云南大学学报(自然科学版), 2022, 44(6):1127~1134 DOI: 10.7540/j.ynu.20210642
Journal of Yunnan University: Natural Sciences Edition

抗乳腺癌活性化合物的 ADMET 性质预测模型

秦雅琴, 夏玉兰, 卢梦媛, 王锦锐, 谢济铭**
(昆明理工大学 交通工程学院, 云南 昆明 650500)

摘要: 为提升抗乳腺癌药物虚拟筛选过程中吸收(absorption)、分配(distribution)、代谢(metabolism)、排泄(excretion)、毒性(toxicity)等属性的预测能力, 提出一种抗乳腺癌药物定量结构−ADMET 性质预测模型. 首先, 从化合物的分子描述符数据中遴选出对 ADMET 性质具有影响的 319 个特征变量; 然后, 以逻辑回归(Logistic Regression, LR)、朴素贝叶斯(Naïve Bayes, NB)、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)作为 ADMET 分类预测的候选模型, 筛选出 GBDT 模型为最优模型; 最后, 针对 GBDT 模型训练成本较高的问题, 借助概率代理模型拟合超参数与预测精度之间的关系 (即黑箱模型) 构建 GBDT*模型. 结果显示, GBDT*集成学习模型整体表现最优, 准确率、精准率、灵敏度、AUC 指标分别达 90%、88%、89%、0.95 以上, 误报率低于 15%, 表明 GBDT*集成机器学习模型在抗乳腺癌活性化合物的 ADMET 性质预测方面具有良好的性能.

关键词: 计算机应用; 集成学习; 特征筛选; 超参数优化; 抗癌候选药物
中图分类号: TP399 **文献标志码:** A **文章编号:** 0258−7971(2022)06−1127−08

在抗乳腺癌药物的研发过程中, 由于某些化合物的一些药代动力学性质(ADMET)无法被预见, 即药物的吸收(Absorption)、分配(Distribution)、代谢(Metabolism)、排泄(Excretion)、和毒性(Toxicity), 可能会降低药物研发效率, 造成大量的资源浪费[1-2]. 而常规的生物试验方法常常代价高昂且耗时长[3-4], 随着药物信息学技术及数据挖掘技术的不断发展, 利用累积的药物实验数据进行 ADMET 性质建模可快速对批量化合物进行处理和预测[5].

机器学习因其能有效捕获非线性数据的内在规律, 从复杂的 ADMET 数据中学习化学结构与药效学的关联, 成为药物化学领域用来解决复杂化合物性质预测问题的一个重要方法[6-7]. Chi 等[5] 使用支持向量回归方法(Support Vector Regression, SVR)解决了药物吸收预测时输入和输出之间的非线性问题. 李晓等[8] 针对人体小肠吸收、血脑屏障透过等多个 ADMET 相关的性质, 使用支持向量机(Support Vector Machine, SVM)分别建立适用于小分子化合物的预测模型. 但 SVM 对参数调节和和函数的选择相当敏感, 极易影响预测精度. 莫贤炜等[9] 对苯基哌嗪类 5-HT_7 受体拮抗剂进行三维定量构效关系 (Quantitative Structure-Activity Relationship, QSAR) 分析及 ADMET 相关性质的预测, 用于受体拮抗药物的预测与筛选. Tsou 等[10] 将深度神经网络 (Deep Neural Networks, DNN) 用于三阴性乳腺癌 (Triple-Negative Breast Cancer, TNBC) 抑制剂药物的虚拟筛选, 由于 DNN 结构里下层神经元和所有上层神经元都能够形成连接, 若调参不当, 易导致参数数量膨胀. Feinberg 等[11] 利用包含多个图卷积层的图卷积神经网络学习每个化合物分子式的图示化特征向量, 并应用于 ADMET 性质预测, 取得了较好的准确性. Dahl 等[12] 使用分子描述符和药效团指纹作为分子特征, 训练并构建了基于随机森林 (Random Forest, RF) 和逻辑回归 (Logistic Regression, LR) 的 ADMET 预测模型. Dong 等[13] 利用多个数据库得到的海量化合物数据, 分别构建了基于朴素贝叶斯 (Naïve Bayes, NB) 和决策树 (Decision Tree, DT) 的预测模型, 结果表明, 在面对多分类任务或数据特征缺失时, NB 模型能够表现出良好的鲁棒性. 以上研究多使用较为经典的机器

收稿日期: 2022-01-15; **接受日期:** 2022-05-24; **网络出版日期:** 2022-07-15
基金项目: 云南省交投集团科技研发项目 (YCIC-YF-2021-05).
作者简介: 秦雅琴(1972−), 女, 湖南人, 博士, 教授, 主要研究人因安全、机器学习. E-mail: qinyaqin@kust.edu.cn.
** **通信作者:** 谢济铭(1994−), 男, 甘肃人, 博士, 科研助理, 主要研究图像智能分析、系统建模优化. E-mail: xiejiming@kust.edu.cn.

学习模型,其模型结构可以继续优化以更加适应化合物的 ADMET 属性预测任务.

基于此,本文从数据样本与特征约束条件出发,构建基于 LR、NB、GBDT 模型的 ADMET 性质预测模型,并挑选出最优模型 GBDT. 同时考虑上述经典模型中超参数设置对预测精度、调参时间等的影响,提出改进最优模型 GBDT*. 经对比验证,发现超参数调优算法可有效发挥分类 GBDT*模型最优性能,研究成果有助于抗乳腺癌候选药物的 ADMET 性质预测.

1 ADMET 性质预测模型的设计

针对抗乳腺癌活性化合物的 ADMET 性质预测问题,需收集一系列作用于乳腺癌治疗靶标的化合物数据,然后以化合物的诸多分子结构描述符作为输入变量,选取 ADMET 性质中表征人体对化合物渗透吸收能力的 Caco-2 性质、表征化合物在人体内的代谢能力的 CYP3A4 性质、表征化合物对心脏毒副作用的 hERG 性质进行建模,并定义化合物各性质的表现程度为二分类变量,例如'Caco-2=1'代表小肠上皮细胞对该化合物具有较好的渗透吸收能力,'Caco-2=0'代表小肠上皮细胞对该化合物渗透吸收能力较差;'CYP3A4=1'代表人体对该化合物具有代谢能力,'CYP3A4=0'代表对该化合物无代谢能力;'hERG=1'代表该化合物具有心脏毒性,'hERG=0'代表该化合物无心脏毒性. 基于此,构建基于逻辑回归和机器学习方法的化合物 ADMET 预测模型,筛选出最优模型,并采用超参数调优的方法对优选模型进行优化处理,作为最终的 ADMET 性质分类预测模型,进行抗乳腺癌活性化合物的 ADMET 性质预测. 模型框架如图 1 所示.

2 ADMET 性质预测模型

将降维处理后的 N 个化合物分子描述符数据 $\hat{\boldsymbol{X}}_N$ 作为分类器的输入,以实现 ADMET 性质判别.

$$\hat{\boldsymbol{X}}_N = \left(\hat{\boldsymbol{X}}_N^1, \hat{\boldsymbol{X}}_N^2, \cdots, \hat{\boldsymbol{X}}_N^q, \cdots, \hat{\boldsymbol{X}}_N^Q\right), \tag{1}$$

式中,$\hat{\boldsymbol{X}}_N^q$ 为第 N 个化合物的第 q 条分子描述符数据,$\hat{\boldsymbol{X}}_N^q = \left(\hat{x}_1^q, \hat{x}_2^q, \cdots, \hat{x}_\ell^q, \cdots, \hat{x}_{\Upsilon_N}^q\right)$;$\Upsilon_N$ 为不同维度的自变量,包括相对分子质量、脂水分配系数 LogP、氢键供体数量等,$\hat{x}_\ell^q$ 为第 ℓ 个元素值,即相对分子质量、脂水分配系数 LogP、氢键供体数量等变量的取值;Q 为化合物结构式数据集样本量.

逻辑回归 (Logistic Regression, LR)[14] 作为一种基于二项分类的回归分析模型,通过在线性回归的基础上增加一个 Sigmoid 函数映射,实现对定性变量的有效预测. 朴素贝叶斯 (Naïve Bayes, NB)[15] 通过给定独立的目标值属性之间的相互条件,假定模型的变量遵循某种概率分布,对样本数据集进行分类. 两者均具有形式简单、性能稳定、鲁棒性强等优点,广泛应用于文本分类、入侵检测、故障诊断等领域[16]. 随着深度学习在模式识别中的广泛应用,梯度提升决策树 (Gradient Boosting Decision Tree, GBDT)[17] 基于梯度提升学习策略,对决策树中的回归树的迭代优化,寻找最佳划分特征,进而学习样本路径实现分类,是近年来一种模型复杂度

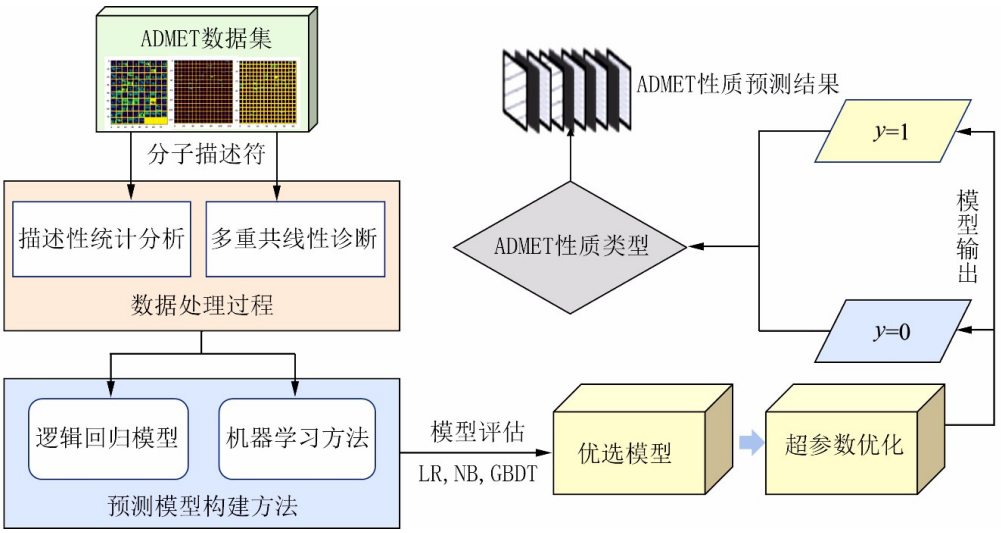

图 1 ADMET 性质预测模型框架

Fig. 1 A predictive modelling framework for the nature of ADMET

较高、参数随机性较强的学习器. 因此, 本文选取 LR 模型、NB 模型及 GBDT 模型进行 ADMET 性质分类预测模型的构建.

此外, 由于上述经典模型中人工设置超参数(如学习速率、层数以及每层的神经元数等参数)对预测性能的影响较大, 训练时间较长[18]. 为使算法获得最优性能, 采用概率代理模型拟合超参数 x 与预测精度 y 之间的关系 (即黑箱模型), 再通过采集函数扩大数据集 $D=\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\cdots,(x_n,y_n)\}$的规模, 更新代理模型的后验分布, 直到后验分布基本贴合于真实分布, 从而筛选出优选模型的最优超参数, 能有效地搜索可能的超参数空间, 提升模型的训练速度.

3　实验及结果分析

3.1　数据处理　实验数据来源于公开数据集“2021 年中国研究生数学建模竞赛”. 数据集包含: ①分子描述符: 1 974 个化合物的 729 个分子描述符信息, 分子描述符是一系列用于描述化合物的结构和性质特征的参数, 包括物理化学性质与拓扑结构特征等; ② ADMET 性质: 1 974 个化合物的 ADMET 性质(Caco-2、CYP3A4、hERG)的数据, 该性质可表征候选药物在人体内是否具备良好的药代动力学性质.

根据以往的研究可知, 该数据集化合物样本量较小(1 974 个化合物), 分子描述符(729 个特征变量)特征冗余, 具有有效特征不明显、难以直接预测应用等特点[19]. 首先通过描述性统计分析, 剔除原始数据中的“0”值样本, 如表 1 所示; 然后为客观评价模型性能, 避免模型忽略不同量纲指标的潜在信息, 利用多重共线性诊断[20]、极值归一化处理等方法, 对数据做深层次的处理分析; 最后从 729 个分子描述符中, 遴选出对 ADMET 性质具有影响的 319 个特征变量, 为构建化合物的 ADMET 性质预测模型提供有效的数据基础.

表 1　描述性统计分析结果示例

Tab. 1　Example of descriptive statistical analysis results

分子描述符	N	最小值	最大值	均值	标准偏差
nAcid	1 974	0	4	0.11	0.35
ALogP		−23.11	5.18	1.11	1.43
ALogp2		0	533.84	3.29	12.83
AMR		54.07	517.43	116.56	31.57
apol		30.66	359.66	60.63	19.45
nH		5	180	22.65	10.78
nB		0	0	0	0
Zagreb		62	748	150.72	41.45

3.2　模型预测结果分析　为降低由于数据样本量受限导致的预测偶然性, 提高模型泛化能力及数据使用率, 采用小型数据集适用的 k 折交叉验证方法对各预测模型验证. 本文通过将数据集划分为 5 折, 即将样本集分为 5 份, 每次选择 1 份样本集用于验证, 将剩余的 4 份样本集用于测试.

3.2.1　Caco-2 性质预测结果　ADMET 性质中 Caco-2 性质预测混淆矩阵如图 2 所示. 结合表 2 可以发现, NB 模型对 Caco-2 性质的预测效果在准确率(Accuracy, 评价总体预测效果)、精准率(Precision, 反映预测的精确性)、灵敏度(True Positive Rate, TPR)、误报率(False Positive Rate, FPR)方面表现最差, 而 GBDT 模型优于 LR 和 NB 模型.

具体来看, GBDT 模型相比 LR 模型在准确率、精准率、灵敏度、误报率方面依次提升了 3.9%、5.8%、3.3%、4.2%; GBDT 模型相比 NB 模型在准确率、精准率、误报率方面则依次提升了 11.7%、20.5%、22.3%, 灵敏度虽然下降了 5.4%, 但也达到

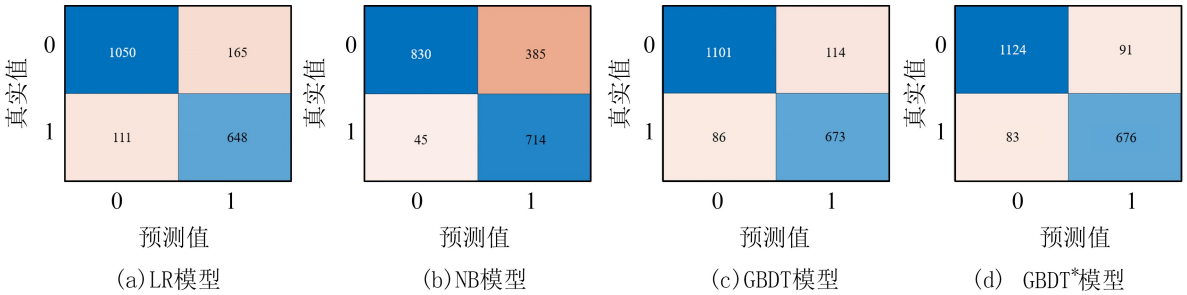

图 2　Caco-2 性质各预测模型混淆矩阵

Fig. 2　Confusion matrix of each prediction model of Caco-2

表 2 模型指标对比(Caco-2)

Tab. 2 Comparison of model indicators (Caco-2)

模型	准确率/%	精准率/%	灵敏度/%	误报率/%	AUC-0	AUC-1
LR	86.0	79.7	85.4	13.6	0.86	0.86
NB	78.2	65.0	94.1	31.7	0.85	0.85
GBDT	89.9	85.5	88.7	9.4	0.96	0.96
GBDT*	**91.2**	**88.1**	**89.1**	**7.5**	**0.97**	**0.97**

注: 下划线表示LR、NB、GBDT三者最优结果, 黑体加粗表示所有模型最优结果, AUC-0和AUC-1为模型精度评价指标

了 85% 以上, 同时与精准率保持均衡, 表明 GBDT 模型对 ADMET 性质中的 Caco-2 性质的预测精度良好, 优选模型即基于 GBDT 的 ADMET 性质预测模型. 对其进行超参数优化过后, 基于 GBDT*的 ADMET 性质预测模型的准确率达到 91.2%. 其相比基准 GBDT 模型在准确率、精准率、灵敏度、误报率方面则依次提升了 1.3%、2.6%、0.4%、1.9%. GBDT*模型准确率的进一步上升, 验证了本文超参数优化方法的有效性. 同时也说明 GBDT*更适用于 ADMET 性质的预测问题.

考虑到 ADMET 性质预测问题中样本数据数量不平衡会对模型的预测效果产生影响, 而工作特性曲线(Receiver Operating Characteristic curve, ROC)能够综合客观衡量模型本身整体性能, 具有避免不同测试集带来的干扰, 不受样本不均影响等特点. 因此, 为客观反映模型的预测性能, 选取 ROC 作为 ADMET 性质预测效果的进一步评价指标. 在显著性水平为 0.05 的情况下, 计算 ROC 曲线下面积(Area Under Curve, AUC), 研究所构建的预测模型是否适用于 ADMET 不同性质的判别.

从图 3 可以看出, 在对 Caco-2 性质进行预测时, 基于 GBDT 的 ADMET 预测模型 AUC 最大(AUC=0.96). 相比 LR 和 NB 算法的预测模型, GBDT 模型 AUC 指标分别提高了 0.10 和 0.11. 再次说明在对分子描述符数据进行统一清洗处理的条件下, 基于 GBDT 算法构建的 ADMET 性质预测模型对 Caco-2 性质具有较好的预测能力. 同时也说明基于 GBDT 的 ADMET 预测模型更适合处理低维非线性分析描述符数据, 对其进行超参数优化后, GBDT*与 GBDT 模型的 AUC 指标虽相差不大, 但准确率、精准率、灵敏度、误报率均有效提升. 总体来看, 基于 GBDT*算法构建的 ADMET 性质预测模型能有效提升预测精度, 具有应用于 ADMET 性质预测的潜力.

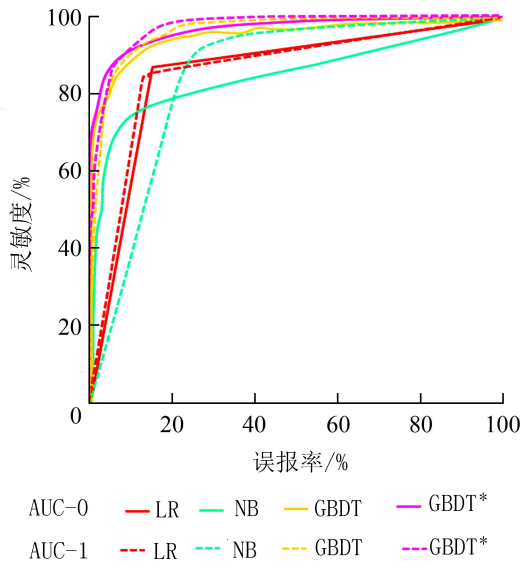

图 3 Caco-2 性质各预测模型 ROC 曲线

Fig. 3 ROC curves of each prediction model of Caco-2

3.2.2 CYP3A4 性质预测结果 CYP3A4 结果与 Caco-2 类似, 如表 3 所示. 具体表现为: 与基于 LR 的 ADMET 性质预测模型相比, GBDT 在准确率、精准率、灵敏度、误报率方面分别提升了 3.2%、0.4%、4.2%、0.6%; 与基于 NB 的 ADMET 性质预测模型相比, GBDT 模型在精准率和误报率方面较弱, 这是因为原始数据集中 CYP3A4 样本类别不均衡, 无代谢能力的样本(CYP3A4=0)占有代谢能力的样本(CYP3A4=1)的 35%, 导致模型对无代谢能力的样本(CYP3A4=0)判断不准确; 但在模型总体预测效果方面, GBDT 模型的准确率较 NB 模型提升了 5.8%, AUC 综合评估指标提升了 0.07, 并且精准率和灵敏度也得到了兼顾. 因此, 从全局考虑, 仍选用 GBDT 模型作为优选模型, 进行 ADMET 性质预测, 在对其进行超参数优化过后, GBDT*与 GBDT 各评价指标相差不大, 但有效缩减了基于 GBDT 的 ADMET 性质预测模型的训练时间.

3.2.3 hERG 性质预测结果 hERG 性质判别结

表 3　模型指标对比(CYP3A4)

Tab. 3　Comparison of model indicators (CYP3A4)

模型	准确率/%	精准率/%	灵敏度/%	误报率/%	AUC-0	AUC-1
LR	90.0	94.7	91.6	14.6	0.91	0.91
NB	87.4	**96.9**	85.7	**7.8**	0.91	0.91
GBDT	93.2	95.1	95.8	14.0	0.98	0.98
GBDT*	**93.3**	95.0	**96.0**	14.4	**0.98**	**0.98**

果也与上述类似, 如表 4 所示. 在 hERG 性质预测过程中, 集成学习方法 GBDT 相较于 LR 模型与 NB 模型在准确率、精准率等方面均取得了最佳的预测结果, 成为优选模型. 且 GBDT*较 GBDT 在准确率、精准率、灵敏度、误报率方面提升了 0.9%、1%、0.5%、1.4%; 在 AUC 综合评价指标方面提升了 0.01, 体现出基于 GBDT*的 ADMET 预测模型的优越性.

表 4　模型指标对比 (hERG)

Tab. 4　Comparison of model indicators (hERG)

模型	准确率/%	精准率/%	灵敏度/%	误报率/%	AUC-0	AUC-1
LR	85.4	87.2	86.4	15.9	0.86	0.86
NB	83.2	84.5	85.4	19.7	0.88	0.88
GBDT	89.7	89.4	92.4	13.7	0.96	0.96
GBDT*	**90.6**	**90.4**	**92.9**	**12.3**	**0.97**	**0.97**

3.2.4　ADMET 性质的特征筛选　基于模型精度分析, 选择优选模型 GBDT 预测模型探究不同特征变量对各 ADMET 性质的影响, 采用经验阈值法(特征权重大于 0.015 的变量)筛选出显著变量. 按重要性百分比从大到小依次排序, 结果如图 4 所示. 三类性质的特征变量在权重数值层面较为集中分布于某一种或几种变量上. 例如 Caco-2 特征重要性指标中, 大于 0.015 的指标有 8 项, 其中 ECCEN 特征对 Caco-2 性质起到绝对控制作用, 占比 50.30%. CYP3A4 各特征重要性指标中, VP-7、Zagreb、SP-6 是对 CYP3A4 影响程度较大的变量, 分别占比 27.00%、13.97%、10.97%. hERG 各特征重要性指标中, ECCEN 是影响 hERG 预测的关键特征变量, 占比 31.40%.

可见, ADMET 性质受不同特征因素影响差异大, 导致其预测效果的随机性. 传统的最优权重阈值方法只能筛选出 ADMET 性质的明显特征, 而难以确定最有效的特征变量. 因此, 本文采用概率代理模型拟合超参数与预测精度之间的关系 (即黑箱模型), 及时调整模型最佳超参数, 获取有效特征因子, 以适应各性质的预测需求.

3.3　GBDT*模型优化效果分析　为进一步验证本文 GBDT*算法的优势, 设置 GBDT*模型最大迭代次数为 30 次, 参数调整范围为: 树的数量为 (0,1 200), 学习率为 (0,1), 最大特征数为 (0,100), 经超参数自动寻优后, 输出结果如表 5 所示. 同时找到适合 Caco-2、CYP3A4、hERG 预测的有效特征数分别为 49、14、36 个. 调参可视化过程如图 5 所示, 可以看出, GBDT*对 Caco-2 性质和 hERG 性质预测模型的优化效果显著, 对 CYP3A4 性质预测模型的优化效果稍弱, 可能是数据样本量太小、有效特征不明显所致. 总的来说, GBDT*模型能够针对不同输入及时调整所需超参数, 提升模型快速找到不同 ADMET 性质的有用特征, 有效改善数据特征不明显、维度过高导致特征冗余等情况, 提升模型训练的效率.

4　结论

本文以抗乳腺癌活性化合物的 ADMET 性质中的吸收、代谢、毒性属性的分类预测为研究方向, 提出基于 GBDT*算法的 ADMET 性质预测方法. 然后对抗乳腺癌活性化合物的物理化学性质、拓

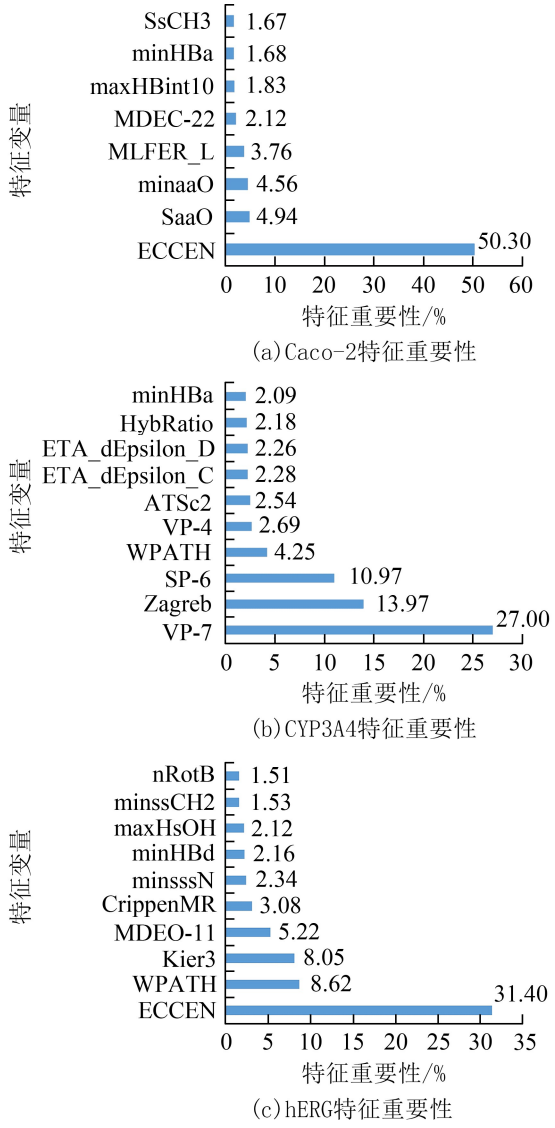

图 4 ADMET 性质的特征重要性

Fig. 4 Characteristic importance of ADMET

表 5 GBDT* 调参结果

Tab. 5 Hyperparameters tuning results of GBDT*

ADMET性质	参数	调参结果
Caco-2	树的数量	499
	学习率	0.005
	最大特征数	49
CYP3A4	树的数量	13
	学习率	0.071
	最大特征数	14
hERG	树的数量	54
	学习率	0.005
	最大特征数	36

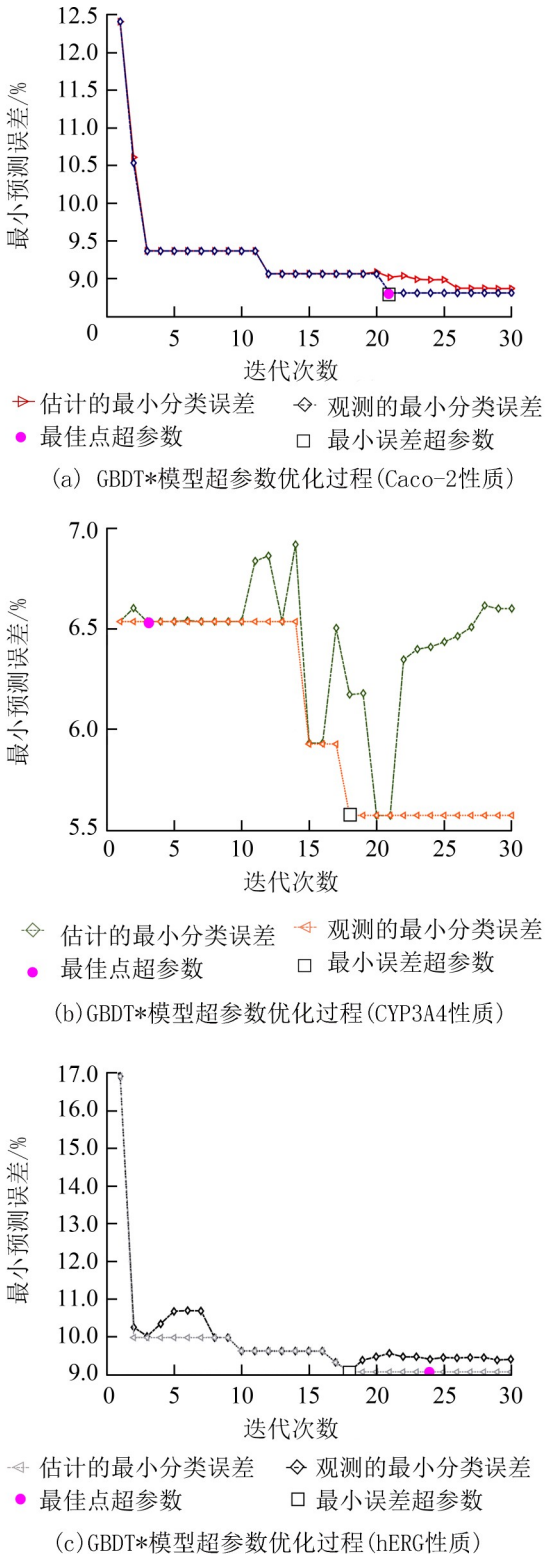

图 5 GBDT*模型超参数优化过程可视化

Fig. 5 Visualization of hyperparaments optimization process for GBDT* model

扑结构特征等数据进行清洗处理,获取丰富的状态信息,选取 LR、NB、GBDT 作为 ADMET 分类预测的候选模型,针对经典算法训练成本较高的问题,

对 GBDT 模型进行超参数寻优, 提出 ADMET 性质分类预测模型为最优模型 GBDT*, 有效改善浅层机器学习调参时间久、局部最小化以及过拟合等缺陷, 能更好地根据小样本、多特征条件下分子描述符变量对 ADMET 性质进行预测, 有助于抗乳腺癌候选药物的虚拟筛选研究. 本文在预测时仅以分子描述符特征作为自变量, 未来将综合考虑各类因素, 建立更加通用且稳定的 ADMET 性质预测模型.

参考文献:

[1] Wang S Q, Sun H Y, Liu H, et al. ADMET evaluation in drug discovery 16 predicting herg blockers by combining multiple pharmacophores and machine learning approaches[J]. Mol Pharm, 2016, 13(8): 2 855-2 866. DOI: 10.1021/acs.molpharmaceut.6b00471.

[2] Sheridan R P, Culberson J C, Joshi E, et al. Prediction accuracy of production ADMET models as a function of version: Activity cliffs rule[J]. Journal of Chemical Information and Modeling, 2022, 62(14): 3 275-3 280. DOI: 10.1021/acs.jcim.2c00699.

[3] Wang Y L, Xing J, Xu Y, et al. In silico ADME/T modelling for rational drug design[J]. Quarterly Reviews of Biophysics, 2015, 48(4): 488-515. DOI: 10.1017/S0033583515000190.

[4] Patel C N, Kumar S P, Rawal R M, et al. A multiparametric organ toxicity predictor for drug discovery[J]. Toxicology Mechanisms and Methods, 2020, 30(3): 159-166. DOI: 10.1080/15376516.2019.1681044.

[5] Chi C T, Lee M H, Weng C F, et al. In silico prediction of PAMPA effective permeability using a two-QSAR approach[J]. International Journal of Molecular Sciences, 2019, 20(13): 3 170. DOI: 10.3390/ijms20133170.

[6] Ferreira L L G, Andricopulo A D. ADMET modeling approaches in drug discovery[J]. Drug Discovery Today, 2019, 24(5): 1 157-1 165. DOI: 10.1016/j.drudis.2019.03.015.

[7] Wu F X, Zhou Y Q, Li L H, et al. Computational approaches in preclinical studies on drug discovery and development[J]. Frontiers in Chemistry, 2020, 8: 726. DOI: 10.3389/fchem.2020.00726.

[8] 李晓, 李达, 周雪松, 等. 化合物 ADMET 性质预测平台的构建[J]. 生物信息学, 2017, 15(3): 179-185. DOI: 10.3969/j.issn.1672-5565.201704003.
Li X, Li D, Zhou X S, et al. Development of the platform for prediction of chemical ADMET properties[J]. China Journal of Bioinformatics, 2017, 15(3): 179-185.

[9] 莫贤炜, 周海燕, 李晓雷, 等. 基于 Topomer CoMFA 方法的苯基哌嗪类 5-HT_7 受体拮抗剂的 3D-QSAR 研究[J]. 计算机与应用化学, 2018, 35(8): 667-679.
Mo X W, Zhou H Y, Li X L, et al. 3D-quantitative structure-activity relationships study of phenyl-piperazines as 5-HT7 receptor antagonist based on topomer comfa method[J]. Computers and Applied Chemistry, 2018, 35(8): 667-679.

[10] Tsou L K, Yeh S H, Ueng S H, et al. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery[J]. Scientific Reports, 2020, 10(1): 1-11. DOI: 10.1038/s41598-019-56847-4.

[11] Feinberg E N, Joshi E, Pande V S, et al. Improvement in ADMET prediction with multitask deep featurization[J]. Journal of Medicinal Chemistry, 2020, 63(16): 8 835-8 848. DOI: 10.1021/acs.jmedchem.9b02187.

[12] Pires D E V, Blundell T L, Ascher D B. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures[J]. Journal of Medicinal Chemistry, 2015, 58(9): 4 066-4 072. DOI: 10.1021/acs.jmedchem.5b00104.

[13] Dong J, Wang N N, Yao Z J, et al. ADMETlab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database[J]. Journal of Cheminformatics, 2018, 10(1): 1-11. DOI: 10.1186/s13321-017-0256-5.

[14] Song C Y, Wang L G, Xu Z S. An optimized logistic regression model based on the maximum entropy estimation under the hesitant fuzzy environment[J]. International Journal of Information Technology & Decision Making, 2022, 21(1): 143-167.

[15] Jackins V, Vimal S, Kaliappan M, et al. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes[J]. The Journal of Supercomputing, 2021, 77(5): 5 198-5 219. DOI: 10.1007/s11227-020-03481-x.

[16] 谢济铭, 秦雅琴, 彭博, 等. 多车道交织区车辆跟驰行为风险判别与冲突预测[J]. 交通运输系统工程与信息, 2021, 21(3): 131-139. DOI: 10.16097/j.cnki.1009-6744.2021.03.016.
Xie J M, Qin Y Q, Peng B, et al. Risk discrimination and conflict prediction of vehicle-following behavior in multi-lane weaving sections[J]. Journal of Transportation Systems Engineering and Information Technology, 2021, 21(3): 131-139.

[17] Li S M, Lin Y L, Zhu T, et al. Development and external evaluation of predictions models for mortality of COVID-19 patients using machine learning method[J]. Neural Computing and Applications, 2021(11): 1-10.

[18] Wu J, Chen X Y, Zhang H, et al. Hyperparameter optimization for machine learning models based on Bayesian optimization[J]. Journal of Electronic Science and Technology, 2019, 17(1): 26-40.

[19] 夏玉兰, 谢济铭, 王雅婧, 等. 抗癌候选药物 ERα 抑制剂活性预测[J]. 深圳大学学报(理工版), 2022, 39(5): 529-537.
Xia Y L, Xie J M, Wang Y J, et al. Activity prediction of anti-cancer drug candidate ERα inhibitor[J]. Journal of Shenzhen University (Science and Engineering), 2022, 39(5): 529-537.

[20] 朱钰, 郑屹然, 尹默. 统计学意义下的多重共线性检验方法[J]. 统计与决策, 2020, 36(7): 34-36. DOI: 10.13546/j.cnki.tjyjc.2020.07.007.
Zhu Y, Zheng Y R, Yin M. Multicollinearity test under statistical significance[J]. Statistics and Decision, 2020, 36(7): 34-36.

Predictive modeling of ADMET properties of anti-breast cancer active compounds

QIN Ya-qin，XIA Yu-lan，LU Meng-yuan，Wang Jin-rui，XIE Ji-ming**

(School of Transportation Engineering, Kunming University of Science and Technology, Kunming 650500, Yunnan, China)

Abstract: A quantitative structure-ADMET prediction model is proposed to improve the prediction of absorption, distribution, metabolism, excretion and toxicity of anti-breast cancer drugs in the virtual screening process. Firstly, 319 variables are selected from the molecular descriptors of the compounds. Then Logistic Regression (LR), Naïve Bayes (NB) and Gradient Boosting Decision Tree (GBDT) are used to predict the properties of ADMET. Finally, in order to address the problem of high training cost of GBDT models, the GBDT* model is constructed by fitting the relationship between hyperparameters and prediction accuracy (i.e. black box models) with the help of a probabilistic agent model. The results show that the GBDT* integrated learning model performs best overall. The results show that the overall performance of the GBDT* integrated learning prediction model is optimal. The accuracy, precision, sensitivity and AUC of GBDT* reach over 90%, 88%, 89%, and 0.95, respectively, and the false alarm rate is less than 15%, indicating that the GBDT* integrated machine learning model has good performance in predicting the ADMET properties of anti-breast cancer active compounds.

Key words: computer applications; integrated learning; feature screening; hyperparametric optimization; anti-cancer drug candidates