Modeling and Simulation 建模与仿真, 2023, 12(2), 1583-1592
Published Online March 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.122147

Hans汉斯

基于RFR模型的抗乳腺癌候选药物优化

宛翔天，杨家麒，刘兴宇，李江涛

上海理工大学，机械工程学院，上海

收稿日期：2023年2月21日；录用日期：2023年3月22日；发布日期：2023年3月29日

摘　要

在抗乳腺癌药物研发中，为了节省时间成本，建立化合物预测模型来筛选活性化合物是一种有效的方法。本文根据提供的乳腺癌治疗靶标雌激素受体α亚型(Estrogen receptors alpha, ERα)拮抗剂信息，利用Lasso回归与随机森林相结合的方法，对数据降维并筛选出影响生物活性的主要变量，明确了模型建立的方向；在此基础上，建立了关于生物活性的随机森林回归模型并进行了预测。

关键词

抗乳腺癌药物研发，预测模型，Lasso回归，随机森林回归

Optimization of Candidate Drugs for Anti-Breast Cancer Based on RFR Model

Xiangtian Wan, Jiaqi Yang, Xingyu Liu, Jiangtao Li

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 21st, 2023; accepted: Mar. 22nd, 2023; published: Mar. 29th, 2023

Abstract

In the research and development of anti-breast cancer drugs, it is an effective method to establish a compound prediction model to screen active compounds in order to save time and cost. According to the information of Estrogen receptors alpha (ERα) antagonist, which is the target of breast cancer treatment, this paper uses Lasso regression and random forest to reduce the dimension of the data and screen out the main variables that affect the biological activity, and makes clear the direction of establishing the model. On this basis, a random forest regression model about biological activity was established and predicted.

文章引用: 宛翔天, 杨家麒, 刘兴宇, 李江涛. 基于 RFR 模型的抗乳腺癌候选药物优化[J]. 建模与仿真, 2023, 12(2): 1583-1592. DOI: 10.12677/mos.2023.122147

Keywords

Anti-Breast Cancer Drug Development, Prediction Model, Lasso Regression, Random Forest Regression

1. 引言

根据世界卫生组织国际癌症研究署发布的《2020 年全球癌症负担报告》显示：全球乳腺癌新发病例高达 226 万例，已经超过肺癌(221 万例)成为全球第一大癌，占全球新发癌症病例的 11.7%。乳腺癌是目前世界上最常见、致死率较高的癌症之一[1]。而我国乳腺癌病症情况更加严峻。据统计，2020 年我国乳腺癌新发病例约 42 万，乳腺癌发病数高居全球第一，死亡人数约 11.7 万，居女性癌症死亡首位[2]。乳腺癌发病率之高，给我国女性带来了沉重的疾病负担。

因此，我国对乳腺癌药物的研发进程优化势在必行。研究发现，乳腺癌的发展与雌激素受体密切相关，雌激素受体 α 亚型(Estrogen receptors alpha, ERα)在小于等于 10%的正常乳腺上皮细胞中表达，但大约在 50%~80%的乳腺肿瘤细胞中表达；利用 ERα 基因缺失小鼠进行实验验证，发现 ERα 确实在乳腺发育过程中扮演了十分重要的角色。目前，抗激素治疗常用于 ERα 表达的乳腺癌患者，其通过调节雌激素受体活性来控制体内雌激素水平。因此，ERα 被认为是治疗乳腺癌的重要靶标，能够拮抗 ERα 活性的化合物可能是治疗乳腺癌的候选药物。当前，在药物研发中，通过建立化合物活性预测模型的方法来筛选有效活性化合物可以加快研发进展和降低研发成本。此处对乳腺癌的具体做法是：针对相关靶标(ERα)，收集一系列作用于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。

2. 模型建立

本文首先对 1974 个化合物的 729 个分子描述符(即变量)进行变量选择，根据变量对生物活性影响的重要性进行排序。通过观察处理后的数据发现，变量数大于测量次数，也就说明当用所有的变量去表示目标值的时候，数据矩阵无法做到列满秩，某些变量可以通过其他变量进行表示，即变量与变量之间存在着多重共线性。Lasso 回归在解决多重共线性的问题上具有优势。应用 Lasso 回归筛选出的变量能够很好地对模型进行表达，且各变量之间相互独立，因此选用 Lasso 回归方法进行主要变量的筛选。随机森林(RF)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 样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果。它具有很高的预测准确率，并可以根据变量的重要性进行排序。具体流程如图 1 所示，其中 x_n 为样本个数，$n=1,2,\cdots,729$；x_m 为 Lasso 初步降维筛选所得自变量，$m=1,2,\cdots,157$；x_r 为随机森林回归二次降维筛选所得自变量，$r=1,2,\cdots,20$。

2.1. Lasso 回归初步降维

根据经验规则，如果方差膨胀因子 VIF > 10，则认为该回归方程存在严重的多重共线性。经检验，数据材料所提供初始样本 VIF = 1.55e+06，存在高度多重非线性问题。如果多个解释变量之间高度相关，

则不容易区分它们各自对被解释变量的单独影响力。在应对多重共线性的问题中，Lasso 回归能够客观的筛选出有效变量从而解决多重共线性问题。在进行建模之前，需要将原始数据表格中变量全部为 0 的参数删除。从剩余 504 个变量中，找出主要的操作变量，同时，这些变量之间应不具有多重共线性，它们相互独立，具有代表性。而考虑到 Lasso 回归收缩估计量存在偏差，因此需要对这些选择出来的参数再次进行最小二乘法回归，并以此回归系数判定这些主要变量对于目标值的实际权重，根据权重系数去除趋近于 0 的系数的参数，最终就可以得出对于目标值最具影响力的参数变量。

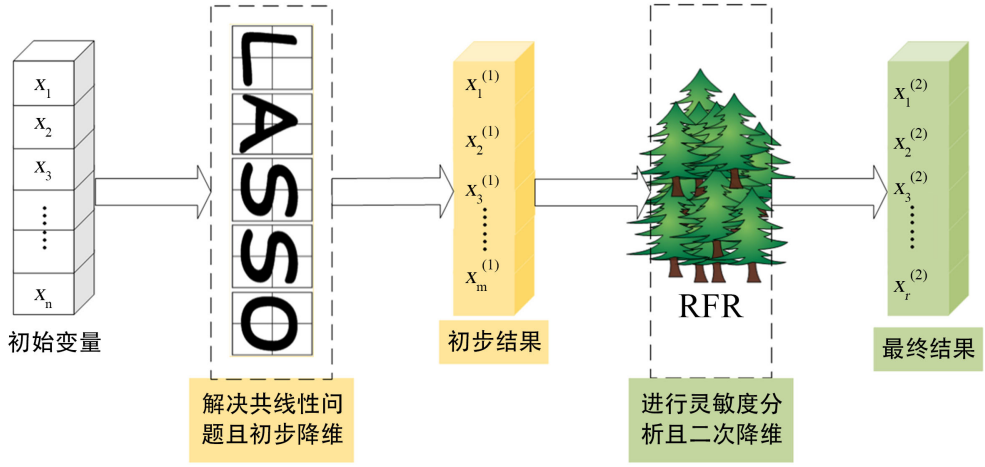

Figure 1. Flow chart
图 1. 思路流程图

Lasso 回归是 1997 年由 Tibshirani 提出的一种压缩估计方法，它的本质其实是正则化的最小二乘法，通过给最小二乘法添加正则化的惩罚函数[3]，让回归系数的绝对值之和在小于一个常数的约束条件下，使得回归模型残差平方和最小，产生严格等于零的回归系数，从而判定某些变量对于目标值的有效程度，筛选出有效变量[4]。

普通的最小二乘法线性回归对应的线性模型如下：

$$f(x)=w_1x_1+w_2x_2+\cdots+w_dx_d+b \tag{1}$$

用向量表示为，当用空间里的一个直观的模型取逼近这些数据点时，往往无法完全经过这些点，即模型理论值与数据值存在一定的偏差，其中 b 为偏差值[5]。此时我们需要这些误差的总和最小，而由于实际值分布较为分散，往往在理论模型的上下两侧，所以若是将这些误差直接相加，则会出现正负抵消的现象，从而无法正确地估计误差的大小。因此，需要将每一项的误差项平方，即将误差项正向化，根据这个方法，得出最小二乘法的误差项公式：

$$L(w)=\sum_{i=1}^{N}\left\|f(x_i)-y_i\right\|^2=\sum_{i=1}^{N}\left\|w^{\mathrm{T}}x_i-y_i\right\|^2=\sum_{i=1}^{N}\left(w^{\mathrm{T}}x_i-y_i\right)^2=\left(w^{\mathrm{T}}X^{\mathrm{T}}-Y^{\mathrm{T}}\right)\left(wX-Y\right) \tag{2}$$

将上式展开得，

$$L(w)=\left(w^{\mathrm{T}}X^{\mathrm{T}}-Y^{\mathrm{T}}\right)\left(wX-Y\right)=w^{\mathrm{T}}X^{\mathrm{T}}-2w^{\mathrm{T}}X^{\mathrm{T}}Y+Y^{\mathrm{T}}Y \tag{3}$$

为了得出使误差项最小得 w，对其求偏导，得出

$$w=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y \tag{4}$$

但是在这种情况下，为了满足矩阵可逆的条件下，只对于目标值 y 的个数大于变量值 x 的个数，一旦变量值 x 的个数过多，就会引起多重共线性和过拟合的现象[6]。

Lasso 回归为了解决这样得问题引入了扰动项构建正则化框架，使得矩阵变为可逆矩阵

$$\arg\min_{w}\left[L(w)+\lambda p(w)\right] \tag{5}$$

其中，$P(w)=\|w\|$，此时 Lasso 得回归模型为

$$\hat{w}(\text{lasso})=\arg\min_{w}\|y-Xw\|_2^2+\lambda\|w\|_1 \tag{6}$$

其中 $\|w\|_1=\sum_{j=1}^{P}|w_j|$ 为 w 上的 1 范数罚分，它可以稀释解的个数，λ 是一个调整参数，选择合适的 λ 能够调整最小二乘法拟合，并将一些分量得系数参数收缩为 0。通过这样得方法，Lasso 回归就将所有的变量筛选出对于目标值影响较大的显著变量，将影响较小的变量系数直接压缩至 0。

在对数据进行 Lasso 回归之前，由于数据变量较多，且量纲都不统一，因此，需要对数据先通过 Matlab 进行标准化，然后用 Stata 对标准化后的数据进行 Lasso 回归。当 λ 作为函数，在不同的 λ 约束条件下，回归系数也会发生很大的改变，当 λ 过大时，惩罚力度也会变大，使得所有的回归系数都归于 0。

紧接着，通过使用 K 折交叉验证的方法来选择最佳参数。K 折交叉验证，就是将样本数据随机分为 K 个等分。将第一个子样本作为“验证集”保留不动，其余的子集作为“训练集”来估计这个模型，再以此预测保留的第一个子样本，并计算这个子样本的均方误差。依次类推，将所有的子样本的均方误差加总，再通过调整参数，使得整个样本的均方误差最小，从而具有最佳的预测能力。

Lasso 回归不仅解决了高度多重共线性问题，还进行了样本变量的初步降维，一次降维后，筛选出 157 个变量。

2.2. 随机森林二次降维与结果分析

在 Lasso 回归降维结果的基础上进行随机森林二次降维。RFR 是集成学习的一种方法，它将许多决策树集成在一起以提高泛化能力。在训练阶段，RFR 使用 bootstrap 抽样从输入的训练数据集中收集几个不同的子训练数据集，依次训练几个不同的决策树。目的是使用几个不同的子模型，以增加最终模型预测结果的稳定性。对于切片变量和切片点的选择，我们遍历每个特征和每个特征的所有值，然后通过测量分割后节点的杂质(即每个子节点 $G(x, v)$的杂质的加权和)来确定最佳的分割变量和分割点。$G(x, v)$的公式由式(7)给出：

$$G\left(x_i+v_{ij}\right)=\frac{n_L}{N_s}H\left(X_L\right)+\frac{n_R}{N_s}H\left(X_R\right) \tag{7}$$

其中 x_i 是分割变量之一；v_{ij} 是分割变量的一个值；n_L、n_R、N_s 分别表示分割后左右子节点的训练样本数和当前节点的所有训练样本数；X_L 和 X_R 分别为左右子节点的训练样本集；$H(X)$是杂质函数。分类和回归任务的杂质函数 $H(X)$不同，本文采用均方误差(MSE)作为杂质函数。因此，$G(x, v)$的表达式由式(8)给出：

$$G(x,v)=\frac{1}{N_s}\left(\sum_{y_i\in X_L}\left(y_i-\bar{y}_L\right)^2+\sum_{y_j\in X_R}\left(y_j-\bar{y}_R\right)^2\right) \tag{8}$$

而决策树中节点的训练过程等价于寻找分割变量和 $G(x, y)$最小的分割点。

在 RFR 的基础上进行特征提取，计算特征的重要性。节点 k 的特征重要性由式(9)给出：

$$n_k = w_k \times G_k - w_L \times G_L - w_R \times G_R \tag{9}$$

其中 w_k、w_L、w_R 分别为节点 k 及其左右子节点的训练样本数与总训练样本数的比值；G_k、G_L、G_R 分别为节点 k 及其左右子节点的杂质。在得到每个节点的重要性后，我们可以通过式(10)计算特征的重要性：

$$f_i = \frac{\sum_j n_j}{\sum_m n_m} \tag{10}$$

其中 f_i 为特征 i 的重要性，j 为特征 i 上划分的节点，m 为所有节点。由式(11)给出归一化后的特征重要性公式：

$$f_{norm,i} = \frac{f_i}{\sum_{j \in all\ features} f_j} \tag{11}$$

将 f_{norm} 由大到小排序后，可以得到 f^*_{norm} 和累积重要性 C_i。

与传统方法不同，该方法可以对不同输出的重要变量进行筛选。由于在 RFR 中充分利用了不足的样本，特征提取的结果比传统的灵敏度分析更加可信。那些累积重要性通常大于 0.95 的特征可以被认为是各种输出的重要特征，构建 RFR 的总体过程如图 2 所示。

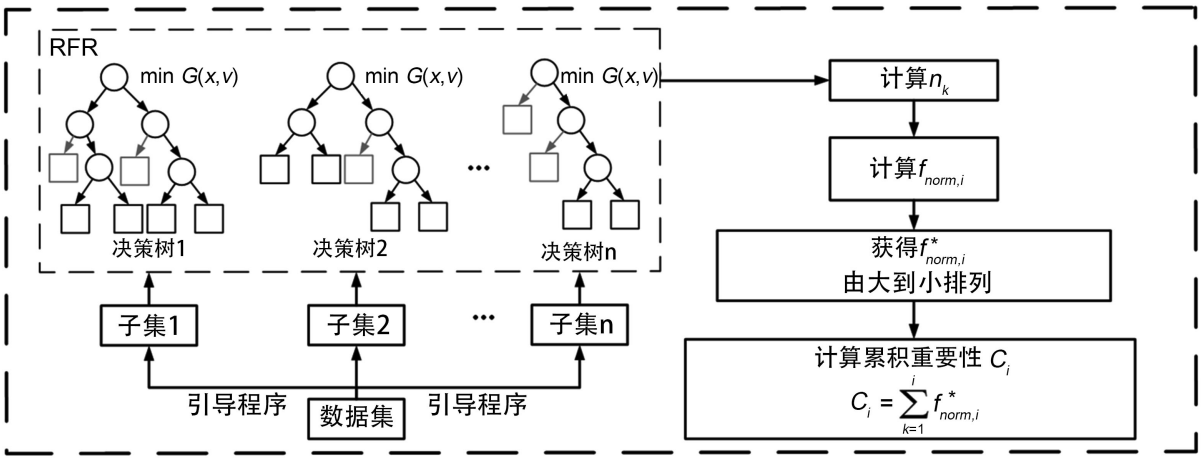

Figure 2. The overall structure process of RFR
图 2. RFR 的总体结构过程

这里，在 Lasso 回归降维基础上进行随机森林回归二次降维，根据变量对生物活性影响的重要性进行排序，并给出前 20 个对生物活性最具有显著影响的变量，结果如表 1 和图 3 所示。

3. 模型求解

本节利用上小节中筛选出来的不超过 20 个分子描述符变量，构建化合物对 ERα 生物活性的定量预测模型。然后使用构建的预测模型，对 50 个化合物进行 IC50 值和对应的 pIC50 值预测。本质上就是根据已有的变量数据，通过建立的模型去预测对应的 pIC50 值，同时用已知的样本数据验证所建立的模型的合理性。因为 IC50 值和对应的 pIC50 值之间存在函数关系，且 pIC50 值与生物活性值具有正相关性，即 pIC50 值越大表明生物活性越高；实际 QSAR 建模中，一般采用 pIC50 来表示生物活性值，且 pIC50 值幅度较小、精度更高，因此预测模型选择针对 pIC50，然后利用函数关系求解 IC50 值。

Table 1. The top 20 variables with the most significant effects on biological activity
表 1. 前 20 个对生物活性最具有显著影响的变量

	Name	Importance
1	maxHsOH	0.0425
2	minsssN	0.0382
3	MLFER_A	0.038
4	maxsOH	0.0365
5	ATSp5	0.0307
6	minssO	0.0302
7	BCUTp-1h	0.0265
8	hmin	0.0253
9	SaaCH	0.0229
10	MDEC-22	0.0228
11	C2SP2	0.0216
12	ECCEN	0.0214
13	CrippenLogP	0.0211
14	nHBAcc	0.0193
15	mindO	0.0163
16	SP-7	0.016
17	VPC-6	0.0155
18	ATSc3	0.0154
19	maxHBd	0.0151
20	BCUTc-1h	0.015

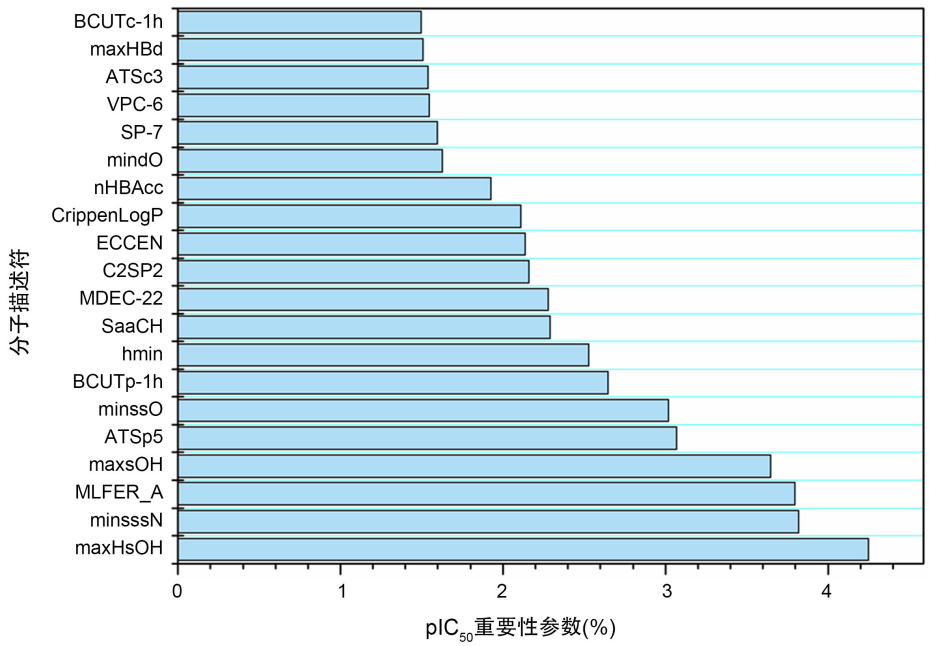

Figure 3. Importance ranking diagram of variables
图 3. 变量重要性排序图

3.1. 预测方法

在数学建模中，预测数据的模型有着很多不同的预测方法，并且不同的方法各有优劣，此处选择了三种不同的预测模型对数据进行预测，分别为径向基函数(RBF)、响应面法(RSM)和随机森林回归(RFR)，并通过三种方法的 MAPE 值进行对比寻找出最优的预测模型。

径向基函数(RBF)是一个取值仅仅依赖于离原点距离的实值函数。构造神经网络的基本方法为假设某种过程是属于某种函数空间的函数，然后连接成神经网格，运行一段时间该网络的电势趋于最小达到某种动态的平衡，从而可以求出该函数，而选择径向基函数空间是一个比较简单的容易用神经网络实现的方法。

响应面法(RSM)通过较少的试验在局部范围内比较精确的逼近函数关系，并用简单的代数表达式展现出来，计算简单。同时，通过回归模型的选择，可以拟合复杂的响应关系，具有良好的鲁棒性，并且还能获得显式表达。

随机森林回归(RFR)是集成学习的一种方法，它将许多决策树集成在一起以提高泛化能力。在训练阶段，RFR 使用 bootstrap 抽样从输入的训练数据集中收集几个不同的子训练数据集，依次训练几棵不同的决策树。目的是使用几个不同的子模型，以增加最终模型预测结果的稳健性和稳定性。随机林的优点是：它具有非常高的预测准确率，对异常值和干扰具良好的容忍度，且不易出现过度拟合。

RBF 神经网络[7]由输入层接收训练样本 $x_i \in R^d$，隐含层将输入样本通过径向基函数映射到新的空间，设隐含层节点数为 M，若径向基函数为高斯函数，则 $c_i \in R^d$ 表示高斯函数的中心向量，δ_i 表示高斯函数的核宽，由式(1)实现输入空间到新空间的映射：

$$\varphi\left(\left\|x-c_i\right\|\right)=\exp\left(-\frac{\left\|x-c_i\right\|^2}{\delta_i}\right) \tag{12}$$

输出层在新空间实现线性加权求和，设 w_i 为隐含层与输出层的连接权值，φ 为径向基函数，$y \in R$ 为输出结果，$R^d \to R$ 的映射函数为：

$$y=f\left(x\right)=\sum_{i=1}^{M} W_i \phi\left(\left\|x-c_i\right\|\right),\ i=1,2,\cdots,M \tag{13}$$

由此，RBF 神经网络完成 $f: R^d \to R$ 的非线性映射。

响应面法先选定采取的近似数学多项式模型的类型，然后将样本数据输入进该数学多项式模型中进行求解，得出一个初步的近似化模型，再用原始的样本数据带入计算得出的数学模型中求解新的输出，与原输出进行比较，计算其残差平方和，并以残差平方和最小为目标，进行数学模型中最佳项的选择条件。

这里采用的是二阶的响应面多项式方程，因为一阶的响应面多项式方程无法拟合复杂的模型，因此不选取一阶响应面模型。而对于高阶的响应面方程需要较多的数据变量个数来满足，此时也不满足，因此我们选用二阶响应面方程模型来进行模型预测。二阶响应面通式如下表示：

$$\hat{y}=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_M x_M+\beta_{M+1} x_1^2+\beta_{M+2} x_2^2+\cdots+\beta_{2M} x_{2M}^2+\sum_{i\neq j}\beta_{ij} x_i x_j \tag{14}$$

$$\mathrm{SSR}=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{15}$$

其中 $\hat{y}$ 为响应近似值，y_i 为响应实际值，n 是构造响应面模型的样本数，x_i 为样本参数值。

在数学模型确定后，还需要确定反馈方法，这里采用的是顺序替换法，即从常数项开始拟合，每次

增加一个项使回归平方和(SSR)最小，每增加一个项后，检查是否可以去掉或替换已经存在的项，同时使SSR 更小。顺序替换法的计算量较小，方便快捷。

随机森林回归原理在上文已经介绍，这里就不在赘述。不同的是，前文随机森林回归计算的是 Lasso 回归降维之后的 157 个变量，而这里的随机森林回归基于上节给出的前 20 个对生物活性值影响最大的变量。对五十个化合物进行生物活性值预测，图 4 为随机森林树、节点图的部分展示。

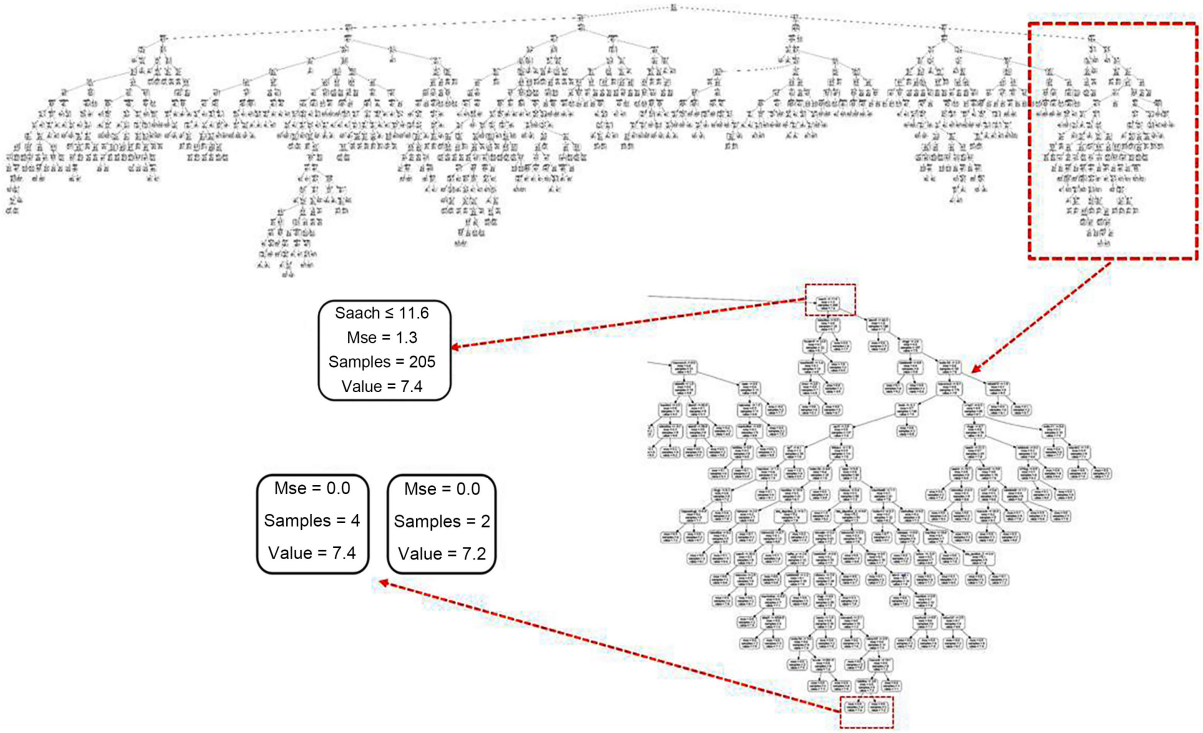

Figure 4. Random Forest-Tree and node diagram (part)
图 4. 随机森林－树、节点图(部分)

3.2. 精度验证

径向基函数和响应面法，误差与准确率比较接近。相比之下，随机森林回归具有更高的准确率，如表 2 和图 5 所示，将所有样本中 75%的样本作为训练集，25%的样本作为测试集，误差验证方式为 MAPE (8.34%)，准确率为 91.66%。

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{16}$$

Table 2. Comparison of accuracy rates of the three methods
表 2. 三种方法的准确率比较

	MAPE	准确率
径向基函数(RBF)	20.76%	79.24%
响应面法(RSM)	27.30%	72.70%
随机森林回归(RFR)	8.34%	91.66%

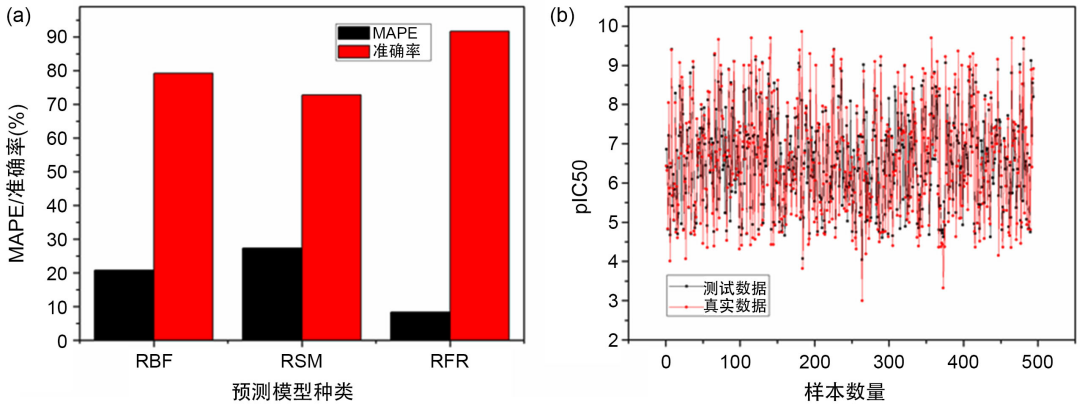

Figure 5. Model accuracy test. (a) Comparison of the accuracy of the three methods; (b) Comparison of the error of the RFR model
图 5. 模型精度检验 (a) 三种方法的准确率比较；(b) RFR 模型误差对比图

平均绝对百分比误差(MAPE)范围[0, +∞)，MAPE 为 0%表示完美模型，MAPE 大于 100%则表示劣质模型。

由图 5 可知，RFR 模型在对化合物的 pIC50 预测值与真实值有较高的拟合度，其 8.34%的平均百分比误差与 91.66%的准确率明显优于 RBF 模型和 RSM 模型。因此，使用基于 RFR 构建的预测模型，对 50 个化合物进行 pIC50 值预测，根据预测的 pIC50 值计算相对应的 IC50 值，其中 IC50 值是 pIC50 值的负对数，即 IC50 = 10^(9-pIC50)，结果如表 3 和图 6 所示。

Table 3. Forecast results
表 3. 预测结果

SMILES	IC50	pIC50
1	10.64902	7.97269
2	90.78559	7.041983
3	65.91116	7.181041
4	53.06057	7.275228
5	58.0825	7.235955
6	85.94277	7.065791
7	89.75984	7.046918
8	87.38266	7.058575
9	107.6548	6.967967
...	...	...
42	356.3759	6.448092
43	463.1999	6.334232
44	600.3307	6.221609
45	357.3254	6.446936
46	60.04358	7.221533
47	47.19953	7.326062
48	48.3786	7.315347
49	59.04007	7.228853
50	10.27856	7.988068

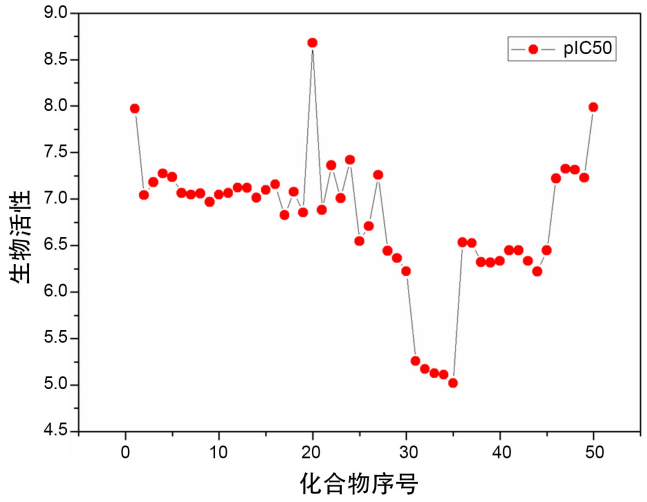

Figure 6. Prediction results of pIC50
图 6. 目标化合物 pIC50 预测结果

4. 结论

本文对 1974 个化合物的 729 个分子描述符进行变量选择，利用 Lasso 回归与随机森林回归对数据进行降维，再选择径向基函数、响应面法和随机森林回归这三种不同的预测模型对数据进行预测，并通过三种方法的 MAPE 值进行对比寻找出最优的预测模型，其中随机森林回归模型最优，并且其精度最好。在建模过程中，已经证明了筛选出的变量有足够的能力去构建预测模型，并且通过调参提高了模型的精度。从优化的结果不难看出药物研发所用化合物的生物活性存在不小差距，而根据预测结果可以得出抗乳腺癌候选药物的目标化合物 pIC50 的生物活性，以此来筛选出有效活性化合物进行药物研发，对加快研发进展和降低研发成本具有一定的帮助。

参考文献

[1] World Health Organization (2021) World Cancer Report 2020. http://www.iarc.who.int/featured-news/new-world-cancer-report/

[2] Cao, W., Chen, H., Yu, Y., *et al.* (2021) Changing Profiles of Cancer Burden Worldwide and in China: A Secondary Analysis of the Global Cancer Statistics 2020. *Chinese Medical Journal*, **134**, 783-791. https://doi.org/10.1097/CM9.0000000000001474

[3] 喻沩舸, 吴华瑞, 彭程. 基于 Lasso 回归的北京地区黄瓜价格波动原因分析[J]. 北方园艺, 2020(12): 165-170.

[4] 姜志堂, 李永. 基于 Lasso 模型的农业天气敏感性与服务效益研究[J]. 现代商贸工业, 2020, 41(30): 146-147.

[5] 胡聿文. 基于 LASSO 回归的江西省高校 R&D 投入强度的影响因素分析[J]. 科技和产业, 2020, 20(5): 84-88.

[6] 王纯杰, 温丽男, 马元嘉. 基于岭回归和 Lasso 回归的螺纹钢期货价格实证分析[J]. 吉林师范大学学报(自然科学版), 2020, 41(1): 36-41.

[7] 许敏, 胡丽丹. 径向基函数神经网络快速算法及其应用[J]. 统计与决策, 2021, 37(16): 52-56.