Modeling and Simulation 建模与仿真, 2025, 14(4), 73-85
Published Online April 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.144266

Hans汉斯

抗乳腺癌候选药物的优化建模

牛　寅

上海外国语大学贤达经济人文学院数据科学学院，上海

收稿日期：2025年3月2日；录用日期：2025年3月25日；发布日期：2025年4月7日

摘　要

女性乳腺癌发病人数首超肺癌，成为全球最常见癌症，了解乳腺癌的发病机理十分关键。经过研究ERα是治疗乳腺癌的重要靶标，如果能够对ERα及其相关性质进行深入研究，将会为患者带来希望和福音。本研究主要采用自适应LASSO、随机森林算法、BP神经网络算法等数据挖掘方法，建立了一系列变量筛选模型、pIC_{50}生物活性值预测模型，具体做法如下：针对问题一，为消除量纲影响，标准化处理1974个化合物的729个分子描述符。模型一采用随机森林算法直接对特征进行筛选排序。模型二先用LASSO方法从原始变量中筛选出105个重要变量，剔除无关或者影响不太显著的部分变量，再借助随机森林算法对保留下来的变量进行重要性排序。对比两个模型，发现两个模型筛选的主要变量中有5个相同；基于组合模型筛选的变量均通过显著性检验且拟合优度较高；单一模型的筛选结果中存在3个不显著变量。因此，选定LASSO-随机森林算法组合模型对变量的筛选排序结果作为最终输出。针对问题二，将数据集按照8:2的比例划分为训练集、测试集。首先，基于问题一的结果，采用逐步回归法和自适应LASSO法在训练集上分别进行二次筛选，根据投票原则综合确定17个变量作为最终筛选结果和后续神经网络的输入层变量。其次，由于因变量和自变量之间存在非线性关系，采用BP神经网络进行建模，把已知的测试集样本代入训练好的模型中发现拟合效果较好，表明所构建模型的有效性。

关键词

pIC_{50}，随机森林，自适应LASSO，BP神经网络

Optimization Modeling of Candidate Drugs for Breast Cancer Treatment

Yin Niu

School of Data Science, Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai

Received: Mar. 2nd, 2025; accepted: Mar. 25th, 2025; published: Apr. 7th, 2025

Abstract

The number of female breast cancer cases has exceeded that of lung cancer for the first time, making

breast cancer the most common cancer globally. Therefore, understanding the pathogenesis of breast cancer is crucial. Through research, it has been found that ERα is an important target for breast cancer treatment. If in-depth research can be carried out on ERα and its related properties, it will bring hope and good news to patients. The study mainly uses data mining methods such as adaptive LASSO, random forest algorithm, and BP neural network algorithm to establish a series of variable screening models and pIC_{50} bio-activity value prediction models. The specific approaches are as follows: For problem 1, to eliminate the influence of dimensionality, 729 molecular descriptors of 1974 compounds were standardized. In Model 1, the random forest algorithm was directly used to screen and rank the features. In Model 2, the LASSO method was first used to select 105 important variables from the original variables, eliminating some irrelevant or less significant variables. Then, the random forest algorithm was used to rank the importance of the retained variables. By comparing the two models, it was found that there were 5 identical main variables selected by the two models. The variables selected by the combined model all passed the significance test and had a relatively high goodness-of-fit. There were 3 insignificant variables in the screening results of the single model. Therefore, the variable screening and ranking results of the LASSO - random forest algorithm combined model were selected as the final output. For problem 2, the dataset was divided into a training set and a test set at a ratio of 8:2. First, based on the results of problem 1, step-wise regression and adaptive LASSO methods were used for secondary screening on the training set respectively. According to the voting principle, 17 variables were comprehensively determined as the final screening results and the input-layer variables for the subsequent neural network. Second, since there is a non-linear relationship between the dependent variable and the independent variables, a BP neural network was used for modeling. Substituting the known test - set samples into the trained model showed a good fitting effect, indicating the effectiveness of the constructed model.

Keywords

pIC_{50}, Random Forest Algorithm, Adaptive LASSO, BP Neural Network

1. 问题重述

1.1. 问题的背景

在刚过去的2020年，世界卫生组织国际癌症研究机构(IARC)发布了2020年全球最新癌症负担数据：女性乳腺癌发病人数首超肺癌，成为全球最常见癌症。乳腺癌是一种高发且困扰着许多女性的疾病，2020年中国有42万新发病例，且发病率也在逐年上升[1]。但是，在常见的恶性肿瘤之中，乳腺癌的5年生存率是相对较高的，据2018年《柳叶刀》上发表的全球癌症生存状况监测数据显示，2010~2014年间，我国的乳腺癌5年生存率为83.2%。而在医疗发达的国家，乳腺癌的5年生存率可达90%以上。

随着科学的不断进步，乳腺癌的治疗方法也开始向分子学和基因技术上发展[2]。人们对乳腺癌病发的癌基因、肿瘤生成、细胞凋零等的认知越来越高，并且建立了相关的衡量指标[3] [4]。有研究发现，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。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[5]。

目前，在药物研发中，为了节约时间和成本，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[6]。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(Quantitative Structure-Activity Relationship, QSAR)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[7] [8]。

1.2. 要解决的问题

根据提供的 ERα 拮抗剂信息(1974 个化合物样本，每个样本都有 729 个分子描述符变量，1 个生物活性数据)，构建化合物生物活性的定量预测模型，从而为优化 ERα 拮抗剂的生物活性提供预测服务。我们需要解决以下两个问题：

问题一：根据文件“Molecular_Descriptor.xlsx”和“ERα_activity.xlsx”提供的数据，针对 1974 个化合物的 729 个分子描述符进行变量选择，借助数学建模方法建立基于数据分析的特征选择模型，进行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分子描述符。

问题二：在问题一的基础上，选择不超过 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，对文件“ERα_activity.xlsx”的 test 表中的 50 个化合物进行 IC_{50} 值和对应的 pIC_{50} 值预测。

2. 模型符号说明

1) 符号 $X_i\left(i=1,2,\cdots,729\right)$，表示分子描述符变量；

2) 符号 $ZX_i\left(i=1,2,\cdots,729\right)$，表示标准化后的分子描述符变量。

3. 问题一的模型建立与求解模型假设

3.1. 问题分析

针对乳腺癌治疗靶标 ERα，根据文件“Molecular_Descriptor.xlsx”和“ERα_activity.xlsx”提供的 1974 个化合物的 729 个分子描述符信息和 1974 个化合物对 ERα 的生物活性数据，我们首先对数据进行初步的分析，发现给出的数据存在量纲不同和变量自身变异大小等数据缺陷问题，进行标准化处理后，分别采用随机森林的特征提取方法和基于 LASSO-随机森林组合的特征提取方法对 1974 个化合物的 729 个分子描述符进行变量选择。根据拟合效果，判断最优特征子集。

3.2. 数据标准化

在文件“ERα_activity.xlsx”的 training 表(训练集)中，提供了 1974 个化合物的结构式，用一维线性表达式 SMILES (Simplified Molecular Input Line Entry System)表示。SIMLES 表达式每一个都是单独的实验，所以我们认为提供的数据都是真实观测到的，不需要进行异常样本的剔除工作。

标准化处理的实质是将数据按一定比例缩放，使之落在一个特定的区间，便于不同单位或量级的数据能够进行比较和加权。分析文件“Molecular_Descriptor.xlsx”和“ERα_activity.xlsx”提供的数据发现，分子描述符之间量纲不同且数值大小变化有差异，直接进行特征提取，将会影响变量选择的准确性和可靠性。为了消除数据不同量纲的限制，在此需要采用 Z-Score 标准化方法对部分变量差异比较大的进行标准化处理，将其转化为无量纲的数值。

$$Z=\frac{X-\bar{X}}{s},\ \bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\ s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}.$$

其中，Z 是标准化后的数据，X 为原始数据，$\bar{X}$ 为均值，s 为标准差。

经过标准化处理的数据消除了不同量纲和取值范围的影响，但保留原来数据存在的关系。获得的问题一标准化数据，处理后的数据共包含有 729 组与 1974 个化合物的变量，用于后续特征的提取。

3.3. 特征提取

3.3.1. 特征选择方法调研

特征选择是模型建立前的重要步骤。过多的变量特征会造成维数灾难，不仅影响机器学习的学习效率，而且也会影响模型学习的效果。删除不相关的特征变量能够在减少模型的训练时间的同时防止模型过拟合，提高模型的精确度。

机器学习中的特征选择有过滤式、包裹式和嵌入式三种[9]。对这些特征选择方法的优点与缺点进行调研，得到结果如表 1 所示。

Table 1. Comparison of feature selection
表 1. 特征选择比较

特征选择方法	优点	缺点
过滤式	简单并且能够快速地搜索出有效的特征子集	特征选择过程独立于机器学习模型，因此对提高机器学习模型的性能效果一般
包裹式	选择的特征能够提升机器学习模型性能，使模型取得更理想的学习效果	计算开销大、效率低，同时容易产生过拟合，因此在高维特征筛选时的性能较差
嵌入式	能够较好、较快的完成特征选择，且计算复杂度介于过滤式和包裹式方法之间	只有部分模型有这个功能，如：随机森林、XGBoost、内置正则化的回归模型

综上，由于包裹式方法需要不断地筛选变量和训练学习器，使得其计算效率和计算开销都相对较差，本文考虑综合过滤式方法和嵌入式方法对本问题进行研究。如图 1 所示，本问题首先采用过滤式方法中的 LASSO 算法对 729 个变量进行初步筛选来剔除部分不重要的变量，然后采用嵌入法中的随机森林算法对筛选后的重要变量进行重要性评估。

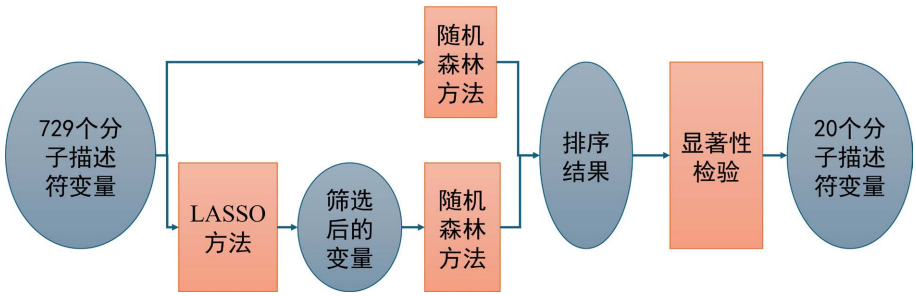

Figure 1. Flow chart of problem one
图 1. 问题一流程图

3.3.2. 基于随机森林的特征选择方法

近年来，机器学习方法在各领域都有广泛的应用，且随机森林算法在特征选择方面效果较好。随机森林算法是一种以决策树为学习器的集成学习方法。该方法的原理是随机生成多棵决策树并进行训练，最后结合其预测结果得到最终输出[10]。因此，随机森林算法相比于单一决策树算法来说，有更强的鲁棒性、泛化能力强且不易发生过拟合现象。

本研究建立随机森林模型对特征的重要性进行排序，将初始的 729 个变量嵌入模型中，结合这些变量和因变量之间的关系评估特征的重要性。本研究的 729 个初始变量与因变量之间并不是简单的线性关系，进一步说明了本研究选用随机森林算法的合理性。

随机森林主要运用 Gini 指数法或袋外数据错误率度量变量的重要性，其中，Gini 指数法使用最广泛的一种分割规则。假设集合 T 包含 k 个类别的记录，那么 Gini 指数为

$$\mathrm{Gini}(T)=1-\sum_{j=1}^{k}p_j^2.$$

其中，p_j 表示类别 j 出现 T 的频率。当 $\mathrm{Gini}(T)$ 最小为 0 时，即在此节点中所有记录都属于同一类别，表示此时能得到最大的有用信息；当此节点中的所有记录类别字段来说是均匀分布时，$\mathrm{Gini}(T)$ 最大，表示得到最小的有用信息。而袋外数据是不能从初始数据集中抽取出来的样本组成的集合。使用袋外数据来估计随机森林算法的泛化能力，称之为袋外数据估计[11]。

本研究主要选取 Gini 指数作为特征重要性排序的依据，排序结果如表 2 所示。

Table 2. Ranking results using the random forest method
表 2. 应用随机森林法的排序结果

排序	变量代码	第一次	第二次	第三次	平均值
1	ZX660	0.1977	0.2015	0.1819	0.1937
2	ZX588	0.0392	0.0554	0.0591	0.051233333
3	ZX477	0.0308	0.0446	0.0443	0.0399
4	ZX532	0.0355	0.0352	0.0349	0.0352
5	ZX57	0.0362	0.0322	0.0362	0.034866667
6	ZX407	0.0352	0.0304	0.0294	0.031666667
7	ZX358	0.0284	0.0147	0.0278	0.023633333
8	ZX40	0.0245	0.0157	0.0201	0.0201
9	ZX411	0.0225	0.0174	0.0146	0.018166667
10	ZX674	0.0123	0.0196	0.0181	0.016666667
⋮	⋮	⋮	⋮	⋮	⋮

由表 2 可看出，在 729 个初始变量中，分子描述符 ZX660 的重要性最高，说明它与生物活性的变化存在显著相关关系；分子描述符 ZX588 的重要性排序为第二，说明与生物活性的变化存在明显相关的关系；分子描述符 ZX777、ZX532、ZX57 等其他变量和生物活性的变化之间也存在较为强烈的相关关系，因此其重要性排序也靠前。

3.3.3. 基于 LASSO-随机森林的特征选择方法

LASSO 方法(最小绝对值压缩与选择方法)是 Tibshirani R 在 1996 年提出来的一种有偏估计方法，其本质是通过添加约束条件对模型系数进行压缩，将没有影响或影响较小的自变量的回归系数自动压缩到零，这不仅在一定程度上消除多重共线性的影响，而且在对参数进行估计的同时也实现对变量的选择[12] [13]。

对于线性模型：

$$y=\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon,$$

常数项和回归系数的 LASSO 估计定义为：

$$\left(\hat{\alpha},\hat{\beta}\right)=\arg\min\left\{\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}\beta_jx_{ij}\right)^2\right\},s.t.\sum_{j=1}^{p}\left|\beta_j\right|\le s.$$

其中，$s\ge0$ 是调和参数。令 $\hat{\beta}_j^0$ 表示 β_j 的最小二乘估计，$s_0=\sum_{j=1}^{p}\left|\hat{\beta}_j^0\right|$，当 $s<s_0$ 时，模型中的回归系数的 LASSO 估计的绝对值就会小于其最小二乘估计的绝对值，随着 s 的进一步减少，某些系数的 LASSO 估计值会很小，甚至等于零，这些估计值为零的系数所对应的变量将会被删除，从而达到变量选择的目的。当 $s\ge s_0$ 时，LASSO 估计就是最小二乘估计，所有的变量均被选入模型中，将不再具有约束作用。

该方法能在一定程度上消除变量之间的多重共线性，实现更高效地收缩子集效果，从而降低模型复杂度。因此本文运用 LASSO 方法将 729 个原始变量进行初次筛选，剔除无关或者影响不太显著的部分变量，再采取随机森林算法对保留下来的变量进行重要性排序。

组合模型对变量进行筛选和排序的具体步骤如下：

1) 运用 LASSO 方法对标准化后的 729 个变量指标进行参数估计和变量选择，根据 Cp 准则确定最优解。LASSO 回归的参数估计结果如表 3 所示。从表 3 可以看到，该步骤最终选取了 ZX16、ZX25、ZX39、……、ZX728 共 105 个变量，剔除了 624 个与生物活性不相关或者相关性不显著的变量。

Table 3. LASSO regression results
表 3. LASSO 回归结果

序号	变量	LASSO 估计参数	LASSO 估计参数绝对值	序号	变量	LASSO 估计参数	LASSO 估计参数绝对值
1	ZX16	0.006861853	0.006861853	16	ZX104	−0.00267442	0.00267442
2	ZX25	0.060010197	0.060010197	17	ZX106	0.054790579	0.054790579
3	ZX39	0.062605863	0.062605863	18	ZX113	−0.014112875	0.014112875
4	ZX43	0.166901578	0.166901578	19	ZX116	−0.065351001	0.065351001
5	ZX56	0.005004393	0.005004393	20	ZX125	0.010672012	0.010672012
6	ZX57	−0.0254005	0.0254005	21	ZX152	0.086150767	0.086150767
7	ZX58	0.020201356	0.020201356	22	ZX164	0.000341475	0.000341475
8	ZX59	0.230539413	0.230539413	23	ZX167	−0.016579498	0.016579498
9	ZX60	0.140812393	0.140812393	24	ZX187	0.006816475	0.006816475
10	ZX62	−0.014188471	0.014188471	25	ZX194	0.00208978	0.00208978
11	ZX64	−0.012792686	0.012792686	26	ZX236	0.07417386	0.07417386
12	ZX69	−3.2466E−15	3.2466E−15	27	ZX238	−0.001805145	0.001805145
13	ZX71	0.093811891	0.093811891	28	ZX244	0.049941695	0.049941695
14	ZX76	−0.015277208	0.015277208	⋮	⋮	⋮	⋮
15	ZX81	0.00848806	0.00848806	105	728	0.045731374	0.045731374

2) 运用随机森林算法对上一步骤保留的变量进行二次筛选和排序，同样选用 Gini 指数作为特征重要性排序的依据，前 20 个对生物活性最具有显著影响的分子描述符及其排序如表 4 所示。

Table 4. LASSO-random forest variable ranking results
表 4. LASSO-随机森林变量排序结果

排序	分子描述符变量	第一次 LASSO-随机森林	第二次 LASSO-随机森林	第三次 LASSO-随机森林	平均值
1	ZX407	0.2299	0.1983	0.227	0.2184
2	ZX477	0.0934	0.0873	0.0805	0.087066667
3	ZX674	0.0531	0.0589	0.0691	0.060366667
4	ZX43	0.0534	0.0523	0.0574	0.054366667
5	ZX413	0.0287	0.0471	0.029	0.034933333
6	ZX640	0.0332	0.0412	0.0294	0.0346
7	ZX106	0.0289	0.0272	0.0272	0.027766667
8	ZX104	0.0251	0.0251	0.0279	0.026033333
9	ZX586	0.0266	0.0206	0.0269	0.0247
10	ZX357	0.0208	0.0202	0.0229	0.0213
11	ZX347	0.0188	0.0211	0.0201	0.02
12	ZX412	0.02	0.0166	0.0209	0.019166667
13	ZX59	0.0173	0.0211	0.0136	0.017333333
14	ZX728	0.0162	0.0178	0.0163	0.016766667
15	ZX666	0.0163	0.0167	0.0161	0.016366667
16	ZX352	0.0162	0.01	0.0179	0.0147
17	ZX25	0.0143	0.0147	0.0138	0.014266667
18	ZX238	0.0141	0.0127	0.013	0.013266667
19	ZX587	0.0136	0.0121	0.0119	0.012533333
20	ZX270	0.0118	0.014	0.0116	0.012466667

由表 4 可看出，在 105 个 LASSO 方法选择的变量中，分子描述符 ZX407 的重要性最高，说明它与生物活性的变化存在显著相关关系；分子描述符 ZX477 的重要性排序为第二，说明它与生物活性的变化存在明显相关的关系；分子描述符 ZX674、ZX43、ZX413 等其他 18 个变量和生物活性的变化之间也存在较为强烈的相关关系，因此其重要性排序也靠前。

3.3.4. 特征提取结果的对比与分析

基于单一的随机森林算法的特征排序结果和基于 LASSO-随机森林算法的组合模型对特征进行排序的结果如表 5 所示。

Table 5. Feature ranking of the two models
表 5. 两种模型的特征排序

随机森林方法			LASSO-随机森林方法		
排序	分子描述符变量	平均值	排序	分子描述符变量	平均值
1	ZX660	0.1937	1	ZX407	0.2184
2	ZX588	0.051233333	2	ZX477	0.087066667

续表

3	ZX477	0.0399	3	ZX674	0.060366667
4	ZX532	0.0352	4	ZX43	0.054366667
5	ZX57	0.034866667	5	ZX413	0.034933333
6	ZX407	0.031666667	6	ZX640	0.0346
7	ZX358	0.023633333	7	ZX106	0.027766667
8	ZX40	0.0201	8	ZX104	0.026033333
9	ZX411	0.018166667	9	ZX586	0.0247
10	ZX674	0.016666667	10	ZX357	0.0213
11	ZX12	0.013766667	11	ZX347	0.02
12	ZX640	0.013133333	12	ZX412	0.019166667
13	ZX352	0.012733333	13	ZX59	0.017333333
14	ZX653	0.012366667	14	ZX728	0.016766667
15	ZX80	0.011466667	15	ZX666	0.016366667
16	ZX24	0.0101	16	ZX352	0.0147
17	ZX666	0.0086	17	ZX25	0.014266667
18	ZX717	0.008366667	18	ZX238	0.013266667
19	ZX239	0.008033333	19	ZX587	0.012533333
20	ZX41	0.0074	20	ZX270	0.012466667

由表 5 可以看到，单一方法和组合方法选择的前 20 个对生物活性最具有显著影响的分子描述符均包含了 ZX477、ZX666、ZX352、ZX640 和 ZX407 共 5 个变量，其余 15 个变量则不同。因此，有必要对两种方法的排序结果进行比较。

本问题将两个方法选择的 20 个变量作为自变量，生物活性作为因变量，以 SPSS 作为实验平台构建多元线性回归模型，变量的显著性水平和模型的拟合程度如表 6 所示。

Table 6. Significance level and fitting degree of molecular descriptor variables
表 6. 分子描述符变量的显著性水平及拟合程度

序号	随机森林方法		LASSO-随机森林方法	
	分子描述符变量	显著性水平	分子描述符变量	显著性水平
1	ZX660	0.000	ZX407	0.000
2	ZX588	0.905	ZX477	0.000
3	ZX477	0.000	ZX674	0.000
4	ZX532	0.776	ZX43	0.000
5	ZX57	0.000	ZX413	0.001
6	ZX407	0.000	ZX640	0.000
7	ZX358	0.092	ZX106	0.000
8	ZX40	0.002	ZX104	0.000

续表

9	ZX411	0.000	ZX586	0.000
10	ZX674	0.000	ZX357	0.000
11	ZX12	0.000	ZX347	0.000
12	ZX640	0.000	ZX412	0.000
13	ZX352	0.000	ZX59	0.000
14	ZX653	0.000	ZX728	0.000
15	ZX80	0.000	ZX666	0.000
16	ZX24	0.123	ZX352	0.000
17	ZX666	0.000	ZX25	0.000
18	ZX717	0.038	ZX238	0.001
19	ZX239	0.000	ZX587	0.000
20	ZX41	0.002	ZX270	0.188
	R^2 拟合度：0.580		R^2 拟合度：0.601	

由表 6 可以看出，一方面，基于组合模型选择的 20 个分子描述符和生物活性构建的多元线性回归模型中，各变量均在 1%的显著性水平下显著，而基于单一模型选择的分子描述符构建的多元回归模型中，存在 3 个变量不显著；另一方面，基于组合模型构建的回归模型拟合度为 0.580，而基于单一模型构建的回归模型拟合度为 0.601。

综上，可以说明，结合 LASSO 方法和随机森林算法的组合特征选择模型能够弥补两个方法的缺点，得到的变量重要性排序结果相对于单一模型更有效且合理。

最终得到由 Lasso-随机森林组合筛选的分子描述符及其排序为 ZX407、ZX477、ZX674、ZX43、ZX413、ZX640、ZX106、ZX104、ZX586、ZX357、ZX347、ZX412、ZX59、ZX728、ZX666、ZX352、ZX25、ZX238、ZX587、ZX270。

4. 问题二的模型建立与求解模型假设

4.1. 问题分析

针对问题二，要求基于问题一选择的 20 个变量中再一次筛选主要建模变量，由于个别变量存在高度非线性且多个变量之间涉及多重共线性问题，仅使用传统的线性降维方法筛选变量是不可靠的。

因此，本问题首先对原始数据集进行训练集和检验集的划分；其次，通过逐步回归法、自适应 LASSO 回归法基于训练集分别建立变量选择模型，综合比较 2 种方法输出的变量选择结果，确定最终筛选的变量；然后，将综合筛选得到的变量作为 BP 神经网络预测模型的输入层变量，基于训练集数据对模型进行训练后，把已知的测试集样本代入模型中，进一步验证所构建模型的有效性；最后，对未知的 50 个化合物进行生物活性 pIC_{50} 以及相应的 IC_{50} 的预测。在此说明，本研究选用 pIC_{50} 作为相应问题的建模对象。给出该问题的思路如图 2 所示。

4.2. 数据集划分

将原始的 1974 个化合物数据集按照 8:2 的比例划分为训练集和测试集，其中基于训练集数据来训练模型，在测试集上根据所给的性能评价指标来评估所构建模型的预测效果。最后，将待预测的 50 个化合物数据集作为预测集代入所构建的模型中，得到预测结果。

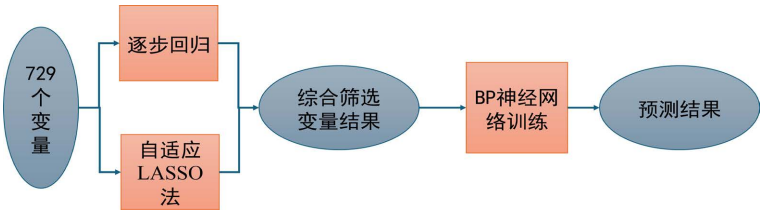

Figure 2. Flow chart of question 2
图 2. 问题二流程图

4.3. 变量筛选

在对模型的变量进行筛选和回归预测的文献中，逐步回归方法得到了较广泛的应用。然而，逐步回归法极易陷入局部最优解，且无法彻底解决变量之间严重多重共线性的问题。

因此，针对问题二，本研究使用逐步回归法和自适应 LASSO 法对变量进行综合选择。

4.3.1. 逐步回归法

逐步回归法将自变量逐个引入模型中，依据引入后的偏回归平方和以及参数的显著性水平判断该变量是否需要剔除，该步骤不断进行直到没有新变量引入也无旧变量剔除为止[14]。

以 SPSS 为实验平台，通过不断地对变量进行引入和剔除后，得到基于逐步回归法所确定的最终模型的变量个数为 16 个时，模型回归系数 R^2 为 0.637，调整后的 R^2 为 0.633，并在 1%的显著性水平下显著。变量筛选结果如表 7 所示。

Table 7. Variables and coefficients determined by stepwise regression method
表 7. 逐步回归法所确定的变量及系数

变量	X407	X477	X43	X412	X59	X674	X357	X728
系数	0.198	0.857	0.210	−0.012	0.164	0.476	0.056	0.070
变量	X347	X25	X640	X106	X413	X586	X352	X666
系数	−0.054	1.137	−0.146	0.000	−0.063	−0.896	0.055	−0.172

4.3.2. 自适应 LASSO 法

由于 LASSO 方法对重要变量的参数估计是有偏的，Zou 提出了具有 Oracle 性质的自适应 LASSO 方法。定义 $\hat{\beta}^*(n)=\arg\min_{\beta}\left\|y-\sum_{j=1}^{p}x_j\beta_j\right\|^2+\lambda_n\sum_{j=1}^{p}\hat{\omega}\left|\beta_j\right|$，其中权重 $\hat{\omega}_j=\dfrac{1}{\left|\hat{\beta}\right|^{\gamma}}$，$\gamma>0$，$j=1,2,\cdots,p$，$\hat{\beta}_j$ 为普通最小二乘法估计的系数[15]。

以 R 语言为实验平台进行自适应 LASSO 回归计算，得到基于自适应 LASSO 方法确定的最终模型的变量个数为 17 个，变量筛选和参数估计结果如表 8 所示。

Table 8. Variables and coefficients determined by adaptive LASSO method
表 8. 自适应 LASSO 法所确定的变量及系数

变量	X25	X43	X59	X106	X270	X347	X352	X357	X407
系数	1.1255	0.21559	0.15008	0.0004	0.00678	−0.04814	0.04826	0.05374	0.19228
变量	X412	X413	X477	X586	X640	X666	X674	X728	
系数	−0.01068	−0.06026	0.78915	−0.8184	−0.14345	−0.12927	0.51532	0.6714	

4.3.3. 综合两类方法的变量选择

综合逐步回归法和自适应 LASSO 法对变量选择的输出结果，确定 BP 神经网络的输入层变量，如表 9 所示。

Table 9. Variable selection of the two methods
表 9. 两种方法的变量选择

分子描述符变量	逐步回归法	自适应 LASSO 法	最终确定结果	分子描述符变量	逐步回归法	自适应 LASSO 法	最终确定结果
X25	√	√	√	X407	√	√	√
X43	√	√	√	X412	√	√	√
X59	√	√	√	X413	√	√	√
X104				X477	√	√	√
X106	√	√	√	X586	√	√	√
X238				X587			
X270		√	√	X640	√	√	√
X347	√	√	√	X666	√	√	√
X352	√	√	√	X674	√	√	√
X357	√	√	√	X728	√	√	√

由表 9 可以看出，分子描述符 X104、X238、X587 这 3 个变量均仅有 1 个模型对其进行保留，除此之外的 17 个变量都至少有 2 个及以上变量对其进行保留。因此，综合考虑上述 3 种方法的变量选择结果，可以确定最终变量为 X25、X43、X59、X106、X270、X347、X352、X357、X407、X412、X413、X477、X586、X640、X666、X674 以及 X728 共 17 个。

4.4. 基于 BP 神经网络的生物活性值预测模型

由于分子描述符与生物活性值之间呈现出较强的非线性关系，因此，本研究采用 BP 神经网络对其进行建模分析[16]。

4.4.1. 模型构建

1) 训练集、测试集和预测集。本研究将 1580 个化合物的生物活性值及其筛选出的 17 个分子描述符数据作为训练集，对模型进行训练；将 394 个化合物的生物活性值及其筛选出的 17 个分子描述符数据作为测试集，将其代入训练好的 BP 神经网络模型中确定所构建模型的有效性；将待预测的 50 个化合物数据作为预测集，输入到模型中进而得到预测结果。

2) 输入层和输出层节点数确定。由于在隐含层节点数足够多的情况下，3 层 BP 神经网络模型能够逼近任意非线性函数。因此，本文选用仅含有 1 个隐含层的 3 层 BP 神经网络模型。根据前文的变量筛选结果，网络输入层节点数设置为 17；输出层变量为该化合物的生物活性值 pIC_{50}，网络输出层节点数设置为 1。

3) 隐含层节点数确定。隐含层节点数的设定会对 BP 神经网络的预测性能产生影响，因此本研究根据经验公式 $a=\sqrt{b+c}+d$ 和 $\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_{test}^{(i)}-\hat{y}_{test}^{(i)}\right)^2$ 来进一步确定，其中 a 为隐含层节点数，b 为输入层

节点数，c 为输出层节点数，d 为 1~10 之间的常数。由于本研究的输入层节点数 b 为 17，输出层节点数 c 为 1，因此 a 的取值范围为 5~15。本研究以 MATLAB2016a 为实验平台，调用人工神经网络工具箱，设置不同的隐含层节点数，基于训练集数据分别进行 3 次训练。得到的结果如表 10 所示。

Table 10. Three training results of training set data
表 10. 训练集数据三次训练结果

隐含层节点数	第一次 MSE	第二次 MSE	第三次 MSE	平均 MSE
5	0.5484	0.5591	0.6022	0.5699
6	0.6168	0.5704	0.5389	0.575366667
7	0.5228	0.4919	0.5138	0.5095
8	0.5355	0.4872	0.4988	0.507166667
9	0.5104	0.5245	0.5629	0.5326
10	0.4823	0.4764	0.5475	0.502066667
11	0.4799	0.6683	0.5292	0.559133333
12	0.613	0.4237	0.4776	0.504766667
13	0.4868	0.5743	0.5523	0.5378
14	0.501	0.5492	0.5824	0.5442
15	0.5562	0.5036	0.5189	0.526233333

由表 10 可以看出，当隐含层节点数为 10 时，对同一数据集所构建的模型其平均 MSE 最小。因此，确定隐含层节点个数为 10。

4.4.2. 模型训练及仿真

本实验以 MATLAB2016a 为实验平台，调用人工神经网络工具箱，选择 Levenberg-Marquardt 算法为训练函数，对选定的样本数据进行学习训练后对 394 个化合物数据组成的测试集进行预测。部分预测结果如表 11 所示。

Table 11. Forecast results
表 11. 预测结果

SMILES	pIC50	检验结果	SMILES	pIC50	检验结果
测试样本 1	5.430	7.155	测试样本 8	7.155	7.616
测试样本 2	5.930	6.459	测试样本 9	6.292	8.286
测试样本 3	5.509	4.605	测试样本 10	6.678	7.414
测试样本 4	4.609	6.430	测试样本 11	5.593	8.000
测试样本 5	6.921	6.388	⋮	⋮	⋮
测试样本 6	4.796	4.514	测试样本 393	7.886	6.428
测试样本 7	6.469	7.448	测试样本 394	7.569	7.513

由表 11 可得，BP 神经网络的表现较好。进一步分析后发现，拟合优度 R^2 为 0.825，说明本实验采用 BP 神经网络对化合物的生物活性值进行预测是可行的。

5. 结论

本研究围绕乳腺癌候选药物涉及的两个问题进行建模并得出有效结论。针对问题一，对 1974 个化合物的 729 个分子描述符进行标准化处理以消除量纲影响，构建了两个模型：模型一用随机森林算法直接筛选排序特征，模型二则先通过 LASSO 方法从原始变量中选出 105 个重要变量，再用随机森林算法对保留变量进行重要性排序。对比发现两模型筛选的主要变量有 5 个相同，组合模型筛选的变量通过显著性检验且拟合优度高，故选定 LASSO-随机森林算法组合模型的变量筛选排序结果作为最终输出。针对问题二，将数据集按 8:2 划分为训练集和测试集，基于问题一结果，利用逐步回归法和自适应 LASSO 法在训练集上二次筛选，按投票原则确定 17 个变量作为最终筛选结果及后续神经网络输入层变量；鉴于因变量和自变量间的非线性关系，采用 BP 神经网络建模，将测试集样本代入训练好的模型，拟合效果良好，证明了模型的有效性。

参考文献

[1] 鲍娇玉, 刘存, 陈文君, 等. 1990-2021 年中国女性乳腺癌疾病负担趋势及危险因素分析[J]. 肿瘤预防与治疗, 2025, 38(2): 98-106.

[2] Chae, E.Y., Jun, M.R., Cha, J.H., 等. 早期手术治疗的年轻女性病人乳腺癌复发的 MRI 和临床病理特征预测模型[J]. 国际医学放射学杂志, 2025, 48(1): 116-117.

[3] 曹明, 徐曼, 杨小苗, 等. 不同生命周期女性乳腺癌临床病理特征研究[J]. 分子诊断与治疗杂志, 2024, 16(11): 2086-2089.

[4] 毛煦. CYP1 介导的雌二醇生物活化对女性乳腺癌发展影响的研究进展[J]. 中国药理学与毒理学杂志, 2023, 37(8): 640-644.

[5] 吴琪. 全新雌激素受体亚型 ER-α25 在乳腺癌细胞中的作用及机制研究[D]: [硕士学位论文]. 桂林: 桂林医学院, 2023.

[6] 沈家成, 姜筝, 孙盛鑫, 等. 吲哚类 CB2 配体 3D-QSAR 模型的构建与比较[J]. 广州化工, 2024, 52(23): 109-113.

[7] 崔春利, 闫浩晨, 王敏, 等. 基于衰老视角的 3 种常见慢病共有机制与中药发现[J]. 西安交通大学学报(医学版), 2025, 46(1): 101-111.

[8] 孙娜, 王谕靖, 周彩红, 等. 基于综合网络药理策略研究夏枯草降压作用机制[J]. 中国药学, 2024, 33(11): 1068-1081.

[9] 曾婷婷. 基于机器学习的房价预测模型研究[D]: [硕士学位论文]. 廊坊: 西南科技大学, 2020.

[10] 刘云翔, 陈斌, 周子宜. 一种基于随机森林的改进特征筛选算法[J]. 现代电子技术, 2019, 42(12): 117-121.

[11] 吴孝情, 赖成光, 陈晓宏, 任秀文. 基于随机森林权重的滑坡危险性评价: 以东江流域为例[J]. 自然灾害学报, 2017, 26(5): 119-129.

[12] 曹正凤. 随机森林算法优化研究[D]: [博士学位论文]. 北京: 首都经济贸易大学, 2014.

[13] 马臣, 邹进美, 夏孟红. 运用 Lasso 回归模型预测重庆南川地区呼吸道感染的研究[J]. 检验医学与临床, 2025, 22(4): 480-484.

[14] 焦伟, 陈亚宁, 李稚, 李玉朋, 黄晓然, 李海霞. 基于多种回归分析方法的西北干旱区植被 NPP 遥感反演研究[J]. 资源科学, 2017, 39(3): 545-556.

[15] 郭文军. 中国区域碳排放权价格影响因素的研究——基于自适应 Lasso 方法[J]. 中国人口·资源与环境, 2015, 25(S1): 305-310.

[16] 张青, 王学雷, 张婷, 杨超, 吕晓蓉. 基于 BP 神经网络的洪湖水质指标预测研究[J]. 湿地科学, 2016, 14(2): 212-218.