Operations Research and Fuzziology 运筹与模糊学, 2024, 14(4), 514-523
Published Online August 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.144419

Hans汉斯

基于Light-GBM算法对生物活性的定量预测

毛轩晴

浙江理工大学启新学院，浙江 杭州

收稿日期：2024年6月29日；录用日期：2024年8月14日；发布日期：2024年8月21日

摘　要

雌激素受体α亚型(ERα)作为乳腺癌内分泌疗法的重要靶点，拮抗ERα活性的化合物可能是治疗乳腺癌的候选药物。本文首先对数据进预处理，包括使用肘部法则和轮廓系数确定K-means聚类K值，再进行聚类处理，并用安德鲁斯曲线可视化。采用方差过滤法和随机森林法对分子描述符进行重要性排序，并对初筛变量进行皮尔逊相关性分析，得到对生物活性影响最显著且独立性较强的20个分子描述符。接着，基于Light-GBM算法建立化合物ERα生物活性的定量预测模型，将数据集按照4:1的比例划分为训练集和测试集。测试集的MSE为0.468、RMSE为0.684、MAE为0.499、R-square为0.788。本文的模型具有较高的预测精度，能加快新药的研发速度，有助于研究乳腺癌的发生和发展机制。

关键词

随机森林，方差过滤法，Light-GBM，距离相关系数，K-Means聚类，肘部法则，轮廓系数

Quantitative Prediction of Biological Activity of Anti-Breast Cancer Drug Candidates Based on Light-GBM Algorithm

Xuanqing Mao

Qixin College, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: Jun. 29th, 2024; accepted: Aug. 14th, 2024; published: Aug. 21st, 2024

Abstract

As an essential target for endocrine therapy of breast cancer, estrogen receptor (ERα) subtypes may be candidates for drug discovery against breast cancer if the compounds can antagonize ER activity. This study initially preprocesses the data, including determining the K value of K-means clustering using the elbow method and silhouette coefficient, conducting clustering, and visualiz-

文章引用: 毛轩晴. 基于 Light-GBM 算法对生物活性的定量预测[J]. 运筹与模糊学, 2024, 14(4): 514-523.
DOI: 10.12677/orf.2024.144419

ing the results with Andrews curves. Then, variance filtering and random forest methods are used to rank the molecular descriptors in terms of importance. Pearson correlation analysis is further applied to the initially screened variables, resulting in 20 molecular descriptors that have the most significant and independent impacts on biological activity. Subsequently, a quantitative prediction model for ER bioactivity of compounds is built based on the Light-GBM algorithm. The dataset is divided into a training set and a test set at a ratio of 4:1. The model performance on the test set shows an MSE of 0.468, RMSE of 0.684, MAE of 0.499, and R-square of 0.788. This model exhibits high prediction accuracy, which can accelerate the development of new drugs and contribute to the research on the occurrence and development mechanisms of breast cancer.

Keywords

Random Forest, Variance Filtering, Light-GBM, Distance Correlation Coefficient, K-Means Clustering, Elbow Rule, Silhouette Coefficient

1. 引言

癌症因其高致死率备受人们关注，根据世界卫生组织属下的国际癌症研究所发布的最新数据，2020年全球乳腺癌病例首次超过了肺癌，以第一的比例占全球所有新发癌症病例的11.7% [1]。对于治疗乳腺癌的药物研究，国内外已有不少学者在乳腺癌分子靶点和靶向治疗方向上取得显著进展，已发现不少抗乳腺癌活性表现良好的化合物，且在临床实践中取得明显疗效，例如查耳酮类化合物、他莫昔芬和雷诺昔芬。特别是雌激素受体 α 亚型(ERα)作为乳腺癌内分泌疗法的重要靶点[2]，在超过 70%的乳腺癌患者中过度表达，因此拮抗 ERα 活性的化合物可能是治疗乳腺癌的候选药物[3]。

近年来随着药物大数据平台的实现，不少学者在研究治疗乳腺癌过程中运用教据挖掘方法得到重要结论。例如秦璞等[4]应用随机森林和支持向量机对三阴性乳腺癌基因数据的降维和筛选，得到部分基因和三阴性乳腺癌的转移或者预后有相关性等；王江翔[5]等提出粒子群算法优化能够提升模型的预测性能，优化后的 Light-GBM 模型预测效果更好。随着抗乳腺癌药物的生物活性被逐渐深入研究，评价抗乳腺癌药物的副作用的研究也越发受到关注，例如魏静[6]通过实验研究得到羧甲基 β-葡聚糖联合阿霉素具有协同抗乳腺癌以及减轻心脏毒性的功能。

因此，本文旨在筛选出最显著且不相关的 20 个分子描述符，构建一个更为精简且高效的生物活性定量预测模型，从而提高模型对未知化合物生物活性的预测精度，快速评估大量潜在乳腺癌药物的生物活性，加快药物筛选和评估的速度，推动药物研发领域的不断进步。

2. 数据来源与预处理

2.1. 数据描述

本文数据来自2021年华为杯中国研究生数学建模竞赛D题中1974个化合物对ERα的生物活性数据，1974 个化合物的 729 个分子描述符信息和 1974 个化合物的 5 种 ADMET 性质的数据。IC50 表示的化合物对的生物活性值，是实验测定值。它的单位是 nM，值越小代表生物活性越大，对抑制活性越有效。将 IC50 值进行负对数转化而得到的 pIC50。该值通常与生物活性具有正相关性，即 pIC50 值越大表明生物

活性越高。在建模过程中，本文采用 pIC50 来表示生物活性值。

2.2. 数据预处理

考虑到由于测量误差、仪器故障、人为疏忽等各种原因，原始数据可能会存在数据缺失、数据异常、数据偏差等问题。因此，分析之前有必要对 1974 个样本以及 729 个分子描述符(即自变量)的数据进行预处理，确保数据的质量和完整性。首先，本文通过遍历查找数据中是否有缺失值，发现数据完整。接着，根据拉依达准则检测是否有异常值，发现没有粗大误差值，即依拉达准则下没有异常值。

为了进一步检查数据是否有异常值，提高模型预测结果的稳健性。我们采用聚类分析的方法，将相似性强的对象尽可能聚集到一起，分离不同数据。常见的聚类方法有基于层次的凝聚层次聚类、基于划分的 K-means 聚类、基于密度的 DBSCAN 聚类等等[7]。由于 K-means 对噪声和离群值非常敏感，因此，本文采用 K-means 聚类检测异常样本。

K-means 聚类是最常用的基于欧式距离的聚类算法，它的目标是将数据点分组到 K 个簇中，以使簇内的点尽可能相似，而簇间的点尽可能不同。它的核心思想是通过迭代优化簇中心的位置，以最小化簇内的平方误差总和。它的优点是对于大型数据集也是简单高效，时间复杂度、空间复杂度低。但它需要预先设定 K 值，对最先的 K 个点选取很敏感。因此，为了确定最佳的 K-means 聚类簇数 K，本文结合肘部法则和轮廓系数来决定。

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。当聚类数 k 较小时，每个簇的聚合程度较低，误差平方和 SSE 下降幅度较大；随着 k 越接近真实聚类数，每个簇的聚合程度变高，SSE 下降幅度逐渐减小；当 k 大于真实聚类数时，SSE 下降幅度趋于平缓。手肘法[8]的核心指标是误差平方和(SSE)，公式如下：

$$SSE = \sum_{i=0}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{1}$$

其中，C_i 是第 i 个簇，p 是 C_i 中的样本点，m_i 是 C_i 的质心(C_i 中所有样本的均值)，SSE 是所有样本的聚类误差，代表了聚类效果的好坏。

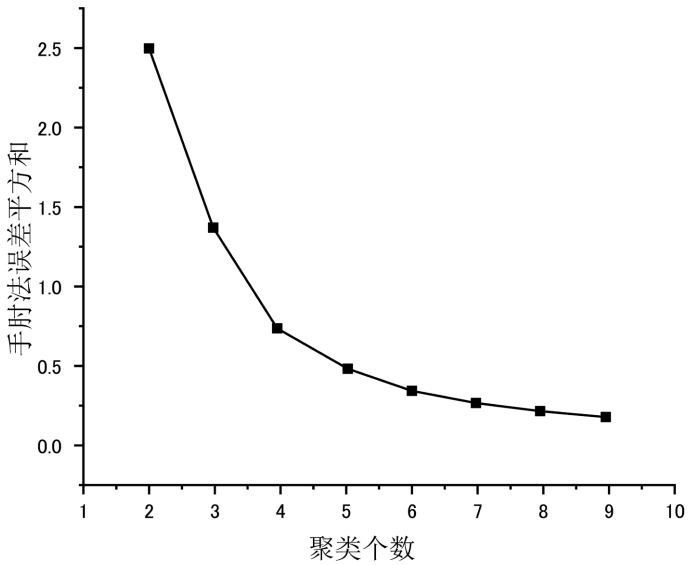

Figure 1. Elbow rule determines the number of clusters
图 1. 肘部法则确定聚类个数

如图 1 所示，通过观察 *SSE* 随 *k* 的变化曲线，在 $k = 4$ 时，畸变程度得到大幅改善，之后缓慢下降，考虑选取临界点 $k = 4$ 作为聚类数量。

轮廓系数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。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 *S*。第 *i* 个对象的轮廓系数 *S* 的计算公式为：

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2}$$

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。

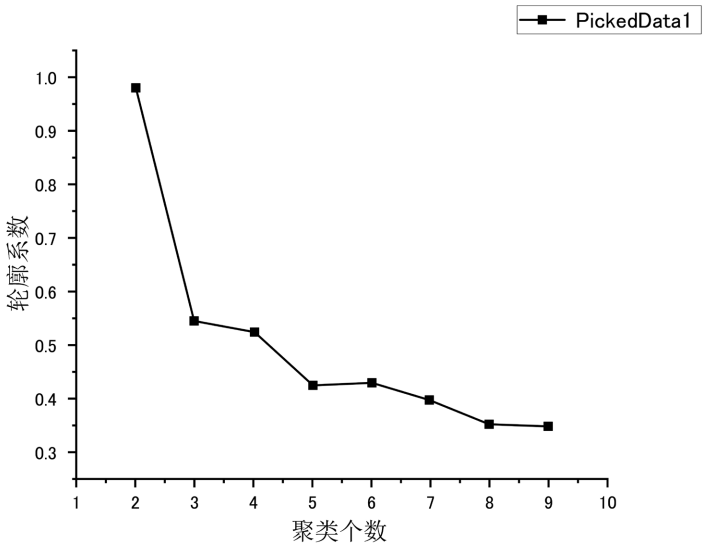

Figure 2. The contour coefficient determines the number of clusters
图 2. 轮廓系数确定聚类个数

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，本文最终选择 *k* 值为 4。经过 K-means 聚类算法，得到分类结果：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。

安德鲁斯曲线是一种用于显示数据分布密度的统计工具，它将数据的分布范围分为几个区间，然后根据数据落在每个区间的频率进行绘制。本文借助安德鲁斯曲线直观展示各个簇的分布情况。具体来说，横轴代表数据值，纵轴代表频率或者概率密度，每个簇的数据会被归一化到同一范围，然后以柱状图的形式展示在曲线上。其公式为：

$$f(t) = \frac{x_1}{sqrt(2)} + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + x_5\cos(2t) + \cdots \tag{3}$$

结果如图 3 所示，观察安德鲁斯曲线的形状和簇的分布情况能判断是否存在异常样本。如果某个簇的数据在大部分区间内都没有分布，或者分布极其不均匀，那么这个簇就可能包含异常样本。反之，如果各个簇的数据在多个区间内有较为均匀的分布，那么这个数据集就可能是正常且稳定的。我们可以发现第 1 类样本曲线波动幅度大，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。

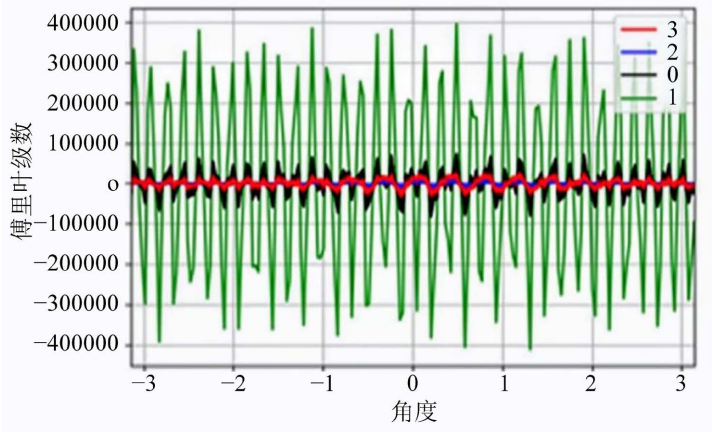

Figure 3. Diagram of Andrews curve
图 3. 安德鲁斯曲线图

3. 筛选分子描述符

为了从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，即为特征选择(变量选择)问题。考虑到分子描述符与生物活性之间的关系可能是非线性的，传统的统计方法可能无法识别这些非线性的关系。且为了更有效率地处理大量的数据，本文选择利用机器学习算法解决。依据特征选择和学习器的不同结合方式，可以将变量筛选选择方法大致分为四类[9]：Filter、Wrapper、嵌入式和混合式。为了评价自变量的离散程度和特征与目标之间的相关性，本文分别选择了过滤式方法中的方差过滤法以及嵌入式方法中的随机森林法。

3.1. 计算方差过滤筛选分子描述符

方差过滤法是一种基于变量离散程度的特征选择方法，其通过特征本身的方差来筛选特征。方差越大，说明数据的离散程度越大，特征越明显。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，我们对方差进行归一化处理，使取值都在(0, 1)之间，从而使得各个分子描述符的方差具有可比性。数据归一化处理公式如下：

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

将归一化处理后的方差值作为变量重要性的衡量标准，方差的计算公式如下：

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n} \tag{5}$$

最后，将归一化处理后的方差值降序排列从而得到一个变量重要性排序。

3.2. 随机森林模型筛选分子描述符

如图4所示，随机森林模型会在原始数据集中随机抽样，构成 n 个不同的样本数据集，搭建 n 个不同的决策树模型，最后根据决策树模型的平均值或者投票情况获取最终结果。为了保证模型的泛化能力，随机森林模型在建立每棵树时遵循“数据随机”和“特征随机”原则[10]。我们引入一个参数 k 来控制随机性的引入程度 $k = \log_2 d$。

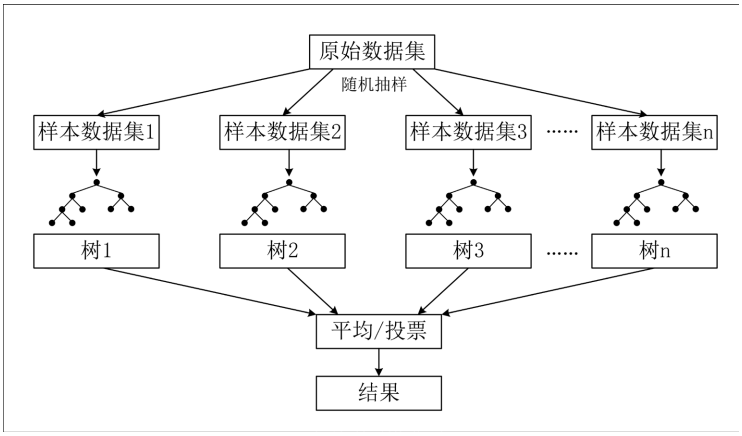

Figure 4. Schematic diagram of the random forest model
图 4. 随机森林模型示意图

假设集成包含 T 个基学习器 $\{h_1, h_2, \cdots, h_T\}$，其中 h_i 在实例 x 上的输出为 $h_i(x)$。使用简单平均法对于若干和弱学习器的输出进行平均得到最终的预测输出。即最终预测为

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x) \tag{6}$$

为了剔除掉相关性特别高的特征，本文采用距离相关系数法。两个变量之间的总体的皮尔逊相关系数定义为两个变量之间的协方差和标准差之积的商，公式为：

$$\rho_{xy} = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}|x| \operatorname{var}|y|}} \tag{7}$$

其中，$\operatorname{cov}(x, y)$ 为 x 与 y 的协方差，$\operatorname{var}|x|$ 为 x 的方差，$\operatorname{var}|y|$ 为 y 的方差。ρ_{xy} 表示 x 与 y 之间的相关系数。$|\rho_{xy}| \neq 0$ 时，$|\rho_{xy}|$ 的值越大，表示两个特征 x 和 y 之间的相关程度就越大。$|\rho_{xy}| = 0$ 时，表示 x 与 y 相互独立，不存在线性关系，x 与 y 之间没有相关性。

3.3. 去除强相关变量的距离相关系数法

本文应用方差过滤法和随机森林法，将筛选出的变量根据重要性降序排列，得分越高说明变量的价值越大。每种方法选取重要性前 40 的变量，表 1 展示前 20 个变量的重要性排序。

Table 1. The importance of the twenty groups of variables corresponding to the two methods
表 1. 两种方法相对应的变量重要性排序前 20 组

	方差过滤法		随机森林法	
序号	Variable	Importance	Variable	Importance
1	WPATH	1.000000	MDEC-23	1.000
2	fragC	0.143560	LipoaffinityIndex	0.362
3	ATSp5	0.037091	maxHsOH	0.269
4	ATSp4	0.034041	minsssN	0.203
5	ATSp3	0.026320	C1SP2	0.188
6	ATSp2	0.012247	maxssO	0.183
7	ATSp1	0.008582	minHsOH	0.164

续表

8	ECCEN	0.005179	BCUTc-1l	0.131
9	VABC	0.000248	nC	0.114
10	MW	0.000241	minHBint5	0.087
11	TopoPSA	0.000038	minsOH	0.082
12	Zagreb	0.000033	nHBAcc	0.073
13	AMR	0.000019	MLogP	0.073
14	CrippenMR	0.000019	VC-5	0.066
15	ETA_Eta_R	0.000011	TopoPSA	0.058
16	SHBa	0.000009	MLFER_A	0.053
17	apol	0.000007	SHsOH	0.052
18	sumI	0.000007	MDEO-12	0.052
19	nBonds2	0.000007	MDEC-33	0.041
20	nAtom	0.000006	ATSc3	0.038

对比筛选出的变量，每种方法筛选出来的变量差异较大。为了验证方法的有效性，分别采用方差过滤法和随机森林法得到的每一组序列的前 40 个分子描述符作为模型特征，以抑制分子活性程度值 PIC50 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。MSE 计算公式如下：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2 \tag{8}$$

其中，m 表示样本量，$\hat{y}_i$ 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。

得到了方差过滤法和随机森林法两组特征在测试集上得到的 MSE 结果，保留两位小数的结果分别是 1.06 和 0.78。随机森林法的 MSE 值远远小于方差过滤法的 MSE 值，表明在预测分子活性程度上，随机森林法的筛选出的变量表现最好。观察数据可以发现方差过滤法筛选出的变量 WPATH 和 fragC，分别以重要性 1.0 和 0.14356 远远高于其他分子描述符，因此，选择 WPATH 和 fragC 作为最终的变量。再选取随机森林筛选出的重要性前 28 的变量，得到 30 个分子描述符进行相关性分析。

由于采用距离相关系数法来进行筛选，本文将阈值设为 0.7，认为相关性值大于 0.7 的两个变量具有高相关性，将其中一个变量进行剔除，保留重要性较高的；如果小于 0.7，说明两个变量相关性不强。最后在此基础上按照优先删除高相关、低重要性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，分别是：MDEC-23、ndssC、BCUTc-1l、BCUTc-1h、SHsOH、TopoPSA、VC-5、MLFER_A、C2SP2、minHBa、MDEC-23、CrippenLogP、minHBint10、MDEO-12、maxssO、maxdssC、VCH-5、minHsOH、minHBint5、ETA_BetaP_s。

3.4. 相关性分析

为了验证筛选出的 20 个分子描述符相关性，将这 20 个分子描述符的相关性系数绘制热力图进行可视化。热力图中单元格颜色越接近色阶的顶端(红色)，表示正相关越强；颜色越接近色阶的底端(深蓝色)，表示负相关越强。图 5 表示这 20 个描述符之间的相关性较低，有较强的独立性，保证了模型的泛化能力。

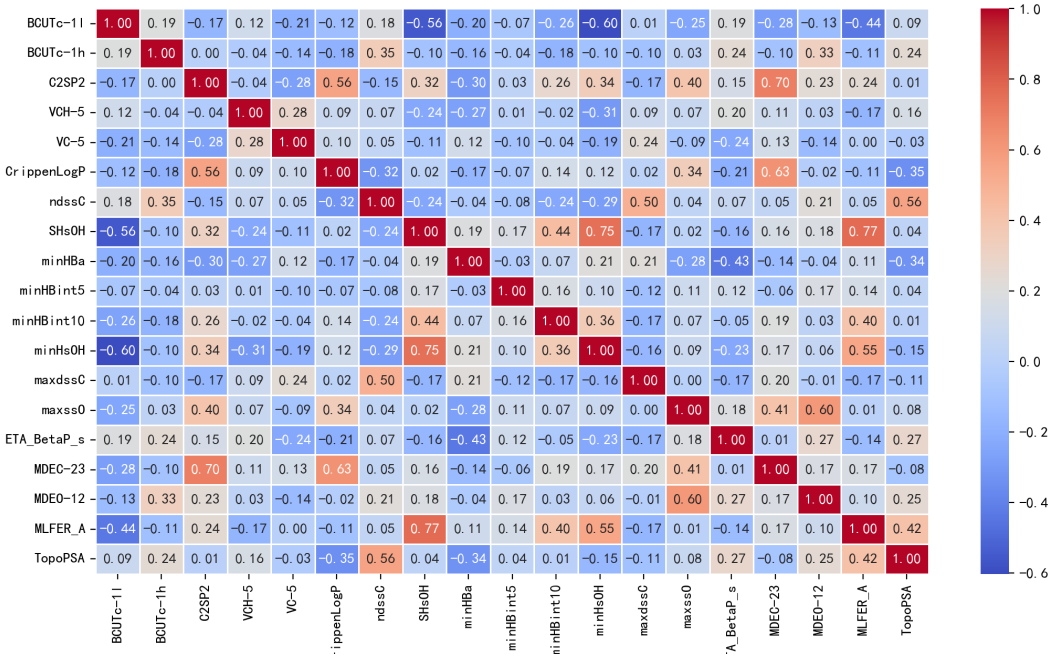

Figure 5. Correlation heat map of the main molecular descriptors
图 5. 主要分子描述符的相关性热力图

4. 基于 LightGBM 算法的生物活性定量预测

4.1. 基于 LightGBM 算法的预测模型

LightGBM 算法通过使用一种新的分裂算法，称为直方图分裂算法，能更有效地找到数据中的分裂点；使用新的叶子生长策略，称为极端梯度提升，这种叶子生长策略可以更有效地拟合数据，从而提高决策树的预测精度；使用新的正则化方法，称为 L1 正则化，防止决策树过拟合数据，从而提高决策树的泛化能力[11]。据此，本文使用 LightGBM 算法对生物活性进行定量预测。

LightGBM 梯度提升算法的核心是最小化损失函数。损失函数 $L\left(y,F\left(x\right)\right)$ 衡量真实值 y 和模型输出 $F(x)$之间的差异[12]。常使用平方损失，为

$$L\left(y,F\left(x\right)\right)=\frac{1}{2}\left(y-F\left(x\right)\right)^{2} \tag{9}$$

梯度提升通过迭代地构建弱学习器来最小化损失函数。在每次迭代中，计算损失函数对模型输出的梯度，用于构建一棵回归树来逼近负梯度：

$$\text{Gradient}=\frac{\partial L\left(y,F\left(x\right)\right)}{\partial F\left(x\right)} \tag{10}$$

$$\hat{y}_i^t=\hat{y}_i^{t-1}+\text{argmin}_\rho\left(\sum_{i=1}^{n}L\left(y_i,\hat{y}_i^{t-1}+\rho b_t\left(x_i\right)\right)\right) \tag{11}$$

其中，$\hat{y}_i^{t-1}$ 是第$(t-1)$棵树对样本(i)的预测输出，$\left(b_t\left(x_i\right)\right)$ 是第(t)棵树的输出。

最后，通过学习率 η 控制新建树对模型输出的贡献，将每颗树的输出累加到当前模型输出中：

$$F\left(x\right)=F\left(x\right)+\eta\sum_{t=1}^{T}b_t\left(x\right) \tag{12}$$

通过不断调试参数，在防止模型过拟合下又能保持一定的收敛速度，最终设置学习率为 0.1、最大迭代次数为 100、叶子节点数设置为 32、最大深度为−1。将生物活性定量预测模型在测试集上取得的成果进行可视化处理，结果如图 6 所示。此时，基于 LightGBM 的定量预测模型的 MSE 为 0.468、RMSE 为 0.684、MAE 为 0.499、R-square 为 0.788。

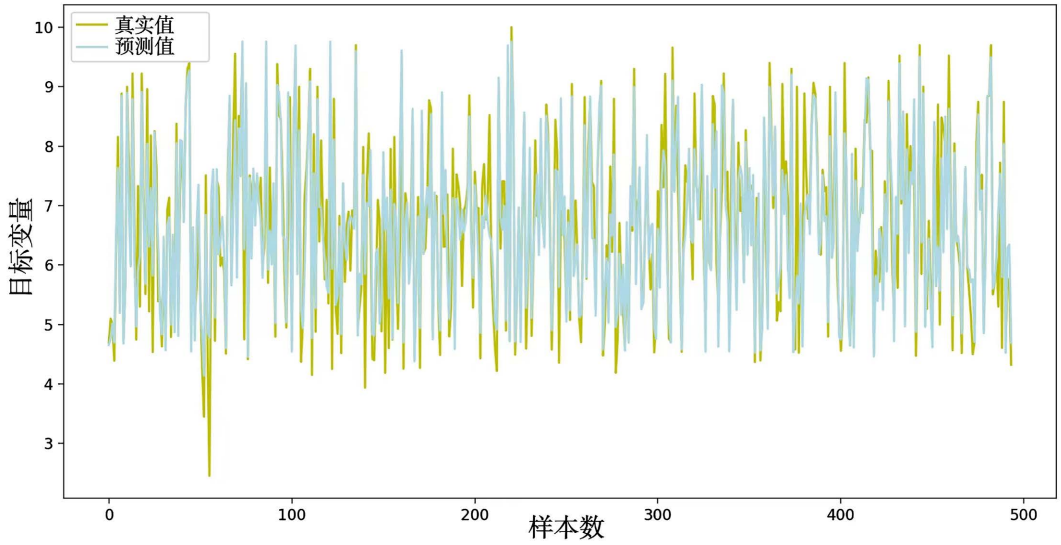

Figure 6. Test set prediction effect of quantitative prediction model based on Light-GBM
图 6. 基于 Light-GBM 的定量预测模型的测试集预测效果

4.2. 预测结果

采用基于 LightGBM 的定量预测模型对 test 表中的 50 个化合物进行 IC50 值和对应的 pIC50 值预测，表 2 展示了前十个化合物的预测值，部分化合物名称过长的部分用省略号代替，索引顺序和原表相同。

Table 2. Predicted values for 50 compounds in the test table "ERα_activity.xlsx"
表 2. 测试表“ERα_activity.xlsx”中 50 种化合物的预测值

SMILES	IC50_nM	pIC50
COc1cc(OC)cc(\C=C\c2ccc(OS(=O)(=O)...	44.92620055	7.347500309
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3ccccc23)c4ccc(O)cc4	17.94817113	7.745979798
COc1ccc2C(=C(CCOc2c1)c3ccc(O)cc3)c4ccc(\C=C\C(=O)O)cc4	13.34744841	7.874601749
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	6.277348282	8.202223775
OC(=O)\C=C\c1ccc(cc1)C2=C(CCSc3cc(F)ccc23)c4ccc(O)cc4	11.86434769	7.925756135
CC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	47.24507998	7.325643411
Oc1ccc(cc1)C2=C(c3ccc(\C=C\c4ccccc4)cc3)c5ccc(F)cc5OCC2	15.60204214	7.806818553
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)c4ccccc4)cc3)c5ccc(F)cc5OCC2	28.2700538	7.548673365
OC(=O)\C=C\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	20.00278735	7.698909482
CCN(CC)C(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	38.47472579	7.414824467

5. 结论与建议

本文关于影响化合物活性的主要分子描述符筛选以及对化合物活性进行定量预测的研究，最终筛选

出 20 个能显著影响生物活性且相互之间相关性较低的分子描述符为：MDEC-23、ndssC、BCUTc-1l、BCUTc-1h、SHsOH、TopoPSA、VC-5、MLFER_A、C2SP2、minHBa、MDEC-23、CrippenLogP、minHBint10、MDEO-12、maxssO、maxdssC、VCH-5、minHsOH、minHBint5、ETA_BetaP_s。这些分子描述符有望为机构在化合物筛选和药物设计过程中提供重要参考，加快新药的研发速度，并为乳腺癌的预防和治疗提供新的思路。本文的模型具有广泛的应用前景，不仅可以应用于乳腺癌治疗靶标生物活性预测，也能对其他癌症的治疗策略研究提供有力支持。

为了简化模型，本文设定的假设与现实情况仍有一定距离，后续可以收集更多的实验数据或采用数据生成技术来扩大数据集规模，以提高模型的泛化能力；调整模型参数或使用其他先进的机器学习算法来进一步提高预测精度[13] [14]。

参考文献

[1] 联合国. 世界癌症日：乳腺癌已超过肺癌成为全球主要新发癌症类型[EB/OL]. https://news.un.org/zh/story/2021/02/1077332, 2024-01-13.

[2] 宁文涛, 胡志烨, 董春娥, 等. 抗乳腺癌双靶点药物研究进展[J]. 中国药物化学杂志, 2020, 30(12): 778-788.

[3] 王斯. 高维数据下基于稀疏神经网络的抗乳腺癌候选药物筛选、预测与优化[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2024.

[4] 秦璞, 郭志旺, 郭维恒, 等. 应用随机森林和支持向量机对三阴性乳腺癌基因数据的降维和筛选[J]. 中国卫生统计, 2020, 37(3): 389-394.

[5] 王江翔, 肖清泉. 基于粒子群算法优化的乳腺癌化合物活性预测研究[J]. 智能计算机与应用, 2023, 13(7): 45-52.

[6] 魏静, 李婷英, 张莹, 等. 羧甲基 β-葡聚糖联合阿霉素抗乳腺癌以及减轻心脏毒性的实验研究[J]. 中国临床药理学杂志, 2021, 37(3): 275-279.

[7] 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.

[8] 林磊, 孙建孟. 基于 K-均值聚类与肘部法则的测井相建立方法研究[C]//中国地球物理学会, 中国地震学会, 等. 2020 年中国地球科学联合学术年会论文集. 青岛: 中国石油大学(华东), 2020.

[9] Sujay, A. and Siva, R.V. (2021) Multimodal Sentiment Analysis Using Relief Feature Selection and Random Forest Classifier. *International Journal of Computers and Applications*, **43**, 1-9.

[10] 汪家清, 韦哲, 张太鹏, 等. 基于随机森林算法的乳腺癌预测模型的研究[J]. 中国医学装备, 2022, 19(1): 119-123.

[11] 孟祥福, 田友发, 张霄雁. 基于 LightGBM 模型的肺腺癌免疫相关基因筛选与患者生存率预测[J]. 生物医学工程学杂志, 2024, 41(1): 70-79.

[12] 吴晖南, 陈淑娇, 陈展峰, 等. 基于 LightGBM 模型的糖尿病预测模型的研究[J]. 中国卫生标准管理, 2023, 14(24): 64-67.

[13] 林瑜, 吴静依, 蔺轲, 等. 基于集成学习模型预测重症患者再入重症监护病房的风险[J]. 北京大学学报(医学版), 2021, 53(3): 566-572.

[14] 郑惠文. 机器学习算法在内科疾病诊断中的应用[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2021.