



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

南京信息工程大学

参赛队号

21103000002

队员姓名

1. 费建伟

2. 戴昀书

3. 周文浩

**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   **两阶段抗乳腺癌候选药物特性预测与优化建模**

**摘要：**

乳腺癌是女性最多发的癌症，也是死亡最多的癌症。预防与治疗乳腺癌对提升女性整体健康有着非常积极的作用。在相关药物的研发中，为了节约时间与成本，通常建立化物结构-活性关系预测模型来预测化合物分子针对靶标（ER $\alpha$ ）的活性数据，或优化其结构。本文建立了**两阶段抗乳腺癌候选药物特性预测与优化模型**，并给出了相关化合物分子描述符同时具有更好 ADMET 性质以及生物活性时的取值。

对于问题一，首先进行数据清洗，依次剔除了方差为 0 的变量，过滤了部分存在缺失值的变量。接着，对清洗后的数据与 pIC50 分别进行 **Z-score 标准化** 和 **0-1 归一化**。分别采用**灰色关联分析**、**最大互信息系数**、**随机森林** 与 **平均精度损失** 计算分子描述符变量对 ER $\alpha$  活性影响重要度，并将结果相融合，以此提出了**基于加权集成重要度分析** 的变量筛选法。同时，考虑了变量间可能存在的**线性相关** 和 **非线性相关性**，对高重要度变量进行独立性分析并利用**自适应迭代变量选择法** 地进行筛选，从而获得最终的 20 个具有**代表性**、**独立性**的变量。

对于问题二，由第一问已得到 20 个对生物活性最重要的变量，直接作为该问的模型输入。首先对训练集进行 **9:1 随机划分**，前者用于算法训练，后者用于验证算法准确率。本文全面评估了**核岭回归**、**支持向量机回归**、**多层感知机**、**梯度提升回归**、**随机森林回归** 以及 **XGBoost** 算法在预测生物活性上的表现，并通过 4 种评估标准自适应地选择了最优的 **XGBoost** 模型，作为本问以及后续的基础模型。**多次实验表明**，收敛后的模型在随机划分的测试上的均方根误差低于 0.07，同时 R2 Score 超过了 0.72。

对于问题三，为了同时满足分子描述符变量对生物活性与 ADMET 具有高重要度的要求，因此重新构建了特征选择方法，构建了**两阶段 XGBoost-MLP 特征选择与并行预测方法**。在阶段一中**借助 XGBoost 选择同时面向 ADMET 与生物活性的最优变量**，在阶段二中，在此基础上构建了五个分别预测 ADMET 的基于 MLP 模型。在随机划分的测试集上，本文方法对五种特性的预测精度均超过了 80%，AUC 均超过了 0.9。

对于问题四，提出了**基于 DE-MLP 的全局优化搜索算法**。首先构建并训练了六个独立的 MLP，分别预测 ADMET 以及 ER $\alpha$  水平。在此基础上，将其作为 DE（差分进化）的目标函数，通过不断的交叉编译，**搜索能够使 ADMET 总体较优，且 ER $\alpha$  水平较好的化合物分子变量取值范围**，为药物研发提供了一定的帮助。

关键词：自适应迭代变量选择，两阶段活性预测，药物特性预测

# 目 录

<b>一、问题重述</b>	4
1.1 问题背景	4
1.2 需要解决的问题	5
<b>二、模型假设</b>	6
<b>三、符号说明</b>	7
<b>四、解题流程图</b>	8
<b>五、问题一</b>	9
5.1 问题分析	9
5.2 数据清洗	9
5.2.1 变量全部为 0 的情况	9
5.2.2 变量部分为 0 的情况	10
5.3 建立自适应迭代变量选择模型	11
5.3.1 数据预处理	11
5.3.2 变量筛选阶段	11
5.3.3 独立性检验阶段	18
5.4 结果展示	21
5.5 小结	22
<b>六、问题二</b>	22
6.1 问题分析	22
6.2 模型建立	23
6.2.1 优化目标	23
6.2.2 单模型算法	23
6.2.3 集成模型算法	26
6.3 模型求解	28
6.3.1 模型参数设置	28
6.3.2 数据集准备与评价指标	29
6.4 实验结果	30
6.4.1 模型性能分析	30
6.4.2 模型预测稳定性分析	30
6.4.3 结果展示	32
6.5 小结	33
<b>七、问题三</b>	33
7.1 问题分析	33
7.2 模型建立	34
7.2.1 基于两阶段的 XGboost-MLP 生物活性特征预测模型	34
7.3 模型求解	35
7.3.1 模型参数	35
7.3.1 评价指标	35
7.2.1 结果分析与对比	35
<b>八、问题四</b>	38
8.1 问题分析	38
8.2 优化模型建立	38

8.2.1 制定优化目标 .....	38
8.2.2 基于 DE-MLP 的分子描述变量控制算法 .....	39
8.2.1 差分进化算法 .....	错误!未定义书签。
8.2 模型求解 .....	41
8.2.2 模型参数设置 .....	41
8.2.2 模型结果展示 .....	42
九、模型的评价 .....	42

## 一、问题重述

### 1.1 问题背景 问题背景，查阅相关文献，从背景引入本问题

癌症严重危害人类健康。2021 年 2 月，国际癌症研究机构发布了对全球 185 个国家、36 种癌症的发病和死亡情况，相比于 2018 年的统计数据，仅对女性而言，乳腺癌位列目前全球女性癌症发病和死亡的第一顺位<sup>[1]</sup>。在整体中国人群中，女性乳腺癌为第四大发病原因（9.1%），仅次于肺癌（17.9%）、结直肠癌（12.2%）和胃癌（10.5%）。相比死亡病例数在全球的排位，中国乳腺癌死亡病例占比 3.9%，在全球癌症死亡原因中居第七位<sup>[2]</sup>。由于亚洲人口众多（占全球总人口的 59.5%），2020 年有 49.3% 的新发病例和 58.3% 的癌症死亡发生在亚洲，欧洲和美洲分列第二、三位。女性乳腺癌在 159 个国家中发病率居首位，在 110 个国家中为死亡的主要原因。其中发病率最高的是澳大利亚/新西兰（95.5/10 万），最低的是中亚南部（26.2/10 万）。女性乳腺癌死亡率最高的地区为美拉尼西亚（27.5/10 万），死亡率最低的地区是 东亚（9.8/10 万）。该调查根据统计学特性对 2020~2024 年亚洲地区女性乳腺癌发病和死亡人数做出了预测，如图 1 所示。

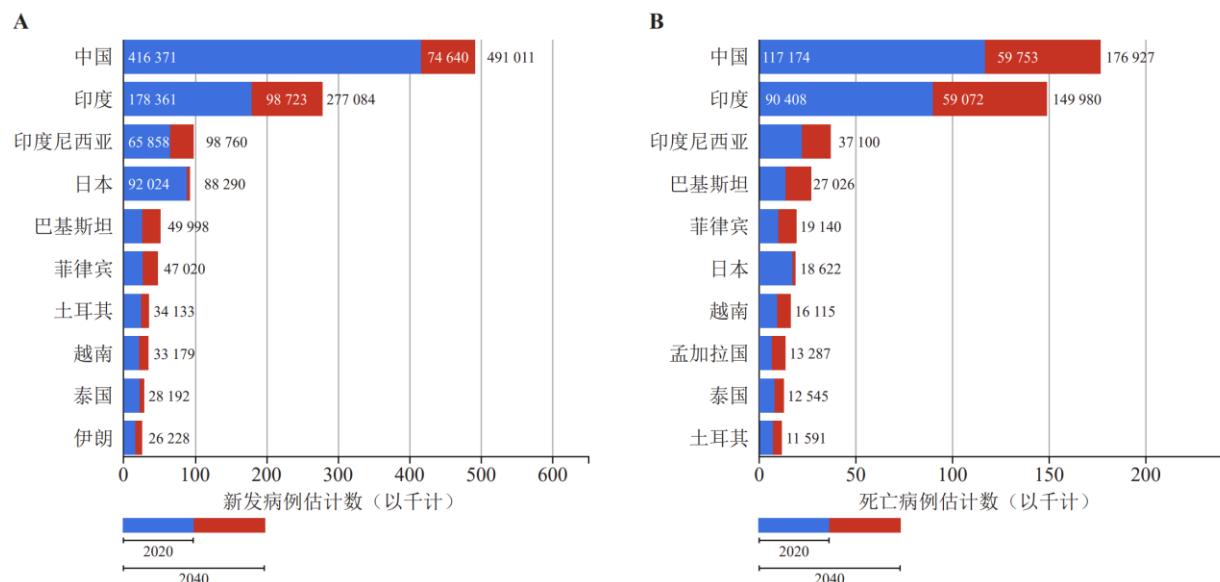


图 1 亚洲部分地区国家女性乳腺癌发病和死亡人数 2020~2024 时间变化趋势预测<sup>[2]</sup>

随着我国的人口增长和老龄化问题，乳腺癌疾病负担不断加重，发病率和死亡率均居于世界前列。我国同时表现出城乡差异大、地区分布不均衡的现状，开展癌症防治工作必要且具有挑战。

乳腺癌是激素依赖性肿瘤，癌细胞的生长受体内多种激素的调控。根据激素受体不同，乳腺癌被分为四种亚型，包括孕激素受体(PR)阳性、雌激素受体(ER)阳性、人类表皮生长因子受体(HER2)阳性及三阴性乳腺癌。据统计，约 70% 的乳腺癌患者表现为雌激素受体 ER $\alpha$ (Estrogen receptors alpha, ER $\alpha$ )阳性。通过调控 ER $\alpha$  的转录活性或者 ER $\alpha$  蛋白的降解调节雌激素受体活性来控制体内雌激素水平，达到抗肿瘤的效果。ER $\alpha$  已经成为治疗乳腺癌最有效的靶点之一。

在目前的药物研发过程中，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。一个化合物想要成为乳腺癌候选靶向药物，除了需要具备良好指抗乳腺癌活性外，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。由于各个变量

的药理特性十分复杂，它们之间可能存在高度非线性和相互强耦合的关系，所以探索各种化合物样本性质，进行生物活性的定量预测和 ADMET 性质的分类预测，具有非常重要的意义。

## 1.2 需要解决的问题

基于上述研究背景，本文需要研究完成以下问题：

### 问题一（数据处理）：

根据题目要求，针对 1974 个化合物的 729 个分子描述符，根据分子描述符对化合物生物活性水平的影响程度对其进行重要度排序，需要筛选前 20 个最重要的分子描述符，并给出依据。题中未给出分子描述符间独立性的说明，因此结题过程中不仅考虑变量的代表性，同时需要考虑变量的独立性。另外，数据中同时包括离散数据以及连续数据，如何处理此类混合数据也需要仔细考虑。灰色关联度分析。通过计算分子描述符变量与化合物生物活性水平的相关性，筛选关键变量。

### 问题二（建立化合物对 ER $\alpha$ 生物活性定量预测模型）：

根据问题一中得出的变量对生物活性影响的重要性排序，选择不超过 20 个分析描述符变量，构建化合物对 ER $\alpha$  生物活性的定量预测模型，对文件“ER $\alpha$ \_activity.xlsx”的 test 表中的 50 个化合物进行 IC50 值和对应的 pIC50 值预测，并记录结果。

### 问题三（建立化合物分类预测模型）：

利用题中提供的 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 的分类预测模型，并使用这 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，并记录结果。

### 问题四（高生物活性高 ADMET 方案优化）：

根据上述三个问题的模型与结果，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 ER $\alpha$  具有更好的生物活性，同时具有更好的 ADMET 性质。

提供的数据见表 1-1

表 1-1 资料目录

资料名称	资料内容
附件一： ER $\alpha$ _activity.xlsx	提供了 1974 种化合物样本 IC50_nM 和 pIC50 数据
附件二： Molecular_Descriptor.xlsx	提供了每个化合物样本的 729 个分子描述符变量
附件三： 分子描述符含义解释.xlsx	提供不同分子描述符的解释
附件四： ADMET.xlsx	提供每个化合物样本的 5 种 ADMET 性质数据

## 二、模型假设

为有效解决上述所有问题，做出如下假设：

**1、变量相关性假设。**根据“分子描述符含义解释”，附件“Molecular\_Descriptor”中所提供的 729 个分子描述符共包含 53 种类别的变量，因此可能存在变量间具有不同程度相关性的情况。在变量选择的过程种，需要注意筛选具有高相关性的变量。

**2、变量缺失假设。**变量缺省/附件“Molecular\_Descriptor”中存在大量全部为 0的变量，考虑此类变量为缺失变量。另外，存在少量部分 0 值且其余值为浮点数的非离散型变量，查阅“分子描述符含义解释”，可知此类变量不属于离散变量，因此考虑为数据缺失。由于数据填充的困难以及可能的负面影响，对以上两类变量进行直接剔除，具体流程在文中详细说明。

**3、预测目标假设。**附件“ER $\alpha$ \_activity”中的 IC50\_nM 以及 pIC50 具有不同的数值区间以及波动程度，pIC50 作为预测目标更加稳定，具体结果在文中详细说明。

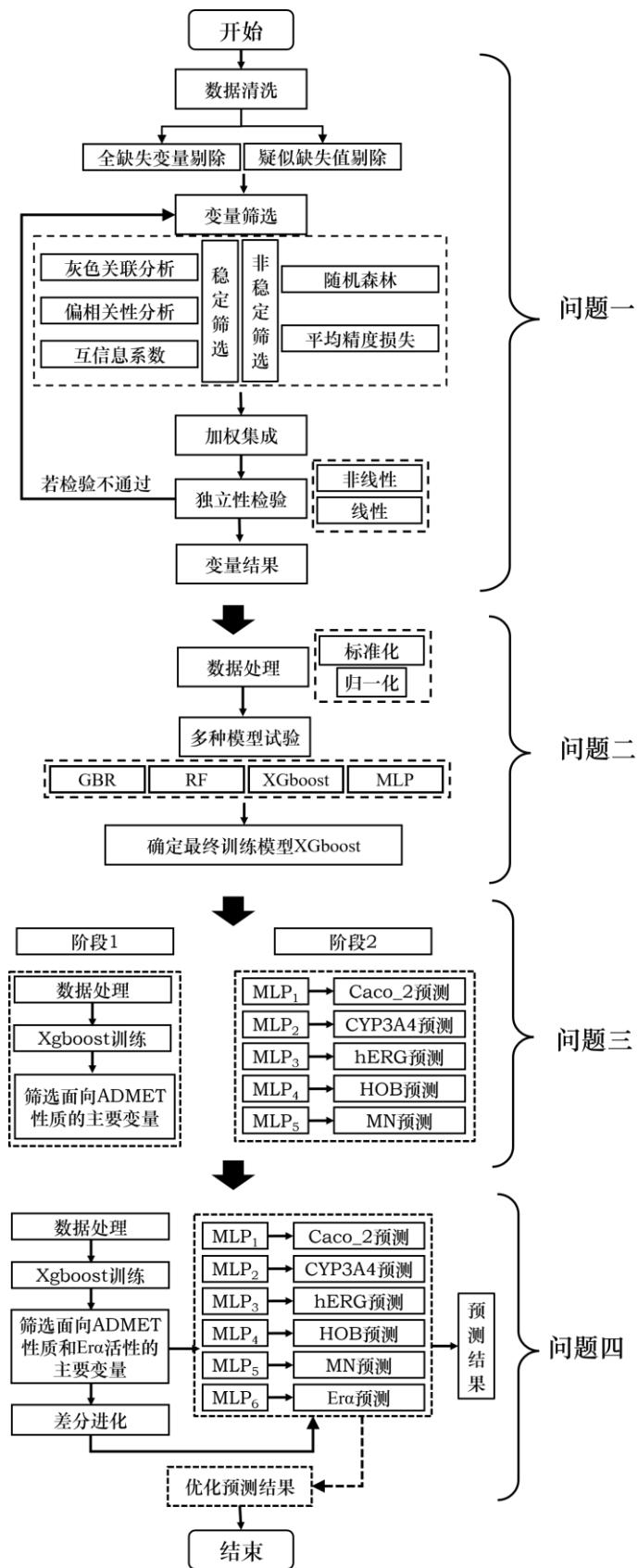
**4、ADMET 独立假设。**假设附件“ADMET”中的五种化合物特性相互独立。因此在同时优化 ADMET 性质时，其相互不影响。

### 三、符号说明

表 3-1 模型符号分析与说明

符号	意义
$X$	全体样本集
$X_{train}$	训练集
$X_{test}$	测试集
$N$	分子描述符变量数量
$\alpha$	变量中 0 值数量占总样本数的比重
$\beta$	变量中不重复的非 0 值的种类
$\gamma$	变量中出现次数第 2 多的数值的数量占总样本数的比重
$f$	基于统计的变量选择算法
$\varphi$	基于模型学习的变量选择算法
$\phi$	预测模型
$\chi^2$	卡方检验统计值
$F$	误差函数
$J_w$	核岭回归目标函数
$\Phi$	核函数

## 四、解题流程图



## 五、问题一

### 5.1 问题分析

根据题目要求，针对 1974 个化合物的 729 个分子描述符变量（以下简称变量），根据分子描述符对化合物生物活性水平的影响程度对其进行重要度排序，需要筛选前 20 个最重要的分子描述符，并给出依据。首先，虽然题中未明确要求数据清洗，但通过分析，可以发现数据存在清洗的必要，因此在挑选变量前进行清洗。其次，虽然题中未明确变量相关的特性，附件表明存在可能，因此还需要独立性检验剔除高相关变量。最终筛选出 20 个最具有代表性、独立性的重要变量。

问题一的总体思路如图 5.1 所示，主要可分为三个部分：

- (1) 数据清洗，筛选并剔除缺失变量；
- (2) 基于集成自适应筛选的重要度分析，选择前 20 个主要变量；
- (3) 变量相关性分析筛选，对(2)中选择的主要变量进行相关性分析与独立性检验，剔除高度相关的变量，并根据(2)的结果选择新的候补变量。

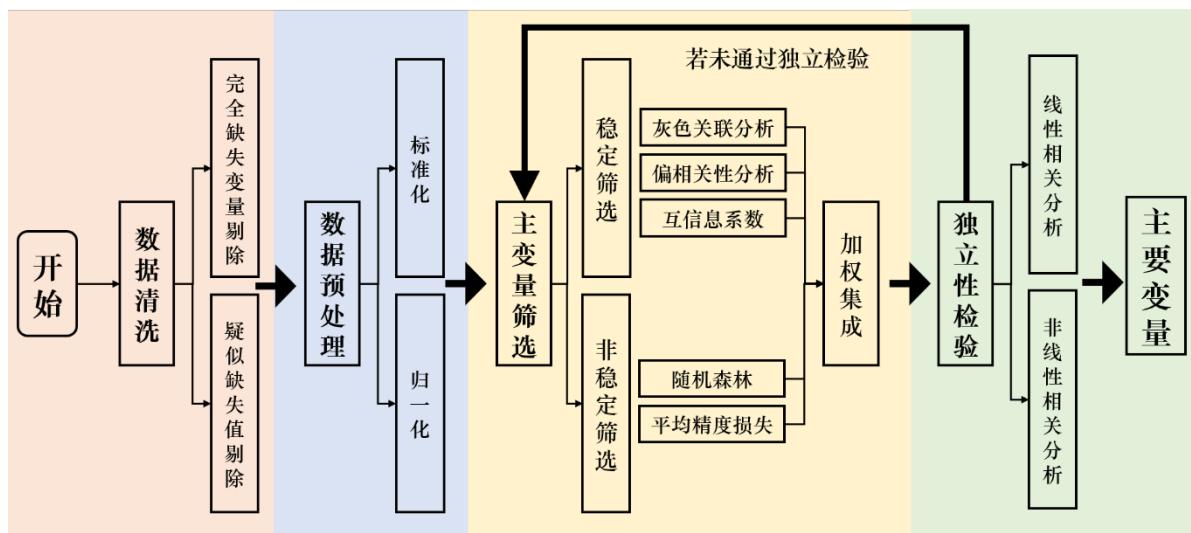


图 5.1 问题一思路流程图

### 5.2 数据清洗

通过分析附件二“Molecular\_Descriptor”与附件三“分子描述符含义解释”中的数据，我们发现了数据集的两个特性，如下：

- (1) 变量存在离散型与连续型两种类型，如变量“naAromAtom”属于离散型变量，意义为 Number of aromatic atoms；而变量 AMR 则属于连续型变量，意义为 Molar refractivity。
  - (2) 变量共有 53 种类型，每个类型包含的变量数量从 1 到 488 种不等。
- 基于这两个特性，下文我们进行了数据清洗和变量筛选。

#### 5.2.1 变量全部为 0 的情况

原始附件“Molecular\_Descriptor”中提供了 1947 条数据，每条数据中包含 729 个变量。通过遍历分析法对该附件进行分析后，我们发现 729 个变量中有 225 个在 1947 条数据上的值均为 0（方差为 0），认为该变量数据完全缺失。所采用的遍历法剔除异常数据伪代码

## 即便是最简单的遍历，也需要详细的描述算法过程

如下：

算法 1 遍历法寻找全 0 变量

---

### 算法 1 遍历法寻找全 0 变量

---

输入：数据集  $X \in \mathbb{R}^{M \times N}$

输出：缺失变量  $id$  与名称

```
for n from 1 to N:          #遍历 N 个变量
    zero_count = 0           #初始化第 n 个变量包含 0 的数量
    for m from 1 to M:       #遍历 M 个变量
        if Xm,n = 0:         #判断该变量值是否为 0
            zero_count = zero_count + 1   # 第 n 个变量 0 值数量+1
    if zero_count = N:        # 判断第 n 个变量是否全为 0
        id.append(n)           # 若是，将第 n 个变量剔除
```

---

根据搜寻到的全 0 变量 ID，对原始数据进行剔除，得到剩余 504 个变量。虽然存在一种可能，即部分数据全部为 0 的变量属于离散型变量，数值为 0 确实存在实际意义，但考虑到其数值完全相同，不具有作为因变量的基本属性，因此这里仍然进行剔除。

### 5.2.2 变量部分为 0 的情况

通过分析数据集，我们发现部分变量具有大量的 0，但考虑到前文（5.2）提及的属性（1），需要仔细考虑这些变量中 0 的意义，即这些变量是否属于离散变量，或者作为连续变量时 0 值是否具有实际意义。结合附件观察数据，可以发现部分变量具有以下属性：

- ① 有一定数量的值为 0；
- ② 属于连续型变量；
- ③ 非 0 值的种类极多。

结合以上 3 个特点，可以认为此类变量中的 0 值为缺失值。由于信息不全导致的数据填充困难以及不精确填充可能带来的负面影响，因此这里对该类变量直接进行剔除处理。各个变量进行剔除处理的指标和意义如表 5.3 所示：

表 5-1 变量的剔除指标

指标	意义
变量中 0 值数量占总样本数的比重不低于 $\alpha$	缺失值比例足够高
变量中不重复的非 0 值的种类不少于 $\beta$	变量为离散变量且离散度高
变量中出现次数第二多的数值数量占总样本数的比重不低于 $\gamma$	0 值属于异常值

实验筛选结果如表 5.2。在实际实验中，通过观察发现这三个指标被设置为  $\alpha=20\%$ 、 $\beta=50$  以及  $\gamma=3\%$  时可以较好的筛选出缺失的 0 变量，且能够防止具有大量真实值为 0 的变量误检，最终使用的变量数为 252 个。

表 5-2 不同筛选指标下筛选数量

$\alpha$	$\beta$	$\gamma$	剔除变量数	总剔除变量数	剩余变量数	误检变量数
10%	50	3%	241	466	250	2
10%	50	5%	250	475	254	7
10%	99	3%	241	466	263	4
10%	99	5%	250	475	254	7

20%	50	3%	241	466	263	5
20%	50	5%	248	473	256	3
20%	99	3%	241	466	263	9
20%	99	5%	245	470	259	15

### 5.3 建立自适应迭代变量选择模型

根据问题一的要求，需要在 729 个变量中选择出对生物活性影响最显著的 20 个变量，即变量与生物活性的关联度要大。由于变量的筛选是为了降低特征维数，为后续的学习算法提供可靠有效的特征，因此在结合多种关联度计算方式的同时，以基于学习的特征选择方法为主，基于统计的方法为辅，设计了集成变量筛选模型。

同时，在考虑关联度时，不仅要考虑变量与生物活性的线性关联度，还要考虑它们的非线性关联度。另外，根据数据集的特性，还需要注意所挑选变量间的独立性。因此筛选变量的过程分为两部分，（1）挑选主要变量；（2）独立性检验，如表 5-3 所示：

表 5-3 主要变量筛选流程

- 
- ① 通过模型集成投票机制选择对生物活性影响最大的 20 个主要变量
  - ② 对选择的主要变量进行独立性分析，删选高相关的冗余变量
  - ③ 选择（1）中的候补的主要变量，添加进主要变量集合中
  - ④ 迭代以上步骤，直到选择出 20 个最具有代表性、独立性的变量
- 

#### 5.3.1 数据预处理

经过 5.2 中的数据清洗，共剩余 263 个变量，这些变量的类型（离散/连续）、取值区间（量纲）、变化程度各不相同，且差异较大。因此在筛选变量时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-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-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：

$$X_i^n = \frac{X_i^n - \mu}{\sigma}, \quad (5.1)$$

其中， $\mu$  于  $\sigma$  分别表示第  $n$  个变量在  $M$  个样本上的均值与方差。 $X_i^n$  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：

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

对于连续类型的分子变量，由于其量纲差异大，极值极端情况多，因此选择 Z-score 标准化进行数据处理。对于生物活性指标，IC50\_nM 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 作为后续的变量筛选的目标值，同时利用 0 - 1 归一化进行数据处理。

什么时候选择标准化，什么时候选择归一化，为什么选择都需要说清楚

#### 5.3.2 变量筛选阶段

在变量筛选阶段，需求解的是变量与生物活性的相关性，并给出相关性程度以选择最

## 考虑相关性时，不仅要考虑线性相关，还要考虑非线性相关

重要的变量。但在考虑相关性时，需要考虑两者的线性相关以及非线性相关度，因此选择多种模型，并联合其结果选定最终符合要求的重要变量。同时，计算变量重要度时还需要考虑特征筛选过程中的稳定性，即每次实验得出的结果是否统一，以及如何消除这类不稳定性。实验发现不同相关性计算方法的结果并不统一，因此采用集成的方法对多种计算结果进行集成。

### 5.3.2.1 稳定筛选

利用几种基于统计技术的稳定特征选择方法可以计算出变量与生物活性间的相关性，采用了包括灰色关联分析、偏相关性分析以及最大互信息系数的方式，发现这几类方式可以显著的计算出部分变量与生物活性具有更高的相关性，但几种算法的结果并不统一。3种方式的计算结果如图 5-2, 5-3 以及 5-4 所示。

(1) 灰色关联分析 (Grey Relation Analysis, GRA)<sup>[3]</sup> 是一种应用广泛的关联分析方法，它的中心思想是通过计算参考数据列与若干个对比数据列的几何形状的相似程度来确定其关联是否紧密，其关联系数反映各变量与目标变量的接近程度，关联系数越大则变量越优。而且，不管样本量的大小和规律性的强弱，灰色关联算法都可以适用。

在本文中，我们使用灰色关联分析算法进行变量筛选。从本质上讲，灰色关联算法就是一个度量两个向量之间距离的方法。对于有时间性的因子，向量可以看成一条时间曲线，通过度量两条曲线的形态和走势是否相近来进行相关性判别。为了避免其他干扰，突出形态特征，我们采用均值化方法进行对数据进行归一化处理。

$$x_i(k)' = \frac{x_i(k)}{mean(x_i)} \quad (5.3)$$

$$mean(x_i) = \frac{1}{n} \sum_{k=1}^n x_i(k) \quad (5.4)$$

均值化就是把这个序列的数据除以均值，由于数量级大的序列均值比较大，所以除掉之后就能归一化到 1 的量级附近，从而符合系数的一般定义，以便于计算灰色关联系数

$$\xi_i(k) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|} \quad (5.5)$$

也就是说对于某一个因素，其中的每个维度进行计算，得到一个新的序列，这个序列中的每个点就代表着该子序列与母序列对应维度上的关联性（数字越大，代表关联性越强）。由于在公式(5.3)中，分子上是取距离的全局最小值，所以下面的分母必然大于分子，因而如果分母非常大，曲线距离就非常远， $\xi_i(k)$  接近 0，反之，如果  $x_i$  和  $x_0$  在所有维度上的差几乎一样，那么  $\xi_i(k)$  接近 1。这样  $\xi_i(k)$  取值范围在 0~1 之间，0 表示不相关，1 表示强相关。 $\rho$  是一个取值为(0,1)的可调节系数，目的是为了调节输出结果的差距大小， $\rho$  越小，区分度越大。

### (2) 相关性分析与偏相关性分析

皮尔逊相关系数(Pearson correlation coefficient, PCC)是用于度量两个变量 X 和 Y 之间的相关（线性相关），其值介于 -1 与 1 之间。两个变量之间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的商，总体相关系数通常表示为  $\rho_{X,Y}$ ，

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (5.6)$$

通过估算样本的协方差和标准差，就可以得到皮尔逊相关系数  $r$ ，

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (5.7)$$

在本文中，我们使用皮尔逊相关系数对筛选的 20 个变量进行独立性检验，确保各个变量之间是独立的。**两变量具有相关性，不代表其是因果关系，故进行偏相关分析-确定因果关系**

**偏相关性分析**与相关性分析一同使用，在剔除相关性较低的变量后，为了真实地刻画变量与生物活性间的相关性，还利用偏相关性分析中的控制变量法，在计算某变量相关性的同时，控制（固定）其余变量，以计算真实反映变量相关性程度的统计量。我们延续利用皮尔逊相关系数作为相关性计算方式，在计算变量偏相关性系数的同时，计算其 95% 参数置信区间以及 p-value 以表明结果的精确与合理性。

### (3) 最大互信息系数 (Maximal Information Coefficient, MIC)

最大互信息系数是用于度量两个变量之间的关联程度的度量方式。MIC 利用分箱技术将互信息 (Mutual information MI) 应用于连续随机变量。其中互信息也是一种信息度量，它描述了随机变量包含关于另一个随机变量的信息量。令两个分子描述符变量  $(X, Y)$  的联合分布为  $p(x, y)$ ，边缘分布分别为  $p(x), p(y)$ ，则互信息  $I(X; Y)$  被定义为：

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (5.8)$$

MIC 首先将两个连续随机变量离散到二维空间中，然后不断地用箱子（小方格）去分割，最后计算散点落入各箱子地概率，即为联合概率。mic 值的计算公式为：

$$\text{mic}(x; y) = \max_{a,b < B} \frac{I(x; y)}{\log_2 \min(a, b)} \quad (5.9)$$

其中， $a, b$  分别为箱子在水平和垂直方向上的数量， $B$  是箱子总数。

MIC 计算可分为三个步骤：

- ① 将  $X, Y$  构成的散点图划分成  $i$  行  $j$  列，并求出最大的 mic 值
- ② 对最大得 mic 值归一化
- ③ 选择不同  $i, j$  划分下的最大的 mic 值作为 MIC 值

图 5-2、5-3 以及 5-4 是分别基于以上三种关联度计算方式计算得到的变量与 ERa 活性相关性。可以发现，三种方法能够一定程度的寻找到关联度更高的一些变量，这些变量即使在不同的量纲下也显示出高于平均值的水平。其中，黑色横虚线表示全体变量的重要度平均值，而红色线表示前 20 名重要度的平均值。然而，三种方法的预测结果各不相同，另一方面，这类方法的预测又较为稳定。考虑到以上要素，本文还借助基于模型学习的关联度计算方法。

**基于统计的方法，可以显著找到“重要”变量，虽然无论试多少次，排出来的结果都是一样的，即“稳定筛选”，但3种方式的结果差异较大，可信度低，效果较差。所以需要引出下面基于机器学习的筛选方法。**

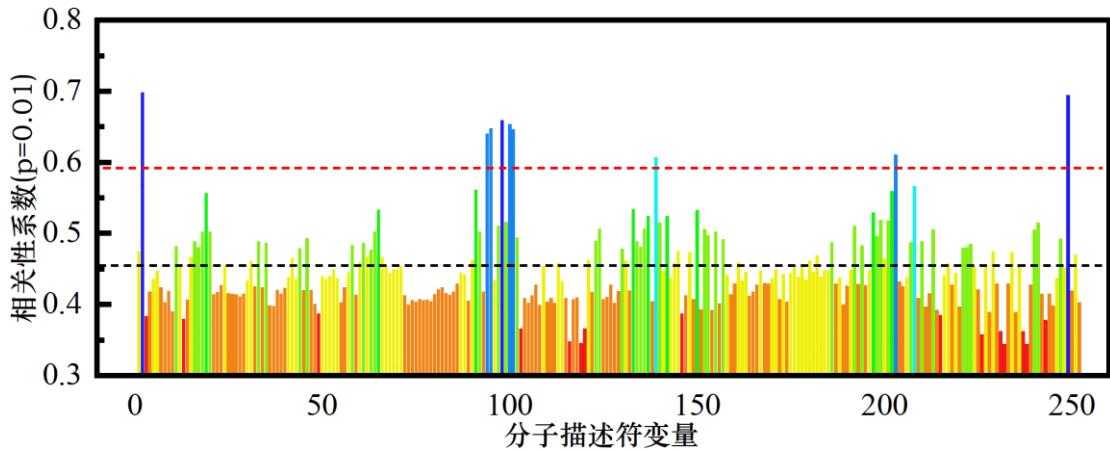


图 5.2 基于灰色关联度算法的变量重要性

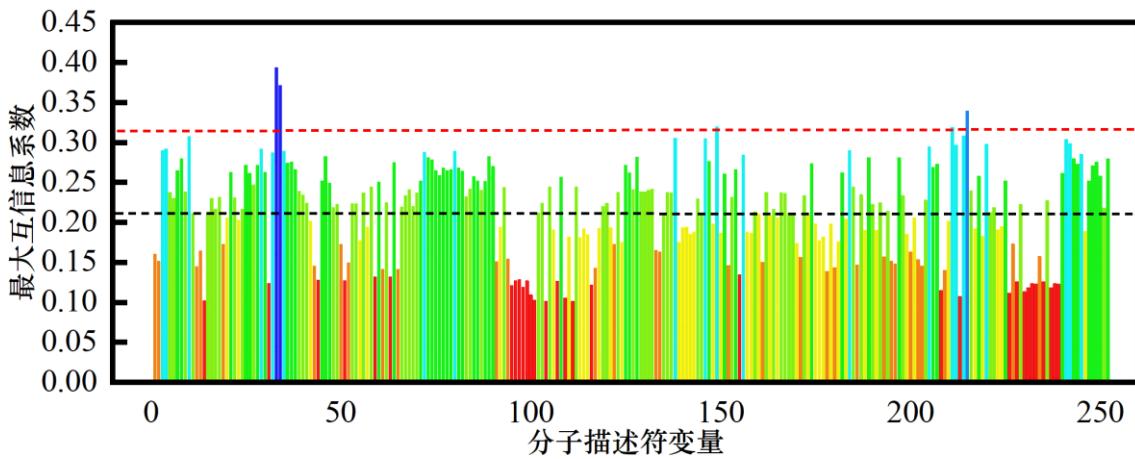


图 5.3 基于最大互信息系数的变量重要性

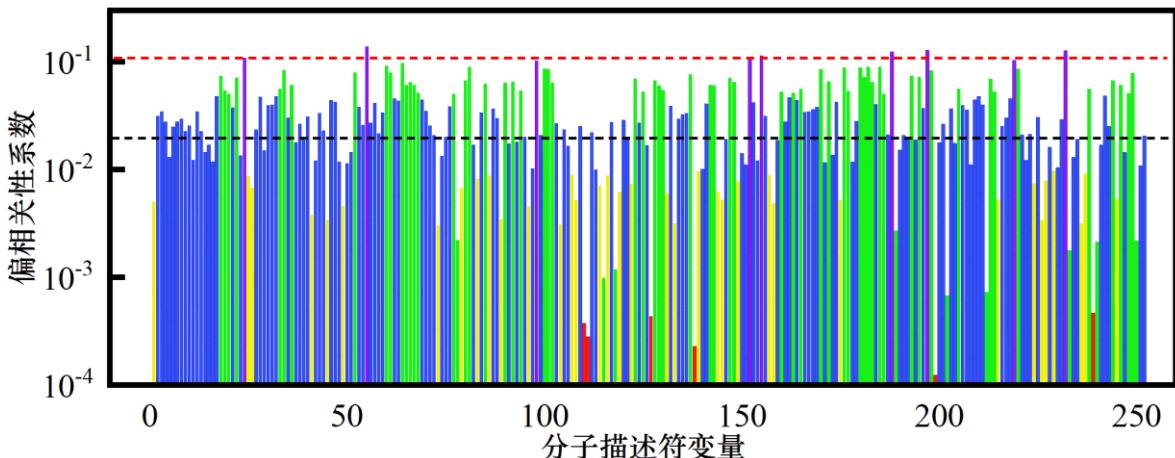


图 5.4 基于偏相关的变量重要性

### 5.3.2.2 非稳定筛选

5.3.2.1 中的方法通过基于统计的方法计算出的变量重要/相关度，对于固定的数据集，其结果是稳定的。然而对于如随机森林等方法而言，其结果是不稳定的。这是由于对于随机森林中的树，其训练数据是通过 Bootstrap Aggregating 得到的全体数据集的子集。在计算重要度得分时，若是通过计算 Gini 系数 (Gini coefficient)，则某些变量变选中后，基尼指数下降，其他相关变量的重要性将下降，导致存在对于变量选择顺序的偏差。若是通过 Permutation Error 的方式随机打乱带袋外特征并计算精度损失的方法，则需要多次测试，以因随机森林的特性，结果不稳定，所以，需要进行多次测试，以减少偏差

提高随机数据划分的偏差。

(1) 随机森林 (Random Forest) 是一种包含多个决策树的分类器。决策树算法采用树形结构，主要由三部分构成，分别是：根节点（包含样本的全集）、内部节点（对应特征属性测试）和叶节点（代表决策的结果），在推理阶段，数据流不断地由树内部节点地某一属性值判断，从而导向下一个分支节点，指导叶节点处，即分类结果。

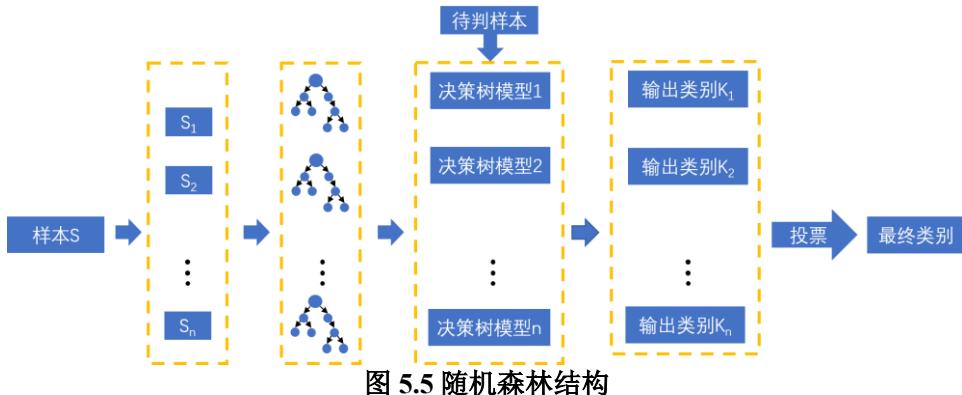


图 5.5 随机森林结构

决策树使用层递归判断分支推理来得到最终的分类。而随机森林的输出类别则由各个树输出的类别众数决定。并且，决策树是一种基于 if-else 链的有监督机器学习算法，这些 if-else 规则是由训练得到，而非人工标注。而构成随机森林的决策树，是互相独立的。当使用随机森林进行分类任务时，新的输入样本进入，就让森林的每一棵决策树分别进行判断和分类，得到各自的分类结果，最后投票决定最终的随机森林分类结果。随机森林变量选择过程如表 5-4 所示。

表 5-4 随机森林变量选择过程

---

输入：1947 个样本的 729 维特征向量数据；标签数据（生物活性）

输出：20 个对生物活性最具有显著影响力的变量

---

对向量数据进行单位标准化

初始化随机森林

利用标签数据对随机森林进行训练

计算特征重要度

计算特征重要度排序索引

While True:

    初始化新随机森林

    利用标签数据对随机森林进行训练

    计算特征重要性

    计算特征重要度排序索引

    计算 N 次平均特征重要度

    计算 N 次平均特征重要度排序索引

    If 平均特征重要度排序索引三轮未改变：

        Return 特征重要度排序索引

根据特征重要度排序索引选择特征

---

(2) 平均精度损失 (Mean Decrease Accuracy) 是一种根据特征变化时模型精度变化情况统计得到的基于模型学习的特征重要性指标。该类方法首先需要指定目标模型，并且通过随机特征扰动的方法依次改变特征，并测试模型在扰动后特征上的效果。虽然该方法同样可以指定目标模型为随机森林，但与直接利用随机森林输出的特征重要度不同的是，前者利用的是叶子节点的 Gini

两者图形如此相似，有个很重要的原因就是，基础模型都是随机森林。  
虽然一个算gini，一个算精度损失，但本质上都是随机森林训练出来的。

不纯度，而平均精度损失是基于模型的最终预测结果。

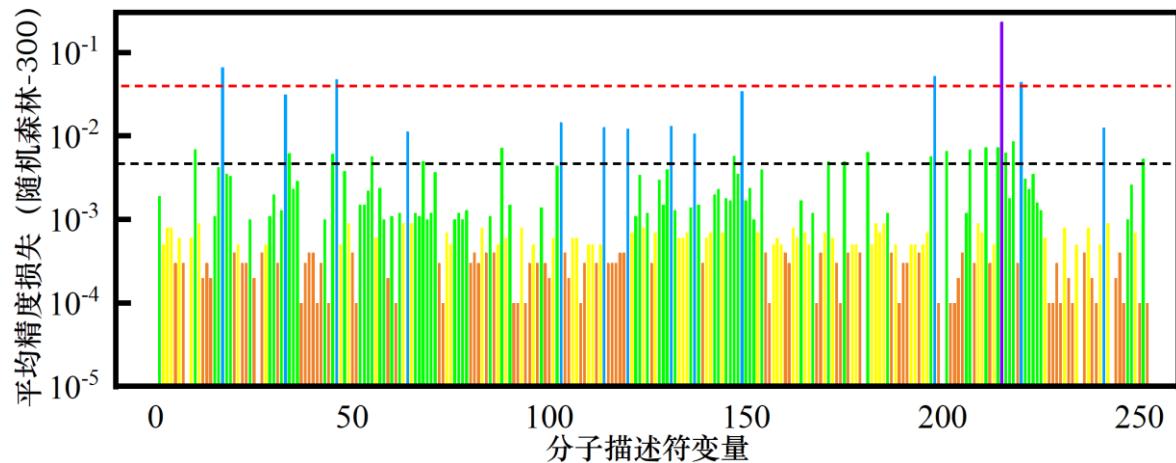


图 5.6 基于平均精度损失（以随机森林为目标模型）的变量重要性

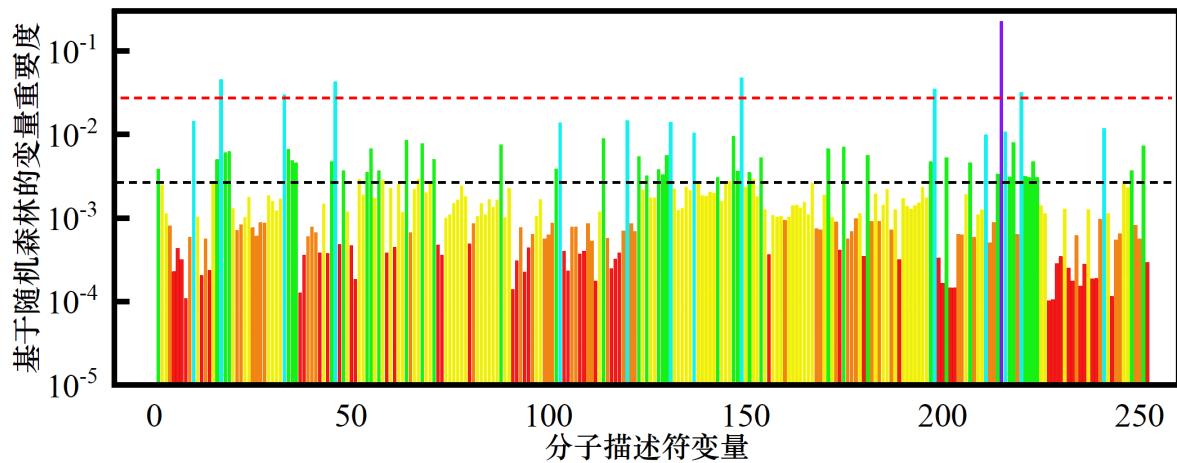


图 5.7 基于随机森林的变量重要性

图 5.6 以及 5.7 中展示了基于平均精度损失以及随机森林的变量重要度，可以发现，这两个方法在计算重要度方面同样具有优势，高重要度变量较一般变量十分具有区分度。

### 5.3.2.3 基于加权集成预测的变量筛选

以上结果表明五种方法在筛选主要变量时均具有一定的成效，然而五种方法的结果确各不相同，为了得到更加精确的主要变量，拟采用集成的方法，对五种方法采取加权平均，计算出某个变量的重要度。图 5-8 中展示了以上五种方法的变量重要度在经过标准化后的情况，可以发现，全部五种方法都能够找到少量的、重要度显著高于其他变量的主要变量。

五种方法都能有效找到少量重要变量，但因为结果不尽相同  
因此，引出加权集成方法。

根据分析5种方法的区分度，平均精度损失区分度大，效果好，偏相关性区分度小。因此，以基于机器学习的方法作为主导，基于统计的方法作为辅助进行集成。

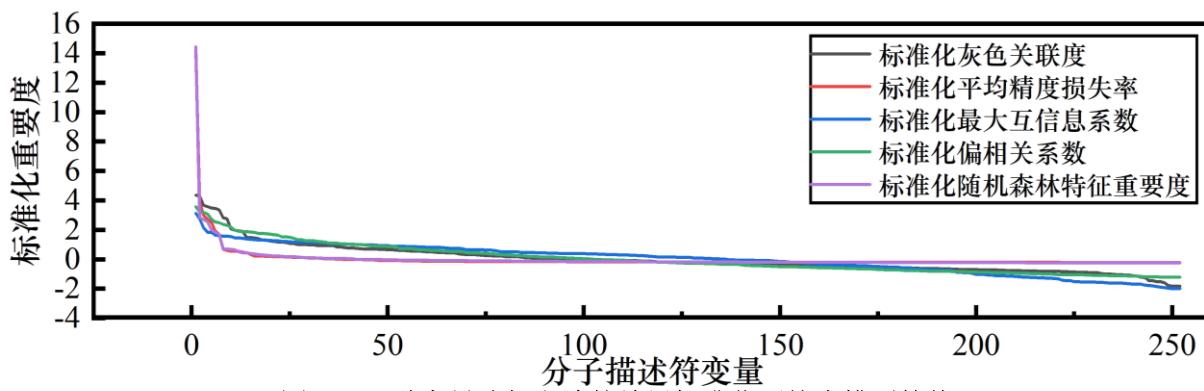


图 5.8 五种变量选择方法的结果标准化后按序排列趋势

结合表 5-2 可知，在选择前 20 个与生物活性最相关的变量时，基于随机森林 300 轮迭代的平均精度损失算法在选择主要特征时能够以更高的置信度选择出主要变量，使得在标准化后，主要变量的重要度高于其他算法一个数量级。通过图 5-8 与表 5-2 所反应的结果可知，**基于随机森林以及基于平均精度损失的方法在选择变量时更具优势**。而同时，图 5 种表明基于**偏相关系数**的方法在变量重要度预测方面的结果更加不具有**区分度**，且排名前 20 变量的重要度均值与全体变量的重要度均值差异要小于其他算法，因此在设计本文的继承算法是对其不进行考虑。**在对剩余 4 个结果进行集成的同时，将基于统计与基于模型学习的预测结果作为主导因素，给予更高的权限，而前 2 者的预测结果作为参考值。**

表 5.5 五种选择算法的筛选结果定量比较（标准化后）

选择算法	全体变量重要度均值	主要变量重要度均值	差异
灰色关联分析	-4.05319516e-16	2.46783470	+6.08861552e16%
平均精度损失	7.04903507e-18	1.75394783	+2.48820982e17%
最大互信息系数	-3.27780131e-16	1.66430338	+5.07749930e16%
偏相关性	-2.07946534e-16	2.29929123	+1.10571269e16%
随机森林	2.11471052e-17	1.70601330	+8.06736094e16%

对于灰色关联分析、最大互信息系数、平均精度损失以及随机森林等五种方法，以  $f_{i, i=1,2}$  与  $\varphi_{i, i=1,2}$  分别进行表示， $f_1(X^n)$ ， $f_2(X^n)$ ， $\varphi_4(X^n)$  与  $\varphi_5(X^n)$  分别表示经过标准化后的对于第  $X^n$  个变量的重要度预测。本文设计了一种加权集成预测的重要度度量方式，首先对  $f_1(X^n)$ ， $f_2(X^n)$  的结果进行平均以保证预测稳定性，同时对  $\varphi_4(X^n)$  与  $\varphi_5(X^n)$  进行指数增强，将其结果进行缩放，得到  $e^{-\varphi_1(X^n)}$  与  $e^{-\varphi_2(X^n)}$ ，并通过增加常数 1 保证该两者系数均大于 1，最终对于变量  $X^n$  而言，其与生物活性特征 Era 的相关程度以公式进行表示：

$$\frac{f_i(X^n) + f_2(X^n)}{3} \cdot \frac{(1 + e^{-\varphi_1(X^n)})}{(1 + e^{-\varphi_2(X^n)})}$$

该预测优化算法在考虑 2 种稳定筛选算法结果的同时，还将 2 种基于模型学习的算法的结果作为系数，通过以自身负数的  $e$  底数来将普遍变量与重要变量的差异进一步增大，不仅排除了  $\varphi_i$  的不稳定性，也保证了  $f_i$  的参与度。利用所提出的加权集成算法能够预测出的高重要性变量如表 5.3 所示。

该模型非常贴合数据，不仅保证了统计方法的稳定性，而且抵消了机器学习方法的非稳定性，同时增强了机器学习方法的权重

表 5.3 利用加权集成算法的 20 个主要变量筛选结果（归一化后）

编号	变量名称	重要性指数	编号	变量名称	重要性指数
1	nH	0.6983	11	hmin	0.5329
2	ATSc1	0.6947	12	DELS	0.5295
3	ATSc5	0.6593	13	DELS2	0.5244
4	BCUTw-1h	0.6109	14	ETA_Beta	0.5243
5	C1SP2	0.6069	15	ETA_BetaP	0.5195
6	SCH-7	0.5669	16	ETA_EtaP_B	0.5184
7	VCH-7	0.5615	17	ETA_EtaP_B_RC	0.5166
8	SP-7	0.5594	18	nHBAcc2	0.5157
9	ECCEN	0.5573	19	MDEC-33	0.5154
10	maxHBd	0.5339	20	MLFER_BO	0.5113

### 5.3.3 独立性检验阶段 阐述独立性检验的必要性，以及线性相关/非线性相关检验的原因

在选择出主要变量后，需要对其独立性进行检验。而由于题中未给出分子描述符间独立性的说明，因此解题过程中不仅需要考虑变量的代表性，同时需要考虑变量的独立性。另外，在独立性检验阶段方面，需要同时考虑变量间的线性相关性以及非线性相关性。线性相关分析在挖掘变量间线性相关关系上具有高效、精确的特定。而非线性相关分析能够捕捉变量间的非线性相关关系，这种隐藏的相关性可能可能导致模型训练时精度收到一定程度的干扰，因此本文将结合两种相关性分析方法对所选择的变量进行检测。

#### 5.3.3.1 线性相关分析

在线性相关性分析方面，可以直接使用皮尔逊相关系数对各个方法筛选的 20 个变量进行检验，确保各个变量之间是独立的。如图 5.9 所示，计算四个不同筛选方法选择得到与生物活性最相关的 20 个变量间皮尔逊相关系数。可以发现基于统计方法计算到的重要变量间存在严重的相关性。如在灰色关联分析的结果中，nHBint2、nHBint3、nHBint6、nHBint8 以及 nHBint9 几个变量具有高度的相关性。而“分子描述符含义解释”附件也表明了这几个变量数以同一种类型，即“Atom Type Electro topological State”。这也侧面印证了该类变量对于生物活性具有显著影响。而对于基于模型学习的两种变量选择方法，虽然变量间的相关程度要明显低于以上两者，但同样存在高相关的冗余变量，如 MDEC-23 与 MDEC-33，在可选择变量数量有限的情况下，选择具有高度线性相关性的变量，会对后续的模型学习带来冗余变量，造成负面影响，因此在独立性检验阶段，需要剔除高度线性相关的变量。

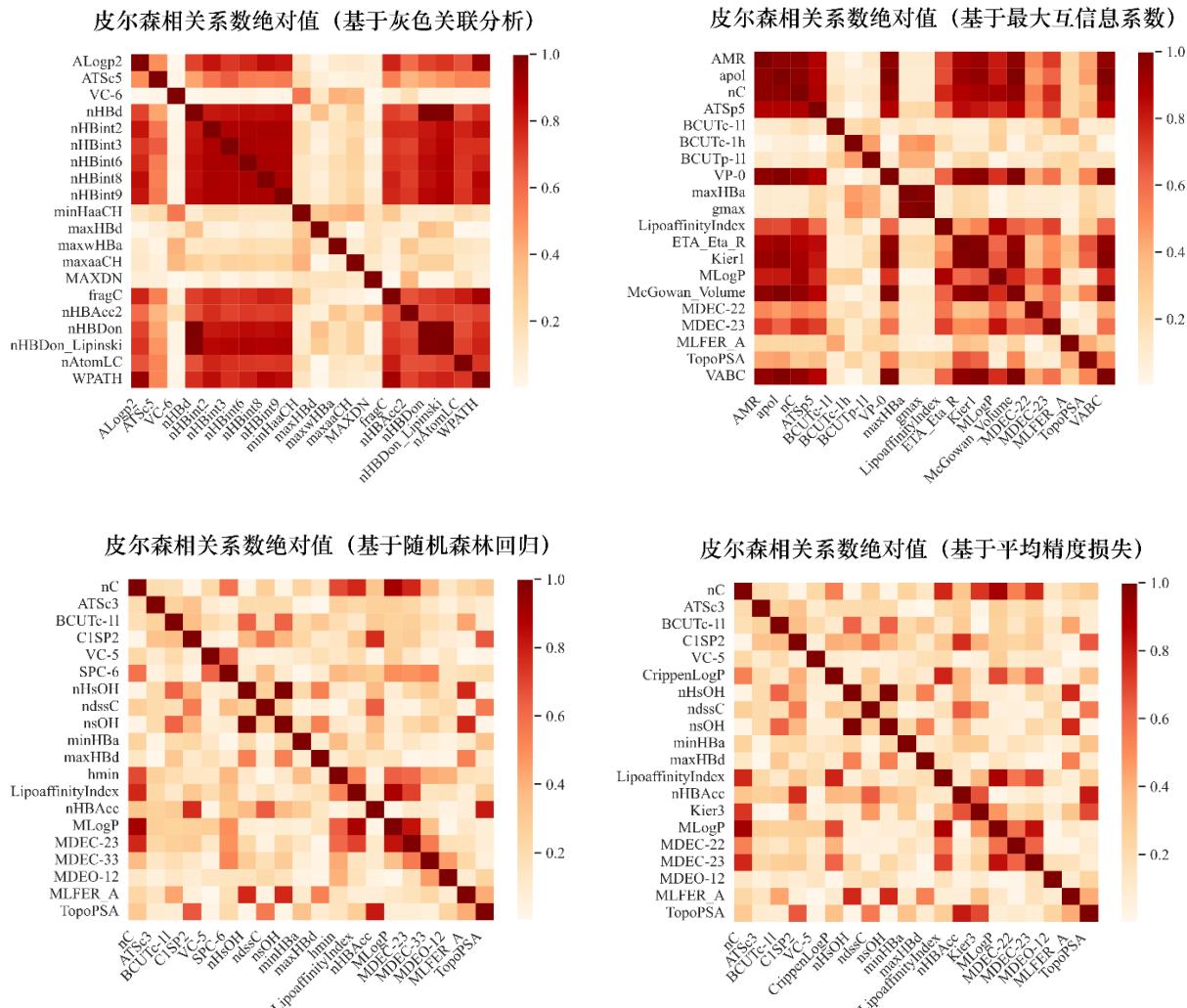


图 5.9 不同方法筛选变量的皮尔森相关系数（绝对值）。

### 5.3.3.2 非线性相关分析

上一节中分析了所选择变量间的线性相关性，然后在独立性检验的过程中还有必要进行非线性相关的分析，在这一部分，本文采用卡方独立性检验来计算变量间可能存在的相关性，并用 p-value 定量的分析不同变量筛选方法的结果。在统计学中，独立性检验是一种假设检验方法，即先假设再推翻，也属于卡方检验的一种。根据次数来检验两类因子之间是否相互独立。值得注意的是，独立性检验必须建立在零假设上：一个样本的易发生事件频次受某个特定理论分配。其中，事件要求互斥，且时间发生概率总和为 1。独立性检验步骤见表 5-2。

表 5-7 独立性检验步骤

- 
- ① 计算卡方检验的统计值  $\chi^2$ ，即把每一个观察值和理论值的差依次做平方、除以理论值、加总。
  - ② 计算  $\chi^2$  统计值的自由度  $df$
  - ③ 根据预设的置信水平，查出自由度为  $df$  的卡方分配临界值，比较它与第一步得出的  $\chi^2$  统计值，推论能否拒绝零假设。
-

图 5-10 展示了基于卡方检验的非线性关系程度，以 p-value 作为参考，其取值范围在 0-1 之间，其越小表明两个变量间的相关程度越高。该部分结果与 5.3.3.1 中相同，对于基于统计方法，其变量间的非线性相关程度也较高。而对于基于模型学习的方法，这种现象虽然有所缓解，但也需要仔细考量，因为能够观察到存在着一定数量的变量间是具有高度相关性的。

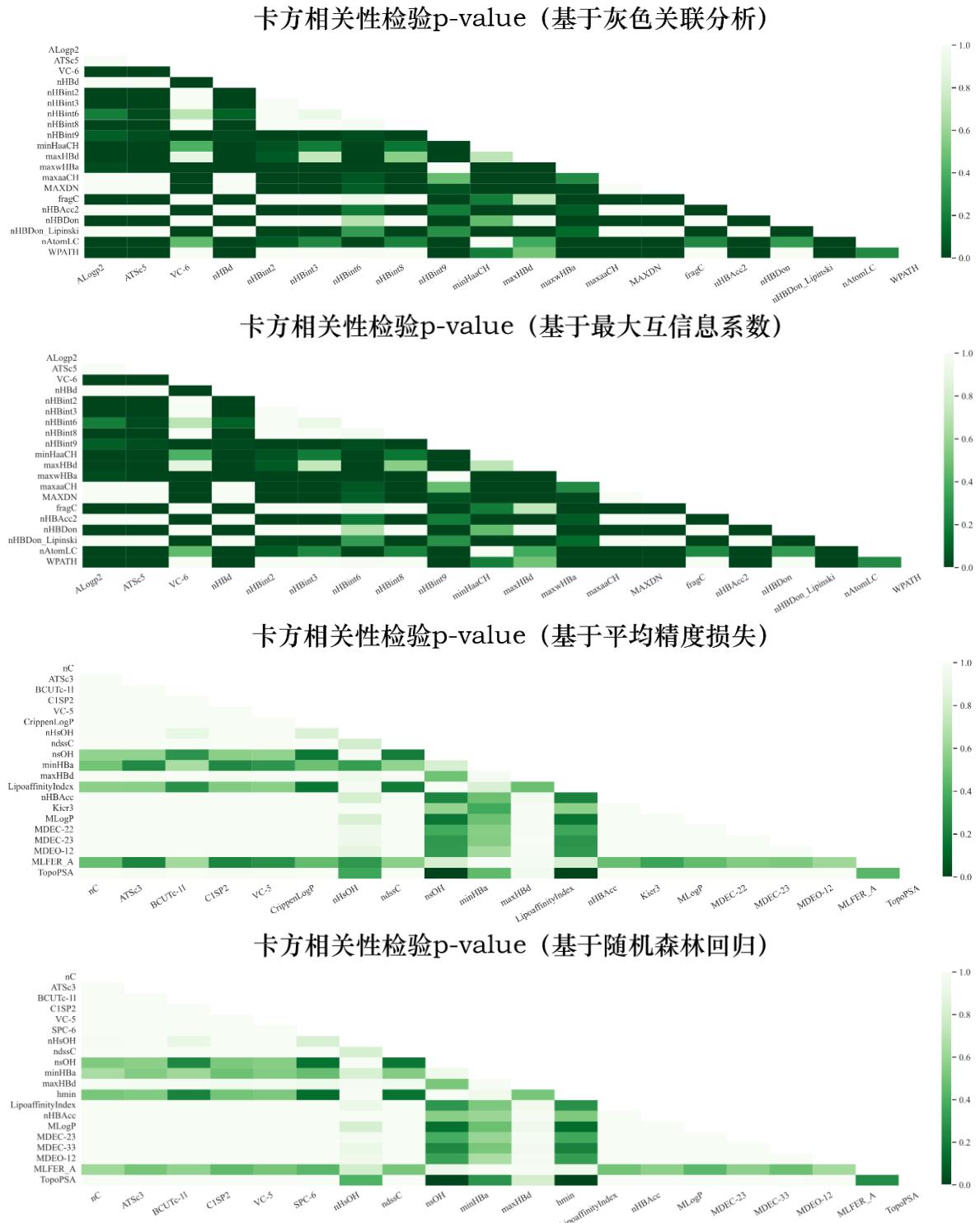


图 5.10 不同方法筛选变量的卡方相关性检验 p-value（越小越相关）。

独立检验阶段的结果也表明，基于模型学习的方法在选择变量上相比基于统计的方法，更能够选择出有效、独立的特征，进一步佐证了所提出的基于加权集成的变量重要度算法的合理性。

### 5.3.3.3 主要变量去相关

通过以上分析可以发现，无论是稳定的变量选择方法还是不稳定的变量选择方法，其结果始终会出现大量的高相关变量，这也符合问题分析中对于数据集相关性的假设。因此，需要进行去相关操作，对所选择的主要变量进行剔除，并添加候补的重要变量，以选择出不仅具有代表性，且互相独立的变量，防止在预测的过程中产生数据冗余。

如图 5-10 后，利用所提出的自适应迭代特征选择方法，在经过数轮的迭代后选择出了新的重要变量集合，再次计算其皮尔森相关系数可以发现，两两变量间具有较低的相关性水平，较为符合独立性的标准，为后续模型学习剔除了冗余特征。

主要独立变量间相关性

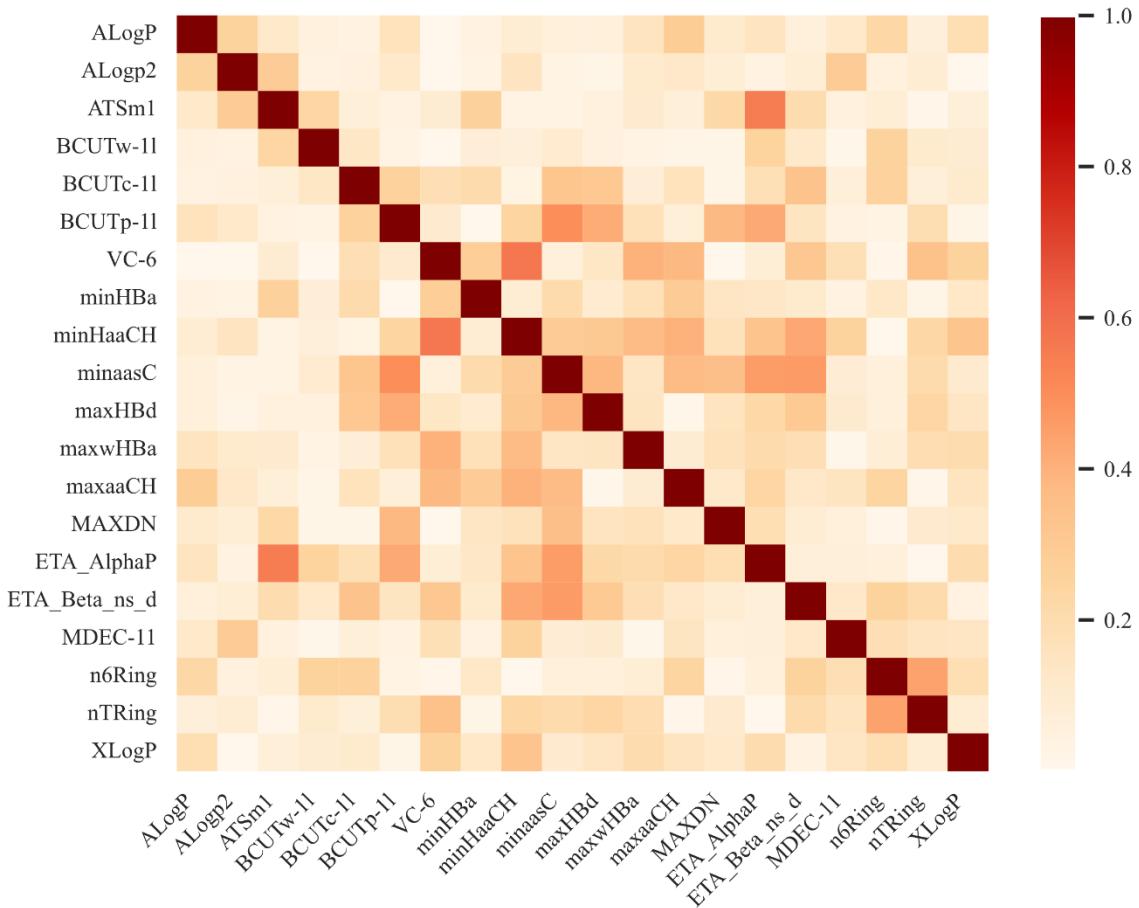


图 5.10 去相关后主要变量间的相关性。

## 5.4 结果展示

按照重要性排序，得到了对生物活性最具有显著影响力的前 20 个变量（即 20 个分子描述符）。这些变量如表所示。

表 5.8 20 个主要变量的最终筛选结果

变量编号	变量名称	重要性指数	变量编号	变量名称	重要性指数
1	nH	0.6983	11	hmin	0.5329
2	ATSc1	0.6947	12	DELS	0.5295
3	ATSc5	0.6593	13	DELS2	0.5244
4	BCUTw-1h	0.6109	14	ETA_Beta	0.5243
5	C1SP2	0.6069	15	XlogP	0.5195
6	SCH-7	0.5669	16	ETA_EtaP_B	0.5184
7	VCH-7	0.5615	17	TopoPSA	0.5166
8	SP-7	0.5594	18	nHBAcc2	0.5157
9	ECCEN	0.5573	19	MDEC-33	0.5154
10	maxHbd	0.5339	20	MLFER_BO	0.5113

## 5.5 小结

首先通过数据清洗，首先将无意义变量（值均相同且为 0）和高度疑似存在大量异常值的数据（部分值为 0 的离散变量）进行剔除，从而使初始的 729 个变量缩减到 252 个。由于变量的离散类型、取值区间、分布情况各不相同，因此采用 Z-score 标准化对变量做了进一步处理，同时我们采用 pIC50 作为目标值，并进行 0-1 归一化。通过利用基于统计与基于模型学习的方法对变量数据进行筛选，包括灰色关联分析算法、皮尔逊相关系数、偏相关性分析、随机森林与平均精度损失方法。

# 六、问题二

## 6.1 问题分析

依据题意，需要根据问题 1 所得前 20 个对生物活性最具有显著影响的分子描述符变量，构建化合物对 ER $\alpha$  生物活性的定量预测模型，并使用该模型对测试数据集中的 50 个化合物进行 ER $\alpha$  生物活性定量预测。考虑到 ER $\alpha$  生物活性的 IC50 值和 pIC50 值之间是算术转换，由于 pIC50 的分布非常集中，可使学习尽快收敛，因此我们选择 pIC50 作为目标。在经过问题 1 筛选到的 20 个变量的条件下，本文对该问题的解答可具体分为三个步骤：

- ① 采用标准化处理对分子描述符进行量纲消除，使模型输入数值属于同一级别。同时，采用归一化操作对 pIC50 进行处理，
- ② 对分子描述符数据集进行 9: 1 的随机划分，前者用于算法的训练，后者用于验证该算法的准确率。当算法对测试划分集的预测误差小于某一阈值时，则认为已得到一个可用的 ER $\alpha$  生物活性定量预测模型。
- ③ 第 2 步训练得到的模型对测试集进行预测并记录结果。
- ④ 联合多种评估指标，选择精度最优的模型，并作为后续模型的 ER $\alpha$  生物活性预测模型。

该问题的解题细节由流程图 6.1 所示。

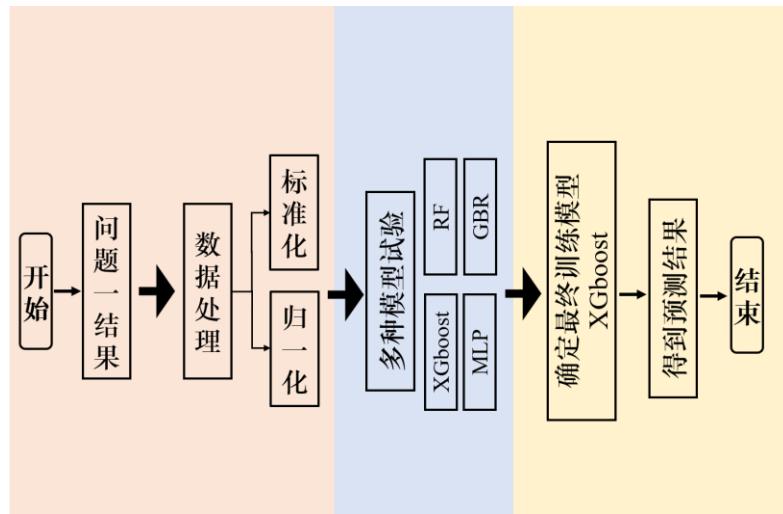


图 6.1 问题 2 的解题思路流程图

## 6.2 模型建立

求解此题时,  $\text{pIC}_{50}$  由于其数值区间稳定且较小, 因此选择  $\text{pIC}_{50}$  作为预测目标, 而  $\text{IC}_{50}$  在预测  $\text{pIC}_{50}$  直接进行算数转换计算得到。本文题可以归纳为回归问题, 在模型建立的过程中, 本文评估了一系列模型并根据其在多项指标上的性能, 包括精度以及稳定性, 来选择最优的模型。首先, 选用了两组不同类型的模型, 分别是单体模型与集成模型, 前者指利用单个算法拟合训练集, 而后者则借助多个子模型, 通过拟合数据集的子集来达到学习的效果。在考虑单模型时, 首先选用了传统的回归模型。注意, 分子描述变量与  $\text{pIC}_{50}$  间可能存在高度的非线性, 线性甚至多项式回归的效果可能较差, 因此直接选用核岭回归模型。同时选择了传统的支持向量机分类器的回归版本, 即支持向量机回归预测器。集成模型部分, 则使用了梯度提升回归。

### 6.2.1 优化目标

本体要求预测化合物的  $\text{IC}_{50}$  水平以及  $\text{pIC}_{50}$  水平, 该任务属于回归任务, 因此在优化目标模型是, 实际是优化该模型的输出与对应真实值的误差。一般而言, 为了防止噪声, 误差由损失函数  $F$  定义:

$$F = \frac{\sum_{i=1}^N (X_i - \phi(X_i))^2}{n}, \quad (6.1)$$

其中,  $X_i$  表示第  $i$  条化合物的分子描述符变量数据,  $\phi$  表示预测模型, 问题 2 的模型就是尽可能的优化  $F$ , 使  $\phi(\cdot)$  的输出与真实值的差异尽可能的小。

### 6.2.2 单模型算法

#### 6.2.2.1 核岭回归

核岭回归是数据预测与分类领域常用的监督型算法，通过利用测试集数据来建立模型，再利用模型训练集中的数据进行处理，旨在寻找一个拟合曲线，使得所有样本点到达这个曲线的距离的和最小，从而达到预测与分类的目的。在经典的线性回归问题中，最小二乘法常被用来计算模型回归模型的参数  $w$ ，但最小二乘法不适用于大规模数据，且对数据噪声十分敏感。当输入数据存在多重线性时，由最小二乘法计算出的参数  $w$  在数值上会非常大，使得在线性模型  $y = w^T x$  中，输出变量  $y$  会随着输入变量  $x$  的微小扰动而产生巨大变化。此时回归系数趋向于无穷大，也就失去了回归的意义。如果能限制参数  $w$  的增长，使  $w$  不会变得特别大，那么模型对噪声的敏感度就会降低。为了解决这个问题，可以对模型原来的目标函数上加上一个惩罚项，来限制模型参数  $w$  的大小。这种做法也称为正则化(Regularization)。如果惩罚项是参数的  $l_1$  范数，就是套索回归(Lasso Regression)；如果惩罚项是参数的  $l_2$  范数，就是岭回归(Ridge Regression)，目标函数为

$$J_w = \min_w \left\{ \|y - Xw\|^2 + \alpha w^T w \right\}, \quad (6.2)$$

其中， $\alpha$  是用来平衡损失和正则化的一个可调节系数， $\alpha$  的数值越大，那么正则项的作用就越明显； $\alpha$  的数值越小，正则项的作用就越弱。极端情况下， $\alpha=0$  则岭回归的目标函数和原来的损失函数是一样的，如果  $\alpha=\infty$ ，则损失函数只有正则项，此时其最小化的结果必然是  $w=0$ 。

然而，在实际应用中，由于数据可能是非线性的，而当数据表现出较强非线性的时候，岭回归模型的精度就会下降，因此，可以引入核机制来缓解这个问题。具体来说，就是把数据映射到一个核空间，使得数据在这个核空间里线性可分。设核函数为  $\Phi_i = \Phi(x_i)$ ， $\Phi_i$  是一个  $n$  维空间中的向量（可接近于无穷维），可以认为  $\Phi_i$  是核空间  $x_i$  的一组特征，在此核空间中对这组特征进行线性回归。核岭回归的目标函数为

$$J_w = \min_w \left\{ \|y - \Phi w\|^2 + \alpha w^T w \right\}, \quad (6.3)$$

由正规方程解得

$$w = (\Phi^T \Phi + \alpha I_n)^{-1} \Phi^T y, \quad (6.4)$$

由于  $\Phi$  维度可达无穷，难以求逆，故用(5.4)进行等式变换，

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}, \quad (6.5)$$

其中， $B = \Phi$ ,  $P = \frac{1}{\alpha} I_d$ ,  $R = I_n$ ，令  $\lambda = (\Phi \Phi^T + \alpha I_n)^{-1}$   $y \in \mathbb{R}^{n \times 1}$ ，则有

$$w = \Phi^T \lambda = [\Phi_1, \Phi_2, \dots, \Phi_n] \lambda = \sum_{i=1}^n \lambda_i \Phi_i, \quad (6.6)$$

又因为  $K = \Phi \Phi^T \in \mathbb{R}^{n \times n}$  即为 Gram 矩阵，且  $K_{ij} = \Phi_i^T \Phi_j$ ， $K_i$  是  $K$  的第  $i$  列，则

$$y_i = w^T \Phi_i = y^T (K + \alpha I_n)^{-1} \Phi \Phi_i = y^T (K + \alpha I_n)^{-1} K_i \quad (6.7)$$

在本题中，也就是说  $y_i$  就是 ERα 生物活性无侧置预测值函数。

### 6.2.2.1 支持向量机回归

(Support vector machines, SVM) 是一种经典的二分类机器学习算法。其基本思想在于通过训练数据集构建几何分离超平面，并使离超平面的最近点（以下简称最近点）到该平面的距离最大。存在无穷多个超平面将数据集线性分离，然而几何间隔最大的超平面是

唯一的。SVM 正是通过该超平面使得二分类问题求解达到最优，如图所示：

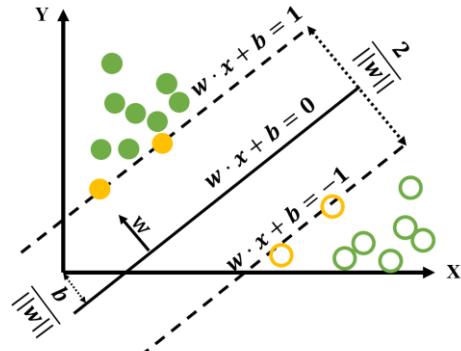


图 6.2 SVM 原理示意图

其中， $w \cdot x + b = 0$  为分离超平面， $w \cdot x + b = 1$  和  $w \cdot x + b = -1$  中的 1 和 -1 是二分类标签值，是为了保证  $y_i(wx_i + b)$  始终大于 0。SVM 事实上是求解以下的最优问题：

$$\max_{w,b} \left( \min_{x_i} \left( \frac{y_i(wx_i + b)}{\|w\|} \right) \right) \quad (6.8)$$

令最近点到分离超平面的函数距离（即  $\frac{y_i(wx_i + b)}{\|w\|}$ ）设为 1，从而最优问题转化为：

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{s.t. } y_i(wx_i + b) \geq 1 \quad (6.9)$$

把约束条件融合到优化目标函数中，并建立拉格朗日公式：

$$L(w,b,\alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i(wx_i + b) - 1) \quad \text{s.t. } \alpha_i \geq 0 \quad (6.10)$$

从而最优问题可以转化为

$$\min_{w,b} \left( \max_{\alpha_i \geq 0} (L(w,b,\alpha)) \right) \quad (6.11)$$

如满足 KKT 条件（拉格朗日乘子法的泛化），则需要求对偶问题：

$$\max_{\alpha_i \geq 0} \left( \min_{w,b} (L(w,b,\alpha)) \right) \quad (6.12)$$

对  $w, b$  求极小值，得：

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad \sum_{i=1}^n \alpha_i y_i = 0 \quad (6.13)$$

把  $w, b$  代回  $L(w,b,\alpha)$ ，得：

$$\max_{\alpha_i \geq 0} W(\alpha) = \max_{\alpha_i \geq 0} \left( \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j x_i x_j \right), \quad \text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0 \quad (6.14)$$

对  $\alpha_i$  求导，从而得出  $W(\alpha)$  的极大值。

### 6.2.2.1 多层感知机 (Multi-Layer Perceptron)

MLP 是最基础的神经网络架构，其结构如图 6 所示。其中包含了输入层、隐藏层以及输出层。由于本文的输出是单个数值，则模型的输出层也只有 1 一个输出，输入层的数量根据输入维数相关，在本文中被设置为 20。每一个神经元的运算可以被记为：

$$O_j = \sigma\left(\sum w_{ij}x_i + b_j\right)$$

其中， $j$  表示输出层的索引， $i$  表示输出层的索引， $w_{ij}$  是对应的隐藏层的节点的权重， $b_j$  表示偏执， $\sigma$  表示激活函数，在本文中，用于 pIC50 均为正数，因此所有的激活函数均设置为 ReLU。

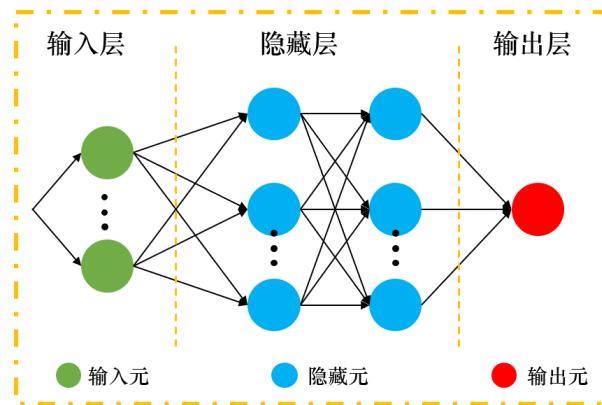


图 6.3 MLP 模型结构

### 6.2.3 集成模型算法

#### 6.2.3.1 梯度提升回归 (Gradient Boosting)

梯度提升回归算法由两方面组成：提升（boosting）和梯度（gradient）。Boosting 与 Bagging 并称增强基础学习回归算法，它们皆由多个基学习器组成，其中 Boosting 由于其基学习器之间的强依赖关系，因此是串行提升基学习器的效果，而 Bagging 是并行提升。Boosting 可主要分为以下三个步骤：

1. 训练一个初始基学习器，根据该基学习器的效果对训练集进行调整，增强在当前基学习器中处于弱势的样本使其在后续基学习器中处于强势。
2. 根据调整后的训练集训练下一个基学习器。
3. 重复以上过程直到基学习器的数量达到预先设定的值  $T$ ，Boosting 的输出即为  $T$  个基学习器的加权和。

而 gradient 则是指基学习器的补偿增强是由损失函数的负梯度来决定的。

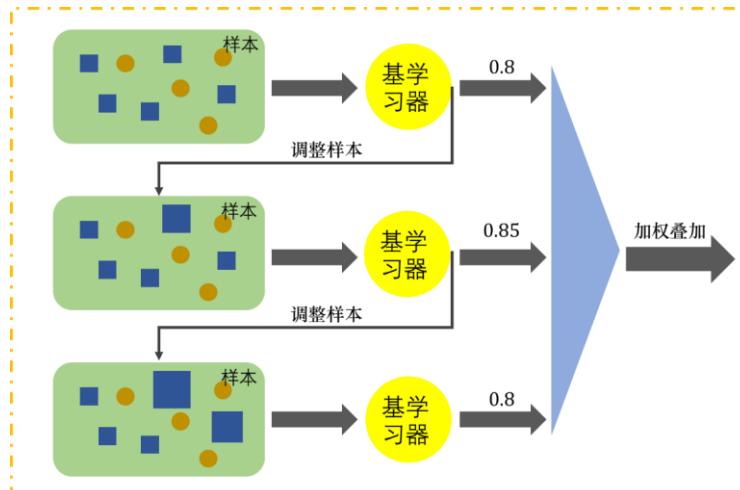


图 6.4 MLP 模型结构

### 6.2.3.2 随机森林回归 (Random Forest Regression)

章节 5.3.2.2 中已经介绍了随机森林算法的基本原理，在此不再介绍其原理。不同点在于，随机森林回归中，若每个叶子节点的不纯度的总和为  $G(x, v)$ ，则其优化目标一般均方误差 MSE， $G(x, v)$  即为：

$$G(x, v) = \frac{1}{N_s} \left( \sum_{y_i \in X_{\text{left}}} (y_i - \bar{y}_{\text{left}})^2 + \sum_{y_j \in X_{\text{right}}} (y_j - \bar{y}_{\text{right}})^2 \right)$$

### 6.2.3.3 极限梯度 boost (XGBoost)

XGBoost 是 GBDT 算法的一个工程实现，它对 GBDT 进行了很多工业级的优化。设 XGBoost 由  $k$  个基学习器组成，令第  $t$  次迭代训练的树模型是  $f_t(x)$ ，则：

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (6.15)$$

其中  $\hat{y}_i^{(t)}$  是第  $t$  次迭代后样本  $i$  的预测结果， $\hat{y}_i^{(t-1)}$  是前  $t-1$  个基学习器的预测结果， $f_t(x_i)$  是第  $t$  个基学习器。损失函数  $L$  为：

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) \quad (6.16)$$

目标函数  $Obj$  由损失函数和抑制模型复杂度的正则项  $\Omega$  组成：

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^t \Omega(f_i) \quad (6.17)$$

又由于  $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$ ，目标函数可转化为：

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i) \quad (6.18)$$

函数  $f(x + \Delta x)$  在点  $x$  处进行泰勒二阶展开，有：

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \quad (6.19)$$

将损失函数二阶展开有：

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) = l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \quad (6.20)$$

其中， $g_i$  为损失函数的一阶导数， $h_i$  为其二阶导。假设损失函数为平方损失，即：

$$l(y_i, \hat{y}_i^{(t-1)}) = (y_i - \hat{y}_i^{(t-1)})^2 \quad (6.21)$$

则目标函数为：

$$Obj^{(t)} = \sum_{i=1}^n \left[ l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{Constant} \quad (6.22)$$

## 6.3 模型求解

### 6.3.1 模型参数设置

表 6-1 各个模型的参数等基本设置

模型名称	变量名称	参数值
核岭回归	核函数类型 Kernel 核系数 ma	径向基函数('rbf') 1.0
支持向量机回归	核函数类型 Kernel 核系数 Gamma 核函数中的独立项 coef0 容忍停止标准 tol 惩罚参数 C epsilon-tube 收缩启发式 求解器迭代限制 max_iter	径向基函数('rbf') 1/20 0.0 1e-3 0.1 0.1 True -1
多层感知机	层的性质 层数及神经元个数(包括输入输出层) MLP 每层的激活函数 优化器 学习率 批次大小 迭代轮数	全连接层 [20,256,256,1] [ReLU, ReLU, ReLU] SGD 0.01 64 100
梯度提升回归	最大迭代次数 n_estimators 学习率 learning rate 子采样 subsample 损失函数 loss	100 0.01 0.8 ls
XGBoost	决策树的个数 学习率 每个决策树的最大深度 子样本的比例 每棵树随机选取的特征的比例 损失阈值 L1 的正则化参数 L2 的正则化参数	600 0.3 5 0.6 0.8 0.2 0.05 0.1
随机森林回归	弱分类器的个数 n_estimators 子树评估标准 criterion 树的最大深度 max_depth 叶子节点最少样本数 min_samples_leaf	1000 'gini' 20 1

### 6.3.2 数据集准备与评价指标

#### 6.3.2.1 数据集准备 由于数据量少，所以选择9-1划分

根据问题一的结果，目前的训练数据集包含 1947 条样本，每条样本的维度为 252，即  $X \in \mathbb{R}^{1947 \times 252}$ 。为了评估模型性能，本文采用随机划分的方法，将数据集拆分成训练集以及测试集。模型在训练集上进行训练，并在测试集上验证评价指标，并选择最终的预测模型。**考虑到，数据集规模较少，因此本文采用 9: 1 的随机划分方式**获得包含 1752 条数据的训练集  $X_{train} \in \mathbb{R}^{1752 \times 252}$  以及包含 195 条数据的测试集  $X_{test} \in \mathbb{R}^{195 \times 252}$ 。

对于数据预处理， $X_{train}$  与  $X_{test}$  分别使用标准化进行预处理，优化目标则取 pIC50，并利用归一化进行预处理。

#### 6.3.2.2 评价指标

对于本问题，生物活性值 pIC50 属于离散变量，因此预测该变量属于回归任务。为了全面的评估预测模型的性能，本文采取了多种评估方式，包括：

**平均绝对误差 (Mean-Absolute Error, MAE)** 是一种衡量多对预测值与真实值误差的度量方法，又称为 L1 范数，其计算公式为：

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}, \quad (6.23)$$

其中， $y_i$  为真实值， $\hat{y}_i$  预测值，由于绝对值化，因此避免了正负互相抵消。

**均方误差 (Mean-Square Error, MSE)** 也是一种衡量多对预测值与真实误差的度量方式，又称为 L2 范数，其计算公式为：

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}, \quad (6.24)$$

与 MAE 相比，平方误差对偏离较多的预测值给予了更严重的惩罚。

**均方误差 (Mean-Square Error, MSE)** 是 MSE 的算术平方根，其计算公式为：

$$RMSE = \sqrt{MSE}, \quad (6.25)$$

与 MSE 相比，它使误差的量纲与数据保持一致，以达到更精细的结果。

**R2 Score** 称决定系数 (coefficient of determination) 是一种度量特征中可由自变量解释的部分所占比例，从而衡量回归模型的表现。令平均真实值  $\bar{y}$  为：

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad (6.26)$$

总平方和  $SS_{tot}$  为：

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad (6.27)$$

残差平方和 $SS_{res}$ 为:

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (6.28)$$

从而, R 方 ( $R^2$ ) 为:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (6.29)$$

## 6.4 实验结果

### 6.4.1 模型性能分析

表 1-1 中展示了六种模型在 4 种评估标准下的效果, 其中每个指标的最优值由加粗字体给出。由于归一化后的 pIC50 数值本身不大, 因此 MAE, MSE, RMSE 的数值不能够直接反映出模型的精度, 因此需要结合 R2 来进行考量。可以发现, 基于集成模型的预测方法在 R2 上以及前 3 个参数上均优于单个模型, 甚至超越了 MLP 的结果。而 XGboost 不仅在所有指标上达到了极优的效果, 而且在 MAE 和 R2 上以较大差距超越了其他模型。因此在后续的任务中, 本文采用 XGboost 作为求解模型。

表 6-2 各个模型在不同评估标准下的结果

模型名称	MAE(训练 /测试)	MSE(训练 /测试)	RMSE(训练 /测试)	R2(训练 /测试)
核岭回归	0.1194(0.1205)	0.0216(0.0231)	0.1470(0.1368)	0.3587(0.3421)
支持向量机回归-rbf	0.0975(0.1196)	0.0149(0.0132)	0.1220(0.1276)	0.5204(0.4976)
多层感知机	0.0859(0.0965)	0.0135(0.0165)	0.1162(0.1036)	0.6496(0.6177)
梯度提升回归	0.0780(0.0816)	<b>0.0097</b> (0.1104)	0.0985(0.1016)	0.6926(0.6582)
随机森林回归	0.0722(0.0730)	0.0098(0.1017)	<b>0.0990</b> (0.1063)	0.6940(0.6489)
XGboost	<b>0.0704</b> (0.0739)	0.0099(0.1005)	0.0996(0.9683)	<b>0.7262</b> (0.7193)

### 6.4.2 模型预测稳定性分析

由于每次训练/测试数据的拆分不同, 且加之基于学习的方法的随机性。算法表现的稳定性也是需要考量的重要因素之一, 如图 1-2 所示, 本文展示了六种算法在三种指标(MAE, RMSE 以及 R2 Score) 下测试 100 次的结果。结果以箱线图的形式表现。其中, 六种颜色的矩形表示第 25% 的误差值与第 75% 的误差值, 越宽表明其算法随机波动性越强。其中的实心黑色横线表明了其 100 次实验误差的中值, 空心小矩形表示 100 次实验误差的均值, 均值越小, 表明平均而言, 该模型具有更好的效果。箱外的黑色实线表示误差的合理上下界, 超出该范围说明此次实验的误差不属于该模型的实际表现范围 (这也是需要通过大量实验验证模型性能的原因)。

总体而言, 三种基于集成学习的方法的表现在稳定性方面具有更高的水准。虽然神经网络被成为黑箱模型, 但在实际使用中, 它的稳定性也较为可靠。其次, XGboost 虽然具有一个稍大的波动范围, 但其平均精度要更优。

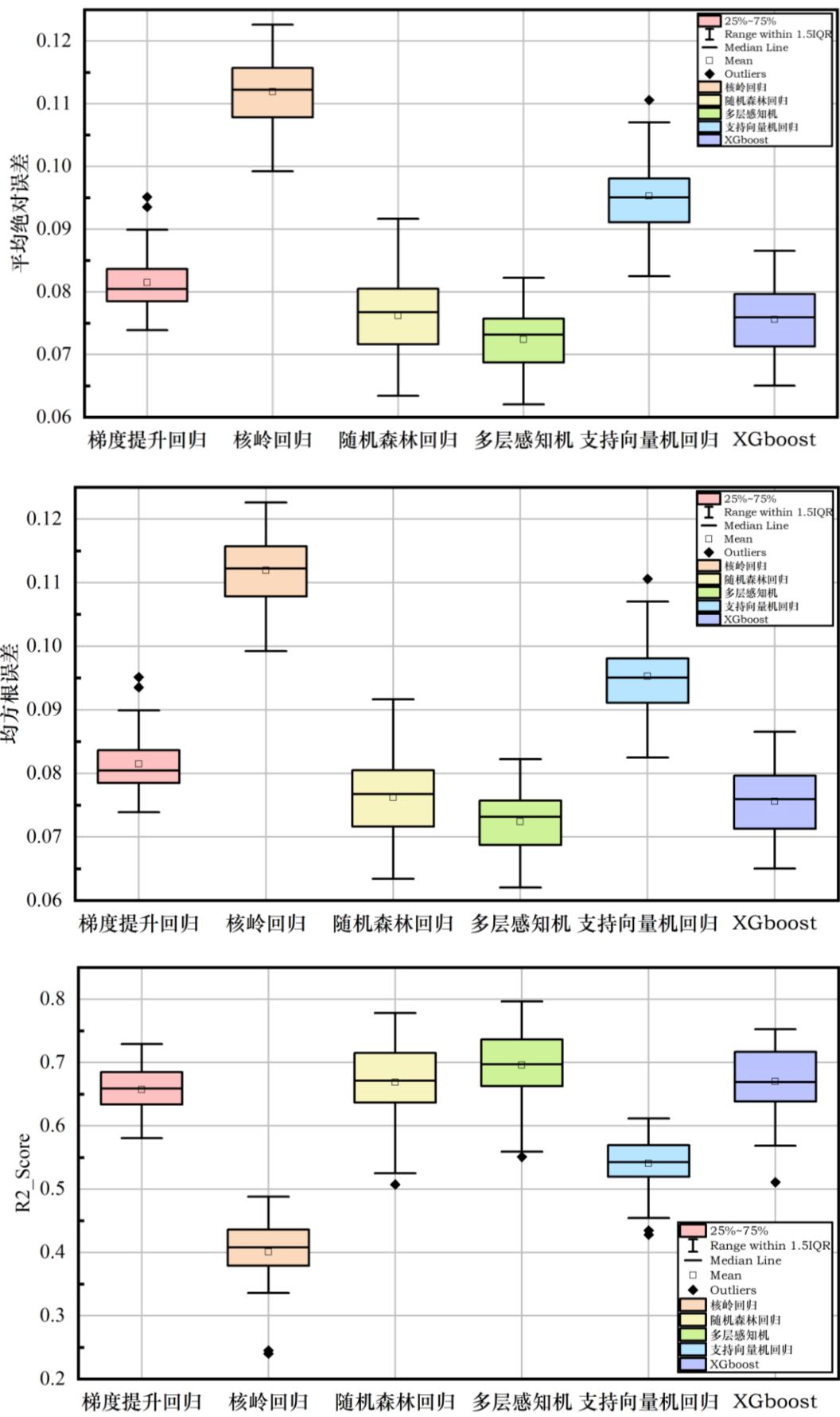


图 6.5 不同回归模型预测稳定性对比箱线图

### 6.4.3 结果展示

通过对 3 种单体模型以及 3 种集成模型实验，从精度以及稳定性上对其进行了全面的对于考量。XGboost 以其更优的测试精度以及测试稳定性超越了其他几个模型。因此在第二题的预测 pIC50 上，本文选择 XGboost 作为预测模型，对于 50 个测试样本的预测结果如表 6.3 所示。根据公式

$$IC50 = -\log_{10}(10^{-9} \cdot pIC50) \quad (6.30)$$

可以求得

$$pIC50 = 10^{-IC50} / 10^{-9} \quad (6.31)$$

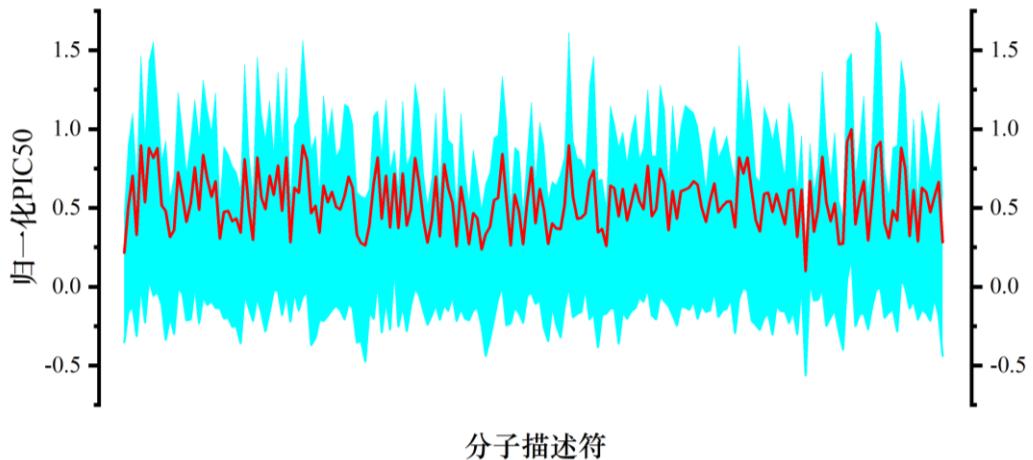


图 6.6 100 次随机实验在 198 个固定样本上的预测误差上下界

表 6-3 50 个测试样本的 IC50 以及 pIC50 预测值

编号	IC50 预测值	pIC50 预测值	编号	IC50 预测值	pIC50 预测值
1	3.1	8.509	26	1040	5.983
2	22.3	7.652	27	582	6.235
3	21.7	7.664	28	1640	5.785
4	14.5	7.839	29	3928	5.406
5	4.2	8.377	30	2823	5.549
6	59.4	7.226	31	6220	5.206
7	23152	4.635	32	5379	5.269
8	109	6.963	33	5267	5.278
9	74.8	7.126	34	4740	5.324
10	326	6.487	35	2860	5.544
11	71	7.149	36	80	7.097

12	1159	5.936	37	226	6.646
13	14.5	7.839	38	10425	4.982
14	57	7.244	39	173	6.762
15	69.5	7.158	40	11	7.959
16	51.2	7.291	41	179	6.747
17	11.7	7.932	42	5.5	8.260
18	634	6.198	43	173	6.762
19	407	6.390	44	93	7.032
20	4.4	8.357	45	179	6.747
21	147	6.833	46	1120	5.951
22	429	6.368	47	820	6.086
23	4677	5.330	48	1940	5.712
24	1129	5.947	49	800	6.097
25	195	6.710	50	21.7	7.664

## 6.5 小结

本题要求在不选择超过 20 个分子描述符变量的情况下，构建化合物对 ER  $\alpha$  生物活性的定量预测模型。第一问已得到 20 个对生物活性最相关的变量，可以直接作为该问的模型输入。由于 pIC50 与 IC50 是恒等变换，且 pIC50 分布更加集中，因此选择 pIC50 作为生物活性的目标值，同时对 train 数据集进行 9:1 随机划分，前者用于算法训练，后者用于验证该算法准确率。由于样本数量的限制以及分子描述符变量和 pIC50 之间可能存在高度的非线性关系，因此选用核岭回归与支持向量机回归预测器作为基学习器，并结合梯度提升回归集成学习构建训练模型。收敛后的模型在 train 数据集测试部分上的性能表明了我们模型的优越性。

选择这些模型的原因，及其结果的优越性

## 七、问题三

### 7.1 问题分析

问题三要求利用附件“Molecular\_Descriptor.xlsx”中所提供的 729 个分子描述符，针对附件“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。

根据题目要求，为了预测 Caco-2、CYP3A4、hERG、HOB、MN 等 5 个变量，需要分别构建其预测模型。与此同时，问题 1 中所选择的主要变量，由于其是与生物活性水平 Era 所相关的，所以在本题中不再使用。虽然为了每个 ADMET 数据可以重复问题一中的步骤单独构建变量选择流程，但为了后续的问题四，本题中摒弃了问题一的方法。而是重新构

## 阐明为什么没有沿用问题一的特征筛选方法

建了特征选择方法，使其对 Caco-2、CYP3A4、hERG、HOB、MN 等 5 个变量均具有高度的相关性。完整的流程如图 1-1 所示。首先，重复问题一中的数据筛选与与处理流程，然后借助 Xgboost 选择面向 ADMET 整体的最优变量，并且对于这五个特性，分别构建五个单独的预测模型。在随机划分的训练/测试集上验证模型后，选择最优模型，对 test 数据集中的变量进行 ADMET 的预测

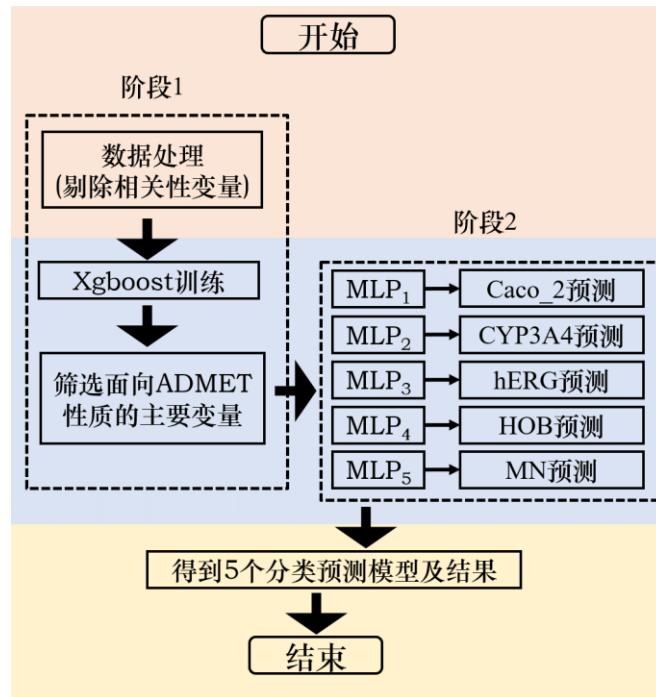


图 7.1 问题三的求解思路

## 7.2 模型建立

### 7.2.1 基于两阶段的 XGboost-MLP 生物活性特征预测模型

#### 阶段 1：XGboost 变量选择

在 5.3.2 中，本文已经详细介绍了 XGboost 最为回归预测模型的原理。在变量选择的训练过程中，为了提高生成新树的效率，XGboost 给出了每个变量在每次迭代中的重要性分数，以表示每个变量对模型训练的重要性，并为下一次迭代中建立具有梯度方向的新树提供依据。特征的统计显著性可以直接作为特征选择的依据。本文首先根据所有特征对 XGboost 进行分类，然后获得特征变量的重要性，并根据生成模型过程中的信息按降序对其进行排序。最后，将过滤后的特征输入二阶段的 MLP 分类器，以构建分类模型。

#### 阶段 2：基于并行 MLP 的生物活性特征预测

第一阶段的 XGboost 变量选择过滤了非重要特征。在第二阶段中，利用五个并行且不同的 MLP 对五个生物活性特征进行分别预测，值得注意的是，这五个特征属于离散变量，其取值为 0 或者 1。因此在优化模型时选择使用是二至交叉熵函数，如公式所示

$$F = -\sum_i t_i \log(MLP_n(X_i)) + (1-t_i) \log(1-MLP_n(X_i))$$

其中  $t_i$  表示某生物特性的真值， $MLP_n(X_i)$  表示对应的 MLP 预测模型对于第  $i$  条样本对该生物特征的预测值。  
**每个模型根据数据特点选择不同的目标函数，需要将目标函数的数学方程展示出来**

### 7.3 模型求解

#### 7.3.1 模型参数

表 7-1 所使用的 MLP 与 XGBoost 的模型参数设置

多层感知机	层的性质 层数及神经元个数(包括输入输出层)	全连接层 [20,256,256,1] [ReLU, ReLU, ReLU]
	MLP 每层的激活函数	
	优化器	SGD
	学习率	0.01
	批次大小	64
XGBoost	迭代轮数	100
	决策树的个数	600
	学习率	0.3
	每个决策树的最大深度	5
	子样本的比例	0.6
	每棵树随机选取的特征的比例	0.8
	损失阈值	0.2
	L1 的正则化参数	0.05
	L2 的正则化参数	0.1

#### 7.3.2 评价指标

在实验指标方面，本文使用二分类精度（Binary Classification Accuracy, Acc）以及曲线下面积（Area Under Curve, AUC）对算法进行评估。Acc 度量的是正确预测的数量占总预测数的比，如（1）所示：

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (7.1)$$

其中，TP(True Positive)表示正确预测的真样本，TN(True Negative)表示正确预测的假样本，FP(False Positive)表示错误预测的假样本，FN(False Negative)表示错误预测的真样本。

AUC 是比 Acc 更加综合且鲁棒的评价指标，其定义是受试者工作特征曲线（Receiver Operating Characteristic Curve, ROC）下的面积，ROC 曲线描述的是分类阈值变化时，真阳性率 TPR(True Positive Rate)和伪阳性率 FPR(False Positive Rate)的变化，可以通过式（2）与式（3）计算得到

$$TPR = \frac{TP}{TP + FN} \quad (7.2)$$

#### 7.3.3 结果分析与对比

如图 7.2 所示，为了证明所选择的 MLP 在二类分类任务上的优越性，本文对比了六种不同的模型在预测五种生物特性上的 ROC 曲线。该曲线下面积为 AUC 值，在表 7-2 中报告。可以发现 MLP 在多项特性的预测上均达到了最优水平。而根据表 7-2 可知，即使在类别不平衡的条件下，MLP 仍然能够达到很高的 AUC 值，没有受到类别失衡的影响。在 ACC 方面，其训练集平均准确接近 90%，而测试准确度仅略微低于训练精度，表明该模型

对于本问题，作者没有进行不均衡处理，而是从结果出发，认为 MLP 能够比较好的处理不均衡问题<sup>35</sup>

的过拟合风险也比较低。表 6-3 中报告了训练好的 MLP 模型对于附件中 test 数据集的预测结果。

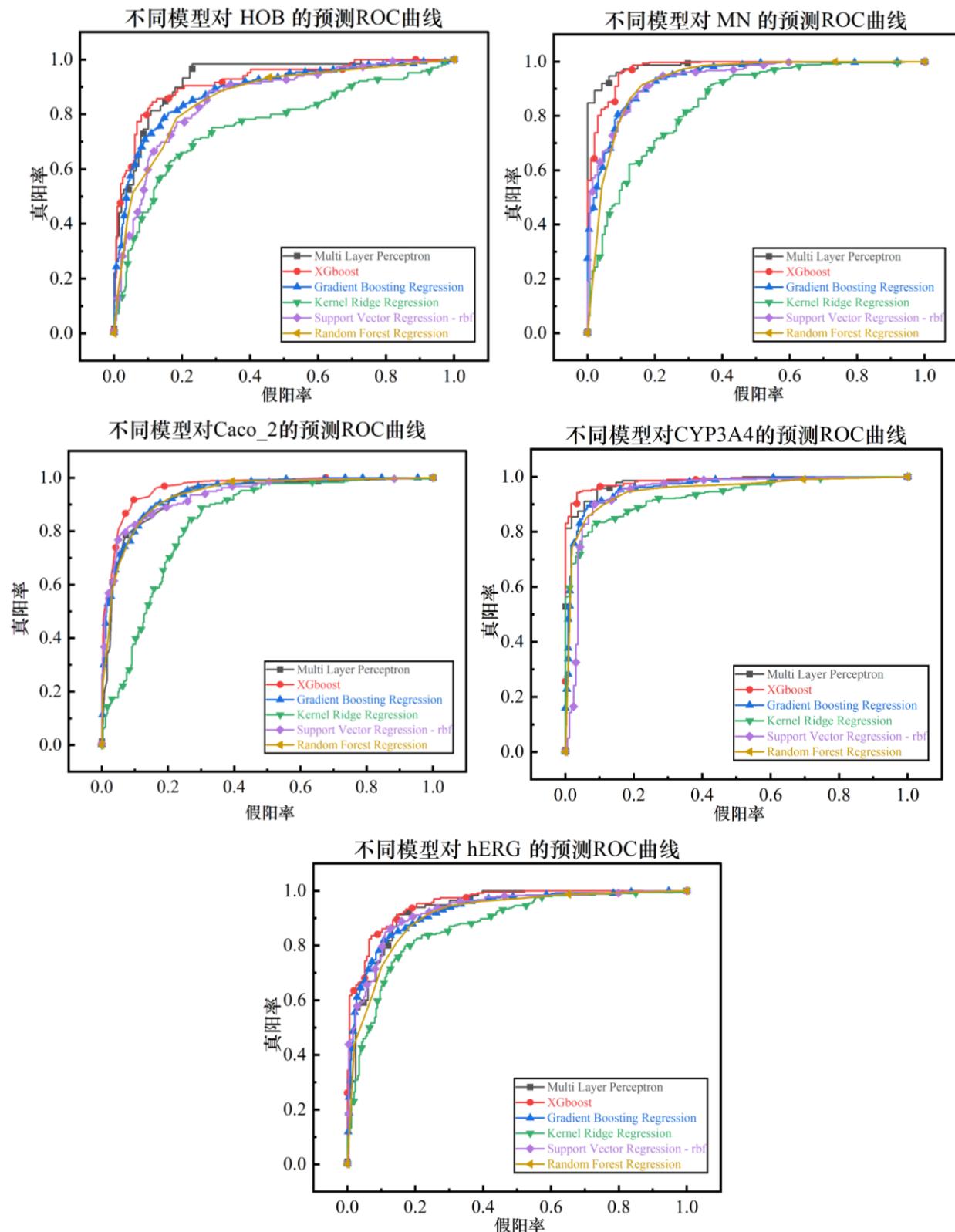


图 7.2 六种模型对五种生物活性特征的预测 ROC 曲线

表 7-2 MLP 对于五个生物特性预测的 ACC 以及 AUC

模型名称	训练 ACC	测试 ACC	训练 AUC	测试 AUC
Caco_2	87.37%	87.94%	0.9437	0.9451
CYP3A4	93.94%	93.46%	0.9749	0.9734
hERG	84.85%	85.12%	0.9414	0.9445
HOB	84.85%	83.49%	0.9179	0.9016
MN	96.46%	96.07%	0.9824	0.9814

表 7-3 50 个测试样本的 ADMET 预测值

编号	Caco-2	CYP3A4	hERG	HOB	MN	编号	Caco-2	CYP3A4	hERG	HOB	MN
1	0	1	1	0	1	26	1	1	0	1	0
2	0	1	0	0	1	27	1	1	1	1	0
3	0	1	1	0	1	28	1	0	1	1	1
4	0	1	1	0	1	29	1	0	1	0	1
5	0	1	1	0	1	30	0	0	1	0	1
6	0	1	1	0	0	31	0	0	1	0	1
7	0	1	1	0	1	32	0	0	1	0	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	1	0	1	0	1
11	0	1	1	0	1	36	1	0	1	0	1
12	0	1	1	0	1	37	0	0	1	0	1
13	0	1	1	0	1	38	0	0	1	1	0
14	0	1	1	0	0	39	0	1	1	1	1
15	0	1	1	0	0	40	0	1	1	1	1
16	0	0	1	1	1	41	0	1	1	0	1
17	0	1	1	0	1	42	0	0	1	0	1
18	0	1	0	0	0	43	0	0	1	0	1
19	1	1	0	0	0	44	0	0	1	0	1
20	0	1	1	0	0	45	0	0	1	0	1
21	0	0	0	0	0	46	0	0	1	0	1
22	0	1	1	0	0	47	0	0	1	0	1
23	0	1	1	0	0	48	0	0	1	0	1
24	0	0	1	00	0	49	0	0	1	0	1
25	0	0	1	0	0	50	0	0	1	0	0

## 7.4 小结

问题三要求利用 729 个分子描述符，构建化合物对五个 ADMET 数据 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型。虽然题目未限制模型输入，但是第一问的结果表明，仍然需要在训练之间对变量进行筛选。为了同时满足 ADMET 对变量的要求，因此重新构建了特征选择方法，构架了两阶段 XGboost-MLP 特征选择与并行预测方法，借助 Xgboost 选择面向 ADMET 整体的最优变量，使其对 Caco-2、CYP3A4、hERG、HOB、MN 等 5 个变量均具有高度的相关性，并且对于这五个特性，分别构建五个单独的基于 MLP 预测模型。在随机划分的训练/测试集上验证模型后，选择最优模型，对 test 数据集中的变量进行 ADMET 的预测。

## 八、问题四

### 8.1 问题分析

问题描述：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 ER $\alpha$  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）。

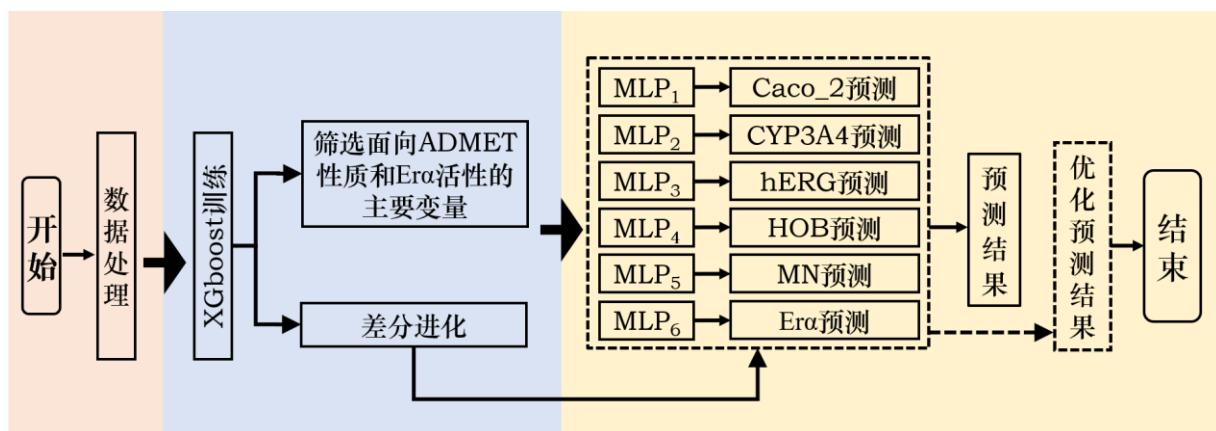


图 8.1 问题四解题思路

### 8.2 优化模型建立

#### 8.2.1 制定优化目标

对于问题 4，需要在保证化合物对抑制 Er $\alpha$  效果（即 pIC50 的值越大）的同时，使化合物具有更好的 ADMET 性质，即 Caco-2、CYP3A4、hERG、HOB、MN 分别为 1、1、0、1、0，或者至少有三个性质为以上值。对于任意一个化合物样本  $X_i$ ，根据数据清洗与变量选择算法，其最终是一个  $n$  维的向量： $X_i = [X_i^1, X_i^2, \dots, X_i^n]$ ，该化合物样本的 Caco-2、CYP3A4、hERG、HOB、MN 以及 pIC50 性质分别为

$$\begin{cases} \text{Caco-2} = \text{MLP}_1(X_i) \in \{0,1\} \\ \text{CYP3A4} = \text{MLP}_2(X_i) \in \{0,1\} \\ \text{hERG} = \text{MLP}_3(X_i) \in \{0,1\} \\ \text{HOB} = \text{MLP}_4(X_i) \in \{0,1\} \\ \text{MN} = \text{MLP}_5(X_i) \in \{0,1\} \\ \text{pIC50} = \text{MLP}_6(X_i) \in \mathbb{R} \end{cases} \quad (8.1)$$

该问题可以表述为，求解  $X_i = [X_i^1, X_i^2, \dots, X_i^n]$ ，约束条件为：

$$s.t. \begin{cases} \max(\text{MLP}_6(X_i)) \\ \sum_{n=1}^5 \text{MLP}_n(X_i) \geq 3 \end{cases} \quad (8.2)$$

### 8.2.2 基于 DE-MLP 的分子描述变量控制算法

差分进化算法 (Differential Evolution, DE) 是一种多目标优化算法，是在遗传算法 (Genetic Algorithm, GA) 的基础上提出的，利用模拟生物进化的方法，通过反复迭代，使得更适应环境的个体被保存下来。具体流程图如图 8.2 所示。

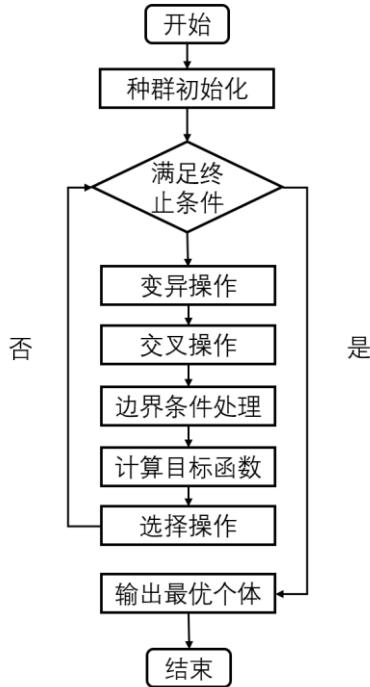


图 8.2 差分进化算法流程图

特别的是，差分进化算法保留了基于种群的全局搜索策略，基于差分的简单变异操作和一对一的竞争生存策略，降低了遗传操作的复杂性。同时，差分进化算法可以通过记忆动态跟踪当前的搜索情况，以调整其搜索策略，具有较强的全局收敛能力和鲁棒性。具体来说，差分遗传算法采用实数编码随机生成种群个体，种群由突变（包括变异和交叉操作）和选择过程驱动。其中，变异和交叉这两步操作被设计用于利用或探索搜索空间，而选择过程被用于确保有希望的个体的信息可以进一步利用。具体操作步骤见下表：

表 8-1 差分进化算法操作步骤

- 
- ① 从随机产生的初始父代群体中随机选取两个个体的差分矢量作为第三个个体的随机变化源
  - ② 将差分矢量加权后按一定规则与第三个个体求和而产生变异个体（**变异操作**）
  - ③ 变异个体与某个预先决定的目标个体进行参数混合，生成试验个体（**交叉操作**）
  - ④ 如果试验个体的适应度值优于目标个体的适应度值，则在下一代中试验个体取代目标个体，否则目标个体仍保存下来（**选择操作**）
  - ⑤ 在每一代的进化过程中，每一个体矢量作为目标个体一次，算法通过不断地迭代计算
  - ⑥ 保留优良个体，淘汰劣质个体，引导搜索过程向**全局最优解**逼近
- 

对于优化问题：

$$\min f(x_1, x_2, \dots, x_D) \quad \text{s. t. } x_j^L \leq x_j \leq x_j^U, j = 1, 2, \dots, D, \quad (8.3)$$

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。

算法流程如下：

### (1) 种群 $\{v_i(g+1)\}$ 的初始化

初始种群  $\{x_i(0) | x_{j,i}^L \leq x_{j,i}(0) \leq x_{j,i}^U, i = 1, 2, \dots, NP; j = 1, 2, \dots, D\}$  随机产生：

$$x_{j,i}(0) = x_{j,i}^L + \text{rand}(0,1) \cdot (x_{j,i}^U - x_{j,i}^L) \quad (8.4)$$

其中， $x_i(0)$  表示种群中第 0 代的第  $i$  个个体， $x_{j,i}(0)$  表示第  $i$  个个体的第  $j$  个基因。 $NP$  表示种群大小， $\text{rand}(0,1)$  表示在  $(0,1)$  区间均匀分布的随机数。

### (2) 变异操作

对种群中样本进行随机选取，每次选取两个不同的个体，将其向量差缩放后与待变异个体进行向量合成，即

$$(g+1) = x_{r_1}(g) + F \cdot (x_{r_2}(g) - x_{r_3}(g)), \quad i \neq r_1 \neq r_2 \neq r_3, \quad (8.5)$$

其中， $F$  为缩放因子， $x_i(g)$  表示第  $g$  代种群中第  $i$  个个体。

为了保证所求解的有效性，在进化的过程中必须判断个体中各基因是否满足边界条件，倘若未满足边界条件，则应用随即方法重新生成基因（与初始种群的生成方法相同）。

第  $g$  代种群  $\{x_i(g) | x_{j,i}^L \leq x_{j,i}(g) \leq x_{j,i}^U, i = 1, 2, \dots, NP; j = 1, 2, \dots, D\}$  通过变异后，产生一个中间体

$$\{v_i(g+1) | v_{j,i}^L \leq v_{j,i}(g+1) \leq v_{j,i}^U, i = 1, 2, \dots, NP; j = 1, 2, \dots, D\} \quad (8.6)$$

### (3) 交叉操作

对第  $g$  代种群  $\{x_i(g)\}$  及其变异的中间体  $\{v_i(g+1)\}$  进行个体间的交叉操作：

$$u_{j,i}(g+1) = \begin{cases} v_{j,i}(g+1), & \text{if } \text{rand}(0,1) \leq CR \text{ or } j = j_{rand} \\ x_{j,i}(g), & \text{otherwise} \end{cases}$$

### (4) 选择操作

差分进化使用贪心算法来选择进入下一代种群的个体：

$$x_i(g+1) = \begin{cases} u_i(g+1), & \text{if } f(u_i(g+1)) \leq f(x_i(g)) \\ x_i(g), & \text{otherwise} \end{cases}, \quad (8.7)$$

同时，为了避免早熟现象的出现，引入具有自适应变异的算子：

$$\lambda = e^{-\frac{G_m}{G_m+1-G}}, F = F_0 \cdot 2^\lambda \quad (8.8)$$

$$\lambda = e^{-\frac{G_m}{G_m+1-G}}, F = F_0 \cdot 2^\lambda \quad (8.9)$$

其中,  $F_0$  为变异算子,  $G_m$  代表最大进化代数,  $G$  代表当前进化代数。在算法进行的初始阶段, 自适应变异算子为  $F_0 \sim 2F_0$ , 具有较大值, 在初期保持个体多样性以避免早熟现象。随着算法的进行, 自适应变异算子逐步降低, 到后期接近  $F_0$ , 保留优良信息, 以增加最终搜索到全局最优解的概率。

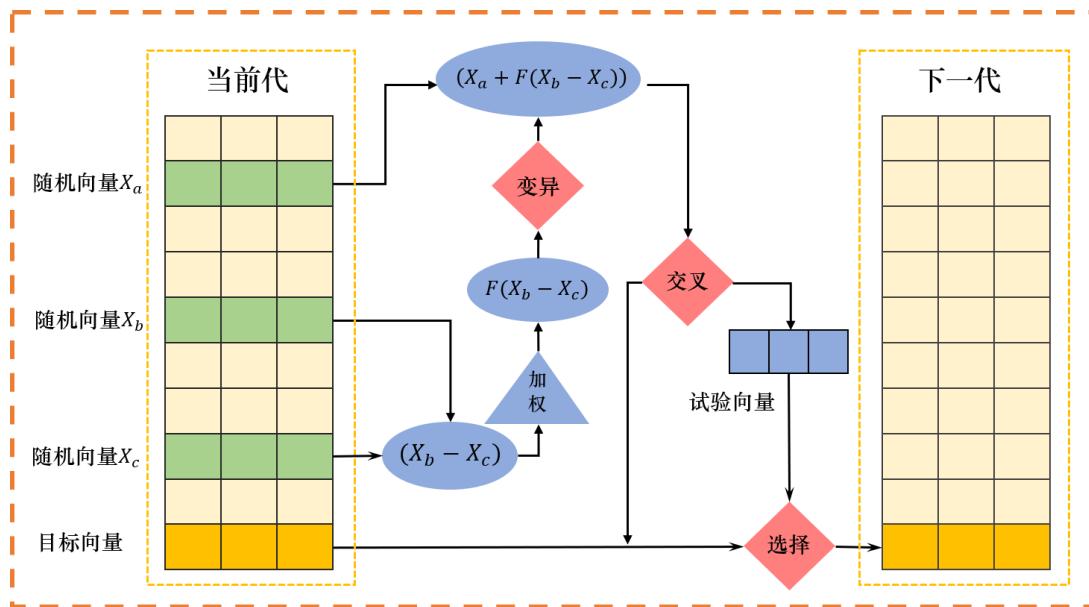


图 8.3 差分进化结构图

### 8.3 模型求解

#### 8.3.1 模型参数设置

表 8-2 所使用的 MLP 与 XBboost 的模型参数设置

	层的性质 层数及神经元个数(包括输入输出层) MLP 每层的激活函数	全连接层 [20,256,256,1] [ReLU, ReLU, ReLU]
多层感知机	优化器 学习率 批次大小 迭代轮数	SGD 0.01 64 100
差分进化	种群规模 变异系数	600 0.3

### 8.3.2 模型结果展示

表 8-3 符合要求的变量与取值

变量	取值(范围)	变量	取值(范围)
MDEC-23	47.8~50.8	C3SP2	0~6.8
LipoaffinityIndex	11	VCH-5	0.18~0.19
minsssN	2.5	ATSc2	-0.17~-0.16
minsOH	9.5~10.5	VPC-6	6~9.8
C1SP2	2.94~19.7	minHBint10	6~6.3
maxssO	5.3~6.4	ETA_Shape_Y	0.225~0.445
nHBAcc	26.4~66.4	maxHBint8	10.1
minHBint5	-1.2~-1	nHBint6	1.1~86
XLogP	9.1~9.5	mindO	14.4
ndssC	0~2.8	ALogP	1.9~5

### 8.4 小结

问题四要求在 ADMET 以及 ERA 水平都保持较优状态时，确定部分变量的取值范围。延续问题三，提出了基于 DE-MLP 的全局优化搜索算法，首先构建并训练了六个独立的 MLP，分别预测 ADMET 以及 Era 水平，在此基础上，将其作为差分进化算法的目标函数，通过不断的交叉编译，搜索能够使 ADMET 总体较优，且 Era 水平较好的化合物分子变量取值范围，为药物研发提供了一定的帮助。

## 九、模型的评价

本文提出了一种自适应迭代变量选择模型，依次通过数据标准化、多种变量筛选机制联合投票和主要变量独立性分析，有效地从 729 维变量中选择出前 20 个对生物活性影响最显著的变量。实际上，此举是为了降低特征维数，为后续的模型学习提供健壮的数据。为此并且，后续对 ERA 预测的实验所展现的高精度有力支持了该变量选择模型。此外，由于所提出的变量选择模型与本次数据集本身解耦，因此可以理论上也适用于其他需要对高维输入变量筛选的场景，具有一定的现实意义。本文还使用了两阶段 XGBoost-MLP 模型应用于生物活性特征预测。在使用常规的 MLP 建立预测模型之前，首先通过 XGBoost 技术对模型输入进行了筛选。与其他方法横向对比的实验数据中，该联合模型在正确率上都表现出了显著的优势。由此可见，面对其他类似高维度输入、低样本的数据预测任务，该方法也具有良好的可适用性。本文还提出了基于 DE-MLP 的分子描述变量控制算法。面对带有约束条件的多目标优化的情况，使用差分进化和多层感知机技术进行分别处理。特别是差分进化能够保留基于种群的全局搜索策略，在 MLP 进行生物活性 ADMET 预测的时候，依然能够提供近似全局最优的解。

总的来看。本文提出的模型在实验技术上都表现出了优秀的性能，且对其他类似的问题都有良好的可推广性。

## 参考文献

- [1] 曹毛毛 and 陈万青, "GLOBOCAN 2020 全球癌症统计数据解读," *中国医学前沿杂志*, vol. 13, no. 3, pp. 63-69, 2021.
- [2] 张雅聪, 吕章艳, 宋方方, and 陈可欣, "全球及我国乳腺癌发病和死亡变化趋势," *肿瘤综合治疗电子杂志*, vol. 7, no. 2, pp. 14-20, 2021.
- [3] 贺利军, 李文锋, and 张煜, "基于灰色综合关联分析的多目标优化方法," *控制与决策*, vol. 35, no. 5, pp. 1134-1142, 2020.
- [4] Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms[J]. *J. Mach. Learn. Res.*, 2017, 18(1): 6345-6398.
- [5] 刘波,王凌,金以慧.差分进化算法研究进展[J].*控制与决策*,2007(07):721-729.
- [6]李欣海.随机森林模型在分类与回归分析中的应用[J].*应用昆虫学报*,2013,50(04):1190-1197.