度的最佳 k 值,通过程序循环,根据不同 k 值训练模型后比较 k 值与误差的关系,得到超参数 k 取值如表 6 所示。

表 6 KNNR 模型的超参数调优

Table 6 Super parameter optimization of KNNR model

离子	超参数设置
	k
K^+	11
Na^+	10

KNNR 模型对训练样本和测试样本的预测结果如表 7 所示。在训练样本中,KNNR 对 Na^+ 量取得的决定系数最大, R^2 为 0.919。在测试样本中,KNNR 对 K^+ 量取得的决定系数最大, R^2 为 0.430,但是对 Na^+ 量取得的决定系数较小, R^2 为 0.219。综合来看,KNNR 对 K^+ 量的预测效果较好。

表 7 KNNR 模型的预测结果

Table 7 Prediction results of KNNR model

离子	训练样本			测试样本		
	$MAE/(g \cdot kg^{-1})$	$RMSE/(g \cdot kg^{-1})$	R^2	$MAE/(g \cdot kg^{-1})$	$RMSE/(g \cdot kg^{-1})$	R^2
K^+	0.049	0.064	0.877	0.169	0.218	0.430
Na^+	0.262	0.322	0.919	2.826	3.630	0.219

2.6 GBRT 模型的建立与验证

在研究使用 R 程序对训练样本训练模型时, 需修改 `h2o.gbm()` 函数的参数, 参数 x 和 y 分别为训练样本的自变量和因变量, 参数 `training_frame` 为训练样本数据, 超参数 `ntrees` 和 `max_depth` 分别表示树的个数和树的最大深度。通过试错法结合 `h2o.grid()` 函数进行超参数网格搜索, 使用训练样本对 GBRT 参数调优, 得到超参数 `ntrees` 和 `max_depth` 取值如表 8 所示。

表 8 GBRT 模型的超参数调优

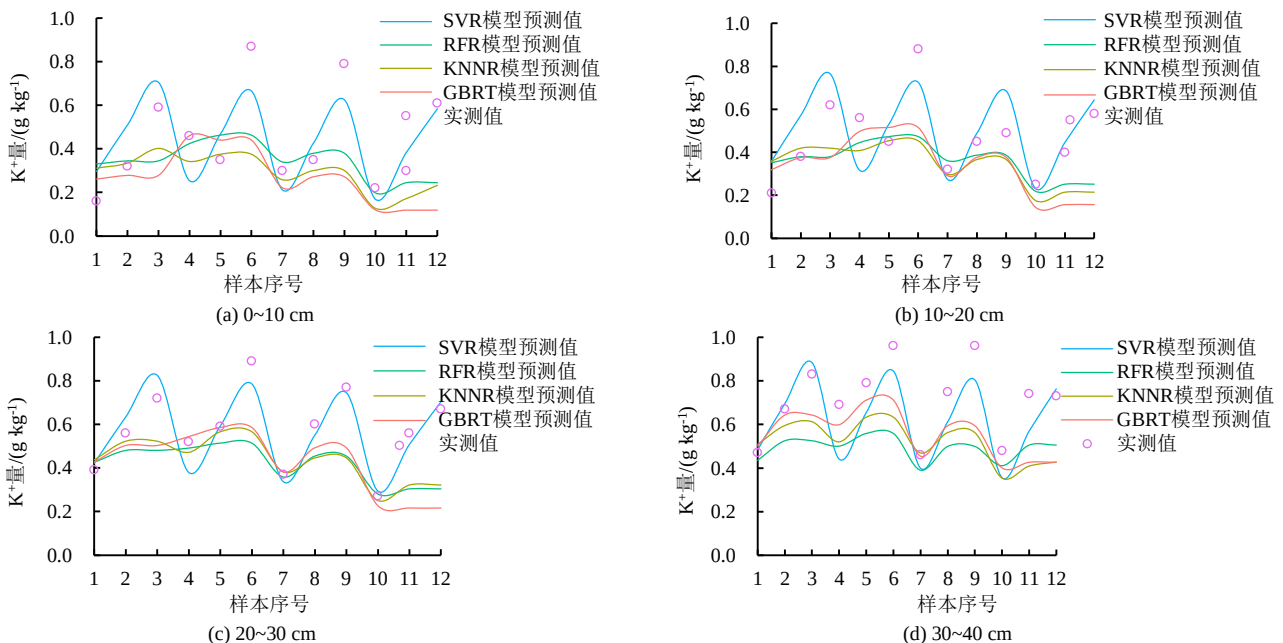
Table 8 Super parameter optimization of GBRT model

离子	超参数设置	
	<i>ntrees</i>	<i>max_depth</i>
K ⁺	100	5
Na ⁺	120	2

GBRT 模型对训练样本和测试样本的预测结果如表 9 所示。在训练样本中, GBRT 对 K⁺量取得的决定系数最大, R^2 为 0.946。在测试样本中, GBRT 对 K⁺量取得的决定系数最大, R^2 为 0.395。综合来看, GBRT 对 K⁺量的预测效果较好。

表 10 不同土层深度 K⁺量模型预测精度对比Table 10 Comparison of prediction accuracy of K⁺ content model in different soil depths

土层深度/cm	MAE/(g·kg ⁻¹)				RMSE/(g·kg ⁻¹)			
	SVR 模型	RFR 模型	KNNR 模型	GBRT 模型	SVR 模型	RFR 模型	KNNR 模型	GBRT 模型
0~10	0.122	0.160	0.181	0.202	0.135	0.219	0.247	0.269
10~20	0.114	0.137	0.152	0.153	0.135	0.184	0.197	0.200
20~30	0.056	0.162	0.148	0.157	0.069	0.211	0.195	0.216
30~40	0.106	0.218	0.195	0.157	0.126	0.251	0.228	0.197

图 3 不同土层深度 K⁺量预测值与实测值的关系Fig.3 Relationship between predicted value and measured value of K⁺ content in different soil depths表 11 不同土层深度 Na⁺量模型预测精度对比Table 11 Comparison of prediction accuracy of Na⁺ content model in different soil depths

土层深度/cm	MAE/(g·kg ⁻¹)				RMSE/(g·kg ⁻¹)			
	SVR 模型	RFR 模型	KNNR 模型	GBRT 模型	SVR 模型	RFR 模型	KNNR 模型	GBRT 模型
0~10	0.540	2.149	2.381	2.534	0.636	2.783	2.967	3.168
10~20	0.619	2.523	2.431	2.496	0.748	3.053	2.981	3.032
20~30	0.835	3.069	3.022	3.035	1.198	3.840	3.817	3.821
30~40	1.371	3.621	3.469	3.358	1.710	4.658	4.524	4.410

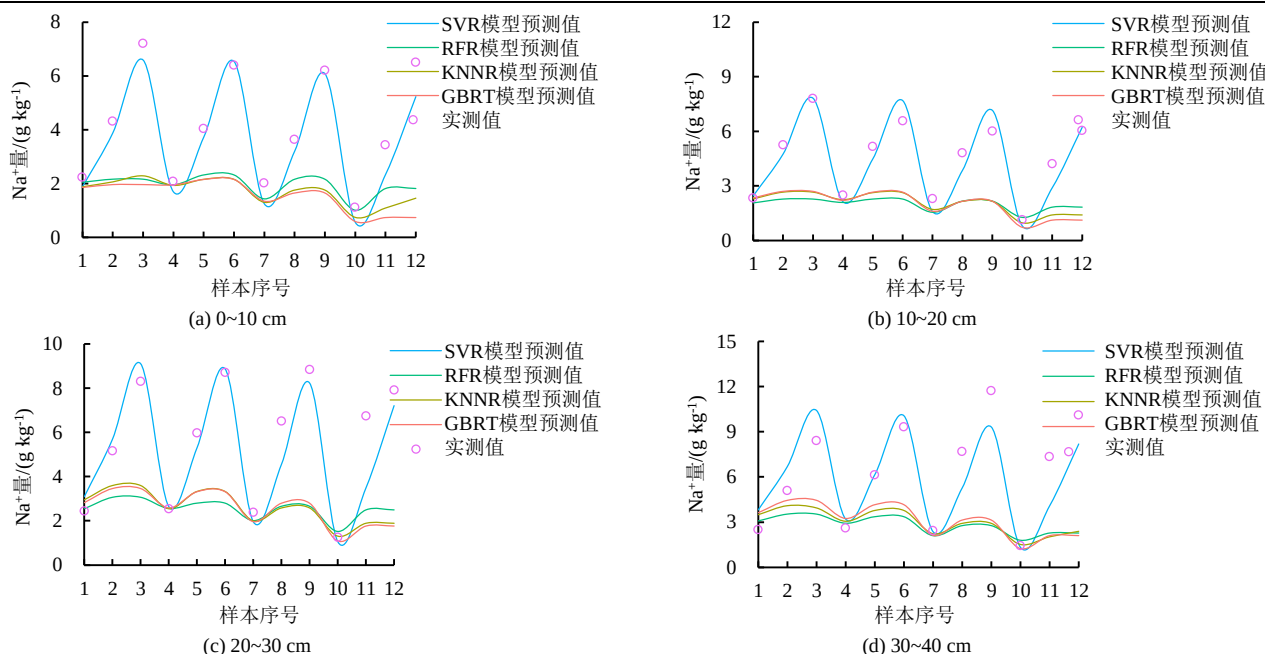
表 9 GBRT 模型的预测结果

Table 9 Prediction results of GBRT model

离子	训练样本			测试样本		
	MAE/(g·kg ⁻¹)	RMSE/(g·kg ⁻¹)	R^2	MAE/(g·kg ⁻¹)	RMSE/(g·kg ⁻¹)	R^2
K ⁺	0.027	0.038	0.946	0.167	0.223	0.395
Na ⁺	0.216	0.275	0.924	2.856	3.650	0.200

2.7 不同土层深度钾、钠离子量的模型预测精度对比

不同土层深度 K⁺量的模型预测精度如表 10 和图 3 所示, 将 K⁺量预测误差进行对比时发现, SVR 模型均取得最低误差, 4 个土层深度 (0~10、10~20、20~30 cm 和 30~40 cm) 的 MAE 分别为 0.122、0.114、0.056 g/kg 和 0.106 g/kg, RMSE 分别为 0.135、0.135、0.069 g/kg 和 0.126 g/kg, 表现最优。不同土层深度 Na⁺量的模型预测精度如表 11 和图 4 所示, 将 Na⁺量预测误差进行对比时发现, SVR 模型均取得最低误差, 4 个土层深度的 MAE 分别为 0.540、0.619、0.835 g/kg 和 1.371 g/kg, RMSE 分别为 0.636、0.748、1.198 g/kg 和 1.710 g/kg, 表现最优。因此, SVR 模型对以深度分层的土壤 K⁺、Na⁺量预测取得理想效果。

图4 不同土层深度 Na^+ 量预测值与实测值的关系Fig.4 Relationship between predicted and measured Na^+ content in different soil depths

3 讨论

机器学习算法在土壤理化性质预测中具有应用潜力, 相关研究表明, 通过机器学习算法可以建立可靠有效的模型^[27-29]。本研究使用 SVR、RFR、KNNR、GBRT 模型对棉田土壤的 K^+ 、 Na^+ 量进行预测, 结果表明, SVR 模型对 K^+ 、 Na^+ 量整体测试样本的预测精度最佳, MAE 分别为 0.100 g/kg 和 0.841 g/kg, RMSE 为 0.119 g/kg 和 1.154 g/kg, R^2 分别达到 0.687 和 0.838。

有研究表明, 将不同试验处理^[30]、不同土层深度^[31]、不同生育期^[32]的样本数据作为输入变量建立模型是可行的, 这与本文的模型构建思路一致。SVR 模型和 RFR 模型可以实现对土壤有机碳的空间分布预测^[33], 以及反向传播神经网络适用于土壤墒情预测^[34], 本文使用 4 种机器学习算法可以实现对不同土壤离子量的预测, SVR 模型对钾钠离子的预测精度较高, 而 RFR、KNNR、GBRT 模型的预测效果相对较差。由于本文仅使用 2 a 试验数据, 其他模型可能存在样本数量的限制而表现的预测精度较低。因此, 在后续的试验中, 可连续多年在不同的棉田内采集样本数据, 使机器学习回归预测模型在大量样本数据的训练下进一步提升预测精度。

本研究位于南疆棉花种植区, 由于南疆独特的土壤理化性质导致土壤含盐量较高且棉田盐分空间变异性大, 盐分离子对作物生长产生较为严重的影响。土壤母质含钾钠矿物比较多是土壤钾钠离子主要来源, 施肥和灌溉也是影响土壤中钾钠离子量的主要因素之一, 由于土壤浸提液中盐分一般以离子

的形式存在, 土壤电导率可以综合反映出土壤浸提液中各种阴阳离子量之和, 也可以在一定程度上表征土壤的含盐量^[35]。在对盐渍化土壤进行改良时, 精准预测土壤盐分离子量, 对制定适宜的改良措施具有重要的参考价值。本文通过对 4 种机器学习方法预测南疆棉田土壤 K^+ 、 Na^+ 量的结果进行比较分析, 认为 SVR 模型能够较为准确地预测 K^+ 、 Na^+ 量。在实际应用时, 可在棉花播前采集土样测定土壤 K^+ 、 Na^+ 量的本底值, 再结合 SVR 模型即可预测出棉花不同生育期各土层的 K^+ 、 Na^+ 量。

4 结论

1) K^+ 量和 Na^+ 量受盐分影响明显, 土壤盐分越高则 K^+ 量和 Na^+ 量越高, 且在棉花生育期内, K^+ 量随着棉花生育期呈先增加后减少的趋势, Na^+ 量随着棉花生育期呈下降趋势。

2) 在盐分本底值、棉花生育期及土层深度的多重影响下, 建立的 SVR 模型对土壤 K^+ 、 Na^+ 量的预测精度最高, 可作为南疆棉田土壤 K^+ 、 Na^+ 量预测的一种方法。

(作者声明本文无实际或潜在的利益冲突)

参考文献:

- 王兴鹏. 冬春灌对南疆土壤水盐动态和棉花生长的影响研究[D]. 北京: 中国农业科学院, 2018.
WANG Xingpeng. Effects of winter-spring irrigation on soil water-salt dynamics and cotton growth[D]. Beijing: Chinese Academy of Agricultural Sciences, 2018.
- 杨涛, 李生梅, 黄雅婕, 等. 海岛棉资源自然复合盐胁迫综合评价[J]. 核农学报, 2021, 35(7): 1 507-1 521.
YANG Tao, LI Shengmei, HUANG Yajie, et al. Comprehensive evaluation

- of natural compound salt stress of sea-island cotton resources[J]. Journal of Nuclear Agricultural Sciences, 2021, 35(7): 1 507-1 521.
- [3] 屈忠义, 孙慧慧, 杨博, 等. 不同改良剂对盐碱地土壤微生物与加工番茄产量的影响[J]. 农业机械学报, 2021, 52(4): 311-318, 350.
QU Zhongyi, SUN Huihui, YANG Bo, et al. Effects of different amendments on soil microorganisms and yield of processing tomato in saline alkali soil[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(4): 311-318, 350.
- [4] 窦旭, 史海滨, 李瑞平, 等. 盐渍化土壤剖面盐分与养分分布特征及盐分迁移估算[J]. 农业机械学报, 2022, 53(1): 279-290, 330.
DOU Xu, SHI Haibin, LI Ruiping, et al. Distribution characteristics of salinity and nutrients in salinized soil profile and estimation of salt migration[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(1): 279-290, 330.
- [5] 郑复乐, 姚荣江, 杨劲松, 等. 改良材料对微咸水滴灌农田土壤盐分分布与离子组成的影响[J]. 灌溉排水学报, 2020, 39(8): 60-71.
ZHENG Fule, YAO Rongjiang, YANG Jinsong, et al. The effects of soil amendment with different materials on soil salt distribution and its ion composition under brackish-water drip irrigation[J]. Journal of Irrigation and Drainage, 2020, 39(8): 60-71.
- [6] 王航, 周青云, 张宝忠, 等. 不同灌水处理下滨海盐碱地土壤-玉米阳离子变化规律及相关关系研究[J]. 灌溉排水学报, 2021, 40(12): 36-43.
WANG Hang, ZHOU Qingyun, ZHANG Baozhong, et al. Irrigation affects the translocation of cations from soil to maize roots in saline-alkaline soil[J]. Journal of Irrigation and Drainage, 2021, 40(12): 36-43.
- [7] 王铭鑫, 范超, 高秉博, 等. 融合半变异函数的空间随机森林插值方法[J]. 中国生态农业学报(中英文), 2022, 30(3): 451-457.
WANG Mingxin, FAN Chao, GAO Bingbo, et al. A spatial random forest interpolation method with semi-variogram[J]. Chinese Journal of Eco-Agriculture, 2022, 30(3): 451-457.
- [8] 马国林, 丁建丽, 韩礼敬, 等. 基于变量优选与机器学习的干旱区湿地土壤盐渍化数字制图[J]. 农业工程学报, 2020, 36(19): 124-131.
MA Guolin, DING Jianli, HAN Lijing, et al. Digital mapping of soil salinization in arid area wetland based on variable optimized selection and machine learning[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(19): 124-131.
- [9] 刘明杰, 徐卓揆, 郜允兵, 等. 基于机器学习的稀疏样本下的土壤有机质估算方法[J]. 地球信息科学学报, 2020, 22(9): 1 799-1 813.
LIU Mingjie, XU Zhuokui, GAO Yunbing, et al. Estimating soil organic matter based on machine learning under sparse sample[J]. Journal of Geo-information Science, 2020, 22(9): 1 799-1 813.
- [10] 杨玮, 兰红, 李民赞, 等. 基于图像处理和 SVR 的土壤容重与土壤孔隙度预测[J]. 农业工程学报, 2021, 37(12): 144-151.
YANG Wei, LAN Hong, LI Minzan, et al. Predicting bulk density and porosity of soil using image processing and support vector regression[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(12): 144-151.
- [11] ACHIENG K O. Modelling of soil moisture retention curve using machine learning techniques: Artificial and deep neural networks vs support vector regression models[J]. Computers & Geosciences, 2019, 133: 104 320.
- [12] 陈颖, 张灿, 肖春艳, 等. 基于 GWO-SVR 的土壤镉元素含量含水率校正预测模型研究[J]. 光学学报, 2020, 40(10): 180-187.
CHEN Ying, ZHANG Can, XIAO Chunyan, et al. Study on prediction model of soil cadmium content moisture content correction based on GWO-SVR[J]. Acta Optica Sinica, 2020, 40(10): 180-187.
- [13] 尉芳, 刘京, 夏利恒, 等. 陕西渭北旱塬区农田土壤有机质空间预测方法[J]. 环境科学, 2022, 43(2): 1 097-1 107.
WEI Fang, LIU Jing, XIA Liheng, et al. Spatial prediction method of farmland soil organic matter in Weibei Dryland of Shaanxi Province[J]. Environmental Science, 2022, 43(2): 1 097-1 107.
- [14] TZIACHRIS P, ASCHONITIS V, CHATZISTATHIS T, et al. Comparing machine learning models and hybrid geostatistical methods using environmental and soil covariates for soil pH prediction[J]. ISPRS International Journal of Geo-Information, 2020, 9(4): 276.
- [15] 杨练兵, 陈春波, 郑宏伟, 等. 基于优化随机森林回归模型的土壤盐渍化反演[J]. 地球信息科学学报, 2021, 23(9): 1 662-1 674.
YANG Lianbing, CHEN Chunbo, ZHENG Hongwei, et al. Retrieval of soil salinity content based on random forests regression optimized by Bayesian optimization algorithm and genetic algorithm[J]. Journal of Geo-information Science, 2021, 23(9): 1 662-1 674.
- [16] 胡贵贵, 杨粉莉, 杨联安, 等. 基于主成分和机器学习的土壤有机质含量空间预测建模[J]. 干旱区地理, 2021, 44(4): 1 114-1 124.
HU Guigui, YANG Fenli, YANG Lian'an, et al. Spatial prediction modeling of soil organic matter content based on principal components and machine learning[J]. Arid Land Geography, 2021, 44(4): 1 114-1 124.
- [17] 刘宏超, 马俊杰, 李韧. 基于 KNN 机器学习方法对青藏高原唐古拉地区表层土壤水热状况的模拟[J]. 冰川冻土, 2021, 43(4): 1 243-1 252.
LIU Hongchao, MA Junjie, LI Ren. Simulation of the water-thermal features within the surface soil in Tanggula region, Qinghai-Tibet Plateau, by using KNN model[J]. Journal of Glaciology and Geocryology, 2021, 43(4): 1 243-1 252.
- [18] 曹肖奕, 丁建丽, 葛翔宇, 等. 基于不同卫星光谱模拟的土壤电导率估算研究[J]. 干旱区地理, 2020, 43(1): 172-181.
CAO Xiaoyi, DING Jianli, GE Xiangyu, et al. Estimation of soil conductivity based on spectral simulation of different satellites[J]. Arid Land Geography, 2020, 43(1): 172-181.
- [19] 田美玲, 葛翔宇, 丁建丽, 等. 耦合机器学习和机载高光谱数据的土壤含水量估算[J]. 激光与光电子学进展, 2020, 57(9): 232-241.
TIAN Meiling, GE Xiangyu, DING Jianli, et al. Coupled machine learning and unmanned aerial vehicle based hyperspectral data for soil moisture content estimation[J]. Laser & Optoelectronics Progress, 2020, 57(9): 232-241.
- [20] 傅邦杰, 牛瑞卿, 王春胜. 丹江口库区土壤镍含量高光谱反演方法[J]. 遥感信息, 2021, 36(3): 44-49.
FU Bangjie, NIU Ruiqing, WANG Chunsheng. Soil nickel metal content estimation based on hyper-spectrum in Danjiangkou Reservoir area[J]. Remote Sensing Information, 2021, 36(3): 44-49.
- [21] 任必武, 陈瀚阅, 张黎明, 等. 机器学习用于耕地土壤有机碳空间预测对比研究: 以亚热带复杂地貌区为例[J]. 中国生态农业学报(中英文), 2021, 29(6): 1 042-1 050.
REN Biwu, CHEN Hanyue, ZHANG Liming, et al. Comparison of machine learning for predicting and mapping soil organic carbon in cultivated land in a subtropical complex geomorphic region[J]. Chinese Journal of Eco-Agriculture, 2021, 29(6): 1 042-1 050.
- [22] WANG Xinxin, HAN Jigang, WANG Xia, et al. Estimating soil organic matter content using sentinel-2 imagery by machine learning in Shanghai[J]. IEEE Access, 2021, 9: 78 215-78 225.
- [23] 张万涛, 吉静怡, 李彬彬, 等. 黄土高原不同地貌区农田土壤有机质预测方法研究[J]. 植物营养与肥料学报, 2021, 27(4): 583-594.
ZHANG Wantao, JI Jingyi, LI Binbin, et al. Spatial prediction of soil organic matter of farmlands under different landforms in the Loess Plateau, China[J]. Journal of Plant Nutrition and Fertilizers, 2021, 27(4): 583-594.
- [24] 段维纳, 竞霞, 刘良云, 等. 融合 SIF 和反射光谱的小麦条锈病遥感监测[J]. 光谱学与光谱分析, 2022, 42(3): 859-865.
DUAN Weina, JING Xia, LIU Liangyun, et al. Monitoring of wheat stripe rust based on integration of SIF and reflectance spectrum[J]. Spectroscopy and Spectral Analysis, 2022, 42(3): 859-865.
- [25] ABEDI F, AMIRIAN-CHAKAN A, FARAJI M, et al. Salt dome related soil salinity in Southern Iran: Prediction and mapping with averaging machine learning models[J]. Land Degradation & Development, 2021, 32(3): 1 540-1 554.

- [26] 金则澎, 毛峰, 程乾, 等. 梯度提升回归树在千岛湖水体 CDOM 反演中的应用[J]. 遥感信息, 2022, 37(1): 110-118.
JIN Zepeng, MAO Feng, CHENG Qian, et al. Application of gradient boosting regression tree in CDOM inversion of Qiandao Lake[J]. Remote Sensing Information, 2022, 37(1): 110-118.
- [27] ZHOU Tao, GENG Yajun, JI Cheng, et al. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images[J]. Science of the Total Environment, 2021, 755: 142-661.
- [28] WANG Zong, DU Zhengping, LI Xiaoyan, et al. Incorporation of high accuracy surface modeling into machine learning to improve soil organic matter mapping[J]. Ecological Indicators, 2021, 129: 107-975.
- [29] TAGHIZADEH-MEHRJARDI R, SCHMIDT K, TOOMANIAN N, et al. Improving the spatial prediction of soil salinity in arid regions using wavelet transformation and support vector regression models[J]. Geoderma, 2021, 383: 114-793.
- [30] 王兴鹏, 蒋富昌, 王洪博, 等. 基于 AquaCrop 模型的南疆无膜滴灌棉花灌溉制度优化[J]. 农业机械学报, 2021, 52(4): 293-301, 335.
WANG Xingpeng, JIANG Fuchang, WANG Hongbo, et al. Irrigation scheduling optimization of drip-irrigated without plastic film cotton in South Xinjiang based on AquaCrop model[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(4): 293-301, 335.
- [31] 牛晓倩, 贾小旭, 刘成功, 等. 关中平原农田土壤水力参数空间分异与模拟[J]. 水土保持学报, 2021, 35(1): 198-204.
NIU Xiaoqian, JIA Xiaoxu, LIU Chenggong, et al. Spatial variations and simulations of farmland soil hydraulic parameters in the Guanzhong Plain[J]. Journal of Soil and Water Conservation, 2021, 35(1): 198-204.
- [32] 谭先明, 张佳伟, 王仲林, 等. 基于 PLS 的不同水氮条件下带状套作玉米产量预测[J]. 中国农业科学, 2022, 55(6): 1 127-1 138.
TAN Xianming, ZHANG Jiawei, WANG Zhonglin, et al. Prediction of maize yield in relay strip intercropping under different water and nitrogen conditions based on PLS[J]. Scientia Agricultura Sinica, 2022, 55(6): 1 127-1 138.
- [33] MAHMOUDZADEH H, MATINFAR H R, TAGHIZADEH-MEHRJARDI R, et al. Spatial prediction of soil organic carbon using machine learning techniques in western Iran[J]. Geoderma Regional, 2020, 21: e00 260.
- [34] 薛明, 韦波, 李娟, 等. 基于改进 BP 神经网络与支持向量机的土壤墒情预测方法研究[J]. 土壤通报, 2021, 52(4): 793-800.
XUE Ming, WEI Bo, LI Juan, et al. Forecast method of soil moisture based on improved BP neural network and support vector machine[J]. Chinese Journal of Soil Science, 2021, 52(4): 793-800.
- [35] 刘国锋, 徐增洪, 么宗利, 等. 冲水灌溉对西北硫酸盐型土壤中盐分离子变化的影响研究[J]. 干旱区资源与环境, 2019, 33(3): 118-123.
LIU Guofeng, XU Zenghong, YAO Zongli, et al. Effects of irrigation on the salt ions in sulfate-type saline-alkali soil[J]. Journal of Arid Land Resources and Environment, 2019, 33(3): 118-123.

Predicting Soil K^+ and Na^+ Contents in Cotton Field Using Machine Learning Algorithm

TANG Maosong^{1,2}, ZHANG Nan^{1,2}, LI Guohui^{1,2}, ZHAO Zeyi^{1,2}, LI Mingfa³, WANG Xingpeng^{1,2,4*}

(1. College of Water Resource and Architecture Engineering, Tarim University, Alaer 843300, China;

2. Laboratory of Modern Agricultural Engineering, Tarim University, Alaer 843300, China;

3. Hydrology and Water Resources Management Center of the First Division of Xinjiang Production and Construction Corps,

Alaer 843300, China; 4. Key Laboratory of Northwest Oasis Water-saving Agriculture,

Ministry of Agriculture and Rural Affairs, PR China, Shihezi 832000, China)

Abstract: 【Objective】The contents of K^+ and Na^+ in soil affect soil fertility and quality, and understanding their spatiotemporal changes and the factors influencing their changes is critical to improving soil management and alleviating soil alkalization. We propose a machine learning method to predict changes in K^+ and Na^+ content in soils.

【Method】Taking data measured from a cotton field in Southern Xinjiang as an example, we compared four machine learning algorithms: support vector regression (SVR), random forest regression (RFR), K-nearest neighbor regression (KNNR), and gradient lifting regression tree (GBRT). All algorithms were first trained based on K^+ and Na^+ measured in 2020, and the trained models were then tested against the data measured in 2021. The accuracy and robustness of the models were evaluated using the mean absolute errors (MAE), root mean square error (RMSE), and the determination coefficient (R^2). 【Result】The MAE of SVR, RFR, KNNR and GBRT for predicting K^+ content was 0.100, 0.169, 0.169 and 0.167 g/kg, respectively; their associated RMSE was 0.119, 0.218, 0.218 g/kg and 0.223 g/kg, respectively, and their R^2 was 0.687, 0.437, 0.430, and 0.395, respectively. For predicting Na^+ content, the MAE of SVR, RFR, KNNR and GBRT was 0.841, 2.841, 2.826 g/kg, and 2.856 g/kg, respectively; and their RMSE was 1.154, 3.658, 3.630 g/kg, and 3.650 g/kg, respectively, and R^2 was 0.838, 0.299, 0.219, and 0.200, respectively. SVR model is most accurate for predicting soil K^+ and Na^+ in the depths of 0~10, 10~20, 20~30 and 30~40 cm, with its MAE for K^+ at the four depths being 0.122, 0.114, 0.056 g/kg and 0.106 g/kg, respectively, and RMSE being 0.135, 0.135, 0.069 g/kg and 0.126 g/kg, respectively. The MAE of SVR for predicting Na^+ at the four depths was 0.540, 0.619, 0.835 g/kg and 1.371 g/kg, respectively, and its RMSE was 0.636, 0.748, 1.198 g/kg and 1.710 g/kg, respectively. 【Conclusion】Among the four algorithms we compared, SVR is most accurate for predicting soil K^+ and Na^+ at depth from 0 to 40 cm, and it can be used to predict variation in K^+ and Na^+ in response to environmental change in the cotton fields in Southern Xinjiang.

Key words: South Xinjiang cotton field; soil salt ions; machine learning; regression prediction model

责任编辑: 赵宇龙