Operations Research and Fuzziology 运筹与模糊学, 2022, 12(1), 68-80
Published Online February 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.121007

Hans汉斯

基于机器学习的抗乳腺癌候选药物的优化

李荟霖，李明英

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2022年1月5日；录用日期：2022年1月30日；发布日期：2022年2月9日

摘　要

能够拮抗ERα活性的化合物可能是治疗乳腺癌的候选药物，研究这类化合物对乳腺癌的攻克具有重要意义。本文提出了对治疗乳腺癌的候选药物的实验数据进行数据预处理、特征选择、模型预测的一系列方法。目的：获得具有更好生物活性的新化合物分子。基于k-means聚类与安德鲁斯曲线的异常样本检测模型对异常样本进行剔除；对样本中729个分子描述符进行筛选，保留20个对生物活性最具有显著影响的分子描述符。使用基于三类特征筛选方法的五种方法，基于此建立了多维特征加权提取模型。构建化合物对ERα生物活性的QSAR模型。以PIC_{50}为因变量，对筛选出的20个分子描述符作为自变量，建立了XGBoost，LightGBM机器学习模型，利用网格搜索法获取模型最优参数，保留更有效的模型预测结果。

关键词

特征提取，QSAR模型，机器学习

Optimization of Anti-Breast Cancer Drug Candidates Based on Machine Learning

Huilin Li, Mingying Li

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Jan. 5th, 2022; accepted: Jan. 30th, 2022; published: Feb. 9th, 2022

Abstract

Compounds that can antagonize the activity of ERα may be candidate drugs for the treatment of breast cancer, and it is of great significance to study such compounds in the fight against breast cancer. This paper proposes a series of methods for data preprocessing, feature selection, and model prediction on experimental data of candidate drugs for the treatment of breast cancer. Objective: To obtain new compound molecules with better biological activity. The abnormal sample detection model based on k-means clustering and Andrews curve eliminated abnormal samples;

文章引用: 李荟霖, 李明英. 基于机器学习的抗乳腺癌候选药物的优化[J]. 运筹与模糊学, 2022, 12(1): 68-80.
DOI: 10.12677/orf.2022.121007

729 molecular descriptors in the sample were screened, and 20 molecular descriptors with the most significant impact on biological activity were retained. Using five methods based on three types of feature screening methods, a multi-dimensional feature weighted extraction model was established based on this. Construct a quantitative prediction model of the compound's biological activity on ERa. Using PIC_{50} as the dependent variable and the 20 molecular descriptors selected as independent variables, the XGBoost and LightGBM machine learning models were established, and the grid search method was used to obtain the optimal parameters of the model to retain more effective model prediction results.

Keywords

Feature Selection, QSAR Model, Machine Learning

1. 研究背景

根据国际癌症研究机构(IARC)调查的最新数据显示，乳腺癌在全球女性癌症中的发病率为 24.2%，位居女性癌症的首位，其中 52.9%发生在发展中国家[1]。乳腺癌是目前世界上最常见，致死率较高的癌症之一。乳腺癌的药物研发具有非常重要的意义。乳腺癌的发展与雌激素受体密切相关，有研究发现，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 确实在乳腺发育过程中扮演了十分重要的角色[2]。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。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[3]。

目前，在药物研发中，为了节约时间和成本，通常采用构建化合物的定量结构－活性关系(Quantitative Structure-Activity Relationship, QSAR)模型的方法来筛选潜在活性化合物[4]。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[5]。

2. 数据预处理

考虑到原始数据可能会因各种原因存在数据缺失、数据异常等问题，分析之前有必要对所获得的样本数据进行预处理。因此我们对 1974 个样本以及 729 个分子描述符(即自变量)的数据进行数据预处理。本文对数据的预处理包含：查找数据中是否有缺失值；检测异常样本并对异常样本进行剔除。

根据所述思路和方法，我们发现题目所提供的数据均是完好的且在依拉达准则下没有异常值出现。根据 K-means 聚类以及安德鲁斯曲线检测到一个异常样本。

基于 K-means 聚类和安德鲁斯曲线的异常样本检测模型

首先在医学领域对样本的异常检测是十分必要的。如果存在异常样本，分析该问题得出的结论很有可能与真实情况大相径庭，那么对问题的分析也就是徒劳的。同时对异常样本进行检测有利于提高模型预测结果的稳健性。K-means 聚类是我们最常用的基于欧式距离的聚类算法，其认为两个目标的距离越近，相似度越大。它的优点是容易理解，可以达到局部最优，算法复杂度低。一个很显著的缺点是 K 值

是人为确定的。K-means 具体的算法步骤如下：

1) 选择初始化的 k 个样本作为初始聚类中心 $a=a_1,a_2,\cdots,a_k$；

2) 针对数据集中每个样本 x_1 计算它到 k 个聚类中心的距离,并将其分到距离最小的聚类中心所对应的类中；

3) 针对每个类别 a_1，重新计算它的聚类中心 $a_j=\frac{1}{|c_j|}\sum_{x\in c_j}x$ (即属于该类的所有样本的质心)；

4) 重复上面两步操作，直到达到某个中止条件(迭代次数、最小误差变化等) [6]。

对样本进行有效的异常检测的具体步骤如下：

1) K 值的确定

采用肘部法则和轮廓系数来决定 K-means 聚类簇数。

肘部法则(Elbow Method)原理：对于一个簇，它的畸变程度越低，代表簇内成员越紧密，畸变程度越高，代表簇内结构越松散。畸变程度会随着类别的增加而降低，但对于有一定区分度的数据，在达到某个临界点时畸变程度会得到极大改善，之后缓慢下降，这个临界点就可以考虑为聚类性能较好的点。

如图 1 所示，在 $k=4$ 时，畸变程度(y 值)得到大幅改善，可以考虑选取 $k=4$ 作为聚类数量。

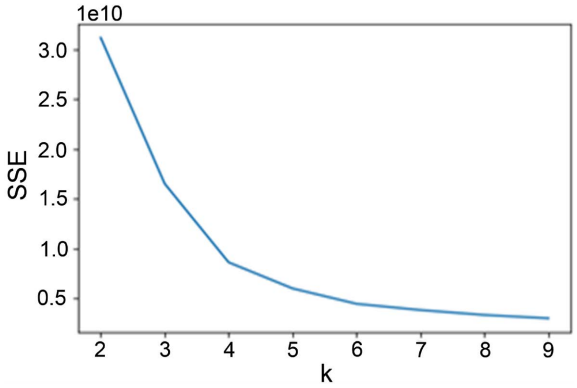

Figure 1. Diagram of the elbow rule
图 1. 肘部法则图

轮廓系数(Silhouette Coefficient)原理：它结合内聚度和分离度两种因素，是聚类效果好坏的一种评价方式[2]。聚类结果的轮廓系数的取值在(−1, 1)之间，值越大，说明同类样本相距越近，不同样本相距越远，则聚类效果越好。负值通常表示样本已分配给错误的聚类，因为不同的聚类更为相似。

对于簇中的每个向量，分别计算它们的轮廓系数。第 i 个对象的轮廓系数的计算公式为：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{1}$$

其中 $a(i)$ 为簇内不相似度，表示 i 向量到同簇内其他点不相似程度的平均值，体现内聚度；$b(i)$ 为簇间不相似度，表示 i 向量到其他簇的平均不相似程度的最小值，体现分离度。所有样本的 $s(i)$ 的均值称为聚类结果的轮廓系数，定义为 S，是该聚类是否合理、有效的度量。

如图 2 所示，$K=2, 4$ 时轮廓系数都是取比较大的值，综合肘部法则的结果，我们最终选择 k 值为 4。

2) 对样本进行 K-means 分类。

对 1974 个化合物的分类结果如表 1 所示：第 0 类 1040 个化合物，第 1 类 1 个化合物，第 2 类 8 个化合物；第 3 类 925 个。

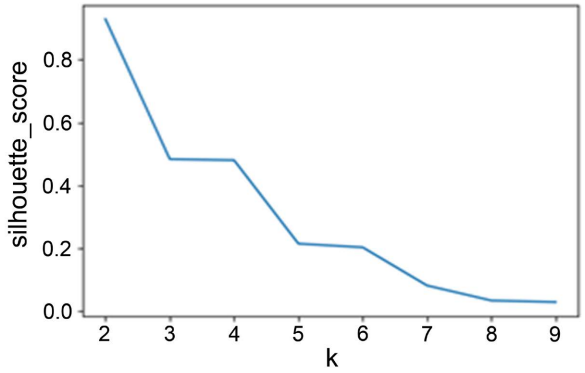

Figure 2. Contour coefficient diagram
图 2. 轮廓系数图

Table 1. K-means classification results table
表 1. K-means 分类结果表

类别	分类结果
0	1040
1	1
2	8
3	925

3) 采用安德鲁斯曲线对分类结果进行可视化处理。安德鲁斯曲线可用于可视化高维数据，起到聚类作用；异常样本检测，同一类别的曲线基本一致，若有不一致曲线则为异常记录。其公式为：

$$f(t)=\frac{x_1}{sqrt(2)}+x_2\sin(t)+x_3\cos(t)+x_4\sin(2t)+x_5\cos(2t)+\cdots \tag{2}$$

结果如图 3 所示，发现第 1 类样本与第 0 类和第 2、3 类相差较大，第 1 类化合物为 1562 号样本，该化合物各分子描述符与其他化合物相比差异明显，因此将其定义为异常样本并剔除。

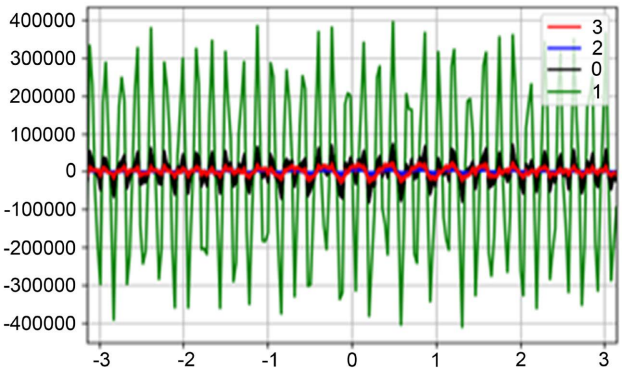

Figure 3. Andrews curve
图 3. 安德鲁斯曲线

3. 特征选择

3.1. 问题分析

需要从 729 个分子描述符里面通过一系列方法选出 20 个对化合物生物活性值有显著影响的分子描述

符，以降低数据集维度方便接下来问题的分析讨论。通常来说，从两个方面考虑来选择特征：

1) 变量自身的离散程度；

2) 变量同因变量之间的相关性。

可以将变量筛选选择方法大致分为三类：Filter (过滤式)、Wrapper (包裹式)、Embedding (嵌入式) [7]。

过滤式方法的通用性强，省去了分类器的训练步骤，算法复杂性低，因而适用于大规模数据集，可以快速去除大量不相关的变量。缺点是由于算法的评价标准独立于特定的学习算法，所选的变量子集在分类准确率方面通常低于其他两种方法。包裹式和嵌入式特征选择方法则考虑到了不同变量组合产生的效果来评价变量的价值，但是时间开销往往更大[8]。

鉴于每种方法都有其考量的重点以及合理性，为综合考虑自变量自身的发散程度，各自变量与因变量之间的相关性，不同自变量组合同因变量之间的相关性等问题，本文决定分别在三类变量筛选方法下各选取 1~2 种方法分别对 729 个自变量的重要性进行量化排序。本文决定采用方差过滤法、互信息法、递归特征消除法、基于带 L1 惩罚项的贝叶斯岭回归基模型的特征选择法、随机森林法对 1973 个化合物的 729 个分子描述符进行重要性排序，分别得到五组变量重要性序列。先分别采用每一组序列的前 30 个分子描述符作为自变量，以抑制分子活性程度值 PIC_{50} 作为模型标签，采用 LightGBM 集成学习算法进行训练。根据训练结果来给予每种方法一个权重，最后用每种方法的加权平均得分重新筛选变量，筛选出 30 个变量。最后在此基础上按照优先删除高相关、低权重分子描述符的原则进行变量的二次提取，筛选出最终的 20 个变量。问题一的思维导图如图 4 所示：

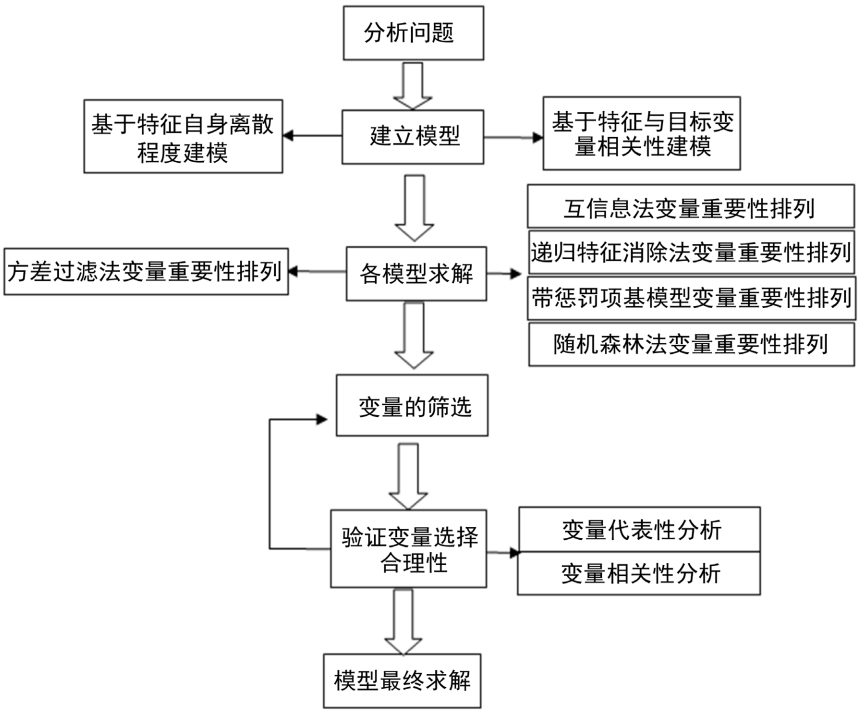

Figure 4. Mind map of feature selection
图 4. 特征选择的思维导图

3.2. 建立模型

针对变量自身的离散程度的评价，我们采用了方差过滤法。观察样本数据我们可以看到有些变量的取值基本都为 0，或者取值相差不大，那么该变量的变化对于因变量的影响就会很小，也可以粗略地认为该

种变量的价值不高。本文通过对各自变量的值求方差，根据所求方差大小降序排列从而得到一个变量重要性排序。为了解决有些变量本身取值较大而导致的方差较大的问题，将所得到的 729 个自变量方差进行归一化处理，取值在(0, 1)之间，再对变量进行重要性排序。方差的计算以及数据归一化处理公式如下：

方差：

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - x\right)^2}{n} \tag{3}$$

归一化：

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

针对特征与目标相关性，我们采用两种包裹式和两种嵌入式方法来评价特征。两种包裹式特征评价方法：互信息法和递归特征消除法。互信息法是用来评价一个事件的出现对于另一个事件的出现所贡献的信息量，皮尔逊系数只能衡量线性相关性而互信息系数能够很好地度量各种相关性，但是计算相对复杂一些，得到相关性之后就可以排序选择特征了。具体的计算公式为：

$$I\left(U;C\right) = \sum_{e_t \in (1,0)} \sum_{e_c \in (1,0)} P\left(U = e_t, C = e_c\right) \log_2 \frac{P\left(U = e_t, C = e_c\right)}{P\left(U = e_t\right) P\left(U = e_c\right)} \tag{5}$$

其中 U、C 代表两个事件，e 的取值可以为 0 或者 1，1 代表出现这个事件，0 代表不出现。递归特征消除的主要思想是反复的构建模型(如 SVM 或者回归模型)然后选出最好的(或者最差的)的特征(可以根据系数来选)，把选出来的特征选择出来，然后在剩余的特征上重复这个过程，直到所有特征都遍历了。这个过程中特征被消除的次序就是特征的排序。因此，这是一种寻找最优特征子集的贪心算法。

两种包裹式特征评价方法：基于带 L1 惩罚项的贝叶斯岭回归基模型的特征选择法、随机森林法。L1 惩罚项降维的原理在于保留多个对目标值具有同等相关性的特征中的一个。随机森林法算法自带特征选择功能，它可以评估每个特征在相应问题上的重要性[9]。

应用以上五种方法进行变量筛选，根据重要性降序排列，得分越高说明变量的价值越大，因此每种方法选取排列后的前 30 个变量，共筛选出五组自变量。下图只截取了五种方法的前 20 个变量的重要性排序，结果如图 5 所示：

	方差过滤法		互信息法		递归特征		带惩罚项		随机森林	
排名	var	im	var	im	var	im	var	im	var	im
1	WPATH	1.000	SsOH	1.000	rippenLog	1.000	A_Epsilon	1.000	MDEC-23	1.000
2	fragC	0.533	BCUTc-11	0.976	nHother	1.000	SsI	0.787	maxHsOH	0.176
3	ATSp5	0.279	maxHsOH	0.933	nssCH2	1.000	minsI	0.770	maxssO	0.153
4	ATSp4	0.260	BCUTc-1h	0.926	nsssCH	1.000	A_EtaP_B	0.768	minsssN	0.116
5	ATSp3	0.201	SHsOH	0.905	ndssC	1.000	VCH-5	0.754	minsOH	0.103
6	ATSp2	0.092	minHsOH	0.872	naasC	1.000	nG12Ring	0.743	C1SP2	0.099
7	ATSp1	0.063	WTPT-3	0.837	naaaC	1.000	WTPT-2	0.741	BCUTc-11	0.068
8	ECCEN	0.025	maxsOH	0.833	nssS	1.000	VCH-4	0.739	MLogP	0.058
9	VABC	0.002	minHBa	0.806	naaS	1.000	MDEN-13	0.738	ATSc3	0.058
10	MW	0.002	WTPT-5	0.789	SaaN	1.000	n12Ring	0.738	MLFER_A	0.054
11	Zagreb	0.000	maxssO	0.787	minsOH	1.000	SCH-4	0.737	VC-5	0.045
12	TopoPSA	0.000	SssO	0.785	minssO	1.000	maxHBd	0.735	minHsOH	0.045
13	AMR	0.000	minsOH	0.770	minaaO	1.000	TA_Shape_	0.734	ndssC	0.043
14	CrippenMR	0.000	MDEC-23	0.757	minsOm	1.000	A_dEpsilor	0.733	Acc_Lipin	0.039
15	ETA_Eta_R	0.000	minwHBa	0.751	minsF	1.000	minHBd	0.733	affinityI	0.036
16	SHBa	0.000	SaaCH	0.732	maxwHBa	1.000	maxsssCH	0.733	nHBAcc	0.035
17	apol	0.000	maxHBa	0.728	maxHBint4	1.000	A_dEpsilor	0.732	SHsOH	0.031
18	nBonds2	0.000	minaasC	0.726	maxHBint8	1.000	ATSc4	0.732	MDEO-12	0.029
19	sumI	0.000	MLFER_A	0.723	maxHBint9	1.000	ndssS	0.732	minHBint5	0.026
20	ATSm5	0.000	SHBd	0.714	maxHBint1(	1.000	maxdS	0.732	XLogP	0.022

Figure 5. The importance of the five groups of variables corresponding to the five methods
图 5. 五种方法对应的五组变量重要性排序

从图 5 中我们可以看出，每种方法筛选出来的变量差异较大。

3.3. 模型求解

分别采用方差过滤法、互信息法、递归特征消除法、基于带 L1 惩罚项的贝叶斯岭回归基模型的特征选择法、随机森林法得到的每一组序列的前 20 个分子描述符作为模型特征，以抑制分子活性程度值 PIC_{50} 作为模型标签，采用 LightGBM 集成学习算法进行训练，将 1973 个化合物分为 75%训练集，25%测试集，并利用网格搜索法进行参数调优。分别得到了五组特征在测试集上得到的 MSE 结果如表 2 所示，保留两位小数的结果分别是 1.05，0.79，0.87，1.05，0.78。MSE (Mean Squared Error)计算公式如下：

$$\mathrm{MSE}=\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\widehat{y}_i\right)^2 \tag{6}$$

m 表示样本量，$\widehat{y}_i$ 表示预测值，y_i 表示真实值。MSE 的值越大说明模型的预测效果越差，反之越好。

Table 2. Comparison of model fitting accuracy of five groups of variables
表 2. 五组变量的模型拟合精度对比

方法	MSE
方差过滤法	1.06
互信息法	0.79
递归特征消除法	0.87
带 L1 惩罚项	1.05
随机森林法	0.78

根据表 2 所示的模型预测的 MSE 值分别给予五组重要性排序不同的权重，MSE 值越大的那种方法被赋予的权重越小。由于对各组重要性得分进行归一化处理后，各个变量的分数差异非常小，为了能够使差距更明显，本文赋予的每种方法的权重相对较大，方差过滤法得到的重要性程度权重为 2，互信息法、递归特征消除法、基于带 L1 惩罚项的贝叶斯岭回归基模型的特征选择法、随机森林法重要性程度权重依次为 4、3、1、5，并将五组重要性加权求和得到最终 729 个分子描述符的重要性程度以及排序，在这里只截取了前 12 个重要变量。结果如表 3 所示：

Table 3. Weighted scores of five groups of variables
表 3. 五组变量的加权得分

	var	$\omega_1\cdot s_1$	$\omega_2\cdot s_2$	$\omega_3\cdot s_3$	$\omega_4\cdot s_4$	$\omega_5\cdot s_5$	Score
1	MDEC-23	2.78E−05	3.028	2.6	0.721	5	11.35
2	maxssO	2.76E−06	3.147	2.8	0.721	0.767	7.44
3	minsOH	4.89E−06	3.080	3	0.722	0.516	7.32
4	fragC	1.066	2.307	2.8	0.721	0.015	6.91
5	WPATH	2	2.420	1.4	0.721	0	6.54
6	SssO	1.04E−05	3.141	2.6	0.721	0	6.46
7	maxHBint10	2.36E−06	2.702	3	0.722	0.019	6.44

Continued

8	minssO	2.54E−06	2.685	3	0.721	0.006	6.41
9	SHBint10	1.12E−05	2.775	2.8	0.721	0.089	6.38
10	maxHBa	1.45E−06	2.916	2.4	0.721	0	6.03
11	gmax	1.18E−06	2.817	2.4	0.721	0.005	5.94
12	mindssC	9.9E−08	2.760	2.4	0.723	0.036	5.92

多维特征加权提取模型计算公式为：

$$\text{Score} = \omega_1 \cdot s_1 + \omega_2 \cdot s_2 + \omega_3 \cdot s_3 + \omega_4 \cdot s_4 + \omega_5 \cdot s_5 \tag{7}$$

其中 $\omega_i \cdot s_i, i = 1,2,3,4,5$ 表示每种方法变量重要性得分乘上每种方法的权重 $\omega_i, i = 1,2,3,4,5$ 表示每种方法的权重，依次是方差过滤法、互信息法、递归特征消除法、基于带 L1 惩罚项的贝叶斯岭回归基模型的特征选择法、随机森林法被赋予的权重，$s_i, i = 1,2,3,4,5$ 表示的是变量依次在五种方法上的重要性得分。Score 为每种变量最后的加权得分，本文依据这个分数进行降序排列，筛选出排在前面的 20 个变量。

3.4. 验证变量筛选的合理性

对通过多维特征加权提取模型初步筛选出的前 20 个变量的合理性从两个方面进行验证。

1) 变量的代表性分析：以初步筛选出的 20 个分子描述符作为模型特征，以抑制分子活性程度值 PIC_{50} 作为模型标签带入 LightGBM 集成学习算法进行训练，并采用网格搜索交叉验证进行参数调优，得到最终 MSE 结果为 0.6122，与其他变量组的模型预测结果相比该组 MSE 值有显著降低且较为稳定。预测结果对比如图 6 所示：

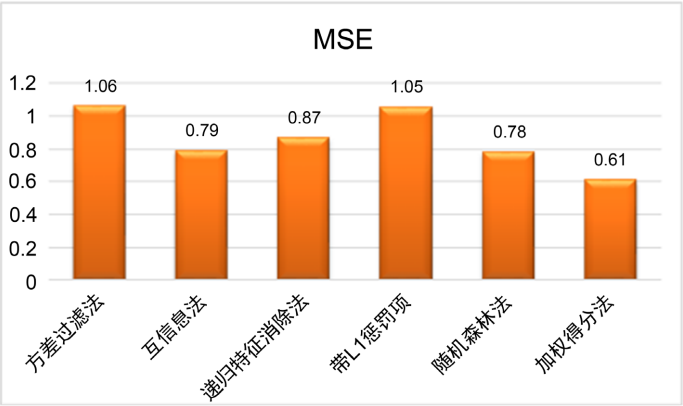

Figure 6. Comparison of model fitting effects of six groups of variables
图 6. 六组变量的模型拟合效果对比图

从图 6 中我们可以看到通过用加权得分法筛选出来的变量带入 LightGBM 集成学习算法进行训练，训练好的模型在测试集上的预测效果更优了，也就说明通过该种方法筛选出来的变量更具代表性。

2) 变量的相关性分析：

利用皮尔森相关系数得到了一个相关关系数热力图来检验用多维特征加权提取模型筛选出的 20 个变量之间是否具有相关性。随机选取了这初步筛选出的 20 个变量中的十个进行相关关系的分析。发现还存在变量之间的强相关性问题，说明我们所筛选的变量不够独立。原因可能是数据还有降维的空间，接下来需要进行对变量二次筛选。

3.5. 模型的最终求解

我们再次截取了变量进行加权得分降序排列后的前 30 个变量并作出其相关系数热力图。部分变量间存在高相关性，决定按照优先删除高相关、低权重分子描述符的原则进行特征二次提取，因此删除保留的 30 个变量中的 10 个，得到最终的 20 个变量。

3.6. 模型的评价

利用最终保留的 20 个变量进行 LGBM 模型训练，采用网格搜索进行参数优化，

最终得到的 MSE = 0.53799，明显优于变量二次提取之前的模型拟合效果，说明经过二次提取筛选出的 20 个变量更具代表性。拟合效果如图 7 所示：

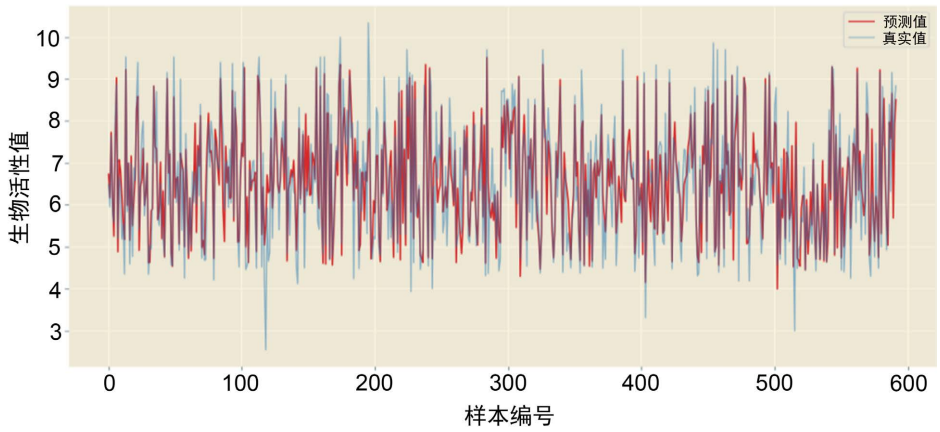

Figure 7. The final model fitting effect diagram of the 20 variables selected
图 7. 最终筛选出的 20 个变量的模型拟合效果图

模型在这 20 个变量的拟合效果非常好，验证了筛选变量的合理性。

4. 定量预测模型建模及预测

4.1. 问题分析

建立化合物对 ERα 生物活性的定量预测模型。本文根据数据预处理之后的样本进行分析，根据问题 1 筛选出来的 20 个分子描述符作为模型的自变量。考虑到在实际 QSAR 建模中，一般采用 PIC_{50} 来表示生物活性值。我们用 PIC_{50} 作为因变量进行模型的拟合。本文决定利用基于集成学习的 XGBoost 算法和 LightGBM 算法，建立对 ERα 生物活性进行精确预测的预测模型，然后进行对比分析。问题二的思维导图如图 8 所示。

4.2. 模型建立

4.2.1. 基于 XGBoost 集成学习算法的 ERα 生物活性预测模型

XGBoost 是集成学习方法的一种。XGBoost 中的基学习器除了可以是 CART 也可以是线性分类器。Boosting 是一种常用的统计学习方法，在训练过程中，通过改变训练样本的权重，学习多个分类器，最终获得最优分类器。XGBoost 在传统 Boosting 的基础上，引入正则化项，加入剪纸，控制了模型的复杂度，同时支持列抽样使学习出来的模型更加简单，借鉴了随机森林的做法，防止过拟合，这也是 XGBoost 优于传统 GBDT 的一个特性[10]。

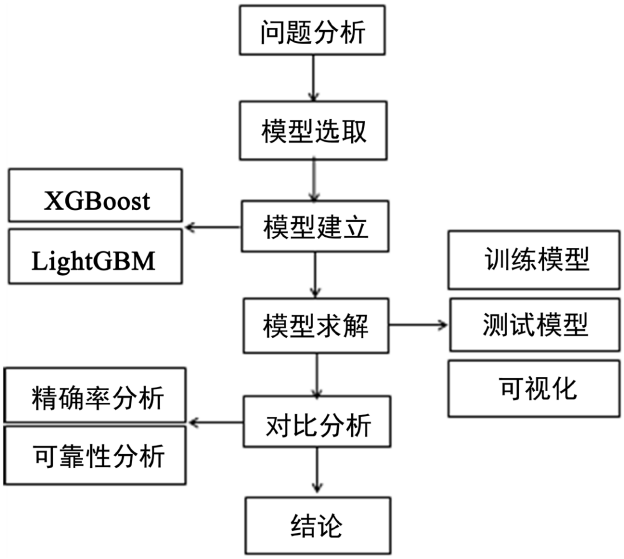

Figure 8. Modeling and prediction mind map
图 8. 建模及预测思维导图

其目标函数可以分为两个部分，一部分是损失函数，一部分是正则(用于控制模型的复杂度)。XGBoost属于一种前向迭代的模型，会训练多棵树。对于 XGBoost 的预测模型可以表示为：

$$L(\Phi)=\sum_i L(\Phi)=\sum_i l\left(\widehat{y}_i-y_i\right)+\sum_k \Omega(f_k) \tag{8}$$

其中 i 表示第 i 个样本，$l\left(\widehat{y}_i-y_i\right)$ 表示第 i 个样本的预测误差，误差越小越好，也就是模型的损失函数。$\sum_k \Omega(f_k)$ 表示树的复杂度的函数，也就是正则项，越小复杂度越低，泛化能力越强[11]。模型学习过程为每一次保留原来的模型不变，加入一个新的函数 f 到模型中，如下式所示：

$$\begin{aligned}
&\widehat{y}_i^{(0)}=0\\
&\widehat{y}_i^{(1)}=f_1(x_i)=\widehat{y}_i^{(0)}+f_1(x_i)\\
&\widehat{y}_i^{(2)}=f_1(x_i)+f_2(x_i)=\widehat{y}_i^{(1)}+f_2(x_i)\\
&\widehat{y}_i^{(t)}=\sum_{k=1}^{t} f_k(x_i)=\widehat{y}_i^{(t-1)}+f_t(x_i)
\end{aligned} \tag{9}$$

由式(9)可知，第 t 轮的模型预测保留了前面 $t-1$ 轮的模型预测，f 函数选择的标准是使目标函数最小化。在通过二阶泰勒展开，得到了最终的目变函数：

$$Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{10}$$

式中的 G、H 与数据点在误差函数上的一阶、二阶导数有关，T 表示叶子的个数，然后不断地枚举不同树的结构，根据目标函数来寻找出一个最优结构的树，加入到我们的模型中，再重复这样的操作。常用的枚举方法是贪心法，每一次尝试去对已有的叶子加入一个分割。对于一个具体的分割方案，我们可以获得的增益(即每个分割方案的分数) [12]可以由如下公式计算：

$$\text{Gain}=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_R+H_L+\lambda}\right]-\gamma \tag{11}$$

每轮迭代时，都需要遍历整个训练数据多次，虽然这样可以找到精确的划分条件。但是计算量较大，时间成本较高。

4.2.2. 基于 LightGBM 集成学习算法的 ERα 生物活性预测模型

LightGBM 是个快速的，分布式的，高性能的基于决策树算法的梯度提升框架[3]。LightGBM 使用的是 Histogram 算法，其思想是将连续的浮点特征离散成 k 个离散值，并构造宽度为 k 的 Histogram。然后遍历训练数据，统计每个离散值在直方图中的累计统计量。在进行特征选择时，只需要根据直方图的离散值，遍历寻找最优的分割点。占用的内存更低，数据分隔的复杂度更低[13]。

同时它带有深度限制的 Leaf-wise 的叶子生长策略。Leaf-wise 则是一种更为高效的策略，每次从当前所有叶子中，找到分裂增益最大的一个叶子，然后分裂，如此循环。可以降低更多的误差，得到更好的精度。Leaf-wise 的缺点是可能会长出比较深的决策树，产生过拟合。因此 LightGBM 在 Leaf-wise 之上增加了一个最大深度的限制，在保证高效率的同时防止过拟合[14]。如图 9 所示：

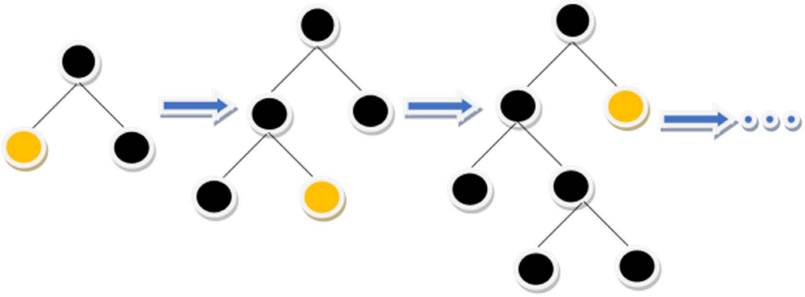

Figure 9. Leaf-wise leaf growth strategy
图 9. Leaf-wise 的叶子生长策略

4.3. 模型求解

由于我们使用 PIC_{50} 作为模型的因变量，因此要得到 IC_{50} 的测试结果需要如下公式转化：

$$J = \frac{P}{P+q+r} \tag{12}$$

将训练样本划分为 70%训练集和 30%测试集并进行网格搜索参数优。

4.3.1. 基于 XGBoost 算法的 ERα 生物活性预测模型求解

对 XGBoost 进行网格搜索参数优化，取值范围为：

最优参数取值为：

{'colsample_bytree': 1.0,

'gamma': 0.0,

'learning_rate': 0.12,

'max_depth': 10,

'min_child_weight': 2,

'n_estimators': 35,

'reg_alpha': 0.1,

'subsample': 1.0}

最终模型的拟合效果如图 10 所示：

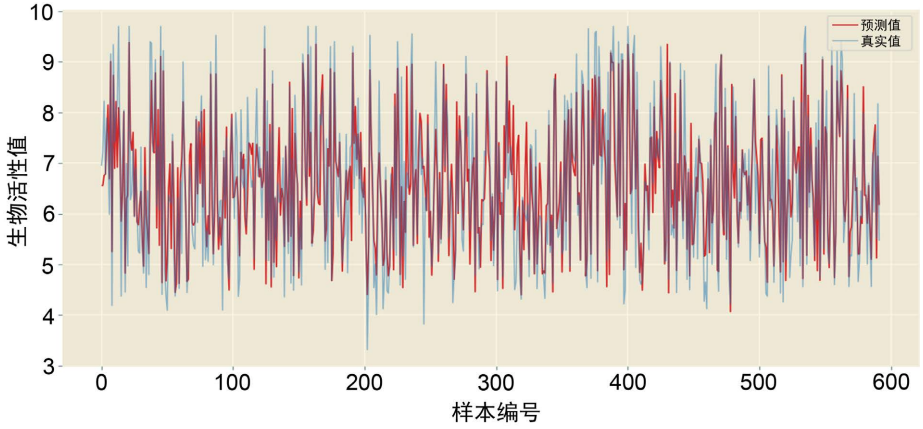

Figure 10. Fitting effect of XGBoost model
图 10. XGBoost 模型拟合效果

4.3.2. 基于 LightGBM 算法的 ERα 生物活性预测模型求解

对 LightGBM 进行网格搜索参数优化最优参数取值为：

{'learning_rate': 0.132,

'max_depth': 10,

'n_estimators': 60,

'num_leaves': 15}

最终模型的拟合效果如图 11 所示：

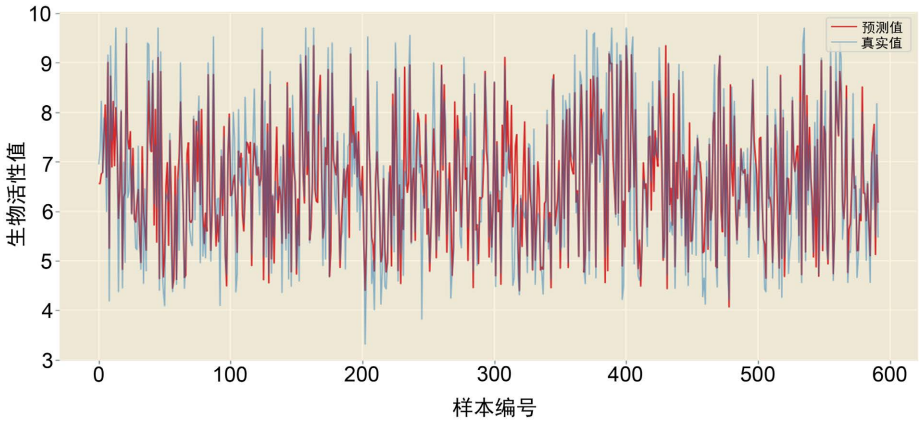

Figure 11. Fitting effect of LightGBM model
图 11. LightGBM 模型拟合效果

5. 结论

5.1. 问题一的结论

1) 筛选出的 20 个变量：MDEC-23，maxssO，minsOH，fragC，SssO，maxHBint10，maxHBa，mindssC，minHBa，minaasC，Zagreb，MDEC-22，CrippenLogP，nwHBa，minsssN，nHother，minssCH2，C1SP2，SaasC，SPC-6。

2) 通过 LightGBM 模型训练，得到了更优的拟合结果，对经过一系列操作最终筛出的 20 个变量的代表性和独立性进行了验证，同时也验证了筛选过程的合理性。

5.2. 问题二的结论

得到在 XGBoost 模型测试集上的评估结果为 MSE = 0.5829、MAE = 0.5625、RMSE = 0.7635、R-square = 0.7118；得到在 LightGBM 模型测试集上的评估结果为 MSE = 0.5273、MAE = 0.5525、RMSE = 0.7262、R-square = 0.7549。结果如表 4 所示：

Table 4. Comparison of model fitting effects
表 4. 模型拟合效果对比

	MSE	MAE	RMSE	R-square
XGBoost	0.5829	0.5625	0.7635	0.7118
LightGBM	0.5273	0.5525	0.7262	0.7549

如表 4 所示，MSE，MAE，RMSE 的值越小越好，R-square 越大越好，所以 LightGBM 模型的拟合效果更好。选取 LightGBM 模型对问题给出的 50 个样本的预测结果作为最终预测值。

致　谢

感谢在我完成这篇论文的过程中给予我帮助的老师和同学，感谢这篇论文所引用的文献的作者。

参考文献

[1] 袁文芳, 张艳琼. 雌激素在妇产疾病中的作用[J]. 医学信息, 2021, 34(10): 54-58.

[2] 王嘉铭, 李宁宁, 唐毅, 苏榕. 雄激素受体与乳腺癌内分泌耐药的研究进展[J]. 现代肿瘤医学, 2021, 29(3): 524-527.

[3] 高世勇, 吕凤, 许东旭. 雌激素受体及其与乳腺癌相关性研究进展[J]. 药学进展, 2020, 44(11): 861-868.

[4] Yang, L., Sang, C.H., Wang, Y.H., Liu, W.T., Hao, W.Y., Chang, J. and Li, J.Z. (2021) Development of QSAR Models for Evaluating Pesticide Toxicity against *Skeletonema costatum*. *Chemosphere*, **285**, 131456. https://doi.org/10.1016/j.chemosphere.2021.131456

[5] Zhang, H., Shen, C., Zhang, H.R., Chen, W.X., Luo, Q.Q. and Ding, L. (2021) Discovery of Novel DGAT1 Inhibitors by Combination of Machine Learning Methods, Pharmacophore Model and 3D-QSAR Model. *Molecular Diversity*, **25**, 1481-1495. https://doi.org/10.1007/s11030-021-10247-x

[6] 徐爱兰, 朱晏民, 孙强, 於香湘, 彭小燕. 基于 K-means 划分区域的深度学习空气质量预报[J]. 南通大学学报(自然科学版), 2021, 20(3): 49-56.

[7] 魏东, 张天祎, 冉义兵. 基于特征选择及机器学习的犯罪预测方法综述[J]. 科学技术与工程, 2021, 21(28): 11910-11920.

[8] 李郅琴, 杜建强, 聂斌, 熊旺平, 黄灿奕, 李欢. 特征选择方法综述[J]. 计算机工程与应用, 2019, 55(24): 10-19.

[9] 雷惠敏, 张和生. 最优特征选择下多层次分割的城市道路提取[J]. 中国空间学术, 2021: 1-9. http://kns.cnki.net/kcms/detail/11.1859.V.20211025.1101.002.html

[10] 廉睿玲. XGBoost 算法在四川省 GPM 降水数据降尺度中的应用[J]. 水电能源科学, 2021, 39(10): 14-17.

[11] 王晓东, 安瑞东. 基于机器学习的热轧带钢力学性能预测模型及应用[J]. 塑性工程学报, 2021, 28(10): 155-165.

[12] 齐巧娜, 刘艳, 陈霁晖, 刘昕竹, 杨锐, 张津源, 崔梦璇, 谢艺萌, 王则远, 于泽, 高飞, 张健. 机器学习 XGBoost 算法在医学领域的应用研究进展[J]. 分子影像学杂志, 2021, 44(5): 856-862.

[13] 宫鹏, 王德兴, 袁红春, 陈冠奇, 吴若有. 基于 LightGBM 的南太平洋长鳍金枪鱼渔场预报模型研究[J]. 水产科学, 2021, 40(5): 762-767.

[14] 上官艺, 王孟, 王春娟, 谷鸿秋, 赵性泉, 王伊龙, 王拥军, 李子孝. 基于机器学习的缺血性卒中功能预后预测模型研究[J]. 中国卒中杂志, 2021, 16(9): 895-900.