

DOI:10.13409/j.cnki.jdpme.20240531003

基于机器学习算法的非饱和土水特征曲线预测*

张夏阳¹, 高游¹, 于响², 何伟¹

(1. 宁波大学土木工程与地理环境学院, 浙江宁波 315211; 2. 湖南智谋规划工程设计咨询有限责任公司, 湖南株洲 412000)

摘要: 土水特征曲线(SWCC)是研究非饱和土渗透、强度预测与本构关系的基础。机器学习算法具有高效处理大量数据和特征提取等特点。采用六种机器学习算法(四种集成学习和两种传统机器学习算法)对美国非饱和土数据库中的154条SWCC包含1976个数据点进行模拟;并使用四个性能评价指标(R^2 、EVS、MAE和RMSE)评价算法的性能。选取两种数据输入的方式:对压力水头进行对数处理和未处理两类。结果表明,在两种输入情况下,对LightGBM、XGB、RF和AdaBoost算法的影响很小;但是对GPR和SVM两种传统机器学习算法的影响很大,在未进行对数处理情况下, R^2 降低明显甚至会出现无法模拟SWCC的情况。此外,LightGBM对SWCC测试集的模拟效果上均优于其他模型,拥有高的趋势评价指标(R^2 和EVS)和低的误差测量指标(MAE和RMSE);六种算法对SWCC模拟的优劣的排列顺序依次为:LightGBM、GPR、XGB、RF、AdaBoost和SVM。最后,利用已训练好的LightGBM模型对9条不包含在数据库内的SWCC数据进行预测,结果显示LightGBM能够较好地预测非饱和土的土水特性。研究结果对提升不同类型土的SWCC预测具有重要的指导意义。

关键词: 非饱和土; 数据库; 土水特征曲线; 机器学习; 预测

中图分类号: TU411 **文献标识码:** A **文章编号:** 1672-2132(2025)01-0104-06

Prediction of Unsaturated Soil Water Characteristic Curve Based on Machine Learning Algorithms

ZHANG Xiayang¹, GAO You¹, YU Xiang², HE Wei¹

(1. School of Civil & Environmental Engineering and Geography Science, Ningbo University, Ningbo 315211, China;
2. Hunan Zhimo Planning and Engineering Design Consulting Co. Ltd., Zhuzhou 412000, China)

Abstract: The soil water characteristic curve (SWCC) is fundamental for studying the permeability, strength prediction, and constitutive relationships of unsaturated soils. Machine learning algorithms are characterized by their efficiency in large dataset processing and feature extraction. This study used six machine learning algorithms (four ensemble learning and two traditional machine learning algorithms) to simulate 154 SWCCs with 1976 data points from the United States Unsaturated Soil Database. Four performance evaluation indicators (R^2 , EVS, MAE, and RMSE) were used to assess the algorithms' performance. Two types of data input methods were selected: one with logarithmic processing of matric suction, and the other without any transformation. The results showed that, under

* 收稿日期:2024-05-31;修回日期:2024-07-11

基金项目:国家自然科学基金项目(42272312)资助

作者简介:张夏阳(1999—),男,硕士研究生。主要从事非饱和土力学方面的试验研究。E-mail:2219950752@qq.com

通信作者:高游(1989—),男,副教授,博士。主要从事非饱和土力学方面的研究。E-mail:gaoyou@nbu.edu.cn

both input types, the effect on the LightGBM, XGB, RF, and AdaBoost algorithms was minimal. However, the two traditional machine learning algorithms, GPR and SVM, were significantly affected. Without logarithmic transformation, R^2 decreased noticeably, and in some cases, the SWCC could not be simulated. Additionally, LightGBM outperformed other models in simulating the SWCC for the test set, with higher trend evaluation indicators (R^2 and EVS) and lower error measurement indicators (MAE and RMSE). The ranking of the six algorithms in terms of SWCC simulation performance was as follows: LightGBM, GPR, XGB, RF, AdaBoost, and SVM. Finally, the trained LightGBM model was used to predict 9 SWCC datasets not included in the original database. The results showed that LightGBM could effectively predict the soil water characteristics of unsaturated soils. The findings provide important guidance for improving SWCC predictions for different types of soils.

Keywords: unsaturated soil; database; soil water characteristic curve; machine learning; prediction

0 引言

非饱和土的持水特性是影响边坡稳定和水土流失的重要因素之一,直接影响其抗冲刷能力和抗滑性能^[1-2]。具有较好的持水特性的土体能够在一定程度上减缓水流速度,从而减少了冲刷和侵蚀的可能性。因此,在考虑边坡稳定和水土流失问题时,需要充分考虑土体的持水特性。

非饱和土的持水特性通常用土水特征曲线(SWCC)表示,即土体孔隙中水分含量和吸力之间的关系^[3-4]。广吸力范围内的SWCC一般需要联合压力板法、滤纸法和饱和盐溶液蒸汽平衡法等试验方法量测^[5-7]。然而,这些方法的试验过程繁琐且耗时。为此,学者们提出各种SWCC拟合方程。比如,经典的VG与FX的土水模型^[8-9]、考虑毛细水和吸附水影响的土水模型^[10-11]和考虑粒径级配影响的土水模型^[12]等。这些模型通常包含许多拟合参数,且不同类型土的拟合参数差异很大。这也导致这些理论模型在模拟不同类型土的SWCC效果不佳。

近年来,由于机器学习算法具有高效处理大量数据和特征提取等特点,在数据处理、图像识别和语音识别等领域均被广泛应用^[13-14]。许多学者将其应用于非饱和土水力特性领域的研究,并取得了良好的研究成果^[15-19]。如,M. G. Schaap等^[15]使用神经网络对砂土的脱湿SWCC进行建模,获得较好的模拟结果。K. Lamorski等^[16]基于波兰矿物土的SWCC数据库,采用支持向量机(SVM)和人工神经网络(ANN)对其进行训练和测试,指出SVM预测SWCC的效果更好。陈勇等^[17]整理了国内外350余

组SWCC,采用多元回归、XGBoost和随机森林3种回归模型对数据集进行验证计算,发现XGBoost模型预测效果最好。W. Qin等^[18]通过引入粒子群算法对SVM参数进行优化,并对黄土SWCC进行预测,得出的新方法比以往报道的模型具有更高的准确性。J. Zhan等^[19]使用随机森林算法(RF)对黄河下游冲积平原区土的SWCC进行建模分析,通过实测值验证了算法的有效性。为了充分评估不同算法预测SWCC的性能,有必要对其进行详细分析。

本文采用六种机器学习算法(四种集成学习算法和两种传统机器学习算法)模拟美国非饱和土数据库^[20]的SWCCs,所选数据集为154条SWCCs包含1976个数据点。为了对比不同算法模拟性能,使用决定系数(R^2)、可释方差得分(EVS)、平均绝对误差(MAE)和均方根误差(RMSE)对每种算法的模拟结果进行对比和分析。最终,通过利用训练好的模型对训练集之外的SWCC进行预测和分析,进一步验证不同类型的非饱和土SWCCs预测能力。

1 数据分析

UNSODA数据库包含了不同地区和不同类型的非饱和土SWCCs。本文从UNSODA数据库中选取154条SWCCs包含1976个数据点作为数据集^[21]。选择干密度(ρ_d)、比重(G_s)、饱和含水量(θ_s)、黏粒含量(F_{clay})、粉粒含量(F_{silt})与压力水头(H)作为输入变量,含水量(τw)作为输出变量。表1为本文所使用数据集的统计描述,为全面、有效地反映数据集的特征,本文以平均值、标准差、最小值、Q1、Q2和Q3统计数据集的输入变量。

表1 数据集的统计描述

Table 1 Statistical description of the database

统计描述	输入变量					
	$\rho_d/(g \cdot cm^{-3})$	G_s	θ_s	$F_{clay}/\%$	$F_{silt}/\%$	H/cm
平均值	1.48	2.65	0.42	21.57	29.74	1747
标准差	0.19	0.04	0.09	17.85	20.61	4026
最小值	1.01	2.50	0.23	0	0	0.01
Q1(25%)	1.34	2.62	0.36	4.92	16.48	30
Q2(50%)	1.49	2.65	0.41	17.90	27.80	80
Q3(75%)	1.61	2.66	0.48	33.82	42.70	690
最大值	1.97	2.80	0.66	76.36	85.46	16 540

注: Q1、Q2和Q3分别为第一、第二和第三四分位数

图1为6个输入变量之间的斯皮尔曼(Spearman)相关系数(r_s)矩阵图, $|r_s|$ 值越接近1,表明两个变量之间存在越强的正相关(或负相关)关系。从图中可知:比重与粉粒含量之间相关性最弱($|r_s|=0.01$);压力水头与黏粒含量之间的相关性较弱($|r_s|=0.20$);干密度与黏粒含量之间的相关性适中($|r_s|=0.49$);而干密度与饱和含水量之间的相关性最强($|r_s|=0.83$)。根据 r_s 的绝对值大小可将相关性程度划分为4个等级^[22],分别为很弱、较弱、中等与很强,见表2。

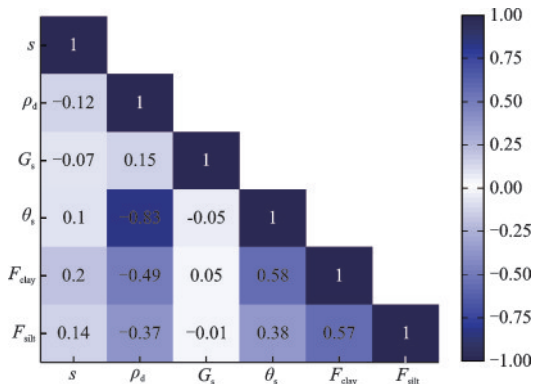


图1 输入变量的斯皮尔曼相关系数矩阵

Fig.1 Spearman correlation matrix of input variables

表2 基于 r_s 值的相关性划分

Table 2 Correlation classification based on r_s values

r_s (绝对值)	等级	输入变量
0.00~0.19	很弱	H 与 ρ_d , H 与 G_s , H 与 θ_s , H 与 F_{silt} , ρ_d 与 G_s , G_s 与 θ_s , G_s 与 F_{clay} , G_s 与 F_{silt}
0.20~0.39	较弱	H 与 F_{clay} , ρ_d 与 F_{silt} , θ_s 与 F_{silt}
0.40~0.59	中等	ρ_d 与 F_{clay} , θ_s 与 F_{clay} , F_{clay} 与 F_{silt}
0.80~1.00	很强	ρ_d 与 θ_s

2 预测结果与讨论

2.1 理论基础

为了评价六种机器学习算法的预测性能,选取四种性能评价指标:决定系数(R^2)、可释方差得分(EVS)、平均绝对误差(MAE)和均方根误差(RMSE)。表3和表4分别列出了本文使用的机器学习算法和性能评价指标公式。

表3 本文使用的机器学习算法

Table 3 Machine learning algorithms used in this study

名称	特征
LightGBM	基于决策树的学习算法,主要通过构建多个决策树提高模型的预测能力;适用于大规模、高维度数据集
XGB	加权决策树构建强模型,逐步提高整体模型的预测能力;适用于大规模、高维度数据集
Ada Boost	通过迭代,将前一轮强学习器的学习结果与当前一轮弱学习器的学习结果加权;适用于中小型数据集
RF	若干决策树学习器的集合,该集合中的每个决策树学习器均可估计一个分类;适用于高维,大规模数据集
GPR	利用高斯过程先验知识对样本数据进行回归分析的非参数回归模型;适合小规模数据集
SVM	最大间隔分类,使用核函数,支持高维特征空间,稀疏性;适用于样本量相对较少的高分类精度数据集

表4 本文使用的性能评价指标

Table 4 Performance evaluation indicators used in this study

评价指标公式	附注
$R^2 = 1 - \frac{\sum_{i=1}^m (y_p - y_m)^2}{\sum_{i=1}^m (y_p - y_{mc})^2}$	当 $R^2=1$,完全预测; $R^2=0$, $y_{mc}=y_m$; $R^2<0$, y_{mc} 优于 y_m
$EVS = 1 - \frac{var(y_p - y_m)}{var(y_m)}$	当EVS接近100%时, 该模型表现更好
$MAE = \frac{\sum_{i=1}^m y_p - y_m }{m}$	当MAE值越接近0表示 预测性能越好
$RMSE = \sqrt{\frac{\sum_{i=1}^m (y_p - y_m)^2}{m}}$	当RMSE值越接近0表示 预测越准确

注: y_m 为实测值; y_p 为预测值; var 为方差; y_{mc} 为实测值的平均值; m 为总数据量

2.2 六种机器学习算法的性能

在本文中,将数据集的80%作为训练集,剩余

的20%为测试集^[23-24]。80:20的划分比例在数据利用率和评估可靠性方面取得了较好的平衡。这种比例确保了模型有足够的训练数据,同时测试集的规模也足够大,以提供可靠的性能评估。表5为数据集的输入与输出参数。由表可知,本研究选取两种数据输入的方式:对前期输入数据的压力水头进行对数处理和未处理两类,分别标号为No.1和No.2。

图2为利用六种机器学习算法训练得到含水量预测值与实测值的对比。压力水头输入为对数处理后的,即类别No.1。从图中可以看出,六种算法训练集的 R^2 都在0.99及以上。这表明训练集数据的预测值均接近实测值,而数据越贴近1:1回归线,表明算法性能越好。从测试集的 R^2 值大小对算法进行评价可知,LightGBM模拟SWCCs的性能最好,其测试集 R^2 为0.985;次之为GPR与XGB的测试集 R^2 分别为0.984与0.982,排第二与第三;随后为RF与AdaBoost,其测试集 R^2 分别为0.974与0.973,排第四与第五;SVM的测试集 R^2 为0.953,在六种算法中最低。

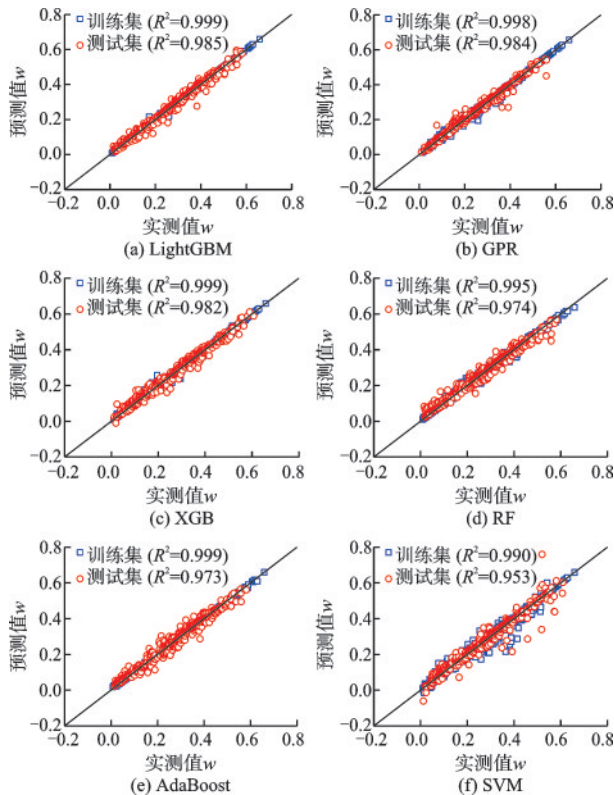


图2 基于压力水头对数形式下六种算法的预测值与实测值对比

Fig.2 Comparison of predicted and measured values of six algorithms based on logarithmic processing of matrix suction

表5 六种机器学习算法的输入和输出参数

Table 5 Input and output parameters of six machine learning algorithms

输入形式	训练集/ %	测试集/ %	输入	输出
No.1	80	20	$\ln(H), \rho_d, G_s, \theta_s, F_{clay}, F_{silt}$	w
No.2	80	20	$H, \rho_d, G_s, \theta_s, F_{clay}, F_{silt}$	w

表6为六种算法分别在对数/非对数压力水头两种情况下测试集的四项性能指标的统计结果。由表可知,在表5中两种输入情况下,四种集成学习算法的趋势评价指标(R^2 和EVS)和误差测量指标(MAE和RMSE)变化并不明显,其中LightGBM与XGB的 R^2 并未降低,而RF与AdaBoost的 R^2 只降低了0.002和0.003,说明将压力水头进行对数处理与否并不影响集成学习算法的预测性能。然而,对于传统机器学习算法,当压力水头不取对数后,GPR的 R^2 从0.982降低到了0.901,降低幅度为8.25%,SVM从0.954降低到了0.490,降低幅度为48.64%,说明将压力水头对数处理能大幅提高GPR与SVM的预测能力。

表6 不同输入形式下六种算法测试集的性能指标值
Table 6 Performance indicators of test sets for six algorithms under different input types

算法	输入形式	R^2	EVS	MAE	RMSE
LightGBM	No.1	0.985	0.985	0.011	0.017
	No.2	0.985	0.985	0.011	0.017
GPR	No.1	0.984	0.984	0.009	0.017
	No.2	0.901	0.901	0.026	0.041
XGB	No.1	0.982	0.982	0.013	0.018
	No.2	0.982	0.982	0.013	0.018
RF	No.1	0.974	0.974	0.015	0.022
	No.2	0.972	0.970	0.015	0.022
AdaBoost	No.1	0.973	0.973	0.016	0.022
	No.2	0.962	0.968	0.017	0.024
SVM	No.1	0.953	0.953	0.015	0.029
	No.2	0.490	0.493	0.058	0.098

基于上述分析,建议取对数下的压力水头数据进行后期数据的训练和测试。图3为基于压力水头对数形式下四个性能指标对六种机器学习算法的评估统计结果。从图中可知,LightGBM在预测含水量方面表现最好,拥有最高的趋势测量指标(R^2 和EVS),以及较低的误差测量指标(MAE和

RMSE),其次是GPR、XGB、RF与AdaBoost。而SVM在预测含水量方面表现出最差的预测性能,其趋势评价指标(R^2 和EVS)在所有算法中最低并且误差测量指标(MAE和RMSE)最高。

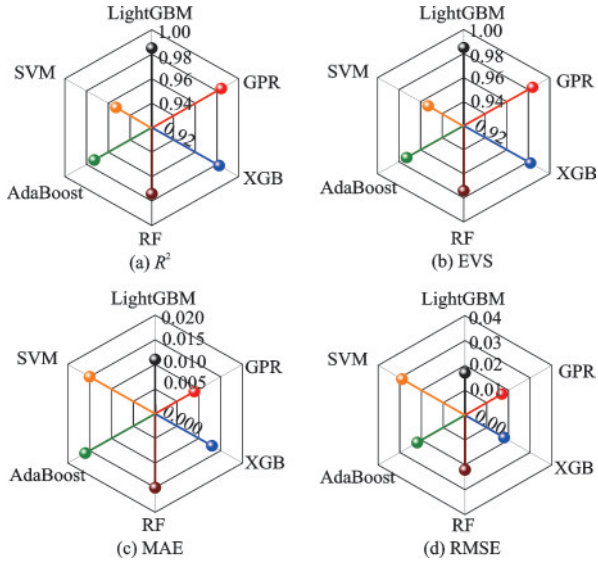


图3 基于压力水头对数形式的机器学习算法性能评估
Fig.3 Performance evaluation of machine learning algorithms based on logarithmic processing of matric suction

图4为基于压力水头对数形式下六个输入变量的特征重要性,从图中可以看出,压力水头在所有算法的预测过程中均表现出最高的重要性,相对重要性在39.74%到48.97%范围内,平均重要性为44.08%,其次是黏粒含量(24.97%)、饱和含水量(13.87%)、粉粒含量(7.90%)、干密度(5.78%)和比重(3.39%)。

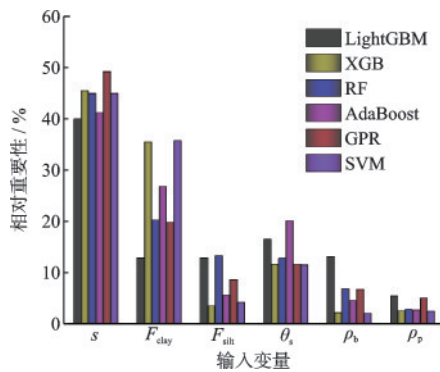


图4 基于压力水头对数形式的输入变量特征重要性分析
Fig.4 Feature importance analysis of input variables based on logarithmic processing of matric suction

为了进一步验证训练模型的预测效果。从UNSODA中随机选取9组未经训练的SWCCs数据(包含133个数据点)作为验证数据,验证数据的基

本物理特性如表7所示。使用训练好的LightGBM模型对这9组SWCC数据进行预测,预测结果如图5所示。从图中可以看出,整体预测的 R^2 与EVS分别为0.888与0.892,MAE为0.040, RMSE为0.047。表明LightGBM表现出良好的预测能力,预测值与实测值都与回归线贴近。

表7 验证数据的基本物理特性
Table 7 Basic physical properties of validation data

编码	输入变量				
	$\rho_d/(g \cdot cm^{-3})$	G_s	θ_s	$F_{clay}/\%$	$F_{silt}/\%$
2611	0.92	2.49	0.507	32.36	37.94
2614	1.72	2.83	0.383	37.39	38.51
2680	1.14	2.49	0.594	26.91	28.09
2740	0.73	2.48	0.698	27.3	41.16
2742	0.75	2.47	0.66	35.71	31.59
2743	0.71	2.48	0.701	43.25	33.1
4680	1.1	2.46	0.555	55.82	35.18
4681	1.08	2.48	0.578	62.11	28.49
4790	0.96	2.40	0.504	29.22	24.08

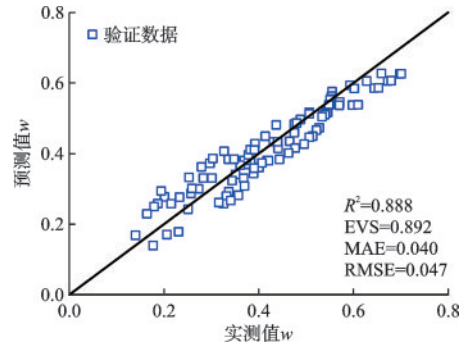


图5 LightGBM预测验证数据
Fig.5 LightGBM prediction of validation data

3 结论

(1) LightGBM对SWCCs测试集的模拟效果上均优于其他模型,拥有较高的趋势评价指标($R^2=0.985$ 和EVS=0.985)和较低的误差测量指标(MAE=0.011和RMSE=0.017);六种算法依次对SWCCs模拟优劣的排列顺序为:LightGBM、GPR、XGB、RF、AdaBoost和SVM。

(2) 在两种压力水头输入情况下,对LightGBM、XGB、RF和AdaBoost算法的影响很小;但是对GPR和SVM两种传统学习方法的影响很大,在

未进行对数处理情况下, R^2 降低明显甚至会出现无法模拟 SWCCs 的情况。

(3) 基于特征重要性分析可知, 压力水头、黏粒含量、饱和含水量和粉粒含量是影响机器学习算法预测 SWCC 的重要因素。其中压力水头是最重要的影响因素, 特征重要性在 39.74% 至 48.97% 之间, 平均重要性为 44.08%。

参考文献:

[1] 毕银强, 张茂省, 谷天峰, 等. 非饱和黄土土水特征及其强度指标[J]. 防灾减灾工程学报, 2018, 38(3): 520-527.
Bi Y Q, Zhang M S, Gu T F, et al. Soil-water characteristics and strength indexes of unsaturated loess[J]. Journal of Disaster Prevention and Mitigation Engineering, 2018, 38(3): 520-527. (in Chinese)

[2] 黎澄生, 孔令伟, 柏巍, 等. 土-水特征曲线滞后阻塞模型[J]. 岩土力学, 2018, 39(2): 598-604.
Li C S, Kong L W, Bai W, et al. Hysteresis model of soil-water characteristic curve[J]. Rock and Soil Mechanics, 2018, 39(2): 598-604. (in Chinese)

[3] 王叶娇, 王有为, 靳奉雨, 等. 石灰改良土的土水特征曲线及其冻结特征曲线[J]. 防灾减灾工程学报, 2020, 40(6): 967-973.
Wang Y J, Wang Y W, Jin F Y, et al. Soil-water characteristic curve and freezing characteristic curve of lime improved Soil[J]. Journal of Disaster Prevention and Mitigation Engineering, 2020, 40(6): 967-973. (in Chinese)

[4] Fredlund D G, Rahardjo H. Soil mechanics for unsaturated soils[M]. New York: John Wiley and Sons Inc, 1993.

[5] 刘倩倩, 李舰, 蔡国庆, 等. 全吸力范围的盐渍土持水特性的试验研究[J]. 岩土力学, 2021, 42(3): 713-722.
Liu Q Q, Li J, Cai G Q, et al. Experimental study on water retention characteristics of saline soil in the full suction range[J]. Rock and Soil Mechanics, 2021, 42(3): 713-722. (in Chinese)

[6] 桂跃, 刘锐, 赵振兴, 等. 高分解度泥炭土-土水特征曲线研究[J]. 防灾减灾工程学报, 2021, 41(3): 622-628.
Gui Y, Liu R, Zhao Z X, et al. Study on soil water characteristic curve of peaty soil with high degree of decomposition[J]. Journal of Disaster Prevention and Mit-

igation Engineering, 2021, 41(3): 622-628. (in Chinese)

[7] Gao Y, Sun D A. Soil-water retention behavior of compacted soil with different densities over a wide suction range and its prediction[J]. Computers and Geotechnics, 2017, 91: 17-26.

[8] Van Genuchten M T. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils[J]. Soil Science Society of America Journal, 1980, 44(5): 892-898.

[9] Fredlund D G, Xing A. Equations for the soil-water characteristic curve[J]. Canadian Geotechnical Journal, 1994, 31(4): 521-532.

[10] Or D, Tuller M. Liquid retention and interfacial area in variably saturated porous media: Upscaling from single-pore to sample-scale model[J]. Water Resources Research, 1999, 35(12): 3591-3605.

[11] 刘樟荣, 叶为民, 崔玉军, 等. 基于微孔填充和毛细管凝聚理论的持水曲线模型[J]. 岩土力学, 2021, 42(6): 1549-1556.
Liu Z R, Ye W M, Cui Y J, et al. Water retention curve model based on micro-pore filling and capillary condensation theories[J]. Rock and Soil Mechanics, 2021, 42(6): 1549-1556. (in Chinese)

[12] Zhai Q, Rahardjo H, Satyanaga A, et al. Estimation of wetting hydraulic conductivity function for unsaturated sandy soil[J]. Engineering Geology, 2021, 285: 106034.

[13] Hafez H, Teirelbar A, Kurda R, et al. Pre-bcc: A novel integrated machine learning framework for predicting mechanical and durability properties of blended cement concrete[J]. Construction and Building Materials, 2022, 352: 129019.

[14] 黄志, 周芙蓉, 陈娟, 等. 基于 IPOA 的巨型组合框架结构震损快速预测模型研究[J]. 防灾减灾工程学报, 2024, 44(2): 263-272.
Huang Z, Zhou F R, Chen J, et al. Rapid prediction model for seismic damage in mega composite frame structures based on IPOA methods[J]. Journal of Disaster Prevention and Mitigation Engineering, 2024, 44(2): 263-272. (in Chinese)

[15] Schaap M G, Bouten W. Modeling water retention curves of sandy soils using neural networks[J]. Water Resources Research, 1996, 32(10): 3033-3040.

(本文编辑: 苏泽云)

(下转第 118 页)