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	电子科技大学
参赛队号	21106140116
队员姓名	**1.**张居虹
	2.杨洁
	3.白强

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：

雌激素受体 α 亚型在乳腺癌的治疗过程中具有重要的作用，从能够拮抗 ERα 活性的化合物中，挑选同时具备良好的生物活性和 ADMET 性质的化合物作为治疗乳腺癌的候选药物。因此，在乳腺癌的药物研发中，需要根据分子描述符变量，采用有效的数据挖掘方法建立对于化合物生物活性的定量预测模型，以及关于 ADMET 性质的分类预测模型，并且对 ERα 拮抗剂的生物活性进行合理的优化和对 ADMET 性质科学的预测成为乳腺癌治疗过程中的重要组成部分。

针对问题一，对具有 1974 个化合物样本的原始数据集进行**数据清洗**，发现全部为 0 的冗余分子描述符，因此将其剔除，然后对数据**标准化**处理。将数据集以 7:3 的比例分为训练集和验证集，采用**随机森林**的算法对分子描述符变量进行挑选，得到在化合物训练集和测试集上的准确度分别为 **98.3%、75.9%**，由于因变量是生物活性值（连续变量），所以采用**均方差（MSE）值**大小判断变量的重要程度，挑选出前 20 名对生物活性值影响最显著的变量。

针对问题二，需要对 50 个化合物进行 IC_{50} 值和对应的 pIC_{50} 值预测，使用问题一提取的 20 个重要变量，对化合物数据标准化后，以 7 : 3 分为训练集和验证集，采用生物活性 pIC_{50} 值，作为预测模型的因变量，比较**随机森林回归树**、**梯度提升回归树**和**支持向量机回归（SVR）模型**在训练集和验证集上的真实值与预测值的分布差异，准确度分别为：**72.8%**、**73.0%**和 **69.2%**，梯度提升树表现最好，因此使用梯度提升树对 50 个化合物 pIC_{50} 值进行预测，然后再通过数值转换得到 IC_{50} 值。

针对问题三，对 50 个化合物进行五个 ADMET 性质的分类性能预测，先对样本数据做最大最小值标准化处理，由于变量存在多余的信息，使用主成分分析（PCA）法对分子描述符变量进行提取主要信息，提取使得累计方差贡献率达到 90%的 37 个主要成分，对每个 ADMET 性质，都分别采用 **logistic 分类**、**决策分类树**、**随机森林分类树**、梯度提升**分类树**、**支持向量机分类**（SVM）和 BP 神经网络分类算法进行预测，其中 Caco-2、CYP3A4、hERG、HOB、MN 的最高准确度分别为：**91.7%**，**94.9%**，**91.3%**，**85.6%**，**96.1%**。

针对问题四，为能够使化合物对抑制 ERα 具有更好的生物活性，确定分子描述符的取值，首先根据生物活性 **IC_{50} 值的转折点**（对应的 pIC_{50} 等于 7 ），挑选前 700 个 pIC_{50} 大于 7 的化合物，满足化合物对抑制 ERα 具有较好的生物活性。结合问题一筛选出的 20 个分子描述符变量作为自变量，将五个 ADMET 性质做积分累计处理转换为一个优化目标属性作为因变量，利用**随机森林分类算法**对数据进行训练得到优化目标函数，以前 700 个 pIC_{50} 大于 7 的化合物对应分子描述符变量的最大最小值作为变量的约束条件，从而构建一个多约束优化问题，采用**粒子群算法**对该问题进行求解，结果显示在优化得到的区间范围内，对应至少三个 ADMET 性质表现好。

关键词：生物活性值，随机森林，梯度提升决策树，BP 神经网络、粒子群优化

目录

1 问题重述

1.1 问题背景

乳腺癌（breast cancer, BC）是女性中常见的一种恶性肿瘤疾病[1], 也是目前世界上最常见，致死率较高的癌症之一，并且乳腺癌的发病率不断的呈现上升趋势，找寻相关药物有效防治乳腺癌是热门前沿的科学问题。

首先，乳腺癌的发展与雌激素受体密切相关[2]，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。对 ERα 基因缺失小鼠的实验结果表明，ERα 确实在乳腺发育过程中扮演了十分重要的角色，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。其次，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[3]（Absorption 吸收、Distribution 分布、Metabolism 代谢、Excretion 排泄、Toxicity 毒性）性质。其中，ADME 主要指化合物的药代动力学性质，描述了化合物在生物体内的浓度随时间变化的规律，T 主要指化合物可能在人体内产生的毒副作用。一个化合物的活性再好，如果其 ADMET 性质不佳，比如很难被人体吸收，或者体内代谢速度太快，或者具有某种毒性，那么其仍然难以成为药物，因而还需要进行 ADMET 性质优化。这里仅考虑化合物的 5 种 ADMET 性质，如下表所示：

ADMET 性质	作用
小肠上皮细胞渗透性（Caco-2）	度量化合物被人体吸收的能力
细胞色素 P450 酶（Cytochrome P450, CYP）3A4 亚型（CYP3A4）	人体内的主要代谢酶，可度量化合物的代谢稳定性
化合物心脏安全性评价（human Ether-a-go-go Related Gene, hERG）	度量化合物的心脏毒性
人体口服生物利用度（Human Oral Bioavailability, HOB）	度量药物进入人体后被吸收进入人体血液循环的药量比例
微核试验（Micronucleus，MN）	检测化合物是否具有遗传毒性

因此面临的一个重要实际问题是：**如何找到拮抗 ERα 活性且在人体内具备良好的 ADMET 性质的化合物？**

为了评估化合物生物活性的治疗效果，需要观测一系列化合物的分子描述符变量，以及化合物的 ADMET 性质，然后由研究者对其进行分析，在这些观测指标中，化合物的生物活性和 ADMET 性质用来衡量拮抗 ERα 活性的化合物可能是治疗乳腺癌的候选药物，化合物对 ERα 的生物活性值（用 IC50 表示，为实验测定值，单位是 nM，值越小代表生物活性越大，对抑制 ERα 活性越有效）；在实验研究中，通常将 IC50 值转化而得的 pIC50（即 IC50 值的负对数，该值通常与生物活性具有正相关性，即 pIC50 值越大表明生物活性越高；实际 QSAR 建模中，一般采用 pIC50 来表示生物活性值）。

建模目标：根据提供的 ERα 拮抗剂信息（1974 个化合物样本，每个样本都有 729 个分子描述符变量，1 个 pIC50 数据，5 个 ADMET 性质数据），构建化合物 pIC50 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 pIC50 和 ADMET 性质提供预测服务，以此来预测更好的 pIC50 的新化合物分子，或者指导已有活性化合物

的结构优化。

1.2 问题解析

1.2.1 问题一解析

针对文件“Molecular_Descriptor.xlsx”和“ERα_activity.xlsx”提供 1974 个化合物的 729 个分子描述符进行变量**特征选择**，对化合物样本数据进行预处理，结合分子描述符特征选择的降维方法，挑选出能够对拮抗 ERα 活性的化合物大小影响最显著的的分子描述符变量，需要尽可能使被选择的变量对回归预测问题具有显著的影响，且降维后的特征具有可解释性，要求寻找重要程度排名前 20 个分子描述符变量，并且需要说明分子描述符详细的筛选过程及其合理性。

步骤一：读取化合物样本数据，对数据进行清洗，检查数据有缺失值、离群点、冗余数据等异常数据，分子描述符变量是否存在量纲、数量级的差异。

步骤二：对数据进行预处理后，得到初步的自变量数据，将化合物对应的 pIC50 作为因变量，尝试构建自变量与因变量之间关系的描述方式，从而筛选出描述变量符中对于 pIC50 值重要的变量符。

步骤三：建立分子描述符变量和化合物 pIC50 之间的潜在对应关系，采用决策树、随机森林和梯度提升树分别构建模型，为防止模型过拟合将数据按照 7:3 均匀采样分成训练集和测试集，选择预测准确度最高的模型进行分子描述符变量的筛选。

步骤四：利用选择的模型，使用均方差（MSE）衡量分子描述符变量的贡献程度，挑选贡献程度排名前 20 个的分子描述符变量作为对 pIC50 值影响最为显著的变量从而解决问题一。

1.2.2 问题二解析

结合问题一挑选出的变量，需要选择出不超过 20 个分子描述符变量，因此需要通过数据挖掘技术和机器学习方法，针对已知 1974 个化合物，构建问题一中挑选出的 20 个分子描述符变量和 pIC50 之间的潜在模型关系，通过分析题目可以得知这是典型的**有监督回归问题**，可以通过构建相关有监督回归问题的模型从而有效解决问题。

步骤一：首先利用问题一挑选出的 20 个重要变量，对 1974 个化合物分子描述符的原始数据集进行处理的到 1974×20 的新数据集，为后续模型的构建方便，对新数据集进行标准化处理。

步骤二：对化合物数据集按 7:3 比例分为训练集和验证集，分别使用随机森林、梯度提升树和 SVM（支持向量机）对训练集进行拟合训练，同时需要不断的调试模型的相关参数，的到相对较好的模型。

步骤三：分别训练好的三个模型运用到化合物验证集数据集中，比较随机森林、梯度提升树和 SVM（支持向量机）的准确度大小，选择准确度高的模型，对 50 个化合物的 pIC50 值进行预测。

1.2.3 问题三解析

需要建立化合物样本的分子描述符变量数据和 ADMET 性质（包括：Caco-2、CYP3A4、hERG、HOB、MN）数据集之间的模型对应关系，并为“ADMET.xlsx”的 test 表中的 50 个化合物进行相应分类预测。由于自变量过多，需要采用降维算法剔除原始数据中的冗余信息，然后分别建立降维后的自变量和 ADMET 五个性质的分类预测模型，并对模型进行验证说明模型构建的合理性和有效性，这是一个**有监督分类问题**。

步骤一：首先采用最大最小值处理方法对化合物样本数据集进行预处理，由于分子描述符变量过多，通过利用主成分分析法（PCA）对化合物数据集中的分子描述符变量进行降维处理，有利于后续关于 ADMET 性质分类预测模型的构建。

步骤二：选取使得化合物数据集的累计方差贡献率大于 90%以上的主成分，作为构建 ADMET 性质分类预测模型的自变量。

步骤三：以 7:3 比例将降维后的数据分为训练集和验证集，对每一个二分类问题分别使用 Logistic、决策树、随机森林、梯度提升树、SVM、BP 神经网络 6 中及其学习算法对训练集进行拟合，并比较模型在验证集上的效果。

步骤四：选择对化合物 ADMET 性质预测准确度最高的模型，完成对 50 个化合物的 5 个 ADMET 性质的预测。

1.2.4 问题四解析

寻找化合物的分子描述符变量，使得化合物对抑制 ERα 具有更好的 pIC50，并且确定分析描述符的取值（或取值范围），同时要求给定的五个 ADMET 性质中，至少三个性质表现得较好。这是一个**多约束优化问题**，以化合物的 pIC50 和五个 ADMET 性质同时作为因变量，构建分子描述符—化合物活性和 ADMET 性质的之间的目标函数，并使用相关优化算法得到自变量的取值区间。

步骤一：使用问题一的 20 个有重要影响程度的分子描述符变量作为本题的自变量，以及 pIC_{50} 值排名靠前的化合物样本，从而得到对应 20 个变量的取值范围即设置为优化问题中的约束条件，从而完成分子描述符的筛选工作。

步骤二：对 ADMET 性质进行转换，以转换后的 ADMET 性质积分和作为优化目标，采用随机森林分类算法对其积分进行预测，得到多约束优化问题中的目标函数。

步骤三：以挑选出的分子描述符变量值的上下界，作为化合物 ADMET 性质模型的约束条件，以随机森林模型作为优化目标，采用粒子群算法对模型进行优化，保证至少三个 ADMET 性质表现得较好，即：优化结果显示转换后的 ADMET 性质的总积分和大于或等于 3。

1.2.5 总技术路线图

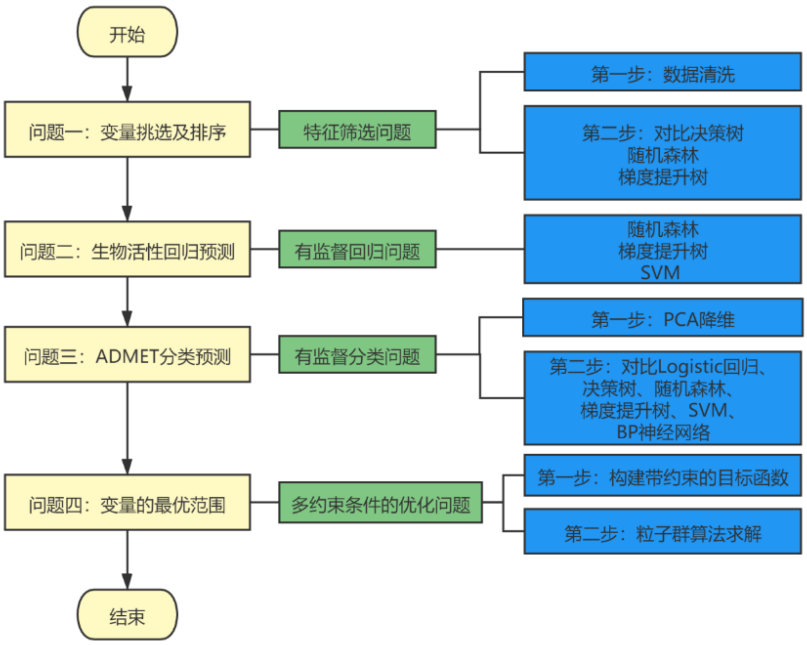


图 1-1.问题的研究总思路

2 模型假设

假设 1：所有数据的测量均正确，数据处理步骤均无误。

假设 2：分子描述符变量的范围均在合理范围内。

假设 3：生物活性和 ADMET 性质仅与题中所给的 729 中分子描述符有关，不再有其他变量的干扰。

假设 4：假设每修过一次涉及到的分子描述符极其数值，生物活性和 ADMET 性质均会及时发生相应的变化。

3 符号说明

符号	含义
x	样本数据值
$\bar{x}$	对样本数据做标准化处理后的值
μ	原样本数据的均值
σ	原样本数据的方差
γ_1	某一种分子描述符在实际中可能取得的最小值
γ_2	某一种分子描述符在实际中可能取得的最大值
c_1, c_2	两个$[0,1]$上的随机数
r_1, r_2	表示学习因子，用以调节学习的最大步长
ω	表示惯性因子，用以调节解空间的搜索范围
x_i	筛选出来的第i个分子描述符

4 主要变量的筛选与合理性验证

4.1 问题一的分析

第一步，对数据进行预处理。对化合物的分子描述符变量进行挑选，并对其进行重要性排序的时候，首先需要对数据集中的脏数据进行一些初步检验，查看化合物样本数据是否存在缺失值等异常情况，通过简单的观测和绘图等方法，可以很容易发现，化合物数据不存在异常数据。原文件中的数据集包括共 1974 个化合物样本，每个化合物样本数据中

共包括有 729 条分子描述符变量，而在 729 条分秒描述变量中可观察到大量描述变量对应的属性值其平均值接近 0 且方差极小的情况（如图 4-1 所示）。因此这些描述变量是冗余的，需要对数据首先进行预处理——清洗工作。

SMILES	nB	nP	nF	nCl	nBr	nI	nX	nBondsT
Oc1ccc2O[C@H]([C@H](Sc2c1)C3CCCC3)c4ccc(OCCN5CCCCC5)c	0	0	0	0	0	0	0	0
Oc1ccc2O[C@H]([C@H](Sc2c1)C3CCCCCC3)c4ccc(OCCN5CCCCC	0	0	0	0	0	0	0	0
Oc1ccc(cc1)[C@H]2Sc3cc(O)ccc3O[C@H]2c4ccc(OCCN5CCCCC5)c	0	0	0	0	0	0	0	0
Oc1ccc2O[C@H]([C@@H](CC3CCCCC3)Sc2c1)c4ccc(OCCN5CCCC	0	0	0	0	0	0	0	0
Oc1ccc2O[C@H]([C@@H](Cc3ccccc3)Sc2c1)c4ccc(OCCN5CCCCC	0	0	0	0	0	0	0	0
Oc1ccc2O[C@H]([C@H](Sc2c1)c3ccncc3)c4ccc(OCCN5CCCCC5)cc	0	0	0	0	0	0	0	0
Oc1ccc(cc1)C2=Cc3cc(O)ccc3C24Cc5ccc(OCCN6CCCCC6)cc5C4	0	0	0	0	0	0	0	0
Oc1ccc2O[C@H]([C@H](Sc2c1)c3ooccc3)c4ccc(OCCN5CCCCC5)cc4	0	0	0	0	0	0	0	0
Oc1ccc(cc1)C2=Cc3cc(O)ccc3C24Cc5ccccc5C4	0	0	0	0	0	0	0	0
Oc1ccc(cc1)C2=Cc3cc(O)ccc3C24C(=O)c5ccccc5C4=O	0	0	0	0	0	0	0	0
Oc1ccc(cc1)C2=Cc3cc(O)ccc3C24C(=O)c5ccc(OCCN6CCCCC6)cc5	0	0	0	0	0	0	0	0
Oc1ccccc(c1)C2=Cc3cc(O)ccc3C24Cc5ccccc5C4	0	0	0	0	0	0	0	0
CC(C)(C)[C@H]1Sc2cc(O)ccc2O[C@H]1c3ccc(OCCN4CCCCC4)cc3	0	0	0	0	0	0	0	0

图 4-1 存在多种性质变量均为零的现象

其次，变量数据之间有无量纲和数量级的差异，如果出现该问题，需要对数据进行标准化或归一化处理。

第二步，寻找回归预测模型的显著影响变量。本问希望对 725 个描述变量进行降维操作，从所有分子结构描述符变量中提取出寻找建立生物活性预测模型的前 20 个显著影响变量，使挑选出的显著影响变量具有代表性。首先，通过已知的化合物样本数据，首先需要初步诊断数据类型，可以发现化合物样本数据的自变量是连续型数值型变量，对应的样本数据的因变量也是连续型数值型变量，因此可以建立有监督的回归预测模型，通过分析预测模型中的具有关键影响力的变量从而得到前 20 个显著影响变量；其次，本题希望可以从 729 条分子描述符信息中，寻找对化合物作用效果起到显著影响的分子描述符信息，即：对分子描述符进行降维。常见的降维方法之一——PCA（主成分分析）方法，虽然 PCA 方法可以对数据降维，提取分子描述符变量的主要信息成分，但是主成分的可解释性不强，而且本题要求对变量的重要程度进行排序处理，这需要对原有数据的变量进行操作处理，而不能对变量本身进行空间维度变换，因此像 PCA 这类的降维方法在这里并不适用，另外，本题要求结合挑选出来 20 个变量具有很强的可解释性和并与 pIC50 高度相关，因此需要可选用随机森林回归、梯度提升树回归等挑选显著影响变量。

针对难点（1）——代表性变量选择问题，关于高维数据的降维算法主要分为两大类别，一类是特征变化，另一类是特征选择，特征选择是从原有的数据中，通过某种判别准则或筛选条件，直接从里面选择出可用于建模，具有高度代表性的重要特征。由于问题一需要筛选出来的特征具有强的可解释性，因此我们使用特征选择——随机森林的降维方法进行处理。

针对难点（2）——在已知数据中，由于生物活性数据有两种表达形式，其中 pIC50 值（IC_{50} 值的负对数）是常用数据，因为 pIC50 值值越大，说明生物化合物活性越好，对抑制 ERα 活性越有效，这样更容易用坐标图表述，并且在后面问题的分析中，使用回归分析的条件之一就是其残差服从正态分布，而使用负对数值的话，正好可以满足该条件。尤其是建模过程中，负对数变换可以使数据平稳化，减弱数据模型的共线性，异方差性。

最后，从变量降维过程中采用的算法及处理流程以及变量降维的最终结果两方面对所选择变量的合理性进行评价。问题一的思路流程图如图 4-2 所示。

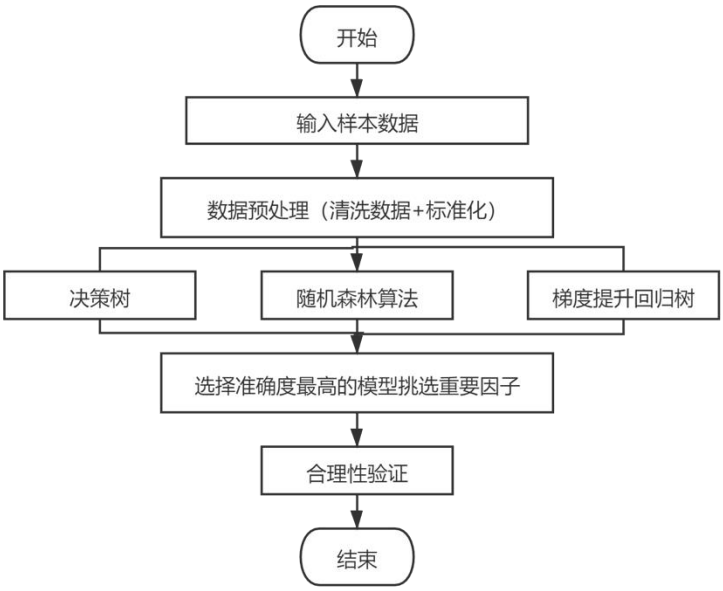


图 4-2 问题一的流程图

4.2 数据预处理

4.2.1 数据清洗

首先使用 python 编程语言对数据进行可视化，观察每一个描述自变量和对应预测因变量之间的分布关系。

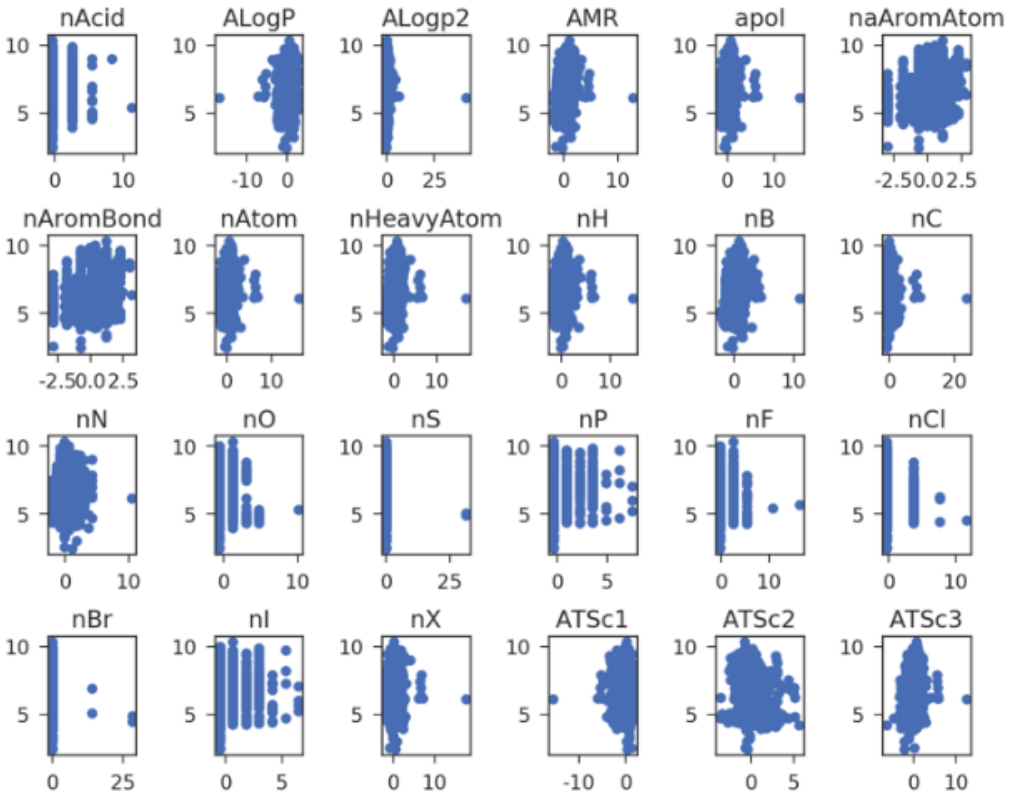


图 4-3 自变量和对应预测因变量之间的分布关系

通过对可视化发现，有些自变量在因变量发生变化时几乎保持不变，如图 4-3 中自变量 nS 和 nBr，这表明该特征数据是冗余数据，因此需要将这些自变量特征剔除。

	nAcid	ALogP	ALogp2	AMR	apol	naAromAtom	nAromBond	nAtom	nHeavyAtom	nH	...	MW
count	1974.000000	1974.000000	1.974000e+03	1974.000000	1974.000000	1974.000000	1974.000000	1974.000000	1974.000000	1974.000000	...	1974.000000
mean	0.108409	1.110164	3.288495e+00	116.557106	60.626471	15.446809	16.189463	50.761905	28.112462	22.649443	...	391.056697
std	0.347900	1.434250	1.283292e+01	31.567455	19.449748	5.155854	5.635271	18.089182	8.073881	10.775491	...	111.596103
min	0.000000	-23.105000	3.600000e-07	54.067000	30.661930	0.000000	0.000000	21.000000	14.000000	5.000000	...	194.094294
25%	0.000000	0.376300	4.052598e-01	88.303700	44.432102	12.000000	12.000000	36.250000	21.000000	14.000000	...	303.981576
50%	0.000000	1.170950	1.560251e+00	114.837500	59.901376	16.000000	18.000000	50.000000	28.000000	22.000000	...	386.036615
75%	0.000000	1.948100	4.018823e+00	141.423650	74.421376	18.000000	18.000000	62.000000	34.000000	29.000000	...	463.195900
max	4.000000	5.181700	5.338410e+02	517.429400	359.662740	30.000000	34.000000	343.000000	163.000000	180.000000	...	2349.392344

8 rows × 729 columns

图 4-4 自变量对应的统计特性

如图 4-4 所示，通过计算特征变量的相关统计特性（如平均值、标准差、最大值、最小值等），若最大值等于最小值则表明该特征变量为无效特征自变量，经统计发现有 225 个无效变量，将其剔除后我们得到了 504 个剩余自变量，方便后续继续试验。

4.2.1 数据标准化

根据上一节的分析，为后面处理化合物样本数据方便，对化合物数据进行标准化处理，经过处理后数据符合标准正态分布。这样处理的好处有两点：能够提升模型精度的同时达到提高求解中的收敛速度的效果。标准化的计算过程如下：

$$\bar{x} = \frac{x-\mu}{\sigma} \tag{4-1}$$

其中，x 表示标准化之前的数据，μ,σ 表示原数据的均值和方差，$\bar{x}$ 表示标准化处理后的数据。

4.3 数据特征筛选

随机森林算法是以 K 个决策树$\{H(X,\theta_i), i=1,2,\cdots,K\}$为基本分类器，在本文中，设定 K=500，对化合物样本进行集成学习后，可以得到的一个关于分子描述符变量组合的分类器[4]。当时输入待分类化合物样本时，该算法输出的关于化合物样本的分类结果，会根据每个决策树的分类进行简单的分子描述符选择给出。其中，$\{\theta_i, i=1,2,\cdots,K\}$是随机的，由随机森林的下面两种思想得到（结合题目描述为）：

（1）Bagging 思想：从原化合物样本集 X 中，随机取 K 个与原化合物样本集同样规模的化合物训练集，这是可放回的抽化合物样本，通过每个获得的化合物样本集，构造一个对应的决策树。（本文设定 K=500）

（2）特征子空间思想：对每个决策树进行分裂时，从全部属性中等概率的随机抽取一个属性集（通常取$\log_2 m+1$个属性，m 为特征总数），再从这个子集中选取一个最优属

性来分裂节点。（本文选择 m=10）

在对化合物样本数据集训练随机森林时，类似于决策树其。第 K 个决策树的训练过程如图*所示，依次类推得 K 个决策树，进行组合得到一个随机森林。

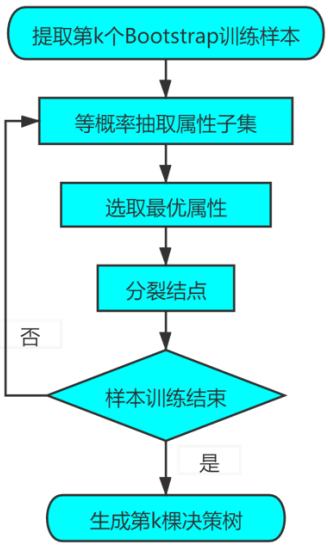


图 4-5 第 K 个决策树的训练过程

随机森林采用了两大随机化思想，尽可能避免对每一个化合物样本都考虑的现象，即：过拟合，同时在一定程度上，减弱噪声的影响，使得化合物样本的分子描述符变量在不减少的情况下，该算法可以获得每个分子描述符的重要性。

本题的因变量是一个连续变量，因此随机森林算法是在分类基础上进行回归分析，将样本分类的结果进行一定的运算，获得每个分子描述符变量对化合物 pIC50 的影响程度，随机森林算法对分子描述符的评价是根据内部所有决策树对分子描述符的平均值。而在决策树中，应对连续型因变量的回归问题，计算特征重要性采用的原则是 MSE（最小均方差）。

即对于任意划分分子描述符集合 A，任意划分点 s 两边划分成化合物 pIC50 值的数据集 D1、D2，求出使 D1、D2 各自化合物 pIC50 数据集的均方差最小，同时 D1 和 D2 的均方差之和最小所对应的分子描述符特征和对应值划分点。计算公式如下所示：

$$\min_{A,s}\left[\min_{c_1}\sum_{x_i\in D_1(A,s)}\left(y_i-c_1\right)^2+\min_{c_2}\sum_{x_i\in D_2(A,s)}\left(y_i-c_2\right)^2\right] \tag{4-2}$$

其中，c1 为 D1 数据集的样本输出均值，c2 为 D2 数据集的样本输出均值。

随机森林的预测是所有树的预测值的平均值。

使用 python 搭建随机森林回归模型，为防止模型过拟合，将输入的数据按照 7:3 分为化合物训练集和化合物测试集，得到训练集的大小为（1381,504）(1381,504)、测试集的大小为(593,504)。然后使用随机森林训练模型。在化合物训练集和化合物测试集上的准确度分别为 98.3%、75.9%，如图 4-6 所示，在测试集上整体拟合效果较好。

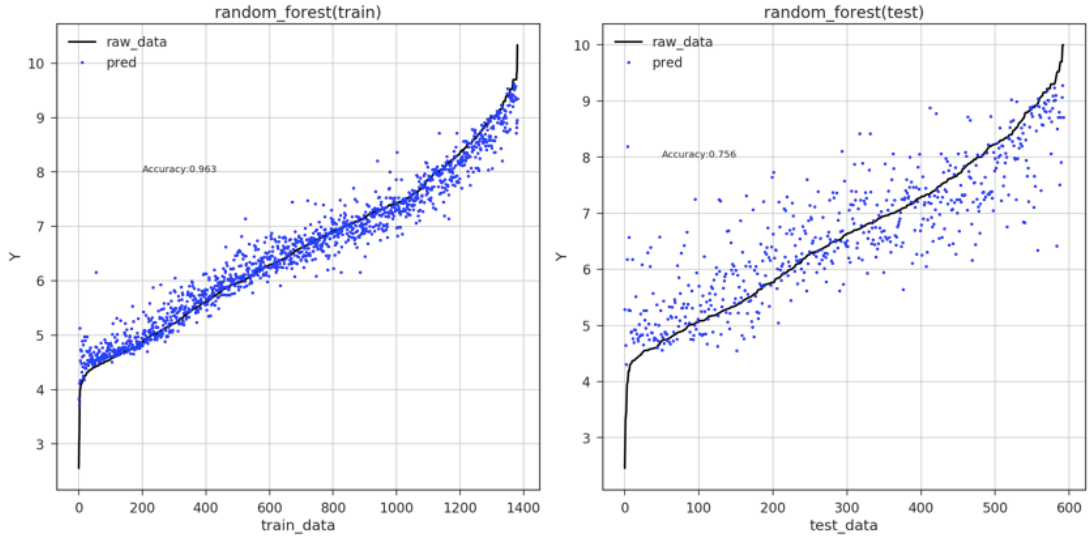


图 4-6 模型在训练集和测试集上的效果

将训练好的随机森林回归模型做特征挑选时，因为问题中因变量时连续变量，因此使用 MSE 进行特征重要性衡量。

根据上述特征筛选法则，筛选出前 20 名的显著影响变量，其对应的影响得分如图所示。

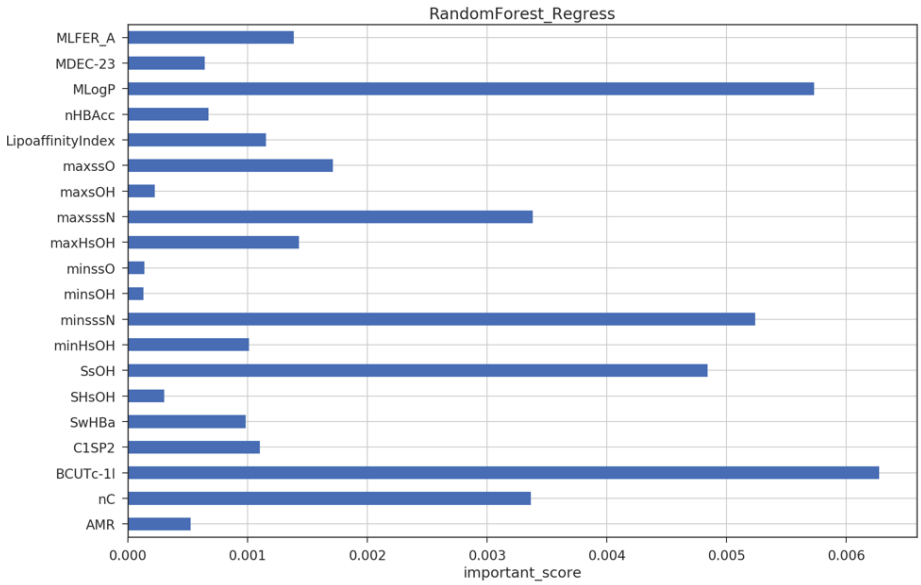


图 4-7 前 20 名显著影响变量对应的影响度

经过特征选择算法计算筛选出的前 20 名特征在随机森林回归模型中都具有较高的重要得分，同时选择特征的回归模型在测试集上具有相对较高的准确率，因而说明我们特征选择的合理性。

5 生物活性的定量预测模型

5.1 问题二的分析

根据问题二的要求，首先需要挑选出对 ERα 生物活性有显著性影响的主要的分子描述符变量，根据提取出来的主要分子描述符变量构建的化合物对 ERα pIC50 的定量预测模型，模型的预测效果才能表现得更好，将已知数据分为两大类：采用 70%的化合物样本数据作为测试集，30%的化合物样本作为验证集。

首先需要进行筛选变量，此题无法使用常见的某些降维方法（如：奇异值分解），因此通过使用随机森林算法，它可以选择 20 个最重要且具有可解释性的，对化合物 pIC50 影响最显著的分子描述符变量。建立对 ERα pIC50 的预测模型。首先处理化合物样本数据，因为化合物样本对应的不同分子描述符变量之间的数据大小存在一定差异，为后续建模的顺利进行，对化合物样本数据进行标准化根据筛选后的化合物样本数据集进行构建化合物 pIC50 预测模型，选择机器学习算法中的随机森林方法进行预测模型的构建，当化合物样本数据集中出现不均匀的数据时，随机森林方法可以平衡模型误差，而且当提取的分子描述符变量出现非线性相关关系，采用随机森林构建预测模型，可以通过增加决策树，尽可能减弱过拟合。同样也可以采用梯度提升决策树（GBDT）和支持向量机（SVM），但是梯度提升决策树和随机森林方法不同，它不需要太多的树，使用上述的多种算法得到测试集中 50 个化合物样本对应的 pIC50 值，并进行对比验证，问题二的思路流程图如图 5-1 所示。

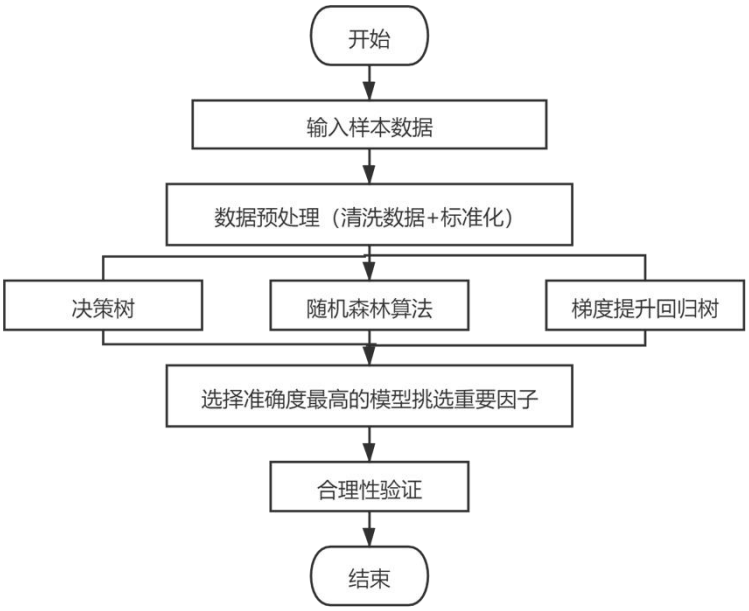


图 5-1 问题二的求解流程图

5.2 定量预测模型的建立

为较好的设计一种化合物对生物活性的定量预测模型，在求解的过程中主要用到了随机森林、梯度上升树和支持向量机这三种方法进行定量预测，其中随机森林在问题一中已经说明，这里就不再赘述。

5.2.1 梯度提升回归树

梯度提升回归树[5]通过合并多个决策树来构建既可用于化合物生物活性值的回归，也可用于化合物 ADMET 性质的分类的预测模型。一般情况下，梯度提升采用连续的方式构

造树，每棵树都试图纠正前一棵树的错误。默认情况下，梯度提升回归树中没有随机化，而是用到了强预剪枝。梯度提升树通常使用深度很小（1 到 5 之间）的树，这样模型占用的内存更少，预测速度也更快。

梯度提升树的特征重要性与随机森林的特征重要性有些类似，由于梯度提升和随机森林两种方法在类似的数据上表现的都很好，因此一种常用的方法就是先尝试随机森林，它的鲁棒性很好。如果随机森林效果很好，但预测时间太长，或者机器学习模型精度小数点后第二位的提高也很重要，那么切换成梯度提升通常会有用。

下面给出梯度提升回归树的算法：

输入是训练集样本$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，最大迭代次数为$T$，损失函数为$L$；

输出是强学习器$f(x)$，获得的过程分为以下几步：

1）初始化弱学习器

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

2）对迭代轮数$t=1,2,\cdots,T$，有

Step1：对样本$i=1,2,\cdots,m$，计算负梯度

$$r_{ti}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)} \tag{5-2}$$

Step2：利用$(x_i,r_{ti}),i=1,2,\cdots,m$，拟合一颗决策回归树，得到第 t 棵回归树，其对应的叶子结点区域为$R_{tj},j=1,2,\cdots,J$。其中 J 为回归树 t 的叶子节点的个数。

Step3：对叶子区域$j=1,2,\cdots,J$，计算最佳拟合值

$$c_{tj}=\arg\min_c\sum_{x_i\in R_{tj}}L(y_i,f_{t-1}(x_i)+c) \tag{5-3}$$

Step4：更新强学习器：

$$f_t(x)=f_{t-1}(x)+\sum_{j=1}^{J}c_{tj}I,x\in R_{tj} \tag{5-4}$$

3）得到强学习器$f(x)$的表达式

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

5.2.2 基于支持向量机的定量（回归）预测模型

支持向量机是有监督回归问题中的一种常用机器学习方法，其在拟合数据样本对未知数据进行相关性质预测方面有着普遍的应用[6]。其核心思想是寻找一个满足分类要求的最优分类超平面，使得该平面在保证分类精度，同时使超平面两侧的空白区域最大化。样本中距离超平面最近的一些点组成的向量叫做支持向量，判断分类效果好坏的依据是不同样本都能归属到合适的类别中，同时各类样本点到超平面的距离最远。

间隔超平面所在位置点可以用下面的方程来描述

$$w^T x+b=0 \tag{5-6}$$

在二维空间内点(x,y)到直线$Ax+By+C=0$的距离d_2的计算公式如下：

$$d_2=\frac{|Ax+By+C|}{\sqrt{A^2+B^2}} \tag{5-7}$$

将其思路扩展到 n 维空间中，点$(x_1,x_2,\cdots,x_n)$到直线$w^T x+b=0$的距离d_n的计算公式为

$$d_n=\frac{|w^T x+b|}{\|w\|} \tag{5-8}$$

w,b是编程求解时设计的训练参数，形成了两条间隔线$w^T x+b=1$和$w^T x+b=-1$，正确的分类结果应该是所有样本点正确分布在处于自己的那一侧，并且不在两条间隔线之间出现。对于样本中任意点x_i有以下约束：

$$\begin{cases} w^T x_i+b\geq 1, y_i=1 \\ w^T x_i+b\leq -1, y_i=-1 \end{cases} \tag{5-9}$$

已知点x_i到超平面(w,b)的距离为d，只有处于间隔线上的点对间隔线有影响，比如下图中的一个正点和两个负点，对于间隔线上的点来说：$|w^T x+b|=1$，次数距离可简化为

$$d_3=\frac{1}{\|w\|} \tag{5-10}$$

支持向量机算法的优化目标是在约束条件下求得d_3最大，经过一系列线性和对偶变换，最终转化成求一下规划问题：

$$\begin{aligned} &\min_{w,b} \quad \frac{1}{2}\|w\|^2 \\ &s.t. \quad y_i\left(w^T x_i+b\right)\geq 1, \ \ i=1,2,,m \end{aligned} \tag{5-11}$$

如图所示，当数据完全线性不可分时，即无法找到一条间隔线或间隔超平面，此时需要蒋书记投影到高维线性可分空间，在新映射空间内寻找超平面，通过间隔最大化的方式，

学习得到支持向量机，此时的分隔超平面表示为 $f(x)=w\phi(x)+b$，优化问题经一系列线性和对偶转换，最终化为以下规划问题：

$$\begin{aligned} \min_{\lambda} \quad & \left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j\left(\phi(x_i)\phi(x_j)\right)-\sum_{j=1}^{n}\lambda_i\right] \\ s.t. \quad & \sum_{i=1}^{n}\lambda_i y_i=0, \lambda_i\geq 0, C-\lambda_i-\mu_i=0 \end{aligned} \tag{5-12}$$

5.3 模型求解与分析

5.3.1 数据整理

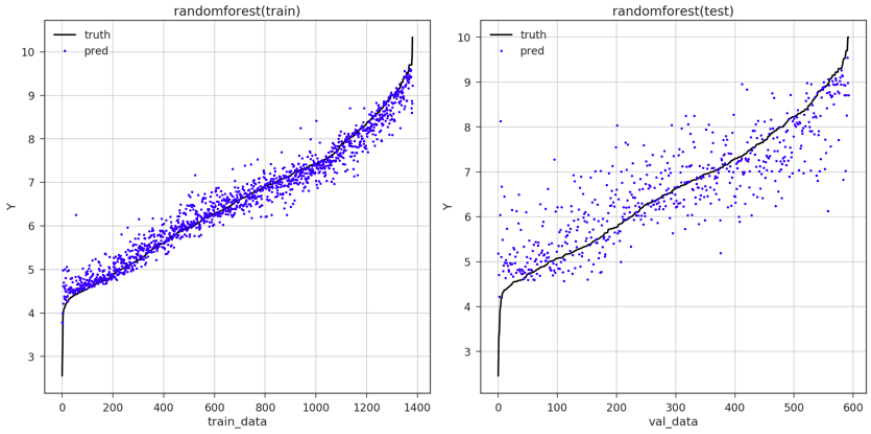
此步需对“Molecular_Descriptor.xlsx”和“ERα_activity.xlsx”文件中的 1974 组数据进行整理，对数据清洗得到 504 维的数据，保留前文提取出的 20 项主要变量及 pIC50 这一变量，用 Python 对这些数据进行标准化处理，将每一行作为一组输入训练集，得到数据集并将数据集中的 1974 组数据按 70%和 30%比例分为训练集数据（70%）和验证集数集（30%），以防止模型的过拟合并利用在验证集上的表现建模模型的效果，方便对模型进行选择。

5.3.2 参数设置

使用 Python 应用随机森林回归算法，综合考虑到算法速度和算法准确率，设定 K=500，M=150；梯度上升树回归算法的最大迭代次数为 7；支持向量机回归算法使用高斯核技巧，且设置正则化参数为 5 以防止模型过拟合。

5.3.3 模型验证集结果对比

如下图 5-2 所示，分别展示随机森林、梯度上升树、支持向量机在训练集和验证集上的真实值和预测值之间的分布差异。



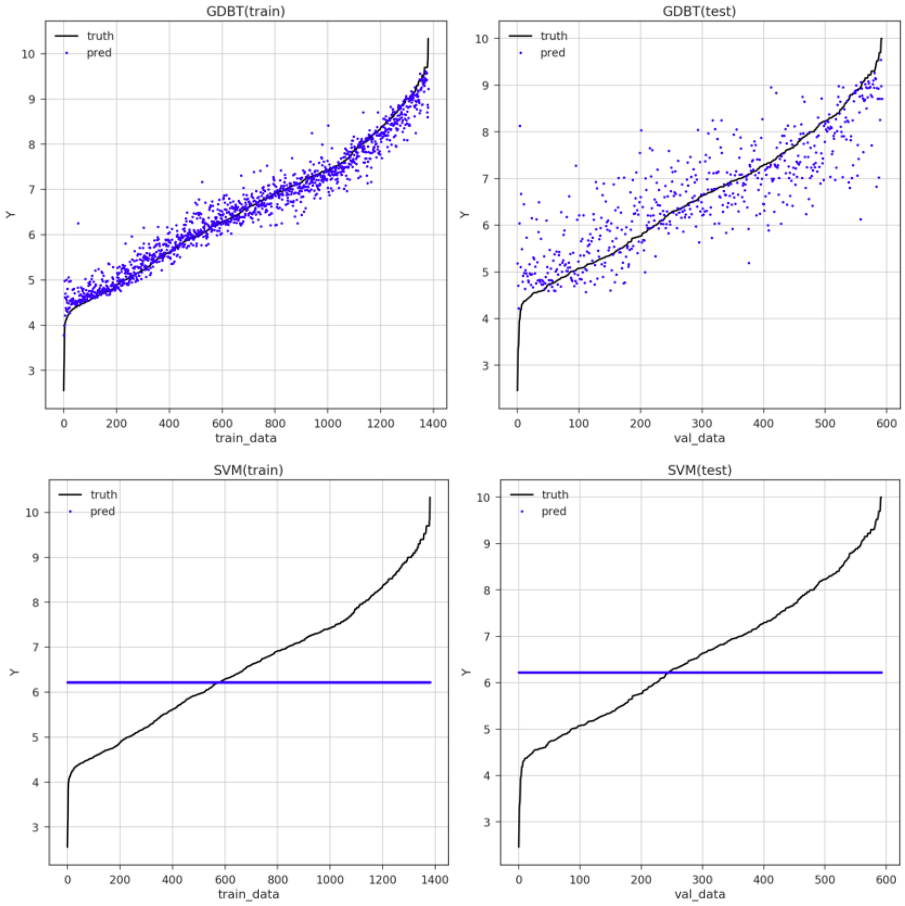


图 5-2 三种模型在训练集和验证集上的表现

如图所示，在 SVM 模型上，虽然其准确度和随机森林、梯度上升树相比，相差不大，仅仅只有 3-4%的区分，但是明显 SVM 模型在预测时几乎使用了平均值作为预测结果，显然是不正确的，而相对而言，随机森林和梯度上升树模型明显表现的更好。

表 5-1 三种模型在验证集上的准确度对比

模型	随机森林	梯度上升树	支持向量机
准确度	72.8%	73.0%	69.2%

如上表所示，对比三个模型在验证集上的结果，梯度上升树表现最好，在验证集上达到了 73.0%的效果，因此后续在模型测试集上进行回归预测时，我们选择梯度上升树进行回归预测。

5.3.4 模型测试集预估结果

将“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，预测过程使用梯度上升树对 pIC_{50} 值进行预测值，进而计算得到对应的 IC_{50} 值。下表将按化合物在表格中的顺序展示部分的预测结果，全部预测结果见附件。

SMILES	IC50_nM	pIC50
COc1cc(OC)cc(\C=C\c2ccc(OS(=O)(=O)[C@@H]3C[C@@H]4O[C@H]3C(=	20.97413	7.678316
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3ccccc23)c4ccc(O)cc4	26.30456	7.579969
COc1ccc2C(=C(CCOc2c1)c3ccc(O)cc3)c4ccc(\C=C\C(=O)O)cc4	26.52107	7.576409
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	18.16712	7.740714
OC(=O)\C=C\c1ccc(cc1)C2=C(CCSc3cc(F)ccc23)c4ccc(O)cc4	17.82525	7.748964
CC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	13.18985	7.87976
Oc1ccc(cc1)C2=C(c3ccc(\C=C\c4ccccc4)cc3)c5ccc(F)cc5OCC2	17.1119	7.766702
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)c4ccccc4)cc3)c5ccc(F)cc5OCC2	15.35679	7.8137
OC(=O)\C=C\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	99.77058	7.000998
CCN(CC)C(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	24.69151	7.607452
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)N4CCCC4)cc3)c5ccc(F)cc5OCC2	22.94187	7.639371
CCN(CC)CCNC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	31.09462	7.507315
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)N4CCNCC4)cc3)c5ccc(F)cc5OCC2	24.70445	7.607225
CN1CCN(CC1)C(=O)\C=C\c2ccc(cc2)C3=C(CCOc4cc(F)ccc34)c5ccc(O)cc5	24.54544	7.610029
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)N4CCN(Cc5ccccc5)CC4)cc3)c6ccc(F)c	18.57358	7.731104
Cc1ccc(cc1)N2CCN(CC2)C(=O)\C=C\c3ccc(cc3)C4=C(CCOc5cc(F)ccc45)c6	18.44425	7.734139
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)Nc4ccccc4)cc3)c5ccc(F)cc5OCC2	27.2063	7.56533
OC(=O)COc1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	117.4835	6.930023
Oc1ccc(cc1)C2=C(c3ccc(C=O)cc3)c4ccc(F)cc4OCC2	103.0988	6.986747
CCC(=C(c1ccc(O)cc1)c2ccc(\C=C\C(=O)O)cc2)c3ccccc3	2.211789	8.655256
OC(=O)CCCOc1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	131.6424	6.880604
COc1ccc2C(=C(CCOc2c1)c3ccc(O)cc3)c4ccc(OCC(=O)O)cc4	28.34067	7.54759
CCCC(CCC)(c1ccc(O)c(C)c1)c2cccs2CCCC(CCC)(c1ccc(O)c(C)c1)c2cccs2	393.8784	6.404638
CCCC(CCC)(c1ccc(O)c(C)c1)c2cc(C)cs2	295.5238	6.529408
CCCC(CCC)(c1ccc(O)c(C)c1)c2ccc([nH]2)C(=O)OCC	497.2259	6.303446
CCCC(CCC)(c1ccc(O)c(C)c1)c2ccc(C(=O)OCC)n2C	499.291	6.301646
CCCC(CCC)(c1ccc(O)c(C)c1)c2c[nH]c3cc(OCc4ccccc4)ccc23	22.9873	7.638512
Oc1ccc(cc1)c2nc(Cl)c(c(Oc3ccc(OCCN4CCCC4)cc3)n2)c5ccccc5	67.02247	7.17378
CN(C)CCOc1ccc(Nc2nc(nc(Cl)c2c3ccccc3)c4ccc(O)cc4)cc1	45.94106	7.337799

图 5-4 最优模型在测试集上的预测结果（展示部分）

6 问题三

6.1 问题三的分析

根据问题三的要求，本题需要建立化合物的 729 个分子描述符和 ADMET 数据中的五个特征（分别为：Caco-2、CYP3A4、hERG、HOB、MN）潜在的对应关系，共需要构建 5 个分类预测模型。

首先对数据进行归一化处理方式，因使每个分子描述符变量对化合物样本的生物活性的影响因子权重值一致的。如果直接使用原始化合物样本数据来构建分类预测模型，模型的训练效果会表现很差，这使得分别构建化合物的 5 个 ADMET 性质的分类预测模型较为困难。所以采用 PCA（主成分分析）方法处理提取关键的维度信息，将 729 维的信息压缩到低维空间中，相关性高的分子描述符变量会被合并起来，无关信息会被剔除。通过把方差贡献度作为提取主要成分个数的标准，通常选择累计方差贡献率大于等于 90%的主成分，然后分别建立化合物的 5 个 ADMET 性质的分类预测模型。对化合物数据进行构建分类预测模型，常见的分类预测模型有 Logistic（逻辑）回归模型、决策树、随机森林、梯度提升决策树、支持向量机、SVM 和 BP 神经网络。在本题的构建分类模型的过程中，考虑将上述几种常见分类预测模型都用来训练，通过化合物验证集的预测准确率来筛选分类模型，最后选定预测准确率最高的分类模型用于测试集预测。问题三的思路流程图如图 6-1 所示。

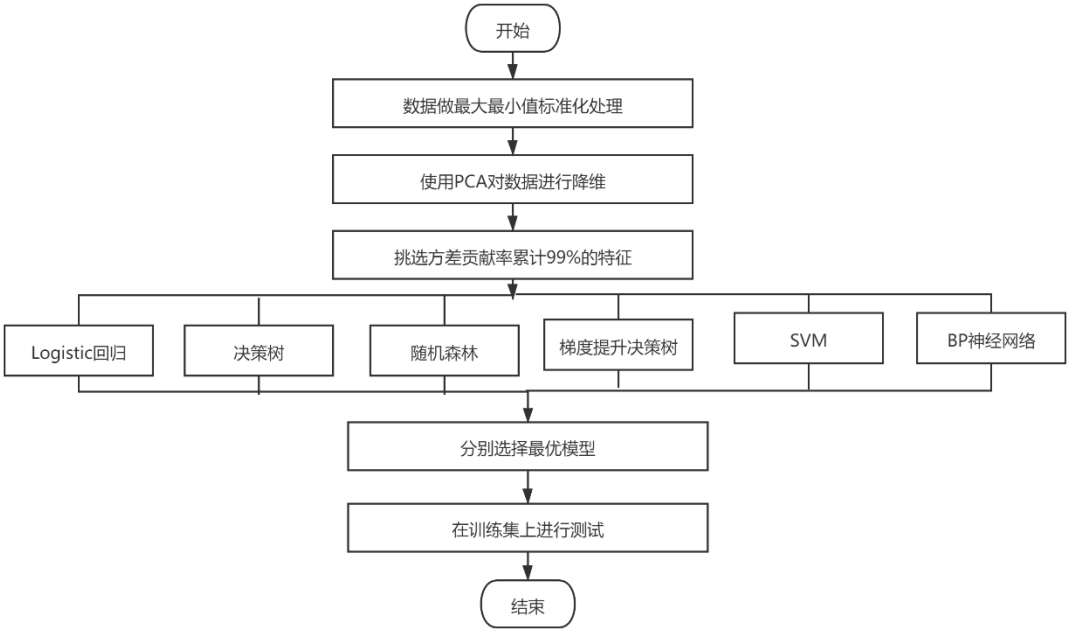


图 6-1 问题三的流程图

6.2 分类预测模型的建立

6.2.1 主成分分析进行降维

经过以上分析，需要对样本数据做降维处理，本文中选择主成分分析法[7]进行降维，筛选掉数据集中特征一样或者特征极其相似的数据。

主成分分析法主要有几以下个步骤[8]：

（1）将附件一中的数据集写成矩阵形式，

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{bmatrix} \tag{6-1}$$

其中 a_{ij} 表示第 i 个样本的第 j 个观测指标，对该矩阵做标准化处理，得到矩阵 $X_{n\times p}$：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{6-2}$$

（2）求数据集的相关系数矩阵 R，$R=\left(r_{ij}\right)_{p\times p}$，$r_{ij}$ 的计算公式如下：

$$r_{ij}=\frac{\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_i\right)\left(x_{kj}-\bar{x}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_i\right)^2\sum_{k=1}^{n}\left(x_{kj}-\bar{x}_j\right)^2}} \tag{6-3}$$

其中，$r_{ij}=r_{ji},r_{ii}=1$，

（3）求矩阵 R 的特征方程 $\det\left(R-\lambda E\right)=0$ 的特征根；并将特征根由大到小排序，记作：$\lambda_1\geq\lambda_2\geq\lambda_3\geq\cdots\geq\lambda_p>0$。

（4）确定主成分个数 m，m 满足如下关系式：

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\geq\alpha \tag{6-4}$$

其中，α 根据实际问题确定。

（5）计算 m 个相应的单位特征向量：

$$e_1=\begin{bmatrix}e_{11}\\e_{21}\\\vdots\\e_{p1}\end{bmatrix},e_2=\begin{bmatrix}e_{12}\\e_{22}\\\vdots\\e_{p2}\end{bmatrix},\cdots,e_m=\begin{bmatrix}e_{1m}\\e_{2m}\\\vdots\\e_{pm}\end{bmatrix} \tag{6-5}$$

（6）计算贡献率以及写出主成分：

$$\text{贡献率}=\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k},i=1,2,\cdots,p \tag{6-6}$$

$$\text{累计贡献率}=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k},i=1,2,\cdots,p \tag{6-7}$$

第 i 个主成分：

$$F_i=e_{1i}x_1+e_{2i}x_2+\cdots+e_{pi}x_p \tag{6-8}$$

6.2.2 搭建各类分类预测模型

本题需要利用附近提供的分子描述符和化合物的 ADMET 数据构建化合物的分类预测模型，本次研究大致思路是建立多种学习模型，如 logistic 分类、决策树分类、随机森林分类模型、梯度提升树分类以及 BP 神经网络学习模型通过编写程序，训练出一个最优模型，用于完成对化合物的分类预测。

（1）logistic 分类

logistic 分类[9]是一种用于解决二分类问题的机器学习方法，用于估计某种特征的可能性。logistic 分类能将数据分成 0 和 1 两类。logistic 分类过程比较简单，但很经典，主要有线性求和，sigmoid 函数激活，计算误差，修正参数四个步骤，前两步用于判断，后两步用于参数修正。

1）判断过程

设定一个 N 维的输入列向量 x，n 维参数列向量 h，以及偏置量 b，通过线性求和得到变量 z：

$$z = h^T x + b \tag{6-9}$$

由于 z 的值域是 $[-\infty, +\infty]$，故需要通过建立一种 sigmoid 函数，将其映射到 $[0,1]$ 上。sigmoid 函数的形式如下：

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{6-10}$$

该函数具有性质：

$$\sigma'(x) = \sigma(x)\left(1-\sigma(x)\right) \tag{6-11}$$

故而 sigmoid 函数的图像如图所示

由图可知，当 x 的值越大，$\sigma(x)$ 的值就越接近于 1。将前面的变量 z 代入到 sigmoid 函数，可以得到

$$a = \sigma(z) = \sigma\left(h^T x + b\right) \tag{6-12}$$

由此，当 a 大于 0.5 时，可将 x 归属为 1 类，当 a 小于 0.5 时，可将 x 归属为 0 类。

2）修正过程

在判断过程中，用到了参数向量和偏置量。初始时刻，参数向量 h 的值是随机的，偏置量 b 为 0。为此，通过训练来使得 h,b 能够尽可能的达到较优的值。

设输入变量 x 的期望判定是 y，而实际得到的判定值是 a。为此建立损失函数 $C(a,y)$，通过修正 h,b，使得 C 最小化。由凸优化理论可以得到每次迭代的参数更新公式为：

$$\begin{cases} h = h - \eta a(a-y)(1-a)x \\ b = b - \eta a(a-y)(1-a) \end{cases} \tag{6-13}$$

（2）在该问题中应用随机森林和梯度提升树做分类预测，和前文所做的回归分析方法基本一致，但也有以下不同之处：

对于随机森林，在处理分类问题时，对于测试样本，森林中每棵决策树会给出最终类别，最后综合考虑森林内每一棵决策树的输出类别，以投票表决的方式来决定测试样本的类别，而在处理回归问题时，则以每棵决策树输出的均值为最终结果。

（3）基于 BP 神经网络的预测模型

BP 神经网络[10]是一种按照误差逆向传播算法训练的多层前馈神经网络，有着极其广

泛的应用，其基本原理为：输入数据 Xi 通过中间节点（隐层点）作用于输出节点，经过非线形变换，产生输出数据 Yk，网络训练的每个样本包括输入向量 X 和期望输出量 t，网络输出值 Y 与期望输出值 t 之间的偏差，通过调整输入节点与隐层节点的联接强度取值 Wij 和隐层节点与输出节点之间的联接强度 Tjk 以及阈值，使误差沿梯度方向下降，经过反复学习训练，确定与最小误差相对应的网络参数（权值和阈值），训练即告停止。此时经过训练的神经网络即能对类似样本的输入信息，自行处理输出误差最小的经过非线形转换的信息。

X 为 n 维列向量，网络层由权值矩阵 W 和阈值矢量 b 组成，S 个神经元的输出组成了 S 维的神经网络输出矢量 y：

$$y = f\left(wx + b\right) \tag{6-14}$$

其中，输入层网络权值矩阵 W 和阈值矢量 b 的形式如下：

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{s1} & w_{s2} & \cdots & w_{sn} \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_s \end{bmatrix} \tag{6-15}$$

BP 算法

BP 算法由正向传播和逆向传播两部分组成。在正向传播过程中输入数据从输入层进入，经过隐层处理后传至输出层。如果输出层得不到期望的输出就转为逆向传播，即吧误差数据沿连接路径返回并通过修改各层神经元之间的连接权值使误差数据最小。

设 bp 神经网络有 M 层，第一层是输出节点，第 M 层仅喊输出节点。假设有 N 个样本对 $\left\{x_1(p), x_2(p), \cdots, x_n(p), p = 1, 2, \cdots, N\right\}$，第 M 层的输出为 $H(p)$。选择 Sigmoid 函数作为隐层节点和输出节点的作用函数，其数学表达式如下：

$$f(x) = \frac{1}{1 + e^{-\alpha x}} \tag{6-16}$$

其中，α 表示神经元非线性的参数。

对某一输入 $x(p)$，节点 i 的输出为 x_i，节点 j 的输入为 $t_j = \sum_i w_{ji} x_i$，以及输出为 $y_i = f\left(t_j\right)$。现定义误差函数为：

$$e = \frac{1}{2N} \sum_{p=1}^{N} \left(H_p - T_p\right)^2 \tag{6-17}$$

其中，T_p 为神经网络的目标输出。

再定义 $e = \frac{1}{2}\left(H(p) - T(p)\right)^2$，$\varsigma_j = \frac{\partial e}{\partial t_j}$，并且有 $y_i = f\left(t_i\right)$，于是有：

$$\frac{\partial e}{\partial W_{ji}}=\frac{\partial e}{\partial t_j}\frac{\partial t_j}{\partial W_{ji}}=\frac{\partial e}{\partial t_j}x_i=\varsigma_j x_i \tag{6-18}$$

然后分两种情况讨论：

（i）当 j 为输出节点时， $y_i=H$ ，有

$$\varsigma_j=\frac{\partial e}{\partial t_j}=\frac{\partial e}{\partial H}\frac{\partial H}{\partial t_j}=-(T-H)\,f'(t_j) \tag{6-19}$$

（ii）当 j 为隐层节点时，有

$$\varsigma_j=\frac{\partial e}{\partial t_j}=\frac{\partial e}{\partial y_j}\frac{\partial y_j}{\partial t_j}=\frac{\partial e}{\partial y_j}\,f'(t_j) \tag{6-20}$$

其中， y_j 为传送到下一层 $(k+1)$ 的输入，计算 $\frac{\partial e}{\partial y_i}$ 要从 $(k+1)$ 层送回。

在 $(k+1)$ 层第 m 个节点有：

$$\begin{aligned}\frac{\partial e}{\partial y_j}&=\sum_m\frac{\partial e}{\partial t_j}\frac{\partial t_j}{\partial y_j}\\&=\sum_m\frac{\partial e}{\partial t_m}\cdot\frac{\partial}{\partial y_j}\sum_j W_{jm}H_j\\&=\sum_m\frac{\partial e}{\partial t_m}\cdot\sum_j W_{jm}\\&=\sum_m\varsigma_m W_{jm}\end{aligned} \tag{6-21}$$

由上式可以得到

$$\begin{cases}\varsigma_j=f'(t_j)\sum_m\varsigma_m W_{jm}\\\dfrac{\partial e}{\partial W_{jm}}=\varsigma_j y_j\end{cases} \tag{6-22}$$

于是有：

$$\begin{aligned}f'(t_j)&=\frac{\mathrm{d}e}{\mathrm{d}t_j}\left(\frac{1}{1+e^{-\alpha t_j}}\right)\\&=\alpha\left(1+e^{-\alpha t_j}\right)^2 e^{-\alpha t_j}\\&=\alpha f(t_j)\left(1+f(t_j)\right)\\&=\alpha y_j(1+y_j)\end{aligned} \tag{6-23}$$

根据上面的推导，可以将 BP 神经网络算法总结如下：

Step1：选取初始权值、阈值；
Step2：重复下面两个过程，直到满足实践要求为止：

Move1：对于学习样本 p 从 1 到 N，包括：

（1）正向过程：计算每层各节点 j 的 y_j, t_j 和 H 的值

（2）逆向过程：对各层，第 M 层到第二层，对每个节点，逆向计算 ς_j；

Move2：修正权值：

$$W_{ji}(t+1)=W_{ji}(t)-\eta\frac{\partial e}{\partial W_{ji}},\eta>0 \tag{6-24}$$

（1）设计网络结构层

输入输出层：将提取的 37 个主要变量作为输入，生物活性和性质作为输出。故输入层神经元个数 n=37，输出层 m=6。因此，本文设计的网络结构图如下图所示：

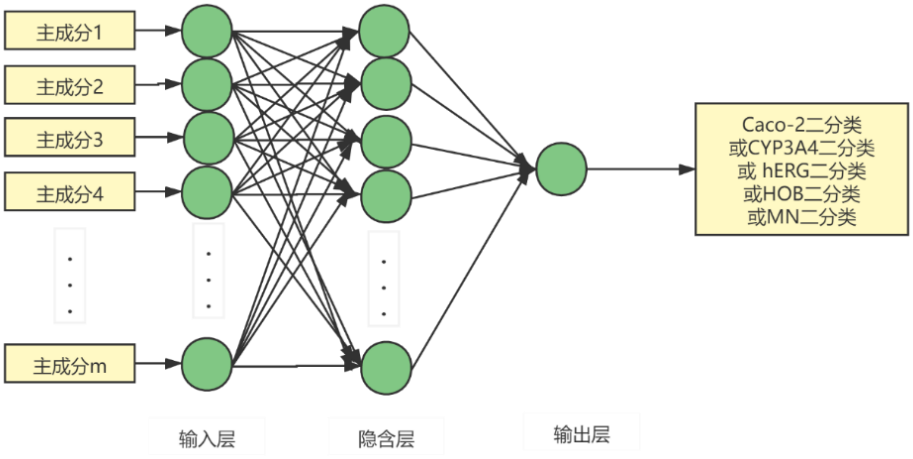


图 6-2 问题求解的 BP 神经网络结果示意图

（2）激励函数的选取

为了保证神经网络可以无限制的逼近任意非线性函数，模拟神经元的状态变化，就必须选择合适的激励函数。在编写程序求解时，选择的是 ReLU 激励函数，其形式如下

$$f(x)=\begin{cases}0 & x\le 0\\ x & x>0\end{cases} \tag{6-25}$$

6.3 模型求解与分析

6.3.1 数据整理

此步需对“Molecular_Descriptor.xlsx”和“ADMET.xlsx”文件中的 1974 组数据进行整理，用 Python 对清洗过的 504 维的数据进行标准化处理，使用 PCA 将原来 504 维变量降维降至 37 维（根据方差累计贡献率而得），保留 Caco-2、CYP3A4、hERG、HOB、MN

作为目标分类变量，将每一行作为一组输入训练集，得到数据集并将数据集中的 1974 组数据按 70%和 30%比例分为训练集数据（70%）和验证集数集（30%），以防止模型的过拟合并利用在验证集上的表现建模模型的效果，方便对模型进行选择。

6.3.2 模型验证集结果对比

（1）PCA 降维

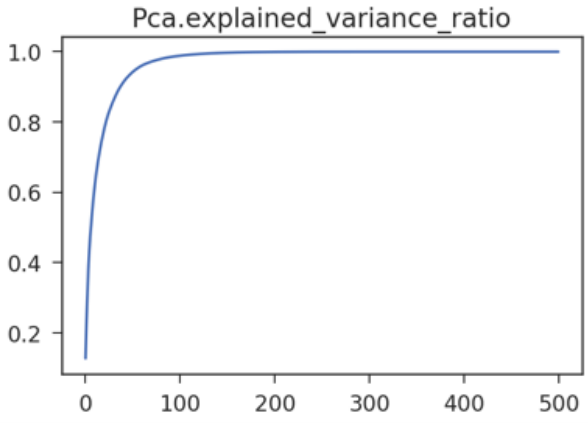


图 6-3 PCA 方差累计贡献图

使用 Python 应用 PCA 模型对数据进行降维时，设定方差累计贡献率为 90%的标准，经实验确定经 PCA 变换维度空间后，如图所示几乎前 100 个主成分的累计贡献率已经占据了大部分，实际计算显示前 105 个主成分的累计贡献率达到了 90%。因此选用这 37 个主成分代表全部特征。

（2）Caco-2 分类问题

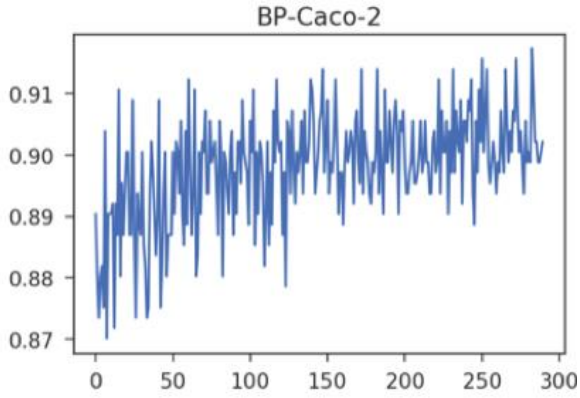


图 6-4 Caco-2 分类 BP 神经网络调参过程

logistic 分类算法正则化参数 C=100000；决策树分类算法设定最大深度为 14；随机森林分类算法，综合考虑到算法速度和算法准确率，设定 K=1000，M=10；梯度上升树分类算法的最大迭代次数为 4；支持向量机分类算法使用高斯核技巧，且设置正则化参数 C 为 14 以防止模型过拟合。BP 神经网络分类算法隐藏层神经元个数为 292，隐藏层参数经过参数调试得到。

如上图所示，隐藏层经元数设置为 292 时，模型在验证集上表现最好。

表 6-1. Caco-2 分类问题中 6 种机器学习模型的求解准确度

模	logis	决	随机	梯度上	支持向	B

型	tic	策树	森林	升树	量机	P
准确度	90.4%	83.8%	89.0%	90.1%	90.4%	91.7%

如上表所示，在 Caco-2 分类问题中 BP 模型表现最好，在验证集上的准确度为 91.7%。

（3）CYP3A4 分类问题

logistic 分类算法正则化参数 C=1；决策树分类算法设定最大深度为 14；随机森林分类算法，综合考虑到算法速度和算法准确率，设定 K=1000，M=10；梯度上升树分类算法的最大迭代次数为 5；支持向量机分类算法使用高斯核技巧，且设置正则化参数 C 为 50 以防止模型过拟合。BP 神经网络分类算法隐藏层神经元个数为 171，隐藏层参数经过参数调试得到。

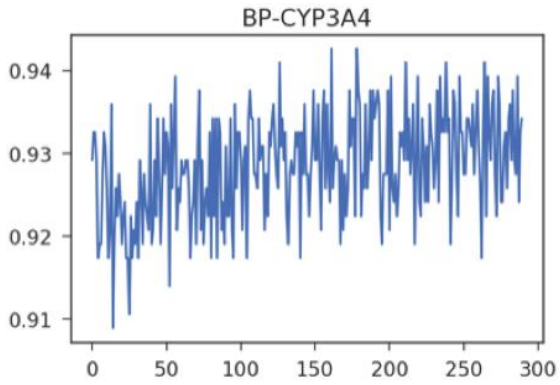


图 6-5 CYP3A4 分类 BP 神经网络调参过程

如上图所示，隐藏层经元数设置为 171 时，模型在验证集上表现最好。

表 6-2. CYP3A4 分类问题中 6 种机器学习模型的求解准确度

模型	logistic	决策树	随机森林	梯度上升树	支持向量机	BP
准确度	94.9%	88.4%	92.7%	93.8%	95.1%	94.2%

如上表所示，在 CYP3A4 分类问题中支持向量机模型表现最好，在验证集上的准确度为 95.1%。

（4）hERG 分类问题

logistic 分类算法正则化参数 C=100000；决策树分类算法设定最大深度为 11；随机森林分类算法，综合考虑到算法速度和算法准确率，设定 K=100，M=10；梯度上升树分类算法的最大迭代次数为 6；支持向量机分类算法使用高斯核技巧，且设置正则化参数 C 为 150 以防止模型过拟合。BP 神经网络分类算法隐藏层神经元个数为 142，隐藏层参数经过参数调试得到。

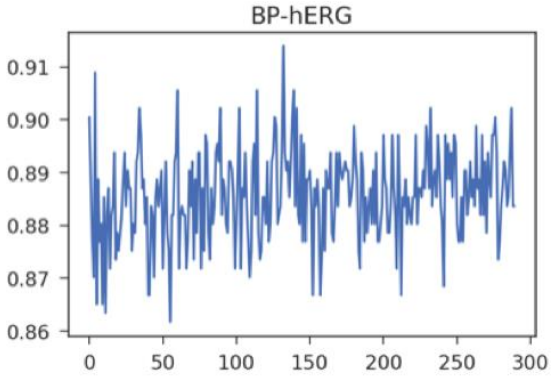


图 6-6 hERG 分类 BP 神经网络调参过程

如上图所示，隐藏层经元数设置为 142 时，模型在验证集上表现最好。

表 6-3.CYP3A4 分类问题中 6 种机器学习模型的求解准确度

模型	logistic	决策树	随机森林	梯度上升树	支持向量机	BP
准确度	89.9%	83.8	88.5	88.0	88.7	91.3%

如上表所示，在 hERG 分类问题中 BP 模型表现最好，在验证集上的准确度为 92.2%。

（5）HOB 分类问题

logistic 分类算法正则化参数 C=100；决策树分类算法设定最大深度为 10；随机森林分类算法，综合考虑到算法速度和算法准确率，设定 K=1000，M=10；梯度上升树分类算法的最大迭代次数为 7；支持向量机分类算法使用高斯核技巧，且设置正则化参数 C 为 250 以防止模型过拟合。BP 神经网络分类算法中隐藏层神经元为 108，隐藏层参数经过参数调试得到。

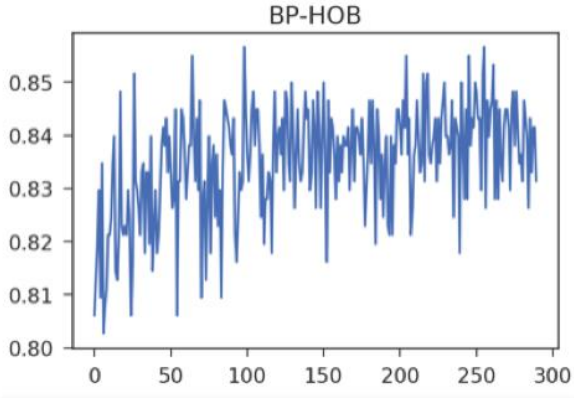


图 6-7 HOB 分类 BP 神经网络调参过程

如上图所示，隐藏层经元数设置为 108 时，模型在验证集上表现最好。

表 6-4 CYP3A4 分类问题中 6 种机器学习模型的求解准确度

模型	logistic	决策树	随机森林	梯度上升树	支持向量机	BP
准确度	82.1%	78.6%	83.0%	84.7%	84.5%	85.6%

如上表所示，在 HOB 分类问题中 BP 模型表现最好，在验证集上的准确度为 85.6%。

（6）MN 分类问题

logistic 分类算法正则化参数 C=1；决策树分类算法设定最大深度为 10；随机森林分类算法，综合考虑到算法速度和算法准确率，设定 K=1000，M=10；梯度上升树分类算法的最大迭代次数为 6；支持向量机分类算法使用高斯核技巧，且设置正则化参数 C 为 90 以防止模型过拟合。BP 神经网络分类算法的隐藏层设置 216 个神经元，隐藏层参数经过参数调试得到。

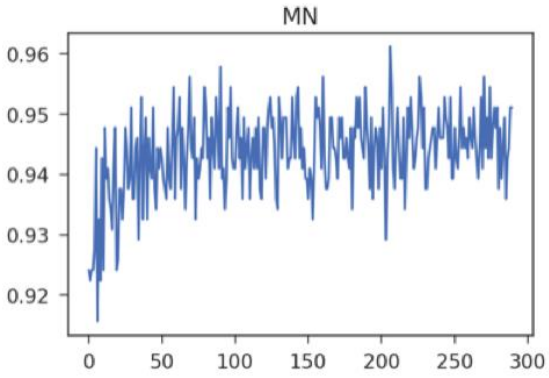


图 6-8 MN 分类 BP 神经网络调参过程

如上图所示，隐藏层神经元数设置为 216 时，模型在验证集上表现最好。

表 6-5. CYP3A4 分类问题中 6 种机器学习模型的求解准确度

模型	logistic	决策树	随机森林	梯度上升树	支持向量机	BP
准确度	92.1%	90.4%	92.7%	95.4%	95.1%	96.1%

如上表所示，在 MN 分类问题中 BP 模型表现最好，在验证集上的准确度为 96.1%。

同时，综合上述五个分类问题可以看出，BP 神经网络在各个问题上整体表现最好，只有在 CYP3A4 分类问题上表现略弱于支持向量机模型。

6.3.3 模型测试集预估结果

将“ADMET.xlsx”的 test 表中的 50 个化合物进行 Caco-2、hERG、CYP3A4、HOB、MN 分类预测，预测过程使用训练好的 BP 神经网络模型进行预测， CYP3A4 的分类使用训练好的支持向量机模型进行预测。下表将按化合物在表格中的顺序展示部分 Caco-2、hERG、CYP3A4、HOB、MN 的预测结果，全部预测结果见附件。

A	B	C	D	E	F
SMILES	Caco-2	CYP3A4	hERG	HOB	MN
COc1cc(OC)cc(\C=C\c2ccc(OS(=O)(=O)[C@@H]3C[C@@H]4O[C@H]3C(=	0	1	1	0	1
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3ccccc23)c4ccc(O)cc4	0	1	1	0	1
COc1ccc2C(=C(CCOc2c1)c3ccc(O)cc3)c4ccc(\C=C\C(=O)O)cc4	0	1	1	0	1
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	0	1	1	0	1
OC(=O)\C=C\c1ccc(cc1)C2=C(CCSc3cc(F)ccc23)c4ccc(O)cc4	0	1	1	0	1
CC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	0	1	1	0	1
Oc1ccc(cc1)C2=C(c3ccc(\C=C\c4ccccc4)cc3)c5ccc(F)cc5OCC2	0	1	1	0	1
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)c4ccccc4)cc3)c5ccc(F)cc5OCC2	0	1	1	0	1
OC(=O)\C=C\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	0	0	0	0	1
CCN(CC)C(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	0	1	1	0	1
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)N4CCCC4)cc3)c5ccc(F)cc5OCC2	0	1	1	0	1
CCN(CC)CCNC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	0	0	0	0	1
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)N4CCNCC4)cc3)c5ccc(F)cc5OCC2	0	1	1	1	1
CN1CCN(CC1)C(=O)\C=C\c2ccc(cc2)C3=C(CCOc4cc(F)ccc34)c5ccc(O)cc5	0	1	1	0	1
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)N4CCN(Cc5ccccc5)CC4)cc3)c6ccc(F)c	0	1	1	0	1
Cc1ccc(cc1)N2CCN(CC2)C(=O)\C=C\c3ccc(cc3)C4=C(CCOc5cc(F)ccc45)c6	0	1	1	0	1
Oc1ccc(cc1)C2=C(c3ccc(\C=C\C(=O)Nc4ccccc4)cc3)c5ccc(F)cc5OCC2	0	1	1	0	1
OC(=O)COc1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	0	1	1	1	1
Oc1ccc(cc1)C2=C(c3ccc(C=O)cc3)c4ccc(F)cc4OCC2	1	1	1	0	1
CCC(=C(c1ccc(O)cc1)c2ccc(\C=C\C(=O)O)cc2)c3ccccc3	1	1	1	1	1
OC(=O)CCCOc1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	0	1	1	0	1
COc1ccc2C(=C(CCOc2c1)c3ccc(O)cc3)c4ccc(OCC(=O)O)cc4	0	1	1	0	1
CCCC(CCC)(c1ccc(O)c(C)c1)c2cccs2CCCC(CCC)(c1ccc(O)c(C)c1)c2cccs2	1	0	0	0	0
CCCC(CCC)(c1ccc(O)c(C)c1)c2cc(C)cs2	1	1	1	0	0
CCCC(CCC)(c1ccc(O)c(C)c1)c2ccc([nH]2)C(=O)OCC	1	0	0	0	1
CCCC(CCC)(c1ccc(O)c(C)c1)c2ccc(C(=O)OCC)n2C	1	0	0	0	1
CCCC(CCC)(c1ccc(O)c(C)c1)c2c[nH]c3cc(OCc4ccccc4)ccc23	0	0	0	0	1
Oc1ccc(cc1)c2nc(Cl)c(c(Oc3ccc(OCCN4CCCC4)cc3)n2)c5ccccc5	0	0	0	1	1
CN(C)CCOc1ccc(Nc2nc(nc(Cl)c2c3ccccc3)c4ccc(O)cc4)cc1	0	0	0	1	1

图 6-9 五种性质的预测结果

7 问题四

7.1 问题四的分析

本题需要对化合物的分子描述符变量组合方案进行优化选择，寻找对使化合物对抑制 ERα 具有更好的生物活性的分子描述符变量，又可以使得至少三个 ADMET 性质表现较好，同时确定化合物分子描述符的取值（或取值范围）。

本题的难点在于：需要确定分子描述符变量的取值（或取值范围），同时变量的取值还需要满足至少三个 ADMET 性质表现良好，可能涉及到需要将 ADMET 性质作为回归方程模型的约束条件。

针对难点的分析：已知样本数据集中共有 1974 个原始化合物样本，729 个分子描述符变量，通过观察数据文件发现，化合物生物活性的 pIC_{50} 值等于 7 时，对应的 IC_{50} 值有很大的数值变化，为后续构建的模型优化处理方便，使化合物对抑制 ERα 具有更好的生物活性，选择 pIC_{50} 值大于 7 对应的化合物样本数据进行后面的模型构建，由于化合物分子描述符变量过多，存在冗余的信息，需要对分子描述符变量进行降维，这里采用问题一挑选出的，使化合物对抑制 ERα 具有更好的生物活性的分子描述符变量，排名前 20 的分子描述符变量作为后续建模的自变量数据，通过这一系列的数据筛选处理后，得到的数据集都是已知数据中，能够使得化合物保持较好活性的样本集，紧接着，以五个 ADMET 性质作为因变量，对五个 ADMET 性质分类变量做处理，使得它们在对应性质表现好的时候，取值为 1，例如：Caco-2 代表：化合物的小肠上皮细胞渗透性能力，当渗透性好的时候，设定取值为 $y_1=1$，渗透性不好的时候，设定取值为 $y_1=0$，CYP3A4 代表：化合物能否被 CYP3A4

代谢，当不能被 CYP3A4 代谢时，设定取值为 $y_2=1$，以此类推。

ADMET 性质	性质的具体含义	0 表示含义	1 表示含义	性质表现好
Caco-2	化合物的小肠上皮细胞渗透性能力	渗透性较差	渗透性较好	$y_1=1$
CYP3A4	化合物能否被 CYP3A4 代谢	不能被 CYP3A4 代谢	能够被 CYP3A4 代谢	$y_2=1$
hERG	化合物是否具有心脏毒性	不具有心脏毒性	具有心脏毒性	$y_3=1$
HOB	化合物的口服生物利用度	利用度较差	利用度较好	$y_4=1$
MN	化合物是否具有遗传毒性	不具有遗传毒性	具有遗传毒性	$y_5=1$

采用随机森林回归算法，以挑选出来的 20 个分子描述符作为自变量，经过数据变换处理后的 ADMET 性质（$y_1+y_2+y_3+y_4+y_5$）作为因变量，然后采用粒子群算法对该随机森林回归算法进行求解，在建模求解的过程中，根据目标函数取值情况，选择是否将 $y_1+y_2+y_3+y_4+y_5\geq 3$ 作为约束条件，因为不设置该约束条件，可以使得粒子群算法求解的自变量取值点范围更大，若目标函数值大于等于 3，则符合题意，满足五个 ADMET 性质，至少三个 ADMET 性质表现好的情况，若目标函数值小于 3，则需要添加约束条件，设置 $y_1+y_2+y_3+y_4+y_5\geq 3$ 作为约束条件，使得模型运算结果满足五个 ADMET 性质，至少三个 ADMET 性质表现好。

问题四的思路流程图如图 7-1 所示。

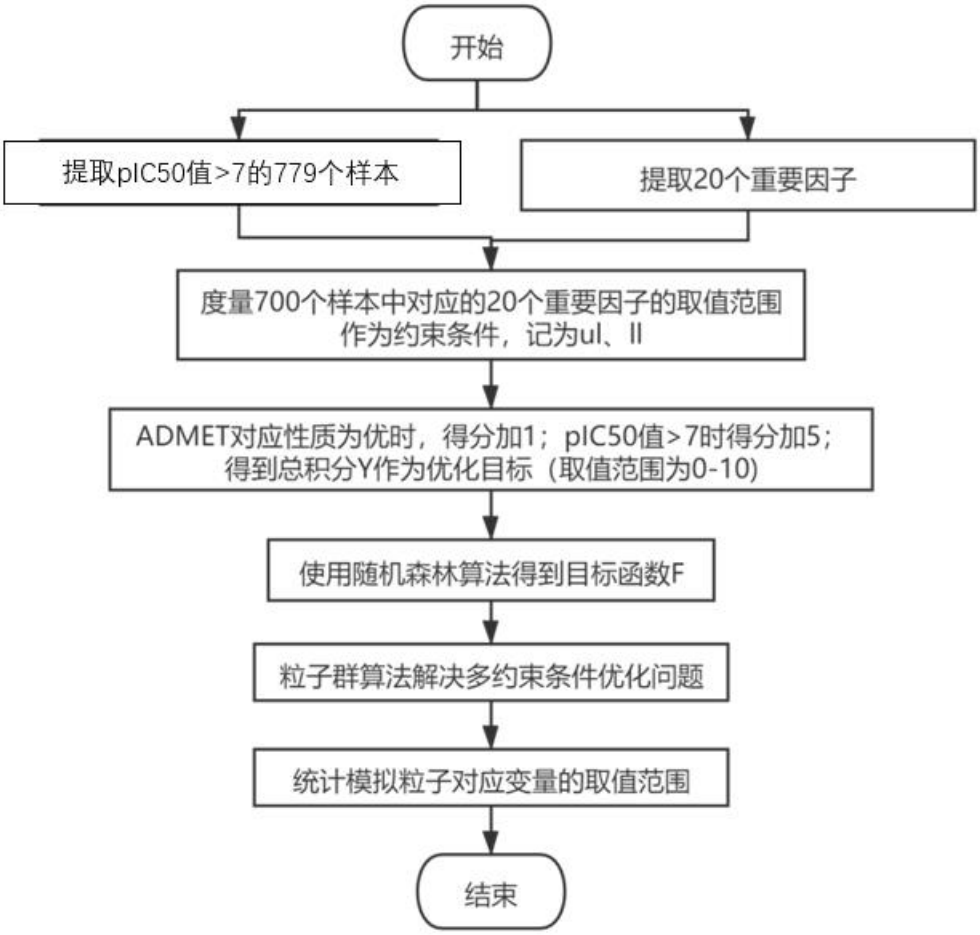


图 7-1. 问题四的流程图

7.2 操作方案建模与优化

7.2.1 建立优化模型

根据问题一的求解，该问题所建立的模型中影响生物活性和较优性质的主要变量一共有 20 个，记为：

$$X=\left(x_1,x_2,\cdots,x_{20}\right) \tag{7-1}$$

（1）目标函数

问题要求在保证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 性质的条件下，找出满足条件的最优的分子描述符，以及其取值范围。对五个 ADMET 性质分别做二分类处理，将各种性质表现较优的记为 1，否则就记为 0，设 Yi 表示五种性质的表现，并对其做求和运算，进一步建立目标函数为此，可以建立如下的目标函数：

$$\max Y=\max\sum_{i=1}^{5}Y_i \tag{7-2}$$

（2）约束条件

对于该优化问题，前面筛选处出来的 20 个分子描述符在实际中具有一定的取值范围，因此存在如下约束：

$$\gamma_i < x_i < \gamma_2, i = 1,2,\cdots,20 \tag{7-3}$$

其中，γ_i 表示某一种分子描述符在实际中可能取得的最小值，γ_2 表示某一种分子描述符在实际中可能取得的最大值。

7.2.2 模型求解算法---粒子群算法

粒子群算法[10]初始化为一群随机粒子，然后通过迭代找到最优解。在每一次迭代中，粒子通过跟踪两个极值来更新自己；第一个就是粒子本身所找到的最优解，这个极值是局部极值；第二个是郑哥种群目前找到的最优解，这个极值是全局极值。

假设在一个 D 维目标搜索空间中，有 N 个粒子组成一个种群，其中第 i 个粒子表示为一个 D 维的向量：

$$X_i = (x_{i1}, x_{i2},, x_{iD}), i = 1,2,\cdots,N \tag{7-4}$$

第 i 个粒子的“飞行”速度也是一个 D 维向量，记作：

$$V_i = (v_{i1}, v_{i2},\cdots, v_{iD}), i = 1,2,\cdots,N \tag{7-5}$$

第 i 个粒子搜索到目前为止的最优位置称为局部极值，记作：

$$P_{best} = (p_{i1}, p_{i2},\cdots, p_{iD}), i = 1,2,\cdots,N \tag{7-6}$$

整个粒子群搜索到目前为止的最优值为全局权值，记作：

$$g_{best} = (p_{g1}, p_{g2},\cdots, p_{gD}), i = 1,2,\cdots,N \tag{7-7}$$

找到这两个最优位置后，粒子将根据下面两个公式更新自己的速度和位置：

$$v_{id} = \omega v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{gd}) \tag{7-8}$$

$$x_{id} = x_{id} + v_{id} \tag{7-9}$$

其中，c_1, c_2 表示学习因子，用以调节学习的最大步长，r_1, r_2 表示两个 $[0,1]$ 上的随机数，ω 表示惯性因子，用以调节解空间的搜索范围。

为此，粒子群算法的流程可总结如下：

Step1：初始化粒子群，包括群体规模 N，每个粒子的位置 x_i 和速度 v_i；

Step2：计算每个粒子的适应度值；

Step3：对每个粒子，用它的适应度值和局部极值比较，如果适应度值大于局部极值，则用适应度值替代局部极值；

Step4：对每个粒子，用它的适应度值和全局极值比较，如果适应度值大于全局极值，则用适应度值替代全局极值；

Step5：根据公式（7-8）和（7-9），更新粒子的速度和位置；

Step6：如果满足结束条件：误差足够好或者达到最大循环次数，则退出，否则返回 Step2。

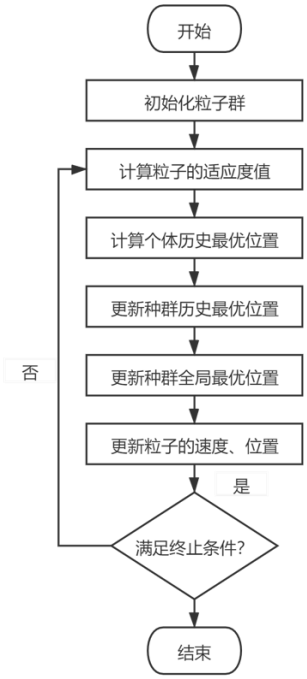


图 7-2 粒子群算法的流程图

7.3 模型求解

7.3.1 实验设置

使用问题一中挑选的 20 个显著影响变量作为目标优化函数中的自变量，同时使用问题二中在有监督回归问题上表现最好的模型——梯度上升树，对 pIC50 值进行回归预测；使用问题三中在有监督二分类问题上分别表现最好的模型——，对 Caco-2、CYP3A4、hERG、HOB、MN 值进行分类预测。

7.3.2 实验结果

表 7-2 20 个显著影响变量的区间范围取值

变量	区间	变量	区间	变量	区间	变量	区间
AMR	71.71～142.23	nC	13.00～28.68	BCUTc-1l	-0.31～-0.24	C1SP2	0.00～10.00
SwHBa	-7.56～43.60	SHsOH	0.00～1.28	SsOH	0.00～26.54	minHsOH	-0.10～0.72

minsssN	0.00～2.20	minsOH	0.00～9.88	minssO	1.54～6.54	maxHsOH	0.17～0.77
maxsssN	0.00～2.54	maxsOH	0.00～7.60	maxssO	0.75～6.72	Lipoaffinity Index	15.72～22.74
nHBAcc	0.00～17.78	MLogP	2.75～5.53	MDEC-23	32.49～54.02	MLFER_A	1.66～3.86

表 7-3 展示了经过粒子群优化算法优化后，对应的 40 个粒子在使得优化目标达到要求时的公共取值范围。如图 7-3 所示，经过 200 轮迭代后，模拟的 40 个粒子在区间范围内使得目标函数到达优化目标。

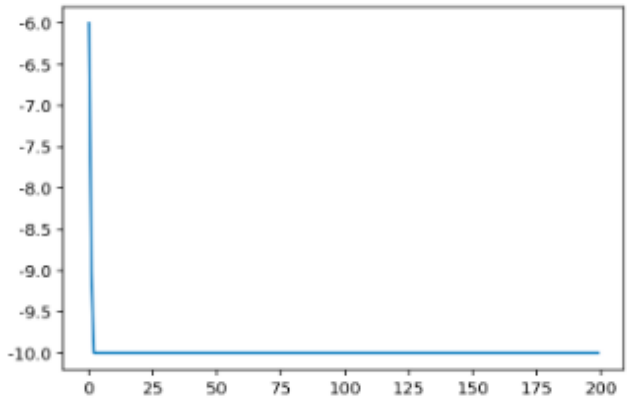


图 7-3 粒子群优化算法的过程

对应 40 个模拟粒子的各个属性预测值展示如下：

[7.0158714] [1] [0] [1] [1] [0]
[7.15856584] [1] [0] [1] [1] [0]
[7.15299126] [1] [0] [1] [1] [0]
[7.13051356] [1] [0] [1] [1] [0]
[7.26055208] [1] [0] [1] [1] [0]
[7.14270671] [1] [0] [1] [1] [0]
[7.34248225] [1] [0] [1] [1] [0]
[7.05351788] [1] [0] [1] [1] [0]
[7.28689033] [1] [0] [1] [1] [0]
[7.01627145] [1] [0] [1] [1] [0]
[7.30056282] [1] [0] [1] [1] [0]
[7.31349986] [1] [0] [1] [1] [0]
[7.11520594] [1] [0] [1] [1] [0]
[7.3266508] [1] [0] [1] [1] [0]
[7.12192305] [1] [0] [1] [1] [0]
[7.32444548] [1] [0] [1] [1] [0]
[7.09514656] [1] [0] [1] [1] [0]
[7.24802276] [1] [0] [1] [1] [0]
[7.27243907] [1] [0] [1] [1] [0]
[7.23331568] [1] [0] [1] [1] [0]
[7.03610514] [1] [0] [1] [1] [0]
[7.31187058] [1] [0] [1] [1] [0]
[7.3190307] [1] [0] [1] [1] [0]
[7.11081744] [1] [0] [1] [1] [0]
[7.16763703] [1] [0] [1] [1] [0]
[7.22787548] [1] [0] [1] [1] [0]
[7.24035167] [1] [0] [1] [1] [0]
[7.23906033] [1] [0] [1] [1] [0]
[7.05249468] [1] [0] [1] [1] [0]
[7.24501502] [1] [0] [1] [1] [0]
[7.26751186] [1] [0] [1] [1] [0]
[7.31699609] [1] [0] [1] [1] [0]
[7.32087645] [1] [0] [1] [1] [0]
[7.24993088] [1] [0] [1] [1] [0]
[7.33310163] [1] [0] [1] [1] [0]
[7.24820247] [1] [0] [1] [1] [0]
[7.24185195] [1] [0] [1] [1] [0]
[7.08386089] [1] [0] [0] [1] [1]
[7.01915807] [0] [1] [1] [0] [1]
[7.33130446] [1] [0] [1] [1] [0]

图 7-4 区间粒子预测结果

如图 7-4 所示，对应每一部分第 1 列为 pIC50 预测值，第 2～第 6 列分别为 Caco-2、

CYP3A4、hERG、HOB、MN 的预测值，其中所有的模拟粒子的 pIC50 值均在 7 以上，Caco-2 为优取 1、CYP3A4 为优取 0、hERG 为优取 0、HOB 为优取 1、MN 为优取 0，模拟粒子在区间范围内均满足至少三个性质为优的条件。

表 7-3 20 个显著影响变量的区间范围缩小率

变量	区间缩小率	变量	区间缩小率	变量	区间缩小率	变量	区间缩小率
minHsOH	6.8%	maxssN	7.0%	maxssO	11.2%	minsOH	15.7%
minssN	19.2%	minssO	25.5%	SwHBa	28.7%	maxHsOH	28.9%
MLogP	35.4%	maxsOH	39.0%	C1SP2	50.0%	SHsOH	53.3%
SsOH	59.5%	MDEC-23	60.2%	BCUTc-1l	66.4%	nHBAcc	73.0%
MLFER_A	74.5%	Lipoafinity Index	74.6%	nC	82.3%	AMR	84.8%

由表 7-3 可知，20 个显著影响 pIC50 的变量中，经过粒子群优化算法后得到的变量取值区间，有一半的区间缩小率在 50%以上，这说明了我们多目标多约束优化问题解决方法的有效性。

8 模型评价和改进

8.1 模型的优点

（1）充分考虑各分子描述符变量之间、各变量与化合物生物活性值之间的潜在相关关系，使用随机森林、梯度提升决策树等方法剔除不重要变量特征，并且可视化模型回归预测结果以验证特征筛选的合理性。

（2）根据化合物样本数据中，分子描述符与生物活性值之间复杂的对应关系，采用随机森林、梯度提升决策树和支持向量机的机器学习算法，同时按照 7:3 的比例分为训练集和测试集，选择预测误差最小的算法用于生物活性值的预测，所选择的模型在验证集上准确度表现较好。

（3）分别建立关于五个 ADMET 性质的二分类预测模型，采用 Logistic 回归、决策树和 BP 神经网络等多种算法，并比较多个模型在验证集上的准确率，根据准确率选择最优的模型进行 ADMET 性质的分类预测，多种方法验证对比保证模型的合理性，且模型在验证集上准确度表现较好。

（4）通过累计评分的方式建立优化目标函数，将多目标多约束优化问题转换为单目标多约束优化问题，方便模型的求解。

8.2 模型的缺点

（1）对问题进行分析时只考虑 729 个分子描述符变量，实际中，可能还有其它变量，会对化合物生物活性和 ADMET 性质产生影响。

（2）问题三中使用 PCA 方法进行降维，提取的主成分可解释性不强，使用随机森林进行预测分类，有可能会出现很多相似的决策树，这可能会降低预测的准确度。

（3）构建的预测模型中，化合物样本数量较少，可以获得更多的化合物样本数据对模型参数进行优化更正。

8.3 模型的改进与推广

（1）本文中提到的学习算法和化合物相关的预测模型可以进行推广，并且应用到生物医学等研究领域中的变量选择和优化问题。

（2）本文中的化合物数据预处理方法、变量的特征筛选和特征降维等方法，都有一定的理论依据作为支撑，在其他的研究领域中涉及数据处理和变量降维等问题，都可以采纳和使用。

（3）本题采用的模型解决的对应问题，可以普遍推广到小样本多特征的问题场景中去。

9 参考文献

[1]李璞玉，冯笑山，降糖药二甲双胍治疗乳腺癌的研究进展[J]，河南科技大学学报(医学版)，2015，000(004):311-314。

[2]袁霞妹，汤婉芬，楼君，卡培他滨治疗晚期复发/转移性乳腺癌 78 例分析[J]， 中国肿瘤，2008，17(6):539-540。

[3]兰丽平，中药单体石蒜碱抗乳腺癌的功能与机制研究，华东师范大学，2017。

[4]赵鹏，陆志，蒋珍华,等，基于随机森林回归分析的脉管制冷机性能预测模型[J]. 红外，42(8):5。

[5]李一蜚,基于梯度提升回归树的中国近地面 O_3 浓度遥感估算[D],中国矿业大学。2020。

[6]张海洋，基于改进支持向量机的混凝土面板堆石坝变形预测模型研究[D]，西安理工大学，2021。

[7]崔雨萌，面向趋势预测的高维数据降维方法研究[D]，北方工业大学，2021。

[8]宦若虹，陶一凡，陈月等，基于多线性主成分分析和张量分析的 SAR 图像目标识别方法 CN106778837A[P]，2017。

[9]高圆圆，危重症产妇行急诊剖宫产终止妊娠危险因素的 logistic 回归分析及母婴结局观察[J]，中国妇幼保健，2021，4800-4803。

[10]王耀东，基于改进粒子群算法的 BP 神经网络优化及应用[D]，西安科技大学。