**抗乳腺癌候选药物的优化建模**

**1、问题重述**

1.1 问题背景

癌症是一种难以治疗和分类的复杂疾病，是一种影响人类健康的主要疾病。全世界癌症的发病率和死亡率都在上升。2018年，新增癌症病例为1810万，癌症死亡人数达到960万。肺癌、乳腺癌、结肠直肠癌、前列腺癌、胃癌、肝癌是[131]型最常见的癌症类型。早期诊断癌症是提高临床疗效的关键。然而，早期癌症的症状很少被发现，导致治疗的延迟，[132]晚期诊断大大降低了患者的生存率。

乳腺癌是目前世界上最为常见，致死率较高的癌症之一，严重威胁着当今世界上女性的身心健康。并且乳腺癌的发展与雌激素受体密切相关，并且研究发现雌激素受体α亚型（Estrogen receptors alpha, ERα）在大约50%-80%的大多数乳腺肿瘤细胞中表达，说明Erα对于乳腺癌的预测具有至关重要的作用，因此ERα被认为是治疗乳腺癌的重要靶标，能够拮抗（微生物菌落之间的对抗作用）ERα活性的化合物可能是治疗乳腺癌的候选药物，ERα是乳腺癌疾病的活性调节因子，它可以治疗乳腺癌及抑制肿瘤血管新生，因此建立化合物活性预测模型的方法来预测具有更好生物活性的新化合物分子或者导已有活性化合物的结构优化起着至关重要的作用。

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，也就是是否能被人体吸收或者在人体中有一个合适的代谢速度，这被称为ADMET。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。

1.2 问题分析

由于ERα被认为是治疗乳腺癌的重要靶标，能够拮抗（微生物菌落之间的对抗作用）ERα活性的化合物可能是治疗乳腺癌的候选药物，因此建立化合物活性预测模型的方法来预测具有更好生物活性的新化合物分子或者导已有活性化合物的结构优化起着至关重要的作用。当然，一款药物除了具备良好的生物活性，

还需要考虑是否在人体中具备良好的药代动力学性质和安全性，这些特性合称为ADMET，并且我们仅考虑化合物的5种ADMET性质:

1. 小肠上皮细胞渗透性（Caco-2），可度量化合物被人体吸收的能力；
2. 细胞色素P450酶（Cytochrome P450, CYP）3A4亚型（CYP3A4），这是人体内的主要代谢酶，可度量化合物的代谢稳定性；
3. 化合物心脏安全性评价（human Ether-a-go-go Related Gene, hERG），可度量化合物的心脏毒性；
4. 人体口服生物利用度（Human Oral Bioavailability, HOB），可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；
5. 微核试验（Micronucleus，MN），是检测化合物是否具有遗传毒性的一种方法。

我们需要检测这5种性质对于人体的作用。

1.3 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个生物活性数据，5个ADMET性质数据），构建化合物生物活性的定量预测模型和ADMET性质的分类预测模型，从而为同时优化ERα拮抗剂的生物活性和ADMET性质提供预测服务。

2、模型假设和符号说明

2.1 模型假设

3、问题一的分析与建模

3.1 问题一的分析

3.2 随机森林算法

3.2.1 随机森林算法的介绍

在集成学习中有很多研究热点，主流的方法有 boosting 算法以及 bagging 算法。在boosting 算法中，后继树模型会对先前的树模型预测错误的样本点赋予更多的权重，使得模型更关注预测错误的样本点，按照任务属性分别使用投票法或者均值进行预测；在bagging 算法中，后继树模型与先前的树模型相互之间是独立的，并且采用了有放回的随机采样，最后根据投票法或者均值进行预测。Breiman 提出了随机森林，在 bagging 算法的基础上将随机森林引入了两层随机性。随机森林除了采用自助采样样本这个随机性以外，还在构造树的时候引入了特征的随机性，选择随机选择的特征来构建最佳分割。相比于其他分类器，随机森林的这两个随机性使得其表现突出，更稳健并有效的防止了过拟合。

3.2.1 随机森林算法的基本原理

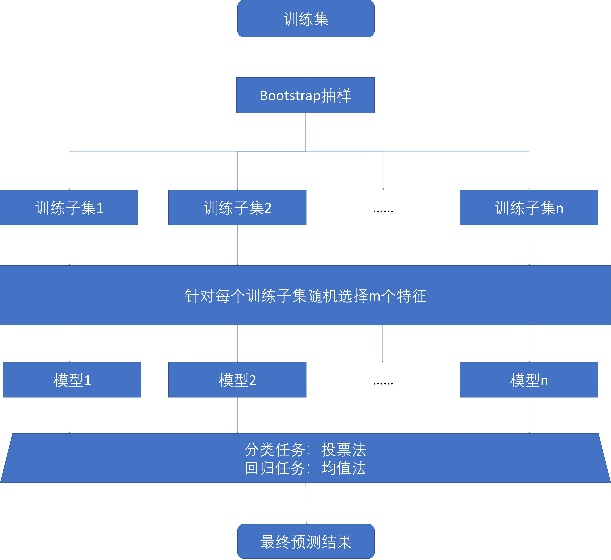
随机森林是一种经典的适用于分类问题和回归分析的机器学习算法，已经在众多领域包括化工行业广泛应用[4]。本文所研究的汽油精制产生辛烷值的生产过程是一个非线性过程[11]，而随机森林也可以解决非线性问题。本题生产数据有缺失值和噪声，而随机森林因为其引入的随机性，对缺失值和噪声比较鲁棒，具有很好的抗噪能力。同时 RF 有一个非常重要的优势，即 RF 不需要使用交叉验证方法来计算误差，RF 可以在生成的过程中建立误差的无偏估计--袋外误差率。因此，在考虑模型时，本文优先考虑了随机森林 RF。

随机森林算法的步骤（包括归类和回归）如下，具体的流程图间图3-1。

STEP1：从原始数据集中有放回的随机采样个训练样本子集；

STEP2：对于每个训练样本子集，生长一个未修剪的分类或回归树，每棵树都尽最大程度的生长，并做了如下修改：在每个节点上，在所有变量中随机采样得到个变量子集，将其作为树分裂的候选变量；

STEP3：通过组合个树的预测结果来预测新样本。（对于分类树来说可用投票法，对于回归树来说，可用均值法）。



3.2 问题一的结论

4、问题二的分析与建模

4.1 问题二的分析

4.2 问题二的结论

5、问题三的分析与建模

5.1 问题分析

候选药物除了需要具备良好的生物活性之外，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性质。在问题（3）中需要对化合物的5种ADMET性质：Caco-2、CYP3A4、hERG、HOB、MIN分别建立分类预测模型，并将待检测的50个化合物通过这5个分类模型进行预测，以此来得到对应的结果值。

根据问题1和问题2得出的结论，我们已经知道了变量对于生物活性影响的重要性排序，并得到了前20个对于生物活性最具有显著影响的分子描述符，并构建出了化合物对于Erα生物活性的定量预测模型。

常见的分类预测方法有二元分类和多元分类。在python种，SVM的构建支持向量机可以对多元分类有很好的预测作用，二在机器学习种，二元分类也是一种比较常见的模型，包括art决策树、随机森林和神经网络等。根据前2个问题得出的结论和数据，初步打算使用SVM和神经网络来作为此次分类预测模型来实现对于5种ADMET性质的分类。

5.2 基于SVM算法的分类预测模型

5.2.1 SVM算法的介绍

支持向量机（Support Vector Machine, SVM）是一种对数据进行二分类的广义线性分类器，其分类边界是对学习样本求解的最大间隔超平面。SVM使用铰链损失函数计算经验风险并在求解系统中加入了正则化项以优化结构风险，是一个具有稀疏性和稳健性的分类器 。支持向量机为我们提供了在众多可能的分类器之间进行选择的原则，从而确保模型对未知数据集具有更高的泛化能力，通过超平面的引入，可以实现对于样本数据的三维分隔。

支持向量机是基于结构风险最小化原理提出的一种机器学习方法。它在解决小样本、非线性、高维数据集的分类问题上有许多独特的优点。支持向量机的分类思想是在线性可分条件下寻找最优分类面从而实现数据的多元分类[1]。

5.2.2 SVM算法的具体实现

给定的训练向量在两个类当中，并且给出了向量。我们的目标是要找到和使得公式对于绝大多数的样本都是分类正确的。

SVM算法的目的是为了解决以下比较初始的问题：



Subject to 



直觉上，我们试图最大化边距（通过最小化），同时在样本被错误分类或在边际边界时产生惩罚。理想情况下，该值对所有样本值都将，这就表明了这是一个完美的预测。但是问题是我们通常并不总是能够通过一个超平面来实现完美的分离，因此我们允许一些样品处在一个与它们正确的边界距离相距的距离。惩罚项控制了这个惩罚的强度，因此，起着反常规参数的作用。

原始的双重问题是：



Subject to 



在这里表示所有元素的向量，然后是一个n阶乘的半正定矩阵。

，其中是核心。被称为双系数，并且它们被所限制。这种双重表现突出了这样一个事实，就是训练向量通过函数隐式映射到更高（可能是无限）维度的空间。

一旦优化问题被解决，给定样本的决策函数的输出变为：



并且决策类对应它的符号。我们只需要对支持向量求和（即位于边界内的样本），因为对于其他样本来说，它们的偶系数是零。

下面有一段内容没有翻译。

5.3基于BP神经网络的分类预测模型

5.3.1 BP神经网络算法介绍

BP神经网络算法是应用最为广泛的神经网络。BP神经网络具有任意复杂的模式分类能力和优良的多维函数映射能力，解决了简单感知器不能解决的异或(Exclusive OR，XOR)和一些其他问题。从结构上讲，BP网络具有输入层、隐藏层和输出层；从本质上讲，BP算法就是以网络误差平方为目标函数、采用梯度下降法来计算目标函数的最小值。

陈：BP神经网络

​ BP 神经网络算法的核心思想：使用梯度下降来搜索可能的权向量的假设空间，以找到最佳的拟合样例的权向量。具体而言，即利用损失函数，每次向损失函数负梯度方向移动，直到损失函数取得最小值。

或者说，反向传播算法，是根据损失函数，求出损失函数关于每一层的权值及偏置项的偏导数，也称为梯度，用该值更新初始的权值和偏置项，一直更新到损失函数取得最小值或是设置的迭代次数完成为止。以此来计算神经网络中的最佳的参数[加一个文献]。

第一层是输入节点，最后一层是输出节点，其间有一层或多层隐含层节点，各层内的节点间无连接，信息仅单方向流动。BP 神经网络的训练方法是通过反向误差传播原理不断调整网络权值，从而使得实际输出与期望输出之间的误差平方和达到最小或小于某个阈值，其训练或学习过程由正向传播和反向传播两部分构成[加一个文献]。当正向传播时，信息从输入层经隐含处理后传入输出层，每一层神经元状态只影响下一层的神经元的状态。如果在输出层得不到希望的 输出，则转入反向传播，将误差信号沿原来的神经元连接通路返回。在返回过程中，逐一修改各神经元连接的权值，经过不断的迭代，最后使信号误差达到允许范围之内。

5.3.2 神经网络的激励函数

函数一般可以用以下方程来表示，大于阈值激活，小于阈值则抑制。

(1)阈值型



从表达式中可以看出函数只有0和1两种状态，即时，函数值为1，神经元被激活；时，函数值为0，神经元被抑制。

(2)伪线型



上式中的c是一个常数，从函数中可以看出，在时函数的输入和输出时满足线性关系的；在或时函数为常数，不满足线性关系，但函数一直连续。

(3)s型（sigmoid）



从表达式中可以看出sigmoid函数时连续并且可导的，同时也有最大值和最小值，反映出了神经元的饱和特性。

通过对以上的几种函数模型的分析可以知道阈值型函数不连续只有0和1这两种状态，伪线性函数是连续的函数，但是函数不光滑且不可导，因此这两种函数都不适合作为神经网络的传递函数。S型函数具有连续、光滑、可导的特性，非常好的描述了神经元接受信息时的工作状态模型。S 型函数是单增函数以及它的反函数也是单增函数，并且此函数能够将变量映射到 0~1 之间，使输出的结果更符合实际的要求。S 函数的函数分布图如图 6-2 所示。

5.3.3 BP神经网络算法

算法的步骤如下：

STEP1:将输入层、隐含层和输出层神经元之间的连接权重、以及神经元阈值、赋予间的随机值，并指定学习系数以及神经元的激励函数。

STEP2：随机选取数据组，提交给网络，用网络的设置计算隐含层各神经元的输出：



STEP3：利用网络的设置，计算输出层神经元的响应为：



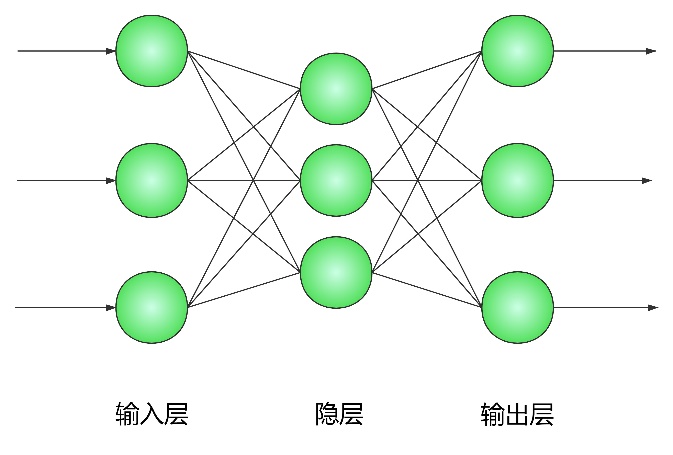
STEP4：利用给定的输出数据，计算输出层神经元的一般化误差和隐含层各神经元的一般化误差



STEP5：利用和隐含层各神经元输出修正隐含层与输出层的连接权重和神经元阈值，且利用和各神经元输入修正和神经元阈值。



STEP6:随机选取另一个数据组，返回到步骤2，重复利用全部数据组进行学习。这是网络样本集完成的一次学习过程，然后重复了下一次学习过程，直到网络的性能符合要求，或学习次数达到设定次数为止。



5.3.4 BP神经网络的评价指标

首先研究的内容是辛烷损失值的预测，该问题属于回归问题。因此本文选取了均方根误差(Root Mean Square Error, RMSE)、平均绝对误差(Mean Absolute Error, MAE)、决定系数（R-square,）以及在设定的误差范围内精度（Accuracy,）。

定义表示测试样本总数，是真实值，是预测值，是实际数据集的均值，表示误差在范围内的预测样本数。标准计算公式如下：









均方根误差 RMSE 可以衡量预测值和真实值之间的偏差大小，若模型的 RMSE 越低，代表该模型表现越好；平均绝对误差对离群值不怎么敏感，能够反映预测值误差的实际情况，如果 MAE 越小，代表该模型表现越好；取值范围为[0*,*1]的决定系数可衡量回归模型的好坏，表征回归模型对真实值的拟合程度，回归模型对真实值的拟合程度越高，决定系数的值越接近于1，一般大于 0.4 就可以说明拟合效果较好；为了更好的反映模型表现，本文设定了一个精度指标，该指标是指误差在某一范围内符合的样本数占总样本数的比例。

问题三的结论

6、问题四的分析与建模

6.1 问题分析

6.2 问题四的结论

7、模型的评价

参考文献

[1] C. Cortes and V. Vapnik, “基于SVM模型的车辆侧翻预测方法,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/bf00994018.

For test