中国研究生创新实践系列大赛

“华为杯”第十八届中国研究生数学建模竞赛

学校	南京林业大学
参赛队号	21102980066
队员姓名	1.钱伟杰
	2.石泽峰
	3.王汉钊

中国研究生创新实践系列大赛
“华为杯”第十八届中国研究生
数学建模竞赛

题目　　抗乳腺癌候选药物的优化建模

摘　要：

研究发现，雌激素受体 α 亚型（ERα）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。一个化合物想要成为候选药物，除了需要具备良好的生物活性外，还需要在人体内具备良好的药代动力学性质和安全性。通常采用建立化合物生物活性预测模型的方法来筛选潜在活性化合物。本文构建化合物生物活性的定量预测模型和 ADMET 性质的分类预测模型，从而为同时优化 ERα 拮抗剂的生物活性和 ADMET 性质提供预测服务。

本文所做的工作可概括为以下几点:

问题一：首先通过低方差滤波去除 225 个单一值特征变量，再对剩余的 504 个变量进行灰色关联分析筛选出前 200 名的特征变量，将样本特征比提高至接近 10:1。接着使用基于随机森林的递归特征消除算法选取前 30 名的特征变量，考虑到算法的随机性影响，将算法试验 50 次，对每次选出的 30 个变量计数，最后得到出现频数最高的 30 个变量。因得到的 30 个变量只有计数，没有得分排名，再对选出的 30 个变量做 10 次随机森林回归，取 10 次回归的平均值作为 30 个变量最终的相关性得分，选出排名靠前的 20 个变量。同时，对得分靠前的 20 个变量分别计算其与 pIC_{50} 的最大互信息系数得分，距离相关性系数得分，皮尔森系数得分，验证变量选取的合理性。

问题二：结合问题 1 递归特征消除选出的和生物活性相关性最高的 30 个特征变量，将变量按对生物活性相关性从高到低排序，求出变量与变量之间的距离相关系数，再通过类似非极大值抑制的方式，对分数高的变量删去和其距离相关系数为强相关的变量(系数>0.6)，从而保证所选变量的独立性，保证选出的特征子集尽可能最优。接着选用 5 种最常用的非线性模型支持向量回归模型，随机森林回归模型，梯度提升回归树模型，XGBoost 模型和 BP 神经网络来建立生物活性预测模型。将 1974 个样本划分成 80%训练集和 20%的测试集，用训练集训练模型，用测试集对模型进行检验，分别得到 5 种模型的三个评价指标 MSE,MAE,R^2，通过比对这三个指标，最终确定了拟合优度R^2为 0.8076 的梯度提升回归树预测模型。使用模型对 test 文件中的 50 个化合物预测 pIC_{50}，并通过 pIC_{50} 与 IC_{50} 之间的转换公式得到 50 个 IC_{50} 的结果。

问题三:首先对每个 ADMET 性质分别进行最优特征子集的选取，每个性质特征子集选取的步骤相同，以 Caco-2 为例，第一步滤去数据集中 225 个单一值特征变量，第二步使用最大互信息系数求取与 Caco-2 相关性最高的 200 个变量，第三步使用基于随机森林的

递归特征消除算法选取变量，试验 50 次，每次选出 40 个变量，挑选出现频数大于 40 的特征变量，第四步，按随机森林得分排序变量，第五步使用问题二中提出的类似非极大值抑制的独立性变量剔除算法选出最优的特征子集。得到了 5 个性质各自的特征子集后，选用 5 种分类预测模型，通过在测试集上的准确率比较，确定最终各 ADMET 性质的分类预测模型。一共选出三个支持向量机分类模型和两个 XGBoost 分类模型，使用模型对 test 文件中 50 个化合物预测 5 个性质的分类结果。

问题四：筛选样本数据，分析主要变量分布，选定需要优化的变量。为满足 ADMET 中至少有三个性质较好及各变量上下限的约束条件下，以最大化 pIC_{50} 为目标，建立单目标优化模型。通过差分进化算法求解，得到满足约束条件下的 pIC_{50} 最优解为 9.5537。进行多次迭代，获得的多组最优解差异浮动最大值仅为 2.06%，验证了模型的稳定性和合理性。

关键词：灰色关联分析；支持向量机；梯度提升回归树；RFE-RF；差分进化算法；pIC_{50}

目录

一、问题重述 5
二、基本假设 7
三、模型符号说明 8
四、问题一模型的建立和与求解 9
4.1 问题分析 9
4.2 分子描述符变量筛选模型的建立 9
4.2.1 灰色关联分析原理(GRA) 10
4.4.2 基于随机森林的递归特征消除算法 RFE-RF 11
4.3 模型评估与合理性验证 14
4.4 模型求解 15
4.4.1 方差分析 15
4.4.2 灰色关联分析 15
4.4.3 基于随机森林的递归特征消除 16
4.4.4 随机森林回归重要性排序 17
4.5 合理性验证 18
4.6 小结与讨论 18
五、问题二模型建立和求解 19
5.1 问题分析 19
5.2 生物活性预测模型的建立 19
5.2.1 变量独立性剔除算法的建立 20
5.2.2 五种非线性模型的建立 20
5.3 模型评估方法 26
5.4 模型的求解 26
5.4.1 最优特征子集的选取 26
5.4.2 数据 Z-score 标准化 27
5.4.3 数据划分 28
5.4.4 模型训练及比较 28
5.4.5 模型训练结果可视化 28
5.4.6 梯度提升回归树的模型参数 29
5.4.7 预测 test 表中的化合物 29
5.5 小结与讨论 29
六、问题三模型的建立与求解 30
6.1 问题分析 30
6.2 各 ADMET 分类预测模型的建立 30
6.3 分类模型的评估方法 31
6.4 模型的求解 31
6.4.1 MIC 相关性分析 31
6.4.2 基于随机森林的递归特征消除试验 31
6.4.3 距离相关系数独立性分析 32
6.4.4 模型训练前数据处理 34
6.4.5 模型训练及比较 34

6.4.6 SVM 模型和 XGBoost 模型参数 35
6.4.7 预测 test 表中的 50 个化合物 36
6.4.8 小结与讨论 36
七、问题四模型的建立与求解 38
7.1 问题分析 38
7.2 筛选样本数据，分析主要变量分布 39
7.3 建立目标优化准则 41
7.4 单目标优化模型的建立 42
7.4.1 优化目标 42
7.4.2 约束分析 43
7.4.3 模型建立 43
7.5 操作方案模型的求解 43
7.6 模型合理性验证 45
7.7 小结与讨论 45
八、模型评价与改进 47
8.1 模型优点 47
8.2 模型缺点 47
8.3 模型的改进与推广 47
参考文献 48
附录 49

一、问题重述

据统计,乳腺癌是世界上第二大常见的癌症,仅次于肺癌。乳腺癌在女性癌症中发病率最高(如图 1.1 所示),图中不同颜色代表全球女性高发癌症的地区分布,其中粉色代表乳腺癌,红色代表宫颈癌,黄色代表肝癌,蓝色代表胃癌,绿色代表甲状腺癌，数字表示全球女性高发癌症分布的国家或地区数[1,2]。中国肿瘤登记中心 2015 年年报中指出,乳腺癌仍居女性癌症发病率首位,并且呈上升趋势[3]。

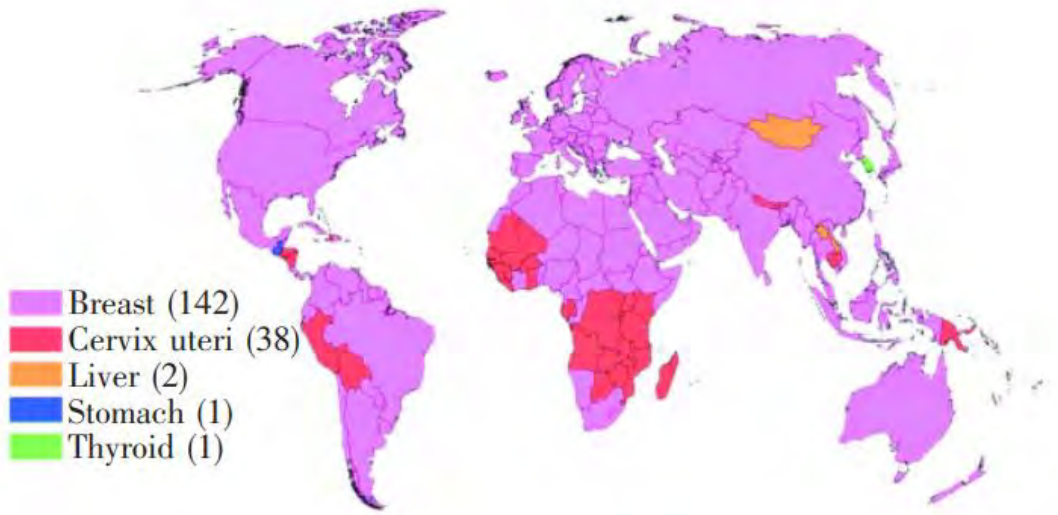


图 1.1 全球女性高发癌症地区分布

有研究发现，雌激素受体α亚型（Estrogenreceptorsalpha,ERα）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。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（QuantitativeStructure-ActivityRelationship,QSAR）模型的方法来筛选潜在活性化合物，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。

一个化合物想要成为候选药物，除了需要具备良好的生物活性（此处指抗乳腺癌活性）外，还需要在人体内具备良好的药代动力学性质和安全性，合称为 ADMET（Absorption 吸收、Distribution 分布、Metabolism 代谢、Excretion 排泄、Toxicity 毒性）性质。本文仅考虑化合物的 5 种 ADMET 性质，分别是：1）小肠上皮细胞渗透性（Caco-2）；2）细胞色素 P450 酶（CytochromeP450,CYP）3A4 亚型（CYP3A4）；3）化合物心脏安全性评价（humanEther-a-go-goRelatedGene,hERG）；4）人体口服生物利用度（HumanOralBioavailability,HOB）；5）微核试验（Micronucleus，MN）。

针对乳腺癌治疗靶标 ERα，由于化合物的分子描述符的复杂性和考虑其 ADMET 性质，通过构建化合物生物活性的定量预测模型和 ADMET 性质的分类预测模型，从而为同时优化 ERα拮抗剂的生物活性和 ADMET 性质提供预测服务。

基于上述研究背景，本文需要研究完成以下问题：

问题一：

根据附件二“Molecular_Descriptor.xlsx”和附件一“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，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分子描述符（即变量）。

问题二：

根据问题一的排序，从 729 个分子描述符筛选出建模主要变量构建化合物对 ERα 生物活性的定量预测模型，使之尽可能具有代表性、独立性，并对附件一“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。

问题三：

利用附件二“Molecular_Descriptor.xlsx”中提供的 729 个分子描述符，针对附件四“ADMET.xlsx”中提供的 1974 个化合物的 ADMET 数据，分别构建化合物的 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型，然后使用所构建的 5 个分类预测模型，对附件四“ADMET.xlsx”的 test 表中的 50 个化合物进行相应的预测。

问题四：

寻找分子描述符，以及这些分子描述符在什么取值或者处于什么取值范围时，能够使化合物对抑制 ERα 具有更好的生物活性，同时具有更好的 ADMET 性质。

二、基本假设

1.假定所有附件中所有提供的数据均为真实数据；
2.假定数据样本的各个变量数值都在规定的数据范围之内；
3.假定化合物抑制 ERα分子活性除题中所给的分子描述符外，不受外界影响；
4.假定各分子描述符均可以独立取值。

三、模型符号说明

序号	符号	含义
1	x_i	特征在优化模型中的变量名
2	A_i	第i个预测模型的特征子集
3	$Caco-2_{pred}$	$Caco-2$的分类预测模型
4	$CYP3A4_{pred}$	$CYP3A4$的分类预测模型
5	$hERG_{pred}$	$hERG$的分类预测模型
6	HOB_{pred}	HOB的分类预测模型
7	MN_{pred}	MN的分类预测模型
8	$pIC50_{pred}$	$pIC50$的回归预测模型

四、问题一模型的建立和与求解

4.1 问题分析

本题的任务为：根据附件二和附件一中提供的数据，针对 1974 个化合物的 729 个分子描述符进行变量选择，并将选出的变量按照重要性程度进行排序，最后给出前 20 个对生物活性最具有显著影响的分子描述符。

本题主要有以下两个难点：

难点一：化合物即样本数的个数为 1974 个，变量数即分子描述符的个数为 729 个，样本数量并没有和特征数量拉开一定的差距，在不做任何处理的情况下，若采用基于模型的回归算法进行特征选择易引起严重的过拟合。针对难点一：先对变量进行预筛选，通过方差分析法去除仅有单一值的变量 225 个，再对剩余的 504 个变量进行灰色关联分析，按照关联度进行排序，初步选出前 200 个变量，缓解直接使用回归模型来做特征选择时会产生的过拟合现象。

难点二：由于目标值生物活性指标 pIC_{50} 与各个变量之间的重要性及相关性并不能直接确定是线性关系还是非线性关系，若二者为非线性关系，可得因、自变量的相关程度低。针对难点二：在灰色关联分析的基础上，为了精准的得到因、自变量的相关程度，本题比对了三类常用的特征选择算法过滤式、包裹式、嵌入式的优缺点，最终选择了包裹式算法基于随机森林回归模型的递归特征消除来选出关联度最高的 30 个变量。由于该算法每次迭代过程时，变量得分会有一定的随机性，将算法试验 50 次，对每次试验选出的变量进行计数，得到了出现频率最高的 30 个变量。最后对第二次筛选后的 30 个变量再做 10 次随机森林回归，按照得分将 30 个变量进行排序，最终得出分数最高的 20 个变量，本题通过多次随机森林以降低平衡算法随机性带来的影响。

4.2 分子描述符变量筛选模型的建立

本文的分子描述符变量筛选模型建立的研究流程如图 4.1 所示，首先使用低方差滤波法，去除 225 个单一值特征变量；其次，对滤波后的变量候选列表与目标 pIC_{50} 之间进行灰色关联分析(Grey Relation Analysis，GRA)，获取前 200 个与目标关联性最高的主要候选变量，再对这些变量使用基于随机森林的递归特征消除算法(Recursive feature elimination-Random Forest Regressor，RFE-RF)，将算法试验 50 次，对每次试验选出的变量进行计数，得到了出现频率最高的 30 个变量。最后，对这 30 个变量使用随机森林回归试验，得到了重要性前 20 的变量的排序，并对筛选出来的变量以最大互信息系数(Maximal Information Coefficient，MIC)、皮尔森相关系数(Pearson correlation coefficient)、距离相关系数(Distance correlation coefficient)进行合理性验证。

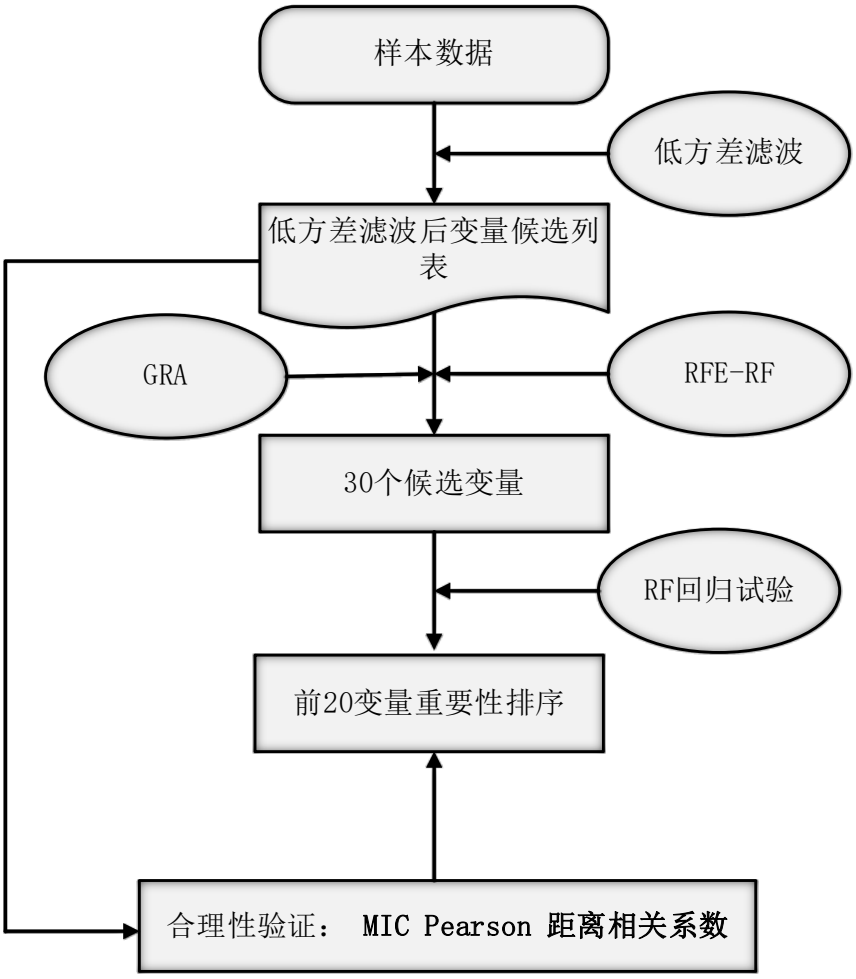


图 4.1 主要变量筛选的研究流程

4.2.1 灰色关联分析原理(GRA)

为了尽可能筛选出与生物活性影响关联性最大的特征，本题先对低方差滤波后的 504 个变量选用灰色关联分析进行初步筛选。基本思想是根据序列曲线几何形状的相似程度来判断其联系是否紧密。曲线越接近，相应序列之间的关联度就越大，反之就越小。灰色综合关联度既体现了折线X_0与X_i的相似程度，也可反映X_0与X_i相对于始点的变化速率的接近程度，是较为全面地表征序列之间联系是否紧密的一个数量指标[4-5]。

灰色关联度$\gamma(X_0,X_i)$常简记γ_{0i}，k点关联系数$\gamma[x_0(k),x_i(k)]$简记为$\gamma_{0i}(k)$,计算步骤如下:

第 1 步:求 pIC_{50} 及 504 个分子描述符的初值像(或均值像)，令

$$X_i' = \frac{X_i}{x_i(1)} = [x_i'(1), x_i'(2), \cdots, x_i'(n)]; i = 0,1,2,\cdots,m \tag{4-1}$$

第 2 步:求差序列，记

$$\Delta_i(k) = |x_0'(k) - x_i'(k)| \tag{4-2}$$

$$\Delta_i = [\Delta_i(1), \Delta_i(2), \cdots, \Delta_i(n); i = 0,1,2,\cdots,m \tag{4-3}$$

第 3 步:求两极最大差与最小差，记

$$M = \max_i \left[\max_k \Delta_i(k)\right], m = \min_i \left[\min_k \Delta_i(k)\right] \tag{4-4}$$

第 4 步:求各分子描述符与 pIC_{50} 的关联系数

$$\gamma_{0i}(k)=\frac{m+\xi M}{\Delta_i(k)+\xi M},\xi\in(0,1);k=1,2,\cdots,n;i=1,2,\cdots,m \tag{4-5}$$

第 5 步:计算各分子描述符与 pIC_{50} 的关联度

$$\gamma_{0i}(k)=\frac{1}{n}\sum_{k=1}^{n}\gamma_{0i}(k);i=1,2,\cdots,m \tag{4-6}$$

4.4.2 基于随机森林的递归特征消除算法 RFE-RF

在得到 504 个分子描述符与 pIC_{50} 的灰色相关性排序后，本题选取前 200 个分子描述符，使用基于随机森林的递归特征消除算法(RFE-RF)对其进一步筛选。

递归特征消除主要思想是针对那些特征含有权重的预测模型，RFE 通过递归的方式，不断减少特征集的规模来选择需要的特征。

第一：给每一个特征指定一个权重，接着采用预测模型在这些原始的特征上进行训练。

第二：在获取到特征的权重值后，对这些权重值取绝对值，把最小绝对值剔除掉。

第三：按照这样做，不断循环递归，直至剩余的特征数量达到所需的特征数量。

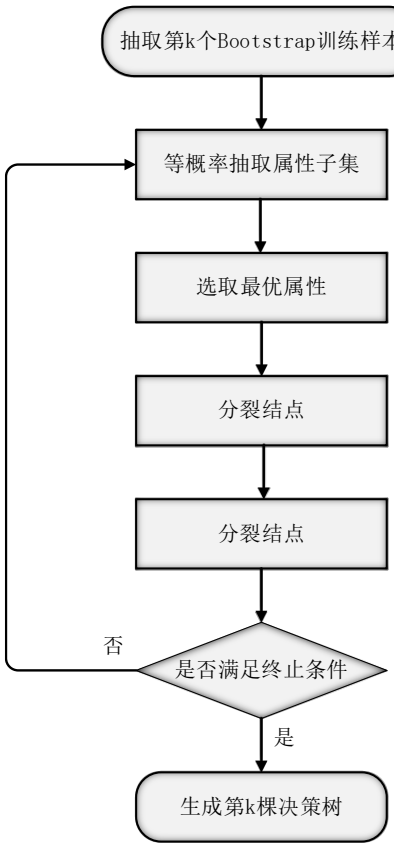
随机森林是以K个决策树$\{h(X,\theta_k),k=1,2,\cdots,K\}$为基本分类器，进行集成学习后得到的一个组合分类器。当输入待分类样本时，随机森林输出的分类结果由每个决策树的分类结果简单投票决定。这里的$\{\theta_k,k=1,2,\cdots,K\}$是一个随机变量序列，它是由随机森林的两大随机化思想决定的：

（1）Bagging思想：从原样本集X中有放回地随机抽取K个与原样本集同样大小的训练样本集$\{T_k,k=1,2,\cdots,K\}$，每个训练样本集T_k构造一个对应的决策树。

（2）特征子空间思想：在对决策树每个节点进行分裂时，从全部属性中等概率随机抽取一个属性子集(通常取$\log_2(M)+1$个属性，M为特征总数），再从这个子集中选择一个最优属性来分裂节点。

训练随机森林的过程就是训练各个决策树的过程，由于各个决策树的训练是相互独立的，因此随机森林的训练可以通过并行处理来实现，这将大大提高生成模型的效率。随机森林中第k个决策树$h(X,\theta_k)$的训练过程如图4.2所示。

将以同样的方式训练得到K个决策树组合起来，就可以得到一个随机森林。当输入待分类的样本时，随机森林输出的分类结果由每个决策树的输出结果进行简单投票（即取众数）决定。随机森林分类流程如图4.3所示。



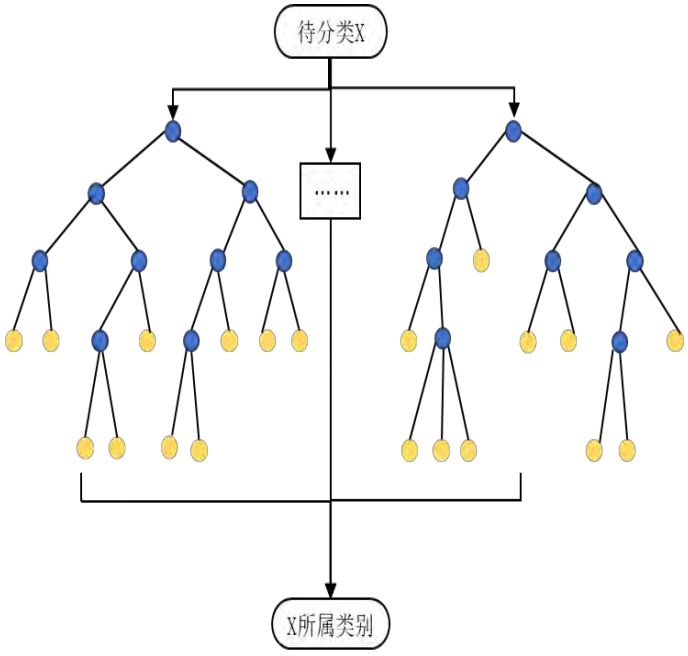


图4.2 随机森林中单个决策树训练过程　　**图4.3** 随机森林分类过程

随机森林对噪声和异常值有较好的容忍性，能够在不需要降维的条件下处理具有高维特征的输入样本，而且能够评估各个特征在分类问题上的重要性，具有良好的可扩展性和并行性。

随机森林算法可以在分类的基础上进行回归分析，通过将样本分类的结果进行一定的运算可以获得各个分子描述符的重要性，表示各分子描述符对预测pIC_{50}的影响程度，某一分子描述符重要性越大，表明该分子描述符对预测pIC_{50}的影响越大，重要性越小，表明该分子描述符对预测pIC_{50}的影响越小。随机森林算法中某一特征的重要性，是该特征在内部所有决策树重要性的平均值，而在决策树中，计算某一个特征的重要性可以采用以下方法：

即某一节点k的重要性为：

$$n_k = \omega_k * G_k - \omega_{left} * G_{left} - \omega_{right} * G_{right} \tag{4-7}$$

其中，$\omega_k, \omega_{left}, \omega_{right}$分别为节点k以及其左右节点中训练样本与总训练样本数目的比例，G_k, G_{left}, G_{right}分别为节点k以及其左右子节点的不纯度。知道每一个节点的重要性之后，即通过公式得出某一特征的重要性。

$$f_i = \frac{\sum_{j \in nodessplitonfeaturei} n_j}{\sum_{k \in allnodes} n_k} \tag{4-8}$$

为了使所有特征的重要性加起来等于 1，需要对每一个特征重要性进行归一化处理，

如下式:

$$f_{ni} = \frac{f_i}{\sum_{j \in allnodes} f_j} \tag{4-9}$$

最终得到这 200 个分子描述符对预测 pIC_{50} 的重要性排序，进行递归消除。

由于随机森林每颗树的训练样本是随机的，树中每个节点的分裂属性集合也是随机选择确定的，导致此算法结果具有一定的随机性，为了尽可能平衡此随机性对结果的影响，本题将对这 200 个特征进行五十次 RFE-RF 算法，对每次试验选出的特征进行计数，得到了出现频率最高的 30 个特征。

最后，本题对这 30 个特征进行了 10 次基于随机森林的重要性排序，取其平均值得到前 20 个对 pIC_{50} 影响最大的特征及其排序。

基于随机森林的 RFE 在本文中的实现可概括如图 4.4 所示。

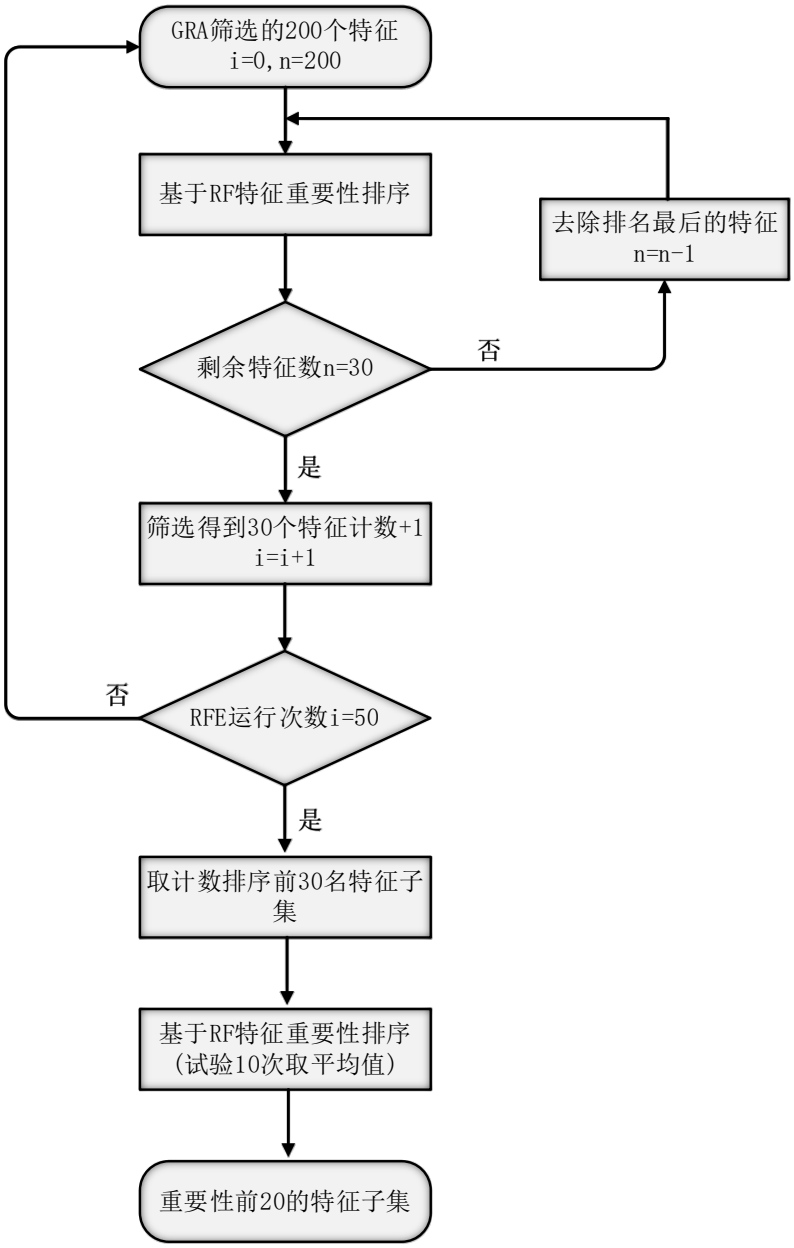


图 4.4 RFE-RF 运行流程

4.3 模型评估与合理性验证

本题除了使用经典的 Pearson 相关系数进行评估以外，还选用了最大互信息系数(Maximal Information Coefficient, MIC)以及距离相关系数(Distance correlation coefficient)，对主要变量筛选进行重要性平稳度的合理性验证。

（1）Pearson 相关系数

Pearson 相关系数[6-7]定义如式 4-10 所示。

$$r=\frac{\sum_{i=1}^{n}(X_i-X)(Y_i-Y)}{\sqrt{\sum_{i=1}^{n}(X_i-X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-Y)^2}} \tag{4-10}$$

其中，X，Y 分别是第一个变量和第二个变量的均值。Pearson 相关系数是衡量两个变量线性相关程度的指标，该指标取值在-1 到 1 之间，越接近 1，表示这两个变量正相关性越强，越接近-1，表示这两个变量的负相关性越强，越接近 0，表示这两个变量线性相关性很弱或者不是线性关系。

（2）距离相关系数

距离相关系数[8-9]定义如式（4-11）所示。

$$R(X,Y)=\sqrt{R^2(X,Y)}=\begin{cases}\sqrt{\dfrac{V^2(X,Y)}{\sqrt{V^2(X)V^2(Y)}}} & V^2(X)V^2(Y)>0\\ 0 & V^2(X)V^2(Y)=0\end{cases} \tag{4-11}$$

其中，距离协方差为：

$$V(X,Y)=\sqrt{V^2(X,Y)}=\sqrt{\left\|\int_{X,Y}(l,s)-\int_X(l)\int_Y(s)\right\|^2} \tag{4-12}$$

距离方差为:

$$V(X)=\sqrt{V^2(X,X)}=\sqrt{\left\|\int_{X,X}(l,s)-\int_X(l)\int_X(s)\right\|^2} \tag{4-13}$$

（3）最大互信息系数MIC

最大互信息系数[10]定义如式（4-14）所示

$$MIC(x;y)=\max_{a*b<B}\frac{I(x;y)}{\log_2\min(a,b)} \tag{4-14}$$

上式中：a,b是在x,y方向上的划分格子的个数，B设置为0.6。

MIC 计算分为三个步骤：

1. 给定 i、j，对 XY 构成的散点图进行 i 列 j 行网格化，并求出最大的互信息值
2. 对最大的互信息值进行归一化
3. 选择不同尺度下互信息的最大值作为 MIC 值

4.4 模型求解

4.4.1 方差分析

使用 Python 编程实现方差分析，在这里只去除单一值特征变量，总计 225 个，部分要删除的变量名称如表 4.1 所示：

表 4.1 去除单一值特征变量

序号	分子描述符
1	nB
2	nBondsQ
3	nHsNH3p
4	nHssNH2p
5	nssBe
6	naOm
7	nSm
8	nsGeH3
9	nssGeH2
10	nsssGeH
⋮	⋮

4.4.2 灰色关联分析

本题对剩余 504 个变量使用 Python 编程实现灰色关联分析，选出前 200 个特征变量，算法参数：分辨系数值选择为 0.5。

灰色关联分析得分情况如表 4.2 所示：

表 4.2 筛选出的分子描述符和灰色关联系数得分

序号	分子描述符	灰色关联系数得分
1	LipoaffinityIndex	0.8034
2	MDEC-23	0.7997
3	SwHBa	0.7930
4	MLogP	0.7880
5	n6Ring	0.7861
6	nwHBa	0.7860
7	nRing	0.7856
8	C2SP2	0.7838
9	ETA_Psi_1	0.7812
10	nT6Ring	0.7809
⋮	⋮	⋮
198	nsCH3	0.5700
199	ATSm2	0.5690
200	nBonds	0.5688
201	MDEO-12	0.5687

202	SP-2	0.5655
⋮	⋮	⋮
500	maxaaN	0.5021
501	minaa0	0.5015
502	maxaa0	0.5013
503	mindsN	0.5003
504	maxdsN	0.5003

由上表可知：当分辨系数值选择为 0.5，灰色关联得分的最大值和最小值之间的差距在 0.3，第一名和第 200 名的得分差距在 0.23，后面从 201 名到最后 504 名的得分差距在 0.068，可以很明显的看出对生物活性影响较大的分子描述符变量在前 200 名内。

4.4.3 基于随机森林的递归特征消除

由于递归特征消除的特征选择结果具有一定的随机性，本题对灰色关联分析得到的前 200 名的特征做 50 次递归特征消除算法试验，每次试验输出 30 个特征变量，对每次试验出现的特征进行计数，选取前 30 个出现频数最高的特征，各变量出现的频数如图 4.4 所示。

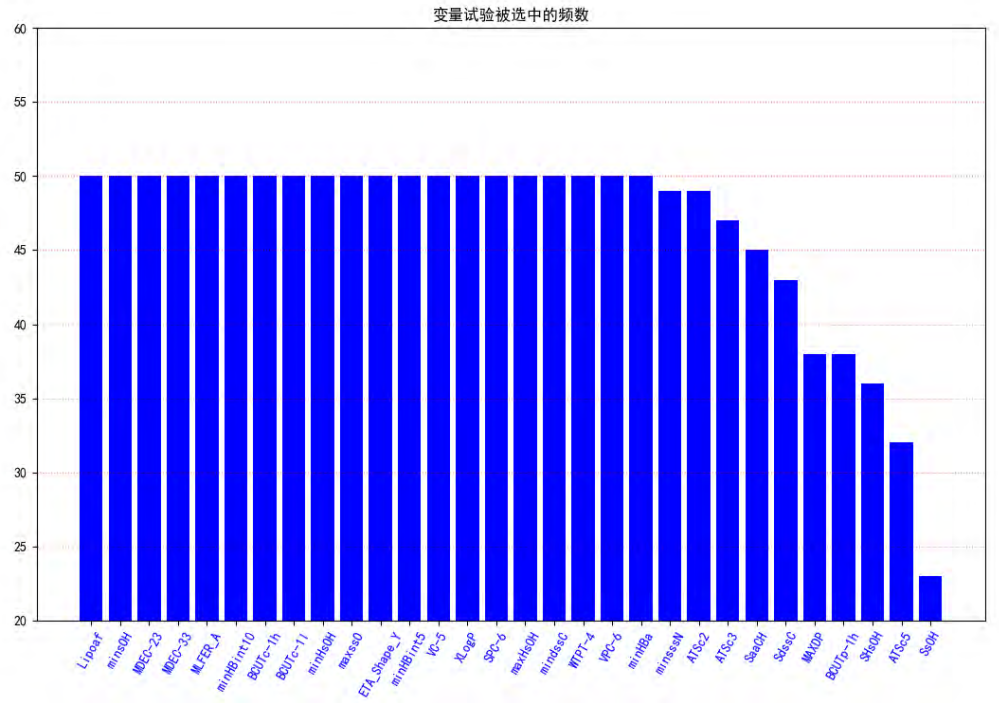


图 4.4 各变量出现的频次如图所示

使用 Python 编程实现递归特征消除试验，具体参数设置见表 4.3 所示：

表 4.3 递归特征消除试验参数设置

参数名	参数值
输出特征数量	30
每轮剔除的变量数	1

决策树个数	500

4.4.4 随机森林回归重要性排序

由于算法递归特征消除可以较准确的得到与生物活性相关性大的特征变量，但其无法对所得的 30 个变量按照重要性进行排序。为此通过对 30 个变量再做随机森林回归，得到 30 个变量的重要性排序。考虑到随机森林算法结果同样具有一定的随机性，通过试验 10 次，对 10 次结果取平均值，得到最终要求的前 20 个分子描述符，具体结果如表 4.4 所示。

表 4.4 基于随机森林回归重要性得分排序

序号	分子描述符名称	对生物活性影响重要性得分
1	MDEC-23	0.2267
2	LipoaffinityIndex	0.0612
3	BCUTc-1l	0.0487
4	minsssN	0.0433
5	maxssO	0.0398
6	maxHsOH	0.0390
7	minHsOH	0.0303
8	MLFER_A	0.0270
9	minsOH	0.0254
10	minHBint5	0.0235
11	BCUTc-1h	0.0224
12	mindssC	0.0209
13	VC-5	0.0207
14	VPC-6	0.0205
15	XLogP	0.0202
16	minHBa	0.0183
17	ATSc3	0.0170
18	MDEC-33	0.0164
19	SHsOH	0.0163
20	minHBint10	0.0161
21	ATSc2	0.0161
22	ETA_Shape_Y	0.0158
23	WTPT-4	0.0156
24	SPC-6	0.0151
25	SdssC	0.0139
26	SsOH	0.0134
27	SaaCH	0.0134
28	BCUTp-1h	0.0129
29	ATSc5	0.0114
30	MAXDP	0.0091

4.5 合理性验证

为了验证特征筛选的合理性及其重要性排名的可靠性，本题使用最大互信息系数(MIC)、距离相关系数、皮尔森系数的方法来对本题确定的 20 个按重要性排序的变量进行合理性验证，结果如表 4.5 所示。

表 4.5 分子描述符合理性验证

序号	分子描述符	最大互信息系数	距离相关系数	皮尔森系数
1	MDEC-23	0.3394	0.5425	0.5380
2	LipoaffinityIndex	0.3198	0.5146	0.4919
3	BCUTc-1l	0.3938	0.399	-0.3193
4	minsssN	0.2639	0.4262	0.4307
5	maxssO	0.2624	0.2586	0.2246
6	maxHsOH	0.3194	0.4653	0.4088
7	minHsOH	0.3300	0.4667	0.3991
8	MLFER_A	0.2981	0.4266	0.3580
9	minsOH	0.3099	0.4759	0.4661
10	minHBint5	0.1569	0.1603	0.1052
11	BCUTc-1h	0.3717	0.3464	-0.2818
12	mindssC	0.2551	0.2437	0.1666
13	VC-5	0.2751	0.2347	0.1032
14	VPC-6	0.2522	0.3593	0.2108
15	XLogP	0.2183	0.3747	0.3285
16	minHBa	0.2409	0.2295	-0.1113
17	ATSc3	0.2167	0.3741	-0.3519
18	MDEC-33	0.2402	0.2985	0.2514
19	SHsOH	0.3257	0.4149	0.3147
20	minHBint10	0.2303	0.3532	0.3206

综合表 4.4、表 4.5 和图 4.4 可以看出 20 个特征在几种衡量特征重要性的评价方法下的异同。可以得到如下规律:

(1)互信息—最大信息数 MIC、距离相关系数、基于随机森林回归的特征重要性排序较为相近，体现了主要变量在不同规则下的平稳度；

(2)此 20 个特征与生物活性之间的距离相关系数和最大信息数 MIC 都较大，体现了其都对目标影响较大，验证了本题变量选择的合理性。

(3)Pearson 相关系数得分中所有变量得分的绝对值都未超过 0.6，可得这些特征与生物活性的线性相关性比较低，所以预测 pIC_{50} 时更适合使用非线性的回归模型进行预测。

4.6 小结与讨论

本题筛选主要变量时，存在两个主要的难题：

(1)变量之间具有高度非线性和相互强偶联性，一般的线性模型无法直接衡量变量的相关性和重要性；

(2)随机森林进行变量重要性评价时会产生一定的随机性，为了尽量平衡这种随机性造成的影响，需要进行多次试验。

五、问题二模型建立和求解

5.1 问题分析

本题的任务：结合问题 1 的分析，选择不超过 20 个分子描述符，构建化合物对 ERα 生物活性的定量预测模型，并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。

本题主要有以下**两个难点**：

难点一：题目要求用来拟合预测模型的变量数不能超过 20 个，按照第一问重要性排序的 20 个变量间可能存在较大的相关性，即 20 个变量之间包含的有用信息较少，建立出的模型效果会较差。为了使选出用于回归的 20 个变量包含最大的信息量，还需充分考虑变量之间独立性的影响。

难点二：由上一章用于模型验证的皮尔森系数可知，各个变量与生物活性指标 pIC_{50} 之间呈线性关系的可能不大，存在高度非线性关系的可能性较大，因此需要选取相应的非线性模型作为本题生物活性预测模型。

5.2 生物活性预测模型的建立

本题模型建立及求解的流程图如图 5.1 所示，首先选取和生物活性相关性最高的 30 个特征变量，求出他们之间的距离相关系数，再通过类似非极大值抑制的方式，将变量按它们对生物活性相关性从高到低排序，对分数高的变量删去和他们距离相关系数为强相关的变量，从而保证所选变量的独立性，保证选出的特征子集含有最大的信息量。

解决完变量的独立性问题后，选用 5 种最常用的非线性模型：支持向量回归模型，随机森林回归模型，梯度提升回归树模型，XGBoost 模型和 BP 神经网络来建立生物活性预测模型。将 1974 个样本划分成 80%训练集和 20%的测试集，用训练集训练，用测试集对 5 种非线性模型进行检验，分别得到 5 种模型的三个评价指标 MSE,MAE,R^2，通过对比这三个指标，选取最终的生物活性 pIC_{50} 预测模型。

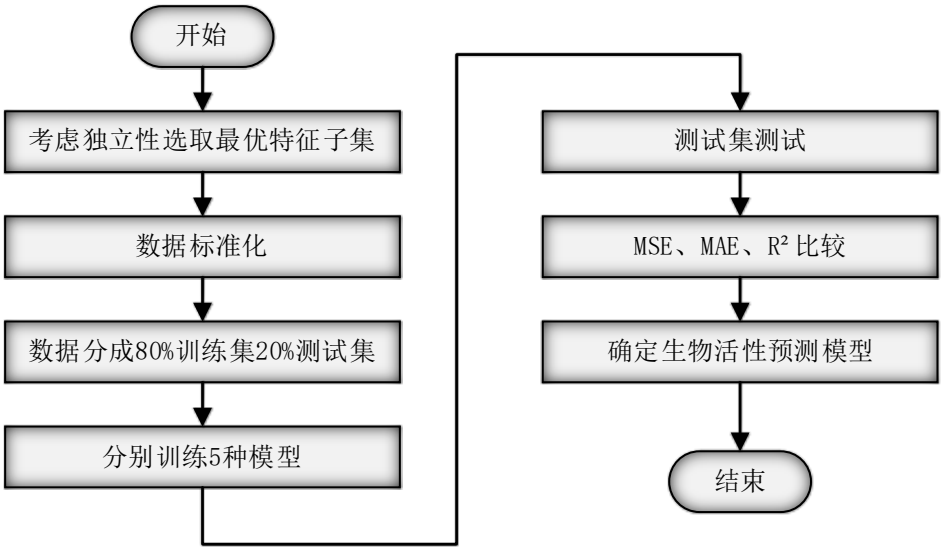


图 5.1 问题二求解流程图

5.2.1 变量独立性剔除算法的建立

在统计学中，对相关性强弱有如下约定，如表 5.1 所示。

表 5.1 相关程度度量表

相关性	相关系数
极强相关	0.8~1.0
强相关	0.6~0.8
中相关	0.4~0.6
弱相关	0.2~0.4
极弱或无相关	0 ~0.2

为了保证变量之间的独立性，本题的首要目标是得到一特征子集，使特征变量之间的相关性小于 0.6。

为此，本题首先选取第一问按随机森林得分排序后的 30 个变量，计算 30 个变量两两之间的距离相关系数，使用类似非极大值抑制的方式对 30 个变量进行变量剔除，剔除的阈值选择为 0.6，具体算法流程如图 5.2 所示。

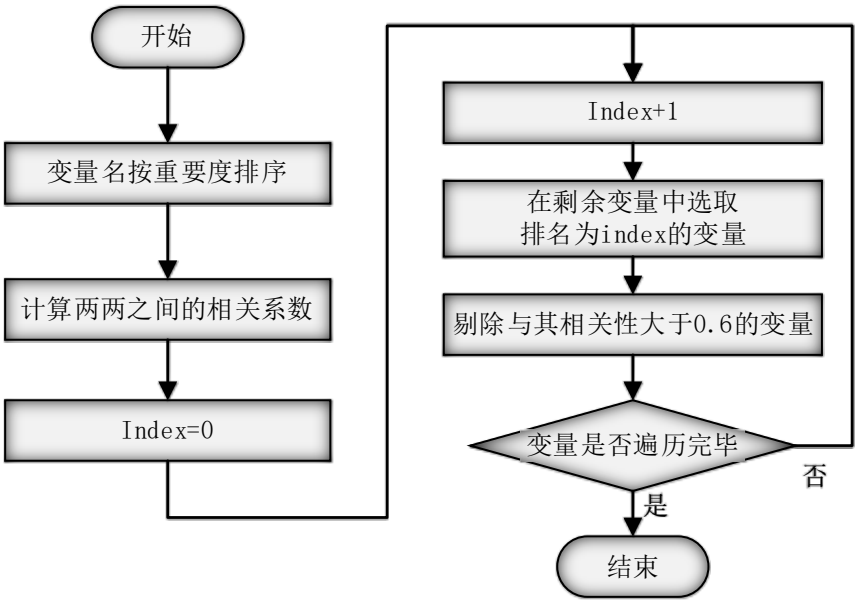


图 5.2 变量独立性剔除算法流程图

5.2.2 五种非线性模型的建立

(1)BP 神经网络模型的建立

本题选用 BP 神经网络作为第一个非线性生物活性预测模型。建立七层网络结构模型，具体网络结构如下表 5.2 所示。

表 5.2 七层 BP 神经网络结构图

网络层号	神经元个数	激活函数
1	19	RELU

2	38	RELU
3	76	RELU
4	38	RELU
5	19	RELU
6	8	RELU
7	1	-

其中第一层为输入层，一共有 19 个分子描述符输入，第二层到第六层为隐藏层，使用 relu 激活函数来提高模型的非线性表达能力，最后一层为输出层，输出生物活性 pIC_{50} 的预测结果。

网络使用的激活函数 RELU 公式如下所示：

$$f(x) = \max(0, x) \tag{5-1}$$

模型训练时，反向传播的优化器选择 adam 优化器，adam 是一种学习率自适应的优化算法，它综合了 GDM 和 RMSprop 优化器的优点。adam 优化器根据目标函数对每个参数的梯度一阶矩和二阶矩估计动态调整针对每个参数的学习率。具体公式如下：

$$\begin{cases} \theta_i^{(k+1)} = \theta_i^{(k)} - g_i^{(k)} \\ g_i^{(k)} = \dfrac{\eta \widehat{v_i}^{(k)}}{\sqrt{\widehat{s_i}^{(k)} + \varepsilon}} \end{cases} \tag{5-2}$$

式中：$g_i^{(k)}$是第k次迭代时；神经网络第i个参数沿梯度方向下降的距离$v_i^{(k)}$和$s_i^{(k)}$分别是第i个参数在第k次迭代时；历史梯度平方的指数衰减平均和历史梯度的指数衰减平均；$\widehat{v_i}^{(k)}$和$\widehat{s_i}^{(k)}$偏差修正，目的是消除迭代初期由于k过小而导致的梯度权值和过小的影响。其计算公式如下：

$$\begin{cases} \widehat{v_i}^{(k)} = \dfrac{v_i^{(k)}}{1 - \beta_1^{(k)}} \\ \widehat{s_i}^{(k)} = \dfrac{s_i^{(k)}}{1 - \beta_2^{(k)}} \\ v_i^{(k+1)} = \beta_1 v_i^{(k)} + (1 - \beta_1) \dfrac{\partial E_R}{\partial \theta_i}(\theta^{(k)}) \\ s_i^{(k+1)} = \beta_2 s_i^{(k)} + (1 - \beta_2)(\dfrac{\partial E_R}{\partial \theta_i}(\theta^{(k)}))^2 \end{cases} \tag{5-3}$$

式中：超参数$0 \le \beta_1 < 1$，$0 \le \beta_2 < 1$，本题分别取为 0.9 和 0.999。

模型的损失函数选择 MSE 指标，具体公式如下：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2 \tag{5-4}$$

本题 BP 神经网络其余参数设置如表 5.3 所示。

表 5.3 BP 神经网络其余参数设置

序号	参数名	参数值
1	Batchsize	64

2	迭代轮数	5000
3	学习率	0.01

(2)支持向量机回归模型的建立

本题选用支持向量回归作为第二个非线性生物活性预测模型。在解决非线性任务中SVM[11-12]表现出了很好的效果。

本题的模型使用**高斯核函数-支持向量回归模型**(rbf-SVR）来作为非线性生物活性预测模型，设定样本数据为$D=(X，Y)$，其对应的模型 pIC_{50} 预测结果记为$f(x)$，使得其与y尽可能接近，与真实输出 y 的差别来计算损失。

假定$f(x)$与y之间最大偏差值为ε，w，b是待确定的参数。模型的理想状态下中，只有当$f(x)$与y完全相同时，其预测数量为完全正确。而当且仅当$f(x)$与y的差的绝对值大于ε时，才计算偏差度，此时相当于以$f(x)$为中心，构建一个宽度为2ε的预测圈，若预测数量落入此预测圈，则认为是被预测正确的。（预测圈内外的松弛程度可有所不同）。

因此，支持向量回归模型（SVR）可转化为（下式左部是正则化项）：

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}l(f(x_i)-y_i) \tag{5-5}$$

l为损失函数

$$l_\varepsilon(z)=\begin{cases}0, & if|z|\le\varepsilon\\ |z|-\varepsilon, & otherwise\end{cases} \tag{5-6}$$

因此引入松弛因子，重写第一个式子为：

$$\min_{w,b,\xi_i,\widehat{\xi_\iota}}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}(\xi_i+\widehat{\xi_\iota})$$

$$s.t.\begin{cases}f(x_i)-y_i\le\varepsilon+\xi_i\\ y_i-f(x_i)\le\varepsilon+\widehat{\xi_\iota}\\ \xi_i\ge0,\widehat{\xi_\iota}\ge0,i=1,2,\cdots,m\end{cases} \tag{5-7}$$

最后引入拉格朗日乘子，可得拉格朗日函数：

$$\begin{aligned}L\left(w,b,\alpha,\hat{\alpha},\xi_i,\widehat{\xi_\iota},\mu,\hat{\mu}\right)=&\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\left(\xi_i+\widehat{\xi_\iota}\right)-\sum_{i=1}^{m}\mu_i\xi_i-\sum_{i=1}^{m}\widehat{\mu_\iota}\widehat{\xi_\iota}\\&+\sum_{i=1}^{m}\alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)+\sum_{i=1}^{m}\widehat{\alpha_\iota}(y_i-f(x_i)-\varepsilon-\widehat{\xi_\iota})\end{aligned} \tag{5-8}$$

对四个遍历求偏导，另偏导数为零，可得：

$$\begin{cases} w = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) x_i \\ 0 = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) \\ C = \alpha_i + \mu_i \\ C = \hat{\alpha}_i + \hat{\mu}_i \end{cases} \tag{5-9}$$

把上边的式子带入，即可求得SVR的对偶问题

$$\max_{\alpha,\hat{\alpha}} \sum_{i=1}^{m} y_i (\hat{\alpha}_i - \alpha_i) - \varepsilon(\hat{\alpha}_i + \alpha_i) - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i)\left(\hat{\alpha}_j - \alpha_j\right) x_i^T x_j \tag{5-10}$$

$$s.t. \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) = 0, 0 \le \alpha_i, \hat{\alpha}_i \le C$$

上式的过程需要满足拉格朗日乘子条件，即

$$\begin{cases} \alpha_i (f(x_i) - y_i - \varepsilon - \xi_i) = 0 \\ \hat{\alpha}_i \left(y_i - f(x_i) - \varepsilon - \hat{\xi}_i\right) = 0 \\ \alpha_i, \hat{\alpha}_i = 0, \xi_i, \hat{\xi}_i = 0 \\ (C - \alpha_i)\xi_i = 0, (C - \hat{\alpha}_i)\hat{\xi}_i = 0 \end{cases} \tag{5-11}$$

最后，可得SVR的解为

$$f(x) = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) x_i^T x + b \tag{5-12}$$

其中b为

$$b = y_i + \varepsilon - \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) x_i^T x \tag{5-13}$$

(3)随机森林回归模型的建立

本题使用随机森林回归模型作为第三个非线性生物活性预测模型。随机森林基本思想是利用 bootstrap 重抽样方法从原始样本中抽取多个样本，对每个 bootstrap 样本构建决策树，然后将所有决策树预测平均值作为最终预测结果[13-14]。随机森林回归可以看成是由很多弱预测器（决策树）集成的强预测器。

本题实现的RFR是将多个二叉决策树打包组合而成的，训练RFR便是训练多个二叉决策树。在训练二叉决策树模型的时候需要考虑怎样选择切分变量、切分点以及怎样衡量一个切分变量、切分点的好坏。针对于切分变量和切分点的选择，采用穷举法，即遍历每个特征和每个特征的所有取值，最后从中找出最好的切分变量和切分点；针对于切分变量和切分点的好坏，一般以切分后节点的不纯度来衡量，即各个子节点不纯度的加权和$G(x_i, v_{ij})$，其计算公式如下：

$$G(x_i, v_{ij}) = \frac{n_{left}}{N_s} H(X_{left}) + \frac{n_{right}}{N_s} H\left(X_{right}\right) \tag{5-14}$$

其中，x_i为某一个切分变量，v_{ij}为切分变量的一个切分值，n_{left}，n_{right}，N_s分别为切分后左子节点的训练样本个数、右子节点的训练样本个数以及当前节点所有训练样本个数，X_{left}，X_{right}分为左右子节点的训练样本集合，$H(X)$为衡量节点不纯度的函数，在本题中**选用MSE作为模型的不纯度函数**。

(4)梯度提升回归树模型的建立

本题使用梯度提升回归树模型作为第四个非线性生物活性预测模型。GBDT模型是Boosting算法的一种，也是Boosting算法的一种改进。他传统的Boosting有着很大的区别，GBDT的核心就在于每一次计算都是为了减少上一次的残差（Residual），而为了减少这些残差，可以在残差减少的梯度（Gradient）方向上建立一个新模型。在GBDT中，每个新模型的建立是为了使得先前模型残差往梯度方向减少，与传统Boosting算法对正确、错误的样本进行加权有着极大的区别。

使用梯度提升回归树建立预测模型

针对梯度提升回归，首先要输入训练集样本：

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m),\} \tag{5-15}$$

其中，最大迭代次数T，损失函数L。

初始化弱学习器：

$$f_0(x) = argmin \sum_{i=1}^{m} L(y_i, c) \tag{5-16}$$

在迭代轮数从 1 到T的过程中，对样本$i = 1，2 \ldots m$，计算负梯度：

$$r_{ti} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{t-1}(x)} \tag{5-17}$$

利用$(x_i, r_{ti})(i = 1，2. \ldots，m)$可以拟合一棵CART回归树，得到了第$t$棵回归树，其对应的叶子节点区域$R_{tj}, j = 1,2, \cdots, J$。其中$J$为回归树叶子节点的个数。

针对每一个叶子节点里的样本，求出使损失函数最小，拟合叶子节点最好的输出值c_{tj}值如下：

$$c_{tj} = argmin \sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x_i) + c) \tag{5-18}$$

更新强学习器：

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I(x \in R_{tj}) \tag{5-19}$$

从而得到最终的强学习器表达式：

$$f(x) = f_0(x) + \sum_{t=1}^{T} \sum_{j=1}^{J} c_{tj} I(x \in R_{tj}) \tag{5-20}$$

(5) XGBoost 回归模型的建立

本题使用 XGBoost 回归模型作为第五个非线性生物活性预测模型。XGBoost 模型的提出解决了梯度提升回归树遇到大而复杂的数据集，运行一次往往会占用巨大计算资源的

问题[15]。

XGBoost的目标函数由损失函数和正则化项两部分和一个常数项组成，公式如下：

$$Obj(\theta) = L(\theta) + \Omega(\theta) + C \tag{5-21}$$

其中损失函数用于衡量模型预测的好坏，而正则化项用于控制模型的复杂度，避免过拟合。

XGBoost的建模过程是在保留原有模型不变的基础上，将上一次预测产生的误差作为参考进行下一棵树的建立。也就是说，它将预测值与真实值的残差作为下一棵树的输入，其加法过程有如下表示：

初始化

$$\hat{y}_i^{(0)} = 0 \tag{5-22}$$

往模型中加入第 1 棵树：

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{5-23}$$

往模型中加入第 2 棵树：

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{5-24}$$

往模型中加入第t棵树：

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5-25}$$

其中 $\hat{y}_i^{(t)}$表示样本i个样本在第t次的预测值，它保留了$t-1$c 次的模型预测结果，之后加入一个新函数$f_t(x_i)$，每一轮加入的新函数可以最大程度的使损失函数减到最小，此时损失函数为：

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \tag{5-26}$$

其中I_j为第j个叶子节点上的样本，w_j为第j个叶子节点的权重，令：

$$G_j = \sum_{i \in I_j} g_i \, , H_j = \sum_{i \in I_j} h_i \tag{5-27}$$

将上式代入并且对w_j求偏导，得到最优权重：

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{5-28}$$

此时，可以得到最优目标函数：

$$Obj(\theta) = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j}{H_j + \lambda} + \Upsilon T \tag{5-29}$$

目标函数值越小，预测误差也越小，则模型效果越好。

5.3 模型评估方法

本题使用采用 MSE、MAE、R^2来比较并选取最优的生物活性预测模型。

(1)均方误差：MSE

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{5-30}$$

其中，y_i和$\hat{y}_i$分别为测试集上的真实值和预测值。

(2)平均绝对误差：MAE

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{5-31}$$

(3)拟合优度：R^2

$$R^2 = 1 - \frac{\sum_i(\hat{y}_i - y_i)^2}{\sum_i(\overline{y}_i - y_i)^2} \tag{5-32}$$

5.4 模型的求解

5.4.1 最优特征子集的选取

本题使用 Python 编程求出对生物活性相关性最高的 20 个变量两两之间的距离相关系数，该系数可以反应变量之间的相关程度，距离相关系数越大，表示两者相关性越强。求出的 20 个变量两两之间的距离相关系数如图 5.3 所示。

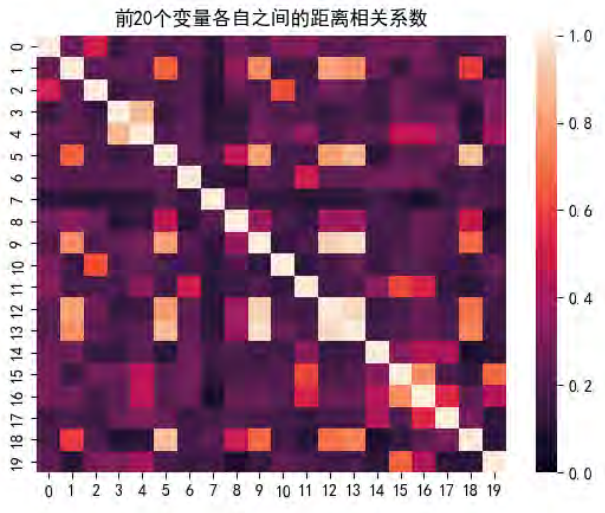


图 5.3 生物活性相关度最高的 20 个变量独立性检验

从图上颜色可以看出前 20 名的某些变量之间相关性达到了极强相关的定义标准，因此，直接选用前 20 名变量的组合并不能最大程度的得到最大量的有用信息，本题还需要重新去选出一个不超过 20 个变量的特征子集。

本题使用类似非极大值抑制的方式，从 30 个变量中剔除了 11 个变量，使得最终变量之间的相关性小于 0.6。用于建模的特征子集如下表 5.4 所示。

表 5.4 距离相关系数得出的最终变量表

序号	分子描述符
1	ATSc2
2	ATSc3
3	ATSc5
4	BCUTc-11
5	BCUTc-1h
6	VC-5
7	SdssC
8	minHBa
9	minHBint5
10	minHBint10
11	minsssN
12	maxssO
13	MAXDP
14	ETA_Shape_Y
15	MDEC-23
16	MDEC-33
17	MLFER_A
18	WTPT-4
19	XLogP

再次检验 19 个变量两两之间的相关性，如图 5.4 所示。

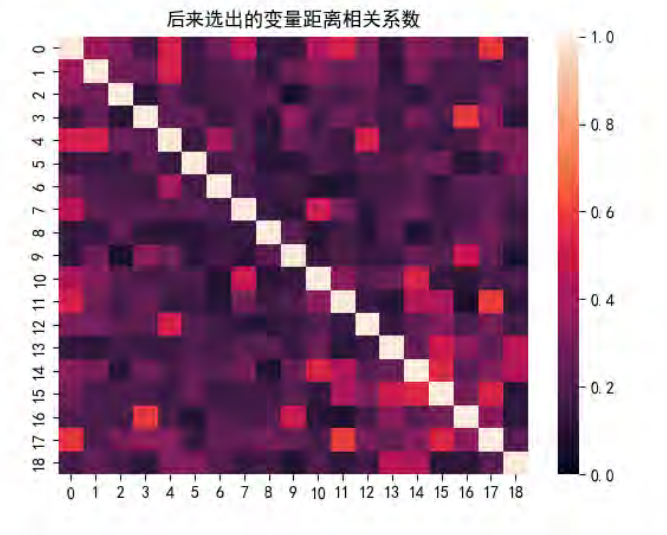


图 5.4 最终的 19 个变量独立性检验

从图中能很明显看出颜色加深，说明筛选后变量之间的独立性较高。

5.4.2 数据 Z-score 标准化

进行数据标准化的原因主要有以下两点：(1)将不同量级的数据统一转化为同一个量

级，保证数据之间的可比性。(2)将数据拉回成均值为 0，标准差为 1 的数据有利于回归模型的收敛。

Z-score 标准化的公式如下所示：

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{5-33}$$

5.4.3 数据划分

按 8:2 的比例将 1974 行数据划分成训练集和测试集，用训练集训练模型，再用训练好的模型在测试集上验证效果。

5.4.4 模型训练及比较

本题使用 Python 编程分别建立 5 个生物活性预测模型，使用训练集训练模型，并在测试集上验证结果。再通过 MAE,MSE, R^2三个评价指标来反映各模型的好坏，结果如下图 5.5 和图 5.6 所示。

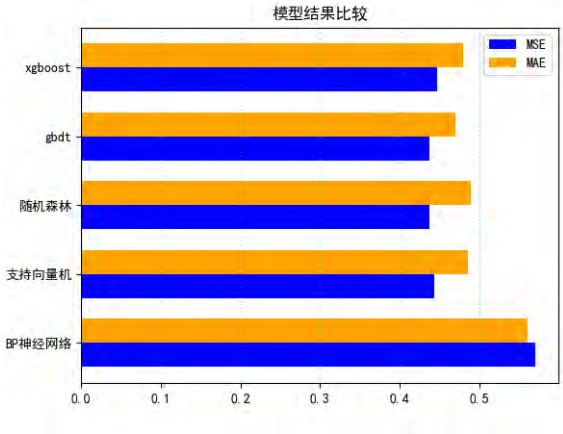


图 5.5 模型 MSE、MAE 比较

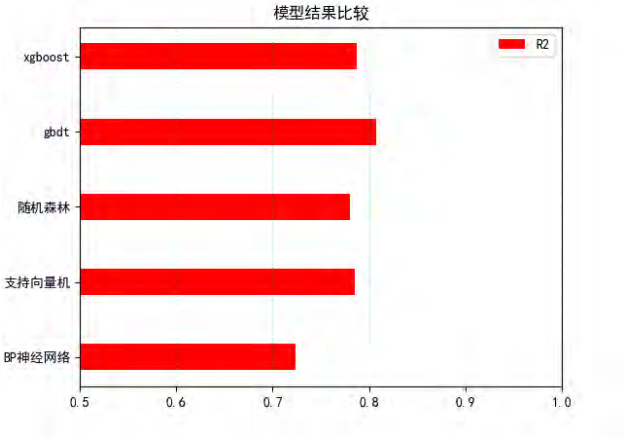


图 5.6 模型R^2比较

从图中可知**梯度提升回归树**的的MSE指标和MAE指标最小，分别为0.4376和0.4704，R^2指标最高为0.8076，所以最终选取梯度提升回归树作为生物活性 pIC_{50} 预测模型。

5.4.5 模型训练结果可视化

梯度提升回归树模型在测试集上的预测效果如图 5.7 所示，将预测值相连，更能直观的反映出预测值和实际值之间的差距。

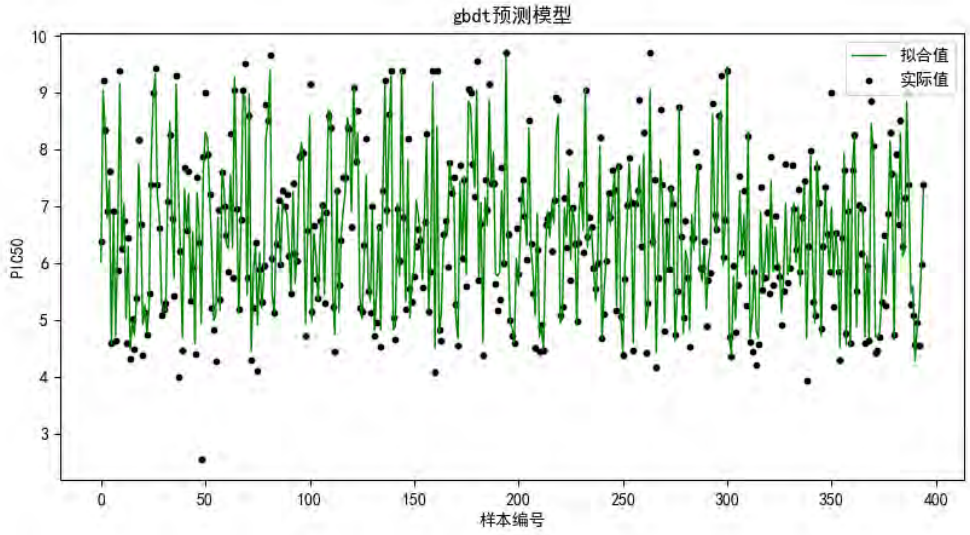


图 5.7 梯度提升回归树在测试集上的测试效果

5.4.6 梯度提升回归树的模型参数

本题使用网格搜索迭代选取梯度提升回归树的最优参数，在本题中各个模型的主要参数设置如表 5.5 所示。

表 5.5 最终模型梯度提升回归树的参数设置

序号	参数名	参数值
1	损失函数	平均绝对误差
2	学习率	0.01
3	决策树个数	2000
4	树最大深度	10
5	叶节点所需的最小样本数	1
6	分割所需的最小样本数	2

5.4.7 预测 test 表中的化合物

本题使用梯度提升回归树模型对 test 表中 50 个化合物进行预测，test 表中部分化合物生物活性预测结果如表 5.6 所示，编号顺序与 test 表中一致。（完整结果见**附件问题二** test.xlsx 和**附录表一**）。

本题只建立 pIC_{50} 生物活性的预测模型，对于 IC_{50} 的求解，通过以下公式进行换算。

$$IC50 = 10^{(9-\mathrm{pIC_{50}})} \tag{5-34}$$

表 5.6 test 表中部分化合物生物活性预测结果

序号	IC_{50}_nM	pIC_{50}
1	26.85868	7.570915
2	68.98395	7.161252
3	55.91154	7.252499
4	36.28374	7.440288

5	14.73488	7.831654
⋮	⋮	⋮
46	46.83408	7.329438
47	57.34258	7.241523
48	71.94671	7.142989
49	153.6082	6.813586
50	53.15635	7.274445

5.5 小结与讨论

本题生物活性预测模型不够精确，可能与数据集的样本数量太少有关，本题所建模型可以作为一项参考，实际应用时，还需要加大样本数量来进行训练提高生物活性的预测精度。

六、问题三模型的建立与求解

6.1 问题分析

本题的任务为：利用 729 个分子描述符，针对 1974 个化合物的 ADMET 数据，分别构建化合物的 Caco-2、CYP3A4、hERG、HOB、MN 性质的分类预测模型，然后使用所构建的 5 个分类预测模型对 test 表中的 50 个化合物进行相应的预测。

本题主要有以下两个**任务**：

任务一：由于不同的 ADMET 性质存在差异，不能使用同一组特征子集来预测 5 种性质的好坏，各个性质之间需要单独进行特征子集的选择。

任务二：因为输入特征子集不同，所以需要对每一个性质都单独建立一个二分类预测模型。

6.2 各 ADMET 分类预测模型的建立

本题模型建立及求解的流程图如图 6.1 所示，首先使用问题一低方差滤波去除单一值特征后的数据，分别对每个 ADMET 性质做最优特征子集的选取。以性质 Caco-2 为例，使用最大互信息系数筛选出前 200 个与目标关联性最高的特征变量，接着分别对这些变量使用基于随机森林的递归特征消除算法，将算法试验 50 次，对每次试验选出的变量进行计数，选出出现频数大于 40 次的变量，总计选出 31 个变量。之后对这 31 个变量进行随机森林回归试验，得到变量关于性质 Caco-2 的相关性排序，最后使用第二问提出的基于距离相关系数的变量独立性剔除算法对剩余的 31 个变量进行处理，得到最终 13 个变量的特征子集，CYP3A4、hERG、HOB、MN 性质的特征子集选取方式和 Caco-2 相同。

选出各 ADMET 性质的特征子集之后，对每个性质建立分类预测模型，使用和问题二相同的方式，分别建立五种分类预测模型：BP、支持向量机、随机森林、梯度提升回归树、XGBoost，使用分类准确率来对各个模型进行评价分析，选出各个 ADMET 性质的最终分类预测模型。

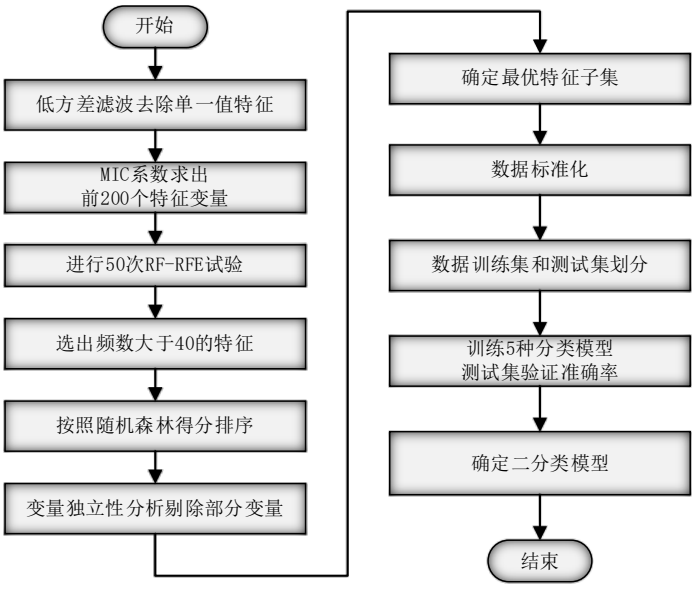


图 6.1 问题三模型建立思路流程图

6.3 分类模型的评估方法

本题使用模型在测试集上的准确率来评价模型的好坏，具体的计算公式如下：

$$准确率 = \frac{测试集预测正确数}{测试样本总数} \times 100\% \tag{6-1}$$

6.4 模型的求解

6.4.1 MIC 相关性分析

本题对剩余的 504 个变量使用 Python 编程实现 MIC 相关性分析，选出前 200 个与 ADMET 性质关联性最高的候选变量，以 Caco-2 为例，其结果如下表 6.1 所示。

表 6.1 性质 Caco-2 的分子描述符 MIC 得分

序号	分子描述符	灰色关联系数得分
1	WTPT-1	0.5608
2	WPATH	0.5318
3	ETA_Eta_R_L	0.5274
4	SP-1	0.5270
5	ECCEN	0.5161
6	MLFER_L	0.5106
7	SP-2	0.5067
8	MW	0.5041
9	SP-0	0.4983
10	CrippenMR	0.4968
⋮	⋮	⋮
198	maxHBint6	0.1917
199	MDEO-11	0.1911
200	nRing	0.1909
201	nHBAcc2	0.1866
202	ETA_Epsilon_5	0.1862
⋮	⋮	⋮
500	SssssNp	0.0004
501	SsSH	0.0004
502	mindssS	0.0004
503	nddssS	0.0001
504	n3Ring	0.0000

从表中可知，排名 200 名之后的分子描述符 MIC 得分已经小于 0.2，按照统计学的定义，变量与性质 Caco-2 成弱相关的关系，从而可以确定对预测性质 Caco-2 有效的变量在 200 名之前。

6.4.2 基于随机森林的递归特征消除试验

因为递归特征消除的特征选择结果具有一定的随机性，所以本题对 MIC 分析得到的前

200 名的特征做 50 次递归特征消除算法试验，每次试验最终输出 40 个特征变量，对每次试验出现的特征进行计数，选取出现频数高于 40 的特征。性质 Caco-2 的各变量出现频数统计如下图 6.2 所示，总计选出 31 个频数高于 40 的特征。

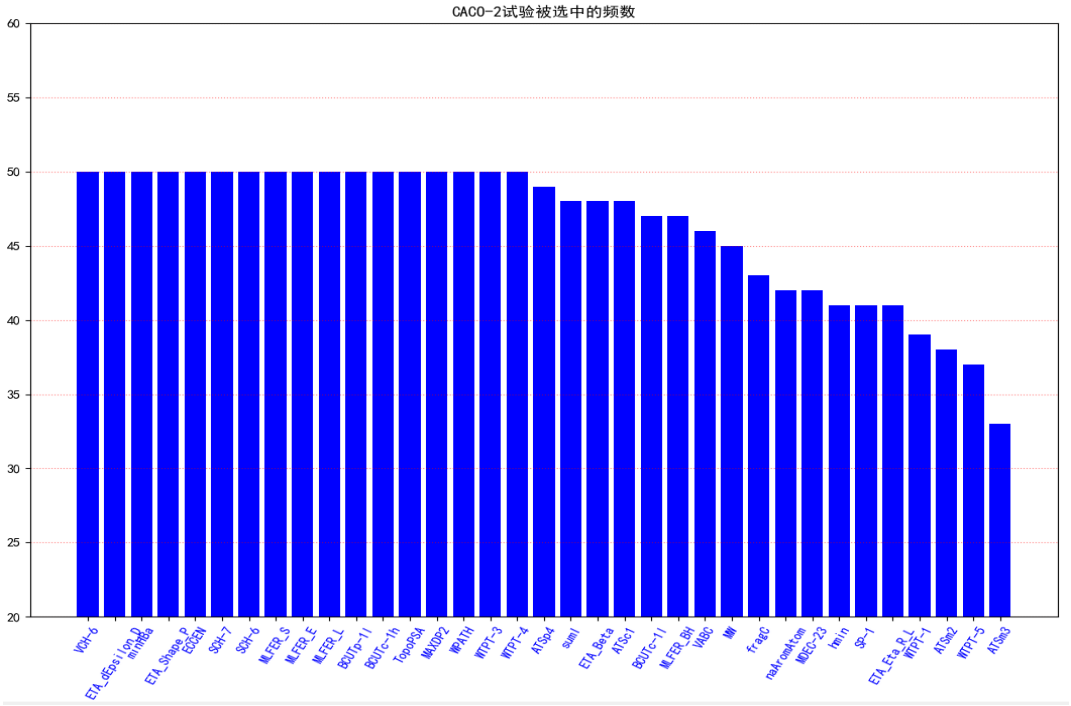


图 6.2 50 次递归特征消除试验各变量的频数统计

6.4.3 距离相关系数独立性分析

以 Caco-2 为例，在上一步骤中选出的 31 个特征只有频数统计，并没有得到他们与性质 Caco-2 的相关性得分。为了方便后续进行关于独立性的变量剔除，这里使用随机森林回归对 31 个变量进行得分排序，为了减小算法随机性的影响，本题进行了 10 次试验，取 10 次得分平均值。Caco-2 的 31 个特征变量随机森林回归得分平均值如下表 6.2 所示。

表 6.2 Caco-2 递归特征消除试验选出的特征变量

序号	分子描述符	Caco-2 相关性得分排序
1	ECCEN	0.0940
2	WPATH	0.0910
3	SP-1	0.0613
4	ETA_Eta_R_L	0.0601
5	MLFER_L	0.0572
6	MW	0.0548
7	ETA_Beta	0.0411
8	VABC	0.0386
9	sumI	0.0331
10	MLFER_S	0.0294
11	SCH-6	0.0270
12	TopoPSA	0.0269

13	MDEC-23	0.0252
14	MLFER_E	0.0251
15	naAromAtom	0.0243
16	MLFER_BH	0.0238
17	VCH-6	0.0226
18	WTPT-4	0.0218
19	BCUTc-1h	0.0207
20	SCH-7	0.0205
21	ETA_Shape_P	0.0204
22	ATSp4	0.0200
23	fragC	0.0199
24	ATSc1	0.0199
25	ETA_dEpsilon_D	0.0197
26	WTPT-3	0.0186
27	minHBa	0.0183
28	MAXDP2	0.0173
29	hmin	0.0164
30	BCUTc-1l	0.0158
31	BCUTp-1l	0.0154

得到得分排序后，本题使用问题二中提出的变量独立性剔除算法，以 Caco-2 为例，对 31 个变量进行变量剔除，得到最终的特征子集。

以 Caco-2 为例，算法总计剔除了 18 个变量，剔除前和剔除后的变量间距离相关系数图如图 6.3 和图 6.4 所示。由颜色可知，变量间的独立性得到提升。

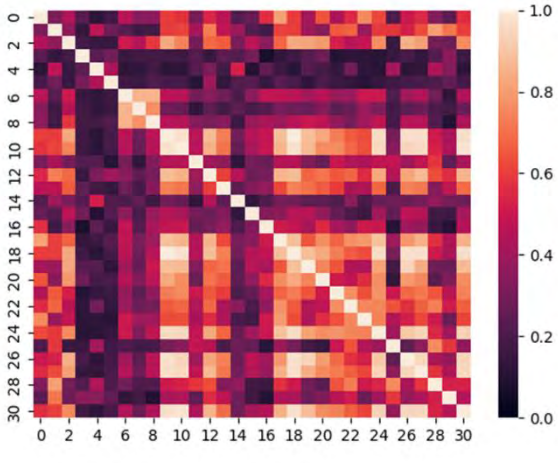


图 6.3 剔除前 Caco-2 变量间的相关系数

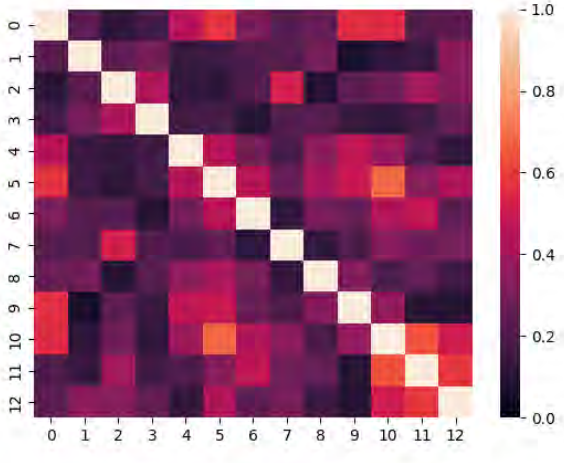


图 6.4 剔除后 Caco-2 变量间的相关系数

最终得到的各个 ADMET 性质的特征子集情况如表 6.3 所示。

表 6.3 各 ADMET 性质最终的特征子集

序号	Caco-2	CYP3A4	hERG	HOB	MN
1	naAromAtom	ATSm1	BCUTc-1l	ATSc3	ATSc2
2	BCUTc-1l	BCUTc-1h	ECCEN	BCUTc-1l	BCUTc-1h
3	BCUTc-1h	SCH-6	SaasC	BCUTc-1h	SCH-7
4	BCUTp-1l	SP-4	minHBd	BCUTp-1l	SsOH
5	SCH-6	minHBa	minHBa	SCH-6	minHBa
6	ECCEN	maxHBd	minsssN	SC-5	mindssC
7	minHBa	ETA_dEpsilon_D	ETA_dEpsilon_B	VP-6	maxsCH3
8	MAXDP2	ETA_dBetaP	ETA_dEpsilon_D	SHsOH	ETA_dEpsilon_C
9	ETA_dEpsilon_D	MLFER_BO	TopoPSA	SdO	ETA_BetaP
10	ETA_Shape_P	WTPT-3	WTPT-4	minHBa	ETA_BetaP_s
11	MLFER_S	WTPT-4	-	minaasC	ETA_Eta_F_L
12	TopoPSA	-	-	hmin	ETA_EtaP_B_RC
13	WTPT-4	-	-	MAXDP2	FMF
14	-	-	-	Kier3	MLFER_BH
15	-	-	-	MDEC-23	TopoPSA
16	-	-	-	MDEC-33	WTPT-4
17	-	-	-	MLFER_BO	WTPT-5
18	-	-	-	WTPT-3	-
19	-	-	-	WTPT-4	-

6.4.4 模型训练前数据处理

本题首先对数据集进行划分，按 8：2 的比例分成训练集和测试集，接着对数据进行 Z-score 标准化，将数据拉成均值为 0，标准差为 1 的数据，保证数据之间可比性的同时提高模型训练的收敛速度。

6.4.5 模型训练及比较

本题对 ADMET 的五个性质分别使用五种模型去做分类预测，通过各个模型在测试集上的准确率比较来选取各 ADMET 性质的最优分类预测模型，具体结果如图 6.5 至图 6.9 所示。

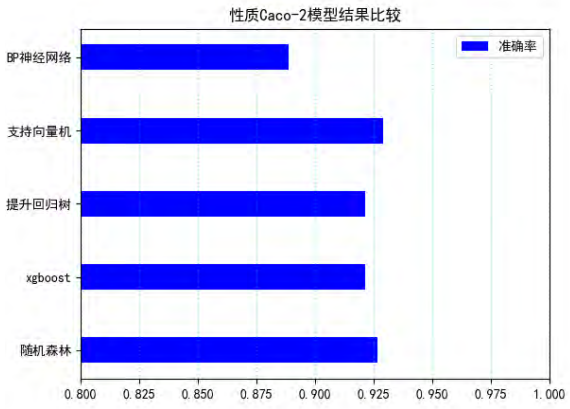


图 6.5 Caco-2 各模型准确率比较

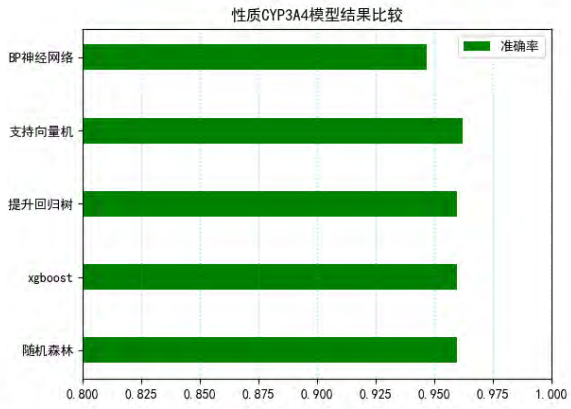


图 6.6 CYP3A4 各模型准确率比较

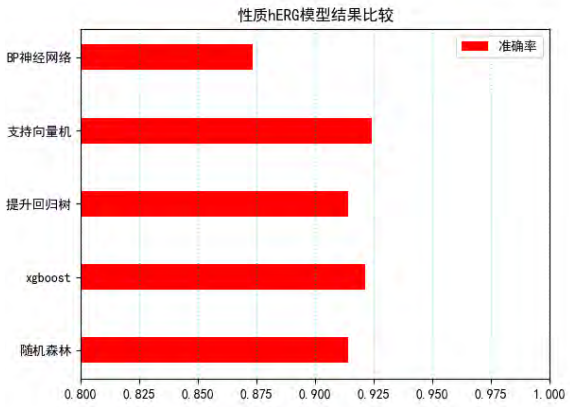


图 6.7 hERG 各模型准确率比较

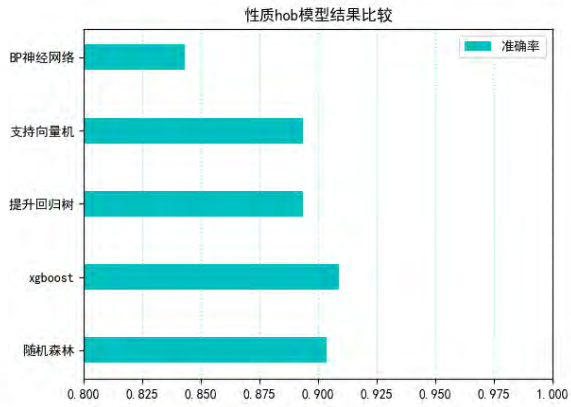


图 6.8 HOB 各模型准确率比较

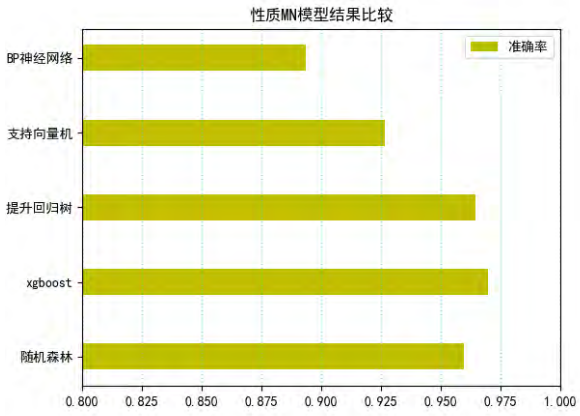


图 6.9 MN 各模型准确率比较

由上面五张模型结果比较图可知，Caco-2，CYP3A4，hERG 的**支持向量机**分类模型在测试集上的准确率最高，分别为 **92.92%，96.25%，92.44%**。HOB 和 MN 性质的 XGBoost 分类模型得分最高，分别为 **90.88%，96.96%**。

本题最终的分类模型选择如下表 6.4 所示。

表 6.4 各 ADMET 性质分类模型选择情况

序号	ADMET 性质	模型选择情况
1	Caco-2	SVM
2	CYP3A4	SVM
3	hERG	SVM
4	HOB	XGBoost
5	MN	XGBoost

6.4.6 SVM 模型和 XGBoost 模型参数

本题使用网格搜索迭代选取 SVM 分类模型和 XGBoost 分类模型的最优参数，在本题中各个模型的主要参数设置如下表 6.5 和表 6.6 所示。

表 6.5 SVM 最优参数

序号	参数名	参数值
1	惩罚系数 C	1.1
2	核函数	rbf

表 6.6 XGBoost 最优参数

序号	参数名	参数值
1	树的个数	10000
2	学习率	0.01

6.4.7 预测 test 表中的 50 个化合物

本题使用选定的模型对 test 表中 50 个化合物的 ADMET 性质进行预测，test 表中部分结果如表 6.7 所示，编号顺序与 test 表中一致。（完整结果见**附件问题三 test.xlsx** 和**附录表二**）。

表 6.7 test 表中部分化合物 ADMET 性质预测结果

序号	Caco-2	CYP3A4	hERG	HOB	MN
1	0	1	1	0	1
2	0	1	0	0	1
3	0	1	0	0	1
4	0	1	0	0	1
5	0	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
46	0	1	1	0	1
47	0	1	1	0	1
48	0	1	1	0	1
49	0	1	1	0	1
50	0	1	1	0	0

6.4.8 小结与讨论

各个 ADMET 性质分类预测模型在测试集上的准确率较优，都超过了 90%，但也存在一定的问题，例如数据的不平衡问题，图 6.10 至图 6.13 是给出的 1974 个化合物各个性质的样本类别情况，从图中可知样本的类别存在数据不平衡的现象，由于样本数量有限，该问题无法解决。实际应用时，应当继续加大样本容量，缓解数据不平衡问题带来的影响。

图 6.10 Caco-2 样本类别图

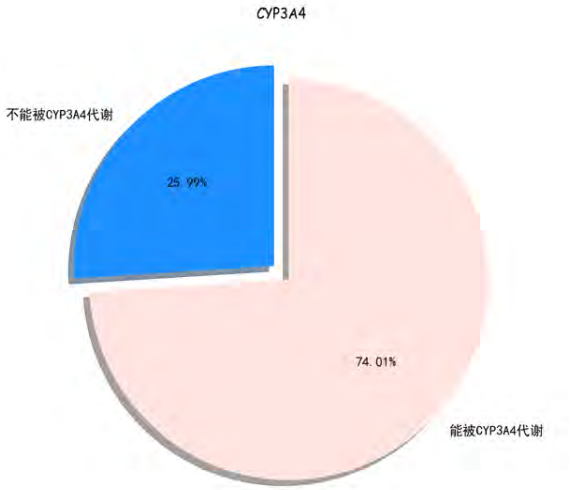


图 6.11 CYP3A4 样本类别图

图 6.12 HOB 样本类别图

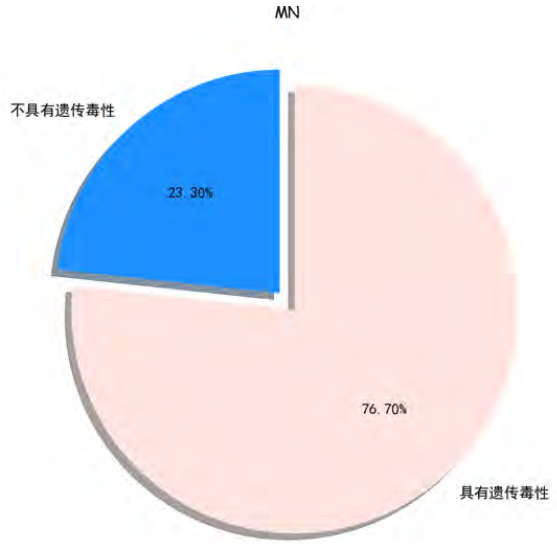


图 6.13 MN 样本类

七、问题四模型的建立与求解

7.1 问题分析

本题的任务：寻找并阐述化合物的哪些分子描述符，以及这些分子描述符在什么取值或者处于什么取值范围时，能够使化合物对抑制 ERα 具有更好的生物活性，同时具有更好的 ADMET 性质（给定的五个 ADMET 性质中，至少三个性质较好）。

本题主要有以下**难点**：

pIC_{50} 的预测模型以及五个 ADMET 的二分类模型为非线性模型，传统的优化算法仅能求出问题的局部最优解，并且求解的结果强烈依赖于初始值。对于此**难点**，本题采用差分进化算法(Differential Evolution，DE)，来进行优化求解，此算法相比传统算法求出全局最优解的可能性更大，但其收敛速度较慢，计算复杂度较高，耗时较长。

问题四模型分析与求解的思路流程图如图 7.1 所示。

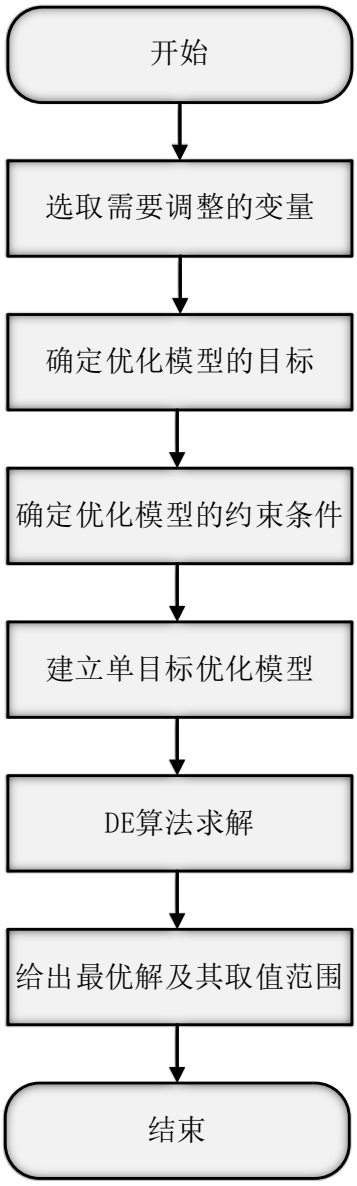


图 7.1 问题四流程图

7.2 筛选样本数据，分析主要变量分布

基于给定的五个 ADMET 性质中至少三个性质较好的约束条件，对 1974 个样本进行筛选，研究化合物满足约束条件时，6 个模型中选取的变量（共 56 个）的主要特征。经过筛选有 1342 个样本超出约束条件，632 个样本满足约束条件。对这 56 个变量的分布进行对比，如图 7.2 所示。

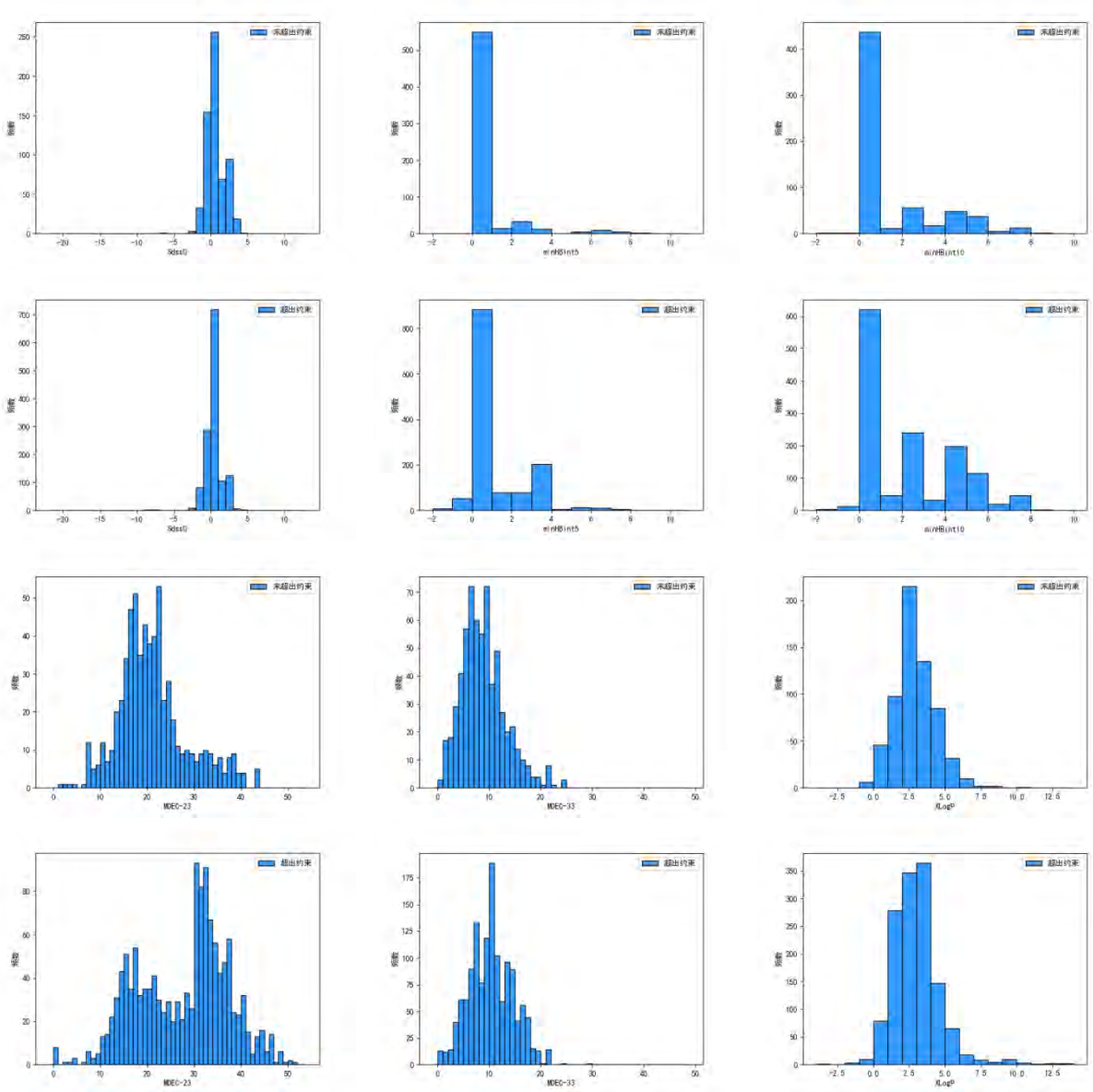


图 7.2 变量样本数据分布（部分）

从 56 个变量未超出约束和超出约束两类数据的分布图来看，总体较为集中，因此不能通过单一调整某一参数，本题在此优化问题中，选择对这 56 个变量同时优化，从变量的分布来看，可以总结以下两点规律：

（1）部分变量符合标准的正态分布，对其操作调整达到最优解的过程中，符合条件的样本量占比可能更高。

（2）部分变量分布存在极端值，说明对目标优化时，可能很快搜索到最优解。

本题选择调整的 56 个变量及其在优化模型中对应的变量名称如表 7.1 所示。

表 7.1 56 个变量名称及其模型中的变量名

序号	特征名称	模型中的变量名
1	ATSc2	x_0
2	ATSc3	x_1
3	ATSc5	x_2
4	ATSm1	x_{28}
5	BCUTc-1h	x_4
6	BCUTc-1l	x_3
7	BCUTp-1l	x_{20}
8	ECCEN	x_{22}
9	ETA_BetaP	x_{49}
10	ETA_BetaP_s	x_{50}
11	ETA_dBetaP	x_{31}
12	ETA_dEpsilon_B	x_{36}
13	ETA_dEpsilon_C	x_{48}
14	ETA_dEpsilon_D	x_{24}
15	ETA_Eta_F_L	x_{51}
16	ETA_EtaP_B_RC	x_{52}
17	ETA_Shape_P	x_{25}
18	ETA_Shape_Y	x_{13}
19	FMF	x_{53}
20	hmin	x_{42}
21	Kier3	x_{43}
22	MAXDP	x_{12}
23	MAXDP2	x_{23}
24	maxHBd	x_{30}
25	maxsCH3	x_{47}
26	maxssO	x_{11}
27	MDEC-23	x_{14}
28	MDEC-33	x_{15}
29	minaasC	x_{41}
30	mindssC	x_{46}
31	minHBa	x_7
32	minHBd	x_{35}
33	minHBint10	x_9
34	minHBint5	x_8
35	minsssN	x_{10}
36	MLFER_A	x_{16}
37	MLFER_BH	x_{54}
38	MLFER_BO	x_{32}
39	MLFER_S	x_{26}
40	naAromAtom	x_{19}
41	SaasC	x_{34}
42	SC-5	x_{37}

43	SCH-6	$\boldsymbol{x_{21}}$
44	SCH-7	$\boldsymbol{x_{44}}$
45	SdO	$\boldsymbol{x_{40}}$
46	SdssC	$\boldsymbol{x_{6}}$
47	SHsOH	$\boldsymbol{x_{39}}$
48	SP-4	$\boldsymbol{x_{29}}$
49	SsOH	$\boldsymbol{x_{45}}$
50	TopoPSA	$\boldsymbol{x_{27}}$
51	VC-5	$\boldsymbol{x_{5}}$
52	VP-6	$\boldsymbol{x_{38}}$
53	WTPT-3	$\boldsymbol{x_{33}}$
54	WTPT-4	$\boldsymbol{x_{17}}$
55	WTPT-5	$\boldsymbol{x_{55}}$
56	XLogP	$\boldsymbol{x_{18}}$

7.3 建立目标优化准则

差分进化算法（DE）采用实数编码方式，其算法原理与遗传算法[16]十分相似，进化流程与遗传算法相同：变异、交叉和选择。DE 算法中的选择策略通常为锦标赛选择，而交叉操作方式与遗传算法也大体相同，但在变异操作方面使用差分策略，即利用种群中个体间的差分向量对个体进行扰动，实现个体变异。DE 的变异方式有效利用群体分布特性，提高算法的搜索能力，避免遗传算法中变异方式的不足[17-18]。

对于无约束优化问题：

$$\boldsymbol{minf(x_1, x_2, \cdots, x_D)}$$
$$\boldsymbol{s.t\ \ x_j^L \le x_j \le x_j^U, j = 1, 2, \cdots, D} \tag{7-1}$$

其中，D 是解空间的维数，x_j^L、x_j^U分别表示第j个分量x_j取值范围的上界和下界。

利用差分进化求解优化问题，主要分为初始化、变异、交叉和选择等几项操作。

1）初始化

先初始化种群:$\{X_i(0) | x_{ij}^L \le x_{i,j}(0) \le x_{i,j}^U; i = 1,2,\cdots,NP; j = 1,2,\cdots,D\}$

其中$x_i(0)$是第i个个体，j表示第j维。

$$\boldsymbol{x_{i,j}(0) = x_{ij}^L + rand(0,1)(x_{i,j}^L - x_{i,j}^U)} \tag{7-2}$$

其中，$x_{i,j}^L$、$x_{i,j}^U$分别表示第j维的上界和下界，$rand(0,1)$表示在区间[0,1]上的随机数。

2)变异

DE 算法通过差分策略实现个体变异，常见的差分策略是随机选取种群中两个不同的个体，将其向量差缩放后与待变异个体进行向量合成。

$$V_i(g+1) = X_{r1}(g) + F(X_{r2}(g) - X_{r3}(g)) \tag{7-3}$$

其中，$r1$，$r2$和$r3$是三个随机数，区间为[1, NP]，F称为缩放因子，为一个确定的常数。g表示第g代。

3)交叉

交叉操作的目的是随机选择个体，因为差分进化也是一种随机算法，交叉操作的方法是：

$$V_{i,j}(g+1)=\begin{cases}V_i(g+1) if rand(0,1)\le CR\\ x_{i,j}(g) otherwise\end{cases} \quad (7\text{-}4)$$

其中，CR称为交叉概率。通过概率的方式随机生成新的个体。

4)选择

在 DE 中采用的是贪婪选择的策略，即选择较优的个体作为新的个体。

$$X_i(g+1)=\begin{cases}U_i(g+1) & if\ \ f(U_i(g+1))\le f(X_i(g))\\ X_i(g) & otherwise\end{cases} \quad (7\text{-}5)$$

DE 算法流程如图 7.3 所示。

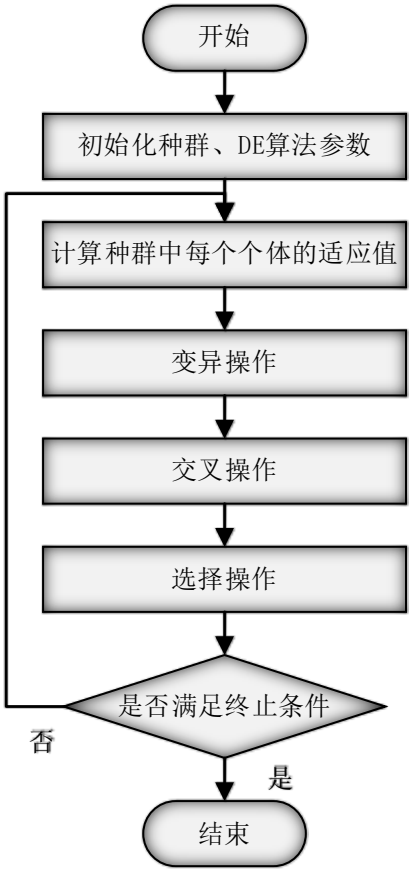


图 7.3 DE 算法流程图

7.4 单目标优化模型的建立

7.4.1 优化目标

为了治疗乳腺癌，本题针对乳腺癌治疗靶标ERα，提出了化合物抑制ERα的生物活性指标 pIC_{50}，因此，使化合物对抑制ERα具有更好的生物活性是本题的核心目标。

在满足化合物具有较好的 ADMET 性质的前提下(给定的五个 ADMET 性质中，至少三个性质较好)，同时调整 56 个主要变量，使得化合物的 pIC_{50} 指标尽可能大，优化目标公式如下：

$$max(pIC50_{pred}) \quad (7\text{-}6)$$

7.4.2 约束分析

（1）主要变量取值范围约束

由于化合物的每个分子描述符性质都有上下限，但题目中未给出明确的上下限，本题使用 1974 个样本中数据变量的最大最小值作为我们各个决策变量的上下界。因此，存在约束如下式所示。

$$x_{imin} \le x_i \le x_{imax}, i = 1,2,\cdots,56 \tag{7-7}$$

（2）化合物 ADMET 性质约束

由于题目要求选出的化合物要保证五个 ADMET 性质中至少有三个性质较好。由于 Caco-2、CYP3A4 和 HOB 数值为 1 时代表其性质好，但 hERG 和 MN 数值为 0 时代表其性质好，所以本文对 hERG 和 MN 的预测值进行取反，使其与另外三个性质统一。因此，存在约束如下式所示。

$$Caco-2_{pred} + CYP3A4_{pred} + F(hERG_{pred}) + HOB_{pred} + F(MN_{pred}) \ge 3 \tag{7-8}$$

其中，$F(x) = |1-x|$，$Caco-2_{pred}$，$CYP3A4_{pred}$，$hERG_{pred}$，HOB_{pred}，MN_{pred}为 5 种 ADMET 性质分类预测模型。

7.4.3 模型建立

结合以上分析，在满足化合物具有 ADMET 五个性质中至少三个较好的前提下，以最大化抑制Erα生物活性为目标，建立单目标优化模型如下：

$$\begin{cases} \boldsymbol{max(pIC50_{pred})} \\ \mathbf{s.t} \begin{cases} \boldsymbol{x_{imin} \le x_i \le x_{imax}, i = 1,2,\cdots,56} \\ \boldsymbol{Caco-2_{pred} + CYP3A4_{pred} + F(hERG_{pred}) + HOB_{pred} + F(MN_{pred}) \ge 3} \end{cases} \end{cases} \tag{7-9}$$

其中,

$$\begin{gathered} F(x) = |1-x| \\ pIC50_{pred} = P(x_i | x_i \in A_1) \\ Caco-2_{pred} = g_1(x_i | x_i \in A_2) \\ CYP3A4_{pred} = g_2(x_i | x_i \in A_3) \\ hERG_{pred} = g_3(x_i | x_i \in A_4) \\ HOB_{pred} = g_4(x_i | x_i \in A_5) \\ MN_{pred} = g_5(x_i | x_i \in A_6) \\ X = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6 \end{gathered} \tag{7-10}$$

A_i为的第i个预测模型的特征子集，i=1，2，…，6，

X为 56 个决策变量的集合

7.5 操作方案模型的求解

本题使用 Python 编程求解，见附件“1_6 个模型选择变量名提取.py”、“2_DE 优化.py”，其中 DE 算法各个参数配置如表 7.2 所示。

表 7.2 DE 算法参数设置

序号	参数名	参数值
1	决策变量维度	56
2	种群数	50
3	最大迭代次数	200
4	交叉和变异的比例	0.3
5	F	0.5

经过200轮迭代，DE算法求得的变量参数值如表7.3所示，其pIC_{50}的值为9.53246618，Caco-2的值为1，CYP3A4的值为1，hERG的值为0，HOB的值为1，MN的值为1。所以其中Caco-2、CYP3A4、hERG、HOB的性质较好，满足问题四要求。

表7.3 变量优化结果

变量	参数名	参数值
x_0	ATSc2	-0.2366
x_1	ATSc3	-0.0534
x_2	ATSc5	0.5659
x_{28}	ATSm1	50.6330
x_4	BCUTc-1h	0.2131
x_3	BCUTc-1l	-0.3396
x_{20}	BCUTp-1l	4.4115
x_{22}	ECCEN	1248.1217
x_{49}	ETA_BetaP	0.9559
x_{50}	ETA_BetaP_s	0.6720
x_{31}	ETA_dBetaP	0.0293
x_{36}	ETA_dEpsilon_B	0.0919
x_{48}	ETA_dEpsilon_C	-0.0731
x_{24}	ETA_dEpsilon_D	0.1071
x_{51}	ETA_Eta_F_L	14.6579
x_{52}	ETA_EtaP_B_RC	0.0472
x_{25}	ETA_Shape_P	0.0783
x_{13}	ETA_Shape_Y	0.4107
x_{53}	FMF	0.4062
x_{42}	hmin	0.3337
x_{43}	Kier3	43.2709
x_{12}	MAXDP	6.7600
x_{23}	MAXDP2	4.7476
x_{30}	maxHBd	0.4659
x_{47}	maxsCH3	1.3541
x_{11}	maxssO	3.0340
x_{14}	MDEC-23	36.9639
x_{15}	MDEC-33	16.2308
x_{41}	minaasC	-0.3756
x_{46}	mindssC	0.1315
x_7	minHBa	-2.2623
x_{35}	minHBd	0.3435
x_9	minHBint10	1.8347

x_8	minHBint5	5.7022
x_{10}	minsssN	2.4483
x_{16}	MLFER_A	1.1742
x_{54}	MLFER_BH	3.6792
x_{32}	MLFER_BO	17.6601
x_{26}	MLFER_S	12.4127
x_{19}	naAromAtom	23.9095
x_{34}	SaasC	-3.8457
x_{37}	SC-5	1.0878
x_{21}	SCH-6	0.5161
x_{44}	SCH-7	0.6930
x_{40}	SdO	141.1725
x_6	SdssC	-16.4122
x_{39}	SHsOH	1.4751
x_{29}	SP-4	22.9705
x_{45}	SsOH	6.2158
x_{27}	TopoPSA	377.8563
x_5	VC-5	0.2349
x_{38}	VP-6	1.4772
x_{33}	WTPT-3	97.1778
x_{17}	WTPT-4	5.5879
x_{55}	WTPT-5	51.4780
x_{18}	XLogP	9.5161

7.6 模型合理性验证

本文对此 DE 单目标优化模型进行了 8 次迭代，得到的 8 个解如表 7.4 所示。

表 7.4 算法 8 次迭代结果

pIC_{50}	Caco-2	CYP3A4	hERG	HOB	MN
9.5325	1	1	0	1	1
9.4922	0	1	0	1	1
9.5537	1	1	0	0	1
9.4479	1	1	1	1	0
9.5033	0	1	0	1	0
9.3566	0	1	1	1	0
9.4129	0	1	0	1	0
9.5138	1	1	1	0	0

由 8 次迭代的结果可知，目标值 pIC_{50} 的差异浮动最大值仅为 **2.06%**，且 8 次迭代的结果均满足 ADMET 性质有三个以上较好的约束，**验证了模型的稳定性和合理性**。

7.7 小结与讨论

求解本题的过程中，本文主要依靠 DE 差分进化算法对此单目标优化模型求解，存在如下问题：

（1）由于差分进化算法本身计算复杂度较高,而本题选出的决策变量多达 56 个，导

致模型求解时间过长，收敛较慢；

（2）由于本次求解时间有限，对差分进化算法的种群个数和最大迭代次数设置的都比较小，而且 6 个模型都是非线性模型，所以最终求得的依然是局部最优解。

如果想要优化此求解模型，可以重点考虑以下几点：

（1）将样本具体划分类别，根据不同类别，收集其分子描述符性质及其抑制ERα活性和 ADMET 性质的多样本大数据信息，以供后续的数据挖掘与建模分析；

（2）获得更多样本后，可以对模型变量进一步进行特征选择与降维，以此来降低优化算法的复杂度，减少单次迭代所需时间，因此就可以设置更大的粒子群个数与更多的迭代次数，以尽可能求得全局最优解。

八、模型评价与改进

8.1 模型优点

1） 充分考虑了各变量与生物活性 pIC_{50} 的非线性关系，使用了灰色关联分析，随机森林回归，最大互信息系数，距离相关系数等适用于处理非线性特征的方法，能充分体现变量与生物活性之间的相关性。

2） 提出了类似非极大值抑制的独立性变量剔除方法，能充分保证变量与变量之间的独立性。

3） 在进行差分进化算法的设计过程种充分考虑了变量范围和 ADMET 性质，迭代搜寻的结果满足所有约束要求。

4）算法速度快，响应性好。

8.2 模型缺点

1）假设所有的分子描述符均可以独立改变，这可能与实际不符。

2）并没有考虑二分类训练样本中数据不平衡对模型带来的影响。

8.3 模型的改进与推广

1)本文提出的方法和模型可推广应用在其他候选药物的优化建模当中。

2)本文中特征选择处理方法的流程和建立分类回归预测模型的流程其他相关研究遭遇数据质 量问题时均可对应采纳、使用。

参考文献

[1]StewartB,WildC.WorldCancerReport2014:InternationalAgencyforResearchonCancer[M].Geneva:WorldHealthOrga－nization,2014.

[2]孙志华,邢晖,刘益巧,etal.乳腺癌靶向治疗的研究进展[J].生命科学研究,2017,21(03):275-82.

[3]陈万青,郑荣寿,张思维,etal.2012 年中国恶性肿瘤发病和死亡分析[J].中国肿瘤,2016,25(01):1-8.

[4]刘思峰, 灰色系统理论及其应用(第五版). 江苏省,南京航空航天大学,2010-10-27.

[5]邓聚龙．灰色系统基本方法［M］．武汉:华中工学院出版社，1987:17－34．

[6]PearsonK.Notesonthehistoryofcorrelation[J].Biometrika,1920,13:24-45.

[7]PearsonK.Mathematicalcontributionstothetheoryofevolution(II):Regression,heredity,andpanmixia[J].PhilosophicalTransactionsoftheRoyalSocietyofLondon.SeriesA,ContainingPapersofaMathematicalorPhysicalCharacter,1895,187:253-318.

[8]SzekelyG.J,RizzoM.L.,BakirovN.K.Measuringandtestingdependencebycorrelationofdistances[J].TheAnnalsofStatistics,2007,35(6):2769-2794.

[9]SzekelyG.J.,RizzoM.L.,BakirovN.K.Browniandistancecovariance[J].TheAnnalsofAppliedStatistics,2009,3(4):1236-1265.

[10]ReshefD.N.,ReshefY.A.,FinucaneH.K,GrossmanS.R.,McVeanG.,TurnbaughP.J.,LanderE.S,MitzenmacherM,SabetiP.C.Detectingnovelassociationsinlargedatasets[J].Science,2011,334(6062):1518-1524.

[11]邓乃杨，田英杰.数据挖掘中的新方法一支持向量机[M].北京.科学出版社，2005：28-35.

[12]徐维维，高凤.灰色算法在股票价格预测中的应用[J].计算机仿真，2007，24（11）：274-276

[13]BREIMAN L.Random forests[J].Machine Learming，2001，45（1）：5-32.

[14]方匡南.随机森林组合预测理论及其在金融中的应用[M].厦门：厦门大学出版社，2012.

[15]CHEN T Q, CARLOS G. XGBoost:A scalable tree boosting system [C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:785794.

[16]Onwubolu GC, Babu B V. New Optimization Techniques in Engi-neering. Berlin, Germany: Springer-Verlag,2004.

[17] Storn R, Price K. Differential evolution -A simple and efficient adaptive scheme for global optimization over continuous spaces [R]. Berkeley: University of California,2006.

[18]Yang Qiwen, Jiang Jingping, Qu Zhaoxia, et al. Improving Genetic Algorithms by Using Logic Operation. Control and Decision, 2002,15(4):510-512.

附录

代码文件说明：

文件名	说明
问题二 test.xlsx	第二问问题的答案
问题三 test.xlsx	第三问问题的答案
2_DE 优化.py	第四问题的代码
2_1 距离相关系数选择特征.py	第二问部分代码
1_3_1 灰色关联分析.py	第一问部分代码

附表一：

序号	IC_{50}_nM	pIC_{50}
1	26.85868	7.570915
2	68.98395	7.161252
3	55.91154	7.252499
4	36.28374	7.440288
5	14.73488	7.831654
6	68.24947	7.165901
7	43.99117	7.356635
8	39.35482	7.405002
9	33.24172	7.478317
10	33.90047	7.469794
11	32.4278	7.489082
12	46.35503	7.333903
13	28.09103	7.551432
14	33.3237	7.477247
15	27.72276	7.557163
16	21.5111	7.667337
17	48.61336	7.313244
18	217.3804	6.66278
19	70.14101	7.154028
20	14.29485	7.84482
21	74.87438	7.125667
22	208.5366	6.680818
23	381.1353	6.418921
24	204.3188	6.689692
25	348.6283	6.457637
26	371.2729	6.430307
27	95.47743	7.020099
28	476.0063	6.322387
29	397.2188	6.40097
30	1817.58	5.740507
31	9911.869	5.003844

32	8743.83	5.058298
33	10230.94	4.990085
34	10205.75	4.991155
35	16024.88	4.795205
36	210.6266	6.676487
37	186.9631	6.728244
38	274.5158	6.561433
39	234.0682	6.630658
40	263.6368	6.578994
41	248.6288	6.604449
42	248.6288	6.604449
43	226.3899	6.645143
44	336.1385	6.473482
45	248.6288	6.604449
46	46.83408	7.329438
47	57.34258	7.241523
48	71.94671	7.142989
49	153.6082	6.813586
50	53.15635	7.274445

附表二：

序号	Caco-2	CYP3A4	hERG	HOB	MN
1	0	1	1	0	1
2	0	1	0	0	1
3	0	1	0	0	1
4	0	1	0	0	1
5	0	1	0	0	1
6	0	1	0	0	1
7	0	1	1	0	0
8	0	1	0	0	1
9	0	1	0	0	1
10	0	1	1	0	1
11	0	1	1	0	1
12	0	1	1	0	1
13	0	1	1	0	1
14	0	1	1	0	1
15	0	1	1	0	1
16	0	1	1	0	1
17	0	1	1	0	1
18	0	1	0	0	1
19	0	1	0	0	1
20	0	0	0	0	1
21	0	1	0	0	0
22	0	1	0	0	1

23	1	0	0	1	0
24	1	1	0	0	0
25	1	1	1	0	0
26	1	1	1	0	0
27	0	1	1	0	0
28	0	1	1	0	1
29	0	1	1	0	1
30	1	1	1	0	1
31	1	1	1	1	1
32	1	1	1	1	1
33	1	1	1	1	1
34	1	1	1	1	1
35	0	1	1	1	1
36	0	1	0	0	1
37	0	1	0	0	1
38	0	1	1	0	0
39	0	1	0	0	1
40	0	1	0	0	1
41	0	1	0	0	1
42	0	1	0	0	1
43	0	1	0	0	1
44	0	1	0	0	1
45	0	1	0	0	1
46	0	1	1	0	1
47	0	1	1	0	1
48	0	1	1	0	1
49	0	1	1	0	1
50	0	1	1	0	0