Modeling and Simulation 建模与仿真, 2023, 12(3), 1820-1828
Published Online May 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.123168

Hans汉斯

抗乳腺癌候选药物的生物活性预测建模

潘伟民

上海理工大学机械工程学院，上海

收稿日期：2023年2月24日；录用日期：2023年5月1日；发布日期：2023年5月8日

摘　要

本文构建化合物对Erα生物活性的定量预测模型，结合集成学习方法与逻辑回归方法，使用自动的参数优化方法，使各算法达到最优泛化性能。首先使用随机森林算法，以信息理论为基础，将化合物的分子描述符对雌激素受体α亚型的活性影响进行特征重要性排序，得到可用于算法判断的20个高效变量；再根据这20个高效分子描述符，利用岭回归算法实现对ERα生物活性的定量预测。结果表明，该模型可以准确预测Erα生物活性，为科学选择抗乳腺癌药物提供了新思路。

关键词

抗乳腺癌，分类预测建模，逻辑回归算法，随机森林算法，岭回归算法

Predictive Modeling of Bioactivity of Anti-Breast Cancer Drug Candidates

Weimin Pan

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

Received: Feb. 24th, 2023; accepted: May 1st, 2023; published: May 8th, 2023

Abstract

In this paper, a quantitative prediction model for ERα biological activity of compounds was constructed, combined with integrated learning method and logistic regression method, and automatic parameter optimization method was used to achieve the optimal generalization performance of each algorithm. First, based on information theory, random forest algorithm was used to rank the characteristic importance of the effects of molecular descriptors of compounds on the activity of estrogen receptor α subtypes, and 20 efficient variables were obtained. Based on these 20 molecular descriptors, ridge regression algorithm was used to quantitatively predict the biological activity of ERα. The results show that this model can accurately predict the biological activity of

文章引用: 潘伟民. 抗乳腺癌候选药物的生物活性预测建模[J]. 建模与仿真, 2023, 12(3): 1820-1828.
DOI: 10.12677/mos.2023.123168

Er*α*, which provides a new idea for scientific selection of anti-breast cancer drugs.

Keywords

Anti-Breast Cancer, Classification Predictive Modeling, Logistic Regression Algorithm, Random Forest Algorithm, Ridge Regression Algorithm

1. 引言

乳腺癌是目前世界上最常见，致死率较高的癌症之一[1]。乳腺癌的发展与雌激素受体密切相关。有研究发现[2]，雌激素受体 α 亚型(Estrogen receptors alpha, ERα)在不超过 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；而对 ERα 基因缺失小鼠的实验结果表明，ERα 在乳腺发育过程中扮演了十分重要的角色[3]。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。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[4]。比如，临床治疗乳腺癌的经典药物他莫昔芬和雷诺昔芬就是 ERα 拮抗剂。而 IC_{50} 和 pIC_{50} 作为影响 ERα 活性的重要指标，准确预测 IC_{50} 和 pIC_{50} 的值对药物的选取具有重要指导意义。因此，构建化合物对 Erα 生物活性的定量预测模型，对科学选取抗乳腺癌药物提供了新思路。

2. 变量对生物活性影响的重要性分析

2.1. 模型选取

本文针对 1974 个化合物的 729 个分子的数据进行预先处理，通过数据清洗发现存在与目标值无相关性的全零列，判断出来共有 225 列全零列。去除全零列可以降低模型的复杂度，提高模型的稳健性。

随机森林算法是最常用也是最强大的监督学习算法之一，它兼顾了解决回归问题和分类问题的能力。随机森林是通过集成学习的思想，将多棵决策树进行集成的算法，能够提供更好的预测性能[5] [6]。许多研究表明[7]，RF 算法对噪声、过拟合和异常值具有较高的预测精度。此外，与 SVR 等其他算法相比，该算法也具有许多优点，如处理高维度的复杂数据和具有更少的参数。给定一个数据集：

$$D=\left\{\left(X_i,y_i\right)\right\}\left(\left|D\right|\right)=n,X_i\in R^m,y_i \tag{1}$$

式中：n 为数据集样本的数量；m 为每个样本的特征个数。将数据集划分为训练集和测试集，其中训练集为：

$$S_L=\left\{\left(X_1,y_1\right),\left(X_2,y_2\right),\cdots,\left(X_L,y_L\right)\right\} \tag{2}$$

式中：L 表示训练集的数量。

2.2. RF 参数优化

随机森林算法把数据分为训练集(Trading Set)、验证集(Validation Set)、测试集(Test Set)，每次随机选出 n 组数据，用训练集训练出 n 个模型，测试集对 n 个模型进行评价，选出最终模型[7]。

在确定决策树算法的划分依据(criterion)时，可选择基尼系数(Gini)、信息熵(Entropy)、信息增益(Gain)

等参数。为使模型获得更好的参数，获得最佳泛化性能，对模型进行预剪枝处理，信息熵计算公式表示为：

$$\text{Entropy}(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|} \tag{3}$$

式中，$\text{Entropy}(D)$ 表示信息熵，k 为类别总数，D 为样本总数，C_k 为属于某个类别的样本数。

为使评估模型更加准确可信，在模型中使用网格搜索和交叉验证(GS-CV)的方式对参数进行优化[8]，GS-CV 首先遍历给定的参数组合来优化模型表现。进而将训练集继续拆分成训练集和测试集，对模型进行交叉验证。最终确定以 entropy 为划分依据，并限制决策数的最大深度 max_depth = 6，决策树数量为 n_estimators = 40，以防止过拟合现象。

2.3. 模型训练

采取随机有放回抽样，抽取 70%的数据作为训练集，30%的数据作为测试集。设置随机数种子为 0 以复现算法。

1) 利用 Bootstrap 方法对训练集随机采样[9]，得到各子训练样本集合和测试样本集合。

2) 利用步骤 1)得到的新的训练集分别建立多个基模型。在每棵树的节点处，先从 m 个特征中随机选取 t 个特征，使用这些特征中的最好的分裂方式对每一个节点进行分裂，使用的分裂方法是 CART [10]。

3) 将测试集带入训练好的决策树模型中，其预测值为：$\text{M1}(\text{X})$, $\text{M2}(\text{X})$, $\text{M3}(\text{X}),\cdots,\text{Mk}(\text{X})$。

4) 将所有决策树的预测值的众数作为随机森林模型的预测结果。

2.4. 重要性分析

通过上述四个步骤，建立模型；根据变量对生物活性影响的重要性进行排序，根据模型计算的重要性结果排序得到前 20 个对生物活性最具有显著影响的分子描述符(即变量)分别别是；LipoaffinityIndex、MDEC-23……MlogP，对生物活性最具有显著影响的前 20 个分子的重要度排序如图 1 所示。

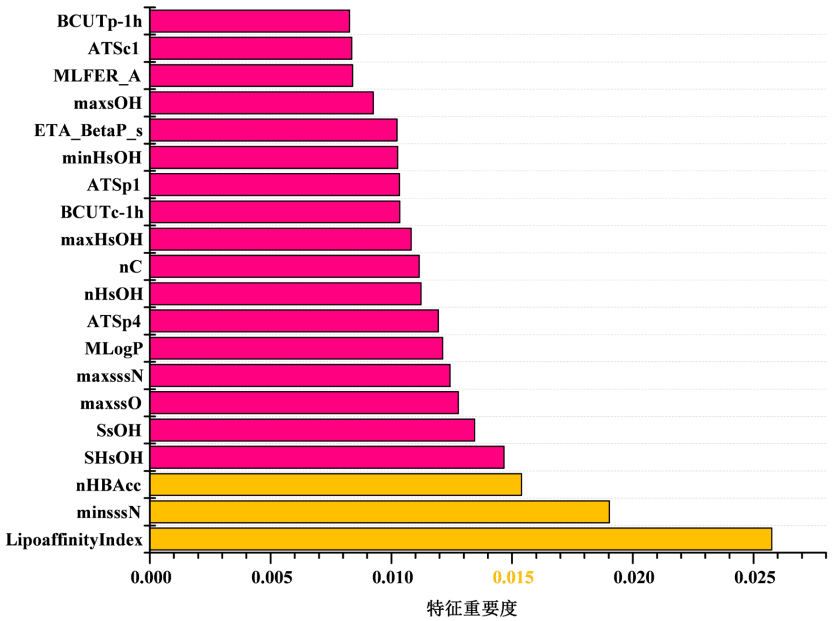

Figure 1. The top 20 descriptions of molecules that have the most significant effects on biological activity

图 1. 前 20 个对生物活性最具有显著影响的分子描述

3. IC_{50} 值及 pIC_{50} 值预测

3.1. 岭回归算法及优化算法

岭回归是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数，它是更为符合实际、更可靠的回归方法，对存在离群点的数据的拟合要强于最小二乘法[11]。不同与线性回归的无偏估计，岭回归的优势在于它的无偏估计，更趋向于将部分系数向 0 收缩。因此，它可以缓解多重共线问题，以及过拟合问题。

在标准线性回归中，通过最小化真实值 y_i 和预测值 $\hat{y}_i$ 的平方误差来训练模型，这个平方误差值也被成为残差平方和(RSS, Residual Sum of Squares)：

$$\mathrm{RSS}=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{4}$$

最小二乘法即最小化残差平方和，为

$$J_\beta(\beta)=\arg\min\sum_{i=1}^{p}\left(y_i-x_i\beta_i-\beta_0\right)^2 \tag{5}$$

将其转化为矩阵形式：

$$J_\beta(\beta)=\arg\min\left(\boldsymbol{y}-\boldsymbol{X}\beta\right)^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\beta\right) \tag{6}$$

求解为：

$$\beta=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{7}$$

将不适定问题转化为适定问题，在矩阵 $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)$ 的对角线元素上加入一个小的常数值 λ，然后取其逆得系数：

$$\hat{\beta}_{ridge}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda\boldsymbol{I}_{\boldsymbol{n}}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\right) \tag{8}$$

$\boldsymbol{I}_n$ 是单位矩阵，对角线全是 1，类似于“山岭”，λ 是岭函数，改变其数值可以改变单位矩阵对角线的值。

随后，代价函数 $J_\beta(\beta)$ 在 RSS 的基础上加入了对系数值的惩罚，矩阵形式为：

$$J_\beta(\beta)=\sum_{i=1}^{p}\left(y_i-X\beta\right)^2+\lambda\sum_{j=0}^{n}\beta_i^2=\sum_{i=1}^{p}\left(y_i-X\beta\right)^2+\lambda\|\beta\|^2 \tag{9}$$

损失函数：

$$J(a)=\left(h(x_1)-y_1\right)^2+\left(h(x_2)-y_2\right)^2+\left(h(x_3)-y_3\right)^2+\cdots+\left(h(x_m)-y_m\right)^2 \tag{10}$$

式中，$h(x_1)$ 为第 i 个训练样本特征值组合预测函数，y_i 为第 i 个训练样本的真实值。

通过梯度下降的方法[12]，以损失函数限制，优化模型精度，进行迭代求解，需要手动指定超参数，迭代求得最优参数的过程如图 2 所示。

$$w_1=w_1-\alpha\frac{\vartheta\cos t\left(w_0+w_1x_1\right)}{\vartheta x_1} \tag{11}$$

$$w_0=w_0-\alpha\frac{\vartheta cost\left(w_0+w_1x_1\right)}{\vartheta x_1} \tag{12}$$

α 为学习速率，α 旁边的整体表示方向沿着这个函数下降的方向线，最后就能找到山谷的最低点，

然后更新 w 值使用，面对训练数据规模十分庞大的任务，能够找到较好的结果。

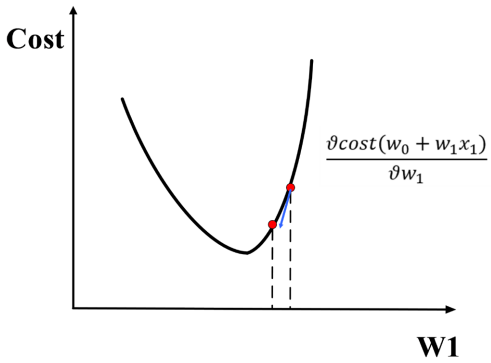

Figure 2. Gradient descent method
图 2. 梯度下降法

回归性能评估：

$$\mathrm{MSE}=\frac{1}{m}\sum_{i=1}^{m}\left(y^{i}-\bar{y}\right)^{2} \tag{13}$$

式中，y^{i} 为预测值，$\bar{y}$ 为真实值。

3.2. 模型训练

使用第二节提取的 20 个分子特征和及分子 $\mathrm{pIC_{50}}$ (相比于 $\mathrm{IC_{50}}$_nM，$\mathrm{pIC_{50}}$ 的数量级相差较小，更适合于回归模型的建立)为数据集，对岭回归模型进行训练。

岭回归算法中正则化力度越大，惩罚项占据主导地位，使得每个自变量的权重系数趋近于零；正则化力度越小，惩罚项的影响也越来越小，导致每个自变量的权重系数震荡的幅度变大，如图 3 所示。

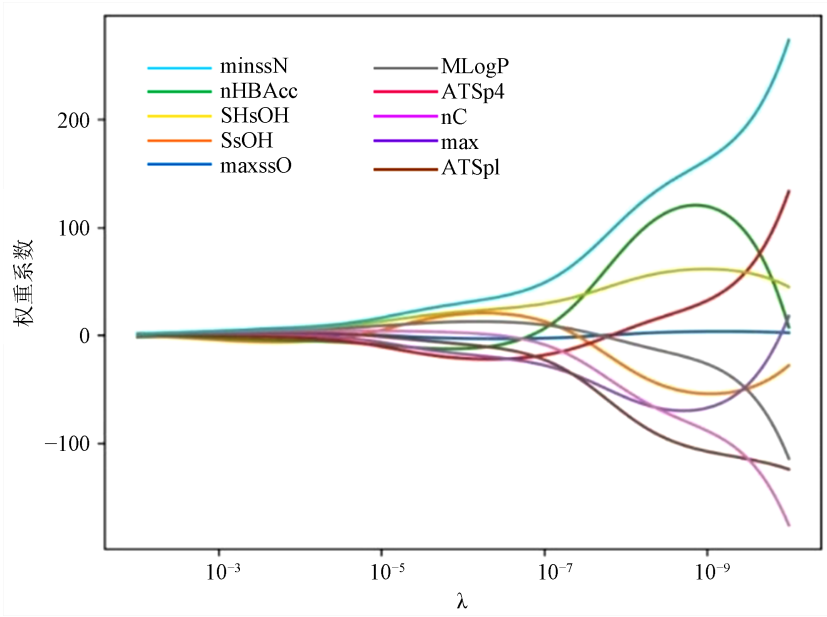

Figure 3. Relationship between regularization force and weight coefficient
图 3. 正则化力度与权重系数关系

另外，对迭代次数、回归偏置设置(模型结果更具稳健性)、是否进行标准化处理(进一步减小因噪声造成的误差)、优化器选择(在数据量增大的时候可以选择随机平均梯度法，进一步对收敛速度进行优化)设置见表 1。

Table 1. Predictor parameter settings
表 1. 预估器参数设置

参数设置	参数意义
alpha = 0.001	正则化力度 0.001
max_iter = 10000	迭代次数 10,000 次
fit_intercept = True	增加回归偏置
normalize = False	将数据标准化处理
solver = auto	优化器选择

3.3. 模型结果

将化合物的分子描述符与 pIC_{50} 指标建立映射关系，以 80%的数据作为训练集，20%的数据作为测试集。得到岭回归输出模型，保留小数点后三位有效数字后的回归模型如式所示。20 个分子描述(根据第二节所得)的权重系数如图 4 所示，(仅标注大于等于 1%的权重系数)，偏置为 4.924，计算模型的均方误差仅为 0.808。

$$\begin{aligned} h(x) &= w^{\mathrm{T}}x + b \\ &= -1.367x_1 + 1.238x_2 - 3.566x_3 - 1.298x_4 - 4.362x_5 - 2.997x_6 + 1.361x_7 + 1.909x_8 \\ &\quad + 2.439x_9 + 4.353x_{10}3.054x_{11} + 1.664x_{12} - 1.904x_{13} + 2.392x_{14} - 4.864x_{15} + 3.559x_{16} \\ &\quad - 3.891x_{17} + 1.123x_{18} - 8.451x_{19} + 2.145x_{20} + 4.924 \end{aligned} \tag{14}$$

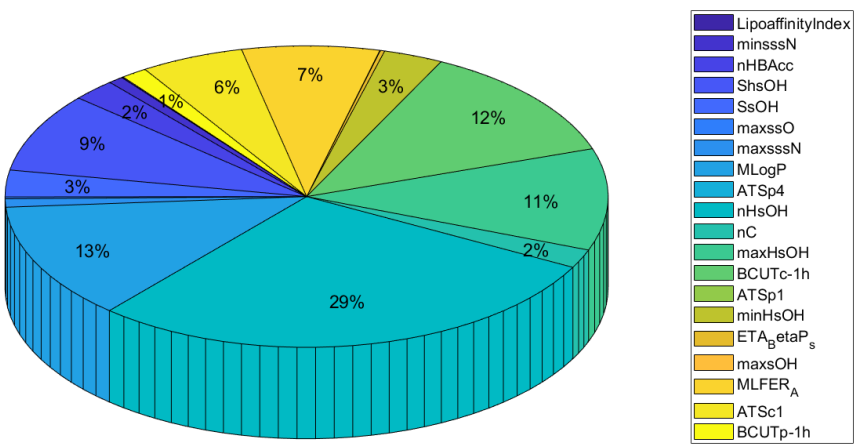

Figure 4. Molecular description weight coefficient
图 4. 分子描述权重系数

3.4. 模型验证

通过训练好的模型，对已有 50 个化合物进行 pIC_{50} 值和对应的 IC_{50} 值进行计算(表 2)。从图 5 中可以看到，预测值可随样本变化迅速并准确地靠近真实值，并且从图 6 和图 7 中可以看到，两个图的图像走向是对应的，这也印证了所建立的模型的准确性。

根据训练好的模型，对已有 50 个化合物进行 IC_{50} 值得预测，根据得到得 pIC_{50} 得值，按数学关系对 IC_{50} 值进行求解：

$$IC_{50} = 10^{(9-pIC_{50})} \tag{15}$$

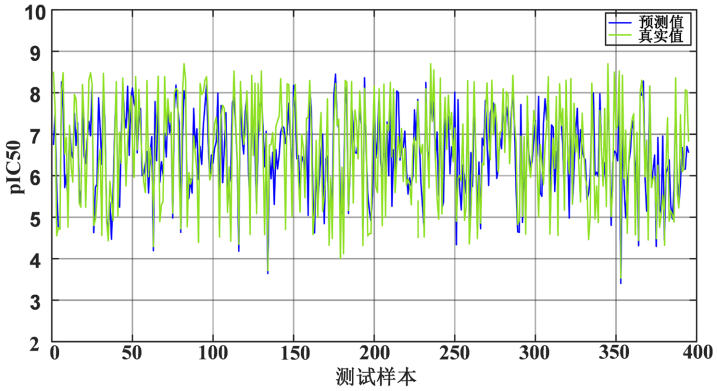

Figure 5. The comparison of the predicted value and the true value of pIC_{50}
图 5. pIC_{50} 预测值与真实值对比

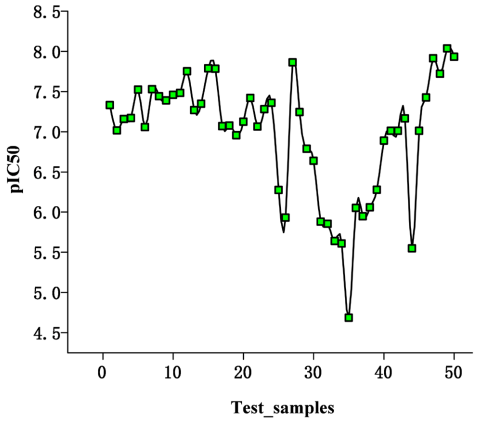

Figure 6. The predicted value of pIC_{50}
图 6. pIC_{50} 值预测值

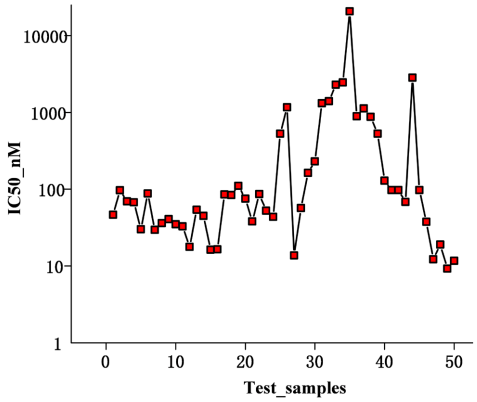

Figure 7. The calculated value of IC_{50}_nM
图 7. IC_{50}_nM 计算值

Table 2. Predicted values of IC_{50}_nM and pIC_{50}
表 2. IC_{50}_nM 和 pIC_{50} 预测值

序号	IC_{50}_nM	pIC_{50}
1	46.55346	7.33205
2	96.63147	7.01488
3	69.54647	7.15773
4	67.56091	7.1703
5	30.01964	7.52259
6	87.73265	7.05684
7	29.65402	7.52792
8	36.07385	7.44281
9	40.86543	7.38864
10	34.85911	7.45768
11	32.84461	7.48354
12	17.74137	7.75101
13	53.97665	7.26779
14	44.97879	7.34699
15	16.26201	7.78883
16	16.43805	7.78415
17	85.29723	7.06907
18	83.97985	7.07583
19	110.52438	6.95654
20	75.25369	7.12347
21	37.99344	7.42029
22	86.42777	7.06335
23	52.44823	7.28027
24	43.70377	7.35948
25	529.67998	6.27599
26	1168.39018	5.93241
27	13.71365	7.86285
28	56.94644	7.24453
29	163.2972	6.78702
30	230.00743	6.63826
31	1316.17499	5.88069
32	1398.45553	5.85435
33	2285.67959	5.64098
34	2460.29961	5.60901
35	20657.72462	4.68492
36	890.24088	6.05049
37	1124.31081	5.94911
38	872.50167	6.05923
39	528.03004	6.27734

Continued

40	129.36861	6.88817
41	97.21344	7.01227
42	97.40854	7.0114
43	68.35348	7.16524
44	2826.3823	5.54877
45	97.21344	7.01227
46	37.58566	7.42498
47	12.2468	7.91198
48	19.0186	7.72082
49	9.2282	8.03488
50	11.69484	7.93201

4. 结论

本文采用了随机森林算法，以信息理论为基础，将化合物的分子描述符对雌激素受体 α 亚型的活性影响进行特征重要性排序，得到可用于算法判断的 20 个高效分子描述符，使用岭回归算法实现对 ERα 生物活性的定量预测，选择线性度较好的 pIC_{50} 指标作为算法训练的目标值，相较于 IC_{50}，可获得更稳定的求解结果，所用建模方法在具有目标值的数据集上可以获得较好的表现，为科学选择抗乳腺癌药物提供了新思路。

参考文献

[1] 蒲星月, 马原, 钟志刚. 2006-2020 年中国女性乳腺癌死亡趋势分析——基于年龄-时期-出生队列模型[J]. 卫生经济研究, 2023, 40(2): 28-33.

[2] 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.

[3] Zhang, X.M., Wang, Y.Z., Li, X., Wu, J., Zhao, L.W., Li, W. and Liu, J. (2021) Dynamics-Based Discovery of Novel, Potent Benzoic Acid Derivatives as Orally Bioavailable Selective Estrogen Receptor Degraders for ERα+ Breast Cancer. *Journal of Medicinal Chemistry*, **64**, 7575-7595. https://doi.org/10.1021/acs.jmedchem.1c00280

[4] Alhammad, R. (2022) Bioinformatics Identification of TUBB as Potential Prognostic Biomarker for Worse Prognosis in ERα-Positive and Better Prognosis in ERα-Negative Breast Cancer. *Diagnostics*, **12**, 2067. https://doi.org/10.3390/diagnostics12092067

[5] 宋述芳, 何入洋. 基于随机森林的重要性测度指标体系[J]. 国防科技大学学报, 2021, 43(2): 25-32.

[6] 马骊. 随机森林算法的优化改进研究[D]: [硕士学位论文]. 广州: 暨南大学, 2016.

[7] 黄梅, 朱焱. 基于随机森林特征重要性的 K-匿名特征优选[J]. 计算机应用与软件, 2020, 37(3): 266-270.

[8] Shi, T. and Horvath, S. (2006) Unsupervised Learning with Random Forest Predictors. *Journal of Computational & Graphical Statistics*, **15**, 118-138. https://doi.org/10.1198/106186006X94072

[9] 于玲, 吴铁军. 集成学习: Boosting 算法综述[J]. 模式识别与人工智能, 2004(1): 52-59.

[10] 简治平. 基于集成学习的特征选择及稳定性分析[D]: [硕士学位论文]. 广州: 中山大学, 2010.

[11] 李炜. 机器学习概述[J]. 科技视界, 2017(12): 149.

[12] 吴冲, 潘启树, 李汉铃. 模糊线性回归预测[J]. 西安交通大学学报, 2000, 34(9): 100-102.