Operations Research and Fuzziology 运筹与模糊学, 2023, 13(2), 778-788
Published Online April 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.132080

Hans汉斯

基于随机森林回归的乳腺癌抗药生物活性预测

龙荣进，袁　松，杨丽鑫，王飞云，周　洁

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2023年2月26日；录用日期：2023年4月9日；发布日期：2023年4月17日

摘　要

在药物研发中，雌激素受体α亚型(Estrogen receptors alpha, ERα)被认为是治疗乳腺癌的重要靶标，能拮抗ERα活性的化合物可能是治疗乳腺癌的候选药物。因此本文旨在以生物活性值pIC50作为因变量，作用于ERα靶标化合物的分子结构描述符作为自变量，构建关于ERα靶标化合物的生物活性预测模型，进而挑选出有效的抗癌候选药物。首先采用方差过滤法、随机森林、XGBoost以及灰色关联分析对自变量进行筛选，得到MDEC-23等16个与pIC50相关性强，且变量间相关性弱的分子结构描述符。其次建立随机森林回归生物活性预测模型，将预测结果与支持向量回归、梯度提升回归树、XGBoost模型和MLP回归模型预测结果进行对比分析，结果表明随机森林回归模型能更好地拟合数据，在R^2、*MAE*、*MSE*上优于其它模型，更适应于对生物活性pIC50值的预测，同时也表明筛选出的分子结构描述符在一定程度上能治疗乳腺癌。

关键词

生物活性预测模型，灰色关联，随机森林回归

Prediction of Antibiotic Activity of Breast Cancer Drug Resistance Based on Random Forest Regression

Rongjin Long, Song Yuan, Lixin Yang, Feiyun Wang, Jie Zhou

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Feb. 26th, 2023; accepted: Apr. 9th, 2023; published: Apr. 17th, 2023

Abstract

In drug development, the Estrogen receptors α subtype (ERα) is considered to be an important

文章引用: 龙荣进, 袁松, 杨丽鑫, 王飞云, 周洁. 基于随机森林回归的乳腺癌抗药生物活性预测[J]. 运筹与模糊学, 2023, 13(2): 778-788. DOI: 10.12677/orf.2023.132080

target for the treatment of breast cancer, and compounds that antagonize ERα activity may be candidates for the treatment of breast cancer. Therefore, this paper aims to use the biological activity value pIC50 as the dependent variable and the molecular structure descriptor acting on the ERα target compound as the independent variable to construct a prediction model for the biological activity of the ERα target compound, and then select effective anti-cancer drug candidates. Firstly, the independent variables were screened by variance filtering, random forest, XGBoost and gray correlation analysis, and 16 molecular structure descriptors such as MDEC-23 with strong correlation with pIC50 and weak correlation between variables were obtained. Secondly, a random forest regression biological activity prediction model is established, and the prediction results are compared and analyzed with the prediction results of support vector regression, gradient boosting regression tree, XGBoost model and MLP regression model, and the results show that the random forest regression model can better fit the data, is better than other models in R^2, *MAE* and *MSE*, and is more suitable for predicting the bioactive pIC50 value, and also shows that the screened molecular structure descriptors can treat breast cancer to a certain extent.

Keywords

Biological Activity Prediction Model, Grey Association, Random Forest Regression

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

乳腺癌是最常见的恶性肿瘤类型，也是全球女性癌症死亡的第二大原因[1]。2020 年全球新发乳腺癌 226.14 万例，占总体癌症发病的 11.7%，死亡 68.50 万例，占总体癌症死亡的 6.9% [2]，图 1 是 2020 年全球位于前十位癌症的数据统计图，由该图可知乳腺癌已成为全球范围内发病率最高的癌症，且死亡率已居所有恶性肿瘤的第五位[3]。据“我国进展期乳腺癌共识指南 2020 (CABC3)”报道，国内乳腺癌发病率逐年升高，每年 10 万人中约有 545.29 人患乳腺癌[3] [4]。因此研发治疗乳腺癌的药物是十分必要的。

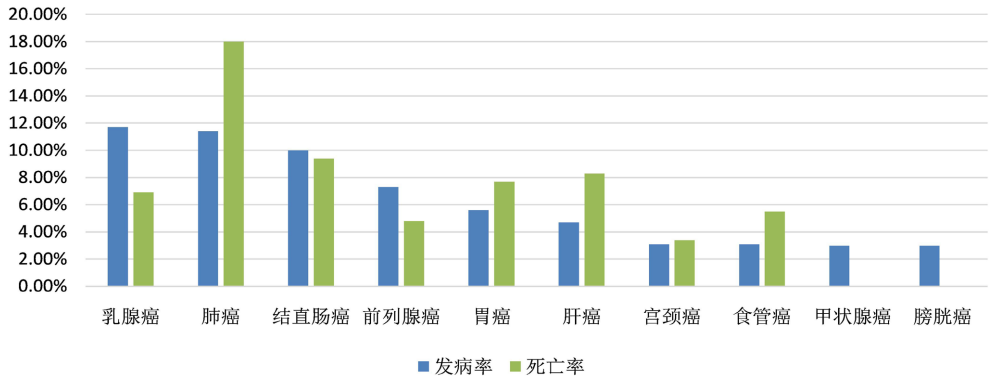

Figure 1. Statistics chart of the top 10 cancer in the world in 2020
图 1. 2020 年全球前十位癌症数据统计图

在药物研发中，雌激素受体 α 亚型(Estrogen receptors alpha, ERα)被认为是治疗乳腺癌的重要靶标[5] [6] [7]，因此能够拮抗 ERα 活性的化合物可能是治疗乳腺癌的候选药物。其次，为了节约时间与成本通

常采用建立“化合物活性预测模型”的方法来筛选潜在活性化合物[8]。具体做法是：针对与疾病相关的某个靶标(此处为 ERα)，收集一系列作用于该靶标的化合物及其生物活性数据，然后以一系列分子结构描述符作为自变量，化合物的生物活性值作为因变量，构建化合物的“定量结构－活性关系模型”，然后使用该模型预测具有更好生物活性的新化合物分子，或者指导已有活性化合物的结构优化。传统机器学习方法尤其是随机森林(rand forest, RF) [9]、支持向量机和人工神经网络在药物活性方面能够达到较高的预测精度[10]。

综上所述，本文以作用于 ERα 靶标化合物的分子结构描述符作为自变量，生物活性 pIC50 作为因变量，利用数据挖掘技术、机器学习等方法，先对变量进行筛选，变量合理验证等过程后构建关于 ERα 靶标化合物的生物活性预测模型，挑选出有效的抗癌候选药物，进而对生物活性 pIC50 进行预测，为乳腺癌的药物研发上提供理论支撑。

2. 数据描述与自变量筛选

2.1. 数据描述

共有 1975 × 730 个数据，如表 1 所示，其中第一列为化合物的结构，最后一列为生物活性 pIC50 值，其余列均为化合物的分子结构描述值。

Table 1. The first 10 rows of data display table
表 1. 前 10 行数据展示表

化合物结构	nAdd	ALogP	ALogP2	……	pIC50
Oc1ccc2O[C@H]([C@H](Sc2c1)C3CCCC3)c4ccc(OCCN5CCCCC5)cc4	0	−0.286	0.081796	……	8.602
Oc1ccc2O[C@H]([C@H](Sc2c1)C3CCCCCC3)c4ccc(OCCN5CCCCC5)cc4	0	−0.862	0.743044	……	8.125
Oc1ccc(cc1)[C@H]2Sc3cc(O)ccc3O[C@H]2c4ccc(OCCN5CCCCC5)cc4	0	0.7296	0.53231616	……	8.509
Oc1ccc2O[C@H]([C@@H](CC3CCCCC3)Sc2c1)c4ccc(OCCN5CCCCC5)cc4	0	−0.3184	0.10137856	……	8.409
Oc1ccc2O[C@H]([C@@H](Cc3ccccc3)Sc2c1)c4ccc(OCCN5CCCCC5)cc4	0	1.3551	1.83629601	……	8.131
Oc1ccc2O[C@H]([C@H](Sc2c1)c3ccncc3)c4ccc(OCCN5CCCCC5)cc4	0	−0.3921	0.15374241	……	6.310
Oc1ccc(cc1)C2=Cc3cc(O)ccc3C24Cc5ccc(OCCN6CCCCC6)cc5C4	0	1.5155	2.29674025	……	9.000
Oc1ccc2O[C@H]([C@H](Sc2c1)c3occc3)c4ccc(OCCN5CCCCC5)cc4	0	−0.1014	0.01028196	……	7.456
Oc1ccc(cc1)C2=Cc3cc(O)ccc3C24Cc5ccccc5C4	0	2.9208	8.53107264	……	8.367
Oc1ccc2O[C@H]([C@@H](Cc3ccccc3)Sc2c1)c4ccc(OCCN5CCCCC5)cc4	0	1.3551	1.83629601	……	7.041

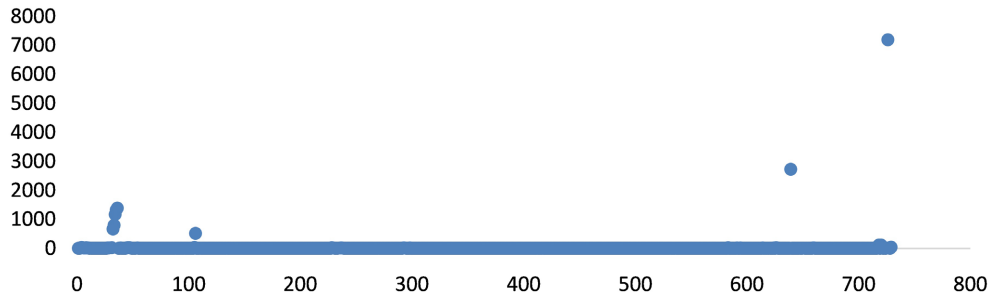

Figure 2. The standard deviation of the independent variable
图 2. 自变量的标准差图

数据描述性分析

计算出 729 个自变量的方差、标准差、均值、四分位点值，部分值如表 2 所示，由这些统计量可知，自变量之间的取值范围较大，建模前应该对数据进行标准化处理，再结合图 2，存在大量数据波动范围小的自变量，其自变量对建模效果产生的影响较小，可利用方差过滤法去除。

Table 2. Table of statistical values for the top 10 independent variables
表 2. 前 10 个自变量的统计量值表

分子结构描述符	方差	均值	最大值	最小值	中位数	1/4 分位数	3/4 分位数
nAcid	0.12103416	0.108409	4	0	0	0	0
ALogP	2.05707373	1.110164	5.1817	−23.105	1.17095	0.3763	1.9481
ALogp2	164.683756	3.288495	533.841	3.6E−07	1.56025081	0.40525983	4.018823
AMR	996.504243	116.5571	517.4294	54.067	114.8375	88.3037	141.4237
apol	378.292691	60.62647	359.6627	30.66193	59.901376	44.432102	74.42138
naAromAtom	26.5828256	15.44681	30	0	16	12	18
nAromBond	31.7562802	16.18946	34	0	18	12	18
nAtom	327.218497	50.7619	343	21	50	36.25	62
nHeavyAtom	65.1875486	28.11246	163	14	28	21	34
nH	116.111209	22.64944	180	5	22	14	29

2.2. 自变量筛选

在做预测模型时常常需要对变量进行筛选以达到降低变量个数、提升模型预测效果以及更便于实际应用的目的。本文中自变量共有 729 个，对于变量的筛选是很有必要的。

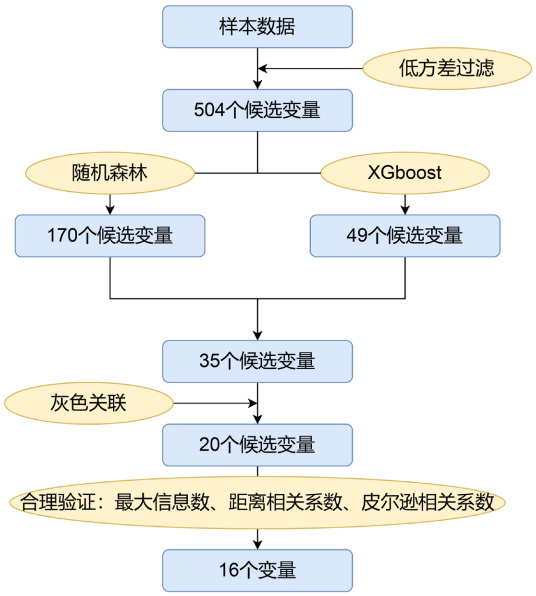

Figure 3. Variable filtering flowchart
图 3. 变量筛选流程图

2.2.1. 筛选流程

变量筛选流程如图 3 所示，通过简单观察数据，发现存在大量取值为 0 的列，以及存在较多数据波动不大的列(见图 2)，所以首先选择低方差过滤法进行变量选择，共过滤掉 225 个单一值特征变量，取得 504 个候选变量；其次利用随机森林、极致梯度提升(eXtreme Gradient Boosting, XGBoost) [11] [12]对 504 个候选变量分别进行筛选，再取交集，共得到 35 个候选变量，其中随机森林和 XGBoost 的变量累积重要度的取值为 80%；最后使用灰色关联分析得到 20 个候选变量(35 个候选变量的灰色关联度值见表 3)，其中灰色关联算法的分辨系数值选择 0.5；最终利用最大信息数(Maximal Information Coefficient, MIC)、距离相关系数、以及皮尔逊相关系数确定建模变量，共得到 16 个变量，其中 MIC 基于信息的非参数性探索，用于衡量两个变量 X 和 Y 之间线性或非线性的关联强度，常用于机器学习的特征选择性，皮尔逊相关系数删除与生物活性 pIC50 存在较大线性相关的变量，距离相关系数删除与 pIC50 存在较大非线性相关的变量。

根据图 3 了解到在变量筛选上：

1) 当累积重要度选择为 0.8 时，XGBoost 筛选的效果比随机森林的好，首先在筛选变量个数上少于随机森林筛选的变量个数，其次在 49 个变量中就有 35 个变量与随机森林筛选结果相同，保留了随机森林 20%的结果；

2) 利用机器学习筛选出的变量还存在较高的关系，如自变量之间、自变量与因变量之间存在线性相关或着非线性相关，必须对挑选的变量进行合理性验证。

Table 3. Grey correlation values for 35 candidate variables
表 3. 35 个候选变量的灰色关联度值

序号	分子描述符	灰色关联度	序号	分子描述符	灰色关联度
1	TopoPSA	0.769646	19	ndssC	0.739748
2	MDEC-23	0.748142	20	C1SP2	0.739685
3	nC	0.747339	21	minHBint4	0.739672
4	MDEC-33	0.742811	22	minHBint5	0.739632
5	WTPT-4	0.742606	23	MLFER_A	0.739591
6	LipoaffinityIndex	0.742298	24	SHsOH	0.739582
7	minsOH	0.742176	25	minsssN	0.739555
8	SP-6	0.741103	26	maxHsOH	0.739459
9	Kier3	0.741095	27	minHsOH	0.739443
10	SHBint10	0.740857	28	MDEO-12	0.739428
11	MLogP	0.740521	29	ETA_dEpsilon_A	0.739358
12	maxssO	0.740353	30	VC-5	0.739351
13	VPC-5	0.740306	31	VCH-5	0.739303
14	nHBAcc	0.740254	32	mindssC	0.739257
15	maxHBint10	0.740154	33	BCUTc-1l	0.739149
16	VPC-4	0.739998	34	gmin	0.738953
17	minHBint10	0.739992	35	ATSp5	0.388834
18	maxHBint5	0.739767			

2.2.2. 变量合理性验证

20 个变量与 pIC50 的 MIC 值、皮尔逊相关系数值、距离相关系数值如表 4 所示，由该表可知：

1) 此 20 个特征变量与生物活性 pIC50 之间的距离相关系数和最大信息数 MIC 都较大，体现了其对建模目标影响较大，验证了本题变量选择的合理性；

2) 所有变量的 Pearson 相关系数绝对值都未超过 0.6，可知这些特征变量与生物活性 pIC50 的线性相关性比较低，则在预测模型的选择上，非线性的回归模型比传统线性回归模型会有更好的表现；

3) 根据距离相关系数判别，可知还存在与生物活性 pIC50 中相关的特征变量，如 nHBAcc、C1SP2。

Table 4. Verify the information table properly
表 4. 合理验证信息表

序号	分子描述符	最大信息数(MIC)	距离相关系数	Pearson
1	MDEC-23	0.339381	0.241800	0.127535
2	LipoaffinityIndex	0.319822	0.193991	0.052755
3	MLogP	0.319369	0.281219	−0.242219
4	minsOH	0.309907	0.408549	0.053509
5	nC	0.308011	0.245233	−0.200817
6	TopoPSA	0.304106	0.467284	−0.444578
7	C1SP2	0.282867	0.553946	−0.514372
8	Kier3	0.273329	0.270442	−0.192635
9	WTPT-4	0.271333	0.327053	−0.265328
10	SP-6	0.265365	0.293177	−0.232848
11	maxssO	0.262426	0.256147	−0.03864
122	SHBint10	0.242552	0.261871	0.100614
13	MDEC-33	0.240234	0.18475	−0.110014
14	VPC-5	0.237669	0.143028	−0.068716
15	maxHBint10	0.235601	0.287701	0.134064
16	nHBAcc	0.234028	0.570577	−0.532135
17	minHBint10	0.230301	0.290532	0.147208
18	VPC-4	0.220728	0.165170	−0.146637
19	ndssC	0.192219	0.337085	−0.330453
20	maxHBint5	0.181000	0.151317	−0.068646

注：距离相关系数判别：中相关 0.4~0.6，弱相关 0.2~0.4，极弱或无相关 0~0.2。

为了避免自变量之间存在较高的相关性而对建模结果产生的影响，计算出这 20 个变量之间的距离相关系数，绘制得到图 4，由该图可知 20 个自变量之间存在距离相关系数大于 0.8 的现象，存在高度相关关系，如 vpc-5 和 vpc-4，maxHBint10 和 SHBint10，maxHBint10 和 minHBint10 等等，因此还需要对变量进一步筛选。

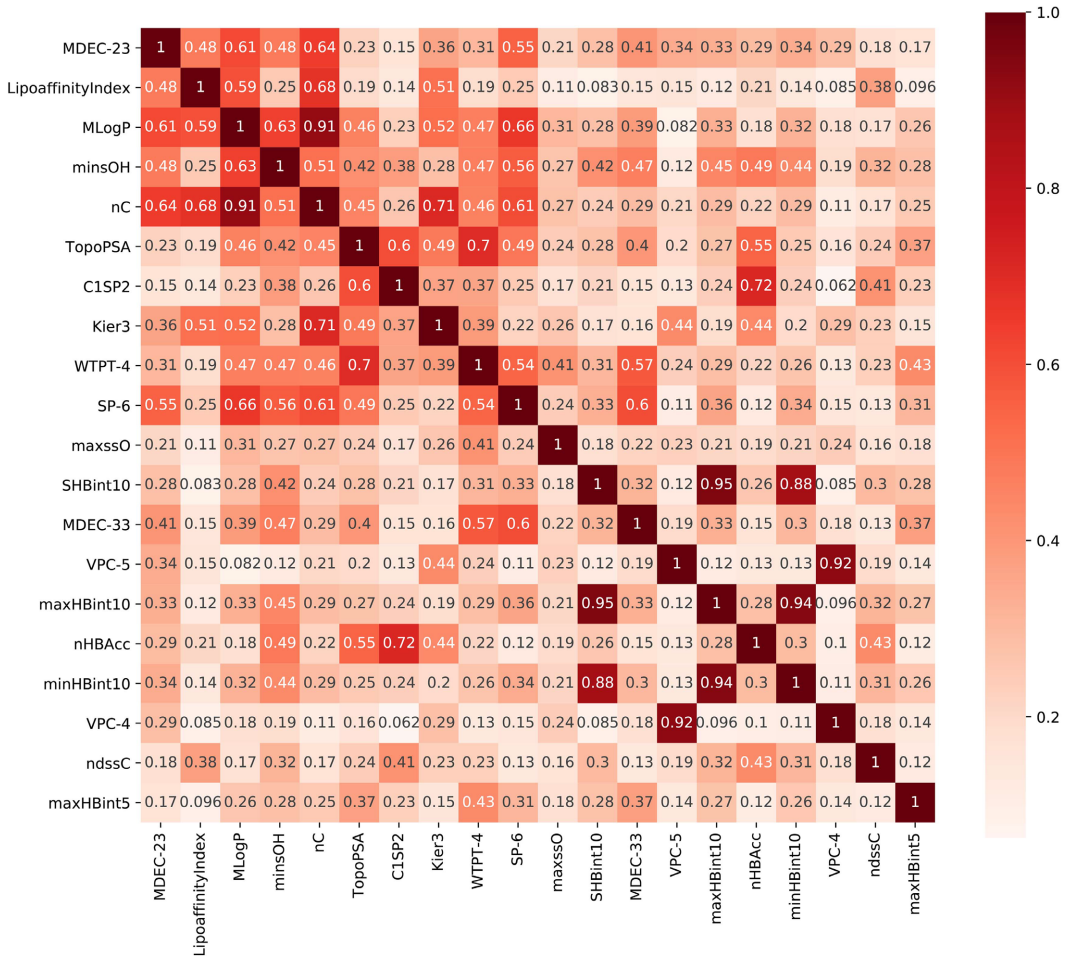

Figure 4. Graph of distance correlation coefficient between variables
图 4. 变量间的距离相关系数图

2.2.3. 变量最终确认

在选择出的对生物活性影响最大的 20 个自变量(见表 4)重，由于变量之间含有部分高度相关，若直接选用这 20 个变量参与建模，则并不能最大程度的得到最大量的有用信息，造成信息的浪费，因此还需要重新挑选出一个不超过 20 个变量的特征子集。根据变量之间距离相关系数图(见图 4)，对于变量之间相关性大于 0.8 的变量，考虑将其中一个自变量删除，最终剩余 16 个变量如表 5 所示。

Table 5. 16 molecular descriptors
表 5. 16 个分子描述符

序号	分子描述符	序号	分子描述符	序号	分子描述符
1	MDEC-23	7	Kier3	13	VPC-5
2	LipoaffinityIdex	8	WTPT-4	14	nHBAcc
3	MLogP	9	SP-6	15	ndssC
4	minsOH	10	maxssO	16	maxHBint5
5	TopoPSA	11	SHBint10		
6	C1SP2	12	MDEC-33		

3. 基于随机森林回归生物活性预测模型

3.1. 随机森林回归的基本思想

首先随机森林回归模型作为非线性生物活性预测模型。随机森林[13] [14]的基本思想是利用 bootstrap 重抽样方法从原始样本中抽取多个样本，对每个 bootstrap 样本构建决策树，然后将所有决策树预测平均值作为最终预测结果。随机森林回归可以看成是由很多弱预测器(决策树)集成的强预测器。

本文实现的 RF 是将多个二叉决策树打包组合而成的，训练 RF 便是训练多个二叉决策树。在训练二叉决策树模型的时候需要考虑怎样选择切分变量、切分点以及怎样衡量一个切分变量、切分点的好坏。针对于切分变量和切分点的选择，采用穷举法，即遍历每个特征和每个特征的所有取值，最后从中找出最好的切分变量和切分点；针对于切分变量和切分点的好坏，一般以切分后节点的不纯度来衡量，即各个子节点不纯度的加权和，其计算公式如下：

$$G\left(x_i, v_{ij}\right)=\frac{n_{left}}{N_s}H\left(X_{left}\right)+\frac{n_{right}}{N_s}H\left(X_{right}\right) \tag{1}$$

其中，x_i 为某一个切分变量，v_{ij} 为切分变量的一个切分值，n_{left}、n_{right}、N_s 分别为切分后左子节点的训练样本个数、右子节点的训练样本个数以及当前节点所有训练样本个数，X_{left}、X_{right} 分别为左右子节点的训练样本集合，$H(X)$为衡量节点不纯度的函数，在本题中选用 MSE 作为模型的不纯度函数。

3.2. 随机森林回归模型的建立

3.2.1. 建模过程

1) 论最优特征子集的选取

根据最终变量的确定，选择了 16 个变量参与构建生物活性预测模型，具体变量见表 5。

2) 数据标准化

数据标准化的公式如下所示：

$$x^*=\frac{x-\bar{x}}{\sigma} \tag{2}$$

进行数据标准化的原因有：① 将不同量级的数据统一转化为同一个量级，保证数据之间的可比性；② 将数据拉回成均值为 0，标准差为 1 的数据有利于回归模型的收敛。

3) 数据划分

按 8:2 的比例将 1974 行数据划分成训练集和测试集，样本比为 1580:394。用训练集训练模型，再用训练好的模型在测试集上验证效果。

4) 建立模型

5) 模型预测效果评估

本文采用 R^2、MSE、MAE 评价模型预测效果，公式如下：

$$R^2=1-\frac{\sum_i\left(\hat{y}_l-y_i\right)^2}{\sum_i\left(\bar{y}_l-y_i\right)^2} \tag{3}$$

$$MSE=\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_l\right)^2 \tag{4}$$

$$MAE=\frac{1}{m}\sum_{i=1}^{m}\left|y_i-\hat{y}_l\right| \tag{5}$$

其中，y_i，$\hat{y}_i$分别是测试集上的真实值和预测值。

3.2.2. 建立随机森林回归模型

对随机森林参数进行设置，设置树的个数为500棵，最大深度为10，得到随机森林回归模型，将其在测试集上的预测效果显示如图5所示，从中可以看出预测值与真实值走势大致相同，因此认为建立的模型有效。

4. 模型预测效果对比

为了说明所建立的随机森林模型有效性与优越性，按照同样的步骤对同一个训练集建立支持向量回归(Support Vector Regression, SVR)、梯度提升回归树、XGBoost回归以及多层感知机(MLP)回归四个模型。将五个模型的真实值与预测值对比图绘制如图6所示，由该图也可以看到随机森林回归很大程度上能更好地拟合真实值，更适应于ERα生物活性的预测。

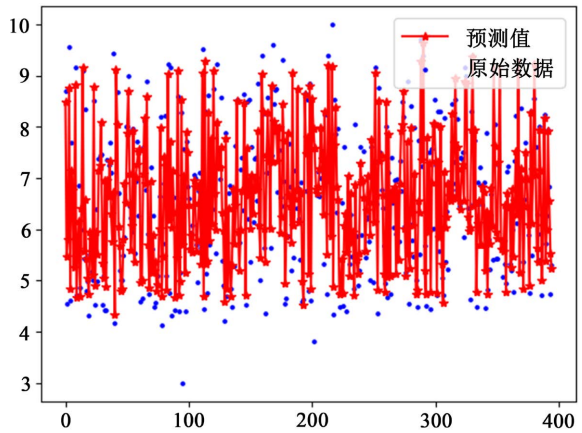

Figure 5. True value vs predicted value
图5. 真实值 vs 预测值

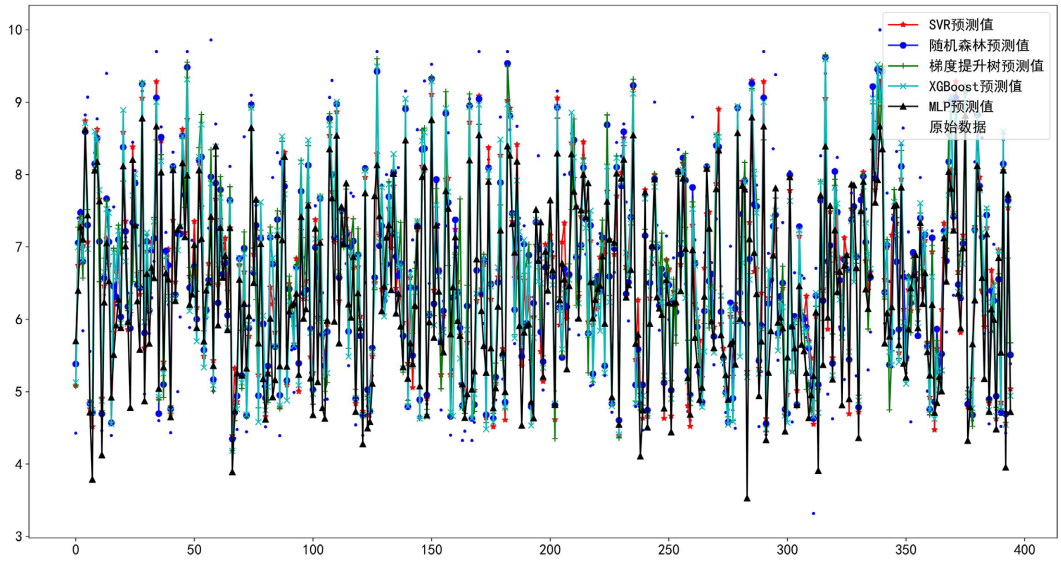

Figure 6. Model comparison diagram
图6. 模型对比图

Table 6. Indicator evaluation table
表 6. 指标评价表

模型	R^2	*MSE*	*MAE*
随机森林回归	0.7207	0.5705	0.5539
支持向量回归	0.6639 (↓7.881%)	0.6869	0.6226
梯度提升树回归	0.6791 (↓5.772%)	0.6553	0.5820
XGBoost	0.7117 (↓1.249%)	0.5885	0.5591
MLP 回归	0.5980 (↓17.025%)	0.8208	0.6971

为了避免偶然性，每个模型训练 50 次，计算五个模型的 R^2 值、*MSE* 值、*MAE* 值(见表 6)，发现随机森林模型的 R^2 值最高，相对于支持向量回归、梯度提升回归树、XGBoost 回归以及 MLP 回归依次提高了 8.557%、6.126%、1.265%、20.518%，同时与文献[15]建立的 GWO-KELM 模型和支持向量回归相比，本文的 R^2 分别提高了 1.722%、13.22%。从 *MSE* 值、*MAE* 值上，随机森林回归模型的值也是最小的，与文献[15]建立的 GWO-KELM 模型相比分别降低了 4.583%、3.586%。说明了本文所选的分子结构描述符更能准确地表达 pIC50 值，以及随机森林回归在生物活性 pIC50 上的预测效果更好。

5. 总结

基于机器学习方法在预测药物活性方面能够达到较高的预测精度的优点，本文利用随机森林等方法实现抗癌候选物的研发，与传统方法相比极大地节约了时间和成本，以及减少人工带来的误差。为了保证模型的可用性以及提高模型的预测效果，第一步选择对变量进行特征选取。在变量筛选上，本文层层递进，先挑选了对靶标 ERα 影响较大的 20 个化合物，再利用 MIC、距离相关系数、皮尔逊相关系数对变量进行合理性选择，避免多重共线性问题，最终挑选出 16 个分子结构描述符，其中约 1/2 的变量与何毅[9]和秦雅琴等人[8] (见表 5 标红色的分子结构描述符)所筛选的变量一致，保留了大部分能治疗乳腺癌的分子结构。第二步建立随机森林回归模型，同时为了说明该模型的有效性以及优越性，建立了支持向量回归[16]、梯度提升回归树、XGBoost 回归以及 MLP 回归模型。最后结合评价指标值、预测效果图的结果以及参考文献可知随机森林回归模型的预测效果在这五个模型当中是最优的，表明随机森林更适应于生物活性值的预测，该结论与叶丹等人[17]所得的结论相同，所以在对 ERα 的生物活性 pIC50 值的预测时，本文建议可选择随机森林回归。

参考文献

[1] 王三六. 术前全身炎症反应指数和纤维蛋白原/清蛋白比值联合分析对乳腺癌患者的意义[J]. 国际检验医学杂志, 2023, 44(3): 326-330+335.

[2] 中央人民政府. 74 种新药进医保谈判成功率再创新高[EB/OL]. http://www.gov.cn/zhengce/2021-12/04/content_5655779.htm, 2022-07-27.

[3] 刘宗超, 李哲轩, 张阳, 周彤, 张婧莹, 游伟程, 潘凯枫, 李文庆. 2020 全球癌症统计报告解读[J]. 肿瘤综合治疗电子杂志, 2021, 7(2): 1-14.

[4] 中国女医师协会乳腺疾病研究中心. 中国进展期乳腺癌共识指南 2020 (CABC3) [J]. 癌症进展, 2020, 18(19): 1945-1964.

[5] 刘昭国, 廖永德, 唐和孝. 雌激素受体在乳腺癌中的研究进展[J]. 肿瘤防治研究, 2012, 39(7): 869-871.

[6] 黄燕红, 李静, 董文武, 张浩, 单忠艳, 滕卫平. 雌激素受体 α、β 亚型在乳头状甲状腺癌中表达的临床及生物学意义研究[C]//中华医学会第十一次全国内分泌学学术会议论文汇编. 2012: 227.

[7] 刘训德. 雌激素受体 α 基因 XbaI 和 PvuII 多态性与乳腺癌及其不同分子亚型易感性的关系[D]: [硕士学位论文].

遵义: 遵义医科大学, 2019.

[8] 夏玉兰, 谢济铭, 王雅婧, 卢梦媛, 王锦锐, 秦雅琴. 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.

[9] 何毅, 马双宝, 孙彪. 基于随机森林的 ERα 生物活性预测研究[J]. 武汉纺织大学学报, 2022, 35(4): 54-56.

[10] 刘利梅, 陈晓晋, 孙世伟, 王宇, 王辉, 梅树立, 王耀君. 深度学习在药物活性预测研究中的应用[J]. 生物化学与生物物理进展, 2022, 49(8): 1498-1519.

[11] Chen, T. and Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[12] 李占山, 刘兆赓. 基于 XGBoost 的特征选择算法[J]. 通信学报, 2019, 40(10): 101-108.

[13] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[14] 方匡南, 吴见彬, 朱建平, 谢邦昌. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[15] 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/OL]. 重庆工商大学学报(自然科版): 1-12. http://kns.cnki.net.http.gzlib.proxy.chaoxing.com/kcms/detail/50.1155.N.20220928.1913.002.html, 2023-04-11.

[16] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000(1): 36-46.

[17] 叶丹, 胡二琴. 基于嵌入式特征选择算法下的抗乳腺癌药物分子活性预测[J]. 电脑知识与技术, 2022, 18(34): 8-10.