

基于机器学习算法的生物质水热炭特性的预测与评价

孙亮¹, 马中青¹, 张志霄², 胡艳军³, 王树荣⁴

(1.浙江农林大学 化学与材料工程学院 国家木质资源综合利用工程技术研究中心, 浙江 杭州 311300; 2.杭州电子科技大学 机械工程学院, 浙江 杭州 310018; 3.浙江工业大学 能源与动力工程研究所, 浙江 杭州 310023; 4.浙江大学 能源清洁利用国家重点实验室, 浙江 杭州 310027)

摘要: 文章利用从文献中收集的305组水热炭的基础特性数据,采用决策树、随机森林、梯度提升树3种机器学习算法建立水热炭基础特性的单任务和多任务预测模型,并利用SHAP法研究输入特征参数对水热炭基础特性影响的差异。结果表明:在3种机器学习算法中,梯度提升树模型在水热炭基础特性的多任务和单任务预测过程中,均体现出最高的准确性,测试集的平均相关系数分别为0.89和0.87,均方根误差分别为0.34和0.37;通过SHAP法对梯度提升树模型的输入特征参数进行评价,发现水热反应温度和原料中C元素含量是影响水热炭产率、高位热值和C元素含量的最主要参数。通过构建水热炭基础特性的预测模型,有利于优化水热炭制备工艺,降低实验成本,提高水热炭制备工艺的经济效益。

关键词: 生物质; 水热转化; 水热炭; 机器学习; 基本特性

中图分类号: TK6 **文献标志码:** A **文章编号:** 1671-5292(2024)11-1431-09

0 引言

水热炭是由生物质原料在水热条件下制备而成的新型炭材料,具有热值高、能量密度大和碳含量丰富等特点,被广泛制成燃料、污染物吸附剂和催化剂载体等^[1]。然而,水热炭化的制备工艺复杂多样,水热炭的性质极易受水热反应温度、时间、压力、催化剂类型和原料种类等因素的影响。因此,通常情况下需要通过大量的实验优化反应条件,以制备出性能较好的水热炭。然而,该方法存在周期长、成本高、结果不可控等缺陷,因此,通过构建水热炭的预测模型,对水热炭的制备工艺进行优化,将有利于降低实验成本,提高水热炭的经济效益。

近年来,机器学习方法已逐步成为生物质能源转化和利用方面的研究热点。以决策树集成算法(包括决策树、随机森林、梯度提升树等方法)和神经网络算法(包括人工神经网络、深度神经网络等方法)为代表的前沿机器学习方法^[2],在生物质能源领域的预测中取得了良好的应用效果,且相比传统方法能够得到更精确的结果。Zhu X Z^[3]采用随机森林和人工神经网络算法预测生物炭吸附

重金属离子的吸附效率,结果表明,随机森林算法的预测准确度更高,测试集的相关系数和均方根误差分别为0.97和0.06。Zhao S^[4]采用人工神经网络、支持向量机和随机森林等算法,预测生物质超临界水气化制氢的产量,结果表明,随机森林算法对H₂产量的预测准确度最高。

以上文献表明,相较于神经网络算法,在针对中小型数据集上,决策树集成算法具有适用性强、准确度高等优点。目前,决策树集成算法已在水热炭基本特性的预测中开展了初步应用。Jiang W^[5]使用水热炭化-热解共活化法制备了30组秸秆活性碳,采用线性回归、支持向量机、随机森林3种回归算法建立预测活性炭吸附性能的模型,结果表明,随机森林算法的预测准确度最高。Djandja O S^[6]收集了185条污泥水热炭化数据,采用随机森林算法预测水热炭的P含量,结果表明,随机森林算法预测精度较高,相关系数为0.92~0.95,生物质中P含量是影响模型预测性能的主要特征。

目前,关于水热炭特性预测的模型存在采集数据少、预测目标单一、模型解释性不足等问题,

收稿日期: 2023-06-21。

基金项目: 浙江省“领雁”研发攻关计划项目(2024C03225);国家自然科学基金面上项目(52376214)。

通信作者: 马中青(1987-),男,博士,副教授,研究方向为生物质热解和气化技术。E-mail: mazq@zafu.edu.cn

导致模型的预测能力和实际应用能力降低。本文从文献中收集了 305 组水热炭的基础特性数据, 采用决策树、随机森林、梯度提升树 3 种决策树集成算法建立了单任务和任务预测模型, 以生物质原料的基本特性和水热炭化反应条件作为输入特征, 对水热炭的质量产率、高位热值和 C 元素含量进行预测和评价, 同时利用 SHAP 法对预测准确度最高的机器学习模型进行解释, 阐述输入特征对水热炭性质影响的差异。

1 试验方法

1.1 数据搜集与预处理

利用 Web of Science 数据库, 检索到 28 篇生物质水热炭制备的文献, 并收集到 305 组水热炭实验结果, 将其用于构建数据集。水热炭特性的主要影响因素为生物质原料的基本特性 (元素分析和工业分析) 和水热炭化反应条件 (温度、时间和固液比), 水热炭的主要预测指标为水热炭的质量产率、高位热值和 C 元素含量。表 1 为数据集的

表 1 数据集的统计参数

Table 1 Statistical parameters of the collected data

数值类型	元素分析/%					工业分析/%			反应条件			水热炭特性		
	C	H	O	N	S	挥发分	灰分	固定碳	温度 ℃	时间 min	固液比	质量产率 %	高位热值 MJ/kg	C 元素含量 %
最大值	53.86	8.09	48.09	7.55	1.52	90.40	19.76	40.90	375.00	480.00	1.00	93.01	32.33	78.20
最小值	33.02	3.02	31.99	0.07	0.00	54.05	0.10	5.04	120.00	0.00	0.05	21.13	16.81	42.70
平均值	45.32	5.94	41.44	1.28	0.17	77.51	5.85	16.63	216.24	89.05	0.19	56.71	23.90	56.62
中位数	46.40	6.00	41.95	0.77	0.10	78.40	4.34	15.99	220.00	60.00	0.14	56.48	23.28	55.30
标准差	4.03	0.76	3.93	1.25	0.29	6.20	5.14	5.67	38.33	89.73	0.18	13.63	3.30	7.44

注: 元素分析以干燥无灰基为准; 工业分析以干燥基为准。

特征信息和相应的统计参数。

为了消除不同维度的特征数据对机器学习模型预测结果的影响, 在模型训练之前, 基于式 (1) 对收集的数据进行标准化。

$$x_i^* = \frac{x_i - x_{\min}}{x_{\text{std}}} \quad (1)$$

式中: x_i 为特征数据的初始值; x_{\min} 为 x_i 的平均值; x_{std} 为 x_i 的标准差; x_i^* 为标准化后的 x_i , 数据标准化处理后, 其均值为 0, 标准差为 1。

皮尔逊相关系数 (Pearson Correlation Coefficient, 记为 C_{PC}) 可以衡量两个连续性变量之间的线性相关性, C_{PC} 的取值为 $[-1, 1]$, 正值表示正相关, 负值表示负相关, 0 表示两个特征不相关, C_{PC} 的绝对值越接近 1, 表示相关性越强。 C_{PC} 的计算式为

$$C_{\text{PC}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

式中: \bar{X} 和 \bar{Y} 分别为样本 x_i 和 y_i 的平均值; n 为样本数量。

在开发机器学习模型之前, 使用 Python 编程

语言将整理好的数据集进行随机排列, 并将其随机拆分为训练集和测试集。选取 80% 的数据点作为训练集, 用于构建训练模型, 剩余 20% 的数据点作为测试集, 用于评估训练后机器学习模型的预测性能。

1.2 机器学习方法

本研究采用决策树、随机森林、梯度提升树 3 种机器学习算法构建水热炭特性的回归模型。决策树^[7]是一种用节点和分支来呈现数据分类和决策规则的算法, 表现为流程图式的分叉树结构。随机森林^[8]采用自助法将训练数据集随机划分成多个子集, 每个子集中的数据用于训练单个基本分类或回归树模型, 最终通过平均法或投票法获得预测结果。梯度提升树^[9]是将决策树作为弱学习器的集成算法, 通过梯度提升机制进行迭代训练, 不断调整训练样本权重以提高模型性能, 最终组合所有的弱学习器成为一个强学习器, 实现高精度的预测和回归分析。此外, 本文使用网格搜索法结合 5 倍交叉验证对机器学习算法中的超参数进行调优, 在成功建立具有最优超参数的机器学习模型后, 利用 SHAP 法对模型进行特征分析^[10]。SHAP 法通过计算数据中每个特征对模型输出的边际贡献, 衡量特征对预测结果线性和非线性的

影响规律。

选取相关系数(R^2)和均方根误差(Root Mean Square Error, 记为 E_{RMS})对机器学习模型的准确性进行评价。 R^2 是衡量模型拟合数据优劣程度的指标,取值为[0, 1], R^2 值越接近 1,表示模型拟合数据的效果越好,模型精度越高。 E_{RMS} 是衡量模型预测值和实际值差异的指标,反映了误差分布的离散程度, E_{RMS} 值越接近 0,表示预测值与实际值偏差越小,模型预测效果越好。 R^2 和 E_{RMS} 的计算式分别为

$$R^2=1-\frac{\sum_{i=1}^m (y_i^*-y_i)^2}{\sum_{i=1}^m (y_i^*-\bar{y}_i)^2} \quad (3)$$

$$E_{RMS}=\sqrt{\frac{1}{m} \sum_{i=1}^m (y_i^*-y_i)^2} \quad (4)$$

式中: m 为样本数量; y_i^* 为预测值; y_i 为实际值; \bar{y}_i 为 y_i 的平均值。

2 结果与讨论

2.1 数据集描述和分析

图 1 为数据集中输入和输出特征的统计数据箱型图,反映了生物质水热炭化过程中原始数据的分布情况。

由图 1(a)可知:生物质原料中 C 元素和 O 元素含量的中位数分别为 46.40%和 41.95%,而 N 元素和 S 元素含量的中位数仅为 0.77%和 0.10%,表明 C 元素和 O 元素含量远高于 N 元素

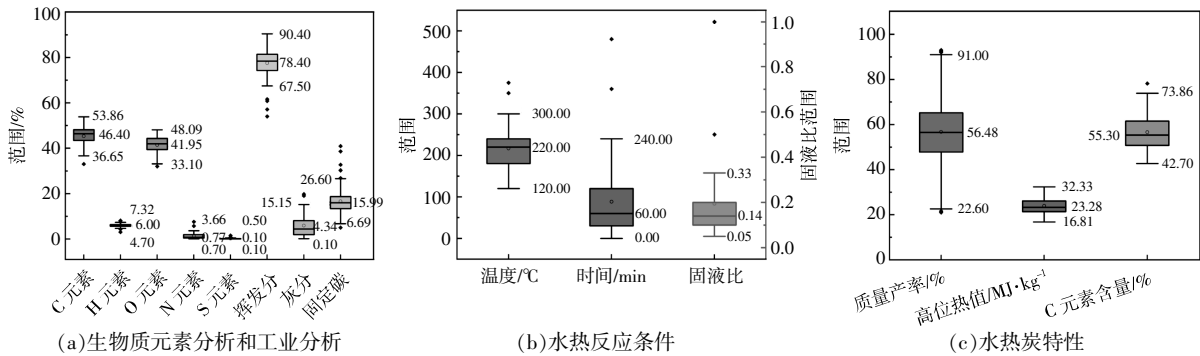


图 1 统计数据分布

Fig.1 The statistical data distribution

和 S 元素含量;挥发分含量主要在 67.50%~90.40%变化,波动范围最大,主要原因是数据集中包含多种类型的生物质,而不同类型的生物质挥发分含量差异较大;灰分含量主要在 0.10%~15.15%内波动,生物质中较高的灰分含量也会导致水热炭的灰分含量较高,将显著降低水热炭的热值^[1]。

由图 1(b)可知,水热反应温度和反应时间的取值分别为 120~375 °C和 0~480 min,固液比的取值为 0.05~1,固液比越大,底物浓度越高。数据集中采集的数据范围越广,表明输入特征多元化程度更高,可以提供更多的信息和多样性。由图 1(c)可知:水热炭的质量产率和高位热值的中位数分别为 56.48%和 23.28 MJ/kg,表明水热反应可以制取较高质量产率和热值的水热炭,因为在超临界的水热反应状态下,生物质可通过热降解反应转化为高能量密度的水热炭产物^[12];相比于生物

质原料,水热炭中的 C 元素含量显著提高,平均 C 元素含量为 56.62%,且随着水热反应温度的升高,C 元素含量逐渐增加,最高达到 78.20%,主要原因是较高的水热反应温度促进了生物质中纤维素、半纤维素和木质素的热降解,水热炭的石墨化程度逐渐提高。

图 2 为任意两个变量之间的皮尔逊相关系数矩阵。

由图 2 可知:绝大部分特征变量之间的 C_{PC} 值较低(-0.3~0.3),呈现弱相关性,表明特征变量彼此之间是独立的,有助于提高模型的稳定性和可靠性;C 元素和 H 元素含量之间呈现正相关性, C_{PC} 为 0.33,主要原因是生物质中存在大量的 C-H 键;C 元素和灰分含量之间呈现负相关性, C_{PC} 为 -0.76,这是因为生物质中的 C 通常来源于有机物质,而灰分则是由生物质中的无机盐和矿物质等组成;固定碳含量分别与灰分和挥发分含量之间

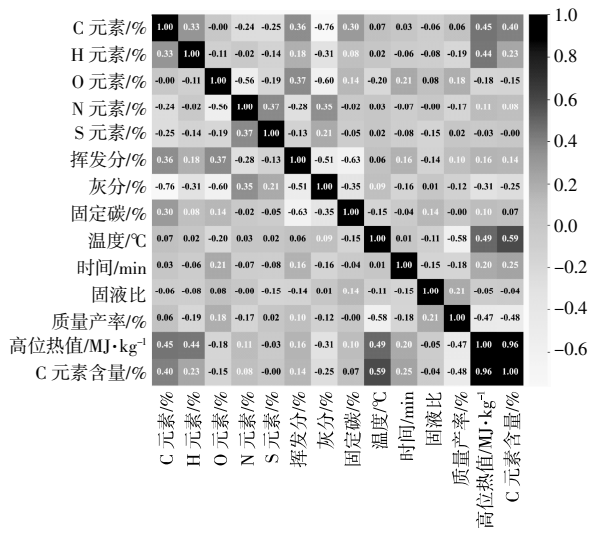


图2 皮尔逊相关系数矩阵

Fig.2 Matrix of pearson correlation coefficient

存在显著的负相关性, C_{PC} 分别为-0.35和-0.63, 主要是因为固定碳含量总是通过总含量减去挥发分和灰分含量获得; 水热炭的高位热值和C元素含量与生物质原料的C, H元素含量和挥发分含量均呈现正相关, C_{PC} 分别为0.45, 0.44, 0.11和0.40, 0.23, 0.14, 而与O元素含量和灰分含量呈现负相关, C_{PC} 分别为-0.18, -0.31和-0.15, -0.25, 表明生物质原料中较高的C, H元素和挥发分含量有利于提升水热炭的高位热值和C元素含量; 水热炭的质量产率与反应温度和反应时间之间的 C_{PC} 分别为-0.58和-0.18, 呈现负相关性, 而与固液比的 C_{PC} 为0.21, 呈现正相关性, 水热炭的高位热值和C元素含量则表现出相反的趋势; 水热炭的高位热值和C元素含量之间呈现强的正相关性, C_{PC} 高达0.98, 而与水热炭质量产率则呈现明显的负相关关系, C_{PC} 为-0.50。总体而言, 大多数目标变量之间存在一定的相关性, 表明可以采用多任务机器学习模型预测水热炭特性。

2.2 模型性能评估

单任务预测中各个机器学习模型的预测结果见表2。

由表2可知: 梯度提升树和随机森林模型对水热炭特性的预测准确度较高, 训练集和测试集的 R^2 值均在0.95和0.80以上, E_{RMS} 值均在0.21和0.50以下, 表明模型能够很好地拟合原始数据; 在3种模型中, R^2 值的大小依次为梯度提升树模型(0.85~0.9)>随机森林模型(0.80~0.85)>

表2 单任务机器学习预测模型性能总结
Table 2 Results of the model performance of the single-target machine learning prediction

机器学习模型	预测目标	训练集		测试集	
		R^2	E_{RMS}	R^2	E_{RMS}
决策树	质量产率	0.97	0.17	0.77	0.54
	高位热值	0.96	0.20	0.65	0.61
	C元素含量	0.95	0.21	0.74	0.56
随机森林	质量产率	0.96	0.20	0.80	0.50
	高位热值	0.96	0.19	0.85	0.37
	C元素含量	0.95	0.21	0.85	0.40
梯度提升树	质量产率	0.98	0.12	0.85	0.44
	高位热值	0.98	0.14	0.88	0.33
	C元素含量	0.98	0.15	0.90	0.33

决策树模型(0.65~0.77), 而 E_{RMS} 值的大小依次为梯度提升树模型 (0.54~0.61)>随机森林模型 (0.37~0.50)>决策树模型(0.33~0.44), 表明采用梯度提升树模型预测水热炭特性的准确率最高, 而决策树模型的准确率最低; 当采用梯度提升树模型预测水热炭的3个指标时, C元素含量的 R^2 值最高(0.90), E_{RMS} 值最低(0.33), 表明C元素含量的预测准确度最高, 而质量产率的 R^2 最低(0.85), E_{RMS} 值最高(0.44), 表明质量产率的预测准确度最低。总之, 机器学习模型对水热炭质量产率预测的准确度相对较低, 而对水热炭高位热值和C元素含量预测的准确度相对较高, 主要原因是水热炭产率的样本数据存在较大差异, 模型在训练过程中可能更偏向于处理典型的样本数据, 而对于具有差异性的样本数据处理效果不佳, 进而影响模型预测的准确性。因此, 在进行水热炭特性的预测时, 需要充分考虑不同类型的特征, 以及特征与模型的适配情况, 从而提高模型预测的准确性和可靠性。

相比于单任务预测, 多任务预测利用多个目标任务共同训练模型, 该方法可以提高模型的鲁棒性和预测能力, 因为模型可以从多个任务中挖掘出更多的信息和特征。不同机器学习模型在多任务预测中的预测结果见表3。

由表3可知: 在水热炭特性的多任务预测方面, 3种机器学习模型的预测准确度与单任务预测类似, 训练集和测试集的平均 R^2 值分别大于0.94和0.62, 平均 E_{RMS} 值分别小于0.25和0.61; 在3种机器学习模型中, 平均 R^2 值的大小依次为

表3 多任务机器学习预测模型性能总结

Table 3 Results of the model performance of the multi-target machine learning prediction

机器学习模型	预测目标	训练集 R^2	平均值	训练集 E_{RMS}	平均值	测试集 R^2	平均值	测试集 E_{RMS}	平均值
决策树	质量产率	0.93		0.26		0.71		0.60	
	高位热值	0.95	0.94	0.23	0.25	0.63	0.66	0.59	0.25
	C元素含量	0.94		0.24		0.62		0.64	
随机森林	质量产率	0.96		0.20		0.76		0.54	
	高位热值	0.96	0.96	0.20	0.21	0.84	0.82	0.39	0.21
	C元素含量	0.95		0.21		0.84		0.41	
梯度提升树	质量产率	0.96		0.20		0.85		0.43	
	高位热值	0.96	0.96	0.20	0.20	0.91	0.89	0.29	0.20
	C元素含量	0.95		0.22		0.92		0.29	

梯度提升树模型(0.89)>随机森林模型(0.82)>决策树模型(0.66),而平均 E_{RMS} 值的大小依次为梯度提升树模型(0.34)<随机森林模型(0.45)<决策树模型(0.61),表明梯度提升树模型在水热炭特性的多任务预测中准确率最高,而决策树模型的准确率最低,主要原因是梯度提升树模型能够更好地适应复杂的样本特征,并且在预测过程中可以较好地处理多个任务。结合表2,3可以看出,当采用梯度提升树模型预测水热炭的3个指标时,

多任务预测模型中的质量产率、高位热值和C元素含量的 R^2 值均高于单任务预测模型,表明多任务预测模型更稳定、准确度更高。因此,梯度提升树模型的预测准确度最高,更适用于水热炭特性的多任务预测。

图3为原始实验数据与梯度提升树模型预测值之间的分布散点图(粗虚线为预测值等于实际值时的最佳拟合线,细虚线区域内为10%的误差范围的数值)。

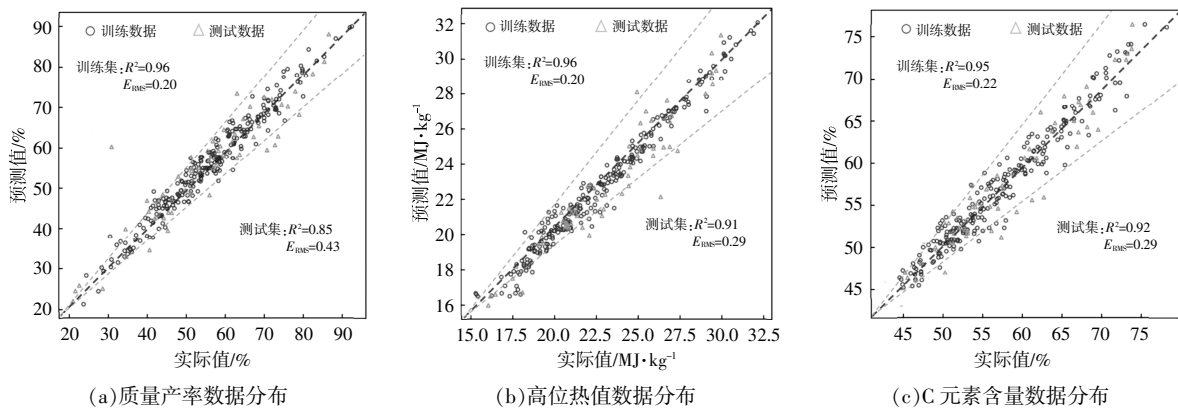


图3 原始实验数据与预测数据的散点图

Fig.3 Scatter plots of predicted data and original experimental data

由图3可以看出,数据点分布较为密集,表明模型预测结果具有较好的稳定性。对于训练数据集,梯度提升树模型在水热炭特性的预测过程中的 R^2 值均高于0.95,训练数据点也主要分布在10%误差范围内,位于最佳拟合线两侧,表明梯度提升树模型在训练数据集上拟合度较好,稳定性较高。对于测试数据集,模型预测的 R^2 值较低(0.85~0.92), E_{RMS} 值较高(0.29~0.43),测试数据点分布较为分散,这是机器学习模型的固有特性导致,该结果与 Mu L^[13]和 Kapetanakis T N^[14]对水热

炭特性的预测结果相近。

2.3 模型解释和分析

将输入变量分为水热反应条件、元素分析和工业分析,采用SHAP法对多任务预测的梯度提升树模型进行特征分析。图4为水热炭输入特征参数绝对SHAP值的平均值分布图,绝对SHAP值的平均值越高,表明该特征参数对预测结果的影响越大。

由图4可知,在水热炭特性的预测中,3类特征类别的重要性程度依次为水热反应条件

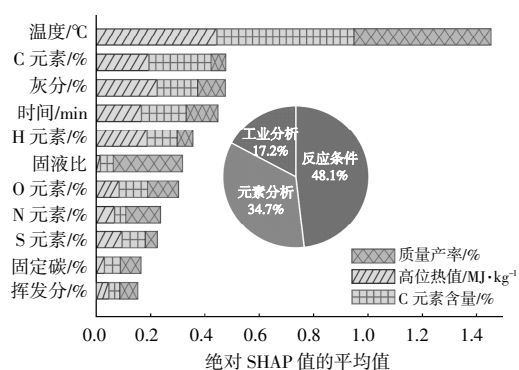
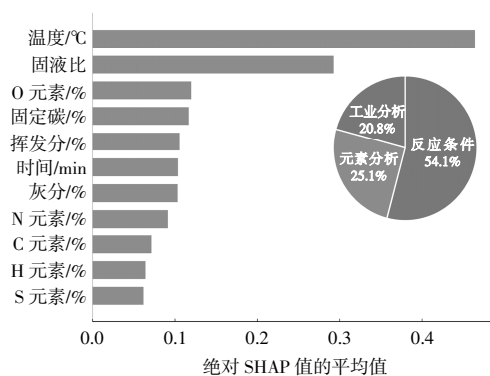


图4 各输入特征参数绝对 SHAP 值的平均值

Fig.4 The mean of absolute SHAP value for characteristic parameters

(48.1%)>元素分析(34.7%)>工业分析(17.2%),表

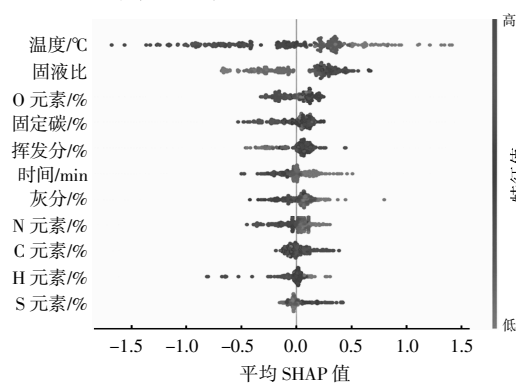


(a)特征重要性排序

明水热反应条件对水热炭的性能影响最大。在水热反应条件下,水热反应温度绝对 SHAP 值的平均值高于反应时间和固液比,表明水热反应温度对水热炭的特性影响最大,主要原因是水热反应温度直接影响生物质的水解速率和程度^[15]。在元素分析中,生物质原料的 C 元素含量对水热炭的特性影响最大,其次为 H 和 O 元素含量,因为在水热炭化过程中,生物质极易发生脱水和脱羧反应。在工业分析中,生物质原料的灰分含量对水热炭的特性影响最大,其次为固定碳和挥发分含量。

图5为输入特征的相对重要性及其对水热炭质量产率的总体影响。

由图5(a)可知,水热反应条件对水热炭质量



(b)特征总体影响趋势

图5 输入特征的相对重要性及其对水热炭质量产率的总体影响

Fig.5 Effect of each feature and its overall impact on the yield of hydrochar

产率预测的重要性达到54.1%,而生物质元素分析和工业分析的重要性分别为25.1%和20.8%,表明水热反应条件对于水热炭质量产率的影响显著。而在水热反应条件中,水热温度是预测水热炭产率过程中最重要的特征参数,固液比对预测水热炭产率也有较大的影响;在生物质元素分析中,O元素含量对预测结果影响较大,H和S元素含量的影响相对较小;在生物质工业分析中,相比灰分含量,固定碳和挥发分含量具有更高的重要性。

由图5(b)可知,水热反应温度和水热炭质量产率的预测结果呈现出明显的线性关系,水热反应温度越高,其SHAP值越低,水热炭质量产率的预测值越低,因为在更高的水热温度条件下,生物质原料的水解程度增强,从而导致原料中更多的有机物分解并转化为气相和液相产物^[16]。固液比对水热炭质量产率的预测起正向作用,固液比越高,其SHAP值越高,水热炭质量产率的预测值也

越高,这是因为较高的底物浓度严重抑制了水解反应,导致水热反应不充分,使得水热炭质量产率较高。而质量产率高的水热炭的热值较低,表明水热炭质量产率与高位热值呈现出相反的变化趋势,此结果与皮尔逊相关系数矩阵图(图2)得到的结果相一致。生物质中O元素含量对水热炭质量产率的预测起负向作用,生物质中O元素含量越高,水热炭的质量产率预测值越低。生物质中固定碳和挥发分含量对水热炭质量产率的影响不同,固定碳含量越低,挥发分含量越高,水热炭质量产率的预测值越高。水热反应时间对水热炭质量产率的预测起负向作用,水热反应时间越长,水热炭质量产率的预测值越低,主要是因为充分的反应时间促使生物质原料中的半纤维素和纤维素发生脱水和解聚反应,生成许多挥发性物质和中间产物,如酸、醇、糠醛等。

图6为输入特征的相对重要性及其对水热炭

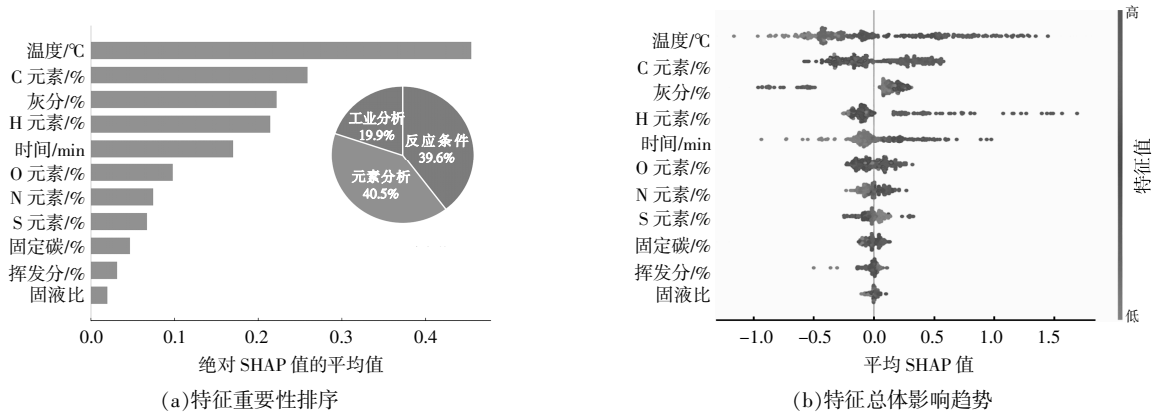


图 6 输入特征的相对重要性及其对水热炭高位热值的总体影响

Fig.6 Effect of each feature and its overall impact on the higher heating value of hydrochar

高位热值的总体影响。

由图 6(a)可知,生物质元素分析是水热炭高位热值预测模型最重要的特征类别,重要性达到 40.5%,而水热反应条件和生物质工业分析的重要性仅分别为 39.6%和 19.9%。在水热反应条件中,反应温度是预测水热炭高位热值过程中最重要的特征参数,其次是反应时间;在生物质元素分析中,C,H 和 O 元素含量对水热炭高位热值的预测结果影响较大,N 和 S 元素含量的影响相对较小;而在工业分析中,灰分含量对水热炭高位热值的预测具有最高的重要性。

由图 6(b)可知,水热反应温度对水热炭高位热值的预测有正面影响,水热反应温度越高,其 SHAP 值越高,水热炭高位热值的预测值也越高,主要原因是原料中的纤维素和半纤维素随着水热炭化程度的加深而逐渐分解,从而留下含量更高的木质素成分,基于木质素更好的热稳定性,水热炭的高位热值显著提高^[17,18]。生物质中 C 和 H 元

素含量对水热炭高位热值的预测有正面影响,生物质中 C 和 H 元素含量越高,水热炭高位热值的预测值越高,主要是因为几乎所有的热量都是由 C 和 H 的氧化而产生,且水热炭高位热值的升高也是由 H-C,O-C 等低键能化学键的减少和 C-C 等高键能化学键的增加引起^[19]。生物质中灰分含量对水热炭高位热值的预测有负面影响,生物质中灰分含量越高,水热炭高位热值的预测值越低,说明原料中过高的灰分含量不利于高热值水热炭的形成。水热反应时间的延长也可提高水热炭的高位热值,主要原因是反应时间越长,组分热降解越充分,碳化程度越深。

图 7 为输入特征的相对重要性及对水热炭 C 元素含量的总体影响。

由图 7(a)可知,水热反应条件是水热炭 C 元素含量预测模型最重要的特征类别,特征重要性达到 46.4%,而生物质元素分析和工业分析的重要性仅分别为 37.3%和 16.3%。在水热反应条件

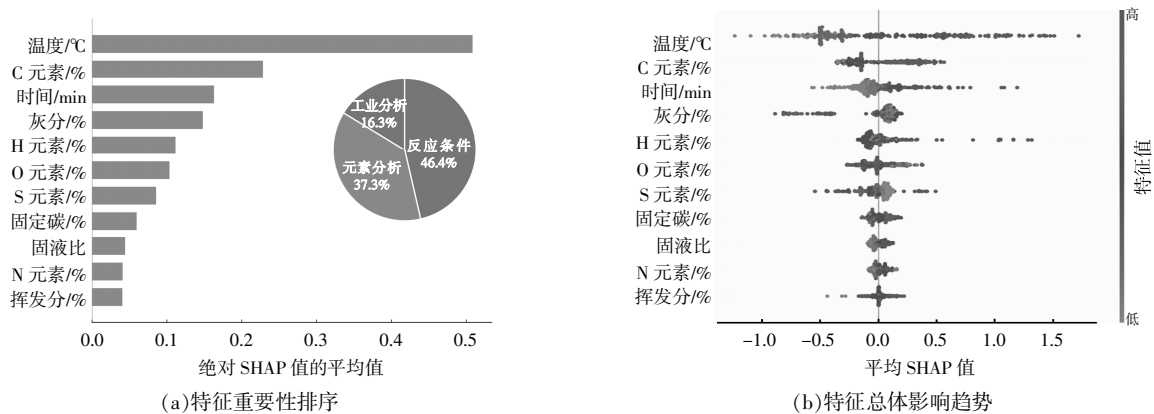


图 7 输入特征的相对重要性及其对水热炭碳含量的总体影响

Fig.7 Effect of each feature and its overall impact on the carbon content of hydrochar

中,反应温度是预测水热炭 C 元素含量过程中最重要的特征参数,其次是反应时间;在生物质元素分析中,C,H 和 O 元素含量对水热炭 C 元素含量的预测结果影响较大,N 元素含量的影响相对较小;而在工业分析中,相比固定碳和挥发分含量,灰分含量对水热炭 C 元素含量的预测具有更高的重要性。水热炭 C 元素含量的特征重要性排名与水热炭高位热值的预测结果高度相似,表明水热炭的高位热值和 C 元素含量之间存在密切的相关性,并且受到相似的特征因素影响。

由图 7(b)可知,水热反应温度和反应时间对水热炭 C 元素含量的预测有正面影响,主要原因是水热炭化过程会发生大量的脱水、脱羟基和脱羧基反应,导致生物质中的大部分 H 和 O 元素以 H_2O 和 CO_2 的形式脱除^[20]。与反应时间相比,反应温度对水热炭 C 元素含量影响更显著,主要是因为脱水反应仅在部分时间范围内受到明显的影响。生物质中 C 元素含量对水热炭 C 元素含量有正面影响,生物质中 C 元素含量越高,对应的 SHAP 值越大,生成的水热炭的 C 元素含量也越高,说明水热炭化过程中生物质中的大部分 C 元素会得到保留,同时表明不同种类的生物质原料对水热炭特性具有不同影响。生物质中灰分含量对水热炭 C 元素含量有负面影响,原因是生物质中较高的灰分含量可能会影响水热炭化过程中的脱羧和芳构化反应^[21]。生物质中 H 和 O 元素含量对水热炭 C 元素含量的预测结果具有不同的影响,更高的 H 含量和更低的 O 含量有利于高 C 含量的水热炭的形成。

3 结论

本文利用文献数据构建水热炭的数据集,使用 3 种决策树集成算法建立水热炭特性的单任务和多任务预测模型,并利用 SHAP 法对预测准确度最高的机器学习模型进行解释分析。研究结果表明:梯度提升树在水热炭特性的多任务预测上表现出最好的预测性能,其平均相关系数和均方根误差分别为 0.89 和 0.34;基于 SHAP 法研究最优模型的输入特征,发现水热反应条件中的反应温度和生物质原料中的 C 元素含量对预测结果影响最大。因此,基于现有的文献数据,机器学习技术可有效地预测出水热炭的特性,从而优化水热炭的制备工艺。

参考文献:

- [1] 王一宁,石岩,李恒,等.生物质废弃物水热碳化资源化应用研究进展[J].中外能源,2022,27(9):23-29.
- [2] Khan M,Naqvi S R,Ullan Z,et al. Applications of machine learning in thermochemical conversion of biomass-A review[J].Fuel,2023,332:126055.
- [3] Zhu X Z,Wang X N,Ok Y S. The application of machine learning methods for prediction of metal sorption onto biochars[J].Journal of Hazardous Materials, 2019,378:120727.
- [4] Zhao S,Li J,Chen C,et al. Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass [J].Journal of Cleaner Production,2021,316:128244.
- [5] Jiang W, Xing X J, Li S, et al. Synthesis, characterization and machine learning based performance prediction of straw activated carbon [J]. Journal of Cleaner Production,2019,212:1210-1223.
- [6] Djandja O S,Salami A A,Wang ZC,et al. Random forest -based modeling for insights on phosphorus content in hydrochar produced from hydrothermal carbonization of sewage sludge [J].Energy,2022,245:123295.
- [7] Charbuty B,Abdulazeez A. Classification based on decision tree algorithm for machine learning[J].Journal of Applied Science and Technology Trends,2021,2(1):20-28.
- [8] Probst P,Wright M N,Boulesteix A L. Hyperparameters and tuning strategies for random forest [J].Wiley Interdisciplinary Reviews:Data Mining and Knowledge Discovery,2019,9(3):e1301.
- [9] Leng L,Zhang W,Chen Q Y,et al. Machine learning prediction of nitrogen heterocycles in bio-oil produced from hydrothermal liquefaction of biomass[J].Bioresour Technol,2022,362:127791.
- [10] Onsree T,Tippayawong N,Phithakkitnukoon S,et al. Interpretable machine -learning model with a collaborative game approach to predict yields and higher heating value of torrefied biomass [J].Energy, 2022,249:123676.
- [11] Nunes L,Matias J,Catalao J. Biomass combustion systems:A review on the physical and chemical properties of the ashes [J].Renewable and Sustainable Energy Reviews,2016,53:235-242.
- [12] Zhuang X,Song Y,Zhan H,et al. Pyrolytic conversion

- of biowaste -derived hydrochar: decomposition mechanism of specific components[J].*Fuel*,2020,266: 117106.
- [13] Mu L, Wang Z, Wu D, et al. Prediction and evaluation of fuel properties of hydrochar from waste solid biomass: Machine learning algorithm based on proposed PSO-NN model[J].*Fuel*,2022,318:123644.
- [14] Kapetanakis T N, Vardiambasis I O, Nikolopoulos C D, et al. Towards engineered hydrochars: Application of artificial neural networks in the hydrothermal carbonization of sewage sludge [J].*Energies*,2021,14 (11):3000.
- [15] Wang T, Zhai Y, Zhu Y, et al. A review of the hydrothermal carbonization of biomass waste for hydrochar formation: Process conditions, fundamentals, and physicochemical properties [J].*Renewable and Sustainable Energy Reviews*,2018,90:223-247.
- [16] 孙畅.FeCl₃ 改性水热炭用于生物滞留池脱氮除磷效果[D].天津:天津大学,2018.
- [17] Parshetti G K, Hoekman S K, Balasubramanian R. Chemical, structural and combustion characteristics of carbonaceous products obtained by hydrothermal carbonization of palm empty fruit bunches [J].*Bioresource Technology*,2013,135:683-689.
- [18] 陈鹏辉,郭建斌,李芳,等.水热预处理对木质纤维素类废弃物厌氧发酵产甲烷及能量平衡的影响[J].*可再生能源*,2022,40(10):1279-1287.
- [19] He M Y, Hu Z Q, Xiao B, et al. Hydrogen-rich gas from catalytic steam gasification of municipal solid waste (MSW): Influence of catalyst and temperature on yield and product composition [J].*International Journal of Hydrogen Energy*,2009,34(1):195-203.
- [20] 刘云云,曹运齐,余强,等.园林废弃物水热炭燃料特性研究[J].*太阳能学报*,2022,43(7):439-444.
- [21] Buss W, Jansson S, Wurzer C, et al. Synergies between BECCS and biochar—Maximizing carbon sequestration potential by recycling wood ash [J].*ACS Sustainable Chemistry & Engineering*,2019,7(4):4204-4209.

Prediction and evaluation of the basic properties of biomass hydrochar using the machine learning algorithms

Sun Liang¹, Ma Zhongqing¹, Zhang Zhixiao², Hu Yanjun³, Wang Shurong⁴

(1.College of Chemistry and Materials Engineering, National Engineering Research Center for Wood-based Resource Comprehensive Utilization, Zhejiang A & F University, Hangzhou 311300, China; 2.School of Mechanical Engineering, Hangzhou Dianzi University, Hangzhou 310018, China; 3.Institute of Thermal and Power Engineering, Zhejiang University of Technology, Hangzhou 310023, China; 4.State Key Laboratory of Clean Energy Utilization, Zhejiang University, Hangzhou 310027, China)

Abstract: In this work, 305 sets of data of hydrochar's basic properties was collected from the references. Then, the single-task and multi-task prediction models of hydrochar's basic properties (mass yield, higher heating value, and carbon content) were established based on three types of the machine learning algorithms (the decision tree, the random forest, and the gradient boosting decision tree). Results showed that among the three types of the machine learning algorithms, the gradient boosting decision tree model was the best algorithm, where the average determination coefficient values of the test set were 0.88 and 0.87, and the root mean square error values were 0.34 and 0.37. The SHAP method was used to evaluate the input characteristic parameters during the modeling by using the gradient boosting decision tree. The dominant influence factors for the prediction of the mass yield, higher heating value, and carbon content of the hydrochar were the hydrothermal reaction temperature and the C content in raw biomass. The construction of the prediction model of the hydrochar's basic properties was favorable to reduce the cost for the optimization of the hydrochar production conditions.

Keywords: biomass; hydrothermal conversion; hydrochar; machine learning; basic properties