2021 年全国研究生数学建模竞赛 D 题

抗乳腺癌候选药物的优化建模

评阅建议

论文摘要撰写规范、引用文献条理清晰、逻辑关系严谨、结论合理，问题求解有新意。有关键词，参考文献格式规范。**（满分 10 分）**

问题 1：描述符选择**（满分 15 分）**

本题涉及描述符 729 个，其中描述符间存在一定的冗余现象，需要筛选出与生物活性最相关的 20 个描述符（即变量）。本题重点考察选取的主要变量是否具有代表性、以及合理性解释。变量选择的过程应当有理有据，并且表述清晰。

本问不宜采用主成分分析（PCA）、t-SNE 等破坏原变量完整性的方法进行降维，如果采用此类做法，本问分值应当低于 7 分。

评阅要点：

1. 所选变量的代表性**（10 分）**。代表性主要考察变量与抗乳腺癌活性之间的关联度要大，并且不应只是线性关联度大。考虑到变量的代表性并作出合理解释得 10 分，可以采用的方法包括但不限于：

（1） 利用灰色关联度进行相关性分析，直接筛选得到与生物活性关联度大的特征变量；

（2） 通过相关性分析，筛除与生物活性弱相关的特征后，再做偏相关性分析，得到主要的特征变量；

（3） 通过相关性分析，筛除弱相关特征后，利用树模型进行回归，利用信息熵或者基尼系数获得特征的重要性打分，进而确定主要的特征变量；

（4） 其它可以用于确定主要特征变量的方法。

2. 所选出的具体变量类别**（5 分）**。最终得到的相关特征变量应包含在以下 11 类描述符中，每识别出一个描述符类别得 1 分，最多 5 分。

序号	描述符类别	描述符
1	AlogP	ALogP, ALogp2, AMR
2	Carbon types	C3SP2, C1SP2
3	Autocorrelation (charge)	ATSc1, ATSc2, ATSc3, ATSc4
4	LipoaffinityIndex	LipoaffinityIndex
5	BCUT	BCUTc-1l, BCUTc-1h, BCUTp-1l, BCUTp-1h
6	XLogP	XLogP
7	Molecular distance edge	MDEC-22, MDEC-23, MDEC-33
8	Atom type electrotopological state	minssCH2, minHBa, mindssC
9	Molecular linear free energy relation	MLFER_A
10	Crippen logP and MR	CrippenLogP
11	Acidic group count	nAcid

3. 关于变量独立性的说明：本问题没有要求考虑变量独立性，但如果考虑了独立性也没有问题，独立性分数放在第二问，因此相应的分数在第二问中给，考虑了的都给分。

问题 2：构建生物活性的定量预测模型（满分 25 分）

本题主要考察参赛者对于构建回归模型的细节把控，本题的整体思路分为四个部分：1）建模前的特征和数据处理（尤其是要考虑所选变量的独立性）；2）建模方法的选择和超参数的调整；3）模型的评价和分析；4）模型的预测性能。

评阅要点：

1. 建模前处理（5 分）

（1）特征变量的选择。参赛者自行选择合适的特征变量构建模型，但需要说明指定特征变量数量的原因；

（2）特征的预处理。例如，参赛者可采用归一化等方式对各个特征变量进行标准化处理；

（3）数据集的划分。参赛者在建模前，应当使用常见的比例（例如：4:1）对数

据集进行划分。参赛者也可以将数据集划分为 5 份或 10 份，用于后面进行的 5 折或 10 折交叉验证。

2. 所选变量的独立性（**5 分**）。独立性主要考察变量之间的相关性大小，考虑到变量的独立性并作出合理解释得 7 分，采用的方法包括但不限于：

（1） Pearson 相关系数或者方差齐性检验（局限性在于无法识别变量间的非线性关系）、互信息法；

（2） Spearman 相关系数法；

（3） 按相关性（系数）聚类；

（4） 独立性检验。

3. 建模方法的选择和超参数的调整（**5 分**）

（1）建模方法的选择包括但不限于朴素贝叶斯、最近邻、决策树、随机森林、支持向量机、神经网络等，应采用多种机器学习方法进行比较，以得到最优模型；

（2）模型训练过程中的参数设置（起码列出最终的模型参数，以 BP 神经网络为例，需列出权值、偏置、激活函数等参数，则此模型不是黑箱模型，可依据这些参数重现 BP 模型）必须展现出来，保证模型的可重现性。

4. 模型的评价和分析（**5 分**）

（1）构建好模型后，对模型进行验证（例如：5 折或 10 折交叉验证、留一交叉验证），必须列出模型最终的训练和测试样本的模型评价指标（例如：相对误差或者决定系数等），需要用文字结合图表的方式表达模型的评价结果；

（2）对模型进行鲁棒性（稳健性）分析，例如采用 Y-random 等方式；

（3）对模型进行过拟合分析，判断模型是否过拟合。如果模型训练和测试样本结果若误差相差较大，这说明模型存在过拟合情况，应不得分。

5. 模型的预测性能（**5 分**）

根据对 50 个验证集化合物的预测结果（参见附件中的 ERα-activity 表格），我们需要计算所有化合物的平均绝对误差（MAE），按照平均绝对误差的结果进行给分，最高 5 分，平均绝对误差越小得分越高。

按照 MAE 值直接规定得分区间：

平均绝对误差（MAE）	得分
0 - 0.3	5
0.3 - 0.5	4
0.5 - 0.7	3
0.7 - 0.9	2
0.9 - 1.1	1
1.1 以上	0

问题 3：构建 5 种 ADMET 性质的分类预测模型（满分 30 分）

本题考查参赛者对于分类模型构建的细节把控。本题同样可以分为 4 个部分：1）建模前的特征和数据处理；2）建模方法的选择和超参数的优化；3）模型的评价和分析；4）模型的预测性能。

评阅要点：

1. 建模前的特征和数据处理（5 分）

包括特征变量选择和预处理，变量选择也要考虑独立性问题。

2. 建模方法的选择和超参数的调整（5 分）

（1）采样方式的选择。ADMET 的数据集存在阴性、阳性化合物比例失调情况，需要参赛者采用过采样或者欠采样等方式解决数据不平衡的情况；

（2）建模方法的选择包括但不限于朴素贝叶斯、最近邻、决策树、随机森林、支持向量机、神经网络等，应采用多种机器学习方法进行比较，并获得最优模型。可以分别针对单一的端点构建模型，也可以构建多任务模型。鼓励使用多任务模型，酌情加分；

（3）模型训练过程中需要表明参数调整的过程，至少需要给出参数优化的方法（例如：格点搜索）和模型的最终参数，保证模型的重现性。

3. 模型的评价和分析（5 分）

（1）构建好模型后，对模型进行合理评价（例如：5 折或 10 折交叉验证、留一交叉验证），必须列出模型在训练样本和测试样本的表现。例如，准确率、

召回率、受试者工作特性曲线下面积（AUC）或者马修斯相关系数（MCC）等（图或表；

（2）模型的过拟合分析。模型的训练和测试结果的误差应是接近。模型训练和测试样本结果若误差相差较大，出现过拟合现象，应不得分。

4. 模型的具体预测性能（**15 分**）

根据对 5 类 ADMET 性质的 50 个验证集化合物的预测结果（参见附件中的 ADMET 表格），分别计算每个端点的预测准确率，当预测准确率为 90%以上时，获得 3 分；准确率 80%以上，获得 2 分；准确率 70%以上得 1 分，70%以下不得分。每个端点最高 3 分，合计 15 分。

问题 4：寻找分子描述符的规律（满分 20 分）

本题为开放式题目，主要考查参赛者对问题的综合分析能力。即在前述问题的基础上，寻找使化合物既具有良好生物活性又具有良好 ADMET 性质的分子结构条件，即最佳分子描述符及其取值范围，为多目标优化问题。本题具有区分度，评委认可的特别好的优化方法，可以**额外加 5 分**，并给出加分理由。

评阅要点：

1. 本题可以认为是多目标优化问题，可以采用的方法包括但不限于遗传算法、粒子群算法、模拟退火等；也可以转化为双目标优化问题，并用多目标算法求解最优解集；甚至可以考虑将多目标优化转化为单目标优化问题。参赛者需要用数学或文字形式描述优化问题，并写清楚所采用的优化算法（**5 分**）。

2. 明确优化目标和约束条件。优化目标应当同时考虑抗乳腺癌活性和 ADMET 性质。约束条件中的分子描述符应当从问题 1 中挑选出的重要的分子描述符出发，各个分子描述符取值范围的设置应当考虑实际情况【例如：其在数据集中的分布情况和数据类型（离散或连续）】。在计算目标函数值过程中，可以采用问题 2 和问题 3 构建的模型，也可以重新构建模型（**5 分**）。

3. 应当使用图表的方式，展现模型优化过程的参数设置情况（以遗传算法为例，

需列出种群规模、交叉概率等参数）。**（5 分）**

4. 对优化结果进行解释分析，鼓励结合生物学和化学知识**（5 分）**。

5. 除最优化方法外，还可采用其它能够得到对化合物抗乳腺癌活性和 ADMET 性质具有重要影响的分子描述符及其取值范围的方法（例如：统计分析或者特征选择等），言之有理即可，可根据实际情况酌情给分，最高不超过 15 分。

6. 部分重要分子描述符及取值范围（仅供参考）：

分子描述符	参考取值范围
ALogP	（0.0239, 2.9208）
ALogp2	（0.0977, 9.3789）
AMR	（82.6879, 146.3967）
ATSc1	（0.1182, 0.5872）
ATSc2	（-0.2964, -0.0364）
BCUTc-1h	（0.0796, 0.2879）
BCUTp-1h	（10.3568, 13.7077）
XLogP	（2.082, 5.432）

附件：ERα_test_results.xlsx