基于 GWO-KELM 与 GBDT 的抗乳腺癌药物性质预测

王 斯[1,2], 张国浩[1,2], 陈义安[1,2]
1. 重庆工商大学 数学与统计学院,重庆 400067
2. 经济社会应用统计重庆市重点实验室,重庆 400067

摘 要:目的 利用人工智能算法辅助药物设计,实现拮抗乳腺癌候选药物的分子描述符筛选、ERα 回归预测、ADMET 分类预测。**方法** 针对乳腺癌候选药物筛选问题,以化合物对抑制乳腺癌靶标的生物活性及其 ADMET 性质出发,基于获取的 1 974 种化合物数据,分别利用稀疏贝叶斯学习与随机森林算法进行两阶段筛选,得到不具备强相关性的前 20 个对生物活性最具显著性影响的分子描述符;随后以筛选后的数据及其 PIC_{50} 值为基础建立了 QSAR 模型,基于灰狼优化的核极限学习机算法对新化合物的生物活性进行了预测,横向对比 11 种常见机器学习算法,同时利用 GBDT 算法构建了 ADMET 分类模型。**结果** GWO-KELM 模型具有更高的拟合优度与更低的均方误差,而且药物性质识别的 4 个模型预测准确率均保持 90%以上。**结论** 所建模型能够有效分析并预测化合物性质,为抗乳腺癌候选药物的研发提供参考。

关键词:乳腺癌;ERα;ADMET;GWO-KELM;GBDT;稀疏贝叶斯学习
中图分类号:TP181 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2023.0006.012

Prediction of Anti-breast Cancer Drug Properties Based on GWO-KELM and GBDT

WANG Si[1,2], ZHANG Guohao[1,2], CHEN Yian[1,2]
1. School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China
2. Chongqing Key Laboratory of Social Economic and Applied Statistics, Chongqing 400067, China

Abstract: Objective This study aimed to achieve molecular descriptor screening, ERα regression prediction, and ADMET classification prediction of antagonistic breast cancer drug candidates by using artificial intelligence algorithms to assist in drug design. **Methods** To address the screening problem of breast cancer drug candidates, starting from the biological activity of the compounds to inhibit the target of breast cancer and their ADMET properties, a two-stage screening was performed based on the obtained data of 1 974 compounds with sparse Bayesian learning and random forest algorithms, respectively, to obtain the top 20 molecular descriptors with the most significant effect on biological activity without strong correlation; subsequently, based on the screened data and its PIC_{50} value, a QSAR model was established, and the biological activity of the new compound was predicted based on the nuclear extreme learning machine algorithm optimized by the gray wolf, and 11 common machine learning algorithms were compared horizontally. The ADMET classification model was constructed. **Results** The results show that the GWO-KELM model has higher goodness of fit and

收稿日期:2022-07-09 **修回日期:**2022-08-16 **文章编号:**1672-058X(2023)06-0093-12
基金项目:重庆市自然科学基金项目(CSTC2020JCYJ-MSXMX0316);重庆市教委科学技术研究计划重大项目(KJZD-M202100801);重庆工商大学研究生创新项目(YJSCXX2022-112-189).
作者简介:王斯(1996—),男,重庆南川人,硕士研究生,CCF 会员(K7643G),从事机器学习与计算机辅助计算研究.
通讯作者:陈义安(1968—),男,四川达州人,硕士,教授,从事大数据分析技术及其应用研究. Email:709590113@qq.com.
引用格式:王斯, 张国浩, 陈义安. 基于 GWO-KELM 与 GBDT 的抗乳腺癌药物性质预测[J]. 重庆工商大学学报(自然科学版), 2023,40(6):93—104.
WANG Si, ZHANG Guohao, CHEN Yian. Prediction of anti-breast cancer drug properties based on GWO-KELM and GBDT[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(6): 93—104.

lower mean square error. The prediction accuracies of the four models were maintained above 90%. **Conclusion** The proposed models can effectively analyze and predict the properties of compounds, which can provide a reference for the development of anti-breast cancer drug candidates.

Keywords: breast cancer; ERα; ADMET; GWO-KELM; GBDT; sparse Bayesian learning

1 引 言

乳腺癌作为女性最常见的癌症,已经跃居世界女性癌症死亡的第二大病因,并且其发病率和死亡率每年仍在不断攀升[1]。为有效治疗该病症,医药与基因学领域进行了大量实验研究,发现人体内雌激素受体 α 亚型(ERα)与该病的发病率密切相关[2],并在乳腺肿瘤细胞中过度表达。因此,良性乳腺上皮细胞中的 ERα 活性升高也就表明患乳腺癌的风险增加,使得科研工作者们不断寻找和研发抑制 ERα 作用的药物[3]。

乳腺癌候选药物研发与临床应用需要的时间和成本巨大。一方面,药物需要有良好的生物活性,相关医药领域通常会为了节约时间与成本,运用计算机与体外研究技术,对可能具有良好表现的化合物进行筛选工作,即收集一系列作用于该靶标的化合物和生物活性数据,应用数学模型,构建定量构效关系[4](Quantitative Structure-Activity Relationship, QSAR),筛选新化合物以及预测药物活性。

另一方面,良好的生物活性虽然有效保证了化合物对抗肿瘤细胞的有效性,但是药物的研发还需要其药代动力学性质和安全性也符合相关政策法规的要求。药代动力学性质即药物吸收、分布、代谢、排泄和毒性的总称,这些性质分别代表着生物体对化合物的各项敏感程度[5]。

随着智能计算的迅速发展,机器学习和深度学习在医疗领域发挥着越来越重要的作用,特别是辅助药物研发方面。顾等[6]构建一种图注意力网络,用于虚拟药物筛选,并将算法横向对比机器学习算法和传统图神经网络算法,均取得良好的结果;谢等[7]基于平均法与堆叠法融合的浅层神经网络模型,通过对药物分子的化学结构进行信息化编码,提高了对药物分子预测的能力,与传统深度学习相比,他们的研究能够保证更好的准确性;Shi 等[8]采用卷积神经网络模型,并将其运用在 ADMET 特性的预测模型上,表明该方法的预测能力与基于手动结构描述和特征选择的可用机器学习模型的预测能力相当;此外,Peng 等[9]提出利用一种改进的图神经网络方法以改进对 ADMET 特性的预测,该方法能够通过将分子键特征与节点特征连接在一起,并应用门单元来调整原子邻域权重以映射中心原子与其相邻原子之间相互作用强度的差异,从而得到更有意义的分子结构模式,探索更好的分子建模。

从上述文献可知:传统药物活性预测方法成本高,时间长,应用范围小,而利用人工智能算法预测候选药物的生物活性和 ADMET 性质已成为当今研究的主流热点,出色的模型可以有效预测候选化合物分子活性并对化合物 ADMET 性质进行分类识别,从而显著地降低研发成本,极大地提高研发成功率,且有效避免因药物产生的副作用和毒性导致的人体疾病。因此,利用更先进的人工智能算法预测抗乳腺癌候选药物的生物活性并进行化合物 ADMET 性质的分别识别极具实践意义。

本文从 UA 的 DrugBanK[10] 数据库中获取了 1 974 种化合物对乳腺癌治疗靶标 ERα 的生物活性和 ADMET 性质数据,采用稀疏贝叶斯学习与随机森林算法进行两阶段筛选,随后基于两阶段筛选后的分子描述符建立了定量预测模型,利用 GWO-KELM 算法构建针对 IC_{50} 与 PIC_{50}(其值用 $Y_{IC_{50}}$, $Y_{PIC_{50}}$ 表示),的定量预测模型,同时利用 GBDT 构建分类预测模型,预测了化合物的 ADMET 性质。本文的研究旨在寻找生物活性较高且尽可能达到更好 ADMET 性质的化合物,以加快抗乳腺癌候选药物的研发进程。

2 变量筛选

2.1 数据描述

通过爬虫技术以及 XML 解析,获取 2 种数据集。第一种是用于定量预测的 ERα 生物活性数据,包含 SMILES 一维线性表达式,以及 $Y_{IC_{50}}$ 和 $Y_{PIC_{50}}$,前者越小越好,后者是前者的负对数变换;另一种是关于 ADMET 性质的类别数据,用于构建分类预测模型。

两种数据中的输入特征是 729 种分子描述符,不失一般性,实际数据通常被认为是稀疏的,所以必须在建模分析前进行特征筛选工作。根据各个特征在不同模型不同阶段的贡献度(特征重要性)进行排序,筛选出前 20 个最显著的分子描述符。常规的变量选择方法包括主成分分析法、LASSO、稀疏贝叶斯学习、随机森林等,但是主成分分析法和 LASSO 这类经典算法对 729 个变量指标进行特征提取时,可能不具备代表性。因此本文选择稀疏贝叶斯与随机森林算法对重要变量进行两阶段评估,以此筛选出对活性值影响大的分子描述符。同时,在筛选前,进行了数据预处理,结果表明原始数据中不存在任何的数据缺失,也无异常点存在。

2.2　稀疏贝叶斯模型一阶段筛选

稀疏贝叶斯模型以贝叶斯理论为基础,其优秀的分类和回归能力可以筛选并寻找包含多个零值的权重向量,同时精确逼近目标向量,从而使得容错与逼近性能更优,泛化误差最小[60-61]。稀疏信号恢复可用式(1)表达。

$$\boldsymbol{C}=\boldsymbol{\omega}\boldsymbol{\varphi}+\boldsymbol{\varepsilon} \tag{1}$$

其中,$\boldsymbol{\varphi}\in\boldsymbol{R}^{N\times M}$ 代表包含 N 个样本的矩阵,每个样本具有 M 个特征,$\boldsymbol{C}=[C_1,C_2,\cdots,C_N]^{\mathrm{T}}$ 代表目标向量,$\boldsymbol{\varepsilon}=[\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_N]^{\mathrm{T}}$ 代表噪声,$\boldsymbol{\omega}=[\omega_1,\omega_2,\cdots,\omega_M]^{\mathrm{T}}$ 代表模型学习用来构成 $\boldsymbol{\varphi}$ 中每一列的权重。

稀疏贝叶斯模型的目标是寻找到一个包含很多零值的 $\boldsymbol{\omega}$ 权重向量,同时结果准确地逼近目标向量 $\boldsymbol{C}$。在 SBL 模型中,为了寻找系数信号恢复的最小范数解,常常使用高斯似然函数模型获取 $\boldsymbol{\omega}$ 的最大似然估计量,具体见式(2)。

$$p(\boldsymbol{C}|\boldsymbol{\omega};\boldsymbol{\sigma}^2)=(2\pi\boldsymbol{\sigma}^2)^{-\frac{N}{2}}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{C}-\boldsymbol{\varphi}\boldsymbol{\omega}\|_2^2\right) \tag{2}$$

为了找到稀疏解,SBL 从数据中估计参数化的先验权重,过程可以用式(3):

$$p(\boldsymbol{\omega};\boldsymbol{\gamma})=\prod_{i=1}^{M}(2\pi\gamma_i)^{-\frac{1}{2}}\exp\left(-\frac{\omega_i^2}{2\gamma_i^2}\right) \tag{3}$$

其中,$\boldsymbol{\gamma}=[\gamma_1,\gamma_1,\cdots,\gamma_M]^{\mathrm{T}}$ 代表 M 个超参数的向量,它控制每个权重的先验方差。

另一方面,在对变量维数众多的特征进行筛选时,除了通过影响程度去寻找重要变量,还应减小变量与变量之间的相关性对影响程度产生的干扰。本文将采用斯皮尔曼相关系数去表示两个变量之间的关联程度,从而将相关性过强的变量做标记并加入二次筛选的随机森林模型中进行相关性分离。

稀疏贝叶斯模型的筛选结果与斯皮尔曼的相关系数结果如表 1 及图 1 所示,一阶段的筛选结果得到了前 40 个对生物活性最具显著性影响的变量,但有个别特征(nF10Ring、nT10Ring、nF、nsF、mindS、SdS、maxdS)的相关性显示为强相关(深色)。

表 1　SBL 变量选择结果

Table 1　Results of SBL variable selection

特征	特征重要性	重要性绝对值	排名	符号	特征	特征重要性	重要性绝对值	排名	符号
nHBAcc	−0.799	0.799	1	−	mindsN	−0.022	0.022	21	−
MLFER-A	0.473	0.473	2	+	minHBint10	0.020	0.020	22	+
nHCsats	0.326	0.326	3	+	MDEO-12	−0.020	0.020	23	−
BCUTp-1h	0.253	0.253	4	+	nT10Ring	−0.018	0.018	24	−
C3SP2	0.222	0.222	5	+	nF10Ring	−0.018	0.018	25	−
ATSc4	0.218	0.218	6	+	minHBa	−0.017	0.017	26	−
MDEC-34	−0.159	0.159	7	−	maxdsN	−0.017	0.017	27	−
SsF	0.137	0.137	8	+	SaaCH	0.016	0.016	28	+
maxHBd	0.134	0.134	9	+	nsF	0.016	0.016	29	+
SwHBa	0.131	0.131	10	+	nF	0.016	0.016	30	+
SHCsats	0.126	0.126	11	+	minHBint7	0.015	0.015	31	+
maxsssCH	0.097	0.097	12	+	maxdS	0.014	0.014	32	+
minHBint5	0.059	0.059	13	+	mindS	0.014	0.014	33	+
minHssNH	−0.054	0.054	14	−	SdS	0.014	0.014	34	+
VCH-5	0.049	0.049	15	+	FMF	0.013	0.013	35	+
mindO	−0.048	0.048	16	−	maxHsOH	0.011	0.011	36	+
minsssN	0.044	0.044	17	+	C1SP2	−0.007	0.007	37	−
mindsCH	−0.039	0.039	18	−	maxHssNH	−0.007	0.007	38	−
C1SP3	0.030	0.030	19	+	C3SP3	−0.007	0.007	39	−
maxHBint8	0.025	0.025	20	+	minwHBa	0.006	0.006	40	+

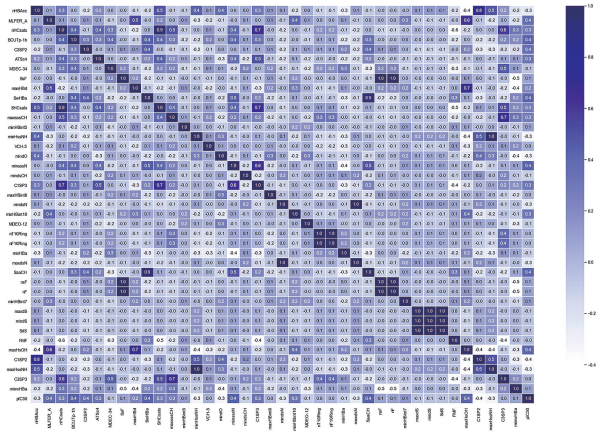


图 1 一阶段筛选变量斯皮尔曼系数

Fig. 1 One-stage screening variable Spearman's coefficient

2.3 随机森林模型二阶段筛选

为处理各变量之间的相关性,以避免后续缩减模型过拟合情形的发生,在二阶段筛选中,本文将一阶段筛选结果选入随机森林模型中,进行新一轮特征分解提取,从而得到对生物活性最具有显著影响的前 20 个变量。最终,得到的生物活性最具有显著影响的前 20 个变量与变量相关系数结果如表 2 及图 2 所示。比较图 2 可以明显看出:通过随机森林模型对变量进行二次筛选后,初次筛选时的强相关性变量相关系数明显减小。

表 2 随机森林筛选结果

Table 2 Results of random forest screening

序号	特征	重要性	序号	特征	重要性
1	minsssN	0.26	11	MDEO-12	0.03
2	maxHsOH	0.09	12	minHBa	0.03
3	BCUTp-1h	0.08	13	ATSc4	0.02
4	MLFER-A	0.07	14	maxHBd	0.02
5	nHBAcc	0.04	15	maxHBint8	0.02
6	SwHBa	0.04	16	SaaCH	0.02
7	minHBint5	0.04	17	FMF	0.02
8	C3SP2	0.03	18	minwHBa	0.02
9	mindO	0.03	19	nHCsats	0.01
10	minHBint1	0.03	20	MDEC-34	0.01

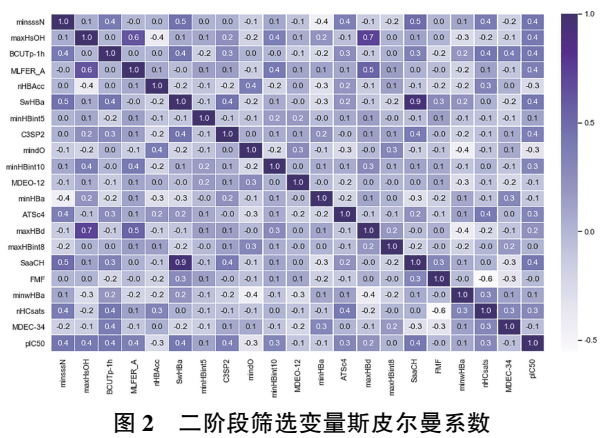


图 2 二阶段筛选变量斯皮尔曼系数

Fig. 2 Two-stage screening variable Spearman's coefficient

3 基于 GWO-KELM 算法的 QSAR 模型预测分析

两阶段筛选后的分子描述符特征已大大减小,考虑到 KELM(Kernel based Extreme Learning Machine)算法具有良好的稳定性、泛化能力、容错能力[12],并且目前广泛应用于辅助医药研发,但该模型避免不了 KELM 神经网络的参数调优问题,特别是正则化系数 C 与核参数 g。因此,本文利用能够进行参数寻优的 GWO(Grey Wolf Optimizer)算法对 KELM 进行改进,从而确定其最优参数,进一步提高模型预测性能。

3.1 KELM 原理

数据集 $\boldsymbol{D}=\{(x_i,y_i),i=1,2,\cdots,n\}$,输入数据 $\boldsymbol{x}_i\in R^n$,输出值为 $y_i\in\mathbf{R}$,向量 $\boldsymbol{h}(x_i)=[h_1(x_i),h_2(x_i),\cdots,h_m(x_i)]$ 的作用是将 x_i 从 n 维输入空间映射到 m 维隐藏层空间,向量 $\boldsymbol{\beta}=[\beta_1,\beta_2,\cdots,\beta_m]^{\mathrm{T}}$ 是用来连结输出节点与隐藏层的权值向量,$\boldsymbol{H}=[h(x_1),h(x_2),\cdots,h(x_n)]^n$ 代表隐含层输出矩阵,正则系数 C 用来减小模型产生的误差。传统 ELM 的输出表达式为式(4):

$$f(x)=\boldsymbol{h}(\boldsymbol{x})\boldsymbol{\beta}=\boldsymbol{h}(\boldsymbol{x})\boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}+\frac{1}{C}\right)^{-1}\boldsymbol{Y} \tag{4}$$

其中,$\boldsymbol{Y}$ 是输出向量。由于传统 ELM 的输出表达式中有矩阵内积存在,因此使用满足条件的核函数来代替矩阵内积,即式(5)—式(6)。

$$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}(i,j)=K(x_i,x_j) \tag{5}$$

$$\boldsymbol{h}(\boldsymbol{x})\boldsymbol{H}^{\mathrm{T}}=[K(x,x_1),K(x,x_2),\cdots,K(x,x_n)]^{\mathrm{T}} \tag{6}$$

其中,核函数是 $K=[K(x_i,x_j)]_{i,j=1}^n$,高斯核函数是 $K(x_i,x_j)=\exp(-\zeta\|x_i-x_j\|^2)$,参数 $\zeta>0$。

得到 KELM 模型的输出为式(7):

$$f(x)=[K(x,x_1),K(x,x_2),\cdots,K(x,x_n)]^{\mathrm{T}}\left(\frac{1}{C}+K\right)^{-1}\boldsymbol{Y} \tag{7}$$

综上可知:KELM 模型中的核映射更稳定,因为其回归预测的泛化性能比常见的预测模型更优。同时,KELM 模型只涉及自身的内积运算,而且不需预先设置隐含层的节点数,这使得模型更加稳定,收敛速度较快。但值得注意的是,KELM 模型有时会因为参数选择不当而导致预测误差偏高。

3.2 灰狼算法改进的 KELM 模型

为了选择合适的算法针对 KELM 模型进行优化,本文进行预实验,选择正余弦优化算法(Sine Cosine Algorithm, SCA)、粒子群优化算法(Particle Swarm Optimization, PSO)、灰狼算法进行实验比较。图 3 可以看出:SCA 算法收敛速度很慢,耗时很长,而 PSO 算法虽然迭代速度收敛较快,但过早陷入局部最优,而 GWO 算法综合表现更好。因此,利用 GWO 对 KELM 算法超参数优化,算法流程图如图 4 所示。

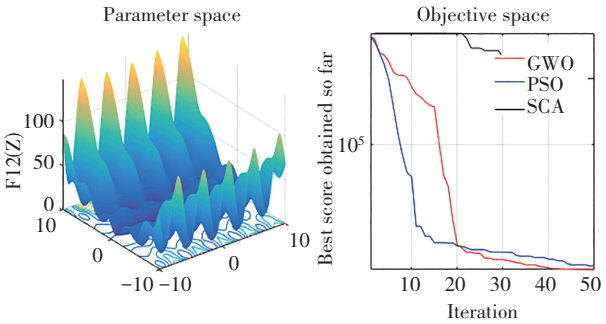


图 3　优化算法参数空间和迭代次数对比

Fig. 3　Comparison of parameter space and number of iterations of optimization algorithm

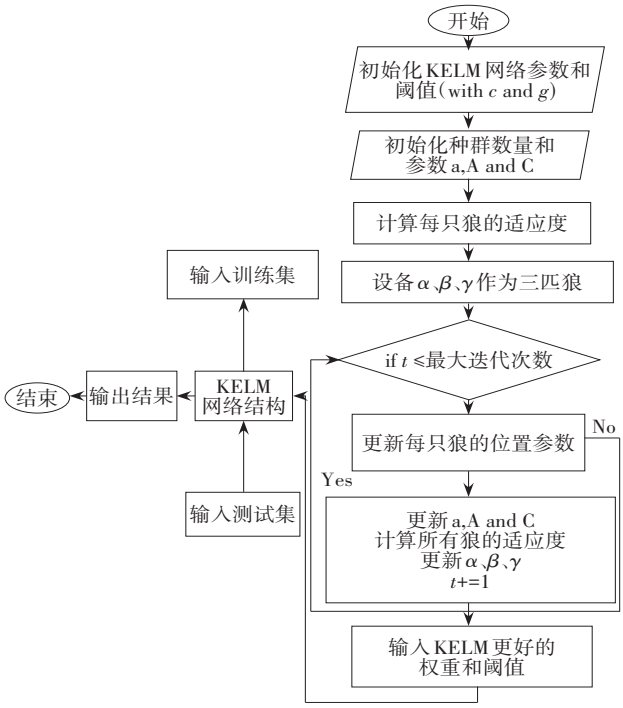


图 4　GWO-KELM 算法流程图

Fig. 4　Flow chart of GWO-KELM algorithm

3.3　实验结果与比较分析

为了科学有效地体现出 GWO-KELM 算法的优越性,本文将基于灰狼优化的 KELM 算法与 11 个常见预测算法进行生物活性预测效果对比,通过图像和数据直观体现该模型的优点。具体对比算法是决策树、线性回归、支持向量机回归、k-近邻、增强学习、梯度提升、装袋算法、极限树、贝叶斯岭回归、自动相关性确定算法和泰尔森估算。

通过对比上面的组图可知,11 个模型均在一定程度上出现预测误差偏大。观察图 5 及图 6 可知,本文算法预测结果与真实值比较吻合,不仅具有最小的误差,而且拟合程度超过 70.85%,拟合程度较好。

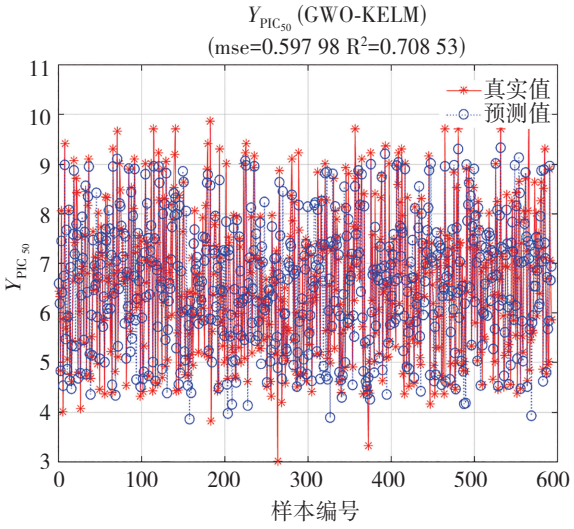
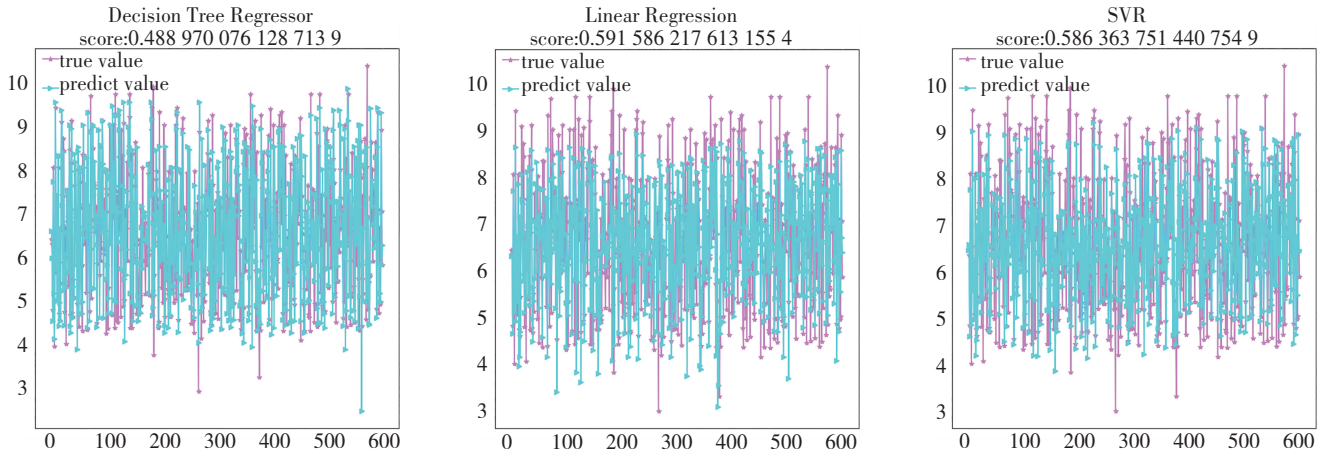


图 5　GWO-KELM 算法预测 $Y_{PIC_{50}}$ 结果

Fig. 5　Prediction of $Y_{PIC_{50}}$ results by GWO-KELM algorithm



(a) DTR　　(b) LR　　(c) SVR

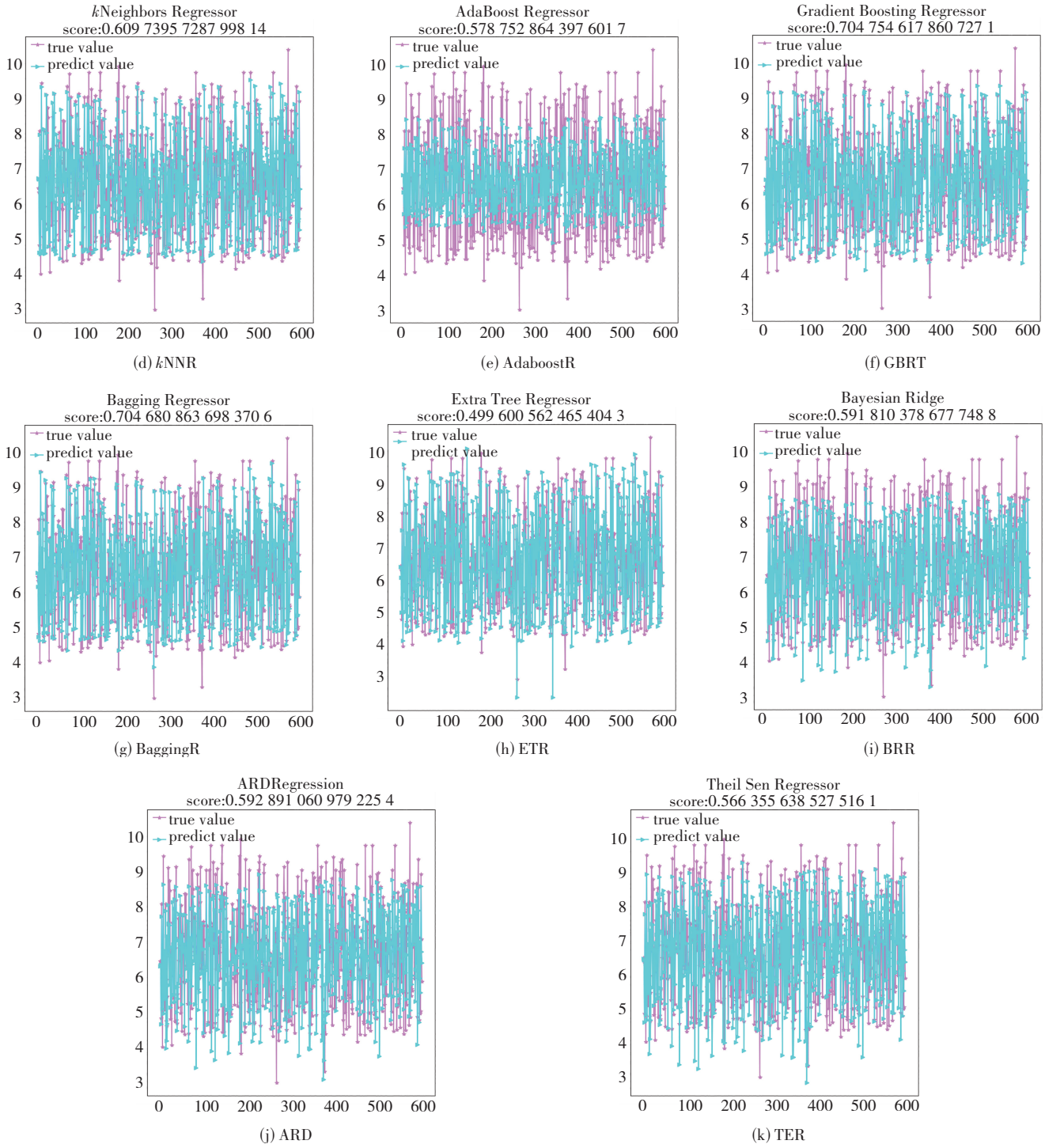


图 6 各类算法预测 $Y_{PIC_{50}}$ 结果图

Fig. 6 Predicted $Y_{PIC_{50}}$ results for each type of algorithm

为了更直观地对比 GWO-KELM 预测算法与其余算法的预测性能,本文共选取了 3 个指标来评价生物活性定量预测有效性,模型主要指标分别为拟合优度 R^2、均方误差、平均绝对误差,计算公式如下:

$$R^2 = 1 - \left(\sum_{i=1}^{m} (\hat{y}_i - y_i)^2 / \sum_{i=1}^{m} (\bar{y} - y_i)^2 \right)$$

$$R_{\mathrm{MSE}} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

$$R_{\mathrm{MAE}} = \frac{1}{m} \sum_{i=1}^{m} |\hat{y}_i - y_i|$$

根据上述预测指标结果,将 GWO-KELM 模型与常见的 12 个预测模型进行比较,模型的主要指标对比如表 3 所示:

表 3　各类算法指标汇总

Table 3　Summary of metrics for each type of algorithm

模型	均方误差	平均绝对误差	似合优度
Our Method	0.597 9	0.574 5	0.708 5
DTR	1.035 3	0.716 4	0.489 0
LR	0.827 4	0.703 6	0.591 6
SVR	0.838 0	0.686 8	0.586 4
*k*NNR	0.791 3	0.643 2	0.609 4
AdaboostR	0.853 4	0.765 7	0.578 8
GBRT	0.598 1	0.584 5	0.704 8
BaggingR	0.598 3	0.568 6	0.704 7
ETR	1.013 8	0.704 4	0.499 6
BRR	0.827 0	0.703 6	0.591 8
ARD	0.824 8	0.703 6	0.592 9
TER	0.878 5	0.720 3	0.566 4

上述结果表明,GWO-KELM 生物活性定量预测模型具有良好的优越性及有效性,能够对生物活性定量预测进行良好的建模;另外,通过与真实值以及 11 个预测模型结果比较,验证了该算法的有效性。其本身模型的特性是在计算时不需要进行迭代,计算速度快,具有出色的泛化能力,能提供更为准确的预测结果;利用 GWO 算法优化 KELM 模型的参数,在参数取值范围内寻求全局最优的参数解,使得 KELM 模型的预测结果更加精确。

3.4　基于 GWO-KELM 模型定量预测结果

以上实验结论证明了 GWO-KELM 定量预测的优秀效果。对新的化合物进行预测,$Y_{PIC_{50}}$ 由负对数变换而来,故无单位,具体可见式(3),预测结果见表 4。

$$Y_{IC_{50}} = 10^{-Y_{PIC_{50}}+9} \tag{3}$$

表 4　$Y_{IC_{50}}$ 值和 $Y_{PIC_{50}}$ 值预测结果

Table 4　Predicted results of $Y_{IC_{50}}$ and $Y_{PIC_{50}}$ values

待测样本	$Y_{IC_{50}}$/(nmol/L)	$Y_{PIC_{50}}$/(nmol/L)	待测样本	$Y_{IC_{50}}$/(nmol/L)	$Y_{PIC_{50}}$/(nmol/L)
1	171.869 9	6.764 8	26	876.321 3	6.057 3
2	18.866 8	7.724 3	27	178.839 7	6.747 5
3	14.935 0	7.825 8	28	73.859 4	7.131 6
4	12.319 5	7.909 4	29	217.281 3	6.663 0
5	7.901 9	8.102 3	30	338.315 7	6.470 7
6	54.756 1	7.261 6	31	16 402.260 0	4.785 1
7	306.263 6	6.513 9	32	12 722.380 0	4.895 4
8	47.059 5	7.327 4	33	16 011.110 0	4.795 6
9	18.850 7	7.724 7	34	28 189.960 0	4.549 9
10	109.720 6	6.959 7	35	17 592.470 0	4.754 7
11	163.604 1	6.786 2	36	5 138.327 0	5.289 2
12	59.039 6	7.228 9	37	7 027.366 0	5.153 2
13	484.941 5	6.314 3	38	2 589.491 0	5.586 8
14	225.971 2	6.645 9	39	12 616.700 0	4.899 1
15	25.415 9	7.594 9	40	16 069.630 0	4.794 0
16	33.910 7	7.469 7	41	20 479.670 0	4.688 7
17	77.262 5	7.112 0	42	20 274.040 0	4.693 1
18	23.154 7	7.635 4	43	18 208.620 0	4.739 7
19	51.764 4	7.286 0	44	3 810.479 0	5.419 0
20	10.867 2	7.963 9	45	20 479.670 0	4.688 7
21	59.901 6	7.222 6	46	25.258 5	7.597 6
22	42.564 2	7.371 0	47	15.521 8	7.809 1
23	131.867 2	6.879 9	48	17.926 3	7.746 5
24	119.821 9	6.921 5	49	29.875 1	7.524 7
25	876.283 6	6.057 4	50	3.997 3	8.398 2

从预测结果来看:样本编号 31—45 的 $Y_{IC_{50}}$ 值,均超过 2 500 nmol/L,其 $Y_{PIC_{50}}$ 低于 6,可以认为这些新化合物对抑制 ERα 活性效果较差,无法成为治疗乳腺癌的候选药物,后续研究可考虑优化分子描述符结构或剔除。

4　基于 GBDT 算法的 ADMET 性质识别

化合物成为治疗乳腺癌的良好药物,必须具备良好的生物活性和 ADEMT 性质。其中,ADME 主要指化合物的药代动力学性质,描述了化合物在生物体内的浓度随时间变化的规律,T 主要指化合物可能在人体内产生的毒副作用。一个符合标准的化合物需具备优良的活性,其次还需要具有容易吸收、代谢适中和无毒等性质。

在选用学习算法进行分类预测建模时,需要考虑算法适用性,分析比较几类常用机器学习算法会发现:*k*NN(*k*-Nearest Neighbor)算法有着低复杂度的优势,但其可解释性不强,且计算时间很长,效率不高;LDA(Linear Discriminant Analysis)算法容易出现过拟合情形,严重影响模型的预测精度,导致泛化能力较低;LR(Logistic Regression)算法简单易行,可解释性强,但是其预测准确率不高;NBC(Native Bayes Classification)算法则需要先验假设相互独立,而文章数据集不符合此假设,因此也不适用;而 GBDT 算法非常适用于文章 ADMET 性质的分类预测分析。首先,文章涉及代表值分类为二元分类问题;其次,算法不需要对数据进行放缩就可以进行分类,同时,该算法损失函数较为稳定,在数据处理时鲁棒性较强。不仅如此,GBDT 分类算法还充分考虑了每个分类器的权重,从而解决了本文的分类任务。因此,本文选择利用 GBDT 算法建立模型进行分类预测,同时选取查准率、*F*1 值、AUC 值 3 个评价指标作如下说明:

$$\text{查准率} = \frac{N_{TP}}{N_{TP}+N_{FP}}$$

$$F1 = \frac{2\times\text{查准率}\times\text{召回率}}{\text{查准率}+\text{召回率}}$$

AUC 值:ROC 曲线右下方的集合面积,一般 AUC 值的范围大于 0.5,在 0.85 以上为较强。

TP 表示被模型预测为正类的正样本,其值用 N_{TP} 表示;FP 表示被模型预测为正类的负样本,其值用 N_{FP} 表示。

4.1 GBDT 原理

GBDT(Gradient Boosting Decision Tree)[13] 是基于 Boosting 的梯度提升算法,采用此算法是因为它在算法可解释性上较强,且容易理解,预测湿度较快、精度较高。具体理论构建如下:

设训练集的特征和标签为

$$T=(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N),x_i\in\mathcal{X},y_i\in\{0,1\}$$

设二分类中的损失函数为 $L(y,f(x))$,则有

$$P(y=1\mid x)=\frac{1}{1+e^{-f(x)}}$$

$$P(y=-1\mid x)=\frac{e^{-f(x)}}{1+e^{-f(x)}}$$

$$L(y,f(x))=-\log P(y\mid x)=\log(1+e^{-yf(x)})$$

$$r_{ti}=-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}=\frac{y_i}{1+e^{y_i f(x_i)}}$$

则由 Newton-Raphson 迭代公式可得:

$$c_{tj}=\sum_{x_i\in R_{ij}} r_{ti}/\sum_{x_i\in R_{ij}}|r_{ti}|(1-|r_{ti}|)$$

4.2 实验结果与比较分析

4.2.1 化合物渗透性识别

针对 Caco-2 的识别,图 7 是常见机器学习算法及本文算法基于训练数据的混淆矩阵,表 5 是各个算法的查准率、AUC 值、F1 得分统计。结果显示:GBDT 查准率为 93.83%,AUC 值为 94.47%,F1 得分为 92.40%,横向对比其余 5 个算法,其具有更好的评估效果与识别能力。

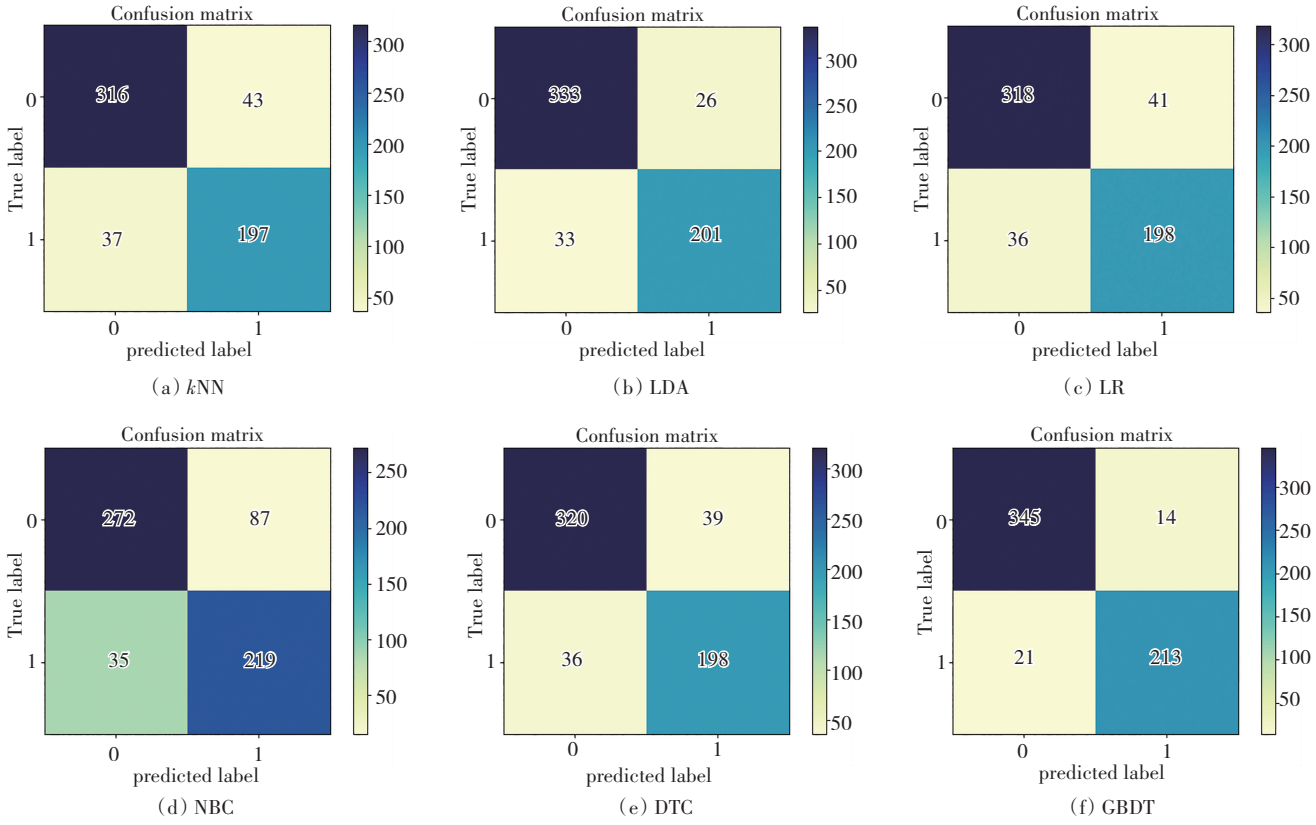


图 7 各类算法预测 Caco-2 混淆矩阵

Fig. 7 Confusion matrix of Caco-2 predicted by various algorithms

表 5 Caco-2 度量表

Table 5 Caco-2 metric scale

模 型	查准率	AUC 值	*F*1
*k*NN	0.820 8	0.858 0	0.831 2
LDA	0.885 5	0.897 6	0.872 0
LR	0.828 5	0.863 4	0.837 2
NBC	0.715 7	0.831 7	0.811 1
DTC	0.835 4	0.867 2	0.840 8
GBDT	0.938 3	0.940 5	0.924 1

4.2.2 化合物代谢能力识别

针对 CYP3A4 识别能力,图 8 和表 6 是常见机器学习算法及本文算法基于 CYPEA4 数据的混淆矩阵,结果显示:GBDT 的测试集表现最优,其查准率可以达到 97.03%,AUC 值为 93.68%,F1 得分为 96.81%,140 种化合物样本被准确分类到 0 类,390 种化合物被分类到 1 类中,识别能力很强。

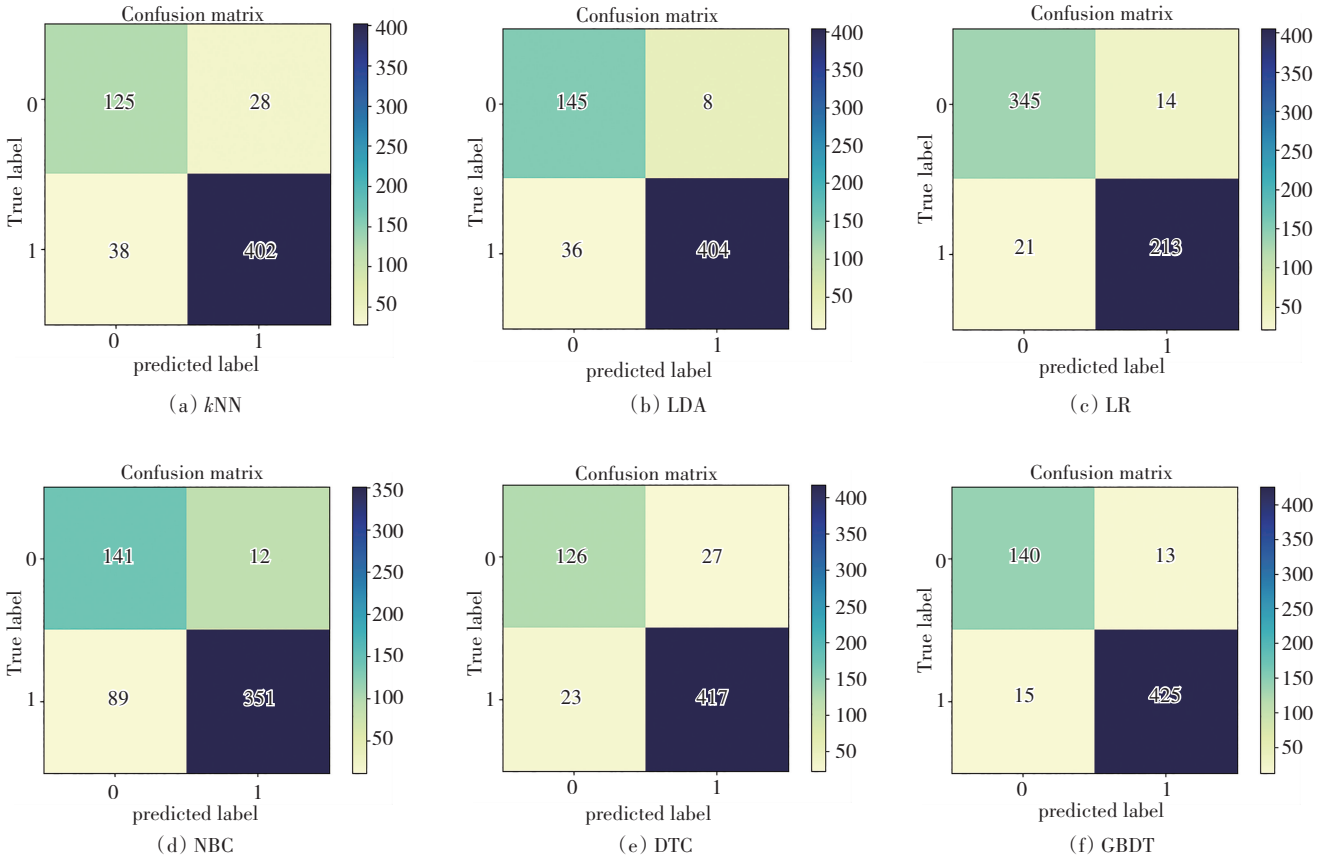


图 8　各类算法预测 CYP3A4 混淆矩阵

Fig. 8　Confusion matrix of CYP3A4 predicted by various algorithms

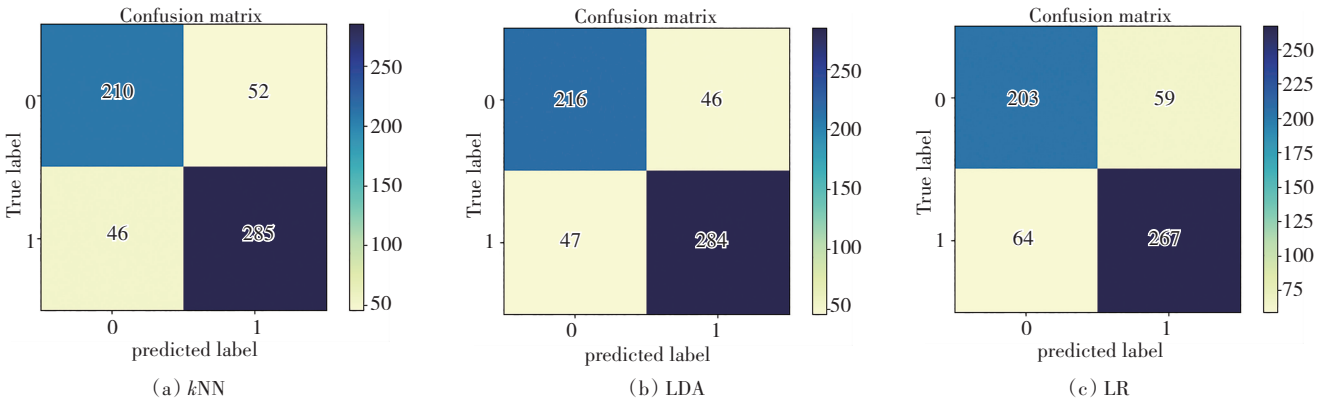
表 6　CYP3A4 度量表

Table 6　CYP3A4 metric scale

模　型	查准率	AUC 值	*F*1
*k*NN	0. 934 9	0. 850 9	0. 924 1
LDA	0. 980 6	0. 890 8	0. 948 4
LR	0. 950 7	0. 870 6	0. 935 3
NBC	0. 966 9	0. 790 0	0. 874 2
DTC	0. 939 2	0. 892 4	0. 943 4
GBDT	0. 970 3	0. 936 8	0. 968 1

4. 2. 3　化合物心脏毒性识别

图 9 和表 7 是常见机器学习算法及本文算法基于 hERG 数据的混淆矩阵。结果显示:GBDT 算法的查准率为 90. 61%、AUC 值为 89. 22%、*F*1 为 90. 47%。在心脏毒性识别中,测试集数据中有 231 种化合物被识别为 0 类,299 中化合物被识别为 1 类,识别能力最优。



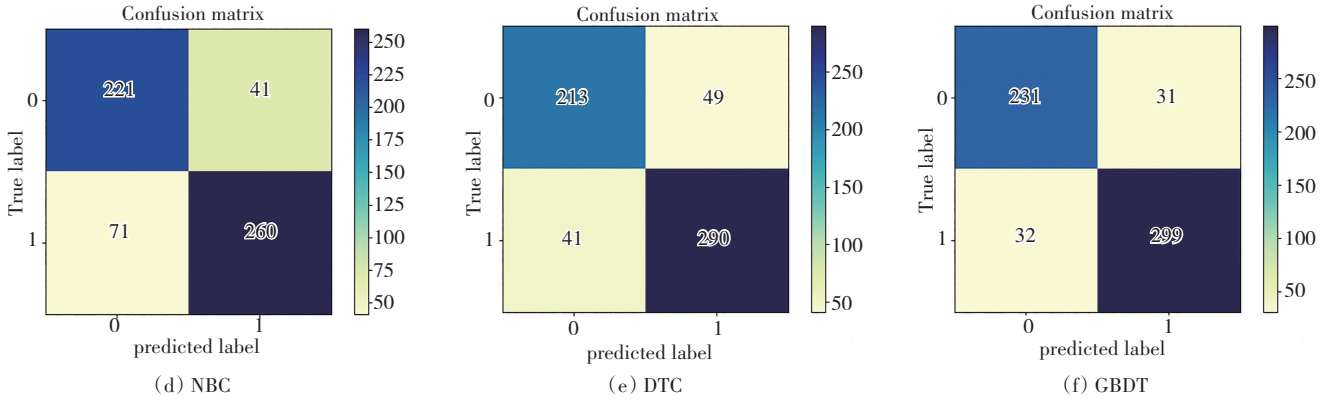


图 9　各类算法预测 hERG 混淆矩阵

Fig. 9　Confusion matrix of hERG predicted by various algorithms

表 7　hERG 度量表

Table 7　hERG metric scale

模　型	查准率	AUC 值	*F*1
*k*NN	0.845 7	0.833 0	0.853 3
LDA	0.860 6	0.840 9	0.859 3
LR	0.819 0	0.789 7	0.812 8
NBC	0.863 8	0.810 3	0.822 8
DTC	0.855 5	0.847 0	0.865 7
GBDT	0.906 1	0.892 2	0.904 7

4.2.4　化合物利用度识别

针对化合物利用度识别,图 10 是常见机器学习算法及本文算法基于 HOB 数据的混淆矩阵,结果显示:GBDT 算法的查准率为 75.00%、AUC 值为 82.86%、*F*1 为 73.17%(表 8)。测试集数据中有 411 种化合物被识别为 0 类,105 中化合物被识别为 1 类,识别效果相对最优。

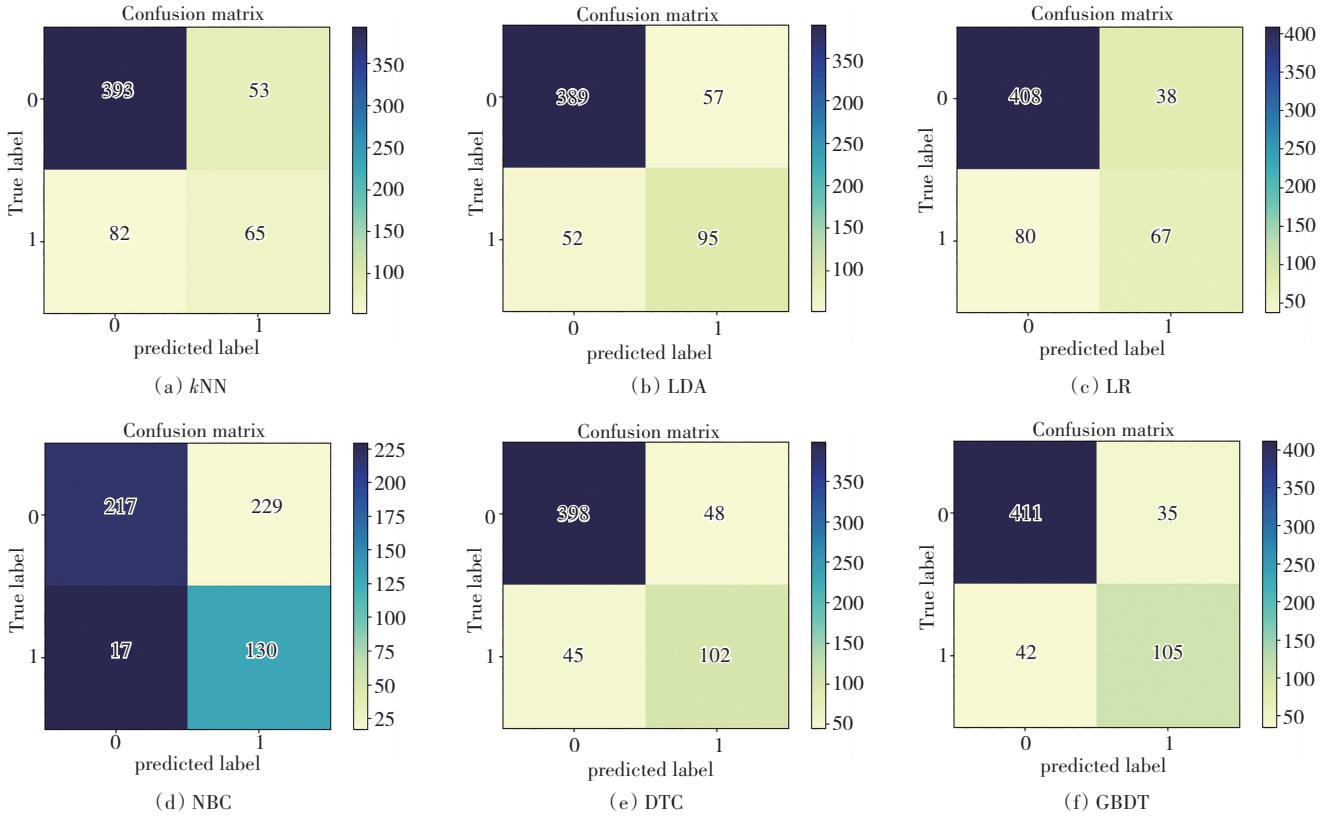


图 10　各类算法预测 HOB 混淆矩阵

Fig. 10　Confusion matrix of HOB predicted by various algorithms

表 8　HOB 度量表

Table 8　HOB metric scale

模　型	查准率	AUC 值	*F*1
*k*NN	0. 550 8	0. 689 1	0. 490 6
LDA	0. 625 0	0. 753 5	0. 635 5
LR	0. 638 1	0. 737 1	0. 531 7
NBC	0. 362 1	0. 644 7	0. 513 8
DTC	0. 680 0	0. 789 2	0. 686 9
GBDT	0. 750 0	0. 828 6	0. 731 7

4.2.5　化合物遗传毒性识别

针对化合物遗传毒性识别,图 11 是常见机器学习算法及本文算法基于 MN 数据的混淆矩阵。结果显示:化合物遗传毒性识别中,GBDT 算法的查准率为 96. 72%、AUC 值为 95. 77%、*F*1 为 97. 58%(表 9)。测试集数据中有 128 种化合物被识别为 0 类,443 中化合物被识别为 1 类,识别能力很强。

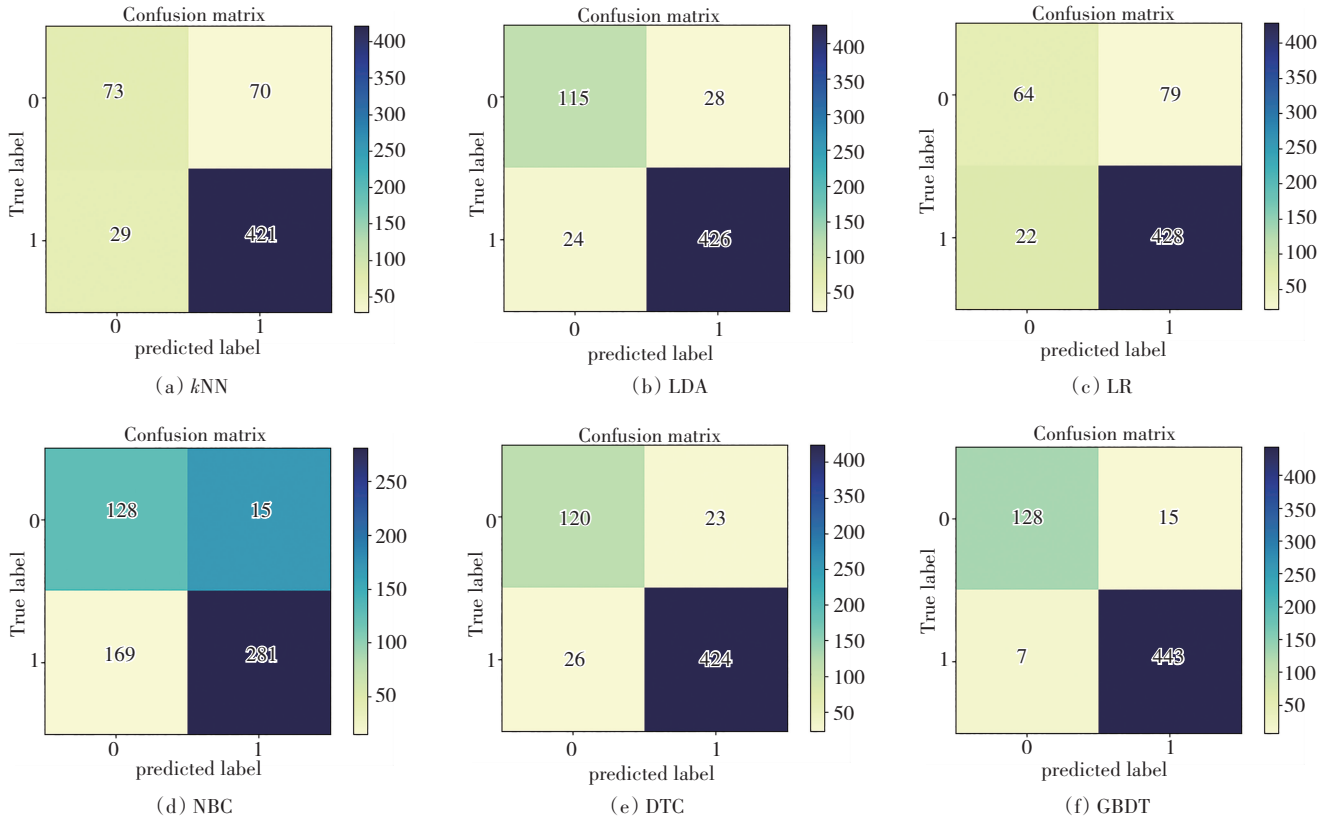


图 11　各类算法预测 MN 混淆矩阵

Fig. 11　Confusion matrix of MN predicted by various algorithms

表 9　MN 度量表

Table 9　MN metric scale

模　型	查准率	AUC 值	*F*1
*k*NN	0. 857 4	0. 786 6	0. 894 8
LDA	0. 938 3	0. 882 8	0. 942 5
LR	0. 844 2	0. 794 2	0. 894 5
NBC	0. 949 3	0. 690 2	0. 753 4
DTC	0. 948 5	0. 885 2	0. 945 4
GBDT	0. 967 2	0. 957 7	0. 975 8

综上,基于 GBDT 算法构建的分类预测模型在测试集中对 ADMET 性质的识别表现优越,且都保持了较高的预测准确性,因此本文将该模型应用在新化合物的 ADMET 性质识别中,从而判断新化合物的代谢能力、心脏毒性等,具体预测结果见表 10。

5　结束语

本文利用机器学习方法辅助实现抗乳腺癌候选药物研发,极大地节约了时间和成本,降低了人工误差。首先基于拮抗 ERα 的生物活性数据,利用稀疏贝叶斯学习以及随机森林算法,实现两阶段的变量筛选,并对 1 974 种化合物进行特征评估,得到 20 个重要特征;其次构建 GWO-KELM 算法进行 $Y_{IC_{50}}$ 与 $Y_{PIC_{50}}$ 的定量预测,并与传统的机器学习算法进行横向对比,证明本文改进算法的优越性,其均方误差最低,为 0. 598,拟合优度为 0. 709;最后利用 GBDT 算法分别构建 ADMET 性质的 5 个分类模型,进而对 50 种化合物做二分类预测,同时也做了机器学习算法的横向对比,其具有最优的预测结果,数据集上测试的分类 *F*1 分别为 92. 40%、96. 81%、

90.47%、73.17%、97.58%。本文算法相比一些传统机器学习算法,具有更好的预测效果,可以为抗乳腺癌候选药物研发提供预测服务,具有一定的实践价值。

表 10 GBDT 算法预测 ADMET 性质结果

Table 10 Results of GBDT algorithm for predicting ADMET properties

待测样本	Caco-2	CYP3A4	hERG	HOB	MN	待测样本	Caco-2	CYP3A4	hERG	HOB	MN
1	○	●	●	○	●	26	●	●	●	●	○
2	○	●	●	○	●	27	○	●	●	○	●
3	○	●	●	○	●	28	○	●	●	○	●
4	○	●	●	○	●	29	○	●	●	○	●
5	○	●	●	○	●	30	○	●	●	○	●
6	○	●	●	○	●	31	○	●	●	●	●
7	○	●	●	○	○	32	●	●	●	●	●
8	○	●	●	○	●	33	●	●	●	●	●
9	○	●	●	○	●	34	○	●	●	●	●
10	○	●	●	○	●	35	○	●	●	●	●
11	○	●	●	○	●	36	○	●	●	○	●
12	○	●	●	○	●	37	○	●	●	○	●
13	○	●	●	○	●	38	○	●	●	○	●
14	○	●	●	○	●	39	○	●	○	○	●
15	○	●	●	○	●	40	○	●	○	○	●
16	○	●	●	○	●	41	○	●	○	○	●
17	○	●	●	○	●	42	○	●	○	○	●
18	○	●	●	○	●	43	○	●	○	○	●
19	○	●	○	○	●	44	○	●	○	○	●
20	○	○	●	○	●	45	○	●	○	○	●
21	○	●	●	○	●	46	○	●	●	○	●
22	○	●	●	○	●	47	○	●	●	○	○
23	●	○	●	○	○	48	○	●	●	○	○
24	●	○	●	○	○	49	○	●	●	○	●
25	●	●	●	○	○	50	○	●	●	○	○

注:○表示该化合物不具备该性质,●表示具备该性质。

在进一步的研究中,拟从如下几个方面进行延伸:

在抗乳腺癌候选药物的筛选过程中,应该同时考虑将化合物 ERα 的生物活性以及 ADMET 性质进行综合评判,在化合物具有较好生物活性的前提下,保证其 ADMET 性质较好,诸如代谢能力、遗传毒性、渗透性等。

在充分挖掘结构性数据信息中,进一步可以采用图神经网络等深度学习方法,对化合物的一维线性表达式 SMILES 进行更深层的数据挖掘。

基于筛选重要化合物的分子描述符,进一步可以通过反向优化算法,确定分子描述符的最优阈值,进而调整化合物结构,使得 ERα 和 ADMET 性质具有更好的表现。

参考文献(References):

[1] LEI S, ZHENG R, ZHANG S, et al. Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020 [J]. Cancer Communications, 2021, 41(11): 83—94.

[2] HE H, SINHA I, FAN R, et al. C-Jun/AP-1 overexpression reprograms ERα signaling related to tamoxifen response in ERα-positive breast cancer[J]. Oncogene, 2018, 37(19): 86—100.

[3] LI X, PENG B, ZHU X, et al. MiR-210-3p inhibits osteogenic differentiation and promotes adipogenic differentiation correlated with Wnt signaling in ERα-deficient rBMSCs[J]. Journal of Cellular Physiology, 2019, 234(12): 75—84.

[4] MURATOV E N, BAJORATH J, SHERIDAN R P, et al. QSAR without borders[J]. Chemical Society Reviews, 2020, 49(11): 25—64.

[5] SHAR P A, TAO W Y, GAO S, et al. Pred-binding: large-scale protein-ligand binding affinity prediction[J]. Journal of Enzyme Inhibition and Medicinal Chemistry, 2016, 31(6): 43—50.

[6] 顾耀文, 张博文, 郑思, 等. 基于图注意力网络的药物 ADMET 分类预测模型构建方法[J]. 数据分析与知识发现, 2021, 5(8): 76—85.
GU Yao-wen, ZHANG Bo-wen, ZHENG Si, et al. Predicting drug ADMET properties based on graph attention network[J]. Data Analysis and Knowledge Discovery, 2021, 5(8): 76—85.

[7] 谢良旭, 李峰, 谢建平, 等. 基于融合神经网络模型的药物分子性质预测[J]. 计算机科学, 2021, 48(9): 251—256.
XIE Liang-xu, LI Feng, XIE Jian-ping, et al. Predicting drug molecular properties based on ensembling neural Networks models[J]. Computer Science, 2021, 48(9): 251—256.

[8] SHI T T, YANG Y W, HUANG S H, et al. Molecular image-based convolutional neural network for the prediction of ADMET properties[J]. Chemometrics and Intelligent Laboratory Systems, 2019, 194(1): 44—53.

[9] PENG Y Z, LIN Y M, JING X Y, et al. Enhanced graph isomorphism network for molecular ADMET properties prediction[J]. Journals & Magazines, 2020, 16 (8): 8344—8360.

[10] LAW V, KNOX C, DJOUMBOU Y, et al. Drug bank 4.0: Shedding new light on drug metabolism[J]. Nucleic Acids Research, 2014, 42(1): 91—97.

[11] WIPF D P, RAO B D. Sparse bayesian learning for basis selection[J]. IEEE Transactions on Signal Processing, 2004, 52(8): 53—64.

[12] MAXWELL A E, WARNER T A, FANG F. Implementation of machine-learning classification in remote sensing: An applied review[J]. International Journal of Remote Sensing, 2018, 39(9): 84—117.

[13] ZHANG J L, XU D, HAO K J, et al. FS-GBDT: Identification multicancer-risk module via a feature selection algorithm by integrating Fisher score and GBDT[J]. Briefings in Bioinformatics, 2020 23(3): 1—14.

责任编辑:李翠薇