引用格式:于天鑫,彭璇. 基于机器学习高通量筛选吸附甲烷的金属有机框架材料[J]. 北京化工大学学报(自然科学版),2021,48(2):100 - 107.

YU TianXin, PENG Xuan. High throughput screening of metal-organic framework materials based on machine learning[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2021,48(2):100 - 107.

基于机器学习高通量筛选吸附甲烷的金属有机框架材料

于天鑫 彭 璇*

(北京化工大学 信息科学与技术学院, 北京 100029)

摘 要: 采用决策树(DT)模型及其衍生的随机森林(RF)模型、极端随机树(ET)模型和梯度提升树(GBDT)模型,对用于甲烷吸附的金属有机框架材料(MOFs)进行了高通量的计算筛选。利用 1 800 种材料的特征向量数据,计算了特征向量之间的相关性并进行重要度分析,发现材料的结构特征与化学信息特征的相关性不大,但是结构特征的重要度较高。将数据库中的 1 260 种材料作为训练集并使用上述 4 种机器学习模型进行训练,再将剩余的 540 种材料作为测试集对模型的筛选结果进行比较和评估。接收者操作特征(ROC)曲线和查准率-查全率(PR)曲线结果表明,GBDT 模型自身稳定性强且预测结果精度高,因而成为筛选吸附甲烷的高性能金属有机框架材料的最佳模型。针对 RF 模型和 GBDT 模型进行参数优化,发现协调单个决策树的个数和决策树节点的分裂特征数量能够有效改善 RF 模型的性能,而调节回归树的学习速率和迭代次数可有效改善 GBDT 模型性能。最后基于 540 种材料的测试集,利用 GBDT 模型筛选出前 20 种高性能吸附材料,分析了它们的主要特征向量与甲烷吸附量之间的关系。

关键词: 甲烷吸附; 金属有机框架材料; 机器学习; 高通量筛选

中图分类号: TM712 **DOI**: 10.13543/j.bhxbzr.2021.02.013

引 言

近年来,甲烷作为一种清洁燃料越来越被人们所重视,而采用金属有机框架材料(MOFs)实现甲烷的吸附[1-3]和储存也引起了较为广泛的关注。随着实验室制备的 MOFs 以及计算机虚拟合成的 MOFs 的数量呈现爆发式的增长,仅仅利用巨正则系综蒙特卡洛模拟(GCMC)方法[4-5]实现高性能吸附材料的高通量计算筛选已经无法满足要求。

基于 GCMC 的高通量筛选方法往往受限于庞大的 MOFs 数据库和有限的计算资源,因此,具有强大数据分析和挖掘能力的机器学习方法被研究者们用来进行高效的 MOFs 高通量筛选研究[6-8]。基于此,本文采用机器学习建模的方法,通过决策树(DT)模型及其衍生的随机森林(RF)模型、极端随机树(ET)模型和梯度提升树(GBDT)模型这 4 种模型对吸附甲烷的 MOFs 材料进行高通量的计算筛选以选择出最佳性能材料;对两种较优模型(RF 模型和 GBDT 模型)的参数优化进行了探究,并推荐了合适的材料结构特征参数。

1 实验部分

1.1 数据库的选择

目前,MOFs 数据库基本上可划分为两类,即由实验合成的 MOFs(eMOFs)所组成的数据库和由计算机合成的 MOFs(hMOFs)所组成的数据库。尽管通过计算机合成的 hMOFs 为 MOFs 的种类提供了无限的可能,但是 hMOFs 数据库中的材料仅有一小部分能够在实验中合成,绝大部分 hMOFs 设计的合理性和可行性存在着很大问题,导致无法通过实验

收稿日期: 2020-08-25
基金项目: 国家自然科学基金(21676006)
第一作者: 女,1995 年生,硕士生
* 通信联系人
E-mail: pengxuan@mail.buct.edu.cn

合成相应的材料。

本文采用 eMOFs 数据库[9-10]，实验数据集中包含 1 800 个真实的 MOFs 数据样本，其中每一种 MOFs 由 9 种特征描述符来表征，即表 1 中的前 6 种结构描述符和后 3 种化学信息描述符。通过 GCMC 模拟计算每种材料在温度 298 K 和压力 35 bar (1 bar = 0.1 MPa) 下的甲烷吸附量，根据美国能源局对吸附甲烷的金属有机框架材料在该条件下的划分标准，将吸附量高于 180 (单位气体吸附量与单位材料的体积比) 的数据样本标记为高性能材料，反之，则标记为低性能材料。

表 1　每种材料特征向量的描述符表示

Table 1　Descriptors used to construct a feature vector for each material

特征向量	描述符
孔体积/($cm^3 \cdot g^{-1}$)	Pv
密度/($g \cdot cm^{-3}$)	Ds
比表面积/($cm^2 \cdot g^{-1}$)	Sa
限制孔径/Å	PLD
最大孔径/Å	MPD
主导孔径/Å	DPD
电负性比率	Er
金属占比(原子数分数)/%	Mp
不饱和度	Du

1.2　数据库的分析

1.2.1　相关性分析

本文计算了每个描述特征之间的相关性，如图 1 所示。

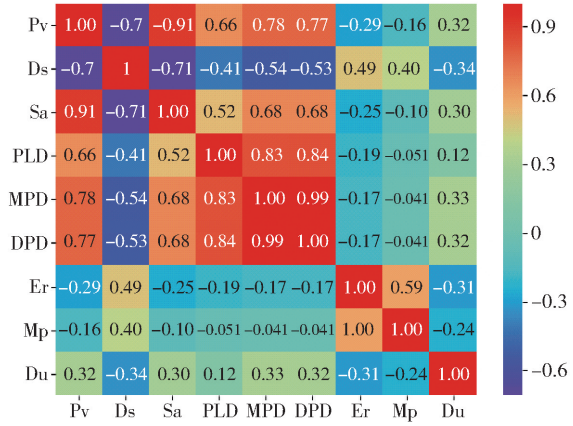

图 1　特征向量的相关性

Fig. 1　Correlation of feature vectors

从图 1 可以看出，材料的最大孔径 (MPD) 和主导孔径 (DPD) 的相关性非常强，达到了 99%。由此可见，绝大多数材料的最大孔径和主导孔径是一致的。其次，可以看出每种材料的孔体积 (Pv) 和比表面积 (Sa) 的相关性也比较强，达到 91%，实际上，当材料的孔径较大时，其相应的比表面积也会增大，以支撑 MOFs 的有机骨架结构，从而更好地实现对甲烷的吸附。与此同时，对于化学信息描述特征来说，它们之间的相关性都不高，而且与结构描述特征的相关性也不强。鉴于两者是从不同的角度对材料信息的提取，因此应该结合结构特征与化学信息特征共同完成材料的筛选。

1.2.2　重要度分析

基于构造决策树时分裂节点的原理[11]，进一步计算每个特征描述符对甲烷吸附能力的重要度。在每棵树的节点分裂时需要选择该节点的分裂特征，通过计算基尼系数来确定节点特征，基尼系数越小，划分的纯度越高，则节点特征越好，特征的重要度就越高。树的节点特征的顺序就是重要度的顺序。从图 2 可以看出，MOFs 材料的孔体积 (Pv) 对材料的吸附能力的重要度最高，这是因为材料的孔体积增大，甲烷的吸附量也会相应增加。除此之外，结构特征描述符对甲烷吸附的重要度较高，影响较大，而由于甲烷是非极性分子，材料的化学信息描述符对于甲烷吸附的重要度较小。因此，结构特征对于甲烷吸附材料性能的影响更大。

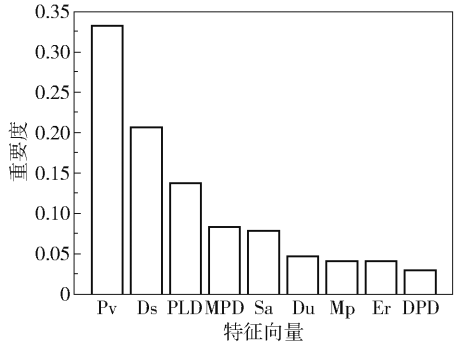

图 2　特征向量对甲烷吸附的重要度

Fig. 2　Importance of feature vectors for methane adsorption

1.3　实验模型的选择

数据库中的很多材料由于结构原因导致某些特征无法测量，存在有缺省值问题，此外当按照分类标准划分时，存在高、低性能材料数量不平衡的问题，极有可能造成数学模型的不稳定。相比于其他机器学习的算法，由单棵决策树衍生出的多

棵决策树是采用集成的学习方法,利用该方法建立模型对数据的要求相对较低,输出的结果更加可靠。为了比较不同机器学习算法的筛选能力,本文选择了决策树基础模型,及由它改进而来的随机森林、极端随机树和梯度提升树 3 种树模型,随机地将数据集划分为训练集和测试集两组,采用普遍的 7∶3的划分方式,即训练集和测试集的材料数分别为 1 260 种和 540 种。利用不同的机器学习方法对训练集进行学习,并使用建立的模型对测试集的数据进行筛选预测。

2 结果与讨论

2.1 模型分析与评价

2.1.1 混淆矩阵计算

通过模型对材料的测试集进行筛选,计算各个模型的混淆矩阵[12-13]。从表 2 中各模型混淆矩阵的计算结果可以看出其分类效果,例如,GBDT 模型在低性能材料的分类结果中,有 375 种材料分类正确,21 种材料分类错误;而在高性能材料的分类结果中,有 135 种材料分类正确,9 种材料分类错误。比较 4 个模型的混淆矩阵,发现它们的错误分类数量大小顺序为 DT > ET > RF > GBDT,GBDT 模型的误分个数明显低于其他模型。

表 2 4 种模型的混淆矩阵

Table 2 Confusion matrix for four models

模型	分类正确数量[a)]	分类错误数量[a)]	分类正确数量[b)]	分类错误数量[b)]	混淆矩阵
DT	382	37	91	30	382 37 30 91
ET	390	37	85	28	390 37 28 85
RF	390	36	91	23	390 36 23 91
GBDT	375	21	135	9	375 21 9 135

a—低性能样本;b—高性能样本。

2.1.2 接收者操作特征(ROC)曲线

图 3 给出了各个模型的 ROC 曲线,该曲线可以用来衡量模型的拟合程度[14]。由图 3 可以看出,随着误诊率的增加,召回率也逐渐增加。召回率 T 与误诊率 F 的计算公式如式(1)、(2)所示。

$$T = \frac{TP}{TP + FN} \tag{1}$$

$$F = \frac{FP}{FP + TN} \tag{2}$$

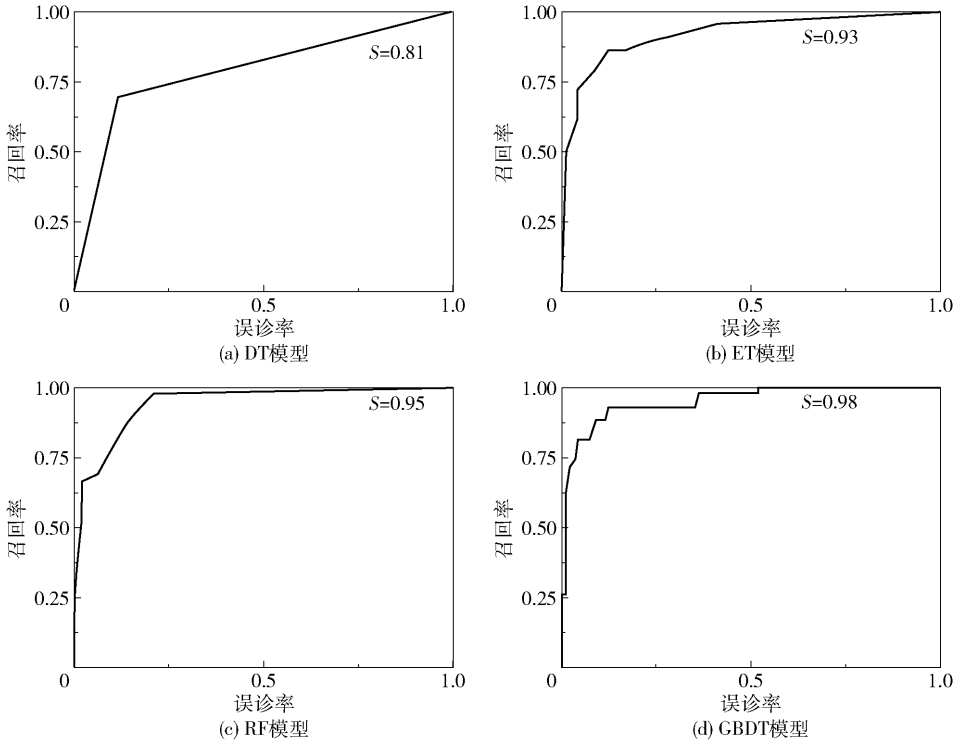

图 3 4 种模型的 ROC 曲线

Fig. 3 ROC curves of four models

式中，*TP* 表示样本的真实类别是正例，并且模型将其预测成为正例的数量；*FN* 表示样本的真实类别是负例，并且模型将其预测成为负例的数量；*TN* 表示样本的真实类别是正例，模型将其预测成为负例的数量；*FP* 表示样本的真实类别是负例，模型将其预测成为正例的数量。对于每一个模型，我们希望其有一个较高的召回率以及较低的误诊率，所以图 3 中每一个图形的拐点越接近左上方则模型的效果越好，也即曲线与横坐标轴围成的面积越大越好。DT、ET、RF 以及 GBDT 这 4 个模型曲线与横坐标轴所围成的面积分别为 0.81、0.93、0.95 和 0.98。从面积上看，GBDT 模型曲线的拐点更加靠近左上方，所围成的面积最大，表明 GBDT 模型比其他模型的拟合效果更好。

2.1.3　查准率–查全率(PR)曲线

由于材料数据库中低性能的材料较多，高性能的材料较少，这种较差的样本均衡性会对模型的筛选造成一定的影响。因此，可以通过 PR 曲线来反映样本均衡性对模型的影响[15]。4 种模型的查准率–查全率曲线如图 4 所示，查全率 *R* 以及查准率 *P* 的计算公式如(3)、(4)所示。

$$R = T = \frac{TP}{TP + FN} \tag{3}$$

$$P = \frac{TP}{TP + FP} \tag{4}$$

可以看出，随着查全率的不断增加，查准率则在不断下降。对于一个较好的模型而言，应该有较高的查全率及查准率，即 PR 曲线的拐点尽量靠近右上方，使曲线与横坐标轴及左边框围成的面积越大越好。4 种模型的 PR 曲线所围成的面积大小顺序为 DT < ET < RF < GBDT，表明 GBDT 模型优于其他模型。对于 GBDT 模型，其 PR 曲线所围面积为 0.94，仅仅比其 ROC 曲线所围面积减少了 0.04，而 RF 模型的相应减少值为 0.09，其他两个模型的减少值也都较大。因此可以表明，GBDT 模型受数据集的样本不平衡的影响较小，模型稳定性较高。

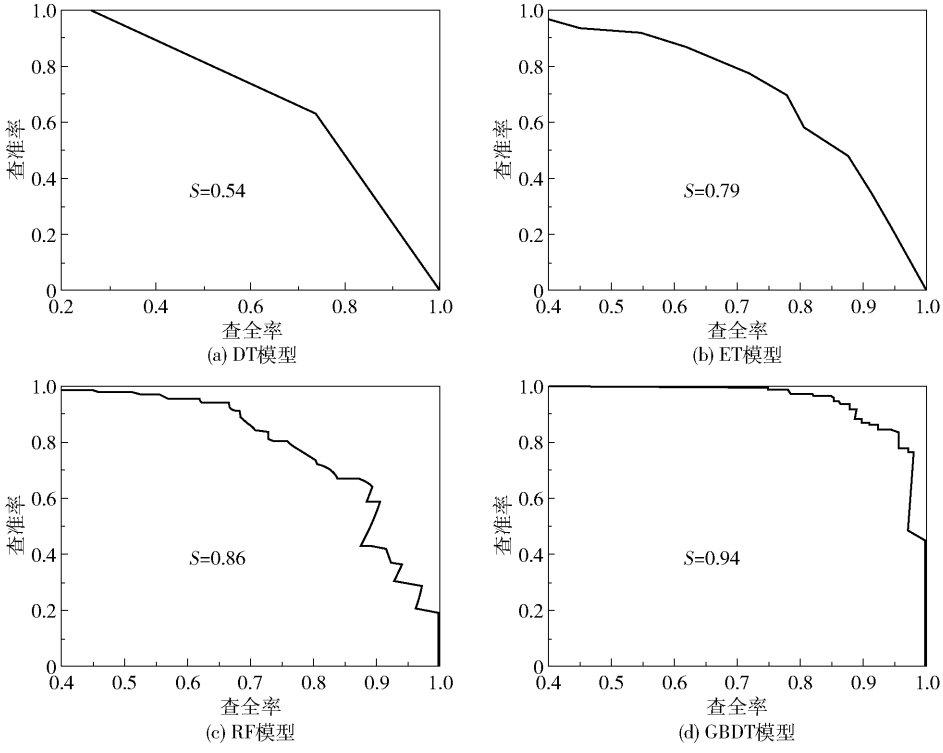

图 4　4 种模型的 PR 曲线

Fig. 4　PR curves of four models

2.2　模型测试结果

2.2.1　测试集

基于 DT、RF、ET 和 GBDT 这 4 种机器学习模型对 540 种材料构成的测试集进行高性能甲烷吸附材料的筛选。从表 3 可以看出，利用 4 种机器学习模型筛选的类别为 0 的低性能材料，其各项指标普遍

比筛选出的类别为 1 的高性能材料要高,原因在于在训练集中进行高低性能的分类时,低性能材料的数量远多于高性能材料的数量,导致 4 种模型对于高性能材料的学习不充分,故而针对高性能材料筛选的效果不明显。4 种模型筛选的准确度大小顺序为 DT < ET < RF < GBDT。可见,相比于由单棵决策树组成的 DT 模型,由 DT 模型衍生出的多棵决策树组成的其他 3 种模型的分类效果更明显。而在由决策树衍生的分类模型中,GBDT 的准确度达到 0.96,筛选效果要好于 ET 和 RF 模型。

表 3 4 种模型测试集的比较

Table 3 Comparison of four models with test set

模型	类别	P	T	$F1$ 分数[a)]	准确度[b)]
DT	0	0.90	0.86	0.88	0.82
	1	0.61	0.70	0.65	
RF	0	0.95	0.91	0.93	0.90
	1	0.76	0.84	0.79	
ET	0	0.94	0.94	0.93	0.88
	1	0.68	0.81	0.74	
GBDT	0	0.97	0.98	0.98	0.96
	1	0.94	0.93	0.93	

a—$\frac{2}{F_1}=\frac{1}{P}+\frac{1}{T}$;b—准确度 $A=\frac{TP+TN}{TP+TN+FP+FN}$。

2.2.2 学习曲线

RF 是基于套袋(bagging)的思想,有放回地均匀取样,而 GBDT 则是基于梯度提升(boosting)的思想,根据训练错误率对样本赋予不同的权重。实验所选取的验证集是在数据训练进行有放回抽取时未被抽取的数据的集合,这些未被抽到的材料数据称作袋外数据[16]。绘制 RF 和 GBDT 这两种较优模型的学习曲线,如图 5 所示。由图可知,GBDT 模型相对于 RF 模型的学习效果更好。在 RF 模型中,训练集的准确度在训练过程中基本保持不变,说明该模型在训练过程中拟合程度较好;而交叉验证集的准确度则是从较低的数值逐渐上升的,且并没有无限接近训练集的准确度,两者之间的间距较大,导致误差比较大。也即在训练过程中,RF 模型的拟合准确度非常高,达到 100%,但是在交叉验证过程中仅达到 90% 左右。这说明 RF 模型对于新的数据集适应性较差,存在过拟合的问题。而对于 GBDT 模型,训练集的准确度在训练过程中有微小的下降,而交叉验证集的准确度则有所上升,且两者有向同一准确度值靠近的趋势(两条数据线趋近的准确度值在 95% 左右)。由此可见,GBDT 模型能够改善 RF 模型中存在的过拟合现象。

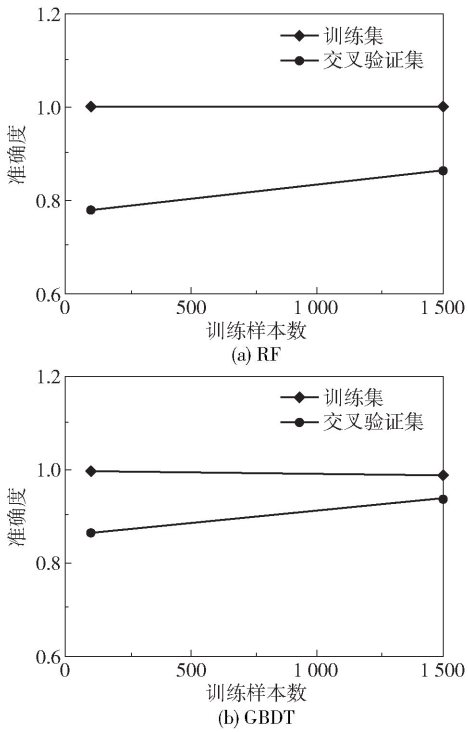

图 5 RF 与 GBDT 模型的学习曲线

Fig. 5 Learning curves of RF and GBDT models

2.3 模型参数讨论

2.3.1 RF 模型参数曲线

图 6 给出了弱学习器的数量 n_e 与特征向量的种类数 F_{max} 对袋外错误率 $R_{e,OOB}$ 的影响。袋外错误率是测试数据误差的无偏估计,用来检验 RF 模型的泛化能力。同时,选择 3 种不同的方式计算每棵树节点上的特征数量,设总特征数为 k,3 种方式分别为 $F_{max}=\log_2 k$,$F_{max}=\sqrt{k}$,$F_{max}=k$。从图 6 可以看出,3 条曲线的变化趋势大致相同,都是随着 n_e 的增加,$R_{e,OOB}$ 不断下降。当 n_e 的值增加到 70 时,曲线 c 的 $R_{e,OOB}$ 要明显高于曲线 a、b,说明在每棵树的分裂节点上选择数据集的全部特征很容易使得模型更加复杂,容易产生过拟合现象,造成 $R_{e,OOB}$ 较高。而在分裂节点上对于特征的数量的选择则可有效降低模型的复杂程度,避免发生模型的过拟合现象。当 n_e 的值大于 100 时,曲线 a 的 $R_{e,OOB}$ 值可以降到最低。当 $n_e=150$,且 $F_{max}=\log_2 k$ 时,就能够有效改善 RF 模型的性能。

2.3.2 GBDT 模型参数曲线

影响 GBDT 模型拟合效果的两个最重要的因素分别为迭代次数 n 和每棵回归树的学习速率 l,因此

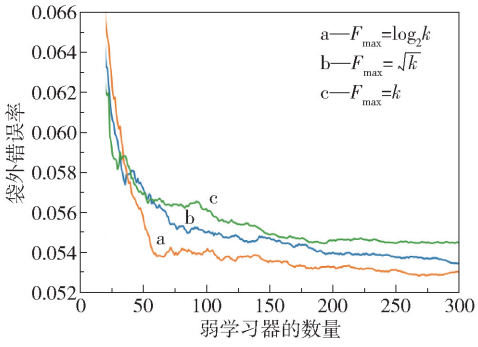

图 6　不同参数对 RF 模型的影响

Fig. 6　Effect of different parameters on the RF model

本文考察了这两个因素对测试集偏差 e 的影响。从图 7 可以看出，不同学习速率下曲线的变化趋势大致相同，即随着 n 的增加，e 值是逐渐减小的。当 n 小于 20 时，3 条曲线的 e 值下降得非常快；而且 $l=0.5$ 时对应的偏差值是最低的，说明在有限的迭代次数内，l 越高，所达到的测试效果越好。而在 n 大于 40 时，$l=0.5$ 曲线对应的 e 值一直维持在较高的水平，而$l=0.1$和 $l=0.2$ 曲线仍然有下降的趋势。$l=0.2$ 曲线在迭代 100 次左右后偏差达到最低，而 $l=0.1$曲线在迭代 140 次左右时偏差达到最低。这说明当回归树的 l 较低时，要增加 n 的值才能保证 e 值降低。因此，在训练模型时，可以调节回归树的学习速率 l 和迭代次数 n 两个参数来改善 GBDT 模型。在本文测试中，当 $n=100$，$l=0.2$ 时，可以达到有效改善模型性能的效果。

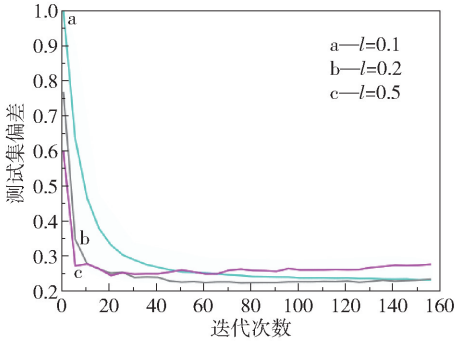

图 7　不同参数对 GBDT 模型的影响

Fig. 7　Effect of different parameters on the GBDT model

2.3.3　高性能吸附材料的特征向量

对影响甲烷吸附量的重要度进行分析发现，影响甲烷气体吸附的主要因素为材料的孔体积、密度、限制孔径及最大孔径。利用 GBDT 模型筛选测试集内的高性能材料，分析前 20 种高性能材料的特征向量与甲烷吸附量之间的关系，结果如图 8 所示。从图中可以看出，当孔体积为 0.5 ~ 0.75 cm^3/g，限制

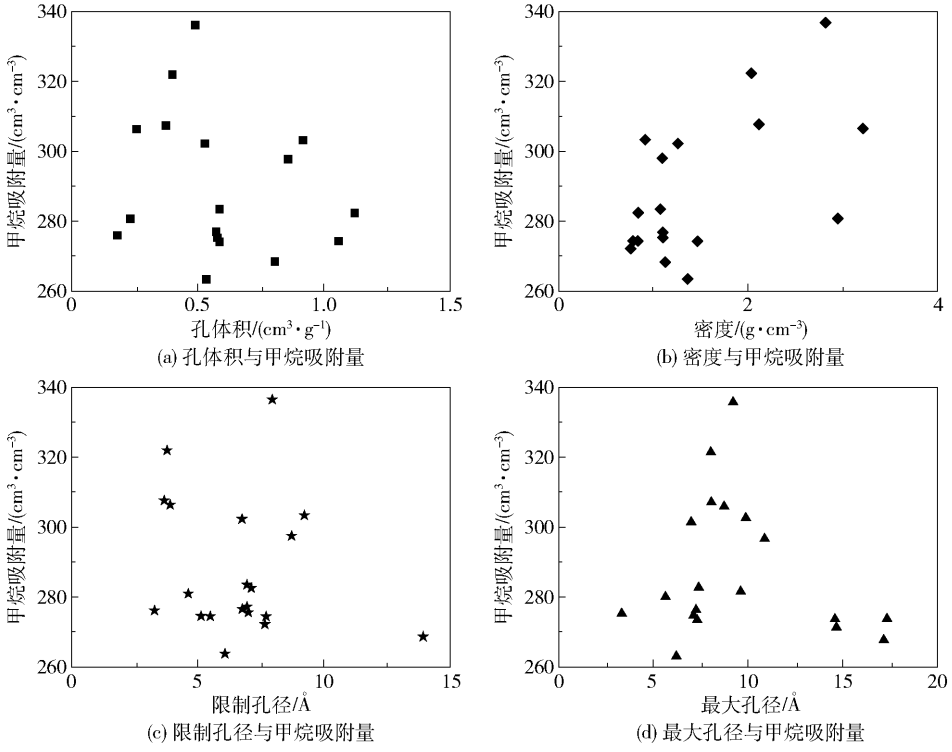

图 8　高性能材料的特征向量与甲烷吸附量的关系

Fig. 8　Relationship between the feature vectors and methane adsorption of high-performance materials

密度为 2 ~ 3 g/cm^3,材料孔径在 4 Å 左右,最大孔径在6 ~ 10 Å 时,甲烷的吸附量较高。

3 结论

本文采用 DT 模型及其衍生的 RF、ET、GBDT 模型对金属有机框架材料进行分类筛选,通过对模型的筛选性能进行比较,得出 GBDT 模型的筛选效果最好。当迭代次数为 100,学习速率为 0.2 时,GBDT 的模型性能最佳。利用 GBDT 模型筛选出的前 20 种金属有机框架材料进行构效关系分析,得出当孔体积为 0.5 ~ 0.75 cm^3/g,材料密度为 2 ~ 3 g/cm^3,材料限制孔径在 4 Å 左右,最大孔径在 6 ~ 10 Å 时,甲烷的吸附量较高。所得结果可望为用于甲烷吸附的金属有机框材料的设计提出合理化建议。

参考文献:

[1] PENG X, CHENG X, CAO D P. Computer simulations for the adsorption and separation of $CO_2/CH_4/H_2/N_2$ gases by UMCM-1 and UMCM-2 metal organic frameworks [J]. Journal of Materials Chemistry, 2011, 21(30): 11259 - 11270.

[2] ALTINTAS C, ERUCAR I, KESKIN S. High-throughput computational screening of the metal organic framework database for CH_4/H_2 separations[J]. ACS Applied Materials and Interfaces, 2018, 10(4): 3668 - 3679.

[3] DAGLAR H, KESKIN S. Computational screening of metal-organic frameworks for membrane-based $CO_2/N_2/H_2O$ separations: best materials for flue gas separation [J]. The Journal of Physical Chemistry C, 2018, 122(30): 17347 - 17357.

[4] AZAR A N V, VELIOGLU S, KESKIN S. Large-scale computational screening of metal organic framework (MOF) membranes and MOF-based polymer membranes for H_2/N_2 separations [J]. ACS Sustainable Chemistry and Engineering, 2019, 7(10): 9525 - 9536.

[5] SEZGINEL K B, UZUN A, KESKIN S. Multivariable linear models of structural parameters to predict methane uptake in metal-organic frameworks[J]. Chemical Engineering Science, 2015, 124: 125 - 134.

[6] SIMON C M, MERCADO R, SCHNELL S K, et al. What are the best materials to separate a xenon/krypton mixture? [J]. Chemistry of Materials, 2015, 27(12): 4459 - 4475.

[7] OHNO H, MUKAE Y. Machine learning approach for prediction and search: application to methane storage in a metal - organic framework [J]. The Journal of Physical Chemistry C, 2016, 120(42): 23963 - 23968.

[8] GUSTAFSON J A, WILMER C E. Intelligent selection of metal-organic framework arrays for methane sensing via genetic algorithms [J]. ACS Sensors, 2019, 4(6): 1586 - 1593.

[9] SARKISOV L, BUENO-PEREZ R, SUTHARSON M, et al. Materials informatics with PoreBlazer v4.0 and the CSD MOF database[J]. Chemistry of Materials, 2020, 32: 9849 - 9867.

[10] GÜLSOY Z, SEZGINEL K B, UZUN A, et al. Analysis of CH_4 uptake over metal-organic frameworks using data-mining tools[J]. ACS Combinatorial Science, 2019, 21(4): 257 - 268.

[11] 胡学钢, 李楠. 基于属性重要度的随机决策树学习算法[J]. 合肥工业大学学报(自然科学版), 2007, 30(6): 681 - 685.
HU X G, LI N. A random decision tree algorithm based on attribute significance[J]. Journal of Hefei University of Technology, 2007, 30(6): 681 - 685. (in Chinese)

[12] 赵存秀. 基于混淆矩阵的分类器性能评价指标比较[J]. 电子技术与软件工程,2020(13):146 - 147.
ZHAO C X. Comparison of performance evaluation indexes of classifiers based on confusion matrix [J]. Electronic Technology and Software Engineering, 2020 (13): 146 - 147. (in Chinese)

[13] RAHMAD F, SURYANTO Y, RAMLI K. Performance comparison of anti - spam technology using confusion matrix classification[J]. IOP Conference Series: Materials Science and Engineering, 2020, 879(1):012076.

[14] 王彦光,朱鸿斌,徐维超. ROC 曲线及其分析方法综述[J]. 广东工业大学学报,2021,38(1):46 - 53.
WANG Y G, ZHU H B, XU W C. A review on ROC curve and analysis[J]. Journal of Guangdong University of Technology, 2021, 38 (1): 46 - 53. (in Chinese)

[15] LIU Z K, BONDELL H D. Binormal precision - recall curves for optimal classification of imbalanced data[J]. Statistics in Biosciences, 2019,11(1): 141 - 161.

[16] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5 - 32.

High throughput screening of metal-organic framework materials based on machine learning

YU TianXin　PENG Xuan*

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: High throughput screening of metal organic frameworks (MOFs) for methane adsorption has been carried out using a decision tree (DT) model and its derived random forest (RF) model, an extreme random tree (ET) model and a gradient lifting tree (GBDT) model. Using the eigenvector data of 1 800 kinds of materials, the correlation and importance were calculated. It was found that the structural characteristics of materials had little correlation with chemical information characteristics, but the importance of the structural characteristics of materials was higher. 1 260 kinds of materials in the database were used as training sets and the four machine learning models were used for training, and the remaining 540 materials were used as test sets to compare and evaluate the screening results of the models. On the basis of the receiver operating characteristic (ROC) curve and the precision recall (PR) curve, it is found that the GBDT model has strong stability and high prediction accuracy, making it the best way to select MOF materials for adsorption of methane. For the parameter optimization of RF and GBDT models, it was found that the coordination of the number of single decision tree and the number of split features of decision tree nodes can effectively improve the performance of RF model, while adjusting the learning rate and iteration times of regression tree can effectively improve the performance of GBDT model. Based on the test set of 540 materials, the relationship between the main feature vector and methane adsorption capacity was analyzed by using the first 20 high performance adsorption materials screened by the GBDT model.

Key words: methane adsorption; metal-organic framework material; machine learning; high throughput screening