抗乳腺癌候选药物的优化建模

摘　要

乳腺癌是女性癌症高发性恶性肿瘤，近年来发病率和死亡率逐年上升，严重危害了女性健康。如何通过科学的方法辅助专家高效研发抗乳腺癌药物对于人类健康发展具有重要意义。本文基于数据挖掘和机器学习方法，研究抗癌药物筛选的优化建模问题，旨在筛选出具有较好的生物活性化合物，同时在人体内具备良好的药代动力学性质和安全性。具体做法如下：

针对问题 1，需要筛选出 1974 个化合物的 729 个分子描述符中对生物活性影响最为显著的前 20 个变量。首先对数据进行预处理；然后建立集成式特征（分子描述符）重要性计算模型，筛选出特征重要性排名前 30 的特征；接着通过 Spearman 相关性分析剔除 10 个相关性系数较高的冗余特征，得到 20 个分子描述符；最后对所筛选的这 20 个特征进行独立性和代表性检验，最终获得影响生物活性的 20 个主要变量。

针对问题 2，构建化合物对 ERα 生物活性的定量回归预测模型。首先，以 pIC50 为因变量，第一问筛选的 20 个变量作为自变量，建立支持向量机、梯度提升和随机森林的回归预测模型，并用随机搜索方法搜索四种模型的最佳超参数。然后，用 MAE、RMSE 和拟合度 3 个评价指标对四种模型进行评价，并画出三种模型对测试集前 20 组数据实际值和预测值的拟合图，观察发现随机森林效果最佳。最后，用随机森林预测题目中给的 50 组化合物的 pIC50，并通过公式求解出 IC50_nM。

针对问题 3，分别构建五种化合物的分类预测模型。首先对数据进行预处理得到 361 个变量，五个化合物的二分类数据作为因变量。然后，构建五种化合物的三个二分类模型：支持向量机、随机梯度下降和随机森林。以 ROC、AUC 和准确率作为模型评价标准，用随机搜索对五个化合物分类模型分别进行最佳超参数搜索，最终发现随机森林对五种化合物的分类效果均优于其他模型。确定使用随机森林作为五种化合物的分类预测模型，并对题中给的 50 组数据进行分类预测。

针对问题 4，

关键词：特征选择、随机森林、支持向量机、梯度提升

目 录

1. 问题重述

1.1 问题背景

据世界卫生组织国际癌症研究机构（IARC）发布的 2020 年全球最新癌症负担数据显示，2020 年乳腺癌首次取代肺癌（220 万）成为全球发病率第一的癌症，占所有新增癌症患者的 11.7%[1]。雌激素受体 α 亚型（Estrogen receptors alpha, ERα）被认为是治疗乳腺癌的重要靶标，能够拮抗 ERα 活性的化合物被认为可能是治疗乳腺癌的候选药物。在治疗乳腺癌药物的研发过程中，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。通过收集到的一系列作用于靶标 ERα 的化合物和其生物活性数据，构建以化合物的分子结构描述符为自变量，化合物的生物活性值为因变量的定量结构-活性关系（Quantitative Structure-Activity Relationship, QSAR）模型，使用该模型不仅可以指导目前已知的生物活性化合物的结构优化，而且可以预测具有更好生物活性的新化合物分子，使其具有更好的生物活性，为乳腺癌的治疗提供新的解决方案。在药物研发领域，评判一个化合物是否能够成为候选药物，除了需要具备良好的生物活性外，还需要在人体内具有良好的药代动力学性质和安全性。即该化合物进入人体后需要具备良好的吸收和代谢性质并且不会对人体产生隐藏的毒性隐患，以上性质被合称为 ADMET 性质（Absorption 吸收、Distribution 分布、Metabolism 代谢、Excretion 排泄、Toxicity 毒性）。因此为抗乳腺癌候选药物构建一个基于生物活性和 ADMET 性质的精确、稳定、高效的化合物分子定量结构-生物活性关系模型对于乳腺癌的治疗具有很重要的实际意义。

表 1-1 题目所给数据

资料名称	含义说明
附件一：ERα_activity	不同化合物样本的 IC50_nM 和 pIC50 数据
附件二：Molecular_Descriptor.xlsx	不同化合物样本的 729 个分子描述符数据
附件三：分子描述符含义解释.xlsx	不同分子描述符解释数据
附件四：ADMET.xlsx	不同化合物的 5 种 ADMET 性质数据

1.2 问题的提出

问题 1：特征选择

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），并请详细说明分子描述符筛选过程及其合理性。

问题 2：构建 pIC50 回归预测模型

结合问题 1，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，请叙述建模过程。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 IC50 值和对应的 pIC50 值预测，并将结果分别填入“ERα_activity.xlsx”的 test 表中的 IC50_nM 列及对应的 pIC50 列。

问题 3：构建 ADMET 分类预测模型

针对文件“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，并简要叙述建模过程。然后使用所构建的 5 个分类预测模型，对文件“ADMET.xlsx”的 test 表中的 50 个化合物进行相应的预测，并将结果填入“ADMET.xlsx”的 test 表中对应的 Caco-2、CYP3A4、hERG、HOB、MN 列。

问题 4：构建最优变量范围模型

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）。

2. 模型假设

假设 1：本题中所提供的 1974 个化合物的 729 个分子描述符数据均是准确的；

假设 2：除了题目提供的 729 个分子描述符之外，不存在其他因素影响化合物活性和 ADMET 性质；

假设 3：题目提供的训练数据和测试数据的原始分布一致；

3. 符号说明

符号表示	含义说明
μ	均值
σ	方差
X_i^n	标准化后的第i个样本的第n变量的值
$IC50$	化合物对 Erα 的生物活性值
$pIC50$	化合物对 Erα 的生物活性值的负对数
p_k	在节点k属于任何一类的概率估计值
G_k	在节点k的 Gini 指数
$I_{\Delta k}$	节点分裂前后 Gini 指数变化量
I_{it}	变量X_i的重要性
$\|\omega\|_2$	L2 范数
$\|\omega\|_1$	L1 范数
λ	正则化参数
r	皮尔逊相关系数
ρ	斯皮尔曼相关系数
τ_B	肯德尔相关系数
$Matrix_\rho$	特征相关性系数矩阵
Y	集成四个结果的特征重要性向量
ω_i	第i个方法在融合中所占权重
RMSE	均方根误差
MAE	平均绝对误差
R^2	拟合度

ROC	受试者工作特征曲线
AUC	ROC 曲线下的面积

4. 数据预处理

考虑到题目给出的 1974 个训练集数据可能会存在数据缺失，数据异常，冗余特征较多，进而会对后期具体问题的建模产生较大影响，因此对数据进行预处理，具体流程图如图 4-1 所示：

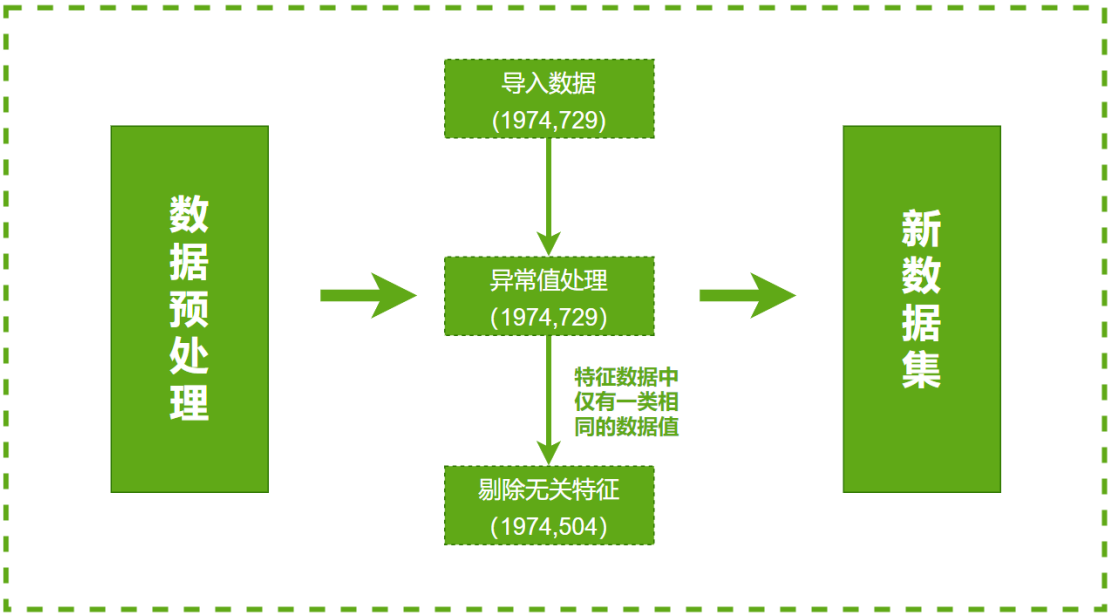

图 4-1 数据预处理流程图

4. 1 异常值处理

首先利用 python 的 pandas 对“Molecular_Descriptor.xlsx”进行数据分析，发现没有缺失值，因此不用进行缺失值处理，且发现只有整型特征 370 个和浮点型特征 359 个数据，两者基本持平。对该数据继续进行基本的统计分析，得到如表 4-1 所示，部分变量的下四分位数（75%）和最大值差距较大，因此初步判断这些数据中存在异常值，可能会影响后期模型的建立和预测。此外，对照分子描述符含义解释掌握分子描述符的基本含义，可以更为清楚的了解到这些整型数据的具体含义，比如化合物中氢原子的个数（Number of hydrogen atoms，nH），酸性基团的数目（Number of acidic groups，nAcid）等，这样则不难理解为何特征数据中存在如此多的整型数据 0，因此对于此类数据具有实际的生物化学性质并不能称为异常值数据，进而不做处理。

表 4-1 部分分子描述符信息统计表

	nAcid	ALogP	ALogp2	AMR	apol	nH	nC
count	1974	1974	1974	1974	1974	1974	1974
mean	0.108	1.110	3.288	116.557	60.626	22.649	22.607
std	0.348	1.434	12.833	31.567	19.450	10.775	6.6313
min	0	-23.105	0	54.067	30.662	5	7
25%	0	0.376	0.405	88.304	44.432	14	17
50%	0	1.171	1.560	114.837	59.901	22	22
75%	0	1.948	4.019	141.424	74.421	29	28
max	4	5.182	533.841	517.429	359.663	180	95

接着，本文采用基于箱线图法对数据异常值进行识别，以部分变量为例，如图 4-2 所示，可以清晰的看到部分变量存在些许离群值，对于这些异常值，使用该位点其它数值平均值代替。

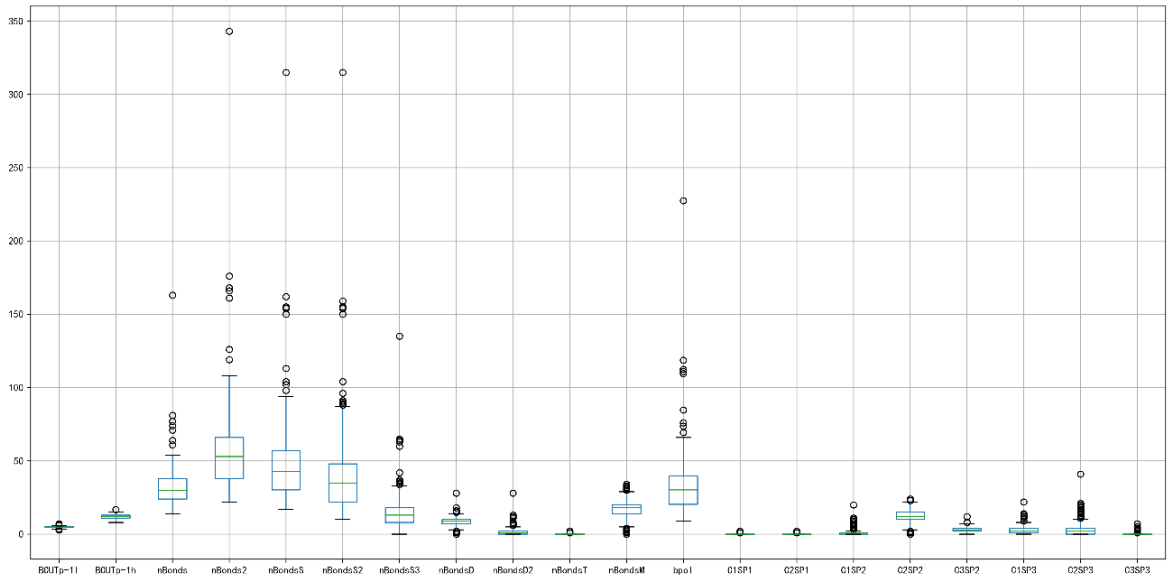

图 4-2 部分分子描述符的箱线图

4. 2 剔除无关特征

通过观察数据可以发现，其中包括一些数据值全部为 0 的变量，尽管数值为 0 也具有实际意义，但预测模型并不能识别其意义，同时这些变量会被认为是无关特征，浪费算力，影响模型的精度。因此，本文通过 Pandas 中的 nunique（）函数来获取唯一值的统计次数，计算出不同分子描述符各自分别有多少个不同的值，如表 4-2 所示。再通过 python 遍历出所有不同个数为 1 的分子描述符将其剔除，算法一见表。

表 4-2 部分分子描述符不同数值的个数

分子描述符	不同数值个数
n10Ring	1
nAcid	5
ALogP	1625
nAtom	87
WTPT-2	1470
MW	1319

表 4-3 算法一遍历法剔除无关特征

算法 1 遍历法剔除无关特征	
输入：原有特征数据集	
输出：剔除无关特征后的数据集	
for s in feature.columns：	#遍历所有特征
if feature[*s*].*nunique*() ==1:	#遍历无关特征
del feature[*s*]	#剔除无关特征

通过表 4-1 部分分子描述符信息统计表可以发现十元环的数目（Number of 10-membered rings，n10Ring）仅存在一种数值 0。所以对此分子描述符进行剔除。再通过算法一遍历后我们剔除了 225 个无关特征，得到了有 504 个特征的新数据集。

5. 问题一

5.1 问题分析

根据问题一的要求，需要对 729 个分子描述符进行变量选择，选出对生物活性 pIC50 具有显著影响的前 20 个分子描述符（即特征），本题可以理解为基于特征与目标性变量的特征选择问题。观察数据，分子描述符多达 729 个，而数据样本只有不到两千个，而且某些分子描述符具有密切相关性，这些冗余描述符可能会导致后续模型的过拟合。因此，为了模型优化，对于这样的高维小样本数据，我们通常需要对计算出的描述符进行预处理，筛选出重要且独立性好的特征。基于此，题目中要求的前 20 个对生物活性最具有显著影响的分子描述符，本文认为是不包含高度相关性的重复特征，是独立性较好同时具有显著重要性的特征。

鉴于每种特征选择方法都有其考量的重点以及合理性，为综合考虑特征自身的发散程度，各特征与目标之间的相关性，不同特征组合同目标之间的相关性等问题，本文将按照如图 5-1 所示流程图进行处理，具体步骤如下：

(1) 首先基于第 4 章数据预处理得到的 504 为数据集建立新的化合物数据集；

(2) 接着分别使用封装式的**递归特征消除法**，嵌入式的**基于 L1 正则化项的 Lasso 回归、基于 L2 正则化项的 Ridge 回归，基于随机森林**的特征选择方法，计算出 504 个特征的重要性；

(3) 再分别对这四种方法的预测能力（R^2 系数）给予权重，加权得到 504 个特征的集成式特征重要性，并筛选出特征重要性排名前 30 的特征（分子描述符）；

(4) 然后对这 30 个特征 **Spearman 相关性方法**分析去除 10 个特征，得到 20 个特征；

(5) 最后对这 20 个特征进行显著变量的独立性验证和代表性验证，说明方法的合理性。

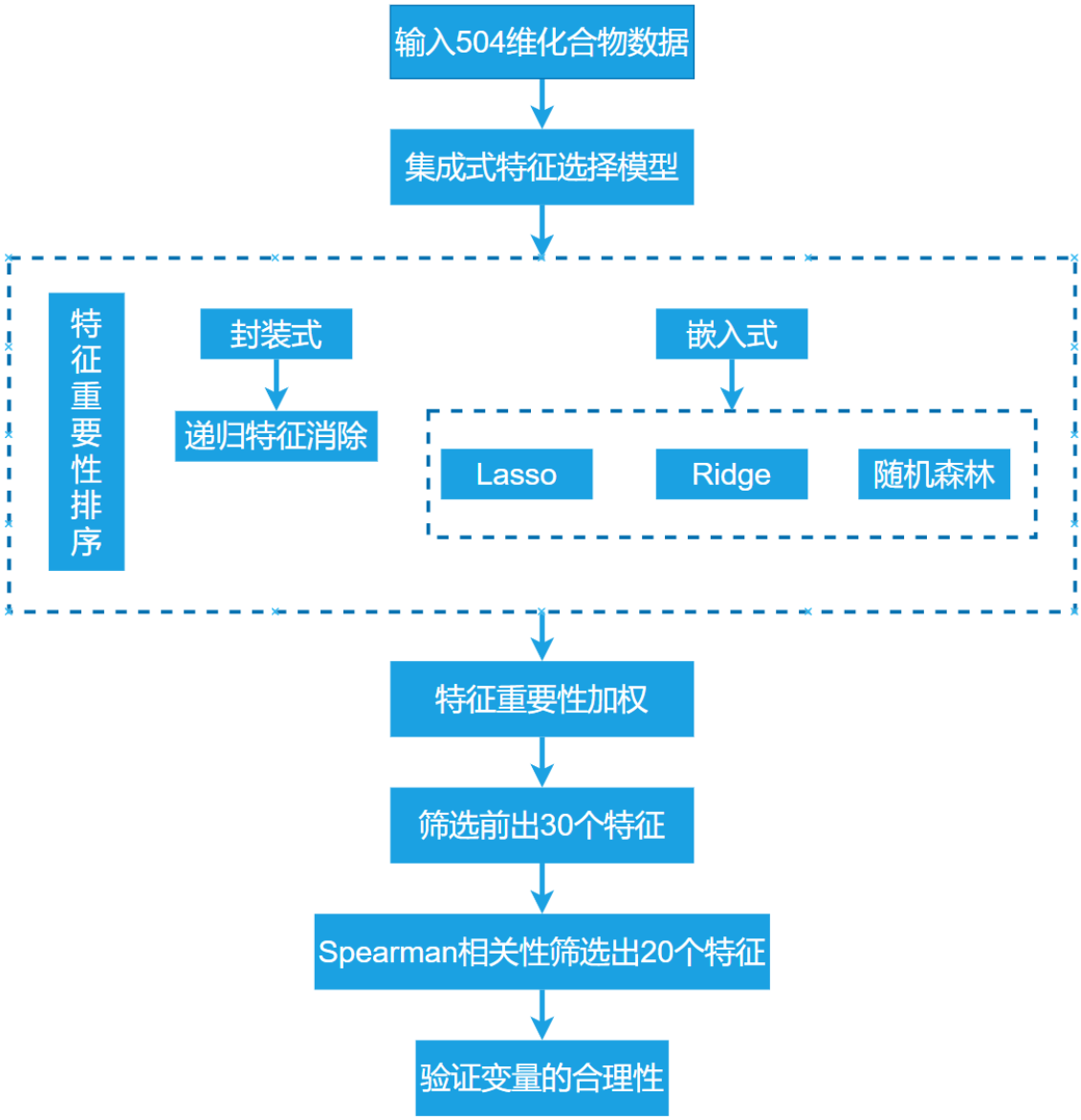

图 5-1 问题一流程图

5.2 数据预处理

经过上一节数据预处理的简单清洗，共剩余 504 个特征，这些特征的类型（离散/连续）、取值区间（量纲）、变化程度各不相同，且差异较大。因此在筛选特征的时候需要根据算法特性确定是否进行标准化操作，以及标准化操作的类型。

$Z-\text{score}$**标准化：**即 0 均值标准化，在保证数据原始分布的同时，将数据缩放为均值为 0，方差为 1 的变量，由于各个变量的量纲存在巨大差异，通过 $Z-\text{score}$ 标准化可以将变量同一量纲。对于每一个变量，都施行标准化操作：

$$X_i^n = \frac{X_i^n - \mu}{\sigma}, \tag{5.1}$$

其中 μ 与 σ 分别表示第 n 个变量在 M 个样本上的均值和方差。 X_i^n 表示标准化后的第 i 个样本的第 n 变量的值。

0－1归一化：将数据的区间缩放到 $[0,1]$ 内：

$$X_i^n = \frac{X_i^n - \min\left(X^n\right)}{\max\left(X^n\right) - \min\left(X^n\right)} \tag{5.2}$$

对于连续类型的分子变量，由于其量纲差异大，因此选择 $Z-\text{score}$ **标准化**进行数据处理。对于生物活性指标，IC50 数值浮动程度大，而 pIC50 经过负对数标准化后数值稳定，且区间紧凑，因此采用 pIC50 作为后续变量筛选的目标值，同时利用 $0-1$ **归一化**进行数据处理。

5.3 模型建立

不同的特征选择方法因为模型方法的不同，相同特征的特征重要性也有所不同，因此，为了综合这些方法对线性、非线性、高耦合特征的建模能力，基于本题的小样本高维数据，因此，本文建立了集成特征筛选模型，期望获得更加灵活，过拟合风险更低的筛选模型。

5.3.1 基于递归特征消除法的特征重要性排序

递归特征消除（Recursive Feature Elimination，简称 RFE）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。

递归特征消除法具体步骤：

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。

在本题中，选择将 504 个特征数据集使用线性回归模型的递归特征消除法对特征重要性进行排序得到对应分子描述符的重要性排序。

5.3.2 基于随机森林的特征重要性

随机森林（Random forest，简称 RF）作为一种新兴起、高度灵活的机器学习算法，其拥有广泛的应用前景，在大量分类以及回归问题中具有极好的准确率。并且，随机森林算法自带特征筛选机制，即随机森林能够评估各个特征在相应问题上的重要性。在本题中，针对分子描述符特征数量多、高度非线性、耦合度高的特点，选择嵌入式特征选择方法随机森林进行数学建模，以期望得到 504 个分子描述符的重要性。

本文采用 Gini 指数作为分子描述符的重要性评判指标，针对数中的每个节点 k，计算其 Gini 指数，如公式(5.3)所示。

$$G_k = 2p_k\left(1-p_k\right), \tag{5.3}$$

其中，p_k 代表样本在节点 k 属于任何一类的概率估计值，一个节点的重要性程度由节点分裂前后 Gini 指数变化量来确定：

$$I_{\Delta k} = G_k - G_{k1} - G_{k2}, \tag{5.4}$$

其中，G_{k1} 和 G_{k2} 分别表示节点 k 产生的子节点，针对每棵树进行递归，最终随机抽样选择样本和变量，产生包含 K 棵树的森林，如果变量 X_i 在 t 棵树中出现 N 次，则变量 X_i 的重要性为：

$$I_{it} = \frac{1}{n}\sum_{t=1}^{T}\sum_{j=1}^{N} I_{\Delta j} \tag{5.5}$$

5.3.3 基于 L2 正则化项的 Ridge 回归的特征重要性

岭回归(ridge regression, Ridge)是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法，本文采用基于 L2 正则化项的 Ridge 回归。

Ridge 回归通过对系数大小施加惩罚来解决普通最小二乘法的一些问题，采

用的是 L2 范数$\|\omega\|_2$，其公式为：

$$min_{\omega}\sum_{i=1}^{m}\left(y_i-\omega^t x_i\right)^2+\lambda\|\omega\|_2 \tag{5.6}$$

其中，$\lambda\geq 0$是控制系数收缩量的复杂性参数：λ的值越大，收缩量越大，这样系数对共线性的鲁棒性也更强。参数λ的选择决定了回归系数被压缩的程度也就是特征重要性的大小，不同的λ可能产生不同的结果。本文采用机器学习领域的交叉验证（Cross Validation）方法，首先将数据分成M个大小相同的样本，通常M可为 5、10 或N（样本量），本文确定为 5；然后选择λ的某个值，将前$M-1$份的数据采用 Lasso 回归方法估计模型，再模型得到的回归系数用于第M份数据的验证，检验模型设立是否正确，并将上述过程重复M次；最后，我们将得到某一λ值下模型的拟合值。交叉验证的方法通常会重复上述过程 100 次，选择 100 个不同的λ，再以均方误差大小决定参数λ值。

线性机器学习算法拟合的模型的预测是输入值的加权和，如线性回归、逻辑回归、加入正则项的 Lasso 回归、Ridge 回归和弹性网络。所有这些算法都是通过找到一组系数用于加权和，然后用于预测。这些系数可以直接用作特征重要性评分的一种粗略类型。本文采用 python 中 sklearn 的 cofe_函数计算基于 L1 正则化项的 Lasso 回归的系数即特征重要性。

5.3.4 基于 L1 正则化项的 Lasso 回归的特征重要性

Lasso(Least absolute shrinkage and selection operator)方法是一种压缩估计。它通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零。在本题中，针对分子描述符特征数量多、高度非线性、耦合度高的特点，选择嵌入式基于 L1 正则化项的 Lasso 回归的数学建模，以期望得到 504 个分子描述符的重要性。

Lasso 方法作为正则化方法的一种，它以 L1 范数$\|\omega\|_1$向量中各个元素绝对值之和作为规则化项，其公式为(5.7)：

$$min_{\omega}\sum_{i=1}^{m}\left(y_i-\omega^t x_i\right)^2+\lambda\|\omega\|_1 \tag{5.7}$$

其中 λ 为正则化参数，Lasso 回归实际上就是线性模型加了一个惩罚项，当 $\lambda=0$ 时就是常见的线性模型。相比较基于 L2 正则化项的 Rigde 回归采用的 L2 范数向量各元素平方和然后求平方根，对于较小的回归系数估计值压缩力度更小，可以减少对重要回归系数的过度压缩。

Lasso 回归的参数 λ 的选择和特征重要性的计算均与 Ridge 类似这里不再赘述。

5.3.5 基于 Spearman 相关系数法冗余特征剔除模型

统计学习方法中常用的相关系数矩阵法主要包括 Pearson、Spearman、Kendall 三种典型算法，下面比较三种算法对本题的适用性。

皮尔逊（Pearson）相关系数：又称为皮尔逊积矩相关系数，是用于度量两个变量 X 和 Y 之间的相关性，其范围为 $[-1,1]$。一般用于分析两个变量之间的关系，是一种线性相关系数，公式定义为两个变量之间的协方差和标准差的商：

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \tag{5.8}$$

Pearson 系数适用于：

(1) 两个变量之间是线性关系，都是连续数据;

(2) 两个变量的总体是正态分布，或接近正态的单峰分布;

(3) 两个变量的观测值是成对的，每对观测值之间相互独立。

斯皮尔曼（Spearman）相关系数: 又称斯皮尔曼秩相关系数，是秩相关系数的一种。“秩”，即秩序，可以理解为一种顺序或排序，根据变量在数据内的位置进行计算，公式为：

$$\rho=1-\frac{6\sum d_i^{\,2}}{n(n^2-1)} \tag{5.9}$$

与皮尔逊相关系数相比，斯皮尔曼相关系数没有那些限制，不需要关心数据如何变化，符合什么样的分布，只需要关心每个变量对应数值的位置。如果两个变量的对应值，在各组内的排列顺位是相同或类似的，则具有显著的相关性。

肯德尔（Kendall）相关系数: 又称肯德尔秩相关系数，它也是一种秩相关系数，不过，它的目标对象是有序的类别变量，它可以度量两个有序变量之间单调关系强弱。肯德尔相关系数使用了“成对“这一概念来决定相关系数的强弱，公式为：

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{5.10}$$

其中 n_c 表示 XY 中拥有一致性的元素对数； n_d 表示 XY 中拥有不一致性的元素对数。与前两种方法相比，其更适用于有序的变量。

因此，三种相关系数都是对变量之间相关程度的度量，由于其计算方法不一样，用途和特点也不一样。

(1) Pearson 相关系数是在原始数据的方差和协方差基础上计算得到，所以对离群值比较敏感，它度量的是线性相关。因此，即使 pearson 相关系数为 0，也只能说明变量之间不存在线性相关，但仍有可能存在曲线相关。

(2) Spearman 相关系数和 Kendall 相关系数都是建立在秩和观测值的相对大小的基础上得到，是一种更为一般性的非参数方法，对离群值的敏感度较低，因而也更具有耐受性，度量的主要是变量之间的联系。

通过图 4-2 部分分子描述符的箱线图我们可以初步发现本题中的特征数据存在非连续型变量、离群值，不符合正态分布，因此选择斯皮尔曼系数进行冗余特征剔除模型的建立。

利用此方法计算两两特征之间的 Spearman 系数，得到特征相关性系数矩阵：

$$Matrix_{\rho} = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1D} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{D1} & \rho_{D2} & \cdots & \rho_{DD} \end{pmatrix} \tag{5.11}$$

其中 ρ_{ij} 表示分子描述符特征 i 与 j 之间的 Spearman 相关系数值， D 表示当前的分子描述符数量。然后找出 $\rho_{ij} \geq 0.8$ 的位置，即可定义为两个特征之间存在强相关性。

5.4 模型求解

5.4.1 特征重要性归一化

四种方法均是以 504 个分子描述符为自变量，生物活性 pIC50 值为因变量建立的数学模型。为了综合四种方法的特征重要性，首先对特征重要性进行归一化处理，得到特征权值，其反应特征的重要程度占比，每一个分子描述符的特征权值，表示为该特征重要性与全体特征重要性之和的比值。归一化的部分分子描述符重要性如表 5-1，可以看到数越大，该特征越重要，同时不同方法得到的特征重要性值还是有很大的差异，因此我们考虑加权融合。

表 5-1 归一化分子描述符重要性

分子描述符	随机森林	Lasso 回归	递归特征消除	Ridge 回归
nAcid	0.008927	0.544725	0.945010	0.639704
ALogP	0.144065	0.595974	0.794297	0.668435
ALogp2	0.105598	0.598920	0.883910	0.594082
AMR	0.418678	0.598920	0.867617	0.494249
apol	0.307342	0.598920	0.103870	0.572045
naAromAtom	0.113447	0.598920	0.260692	0.587306
nAromBond	0.076364	0.598920	0.256859	0.560132
…	…	…	…	…
WPOL	0.074782	0.598920	0.696538	0.573483
XLogP	0.253227	0.739498	0.727088	0.670312
Zagreb	0.135937	0.598920	0.181263	0.623413

5.4.2 集成特征重要性计算

归一化后对每种方法的特征权值进行加权融合，即得到集成的 504 个分子描述符重要性向量：

$$Y=\sum_{i=1}^{4}\omega_i y_i \tag{5.12}$$

其中，Y 为集成四个结果的特征重要性向量，ω_i 为第 i 个方法在融合中所占权重，并满足权重和为 1，y_i 为第 i 个方法归一化的特征权值。这里我们简述下 ω_i

方法融合权重的计算，本文采用分别对这四种方法进行模型建立用此来预测 pIC50 的值，并计算出各个方法的 R^2 的值用来衡量该方法预测能力，再计算出每个方法的权重如表 5-2。

表 5-2 四种方法 R^2 值、权重 ω_i

模型	R^2	ω_i
随机森林	0.754861	0.291923
Lasso 回归	0.646255	0.249922
递归特征消除	0.662262	0.256113
Ridge 回归	0.522444	0.202042

最后得到集成融合后的特征重要性，并对其进行排序得到前 30 个分子描述符具体信息如表 5-3，前通过下图 5-2 的可视化，可以发现前三 maxHsOH、minsssN、MLFER_A 分子描述符的特征重要性较高，后面的特征重要性依次递减。

表 5-3 集成融合后的前 30 个分子描述符重要性

重要性排名	分子描述符	集成特征重要性
1	maxHsOH	0.773179
2	minsssN	0.747391
3	MLFER_A	0.717124
…	…	…
28	maxaaCH	0.550713
29	minHBd	0.546171
30	fragC	0.543504

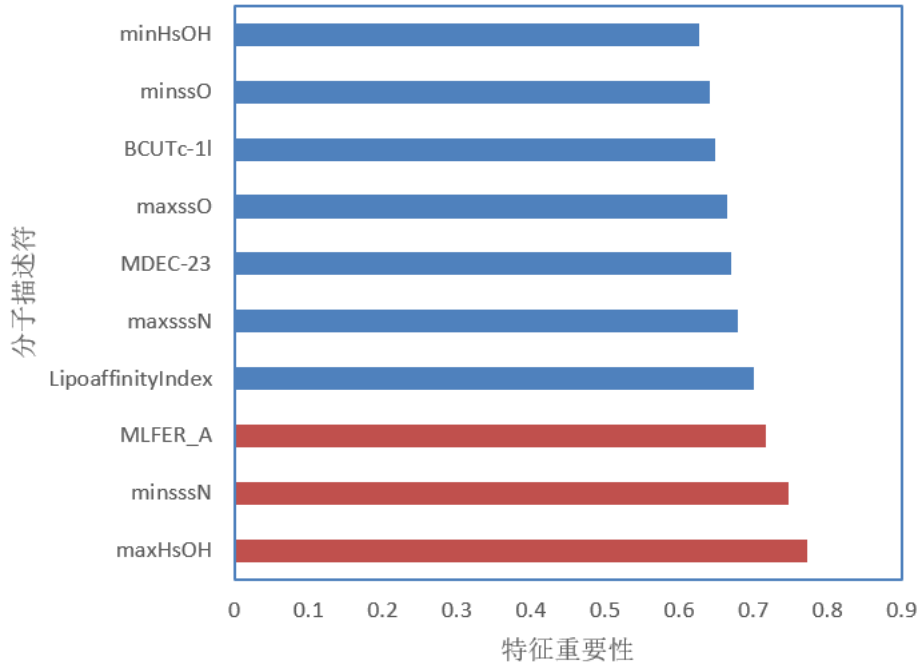

图 5-2 排名前 10 的分子描述符重要性

5.4.3 冗余特征剔除模型的实现

对上节所筛选出来的 30 个特征计算出其相关性系数，得到如图 5-3 的特征相关性热图。

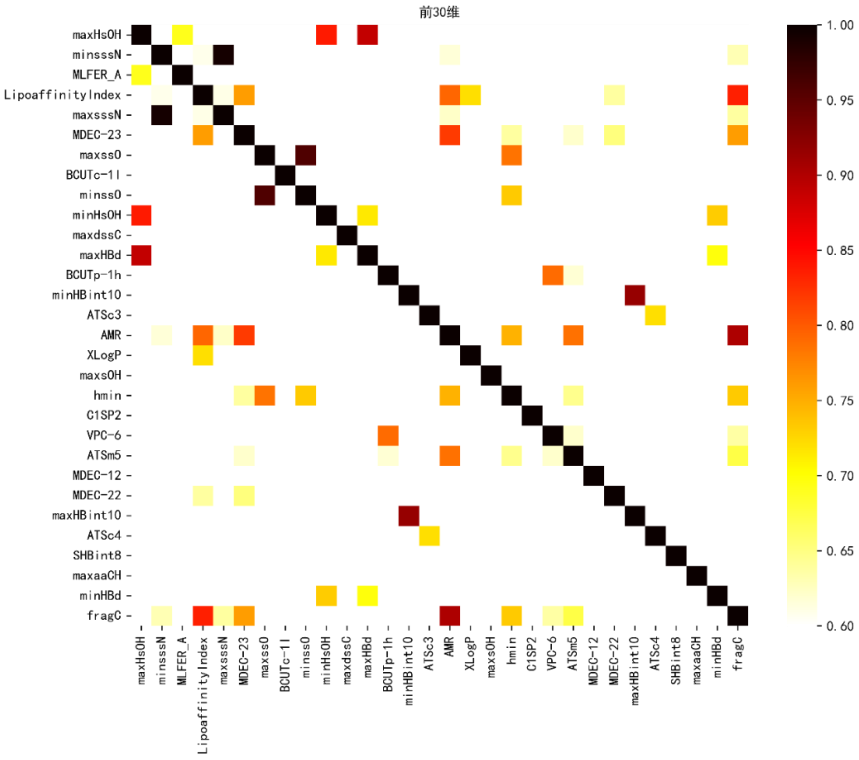

图 5-3 特征重要性排名前 30 的特征相关性热图

通过热图和相关性系数可以很容易的发现前 30 个特征中存在较多冗余特征，部分特征之间冗余性过大，独立性较低。考虑到采用 python 遍历法剔除容易生物化学意义上更为重要的分子描述符，本文根据“附件三：分子描述符含义解释.xlsx”将其前 30 个分子描述符的详细化学信息进行了统计如表 5-4 所示，易知同类别的分子描述符往往具有更强的冗余性，同时考虑类别的多样性，如果特征重要性相关系数、特征重要性都较高则可以去除同类别的分子描述符。因此基于特征重要性，特征相关性系数 $\rho_{ij} \geq 0.8$，分子描述符同类别的优先性手动剔除 10 个特征得到最后 20 个特征，分别剔除了 “fragC”,“maxsssN”,“maxHBd”,“minHBint10”,“hmin”,“AMR”,“minHsOH”,“LipoaffinityIndex”“ATSc4”“minssO”这 10 个特征，得到了题目所要求的 20 个特征。

表 5-4 前 30 个分子描述符含义解释

排序	分子描述符	分子描述符类别	含义
1	maxHsOH	原子型描述符的电拓扑态	最大原子类型: -OH
2	minsssN	原子型描述符的电拓扑态	最小原子类型>N-
3	MLFER_A	MLFER	整体或总和溶质氢键酸度
…	…	…	…
28	maxaaCH	原子型描述符的电拓扑态	最大原子类型::CH:
29	minHBd	原子型描述符的电拓扑态	(强)氢键供体的最小 e 态
30	fragC	片段的复杂性	化合物的复杂性

5.5 特征选择合理性验证

5.5.1 特征代表性验证

为了验证所选特征具有较好的预测能力，我们选择常用的常见的机器学习回归算法 KNN、SVR、RandomForestRegressor、GradientBoostingRegressor 评估了所筛选的 20 个特征的预测性能，采用回归指标 MAE 和 R^2 进行评估如表 5-5，可以发现管是 MAE 值还是 R^2 值都可以证明所选的 20 个特征的有效性。

表 5-5 四种模型的 MAE、R^2 值

模型	MAE	R^2

KNN	0.599362	0.672239
SVR	0.613686	0.67556
RF	0.548393	0.722923
GBoost	0.580251	0.702049

5.5.2 特征独立性验证

同样我们采取 Spearman 相关性方法对所筛选的 20 个特征进行相关系数的计算，得到如图 5-4 的热图，可以清晰的发现该 20 个特征独立性较好，相关性较低，满足独立性验证。

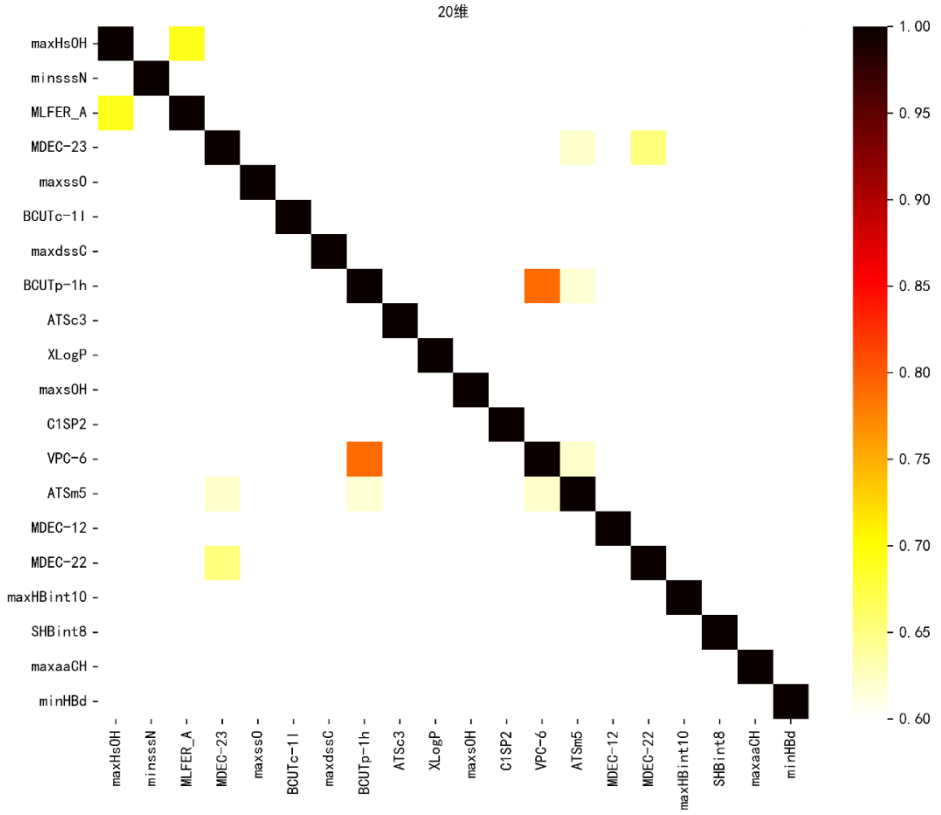

图 5-4 筛选后的 20 个特征相关性热图

6. 问题二

6.1 问题分析

结合问题 1，选择不超过 20 个分子描述符，构建化合物 ERα 生物活性的定量预测模型。本文采用支持向量回归、随机森林、梯度提升三种回归方法分别建立化合物对 ERα 生物活性的定量预测模型，采用随机搜索方法对基于支持向量回归、随机森林和梯度提升的回归预测模型进行超参数搜索，选择出适合模型的最佳超参数并使用统一评价指标对三种模型进行综合对比分析，以确定最佳回归预测模型，使用该模型对附件一“ERα_activity.xlsx”test 表中数据进行 pIC50 和 IC50 值的预测，并可视化预测结果。

问题二的思路流程图如下图 6-1 所示：

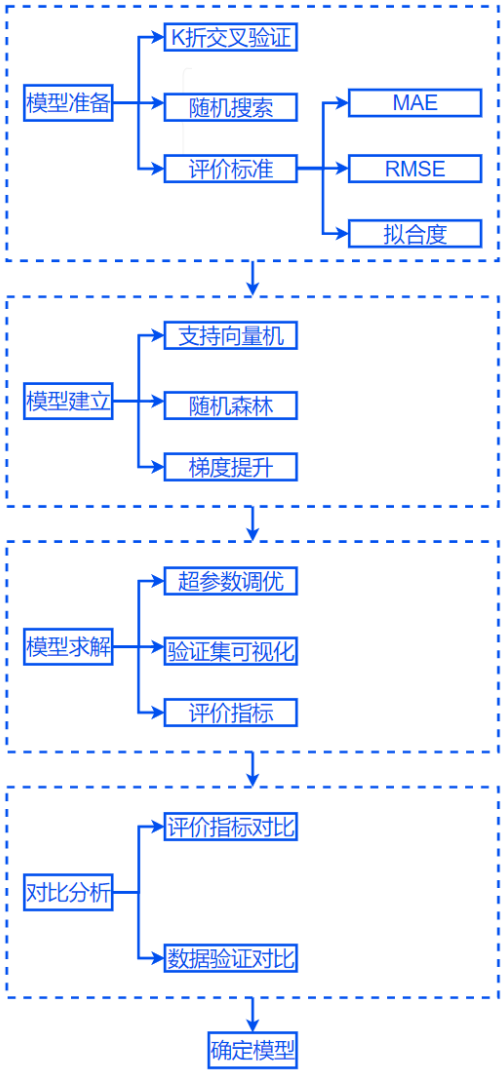

图 6-1 问题 2 流程图

由于 IC50 值的波动较大，不利于模型的建立，因此采用 pIC50 来表示生物活性值，两者的换算关系为：

$$IC_{50} = 10^{(9-pIC_{50})} \tag{6.1}$$

6.2 模型准备

6.2.1 K 折交叉验证

K 折交叉验证使用了无重复抽样技术的好处：为了得到可靠稳定的模型，每次迭代过程中每个样本点只有一次被划入训练集或测试集的机会。10 折交叉验证示意图如下图 6-2 所示，首先将数据集分为 10 等份，然后选取一份作为测试集，另外 9 份作为训练集，一共重复 10 次，每次选取的训练集不同。

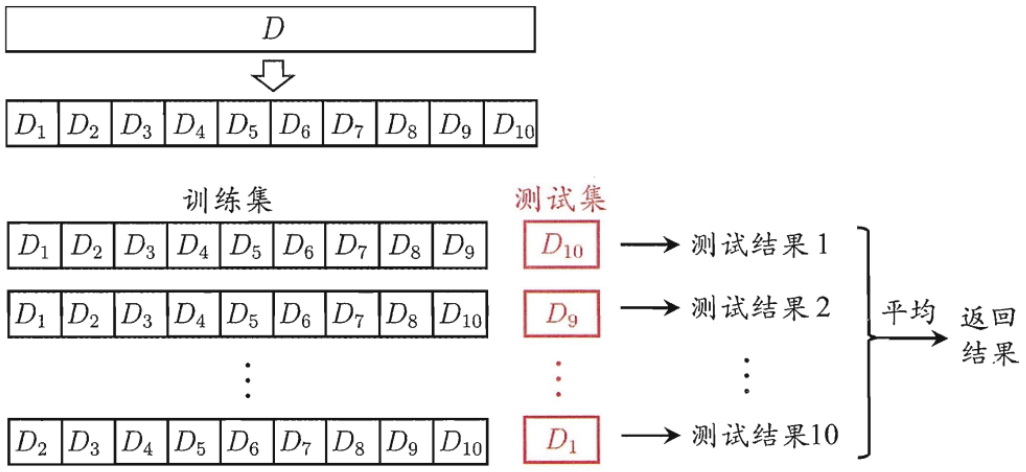

图 6-2 10 折交叉验证示意图

6.2.2 随机搜索

随机搜索（Random Search，RS）是一种常用的机器学习超参数优化的方法。随机搜索就是在给定的参数范围之类进行随机抽样产生参数值，再对多个抽样值选取最优的参数组合。Python 中 sklearn 有 RandomizedSearchCV 的实现，其使用方法如下：

(1)定义一个算法，比如 LogisticRegression，包括自定义算法；

(2)定义参数的分布，如果所有参数都是 list，则等同于网格搜索；

(3)定义 RandomizedSearchCV 对象，传入算法、参数分布、随机种子等参数；

(4)进行模型训练，并进行参数调优。

常见的一种调参方法还有网格搜索（Grid Search），就是对网格中每个交点进行遍历，从而找到最好的一个组合，这种方法的效果不错，适用于需要对整个参数空间进行搜索的情况，但是这种方法计算代价非常大，可能会引起维度灾难。在一些情况下，随机搜索得到的超参数组合的性能稍微差一点。随机搜索的好处在于搜索速度快，相对于复杂的问题比较容易应用，但是这个方法的缺点在于必须通过反复多次实验来对每个问题进行特殊处理，不然容易错过一些重要的信息。

6.2.3 评价标准

为了验证回归模型是否有效，需要在猜测集上对预训练模型进行验证。通常需要一定的指标来判断模型的准确性，常用的指标有均方根误差（RMSE）、平均绝对误差（MAE）、拟合度（R^2）。

(1) 均方根误差 RMSE：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i\right)^2} \tag{6.2}$$

(2) 平均绝对误差 MAE：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i\right| \tag{6.3}$$

其中 y_i 为样本预测值，y_i 为样本真实值，以上的评价指标范围为$[0,+\infty]$，当预测值与真实值完全吻合时等于 0，即完美模型；误差越大，该值越大。

(3) 拟合度 R^2：

$$R^2 = 1 - \frac{\sum_{i}\left(y_i - y_i\right)^2}{\sum_{i}\left(\overline{y_i} - y_i\right)^2} \tag{6.4}$$

其中 $\overline{y_i}$ 为样本平均值，R^2 的取值范围为$[0,1]$：如果结果为 0，说明拟合效果较差；如果结果是 1，说明模型无误，一般来说 R^2 越接近 1 越好。

6.3 模型建立

6.3.1 支持向量回归

支持向量机（Support Vector Machine，SVM）是一种经典的机器学习算法，在小样本数据集的场景中有较为广泛的应用。SVM 主要是将数据进行分类，其模型思想是：假设平面中有两种类型的点，一种是红色一种是蓝色，要求我们将这两种类型的点分开，取一条最优的分界线，既能把两种类型分开，还使得两类样本点离分界线最近的点离分界线的距离尽可能远。对于在高维空间中，不像二维空间中的一条直线，把高维空间中的分界线叫做超平面，而真正决定分割超平面作用的点叫做支持向量。

支持向量回归（Support Vector Regression，SVR）是 SVM 对回归问题的一种运用，主要是升维之后，在高维空间中构造线性决策函数来实现线性回归。SVR 模型的模型函数是非线性高斯核，可以表示为下式(6.5)：

$$\phi_r\left(x,\ell\right)=\exp\left(-\gamma\left\|x-\ell\right\|\right)^2, \tag{6.5}$$

这是一个从 0（离地标差的非常远）到 1（跟地标一样）变化的钟形函数。

6.3.2 随机森林

随机森林（Random Forest，RF），是一种新兴起的、高度灵活的、基于树的机器学习算法，该算法利用多棵树的力量来进行决策。森林中的每棵树并不一样，每棵树都是被随机创造的，每棵树中的每一个节点都是待选特征的一个随机子集，所有树的输出结果整合起来就是森林最后的输出结果。

随机森林重点是“随机”两个字，它包括两个方面：

(1)训练数据的随机选取。即用于训练单棵树的数据应该是随机、有放回的从全部数据中选取和原始数据集相同的数据量；

(2)待选特征的随机选取。使得森林中的各棵树都能够彼此不同，提示系统的多样性，从而提升分类性能。

构造随机森林的步骤可描述为：

Step1：假如有 N 个样本，则有放回的随机选择 N 个样本。用选择好的 N 个样本来训练一个决策树，作为决策树根节点处的样本；

Step2：当每个样本有 M 个属性时，在决策树的每个节点需要分裂时，随即从 M 个属性中选取 m 个属性，满足 m<<M。然后从 m 个属性中采用某种策略（比如信息增益）来选择一个属性作为该节点的分裂属性；

Step3：决策树形成过程中每个节点都要按照 step2 来分裂，一直到不能再分裂为止。注意整个决策树形成过程中没有进行剪枝；

Step4：按照 step1-3 建立大量的决策树，这样就构成了随机森林。

随机森林算法示意图如下图 6-3 所示：

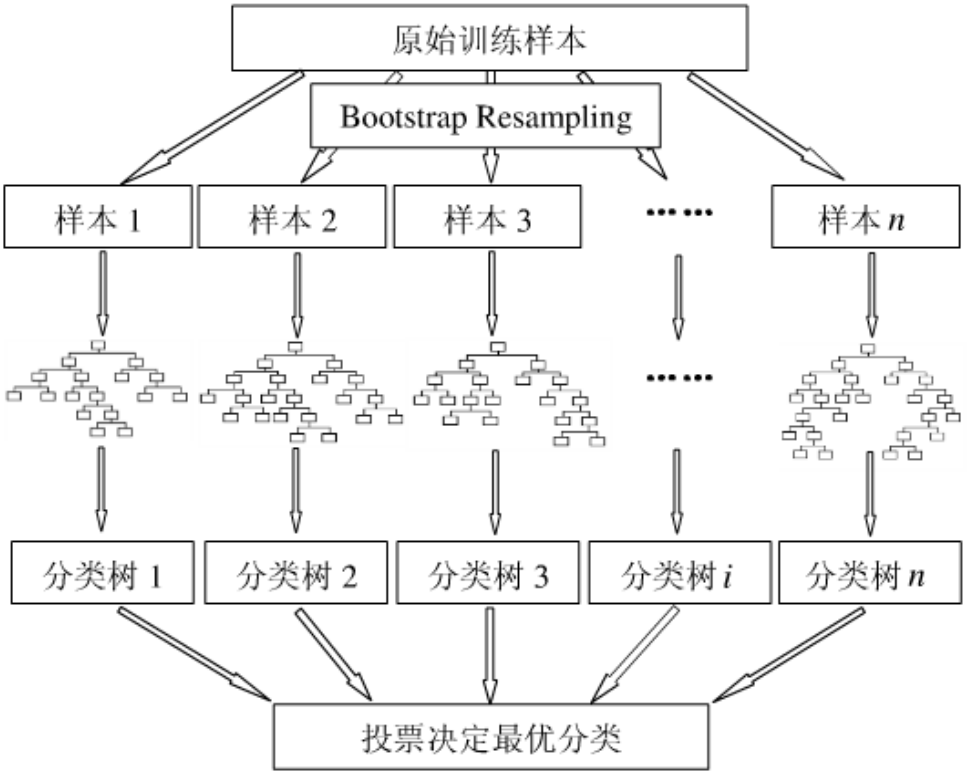

图 6-3 随机森林示意图

6.3.3 梯度提升

梯度提升（Gradient Boosting Regression Tree ，GBRT），是一种迭代的回归树算法，由多棵回归树组成，所有树的结论累加起来得到最终结果。其核心就在于，每一棵树是从之前所有树的残差中来学习的。GBRT 主要有两个部分组成：回归树（RT）和梯度提升（GB）。

(1)RT：GBRT 是迭代的决策树算法，决策树分为两类，回归树和分类树，分类树用于分类标签值，回归树用来预测实际值。GBRT 也是一种迭代的回归树算

法，由多棵回归树组成，所有树的结论累加起来得到最终结果;

(2)GB：Boosting 通过迭代多棵树来共同决策，核心是每一棵树都是学习之前所有树的结论和残差。

梯度提升回归树模型总体来说时非常不错的，它的优势是比较精确，因为在提升过程中是在不断修正前面的决策树。

6.4 模型求解

本文首先对 SVR、GBRT 和 RF 模型进行超参数的确定，使用随机搜索的方式对 SVR、GBRT 和 RF 模型进行 100 次一定范围内的最佳超参数搜索，使用确定的超参数建立 3 个回归预测模型，采用 5 折交叉验证验证模型的鲁棒性并分别对验证集中数据进行预测，使用 RMSE、MAE 和拟合度作为模型的评价指标，分别可视化 3 个模型的预测情况，进一步对比分析模型的优劣，以确定适合本问题的最佳回归预测模型。

6.4.1 支持向量回归

对 SVR 模型最佳超参数进行随机搜索，先在较大范围（C:1-100,gamma：0.0001-0.1）内进行粗略搜索，根据搜索图缩小搜索范围（C:1-10,gamma：0.001-0.01）内进行精确搜索，搜索图如下图 6-4 所示：

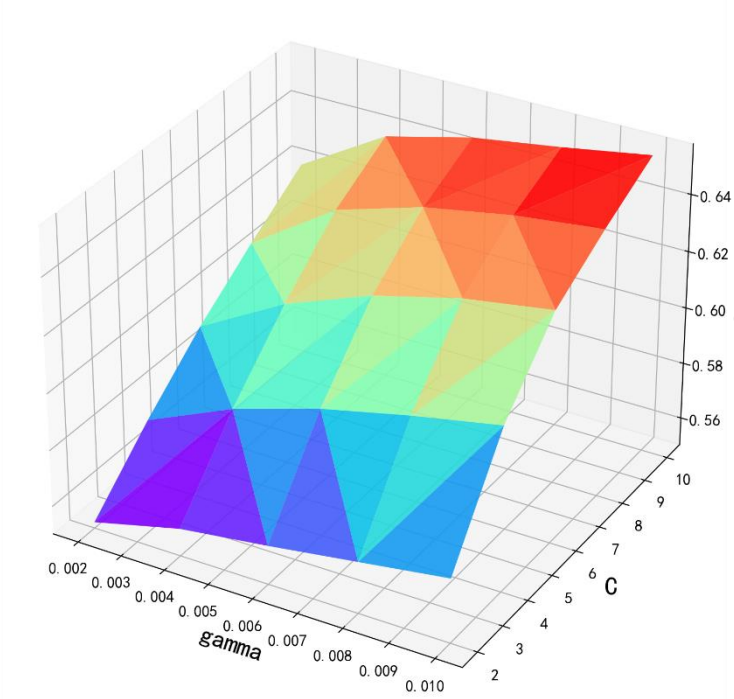

图 6-4 随机搜索算法的 SVR 最佳参数选择图

平面坐标表示两个超参数 C、gamma，纵坐标表示的是模型拟合度，由图可得，颜色越深模型的效果越好，即 C 和 gamma 的值较大时有较高的拟合度，最佳拟合值是 0.650，随机搜索的结果如下表 6-1 所示：

表 6-1 SVR 随机搜索结果

C	gamma
10	0.01

6.4.2 随机森林

对 RF 模型最佳超参数进行随机搜索，先在较大范围（max_features:1-50, n_estimators：1 -1000）内进行粗略搜索，根据搜索图缩小搜索范围（max_features:1-10, n_estimators：1-250）内进行精确搜索，搜索图如下图 6-5 所示：

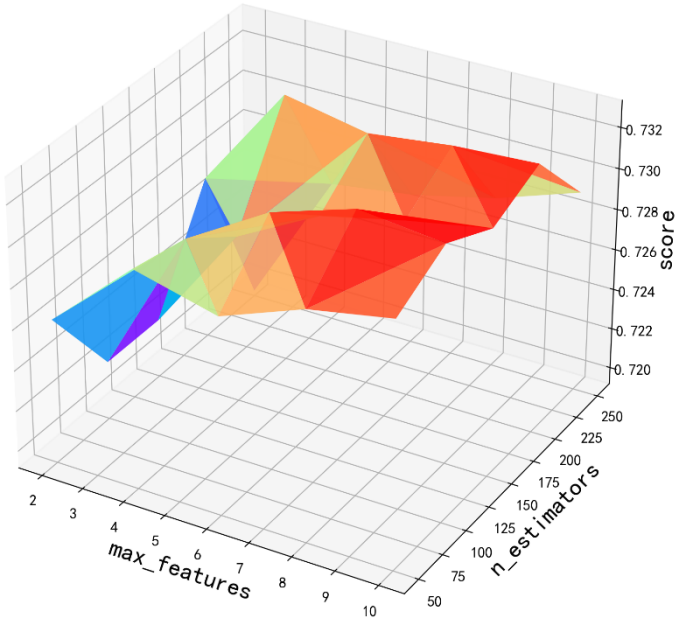

图 6-5 随机搜索算法的 RF 最佳参数选择图

平面坐标表示两个超参数 max_features、n_estimators，纵坐标表示的是模型拟合度，由图可得，颜色越深模型的效果越好，最佳拟合值是 0.740，随机搜索的结果如下表 6-2 所示：

表 6-2 RF 随机预测结果

max_features	n_estimators
4	199

6.4.3 梯度提升

对 GBRT 模型最佳超参数进行随机搜索，先在较大范围（max_depth:1-100,n_estimators：1-500）内进行粗略搜索，根据搜索图缩小搜索范围（max_depth:4-20，n_estimators：40-200）内进行精确搜索，搜索图如下图 6-6 所示：

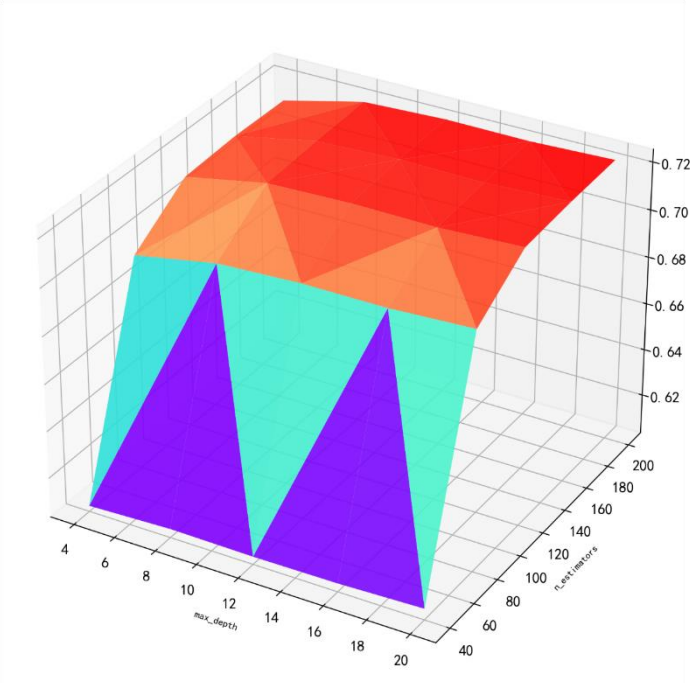

图 6-6 随机搜索算法的 GBRT 最佳参数选择图

平面坐标表示两个超参数 max_depth、n_estimators，纵坐标表示的是模型拟合度，由图可得，颜色越深模型的效果越好，最佳拟合值是 0.724，随机搜索的结果如下表 6-3 所示：

表 6-3 GBRT 随机预测结果

max_depth	n_estimators
20	200

6.5 对比分析

根据 3 种模型验证集的前 30 组数据 pIC50 的实际值和预测值，画出其对比图

如下图 6-7 所示：

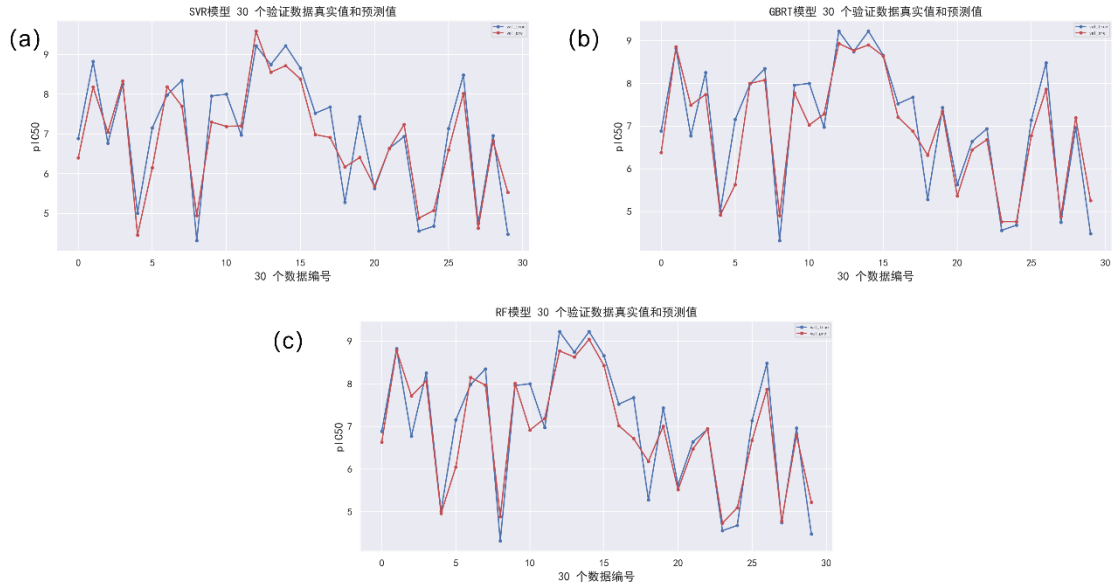

图 6-7 三种模型验证数据真实值和预测值

其中（a）、（b）、（c）分别表示 SVR、GBRT、RF 模型的验证数据集真实和预测值的对比图，从图中可以看出对这 30 个验证集来说三种模型的预测效果差距不大。

根据评价标准中的评价指标对 3 种模型的 RMSE、MAE、拟合度以及运算时间进行计算，得到结果如下表 6-4 所示：

表 6-4 三种模型评价指标

算法名称	RMSE	MAE	拟合度	时间/s
SVR	0.857	0.650	0.650	**48.251**
GBRT	0.761	0.555	0.724	543.296
RF	**0.739**	**0.552**	**0.740**	119.906

四种评价指标中 RMSE 和 MAE 值越小模型预测值越接近实际值；拟合度的 R^2 越大，模型对数据集拟合度越好。所以基于上表的这四种评价指标可以得出：基于随机森林的回归预测模型在验证集上 RMSE、MAE、拟合度均最小，基于支持向量回归的预测模型在验证集上训练时间远小于其他两个模型，所以本文决定采用基于随机森林的回归预测模型作为对 pIC50 值预测的模型。

使用基于随机森林的回归预测模型，对文件“ERα_activity.xlsx”test 表中的 50 个化合物的 pIC50 值预测，并运用公式(6.1)进行转化得到 IC50 的值，结果如

下表 6-5 所示，同时实现了 test 表中 50 组测试数据的 pIC50 预测值的可视化，如下图 6-8 所示。

表 6-5 50 个化合物活性预测值

序号	IC50_nM	pIC50	序号	IC50_nM	pIC50	序号	IC50_nM	pIC50
1	41.287	7.384	18	41.255	7.385	35	40.325	7.394
2	40.165	7.396	19	32.946	7.482	36	102.510	6.989
3	37.600	7.425	20	42.235	7.374	37	101.927	6.992
4	40.747	7.390	21	40.981	7.387	38	51.847	7.285
5	34.583	7.461	22	38.245	7.417	39	159.085	6.798
6	28.811	7.540	23	39.301	7.406	40	154.872	6.810
7	31.238	7.505	24	40.308	7.395	41	165.795	6.780
8	31.238	7.505	25	31.897	7.496	42	161.603	6.792
9	40.209	7.396	26	32.210	7.492	43	154.358	6.811
10	19.562	7.709	27	26.031	7.585	44	157.385	6.803
11	23.792	7.624	28	26.613	7.575	45	165.795	6.780
12	18.146	7.741	29	18.373	7.736	46	18.983	7.722
13	19.373	7.713	30	35.955	7.444	47	18.989	7.722
14	19.236	7.716	31	36.734	7.435	48	18.656	7.729
15	31.523	7.501	32	36.298	7.440	49	19.086	7.719
16	25.483	7.594	33	36.554	7.437	50	24.977	7.602
17	34.062	7.468	34	34.420	7.463			

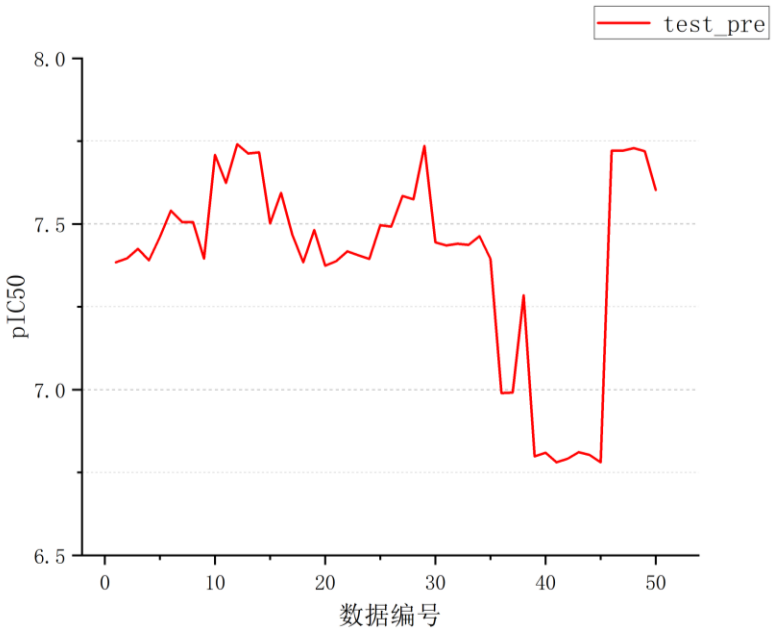

图 6-8 50 个测试数据的 pIC50 值预测走势图

6.6 问题小结

通过对比 3 种模型测试时的具体表现，可以得出：基于随机森林的回归预测模型的效果最好，所以采用该模型来解答第二问的模型，并使用该模型对文件“ERα_activity.xlsx”test 表中的 50 个化合物进行 IC50 值和对应的 pIC50 值预测，并将结果分别填入“ERα_activity.xlsx”test 表中的 IC50_nM 列及对应的 pIC50 列。

7. 问题三

7.1 问题分析

问题三要求针对文件“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，对数据进行预处理将得到 361 个特征数据作为自变量，五个化合物分类数据作为因变量。然后使用 ROC、AUC 和准确率作为评价指标；接着分别用 SVM、SGD、RF 三种模型算法构建化合物的 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型并比较它们的优劣性，确定最优模型。最后用所构建的 5 个分类预测模型，对 50 个化合物进行相应的预测。

问题三的思路流程图如图 7-1 所示：

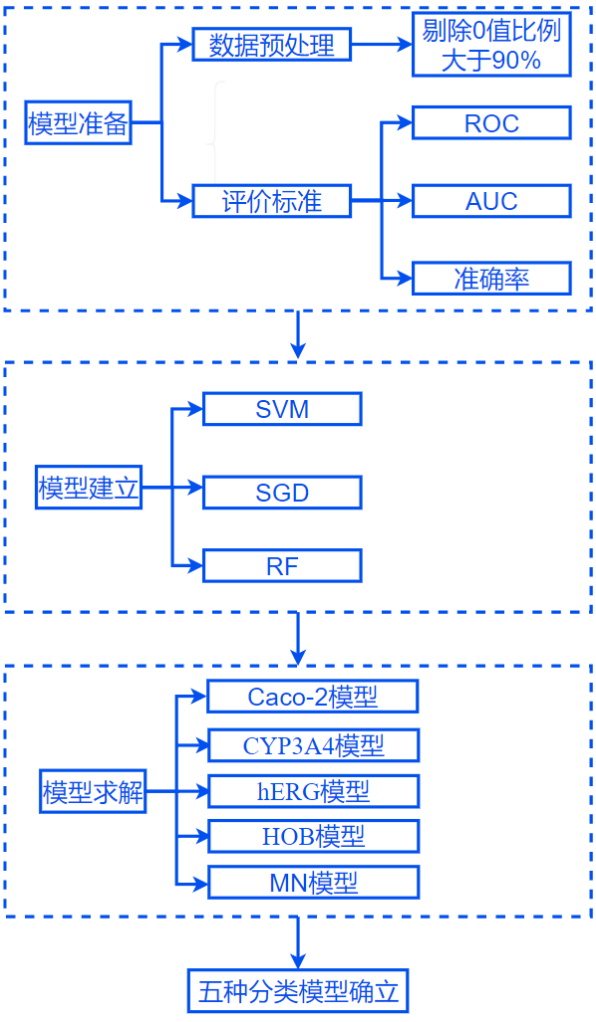

图 7-1 问题 3 流程图

7.2 模型准备

7.2.1 数据预处理

本文对附件“Molecular_Descriptor.xlsx”中提供的 1974 个化合物的 729 个分子描述符数据进行预处理，首先采用依拉达（3σ）准则，剔除特征值不在3σ范围内的异常值，提高数据的稳定性，然后对数据进行排查筛选，剔除特征中 0 值比例大于 90%的数据，该操作得到 361 个特征，数据预处理完毕。

7.2.2 评价标准

(1) ROC

ROC，亦称为受试者工作特征曲线。在二分类问题中，ROC 曲线的每一个点都代表一个阈值，分类器给每个样本一个 score，该分数若大于阈值被定义为正样本，小于阈值则被定义为负样本。

ROC 有个很好的优势：当测试集中的正负样本分布变化时，ROC 曲线能够保持不变。在实际的数据集中经常会出现分类不平衡现象，也就是负样本比正样本多很多或者少很多，而且测试数据中的正负样本分布也可能随着时间变化。ROC 曲线的特点：阈值取值越多，ROC 曲线越平滑；ROC 曲线上的点越靠近左上角越好。

(2) AUC

AUC，指的是 ROC 曲线下的面积，该指标能较好的概括不平衡样本分类器的性能，它的直观含义是任意取一个正样本和负样本，正样本得分大于负样本的概率。AUC 值为 ROC 曲线所覆盖的区域面积，显然，AUC 值越大，反映出正样本的预测结果更加靠前，分类器分类效果越好。

另外，P-R 曲线和 ROC 曲线都是用于分类的评价指标，一般情况下，P-R 曲线用于检索，而 ROC 曲线一般用于分类、识别等。由于在实际问题中，正负样本数通常很不均衡，若选择不同的测试集，P-R 曲线的变化就会非常大，而 ROC 曲线则能够更加稳定地反映模型本身的好坏。在该附件提供的 1974 个化合物的 ADMET 数据中，Caco-2、CYP3A4、hERG、HOB、MN 五种化合物出现的频数如下所示，由图 7-2 可以看出该数据集中的正负样本不均衡，所以在 P-R 曲线和 ROC 曲线的选择中，本文选择 ROC 曲线来作为评价指标之一，放弃了使用 P-R

曲线。

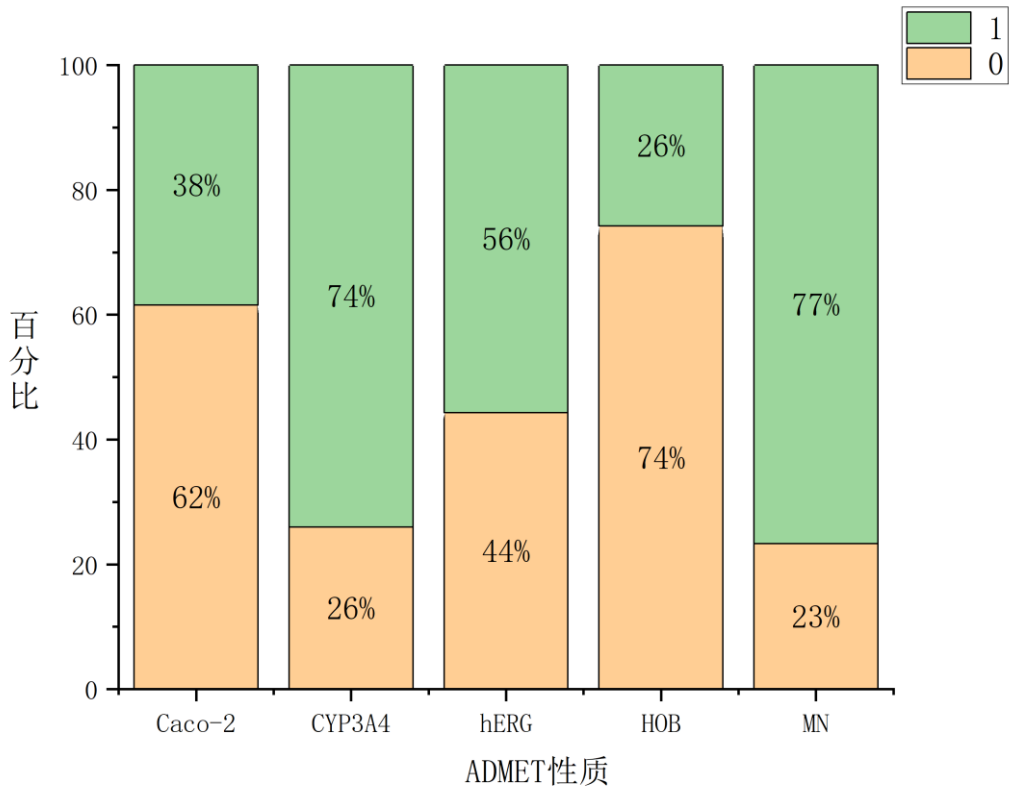

图 7-2 化合物 ADMET 性质中正负样本数据百分比图

(3) 准确率

准确率，表示被预测正确的比例，准确率是我们最常见的评价指标。对于准确率和召回率的选择，二者是矛盾的两个指标，召回率主要强调把样本中的正例样本挑出来，而准确率是强调被挑出来的样本要尽可能的准确，考虑到本题药物的使用具有很低的容错性，所以挑选的样本准确率要尽可能大，因此准确率做为本问的另一个评价指标。

可表示如下式(7.1)所示：

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7.1}$$

其中，TP 表示正确地预测为正例，TN 表示正确地预测为反例，FP 表示错误地预测为正例，FN 表示错误地预测为反例。准确率就是分类预测正确的样本数除以所有的样本数，通常来说，准确率越高，分类器越好。准确率是一个很好很直观的评价指标，但是有时候准确率高并不代表这个算法就一定好，比如在正负

样本不平衡的情况下，准确率这个指标有一定的缺陷。

7.3 模型建立

7.3.1 支持向量机

支持向量机（Support Vector Machine，SVM），SVM 是一种经典的机器学习算法，在小样本数据集的场景中有较为广泛的应用。SVM 主要是将数据进行分类，其模型思想是：假设平面中有两种类型的点，一种是红色一种是蓝色，要求我们将这两种类型的点分开，首先想到的是从两类样本中间画一条直线，但是这种情况下两类样本点离分界线都很近，如果继续输入一个新的样本点，可能会造成分错类的现象，因此要取一条最优的分界线，既能把两种类型分开，还使得两类样本点离分界线最近的点离分界线的距离尽可能远。

7.3.2 随机梯度下降法

梯度法思想的三要素：出发点、下降方向、下降步长。

随机梯度下降法（Stochastic Gradient Descent，SGD）。对于训练速度来说，SGD 一次迭代只用一条随机选取的数据，虽然迭代次数会很多，但是一次学习时间非常快；对于准确度来说，SGD 仅仅用一个样本决定梯度方向，导致求解可能不是最优；对于收敛速度来说，SGD 一次迭代一个样本，导致迭代变化很大，不能很快的收敛到局部最优解。不过，如果目标函数有盆地区域，SGD 会使优化的方向从当前的局部极小值点跳到另一个更好的局部极小值点，对于非凸函数最终收敛于一个较好的局部极值点，甚至全局极值点。

随机梯度下降法的公式归结为通过迭代计算特征值从而求出最合适的值，θ 的求解见公式(7.2)：

$$\theta = \theta - \alpha \frac{\partial J(\theta)}{\partial(\theta)} \tag{7.2}$$

其中，α 是下降系数，即步长，学习率，通俗的说就是计算每次下降的幅度大小，系数越大每次计算的差值越大，系数越小则差值越小，但是迭代计算的时间也会延长，θ 的初值可以随机赋值。

7.3.3 随机森林

随机森林（Random Forest，RF），是一种新兴起的、高度灵活的、基于树的

机器学习算法，该算法利用多棵树的力量来进行决策。森林中的每棵树并不一样，每棵树都是被随机创造的，每棵树中的每一个节点都是待选特征的一个随机子集，所有树的输出结果整合起来就是森林最后的输出结果。

7.4 模型求解

对于模型的求解，分别对五类化合物进行四种模型的求解、评价和预测。求解使用挑选的 361 个变量作为自变量，五个化合物各自的分类数据作为因变量。模型的评价部分，分别用 ROC、AUC、准确率和模型训练时间四个指标作为评价指标确定五个化合物各自的最佳模型。最后，对给出的 50 组测试集数据进行预测。

7.4.1 Caco-2 分类模型求解

使用 python 对三种模型（SVM、SGD、RF）实现在化合物 Caco-2 变量下的 ROC 曲线绘制，对比图如下图 7-3 所示：

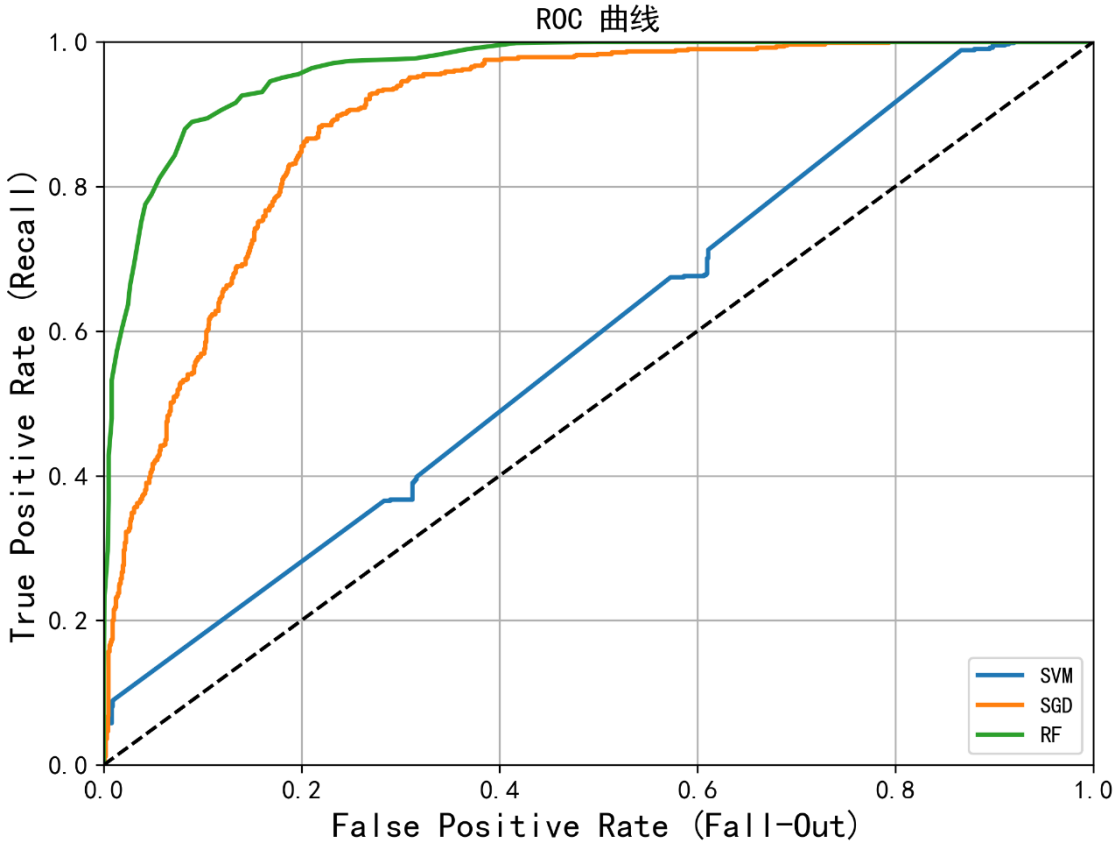

图 7-3 Caco-2 在三种预测模型下的 ROC 曲线对比图

由图 7-3 可以观察出，随机森林 RF 的 ROC 曲线最靠左，AUC 最大，预测效

果最好。而支持向量机 SVM 的 ROC 曲线最靠右，AUC 值最小，预测效果最差。根据上述评价标准中的评价指标对化合物 Caco-2 在三种模型下的 AUC 值、训练集准确率以及测试集准确率进行计算，得到结果如下表 7-1 所示：

表 7-1 Caco-2 在三种模型下的评价指标结果

Caco-2	AUC 值	训练集准确率	测试集准确率
SVM	0.5887	0.6428	0.6076
SGD	0.8934	0.8036	0.7689
RF	**0.9619**	**0.8935**	**0.8455**

根据表 7-1 我们可以看出 RF 的 AUC 值最高，训练集准确率与测试集准确率最高，用随机森林算法构建化合物 Caco-2 的分类预测模型最优。

7. 4. 2 CYP3A4 分类模型求解

使用 python 对三种模型（SVM、SGD、RF）实现在化合物 CYP3A4 变量下的 ROC 曲线绘制，对比图如下图 7-4 所示：

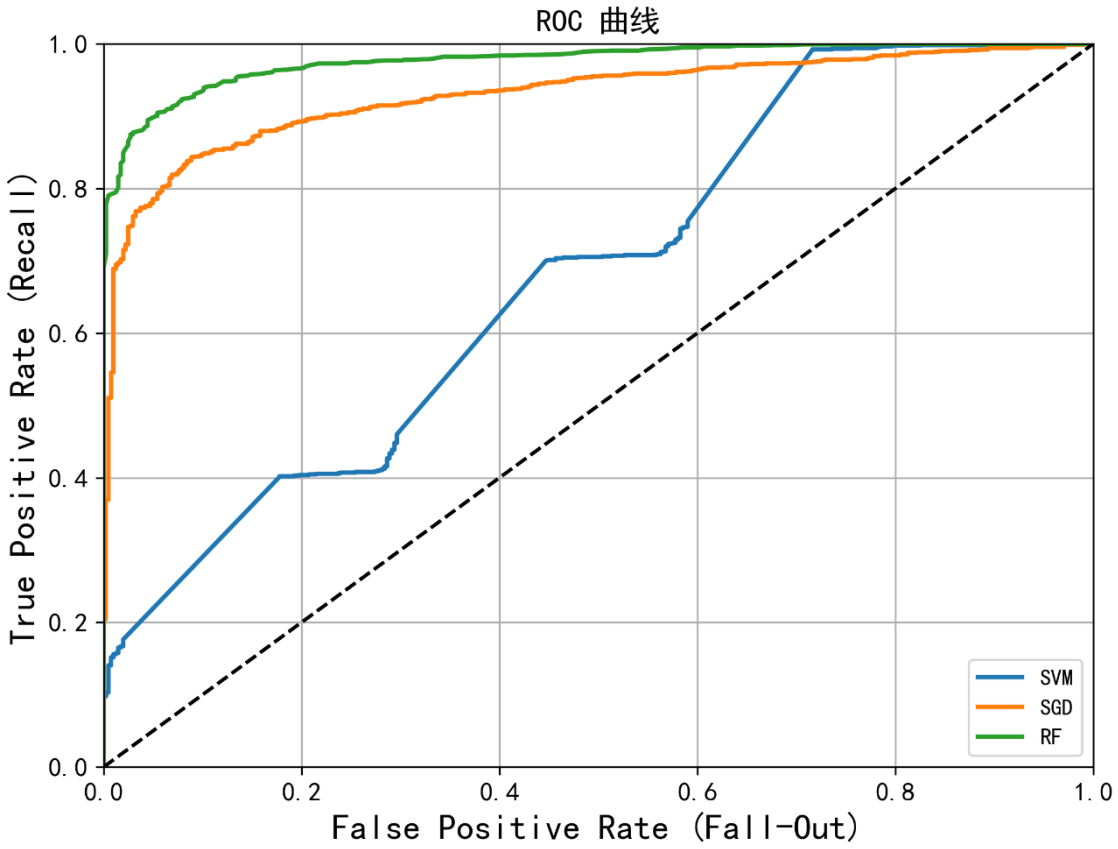

图 7-4 CYP3A4 在三种预测模型下的 ROC 曲线对比图

由图 7-3 可以观察出，随机森林 RF 的 ROC 曲线最靠左，AUC 最大，预测效

果最好。而支持向量机 SVM 的 ROC 曲线最靠右，AUC 值最小，预测效果最差。根据上述评价标准中的评价指标对化合物 CYP3A4 在三种模型下的 AUC 值、训练集准确率以及测试集准确率进行计算，得到结果如下表 7-2 所示：

表 7-2 CYP3A4 在三种模型下的评价指标结果

CYP3A4	AUC 值	训练集准确率	测试集准确率
SVM	0.6917	0.7701	0.7417
SGD	0.9290	0.8575	0.7751
RF	**0.9754**	**0.9322**	**0.9007**

根据表 7-2 我们可以看出 RF 的 AUC 值最高，训练集准确率与测试集准确率最高，用随机森林算法构建化合物 CYP3A4 的分类预测模型最优。

7. 4. 3 hERG 分类模型求解

使用 python 对三种模型（SVM、SGD、RF）实现在化合物 hERG 变量下的 ROC 曲线绘制，对比图如下图 4-1 所示：

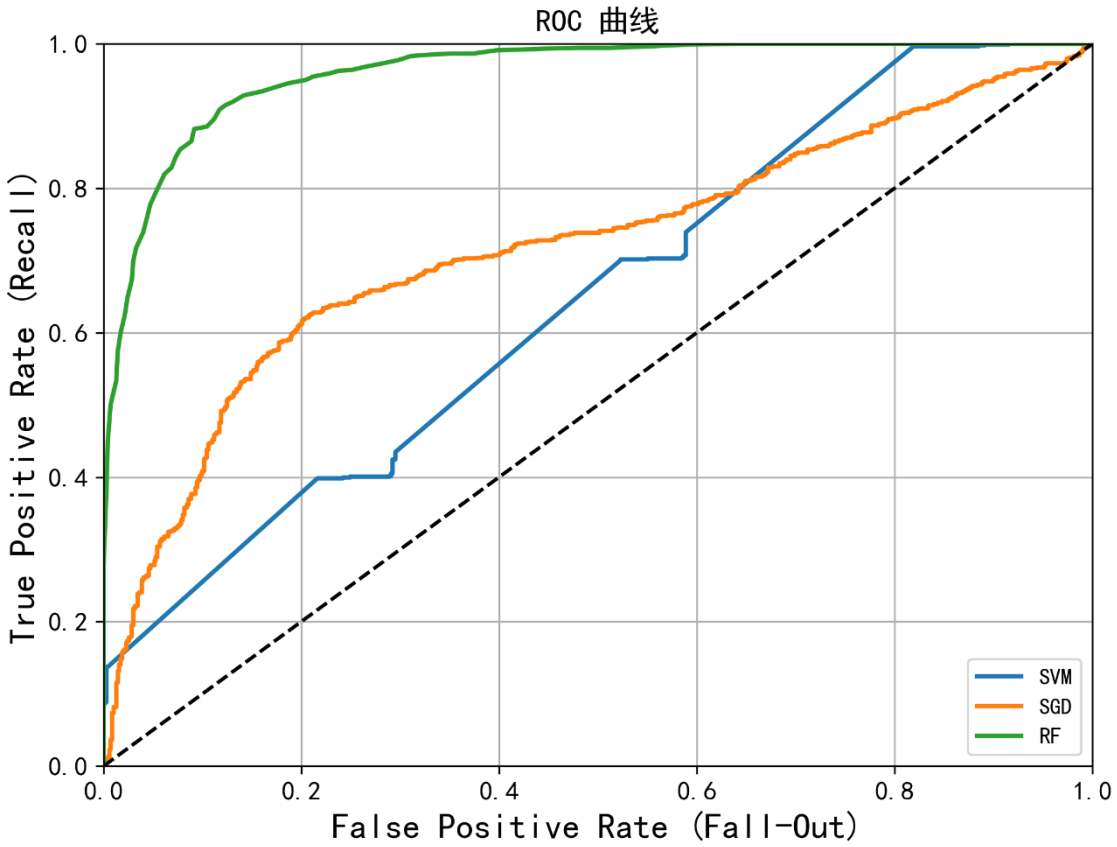

图 7-5 hERG 在三种预测模型下的 ROC 曲线对比图

由图 7-5 可以观察出，随机森林 RF 的 ROC 曲线最靠左，AUC 最大，预测效

果最好。而支持向量机 SVM 的 ROC 曲线最靠右，AUC 值最小，预测效果最差。根据上述评价标准中的评价指标对化合物 hERG 在三种模型下的 AUC 值、训练集准确率以及测试集准确率进行计算，得到结果如下表 7-3 所示：

表 7-3 hERG 在三种模型下的评价指标结果

hERG	AUC 值	训练集准确率	测试集准确率
SVM	0.6448	0.6054	0.5620
SGD	0.7180	0.7808	0.6535
RF	**0.9600**	**0.8955**	**0.8430**

根据表 7-3 我们可以看出 RF 的 AUC 值最高，训练集准确率与测试集准确率最高，用随机森林算法构建化合物 hERG 的分类预测模型最优。

7.4.4 HOB 分类模型求解

使用 python 对三种模型（SVM、SGD、RF）实现在化合物 HOB 变量下的 ROC 曲线绘制，对比图如下图 7-6 所示：

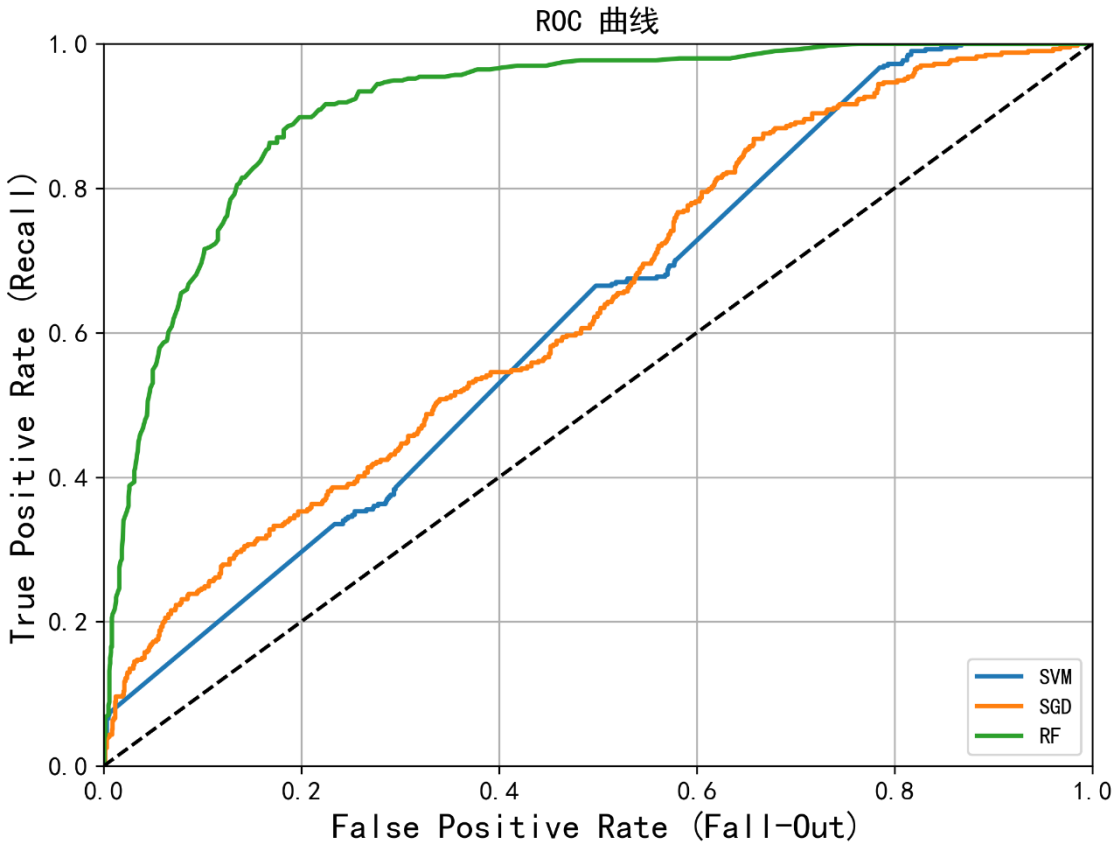

图 7-6 HOB 在三种预测模型下的 ROC 曲线对比图

由图 7-6 可以观察出，随机森林 RF 的 ROC 曲线最靠左，AUC 最大，预测效

果最好。而支持向量机 SVM 和 SGD 的 ROC 曲线更靠右，AUC 值比较小，预测效果比较差。根据上述评价标准中的评价指标对化合物 HOB 在三种模型下的 AUC 值、训练集准确率以及测试集准确率进行计算，得到结果如下表 7-4 所示：

表 7-4 HOB 在三种模型下的评价指标结果

HOB	AUC 值	训练集准确率	测试集准确率
SVM	0.6153	0.7612	0.7139
SGD	0.6368	0.6649	0.6108
RF	**0.9112**	**0.8625**	**0.8023**

根据表 7-4 我们可以看出 RF 的 AUC 值最高，训练集准确率与测试集准确率最高，用随机森林算法构建化合物 HOB 的分类预测模型最优。

7.4.5 MN 分类模型求解

使用 python 对三种模型（SVM、SGD、RF）实现在化合物 MN 变量下的 ROC 曲线绘制，对比图如下图 7-7 所示：

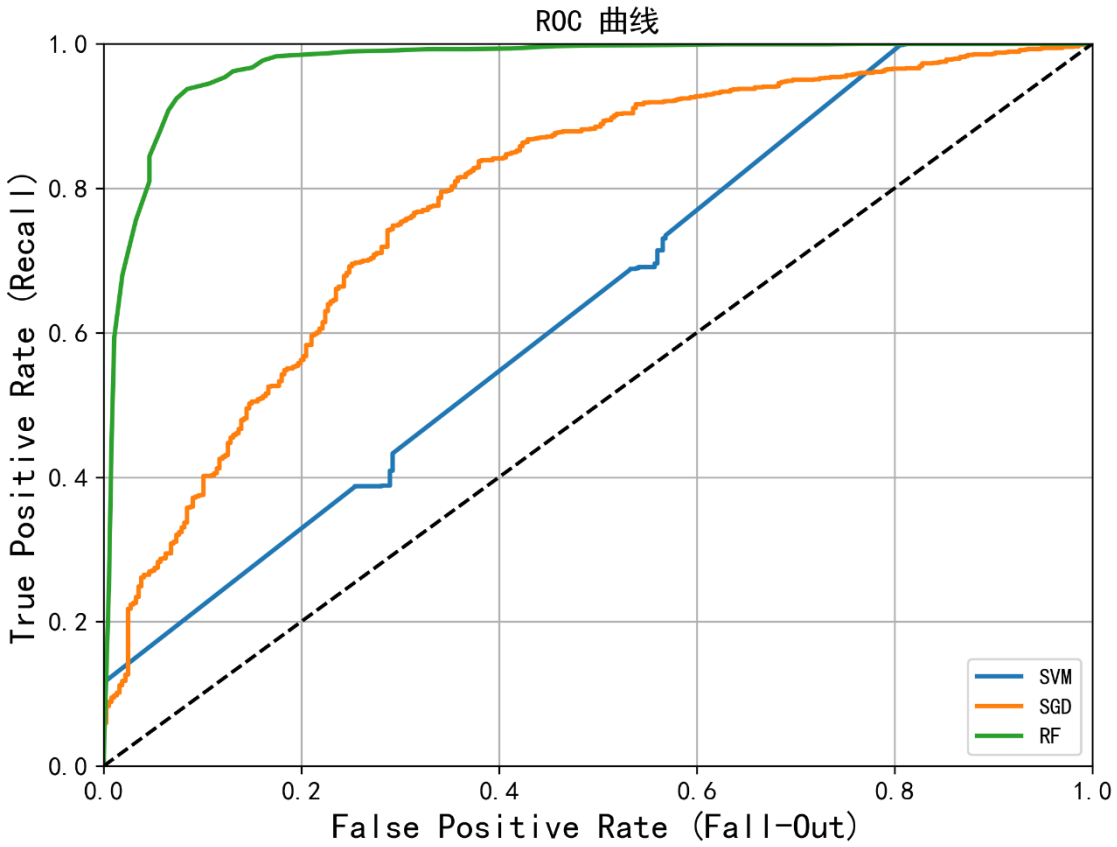

图 7-7 MN 在三种预测模型下的 ROC 曲线对比图

由图 7-7 可以观察出，随机森林 RF 的 ROC 曲线最靠左，AUC 最大，预测效

果最好。而支持向量机 SVM 的 ROC 曲线最靠右，AUC 值最小，预测效果最差。根据上述评价标准中的评价指标对化合物 MN 在三种模型下的 AUC 值、训练集准确率以及测试集准确率进行计算，得到结果如下表 7-5 所示：

表 7-5 MN 在三种模型下的评价指标结果

MN	AUC 值	训练集准确率	测试集准确率
SVM	0.6381	0.8087	0.7747
SGD	0.7846	0.8005	0.7001
RF	**0.9717**	**0.9455**	**0.9117**

根据表 7-5 我们可以看出 RF 的 AUC 值最高，训练集准确率与测试集准确率最高，用随机森林算法构建化合物 MN 的分类预测模型最优。

7.5 模型确定及预测

通过 SVM、SGD、RF 四种模型算法分别构建对化合物 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型，在模型求解一节中我们可以清晰地观察到每种化合物在四种模型下的评价指标对比，均是 RF 算法模型的 ROC 曲线靠近左侧，且均是 RF 算法模型的 AUC 值、训练集准确率以及测试集准确率最高；由于 ROC 曲线越靠近左侧，说明 AUC 值越大，表示分类效果越好，并且 RF 算法模型的训练集准确率和测试集准确率都高于其他三种模型，因此确定用随机森林算法模型作为构建化合物的 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型。最后，用上述构建的 5 个分类预测模型，对文件“ADMET.xlsx”的 test 表中的 50 个化合物进行相应的预测，预测结果如下表 7-6 所示：

表 7-6 RF 算法模型对 50 个化合物进行的预测结果

序号	Caco-2	CYP3A4	hERG	HOB	MN	序号	Caco-2	CYP3A4	hERG	HOB	MN
1	0	1	1	0	1	26	1	1	1	1	1
2	0	1	1	0	1	27	0	0	1	1	0
3	0	1	1	0	1	28	0	0	1	1	0
4	0	1	1	0	1	29	0	0	1	1	0
5	0	1	1	0	1	30	0	0	1	1	0
6	0	1	1	0	1	31	0	0	1	1	1

7	0	1	1	0	1	32	0	0	1	1	1
8	0	1	1	0	1	33	0	0	1	1	1
9	0	1	1	0	1	34	0	0	1	1	1
10	0	1	1	0	1	35	0	0	1	1	1
11	0	1	1	0	1	36	0	0	1	0	0
12	0	1	1	0	1	37	0	0	1	0	0
13	0	1	1	0	1	38	0	0	1	1	0
14	0	1	1	0	1	39	0	0	1	1	0
15	0	1	1	0	1	40	0	0	1	1	0
16	0	1	1	0	1	41	0	0	1	1	0
17	0	1	1	0	1	42	0	0	1	1	0
18	0	1	1	0	1	43	0	0	1	1	0
19	0	1	0	0	1	44	0	0	1	1	0
20	0	0	0	0	1	45	0	0	1	1	0
21	0	1	1	0	1	46	0	0	1	1	0
22	0	1	1	0	1	47	0	0	1	1	0
23	1	0	1	0	0	48	0	0	1	1	0
24	1	0	1	0	0	49	0	0	1	1	0
25	1	1	1	0	0	50	0	0	1	1	0

7. 6 问题小结

问题三用随机森林算法构建化合物 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型，最后对文件“ADMET.xlsx”的 test 表中的 50 个化合物进行相应的预测并列出预测结果。

8. 问题四

9. 模型的评价与改进

参考文献

[1] 郑莹. 中国乳腺癌患者生活方式指南[J]. 中华外科杂志, 2017.

[2] Wenzel, J., Matter, H. and Schmidt, F. (2019) Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. Journal of Chemical Information and Modeling, 59, 1253-1268.

附录