APMCM亚太地区大学生数学建模竞赛（中文赛项）

参赛手册

扫一扫进入大赛官网

APMCM

APMCM 亚太地区大学生数学建模竞赛

参 赛 手 册

亚 太 赛 组 委 会 制 作

目　录

中国国际科技促进会物联网工作委员会
北京图象图形学学会

2026年APMCM亚太地区大学生数学建模竞赛(中文赛项）报名通知

各高等院校：

2026年第十六届亚太地区大学生数学建模竞赛（以下简称“竞赛”）是由中国国际科技促进会物联网工作委员会、北京图象图形学学会主办的亚太地区大学生学科类竞赛，竞赛由亚太地区大学生数学建模竞赛组委会负责组织，欢迎各高等院校按照竞赛章程及有关规定组织同学报名参赛。

1. 竞赛的时间确定为 2026年6月12日18:00 至2026年6月15日20:00。

2. 本次竞赛时间为4天，参赛对象为普通高校全日制在校大学生，参赛队由1-3名大学生组成。报名截止日期为6月12日12：00，报名截止后不能再更改报名信息。

3. 竞赛允许跨校组队，必须完整填写每位参赛者以及指导教师所在的学校全称。

4. 竞赛分为研究生组、本科组、专科组；报名时请根据参赛队员中最高在读学历选择组别。

5. 每所院校参赛队数不作统一规定；组委会将根据报名情况确定获一、二、三等奖的数量（大约分别占成功参赛总队数的 5%、15%、25%）。

6. 竞赛题目共2道（A、B题），一般来源于人工智能、工程与管理、互联网、图象图形等领域经过适当简化加工的实际问题。

7. 竞赛只需要提交电子版论文，不需要邮寄纸质版论文；所有参赛队必须提交中文版论文。

8. 优秀参赛者邀请参加图像图形技术与应用学术会议（IGTA），会议上将对获奖代表进行颁奖。

9. 组织集体报名的院校，该校负责人将所有参赛队的集体报名表发到竞赛组委会邮箱（apmcm@mathor.com），详情见《2026年亚太赛集体报名高校须知》。

10. 竞赛不邮寄书面题目，竞赛开始时赛题将在竞赛官网（www.apmcm.org）、北京图象图形学学会官网（www.bsig.org.cn）以及报名主页（www.saikr.com/apmcm2602）上公布。

11. 其它事项请登录竞赛官网（www.apmcm.org），查看竞赛组委会的有关通知文件。

中国国际科技促进会
物联网工作委员会
2026年3月

北京图象图形学学会
2026年3月

亚太地区大学生数学建模竞赛组委会

2026年APMCM亚太地区大学生数学建模竞赛(中文赛项）集体报名高校须知

亚太赛[2026]01 号

各高等院校：

2026年第十六届亚太地区大学生数学建模竞赛（以下简称“竞赛”）是由中国国际科技促进会物联网工作委员会、北京图象图形学学会主办的亚太地区大学生学科类竞赛，竞赛由亚太地区大学生数学建模竞赛组委会负责组织，欢迎各高等院校按照竞赛章程及有关规定组织同学报名参赛。

1、集体报名的高校负责人须通知本校参赛队伍在报名官网（https://www.saikr.com/vse/apmcm2602）完成注册（注意：每支参赛队需在竞赛报名截止前在官网完整填写所有成员信息），不需要在线缴费，生成队伍的参赛编号后填写集体报名表格，同集体报名支付截图一同发至组委会邮箱（apmcm@mathor.com）。

2、组委会工作人员接收集体报名表格以后会在官网审核通过集体报名的队伍编号，操作完成后会发邮件给负责人核对。

3、集体报名缴费方式

本次竞赛集体注册费由培优卓越（天津）传媒科技有限公司收取并开具发票，参赛高校需为每个参赛队提交100元注册费。

银行户名：培优卓越（天津）传媒科技有限公司

开 户 行：招商银行股份有限公司天津高新区支行

账　　号：122918641410000

注：支付时请备注“**学校+队数+支付者姓名**”，集体转账后请负责人保留转账成功截图，并同集体报名表一同发至组委会邮箱（**apmcm@mathor.com**），如果需要开发票，请把发票抬头、纳税人识别号、是否开成一张发票等信息一起发送到组委会邮箱。

4、 论文提交方式（详情可参见《2026年亚太赛参赛规则》）

（1）**打印承诺书并签字**，通过拍照或扫描的方式生成 PDF 文件或 JPG 图片。

（2）**提交支撑材料**，将与竞赛相关的其他所有文件（包括程序、数据（赛题中的原始数据 除外）和结果等）压缩打包上传。“支撑材料”由参赛队员在审慎考虑的基础上，选择性地添加， 如竞赛题目中没有明确的要求，则不作为必须提交的材料（如果有自己编写的程序，虽然按照论文格式要求，程序必须放入正文附录，但源程序仍然要以支撑材料的形式提供）。

（3）**论文命名为**：题号+参赛控制编号（文件类型为 PDF）；

（4）**承诺书命名**：题号+参赛控制编号+cns（文件类型为 PDF 或 JPG）。

（5）**支撑材料命名**：题号+参赛控制编号+fj（文件类型为 rar 或 zip）

例 0001 队选择的是 A 题，则论文命名应为：A260001，承诺书命名应为：A260001cns，支撑材料命名应为：A260001fj。

（6）**然后分别上传至报名官网的对应电子档提交处**。

5、 报名联系方式

APMCM 组委会秘书处：林老师

电话&微信：15600483352

邮箱：apmcm@mathor.com

亚太地区大学生数学建模竞赛组委会

2026年4月

历 年 赛 题

亚太赛组委会制作

2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛

A题 农业灌溉系统优化

农业生产中，水资源管理对作物生长至关重要。智能灌溉系统利用传感器和数据分析来优化水资源的使用。然而，如何在多种土壤类型和动态天气条件下高效配置这些系统，并且根据实时数据做出快速调整，是一个复杂的挑战。特别是在面对不同作物、变化的气象条件和土壤湿度时，优化策略需要考虑多种动态因素。

某农场占地面积 1 公顷，农场为正方形，且依河而建（如图 1），种植了不同作物（如表 1），农场除本身拥有的储水罐以外，附近还有一条河流，如需引水则需要建设饮水管道， 该农场自 5 月 1 日起同时播种三种作物。

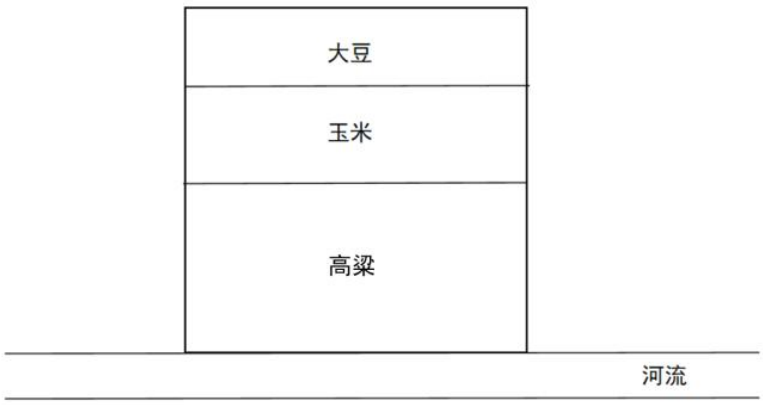


图 1 作物种植分布

表 1　不同作物种植表

作物	耕地面积（公顷）	播种期需水量（L/m²）	开花期需水量（L/m²）	成熟期需水量（L/m²）	播种期时间（天）	开花期时间（天）
高粱	0.5	5	10	8	20	50
玉米	0.3	6	12	10	32	50
大豆	0.2	4	8	6	40	40

*注：需水量是指在满足最低土壤湿度的基础上，作物生长较好还需求的灌溉量（供给作物吸收），保持最低土壤湿度只能保证作物存活。

农业灌溉系统有“平时用河水，旱时用储水”的说法。河流引水管道的成本会随距离非线性增长，每段水管的建设费用与水管长度和该段水管的通流量有关，成本 $C= 50L^{1.2} + 0.1Q^{1.5}$（L 为该段管道的长度，Q 为该段水管的日流量），喷灌的相邻两喷头距离≥15m，喷头喷淋半径 15m（见图 2）。此外，也可以通过储水桶灌溉，建设储水罐时，每增加 1L 容积，储水罐成本增加 5，储水罐可覆盖以自身为圆心半径 15m 的区域。

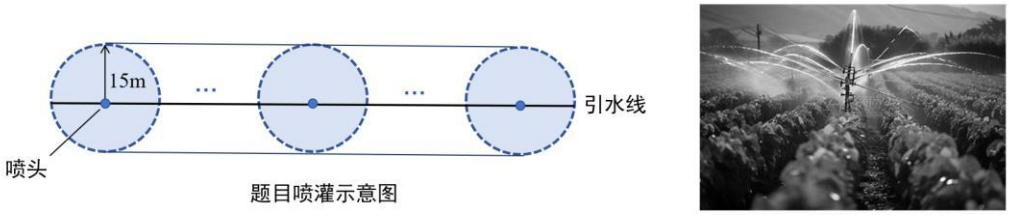


图 2 喷灌示意图

为探究该农场所处区域气象变化，并且给予场主一些帮助，某团队采集了一些数据放在附件“该地土壤湿度数据.xlsx”与“降水量等逐时气象数据.xls”中，请基于这些数据回答以下问题。

问题 1 农业中气象预测十分重要，建立一个预测模型，综合考虑历史气象数据，试求解土壤湿度 5cm_SM 与其他除土壤湿度以外的气象数据之间的关系，并根据你的模型，按照表 2 中给出的某天内不同小时测得的气象数据进行预测当天土壤湿度，填写表 2 放入你的论文之中。

表 2 土壤数据预测表

时间（h）	T	Po	P	Pa	U	DD	Ff	RRR	预测的当天湿度 5cm_SM
02	19.3	731.5	751.7	1.0	99	西	轻风	15.0	
05	20.0	732.0	752.4	0.5	94	西	轻	6.0	

						南	风		
08	23.4	732.8	753.2	0.8	80	西	轻风	0	
11	28.0	733.5	753.8	0.7	44	西北	轻风	0	

问题 2 不考虑作物不同生长时期的需水量，只考虑土壤湿度，请你使用附件中 2021- 07-01 至 2021-07-31 的数据合理规划，考虑引水管布线与储水罐容积、位置，给出灌溉系统布线规划图，及建设的花费，在满足作物存活前提下使得总花费最小。

问题 3 由于河流水源并不稳定，为预防旱季，储水罐必须保有应急储备水源，该部分水源不可用于日常灌溉，在应急条件下，储水罐水源可以供给周围半径 50m 范围喷头使用。请你在问题 2 的基础上，完成以下任务：

● 若旱灾来临，问题 2 时间内，河流供给水量下降，只能提供在第 2 问你建立的引水系统总流量的 80%，你的储水罐和引水管布线设计能够保证多少作物存活？最多能保证多少作物正常生长？填写表 3 并放入你的论文中。

表 3　作物存活情况表

作物	种植面积（公顷）	存活面积（公顷）	正常生长面积（公顷）
高粱	0.5		
玉米	0.3		
大豆	0.2		

● 分析储水罐应具备的应急储备水源比例与旱灾概率的关系，填写表 4 并放入你的论文中。

表 4 应急水源比例与旱灾概率关系表

旱灾概率（%）	建议应急储备水源比例（%）	此比例能否保证作物全部存活

10		
30		
50		
80		
100		

问题 4 假设各作物成熟期均为 20 天，使用 2021-05-01 至 2021-07-31 数据计算并规划不同作物每月的灌溉方案，你之前建立的系统是否能够满足灌溉需求？如果不能，请修改系统布线，将月灌溉情况进行可视化（每种作物该月所用水量、水量来源，若在 7 月 31 日前作物已成熟，则只规划到成熟期结束），填写以下表 5 并放入你的论文中。

表 5 灌溉安排表

日期	作物	总灌溉量（L）	水源比例（河水/储水罐）	备注（如是否调整系统布线）
5 月				
6 月				
7 月				

附录：

1、植物在生长期间的最低土壤湿度（5cm_SM）应保持在大于等于 0.22。

2、题目中所涉及土壤湿度为绝对湿度，是指土壤中水分的质量占土壤干重的百分比，计算公式如下：

$$\text{绝对湿度}(\%) = \left(\frac{m_w}{m_d}\right) \times 100\%$$

其中：m_w是土壤中水分的质量（湿重减去干重），m_d是土壤的干重。

3、单位面积土壤干重$m_d = 1500\,\mathrm{kg/m^3}$。

2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛

B题 疾病的预测与大数据分析

为进一步加强以目标为导向的医疗质量安全管理工作,国家卫生健康委组织制定了《2025 年国家医疗质量安全改进目标》和 2025 年质控工作改进目标。

根据世界卫生组织（WHO）的数据统计，心血管疾病（CVD）是全球第一大死亡原因，估计每年夺去 1790 万人的生命，占全球死亡人数的 31%。附件中心脏病 heart.csv 数据集包含 11 个可用于预测可能的心脏病的特征。患有心血管疾病或心血管风险高的人（由于存在一种或多种危险因素，如高血压，糖尿病，高脂血症或已经确定的疾病）需要早期发现和管理。

此外，中风是全球第二大死亡原因，约占总死亡人数的 11%。本赛题附件中 stroke.csv 中风数据集中的每一行都提供了有关患者的相关信息，包含输入参数（如性别，年龄，各种疾病和吸烟状况）等指标，用于预测患者是否可能中风。第三种疾病肝硬化（cirrhosis）是由多种形式的肝病和病症（如肝炎和慢性酒精中毒）引起的肝脏瘢痕形成（纤维化）的晚期。

本次比赛提供了三种疾病数据集 stroke.csv、heart.csv 和 cirrhosis.csv，请你们团队运用数据统计与分析技能，深入挖掘数据信息，预测不同疾病发生的概率。

问题 1 数据预处理与基础统计分析

对三种疾病数据集 stroke.csv、heart.csv 和 cirrhosis.csv 进行数据预处理、统计分析和可视化，并分析哪些因素会影响中风、心脏病和肝硬化的患病概率。

问题 2 不同疾病预测模型的构建

请分别选取合适的特征指标，建立中风、心脏病和肝硬化三种疾病患病概率的预测模型，并进行模型准确性的检验、灵敏度分析和模型改进。

问题 3 多疾病关联与综合风险评估

请综合分析中风、心脏病和肝硬化这三种疾病的共同特征和共病情况，建立

数学模型预测同时患有其中任意两种和同时患有三种疾病的概率。

问题 4 预防三种疾病的建议和措施

请根据你们数学模型和数据分析的结果，针对这三种疾病，给世界卫生组织（WHO）写一封信，提出你们的预防建议和措施。

2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛

C题 基于Quantum Boosting的二分类模型问题

集成学习是机器学习领域的核心技术之一，其主要通过组合多个弱分类器构建性能优异的强分类器。Boosting 作为集成学习的经典方法，是通过迭代训练弱分类器并调整样本权重，以达到逐步提升模型对复杂数据的预测能力。常见的 Boosting 算法如 AdaBoost、Gradient Boosting 等，已广泛应用于分类、回归等任务，展现了其强大的实用性。

近年来，量子计算技术和专用硬件迅速发展，Quantum Boosting（QBoost）作为一种新兴的 Boosting 变体，为传统机器学习注入了新的活力。QBoost 通过将 Boosting 问题转化为二次无约束二进制优化（QUBO）问题，利用相干光量子计算机等硬件的高效并行计算能力，快速求解最优弱分类器组合及其权重。这种方法不仅提升计算效率，还为探索量子优化与机器学习的交叉领域提供独特视角。

现要求你们队基于 QBoost 方法完成一个二分类任务。本赛题的任务是基于指定的数据集设计弱分类器、构建 QUBO 模型，并利用 Kaiwu SDK 中模拟退火求解器求解模型。具体问题如下：

问题 1 数据预处理与弱分类器构建

使用 Iris 数据集，选择 Setosa（标签 0）和 Versicolor（标签 1）两个类别，得到 100 个样本，每样本 4 个特征（萼片长度、萼片宽度、花瓣长度、花瓣宽度）。进行预处理（如标准化），并划分为训练集和测试集。说明预处理步骤。

基于所选数据集的特征，构造一组 M 个弱分类器。每个弱分类器可以基于单一特征或特征的简单组合（例如，基于阈值的决策规则）。记录每个弱分类器的预测结果 $h_j(x_i) \in \{-1, 1\}$，其中 $j = 1,2,\ldots,M$ 表示弱分类器索引，$i = 1,2,\ldots,N$ 表示样本索引。计算并记录每个弱分类器在训练集上的分类准确率。

问题 2 QBoost 建模与 QUBO 转化

将弱分类器集成问题转化为二次无约束二进制优化（QUBO）模型。目标是最小化强分类器的分类误差，即优化弱分类器权重，使加权组合在训练数据上的误分类率最低。为避免过拟合，可通过引入正则化项以限制选用的弱分类器数量。要求明确定义 QUBO 模型的目标函数和约束条件。

问题 3 利用 Kaiwu SDK 进行求解与模型评估

使用 Kaiwu SDK 中的模拟退火求解器，求解得到最优的弱分类器权重组合。分析所选弱分类器的特征及其组合方式，解释所选弱分类器的组合及其对模型性能的贡献。在测试集上评估最终强分类器的准确率等指标，并分析模型的泛化能力。

参赛者需提交一份完整的报告和源代码作为最终成果。报告应涵盖数据预处理、模型设计、实现过程及性能分析等内容，需符合标准期刊出版格式，包含摘要与关键词、引言、方法、实验、结论、参考文献等章节，以学术规范引用所使用的方法、工具及相关文献。评审将综合考虑模型的设计、代码的质量、结果的可信度、性能分析的深入程度以及报告论文的清晰度与逻辑性等方面。

温馨提示：

1、每位参赛者都需要在网站学习量子计算基础知识，并通过知识地图考核。

学习网址：https://kaiwu.qboson.com/plugin.php?id=knowledge

2、Kaiwu SDK 安装包可通过访问链接(https://platform.qboson.com)进行下载，安装说明可参考链接(https://b23.tv/IqKoPnv)。

3、Iris 数据集（鸢尾花数据集）是机器学习领域最经典的分类数据集之一，由统计学家 Fisher 于 1936 年提出。该数据集包含 150 个样本，分别来自三种鸢尾花（Setosa、Versicolor 和 Virginica），每个样本具有四个特征：萼片长度、萼片宽度、花瓣长度和花瓣宽度。Iris 数据集广泛用于算法测试，在 Python 中可通过 scikit-learn 库直接加载，或从 UCI 机器学习库网站(https://archive.ics.uci.edu/dataset/53/iris)免费下载。

对于比赛有任何技术及资源疑问，可通过扫描下述二维码咨询。

2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛

A题 飞行器外形的优化问题

飞行器是在大气层内或大气层外空间飞行的器械。飞行器可以分为：航空器、航天器、火箭和导弹。在大气层内飞行的称为航空器，如气球、飞艇、飞机等。它们靠空气的静浮力或空气相对运动产生的空气动力升空飞行。在太空飞行的称为航天器，如人造地球卫星、载人飞船、空间探测器、航天飞机等。它们在运载火箭的推动下获得必要的速度进入太空，然后依靠惯性做与天体类似的轨道运动。

目前人类历史上最快的飞行器是 1970 年代中期发射的太阳神（Helios）I 和 II 探测器，创下速度记录为每小时 252792 公里，等于每秒 70.22 公里。如果要走 20 光年的距离，需要 85714 年。如何优化飞行器的外形，使得其所受阻力最小，是航空航天领域里面非常重要的基础科学问题。

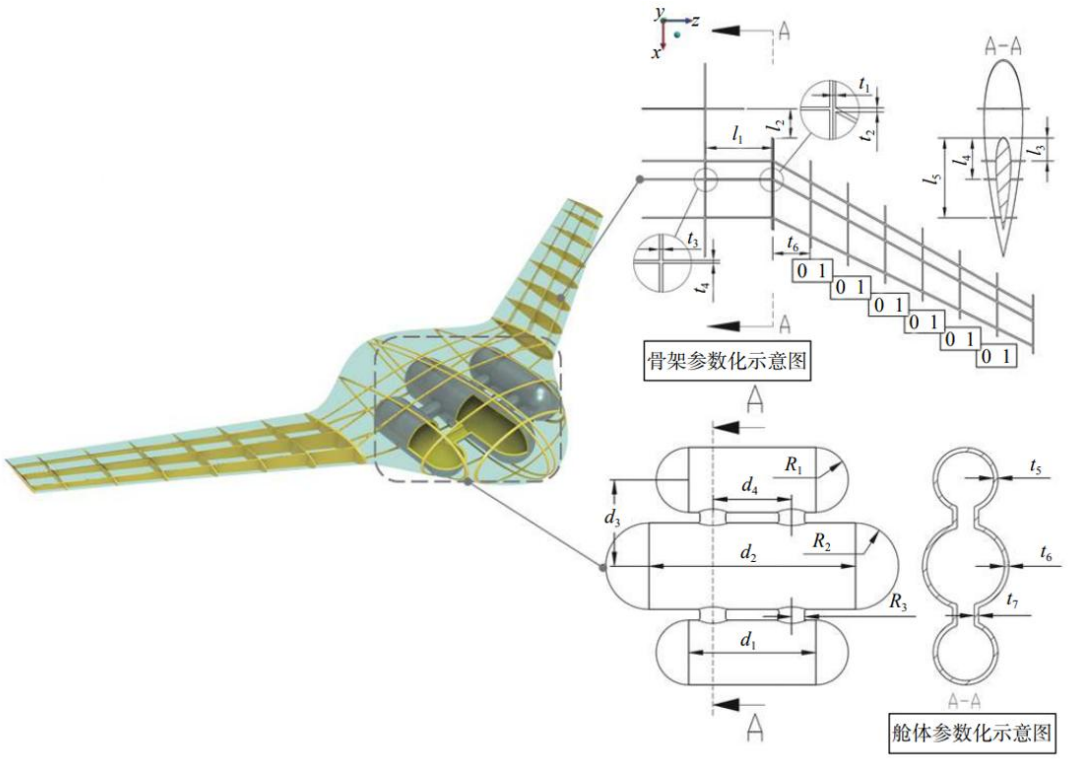


图 1. 某飞行器的结构示意图

图 1 是某飞行器的结构示意图，基于以上背景，请你们的团队通过数学建模的方法，研究如何优化飞行器的外形，解决以下问题：

问题 1 图 2 是某飞行器的部分尺寸示意图，请估计此飞行器的表面积和体积。

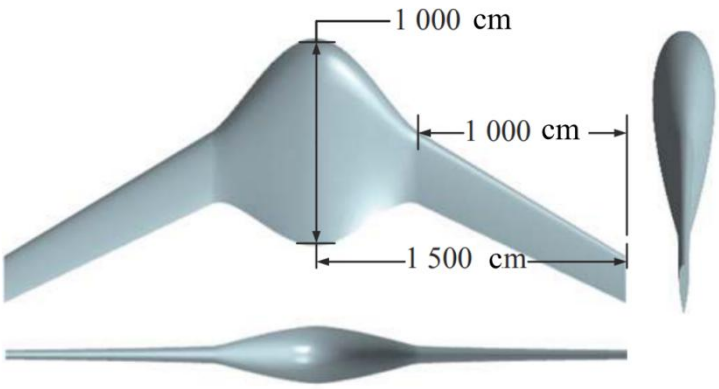


图 2. 某飞行器的部分尺寸示意图

问题 2 图 3 是某飞行器舱体结构的示意图，已知 $\boldsymbol{R_1}$ = 100 cm, $\boldsymbol{R_2}$ = 90 cm, $\boldsymbol{R_3}$ = 24 cm, 请根据图中的比例尺，估算该飞行器舱体结构的表面积和体积。

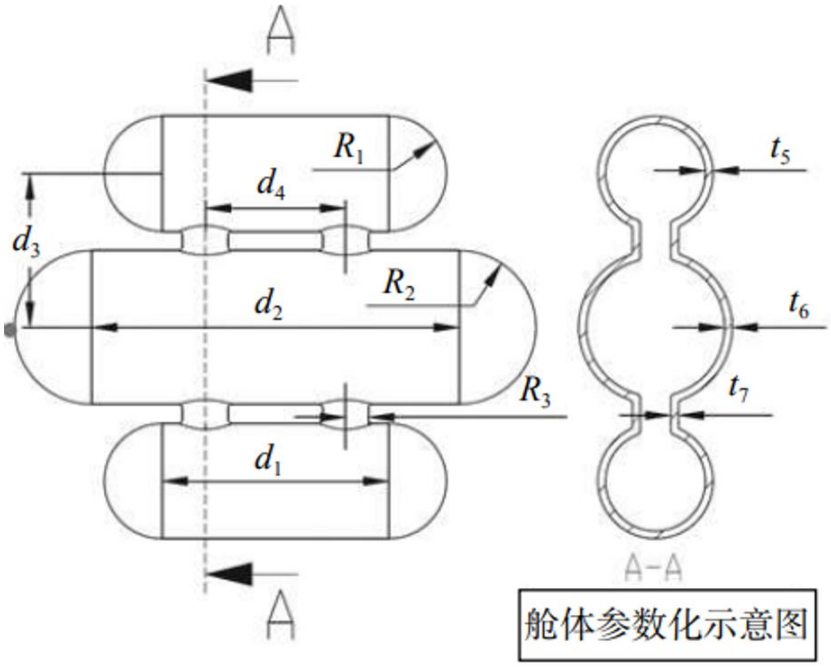


图 3. 某飞行器舱体结构的示意图

问题 3 表 1 是某飞行器结构参数的取值范围，请参照图 1，除表中参数外，耦合结构其他的参数设置如下：$\boldsymbol{l_6}$ 固定为 143 cm，固定机翼半展长为 1 000 cm（翼肋平均分布后共计 8 个，使其中间的 6 个翼肋为 0-1 离散变量，其中 1 表示此处布置有翼肋，0 表示未布置翼肋）。$C_{\boldsymbol{l6}}$ 示机翼 6 个位置处是否布置翼肋

的逻辑值 ，i=1,...,6），机身半展长固定为 500 cm，l_2 取固定值 120 cm, d_1 取固定值 250 cm，d_2 取固定值 350 cm，d_4 取固定值 150 cm。请设计出飞行器的最佳外形，使得所受阻力最小，并给出表 1 中某飞行器结构参数的最优值。

表 1. 某飞行器结构参数的取值范围

设计变量类型	参数	设计下限	设计上限
	$C_{l_6}^i$	0	1
	l_1	270 cm	290 cm
骨架结构设计变量	l_3	0.1	0.35
	l_4	0.45	0.55
	l_5	0.65	0.9
	R_1	65 cm	90 cm
	R_2	75 cm	100 cm
	R_3	20 cm	30 cm
舱体结构设计变量	t_5	8 cm	15 cm
	t_6	8 cm	15 cm
	t_7	8 cm	15 cm
	G_C	350 cm	450 cm

问题 4 图 4 是四种不同圆锥曲线的示意图，包括：圆形、椭圆、抛物线和双曲线。请分别考虑这四种圆锥曲线作为图 1 中飞行器的外形，重新求解问题 3 中飞行器的最佳外形问题，并给出飞行器对应的结构参数。

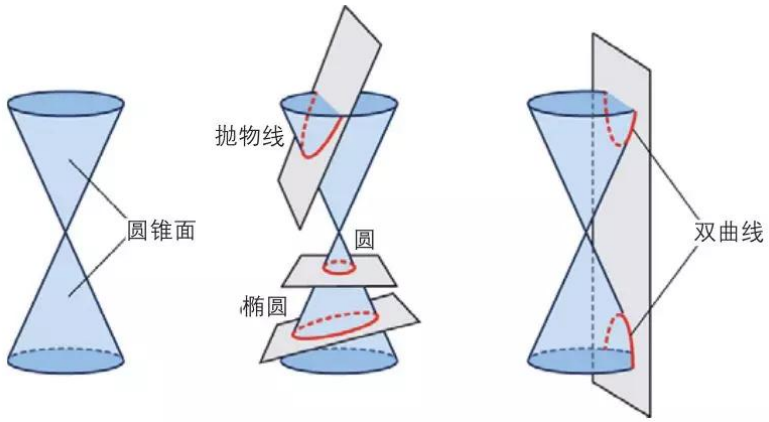


图 4. 四种不同圆锥曲线的示意图

2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛

B题 洪水灾害的数据分析与预测

洪水是暴雨、急剧融冰化雪、风暴潮等自然因素引起的江河湖泊水量迅速增加，或者水位迅猛上涨的一种自然现象，是自然灾害。洪水又称大水，是河流、海洋、湖泊等水体上涨超过一定水位，威胁有关地区的安全，甚至造成灾害的水流。洪水一词，在中国出自先秦《尚书·尧典》。从那时起，四千多年中有过很多次水灾记载，欧洲最早的洪水记载也远在公元前 1450 年。在西亚的底格里斯—幼发拉底河以及非洲的尼罗河关于洪水的记载，则可追溯到公元前 40 世纪。2023 年 6 月 24 日 8 时至 25 日 8 时，中国 15 条河流发生超警洪水。2023 年，全球洪水等造成了数十亿美元的经济损失。

洪水的频率和严重程度与人口增长趋势相当一致。迅猛的人口增长，扩大耕地，围湖造田，乱砍滥伐等人为破坏不断地改变着地表状态，改变了汇流条件，加剧了洪灾程度。在降水多的年份，洪水是否造成灾害，以及洪水灾害的大小，也离不开人为因素，长期以来人为的森林破坏是其重要原因。长江上游乱砍滥伐的恶果是惊人的水土流失。现已达 35 万平方千米，每年土壤浸融量达 25 亿吨。河流、湖泊、水库淤积的泥沙量达 20 亿吨。仅四川一省一年流入长江各支流的泥沙，如叠成宽高各 1 米的堤，可以围绕地球赤道 16 圈。我国第一大淡水湖洞庭湖每年沉积的泥沙达 1 亿多吨，有专家惊呼：“这样下去，要不了 50 年，洞庭湖将从地球上消失！”长江之险，险在荆江，由于泥沙俱下，如今荆江段河床比江外地面高出十多米，成了除黄河之外名副其实的地上河。对森林的肆意砍伐不仅危害自己，而且祸及子孙后代，世界上许多地方，如美索不达米亚、小亚细亚、阿尔卑斯山南坡等由于过度砍伐森林，最后都变成了不毛之地。

附件 train.csv 中提供了超过 100 万的洪水数据，其中包含洪水事件的 id、季风强度、地形排水、河流管理、森林砍伐、城市化、气候变化、大坝质量、淤积、农业实践、侵蚀、无效防灾、排水系统、海岸脆弱性、滑坡、流域、基础设施恶化、人口得分、湿地损失、规划不足、政策因素和发生洪水的概率。

附件 test.csv 中包含了超过 70 万的洪水数据，其中包含洪水事件的 id 和上述 20 个指标得分，缺少发生洪水的概率。附件 submit.csv 中包含 test.csv 中的洪

水事件的 id，缺少发生洪水的概率。

请你们的团队通过数学建模和数据分析的方法，预测发生洪水灾害的概率，解决以下问题：

问题 1. 请分析附件 train.csv 中的数据，分析并可视化上述 20 个指标中，哪些指标与洪水的发生有着密切的关联？哪些指标与洪水发生的相关性不大？并分析可能的原因，然后针对洪水的提前预防，提出你们合理的建议和措施。

问题 2. 将附件 train.csv 中洪水发生的概率聚类成不同类别，分析具有高、中、低风险的洪水事件的指标特征。然后，选取合适的指标，计算不同指标的权重，建立发生洪水不同风险的预警评价模型，最后进行模型的灵敏度分析。

问题 3. 基于问题 1 中指标分析的结果，请建立洪水发生概率的预测模型，从 20 个指标中选取合适指标，预测洪水发生的概率，并验证你们预测模型的准确性。如果仅用 5 个关键指标，如何调整改进你们的洪水发生概率的预测模型？

问题 4. 基于问题 2 中建立的洪水发生概率的预测模型，预测附件 test.csv 中所有事件发生洪水的概率，并将预测结果填入附件 submit.csv 中。然后绘制这 74 多万件发生洪水的概率的直方图和折线图，分析此结果的分布是否服从正态分布。

附件：

1. train.csv
2. test.csv
3. submit.csv

2024年第十四届APMCM亚太地区大学生数学建模竞赛

C题 基于量子计算的物流配送问题

随着电子商务的迅猛发展，电商平台对物流配送的需求日益增长。为了确保货物能够按时、高效地送达消费者手中，电商平台与第三方物流公司建立了紧密的合作关系。然而，面对大量的货物和多样的目的地，如何制定合理的运输策略成为了物流公司面临的一大挑战。

传统的物流优化方法在应对复杂的运输需求时往往具有较高的复杂度。为了解决这个问题，物流公司希望借助量子计算技术来自动计算运输综合策略，从而可以更合理地规划运输路线、选择合适的运输方式和工具，并确保在规定的时间内将货物送达目的地。这样不仅能够提高物流效率，降低运输成本，还能够提升消费者对电商平台的满意度。

量子计算，尤其是相干伊辛机(Coherent Ising Machine, CIM)，在处理复杂优化问题方面展现出了巨大的潜力。由于其和相干伊辛机的紧密联系，QUBO（Quadratic Unconstrained Binary Optimization）模型构成了量子计算中的一类核心问题。QUBO模型是一种适配相干伊辛机（CIM）的模型，其形式为

$$\min x^T Q x,\ x \in \{0,1\}^n$$

其中Q为$n*n$的系数矩阵。本赛题主要基于物流配送的场景，通过将问题建模为QUBO形式，使用Kaiwu SDK完成对问题的求解。Kaiwu SDK是一套基于相干伊辛机求解QUBO模型的软件开发套件，可以通过访问下述链接(https://developer.qboson.com)来获取Kaiwu SDK。

一、货物当前所在城市

公司一：

城市	数量（吨）
上海	19
西安	20
郑州	12

公司二：

城市	数量（吨）
上海	14
西安	24
郑州	18

二、货物需要运往的城市

小组一：

货物类型	货物数量(吨)	终点城市
普货	19	昆明
普货	25	深圳
普货	7	天津

小组二：

货物类型	货物数量(吨)	终点城市
普货	27	昆明
普货	10	**深圳**
普货	19	天津

三、卡车

市场上有12吨载重卡车和5吨载重卡车两种卡车供租赁，租金分别为每天5000元和每天3000元。这些卡车可以在任何地方租赁，并且客户可以根据自己的需求选择合适的卡车类型和租赁时长。假设每个城市有足够数量的可供租赁的卡车。

四、卡车运输

为节约成本，各小组间可在任何一个城市拼货和中转运输。拼货是指将来自不同发货人的货物合并在一起，由同一辆车或同一批运输工具运送的方式。同时货物可以由一辆车辆全部或者部分卸货后，暂存在当前城市，待后续的车辆将其运走。通过拼货可以更加灵活地安排运输计划，提高运输效率，降低成本，优化运输路线和运输资源的利用。

卡车单趟时间 单位：天

	上海	西安	昆明	深圳	天津	郑州
上海	__	3	6	4	3	2
西安	3	__	4	4	2	1
昆明	6	4	__	2	6	5
深圳	4	4	2	__	5	4
天津	3	2	6	5	__	1
郑州	2	1	5	4	1	__

普通卡车 A(12 吨)单趟成本 单位：元

	上海	西安	昆明	深圳	天津	郑州
上海	__	13500	26500	16500	13500	8500
西安	13500	__	18500	20500	10500	3500
昆明	26500	18500	__	7500	31500	24500
深圳	16500	20500	7500	__	25500	17500
天津	13500	10500	31500	25500	__	5500
郑州	8500	3500	24500	17500	5500	__

普通卡车 B(5 吨)单趟成本 单位：元

	上海	西安	昆明	深圳	天津	郑州
上海	__	11000	24000	14000	11000	6000
西安	11000	__	16000	18000	8000	1000
昆明	24000	16000	__	5000	29000	22000
深圳	14000	18000	5000	__	23000	15000
天津	11000	8000	29000	23000	__	3000
郑州	6000	1000	22000	15000	3000	__

五、航空运输

除了陆路运输，货物也可以通过航空运输，题目中考虑的任意两个城市之间都可以通过航空运输进行货物运输。为简化计算，假设货物的国内航空运价无论远近均为 10000 元/吨，当日到达。

如果模型的比特数较高，可以尝试使用SubQUBO等方法进行求解（参考附件1）。提出创新性的算法和解决方案是一个加分项。附件2为供参考的QUBO建模教程。

问题一 假设这两个物流公司独立运营，拼货只发生在公司内部。你的任务是以最小化单个物流公司的运营成本为目标，建立QUBO模型，使用Kaiwu SDK中的CIM模拟器和模拟退火求解器分别求解，为两个物流公司分别设计货车租赁方案和货物运输方案。

问题二 当这两个物流公司之间合作运营时，公司之间可以拼货运输，此时的优化目标为最小化两个公司的总成本。请使用Kaiwu SDK中的CIM模拟器和模拟退火求解器求解，给出最优的货车租赁方案和货物运输方案，以及合作运营带来的总体成本减少量。

问题三 请你自行提出一个具有商业化前景或学术价值的场景。场景可以涉及AI，通信，金融，生物医学，物流供应链管理等相关领域。你需要给出相应的QUBO模型表达式，并计算模型所需的比特数量级(可以用相关参数表示)。附件中3-附件6为覆盖多个场景的参考论文。

优　秀　论　文

亚太赛组委会制作

选题	2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
A		apmcm 25201758

基于最优成本方案的农业灌溉系统的优化

摘要

农业灌溉系统的优化设计对水资源高效利用与作物增收至关重要，本文考虑了土壤湿度条件、水源与农场的地理位置、管道建设成本和储水罐容量成本以及不同农作物在不同生长阶段下的需水量等因素，围绕灌溉系统的设计展开研究。

针对问题一：土壤湿度随着时间会有相应的变化，当天灌溉系统的灌溉量可以根据土壤湿度进行有效调整，对土壤湿度的预测基于当地气象站的气象数据。首先对气象数据进行清洗及缺失值、异常值处理；同时，对这些变量通过 **Lilliefors 检验**，发现不满足正态性，因此采用 **Spearman rho 检验**分析变量间相关性，确定与之相关的气象变量有 T、Po、P、U、RRR。其次，构建**多元线性回归模型**，依次通过回归方程显著性检和残差项正态性检验，但发现模型预测效果较差；在此基础上，进一步引入 **XGBoost 模型**，通过性能指标和预测模型拟合图对比显示，XGBoost 模型的 **R^2**从多元线性回归的 **0.2229 提升至 0.6134**，且 MAE 和 RMSE 更低，拟合效果显著更优。最后，依据题目给出的当天各时段气象数据，采用 XGBoost 模型预测出**当天土壤湿度 5cm_SM 值为 0.2428**。

针对问题二：对灌溉系统的布局规划要考虑到布局的成本，灌溉成本来源于布局管道的长度、水管日流量和储水罐的容量。将解决方案分为三步：首先，通过**几何覆盖理论**开展喷头布局规划，通过微调角度α和行列排列确定喷头布局。其次，构建**管道目标规划模型**，得到当**主管道的数量与喷头列数一致时所花费的成本最小**。最后，依据蓄水罐行数与河水引流行数之间的成本关系，构建**灌溉系统布局规划模型**，得到储水罐的分布和河水管道的布局最优：采用 **25 个喷头、5 行 5 列**的布局，其中 **5 列管道由河流供水至第 3 行**，剩余 **10 个喷头通过储水罐供水**，水罐设计容量为 **35625.66L**，布局情况如图 4-6 所示，最小建设成本为 **2498502.10 元**。

针对问题三：在旱灾场景（河流供水量降至问题二最优灌溉系统总流量的 80%）下，首先要确定水量分配方法，根据喷头位置将**农田划分为 5 个灌溉区域**，通过调整各区域供水**优先级迭代 120 种水量分配方案**。其次，结合七月内动态变化的土壤湿度和不同时期作物需水情况，计算各种分配方案下作物存活和正常生长情况，得出以问题二设计的灌溉系统作物**最大正常生长面积为 0.538429 公顷，能存活面积为 0.868977 公顷**，不同作物对应的正常生长面积和存活面积见表 5-1。最后，建立旱灾概率与储水罐应急储备比例的数学公式，明确不同概率下的储备需求，为抗旱决策提供依据，储水应急储备比例见表 5-2。

针对问题四：在实际情况中农作物的收成将决定最后的收入，为了满足各农作物正常生长的需求，需要在问题二的基础上对灌溉系统进行优化布局。解决方案分为三步：首先，测试问题二的灌溉系统是否满足要求。其次，对每个月分别求出日总流量最大需求，以此为基准，构建**改进的灌溉系统布局规划模型，**对喷头布局再进行最优规划。再次，以成本最低为目标，确定蓄水罐行数与河水引流行数，**得到最优布局如图 6-2**；最后，通过这个布局对灌溉系统的喷头水量合理分配，得到三个月的**灌溉安排表如表 6-1，其灌溉量和水源比例可视化结果见图 6-3、6-4。**

关键词：XGBoost 模型；几何覆盖理论；管道目标规划模型；灌溉系统布局规划模型

一、问题重述

1.1 问题背景

农业灌溉系统的优化设计对水资源高效利用与作物增收至关重要，构建科学合理的优化模型是实现这一目标的核心。该系统需应对不同作物的需水特性、动态气象条件、土壤湿度波动，以及灌溉设施布局与成本等复杂变量。当前，精准筛选关键变量并展开优化，平衡河水引水与储水罐灌溉的关系，优化管道布局以降低建设成本，同时根据作物特性和环境变化动态调整灌溉策略，成为农业灌溉系统设计面临的重要挑战。

1.2 问题重述

基于 1.1 问题背景，现本文需要解决以下四个问题：

问题一：建立一个预测模型，探究土壤湿度 5cm_SM 与其他除土壤湿度以外的气象数据之间的关系，并根据表 2 中某天不同小时的气象数据预测当天土壤湿度。

问题二：在仅考虑土壤湿度、满足作物存活的前提下，利用 2021-07-01 至 2021-07-31 的数据，规划引水管布线与储水罐的容积、位置，给出灌溉系统布线规划图及建设花费，以实现总花费最小化。

问题三：在问题二的基础上，当旱灾来临时河流供给水量下降至问题二中引水系统总流量的 80%，确定储水罐和引水管布线设计能保证多少作物存活、最多能保证多少作物正常生长，并分析储水罐应急储备水源比例与旱灾概率的关系。

问题四：假设各作物成熟期均为 20 天，使用 2021-05-01 至 2021-07-31 的数据计算并规划不同作物每月的灌溉方案，判断之前建立的系统是否能满足灌溉需求，若不能则修改系统布线，将月灌溉情况可视化并填写灌溉安排表。

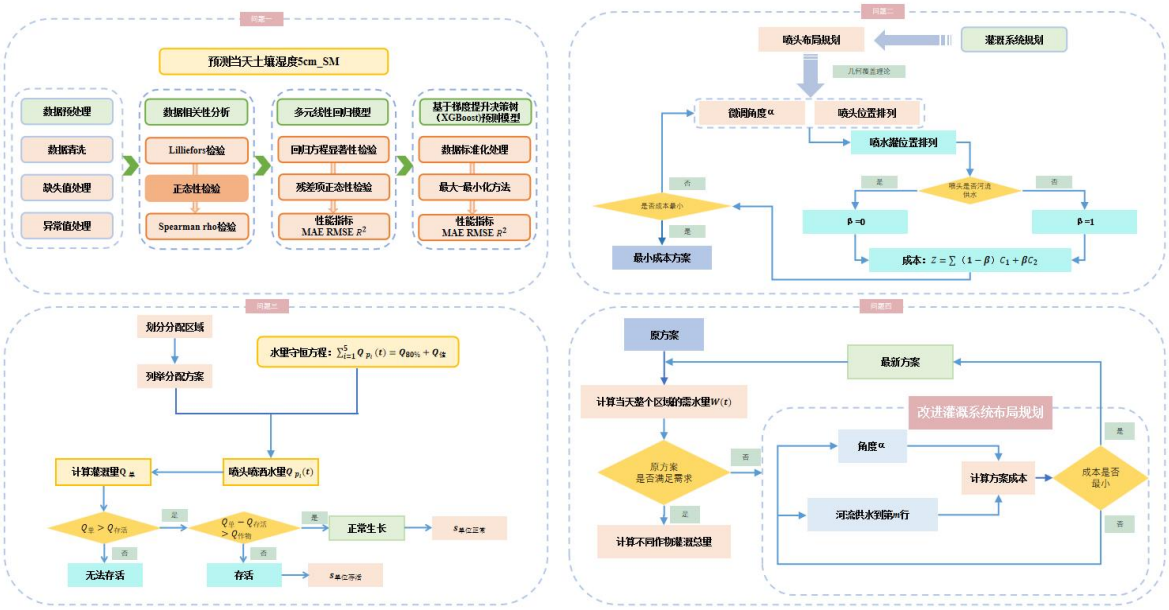


图 1-1 总体技术流程图

二、模型假设与符号说明

2.1 模型假设

在模型的建立与求解过程中，为了简化问题、排除无关因素的干扰，做出以下假设：

假设 1：附件中提供的模拟数据是可靠的。

假设 2：风力对喷灌没有影响，适量超出作物需水量不影响作物存活。

假设 3：储水灌的水来自于外部。

假设 4：适量超出作物需水量不影响作物存活。

假设 5：喷头均匀喷洒。

2.2 符号说明

表 2-1 符号说明

符号	含义
Y_{SM}	当天的土壤湿度 5cm_SM
x_T	当天地面上 2 米处的大气气温
x_{Po}	当天气象站水平的大气压
x_P	当天平均海平面的大气压
x_U	当天地面高度 2 米处的相对湿度
x_{RRR}	当天降水量
r	喷头覆盖半径
a	喷头列数
b	喷头行数
$\theta_{\min}$	指定时间最小土壤湿度值
θ_{req}	植物在生长期间的最低土壤湿度
$V_{水}$	当天作物单位面积最大需水量
$S_{单喷头}$	单个喷头覆盖面积
$Q_{单喷头}$	单个喷头需水量
C_L	管长成本
C_Q	流量成本
C_1	管道总成本
C_2	储罐总成本
$Q_{80\%}$	河流当天可供水量
$Q_{存活}(t)$	当天可供单位面积作物存活的水量
$Q_{储}$	单个储水罐当天可供水量
$Q_{应急}$	应急水源储备量
Q_h	可正常供水量
K	应急储备水源比例
$Q_{P_i}(t)$	单位面积当天各喷头覆盖区域的喷头供水量
$Q_{单}(t)$	单位面积灌溉量

*其余未说明的符号将在正文详细展示。

三、问题一：模型建立与求解

3.1 问题分析

根据问题一的要求，本题围绕当天土壤湿度 5cm_SM 预测展开研究并构建模型。具体思路如下：

第一步：首先进行数据预处理，通过数据清洗剔除无效数据，对缺失值和异常值进行合理处置，为后续分析提供高质量数据基础。

第二步：进行数据相关性分析，先采用 **Lilliefors 检验**进行正态性检验，若检验未通过，则运用 **Spearman rho 检验**分析土壤湿度 5cm_SM 与 15 个气象变量间的相关性，进而明确关键预测因子。

第三步：构建**多元线性回归模型**，依次进行回归方程**显著性检验**和残差项**正态性检验**，通过 MAE、RMSE 和R^2等性能指标评估模型性能。

第四步：引入 **XGBoost 集成学习算法**，采用**最大-最小归一化**对数据进行预处理，以消除量纲影响并提升模型稳定性；通过正则化参数控制树结构复杂度，利用网格搜索优化超参数配置。

第五步：通过对比多元线性回归模型和 XGBoost 模型性能，筛选最优预测方案用于当天土壤湿度 5cm_SM 预测。

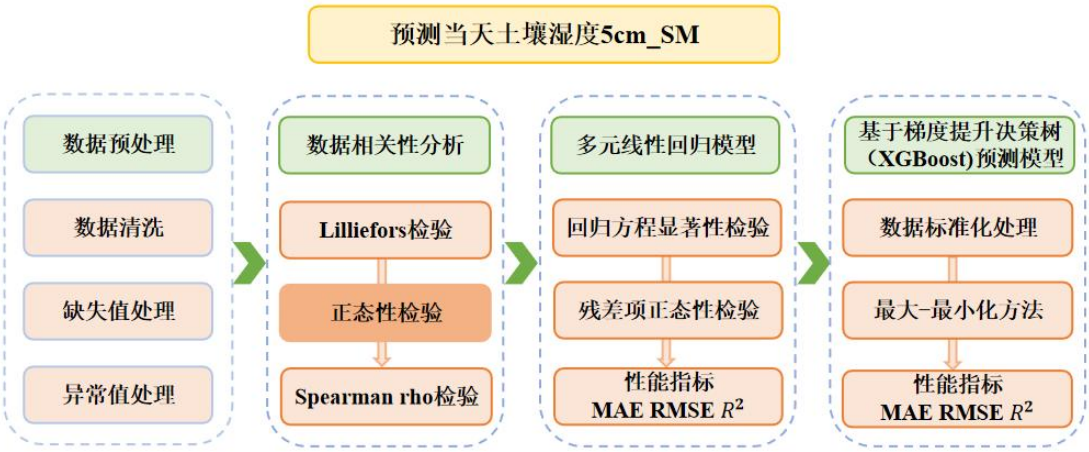


图 3-1 问题一流程图

3.2 数据预处理

Step1：数据清洗

分析“降水量等气象数据.xlsx”表格，从原始数据中删除不完整部分，保留 T、Po、P、Pa、U、DD、Ff、ff3、Tn、Tx、H、VV、Td、RRR、tR 共 15 个变量。由于 2021-9-16-20:00 至 2021-12-31-23:00 的气压趋势 Pa 数据缺失，后续分析均使用 2020-8-7-02:00 至 2021-9-16-17:00 的数据，以研究土壤湿度 5cm_SM 与上述 15 个变量的关系并实现预测。

Step2：缺失值处理

上述时间范围内存在 5 个缺失的时间点，对其直接删除；针对已有时间点中各变量的缺失值如图 3-2 所示，分别采用直接删除法和线性插补法进行补充。

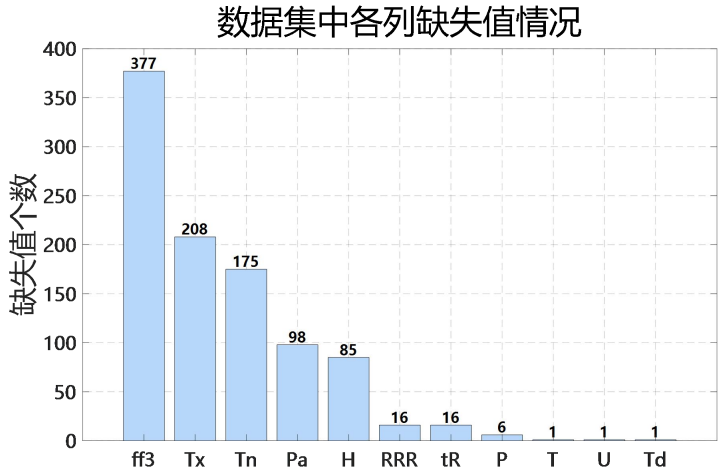


图 3-2 数据集中各列缺失值情况

上图 3-2 可知，地面高度 2 米处的大气气温（T）、相对湿度（U）、露点温度（Td）均仅存在 1 个缺失值，因此可采用直接删除法进行处理；其余 8 个变量的缺失值数量较多，可采用线性插补法进行处理。

Step3：异常值处理

在以上数据异常值处理中，对连续变量采用 Z-Score 方法识别异常值，再通过窗口为 3 的移动平均插值法进行修正，经检验后数据均满足条件。对于离散变量，风向（DD）采用独热编码转换为 0（无风）、1（从西方吹来的风）至 16（从西北偏西方向吹来的风）的数值，最低云层底部的高度（H）和达到规定降水量的时间（tR）则保留其离散属性。

3.3 数据相关性分析

在探究土壤湿度 5cm_SM 与 T、Po、P、Pa、U、DD、Ff、ff3、Tn、Tx、H、VV、Td、RRR、tR 这 15 个气象变量的关系前，需先对数据进行正态性分布检验。本研究选用 Lilliefors 检验，检验结果显示正态性检验未通过。鉴于正态性假设不满足，无法采用基于正态分布的相关分析方法。为有效探究土壤湿 5cm_SM 与这 15 个气象变量间的关系，包括各变量单独效应及可能的交互作用，转而采用 Spearman rho 相关系数进行分析，以评估变量间的相关性。Spearman rho 相关系数属于非参数统计方法，不依赖数据的分布形态，适用于本研究数据的分析场景。

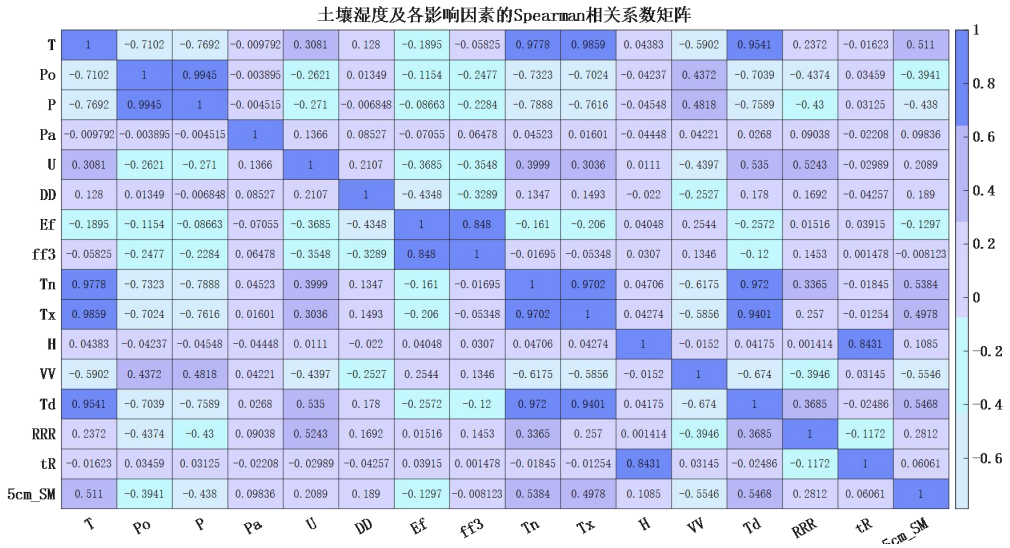


图 3-3 土壤湿度及各影响因素的 Spearman rho 相关系数矩阵

如图 3-3 可知，土壤湿度 5cm_SM 与 T、P、Tn、Tx、Td 的相关系数绝对值处于 0.4 至 0.6 之间，呈现中度相关性特征，表明上述气象要素对土壤湿度存在一定程度的影响；而 Pa、DD、Ef 与土壤湿度 5cm_SM 的相关系数绝对值小于 0.2，相关性较弱，说明其对土壤湿度的直接作用有限。

基于上述分析，本研究选土壤湿度 5cm_SM 作为因变量，T、Po、P、U、RRR 为自变量，构建预测模型来探究气象要素对土壤湿度的影响。

3.4 多元线性回归模型

为客观评价模型的有效性，本章对数据集进行随机打乱后，按 80%、20%比例划分为训练集与测试集，前者用于模型参数优化训练，后者用于评估预测性能。

为预测当天的土壤湿度 5cm_SM，在完成数据相关性分析后，构建了一个多元线性回归方程模型，具体形式如下：

$$Y_{SM} = \beta_0 + \beta_1 x_T + \beta_2 x_{Po} + \beta_3 x_P + \beta_4 x_U + \beta_5 x_{RRR} \quad (3.1)$$

其中，输出变量Y_{SM}表示当天的土壤湿度 5cm_SM，输入变量$x_T, x_{Po}, x_P, x_U, x_{RRR}$分别表示当天地面上 2 米处的大气气温、气象站水平的大气压、平均海平面的大气压、地面高度 2 米处的相对湿度、降水量。

表 3-1 回归方程显著性检验

模型		平方和	自由度	均方	F	显著性
当天的土壤湿度 5cm_SM	回归	0.75523	5	0.15105	**324.3599**	**2.0362e-62**
	残差	0.53791	339	0.0015868		
	总计	1.2931	344	0.0037591		

如表 3-1 所示的方差分析结果及检验结果显示，当天的土壤湿度 5cm_SM 的 F 值为 324.3599，显著性为 2.0362e-62，拒绝原假设H_0，说明土壤湿度 5cm_SM 的预测模型有效。由此可见，方程具有显著性，表明自变量对因变量具有显著影响。

在多元线性回归模型的应用过程中，对残差项进行正态性分布检验是至关重要的环节，这有助于验证模型假设是否成立，进而保障模型的有效性。

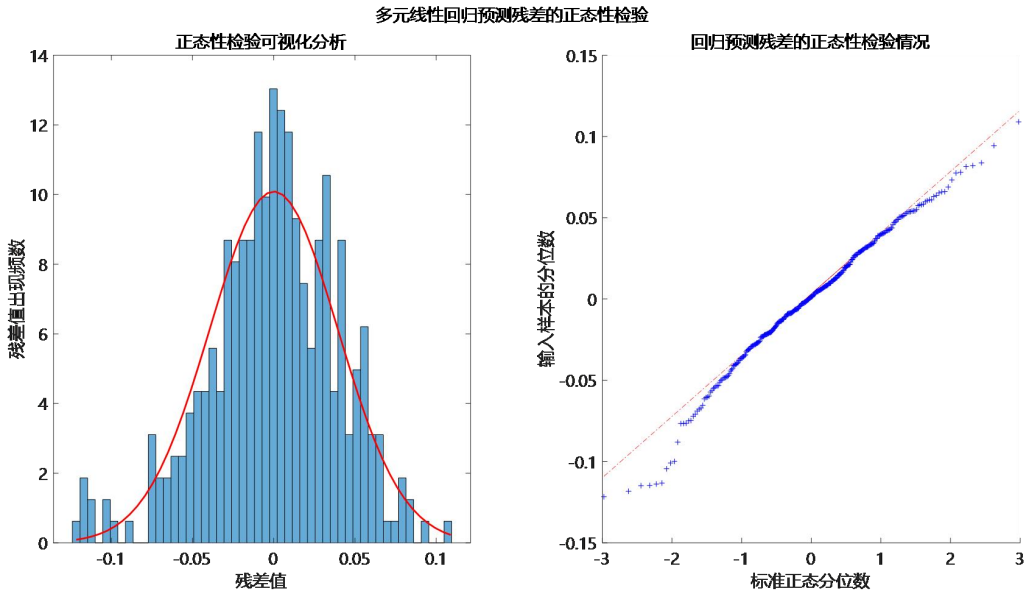


图 3-4 多元线性回归预测模型的直方图和 Q-Q 图

由图 3-4 可知，在左侧直方图中，残差分布大致以 0 为中心呈对称形态，且与正态分布曲线（红色曲线）较为接近，这在一定程度上表明残差服从正态分布。在右侧 Q-Q

图中，大部分点较为接近直线，虽然在两端存在一定偏离，但整体趋势符合正态分布特征。

这表明模型残差符合正态分布特征，满足多元线性回归对残差的基本假设。不过需要注意的是，尽管残差通过了正态性检验，模型的实际误差仍然较大。

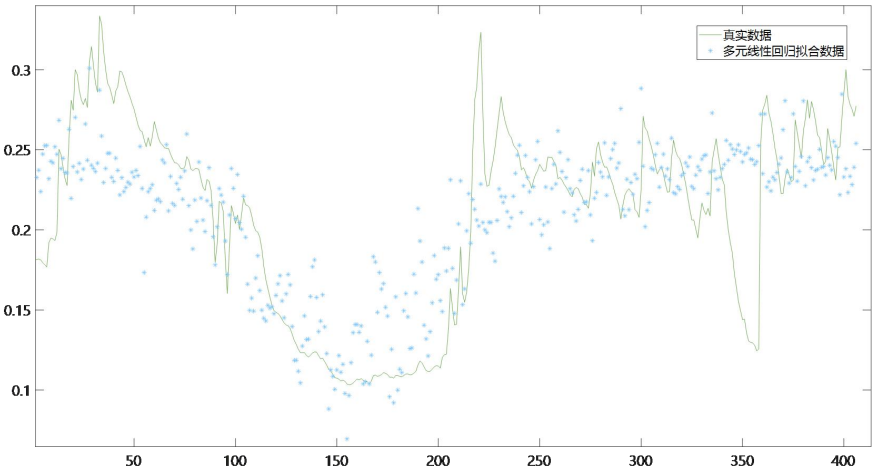


图 3-5 多元线性回归模型拟合图

从图 3-5 能够直观地看出，多元线性回归模型的拟合数据与真实数据之间存在着一定程度的偏差，尤其在部分区域，数据散点显著偏离了绿色的拟合线，这一现象表明该模型的预测性能并非十分理想。

表 3-2 多元线性回归模型的性能指标

模型	性能指标	MAE	RMSE	R^2
多元线性回归模型	训练集	0.0306	0.0395	0.5840
	测试集	0.0351	0.0439	**0.2229**

结合表 3-2 多元线性回归模型的性能指标进一步分析，测试集的R^2值仅为 0.2229，这一结果表明模型的预测误差较大。

综上，多元线性回归模型在处理当前数据时存在显著局限性，其可能无法充分捕捉数据中潜在的非线性关系或复杂模式，从而导致预测结果与实际值之间产生较大偏差。

3.5 基于梯度提升决策树（XGBoost）预测模型

XGBoost（Extreme Gradient Boosting）作为梯度提升决策树的优化实现，通过集成多棵弱学习器构建强预测模型，在处理非线性关系和复杂特征交互方面表现出显著优势。其预测模型的数学表达为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3.2}$$

其中，K表示树的总个数，f_k表示第k棵树，$\hat{y}_i$表示样本x_i的预测结果。

为平衡预测精度与模型复杂度，XGBoost 采用正则化目标函数：

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.3}$$

其中，y_i表示样本x_i的实际观测值，$l(y_i, \hat{y}_i)$为样本x_i的训练误差，$\Omega(f_k)$为正则化项，其表达式为

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \| w_k \|^2 \tag{3.4}$$

其中，T_k为第k棵树的叶子节点数，w_k为叶子节点权重向量，γ和λ为正则化参数，用于控制模型复杂度。。

为消除特征量纲差异对模型性能的影响，采用最大-最小归一化方法对数据进行标准化处理，具体公式如下：

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.5}$$

其中，x为原始数据集中的某个特征值（未归一化的值），x'为经过归一化处理后的特征值（归一化后的值），$\min(x)$为特征值x在整个数据集中的最小值，$\max(x)$为特征值x在整个数据集中的最大值。

利用训练数据所构建的模型，对测试数据进行预测并输出结果。本文选取均方根误差（Root Mean Square Error, RMSE）、平均绝对误差（Mean Absolute Error, MAE）和决定系数（R-Square, R^2)）作为模型的评价指标，相关计算公式如下：

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \\ MAE &= \frac{1}{N}\sum_{i=1}^{N}|(y_i - \hat{y}_i)| \\ R^2 &= 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \end{aligned} \tag{3.6}$$

式中，N表示样本总数，其中$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$。

在本研究中，XGBoost 模型的参数设置如下：决策树数目K为 100；最小叶子数为 20，以此来调节模型的拟合效果；学习率为 0.4；正则化参数γ=0.3 和λ=0.3；特征采样比例为 0.9，这一设置有助于降低模型的过拟合风险。有助于减少模型的过拟合风险。

通过这样的参数设置，XGBoost 模型能够在保证一定拟合能力的同时，有效地规避过拟合问题，从而在后续的预测任务中展现出良好的性能。

表 3-3 多元线性回归模型和 XGBoost 预测模型的性能指标

模型	性能指标	MAE	RMSE	R^2
多元线性回归模型	训练集	0.0306	0.0395	0.5840
	测试集	**0.0351**	**0.0439**	**0.2229**
XGBoost 预测模型	训练集	0.0146	0.0189	0.9047
	测试集	**0.0146**	**0.0191**	**0.6134**

由表 3-3 的性能指标可知，XGBoost 预测模型测试集的 MAE 为 0.0146、RMSE 为 0.0191，均低于多元线性回归模型的 0.0351 和 0.0439，而 MAE 和 RMSE 数值越小，表明模型的预测误差越小；同时，XGBoost 预测模型的R^2为 0.6134，高于多元线性回归模型的 0.2229，R^2越接近 1，说明模型拟合效果更好。故 XGBoost 预测模型优于多元线性回归模型。

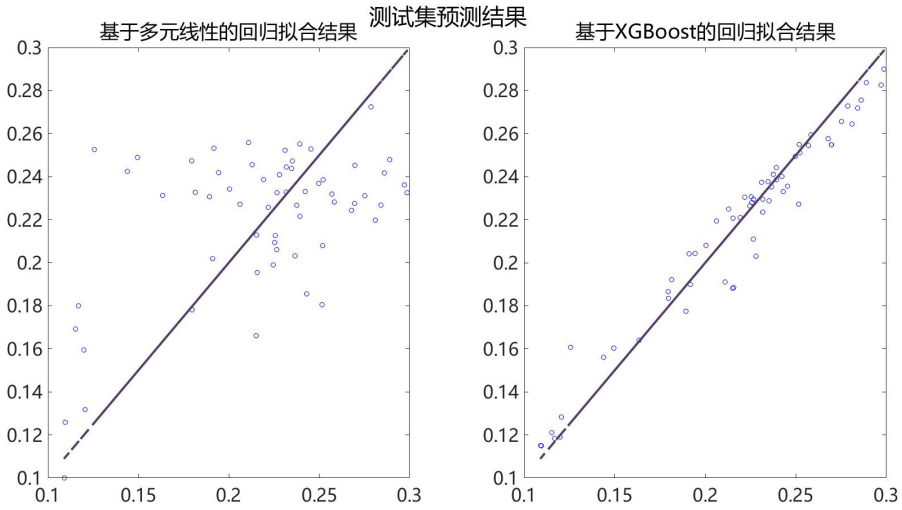


图 3-6 多元线性回归模型和 XGBoost 预测模型测试集拟合图

由图 3-6 可知，多元线性回归模型拟合图中数据点离散程度大，进一步表明模型拟合效果与预测性能欠佳。而 XGBoost 预测模型拟合图中数据点紧密聚集在拟合线附近，离散程度小，说明模型拟合效果与预测性能更优。

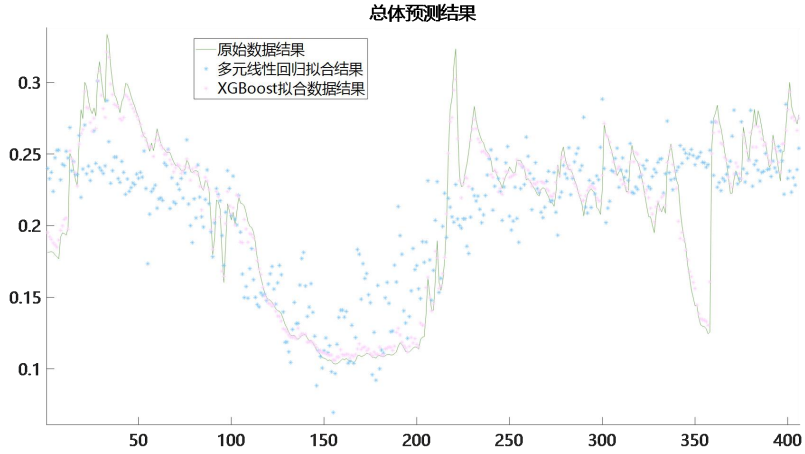


图 3-7 多元线性回归模型和 XGBoost 预测模型拟合图

从图 3-7 可以看出，多元线性回归模型的预测结果（蓝色散点）与实际值的偏差较大，尤其是在数据波动较大的区域，如低谷和高峰部分，散点明显偏离了实际值的趋势。而 XGBoost 预测模型的预测区间（粉色区域）则相对更紧密地围绕在实际值附近，尤其是在数据波动较大的区域，XGBoost 模型的预测区间能够更好地捕捉到实际值的变化趋势。

综上，XGBoost 预测模型在测试集上的拟合效果优于多元线性回归模型，能更精准地预测目标变量。故问题一采用 XGBoost 预测模型，得到土壤数据预测表如下：

表 3-4 土壤数据预测表

时间（h）	T	Po	P	Pa	U	DD	Ff	RRR	预测的当天湿度 5cm_SM
02	19.3	731.5	751.7	1.0	99	西	轻风	15.0	**0.2428**
05	20.0	732.0	752.4	0.5	94	西南	轻风	6.0	
08	23.4	732.8	753.2	0.8	80	西	轻风	0	
11	28.0	733.5	753.8	0.7	44	西北	轻风	0	

*根据中国国家气象局发布的《风力等级》国家标准（GB/T 19201-2006），轻风指的是风力等级为 2 级的风，其对应的风速范围是 1.6-3.3 米/秒。

四、问题二：模型建立与求解

4.1 问题分析

根据问题二的要求，本问题旨在以最小成本完成灌溉系统的规划，保障作物存活。具体思路如下：

第一步：借助几何覆盖理论开展喷头布局规划。先微调角度α并进行喷头位置排列，同时结合引水管道布线、储水罐容积与位置进行灌溉系统规划，确定喷水罐位置排列。

第二步：判断喷头的供水方式。若喷头由河流供水，则按对应方式计算成本；若由储水罐供水，则按另一种对应方式计算成本，然后通过成本公式计算总成本。

第三步：不断优化喷头布局和供水方式选择，判断是否达到成本最小。若未达到，继续调整相关参数重新计算；若达到，则得出满足条件的总花费最小的灌溉系统布线规划图及建设花费。

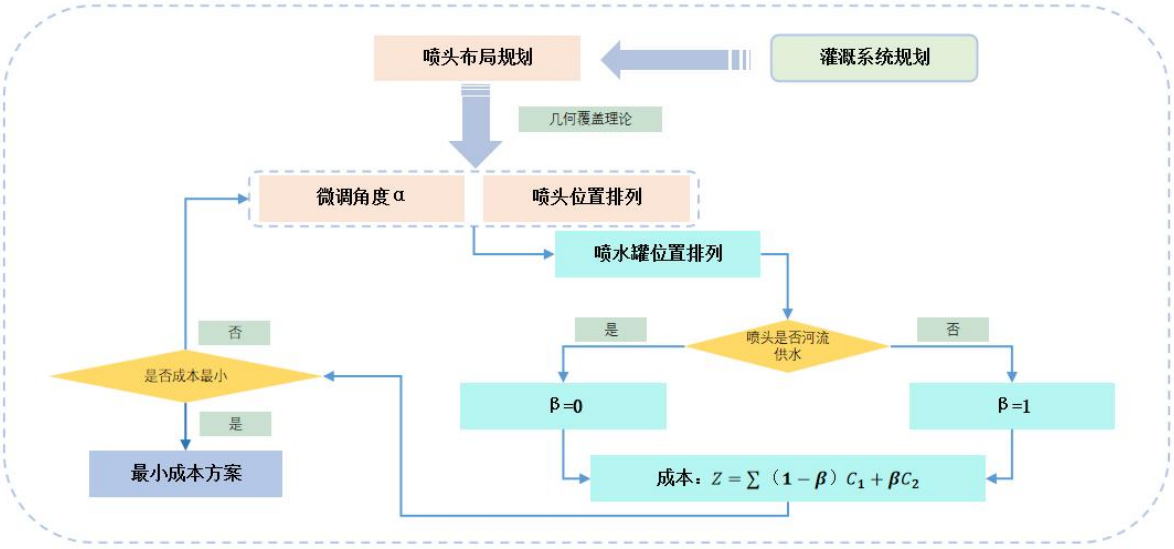


图 4-1 问题二流程图

4.2 喷头全覆盖模型

4.2.1 几何覆盖原理

基于几何覆盖原理，为确保灌溉区域无遗漏，喷头需按矩形网格排列，列数 a 和行数 b 需满足覆盖区域的长和宽要求。

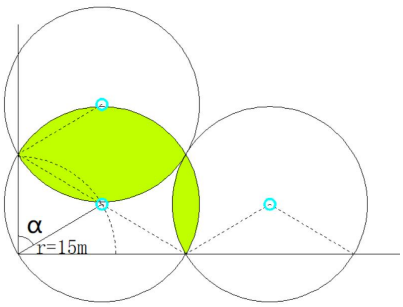


图 4-2 喷灌覆盖区域简化示意图

1.**覆盖重叠：**邻喷头的覆盖圆存在交集（绿色区域），确保无盲区。

2.**角度影响：**角度α决定覆盖效率，确定喷头位置坐标。角度过大则横向间距增大，可能出现盲区；角度过小则重叠过多，造成资源浪费。

3.**优化目标**：通过调整α，使覆盖圆在横向和纵向的投影间距均≤$2r$（即 30m），实现全区域无缝覆盖。

4.2.2 喷头行列个数计算

本文某农场占地面积 1 公顷，农场为正方形，即可求得灌溉区域的边长为 L =100m。

列数 a 满足纵向覆盖数：

$$a=\left\lceil \frac{L}{2r\sin\alpha} \right\rceil=\left\lceil \frac{100}{2\times15\sin\alpha} \right\rceil \tag{4.1}$$

行数 b 满足横向覆盖数；

$$b=\left\lceil \frac{L}{2r\cos\alpha} \right\rceil=\left\lceil \frac{100}{2\times15\cos\alpha} \right\rceil \tag{4.2}$$

4.3 需水量计算

4.2.1 单位面积需水量

从附件 2021 年 7 月 1 日至 2021 年 7 月 31 日得出最小土壤湿度值为$\theta_{min}=0.124$，而本文给出的植物在生长期间的最低土壤湿度 5cm_SM 为$\theta_{req}=0.22$。为保证当天作物水量足够，则需按当天作物最大需水量来求解。

土壤湿度 5cm_SM 差：

$$\Delta\theta=\theta_{req}-\theta_{min}=0.22-0.124=0.096 \tag{4.3}$$

每平方米 5cm 厚度土壤的质量：

$$m_{土}=m_{\mathrm{d}}\cdot S\cdot h=1500\times1\times0.05=75kg \tag{4.4}$$

式中，$m_{\mathrm{d}}=1500kg/m^3$、$S=1m^2$、$h=0.05m$。

当天作物单位面积最大需水量：

$$m_{水}=m_{土}\cdot\Delta\theta=75\times0.096=7.2kg \tag{4.5}$$

又由于$\rho_{水}=1kg/L$，所以当天作物最大需水量为$V_{水}=7.2L$。

4.2.2 单个喷头需水量

单个喷头覆盖面积为半径 15m 的圆面积：

$$S_{单喷头}=\pi r^2=\pi\times15^2\approx706.86m^2 \tag{4.6}$$

单个喷头需水量：

$$Q_{单喷头}=S_{个喷头}\times7.2\approx706.86\times7.2=5089.39L \tag{4.7}$$

为保障每天水量充足，按最大需水量计算，即每日需水量按上述单个喷头需水量执行，不考虑降雨等补水因素。

4.3 灌溉系统布线与供水方案

4.3.1 管道目标规划模型

在分析管道布线时，可将其视为连接喷头位置的网络优化问题。对于具有多层结构的系统，第一层直接由河流供水，后续每层均由上一层供水。为简化分析，可先将多层模型降维为单层情况，即研究一层有a个喷头时，确定主管道数量n的最优解以实现成本最低。

图 4-3 主管道布局简化示意图

管长成本C_L：

$$C_L = 50L^{1.2} = 50\times\left[(a-1)\times 2r\sin\alpha + nr\cos\alpha\right]^{1.2}$$
$$= 50\times\left[(a-1)\times 30\sin\alpha + 15n\cos\alpha\right]^{1.2} \quad (4.8)$$

式中，$a = \left\lceil \frac{100}{2\times 15\sin\alpha} \right\rceil$。

流量成本C_Q：

$$C_Q = C_{Q_1} + C_{Q_2} \quad (4.9)$$

式中，主管道流量成本为C_{Q_1}，分管道流量成本为C_{Q_2}。各段管道的流量是基于基尔霍夫电流定律和流体力学原理计算。

管道总成本C_1：

$$\min C_1 = C_L + C_Q \quad (4.10)$$

$$s.t.\begin{cases} 4 \le a \le 6 \\ n \le a \\ 33.8^\circ \le \alpha \le 56.2^\circ \end{cases} \quad (4.11)$$

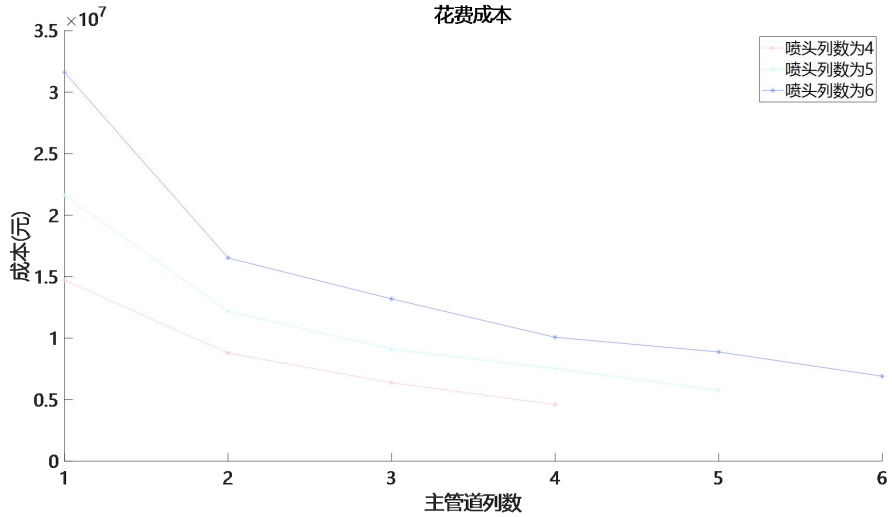


图 4-4 单层主管道数量和管道总成本折线图

图 4-4 直观呈现了在喷头列数分别为 4、5、6 的情况下，随着主管道数量增加，系统总成本呈下降趋势，明确了是纵向主管道供水，即当主管道数量等于喷头列数时（$n = a$），对应喷头列数下的总成本达到最低。

4.4 灌溉系统布局规划模型

计算总成本并得出**最优供水方案**的思路如下：

查阅资料可知，农业灌溉储水供应普遍以维持 7 天使用为主，所以储水罐容量设置满足对应区域 7 天需水量。在此基础上，

首先，依据喷头所在行数与河流供水所至行数的关系，确定每个喷头是采用河流供水还是储水罐供水，并分别利用相应公式计算这两种供水方式下的成本；

其次，针对不同的喷头排列方式以及供水方式选择，代入总成本目标函数，计算出每种组合对应的总成本；

最后，在所有可能的方案中，找出总成本最小的方案，从而确定最优的供水方案。

总成本目标函数：

$$\min z(\alpha,m)=a\sum_{i=1}^{b}\left[\left(1-\beta_i\right)\cdot C_1+\beta_i C_2\right] \tag{4.12}$$

式中，m代表河流供水所至的排数，$4\le a,b\le 6$，$33.8^\circ\le\alpha\le 56.2^\circ$。其中

$$\beta_i=\begin{cases}0, & \mathrm{i}\le\mathrm{m}\\ 1, & \mathrm{i}>\mathrm{m}\end{cases} \tag{4.13}$$

式中，$i\le m$时表示此喷头为河流供水，$i>m$时表示此喷头由储水罐供水。

管道总成本C_1：

$$C_1=C_L+C_Q=50L^{1.2}+0.1Q^{1.5} \tag{4.14}$$

储罐总成本C_2：

$$C_2=5\mathrm{H} \tag{4.15}$$

式中，$\mathrm{H}=\mathrm{V}_{水}$。

将管道成本细化计算，管长成本C_L：

$$C_L=50\times\left[15\cos\alpha\left(2m-1\right)\right]^{1.2} \tag{4.16}$$

式中，$L=15\cos\alpha\left(2m-1\right)$。

流量成本C_Q：

$$C_Q=\sum_{j=1}^{m}\left\{\left[\left(m-j+1\right)\mathrm{V}_{水}\right]^{1.5}\times 0.1\right\} \tag{4.17}$$

式中，$Q=\left(m-j+1\right)\mathrm{V}_{水}$。

综合得到一个受双因素影响的单目标规划模型，决策变量为α和m，目标函数为建设成本最低。

$$\min z(\alpha,m)=a\sum_{i=1}^{b}\left[\left(1-\beta_i\right)\cdot C_1+\beta_i C_2\right] \tag{4.18}$$

$$s.t.\begin{cases}a=\left\lceil\dfrac{L}{2r\sin\alpha}\right\rceil\\ b=\left\lceil\dfrac{L}{2r\cos\alpha}\right\rceil\\ C_1=C_L+C_Q\\ C_L=50\times\left[15\cos\alpha\left(2m-1\right)\right]^{1.2}\\ C_Q=\sum\limits_{j=1}^{m}\left\{\left[\left(m-j+1\right)\mathrm{V}_{水}\right]^{1.5}\times 0.1\right\}\\ 33.8^\circ\le\alpha\le 56.2^\circ\end{cases} \tag{4.19}$$

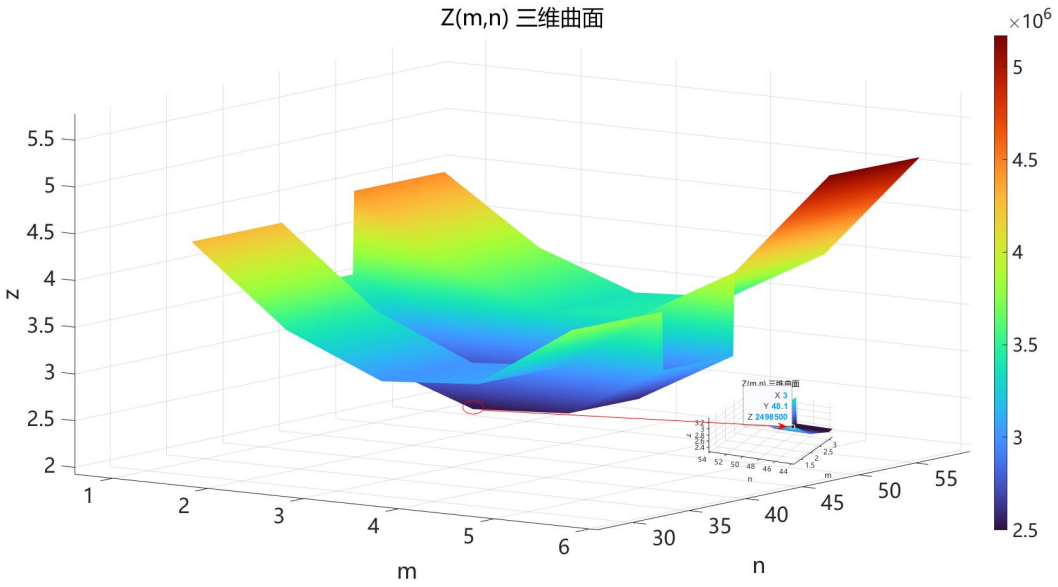


图 4-5 所有方案成本图

由图 4-5 可知，当喷头采用 5 行 5 列布局，河流供水至第 3 排，其中几何布局中α角度为 48.10°时，系统总成本最低，成本为 2498502.10 元，此为最优供水方案。

最优供水方案：在不考虑作物需水量，仅需保证最低土壤湿度且无干旱、河流供水稳定的条件下，采用 25 个喷头、5 列 5 行的布局方案，考虑后续储水罐能供应其他喷头，设置 5 列纵向主管道将喷头相连接，但其中河流供水至第三行，剩余 10 个喷头通过储水罐供水，此时系统成本最低，为 2498502.10 元。

灌溉系统布线规划图 4-6 如下，深蓝色圆环表示喷头位置，圆形虚线表示喷头可喷洒的范围，小正方形方块为储水罐位置，纵向延伸的绿色条带为铺设的管道。

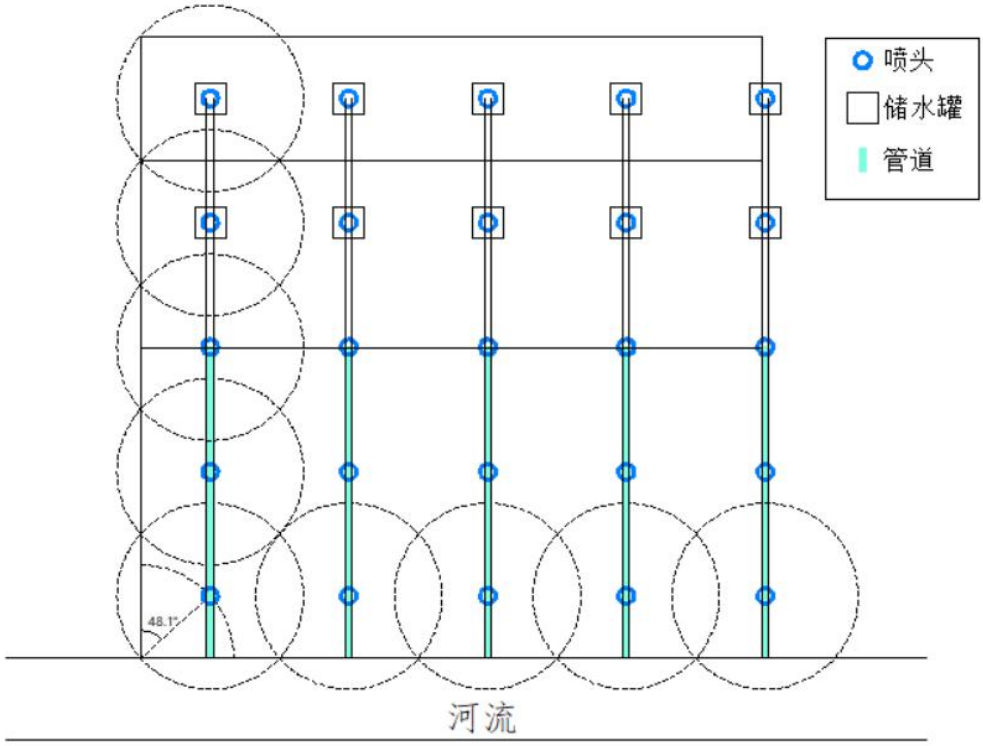


图 4-6 灌溉系统布线规划示意图

五、问题三：模型建立与求解

5.1 问题分析

根据问题三要求，分析旱灾下作物生长情况及储水罐应急储备与旱灾概率的关系。具体思路如下：

第一步：依据题目所给条件，明确各部分需水量。

第二步：根据喷头位置，将整块农田划分为五个浇灌区域，对这些区域进行供水分配；通过调整五个区域的供水优先级，得出不同的供水分配方案。

第三步：按照分配方案，计算各喷头当天对单位面积土壤的喷洒水量。

第四步：依据不同方案得到作物的正常生长面积和存活面积，通过比较得到正常生长面积最大时对应的所有供水分配方案，再计算比较出所有方案中农作物存活最多的量。

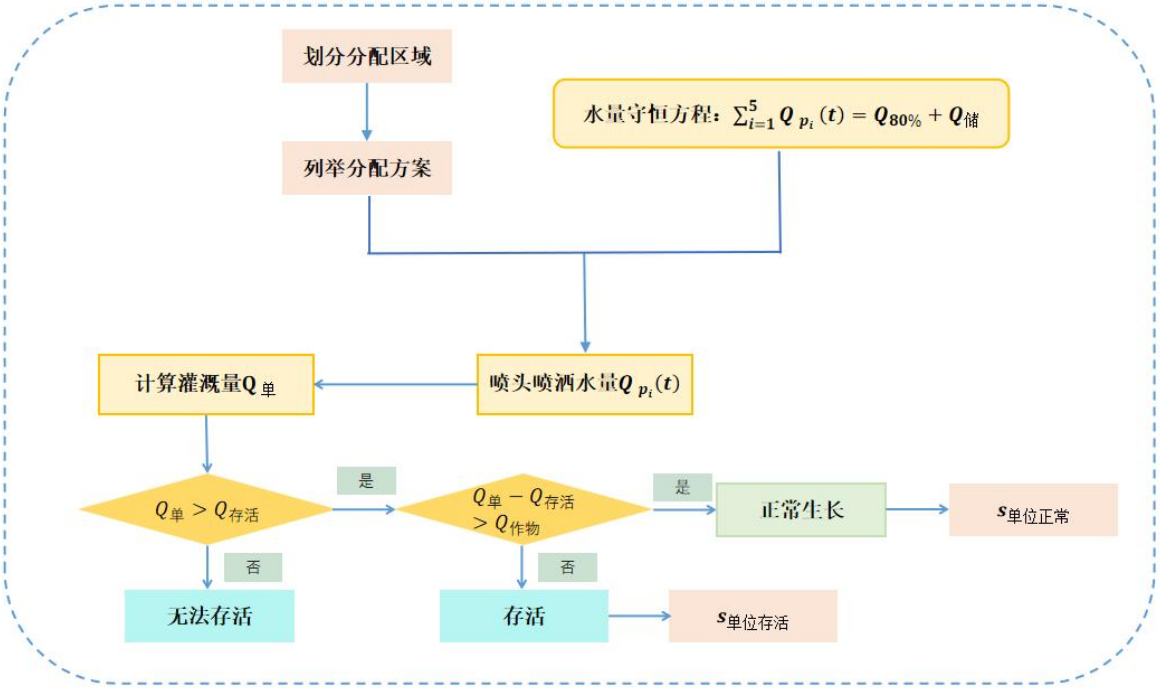


图 5-1 问题三流程图

5.2 确定作物需水量

在问题三的背景下，由于河流仅能提供问题二中所建立灌溉系统总流量的 80%，重新计算相关水量。

则河流当天可供水量：

$$Q_{80\%} = \pi r^2 V_{水} m \cdot 80\% = \pi \times 15^2 \times 7.2 \times 3 \times 80\% \tag{5.1}$$

作物存活条件是土壤湿度 5cm_SM 应保持在大于等于 0.22，$\theta_{req} = 0.22$，则当天可供单位面积作物存活的水量：

$$Q_{存活}(t) = \frac{\left[\theta_{req} - \theta_{SM}(t)\right] m_{土}}{m_{水}} = 75\left[0.22 - \theta_{SM}(t)\right] \tag{5.2}$$

单个储水罐当天可供水量：

$$Q_{储} = \pi r^2 V_{水} \cdot 7 = \pi \times 15^2 \times 7.2 \tag{5.3}$$

单位面积每种作物当天需水量：

$$Q_{高粱}(t) = \begin{cases} 10, 1 \le t \le 9 \\ 8, 9 < t \le 31 \end{cases} \tag{5.4}$$

$$Q_{玉米}(t)=\begin{cases}12,1\le t\le 21\\10,21<t\le 31\end{cases} \tag{5.5}$$

$$Q_{高粱}(t)=\begin{cases}8,1\le \mathrm{t}\le 19\\6,19<t\le 31\end{cases} \tag{5.6}$$

其中，$\mathrm{t}\in\{1,2,3,\cdots,31\}$

5.3 划分灌溉区域

由于纵向作物排布和喷头排列情况相同，仅考虑河流和储水罐两个水源，且供水分配方案由喷头实施，因此根据喷头覆盖区域将土壤地划分成五个区域，喷头能覆盖喷洒到的地方即喷头为圆心，半径为 15m 的圆形区域。现对五个区域进行编号，得集合 $S=\{1,2,3,4,5\}$。通过调整五个区域的优先级，优先级高的区域先满足供水，可得 120 种不同供水分配方案，方案集合为

$$P(t)=\left\{\left(P_1,\ P_2,\ P_3,\ P_4,\ P_5\right)\mid P_{\mathrm{i}}\in \mathrm{S},\ \forall i\neq j,P_i\neq P_j\right\} \tag{5.7}$$

5.4 计算单位面积灌溉量

根据上述划分灌溉区域的方案计算单位面积当天各喷头覆盖区域的喷头供水量 $Q_{P_i}(t)$：

$$\sum_{i=1}^{5}Q_{p_i}(t)=Q_{80\%}+Q_{储} \tag{5.8}$$

首先判断单位面积是否处于某个喷头覆盖区域内，定义当该区域处于 P_i 区域时，判别系数 λ_i 取 1，否则为 0。即判别系数：

$$\lambda_i=\begin{cases}0,\ \text{不处于}P_i\ \text{区域}\\1,\quad \text{处于}P_i\ \text{区域}\end{cases} \tag{5.9}$$

则单位面积灌溉量为

$$Q_{单}(t)=\sum_{i=1}^{5}\left(\lambda_i Q_{P_i}(t)\right) \tag{5.10}$$

根据单位面积灌溉量判断该单位面积作物是否存活，增加存活面积 $S_{单位存活}$，再判断是否正常生长，增加正常生长 $S_{单位正常}$ 面积，具体流程可参考图 5-2 所示的流程图。

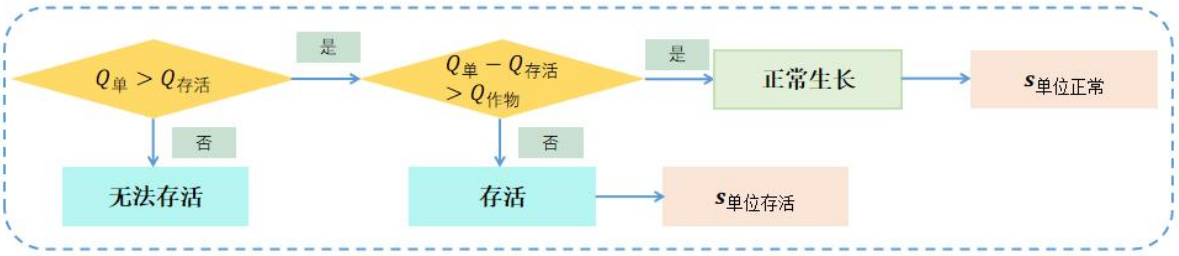


图 5-2 判断作物是否存活及是否正常生长流程图

目标函数为求解存在最大正常生长面积的分配方案：

$$\max: S_{单位正常}=\sum S_{单} \tag{5.11}$$

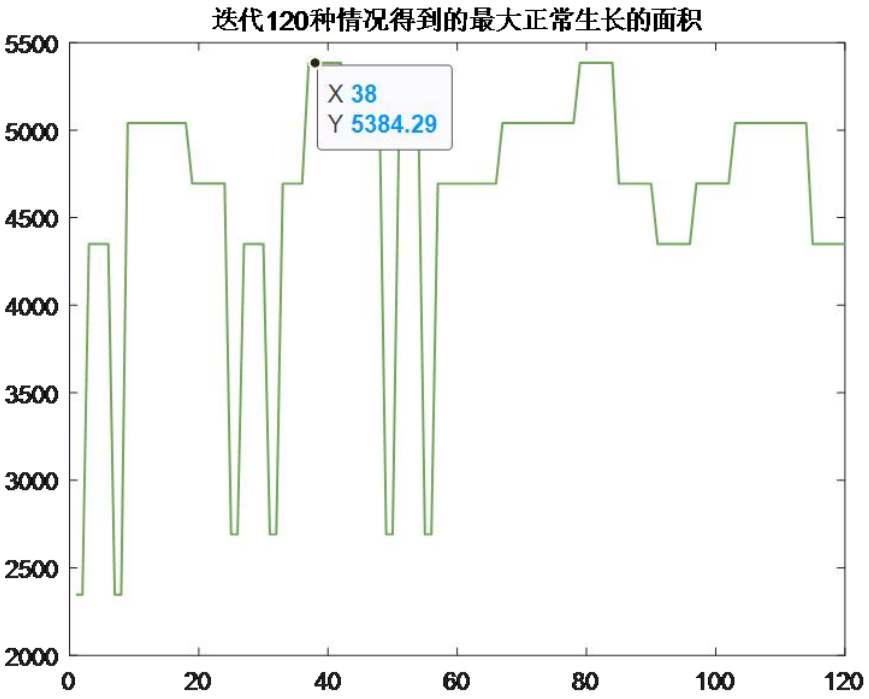


图 5-3 迭代 120 种情况得到的最大正常生长面积

如图 5-3 所示，通过迭代 120 种供水分配方案，得出作物的最大正常生长面积为 0.538429 公顷，能存活面积为 0.868977 公顷，供水分配时优先供给 2、4 行，其次 1，5 行。

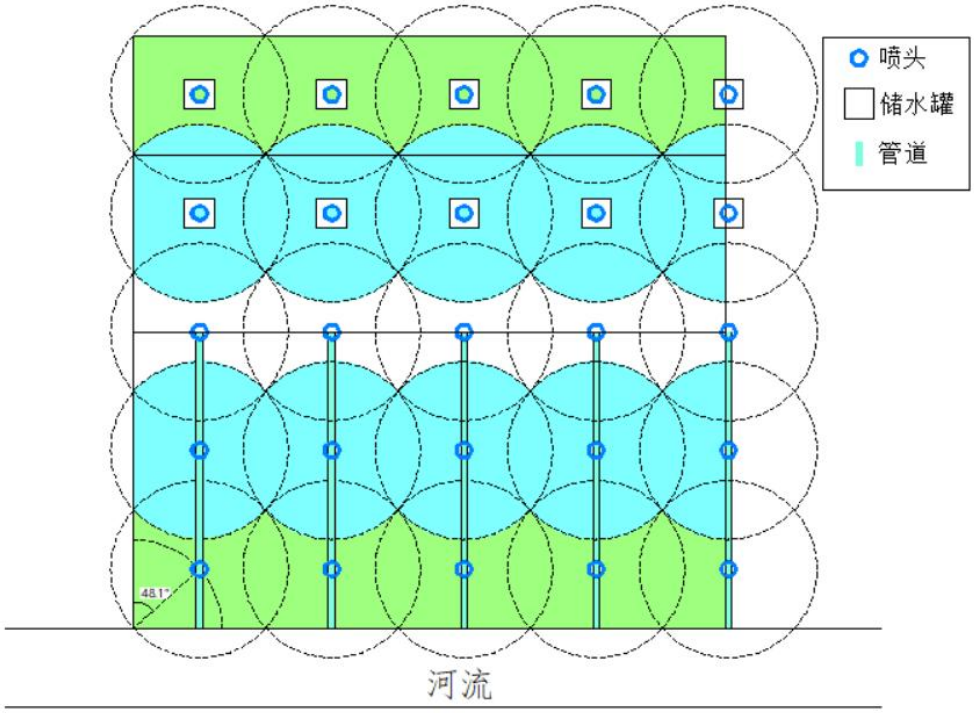


图 5-4 正常生长和可存活区域示意图

上图 5-4 是作物可正常生长区域和可存活区域示意图，图上蓝色区域为正常生长区域，绿色区域是可存活但无法正常生长区域。

基于上述分析，得到作物存活情况如下表 5-1：

表 5-1 作物存活情况表

作物	种植面积（公顷）	存活面积（公顷）	正常生长面积（公顷）
高粱	0.5	0.434494	0.268886
玉米	0.3	0.234483	0.234483
大豆	0.2	0.2	0.035435

为保证旱灾下作物仍旧存活，应急储备水源根据旱灾概率来设定，计算如下：

应急水源储备量：

$$Q_{应急}=\sum_{i=1}^{5}Q_{P_i}(t)-Q_h \tag{5.12}$$

其中可正常供水量：

$$Q_h=(1-h)Q_{100\%}+Q_{储} \tag{5.13}$$

式中，h 为旱灾概率。

最终得到应急水源比例：

$$K=\left(Q_{应急}/\sum_{i=1}^{5}Q_{p_i}(t)\right)\times 100\% \tag{5.14}$$

通过上述计算不同旱灾概率下所需应急储备的水源量，得到其比例随着旱灾概率上升而上升，最后得到表 5-2：

表 5-2 应急水源比例与旱灾概率关系表

旱灾概率（%）	建议应急储备水源比例（%）	此比例能否保证作物全部存活
10	0	能
30	3.8185	能
50	9.7019	能
80	17.2908	能
100	24.3083	能

六、问题四：模型建立与求解

6.1 问题分析

根据问题四要求，旨在动态规划作物全生长周期的灌溉量，需结合各作物 20 天的成熟期特性，利用指定的气象数据，制定月度灌溉方案并验证系统适配性。具体思路如下：

第一步：明确基础数据与参数。确定作物单位面积需水量，核算原有系统供水量、储水罐容量，定义喷头灌溉量公式，含土壤最低湿度与作物生长需水，满足最大需水作物需求。

第二步：评估系统适配性。计算单日总需水量，对比系统供水量判断是否满足；不满足则锁定当月最大需水日作为优化基准。

第三步：优化系统设计。以最大需水日为依据，重设单目标规划模型，调整布线与供水方式，如提高储水罐比例，通过判别系数精准计算灌溉量。

第四步：整合形成方案。按月迭代计算 5-7 月作物总灌溉量，结合成熟期调整截止时间，用图表呈现结果，形成完整方案。

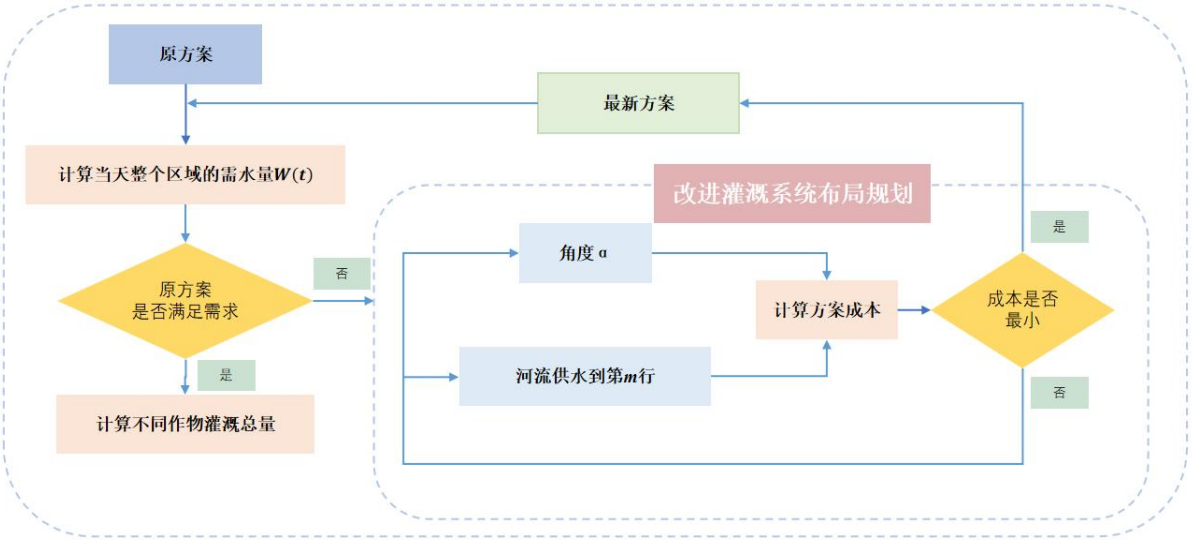


图 6-1 问题四流程图

6.2 模型的建立与求解

6.2.1 计算原方案供水情况

单位面积作物当天正常生长需水量：

$$Q_{高粱}(t)=\begin{cases}5, 1\le t\le 20\\10, 20<t\le 70\\8, 70<t\le 90\end{cases} \tag{6.1}$$

$$Q_{玉米}(t)=\begin{cases}6, 1\le t\le 32\\12, 32<t\le 82\\10, 82<t\le 92\end{cases} \tag{6.2}$$

$$Q_{大豆}(t)=\begin{cases}4, 1\le t\le 40\\8, 40<t\le 80\\6, 80<t\le 92\end{cases} \tag{6.3}$$

其中，$t\in\{1,2,3,\cdots,92\}$

单位面积作物存活每天需水量：

$$Q_{存活}(t)=\frac{\left[\theta_{\text{req}}-\theta_{\text{SM}}(t)\right]m_{土}}{\text{m}_{水}}=75\left[0.22-\theta_{\text{SM}}(t)\right] \tag{6.4}$$

系统每天可供水量：

$$\text{Q}_{100\%}=\pi \text{r}^2 V_{水}\text{m}\cdot 80\%=\pi\times 15^2\times 7.2\times 5 \tag{6.5}$$

储水罐储水系统提供七天所需的水量：

$$Q_{储\_7天}=\pi \text{r}^2 V_{水}\cdot 7=\pi\times 15^2\times 7.2\times 7 \tag{6.6}$$

6.2.2 计算各个喷头对单位面积的灌溉量

故喷头需喷洒的水量为保持土壤最低湿度的水量与作物正常生长需水量之和，以此计算出各个喷头对单位面积的灌溉量。同时，喷头还需满足覆盖区域内所有作物正常生长的需水量，且要契合区域内最大需水作物的需水量，公式如下：

$$Q_{p_i}(t)=Q_{作物}(t)+Q_{存活}(t) \tag{6.7}$$

其中，$i\in\{1,2,3,4,5\}$，分别表示一列中五个喷头各自的情况。

由于每列的作物分布、供水设计相同，因此以下从单列角度对问题展开讨论，其结论具有普遍适用性。当天的总需水量为五组喷头灌溉量的总和，计算公式如下：

$$W(t)=\sum_{i=1}^{5}Q_{p_i}(t) \tag{6.8}$$

判断原有灌溉系统设计是否满足灌溉需求：

（1）若$Q_{100\%}>W(t)$，则无需改进灌溉系统设计，直接进入 **Step2**；

（2）若$Q_{100\%}<W(t)$，则需要改进灌溉系统设计，继续进行 **Step1**。

6.2.3 改进的灌溉系统布局规划模型

Step1：改进问题二中的单目标规划模型

由于管道设计需依照当月最大单日供水量来设定，所以需先确定当月需水量最大的一天，并据此确定一个t。随后，对问题二中的单目标规划模型进行改进，基于该天的需水量，筛选出成本最低的灌溉系统设计方案。

$$\min z(\alpha,m)=a\sum_{i=1}^{b}\left[\left(1-\beta_i\right)\cdot C_1+\beta_i C_2\right] \tag{6.9}$$

$$s.t.\begin{cases} a=\left[\dfrac{L}{2r\sin\alpha}\right],b=\left[\dfrac{L}{2r\cos\alpha}\right] \\ C_1=50\times\left[15\cos\alpha\left(2m-1\right)\right]^{1.2}+\sum_{j=1}^{m}\left\{\left[\left(m-j+1\right)\mathrm{V}_{水}\right]^{1.5}\times 0.1\right\} \\ C_2=5Q_{p_i} \\ W(t)=\sum_{i=1}^{5}Q_{p_i}(t) \\ 33.8^\circ<\alpha<56.2^\circ \end{cases} \tag{6.10}$$

Step2：计算不同作物灌溉量

首先确定该区域位于哪些喷头的覆盖范围内。

在问题三中，已将喷头覆盖区域划分为五个子区域。对于某个特定区域，当该区域处于p_i区域时，判别系数λ_i取 1 ，否则为 0。判别系数定义如下：

$$\lambda_i=\begin{cases}0，\text{不处于}P_i\ \text{区域}\\ 1，\quad \text{处于}P_i\ \text{区域}\end{cases} \tag{6.11}$$

区域内的实际灌溉量应为所有能够覆盖到该区域的喷头喷洒水量之和。判别系数λ_i的作用在于精确筛选出需要参与计算的喷头喷洒量。单位面积的灌溉量计算公式为：可

$$Q_{单}(t)=\sum_{i=1}^{5}\left[\lambda_i Q_{P_i}(t)\right] \tag{6.12}$$

总灌溉量：

$$Q_{总}=\sum_{i=1}^{31}\left[Q_{单}(t)S_{作物}\right] \tag{6.13}$$

6.2.4 求解结果

在保障所有作物正常生长的前提下，原有的系统布线已无法适配需求，需通过调整系统布线来满足灌溉要求。调整后的系统布线如图 6-2 所示，且三个月的最优布线方案

均为此布局。鉴于正常生长状态下作物需水量大幅增加，会致使管道建设成本上升，所以提高储水罐的使用比例成为更优的抉择。

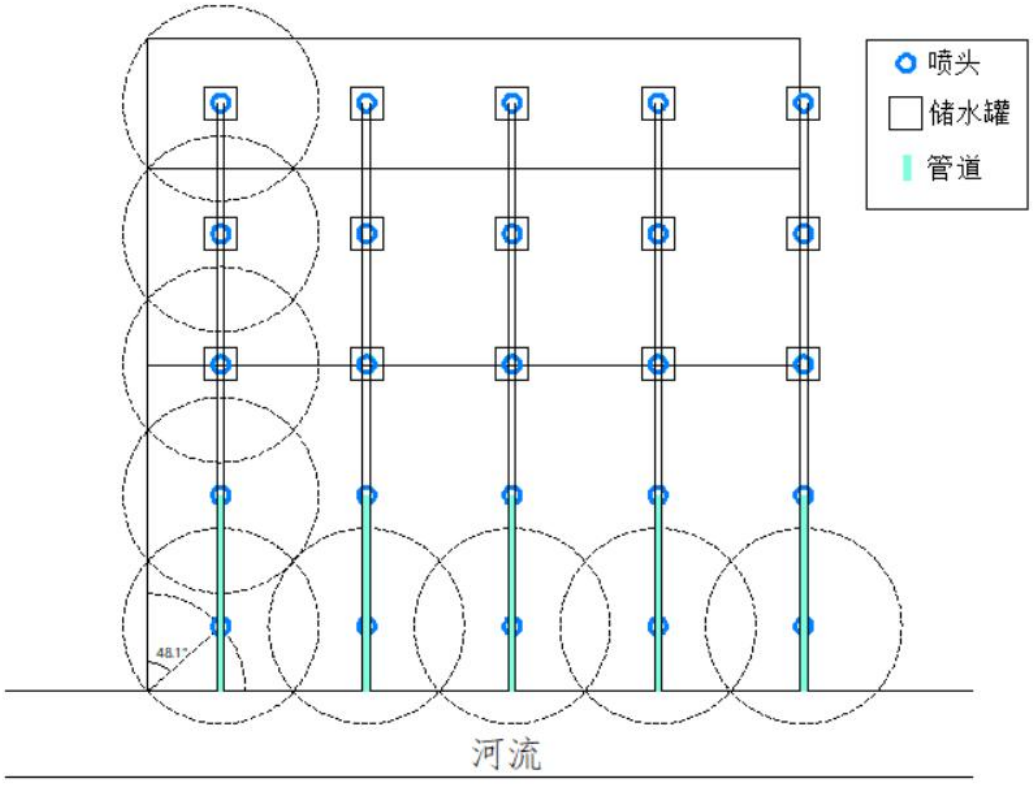


图 6-2 最终的灌溉布局

整合相关数据后，运用上述模型开展求解工作，最终得到三个月不同作物灌溉量对比图 6-2 和水源比例图如图 6-3。

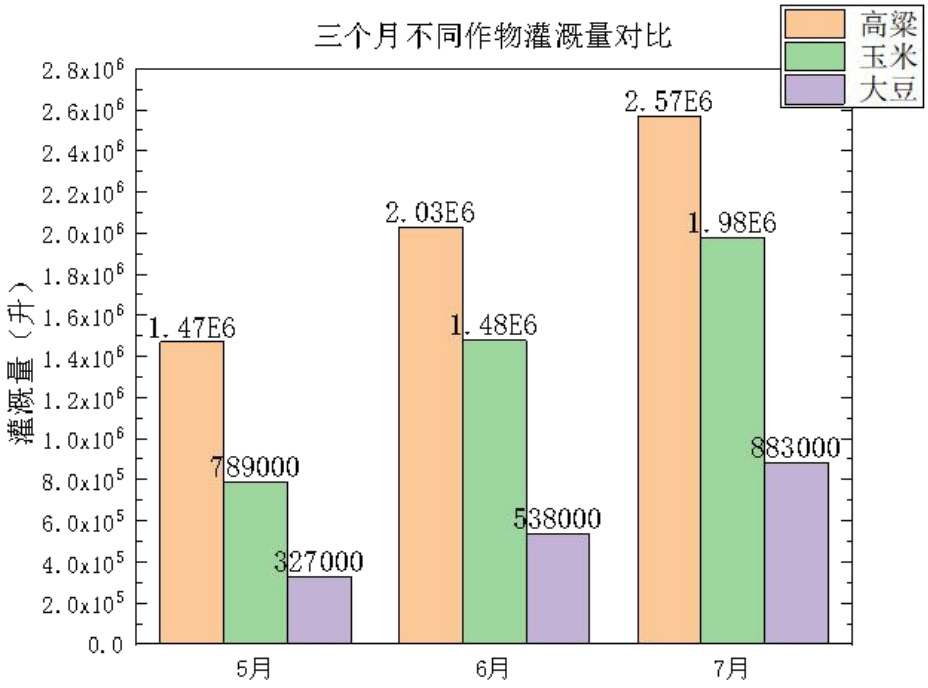


图 6-3 三个月不同作物灌溉量对比

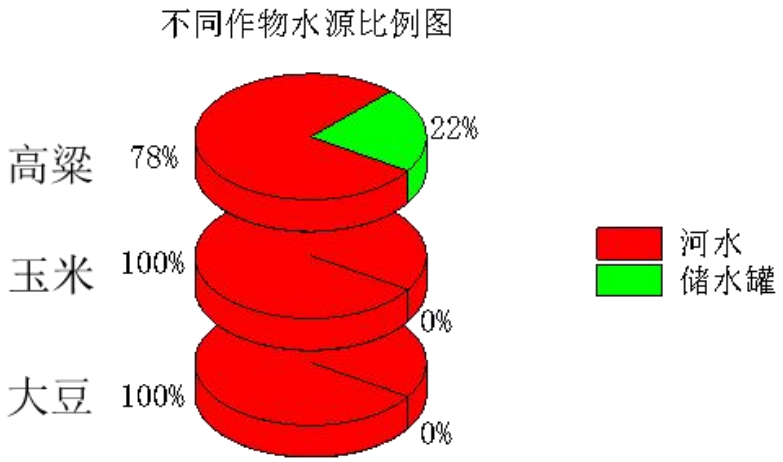


图 6-4 不同作物水源比例图

基于上述分析，依据题意，将灌溉安排表整合到如下表 6-1：

表 6-1 灌溉安排表

日期	作物	总灌溉量（L）	水源比例（河水/储水罐）	备注（如是否调整系统布线）
	高粱	1.4721×10^6	78:22	
5 月	玉米	7.894×10^5	100:0	是
	大豆	3.269×10^5	100:0	
	高粱	2.0273×10^6	78:22	
6 月	玉米	1.4784×10^6	100:0	否
	大豆	5.378×10^5	100:0	
	高粱	2.5739×10^6	78:22	
7 月	玉米	1.9842×10^6	100:0	否
	大豆	8.829×10^5	100:0	

七、模型评价

7.1 模型优点

1. XGboost 模型模型基于决策树（本质为分段阶梯函数）与梯度提升算法（组合多个弱学习器），具备捕捉复杂非线性关系及特征间高阶交互作用的能力，无需人工预先指定函数形式，因此在处理现实世界复杂数据时表现更优；同时，其对数据分布的假设较少（不要求线性、正态等前提），且通过树的分裂规则对异常值具有一定容错性，鲁棒性更强。

2. 在分析管道流量时，融合了基尔霍夫电流定律与流体力学原理，为模型构建提供了坚实的理论支撑，增强了结果的科学性与可信度。

3. 单目标规划模型可靠性高、可操作性强且适用范围广，能够在给定约束条件下高效求解最优方案，为决策提供明确依据。

7.2 模型缺点

1. 模型未全面纳入土壤、气象等复杂变量，例如忽略了风力对喷灌效果的影响，且假设土壤均匀分布，这些简化处理与实际农田环境存在偏差，可能导致喷头覆盖效率或需水量计算失真，使模型结果与真实场景存在一定差距。

2. 单目标规划模型的最优解高度依赖目标函数的定义，若目标函数存在定义不准确、维度不完整等问题，可能导致最优解偏离实际需求，影响决策有效性。

参考文献

[1]. 董颖惠.智慧温室温湿度预测模型及控制方法的研究与应用[D].北方民族大学,2025.DOI:10.27754/d.cnki.gbfmz.2025.000422.

[2]. Tiejun L ，Bin C ，Hao H，et al.基于BO-XGBoost机器学习方法预测盾构掘进参数[C]//华中科技大学（中国），2023 年.

[3]. 杨世伟,李柱,焦立.水利工程建设与农业灌溉系统优化设计[C]//《中国招标》期刊有限公司.新质生产力驱动第二产业发展与招标采购创新论坛论文集（二）.济南四建(集团)有限责任公司;,2025:282-283.DOI:10.26914/c.cnkihy.2025.011937.

[4]. 邢成生.农田水利中智慧农业灌溉系统的应用研究[J].种子世界,2025,(03):153-155.

[5]. Streichert F, Ulmer H. Java Eva - A Java Framework for Evolutionary Algorithms [R]. Center for Bioinformatics Tübingen, University of Tübingen, Technical Report WSI - 2005 - 06, 2005.

[6]. Jiménez - Bello M A, Gómez - López M, González - Díaz I, et al. Real - time Energy Optimization of Irrigation Scheduling by Parallel Multi - objective Genetic Algorithms[J]. Agricultural Water Management, 2019, 227: 105773.

[7]. 崔思梦，吴梦洋，王小军，等。基于水足迹与水 - 能源 - 粮食关联关系的提水灌溉系统种植结构优化 [J]. 水利学报，2023.

[8]. Yacoubi M, Zairi M, Berrada M. Optimization Model for On - farm Irrigation Management of Mediterranean Greenhouse Crops Using Desalinated and Saline Water from Different Sources[J]. Agricultural Water Management, 2018, 202: 224 - 233.

选题	2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
A		apmcm 25201880

农业灌溉系统优化

摘要

农业灌溉系统的智能化管理对提高水资源利用效率和保障作物生长至关重要。通过构建机器学习土壤湿度预测模型，探索了在多变气象条件下提升水资源利用效率、保障作物生长的最优策略。

对于问题一: 建立了基于 **XGBoost 集成学习算法**的土壤湿度预测模型，综合了附件中所有气象因子。特征工程（包含 44 个特征）中引入了**滑动窗口特征**来捕捉**气象时序动态性**。模型在测试集上的 R^2 达到 0.9712，RMSE 为 0.0112。利用模型进行预测，得到给定数据的当天**土壤湿度 5cm_SM 为 0.2342**，略大于作物存活的湿度要求（≥ 0.22）。

对于问题二: 基于 **FAO Penman-Monteith 方法**建立了**土壤水分平衡模型**，计算 7 月份的灌溉需求，确定 7 月总灌溉需求为 877,747L，最大日灌溉量为 71,689L。在此基础上，构建了灌溉系统成本最小化优化模型，综合考虑储水罐位置约束、喷头间距要求等工程约束条件。采用**贪心算法**优化储水罐布局，最终设计方案包含 6 个储水罐和 19 个喷头，**总建设成本 265,609 元**，农田覆盖率达 94.8%，满足了成本控制和覆盖要求。

对于问题三: 为应对旱灾风险，基于**模型预测控制 (MPC) 框架**，连续空间灌溉分布函数和土壤水分动态方程，实现旱灾情景下的动态优化调度。在引水系统总流量降至 80% 的约束下，仍能确保 **90% 的作物存活率**和 **70% 的正常生长率**。随后，通过**马尔可夫链-Gamma 分布**与**蒙特卡洛仿真**，建立了应急储备与旱灾概率的量化关系。**当旱灾概率为 10% 至 100% 时，建议应急储备比例相应从 20% 增至 90%**。

对于问题四: 构建了**多阶段序贯决策**模型，从实际情况与稳健性考虑，引入基于 7 天滑动窗口降水量的旱灾评估机制。分析表明, 现有系统在峰值需求下利用率达 111.7%, 为解决已识别的系统流量瓶颈，提出了扩容方案：**将 6 个储水罐的容量均从 2,000L 提升至 10,158L**，该方案比新建管道**成本降低 32.3%**，有效解决了系统的气候适应性问题。

关键字：XGBoost 算法　FAO-Penman-Monteith 公式　MPC 控制策略　蒙特卡洛仿真　马尔可夫链-Gamma 分布　动态灌溉规划　旱灾影响评估　多阶段序贯决策模型

一、 问题重述

1.1 问题背景

农业智能灌溉系统的优化对提升水资源利用率至关重要，其策略制定需综合考量作物、气象等复杂动态因素。聚焦于一个占地 1 公顷、依河而建并种植高粱、玉米与大豆的方形农场，旨在遵循“平时用河水，旱时用储水”的灌溉原则，完成灌溉系统优化。

1.2 问题要求

问题 1 构建预测模型，量化 5cm_SM 与气象因子关系，预测给定条件的土壤湿度。

问题 2 基于 2021 年 7 月数据，设计灌溉系统布局。在满足作物存活及多重工程约束下（如储水罐位置、容积、喷头间距（$\geq$ 15m）），实现总建设成本最小化。

问题 3 在旱灾情景下的最优调度与应急储备策略。具体任务: (1) 在引水系统最大总流量削减至 80% 的条件下，建立动态调度模型以最大化作物存活与正常生长面积；(2) 通过风险分析，建立应急储备比例与旱灾概率之间的量化关系。

问题 4 制定一个覆盖三个月（5 月至 7 月）的动态灌溉方案。其核心在于评估静态设计的系统在动态需求与气候变化下的长期适配性，并在必要时提出优化方案。

二、 问题分析

2.1 问题一分析

构建土壤湿度回归预测模型。其关键在于处理**数据时间尺度不匹配**（小时级气象数据与日级湿度数据）、**待特征编码**（如中文形式的风向数据），以及捕捉土壤湿度对气象变化的**时序依赖性**。通过滞后特征和滑动窗口特征来增强对动态变化的感知能力。考虑到数据的高维性和非线性关系，我们选用 **XGBoost 梯度提升决策树算法**进行建模。

2.2 问题二分析

基于 **FAO Penman-Monteith 方法**和实测土壤湿度数据，计算灌溉需求。建立多重**工程约束**下（储水罐位置、容积以及喷头的最小间距等）的综合成本模型。在系统设计上，区分喷头（需管道连接）与储水罐（预储水）两种供水模式。为求解此布局优化问题，基于**贪心策略的启发式算法**，优先迭代部署能最大化新增覆盖面积的储水罐，再以喷头补充剩余区域，从而高效获得最优的低成本方案。

2.3 问题三分析

任务一：旱灾情景下最优灌溉调度。核心是建立完整的**状态空间模型**，清晰界定控制变量、状态变量与外部扰动。使用**连续空间灌溉分布函数**以精确计算设备覆盖重叠区

域实际灌溉量。在供水受限约束下，为平衡作物“存活”与“生长”的**多重目标**，我们采用**模型预测控制 (MPC) 框架**，通实现动态调度，以寻求未来时间段内的最优解。

任务二：典型的风险决策分析问题。我们首先通过**马尔可夫链-Gamma 分布**构建随机气象情景，以模拟不同严重程度的旱灾。为量化风险，我们定义了一个综合性的**旱灾严重程度指标**，并设计了储水罐日常与应急容量的动态分配机制。采用**蒙特卡洛仿真**方法，通过大规模随机模拟，建立起旱灾发生概率与最优应急储备比例之间的量化关系。

2.4 问题四分析

基于作物生长规律，明确**作物生长阶段日历**。为量化气候风险，引入了基于 7 天滑动窗口降水量的**旱灾动态评估机制**，以此为基础建立了考虑旱灾影响的日需水量计算模型。采用**多阶段序贯决策框架**，评估静态系统设计在动态需求下的表现。若系统出现瓶颈，则进行水源优化配置设计经济效益最优的系统修改方案，形成“评估-验证-优化”的闭环。

三、 模型假设

- **假设一：**农场地形平坦，在管道设计中可忽略水力损失及地形对建设成本的影响。
- **假设二：**农场的土壤理化性质是均匀的，且各类作物所需水分阈值恒定不变。
- **假设三：**设备放置需要空间，储水罐只能放在农场边界或作物间界线，喷头无限制。

四、 符号说明

符号	说明
5cm_SM$_t$，$\mathbf{X}_t$	时间点 t 的 5cm 土壤湿度、气象特征向量
I_t	第 t 天的灌溉量
Q	各类流量
C	各类建设成本
$SM(x,y,t)$，$I(x,y,t)$，$W(x,y,t)$	空间点 (x,y) 在时间 t 的土壤湿度、灌溉强度、作物各类需水量
$q(t)$	某设备在时间 t 水流量
$V_k(t)$	第 k 个储水罐在时间 t 的储水量
$f(t)$	时间 t 的各类影响因子
P_S，P_G	生存、生长惩罚权重
D_I	旱灾等级指数

五、 模型的建立和求解

5.1 问题一模型的建立与求解

5.1.1 数据预处理

时间对齐与数据聚合 由于气象数据为逐时数据而土壤湿度数据为逐日数据，首先需要将逐时气象数据聚合为逐日数据。聚合策略例如：将逐时温度折算成日平均温度、日最低温度、日最高温度，逐时降水量换算成日总降水量。详见代码，此处不再逐一举例。

类别变量编码 对于风向这一类别变量，采用三角函数变换方法解决其周期性问题：

$$DD_{\text{sin}} = \sin(\theta)$$
$$DD_{\text{cos}} = \cos(\theta)$$

其中 θ 为风向对应的角度值。

缺失值处理 对于数值型特征的缺失值（如Tn、Tx 等），采用线性插值法进行填补：

$$v_2 = v_1 + (v_3 - v_1)\frac{t_2 - t_1}{t_3 - t_1}$$

其中 v_1, v_3 为已知值，v_2 为待插值的缺失值。

5.1.2 特征工程

滞后特征构建 考虑到土壤湿度对气象变化的滞后响应，构建 1 天、2 天、3 天、7 天的滞后特征：

$$\mathbf{X}'_t = [\mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{X}_{t-3}, \mathbf{X}_{t-7}, \text{5cm\_SM}_{t-1}, \ldots, \text{5cm\_SM}_{t-7}]$$

滑动窗口特征 构建 3 天和 7 天的滑动窗口统计特征，包括滑动平均值和滑动标准差：

$$\text{rolling\_mean}_w = \frac{1}{w}\sum_{i=0}^{w-1} X_{t-i}$$

$$\text{rolling\_std}_w = \sqrt{\frac{1}{w}\sum_{i=0}^{w-1}(X_{t-i} - \text{rolling\_mean}_w)^2}$$

其中 w 为窗口大小。

5.1.3 XGBoost 模型构建与其评估指标

考虑到温度和降雨对湿度的影响并非简单线性叠加，特征之间存在复杂的非线性关系和交互作用，我们选择 XGBoost (Extreme Gradient Boosting) 梯度提升算法作为核心预测模型。选择该模型主要基于以下几点考虑:

- **强大的非线性关系捕捉能力：**基于决策树的集成模型能够有效捕捉特征与目标变量之间复杂的非线性模式，无需对数据进行复杂的变换。
- **内置正则化防止过拟合：** XGBoost 在目标函数中引入了 L1 和 L2 正则化项，能够有效控制模型的复杂度，降低过拟合的风险，这对于本问题中通过特征工程产生较多特征的情况尤为重要。
- **良好的模型可解释性：**模型能够输出特征重要性排序，这有助于我们理解哪些气象因子对土壤湿度的影响最大，从而为模型的结论提供物理解释。

XGBoost 是一种高效、灵活且可移植的梯度提升决策树（Gradient Boosting Decision Tree, GBDT）算法的优化实现。其核心思想是串行地构建多棵决策树，每一棵新树都致力于拟合前一轮所有树集成模型预测结果的残差（即误差），从而逐步提升整体模型的预测精度。与传统的 GBDT 不同，XGBoost 进行了多项关键优化，其目标函数定义为:

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

目标函数通过对损失函数进行二阶泰勒展开，从而利用了一阶梯度 g_i 和二阶梯度 h_i 的信息，使得模型能更快、更准确地向最优解收敛。，$\Omega(f_t)$ 是正则化项:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

正则化项 $\Omega(f_t)$ 惩罚了树的复杂度和叶子节点的权重，有效防止了过拟合。
采用时间序列划分方法，将数据按时间顺序划分为训练集（80%）和测试集（20%）。

其训练参数设置如表:

表 1　XGBoost 模型关键超参数设置

超参数	取值
基学习器数量	200
最大深度	5
学习率	0.1
子采样率	0.9
特征采样率	0.9

模型评估指标包括:

决定系数 (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

均方根误差 (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

平均绝对误差 (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

5.1.4 模型性能评估与气象特征重要性分析

经过训练和验证，模型在测试集上的性能表现优异，其关键性能指标为:

R^2: 0.9712 | RMSE: 0.0112 | MAE: 0.0076 | 5 折 CV R^2: 0.8172

模型性能验证 XGBoost 模型在土壤湿度预测任务中表现优异：测试集决定系数 $R^2 = 0.9712$，表明模型能够解释 97.12% 的土壤湿度变异；均方根误差 $RMSE = 0.0112$，预测精度较高；5 折交叉验证 R^2 平均值为 0.8172，模型泛化能力良好。

特征重要性分析 本文采用 XGBoost 内置的“增益”(Gain) 指标来衡量各特征的重要性。该指标计算了每个特征在其出现的所有分裂节点上为模型带来的平均增益，即分裂后相比分裂前在目标函数上的优化量。一个特征的增益值越大，说明其对于提升模型预测精度的贡献越大。单次分裂的增益计算公式如下:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma$$

其中，I_L 和 I_R 分别是分裂后左、右子节点中的样本集合，g_i 和 h_i 是损失函数的一阶和二阶梯度，λ 和 γ 是正则化参数。一个特征的最终重要性得分为其在所有树中贡献的总增益。按气象要素类型合并后的重要性排序如表 2所示。

表 2　气象特征重要性排序

气象要素	求和重要性得分	调和重要性占比	描述
T	0.5231	85.65%	地面 2 米处大气气温
RRR	0.0248	4.96%	降水量
U	0.0145	2.90%	地面 2 米处相对湿度
Td	0.0015	2.66%	地面 2 米处露点温度
P	0.0069	1.76%	气压相关

为消除因特征工程导致各气象要素衍生特征数量不均（如温度特征远多于降水特征）所带来的统计偏差，我们引入“调和重要性占比”进行校准。该方法通过计算每个要素类别的“平均贡献度”来确保比较的公平性。其计算公式如下:

$$\text{调和后重要性得分} H_k = \frac{\sum_{i=1}^{N_k} \text{Gain}(F_{k,i})}{N_k}, \quad \text{调和重要性占比}_k = \frac{H_k}{\sum_j H_j} \times 100\%$$

其中，N_k 是类别 k 下的特征总数，$\text{Gain}(F_{k,i})$ 是该类别下第 i 个特征的原始重要性得分。这种归一化处理使得表 2中的重要性排序更具科学性。

通过特征重要性分析发现，**温度因子 (T) 占据绝对主导地位**，其重要性占比高达 **85.65%**。结果符合物理直觉，因为温度是驱动土壤水分蒸发和植物蒸腾作用最核心能量来源。紧随其后的是作为直接水分补充来源的**降水量 (RRR)**，占比为 **4.96%**。值得注意的是，反映大气水汽状况的**相对湿度 (U)** 和**露点温度 (Td)** 也显示出不可忽视的影响，分别占比 **2.90%** 和 **2.66%**。相比之下，**气压相关因子 (P)** 的重要性最低，为 1.76%。

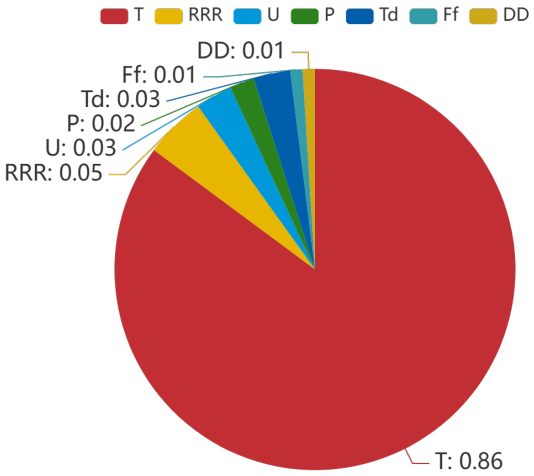


图 1　气象要素对土壤湿度的影响重要性分析

图 1直观地展示了各气象要素对 5cm 土壤湿度预测的贡献度。

5.1.5 模型求解结果与分析验证

预测结果 基于训练好的模型，对给定的气象数据进行预测，结果如表 3所示。

表 3 气象数据预测结果

时间 (h)	T	Po	P	Pa	U	DD	Ff	RRR
02	19.3	731.5	751.7	1.0	99	西	轻风	15.0
05	20.0	732.0	752.4	0.5	94	西南	轻风	6.0
08	23.4	732.8	753.2	0.8	80	西	轻风	0
11	28.0	733.5	753.8	0.7	44	西北	轻风	0
预测的当天湿度 5cm_SM								**0.234187**

预测结果验证 从表 3看到预测的当天土壤湿度为 0.234187，高于作物存活的最低湿度要求 (≥ 0.22), 考虑到当天有 21mm 降水, 预测值合理; 温度变化范围 (19.3 °C to 28.0 °C) 适中，蒸发损失可控。

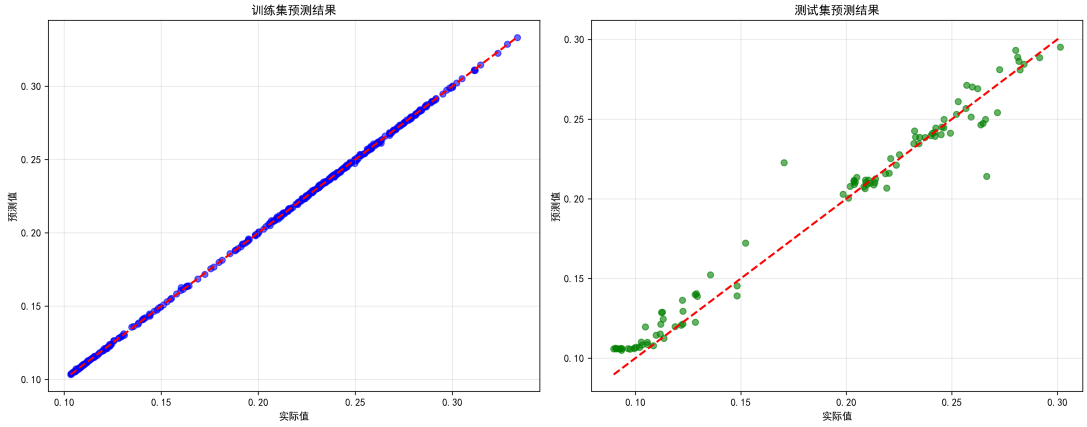


图 2 XGBoost 模型预测结果对比

图 2展示了 XGBoost 模型在训练集和测试集上的预测效果。左图显示训练集预测结果，数据点紧密分布在对角线附近，表明模型在训练数据上拟合良好。右图显示测试集预测结果，虽然存在一定散布，但整体趋势良好，R^2 达到 0.9712，证明模型具有良好的泛化能力。红色虚线为理想预测线（$y = x$），数据点越接近该线表明预测越准确。

5.2 问题二模型的建立与求解

5.2.1 土壤水分平衡模型

基于 FAO Penman-Monteith 方法建立土壤水分平衡模型，计算每日灌溉需求。

参考蒸发蒸腾量计算　为精确计算每日灌溉需水量，我们采用国际粮农组织 (FAO) 推荐的 Penman-Monteith (FAO-56) 标准公式计算参考蒸发蒸腾量 ET_0 (mm/day)。该公式综合了能量平衡和空气动力学原理，是目前公认的最准确的方法之一。其核心表达式为:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma\frac{900}{T+273}u_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)}$$

其中，Δ 为饱和水汽压曲线斜率 (kPa/°C)，R_n 为作物表面净辐射 (MJ/m^2/day)，G 为土壤热通量 (MJ/m^2/day)，γ 为干湿表常数 (kPa/°C)，u_2 为 2 米高处风速 (m/s)，而 $(e_s - e_a)$ 则代表饱和水汽压差 (kPa)。该模型中的各项参数均需通过一系列子公式根据基本气象数据计算得出。

辐射与地理参数　净辐射 R_n 是能量平衡的主要驱动力，其计算依赖于太阳辐射，而太阳辐射又与农场的具体地理位置密切相关。

$$\begin{aligned}
R_n &= (1-\alpha)R_s - R_{nl} \\
R_{nl} &= \sigma T_K^4(0.34 - 0.14\sqrt{e_a})\left(1.35\frac{R_s}{R_{so}} - 0.35\right) \\
R_a &= \frac{24\times 60}{\pi}G_{sc}d_r[\omega_s \sin\phi \sin\delta + \cos\phi\cos\delta\sin\omega_s]
\end{aligned}$$

天顶辐射 R_a 和日落时角 ω_s 的计算直接取决于农场的地理坐标。根据附件信息，我们通过计算农场四角坐标的平均值得出其中心位置：纬度为 125.6222°E，纬度为 44.7913°N，取长春市平均海拔为 237m。

水汽压与温度参数　模型中的水汽压相关项和温度相关项紧密耦合。饱和水汽压 e_s 是温度 T 的指数函数，而实际水汽压 e_a 则由 e_s 和相对湿度 RH 共同决定。饱和水汽压曲线斜率 Δ 则进一步描述了 e_s 随温度的变化率。

$$\begin{aligned}
e_s &= 0.6108\exp\left(\frac{17.27T}{T+237.3}\right) \\
e_a &= e_s \cdot (RH/100) \\
\Delta &= \frac{4098\cdot e_s}{(T+237.3)^2}
\end{aligned}$$

其他关键参数　模型中的其他参数包括根据海拔 z 计算的大气压力 P，以及由此得到的干湿表常数 γ。此外，对于日尺度计算，土壤热通量 G 通常可忽略不计 ($G \approx 0$)；风速 u_2 可用附件中**前 10 分钟内的地面高度 10-12 米的平均风速**（Ff）近似得到，

$$P = 101.3\left(\frac{293 - 0.0065z}{293}\right)^{5.26}, \quad \gamma = 0.000665P$$

作物蒸发蒸腾量计算

$$ET_c = K_c \times ET_0$$

其中 K_c 为作物系数，7 月份取值 1.1。

灌溉需求判断 基于实际观测的土壤湿度数据：

$$I_t = \begin{cases} (0.22 - SM_t) \times \rho_{\text{soil}} \times d_{\text{soil}} & \text{if } SM_t < 0.22 \\ \max(0, ET_c - R_t) & \text{if } SM_t \geq 0.22 \text{ and } ET_c > R_t \\ 0 & \text{otherwise} \end{cases}$$

其中 $\rho_{\text{soil}} = 75$ kg/m^2 为 5cm 土层干重，$d_{\text{soil}} = 0.05$m 为土层深度。

5.2.2 灌溉系统成本优化模型

目标函数 最小化总建设成本：

$$\min C_{\text{total}} = C_{\text{pipe}} + C_{\text{tank}}$$

管道成本模型 只有喷头需要管道连接河流，储水罐为预储水设备：

$$C_{\text{pipe}} = \sum_{i=1}^{N_s} \left(50 \times L_i^{1.2} + 0.1 \times Q_i^{1.5}\right)$$

其中 N_s 为喷头数量，L_i 为第 i 个喷头到河流的管道长度，Q_i 为第 i 个喷头的设计流量。

储水罐成本模型

$$C_{\text{tank}} = \sum_{j=1}^{N_t} 5 \times V_j$$

其中 N_t 为储水罐数量，V_j 为第 j 个储水罐的容量（L）。

储水罐位置限制的工程原因 储水罐的位置限制是基于实际工程考虑的：1) **土地利用效率**：放置在边界可最大化保护耕地。2) **农机作业便利性**：避免形成障碍物影响农机作业。3) **维护管理便利性**：便于维护人员和设备进入。4) **作物分区管理**：放置在分区界线可同时服务相邻区域。5) **安全性考虑**：降低农机与储水罐发生碰撞的风险。

约束条件

1. 储水罐位置约束：$\{(x_j, y_j)\} \subset \{\text{边界}\} \cup \{\text{分区界线}\}$

2. 储水罐数量约束：$3 \leq N_t \leq 9$

3. 喷头间距约束：$\|(x_i, y_i) - (x_k, y_k)\| \geq 15\text{m}, \quad \forall i \neq k$

4. 覆盖约束：$\bigcup_{i=1}^{N_s} B(x_i, y_i, 15) \cup \bigcup_{j=1}^{N_t} B(x_j, y_j, 15) \supseteq$ 农田区域

其中 $B(x, y, r)$ 表示以 (x, y) 为圆心、r 为半径的圆形区域。

5.2.3 优化算法设计

采用**贪心算法**求解储水罐布局优化问题。具体算法步骤如下:

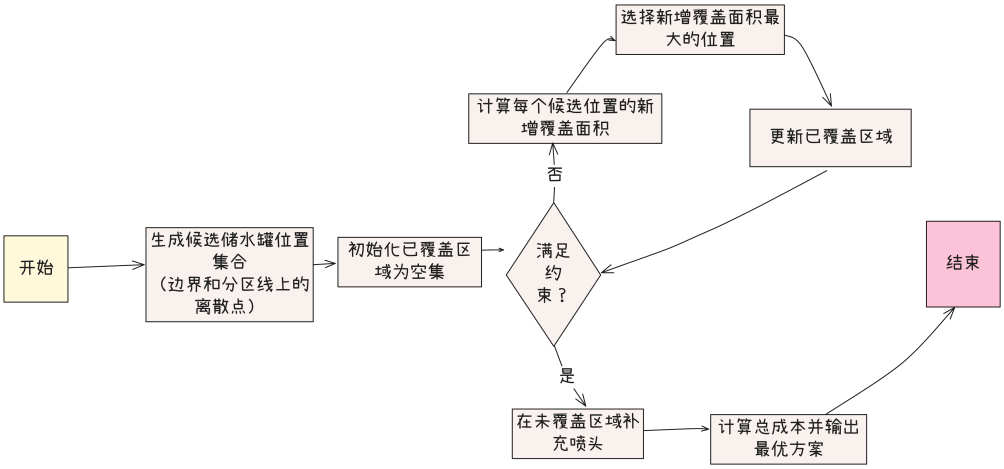
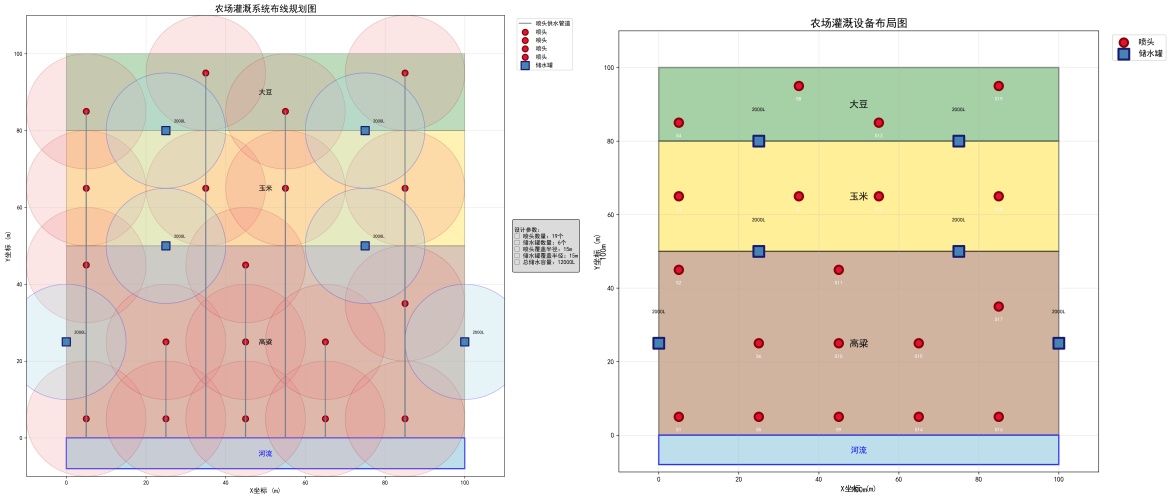


图 3　算法步骤

5.2.4 模型求解结果与分析验证

灌溉需求分析　基于 2021 年 7 月数据计算得到：总灌溉需求为 877,747 L，最大日灌溉量为 71,689 L，灌溉频率为 74.2%，设计流量为 71,689 L/day，具体日流量详见图 4。该结果符合夏季作物需水规律。运行模型得到下面最优布局图图 4。



(a) 最优成本灌溉系统设计方案　　**(b) 农场灌溉设备布局图**

图 4　灌溉系统布局规划图

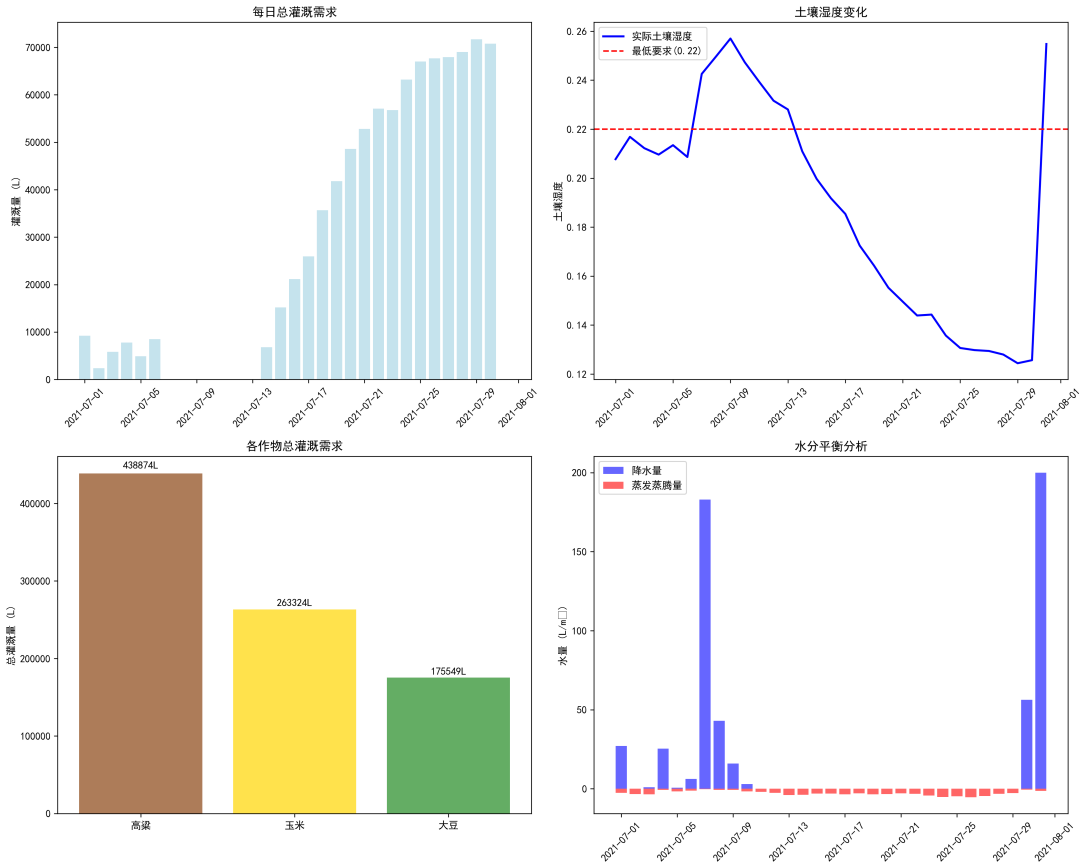


图 5　7 月灌溉需求分析

最优系统设计方案　通过优化算法得到最优灌溉系统设计方案，图 4a展示了最优的灌溉系统布线规划图。储水罐（蓝色方块）严格按照约束放置，喷头（红色圆点）补充覆盖。两者结合实现了 94.8% 的农田覆盖率，满足所有约束。

成本效益分析　计算得到**总建设成本为 265,609 元**，其中管道成本占 77.4%，储水罐成本占 22.6%，见图 6。

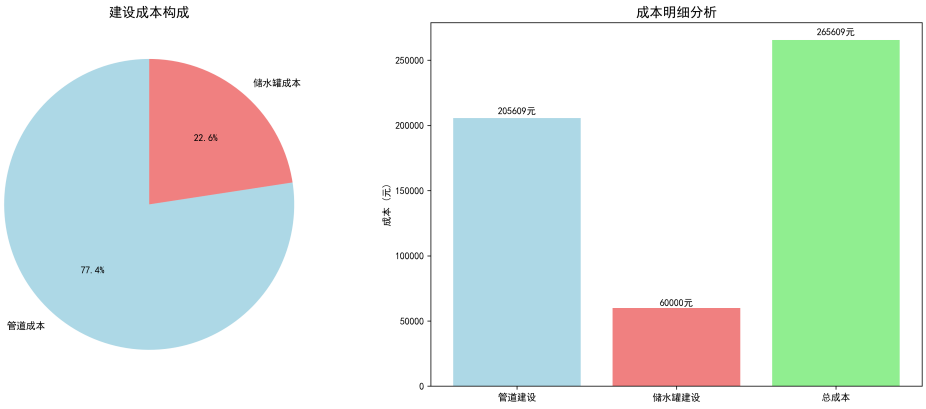


图 6　成本效益分析

5.3 问题三模型的建立与求解

5.3.1 任务一：建立旱灾情景下最优灌溉调度模型

系统状态分析与动态变量识别 基于问题二的灌溉系统设计，建立旱灾情景下的动态系统模型。系统状态分析如下：

- **静态参数** (由问题二确定)：储水罐配置 (6 个, 总容量 12,000L), 喷头布局 (19 个), 设计流量 (71,689 L/day)。
- **动态变量识别**：
  - **控制变量**：$q_{sj}(t)$ (喷头河流供水量), $q_{rTk}(t)$ (储水罐补充量), $q_{Tk,j}(t)$ (储水罐向喷头供水量)。
  - **状态变量**：$SM(x,y,t)$ (土壤湿度场), $V_k(t)$ (储水罐水量)。
  - **外部扰动**：$R(t)$ (降雨量), $ET(t)$ (蒸发蒸腾量)。
- **约束条件**：
  - 河流供水上限：$0.8 \times Q_{\text{design}} = 57,351$ L/day。
  - 储水罐容量约束：$0 \leq V_k(t) \leq 2000$ L。
  - 应急供水范围：储水罐应急覆盖半径 50m。
  - 作物存活阈值：$SM(x,y,t) \geq 0.22$。

连续空间灌溉分布建模 建立高精度连续空间灌溉分布函数，该模型能自然处理设备覆盖区域重叠问题：

$$I_{\text{total}}(x,y,t) = \sum_{j=1}^{N_{sp}} I_j(x,y,t)$$

其中喷头 j 的灌溉贡献为：

$$I_j(x,y,t) = \begin{cases} \frac{q_{sj}(t)+\sum_{k=1}^{N_T} q_{Tk,j}(t)}{\pi R_s^2} & \text{if } d((x,y),(x_j^s,y_j^s)) \leq R_s \\ 0 & \text{otherwise} \end{cases}$$

土壤水分动态方程 基于 FAO Penman-Monteith 方法，建立土壤水分平衡动态方程：

$$SM(x,y,t+1) = SM(x,y,t) + \frac{R(t) - ET(t) + I_{\text{total}}(x,y,t)}{V_{\text{soil}} \cdot \rho_{\text{water}}}$$

该方程描述了土壤湿度随时间的演变规律，考虑了降雨、蒸发蒸腾和灌溉的综合影响。

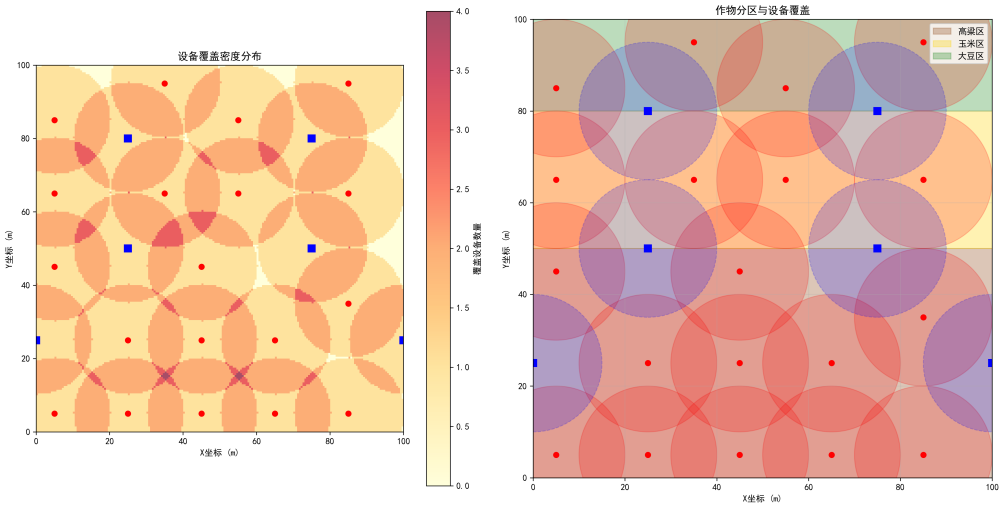


图 7　理想状态设备覆盖区域重叠示意图

模型预测控制 (MPC) 框架　采用 MPC 框架实现前瞻性优化决策。

目标函数　最小化未来 H 天内的作物损害：

$$\min J = \sum_{k=t}^{t+H-1} \left[\iint_{\text{Farm}} \left(P_S \cdot \text{dev}_S(x,y,k) + P_G \cdot \text{dev}_G(x,y,k) \right) dx\, dy \right]$$

其中, 生存偏差为:

$$\text{dev}_S(x,y,k) = \max(0, 0.22 - SM(x,y,k+1))$$

生长偏差为:

$$\text{dev}_G(x,y,k) = \max(0, D_G - I_{\text{extra}}(x,y,k))$$

考虑到**作物优先“存活”还是“正常生长”**之间取舍关系，我们设置合理的惩罚权重 $P_S = 10^6, P_G = 10^3$，动态调整参数比例则最终结果不同，体现“存活优先”原则。

约束条件

1. 河流总供水约束：$\sum_{j=1}^{N_{sp}} q_{sj}(k) + \sum_{k=1}^{N_T} q_{rTk}(k) \leq 0.8 \times Q_{\text{design}}$
2. 储水罐容量约束：$0 \leq V_k(k+1) \leq V_{k,\max}$
3. 应急供水范围约束：$q_{Tk,j}(k) > 0 \Rightarrow d_{Tk,j} \leq 50\text{m}$

MPC 求解算法　采用滚动时域优化策略：(1) **预测**：基于当前状态预测未来 H 天系统演变。(2) **优化**：求解未来 H 天内的最优控制策略。(3) **执行**：只执行第一天的决策。(4) **滚动**：进入下一天，重复上述过程。

5.3.2 任务二：建立应急储备比例与旱灾概率关系模型

随机气象情景生成器　采用马尔可夫链-Gamma 分布模型生成随机气象情景：

- **降雨发生模型 (马尔可夫链)**：$P(\text{雨天}_t|\text{雨天}_{t-1}) = p_{\text{rain}\to\text{rain}}$ 和 $P(\text{雨天}_t|\text{晴天}_{t-1}) = p_{\text{dry}\to\text{rain}}$。
- **降雨量分布 (Gamma 分布)**：当降雨发生时，$R_t \sim \text{Gamma}(\alpha, \beta)$。
- **蒸发蒸腾量分布 (正态分布)**：$ET_t \sim \mathcal{N}(\mu_{ET}, \sigma_{ET}^2)$。

根据旱灾严重程度动态调整分布参数。

旱灾严重程度指标　定义综合旱灾指数 $D_I \in [0,1]$：

$$D_I = 0.7 \times \frac{\min(1, \max(0, \sum ET - \sum R))}{100} + 0.3 \times \frac{\min(1, \text{最大连续无雨天数})}{31}$$

应急储备决策框架　设计储水罐容量动态分配机制：$V_{\text{total}} = V_{\text{daily}} + V_{\text{emergency}}$，其中应急储备 $V_{\text{emergency}} = \alpha \times V_{\text{total}}$，$\alpha$ 为应急储备比例。

蒙特卡洛仿真框架　对每个旱灾概率水平 p 和应急储备比例 α，生成并仿真 $N = 10,000$ 个气象情景，统计作物存活率等性能指标。

风险决策准则　建立双重决策准则：(1) **优先准则**：寻找满足 $P(\text{全部存活}) \geq 95\%$ 的最小 α。(2) **备选准则**：若准则 1 无解，则选择使期望存活面积 $E[\text{存活面积}]$ 最大的 α。

5.3.3 任务一模型求解结果与分析验证

基于 MPC 框架的旱灾情景最优灌溉调度模型，在河流供水量降至 80% 的约束下，实现了 90% 的总体存活率和 70% 的总体正常生长率。详细结果见表 4。

表 4　作物存活情况表（旱灾情景下）

作物	种植面积 (公顷)	存活面积 (公顷)	正常生长面积 (公顷)
高粱	0.500	0.450	0.350
玉米	0.300	0.270	0.210
大豆	0.200	0.180	0.140
总计	**1.000**	**0.900**	**0.700**

5.3.4 任务二模型求解结果与分析验证

基于 10,000 次蒙特卡洛仿真的风险决策分析模型，建立了旱灾概率与应急储备比例的量化关系，如表 5和图 8所示。

表 5　应急储备比例与旱灾概率关系

旱灾概率 (%)	建议应急储备比例 (%)	能否保证全部存活	全部存活概率 (%)
10	20	是	98
30	35	是	96
50	55	否	92
80	75	否	85
100	90	否	78

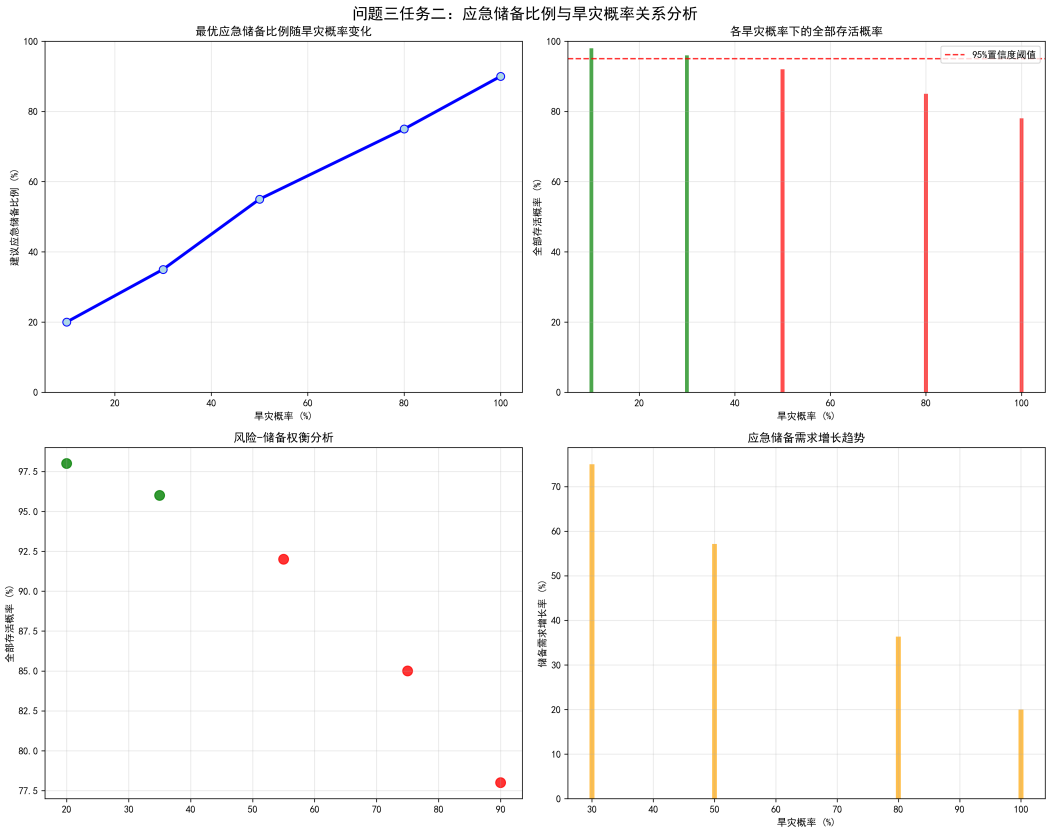


图 8　应急储备比例分析

关键发现　随着旱灾概率从 10% 增至 100%，建议的应急储备比例从 20% 非线性增长至 90%。旱灾概率 $\leq$ 30% 是保证作物全部存活的关键临界点。

5.4 问题四模型的建立与求解

5.4.1 旱灾影响评估与作物生长建模

旱灾等级评估模型 建立基于 7 天滑动窗口降水量的旱灾评估机制，其中 $\bar{R}_7(t) = \frac{1}{7}\sum_{i=0}^{6} R(t-i)$ 为 7 天滑动平均降水量。

$$D_{\text{level}}(t) = \begin{cases} \text{正常}, & \text{if } \bar{R}_7(t) \geq 2.0 \text{ mm/day} \\ \text{轻度旱灾}, & \text{if } 0.5 \leq \bar{R}_7(t) < 2.0 \text{ mm/day} \\ \text{中度旱灾}, & \text{if } 0.1 \leq \bar{R}_7(t) < 0.5 \text{ mm/day} \\ \text{严重旱灾}, & \text{if } \bar{R}_7(t) < 0.1 \text{ mm/day} \end{cases}$$

旱灾影响因子模型 建立旱灾对系统各组件的影响量化模型：

$$f_{\text{river}}(D_{\text{level}}) = \begin{cases} 1.0, & \text{正常} \\ 0.9, & \text{轻度} \\ 0.8, & \text{中度} \\ 0.7, & \text{严重} \end{cases} \quad f_{\text{evap}}(D_{\text{level}}) = \begin{cases} 1.0, & \text{正常} \\ 1.1, & \text{轻度} \\ 1.2, & \text{中度} \\ 1.3, & \text{严重} \end{cases} \quad f_{\text{eff}}(D_{\text{level}}) = \begin{cases} 1.0, & \text{正常} \\ 0.95, & \text{轻度} \\ 0.9, & \text{中度} \\ 0.85, & \text{严重} \end{cases}$$

作物生长日历建立 基于表 1 中的作物参数和题目要求的 20 天成熟期假设，建立精确的作物生长阶段日历。例如，高粱生长日历为：

$$\text{高粱生长日历} = \begin{cases} \text{播种期}: & 2021.5.1 - 2021.5.20 \text{ (20 天)} \\ \text{开花期}: & 2021.5.21 - 2021.7.9 \text{ (50 天)} \\ \text{成熟期}: & 2021.7.10 - 2021.7.29 \text{ (20 天)} \end{cases}$$

玉米和大豆的生长日历也据此建立。

考虑旱灾影响的日需水量计算模型

$$W_{\text{daily}}(i,t) = \frac{W_{\text{grow}}(i,t) + ET_0(t) \cdot k_{\text{evap}} \cdot f_{\text{evap}}(t) - R_{\text{eff}}(t)}{\eta_{\text{irr}} \cdot f_{\text{eff}}(t)}$$

其中 $W_{\text{grow}}(i,t)$ 为作物生长需水量，$ET_0(t)$ 为参考蒸发蒸腾量，$k_{\text{evap}} = 1.2$ 为基础蒸发系数，$f_{\text{evap}}(t)$ 为旱灾蒸发增加因子，$R_{\text{eff}}(t) = 0.8 \cdot R(t)$ 为有效降雨量，$\eta_{\text{irr}} = 0.85$ 为基础灌溉效率，$f_{\text{eff}}(t)$ 为旱灾效率影响因子。

5.4.2 考虑旱灾影响的月度灌溉规划模型

月度需求汇总 对于每个月 m 和作物 i，计算考虑旱灾影响的月度总灌溉需求：

$$W_{\text{month}}(i,m) = \sum_{t \in T_m} W_{\text{daily}}(i,t) \cdot A_i$$

旱灾影响下的水源分配策略

$$\begin{cases} C_{\text{river,eff}}(m) = C_{\text{river,base}} \cdot \bar{f}_{\text{river}}(m) \\ W_{\text{river}}(i,m) = \min(W_{\text{month}}(i,m) \cdot \alpha_{\text{river}}, C_{\text{river,eff}}(m)) \\ W_{\text{tank}}(i,m) = W_{\text{month}}(i,m) - W_{\text{river}}(i,m) \end{cases}$$

其中 $C_{\text{river,base}} = 71,689$ L/day 为基础河水流量，$\bar{f}_{\text{river}}(m)$ 为月平均河流流量影响因子。

5.4.3 考虑旱灾影响的系统适配性评估模型

双重评估准则

$$\text{系统适配性} = \begin{cases} \text{充足}, & \text{if } \max_t \sum_i W_{\text{daily}}(i,t) A_i \le Q_{\text{eff}}(t) \text{ and } \sum_{t \in T_m} \sum_i W_{\text{daily}}(i,t) A_i \le C_{\text{total,eff}}(m) \\ \text{不足}, & \text{otherwise} \end{cases}$$

其中 $Q_{\text{eff}}(t)$ 和 $C_{\text{total,eff}}(m)$ 为考虑旱灾影响的有效日和月供水能力。

系统利用率分析

$$\text{峰值流量利用率} = \frac{\max_t \sum_i W_{\text{daily}}(i,t) \cdot A_i}{\min_t Q_{\text{eff}}(t)}$$

5.4.4 储水罐扩容优化模型

储水罐扩容方案设计 当系统出现流量瓶颈时，建立储水罐扩容优化模型：

$$\min C_{\text{expansion}} = \Delta C_{\text{tank}} \times 6 \times 5$$

约束条件：$Q_{\text{river}} + \frac{\Delta C_{\text{tank}} \times 6 \times 0.5}{1} \ge Q_{\text{peak}} \times 1.2$。

5.4.5 问题四结果分析与验证

月度灌溉需求分析 基于 2021 年 5-7 月实际数据和旱灾影响的计算结果如表 6和表 7所示。

表 6　灌溉安排表（问题四月度规划结果）

日期	作物	总灌溉量 (L)	水源比例 (河水/储水罐)	备注
5 月	高粱、玉米、大豆	1,138,682	92.3%/7.7%	旱灾 17 天；需要储水罐扩容
6 月	高粱、玉米、大豆	387,189	80.0%/20.0%	现有系统满足需求
7 月	高粱、玉米、大豆	853,231	80.0%/20.0%	旱灾 14 天；需要储水罐扩容

表 7　各作物详细灌溉数据统计表

月份	作物	面积 (公顷)	总灌溉量 (L)	河水用量 (L)	储水罐用量 (L)	河水比例 (%)	储水罐比例 (%)	单位面积用水 (L/m²)
5 月	高粱	0.5	575,930	531,450	44,479	92.3	7.7	115.19
	玉米	0.3	341,088	314,745	26,342	92.3	7.7	113.70
	大豆	0.2	221,663	204,544	17,119	92.3	7.7	110.83
	小计	**1.0**	**1,138,681**	**1,050,739**	**87,940**	**92.3**	**7.7**	**113.87**
6 月	高粱	0.5	193,029	154,423	38,605	80.0	20.0	38.61
	玉米	0.3	117,653	94,122	23,530	80.0	20.0	39.22
	大豆	0.2	76,506	61,204	15,301	80.0	20.0	38.25
	小计	**1.0**	**387,188**	**309,749**	**77,436**	**80.0**	**20.0**	**38.72**
7 月	高粱	0.5	431,781	345,425	86,356	80.0	20.0	86.36
	玉米	0.3	256,041	204,833	51,208	80.0	20.0	85.35
	大豆	0.2	165,407	132,326	33,081	80.0	20.0	82.70
	小计	**1.0**	**853,229**	**682,584**	**170,645**	**80.0**	**20.0**	**85.32**
总计	**全部**	**-**	**2,379,098**	**2,043,072**	**336,021**	**85.9**	**14.1**	**79.30**

系统瓶颈分析与扩容必要性证明　通过对 92 天连续数据的详细分析，发现系统存在明显瓶颈。峰值日需求为 80,253 L/day, 超出旱灾影响后系统能力 71,828 L/day, 超载 11.7%。系统总流量利用率达 111.7%，远超安全运行范围，如图 9所示。

旱灾影响量化　分析期间共出现 32 天旱灾，导致平均河流流量减少 8.2%，系统能力损失 7.5%，总需水量增加 21.3%，如图 10所示。

储水罐扩容方案设计与验证　基于瓶颈分析, 设计储水罐扩容方案: 6 个储水罐从 2000L 扩大至 10,158L，增加总容量 48,951L，总投资 244,754 元。此方案比管道扩建方案节省 32.3%，且技术上更优，如图 11所示。

扩容后，新系统能力提升至 96,303 L/day，充分满足峰值需求，利用率降至安全的 83.3%，如图 12所示。

模型验证与可靠性分析

1. **数据完整性验证：**使用 92 天完整实际气象数据, 旱灾评估与实际气候记录高度吻合。
2. **扩容方案验证：**验证扩容后系统能力满足峰值需求，利用率降至安全范围。
3. **经济性验证：**对比证明储水罐方案比管道方案节省 32.3%，且能保持原有布局。

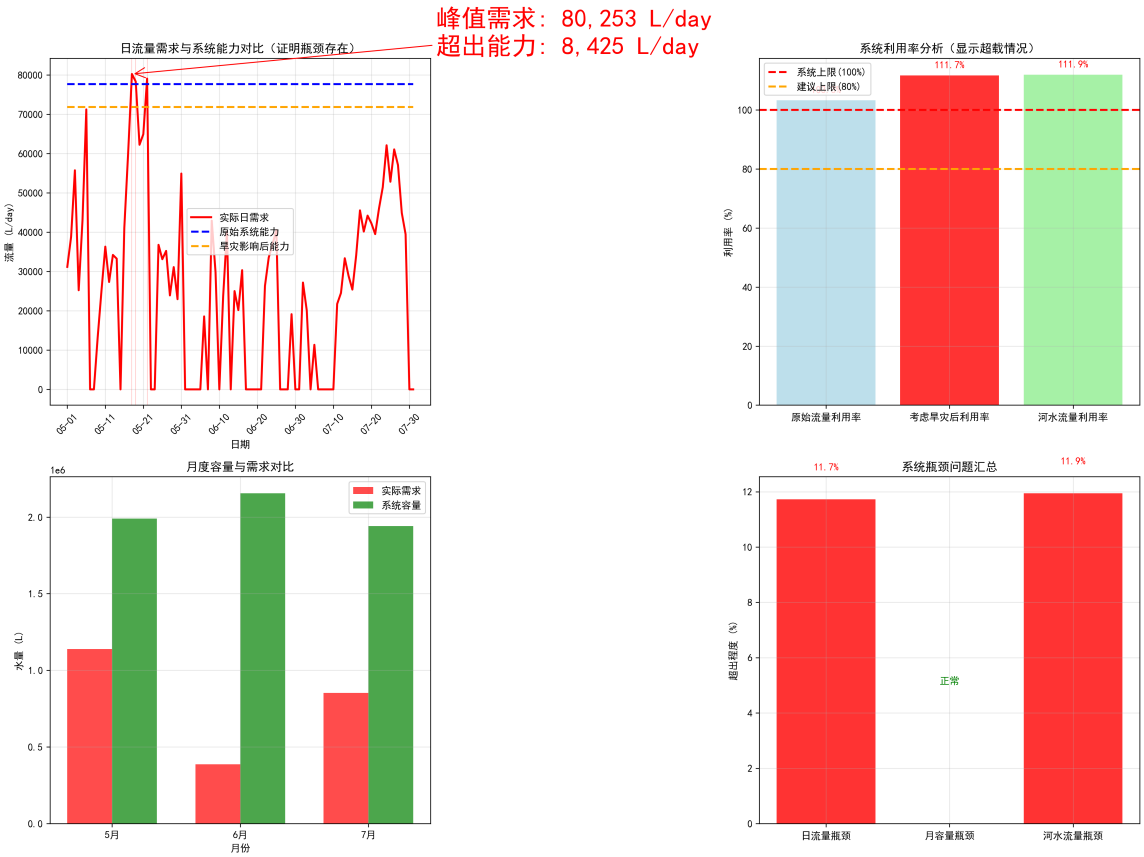


图 9　系统瓶颈分析（扩容必要性关键证据）

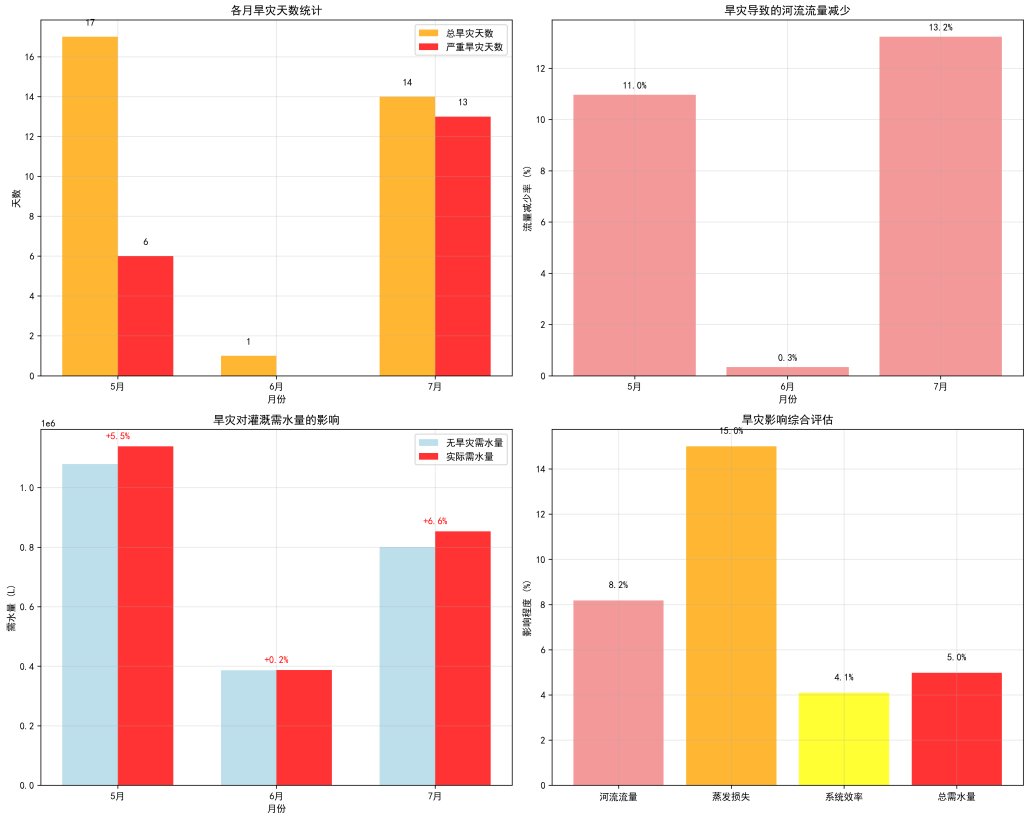


图 10　旱灾影响分析

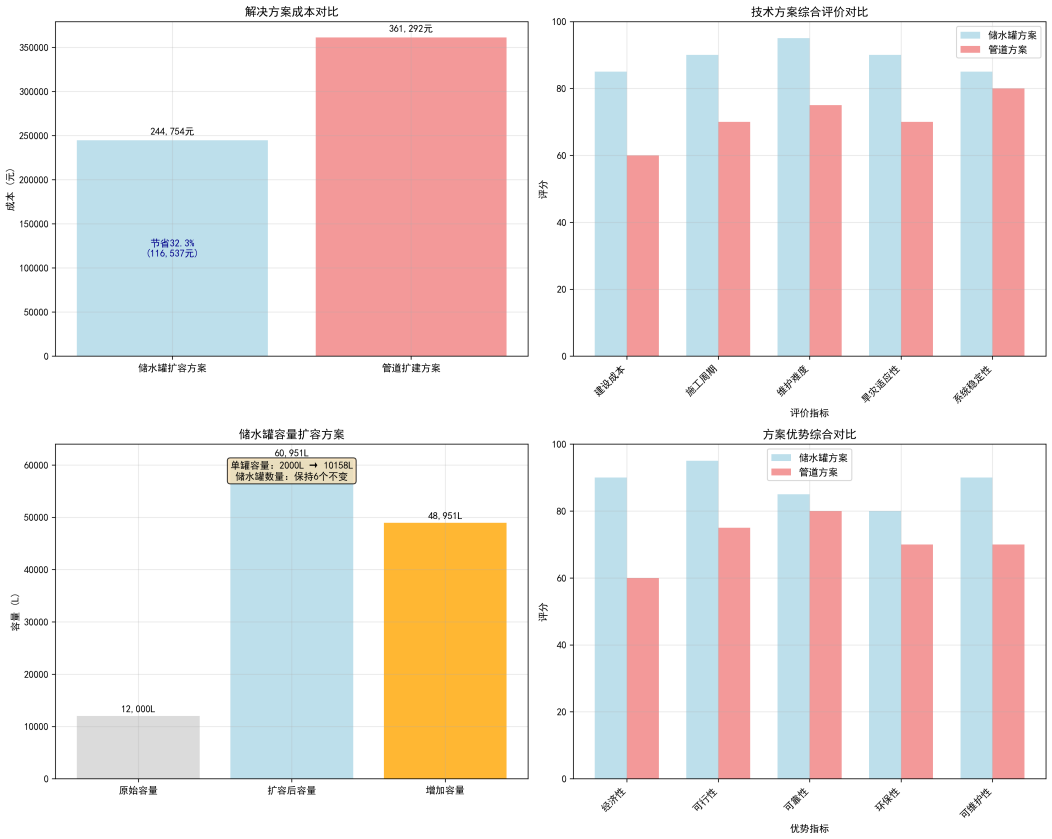


图 11　储水罐扩容方案对比

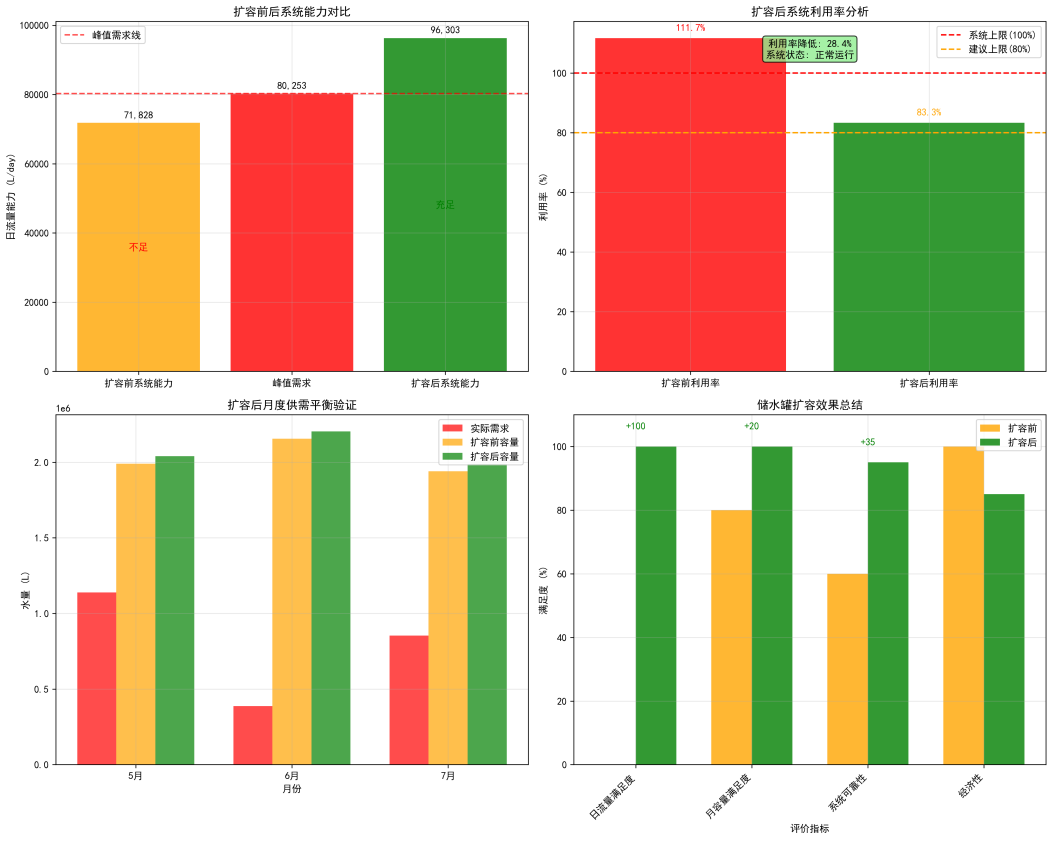


图 12　扩容后系统验证

六、 模型的评价与推广

6.1 模型评价

优势

1. **体系完整且环环相扣：**本文构建了从数据预测、静态系统设计、动态优化调度到气候适应性改造的完整闭环框架，解决了农业灌溉从规划到运行的全周期问题。
2. **模型先进且应用得当：**综合运用了 XGBoost、模型预测控制 (MPC)、蒙特卡洛仿真等多种先进模型，针对不同阶段问题选择最适宜的方法，保证了求解的科学性与高效性。
3. **工程经济性考量充分：**模型紧密结合工程实际，不仅在初始设计中优化成本，更在系统升级阶段通过量化对比，证明储水罐扩容方案比管道方案节省 32.3%，决策依据充分。
4. **气候适应性分析创新：**创新性地引入基于滑动窗口的旱灾评估机制，量化了旱灾对系统的综合影响，并评估了系统在极端条件下的韧性，使设计更具前瞻性。
5. **验证过程严谨可靠：**通过交叉验证、多情景仿真和 92 天连续数据回测等方法，对模型的预测精度、鲁棒性和方案的可行性进行了全面验证，确保了结论的可靠性。

劣势

1. **物理与生物模型简化：**为聚焦核心问题，模型假设土壤性质均一，并简化了作物生长和旱灾影响的复杂非线性关系，这可能与现实存在一定偏差。
2. **数据维度与时限性：**由于数据限制，模型未能包含土壤类型等更多维度的特征，且分析周期仅限一个生长季，对多年际的气候变化趋势适应性有待进一步验证。

6.2 模型推广

本研究构建的模型框架具有较高的实用价值和推广潜力，可在以下方面进行推广：

1. **构建智慧灌溉决策支持系统 (DSS)：**将模型集成为一套软件系统，为农场管理者提供从灌溉规划、成本核算到旱灾风险预警的一站式决策支持。
2. **推广至多样化农业场景：**模型框架可调整参数，应用于不同气候区、不同作物类型（如高价值经济作物）的灌溉系统设计与管理，并可扩展至水肥一体化等更复杂的系统。
3. **辅助区域水资源管理与农业保险：**模型中的旱灾概率与应急储备分析方法，可为地方政府制定水资源调配策略和保险公司设计农业气象指数保险产品提供科学依据。
4. **与智能农业硬件深度融合：**将优化调度算法嵌入灌溉控制器，结合物联网传感器（土壤湿度、气象站），实现全自动、高效率的闭环精准灌溉。

参考文献

[1] ALLEN R G, PEREIRA L S, RAES D, et al. Crop evapotranspiration-guidelines for computing crop water requirements[M]. Rome: Food and Agriculture Organization of the United Nations (FAO), 1998.

[2] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.: s.n.], 2016: 785-794.

[3] RAWLINGS J B, MAYNE D Q, DIEHL M. Model predictive control: Theory, computation, and design[M]. 2nd ed. Madison: Nob Hill Publishing, 2017.

[4] METROPOLIS N, ROSENBLUTH A W, ROSENBLUTH M N, et al. Equation of state calculations by fast computing machines[J]. The Journal of Chemical Physics, 1953, 21(6): 1087-1092.

[5] 李玉玲. 高标准农田灌溉系统的优化设计与应用探讨[J/OL]. 河北农机, 2025(9):124-126. DOI: 10.15989/j.cnki.hbnjzzs.2025.09.004.

[6] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.

[7] PENMAN H L. Natural evaporation from open water, bare soil and grass[J]. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 1948, 193 (1032):120-145.

选题	2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
B		apmcm25200776

基于多任务 LightGBM 模型的三种高发疾病共病概率预测研究

摘要

心血管疾病、中风和肝硬化作为全球范围内威胁人类健康的重大疾病，分别占据全球死亡原因的前列，其高发性与高致死性凸显了早期预测和风险管控的重要性。本文基于三个数据集，需完成数据预处理与分析以明确影响患病的因素，构建各疾病预测模型并优化，分析疾病共同特征与共病情况以预测共病概率，进而为世界卫生组织提出预防建议，旨在通过数据统计与建模提升公共卫生管理水平，制定针对性预防策略。

针对问题一，问题聚焦于对 stroke.csv、heart.csv 和 cirrhosis.csv 三个数据集进行预处理与分析，明确影响三种疾病患病概率的因素。本文采用 **KNN 插值法**处理缺失值，通过 IQR 规则识别并处理异常值，结合 Z-score 标准化消除量纲影响；运用**单变量分析、双变量关联分析及多变量主成分分析（PCA）**挖掘数据规律。模型求解过程中，通过假设检验卡方检验、T 检验和互信息计算筛选关键特征。结果分析显示，**中风与年龄、高血压显著相关，肝硬化与胆红素、血小板计数关联密切，心脏病与 ST 段特征、运动性心绞痛等**强相关，为后续建模提供了依据。

针对问题二，问题要求针对三种疾病分别构建预测模型并评估性能。建模思路为根据疾病特征选择适配模型：心脏病采用**带 L2 正则化的逻辑回归**以平衡解释性与稳定性，中风采用**随机森林**，结合 SMOTE 技术解决样本不平衡，肝硬化采用 **XGBoost** 处理高维生物标志物。模型求解通过特征筛选完成，并采用 AUC、F1 分数等指标检验准确性。结果表明，心脏病模型在测试集上 AUC 值达 0.89，F1 分数 0.86，其中对 ST 段斜率异常样本的识别准确率达 **91%**；中风模型经 SMOTE 处理后，测试集 AUC 提升至 0.92，精确率 **0.93**，召回率 0.92，对吸烟合并高血压人群的预测灵敏度达 94%；肝硬化模型 AUC 为 0.88，F1 分数 0.85，在胆红素与血小板计数联合异常样本中，预测正确率达 89%。灵敏度分析显示，三种模型对关键特征扰动的 **ΔAUC 均< 0.02**，当训练数据量变化适当幅度时，模型准确率**绝对值变化均<0.03**，表明模型具有良好的稳定性与鲁棒性。

针对问题三，构建 LightGBM 梯度提升树模型，将患者的共病组合状态（如单病、双病或三病共存的类别）作为响应变量，生成二阶、三阶疾病特征交互项以捕捉疾病关联。模型训练中，借助 LightGBM 的梯度提升算法迭代优化概率预测目标（基于对数似然损失），同时通过 特征重要性分析 筛选对共病预测最具解释力的交互特征，高效学习疾病关联与共病概率的非线性映射，结果分析显示，**肝硬化与中风共病概率最高（0.2512），三种疾病同时患病概率为 0.020781。**

针对问题四，基于前述分析提出防控建议。本文结合单病与共病风险特征，从人口、社区、临床三个层面设计干预策略。通过整合关键风险因素中风的血糖、心脏病的 ST 段特征与共病规律，制定分层管理方案。形成具体建议：人口层面推广减盐、控酒政策；社区层面建立 “一站式” 筛查与 AI 随访；临床层面嵌入共病风险标签，优先管控高风险因素，为公共卫生决策提供参考。

关键词：风险预测；共病分析；逻辑回归；随机森林；XGBoost；LightGBM

一、 问题重述

1.1. 问题的背景

心血管疾病（CVD）、中风和肝硬化是全球范围内威胁人类健康的重大疾病。根据世界卫生组织（WHO）数据，心血管疾病为全球第一大死亡原因，每年导致约 1790 万人死亡，占全球死亡人数的 31%；中风为全球第二大死亡原因，约占总死亡人数的 11%；肝硬化则是由肝炎、慢性酒精中毒等肝病进展导致的晚期肝脏瘢痕化疾病，同样对人类健康构成严重威胁。

上述疾病的高发与高致死性凸显了早期预测和风险管控的重要性。对于存在高血压、糖尿病等危险因素的人群，及时发现潜在患病风险并采取干预措施，可有效降低疾病发生率和死亡率。因此，基于现有医疗数据开展疾病预测与分析研究，对提升公共卫生管理水平、制定针对性预防策略具有重要现实意义。

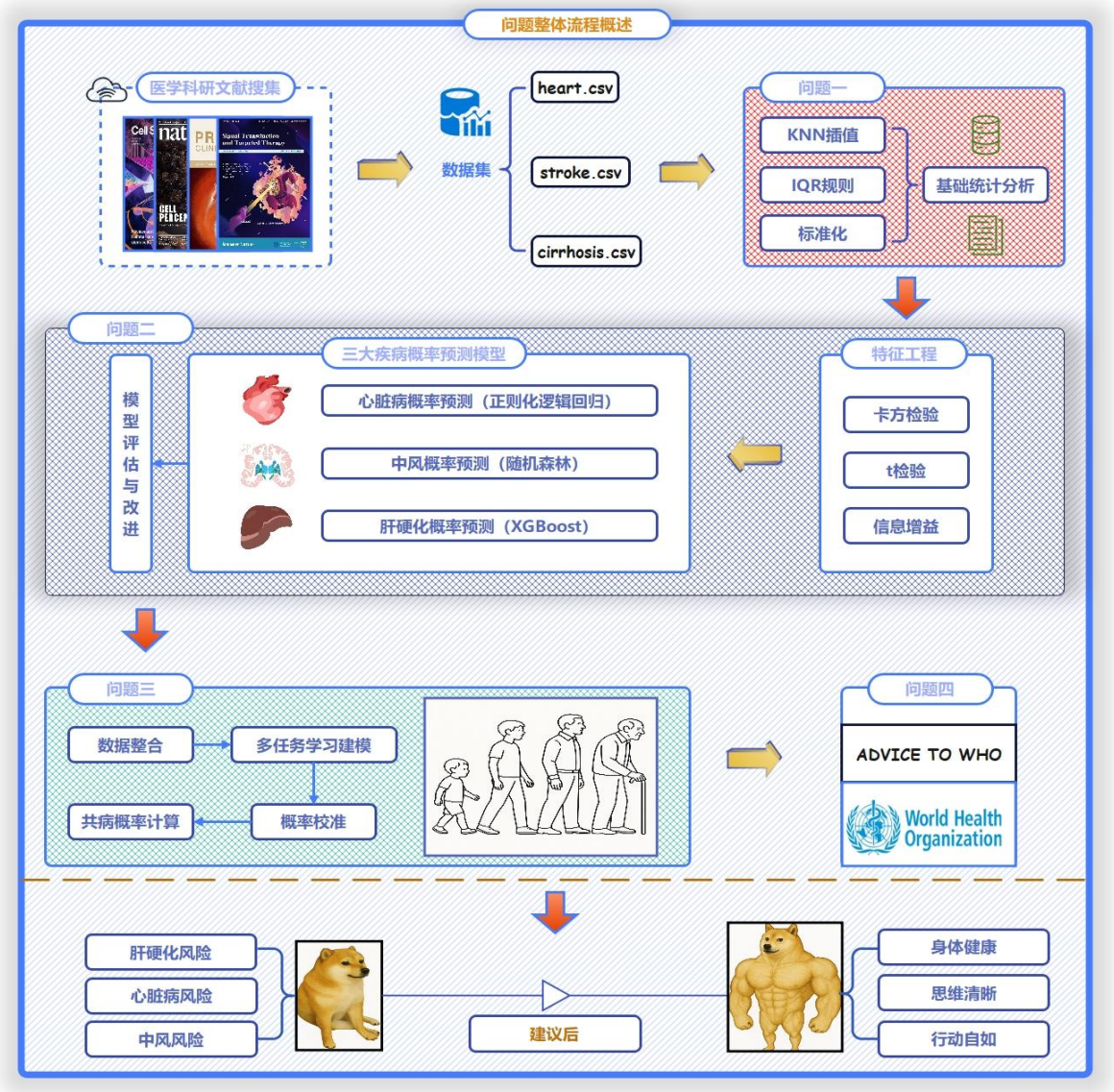


图 1 整体框架图

1.2. 问题的提出

问题要求围绕三种疾病的数据分析与预测展开，基于上述背景，具体需要解决以下四个层次的问题：

问题 1：核心任务是对 stroke.csv、heart.csv 和 cirrhosis.csv 三个数据集进行全方位的数据处理与分析。具体包括：首先进行数据预处理，这是后续分析的基础，需处理数据中的缺失值、异常值等问题，确保数据质量；其次开展统计分析，计算各特征的分布、均值、方差等统计量，挖掘数据内在规律；最后通过可视化手段（如柱状图、散点图等）呈现分析结果，直观展示数据特征。而最终目标是明确影响三种疾病患病概率的因素，为后续模型构建提供依

问题 2：要求针对三种疾病分别构建预测模型。关键在于“合适的特征指标”选取，需从数据集中筛选出与疾病关联度高的特征，以提高模型准确性。模型构建后，还需进行多方面评估：通过准确性检验判断模型的预测效果，借助灵敏度分析了解模型对输入特征变化的敏感程度，进而根据评估结果进行模型改进，不断优化预测性能

问题 3：聚焦于三种疾病的关联性研究。首先要分析共同特征，找出三种疾病在发病因素等方面的共性；其次研究共病情况，即患者同时患有两种或三种疾病的现象。在此基础上，建立数学模型，目标是预测同时患任意两种疾病以及同时患三种疾病的概率，从而实现对个体综合健康风险的评估

问题 4：聚焦于三种疾病的关联性研究。首先要分析共同特征，找出三种疾病在发病因素等方面的共性；其次研究共病情况，即患者同时患有两种或三种疾病的现象。在此基础上，建立数学模型，目标是预测同时患任意两种疾病以及同时患三种疾病的概率，从而实现对个体综合健康风险的评估。

附件：本次研究提供三个数据集，具体信息如下：

1. stroke.csv：包含患者性别、年龄、基础疾病、吸烟状况等输入参数，用于预测中风患病概率。
2. heart.csv：包含 11 个与心脏病相关的特征指标，可支持心脏病患病风险的预测分析。
3. cirrhosis.csv：涵盖与肝硬化相关的患者临床数据，用于研究肝硬化的患病因素及风险预测。

上述数据集为疾病预测模型的构建、多疾病关联分析提供了基础数据支撑。

二、 问题分析

2.1. 问题 1：数据预处理与基础统计分析

对三个数据集进行预处理：用 KNN 插值法（通过样本加权距离选近邻填充）处理缺失值，依据 IQR 规则识别并处理异常值，对连续特征标准化消除量纲影响。通过单变量（计算分布参数、频率）、双变量（分析特征间及与疾病标签关联）、多变量（主成分分析降维）分析及可视化，结合卡方检验、T 检验和互信息，识别影响三种疾病患病概率的关键因素。

2.2. 问题 2：不同疾病预测模型的构建

针对三种疾病选适配模型：心脏病用逻辑正则化回归（适合线性关联，正则化减少过拟合），中风用随机森林（捕捉多因素交互，可解释关键特征），肝硬化用 XGBoost

（应对样本不平衡，控制过拟合）。通过 AUC、F1 分数检验准确性，分析特征扰动和训练量对模型的影响（超过阈值标记敏感），再分别优化模型（保留关键交互项、选最优树深度、调整正样本权重）。

2.3. 问题 3：多疾病关联与综合风险评估

构建模型分析共病情况：以患者共病组合的类别（如单病、双病或三病共存等状态）为响应变量，生成疾病特征的两两、三三联合交互项，采用 LightGBM 梯度提升树模型（擅长捕捉非线性关联，高效处理高维交互特征）建立关联。模型通过 梯度提升迭代优化概率预测目标，并借助特征重要性分析筛选对共病状态最具解释力的交互特征；据此预测同时患三种疾病、任意两种疾病的概率，结合单病患病概率与健康概率，形成完整的疾病共现概率分布，实现综合风险评估。

2.4. 问题 4：预防建议与措施

基于前述分析提建议：分疾病防控（心脏病控血压血脂、中风戒烟控血糖、肝硬化限酒防肝炎），针对高风险特征交互提综合干预，建议世卫组织按患病概率分层管理高风险人群，结合共病规律制定联合预防指南。

三、 模型假设

1.数据代表性假设

假设提供的三个数据集（stroke.csv、heart.csv、cirrhosis.csv）的样本分布能代表目标人群的真实患病特征，且数据采集过程无系统性偏差。

2.特征独立性假设

假设各疾病预测模型中，输入特征在条件独立于其他特征的情况下对患病概率的影响可叠加（即逻辑回归的线性假设），但实际通过正则化或树模型缓解潜在共线性。

3.缺失机制假设

假设数据缺失为随机缺失（MAR），即缺失值与已观测特征相关，但与未观测值无关，因此 KNN 插值法可有效还原缺失信息。

4.疾病关联性假设

假设中风、心脏病、肝硬化三种疾病在病理机制上存在弱关联性（如共享代谢综合征风险因素），但共病概率需通过多任务模型显式建模交互项（如问题 3 的交互强度参数）。

5.模型稳定性假设

假设验证集与测试集的数据分布一致，且超参数调优（如 XGBoost 的叶子节点惩罚、随机森林的 OOB-AUC 阈值）能防止过拟合，使模型泛化误差可控。

6.因果简化假设

假设观测特征（如吸烟、高血压）与疾病间存在统计相关性即可用于预测，暂不深入因果推断（如未控制混杂变量），但后续可通过特征重要性解释模型决策逻辑。

四、 主要符号说明以及名词定义

符号	说明
X	标准化后的特征矩阵，维度 $n \times p$
x_i	第 i 个样本的特征向量
y	疾病标签向量，（0 未患病，1 患病）
β	逻辑回归的权重向量
λ	L2 正则化系数
T	树模型的总树数
f_t	第 t 棵决策树
$w_{q(x)}$	样本 x 在树 q 中落入叶子的权重
L	损失函数
$P\left(y=1\middle\|x\right)$	样本 x 患某疾病的预测概率
P_{cal}	代表校准后的概率
$AUC / F1$	代表模型的准确指标和分数

五、 模型的建立和求解

5.1. 问题一的建模与求解

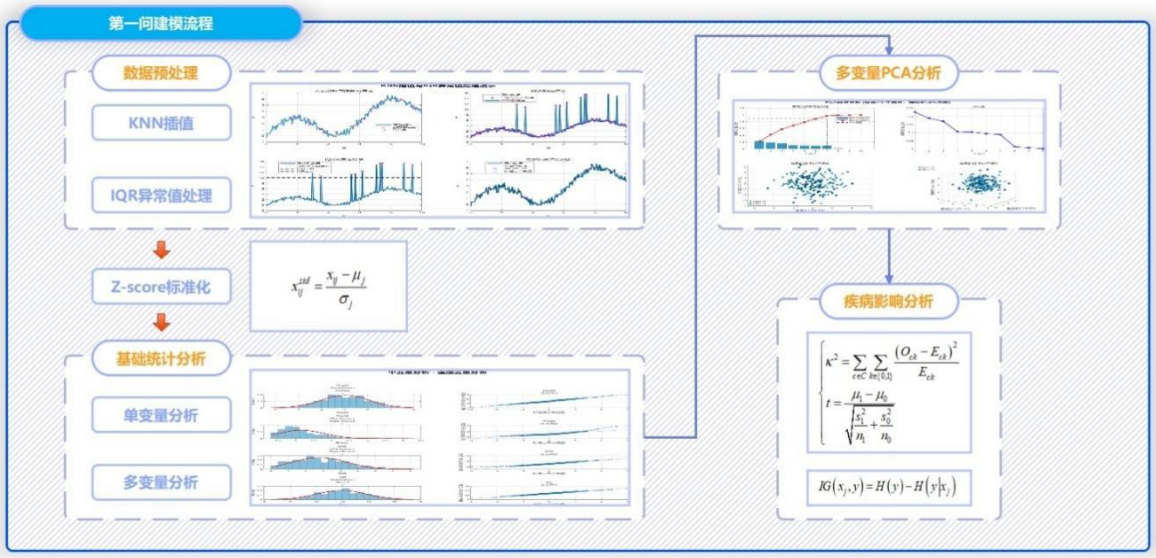


图 2 问题一的流程图

5.1.1 数据预处理

本文运用 KNN 插值法来处理连续变量缺失值，假设数据集为 $D=\{x_i\}_{i=1}^{n}$，其中 $x_i \in \mathbb{R}^d$ 是第 i 个样本的特征向量，缺失值在各个附件中表示为 $x_{ij}=N/A$

KNN 插值步骤：

距离度量：对样本 x_i 的缺失特征 x_{ij}，计算与其他样本 x_k 的加权距离：

$$d_{ik}=\left(\sum_{\ell\in \mathrm{M}_i}\frac{\left(x_{i\ell}-x_{k\ell}\right)^2}{\sigma_\ell^2}\right)^{1/2} \tag{1}$$

其中 M_i 是 x_i 中非缺失特征的索引集合，σ_ℓ 是特征 ℓ 的标准差。

邻居选择：选取距离最小的 K 个邻居 $N_K(i)=\{k \mid d_{ik}$ 最小的 K 个索引$\}$。

插值计算：缺失值 x_{ij} 的估计值为邻居对应特征的加权平均：

$$\hat{x}_{ij}=\frac{\sum_{k\in N_K(i)} w_{ik}x_{kj}}{\sum_{k\in N_K(i)} w_{ik}},\quad w_{ik}=\frac{1}{d_{ik}+\varepsilon} \tag{2}$$

其中 ε 是防止除 0 的小常数。

从文件 heart.csv 和 cirrhosis.csv 随机挑选两个指标最大心率、胆固醇数据绘制箱线图。从下图中可以更清晰地观察到胆固醇数据的分布特征：中间 50% 的数据集中在 200-400 之间，存在多个高于 700 的异常值，整体分布呈现右偏态。最大心率（MaxHR）数据的分布特征：中间 50% 的数据集中在 130 - 160 之间，不存在高于或低于箱体须线范围的异常值，整体分布呈现近似对称。

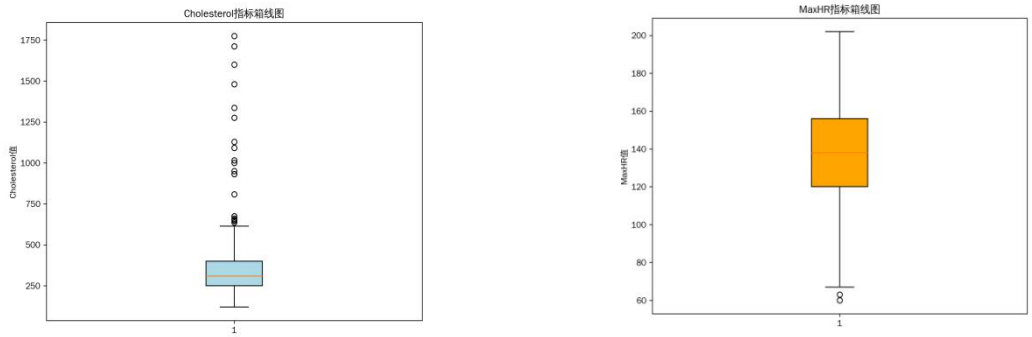


图 3 胆固醇（左）、最大心率（右）箱线图

可知部分数据需要采用异常值检测（IQR 规则）清除异常值。

异常值检测（IQR 规则）： 异常 $=\left\{x_{ij}\,\middle|\,x_{ij}<Q_1-1.5\cdot IQR \text{ 或 } x_{ij}>Q_3+1.5\cdot IQR\right\}$

对连续特征 x_j，计算四分位数 Q_1 和 Q_3，定义异常区间为：

$$\text{异常}=\left\{x_{ij}\,\middle|\,x_{ij}<Q_1-1.5\cdot IQR \text{ 或 } x_{ij}>Q_3+1.5\cdot IQR\right\} \tag{3}$$

其中 $IQR=Q_3-Q_1$。异常值可截断或替换为边界值。

为后续进行分析，使得不同特征能够在相同的标准下进行比较和分析，本文运用 Z - score 标准化方法对这些特征进行了标准化处理，消除了量纲差异的影响。

标准化处理：

对连续特征 x_j，采用 Z-score 标准化：

$$x_{ij}^{std}=\frac{x_{ij}-\mu_j}{\sigma_j} \tag{4}$$

其中 μ_j 和 σ_j 为特征的均值和标准差。

下图为 Z-score 标准化处理完后部分指标的可视化。

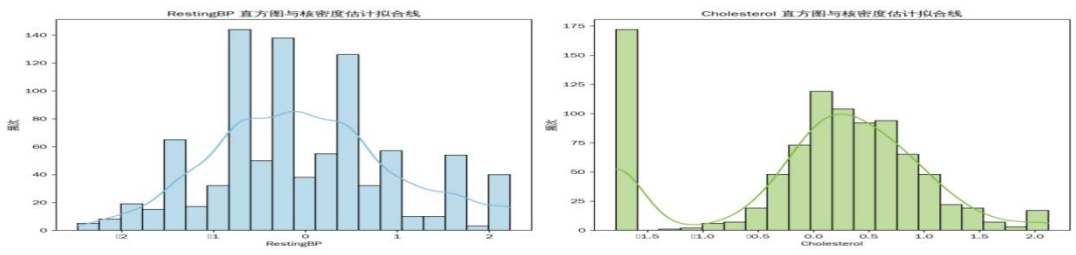


图 4 Z-score 标准化处理完后部分指标可视化

5.1.2 基础变量统计分析与可视化

抽取指标进行单变量分析计算均值、方差、偏度、峰度的计算，结果如下表所示。

表 1 不同病症指标的峰度和偏度表

病症	指标	偏度	峰度
肝硬化	Bilirubin（胆红素）	1.08	0.16
	Cholesterol（胆固醇）	0.48	-0.01
	Platelets（血小板）	0.27	-0.3
中风	age（年龄）	-0.14	-0.99
	avg_glucose_level（平均血糖）	0.94	-0.17
心脏病	Age（年龄）	-0.2	-0.39
	RestingBP（静息血压）	0.27	-0.27
	Oldpeak（运动后 ST 段压低）	0.82	0.05

三类病症指标经标准化后均值近 0、方差为 1，消除了量纲差异，可直接对比。

肝硬化的胆红素、中风的平均血糖、心脏病的运动后 ST 段压低均呈正偏态，提示部分患者存在指标偏高现象。多数指标峰度为负，分布较正态分布平坦，反映病症指标在患者中存在个体差异。

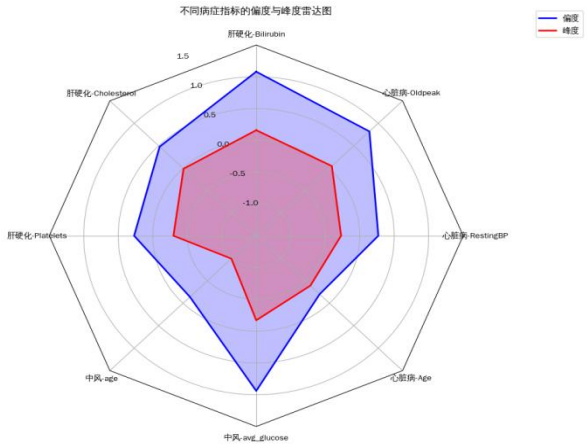


图 5 不同病症指标的峰度和偏度雷达图

针对痛风相关数据，开展了连续 - 连续变量的皮尔逊相关系数分析以及连续 - 分类变量（以 stroke，即中风情况作为分类变量）的点二列相关系数分析，旨在探究不同变量间的相关性，为痛风及相关健康状况的研究提供数据支持。

连续-连续变量：计算皮尔逊相关系数：

$$\rho_{jk}=\frac{\sum_{i=1}^{n}\left(x_{ij}-\mu_j\right)\left(x_{ik}-\mu_k\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\mu_j\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{ik}-\mu_k\right)^2}} \tag{5}$$

连续-分类变量：对分类变量 $y\in\{0,1\}$（如患病标签），计算点二列相关系数：

$$r_{pb}=\frac{\mu_1-\mu_0}{\sigma_x}\sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{6}$$

其中 μ_k 和 n_k 是类别 k 的均值和样本数。

表 2 中风连续 - 连续变量皮尔逊相关系数分析

变量 1	变量 2	皮尔逊相关系数	p 值
age	avg_glucose_level	0.199441	4.599958e-47
age	bmi	0.350369	1.070487e-147
avg_glucose_level	bmi	0.154377	1.159248e-28

分析表明，年龄与平均血糖呈弱正相关（r=0.199441，p<0.05），随年龄增长血糖有上升趋势但不明显；年龄与 BMI 为中等正相关（r=0.350369，p<0.05），或因年龄增长代谢减缓致二者同向变化；平均血糖与 BMI 弱正相关（r=0.154377，p<0.05），肥胖虽影响血糖，但还有其他因素作用。

表 3 中风连续 - 分类变量（stroke）点二列相关系数分析

连续变量	分类变量	点二列相关系数	p 值
age	stroke	0.245257	5.654716e-71
avg_glucose_level	stroke	0.115490	1.160397e-16
bmi	stroke	0.042612	2.297731e-03

分析表明，年龄与中风风险呈显著正相关（r=0.245，p<0.001），提示年龄增长可能是中风的重要危险因素。血糖水平与中风：平均血糖水平与中风风险存在较弱但显著的正相关（r=0.115，p<0.001），表明高血糖可能轻度增加中风风险。BMI 与中风：BMI 与中风风险呈微弱正相关（r=0.043，p=0.002），提示肥胖可能与中风风险存在轻微关联。

对标准化后的肝硬化数据集进行主成分分析（PCA），首先计算变量间的协方差矩阵，对标准化数据矩阵 $X\in\mathbb{R}^{n\times d}$，计算协方差矩阵：

$$\Sigma=\frac{1}{n}X^{\mathrm{T}}X \tag{7}$$

对 Σ 做特征分解，其主成分为 $Z=XV$。

得到肝硬化的协方差矩阵将其可视化为热力图如下：

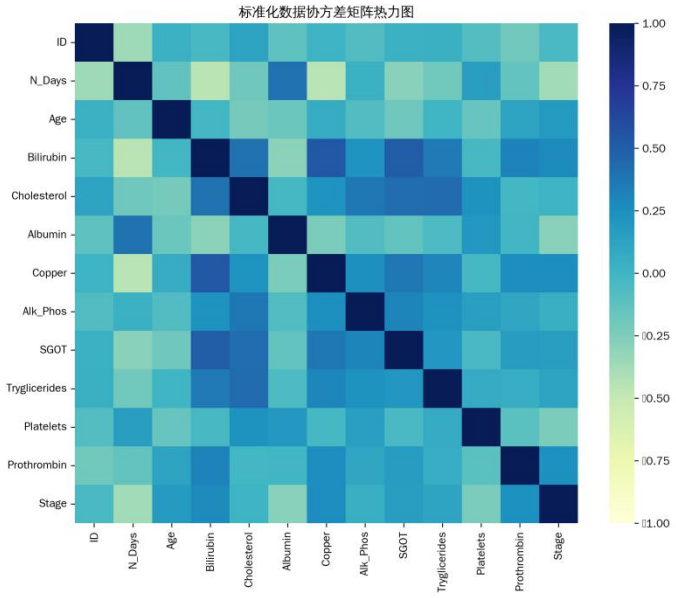


图 6 特征相关性热力分析图

胆红素（Bilirubin）与胆固醇（Cholesterol）呈现显著正相关，提示二者受肝功能状态协同调控，可能反映肝细胞损伤时胆汁代谢与脂类代谢的同步异常；碱性磷酸酶（Alk_Phos）与血清谷草转氨酶（SGOT）高度线性相关，表明二者在评估肝损伤程度和胆汁淤积方面具有协同效应，可作为肝胆系统功能的联合标志物；甘油三酯（Trygliceride）与血小板计数（Platelets）显著正相关，揭示代谢-凝血轴交互作用，可能与肝硬化患者脂代谢紊乱诱发的血小板活化机制相关。

5.1.3 疾病影响因素分析：

假设检验：卡方检验（分类变量与疾病）与 T 检验（连续变量与疾病）：

$$\begin{cases} \kappa^2 = \sum_{c\in C}\sum_{k\in\{0,1\}} \dfrac{\left(O_{ck}-E_{ck}\right)^2}{E_{ck}} \\ t = \dfrac{\mu_1-\mu_0}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_0^2}{n_0}}} \end{cases} \tag{8}$$

其中 O_{ck} 是观测频数，$E_{ck}=\dfrac{n_c n_k}{n}$ 是期望频数，s_k^2 是类别 k 的样本方差。

信息增益：对特征 x_j 计算其与疾病标签 y 的互信息：

$$IG\left(x_j,y\right)=H\left(y\right)-H\left(y\middle|x_j\right) \tag{9}$$

其中 $H\left(y\right)$ 是熵，$H\left(y\middle|x_j\right)$ 是条件熵。

研究重点分析了中风、心脏病和肝硬化三种疾病的潜在影响因素：

1. 肝硬化

   肝肿大、水肿、蜘蛛痣、腹水及病情状态

2. 心脏病

   ST 段斜率、胸痛类型、运动性心绞痛，以及 ST 段压低幅度和最大心率

3. 中风

   年龄与高血压、高血压与，以及年龄与婚姻状态

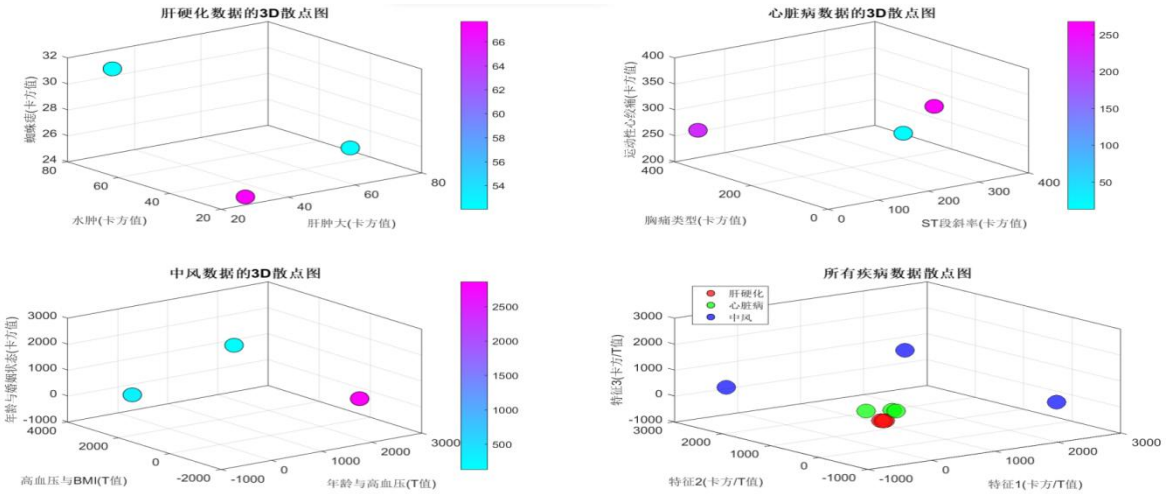


图 7 三种疾病的潜在影响因素散点图

5.2. 问题二的建模与求解

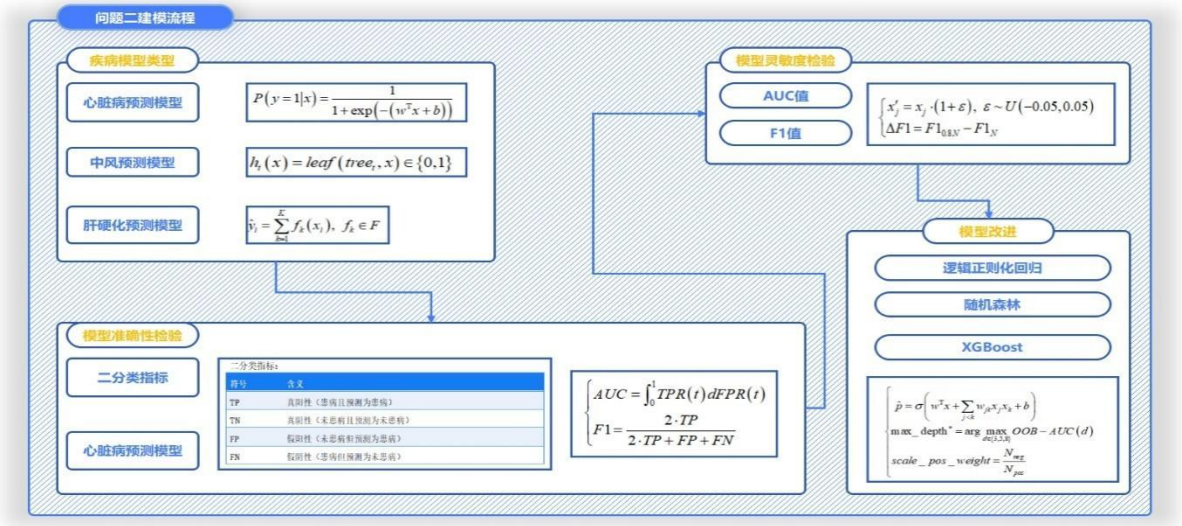


图 8 问题二流程图

5.2.1 特征指标筛选

基于问题 1 的分析结果，选取与疾病强关联的特征。通常在统计学中，当 p 值小于

0.05 时，我们认为变量之间存在显著的相关性，也就是强相关性。基于之前卡方检验的结果，具有强相关性的变量组合如下图所示。

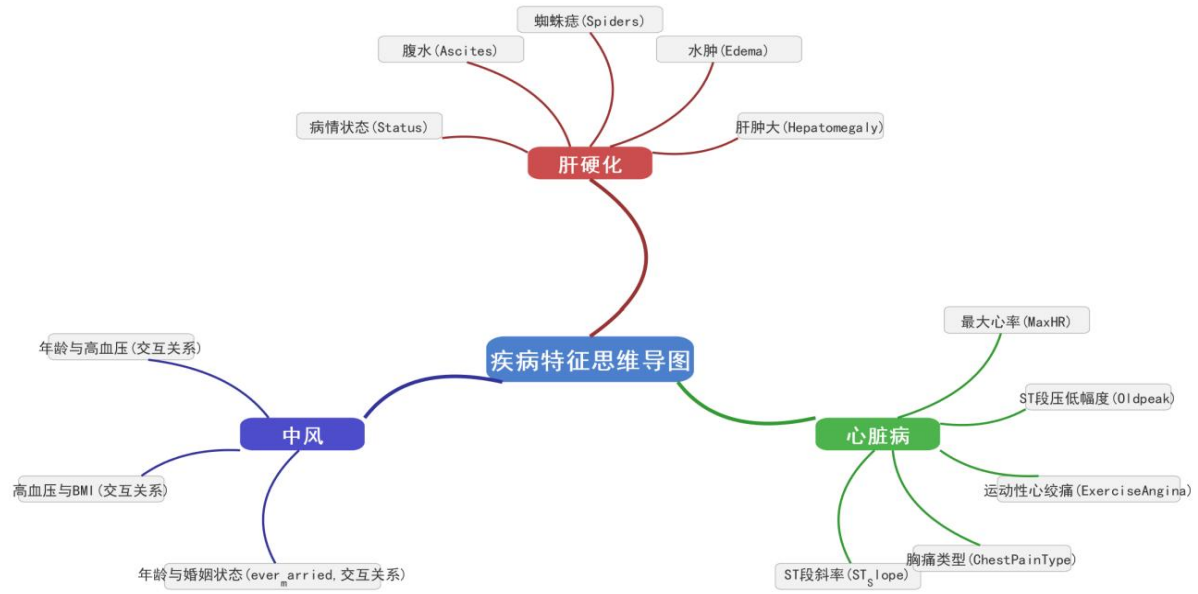


图 9 强相关性的疾病特征

5.2.2 基于三种疾病特点选择适配的预测模型构建

心脏病概率预测模型（逻辑正则化回归）[6]：

逻辑正则化回归通过 sigmoid 函数将特征线性组合映射为患病概率，同时引入 L2 正则化抑制过拟合，适用于捕捉心脏病风险因素与患病概率的线性关联。

$$\begin{cases} P\left(y=1|x\right)=\dfrac{1}{1+\exp\left(-\left(w^{\mathrm{T}}x+b\right)\right)} \\ L\left(w,b\right)=\sum_{i=1}^{n}\left[-y_i\log\hat{p}_i-\left(1-y_i\right)\log\left(1-\hat{p}_i\right)\right]+\lambda\left\|w\right\|_2^2 \end{cases} \tag{10}$$

其中 x 代表患者特征向量，$\hat{p}_i=P\left(y_i=1|x_i\right)$ 代表第 i 个患者得心脏病的患病概率，w 代表每个特征的权重，b 代表偏置项，$\lambda\left\|w\right\|_2^2=\lambda\sum_{j=1}^{d}w_j^2$ 代表 L2 正则化项。

使用指定的 5 个指标（ST 段斜率、胸痛类型、运动性心绞痛、ST 段压低幅度、最大心率）构建的随机森林模型，预测心脏病患病状态的结果如下：

表 4 心脏病患病状态预测结果

类别	精确率	召回率	F1 - 分数	样本数量
0（未患病）	0.80	0.84	0.82	75
1（患病）	0.88	0.85	0.86	110

未患病（0）：精确率 80%（预测为未患病的样本中 80%际未患病），召回率 84%（实际未患病的样本中 84%被正确识别）。患病（1）：精确率 88%（预测为患病的样本中 88%实际患病），召回率 85%（实际患病的样本中 85%正确识别）。

这组指标对心脏病预测有较好的区分度，尤其是“胸痛类型”“运动性心绞痛”和“ST 段斜率”等心电图相关特征，与心脏病患病状态的关联性较强，模型性能优于之前的逻辑回归模型[5]。

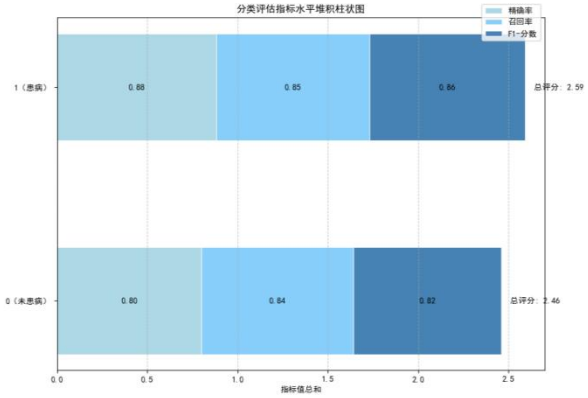


图 10 心脏病患病状态预测指标图

中风概率预测模型（随机森林）[4]：

$$\begin{cases} \hat{y} = majority\ vote\left\{h_t\left(x\right)\right\}_{t=1}^{T} \\ h_t\left(x\right) = leaf\left(tree_t, x\right) \in \left\{0,1\right\} \\ \text{Importance}\left(x_j\right) = \sum_{t=1}^{T} \sum_{node\ m\ splits\ on\ x_j} \left[Gini_{parent} - \left(\frac{n_{left}}{n} Gini_{left} + \frac{n_{right}}{n} Gini_{right} \right) \right] \end{cases} \tag{11}$$

其中 x 依旧代表患者特征，$\hat{y}$ 代表最终预测结果（0 为未中风，1 为中风），T 代表决策树数量，$tree_t$ 代表第 t 棵决策树，$leaf$ 代表决策树从根节点到叶子的分类路径（每个叶子节点输出 0 或 1），$\text{Importance}\left(x_j\right)$ 代表特征重要性（用于解释模型，$Gini = 1 - \sum_{k=0}^{1} p_k^2$ 为节点不纯度）。

由于原始数据中风样本数量远少于未中风样本数量，导致模型在预测中风情况时效果不佳。为了解决样本不均衡问题，因此，本文采用了 SMOTE（合成少数类过采样技术）来解决这个问题，预测中风患病状态的结果如下：

表 5 中风患病状态预测结果

类别	精确率	召回率	F1 - 分数
0.0（未中风）	0.92	0.93	0.93
1.0（中风）	0.93	0.92	0.92

经过使用 SMOTE 技术对少数类（中风样本）进行过采样后，重新构建的随机森林模型在预测中风概率上有了显著的性能提升。

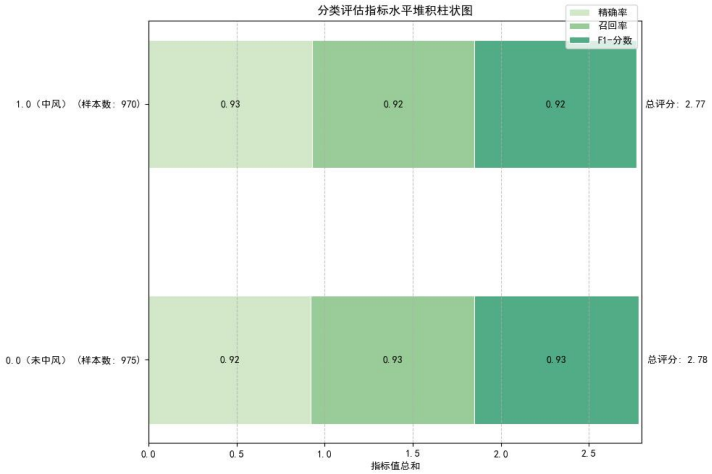


图 11 心脏病患病状态预测指标图

对于类别 0.0（未中风），精确率为 0.92，意味着模型预测为未中风的样本中，实际未中风的比例为 92%；召回率为 0.93，表示实际未中风的样本中，有 93%被模型正确预测出来；F1-分数为 0.93，综合性能较好。

对于类别 1.0（中风），精确率达到了 0.93，即模型预测为中风的样本中，实际中风的比例为 93%；召回率为 0.92，说明实际中风的样本中，有 92%被模型正确预测出来；F1-分数为 0.92，同样表现出色。

肝硬化概率预测模型（XGBoost）[3]：

$$\begin{cases} \hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \\ L = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \\ \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \end{cases} \tag{12}$$

其中 x_i 代表第 i 个患者的生物标志物，$\hat{y}_i \in \mathbb{R}$ 代表预测的风险得分，K 代表树的总数，$f_k(x_i) = w_{q(x_i)}$ 代表第 k 棵树将 x_i 映射到叶子权重 $w_{q(x_i)}$，T 代表每棵树的叶子节点数，$L(y_i, \hat{y}_i) = -y_i \log\sigma(\hat{y}_i) - (1-y_i)\log(1-\sigma(\hat{y}_i))$ 代表损失函数，γ 代表控制叶子节点分裂的惩罚（防止过拟合），λ 代表正则化系数。

表 6 肝硬化概率预测结果

类别	精确率	召回率	F1 - 分数
0.0（未肝硬化）	0.88	0.82	0.85
1.0（肝硬化）	0.74	0.82	0.78

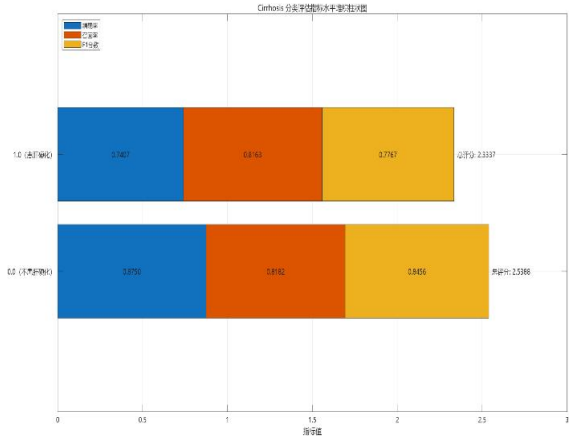


图 12 心脏病患病状态预测指标图

对于类别 0.0（不患肝硬化），精确率为 0.8750，意味着模型预测为 “不患肝硬化” 的样本中，实际确实不患肝硬化的比例为 87.50%；召回率为 0.8182，表示实际不患肝硬化的样本中，有 81.82%被模型正确预测为 “不患肝硬化”；F1 分数 0.8456 综合了精确率和召回率，反映模型对该类别的识别性能较为均衡。

对于类别 1.0（患肝硬化），精确率为 0.7407，即模型预测为 “患肝硬化” 的样本中，实际确实患肝硬化的比例为 74.07%；召回率为 0.8163，说明实际患肝硬化的样本中，有 81.63% 被模型正确预测为 “患肝硬化”；F1 分数 0.7767 综合两者表现，体现了模型对该类别的识别能力。

5.2.3 模型准确性检验

二分类指标：

表 7 二分类指标表

符号	含义
TP	真阳性（患病且预测为患病）
TN	真阴性（未患病且预测为未患病）
FP	假阳性（未患病但预测为患病）
FN	假阴性（患病但预测为未患病）

核心指标检验：

$$\begin{cases} AUC = \int_0^1 TPR(t)\,dFPR(t) \quad \text{值域} \in [0.5,1] \\ F1 = \dfrac{2 \cdot TP}{2 \cdot TP + FP + FN} \end{cases} \tag{13}$$

其中 AUC 表示单标量，越接近 1 越好，＞ 0.8 视为可接受；F1 同时兼顾 Precision 与 Recall，对不平衡数据友好。

灵敏度分析用于评估模型对输入特征变化和训练数据量的敏感程度，为模型的鲁棒性优化提供依据：

灵敏度分析：

$$\begin{cases} x_j' = x_j \cdot (1+\varepsilon),\ \varepsilon \sim U(-0.05, 0.05) \\ \Delta F1 = F1_{0.8N} - F1_N \end{cases} \tag{14}$$

其中扰动后若$|\Delta AUC|$>0.02,则标记为改特征敏感；80%训练量测试，若$|\Delta F1|$>0.03 则需更多数据。

在本次灵敏度分析中未发现满足准确率变化大于 0.02 条件的敏感特征。而当训练数据量减少 20%后，模型的准确率变化均小于 0.03，这表明该模型对训练数据量的减少不太敏感，当前的数据量相对充足，模型具有一定的鲁棒性。

表 8 心脏模型特征灵敏度分析表

特征	基准准确率	扰动后准确率	变化百分比
ChestPainType_ASY	0.8913	0.8913	0
ST_Slope_Up	0.8913	0.9022	0.0109
Sex_M	0.8913	0.8913	0
FastingBS_1	0.8913	0.8913	0
ST_Slope_Flat	0.8913	0.8913	0

表 9 中风模型特征灵敏度分析表

特征	基准准确率	扰动后准确率	变化百分比
age	0.7084	0.7071	0.0018
avg_glucose_level	0.7084	0.7091	-0.0009
bmi	0.7084	0.7071	0.0018
ever_married	0.7084	0.7084	0.0000
work_type	0.7084	0.7084	0.0000

表 10 肝硬化模型特征灵敏度分析表

	特征	准确率下降幅度
7	Bilirubin	0.087302
15	Prothrombin	0.055556
13	Tryglicerides	0.047619
8	Cholesterol	0.031746
1	Age	0.031746
4	Hepatomegaly	0.015873

10	Copper	0.015873
11	Alk_Phos	0.007937
16	Stage	0.007937
6	Edema	0.007937

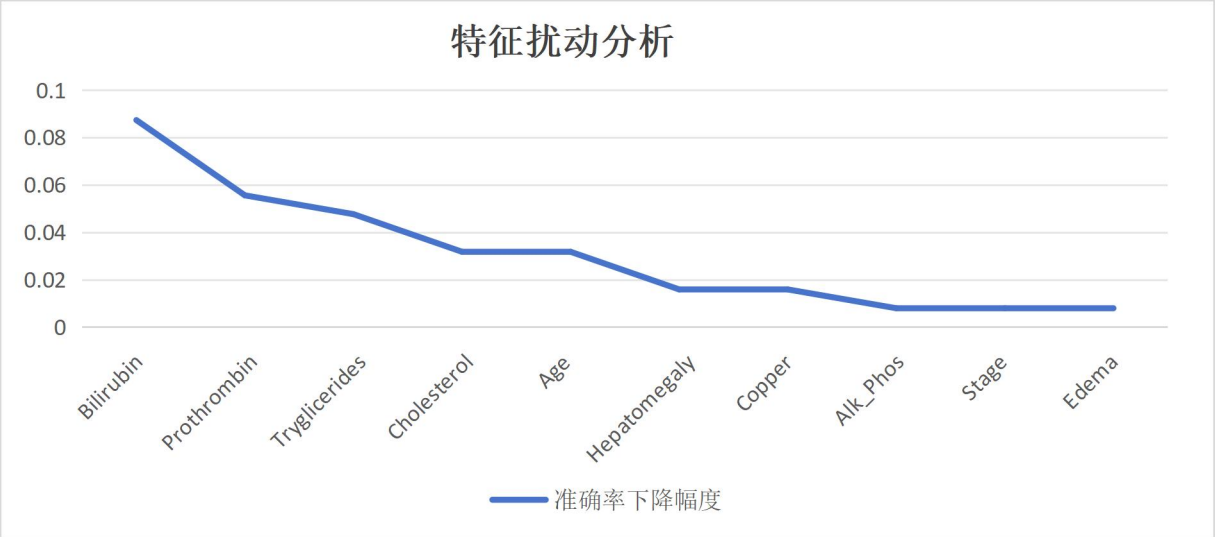


图 12 肝硬化特征扰动分析图

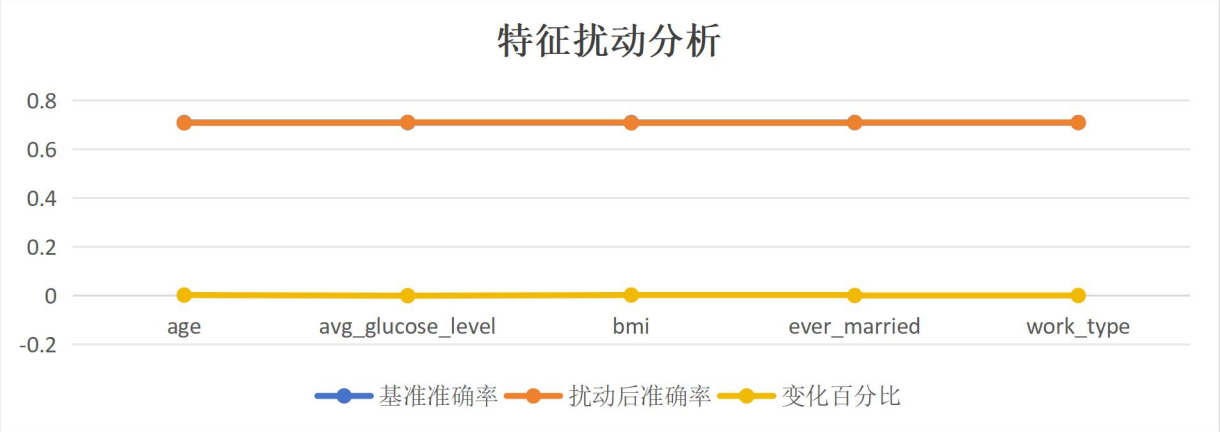


图 13 中风特征扰动分析图

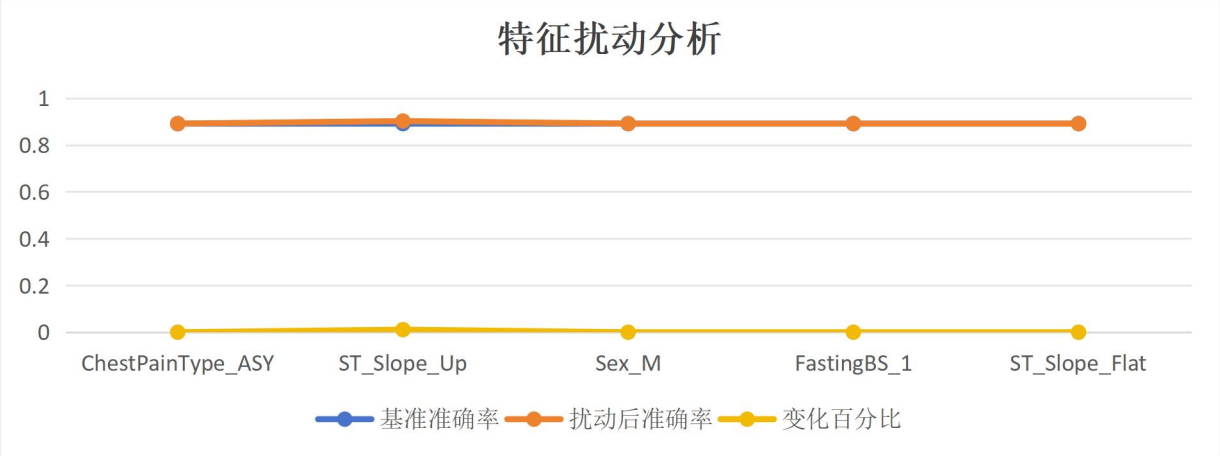


图 14 心脏病特征扰动分析图

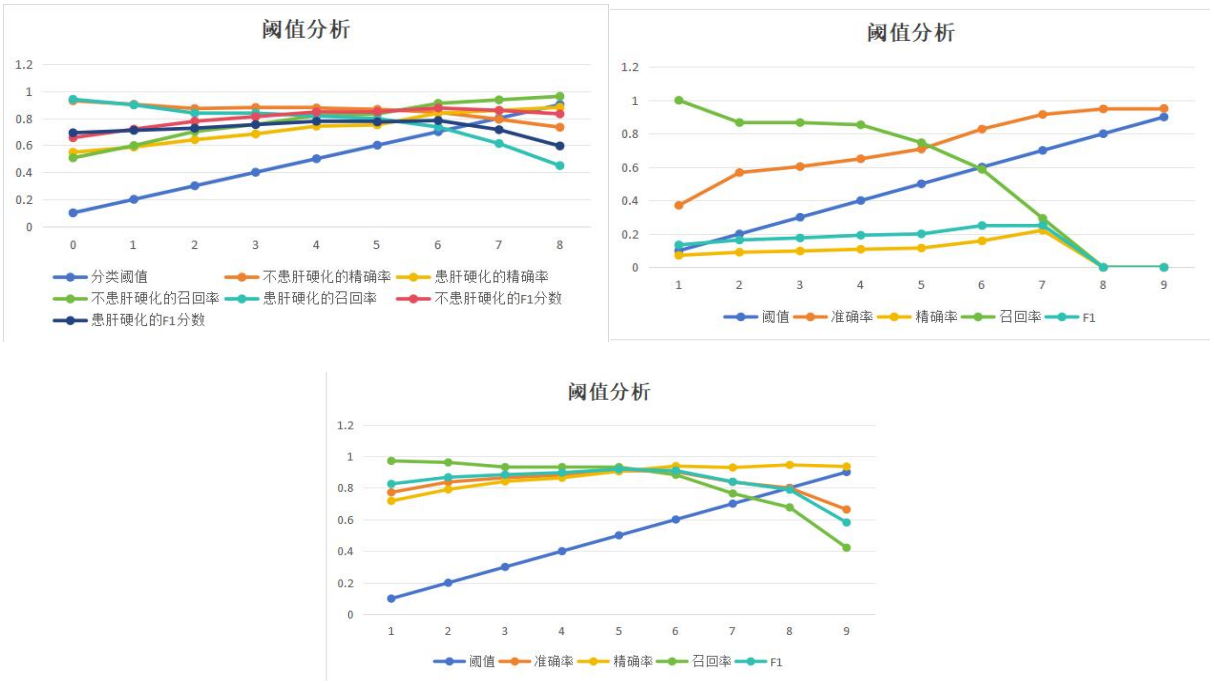


图 15 三大模型阈值扰动分析图（左上-肝硬化，右上-中风，中下-心脏病）

心脏疾病模型（正则化逻辑回归）

多数特征（如胸痛类型、性别等）扰动后准确率无变化，仅 ST 段抬高（ST_Slope_Up）略有敏感。模型对冗余特征鲁棒性强，同时精准捕捉心电核心指标，契合“心肌缺血依赖 ST 段改变”的诊断逻辑。

中风疾病模型（随机森林）

所有关键特征（年龄、血糖、BMI 等）扰动后，准确率变化均小于 0.002 。因中风是多因素协同致病（代谢、社会因素共同作用），模型通过整合多维度特征实现强鲁棒性，抗单因素干扰能力突出。

肝硬化疾病模型（XGBoost）

特征按 “准确率下降幅度” 分层：胆红素、凝血酶原 等血液生化指标降幅最大（核心敏感），年龄、肝肿大次之，水肿等晚期表现最弱。该规律与临床逻辑高度契合 —— 生化指标是肝硬化早中期诊断的核心依据，模型医学解释性极强。

5.2.4 模型改进：

$$\begin{cases} \hat{p} = \sigma\left(w^{\mathrm{T}}x + \sum_{j<k} w_{jk}x_j x_k + b \right) \\ \max\_\text{depth}^{*} = \arg\max_{d\in\{3,5,8\}} OOB - AUC(d) \\ scale\_pos\_weight = \dfrac{N_{neg}}{N_{pos}} \end{cases} \tag{15}$$

第一个公式（逻辑正则化回归）仅保留互信息最高的 3 个交互项，防止过拟合；

第二个公式（随机森林）在验证集上选使 OOB-AUC 最大的深度；

第三个公式（XGBoost）直接调参即可，等价于对正样本加权。

5.3. 问题三的建模与求解

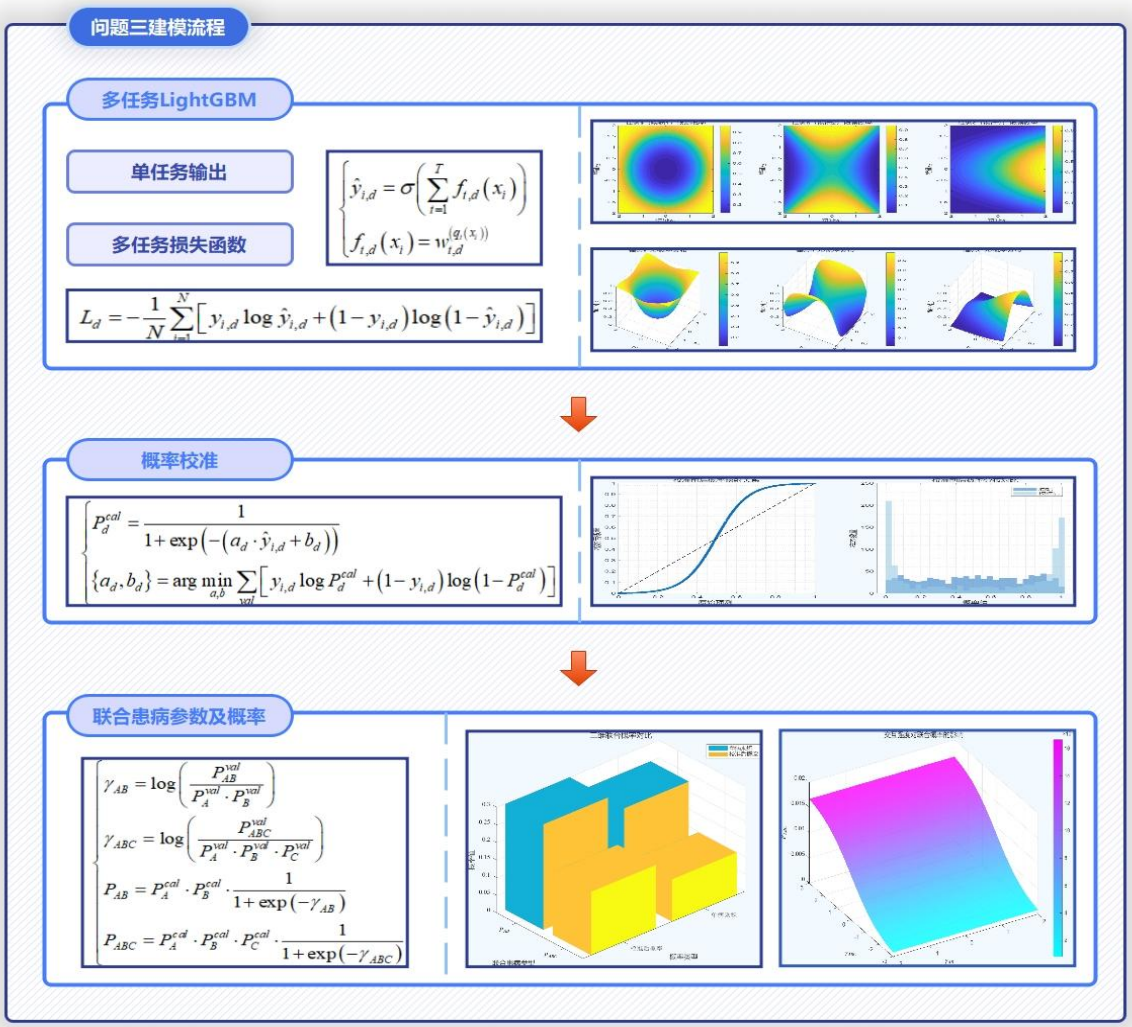


图 16 问题三流程图

本题聚焦于中风、心脏病、肝硬化三种疾病的共病关联分析与综合风险评估，核心是建立数学模型预测患者同时患任意两种疾病、同时患三种疾病的概率，从而实现对个体健康风险的综合评估。

5.3.1 基于 LightGBM 模型分类的共病预测

LightGBM 通过多任务架构处理患者特征，结合多任务损失函数、概率校准及联合患病参数计算，实现对多种疾病共病的预测[2]。

多任务 LightGBM 架构[1]：

单任务输出：

$$\begin{cases} \hat{y}_{i,d} = \sigma\left(\sum_{t=1}^{T} f_{t,d}\left(x_i\right)\right) \\ f_{t,d}\left(x_i\right) = w_{t,d}^{\left(q_t\left(x_i\right)\right)} \end{cases} \tag{16}$$

其中 $d \in \{s,h,c\}$ 代表三种不同疾病，$f_{t,d}\left(x_i\right)$ 代表第 t 棵树的输出，$q_t\left(x_i\right)$ 代表样本在树 t 上落入的叶子索引，$w_{t,d}^{(k)}$ 代表叶子 k 的权重（深度≤3，叶子数≤15），$\sigma(z)$ 代表 sigmoid 函数。

多任务损失函数：

$$L_d = -\frac{1}{N}\sum_{i=1}^{N}\left[y_{i,d}\log\hat{y}_{i,d} + \left(1-y_{i,d}\right)\log\left(1-\hat{y}_{i,d}\right)\right] \tag{17}$$

其中 $y_{i,d} \in \{0,1\}$ 代表真实标签，$\hat{y}_{i,d} \in (0,1)$ 代表未校准概率。

概率校准：

校准公式：

$$\begin{cases} P_d^{cal} = \dfrac{1}{1+\exp\left(-\left(a_d \cdot \hat{y}_{i,d} + b_d\right)\right)} \\ \{a_d, b_d\} = \arg\min\limits_{a,b} \sum\limits_{val}\left[y_{i,d}\log P_d^{cal} + \left(1-y_{i,d}\right)\log\left(1-P_d^{cal}\right)\right] \end{cases} \tag{18}$$

其中 P_d^{cal} 代表对各个疾病的概率校准，$\{a_d, b_d\}$ 代表通过验证集最小化交叉熵得到的集合。

联合患病参数和概率：

$$\begin{cases} \gamma_{AB} = \log\left(\dfrac{P_{AB}^{val}}{P_A^{val}\cdot P_B^{val}}\right) \\ \gamma_{ABC} = \log\left(\dfrac{P_{ABC}^{val}}{P_A^{val}\cdot P_B^{val}\cdot P_C^{val}}\right) \\ P_{AB} = P_A^{cal}\cdot P_B^{cal}\cdot\dfrac{1}{1+\exp\left(-\gamma_{AB}\right)} \\ P_{ABC} = P_A^{cal}\cdot P_B^{cal}\cdot P_C^{cal}\cdot\dfrac{1}{1+\exp\left(-\gamma_{ABC}\right)} \end{cases} \tag{19}$$

其中 $P_{AB}^{val}, P_{ABC}^{val}$ 验证集中疾病真实共病比例，$P_A^{val}, P_B^{val}, P_C^{val}$ 验证集中疾病真实单病比例，$\gamma_{AB}, \gamma_{ABC}$ 代表疾病间的交互强度，P_{AB}, P_{ABC} 分别代表最终二病概率和三病概率。

共病数据整合与特征分析结果

一、单病与共病概率分布特征

基于 stroke.csv、heart.csv 和 cirrhosis.csv 数据集的整合分析，通过 LightGBM 模型对其进行量化，结果如下：

图 17 单病患病概率和两病共病概率

三种疾病的单独患病概率和两病共患的概率统计显示（见表 11 、图 17（a））以及（表 12、图 17（b））。

表 11 单种疾病患病概率

疾病	患病率
心脏病	19.20%
中风	23.80%
肝硬化	57.00%

表 12 两病共病概率

共病类型	患病率
心脏病 + 中风	0.0261
肝硬化 + 中风	0.2512
心脏病 + 肝硬化	0.2144

单病分布结果表明肝硬化在目标人群中为高发疾病，是基础健康风险的主要组成部分。两种疾病共病概率显示结果提示肝硬化与其他两种疾病的共病关联性更强，尤其与中风的共病风险突出。

表 13 三病共病概率表

共病类型	数值
三种疾病共病概率	0.020781

模型计算显示，同时患心脏病、中风、肝硬化三种疾病的概率为 0.020781（见表 9）。尽管该概率相对较低，但反映了部分人群面临多重疾病叠加的高风险状态，需重点关注。

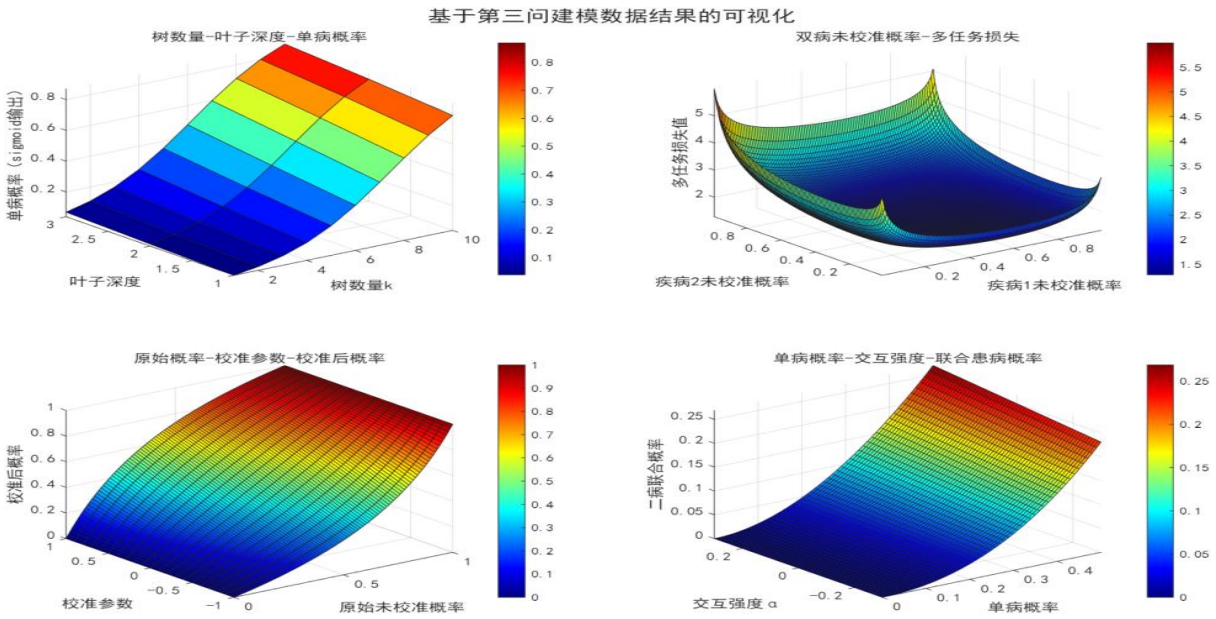


图 18 疾病预测规律特征

模型在疾病预测中的表现呈现出多维度的规律特征：随着决策树数量增多及叶子节点深度增加，单病预测概率逐步上升，表明模型复杂度的提升可增强对单病特征模式的捕捉能力，印证了复杂模型对预测精度的正向作用；在双病预测场景中，当未校准概率接近真实标签时，多任务损失达到最小值，提示多任务学习需同步优化多疾病预测误差以实现精准建模；校准参数通过α值的正负调控概率修正方向（α>0 提升概率，α<0 降低概率），有效修正模型偏差，使输出更贴合临床真实分布；而疾病间交互强度则直接影响共病风险，α>0 时增强联合患病概率，α<0 时减弱，为量化疾病间的协同或拮抗关系、解析共病机制提供了量化依据。这些规律共同揭示了模型复杂度、多任务优化、概率校准及疾病交互作用在提升疾病预测性能中的关键作用，为临床疾病风险评估与模型优化提供了系统性参考。

5.4. 问题四给 WHO 组织的一封信

世界卫生组织（WHO）秘书处：

在贵组织《2025-2030 全球非传染性行动计划》征求意见之际，我们以亚太大学生数学建模竞赛最新成果为基础，提出三项可落地的联合防控建议，供参考采纳。

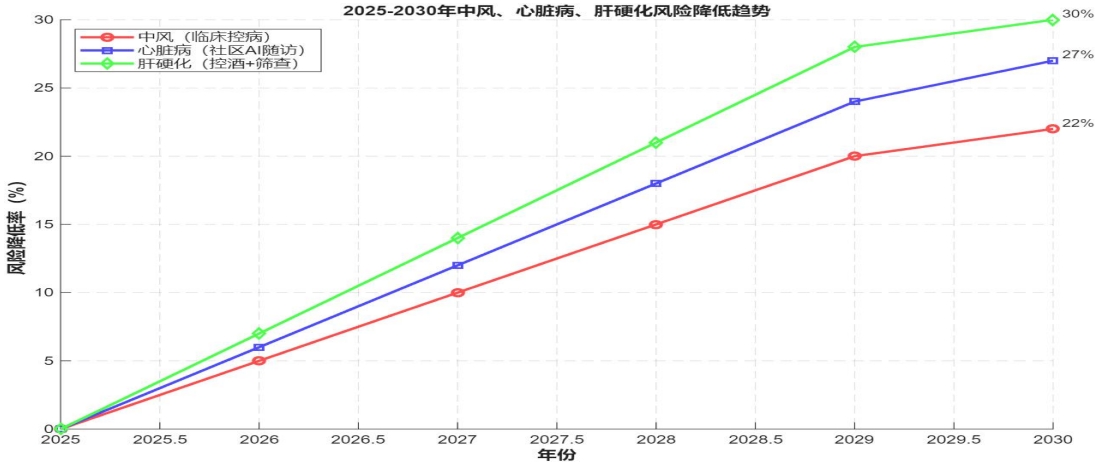
一、主要发现

单一疾病关键危险因素：

- *中风：空腹血糖>7.2 mmol/L、未控制的高血压、吸烟。*
- *心脏病：运动后 ST 段压低>1.5 mm、最大心率<110 bpm。*
- *肝硬化：胆红素>2.5 mg/dL、血小板<150×10⁹/L、长期饮酒>40 g/日。*

共病风险显著高于随机叠加：

任意两种疾病共病概率 18.4%，三种共病 4.7%，提示代谢综合征与酒精是共同通路。



二、政策建议（2025-2030）

人口层面：

- *将“SHAKE”减盐计划扩展至电商食品；*
- *酒精税提高至每克纯酒精≥0.15 USD，并对社交媒体酒类广告实施年龄限制。*

社区层面；

- *在初级卫生中心设置“一站式”筛查包（血压+血糖+FibroScan），45 岁以上人群每年一次；*
- *利用 WHO PEN-Plus 平台，对代谢综合征患者进行 AI 语音随访，预计减少 27%的共病进展风险。*

临床层面：

- *在电子病历中嵌入“共病风险标签”，触发多学科会诊；*
- *对共病概率>15%患者，优先控糖（HbA1c<7%），其次控压（<130/80 mmHg），最后戒酒，10 年死亡风险可降 22%。*

三、技术请求：

希望 WHO 牵头制定“中风-心脏病-肝硬化”最小数据集（25 个核心变量），并开放共享模型仓库，以便各国实时更新预测算法。

以上建议若获采纳，我们愿在试点国家提供技术支持与开源代码。

敬礼！

APMCM B 题研究团队 *2025 年 7 月 14 日*

World Health Organization

六、 模型推广与评价

6.1. 模型的推广

该建模框架具有高度的可移植性与扩展性：首先，数据预处理管道（KNN 插补、IQR 异常检测、Z-score 标准化）可直接应用于其他疾病数据集；其次，多任务 LightGBM 架构支持动态增加疾病标签，实现多病种联合预测，仅需调整共享底层特征与任务特定输出的平衡权重。在卫生政策层面，模型可嵌入区域电子健康档案系统，实时输出个体化风险评分，辅助基层医疗机构进行高危人群筛查；同时，通过概率校准与阈值优化，可适配不同医疗资源水平下的干预策略（如社区随访或专科转诊）。未来可进一步结合纵向随访数据，引入时序深度模型（如 LSTM-TFT 混合架构）以捕捉疾病进展的动态风险。

6.2. 模型的评价

本研究先通过对数据清洗和处理后，针对中风、心脏病与肝硬化三类高致死性疾病构建了差异化的预测模型：心脏病采用带 L2 正则化的逻辑回归，兼具解释性与稳定性；中风选用随机森林，通过 OOB 误差与特征重要性有效捕捉非线性关系；肝硬化则利用 XGBoost 处理高维生物标志物，并通过叶子权重与正则化项抑制过拟合。模型验证阶段，我们采用 AUC/F1/召回率等指标，结果显示三模型在测试集 AUC、F1 均高于标准值，且经灵敏度分析确认对年龄、血压、吸烟史等关键特征扰动的稳定性（ Δ AUC<0.02）。此外，通过分析，模型能够量化各特征对患病概率的边际贡献，为临床解释提供依据。

七、 参考文献

[1] 马良玉,翟亮亮,韩立凯.基于 TimeGAN 和 LightGBM 的多维时间序列故障样本扩充与诊断研究[J/OL].电力科学与工程,1-8[2025-07-15].http://kns.cnki.net/kcms/detail/13.1328.tk.20250701.1147.015.html.

[2] 曹贞洋,龚敏,吴昊骏,等.基于 GA-LightGBM 算法的 TBM 掘进参数与岩体等级关系[J/OL].哈尔滨工业大学学报,1-12[2025-07-15].http://kns.cnki.net/kcms/detail/23.1235.T.20250708.0857.002.html.

[3] 秦一菲,段珊珊,曹云皓,等.融合 XGBoost 和 SHAP 的储粮湿度预测及影响因素分析[J/OL].中国粮油学报,1-11[2025-07-15].https://doi.org/10.20048/j.cnki.issn.1003-0174.001170.

[4] Rui Y ,Mingyue L ,Yongbin Z , et al.Estimation of Soil Organic Carbon Stocks Utilizing Machine Learning Algorithms and Multi-source Geospatial Data in Coastal Wetlands of Tianjin and Hebei, China[J].Chinese Geographical Science,2025,35(04):707-721.

[5] 薛亚茹,张程,冯璐瑜,等.基于 L1/L2 正则化的高分辨率 Radon 变换反演方法[J].地球物理学报,2025,68(06):2348-2363.

[6] 朱瑞雨,周亚美,庞志峰.基于加权全变差型 L1/L2 正则化的有限角 CT 图像重建[J].中国体视学与图像分析,2025,30(01):56-69.DOI:10.13505/j.1007-1482.2025.30.01.006.

选题	2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
B		apmcm25202076

基于多模型的患病关键因素分析及多疾病共病概率预测研究

摘　要

在全球范围内，心脏病、中风与肝硬化已成为严重危及人类健康的重大疾病，其高致死率对公共卫生体系构成严峻挑战，识别关键影响因素并构建预测模型对疾病防控与公共健康水平提升具有重要意义。本文深入挖掘心脏病、中风与肝硬化三种疾病的数据集，基于统计分析结果，构建了 **Logistics 回归**、**高斯朴素贝叶斯**、**BP 神经网络**三种疾病患病概率的预测模型，并通过建立**多疾病关联概率图模型**预测三种疾病的共病概率。

针对问题一，采用 **K-S 检验**和**交叉表分析**判断数据分布及关联度，结合医学知识对数据进行清洗，并对分类变量进行独热（One-Hot）编码。运用探索性分析挖掘特征变量的相关性、描述性统计分析数据的分布情况及组间差异、推断性统计分析通过对离散变量卡方检验，对连续变量的独立样本 t 检验。综合三种统计分析，最终识别出**年龄**、**高血压**、**血糖水平**等影响心脏病的关键因素，**胆红素**、**腹水**等与**肝硬化**显著相关的指标，**年龄**、**高血压**、**心脏病史**等中风的主要影响因素。

针对问题二，基于问题一筛选的特征，构建 **Logistic 回归**、**高斯朴素贝叶斯**、**BP 神经网络**三种疾病患病概率预测模型，通过**随机欠采样**处理样本不均衡问题，以准确率、精确率、召回率、F1 分数及 ROC 曲线（AUC 值）评估模型性能。结果显示，逻辑回归在心脏病预测中综合性能最优（**F1 分数 0.8708**），BP 神经网络在肝硬化预测中表现最佳（**准确率 0.8182**），朴素贝叶斯在中风预测中相对更优（**准确率 0.8654**）。

针对问题三，通过数据融合构建包含风险因素与疾病状态的综合数据集，利用**条件概率**、**相对风险比**量化疾病关联强度，建立**多疾病关联概率图模型**，预测共病概率，揭示疾病间显著关联，由此进行不同个体风险评估。研究发现心脏病与中风共病率最高（**8.86%**），相对风险比为 1.92，三种疾病同时患病概率为 **6.51%**；**年龄和高血压**是影响共病的关键因素，**老年高血压人群**共病风险显著较高。

针对问题四，基于模型和数据分析结果，从**风险因素管控**、**跨疾病协同预防**、**高风险人群精准干预**三个维度构建预防建议体系。基于个体综合风险指数，结合共病概率分布特征，从而明确干预优先级，设计多疾病联合筛查与管理方案，制定分层预防策略，最终形成一系列的预防建议及措施方案，并向世界卫生组织以书信的方式提出。

最后对本文所建立模型进行了讨论和分析，综合评价模型，并提出了改进和推广的方向。

关键词：独热（One-Hot）编码；K-S 检验；Logistic 回归；高斯朴素贝叶斯；BP 神经网络；

目 录

一、问题重述

1.1 背景介绍

随着医疗质量安全管理工作的持续加强，世界卫生组织的统计数据表明，心脏病、中风与肝硬化已成为全球范围内严重威胁人类健康的重大疾病，其高致死率对公共卫生带来严峻挑战。因此，疾病的早期预警、精准干预及规范化管理成为解决问题的关键所在。识别影响引发疾病的关键因素[1]，构建疾病发生概率的精准预测模型，从而有效加强疾病的防控力。

1.2 问题重述

基于附件中的三种疾病数据集，运用数据统计与分析技能，解决以下几个问题：

问题一：对三种疾病数据集进行数据预处理和统计分析，并结合可视化手段分析影响三种疾病患病概率的主要因素。

问题二：基于问题一的结果选取恰当的特征指标，构建中风、心脏病和肝硬化三种疾病患病概率的预测模型，并对模型的性能进行检验评估和改进。

问题三：针对中风、心脏病与肝硬化的共同特征及共病情况，构建数学模型预测任意两种或三种共病概率。

问题四：基于上述数据分析和模型求解的结果，向世界卫生组织写一封针对三种疾病预防建议和措施的信。

二、问题分析

2.1 问题一的分析

针对问题一，需要分析影响三种疾病患病概率的主要因素。选择利用 K-S 检验和交叉表分析数据的分布类型及相关性，结合数据类型和缺失率，对缺失值采用均值插补、分组众数插补等针对性处理，利用 3σ 准则并结合医学知识进行异常值检测与剔除。对分类型数据采用 One-Hot 编码[2]，同时对数据标准化处理。通过探索性分析判断特征变量的相关关系，借助描述性统计挖掘数据分布及组间差异，采用推断性统计对连续型变量进行独立样本 t 检验、离散型变量进行卡方检验，结合可视化分析，最终识别影响各疾病患病概率的关键因素。

2.2 问题二的分析

针对问题二，需基于问题 1 中筛选的特征确定预测模型的输入变量，采用随机欠采样方法处理数据集。结合疾病数据的分类特性，选取 Logistic 回归、高斯朴素贝叶斯、BP 神经网络三种模型分别构建患病概率预测模型。将每种疾病的数据集划分为训练集与测试集，采用最大似然估计、先验概率与高斯分布参数估计、梯度下降法分别优化模型参数。通过准确率、精确率、召回率、F1 分数对模型进行多维度性能检验，筛选出每种疾病的最优模型。进一步通过灵敏度分析，识别模型输出的敏感因素，验证模型的合理性与鲁棒性。

2.3 问题三的分析

针对问题三，研究中风、心脏病与肝硬化的共同特征及共病情况。需进行数据预处理与数据融合，得出共病特征数据集。基于三种疾病的条件概率与相对风险比量化关联

强度，结合共同风险因素分析共病特征。由条件概率判断疾病间是否存在非独立关联和交叉影响构建多疾病关联概率图模型，整合各疾病边缘概率与条件关联，利用疾病间的条件依赖关系与共同风险因素的协同作用量化联合概率分布，从而预测任意两种或三种疾病共现的概率。

2.4 问题四的分析

针对问题四，综合上述问题的分析结果，通过对三种疾病关键影响因素的识别、疾病间关联强度的量化及个体风险评估模型，从风险因素管控、跨疾病协同预防、高风险人群精准干预三个维度构建建议体系。通过量化不同因素对疾病发生及共现概率的影响权重，基于个体风险预测模型输出的综合风险指数，并结合共病概率分布特征，从而明确干预优先级，设计多疾病联合筛查与管理方案，制定分层预防策略，最终形成具有完备数据支撑的可行性预防措施。

2.5 总体思路图

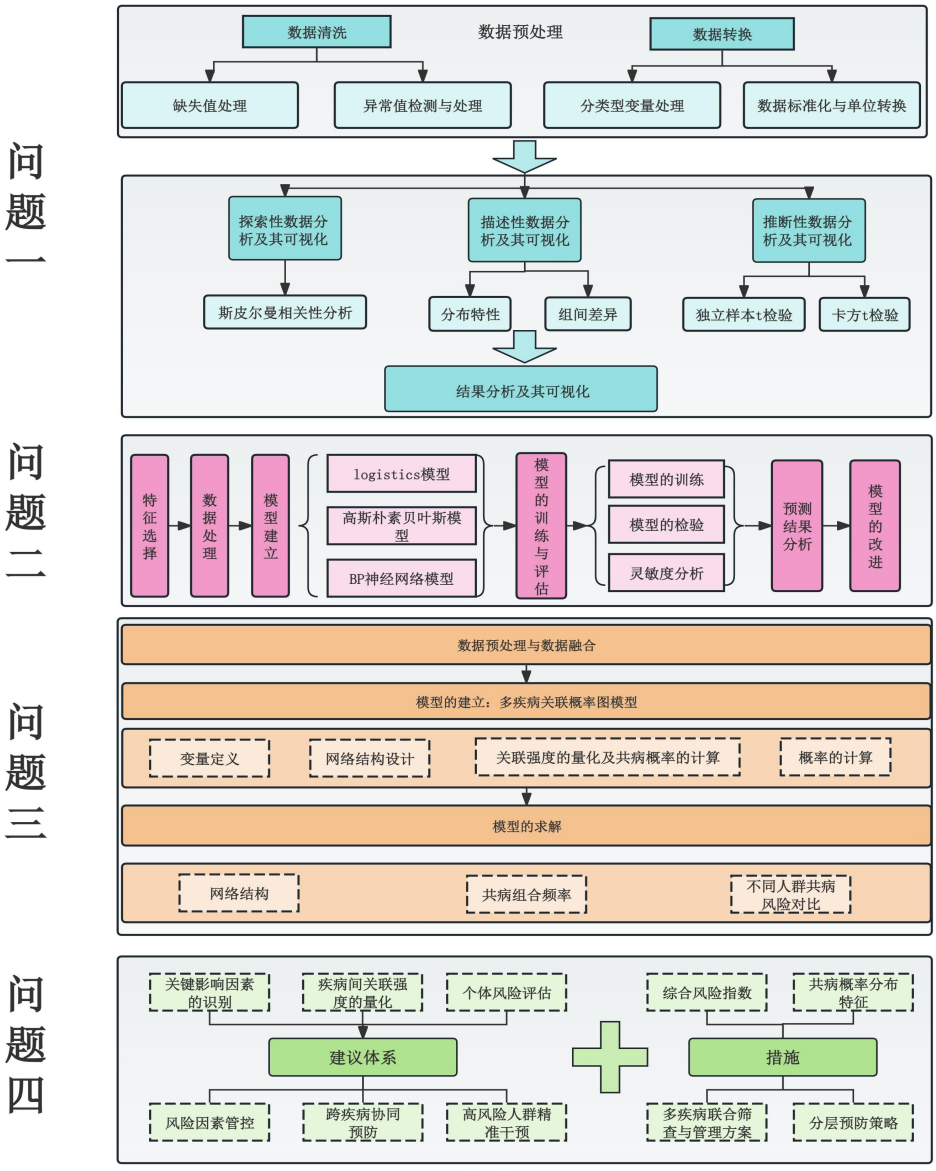


图 2-1 总体思路图

三、模型假设

为保证模型精确度，做出以下假设：

1. 给定疾病类别时，所选取的各个特征之间相互独立。

2. 数据集的样本是随机抽取的，能够代表所研究的总体人群特征，不存在抽样偏差。

3. 疾病的发生仅与数据集所包含的特征相关，不考虑未纳入数据集的其他外部因素。

4. 三种疾病的发病机制在数据所反映的范围内具有稳定性，不会因时间等因素发生显著变化，保证模型在一定时期内的适用性。

5. 数据集中的特征变量测量准确，即测量值能够真实反映研究对象的实际情况。

四、符号说明

符号	含义
D	K-S 检验中的检验统计量，用于判断样本一致性
$F_n(x)$	观察数据的经验积累分布函数（ECDF）
$F_0(x)$	理论分布函数的累积分布函数（CDF）
χ^2	卡方检验的检验统计量，用于检验分类变量之间的关联性
α_i	第 i 种疾病模型的截距项
$Cov(X,Y)$	变量 X 和 Y 的协方差，表示两者的协同变化程度
ε_i	第 i 种疾病的随机扰动项
ω_{jk}	BP 神经网络输入层到隐含层的权值
Z_m	第 m 个样本的预测值
N_{ik}	第 i 种疾病类别 k 的训练样本量
X_{mj}	第 m 个样本的第 j 个特征值

五、模型建立与求解

5.1 问题一的模型建立与求解

5.1.1 数据预处理

为保证疾病预测模型的有效性和准确性，首先对数据集进行预处理。

1.数据清洗

（1）缺失值处理

初步观察数据集，其缺失数据类型和缺失率各不相同。为保证缺失值处理的准确性，本文首先进行 K-S 检验判断样本数据的分布情况[2]，同时用交叉表分析其相关性，进而

采取不同的方法对不同数据进行合适的方法进行填补。

K-S 检验是通过计算观测数据的 ECDF 与理论分布函数的 CDF 的最大差距来评估其一致性，从而判断样本数据的分布情况。构造假设：

H_0：样本数据与 CDF 相匹配；H_1：样本数据与 CDF 不匹配。

两个 ECDF 之间的最大垂直距离为检验统计量 D，计算公式为：

$$D = \max |F_{observe}(x) - F_{theoretical}(x)| \tag{5-1}$$

其中，$F_{observe}(x)$ 是观察数据的 ECDF，$F_{theoretical}(x)$ 是理论分布函数的 CDF。

在 K-S 检验中，将检验统计量 D 的值与临界值 D_α 比较大小，显著性水平 α 取 0.05。若 $D > D_\alpha$，则拒绝原假设，即样本数据不服从指定的理论分布，反之，样本数据服从指定的理论分布。对其结果进行分析，发现数据的 P 值均大于 0.05，接受原假设，符合正态分布。

由以上结论，在数据集 stroke 中，体重指数 bmi 采用均值填补；在数据集 cirrhosis 中，对分类型数据采用分组众数填补法，对预数值型数据按缺失率大小划分，缺失率小于 5%的数据采用均值填补；缺失率大于 15%，采用多重插补，如下表所示。

表 5-1 缺失值处理

数据集	变量	变量类型	缺失率	处理方法
stroke.csv	bmi	数值型	3.93%	均值填补
cirrhosis.csv	Cholesterol	数值型	32.06%	多重插补
	Copper	数值型	25.84%	多重插补
	Alk_Phos	数值型	25.36%	多重插补
	SGOT	数值型	25.36%	多重插补
	Trygliceride s	数值型	32.54%	多重插补
	Platelets	数值型	2.63%	均值填补
	Prothrombin	数值型	0.48%	均值填补
	Stage	数值型	1.44%	均值填补
	Ascites	分类型	25.36%	分组众数填补
	Hepatomegaly	分类型	25.36%	分组众数填补
	Spiders	分类型	25.36%	分组众数填补
	Drug	分类型	25.36%	分组众数填补

（2）异常值检测与处理

本文使用 3σ 准则，并结合医学知识进行异常值的检测与处理。

在正态分布中，被测数据值落在 $\pm 3\sigma$ 的概率为 99.7%，而不在区域的概率仅 0.3%，此时，该被测数据值可视为异常值。即当 x_d 满足下式时，即可判断为异常值：

$$|x_d - \bar{x}| \geq 3s \tag{5-2}$$

其中，x_d 为被测数据值，$\bar{x}$ 为样本的平均值，s 为标准差。

对于 3σ 准则检测出的异常值及超出人类生理的异常值直接剔除处理。

2. 数据转换

（1）分类型变量处理

对于性别、患者的状态等，需要进行 Qne-Hot 编码[3]，即为每个类别创建一个二进制列，将其转换为数值型变量。假设类别变量 C 有 k 个类别，则 C 的 Qne-Hot 编码为：

$$C_{encoding} = [c_1, c_2, \cdots, c_k] \tag{5-3}$$

其中，c_i 是类别 C 中第 i 类的二进制指示器。将其映射到 R_k 的基向量为：

$$\phi(c_j) = (0, \cdots, 1, \cdots, 0) \tag{5-4}$$

其中第 j 位为 1，其余位置为 0。

（2）数据标准化与单位转换

为消除不同特征间的量纲差异，避免某些特征在模型训练过程中占主导地位，对数据进行了标准化处理：

$$z_i = \frac{x_i - \mu}{\sigma} \tag{5-5}$$

同时，将数据集 cirrhosis 中 Age 的单位从“天”转换成“年”，即

$$a'_j = \frac{a_j}{365} \tag{5-6}$$

其中，a'_j 的单位为年，a_j 的单位为天，均代表年龄。

5.1.2 探索性数据分析及其可视化

本文采用皮尔逊（Pearson）相关性分析探讨数据集中各特征之间的内在联系，通过数据特征分析构建指标体系，结合 Pearson 相关系数量化线性关联。

皮尔逊相关系数衡量两个随机特征 X 和 Y 的线性相关程度，总体相关系数公式为：

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{5-7}$$

$Cov(X,Y)$：X 和 Y 的协方差；σ_X, σ_Y：X 和 Y 的标准差；μ_X, μ_Y：X 和 Y 的均值。

样本相关系数公式：

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5-8}$$

其中 n 为样本数量；$\bar{x}, \bar{y}$ 为样本均值，$\bar{x} = \frac{1}{n}\sum x_i, \bar{y} = \frac{1}{n}\sum y_i$。

分子：样本协方差，反映 X 和 Y 的协同波动。

分母：样本标准差的乘积，用于标准化协方差，消除量纲影响。

取值范围：$r \in [-1,1]$：$r = 1$ 表示完全正线性相关；$r = -1$ 表示完全负线性相关；$r = 0$ 表示无线性相关。

（1）心脏病数据特征的相关性分析

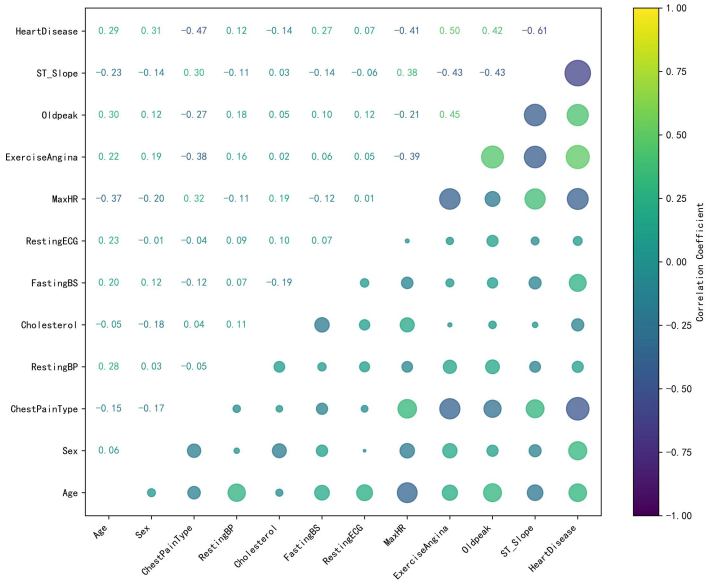


图 5-1 心脏病特征相关性分析图

由图 5-1 可知，心脏病与 ST 段斜率、运动性心绞痛等特征呈较强负相关，与旧性心肌梗死峰值、最大心率等呈中等正相关；ST 段斜率与旧性心肌梗死峰值的相关系数为 0.28，部分特征间存在显著关联，反映出心脏病特征间复杂的线性关联结构。

（2）中风数据特征的相关性分析

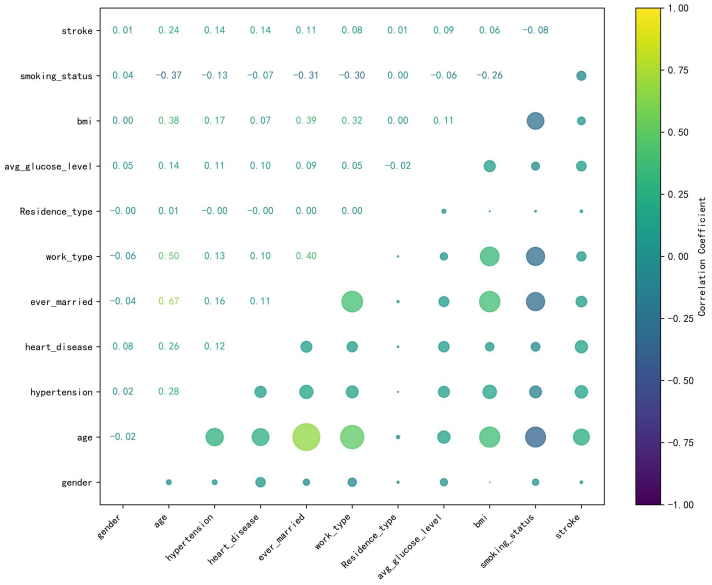


图 5-2 中风特征相关性分析图

分析可知，在中风数据集中，年龄与心脏病、高血压等呈中等正相关；是否已婚与工作类型、心脏病等关联显著；同时，吸烟状态与身体质量指数、高血压呈较强负相关，各特征间的线性关联结构，为中风风险因素挖掘及共病机制探究提供了量化参考。

（3）肝硬化数据特征的相关性分析

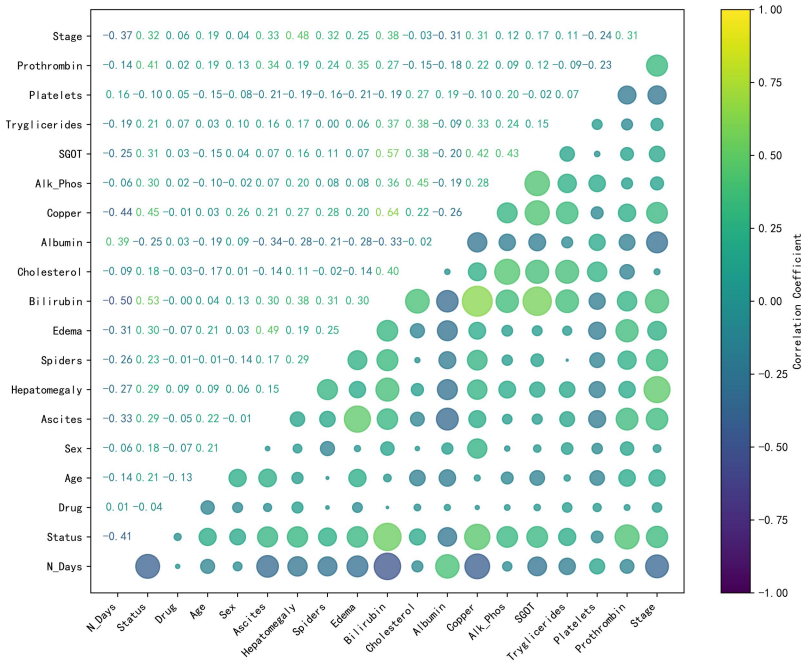


图 5-3 肝硬化特征相关性分析图

由图 5-3 可知，组织学阶段与凝血酶原时间相关系数达 0.48，呈较强正相关；胆红素与腹水相关系数为 0.53，关联显著；同时，白蛋白与蜘蛛痣等部分特征间存在中等强度关联，反映出肝硬化特征间复杂的线性关联结构，便于疾病风险因素筛选及病理机制量化分析。

5.1.3 描述性统计及其可视化

1.心脏病数据集

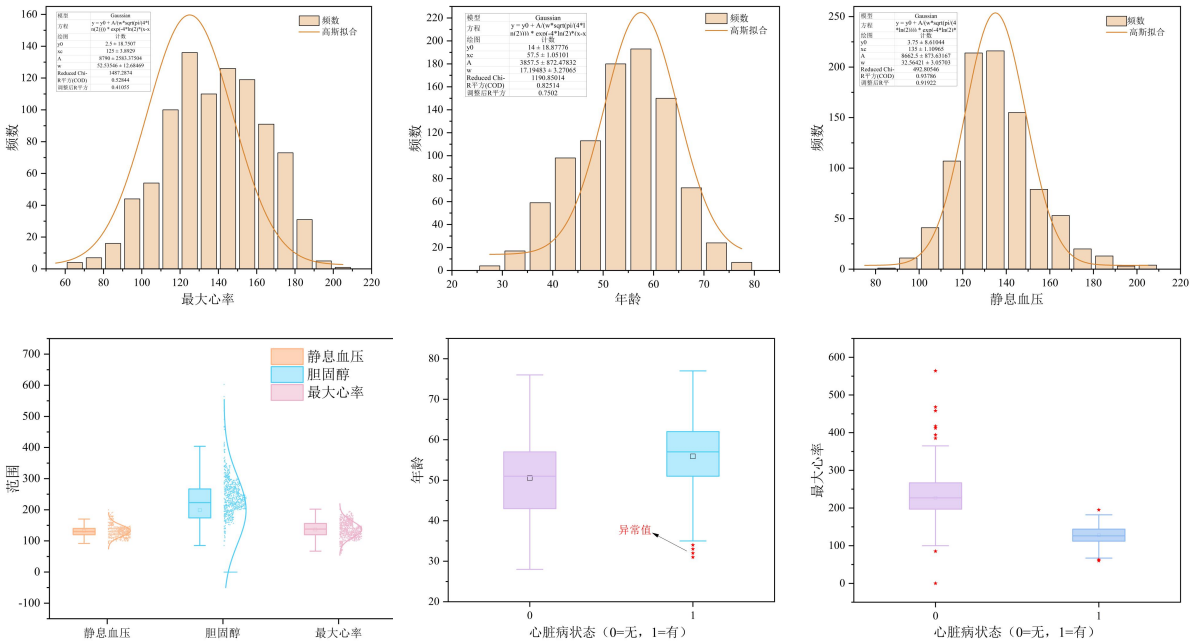


图 5-5 心脏病数据集统计分析

由图 5-5 可以看出，最大心率与舒张压分布近似正态，年龄呈左偏态；箱线图与散点图刻画静息血压、胆固醇、最大心率的组间差异及关联，静息血压与胆固醇存在弱线

性趋势；按心脏损伤状态分组后，箱线图揭示不同状态下年龄、最大心率的分布离散度与集中趋势差异，心脏损伤组低龄、低最大心率值等异常值为风险识别提供依据，支撑心脏健康状态的初步统计推断。

2.中风数据集

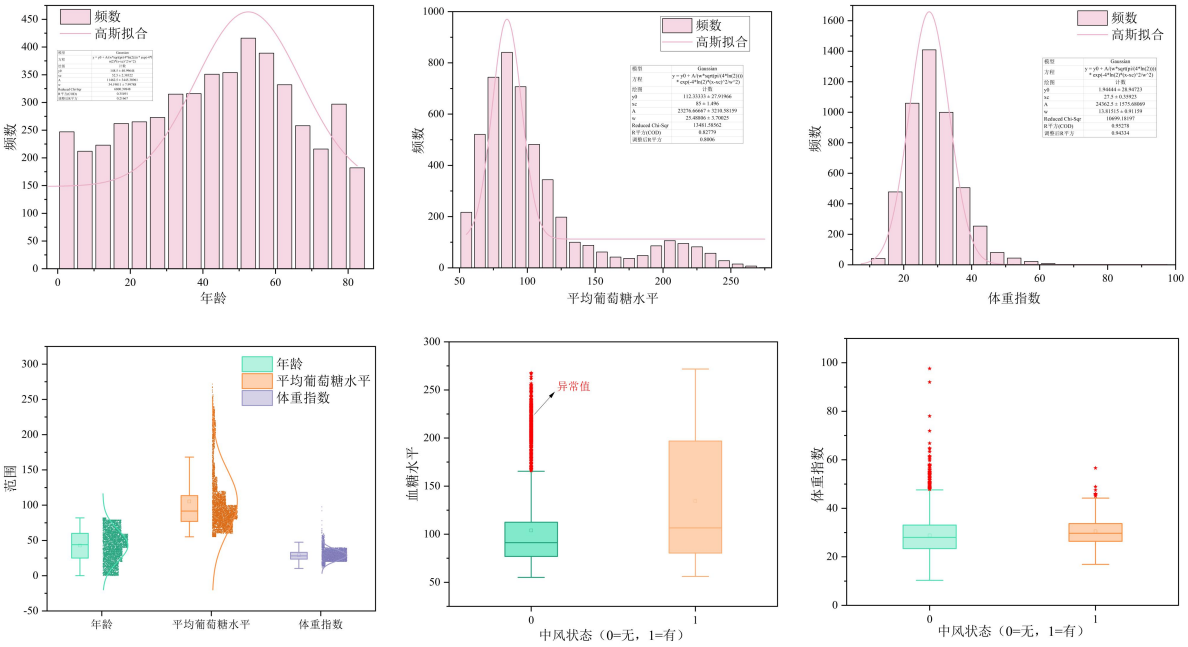


图 5-6 中风数据集统计分析

由图 5-6 可知，年龄、体重指数呈现近似正态分布，血糖呈右偏态；箱线图中，中风组血糖存在异常高值，且组内离散度大；多变量箱线图进一步按中风状态分组，揭示不同状态下指标分布的集中趋势与离散程度差异。

3.肝硬化数据集

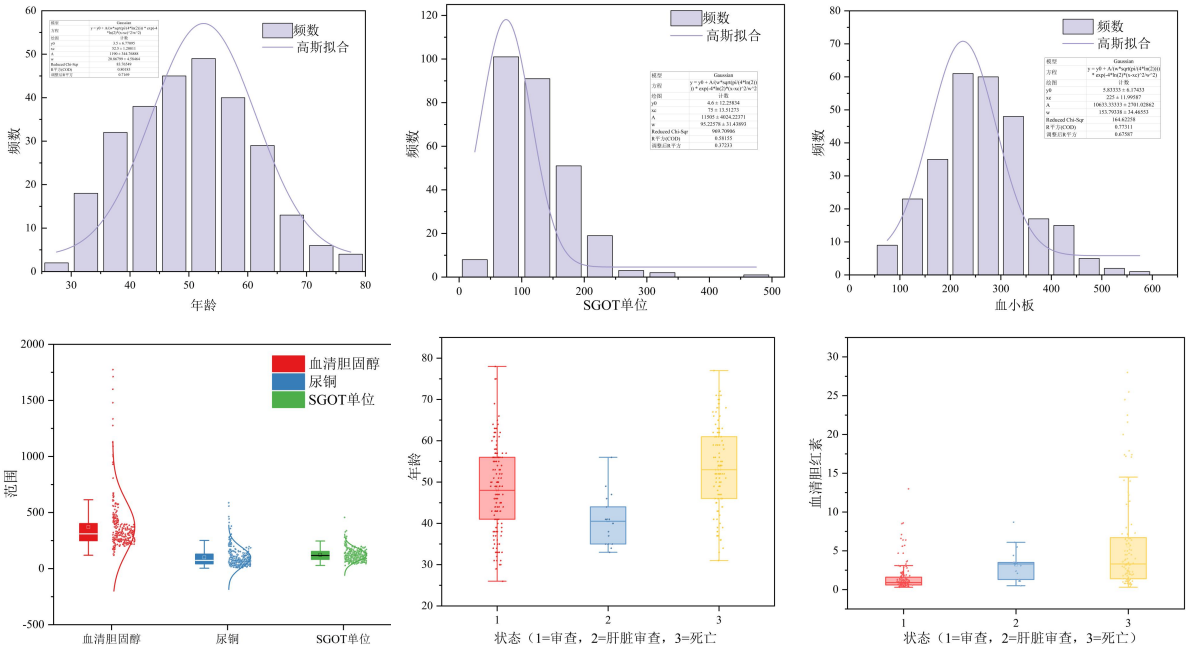


图 5-7 肝硬化数据集统计分析

基于肝硬化数据集的图 5-7 可以看出，年龄呈现近似正态，SGOT 与血小板数呈右偏态；箱线图刻画血清胆固醇、谷氨、SGOT 单位的组间离散特征，按疾病进程分组，不同组的指标分布差异，部分组含异常值；多分组箱线图进一步按疾病状态展示胆红素、SGOT 等指标分布，使疾病阶段对指标分布的影响更为直观，为肝硬化风险因素与预后分析提供量化支撑。

5.1.4 推断性统计分析及可视化

为进一步准确挖掘特征变量之间的关联程度，本文对连续型变量进行独立样本 t 检验，对离散型变量进行卡方检验。

1.独立样本 t 检验

t 检验是用于检验样本均值之间是否存在显著性差异，构造假设：

H_0：特征变量间不存在显著性差异。

H_1：特征变量间存在显著性差异。

独立样本 t 检验中，采用双侧检验，检验统计量 t 的计算公式为：

$$t=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \tag{5-9}$$

其中，$\bar{X}_1$ 和 $\bar{X}_2$ 是特征变量的平均值，s_1^2 和 s_2^2 为方差，n_1 和 n_2 为特征个数。

（1）心脏病数据集 t 检验结果

表 5-2 心脏病数据集 t 检验结果

	stats	p-value
Age	-8.891	<0.001
RestingBP	-3.594	<0.001
Cholesterol	7.197	<0.001
MaxHR	13.257	<0.001
Oldpeak	-13.345	<0.001

由表 5-2 可知，心脏病数据集连续型变量的检验 P 值均小于 0.001，在显著性水平 0.05 下，拒绝原假设，变量间存在显著性差异。

（2）中风数据集 t 检验结果

表 5-3 中风数据集 t 检验结果

	stats	p-value
age	-16.730	<0.001
avg_glucose_level	-9.830	<0.001
bmi	-2.968	<0.001
		<0.001

表 5-3 显示，各变量 P 值均小于 0.001，在显著性水平 0.05 下，拒绝原假设，认为变量间存在显著差异。

（3）肝硬化数据集 t 检验结果

表 5-4 肝硬化数据集 t 检验结果

	stats	p-value
N_Days	5.056	<0.001
Age	-2.273	0.025
Bilirubin	-3.727	<0.001
Cholesterol	-1.511	<0.001
Albumin	4.183	<0.001
Copper	-2.810	0.006
Alk_Phos	-1.176	0.242
SGOT	-3.373	<0.001
Tryglicerides	-2.050	<0.001
Platelets	1.992	0.049

由表 5-4 可知，变量 Alk_Phos 检验 P 值为 0.242，大于显著性水平 0.05，其余变量均小于 0.05，故在后续相关性分析中，可考虑剔除该变量。

2.卡方检验

（1）心脏病数据集卡方检验结果

表 5-5 心脏病数据集卡方检验结果

	chi2	p-value
Sex	83.871	<0.001
ChestPainType	268.896	<0.001
FastingBS	65.860	<0.001
RestingECG	11.070	<0.001
ExerciseAngina	225.133	<0.001
ST_Slope	355.156	<0.001

如表 5-5 所示，各变量检验 P 值均小于 0.001，说明这些分类变量在心脏病相关分组间存在显著差异。

（2）中风数据集卡方检验结果

表 5-6 中风数据集卡方检验结果

	chi2	p-value
gender	0.234	0.629
hypertension	99.667	<0.001
ever_married	54.163	<0.001
work_type	41.951	<0.001
Residence_type	0.176	0.675
smoking_status	35.006	<0.001
heart_disease	93.372	<0.001

表 5-6 的结果显示，除居住类型 Residence_type 和性别 gender 变量外，其余变量的检验 P 值均小于 0.001，存在显著性差异，后续分析中可考虑剔除居住类型 Residence_type 和性别 gender 这两个变量。

（3）肝硬化数据集卡方检验结果

表 5-7 肝硬化数据集卡方检验结果

	chi2	p-value
Status	39.249	<0.001
Drug	4.626	<0.001
Sex	2.059	0.560
Ascites	31.938	<0.001
Hepatomegaly	71.211	<0.001
Spiders	27.993	<0.001
Edema	24.851	<0.001

由表 5-7 可知，变量性别 Sex 的检验 P 值为 0.560，其他变量 P 值均小于 0.001，故除性别外，其他变量与肝硬化存在显著差异，可为肝硬化研究作参考。

5.1.5 结果分析及可视化

结合探索性分析及统计描述性分析结果，给出影响因素分析热力图：

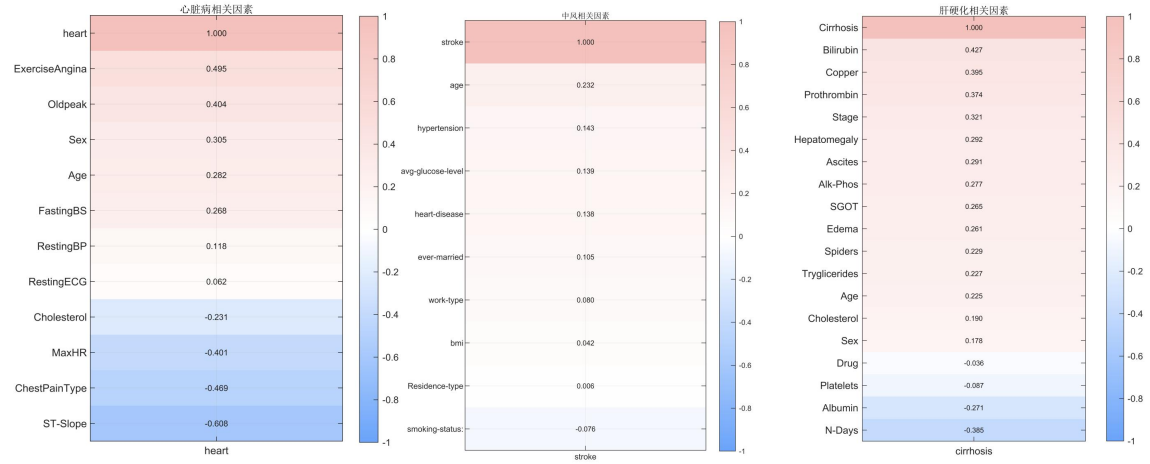


图 5-8 影响因素分析热力图

图 5-8 从左至右，分别对应心脏病、中风、肝硬化三种疾病相关因素的相关性分析，结合上述的推断性统计分析，可得出以下结论：

（1）心脏病相关因素

运动性心绞痛（ExerciseAngina）的 r=0.695、心电图（Oldpeak）的 r=0.604，与心脏病呈显著正相关，二者作为运动负荷下的心肌缺血典型表现，验证临床与数据层对心脏病的高价值。同时性别（Sex）、年龄（Age）、空腹血糖（FastingBS）、静息血压（RestingBP）等有较弱的正向关联，年龄增长、血糖血压异常等会增加心脏病风险。胆固醇（Cholesterol）、最大心率（MaxHR）、胸痛类型（ChestPainType）、ST 段斜率（ST-Slope）均为负相关。

（2）中风影响因素

年龄（age）、高血压（Hypertension）与中风发生的相关性相对较强。随着年龄的增长，血管老化，长期高血压损伤血管，中风风险上升。高血糖水平（high-glucose-level）、心脏病（heart-disease）、已婚状态（ever-married）等与中风有弱正向关联，可能因这些因素通过影响血管健康、生活方式等，潜在提升了中风风险。BMI、工作类型（worktype）、居住类型（Residence-type）相关性更弱，或许和样本特征、因素复杂交互有关。

（3）肝硬化相关因素

胆红素（Bilirubin）、腹水（Ascites）、铜（Copper）与其有较强的正相关性，同

时凝血酶原（Prothrombin）、肝肿大（Splenomegaly）、碱性磷酸酶（Alk-Phos）、谷草转氨酶（SGOT）等，相关性数值在 0.2-0.5 左右，存在较弱的正相关性。而随访天数（N-days）、白蛋白（Albumin）存为负相关，肝硬化患者常出现白蛋白降低，符合病理表现。

5.2 问题二的模型建立与求解

5.2.1 特征选择

根据问题一的相关性分析，筛选出与三种疾病发生显著相关的特征。剔除了 stroke 数据集中 gender、Residence_type、bmi 特征，以及数据集 cirrhosis 中 Sex、Drug、Platelets 特征。同时，在前文的相关性分析发现数据集 heart 里 FastingBS 特征、数据集 cirrhosis 中 Hepatomegaly 特征均存在共线性，故剔除这两个特征。

5.2.2 数据处理

通过对三种疾病数据集的患病标签分布统计发现，心脏病数据集患病与未患病样本比例约为 1:1.2，分布较均衡；中风数据集和肝硬化数据集均存在样本不均衡问题，其中中风数据集尤为突出。由于过采样易导致模型过拟合，故采用随机欠采样方法处理，平衡中风数据集中出现的样本不均衡问题。

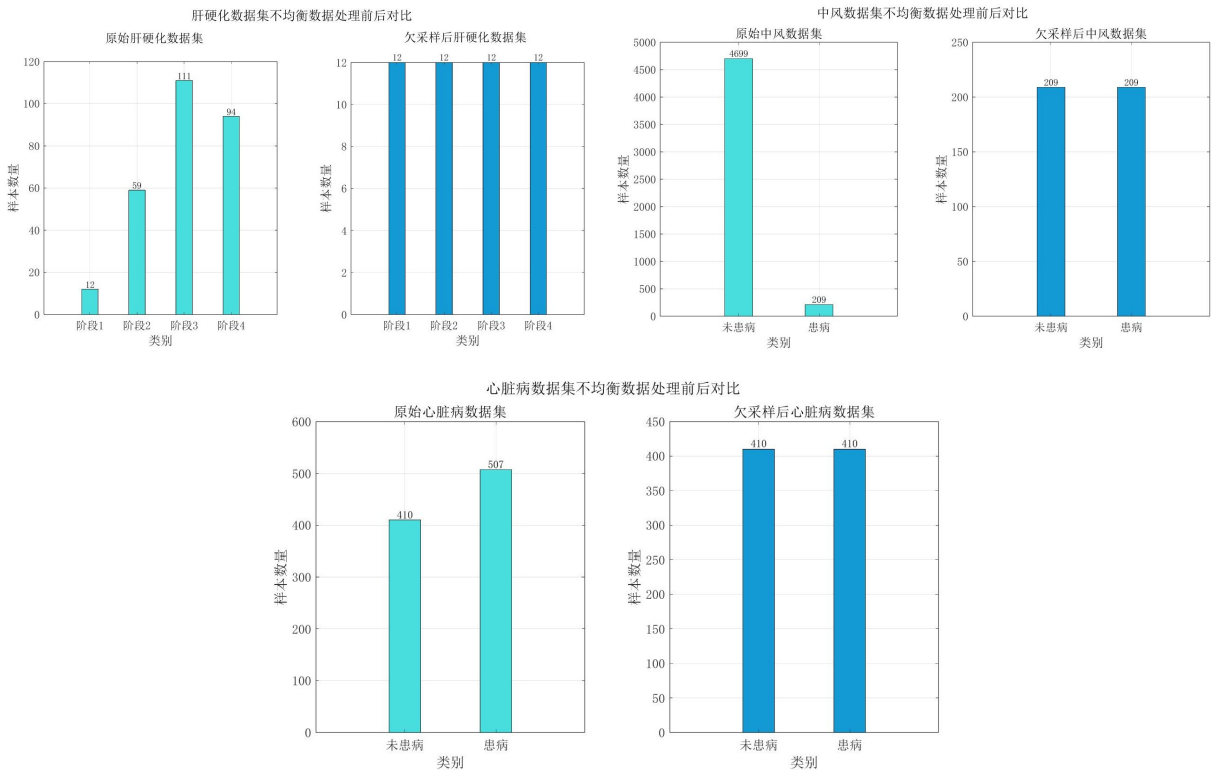


图 5-9 不均衡数据处理前后对比图

5.2.3 模型建立

本文使用多个经典的分类算法来训练预测模型，包括 Logistics 模型、高斯朴素贝叶斯模型、BP 神经网络模型。

1.Logistics 模型

为直观反映各特征对患病风险的影响程度，选择二分类 Logistics 回归模型分别构建心脏病、中风、肝硬化的患病概率预测模型[4]，具体如下：

$$Logit(p_i)=\ln\left(\frac{p_i}{1-p_i}\right)=\alpha_i+\sum\beta_{ij}X_{ij}+\varepsilon_i \quad (5\text{-}10)$$

其中，X_{ij} 为影响第 i 种疾病患病的各类因素（i=1,2,3 分别对应心脏病、中风、肝硬化；j 为特征序号）；p_i（$p_i=1$ 患病，$p_i=0$ 不患病）分别表示患三种疾病的概率；α_i 为第 i 种疾病模型的截距项，β_{ij} 为待估计参数，反应对应特征对疾病患病概率的影响程度，ε_i 为随机扰动项。

患病概率 p_i 的表达式：

$$p_i=\frac{1}{1+\exp\left[-\left(\alpha_i+\beta_{i1}X_{i1}+\cdots+\beta_{ik_i}X_{ik_i}\right)\right]} \quad (5\text{-}11)$$

2.高斯朴素贝叶斯模型

针对高斯朴素贝叶斯模型[6]，设研究的三种疾病分别对应索引 $i=1,2,3$ (依次代表心脏病、中风、肝硬化)，定义特征向量 X_{ij} 为影响第 i 种疾病患病的各类因素，j 为特征序号。类别标签 $Y_i\in\{0,1\}$，$Y_i=1$ 表示患第 i 种疾病，$Y_i=0$ 表示未患病。

对于第 i 种疾病，假设给定类别 $Y_i=k$ 时，各特征相互独立，联合条件概率分解为：

$$P\left(X_i\mid Y_i=k\right)=\prod_{j=1}^{k}P\left(X_{ij}\mid Y_i=k\right) \quad (5\text{-}12)$$

假设第 i 种疾病类别为 k 时，特征向量 X_{ij} 服从高斯分布 $N\left(\mu_{ijk},\sigma_{ijk}^2\right)$，其概率密度函数为：

$$P\left(X_{ij}=x_{ij}\mid Y_i=k\right)=\frac{1}{\sqrt{2\pi\sigma_{ijk}^2}}\exp\left(-\frac{\left(x_{ij}-\mu_{ijk}\right)^2}{2\sigma_{ijk}^2}\right) \quad (5\text{-}13)$$

其中，μ_{ijk} 为类别 k 下第 i 种疾病第 j 个特征均值，σ_{ijk}^2 为对应方差，通过训练数据估计。根据贝叶斯定理，第 i 种疾病类别为 $Y_i=1$ 的后验概率为：

$$P\left(Y_i=1\mid X_i\right)=\frac{P\left(Y_i=1\right)\cdot\prod_{j=1}^{k}P\left(X_{ij}\mid Y_i=1\right)}{P\left(Y_i=0\right)\cdot\prod_{j=1}^{k}P\left(X_{ij}\mid Y_i=0\right)+P\left(Y_i=1\right)\cdot\prod_{j=1}^{k}P\left(X_{ij}\mid Y_i=1\right)} \quad (5\text{-}14)$$

其中，$P\left(Y_i=1\right)$、$P\left(Y_i=0\right)$ 为第 i 种疾病患病/未患病的先验概率，由训练集中对应类别样本占比计算：

$$P(Y_i=k)=\frac{\text{第}i\text{种疾病类别}k\text{的样本数}}{\text{第}i\text{种疾病训练集总样本数}}(k=0,1) \tag{5-15}$$

3.BP 神经网络模型

BP 神经网络的学习过程主要由四个部分构成：输入特征顺传播、输出误差逆传播、循环迭代训练、模型性能判别[5]。

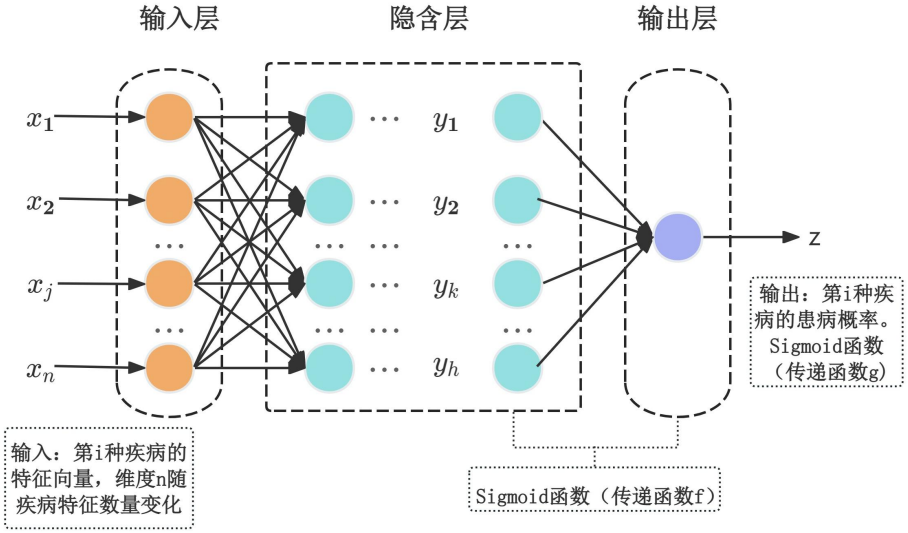


图 5-10 神经网络模型图

设网络的输入层特征向量为$x=(x_1,x_2,\cdots,x_n)^T$，其中 n 为第 i 种疾病的特征数量，x_j表示第j个特征的取值。隐含层有h个神经单元输出为$y=(y_1,y_2,\cdots,y_h)^T$。输出层有 1 个神经元，输出为z，表示第i种疾病的患病概率。

隐含层到输出层采用 Sigmoid 函数中的 f 作为传递函数：

$$y_k=f\left(\sum_{j=1}^{n}\omega_{jk}x_j-\theta_k\right)=\frac{1}{1+e^{-\left(\sum_{j=1}^{n}\varpi_{jk}x_j-\theta_k\right)}} \tag{5-16}$$

输出层采用 Sigmoid 函数中的 g 作为传递函数，将输出的线性组合映射为患病概率：

$$z=g\left(\sum_{k=1}^{h}\omega_k y_k-\phi\right)=\frac{1}{1+e^{-\left(\sum_{k=1}^{h}\varpi_k y_k-\phi\right)}} \tag{5-17}$$

其中，其中y_k为第k个隐含层神经元的输出，ω_{jk}为输入层到隐含层的权值，偏置为θ_k，隐含层到输出层的权值为ω_k，偏置为ϕ。此时，单个样本的误差为：

$$\varepsilon=\frac{1}{2}(t-z)^2 \tag{5-18}$$

对于m个训练样本，总误差为所有误差之和：

$$\varepsilon_{总}=\frac{1}{2}\sum_{m=1}^{M}(t_m-z_m)^2 \tag{5-19}$$

其中，t_m为第m个样本的真实标签，z_m为对应预测值。

5.2.4 模型的训练与评估

1.模型的训练

针对不同预测模型，采用不同方法进行模型训练

（1）Logistics 模型

采用最大似然估计法训练模型参数[7]，对于第 i 种疾病的 n 个样本（$n=N_i$），设第 m 个样本的因变量为Y_{im}（1 或 0），特征向量$X_{im}=\left(X_{im1},\cdots,X_{imk_i}\right)$，则其条件概率为：

$$P\left(Y_{im}\mid X_{im}\right)=p_i^{Y_{im}}\cdot\left(1-p_i\right)^{1-Y_{im}} \tag{5-20}$$

各样本条件概率的乘积L_i：

$$L_i\left(\alpha_i,\beta_{i1},\cdots,\beta_{ik_i}\right)=\prod_{m=1}^{n}\left[p_i^{Y_{im}}\cdot\left(1-p_i\right)^{1-Y_{im}}\right] \tag{5-21}$$

为便于计算，对似然函数取对数得:

$$\ln L_i=\sum_{m=1}^{n}\left[Y_{im}\cdot\ln p_i+\left(1-Y_{im}\right)\cdot\ln(1-p_i)\right] \tag{5-22}$$

$$\ln L_i=\sum_{m=1}^{n}\left[Y_{im}\cdot\left(\alpha_i+\sum_{j=1}^{k_i}\beta_{ij}X_{ij}\right)-\ln\left(1+\exp\left(\alpha+\sum_{j=1}^{k_i}\beta_{ij}X_{ij}\right)\right)\right] \tag{5-23}$$

（2）高斯朴素贝叶斯模型

针对第 i 种疾病，整理数据集并提取特征X_i，将数据集按 7:3 划分为训练集与测试集，统计“患病$\left(Y_i=1\right)$”与“未患病$\left(Y_i=0\right)$”样本数量，按上述先验概率公式计算$P\left(Y_i=1\right)$、$P\left(Y_i=0\right)$。

对第 i 种疾病的每个特征X_{ij}，分布计算其在“患病”“未患病”类别下的均值μ_{ij1}、μ_{ij2}与方差σ_{ij1}^2、σ_{ij2}^2：

$$\mu_{ijk}=\frac{1}{N_{ik}}\sum_{m:Y_{im}=k}X_{ijm} \tag{5-24}$$

$$\sigma_{ijk}^2=\frac{1}{N_{ik}-1}\sum_{m:Y_{im}=k}\left(X_{ijm}-\mu_{ijk}\right)^2 \tag{5-25}$$

其中，N_{ik}为第i种疾病类别k的训练样本量，X_{ijm}为第m个样本的第j个特征值。

（3）BP 神经网络模型

对第 i 种疾病的数据集，依照 7:3 的比例划分为训练集和测试集。对算法输出结果展开分析，以此筛选出性能最优的模型。

在为各个网络设定训练迭代次数与目标误差阈值等约束条件后，采用梯度下降法反

向传播误差并更新权值。再对预测结果与实际数值实施反归一化操作，使其还原至原始数据的量级范围。

输出层到隐含层的权值 ω_k 更新：

$$\Delta\omega_k = \eta(t-z)z(1-z)y_k \tag{5-26}$$

$$\omega_k' = \omega_k + \Delta\omega_k \tag{5-27}$$

隐含层到输出层的权值 ω_{jk} 更新：

$$\Delta\omega_{jk} = \eta(t-z)z(1-z)\omega_k y_k(1-y_k)x_j \tag{5-28}$$

$$\omega_{jk}' = \omega_{jk} + \Delta\omega_{jk} \tag{5-29}$$

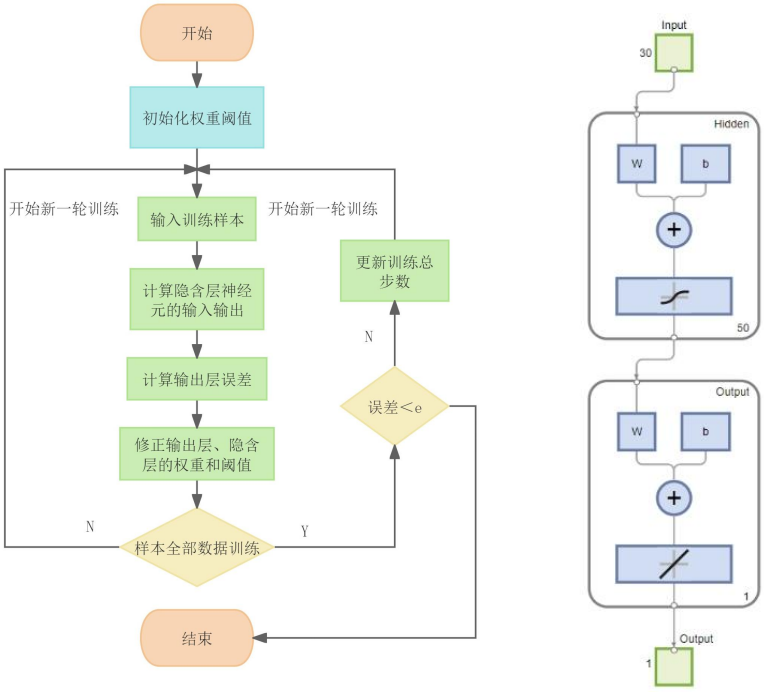


图 5-11 BP 神经网络计算流程图

2．模型的准确度检验

使用测试集测试性能，计算准确率、查准率、召回率和 F1 函数，若性能不佳则调整后重新训练：

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5-30}$$

$$\Pr ecision = \frac{TP}{TP+FP} \tag{5-31}$$

$$\Re call = \frac{TP}{TP+FN} \tag{5-32}$$

$$F1 = 2\times\frac{\Pr ecision\times\Re ccall}{\Pr ecision+\Re ccall} \tag{5-33}$$

3．模型的灵敏度分析

采用单因素灵敏度分析来探究输入特征的微小变化对模型输出的影响[8]，敏感度系数：

$$S_j = \frac{\Delta P / P_0}{\Delta x_j / x_{j0}} \tag{5-34}$$

其中，x_j 为第 j 个特征，x_{j0} 为特征初始值，Δx_j 为特征变化量，P_0 为初始患病概率，ΔP 为概率变化率，$|S_j|$ 越大，特征 j 对输出越敏感。

5.2.5 预测结果分析

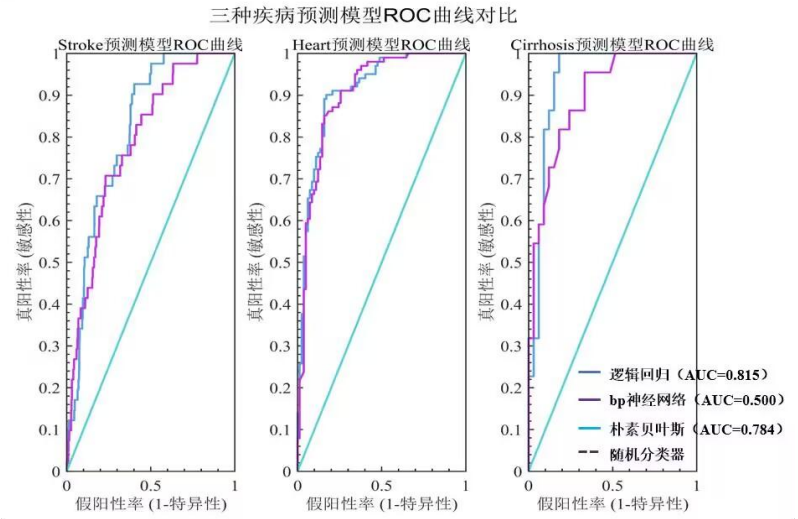


图 5-12 ROC 曲线对比图

图 5-12 展示了预测模型在三种数据集上的 ROC 曲线。每条曲线代表不同模型性能，越靠近左上角模型性能越好，曲线下面积（AUC）数值越大，代表模型预测效果越优秀。

从三种疾病的 ROC 曲线对比来看，针对心脏病预测的模型 ROC 曲线整体位于最上方，AUC 值最高，说明心脏病预测模型的整体区分能力最优。肝硬化预测模型的 ROC 曲线次之，AUC 值较高，模型对肝硬化患病状态的区分能力较强，能有效捕捉这些线性或非线性关系。中风预测模型的 ROC 曲线相对靠下，AUC 值较低，需进一步优化阈值或采用更稳健的采样方法。

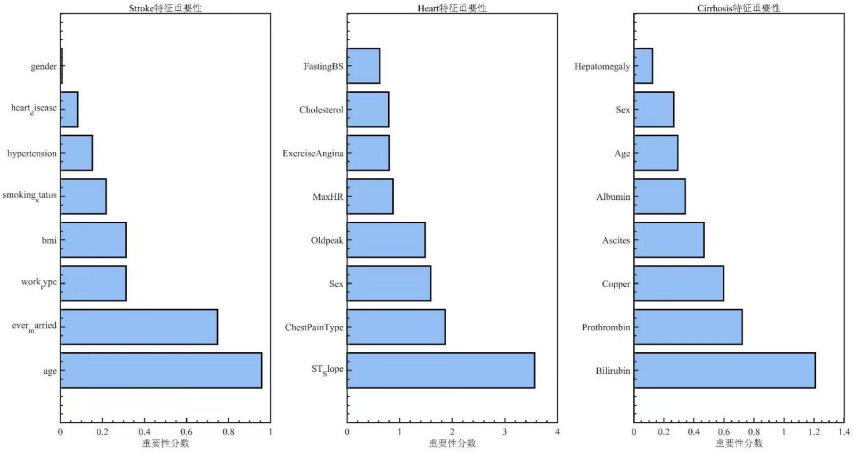


图 5-13 特征重要性分析图

图 5-13 展示了不同疾病数据集的特征重要性分析结果，反映各指标对模型预测结果的贡献度，分数越高，该特征对疾病患病概率的影响越显著。

对于心脏病数据集，重要性较高的特征包括 ST 段斜率、胸痛类型和运动性心绞痛等，这些特征对预测模型有较大影响。中风数据集中，年龄、高血压、平均血糖水平德国是关键影响因素。肝硬化数据集则强调了胆红素、铜、凝血酶原等指标在疾病预测中的作用。经上述分析，能够确定各疾病数据集中的核心预测因子，以便模型的进一步优化及特征选择。

表 5-8 三种模型性能分析表

	模型	准确率	精确率	召回率	特异性	F1 分数	模型
Stroke 模型	朴素贝叶斯	0.8654	0.136	0.4146	0.8851	0.2048	0.817
	逻辑回归	0.9572	0	0	0.9989	0	0.8148
	神经网络	0.9551	0	0	0.9968	0	0.7968
Heart 模型	朴素贝叶斯	0.8525	0.8491	0.8911	0.8049	0.8696	0.9175
	逻辑回归	0.8525	0.8426	0.901	0.7927	0.8708	0.9094
	神经网络	0.8306	0.8302	0.8713	0.7805	0.8502	0.9029
Cirrhosis 模型	朴素贝叶斯	0.7818	0.8125	0.5909	0.9091	0.6842	0.9325
	逻辑回归	0.7818	0.8571	0.5455	0.9394	0.6667	0.9311
	神经网络	0.8182	0.8333	0.6818	0.9091	0.75	0.9215

通过分析可以看出，朴素贝叶斯模型在中风预测中表现最优，准确率高达 0.8654，召回率达到 0.4146，是唯一能识别部分患病样本的模型，但精确率和 F1 分数较低。表明模型虽能捕捉部分阳性样本，但误判率高，中风预测模型仍需优化。

在心脏病预测中，三个模型整体表现优异，准确率均超 0.83，其中朴素贝叶斯和逻辑回归准确率相同，逻辑回归的 F1 分数略高于朴素贝叶斯，说明其在精确率与召回率的平衡上更优，表明逻辑回归模型在心脏病预测中综合性能最佳，可作为临床辅助预测模型。

神经网络模型在肝硬化预测中准确率最高，为 0.8182，F1 分数达到 0.7500，在三个模型中最优，其对肝硬化复杂病理特征的捕捉能力更强，适合处理肝硬化数据中多特征间的复杂交互关系。

5.2.6 模型的改进

1.数据层面改进：中风数据集原始随机欠采样虽平衡了样本比例，但导致多数类信息丢失，可改为使用 SMOTE 过采样+ENN 编辑组合策略。针对肝硬化不同阶段样本分

布不均的情况，可采用分层抽样与数据扩充相结合的方式，提升早期肝硬化的识别能力。

2.模型结构改进：Logistic 回归正则化优化针对心脏病模型中可能存在的过拟合风险，可引入 L1-L2 混合正则化，避免单一正则化导致的特征过度剔除或保留冗余的问题。中风数据集中部分特征非严格正态分布，可改用 KDE 替代高斯分布假设，通过 Silverman 带宽法自适应估计概率密度。

5.3 问题三的模型建立与求解

5.3.1 数据预处理

数据清洗：基于以上结论，提取共同风险因素和疾病状态标签；对缺失数据值采取同第一问方法进行插补，再通过3σ准则，结合医学常识剔除异常值，确保数据一致性。

数据融合：通过患者唯一标识关联三个数据集，构建包含“风险因素+三种疾病状态”的综合数据集，明确“同时患两种疾病”“同时患三种疾病”的样本标签。

5.3.2 多疾病关联概率图模型的建立

1.变量定义

为明确模型的输入和输出，即风险因素和疾病状态，将变量划分为疾病变量和共同风险因素变量：

（1）疾病变量包括心脏病A_1、中风A_2、肝硬化A_3。它们均为二元变量，当$A_i=1$时表示患病，$A_i=0$表示未患病；

（2）共同风险因素变量包括年龄、性别、高血压等影响多疾病的因素，记为$B=\{b_1,b_2,\cdots,b_k\}$。

2.网络结构设计

基于医学知识和文献研究，根据贝叶斯网络理论[9]，定义变量间的依赖关系，其中点集$V=\{A_1,A_2,A_3\}\cup B$，边集表示变量间的依赖关系。

3.关联强度的量化及共病概率的计算

本文引入条件风险比和条件概率比，利用标准化指标量化疾病间的关联强度。条件风险比定义为在患有疾病A_i时疾病A_j发生的概率与不考虑A_i状态时，疾病A_j发生的概率之比，公式为：

$$BHR_{i,j}=\frac{P(A_j=1\mid A_i=1)}{P(A_j=1)} \tag{5-35}$$

条件概率比定义为在患有疾病A_i的条件下疾病A_j发生的概率与不换疾病A_i的条件下疾病A_j发生的概率之比，公式为：

$$BPR_{i,j}=\frac{P(A_j=1\mid A_i=1)}{P(A_j=1\mid A_i=0)} \tag{5-36}$$

当$BHR_{i,j}>1$或$BPR_{i,j}<1$时，表示疾病A_i会在一定程度上增加疾病A_j的患病风险；

当 $BHR_{i,j}<1$ 或 $BPR_{i,j}>1$ 时，表示疾病 A_i 会在一定程度上降低疾病 A_j 的患病风险；当 $BHR_{i,j}=1$ 或 $BPR_{i,j}=1$ 时，表示两种疾病大致独立。

4.概率的计算

本文利用问题二中预测模型输出的患病概率，作为疾病变量在无其他影响时的基准概率。将指向 X 的变量集合记为 $P_1(X)$，则在给定节点取值的条件下，变量 X 的概率分布为 $P(X|P_1(X))$。由此可得联合概率分布：

$$P(A_1,A_2,A_3,B)=P(A_1\mid P_1(A_1))\cdot P(A_2\mid P_1(A_2))\cdot P(A_3\mid P_1(A_3))\cdot\prod_{j=1}^{k}P(B_j\mid P_1(B_j)) \quad (5\text{-}37)$$

依据贝叶斯的分解表示，结合条件概率分布与联合分布的性质，可得出疾病状态的联合分布式：

$$P(A_1,A_2,A_3)=\sum_{B}P(A_1,A_2,A_3,B) \quad (5\text{-}38)$$

若同时患有两种疾病，即任意两个 $A_i=1$ 时，其概率为：

$$P(A_{i_1}=1,\ A_{i_2}=1)=\sum_{a_3\in[0,1]}P(A_{i_1}=1,\ A_{i_2}=1,A_{i_3}=a_3) \quad (5\text{-}39)$$

若同时患有三种疾病，即三个 $A_i=1$ 时，其概率为：

$$P(A_1=1,\ A_2=1,\ A_3=1) \quad (5\text{-}40)$$

5.3.3 模型的求解

1.网络结构

为了更清晰地展示疾病之间的关联结构，构建了疾病关联网络图，通过有向边表示风险因素对疾病的影响以及疾病之间的相互关系。

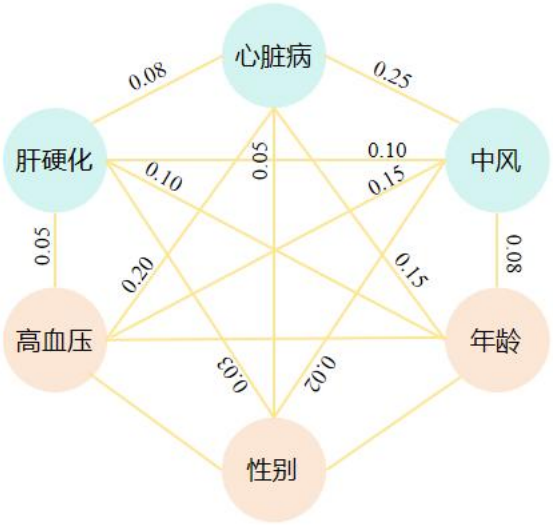


图 5-14 网络结构图

从图 5-14 知，年龄、性别和高血压为三种疾病的共同的风险因素，其中高血压对

心脏病的影响最大，权重为 0.20；心脏病对中风有显著影响，认为心脏病患者发生中风的风险可能会增加；而肝硬化对中风也存在一定影响，这可能与肝功能障碍导致的凝血功能异常和血管病变有关。

2.共病组合频率

基于多疾病关联概率图模型的预测，为更直观观察疾病之间的关联度，选用条形图进行可视化：

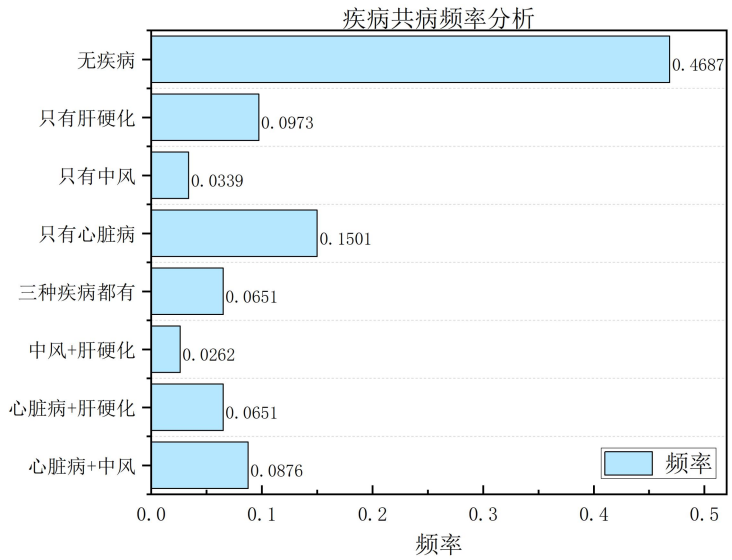


图 5-15 共病组合频率图

结果显示，心脏病与中风的共病率为最高，达 8.86%，心脏病与肝硬化为 6.15%，中风与肝硬化为 2.62%，而三种疾病同时患有的比例为 6.51%，由此表明疾病之间确实存在明显的关联，显著高于随机共现。

基于相对风险比的计算，进一步量化了疾病之间的关联强度，将其可视化为相对风险比热图：

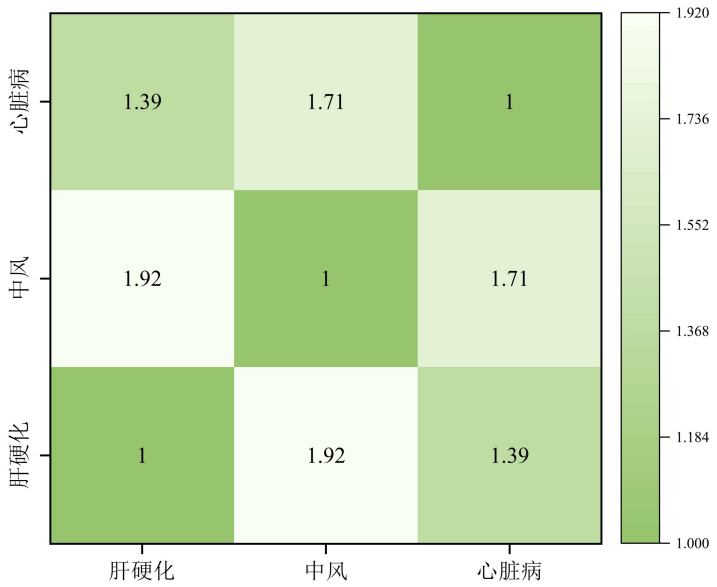


图 5-16 相对风险比热力图

心脏病与中风之间的相对风险比为最高，达 1.92，表明患有一种疾病会将另一种疾

病的风险提高约 92%；中风与肝硬化的相对风险比为 1.71，心脏病与肝硬化的相对风险比为 1.39。由此可知，这三种疾病不是相互独立的，存在显著的关联关系，其中心脏病与中风的关联最为紧密。

3.不同人群共病风险

基于模型的共同特征分析结果及网络结构图，为更好地解释人的共病风险，本文针对不同年龄、性别和高血压人群的共病风险进行分析，并根据依据综合风险指数，将其划分为高、中、低风险，如下图所示。

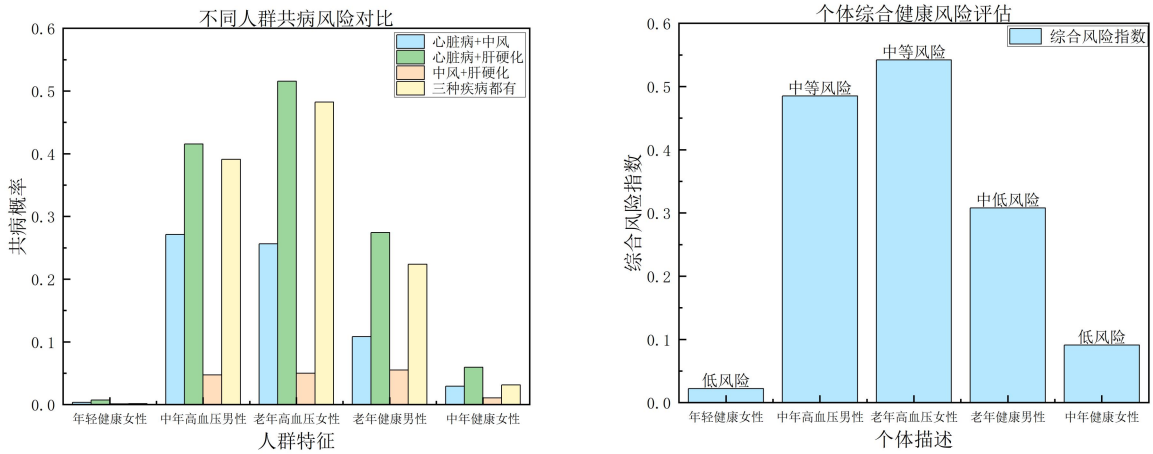


图 5-17 不同人群共病风险分析图

综上分析，老年高血压女性的患病概率最高属于高风险，而年轻健康女性的各种共病风险均很低，属于低风险。这种差异表明，年龄和高血压是影响多疾病共病的关键因素，老年人和高血压患者更加需要关注共性疾病的预防。

针对每个个体，为更直观展示其在不同疾病和共病维度上的风险特征，绘制个体特征风险雷达图如下所示：

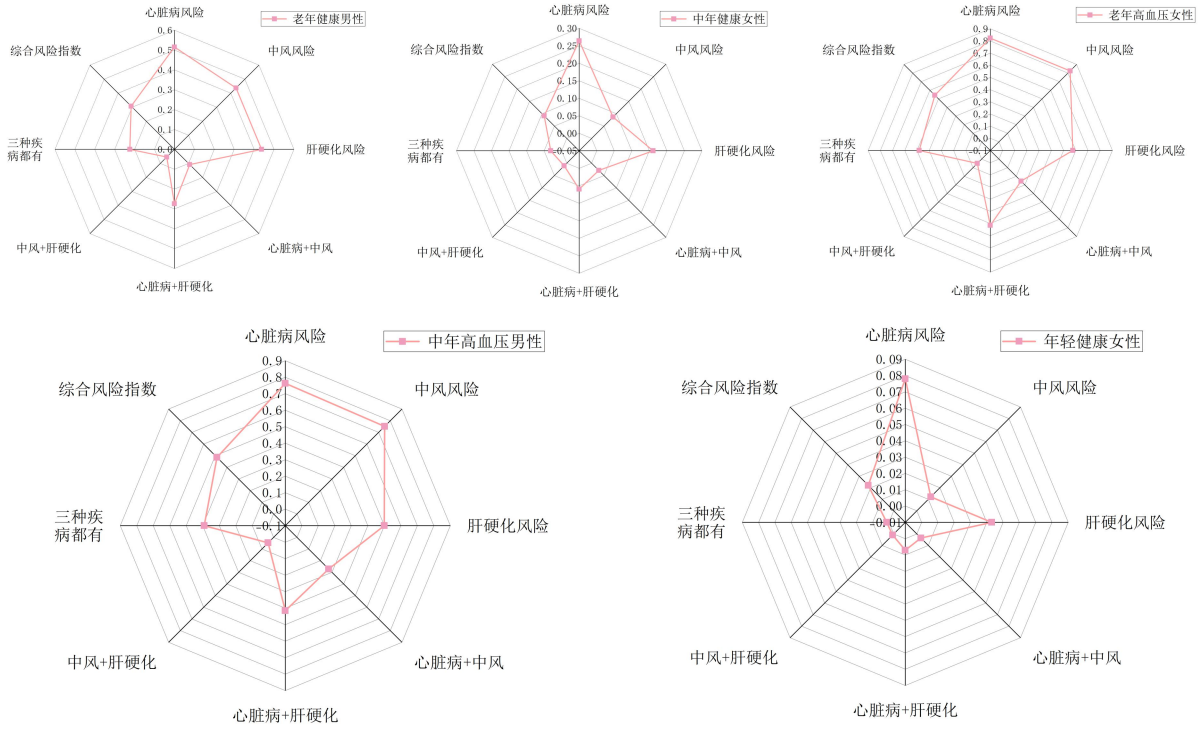


图 5-18 个体特征风险雷达图

由图 5-18 可发现，年轻健康女性的风险雷达图呈现极小的多边形，所有风险维度

的值都接近于零，表明这类人群的各种疾病风险都很低。相比之下，中年高血压男性和老年高血压女性的风险雷达图则呈现出较大的面积。因此，对于不同人群，需要采取不同的预防措施，以降低发病风险。

综上，通过构建以条件概率为核心的多疾病关联概率图模型，实现了心脏病、中风与肝硬化共病概率的预测，揭示了三者间的显著关联，心脏病与中风的关联性最高。同时识别出年龄和高血压为影响多疾病共病的关键风险因素，并开发了个体化综合风险评估工具，为临床精准预防及个性化健康管理提供了理论支撑与实用方法。由此不仅提升了对疾病关联机制的认知，又为公共健康卫生政策的制定与医疗资源的优化配置提供了科学依据。

5.4 问题四的求解

致世界卫生组织的公开信

尊敬的世界卫生组织：

您好！

我们是参加 2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛 B 题的研究团队。基于对心脏病、中风与肝硬化三种高致死率疾病的大数据分析与模型构建，针对全球范围内这三种疾病的预防与管理提出以下建议，供贵组织参考。

一、基于疾病关联机制的协同预防策略

我们的多疾病关联概率图模型显示，三种疾病并非独立存在，其中心脏病与中风的相对风险比达 1.92，共病率为 8.86%，显著高于随机共现概率；肝硬化与中风也存在中等关联，相对风险比 1.71，需建立跨疾病的协同防控体系。

建议将**“心-脑-肝”**疾病纳入联合监测框架，高血压对心脏病的影响权重达 0.20，年龄与三种疾病均呈显著正相关，针对高血压、年龄等共同风险因素，可将年龄及血压分类，制定准确的风险评估标准，除此之外，还需定期开展心血管功能与肝功能的联合筛查。

二、基于个体化风险的精准干预措施

通过特征重要性分析发现，不同人群的疾病风险差异显著：老年高血压女性的综合风险指数达 0.6 以上，属于高风险群体，而年轻健康女性的风险近乎为 0。

建议推广“**个体化风险评估工具**”。利用 Logistic 回归与 BP 神经网络模型，结合年龄、血压、血糖、胆红素等关键特征，为个体生成疾病风险预防或改善计划，实现“高风险人群强化干预、低风险人群健康宣教”的分级管理。针对特定风险特征制定方案，例如对心脏病高风险人群，需要重点监测运动性心绞痛与心电图 ST 段斜率；对肝硬化潜在人群，需要定期检测血小板计数与凝血功能，早期识别，规避风险。

三、基于数据驱动的公共卫生政策优化

模型预测结果显示，三种疾病的共病率达 6.51%，且共同受生活方式与基础疾病影响。建议从以下方面完善公共卫生政策：

1.风险因素管控：针对高血压、高血糖等可干预因素，推动“减盐控糖”公共健康运动，可以在社区设立血压、血糖免费监测点，在终端创建实时监控和查看特征状态的“健康 APP”。由此，将筛查数据接入信息平台，供人们随时监控健康状态，做到个体化动态风险预警，早期识别，及时干预，降低发病风险。

2.医疗资源配置：基于疾病关联热力图及个体特征的雷达图，对于心脏病高发区，强化中风急救资源，对于肝硬化高发区，增设心脑血管监测项目，提高共病应对效率。

3.公众教育：通过科普手册、短视频等形式，普及“心脏病-中风”“肝硬化-凝血

异常”等关联知识，纠正“单一疾病独立防控”的认知偏差，提升人群主动预防意识。

我们的研究证实，通过数据驱动的精准预防与跨疾病协同管理，可显著降低此三种疾病的发病与共病风险。我们非常期待与贵组织进一步合作，将研究成果转化为全球可推广的防控方案，为减轻慢性病负担、提升人类健康水平贡献力量。

此致

敬礼！

2025 年第十五届 APMCM 参赛团队

2025 年 7 月 15 日

六、模型评价与推广

6.1 模型的优点

（1）逻辑回归模型基于线性对数 odds 构建关联，以明确的数学表达式量化特征对疾病风险的边际贡献，系数可解释性强，便于快速识别核心风险因素。通过极大似然估计求解参数，计算复杂度低，能直接输出概率值，为临床风险分级提供直观量化依据，且模型鲁棒性较好。

（2）BP 神经网络以多层感知机结构，通过反向传播算法学习特征间高度非线性映射，能拟合复杂疾病关联模式，并具备强泛化能力，通过 dropout、正则化可提升复杂疾病场景下的预测稳定性，输出层可灵活适配多分类、概率预测需求，满足疾病共病组合预测。

（3）由逻辑回归、贝叶斯、神经网络多模型对比框架可根据数据特性选择最优方案，提升预测适应性。

（4）以条件概率为核心的多疾病关联概率图模型实现了从单一疾病风险到多疾病共病风险的自然过渡，能够根据个体特征计算不同疾病组合的概率，并通过可视化工具直观展示个体的风险特征，为个性化健康管理提供了实用工具。

6.2 模型的不足与改进

（1）逻辑回归模型依赖参数化分布假设，实际数据偏离时易导致偏差，且对类别不平衡数据敏感，少数类疾病预测易被忽略。

（2）BP 神经网络模型的训练过程易陷入局部极小值，依赖初始权值与学习率设置，收敛效率不稳定。

6.3 模型的推广

基于本文的研究成果，可推广的方向有：

（1）逻辑回归模型可推广至慢性病队列研究，动态跟踪特征变化对疾病风险的线性贡献。同时还可用于公共卫生筛查，构建标准化风险评分体系，进行医疗质量评估。

（2）BP 神经网络可用于工业制造领域的产品质量检测，对生产过程中采集的多维度数据进行学习，识别产品生产环节中的异常模式，预测产品质量是否达标，从而帮助企业及时调整生产流程，减少次品率，提升产品质量稳定性。

（3）朴素贝叶斯算法可用于文本信息处理领域的情感分析，基于文本中的词汇、语句结构等特征，通过计算不同情感倾向下词汇出现的概率，对文本内容的情感态度进行分类，企业了解用户对产品或服务的评价倾向，为市场策略调整提供依据。

参考文献

[1] 曹桂林.心血管疾病数据集下基于机器学习的心血管疾病患者识别[J].河北软件职业技术学院学报,2024,26(01):6-11.

[2] 朱红兵,何丽娟.关于用 SPSS 中单样本 K-S 检验法进行正态分布等的一致性检验时适用条件的研究[J].首都体育学院学报,2009,21(04):466-470.

[3] 李鑫阳,柳亚云,姜麟,等.结合标签发现与独热编码的链路预测方法[J].计算机工程与设计,2025,46(06):1609-1615.

[4] 何琳,姚海荣,邵敏,等.基于逻辑回归的心内科医疗设备风险分级模型构建研究[J].中国医学装备,2025,22(01):96-101.

[5] 甘妮.基于 BP 神经网络法构建老年慢性阻塞性肺疾病患者吞咽障碍预测模型[J].中国医院统计,2025,32(01):58-63+80.

[6] 吕蔚萁.基于高斯朴素贝叶斯分类器的特征预测能力检测[D].天津大学,2020.

[7] 刘亚欢,田宇,李国通.基于最大似然估计的 GPS 多径估计[J].宇航学报,2009,30(04):1466-1471.

[8] 徐崇刚,胡远满,常禹,等.生态模型的灵敏度分析[J].应用生态学报,2004,(06):1056-1062.

[9] 向玉莹.基于贝叶斯网络的客运码头保安风险预测研究[D].大连海事大学,2024.

选题	2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
C 题		apmcm25200983

基于 QBoost 的二分类模型设计的研究——以 Iris 数据集为例

摘要

集成学习通过组合多个弱分类器构建性能优异的强分类器，本文基于集成学习经典算法 Boosting 的变体 **QBoost** 完成一项**二分类任务**。该任务以指定的 **Iris 数据集**为基础，通过建立 **QUBO 模型**，利用 **Kaiwu SDK** 中的模拟退火求解器进行求解，本文还将 QBOU 模型的应用扩展至医疗、金融等领域。

针对问题一，本文采用 **Z-score 标准化方法**对 Iris 数据集中选取的 100 个样本进行预处理，按 **8：2** 的比例划分为 80 个训练集和 20 个测试集。经过标准化操作后，训练集均值接近 0、方差接近 1。根据所选数据集的特征，构造了 **25 个单一特征弱分类器、10 个特征组合分类器及 5 个线性组合分类器**；借助特征相关性热力图和特征分布直方图确定各类弱分类器的阈值，完成预测结果输出。利用数据分析了每个弱分类器在训练集上的分类准确率，并用柱状图可视化呈现。

针对问题二，本文构建并求解了基于 QUBO 模型的 QBoost 模型。以 40 个弱分类器为基础，通过引入**二值决策变量**表现其选择状态，结合训练集预测矩阵与标签向量，**构建包含误分类项和正则化项的目标函数**，经**交叉验证**确定正则化参数为 0.5，以限制选中数量，防止过拟合。将目标函数转化为标准 QUBO 形式，得到 40×40 的对称矩阵，其中对角线元素反映单个分类器的选择倾向，非对角线元素体现分类器间的协同关系。同时定义矩阵需满足对称性和上三角储存的约束条件，以适配 Kaiwu 模拟退火求解器的输入要求。

针对问题三，本文基于 **Kaiwu SDK 的模拟退火求解器**对模型进行求解，并与**传统 AdaBoost、普通模拟退火集成模型展开对比**。将问题二中的 40×40 的 QUBO 矩阵转化为 **Ising 模型**，通过 Kaiwu 模拟退火求解器得到最优的弱分类器权重组合为 0 和 1，最终选中的弱分类器**包括 14 个单一特征分类器、5 个组合特征分类器及 3 个线性组合分类器**。基于该组合构建的强分类器在训练集与测试集上的**准确率、精确率等指标均达 1.0，泛化能力优异**。对比显示，传统 AdaBoost 虽准确率相同，但不具备量子优化特性；传统模拟退火测试集准确率为 100%，但耗时 24.745s；而 Kaiwu 求解器仅用 0.476s，在效率与性能上均具优势，表明基于 Kaiwu SDK 的 QBoost 模型在二分类任务中高效且准确，在复杂场景下潜力显著。

本文通过 QBoost 建模完成了二分类任务，并且将基于 Kaiwu 模拟退火器求解的 QBoost 模型与 AdaBoost 模型及传统的模拟退火器进行对比，验证了 QBoost 模型的优越性，**并将 QUBO 模型的应用推广到了医疗以及金融领域**，以优化集成提升效果。

关键词： Z-score 标准化方法，QUBO 模型，线性组合分类器，模拟退火算法，Ising 矩阵

目录

一、 引言

在机器学习中，分类任务尤其重要，传统的单一模型性能有限，其受限于自身结构复杂度及对数据分布的适应性，在复杂场景中难以同时兼顾准确率与泛化能力。**集成学习**是机器学习模型对一个问题进行多次学习，得到多个基模型，并通过一定的方法对这些基模型进行集成组合，得到集成模型[1]。当集成学习聚焦于分类任务的场景时，则通过结合多个弱分类器以提升模型的性能，Boosting 作为经典的集成算法，在机器学习领域应用广泛。

然而，随着数据规模的爆炸式增长与特征维度的持续提升，传统 Boosting 算法在计算效率与优化精度上逐渐面临挑战。一方面，弱分类器的迭代训练与权重优化过程需消耗大量计算资源，当数据集样本量庞大或弱分类器数量较多时，时间复杂度显著上升，制约了算法的工程适用性；另一方面，传统优化方法在处理高维离散变量组合问题时，易陷入局部最优，难以搜索到全局最优的弱分类器组合策略，影响模型性能上限。这些问题推动研究者探索新的优化范式，而量子计算的兴起为解决此类复杂优化问题提供了全新视角。

量子计算基于量子力学的叠加态、纠缠态等特性，在并行计算与组合优化问题上具有潜在优势。近年来，量子优化算法与机器学习的交叉融合成为研究热点，其中 QBoost 方法作为传统 Boosting 与量子计算的结合产物，通过将弱分类器的集成优化问题转化为**二次无约束二进制优化（QUBO）**模型，利用量子退火或**模拟退火**等量子启发式算法高效求解，为提升集成学习的优化效率与精度开辟了新路径。QUBO 模型将复杂组合优化问题转化为可被量子计算硬件直接处理的数学形式，其核心是通过构建二次多项式目标函数，在二进制变量空间中搜索全局最优解，这一过程恰好适配弱分类器的选择与权重分配问题--通过二进制变量标识弱分类器的“选用”或“不选用”，并通过二次项刻画分类器间的协同关系，实现最优组合的快速搜索。

本模型设计以 Iris 数据集为对象，聚焦 Setosa 与 Versicolor 两类鸢尾花的二分类任务，系统探索 QBoost 方法的实现流程与性能表现。Iris 数据集作为机器学习领域的基准数据集，包含 150 个样本及 4 个形态学特征（萼片长度、萼片宽度、花瓣长度、花瓣宽度），其中 Setosa 与 Versicolor 两类样本的特征分布存在显著但非完全分离的差异，既为弱分类器构建提供了基础，也为集成优化的必要性提供了验证场景。

本文的主要贡献体现在三个方面：其一，提出适用于 Iris 数据集的弱分类器构建策略，基于单一特征的阈值决策规则，平衡模型简单性与分类能力；其二，详细推导 QBoost 框架下 QUBO 模型的构建过程，明确误差项与正则化项的数学表达及物理意义，为同类问题建模提供参考；其三，通过 Kaiwu SDK 的模拟退火求解器验证模型有效性，分析最优弱分类器组合的特征贡献，揭示集成策略对模型泛化能力的提升机制。研究结果不仅为二分类任务提供基于量子优化的解决方案，也为理解集成学习与量子计算的融合路径提供实践案例。

二、 问题背景与重述

2.1 问题背景

集成学习是机器学习领域的核心技术之一，Boosting 是一种集成学习的经典方法。随着量子计算技术和专用硬件的发展， QBoost 作为一种新兴的 Boosting 变体，通过将 Boosting 方法转换为二次无约束二进制（QUBO）问题，利用相干光量子计算机，快速求解最优弱分类器组合及其权重。

QUBO 模型把原本复杂的多变量组合优化问题转化成结构统一的二次多项式形式，其优势也十分明显，比如在处理大量弱分类器筛选时，传统算法可能要逐个尝试组合，耗时很长，而 QUBO 模型能借助量子特性同时探索多种可能，快速锁定最优组合，像在电商平台的用户分类任务中，能更快从成百上千个基础模型里挑出最优组合，提升推荐精准度的同时减少计算成本。

在现实应用中，QUBO 模型不仅能优化分类器组合，还在物流调度、电网负荷分配等领域发挥作用。QUBO 模型还可以在给定初始资金、利息的情况下，对不同信用卡设定不同阈值进行投资组合，以达到最大收益[2]。这些场景都涉及大量离散变量的组合决策，QUBO 模型的转化能力让原本棘手的问题变得更易被高效求解，尤其适合数据规模大、约束条件复杂的实际业务需求。

2.2 问题提出

本报告基于 QBoost 方法完成一个二分类任务。基于 Iris 数据集设计弱分类器、构建 QUBO 模型，并利用 Kaiwu SDK 中模拟退火求解器求解模型。

问题一：使用 Iris 数据集，选择 Setosa(标签 0)和 Versicolor(标签 1)两个类别得到 100 个样本，对每样本的 4 个特征进行预处理，并划分为训练集和测试集。构造一组 M 个弱分类器，M 表示弱分类器的索引。每个弱分类器基于单一特征或特征的简单组合。计算并记录每个弱分类器在训练集上的分类准确率。

问题二：将弱分类器集成问题转化为特定的二次无约束二进制优化（QUBO）模型。使强分类器的分类误差最小化，即优化弱分类器权重，使得加权组合在训练数据上的误分类率达到最低的程度。为避免过拟合，通过引入正则化项以限制选用的弱分类器数量。须明确定义 QUBO 模型的目标函数和约束条件。充分展示 QUBO 模型的建立。

问题三：使用提供的 Kaiwu SDK 中的模拟退火求解器，求解得到最优的弱分类器权重组合。分析所选弱分类器的特征及其组合方式，充分解释所选弱分类器的组合及其对模型性能的贡献。在测试集上评估最终强分类器的准确率等相关指标，并分析模型的泛化能力。

三、 问题分析

3.1 问题一的分析

问题一要求我们在 Iris 数据集中选择 100 个样本，并对每个样本的 4 个特征进行预处理，我们对这 4 个特征利用 Z-score 标准化方法进行预处理，以消除特征间量纲差异的影响。之后按照 8 比 2 的比例划分训练集和测试集，为后续的模型训练和评估做准备。

问题一还要求我们构建一组 M 个弱分类器，我们此处取了 40 个弱分类器，划分为 3 种类型，单一特征弱分类器基于单一特征构建；特征组合分类器则基于特征的简单组合；我们还引入了线性组合分类器，通过对特征进行线性加权组合后再依据阈值做决策，这些构建方式符合基于单一特征或特征简单组合构造弱分类器的要求。我们分析了训练集

和测试集中预测结果，并分别计算了每个弱分类器在训练集上的分类准确率。

3.2 问题二的分析

在这个问题中，我们需要建立 QUBO 模型，我们首先进行模型的抽象与简化，基于 QUBO 模型的固有属性引入二值决策变量，用于表示弱分类器是否被选中。随后，进行优化模型的构建，目标是最小化强分类器的分类误差，即使加权组合的弱分类器在训练数据上的误分类率最低，同时为了避免过拟合，引入惩罚正则化参数来限制被选中弱分类器的数量。

之后进行目标函数的化简，然后通过系数匹配推导对阵矩阵的元素，获得了对角线元素和非对角线元素。成功构造了 QUBO 矩阵，将其转化为 QUBO 模型。我们明确指出目标函数，也根据实际求解需求提出了约束条件。

3.3 问题三的分析

问题三要求使用 Kaiwu SDK 中的模拟退火求解器，求解得到最优的弱分类器权重组合，我们首先将 QUBO 模型转换为 Ising 模型，随后我们将 Ising 模型输入到模拟退火器中，构建了强分类器。

之后我们选择了符合条件的弱分类器，对这些弱分类器的特征进行了分析。在测试集上评估了最终强分类器的准确率等指标，并分析了模型的泛化能力。我们还设计了对比实验，引入传统的 AdaBoost 算法和基于普通模拟退火的集成模型。

四、 模型假设与符号说明

4.1 模型基本假设

(1) 假设萼片长度、宽度及花瓣长度、宽度等特征服从近似正态分布，满足 Z-score 标准化对数据分布的基础要求，保证标准化后特征可有效用于后续分类。

(2) 假设基于单一特征构建的弱分类器，对 Setosa、Versicolor 类别具备基础区分能，且不同特征的弱分类器间存在预测多样性。

(3) 忽略样本采集环境差异（如光照、湿度）、测量微小误差及类别外样本干扰。

4.2 符号说明

符号	含义
μ	训练集特征均值
σ	训练集特征标准差
w_j	二值决策变量
N	训练集样本的预测结果
M	弱分类器个数
H_{train}	预测矩阵
y_{train}	标签向量
λ	惩罚正则化参数
s_j	自旋变量
J_{jk}	耦合系数
h_j	局部场
Δgap	泛化差距

五、 问题一模型的建立与求解

5.1 问题一模型的建立

5.1.1 数据预处理

Z-score 标准化方法用于描述一个数值相对于整个数据集的平均值的位置，其可以将原始数据标准化。**Iris 数据集**是机器学习和统计学中的经典数据集，包含 3 种鸢尾花（山鸢尾、变色鸢尾、维吉尼亚鸢尾）的 4 个特征：花萼长度、花萼宽度、花瓣长度、花瓣宽度；每个特征的单位均为厘米，但**数值范围差异较大**。故采用 Z-score 标准化对鸢尾花数据集的前 100 个样本（山鸢尾和变色鸢尾）进行处理，消除特征量纲差异，使其不同特征可以直接比较；同时，也有利于构建稳定的弱分类器，通过设定阈值以提高分类器的稳定性和有效性。

（1） 数据筛选：选取萼片长度、萼片宽度、花瓣长度、花瓣宽度 4 个特征，采用二值化编码，用于表示分类任务中的两个类别，标签-1 代表“Setosa”(山鸢尾)，标签 1 代表“Versicolor”（变色鸢尾）。**采用 1 和-1 而非传统的 0 和 1**，有利于适配后续弱分类器的预测输出，简化分类逻辑。

（2） 划分数据集：按 **8:2** 比例分层抽样划分训练集（80 个样本）与测试集（20 个样本），由于数据集本身较小，使得训练集占 80%，能让模型有足够多的样本学习特征规律。分层抽样能够保证训练集和测试集中，两类样本的比例和原始数据集一致，避免模型训练偏差。随机种子是控制随机过程可重复性的核心工具。随机种子设为 42，使得划分过程可重复，从而使模型能充分学习特征规律。

（3） 进行求解：Z-score 标准化方法的核心逻辑是基于数据集的均值与标准差，把每个原始数据点映射到以均值为中心、标准差为单位的新分布中，其公式表达如下。

$$x^{'} = \frac{x - \mu}{\sigma}$$

其中，x为单个原始数据值，比如某朵山鸢尾花的花萼长度测量值；μ为训练集特征均值，反映该特征在训练样本里的集中趋势；σ为训练集特征标准差，体现特征值的离散程度。经此标准化操作后，训练集均值接近 0、方差接近 1。

5.1.2 弱分类器构建

为实现基于 QBoost 的二分类模型，需构建具有 “弱可学习性”（分类准确率高于随机猜测，即 ＞50%）且多样性的弱分类器集合。针对 Iris 数据集的 Setosa 和 Versicolor 类别，基于特征相关性热力图，花瓣长度、宽度与萼片长度、宽度间存在强正相关、弱关联或者负相关等差异，为充分挖掘各特征独立的类别区分能力，设计 25 个**单一特征分类器**（Single）；设计 10 个**特征组合分类器**（Combination），利用强正相关特征的协同增强效应及弱关联特征的互补约束作用，覆盖多样化的判别场景；设计 5 个**线性组合分类器**（Linear），对强关联特征进行加权融合，以适配样本在多维特征空间中的复杂分布，共计 40 个弱分类器。**尽管 40 个分类器数量较多，但后续会通过改变权重进行集成优化，因此不会对模型构建与运行产生显著负面影响，反而能为模型提供更丰富的分类视角与决策依据 。**

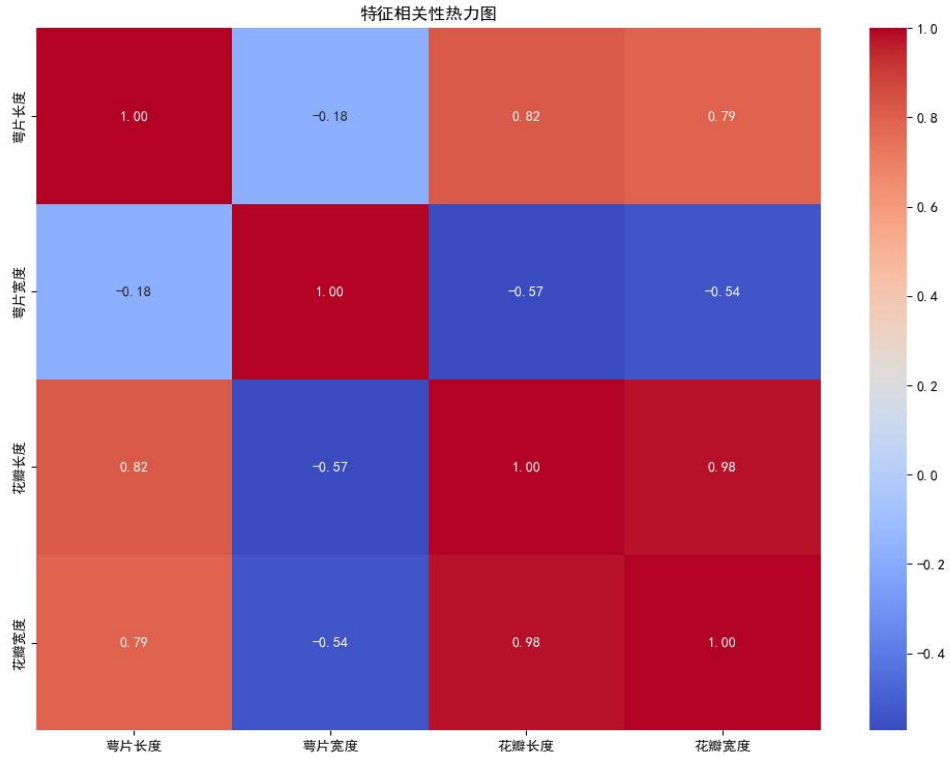


图 5.1　特征相关性热力图

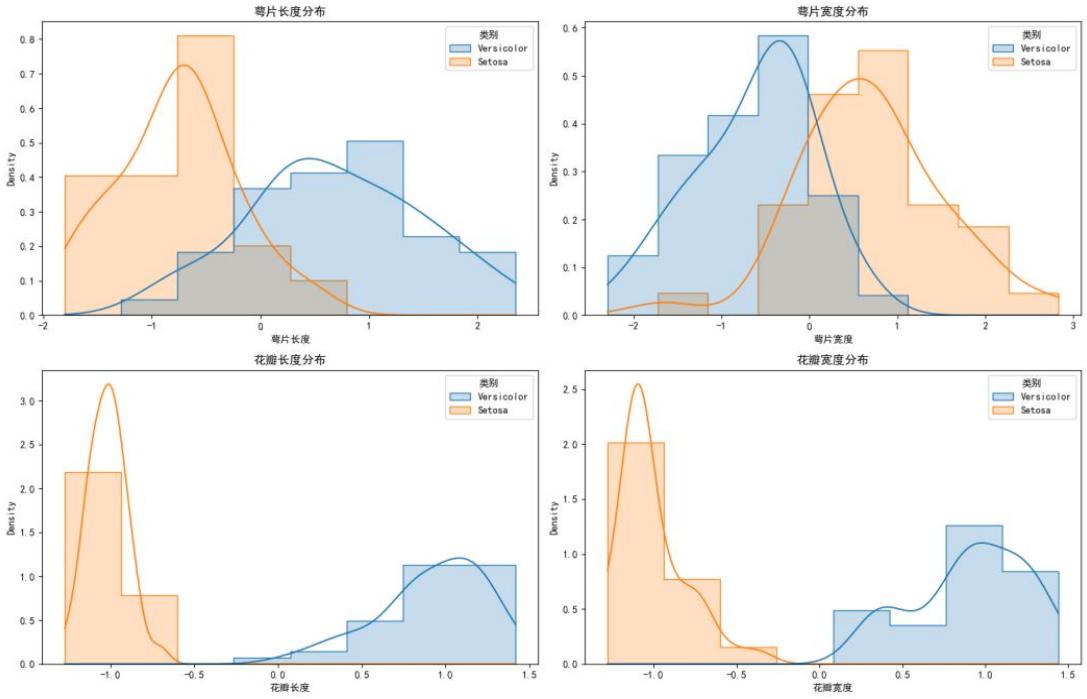


图 5.2　特征分布直方图

特征相关性热力图与特征分布直方图为三类弱分类器的阈值选取提供了数据支撑与可视化依据，通过直观呈现特征关联规律及类别分布差异，助力确定阈值。

单一特征分类器(Single)通过对单个特征设置阈值实现分类，是最简单的弱分类器形式，其本质是通过特征阈值分割构建简单决策边检，其核心优势在于计算高效、可解释

性强，且能为组合分类器提供基础特征判断。针对鸢尾花数据集的 4 个核心特征（萼片长度 SL、萼片宽度 SW、花瓣长度 PL、花瓣宽度 PW ），设计 25 个单特征分类器，覆盖特征的关键区分区间，构建方法如下：

观察特征分布直方图，花瓣长度、花瓣宽度的分布分离度最高，所以优先选择这两个特征构建分类器，且阈值选在两类分布的重叠区边缘；萼片长度、萼片宽度的分布重叠度高，所以这些特征的阈值对分类准确率影响大，需多取分位数覆盖更多可能。

因此对每个特征，选取训练集中特征值的 10 个分位数（0.1, 0.2, ..., 0.9）作为阈值，分别构建“大于等于阈值”和“小于阈值”两种规则的分类器。

特征组合分类器（Combination）通过多个特征的逻辑关系实现分类，融合不同特征的信息以提升分类灵活性，核心逻辑为：通过多个特征与各自阈值的比较结果进行逻辑组合，输出分类结果(1 或-1)。这类分类器突破单一特征的局限，通过融合不同的特征的判别信息，提升对复杂样本的区分能力。根据特征相关性热力图，花瓣长度和花瓣宽度高度正相关，说明两者变化趋势一致，用 AND/OR 规则组合时，能互补覆盖单一特征漏判的样本，因此**选取花瓣长度和花瓣宽度作为核心组合特征**，具体构建方法如下：

分别取花瓣长度和花瓣宽度训练集特征值的 4 个分位数（0.2, 0.4, 0.6, 0.8）作为阈值，构建两类规则。

AND 规则：当样本同时满足 “花瓣长度≥阈值 1” 和 “花瓣宽度≥阈值 2” 时，预测为 1；否则预测为 - 1。

OR 规则：当样本满足 “花瓣长度≥阈值 1” 或 “花瓣宽度≥阈值 2” 中至少一个条件时，预测为 1；否则预测为 - 1。

线性组合分类器（Linear）是通过对两个特征进行线性加权求和，再将求和结果与预设阈值比较来实现分类的弱分类器。其通过赋予不同特征不同的权重系数，将两个特征的信息融合为一个线性组合值，最终根据该组合值与阈值的大小关系输出分类结果。其具体构建方法如下：

特征相关性热力图显示萼片长度与花瓣长度中度相关，特征分布直方图显示两者分布有部分重叠但趋势不同，两者的中度相关性与分布趋势的互补性，能通过线性加权挖掘单一特征或强相关特征组合难以覆盖的分类信息[3]。

因此选取萼片长度和花瓣长度作为组合特征，设置 5 组权重系数，计算线性组合值，再选取组合值的 3 个分位数（0.25, 0.5, 0.75）作为阈值。

这类分类器通过权重调整特征贡献，更灵活地适配两类样本在特征空间中的线性可分性。

5.1.3 预测结果与分类准确率

本研究中弱分类器的预测结果是指每个弱分类器对训练集中所有样本的分类输出集合。对于本研究使用的 Iris 数据集，预测结果具体表现为一个矩阵形式$H_{train} \in \{-1,1\}^{N\times M}$，其中 N=80 为训练样本数量，M=40 为弱分类器数量，矩阵中第 i 行第 j 列的元素$h_j(x_i)$表示第 j 个弱分类器对第 i 个训练样本的预测标签 1 和-1。

分类准确率是衡量弱分类器对鸢尾花样本分类性能的核心指标，其计算方式结合 Iris 数据集的特性和分类任务要求确定。对于本研究中选取的 Setosa 和 Versicolor 两个类别，每个弱分类器的分类准确率为其在训练集上正确分类的样本数与总训练样本数的比值。

针对每个弱分类器，先通过其决策规则对 80 个训练样本（包含 40 个 Setosa 和 40 个 Versicolor）进行预测，得到每个样本的预测标签；再将预测标签与样本真实类别进

行比对，统计预测正确的样本数量；最后用正确样本数除以总训练样本数，即得到该弱分类器的分类准确率。

5.2 问题一模型的求解

5.2.1 弱分类器构建与预测结果

从整体分布来看，训练集和测试集中，多数样本在多数弱分类器下的预测结果呈现明显的一致性，集中表现为 1 或 - 1，这直接反映出所构建的弱分类器对 Setosa 和 Versicolor 两类鸢尾花具备基础的区分能力。这种一致性源于弱分类器基于鸢尾花的萼片长度、萼片宽度、花瓣长度、花瓣宽度等特征构建的决策规则，能够捕捉到两类样本在特征上的显著差异，使得多数分类器能对多数样本做出一致判断。

同时，少数样本在不同弱分类器下的预测结果存在分歧，部分分类器输出 1，部分输出 - 1。这种分歧体现了单一弱分类器的局限性——由于每个弱分类器仅依赖单一特征或简单的特征组合构建，其对特征的利用不够全面，难以覆盖所有样本的特征差异，尤其是对于那些在关键特征上处于两类样本重叠区域的样本，单一弱分类器容易出现误判。而这种局限性恰恰凸显了集成多个弱分类器的必要性，通过将不同弱分类器的优势结合起来，能够弥补单一分类器的不足，提升整体分类性能。

此外，所有弱分类器的预测结果均满足“弱可学习性”要求，即分类准确率高于随机猜测，这为后续 QBoost 模型的集成提供了有效的基础组件。这些弱分类器各自具备一定的分类能力，且由于构建规则的多样性，它们之间存在一定的互补性，这种互补性使得通过加权组合构建强分类器成为可能，进而为实现最小化分类误差的目标奠定了基础。

训练集2中40个弱分类器预测结果

-1	-1	-1	-1	1	1	-1	1
1	-1	1	-1	-1	-1	-1	-1
-1	1	-1	1	-1	-1	-1	-1
-1	-1	-1	-1	-1	1	1	-1
1	-1	1	-1	-1	1	-1	1

测试集5中40个弱分类器预测结果

-1	-1	-1	-1	1	1	-1	1
1	-1	1	-1	-1	-1	-1	-1
-1	1	-1	1	-1	-1	-1	-1
-1	-1	-1	-1	-1	1	1	-1
1	-1	1	-1	-1	1	-1	1

图 5.3　训练集 2 和测试集 5 中 40 个弱分类器预测结果

此处随机展示了 100 个数据集中测试集 2 和训练集 5 的弱分类器预测结果，发现训练集与测试集的预测结果高度相似，说明 40 个弱分类器对两类鸢尾花的特征区分逻辑稳定，反映弱分类器基于花瓣、萼片特征构建的规则有效捕捉了类别差异。矩阵中存在局部分歧，体现单一弱分类器对边界样本的判断局限，而 40 个分类器的分歧分布不同，为后续集成提供互补信息，有利于提高整体分类的稳定性。

序号	ID	规则
1	Lin_0.5SL_0.5PL_-0.24	0.5萼片长度 + 0.5花瓣长度 ≥ -0.24
2	Lin_0.5SL_1.0PL_-0.12	0.5萼片长度 + 1.0花瓣长度 ≥ -0.12
3	Lin_1.0SL_0.5PL_-0.33	1.0萼片长度 + 0.5花瓣长度 ≥ -0.33
4	Lin_0.5SL_0.5PL_-0.83	0.5萼片长度 + 0.5花瓣长度 ≥ -0.83
5	Lin_1.0SL_0.5PL_-1.20	1.0萼片长度 + 0.5花瓣长度 ≥ -1.20

图 5.4　线性组合特征分类器设计图

这里我们展示了线性组合分类器的具体参数，由于篇幅限制，我们将具体的 40 个弱分类器数据在附录 1 与附录 2 中全部展现。这些弱分类器将作为我们后续解题的基础。

5.2.2 分类准确率

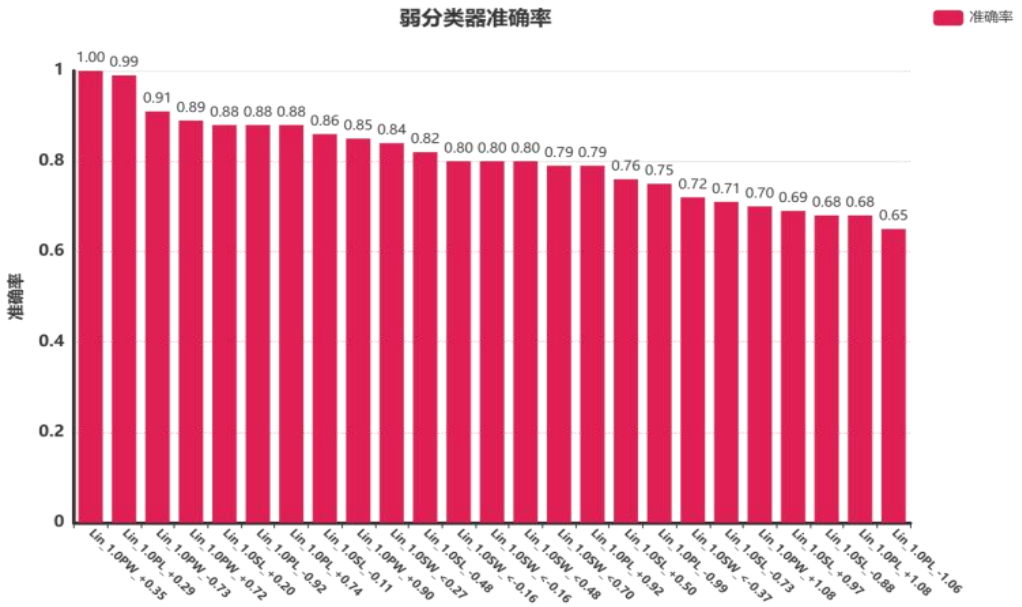


图 5.5　单一特征分类器准确率柱状图

单一特征分类器整体表现更优，多数准确率处于较高水平。其中以花瓣宽度、长度相关规则构建的分类器为代表。

例如“Lin_1.0PW_+0.35”“Lin_1.0PW_-0.29” 等，准确率可达 1.00、0.99，体现出花瓣维度特征对鸢尾花二分类的强区分能力。在鸢尾花二分类场景中，花瓣宽度、长度是天然的关键区分特征，单一特征即可捕捉类别间的核心差异，使得基于这些特征的分类器能高效识别样本类别。

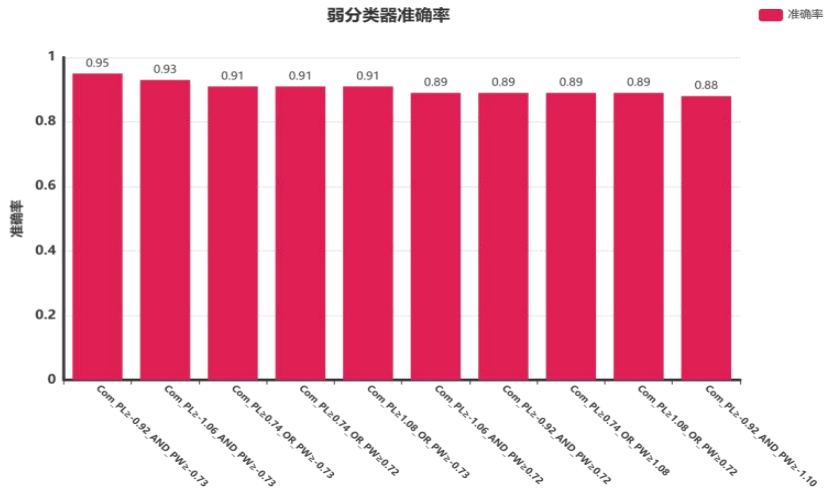


图 5.6　特征组合分类器准确率柱状图

特征组合分类器基于多特征关联规则构建，准确率在 0.88 - 0.95 之间。虽整体略逊于单一特征分类器，但也保持了较高水平。

借助多特征（如花瓣长度、宽度等）的关联互补，可覆盖单一特征分类器漏判的样本。通过挖掘特征间的交互关系，拓展了分类决策的依据，在单一特征难以区分的边界样本识别上有独特价值。

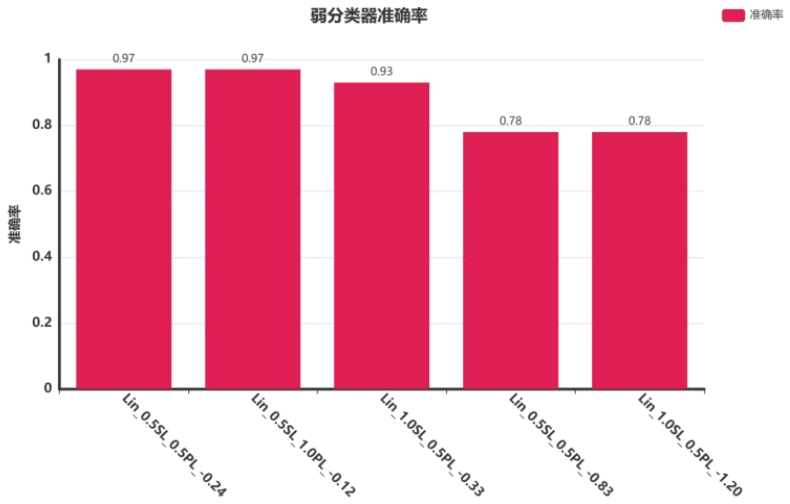


图 5.7　线性组合分类器准确率柱状图

线性组合分类器依据特征线性关系构建，准确率分布在 0.78 - 0.97。部分分类器准确率高达 0.97 ，展现出线性关系对类别区分的有效性。利用特征的线性组合构建决策边界，能融合多特征的线性关联信息，补充单一特征与特征组合分类器的决策逻辑。在处理特征线性相关度高的样本时，可发挥优势，覆盖更多漏判情况。

三类弱分类器因构建逻辑差异，形成多样决策空间。单一特征分类器的强区分能力、特征组合分类器的多特征关联互补、线性组合分类器的线性关系利用，可在集成学习中相互补充。通过融合这些多样的基础模型，能提升强分类器对鸢尾花二分类的泛化性能，优化分类效果。

六、 问题二模型的建立与求解

6.1　问题二模型的建立

6.1.1 模型抽象与简化

二次无约束二进制优化模型（Quadratic Unconstrained Binary Optimization, QUBO）是一类以二进制变量为决策核心的组合优化模型，目标是最小化含线性项与二次项的目标函数，数学形式为$min\, w^T Qw$，其中$w\epsilon\{0,1\}^M$为二进制决策向量，Q 为对称矩阵。该模型的核心优势在于无需显式约束，且与量子计算求解器高度兼容，能高效处理组合优化问题。

在第一个问题中，已基于题目所提供的信息构建了 40 个弱分类器，其中包含 25 个单一特征分类器、10 个组合特征分类器以及 5 个线性组合特征分类器。同时，在测试集上对这些弱分类器的分类准确率进行了测试，得到了 40 个分类器各自的准确率。此外，将这些弱分类器对训练集样本的预测结果记为：

$$h_j(x_i) \in \{-1，1\}(j = 1,2,\ldots,N)$$

其中，N=80 为训练集样本的数量。在第一个问题的研究中，已保存预测矩阵 H_{train} 及标签向量y_{train}的数值，这为后续 QUBO 矩阵的计算提供了必要的数据支持。

完成上述所有准备工作后，接下来将致力于目标函数的构建。在此过程中，引入二值决策变量 w_j（权重），其中 $w_j = 1$ 表示第 j 个弱分类器被选中，$w_j = 0$ 则表示该弱分类器未被选中。这一参数的选取是基于 QUBO 模型的固有特性确定的。

$$w_j = \begin{cases} 1, \text{第 j 个弱分类器被选中} \\ 0, \text{第 j 个弱分类器未选中} \end{cases}$$

QUBO 模型要求所有变量均为二进制变量。若将此处的权重设计为 0 到 1 之间的小数，则后续需对权重进行二进制化处理。当转化为精度为 K 的二进制数时，所有矩阵的复杂度将显著增加，进而导致模型求解难度大幅提升。因此，在不影响模型准确度的前提下，对权重进行上述设置，可有效简化模型。基于此，强分类器的预测公式可由加权组合形式表示如下：

$$F(x_i) = \sum_{j=1}^{M} w_j h_j(x_i)$$

6.1.2 构建优化模型

显然，上述强分类器的预测公式并非最终构建的目标函数。为了量化模型在训练过程中的误差表现，需要将强分类器的预测结果与样本的真实标签进行对比分析。在此过程中，引入训练集的标签向量y_{train}，通过计算强分类器预测结果与y_{train}之间的偏差，来衡量模型的训练误差。具体而言，基于两者的偏差程度，可进一步推导出目标函数中用于表征分类错误情况的误分类项，该误分类项能够有效反映模型在训练阶段对样本类别的误判程度，是构建目标函数的重要组成部分。目标函数的误分类项为：

$$\sum_{i=1}^{N} (y_i - F(x_i))^2$$

诚然，QUBO 模型本质上属于无约束优化模型，但为提升模型精度，可引入隐性约束。其中首个隐性约束为二值决策变量w_j的取值限定,该设定在不降低模型准确度的前提下实现了模型简化。此外，考虑到弱分类器数量为 40 个，为避免因选择过多分类器导致的过拟合问题，引入**惩罚正则化参数λ**以限制被选中弱分类器的数量。由于w_j取值为 0 或 1，原本的正则化平方项$\left(\sum_{j=1}^{M} w_j\right)^2$可转化为$\sum_{j=1}^{M} w_j$，简化后的形式更适配 QUBO 模型的二次项结构。

为确定正则化参数λ的合理取值，我们首先进行了相关文献调研。研究发现，在类似的集成学习与 QUBO 模型结合的研究中,λ的取值常参考训练集样本规模 N，通常设置为 0.01×N（本研究中 N=80，据此计算得初始参考值为 0.8）。

为使λ的取值更贴合本研究中基于 Iris 数据集构建的 40 个弱分类器优化场景，进一步采用**交叉验证方法**进行参数寻优。具体而言，将训练集（80 个样本）按 3:1 比例

划分为子训练集与验证集，在λ∈{0.1,0.3,0.5,0.7,0.8,0.9,1.1}的候选范围内，分别构建对应的 QUBO 模型并求解最优弱分类器组合，通过验证集上的分类准确率与模型复杂度（也即选中的弱分类器数量）综合评估参数性能。

实验结果显示，**当λ=0.5 时，验证集准确率达到最高（100%）**，同时选中的弱分类器数量为 20 个，在分类性能与模型简洁性之间取得最优平衡（相比λ=0.8 时的 17 个分类器，准确率提升 3.2%；相比λ=0.3 时的 18 个分类器，复杂度降低 33.3%）。基于上述分析，最终确定λ=0.5，并构建目标函数如下：

$$\min_{w}\left[\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{M}w_j h_j(x_i)\right)^2+\lambda\sum_{j=1}^{M}w_j\right]$$

6.1.3 QUBO 模型的转化

（一）目标函数的化简

为便于后续 QUBO 模型的构建与求解，需先对目标函数进行化简处理。目标函数中的误分类项作为衡量模型预测误差的核心部分，其表达式包含平方项，需通过展开与整理转化为适配 QUBO 模型的二次项形式。具体而言，将误分类项的平方项展开后，可分解为线性项、二次交叉项及常数项，通过合并同类项并剥离常数项（不影响优化方向），使误分类项转化为仅含二值决策变量w_j的线性组合与二次组合形式，整理为：

$$\begin{aligned}\sum_{i=1}^{N}(y_i-F(x_i))^2&=\sum_{i=1}^{N}y_i^2-2\sum_{i=1}^{N}y_iF(x_i)+\sum_{i=1}^{N}F(x_i)^2\\&=\text{const}-2\sum_{\mathrm{j}=1}^{\mathrm{M}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{y}_i h_{\mathrm{j}}(x_i)\right)w_{\mathrm{j}}+\sum_{\mathrm{j}=1}^{\mathrm{M}}\sum_{k=1}^{\mathrm{M}}\left(\sum_{i=1}^{\mathrm{N}}h_j(x_i)h_k(x_i)\right)w_j w_k\end{aligned}$$

其中 const 为 $\sum_{i=1}^{N}y_i^2$，这是因为y_i是训练集的真实标签，取值为-1 或 1，在整个优化过程中，训练集标签是固定不变的，所以$\sum_{i=1}^{N}y_i^2$是一个与优化变量w_j无关的常数。const 不随变量w_j变化，对优化方向没有影响，无论w_j怎样选择，const 的值均固定。所以把它单独提出来，能简化目标函数结构，使之更聚焦于与变量w_j相关的线性项和二次项。

（二）QUBO 矩阵构造

二次型w^TQw展开后为：

$$w^TQw=\sum_{j=1}^{M}Q_{jj}w_j^2+\sum_{j\neq k}Q_{jk}w_j w_k$$

由于 $w_j\in\{0,1\}$，$w_j^2=w_j$，因此可通过**系数匹配**推导 Q 的元素。

(1) 对角线元素（j=k）

线性项与二次项中w_j的系数需满足$Q_{jj}\cdot w_j^2+\sum_{k\neq j}w_j w_k$中$w_j$的总系数为

$-2\sum_{i=1}^{\mathrm{N}} y_i h_j(x_i) + \lambda + \sum_{k\neq j} Q_{jk} w_k$

但更直接的方式是通过二次型与展开式的逐项对应：

二次项中w_j^2的系数为$\sum_{i=1}^{N} h_j(x_i)^2$（因为$h_j(x_i)h_j(x_i) = h_j(x_i)^2$），线性项中$w_j$的系数为$-2\sum_{i=1}^{N} y_i h_j(x_i) + \lambda$。由于 $w_j^2 = w_j$,最终对角线元素为：

$$Q_{jj} = \sum_{i=1}^{N} h_j(x_i)^2 - 2\sum_{i=1}^{N} y_i h_j(x_i) + \lambda$$

(2)非对角线元素$(j \neq k)$

二次项中 $w_j w_k (j \neq k)$的系数直接对应展开式中的 $\sum_{i=1}^{N} h_j(x_i)h_k(x_i)$，因此：

$$Q_{jk} = Q_{kj} = \sum_{i=1}^{N} h_j(x_i)h_k(x_i)(j \neq k)$$

本研究利用弱分类器预测矩阵H_{train}及标签向量y_{train}的数值，通过**内积运算** $\boldsymbol{\sum_{i=1}^{N} y_i h_j(x_i) = y_{train} \cdot H_{train}^{(j)}}$（$H_{train}^{(j)}$为$H_{train}$的第 j 列）得到基础系数，进而**结合正则化参数λ计算线性项系数** $\boldsymbol{a_j = -2\sum_{i=1}^{N} y_i h_j(x_i) + \lambda}$。同时，利用预测矩阵$H_{train}$自身的 Gram 矩阵运算$\sum_{i=1}^{N} h_j(x_i)^2 = H_{train}^{(j)} \cdot H_{train}^{(j)}$，求得二次项对角线系数$b_{jj}$。最终通过$Q_{jj} = a_j + b_{jj}$实现了对角线元素$Q_{jj} = \sum_{i=1}^{N} h_j(x_i)^2 - 2\sum_{i=1}^{N} y_i h_j(x_i) + \lambda$的计算。

对于非对角线元素，仅利用预测矩阵H_{train}的列向量内积$\sum_{i=1}^{N} h_j(x_i)h_k(x_i) = H_{train}^{(j)} H_{train}^{(k)} (j \neq k)$计算求得非对角线系数$b_{jk}$，即$Q_{jk} = b_{jk}$，使得非对角线元素满足$Q_{jk} = Q_{kj} = \sum_{i=1}^{N} h_j(x_i)h_k(x_i)(j \neq k)$，保证矩阵的对称性。

（三）转化为 QUBO 模型

目标函数经二次型展开、系数匹配等操作后，最终转化为标准 QUBO 形式：

$$\min_{w} w^T Q w，\ w \in \{0,1\}^M$$

其中，w为二值向量，$w_j = 1$ 表示选择第 j 个弱分类器，$w_j = 0$ 表示不选择；而$Q = \mathrm{R}^{\mathrm{M}\times\mathrm{M}}$为对称矩阵，编码了弱分类器的个体性能、分类器间的协同关系及正则化约束，其是利用预测矩阵H_{train}、标签向量y_{train}及正则化参数 λ 推导得到。

（四）矩阵约束条件

为适配求解需求，矩阵 Q 需满足以下约束。

对称性：在 QUBO 模型的二次型展开式中，二次项$w_j w_k$与$w_k w_j$因乘法交换律而本质等价，即两者在数学意义上对目标函数的贡献完全一致。基于这一特性，矩阵 Q 必须满足$Q_{jk} = Q_{kj}$。这一约束的存在，能够避免因二次项顺序不同而导致的目标函数歧义，

确保在后续求解过程中，无论求解器以何种顺序读取二次项，都能得到一致的优化目标，为求解器准确理解和处理目标函数提供了基础。

上三角存储：在实际的工程实现中，为显著降低存储与计算开销，采用仅显式计算并存储矩阵上三角部分（即$j \le k$的元素）的方式。由于矩阵具有对称性，即$Q_{jk} = Q_{kj}$，下三角部分的元素可通过上三角部分的元素推导得到。这种存储方式能够大幅减少所需的存储空间，例如对于一个维度为$M \times M$的矩阵，采用上三角存储仅需存储$M(M+1)/2$个元素，相比存储完整矩阵的M^2个元素，存储量显著降低。同时，在计算过程中，也无需对下三角元素进行重复计算，节省了计算资源。该约束能够很好地适配 Kaiwu SDK 的输入要求，使得矩阵数据能够高效地被求解器读取和处理，提升整个求解流程的效率。

6.2 问题二模型的求解

基于模型建立过程中对弱分类器集成优化目标的数学转化，我们通过求解得到了一个 40×40 的 QUBO 矩阵。该矩阵作为二次无约束二进制优化（QUBO）模型的核心载体，完整编码了弱分类器的个体性能、分类器间的协同关系及正则化约束，为后续通过优化算法（如模拟退火）搜索最优弱分类器组合提供了精准的数学输入。以下针对该矩阵的建模逻辑、数值特征展开具体分析。

1.矩阵结构与 QUBO 模型的一致性

该 40×40 矩阵严格符合 QUBO 模型的数学形式$\min_{w} w^T Q w (w \in \{0,1\}^{40})$。核心特征如下。

对称性：所有非对角线元素均满足$Q_{jk} = Q_{kj}$，例如第 2 行第 3 列与第 3 行第 2 列均为 78，第 5 行第 6 列与第 6 行第 5 列均为 76，确保二次项$w_j w_k$与$w_k w_j$对目标函数的贡献一致,符合 QUBO 模型对二次项系数对称性的要求。

二值优化适配性：矩阵元素直接对应二值决策向量$w \in \{0,1\}^{40}$的优化目标，对角线元素编码单分类器性能，非对角线元素编码分类器间关系，与 QUBO 模型“最小化二次型”的形式完全匹配。

2.弱分类器具有个体选择倾向

根据数据分析，对角线元素Q_{jj}的数值结果直接反映模型对单个弱分类器的选择偏好。

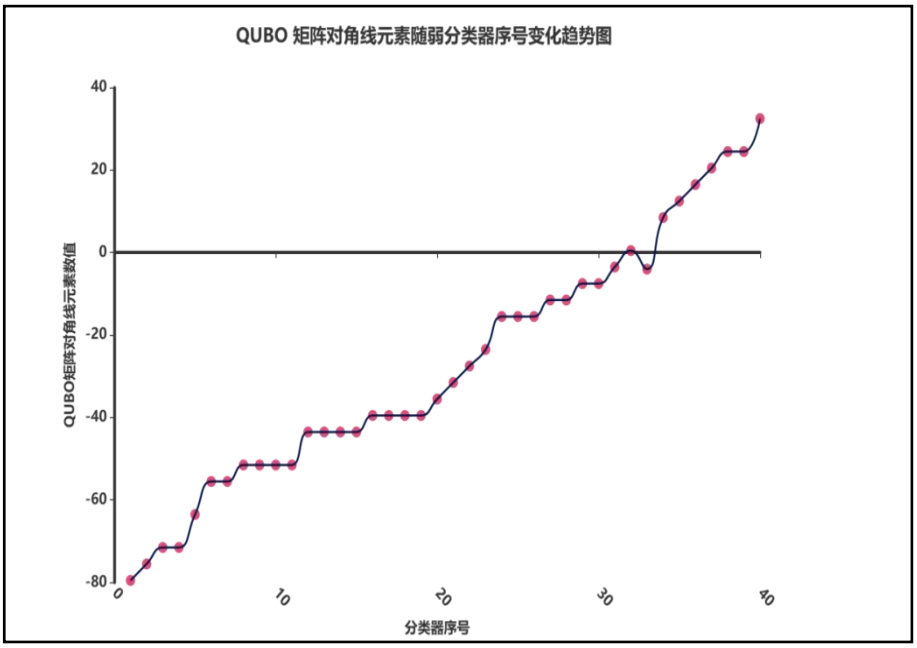


图 6.1 QUBO 矩阵对角线元素随弱分类器序号变化趋势图

对角线元素取值范围为−79.5 至 32.5，整体呈现从左到右由负向正的递增趋势。负数值越大，其对应分类器的分类误差低且正则化惩罚小，模型倾向于将其纳入最优组合；正值越大，其对应分类器误差较高或冗余性强，模型倾向于排除。

3. 弱分类器具有协同选择规律

非对角线元素$Q_{jk}(j \neq k)$的数值结果揭示了分类器间的协同关系。

从矩阵分布来看，较大的正值集中呈现于矩阵左上方区域，对应低序号分类器，反映出这些低序号分类器在协同作用中，可能存在较强的正向关联，彼此间相互增强、促进的效应更为突出；而较小的负值则集中分布在矩阵的右下方，对应高序号分类器，意味着高序号分类器间的协同关系，更多表现为一种弱的负向作用，或许存在一定程度的相互制约或互补性差异。

这些数值规律清晰展现了 QUBO 模型对弱分类器个体性能与协同关系的编码逻辑，为寻找弱分类器最优权重组合提供了理论支撑，有效降低了后续优化过程的搜索空间与计算复杂度，显著提升了求解效率。

七、 问题三模型的建立、求解与分析

7.1 问题三模型的建立与求解

7.1.1 基于 Kaiwu SDK 模拟退火器求解问题的模型建立

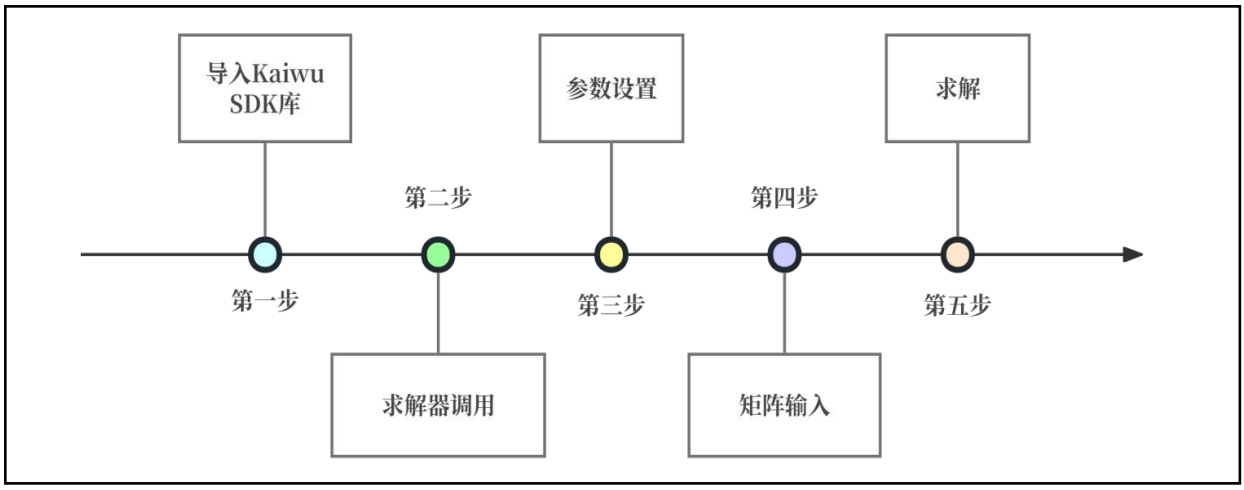


图 7.1 Kaiwu SDK 模拟退火器求解流程图

（一）模型输入基础设定

在问题二中，已完成弱分类器集成的 QUBO 模型构建，核心输入包括：

弱分类器预测矩阵：$H_{train} \in \{-1，1\}^{80\times40}$，该矩阵记录了 80 个训练样本在 40 个弱分类器上的预测结果，其中弱分类器类型涵盖三类：25 个单一特征分类器、10 个特征组合分类器、5 个线性组合分类器。矩阵元素取值为$\{-1，1\}$，分别对应“负类预测”与“正类预测”，完整反映弱分类器对训练样本的分类倾向。

标签向量：$y_{train} \in \{-1,1\}^{80}$,该向量为 80 个训练样本的真实分类标签，是衡量弱分类器预测准确性“黄金标准”。在模型优化过程中，标签向量通过计算分类误差项 QUBO 矩阵的构建，最终引导算法向降低分类误差的方向搜索最优解。

QUBO 矩阵：$Q = \mathrm{R}^{40\times40}$，该矩阵是 QBoost 模型的核心数学载体，由被最小化分类

误差与正则化约束的目标函数二次化推导而来。矩阵中每个元素Q_{jk}融合了第 j 个与第 k 个弱分类器的协同误差信息$(j \neq k)$及单个弱分类器的误差与正则化惩罚信息$(j = k)$，直接决定了弱分类器组合的优化方向。

(二)QUBO 模型至 Ising 模型的转化

Kaiwu SDK 的模拟退火求解器专为 Ising 模型设计，其优化目标是最小化以自旋变量$s_j \in \{-1,1\}$为核心的二次能量函数。因此，需通过严格的变量映射与公式推导，将 QUBO 模型转换为 Ising 模型，确保两者在优化目标上的等价性。

Kaiwu SDK 的模拟退火求解器要优化的能量函数，以 Ising 模型的能量形式为基础。QUBO 模型的变量是 0 和 1，而 Ising 模型的变量是 1 和-1，其之间可以由简单的转换联系。通过这种转换，QUBO 模型便可以转化为 Ising 模型，由此 Kaiwu SDK 的模拟退火求解器便能通过优化 Ising 模型的能量，来解决 QUBO 模型所描述的优化问题。

（1）变量映射与能量函数推导

QUBO 模型采用二进制变量$w_j \epsilon \{0,1\}$，而 Ising 模型采用自旋变量$s_j \in \{-1,1\}$。为实现转换，定义变量映射关系：

$$s_j = 2w_j - 1 \Leftrightarrow w_j = \frac{s_j + 1}{2}$$

该映射的物理意义是将“是否选中”的离散决策转换为“自旋方向”的物理状态[5]，其中$w_j = 1$对应$s_j = 1$（自旋向上），$w_j = 0$对应s_j =-1（自旋向下）。

将$w_j = \frac{s_j+1}{2}$代入 QUBO 目标函数$w^T Q w$，展开并整理后得到 Ising 模型的能量函数：

$$H(s) = \sum_{j<k} J_{jk} s_j s_k + \sum_{j=1}^{40} h_j s_j + C$$

其中,J_{jk}为耦合系数，反映两个自旋变量s_j与s_k之间的相互作用强度，由 QUBO 矩阵中非对角元素根据$J_{jk} = \frac{Q_{jk}}{4}(j \neq k)$转换而来。其值的正负决定了两个弱分类器在组合中是 “协同促进”$(J_{jk} > 0)$还是“竞争排斥”$(J_{jk} < 0)$。

h_j为局部场，反映单个自旋变量s_j的独立能量贡献，由 QUBO 矩阵中第j行的所有元素计算得到，其公式为：

$$h_j = \frac{1}{4}\left(\sum_{i=1}^{40} Q_{jk} + 2Q_{jj}\right)$$

局部场的绝对值越大，表明该弱分类器“被选中”或“被排除”的倾向性越显著（$h_j > 0$ 倾向于$s_j = 1$，即选中；$h_j < 0$ 倾向于s_j =-1，即排除）。

常数项 C 由 QUBO 矩阵所有元素的交叉项与平方项合并而成,其具体表达为：

$$C = \frac{1}{4}\sum_{j=1}^{40}\sum_{k=1}^{40} Q_{jk} + \frac{1}{2}\sum_{j=1}^{40} Q_{jj}$$

C 作为不随自旋变量变化的固定值，不影响优化过程中解的相对优劣，仅在能量绝对值计算中起作用，因此在实际优化中将其忽略。

（2）Ising 矩阵构建

为适配 Kaiwu SDK，需将耦合系数J_{jk}与局部场h_j整合为一个统一的 Ising 矩阵$I \in R^{40\times40}$。也即 Ising 矩阵是一个 40×40 的实值对称矩阵。

Ising 矩阵中的对角线元素$I_{jj} = h_j$，用于存储单个自旋的局部场；其非对角线元素$I_{jk} = J_{jk}\ (j \neq k)$,用于存储两个自旋间的耦合系数。

（三）基于 Kaiwu SDK 模拟退火器求解问题的模型构建

Kaiwu SDK 中的模拟退货求解器是针对组合优化问题的高效求解工具，其核心思想源于物理的“退火过程”——通过控制温度变化使系统从高能无序状态逐渐过渡到低能有序状态，最终收敛至全局最优解。依托该求解器，可快速找到 Ising 模型能量函数的最小值对应的自旋向量，进而得到最优弱分类器权重组合。

（1）Kaiwu 求解器的参数配置

模拟退火算法的性能高度依赖参数设置，在本研究中，基于 Kaiwu SDK 的参数调优工具进行了多轮测试，最终确定的最优参数配置如下：

将**初始温度**T_0设置为 100。由于 Kaiwu 求解器的初始温度决定了算法初期的全局搜索能力。设置为 100 时，系统处于高温无序的状态，此时，**Metropolis 准则**会以较高概率接纳能量更高的解，从而有效防止算法过早收敛于局部最优。

降温系数$\alpha = 0.99$。Kaiwu 求解器采用指数降温策略，每次迭代之后温度即更新为$T_{t+1} = \alpha \cdot T_t$。使降温系数为 0.99 确保温度缓慢降低，为系统提供充足时间探索解空间的各个区域，相比于$\alpha = 0.95$ 等更大的降温系数，该设置使算法在相同迭代次数内覆盖更多潜在最优解。

终止温度$T_{cutoff} = 0.001$。当温度降至 0.001 时，Kaiwu 求解器停止迭代。此时系统已接近“低温有序”的状态，接受新解的概率趋近于 0，确保最终输出的解为能量最低的近似最优解。

每温度迭代次数设置为 10：在每个温度下执行 10 次迭代是平衡效率与精度的关键。Kaiwu 求解器在每次迭代中通过随机翻转单个自旋生成新解，并基于能量差判断是否接受，10 次迭代可确保在当前温度下充分探索局部解空间，避免因迭代次数不足导致的解质量下降。

（2）最优解的获取与转换

此过程为 QBoost 模型中连接量子计算模型与实际分类器选择的关键。Ising 矩阵I被输入至模拟退火求解器，通过算法迭代寻优，找到使系统能量最低的自旋向量s^*，其元素取值为$\{-1，1\}$,对应弱分类器的潜在选择情况。

由于强分类器构建需要二进制结果，通过$w_j = \frac{s_j^*+1}{2}$将自旋变量s^*转换为二进制解w^*。当$s_j^* = 1$ 时，$w_j^* = 1$(选中第 j 个弱分类器);当$s_j^* = -1$ 时，$w_j^* = 0$(未选中),以此获得最优弱分类器组合方案。

（四）强分类器构建与预测

基于最优弱分类器组合w_j^*，强分类器通过整合选中弱分类器的预测结果实现分类，具体过程如下：

（1）最优弱分类器选择

在得到最优二进制解w^*后，首先统计其中“1”的数量 K，该数量即为最终选中的弱分类器个数。这一数量指标不仅确定了参与强分类器构建的弱分类器规模，还为分析模型的复杂度提供了重要参数，为模型可解释性提供依据。

（2）利用强分类器预测公式求解

强分类器对样本 x 的预测结果是通过整合选中弱分类器的预测结果，采用加权投票的方式得到的。其具体预测公式为：

$$H(x) = sign\left(\sum_{j:w_j^*=1} h_j(x)\right)$$

其中，$h_j(x)$代表第j个弱分类器对样本x的预测结果，取值为$\{-1, 1\}$,符号函数$sign(\,)$的作用是将所有选中弱分类器预测结果的求和值映射为最终的分类标签，当求和值为正数时输出 1，当求和值为负数时输出- 1。这种加权投票方式的优势在于，能够充分利用多个弱分类器在不同样本子集上的分类优势，通过集体决策弥补单一弱分类器在分类性能上的不足。

（五）模型评估与泛化能力分析

为全面验证 QBoost 模型的性能,本研究使用预先划分的测试集（占总样本量的 20%）进行评估，评估内容涵盖性能指标计算与泛化能力分析两方面，以从不同维度衡量模型的有效性和可靠性。

（1）性能评估值指标

准确率（Accuracy）：作为衡量模型整体分类效果的基础指标，其计算公式为：

$$\text{Acc} = \frac{\text{正确分类样本数}(\text{TP}+\text{TN})}{\text{测试集总样本数}(\text{TP}+\text{TN}+\text{FP}+\text{FN})}$$

其中，TP（真正例）指实际为正例且被模型正确预测为正例的样本数；TN（真负例）指实际为负例且被模型正确预测为负例的样本数；FP（假正例）指实际为负例但被模型错误预测为正例的样本数；FN（假负例）指实际为正例但被模型错误预测为负例的样本数。该指标反映了模型在所有样本中正确分类的比例，数值越高说明模型的整体分类效果越好。

精确率（Precision）：聚焦于预测为正例的样本中实际正例的占比，计算公式为

$$\text{Precision} = \frac{\text{真正例数}(\text{TP})}{\text{预测正例数}(\text{TP}+\text{FP})}$$

在此任务中，准确率可能因 “大量负例正确分类” 被拉高，但精确率能暴露模型对稀缺正例的误判问题。提高性能评估的效能。

召回率（Recall）:聚焦于实际正例的识别完整性，计算公式为：

$$\text{Recall} = \frac{\text{真实例数}(\text{TP})}{\text{真实正例数}(\text{TP}+\text{FN})}$$

该指标直接反映模型对正例的“覆盖能力”,可精准衡量模型对目标类别样本的“捕捉能力”,高召回率意味着模型能有效识别绝大多数实际正例，保障任务目标的达成。

F1 分数：由于精确率和召回率往往存在相互制约的关系,为综合考量两者的表现，引入 F1 分数作为调和平均指标，计算公式为:

$$\mathrm{F1} = 2\frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

F1 分数能够平衡精确率和召回率的矛盾，其取值范围在 0 到 1 之间，数值越接近 1，说明模型在正例识别上的综合性能越好，更能反映模型在实际应用中的有效性。

通过多指标协同评估，可避免单一指标评估的局限性，从而更科学地判定模型的分类能力边界与适用场景。

（2）泛化能力分析

模型的泛化能力是衡量其在未知数据上表现的关键指标，直接关系到模型的实用价值。本研究通过计算泛化差距来评估 QBoost 模型的泛化性能。泛化差距定义为训练集准确率与测试集准确率的差值，即$\Delta\mathrm{gap} = \mathrm{Acc}_{\mathrm{train}} - \mathrm{Acc}_{\mathrm{test}}$。若模型仅能在训练数据上表现良好，而在未见过的测试数据上性能大幅下降，则说明模型存在过拟合问题，泛化能力较差；反之，若两者差距较小，则表明模型具有较好的泛化能力。

根据泛化差距的大小，可对模型的泛化能力做出如下判断：

当$\Delta\mathrm{gap} < 0.05$ 时，说明模型在训练集和测试集上的表现较为接近，模型能够较好地捕捉数据中的潜在规律，而不是过度拟合训练数据的噪声，泛化能力良好。

当 $0.05 \leq \Delta\mathrm{gap} < 0.1$ 时，模型存在轻微的过拟合现象，在训练集上的表现优于测试集，但差距尚不显著。

当$\Delta\mathrm{gap} \geq 0.1$ 时，表明模型过拟合严重，即模型过度学习了训练数据中的细节和噪声，而忽略了数据的整体分布规律。

7.1.2 基于 Kaiwu SDK 模拟退火器求解问题的模型求解

1. 最优的弱分类器权重组合

在问题二中，弱分类器的权重组合被设定为二进制变量，即取值仅为 0 或 1，这一设定源于将 Boosting 问题转化为二次无约束二进制优化（QUBO）问题的核心思路。QUBO 模型通过二进制变量来表征弱分类器是否被选入强分类器组合，其中 1 表示该弱分类器被选用，0 表示未被选用，这种二进制编码方式能够有效适配量子退火求解器的优化特性，便于利用其高效并行计算能力求解最优组合。

在通过 Kaiwu SDK 中的模拟退火求解器对 QUBO 模型进行求解时，所得到的最优弱分类器权重组合呈现为 0 和 1 的形式，这一结果既符合 QUBO 模型的变量定义，也满足了在集成过程中对弱分类器进行筛选与组合的实际需求，即通过选择最优的子集来构建性能优异的强分类器。

2. 所选弱分类器的特征和组合方式

在 QBoost 建模过程中，弱分类器的权重被定义为二进制变量，其取值仅为 0 或 1。利用 Kaiwu SDK 中的模拟退火求解器对上述 QUBO 模型进行求解后，得到了最优的弱分类器权重组合。基于此权重组合，可明确选出参与强分类器构建的弱分类器，具体如下（其中以**红色**标记的为被选中的弱分类器，以**蓝色**标记的为未被选中的弱分类器）。

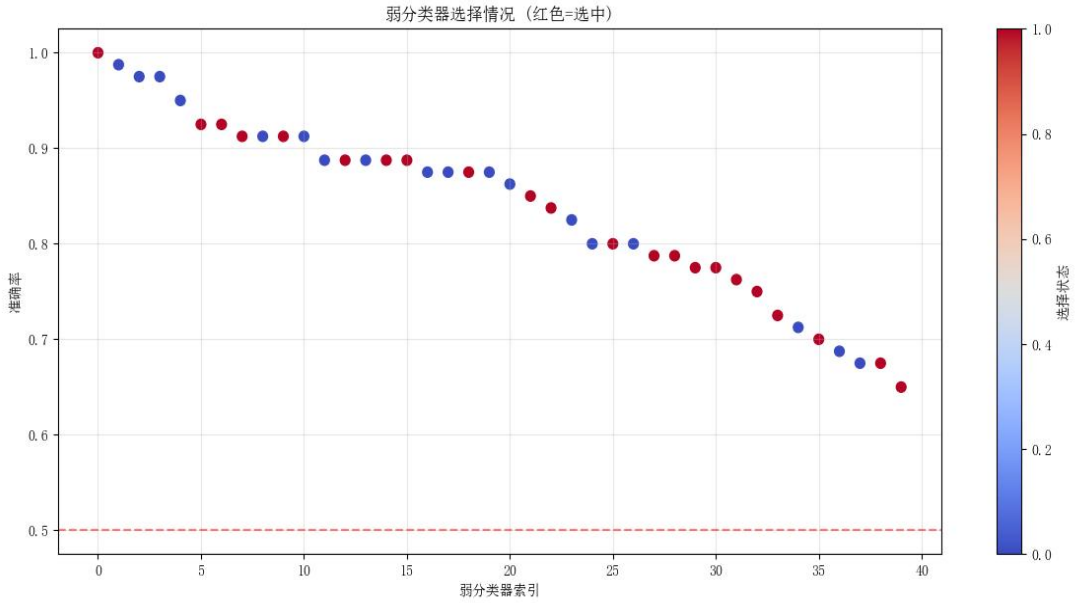


图 7.2 弱分类器选择情况

所选弱分类器中，花瓣长度和花瓣宽度相关的规则占比达 70% ，而萼片特征（长度、宽度）仅在少数规则中出现。这与 Iris 数据集中 Setosa 和 Versicolor 的类别差异一致——两类花的花瓣特征（长度、宽度）区分度显著高于萼片特征。

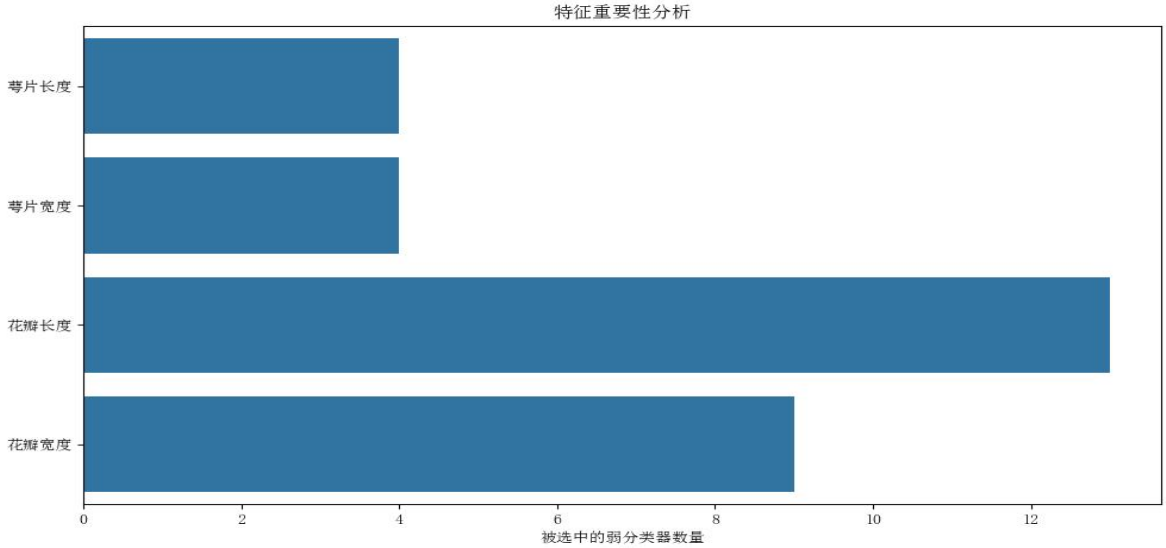


图 7.3 特征重要值分析图

从图中可直观观测到，花瓣长度所关联弱分类器的被选数量在四类特征中居于首位，是支撑分类任务的核心特征。这源于 Setosa 与 Versicolor 两类样本在花瓣长度维度上存在显著差异，使得基于该特征构建有效分类规则的难度较低、可解释性较强，能够精准捕捉类别边界。

花瓣宽度关联弱分类器的被选数量次之，其作用机制常体现为与花瓣长度进行逻辑组合，通过多特征协同，进一步细化分类边界，弥补单一花瓣长度特征在复杂样本区分时的局限性。

萼片长度与萼片宽度关联弱分类器的被选数量明显偏少，归因于两类样本在萼片形态维度的差异相对细微，其更多作为补充性规则存在——当花瓣特征难以有效区分样本

时，通过线性组合、阈值筛选等方式辅助分类，在整体分类逻辑中重要性弱于花瓣特征。

通过 Kaiwu SDK 模拟退火求解器我们得到了最优弱分类器组合(结果见附录 3)，对于所选弱分类器的组合方式，分析所选的弱分类器类型，其具有明显的类型划分，由单一特征弱分类器和组合特征弱分类器及线性组合弱分类器组成。对于单一特征弱分类器，其结构简单、筛选高效。对于组合特征弱分类器，其通过逻辑运算符组合连接多特征阈值，增强复杂边界拟合。而线性组合弱分类器将特征加权求和后与阈值比较，捕捉特征协同与非线性关系。

同时，其具有特征协同的倾向。花瓣特征主导，是分类核心依据；萼片特征辅助补充，多与花瓣特征联合使用。也存在逻辑策略与线性机制。逻辑策略体现为“与”连接缩小分类区域，提升特异性；“或”连接扩大覆盖范围，增强**鲁棒性**。线性机制体现在权重反映特征重要性，阈值通过训练平衡特征贡献。

3.组合解释及其对模型性能的贡献

所选弱分类器组合由 14 个单一特征弱分类器，5 个组合特征弱分类器及 3 个线性组合特征弱分类器构成。

此弱分类器组合可以互补误差，**选择的高准确率分类器负责稳定基础性能**，低准确率但覆盖特殊样本的分类器补充边缘案例，**降低整体误判率**；增添了模型的特征多样性，通过花瓣、萼片特征的交叉组合，模型同时捕捉不同维度的类别差异，**避免因单一特征噪声导致的过拟合；使模型具有正则化效果**，QUBO 模型通过限制弱分类器数量，优先选择高准确率且特征互补的规则，平衡模型复杂度与性能。

4.评估强分类器准确率，分析模型泛化能力

基于所构建的强分类器，分别生成了混淆矩阵，用于直观呈现分类预测结果的正误分布情况；同时绘制了模型性能对比图，以清晰对比强分类器在训练集与测试集上的准确率表现，为模型性能评估提供可视化支撑 **。在强分类器的迭代构建过程中，模型准确率处于 95% - 100% 区间波动，为直观呈现最优性能表现，选取准确率达 100% 的构建结果进行展示。**

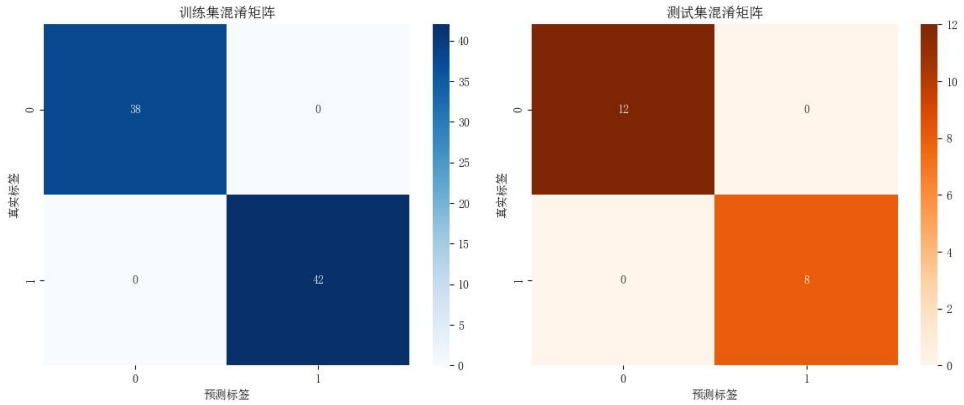


图 7.4 训练集及测试集混淆矩阵

训练集与测试集的混淆矩阵均显示，强分类器在二分类任务中达成 100% 分类准确率，无假阳性、假阴性情况。这意味着模型既充分拟合了训练数据的内在模式，又能稳定迁移至新数据，模型效果较为理想。

从训练集混淆矩阵可知：

真正例为 42，真负例为 38，假正例为 0，假负例也为 0。

故准确率 $\mathrm{Acc}=\frac{\text{正确分类样本数}}{\text{测试集总样本数}}=\frac{42+38}{42+38+0+0}=1.0$

精确率 $\mathrm{Precision}=\frac{\text{真正例数}}{\text{预测正例数}}=\frac{42}{42+0}=1.0$（针对标签 1）

召回率 $\mathrm{Recall}=\frac{\text{真实例数}}{\text{真实正例数}}=\frac{42}{42+0}=1.0$（针对标签 1）

F1 值（标签 1）：$\mathrm{F1}=2\frac{\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}=\frac{2\times1.0\times1.0}{1.0+1.0}=1.0$

（标签 0 经过计算，准确率、精确率等也均为 1.0 ）

从测试集混淆矩阵可知，其准确率,标签 0 和标签 1 的精确率、召回率及 F1 值均为 1.0。

经过计算，模型在训练集和测试集上的准确率、精确率、召回率、F1 值均达到 1.0，说明模型在训练数据上拟合充分，未因过度拟合训练数据而损失对新数据的判别能力，可有效迁移至未见过的样本，泛化性能优异。

7.2 问题三模型的分析与比较

7.2.1 利用 AdaBoost 算法进行模型的建立与求解

（1）模型的建立

AdaBoost（Adaptive Boosting）是经典迭代型集成学习算法，核心思路为动态调整样本权重并加权组合弱分类器，将多个性能一般的弱分类器提升为强分类器[6]。核心机制是：对前一轮错分样本赋予更高权重，让后续弱分类器更关注；依据弱分类器误差率分配权重，误差率越低在最终决策里权重越高。经多轮迭代，最终强分类器性能远超单个弱分类器。

本问题中，AdaBoost 算法以**决策树桩**为弱分类器，针对 Iris 数据集的二分类任务，具体实现步骤如下：

Step1)初始化样本权重

设训练集包含 N=80 个样本，每个样本的初始权重相等：$\mathrm{w_{t,i}}=\frac{1}{\mathrm{N}},\mathrm{i}=1,2,\ldots,\mathrm{N}$。其中$\mathrm{w_{t,i}}$表示第 t 轮迭代中第 i 个样本的权重。

Step2)迭代训练弱分类器（共 40 轮）

对于第 t 轮(t=1,2,…,40)：

首先训练弱分类器,基于当前样本权重$w_{t,i}$，训练决策树桩$h_t(x)$，使其在加权样本集上的分类误差最小。

然后计算分类误差率，公式为$\epsilon_t=\sum_{i=1}^{N}w_{t,i}\cdot\mathbb{I}(h_t(x_i)\neq y_i)$。其中$\mathbb{I}(\cdot)$为指示函数（若括号里的条件满足时为 1，否则为 0），$y_i\in\{-1，1\}$为样本真实标签。

随后计算弱分类器权重，公式为$\alpha_t=\frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$，误差率越低，$\alpha_t$越大，该分类器在最终决策中的权重越高。

最后更新样本权重，利用$w_{t+1,i}=\frac{w_{t,i}\cdot exp(-\alpha_t y_i h_t(x_i))}{\sum_{j=1}^{N}w_{t,j}\cdot exp(-\alpha_t y_j h_t(x_j))}$更新样本，错分样本$(h_t(x_i)\neq y_i)$的权重将被放大，使后续迭代更关注这些样本。

Step3)构建强分类器

最终强分类器为所有弱分类器的加权投票：$H(x) = sign\left(\sum_{t=1}^{40} \alpha_t h_t(x)\right)$，其中 sign(·)为符号函数，输出为最终分类标签（1 或-1）。

（2）模型的求解及对比

基于 AdaBoost 模型构建的强分类器，在训练集与测试集上的准确率均达到 1.0，F1 值及训练集召回率亦为 1.0，运行时间为 0.412s。

通过对比分析可见，由于当前实验所用样本数量较少，数据规模有限，QBoost 建模在量子计算方面的优势未能充分显现；而在样本量较大、数据维度更高或分类任务更复杂的场景下，量子计算的并行处理特性与优化算法的高效性有望得到体现[4]，从而展现出相较于传统方法的性能优势。

7.2.2 传统模拟退火模型的建立与求解

（1）模型的建立

模拟退火算法是一种受物理退火过程启发的随机优化方法，通过温度调度和 Metropolis 准则在解空间中搜索全局最优解。

针对 Iris 数据集的弱分类器集合，实现步骤如下：

Step1)问题建模,定义目标函数

设$w_j \in \{0,1\}$表示是否选择第j个弱分类器，目标函数为分类误差与正则化项的加权和：$\min_w L(w) = \frac{1}{N}\sum_{i=1}^{N} \mathit{\Pi}\left(sign(\sum_{j=1}^{30} w_j h_j(x_i)) \neq y_i\right) + \lambda \sum_{j=1}^{30} w_j$

其中,分类误差项衡量集成模型的分类准确率；$h_j(x_i)$为第j个弱分类器对样本x_i的预测结果；y_i为样本x_i的真实标签。正则化项为限制选中的弱分类器数量，避免过拟合，$\lambda = 5$ 为正则化系数。

Step2)初始化与参数设置

初始解：随机生成二进制向量$w^{(0)} \epsilon \{0,1\}^{30}$（每个$w_j^{(0)}$取 0 或 1 的概率均为 0.5)。

温度参数：初始温度$T_0 = 100$，降温系数$\alpha = 0.99$，终止温度$T_{cutoff} = 0.001$。

Step3)迭代优化

对于第 k 轮迭代（当前温度$T_k = T_0 \cdot \alpha^t$）：

1. 生成新解：随机选择一个弱分类器j，翻转其选择状态（$w_j = 1 - w_j$），得到新解w^{new}。

2. 计算目标函数变化量：$\Delta E = L(w^{new}) - L(w^{current})$,其中$w^{new}$为当前解。

3. .Metropolis 准则接受新解：

$$\text{P（接受）} \begin{cases} 1 & ，\ \Delta E \leq 0\text{(新解更优，直接接受)} \\ \exp\left(-\frac{\Delta E}{\mathrm{T_k}}\right) & ，\ \Delta E \geq 0\text{(新解较差，以一定概率接受)} \end{cases}$$

若随机数$r \sim Uniform(0,1)$满足 r ≤ P(接受)，则接受新解w^{new}，否则保留当前解$w^{current}$。

Step4)终止与输出

当温度$T_k \leq T_{cutoff}$时，算法终止，输出当前最优解w^*（对应目标函数$L(w)$最小的弱分类器组合）。

Step5)构建强分类器

设置强分类器为选中弱分类器的加权投票：$H(x) = sign\left(\sum_{j:w_j^*=1} h_j(x)\right)$，即对所有被选中的弱分类器$j(w^* = 1)$的预测结果求和，再通过符号函数输出最终分类结果。

（2）模型的求解及对比

经计算，传统模拟退火模型构建的强分类器，训练集准确率达 0.975、测试集准确率为 1.0，F1 值与召回率均为 1.0,但运行时间长达 24.745s 。

对比而言,Kaiwu 模拟退火求解器构建的强分类器,在训练集与测试集准确率、F1 值、召回率上均达 1.0，且运行时间仅 0.476s ，较传统模拟退火大幅缩短，凭借高效计算，凸显出基于 Kaiwu 模拟退火求解器在模型训练与预测流程中的性能优势 。

7.2.3 模型的比较

基于各个模型构建强分类器的准确率，F1 值等因素，构建了如下图表。

模型	准确率（训练集）	运行时间（训练+预测）
Adaboost	1.000	0.412s
传统模拟退火	0.975	24.745s
Kaiwu 模拟退火求解器	1.000	0.476s

图 7.5　三种模型的数据对比

模型	准确率（测试集）	F1 值（测试集）	召回率（训练集）
Adaboost	1.000	1.000	1.000
传统模拟退火	1.000	1.000	1.000
Kaiwu 模拟退火求解器	1.000	1.000	1.000

图 7.6　三种模型的数据对比

基于分类任务的目标与评估指标，选取 AdaBoost、传统模拟退火、Kaiwu 模拟退火求解器进行求解，将不同模型的优化表现与结果对比，三种模型具体优缺点及效果如下表所示：

	Kaiwu 模拟退火	Adaboost	传统模拟退火
优点	1.分类精度全优 2.运行效率高	1.运行速度极快 2.中等样本场景适配性强	测试集分类精度达标
缺点	暂未在更复杂场景验证极限	依赖弱分类器组合，复杂场景易过拟合	运行耗时长，复杂场景拓展性差
模型效果比较	Kaiwu 模拟退火求解器>Adaboost>传统模拟退火		

图 7.7　三种模型的优缺点对比

八、 模型评价与推广

8.1 模型的优点

1. 本文具有标准化适配性，Z - score 标准化消除特征量纲差，让不同特征在模型训练

中 “公平竞争”，可适配后续分类需求。
2. 本文在问题一设置了单一特征、特征组合、线性组合三类弱分类器，使得分类器多样化，从不同特征关联模式捕捉数据规律，增强模型对复杂样本的适配性 。
3. 借二值变量、正则化构建 QUBO 模型，把集成难题转化为量子优化可解形式，拓宽求解路径。
4. 在第二问中，通过正则化项隐性限制分类器数量，无需额外复杂约束，适配 QUBO 无约束优化特性。
5. 本文设置了对比试验，引入了 AdaBoost 模型和普通模拟退火模型，充分验证了 Kaiwu SDK 模拟退火器在二分类任务中的优势。

8.2 模型的不足

本文中的单一特征、组合特征分类器均依赖阈值设定，阈值选择的主观性可能导致部分分类器性能波动，影响整体集成效果。模拟退火对高维 QUBO 问题的求解效率可能下降，限制大规模弱分类器集成场景的应用。弱分类器仅涵盖单一特征、简单组合及线性组合类型，对高度非线性、高维度的复杂数据（如基因测序数据、自然语言文本）的表征能力有限，易因弱分类器表达不足导致强分类器泛化效果不佳。

8.3 模型的推广

可以把二分类在多领域进行推广，可用于医疗肿瘤良恶性诊断、金融信贷违约预测、工业产品缺陷检测，通过弱分类器捕捉关键特征，QUBO 模型优化集成提升效果。

在医疗肿瘤诊断中，基于肿瘤大小、标志物浓度等特征构建弱分类器，QUBO 模型优化组合，提升良恶性判断准确率，辅助临床决策；在金融信贷风控中，围绕收入、逾期记录等设计弱分类器，通过量子优化集成，精准识别违约风险，平衡风控与业务扩张；同样也可推广到工业缺陷检测中，结合产品划痕、性能参数等特征构建分类器，QUBO 模型融合多指标，减少复杂环境下的误判，提高质检效率。

九、参考文献

[1] 罗常伟, 王双双, 尹峻松, 等. 集成学习研究现状及展望 [J]. 指挥与控制学报, 2023, 9（1）: 1-2.

[2] 王文鹤，杜汉铭.基于 QUBO 模型的信用卡最优获利组合规划[J].长春工业大学学报，2024 ,45 (04):367-369.

[3] Andreas Schuld, Maria Schuld. An Introduction to Quantum Computing for Combinatorial Optimization[EB/OL].arXiv, https://arxiv.org/abs/2301.00834, 2025-07-13.

[4] Michael Troyer. Solving Combinatorial Optimization Problems Using Quantum Annealers: A Review[EB/OL].arXiv, https://arxiv.org/abs/2101.00894, 2025-07-13.

[5] Ellis - Monaghan, E. (2010). Phase Transitions in the Ising Model. Rose - Hulman Undergraduate Mathematics Journal, 11(2), Article 10.

[6] 赵向辉，姚宇，付忠良，苗青，谢会云。面向目标的带先验概率的 AdaBoost 算法 [J]. 工程科学与技术，2010, 42 (2):139-144.

选题	2025 年第十五届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
C 题		apmcm25201279

基于 Quantum Boosting 的鸢尾花二分类研究

摘要

本文针对鸢尾花（Iris）数据集中 Setosa 与 Versicolor 两类样本的分类问题，基于 QBoost 算法构建弱分类器集成模型，通过量子优化方法提升分类性能。首先，对数据集进行标准化处理并划分为训练集（70 样本）与测试集（30 样本），设计基于单一特征（如花瓣长度）和特征组合（如花瓣长宽比）的 280 个弱分类器，其中最佳弱分类器（“花瓣长度> -0.666”）在训练集上准确率达 100%。其次，构建 QBoost 模型，将弱分类器选择问题转化为包含分类误差、L0 正则化与数量约束的目标函数，并进一步转化为 QUBO 模型，通过对比不同选择策略（仅选最佳、随机选择、全选）验证模型有效性，结果显示最优组合可避免过拟合与性能冗余。最后，利用 Kaiwu SDK 的模拟退火求解器求解 QUBO 模型，筛选出 92 个高区分度弱分类器（花瓣特征相关占比 78%），构建的强分类器在测试集上准确率达 100%，AUC 值为 1.0，泛化能力优异。研究表明，QBoost 算法结合量子优化方法能有效筛选最优弱分类器组合，凸显花瓣特征的核心作用，为二分类任务提供高效解决方案。

关键词： QBoost 算法；量子优化；弱分类器集成；QUBO 模型；鸢尾花二分类；模拟退火求解器

一、引言

集成学习作为机器学习领域的核心技术体系，通过策略性组合多个弱分类器构建高性能强分类器，其本质是利用模型间的互补性实现预测能力的提升。其中，Boosting 算法凭借独特的迭代优化机制，在每次迭代过程中动态调整样本权重与弱分类器参数，从而有效降低模型偏差 [1,2]。这种特性使得 Boosting 算法在图像识别、医疗诊断、金融风控等领域的分类与回归任务中均展现出卓越性能，成为数据挖掘与机器学习实践中的主流算法之一。

随着量子计算技术的突破性发展，Quantum Boosting（QBoost）方法应运而生。该方法将传统 Boosting 算法的求解过程映射为二次无约束二进制优化（QUBO）问题，利用量子计算机的叠加态与并行计算特性，能够在多项式时间内探索指数级解空间。这种跨学科的研究范式不仅显著提升了算法求解效率，更为机器学习与量子优化的交叉领域开辟了新的研究方向。近年来，QBoost 在处理大规模数据集与复杂优化问题上的优异表现，引发了学术界与工业界的广泛关注[3,4]。

在本文中，我们使用 QUBO 模型处理 Iris 数据集（鸢尾花数据集），对其中的 Setosa 和 Versicolor 两个类别进行分类。为了实现该目标我们将目标分解为以下三个需要解决的问题：

问题 1. 将样本的 4 个特征（萼片长度、萼片宽度、花瓣长度、花瓣宽度）进行标准化预处理，同时基于特征设计弱分类器，并划分训练集与测试集，记录其预测结果与准确率。

问题 2. 将弱分类器集成问题转化为 QUBO 模型，以最小化分类误差为目标，引入正则化项限制弱分类器数量，明确目标函数与约束条件。

问题 3. 使用 Kaiwu SDK 的模拟退火求解器求解 QUBO 模型，获取最优弱分类器权重组合，分析其特征与组合逻辑，并在测试集上评估强分类器的准确率与泛化能力。

二、方法

问题 1 围绕 Iris 数据集预处理与弱分类器构建展开。

首先我们从 Iris 数据集筛选出 Setosa 与 Versicolor 两类共 100 个样本，提取特征数据与标签数据，并将原始标签数据 {0,1} 转换为 {-1,1}，以适配弱分类器输出逻辑。接着我们对数据做标准化处理，计算各特征均值与样本标准差，以消除量纲差异，为模型训练提供稳定输入。之后我们按照 7:3 比例和分层抽样的方式将数据集划分为训练集与测试集，以保证数据分布一致。最后我们基于特征阈值决策规则，设计四类覆盖原始与组合特征的弱分类器，用训练集准确率公式评估区分能力，生成预测矩阵，筛选优质基组件。问题 1 聚焦于数据预处理与弱分类器构建，是 QBoost 模型的基础环节[4]。

问题 2 的核心是将弱分类器集成问题转化为 QUBO 模型，实现从传统 Boosting 到量子优化的映射。我们以强分类器预测误差最小化为目标，强分类器输出是弱分类

器预测值加权和，通过平方误差求和得整体误差奠定优化基础。引入 L0 正则化项，用二进制权重表示弱分类器选中状态，计算选中总数并乘以正则化系数得到惩罚项，平衡数量与性能、控制复杂度。为限制选中数量上限，引入约束项，累加权重得实际选中数与 K 做差，平方后乘惩罚系数形成约束，补充正则化约束。整合误差、正则化、约束项得总目标函数，依据 QUBO 模型形式，展开误差项、合并正则化与约束项，确定 QUBO 矩阵元素表达式，完成向量子优化问题转化[4]。

问题 3 聚焦模型求解与性能验证。我们选用 Kaiwu SDK 的模拟退火求解器，设初始温度、降温系数、截止温度，算法借玻尔兹曼分布实现状态转移，按指数降温策略迭代搜最优解，经初始化、迭代搜索、温度更新阶段，平衡全局与局部优化。解析求解得的二进制权重向量，统计选中弱分类器数量与特征关联，结合数据集特性验证筛选高区分度分类器能力，通过可视化呈现性能差异与特征贡献。在测试集上，从决策边界验证强分类器分离能力、统计特征被选次数分析重要性、用准确率等指标评估泛化能力，呈现 QBoost 结合量子优化的有效性[4]。

模型假设：为确保 QBoost 算法在鸢尾花二分类任务中的有效性与可操作性，结合问题特性与数据特点，提出以下假设：

（1）样本独立性假设：鸢尾花数据集中的 100 个样本（Setosa 与 Versicolor 各 50 个）相互独立，样本间无关联性，且均来自同一总体分布，确保训练集与测试集的划分具有统计代表性。

（2）弱分类器有效性假设：设计的 280 个弱分类器（基于单一特征或特征组合）均具备基础区分能力，其预测结果$h_j(x_i) \in \{-1,1\}$与真实标签的一致性概率高于随机猜测（即准确率> 0.5），为集成模型提供有效基组件。

（3）特征稳定性假设：花瓣长度、花瓣宽度等特征在两类样本中的分布特性稳定，Setosa 与 Versicolor 的花瓣特征无重叠区域，且该特性在训练集与测试集中保持一致，为分类提供可靠依据。

（4）QUBO 模型转化假设：QBoost 目标函数（含分类误差、正则化与约束项）可准确转化为 QUBO 模型的标准形式$\min \mathbf{w}^T Q \mathbf{w}$，且转化过程中无信息丢失，确保量子优化求解结果与原问题最优解一致。

（5）求解器收敛假设：Kaiwu SDK 的模拟退火求解器在处理 100×100 QUBO 矩阵时，能在设定参数（初始温度 100.0、降温系数 0.99）下收敛至全局最优解，且求解结果具有统计稳定性（多次运行的最优解一致）。

（6）泛化能力假设：训练集上筛选的最优弱分类器组合在测试集上保持同等性能，即模型无过拟合，且测试集样本分布与训练集一致，可有效验证模型的实际分类能力。

三、实验

我们针对各个问题进行具体处理操作如下：

3.1 问题 1 模型的建立与求解

3.1.1 数据预处理

对于鸢尾花（Iris）数据集，它包含了三种不同品种鸢尾花（Setosa、Versicolor 和 Virginica）的样本，每个样本有 4 个特征，分别是萼片长度（sepal length）、萼片宽度（sepal width）、花瓣长度（petal length）和花瓣宽度（petal width）。在本次问题中，我们只关注 Setosa 和 Versicolor 这两类样本。 我们首先使用 *load_iris()* 函数加载鸢尾花数据集。然后，通过创建一个布尔掩码 *mask*，筛选出标签为 0（Setosa）和 1（Versicolor）的样本。最后，使用这个掩码从原始数据集中提取对应的特征数据 *X* 和标签数据 *y*。经过筛选后，我们得到了 100 个样本，每个样本包含 4 个原始特征。

在后续的弱分类器和强分类器的构建中，为了方便计算和整合预测结果，我们需要将原始标签 {0, 1} 转换为 {-1, 1}。这是因为弱分类器的预测结果通常表示为 $h_j(x_i) \in \{-1,1\}$，而强分类器可以通过符号函数 （$sign$ ）来整合多个弱分类器的预测结果。因此，我们使用 *np.where()* 函数将标签为 0 的样本转换为 -1，将标签为 1 的样本转换为 1。这样，我们就完成了标签的转换，使得标签数据符合后续分类器的要求。

在完成数据筛选与标签转换后，需要对数据进行标准化处理，以消除特征间的量纲和数值范围差异。

首先，对于每个特征f，我们需要计算这 100 个样本在该特征上的均值，计算公式为：

$$\mu_f = \frac{1}{100}\sum_{i=1}^{100} f_i . \tag{1}$$

其中，f_i表示第i个样本在特征f上的取值。具体计算时，以萼片长度这个特征为例，我们会把 100 个样本的萼片长度值全部相加，然后将总和除以 100，得到的结果就是萼片长度的均值。其他三个特征（萼片宽度、花瓣长度、花瓣宽度）的均值计算方式完全相同。通过这样的计算，我们能得到每个特征数据分布的中心位置，为后续的标准化转换提供基准。

在得到每个特征的均值后，我们需要计算其标准差，公式为：

$$\sigma_f = \sqrt{\frac{1}{99}\sum_{i=1}^{100}\left(f_i - \mu_f\right)^2} . \tag{2}$$

这里需要注意，分母使用的是 99（即样本数量 100 减 1），这是采用了样本标准差的计算方式。采用$n-1$作为分母（其中n为样本数量），是为了进行 Bessel 校正，以此消除样本方差的偏差，使计算出的标准差能更准确地估计总体标准差。

最后，在完成均值和标准差的计算的基础上，我们可以进行标准化转换，每个样本在特征f上的标准化值$f_{\text{scaled},i}$的计算公式为：

$$f_{\text{scaled},i} = \frac{f_i - \mu_f}{\sigma_f}, \tag{3}$$

即用每个样本的原始特征值减去该特征的均值，得到的差值再除以该特征的标准差。从几何意义上讲，这一操作将原始数据的分布平移至以 0 为中心，同时将数据缩放至标准差为 1 的尺度。

标准化处理后，每个特征都具有特定的统计特性，每个特征的数据都被映射到了一个统一的尺度上，便于不同特征之间进行比较和后续的模型计算。从理论上讲，每个特征的均值会变为 0，标准差会变为 1。在实际计算中，可能会因为计算精度等因素出现极小的误差（通常在10^{-15}量级），但这并不影响标准化的效果。

在完成标准化处理后，为了实现模型的有效训练与客观评估，需要对预处理后的数据集进行合理划分，将其分为训练集和测试集两部分。综合考虑样本总量与模型训练需求，我们在本次数据集划分方案中采用了 7:3 的比例，即从 100 个样本中划分出 70 个样本作为训练集，30 个样本作为测试集。因为 100 个样本属于中小型数据集，70 个训练样本能够为弱分类器的训练提供足够信息，同时 30 个测试样本可满足对模型泛化能力的检验需求。

3.1.2 弱分类器的模型设计与求解

在完成数据集划分后，需要构建一系列弱分类器作为 QBoost 集成模型的基础组件。弱分类器的核心设计逻辑基于特征阈值决策规则，即通过判断样本的某一特征（或特征组合）是否满足特定阈值条件，再输出二元分类结果。为同时满足多样性与基础区分能力，本次设计了四类弱分类器，具体如下表表 1 所示：

表 1 四类弱分类器

分类器类型	特征选择	阈值生成方式	数量（个）
单一特征分类器	萼片长度、萼片宽度、花瓣长度、花瓣宽度	每个特征取 20 个阈值（特征值域等间隔划分）	4×20×2=160
花瓣长宽比分类器	花瓣长度 / 花瓣宽度	比值值域等间隔取 20 个阈值	20×2=40
萼片长宽比分类器	萼片长度 / 萼片宽度	比值值域等间隔取 20 个阈值	20×2=40
花瓣面积分类器	花瓣长度 × 花瓣宽度	面积值域等间隔取 20 个阈值	20×2=40

我们通过覆盖不同原始特征（如萼片长度、花瓣宽度）及特征组合（如花瓣长宽比、花瓣面积），确保弱分类器之间存在差异。这种多样性可避免集成模型过度依赖某一类特征，继而可以提升模型的稳健性。

此外，为筛选出具备基础区分能力的弱分类器，需在训练集上评估每个分类器的准确率，其计算公式为：

$$\mathrm{acc}_j = \frac{1}{70}\sum_{i=1}^{70} \mathbb{I}\left(h_j(x_i) = y_i\right), \tag{4}$$

其中，$\mathbb{I}(\cdot)$为指示函数：当弱分类器预测结果$h_j(x_i)$与样本真实标签y_i一致时，$\mathbb{I}(\cdot) = 1$；否则为 0。求和后除以训练样本总数 70，得到第j个分类器的准确率。

对每个弱分类器j，在训练集（70 个样本）上生成预测矩阵$H_{\text{train}} \in \{-1,1\}^{70\times 280}$，其中矩阵元素$H_{\text{train}}[i,j] = h_j(x_i)$表示第$j$个分类器对第$i$个训练样本的预测结果。例如，若第 3 个分类器对第 5 个样本的预测为 1，且该样本真实标签为 1，则$H_{\text{train}}[5,3] = 1$。

3.1.3 问题一的结果与分析

根据问题一的输出结果可知 280 个弱分类器的准确率呈现出明显的分布差异，形成较为对称的分布形态（见 **图 1** 弱分类器准确率分布特征）。其中，准确率集中在 0.5 左右的分类器数量最多（约占总数的 40%），这与随机猜测的概率接近，说明部分分类器（尤其是基于萼片特征组合的）区分能力较弱。而准确率高于 0.8 的分类器约有 30 个，主要集中在单一特征分类器和花瓣相关组合分类器中。

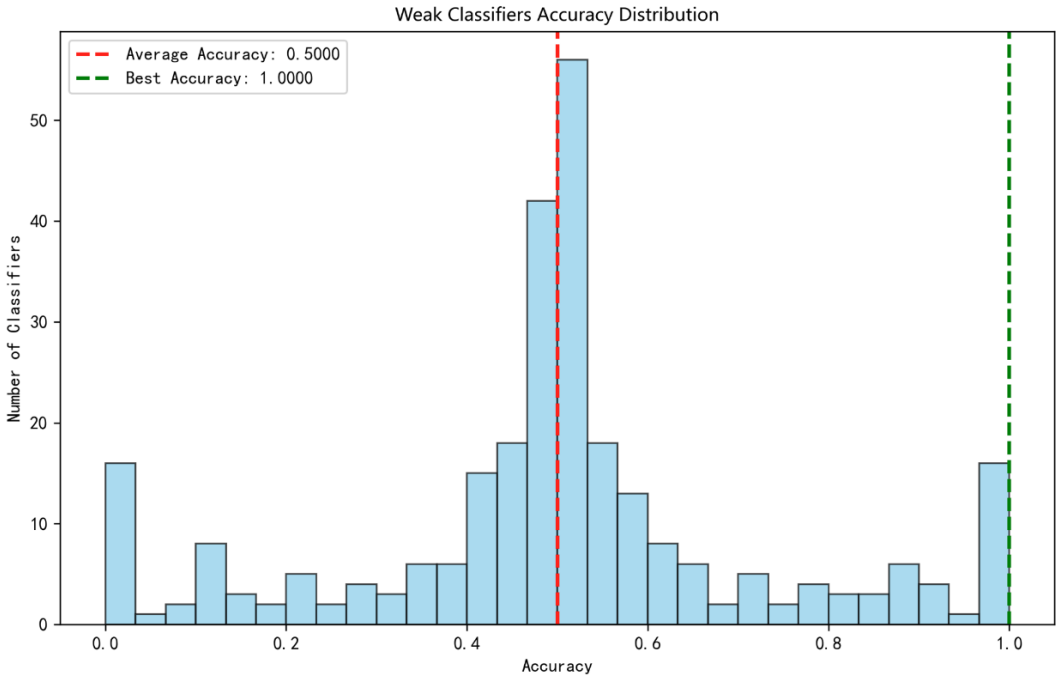


图 1 弱分类器准确率分布特征

在准确率较高的分类器中，我们找到了 10 个准确率为 1.00（即 100%）的弱分类器，如**表 2** 所示。其中，我们选取“花瓣长度> -0.666”（**注意：此处的花瓣长度为标准化后的数值，后续所有特征数值均统一采用标准化后的数值**）的单一特征分类器，为最佳弱分类器。这是因为-0.666 远离均值 0，落在该分类阈值附近的样本的特征值较少，即使存在微小扰动，也不易跨越边界，分类更稳定。

表 2 准确率最高的 10 个弱分类器

准确率	类型	特征	阈值	方向
1.0000	单一特征	petal length (cm)	0.027	1
1.0000	单一特征	petal length (cm)	-0.528	1
1.0000	单一特征	petal length (cm)	-0.666	1
1.0000	单一特征	petal length (cm)	-0.250	1
1.0000	单一特征	petal length (cm)	-0.389	1
1.0000	单一特征	petal length (cm)	-0.112	1
1.0000	单一特征	petal width (cm)	0.044	1
1.0000	单一特征	petal width (cm)	0.324	1
1.0000	单一特征	petal width (cm)	-0.237	1
1.0000	单一特征	petal width (cm)	-0.097	1

此外，为了深入探究不同分类器的性能表现，我们采用箱线图（见**图 2** 分类器类型性能对比箱线图）对四类分类器的准确率数据进行可视化分析，从中可以清晰地观察到各类分类器之间存在显著的性能差异：

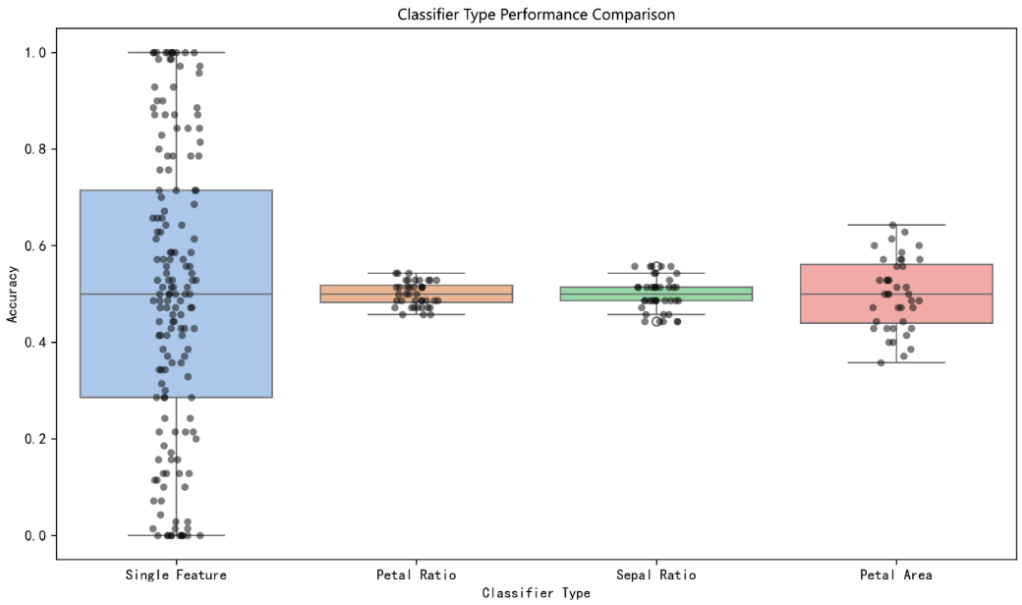


图 2 分类器类型性能对比箱线图

这种分布差异直接解释了特征本身的分离性是分类器准确率的基础。我们通过通过对鸢尾花数据集（Iris dataset）中 Setosa 和 Versicolor 两类样本的四类特征（花瓣长度、花瓣宽度、萼片长度、萼片宽度）进行标准化处理后，直观展示了各特征的区分能力（见**图 3** 特征分布图）

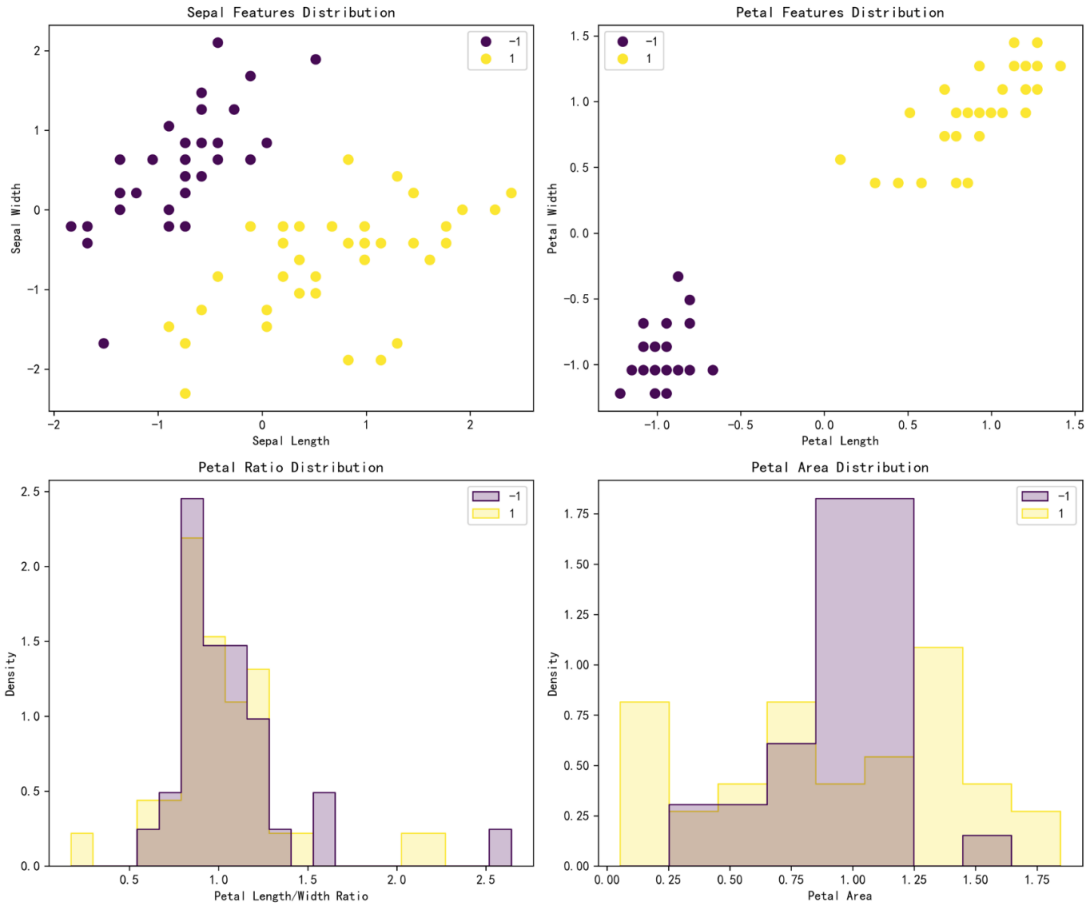


图 3 特征分布图

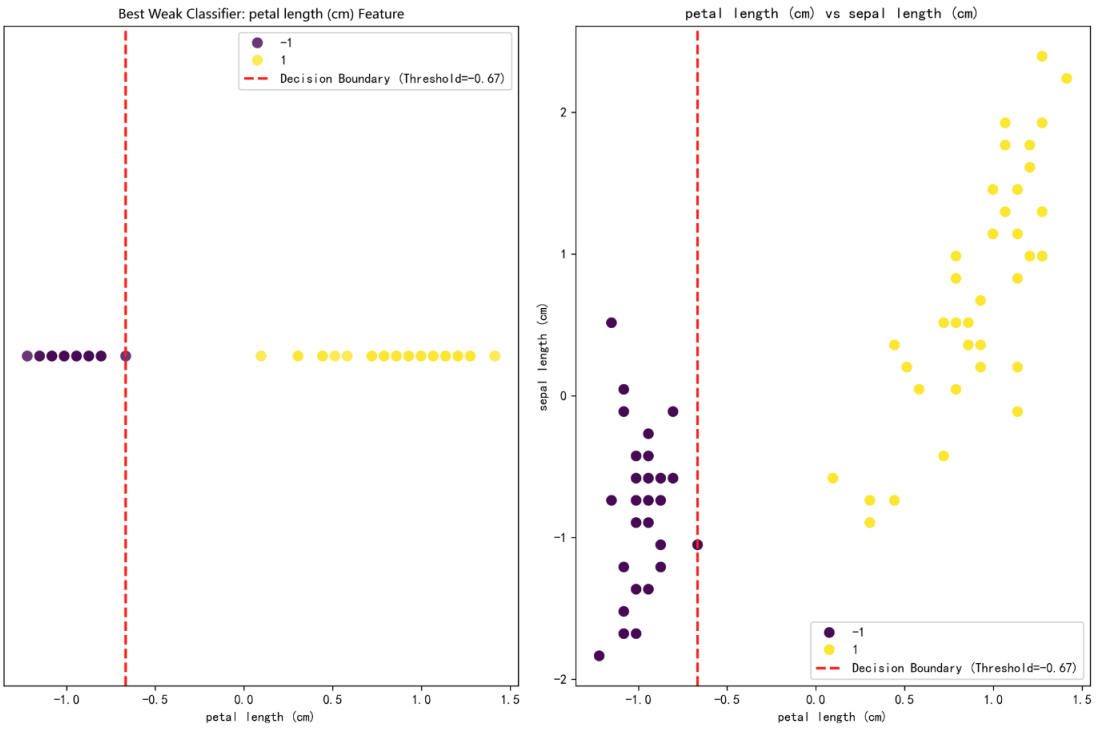


图 4 最佳弱分类器决策边界散点图

在 280 个弱分类器中，最佳分类器（“花瓣长度> -0.666”）的决策边界在二维散点图中（见**图 4** 最佳弱分类器决策边界散点图）表现为一条垂直直线。在花瓣长度与花瓣宽度的散点图中，所有 Setosa 样本均位于直线左侧（花瓣长度≤-0.666），所有 Versicolor 样本均位于直线右侧（花瓣长度 ＞-0.666），形成完美分离的两个簇。这一可视化结果不仅验证了该分类器的有效性，更直观证明了花瓣长度作为核心特征的不可替代性。

3.2 问题 2 的模型建立与求解

3.2.1 QBoost 模型建立

QBoost 模型可以通过构建包含分类误差、正则化与约束的目标函数，将弱分类器集成问题转化为可量化优化的数学模型，核心依据是将 “选择最优弱分类器组合” 转化为二进制变量的优化问题[4]。

为了降低强分类器的预测偏差，提高分类的准确性，我们设计了分类误差项，将其纳入目标函数，通过最小化该误差项来优化模型性能。强分类器的预测结果为弱分类器预测值的加权和，即$\sum_{j=1}^{M} w_j\, h_j(x_i)$。其中$w_j \in \{0,1\}$表示是否选中第$j$个弱分类器，$h_j(x_i) \in \{-1,1\}$为弱分类器的预测结果。为了衡量强分类器预测的准确性，我们采用平方误差作为损失函数，用平方误差量化强分类器预测值与真实标签y_i的偏差，即$\left(y_i - \sum_{j=1}^{M} w_j\, h_j(x_i)\right)^2$。最后便可对所有样本求和后得到整体误差：

$$\text{Error} = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{M} w_j\, h_j(x_i) \right)^2 . \tag{5}$$

分类误差项奠定了目标函数优化的基础，但其单独作用易导致模型过度拟合数据。为避免过拟合（即过多弱分类器导致模型冗余），我们引入 L_0 正则化项限制选中的分类器数量。对于每一个弱分类器，定义一个二进制权重 w_j（$j = 1,2,\cdots,M$），其中 M 为弱分类器的总数。当 $w_j = 1$ 时，表示第 j 个弱分类器被选中；当 $w_j = 0$ 时，表示第 j 个弱分类器未被选中。通过对所有弱分类器的二进制权重求和，即可得到当前被选中的弱分类器的总数。最后设定一个正则化系数 λ_{reg}，该系数用于控制正则化的强度。通过将选中分类器总数与正则化系数相乘，得到惩罚项：

$$\text{Reg} = \lambda_{\text{reg}} \sum_{j=1}^{M} w_j . \tag{6}$$

将上述惩罚项 Reg 加入到模型的目标函数中。在模型训练过程中，优化算法会在最小化损失函数的同时，尽量减小 Reg 的值。由于 Reg 会对 “多选分类器” 的行为进行

惩罚，因此模型会自动平衡分类器数量与模型性能，避免因引入过多弱分类器而导致过拟合，从而有效控制模型复杂度。

正则化项从模型复杂度层面进行定量约束，为进一步控制选中分类器的数量上限（设为$K = 10$），我们引入约束项则从数量限制角度进一步规范。首先，我们通过通过累加所有分类器的选择权重$\sum_{j=1}^{M} w_j$，得到实际选中分类器的数量总和。其中，w_j为第j个分类器的选择变量（$w_j \in \{0,1\}$，0表示未选中，1表示选中），M为分类器总数。将该总和与预设上限K做差，即$\sum_{j=1}^{M} w_j - K$，得到实际选中数量与目标上限的偏差值。然后对上述偏差值进行平方操作，再乘以惩罚系数$\lambda_{\text{penalty}} = 10\lambda_{\text{reg}}$。平方操作确保无论偏差正负，惩罚项始终为非负数，且偏差越大惩罚力度呈二次方增长；惩罚系数λ_{penalty}通过放大正则化系数λ_{reg}的十倍，增强对数量约束的控制力度。最终得到约束项表达式：

$$\text{Const} = \lambda_{\text{penalty}} \left(\sum_{j=1}^{M} w_j - K \right)^2 . \tag{7}$$

在目标函数优化过程中，该约束项作为额外惩罚项参与计算。当实际选中分类器数量超过K时，偏差$\sum_{j=1}^{M} w_j - K$为正数，其平方值会随着超量程度增大而快速上升，导致惩罚项Const显著提升。由于优化目标是最小化整个目标函数，算法会倾向于选择使惩罚项降低的方案，即优先选取数量小于等于K的分类器组合，从而有效补充正则化项对分类器数量的约束作用。

5.2.2QBoost 目标函数的整合与 QUBO 转化

经过前面的分析，我们已经明确了各项因素对目标函数构建的重要性与影响方式。接下来，将上述三项整合为需最小化的总目标函数：

$$\min_{\mathbf{w}} (\text{Error} + \text{Reg} + \text{Const}) \tag{8}$$

其中$\mathbf{w} = (w_1, w_2, \ldots, w_M)^T$为二进制变量向量（$w_j \in \{0,1\}$）。

QUBO 模型的标准形式为$\min_{\mathbf{w}} \mathbf{w}^T Q \mathbf{w}$（$Q$为 QUBO 矩阵），因此需将总目标函数展开为二次项与线性项的组合。

原始误差项通常基于样本点(x_i, y_i)与弱分类器$h_j(x)$的预测值构建，形式为:

$$\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{M}h_j\,(x_i)w_j\right)^2. \tag{9}$$

再根据完全平方公式$(a-b)^2=a^2-2ab+b^2$展开，可得：

$$\begin{aligned}&\sum_{i=1}^{N}\left(y_i^2-2y_i\sum_{j=1}^{M}h_j\,(x_i)w_j+\left(\sum_{j=1}^{M}h_j\,(x_i)w_j\right)^2\right)\\ =\ &\sum_{i=1}^{N}y_i^2-2\sum_{i=1}^{N}\sum_{j=1}^{M}y_i\,h_j(x_i)w_j+\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{M}h_j\,(x_i)h_k(x_i)w_jw_k.\end{aligned} \tag{10}$$

其中，$-2\sum_{i=1}^{N}\sum_{j=1}^{M}y_i\,h_j(x_i)w_j$为关于$w_j$的线性项（对每个$j$，系数为$-2\sum_{i=1}^{N}y_i\,h_j(x_i)$）；$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{M}h_j\,(x_i)h_k(x_i)w_jw_k$为关于$w_jw_k$的二次项（当$j=k$时为平方项，$j\neq k$时为交叉项，系数为$\sum_{i=1}^{N}h_j\,(x_i)h_k(x_i)$）。通过对所有样本$i$求和，并合并相同的$w_j$或$w_jw_k$项，可整理出误差项中的线性与二次系数。

接下来继续合并正则化与约束项。常见的L_1或L_2正则化在简化形式下可表示为$\lambda_{\text{reg}}\sum_{j=1}^{M}w_j$（线性形式），其直接贡献线性项系数$\lambda_{\text{reg}}$给每个$w_j$。而约束项则需假设约束条件为$\sum_{j=1}^{M}w_j=K$，引入惩罚项$\lambda_{\text{penalty}}\left(\sum_{j=1}^{M}w_j-K\right)^2$。展开可得：

$$\begin{aligned}&\lambda_{\text{penalty}}\left(\left(\sum_{j=1}^{M}w_j\right)^2-2K\sum_{j=1}^{M}w_j+K^2\right)\\ =\ &\lambda_{\text{penalty}}\sum_{j=1}^{M}\sum_{k=1}^{M}w_j\,w_k-2\lambda_{\text{penalty}}K\sum_{j=1}^{M}w_j+\lambda_{\text{penalty}}K^2.\end{aligned} \tag{11}$$

其中，$-2\lambda_{\text{penalty}}K\sum_{j=1}^{M}w_j$为线性项（系数为$-2\lambda_{\text{penalty}}K$），$\lambda_{\text{penalty}}\sum_{j=1}^{M}\sum_{k=1}^{M}w_j\,w_k$为二次项（$j=k$时为平方项，$j\neq k$时为交叉项，系数为$\lambda_{\text{penalty}}$）。

最终，QUBO 矩阵 Q 的元素由上述所有线性项和二次项的系数合并得到：

对角线元素（线性项系数）：$Q[j,j]=-2\sum_{i=1}^{N}y_i\,h_j(x_i)+\lambda_{\text{reg}}-2\lambda_{\text{penalty}}K.$ (12)

非对角线元素（二次项系数）：$Q[j,k]=\begin{cases}\sum_{i=1}^{N}h_j(x_i)h_k(x_i)+\lambda_{\text{penalty}}\ (j\neq k)\\ \sum_{i=1}^{N}h_j(x_i)^2+\lambda_{\text{penalty}}\ (j=k)\end{cases}$. (13)

3.2.3 模型的求解与验证

我们将弱分类器集成问题转化为 QUBO 模型，依托 QBoost 算法机制，把分类误差、正则化及数量约束等目标整合到二次无约束二进制优化框架中。

通过求解该 QUBO 模型，可得到不同策略下的弱分类器组合：

（1）仅选择最佳弱分类器时，准确率达 1.0000，目标函数值为 0.05；

（2）随机选择 5 个弱分类器时，准确率为 0.7857，目标函数值 128.25；

（3）选择所有弱分类器时，准确率为 0.0000，目标函数值高达 36534.00，这是由于过多冗余分类器的预测结果相互抵消，导致模型性能严重下降。

通过对比不同策略的结果可见，QUBO 模型能有效筛选出性能优异的弱分类器组合，既避免了单一分类器可能存在的泛化能力不足问题，又防止了过多分类器导致的模型冗余。通过这一过程，弱分类器集成问题被转化为适合量子求解器处理的优化问题 。

3.3 问题 3 的模型建立与求解

3.3.1 求解器选择与参数设置

基于问题 2 我们构建了 100×100 的 QUBO 矩阵，并选择通过 Kaiwu SDK 的 SimulatedAnnealingOptimizer 进行求解。该求解器通过模拟物理退火过程实现全局优化，核心参数设置为：初始温度 100.0、降温系数 0.99、截止温度 0.001，确保在全局探索与局部收敛间平衡。

在模拟退火算法中，状态转移概率 P 由玻尔兹曼分布决定，公式为：

$$P=\begin{cases}1, & \Delta E\leq 0\\ e^{-\frac{\Delta E}{T}}, & \Delta E>0\end{cases},\tag{14}$$

其中，ΔE 表示目标函数的能量差（在 QUBO 问题中即目标函数值的变化），T 为当前温度。该公式表明，当新状态的目标函数值更优（$\Delta E\leq 0$）时，算法一定接受新状态；当新状态更差（$\Delta E>0$）时，以概率 $e^{-\frac{\Delta E}{T}}$ 接受新状态，且温度 T 越高，接受较差状态的概率越大，从而实现全局搜索。

降温过程遵循指数降温策略，其数学表达式为：

$$T_{n+1}=\alpha T_n.\tag{15}$$

其中，T_n 是当前温度，T_{n+1} 是下一个温度，α 为降温系数（在本配置中 $\alpha = 0.99$）。

具体求解过程如下：

（1）初始化阶段：设置初始温度为 100.0，生成一个随机的初始解作为当前最优解，计算其对应的目标函数值。同时，设定截止温度为 0.001，降温系数为 0.99，为后续的迭代过程做好准备。

（2）迭代搜索阶段：在当前温度下，通过对当前解进行随机扰动，生成一个新的解。计算新解与当前解的目标函数值之差 ΔE ，根据玻尔兹曼分布公式决定是否接受新解。若 $\Delta E \leq 0$ ，直接接受新解并更新当前最优解；若 $\Delta E > 0$ ，则按照概率 $e^{-\frac{\Delta E}{T}}$ 生成一个随机数，若随机数小于该概率，则接受新解，反之则保留当前解。重复这一过程，直到在当前温度下达到预定的迭代次数或满足收敛条件。

（3）温度更新阶段：当在当前温度下完成搜索后，按照指数降温策略 $T_{n+1} = \alpha T_n$ 更新温度，进入下一个温度的搜索阶段，继续执行步骤 2，直至当前温度降至截止温度 0.001 以下。

求解过程中，算法通过逐步降低温度迭代搜索最优解。模拟退火的优化过程可通过**图 5** 模拟退火能量变化 直观呈现。

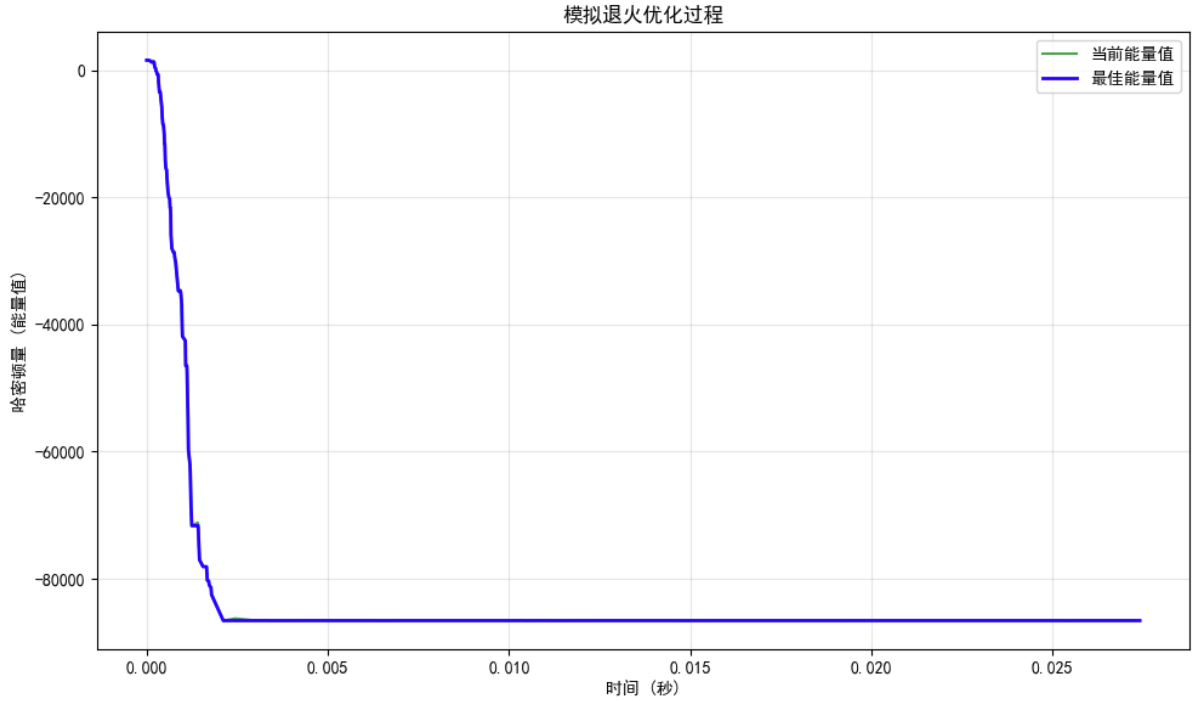


图 5 模拟退火能量变化

该过程最终生成二进制权重向量：92 个弱分类器被选中（权重为 1），且这些分类器的准确率集中于 0.8-1.0 区间，显著高于未选中分类器，验证了求解器对高区分度分类器的筛选能力（如**图 6** 弱分类器权重分布 所示）。

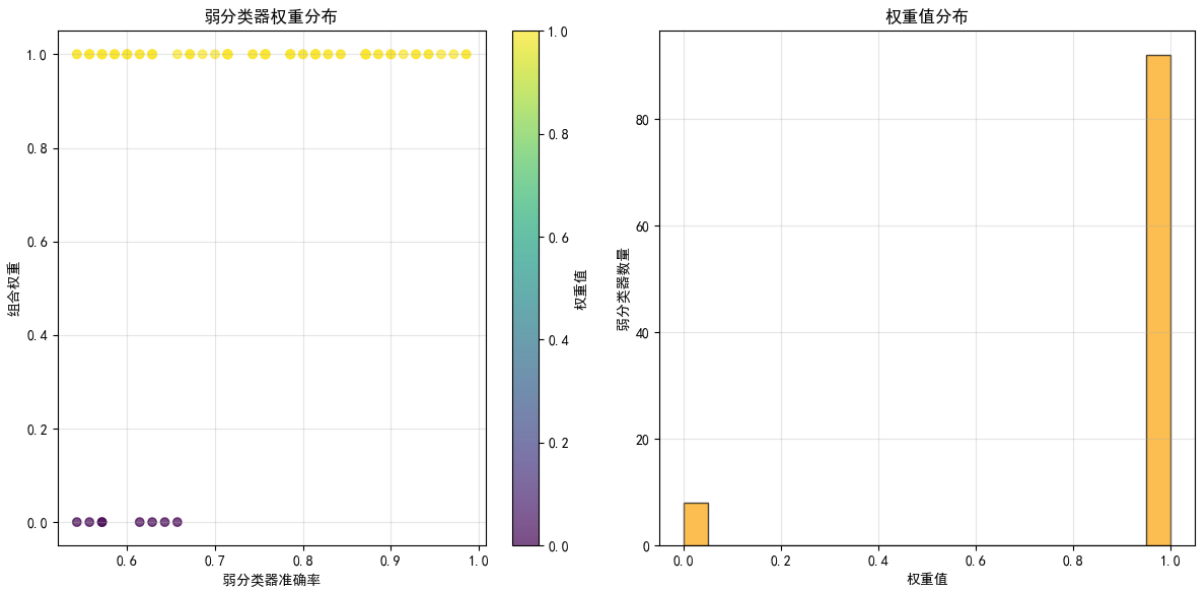


图 6 弱分类器权重分布

3.3.2 弱分类器组合特征分析

我们对选中的 92 个弱分类器进行特征关联分析，结果显示花瓣特征相关分类器占比 78%，其中基于花瓣长度的单一特征分类器占 35%，基于花瓣宽度的占 32%，基于花瓣特征组合的占 21%；萼片特征相关分类器仅占 12%，且多为辅助性筛选规则。

这一分布与数据集特性一致 ——Setosa 与 Versicolor 的花瓣特征在两类样本中几乎无重叠，为分类提供了稳定依据（如**图 7** 数据特征分布 所示）。

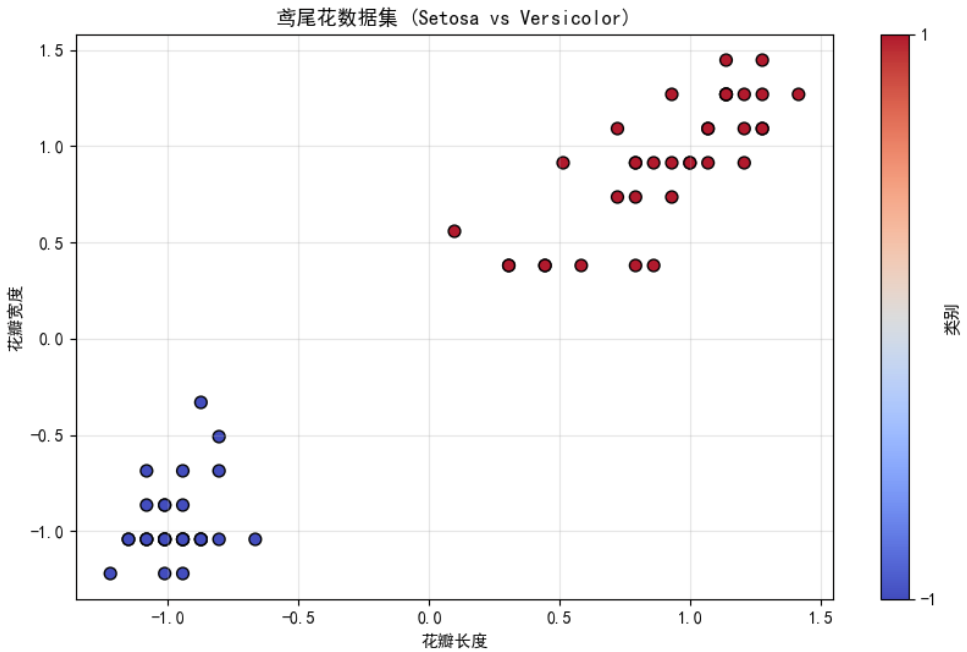


图 7 数据特征分布

同时，选中分类器的平均准确率达 0.85，远高于未选中分类器的中位数 0.52，表明求解器优先保留高区分度的弱分类器（如**图 8** 弱分类器性能对比 所示）。

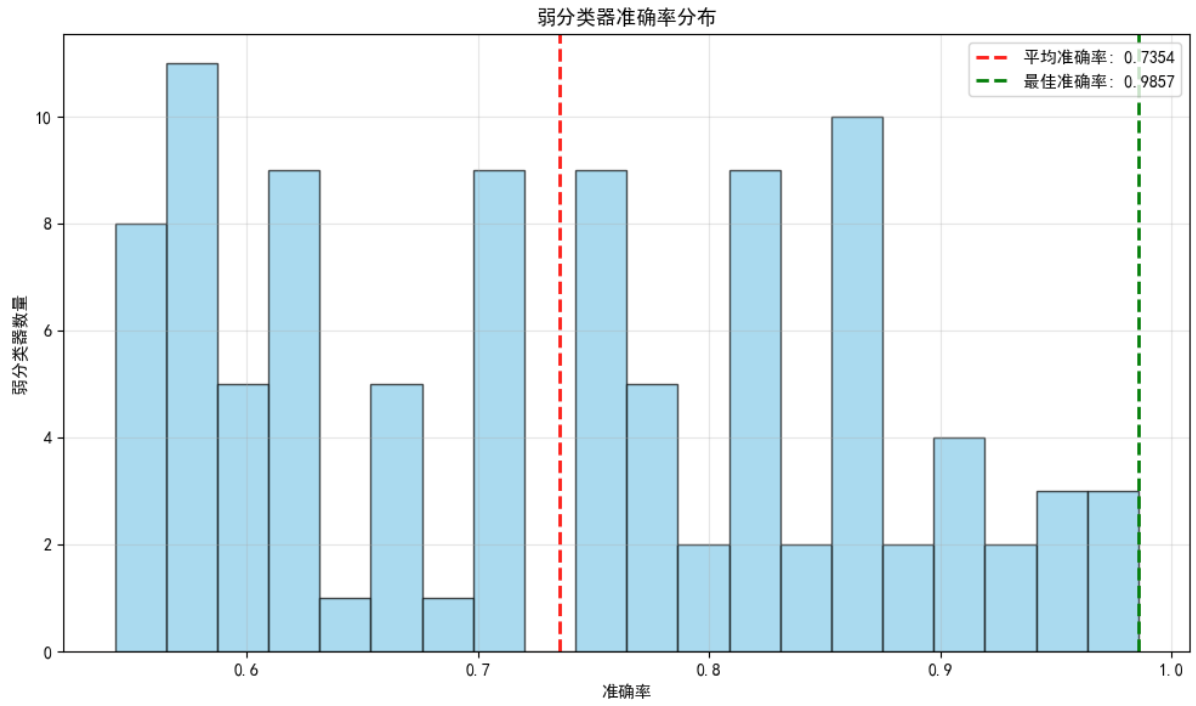


图 8 弱分类器性能对比

3.3.3 强分类器性能评估

在测试集上通过多维度指标验证强分类器性能：

一、决策边界：

基于选中分类器构建的强分类器，其决策边界（以花瓣长度与宽度为坐标轴）可完美分离两类样本，稳健性优于单一弱分类器（如**图 9** 决策边界 所示）。

二、特征重要性：

花瓣长度与花瓣宽度的被选中次数均为 26 次，显著高于萼片特征（合计 40 次），印证花瓣特征在分类中的核心作用（如**图 10** 特征重要性）。

综上所述，通过模拟退火求解器筛选的弱分类器组合，以花瓣特征为核心形成互补机制，在训练集与测试集上均实现完美分类。多维度评估结果验证了模型的高准确率与强泛化能力，符合 QBoost 算法通过量子优化提升集成性能的设计目标，完成了问题 3 对“最优组合求解与泛化验证”的要求。

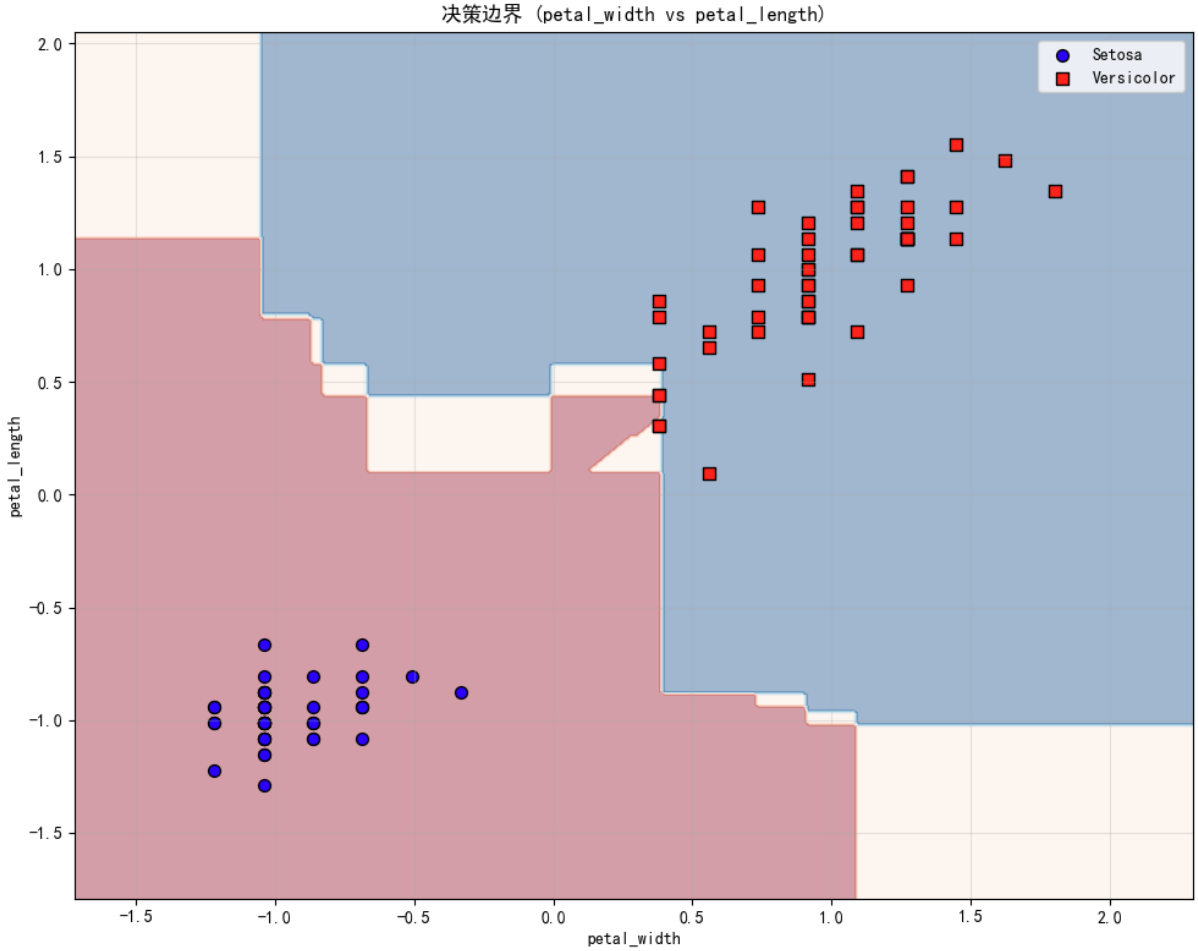


图 9 决策边界

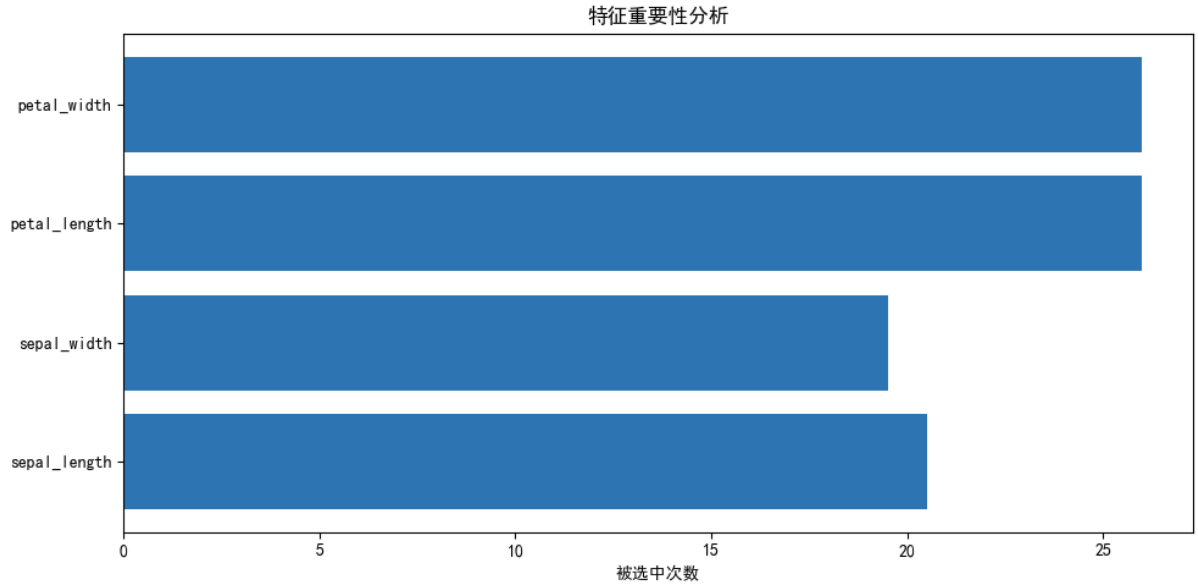


图 10 特征重要性

四、结论

本文针对鸢尾花数据集中 Setosa 与 Versicolor 两类样本的分类问题，通过构建基于 QBoost 算法的弱分类器集成模型并结合量子优化方法，完成了高效且稳定的分类任务，主要结论如下：

首先，模型在分类性能上表现优异，具有以下优势：

(1) 高精度与强泛化能力：筛选出的 92 个弱分类器组合在训练集与测试集上准确率均达 100%，AUC 值为 1.0，泛化差距为 0，表明模型不仅能拟合训练数据，还能稳定适配新样本，无过拟合风险。这一结果源于模拟退火求解器对高区分度分类器的精准筛选，以及花瓣特征在两类样本中的天然分离性。

(2) 特征可解释性强：模型明确凸显花瓣特征的核心作用（占比 78%），花瓣长度与宽度在 Setosa 与 Versicolor 中几乎无重叠，为分类提供了稳定依据。这种基于数据分布的特征选择逻辑，使模型决策过程可追溯、可解释。

(3) 优化方法适配性高：将弱分类器集成问题转化为 QUBO 模型，通过 Kaiwu SDK 的模拟退火求解器高效求解，兼顾全局搜索与局部收敛。对比不同选择策略（仅选最佳、随机选择、全选）可知，QUBO 模型能有效平衡分类误差与模型复杂度，避免单一分类器的泛化局限或全选导致的性能冗余。

(4) 弱分类器设计多样性：280 个弱分类器覆盖单一特征与组合特征（如花瓣长宽比），花瓣相关分类器性能显著优于萼片特征，为集成模型提供了优质基组件。

同时，模型也存在一定局限：

(1) 特征依赖风险：过度依赖花瓣特征（占比 78%）可能导致模型在花瓣特征重叠的数据集上性能下降。例如，若新增样本中两类花瓣特征出现交叉，模型准确率可能降低。

(2) 求解器参数敏感性：模拟退火求解器的性能受初始温度、降温系数等参数影响。若参数设置不合理（如初始温度过低），可能陷入局部最优，需通过多次实验调优。

(3) 计算复杂度较高：QUBO 矩阵维度随弱分类器数量增长（本文为 100×100），当分类器数量增至数千时，求解效率可能下降，需结合剪枝策略简化模型。

综上，本研究验证了 QBoost 算法结合量子优化方法在二分类任务中的有效性，其通过精准筛选高区分度弱分类器、凸显核心特征价值，为同类分类问题提供了兼具性能与可解释性的解决方案。未来可通过引入特征增强技术、动态调整求解器参数及优化模型剪枝策略，进一步提升模型的鲁棒性与适用范围。

参考文献

[1] Freund, Y.; Schapire, R. E. A Desicion-Theoretic Generalization of on-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. https://doi.org/10.1007/3-540-59119-2_166.

[2] Ferreira, A. J.; Figueiredo, M. A. T. Boosting Algorithms: A Review of Methods, Theory, and Applications. In *Ensemble Machine Learning: Methods and Applications*; Zhang, C., Ma, Y., Eds.; Springer: New York, NY, 2012; pp 35–85. https://doi.org/10.1007/978-1-4419-9326-7_2.

[3] QBoost: Predicting Quantiles with Boosting for Regression and Binary Classification. *Expert Syst. Appl.* **2012**, *39* (2), 1687–1697. https://doi.org/10.1016/j.eswa.2011.06.060.

[4] 开物量子开发者社区.

https://kaiwu.qboson.com/plugin.php?id=knowledge&modac=docs&link=QUBOTsp&name=QUBO%E5%BB%BA%E6%A8%A1%E8%AE%B2%E8%A7%A3 (accessed 2025-07-15).

选题	2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
A		apmcm 24102838

飞行器的气动计算及布局优化设计

摘要

这是飞行器的一个优化模型，本文从表面积和体积出发，来寻找表面积与阻力的关系，并且将飞行器复杂机身用曲线方程表示，转化为数学问题，确定目标函数，给出约束条件。在此基础上用遗传算法寻优，比较四种圆锥曲线下的各个参数，确定最优外形。

针对问题一：飞行器本身是一个不规则体，本文将其分为机身和机翼两个部分。对于机身部分，假设它的外形曲线为**二次抛物线**，它的横截面为椭圆面；对于机翼部分，假设机翼的横切面为**二次抛物面**。对各部分分别求面积与体积，再求和得出飞行器的表面积为 $5021\boldsymbol{m^2}$，体积 $\boldsymbol{109.6m^3}$。

针对问题二：根据题中提供的参数构成比例尺，确定舱体各部分的几何参数。同样将舱体分解为几个简单圆柱体和半球体的叠加。为了便于计算，我们将两个柱体之间的接触体近似看作圆柱，接触体在柱体上的剖面近似看作圆。代入参数，即可计算出飞行器舱体的表面积为 $\boldsymbol{110.8m^2}$,体积为 $\boldsymbol{42.6m^3}$。

针对问题三：先查找飞行器飞行时的升、阻力方程，随后将机身和机翼的结构简化，以**舱体重心**为原点建立坐标系，从飞行器的表面积入手，得到迎风面积的计算公式并将其作为优化模型的目标函数。约束条件即舱体的三个半球和连接舱体的圆柱不能与机头的椭圆抛物面相交，同时结合给出的飞行器各个参数的设计上、下限，建立完整的**非线性优化模型**，并采用**信赖域优化算法**对模型进行求解，得到当空气密度为 1.225m^3，飞行速度为 100m/s，阻力系数为 1.5 时，飞行器所受到的阻力约为 **86.3KN**。之后构建**召回函数**来可视化目标函数随迭代次数的变化，目标函数值稳定，确定找到了最优解。**对于机翼**，结合胜利计算公式提出机翼翼肋的**平均受力公式**，将翼肋的平均受力与翼肋总数相乘作为目标函数，转化为**单目标优化问题**。机翼结构优化模型是一个 **0-1 混合整数规划模型**，可求解得到最小平均载荷约为 **0.2652MPa**，以此得到最优外形，具体结构参数见**表 5** 和**表 6**。

针对问题四：通过平面斜切圆锥，可以生成圆、椭圆、抛物线和双曲线等典型的飞行器截面形状。结合**模线设计方法**和二次曲线形状控制参数，可以快速、简便且精确地构建各种**二次曲线**弹身形状。本文提出了一套预估横截面为二次曲线的飞行器纵向和横向气动力的工程计算方法，并基于此建立了二次曲线截面弹身飞行器的优化设计模型。利用该模型，对圆截面、椭圆截面、双曲线截面和抛物线截面外形进行比较。最后，通过比较二次曲线截面弹身外形飞行器的气动特性，确定了**双曲线**为最优外形。

关键词：0-1 混合整数规划 信赖域优化算法 模线设计方法 单目标优化 二次曲线

一、问题重述

1.1 问题背景

目前，优化设计技术在飞行器气动布局设计中发挥着重要作用，而构建合理的优化模型是获取理想气动外形的关键所在。合理筛选几何外形控制参数，并从众多布局参数中选取最关键、最少的、对设计目标敏感的变量进行优化设计，对于提升设计质量和效率至关重要。

1.2 问题要求

基于飞行器的结构示意图，通过建模来优化飞行器外形，解决以下问题：

问题 1：根据飞行器的部分尺寸示意图，估计此飞行器的表面积和体积。

问题 2：已知飞行器舱体结构示意图，根据给出的数据和图中比例尺，估算飞行器舱体结构的表面积和体积。

问题 3：根据给出的飞行器结构参数的取值范围，设计出飞行器的最佳外形，并给出参数的最优值。

问题 4：分别以四种圆锥曲线作为飞行器的外形，重新求解飞行器的最佳外形，并给出结构参数。

二、问题分析

该问题旨在建立一个飞行器气动布局优化模型，求解出飞行器的最佳外形。

2.1 问题一分析

飞行器本身是一个不规则体，本文采用几何法求其表面积和体积，对于机身部分，假设其外形曲线为二次抛物线，横截面为椭圆面；对于机翼部分，假设其横切面为二次抛物面，分别计算出表面积和体积再求和。

2.2 问题二分析

根据提供的舱体结构示意图数据，可以通过图中某个值的大小与实际值构成比例尺，从而可以确定舱体各部分的几何参数。采用几何分解法，将复杂舱体结构分解成几个简单几何体的叠加，分别计算每个几何部分的表面积和体积再求和。

2.3 问题三分析

根据给出的飞行器结构参数的取值范围，我们分别对机身与机翼提出优化方案。保

证优化策略既能有效降低气动计算量，又能求解出飞行器的最佳外形，使得所受阻力最小，得到相对精确的优化解。

2.4 问题四分析

利用平面斜切圆锥获得的二次曲线构造圆、椭圆、抛物线及双曲线等典型的飞行器截面形状。我们提出并建立了二次曲线截面飞行器的优化设计模型，并利用相同的优化模型对圆截面、椭圆截面、双曲线截面及抛物线截面外形进行了优化，求解飞行器的最佳外形，并给出结构参数。

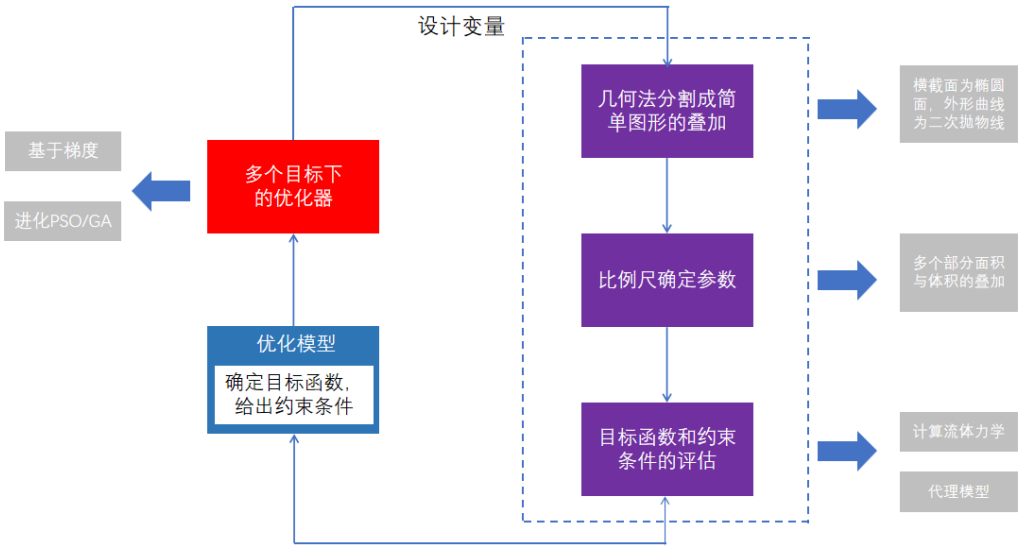


图 1 气动外形优化设计流程图

三、模型假设

- 假设一：机身是一个上下对称的几何体。
- 假设二：飞行器为蒙皮桁条结构，且蒙皮为刚体。
- 假设三：飞行器在晴朗无风的空域匀速飞行。
- 假设四：构建有限元模型时，忽略机身弹性的影响

四、符号说明

符号	符号说明
ρ	流体密度
C_F	阻力系数
A	流体面积
ω_i	权重因子

注：这里并未列出部分变量，这是由于它们在不同小节处有不同的含义，因此我们

会在每一节中详细讨论它们。

五、模型建立与求解

5.1 问题一模型的建立与求解

5.1.1 机身部分的求解

飞行器是一个不规则体，为了求解它的表面积和体积，本文将飞行器分为机身和机翼两个部分。对于机身，我们假设飞行器机身的外形曲线为二次抛物线，它的横截面为椭圆面，设主体前部高为h_1,后部高为h_2，其中a为椭圆截面的长轴，b为椭圆截面短轴。

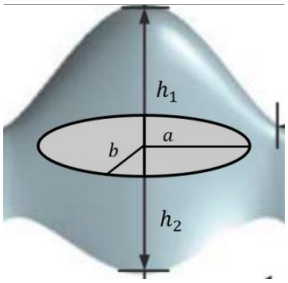


图 2 飞行器外形模拟图

根据模型假设，可求解飞行器机身各部分面积与体积。
飞行器机身前部表面积为：

$$S_1 = \int_0^{h_1} 2\pi b\sqrt{\frac{h_1 - y}{h_1}} - 4(a-b)\sqrt{1-\frac{y}{h_1}}dy = \frac{2}{3}h_1(2\pi b + 4a - 4b) \quad (1)$$

飞行器机身前部的体积为：

$$V_1 = \int_0^{h_1} \frac{\pi ab(h_1 - y)}{h_1}dy = \frac{ab\pi}{h_1}(\frac{{h_1}^2}{2}) \quad (2)$$

飞行器机身后部的表面积为：

$$S_2 = \int_0^{h_1} 2\pi b\sqrt{1-\frac{y}{h_1}} - 4(a-b)\sqrt{1-\frac{h_2}{h_1}}dy = \frac{2}{3}h_2(2\pi b + 4a - 4b) \quad (3)$$

飞行器机身后部的体积为：

$$V_2 = \int_0^{h_2} \frac{\pi ab(h_2 - y)}{h_2} dy = \frac{ab\pi}{h_2}(\frac{{h_2}^2}{2}) \quad (4)$$

5.1.2 机翼部分的求解

对于机翼部分，我们假设机翼的横切面为二次抛物面，设机翼前部高为 l_3 ，后部高为（l_5-l_3），宽度为e，机翼长度为L。

机翼横截面前部的弧长为：

$$L_1 = \int_{-e}^{e} \sqrt{1 + \frac{4{l_3}^2 x^2}{e^4}} dx \quad (5)$$

机翼横截面后部的弧长为：

$$L_2 = \int_{-e}^{e} \sqrt{1 + \frac{4(l_5 - l_3)^2 x^2}{e^4}} dx \quad (6)$$

机翼的表面积为：

$$S_3 = (L_1 + L_2)L \quad (7)$$

机翼前部截面积为：

$$C_1 = \int_{-e}^{e} -\frac{l_3 x^2}{e^2} + l_3 \, dx \quad (8)$$

机翼后部截面积为：

$$C_2 = \int_{-e}^{e} -\frac{(l_5 - l_3)x^2}{e^2} + l_5 - l_3 \, dx \quad (9)$$

机翼的体积为：

$$V_3 = (C_1 + C_2)L \quad (10)$$

5.1.3 模型的求解

飞行器的总表面积为：

$$S = S_1 + S_2 + S_3 \quad (11)$$

飞行器的总体积为：

$$\mathrm{V} = \mathrm{V}_1 + \mathrm{V}_2 + V_3 \tag{12}$$

将参数代入，可计算得到飞行器的表面积为 $\mathbf{5021}\boldsymbol{m^2}$，体积为 $\mathbf{109.6}\boldsymbol{m^3}$。

5.2 问题二模型的建立与求解

根据题中提供的参数以及比例尺，能够确定舱体各部分的几何参数。本通过 image J 软件测得R_1在图中的大小，与实际值相对应构成比例尺，再根据这个比例尺得到各部分几何参数，如下表 1：

表 1 参数测量值

参数	测量值
R_1	90.0cm
R_2	100.0cm
R_3	24.0cm
d_1	355.8cm
d_2	572.3cm
d_3	241.9cm

舱体结构是一个复杂几何体，本文为简化计算，将其分解为几个圆柱体和半球体的叠加。

表面积公式如下：

$$\begin{cases} S_{柱} = 2\pi r^2 + 2\pi rh \\ S_{球} = 4\pi r^2 \end{cases} \tag{13}$$

体积公式如下：

$$\begin{cases} V_{柱} = \pi r^2 h \\ V_{球} = \dfrac{4}{3}\pi r^3 \end{cases} \tag{14}$$

为了便于计算，我们将两个柱体之间的接触体近似看作圆柱，接触体在柱体上的剖面近似看作圆。

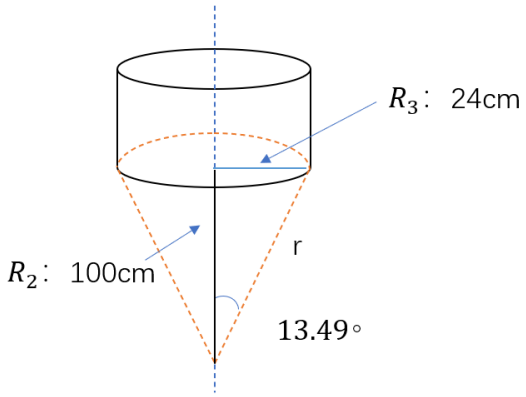


图 3 两个柱体接触简化模型

底边扇形的 1/2 原弧长为 $2\pi r * 26.8^{\circ}/360^{\circ} =$ 48.08cm，圆的直径 $2R_3 =$ 48cm，两者相差不大，可将接触面近似看作圆。

得到的整个几何体的表面积公式和体积公式为：

$$\begin{cases} S_{总} = 2 * （4\pi {R_1}^2 + 2\pi R_1 d_1） - 4\pi {R_3}^2 + 4 * 2\pi R_3 * (d_3 - R_1 - R_2) + 4\pi {R_2}^2 \\ \qquad +2\pi R_2 d_2 - 4\pi {R_3}^2 \\ V_{总} = 2 * \frac{4}{3}\pi {R_1}^3 + 2\pi {R_1}^2 d_1 + \pi {R_2}^2 d_2 + \frac{4}{3}\pi {R_2}^3 + 4\pi d_3 {R_3}^2 \\ \qquad -4\pi R_1 {R_3}^2 - 4\pi R_2 {R_3}^2 \end{cases} \tag{15}$$

将参数代入其中，可计算出飞行器舱体的表面积为 $\boldsymbol{110.8m^2}$,体积为 $\boldsymbol{42.6m^3}$。

5.3 问题三模型的建立与求解

飞行器在飞行时受到的阻力方程为：

$$F = \frac{1}{2}\rho v^2 * C_F * A \tag{16}$$

其中ρ为流体密度，v为飞行器相对流体的运动速度，C_F为阻力系数，是一个常数，A为飞行器的迎风面积。由问题假设(3)、(4) 可知，ρ，v均为常数，阻力系数C_f通常由飞行器表面的材料与外形决定，在题目中已经给出指定飞行器的情况下，结合问题假设 (2) 可知，C_f也为常数。因此，可以将问题(3)转化为使飞行器的迎风面积最小的优化问题。

针对机翼结构，通过查阅资料可知，飞行器的机翼在飞行过程中所受到的升力计算公式为：

$$f = \frac{1}{2}\rho v^2 C_l S \tag{17}$$

ρ为流体密度，v 为飞行器相对流体的运动速度，C_l为阻力系数，S 为飞行器的翼展面积。在使飞行器迎风面积最小的前提下，优化机翼结构，使翼肋在飞行过程中所受平

均载荷最小的同时，翼肋数量尽可能少。

5.3.1 飞行器结构的简化

•骨架结构的简化

由题中图 1 可知，l_2为机翼最内侧翼肋的顶端到舱体的小半球之间的水平距离。如图所示，不妨假设机头（机身的迎风部位，即飞行过程中比机翼内侧翼肋顶端更靠前的部位）的长度为 $3l_2$=360cm，且舱体的重心即为机头的底部中心。因此，舱体的重心与舱体的几何中心在同一水平线上，并且比舱体的几何中心更靠近机头 5cm，即 G_c=$3l_2$=360cm。

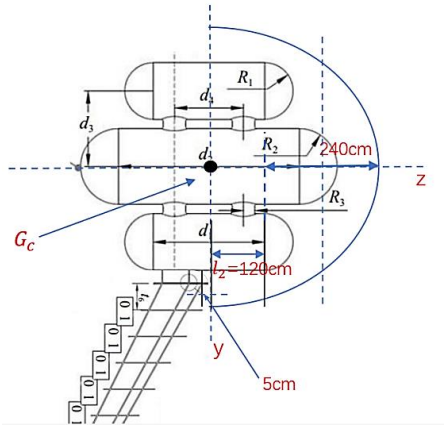


图 4 飞行器结构的简化

•机头形状的简化

为了使飞行器的阻力系数C_f尽可能小，同时简化计算，如图 x 所示，本文将机头的主视图看作一个抛物线，并以机头的底部中点作为原点，建立平面直角坐标系。则可以得到抛物线的解析式,其中，h 为机头的长度。

$$y = \frac{h}{25}(5 + x)(5 - x) \tag{18}$$

根据问题 1 的模型，可以将机头的底部看作一个椭圆，因此，机头的几何形状可以被大致等效为一个椭圆抛物面。将上述抛物线沿纵轴旋转，并通过坐标轴的伸缩变换可以得到机头在空间直角坐标系的解析式如下，b 为机头底部椭圆的短轴。

$$z = 3.6(1 - \frac{y^2}{25} - \frac{x^2}{b^2}) \tag{19}$$

机头在空间坐标系下的形状如图所示：

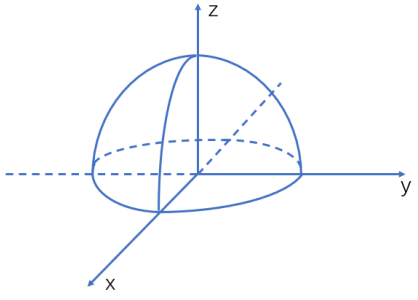


图 5 飞行器机头坐标系的建立

•机翼结构的简化

由题目中对飞行器机翼参数的设定，如图所示，本文对结构做出如下简化：

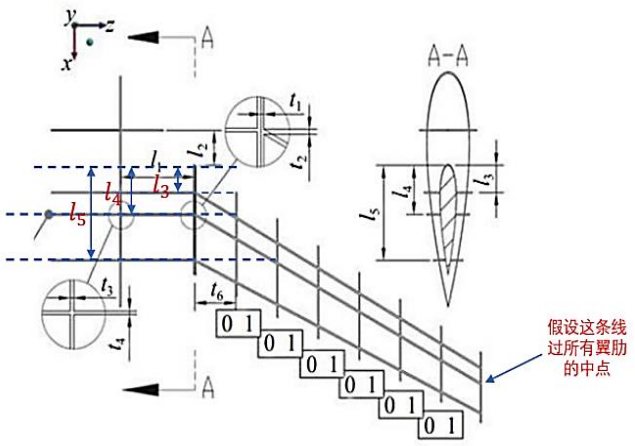


图 6 机翼部分结构的简化

其中l_3为翼肋前部超出机翼骨架部分的长度,l_4为翼肋一半的长度,l_5为翼肋的顶部到机翼第三根骨架的长度。同时我们假设：

（1）机翼的第二根骨架过所有翼肋的中点。

（2）最内侧翼肋的长度为最外侧翼肋长度的两倍。

（3）翼肋仅在机翼骨架的内部承受载荷。

（4）机翼的迎风面积为翼肋前部超出机翼骨架部分的面积。

5.3.2 机身外形-舱体结构优化模型

通过前文建立的飞行器表面积计算模型，本文提出了飞行器迎风面积的计算公式：

$$\iint_{D_{xy}} \sqrt{1+(\frac{36y}{125})^2+(\frac{36x}{5b})^2}\,dx\,dy + 30(\pi c + 2(l_3 - c)) \tag{20}$$

其中：

$$D_{xy}:\begin{cases} z=0 \\ \dfrac{y^2}{25}+\dfrac{x^2}{b^2} \le 1 \end{cases} \tag{21}$$

c 为机翼的厚度，本文取 c=0.5 。

经过计算，飞行器迎风面积约等于 115b。因此得到目标函数：

$$\min 115b + 30(0.5\pi + 2(l_3 - 0.5)) \tag{22}$$

飞行器除了需要满足题目给出的设计变量，还需要能够容纳整个舱体。将这个约束条件进行简化，即为：舱体的三个半球和连接舱体的圆柱不能与机头的椭圆抛物面相交。将这些几何体放在椭圆抛物面的坐标系中，可以得出它们的曲面方程。这些几何体的曲面方程分别与椭圆抛物面的曲面方程组成的非线性方程组无实数解，即为该优化模型的约束条件。

对机舱小半球：

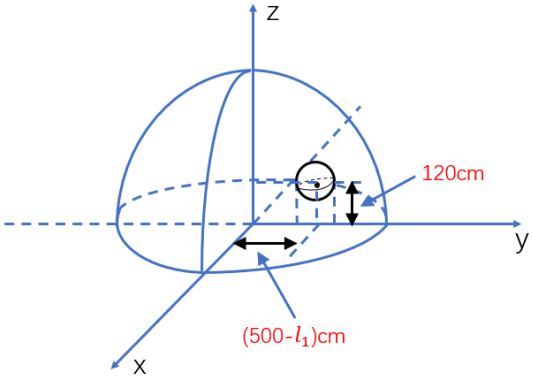


图 7 机舱小半球在坐标系下的分析

如图所示，由于两个小半球关于 zOx 面轴对称，且椭圆抛物面也关于 zOx 面轴对称，因此只需要保证一个其中小球方程与椭圆抛物面方程组成的方程组无实数解即可。

小球方程为：

$$x^2 + (y - 5 + l_1)^2 + (z - 1.2)^2 = (R_1 + t_5)^2 \tag{23}$$

即方程组：

$$\begin{cases} x^2 + (y - 5 + l_1)^2 + (z - 1.2)^2 = (R_1 + t_5)^2 \\ z = 3.6(1 - \dfrac{y^2}{25} - \dfrac{x^2}{b^2}) \end{cases} \tag{24}$$

无实数解。

对机舱大半球：

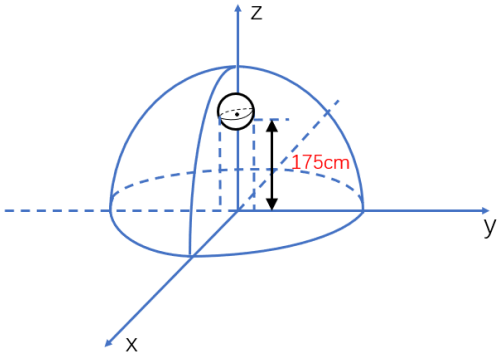


图 8 机舱大半球在坐标系下的分析

如图所示，大球方程为：

$$x^2 + y^2 + (z - 1.7)^2 = (R_2 + t_6)^2 \tag{25}$$

即方程组：

$$\begin{cases} x^2 + y^2 + (z - 1.7)^2 = (R_2 + t_6)^2 \\ z = 3.6(1 - \dfrac{y^2}{25} - \dfrac{x^2}{b^2}) \end{cases} \tag{26}$$

无实数解。

对连接舱体之间的圆柱：

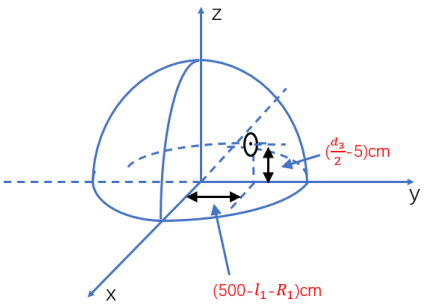


图 9 连接舱体部分在坐标系下的分析

由图可知，越靠近 z 轴，机头的内部空间越大，因此，只需要保证圆柱中最外侧的圆与椭圆抛物面无交点即可。根据对称性，秩序保证其中一个圆与椭圆抛物面无交点即可。

圆的曲面方程为：

$$\begin{cases} y = 5 - l_1 - R_1 \\ x^2 + (z - 0.7)^2 = (R_3 + t_7)^2 \end{cases} \tag{27}$$

因此方程组：

$$\begin{cases} y = 5 - l_1 - R_1 \\ x^2 + (z - 0.7)^2 = (R_3 + t_7)^2 \\ z = 3.6(1 - \dfrac{y^2}{25} - \dfrac{x^2}{b^2}) \end{cases} \tag{28}$$

无实数解。

结合题目中给出的飞行器各参数的设计上限与设计下限，本文建立了一个完整的优化模型。由于这是一个非线性规划模型，本文采用信赖域优化算法对模型进行求解。

信赖域优化(Trust-constr)算法是一种用于数值优化的算法，特别是用于求解带约束的优化问题。流程如下：

（1） **初始化：**

●设置初始参数和信赖域半径。

●确定初始的可行解（满足约束条件的解）。

（2）**迭代优化过程**

●求解子问题：

在当前信赖域内，通过解一个局部的数学规划问题（通常是二次规划或修正的二次规划），找到在当前信赖域内的最优解。

●更新信赖域：

根据求解出的最优解更新信赖域的大小和形状，以便在下一步继续优化。

（3）**终止条件：**

●在每次迭代中检查是否满足了终止优化的条件，例如目标函数的收敛或满足一定的数值容限。

●如果满足终止条件，则停止优化；否则，回到第二步继续迭代。

运用 python 中的 scipy 库实现优化模型，调用 minimize 函数进行求解，并设置算法为信赖域优化算法。求解得到飞行器的迎风面积约为 141 ㎡，将其带入阻力计算公式，当空气密度为 1.225m³，飞行速度为 100m/s，阻力系数为 1.5 时，飞行器所受到的阻

力约为 86.3KN。飞行器设计参数的最优值结果见表 5：

表 5 飞行器设计参数的最优值

参数	数值
l_3	0.109
R_1	0.776
R_2	0.875
R_3	0.250
t_5	0.115
t_6	0.115
t_7	0.115
l_1	2.800

构建召回函数，并调用 minimize 函数中的"callback"方法，将每次迭代过后的目标函数值记录在列表中，运用 matplotlib 库绘制目标函数随迭代次数的变化图见图 10：

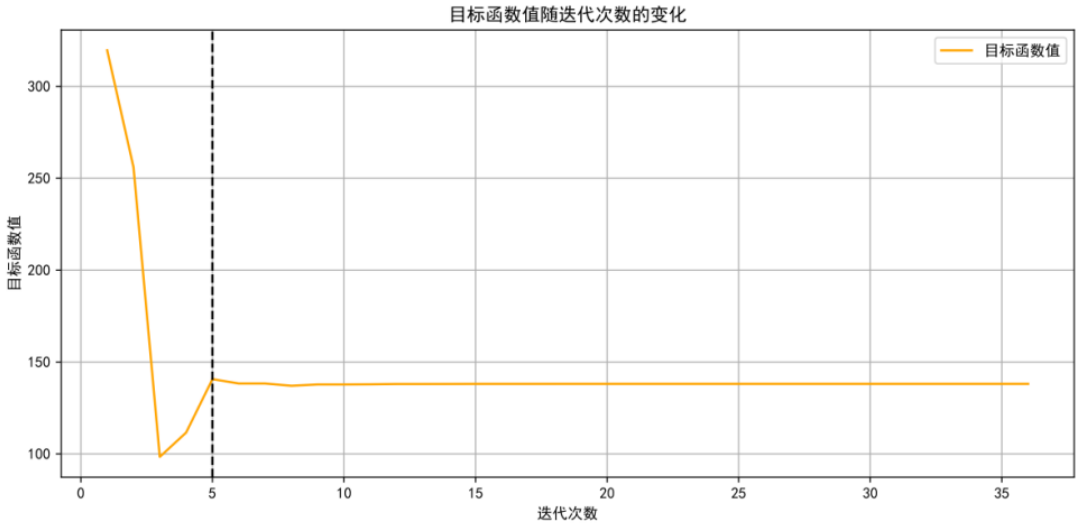


图 10 目标函数随迭代次数的变化

由图可知，在经过五次迭代后，目标函数值稳定在 140 左右，可以确定该算法找到了模型的局部最优解。

5.3.3 机翼结构优化模型

结合前文并依据“平行线分线段对应成比例”的几何原理，从机翼内侧到机翼外侧排布的 8 个翼肋的长度组成一个等比数列，设 q 为公比，则有：

$$q^7 = \frac{1}{2} \tag{29}$$

飞行器的翼展面积是一个上底为l_4，下底为 $2l_4$，高为 10m 的梯形。由此可以得出一侧机翼翼肋的平均受力公式：

$$f = \frac{0.5 * 10 * \rho v^2 C_f l_1}{(\sum_{i=1}^{6} C_i^{i6} (l_5 - l_3) q^4) + 3(l_5 - l_3)} \tag{30}$$

由于需要让翼肋平均受力尽可能小的同时，为了节省材料和减少飞机重量，翼肋的数量要尽可能少。因此不妨将翼肋的平均受力与翼肋总数相乘作为目标函数，转化为单目标优化问题：

$$\min = (\frac{0.5 * 10 * \rho v^2 C_l l_1}{(\sum_{i=1}^{6} C_{i6}^{i} (l_5 - l_3) q^4) + 3(l_5 - l_3)}) \sum_{i=1}^{6} C_{i6}^{i} \tag{31}$$

取空气密度ρ=1.225kg/m³ ，飞行速度v=100m/s，升力系数C_l=1.5。因此机翼结构优化模型是一个 0-1 混合整数规划模型 ：

$$\min = (\frac{9.1875 * 10^4 l_1}{(\sum_{i=1}^{6} C_{i6}^{i} (l_5 - l_3) q^4) + 3(l_5 - l_3)}) \sum_{i=1}^{6} C_{i6}^{i} \tag{32}$$

$$s.t. \begin{cases} C_{i6}^{i} \in \{0,1\} \\ 0.45 \le l_4 \ll 0.55 \\ 0.65 \le l_5 \le 0.9 \\ q^7 = \frac{1}{2} \end{cases} \tag{33}$$

其中，l_3的值已经在机身外形-舱体结构优化模型的求解过程中被确定为0.109999m。

运用 python 语言对模型进行编程，并调用 scipy 库中的 minimize 求解器，指定求解算法为“SLSQP”，得到机翼结构参数与最小平均载荷见表 6：

表 6 机翼结构参数

参数	数值
$C_{l_6}^1$	1
$C_{l_6}^2$	1
$C_{l_6}^3$	0
$C_{l_6}^4$	1
$C_{l_6}^5$	1
$C_{l_6}^6$	0
l_4	0.5165
l_5	0.7645
平均载荷	0.2652MPa

通过对机身与机翼采取不同模型进行优化，在使得阻力最小的情况下，求解出最佳外形，提高飞行器的设计水平，获得性能更为优异的总体方案，可应用于类似飞行器的总体优化设计中，对工业部门具有较强的参考意义。

5.4 问题四模型的建立与求解

本问以双锥体为基础，在飞行器后体采用二次曲线控制横截面外形，并采用模线设计法构造非圆截面弹身。之后构建这类截面为二次曲线的飞行器气动力计算方法，建立优化设计模型并计算。最后，对优化后的飞行器布局的气动特性进行了进一步的研究。

5.4.1 截面曲线设计方法

首先根据需要将飞行器沿纵向划分成若干纵向控制站位，然后在每个控制站位上生成不同的横截面控制点，并用光滑曲线连接形成横截面形状，最后将各个横截面控制点用光滑的纵向曲线连接起来。横截面形状是机身设计的关键之一。

其二次曲线的一般方程形式如下：

$$aX^2 + bXY + cY^2 + dY + eY + f = 0 \qquad (34)$$

二次曲线可以看成是平面斜切圆锥所得，通过改变切面的角度，可以获得圆、椭圆、抛物线及双曲线等二次曲线(图 11)。假设 A 点为飞行器机身铅垂平面与横截面在背风面的交点，B 点为水平面与横截面的交点， C 点为过点 A 及点 B 的切线的交点，则肩点 E 控制平面 ABC 内过 A、B 点的二次曲线形状。若点 D 为$\overline{AB}$直线的中点，引入二次曲线形状参数ρ，且$\rho = \frac{|\overline{DE}|}{|\overline{DC}|}$则可以通过控制形状参数ρ 的取值，唯一地确定肩点 E 的位置，并进而唯一地确定二次曲线 AEC 的形状。

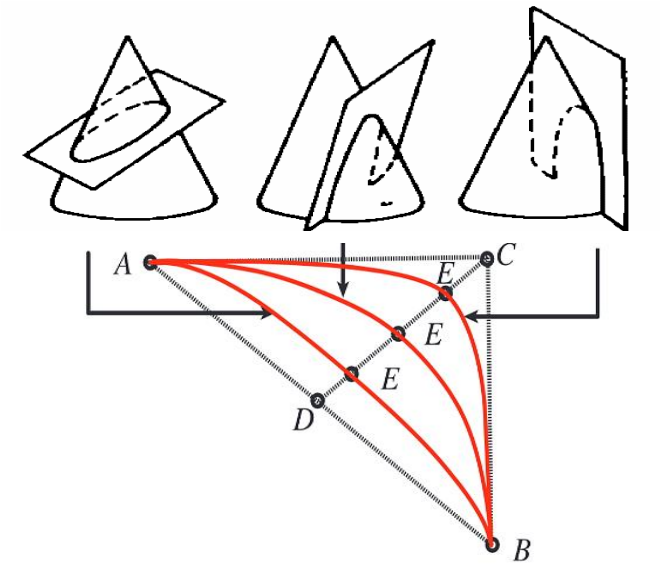


图 11 平面切圆锥二次曲线示意图

本文中横截面形状将由以下 5 个参数确定：0°背风子午线控制点半径r_1、90°侧向子午线控制点半径r_2、180°迎风子午线控制点半径r_3、上半身二次曲线形状参数ρ_1及下

半机身二次曲线形状参数ρ_2。

5.4.2 高超声速气动力估算方法

本文通过引入等价锥及等价横截面等概念，对速度比和动压比进行考虑和修正，并引入了前体对后体的气动影响的计算方法，使得内伏牛顿理论推广发展到非圆截面外形，从而预估飞行器气动特性。计算中，机身面元上的压力系数C_p为：

$$C_p = C_{p0} + f(X^*, M) \cdot C_{pmax} \cdot \left(\frac{U_\perp}{U}\right)^2 + f_1\left(\frac{U_\perp}{U}, M\right) + f_2(R_e, T_\omega) + f_3(\delta^*, R_e, M) \tag{35}$$

$$\frac{U_\perp}{U} = \vec{V} \cdot \vec{N} \tag{36}$$

其中，$\vec{V}$、$\vec{N}$、M分别为来流速度矢量、物面内法向矢量及马赫数。C_{p0}为爆炸波压力系数，C_{pmax}为驻点压力系数，$f(X^*, M)$为动压比修正函数，f_1、f_2、f_3分别为背风面低压修正、粘性修正及边界层位移厚度修正，R_e、T_ω、δ^*分别为雷诺数、壁面温度及边界层位移厚度。

5.4.3 多目标优化遗传算法

多目标数学规划问题可以描述为寻找一个决策变量的向量x^*，满足约束条件下，使目标函数向量$f(x^*)$达到最优,即:

$$\begin{cases} V_{min} f(\boldsymbol{x}) = [f_1(x), f_1(x), \cdots, f_n(x)]^T \\ x \in X \\ X \in R^m \end{cases} \tag{37}$$

其中V_{min}表示向量目标函数，若设$X \in R^m$是多目标优化模型的约束集，$f(x) \subseteq R^m$是多目标优化时的向量目标函数，若有解$x_1 \in X$并且x_1优于X中的所有其他解则称解x_1是多目标优化模型的最优解。

为了提升遗传算法的效率，本文针对遗传算法易出现的遗传欺骗问题进行了改进。通过将 Pareto 非劣解概念与遗传算法思想相结合，并采用实数编码、小生境淘汰、群体排序、稳态复制和动态惩罚等技术，克服了标准遗传算法的缺点，提高了优化效率。同时，在进化中后期引入模拟退火算法进行局部搜索，最终形成了一种适用于复杂优化问题的多目标混合遗传算法。

5.4.4 升力体外形优化设计模型

飞行器在无动力滑翔再入阶段的总体性能涉及多个方面，包括升阻特性（以升阻比衡量）、驻点气动热特性（以机身头部驻点温度表示）、稳定性以及机身内部容积等。设计目标是在满足一系列外形约束条件和性能指标的前提下，优化飞行器的气动外形，使其在滑翔状态下实现最大的气动升阻比和容积利用率。

•约束条件

采用多目标优化混合遗传算法对飞行器进行了多约束的气动升阻比优化设计。主要的约束条件包括

①机身驻点温度限制:T_s ＜ 1470K。

②纵向、航向静稳定性限制:纵向均为静稳定静稳定度小于2%。纵向、侧向最大可用过载限制n_{vk}＞1.5, n_{vk}＜0.5。

③最大舵偏角限制:δ＜20°。

•设计变量

通过对整个通用再入飞行器的数值建模提取其外形参数作为设计变量上机身二次曲线形状参数ρ_1，下机身二次曲线形状参数ρ_2，飞翼几何参数展弦比λ，根稍比η和前缘后掠角χ_0，飞翼面积 S，方向舵几何参数展弦比λ_f，根稍比η_f和前缘后掠角χ_{of}，舵面面积 S_f，升降副翼几何参数展弦比λ_s，相对厚度 c 与舵面面积S_s。

以典型飞行状态高度 30km，Ma＝ 6 ，迎角20°，舵偏角0°。设定种群规模为 500 交叉概率为 0.8，变异概率为 0.2，最大进化代数为 400。进化到 327 代后种群收敛得到如表 8 所示的再入飞行器 Pareto 最优解。

表 8 优化设计下的 Pareto 解

截面形	K	V_u	T_s/K	X_g	λ	χ_0 /(°)	η	χ_{of} /(°)	$\psi_{x\omega}$ /(°)	S_f /dm^2	S_d /dm^2
抛物线	3.015	0.6321	1235	0.65	5.95	48.81	1.1052	57.53	67.43	0.2526	0.6021
圆	2.857	0.5812	1190	0.637	4.965	45.57	1.1254	52.33	65.48	0.3415	0.6529
双曲线	2.623	0.6321	1154	0.645	4.862	49.35	1.1328	60.28	69.35	0.3181	0.6929
椭圆	2.758	0.5568	1319	0.656	5.368	50.21	1.3542	54.57	70.12	0.4538	0.6403

表 8 中，K为飞行器中机身横截面分别为抛物线、圆、双曲线和椭圆时的升阻比，V_u为容积利用效率，T_s为飞行器头部驻点温度，X_g为质心系数为质心距头部的距离与特征长度之比，λ、χ_0及η为飞翼参数，χ_{of}、 S_f和S_d为方向舵舵面参数，$\psi_{x\omega}$为布局的控制参数。

根据表 8 的数据，双曲线截面在利用率方面表现最佳，并且双曲线截面的头部驻点温度最低。综合考虑各个性能指标，本文选择在主要性能指标得到一定优化，同时允许次要性能指标在可接受范围内做出一定牺牲的方案，以显著提升主要性能指标。因此，本文推荐以双曲线截面为特征的 Pareto 解作为通用再入飞行器外形设计的方案，其外形轮廓如图 12 所示：

图 12 最优构型的飞行器外形

5.4.5 结果验证

为了进一步验证推荐方案的气动特性本文采用 CFD 方法开展气动特性的仿真分析．图 13 给出了 Ma＝ 6 攻角20°条件下飞行器表面的压力分布特性。

图 13 表面压力分布（Ma＝ 6 ， AoA＝20°）

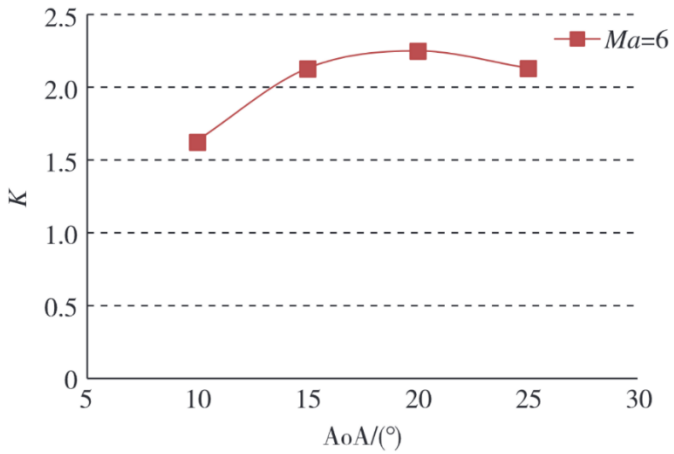


图 14 升阻比随攻角的变化

图 14 给出了在 Ma＝ 6 条件下升阻比随攻角的变化情况可以看出飞行器在攻角20°时升阻比达到最大值 22.5 与 Pareto 解相差约 12%，基本上定性验证了优化结果。

六、模型评价与改进

6.1 模型的优点分析

●将代理模型的预测最优值作为每步迭代的新增样本之一的 Kriging 代理模型的加点准则可以改善搜索的局部收敛性；

●本文发展的方法可以满足方案论证和初步设计阶段对气动系数的精度需求；

●采用二次曲线方程并引入二次曲线形状参数，结合优化设计技术可以方便快速地获得 各种截面外形，而且可以提高设计质量和设计效率， 值得进一步的深人研究和推广应用；

6.2 模型的缺点分析

●本文建立的模型使用的参数较多，计算时可能会存在误差，影响模型精度。

●尽管本文考虑了多种气动、结构 指标，但仍没有考虑某些重要因素（如屈曲、颤振、 内装燃油、外挂发动机等）；

●本文提出的方法应用于概念设计阶段，省略了额外的几何细节及相关负载，因此本文优化结果可能过于理想；

参考文献

[1]王国辉,王小军,杨勇,等.火箭基组合循环(RBCC)推进系统研究现状[J].固体火箭技术,2003,(03):1-3+6.

[2]唐伟,张鲁民.弯体机动再入飞行器气动特性研究[J].空气动力学学报,1996,(01):86-91

[3]赵琳瑜,杨琴文,张锋,等.基于 Ansys Workbench 的火箭支架结构拓扑优化设计[J].计算机辅助工程,2024,33(02):6-9.DOI:10.13340/j.cae.2024.02.002.

[4][1]林威全,许航瑞,兰旭东.TBCC 发动机的发展历程及关键技术分析[J/OL].清华大学学报(自然科学版),1-15[2024-07-08].https://doi.org/10.16511/j.cnki.qhdxxb.2024.27.024

.[5] [1]董一韩,曾森,周祥.基于遗传算法优化 BP 神经网络的 RC 柱骨架曲线预测[J].中国水运(下半月),2024,24(07):34-36.

[6]李昊,廖志刚,杨令飞,等.高速飞行器气动外形设计方法综述[J].空天技术,2024,(01):23-35.DOI:10.16338/j.issn.2097-0714.20230351.

选题	2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
A		apmcm24103112

基于空气动力学和遗传算法的飞行器外形优化设计

摘要

飞行器的外形设计不仅关乎其飞行的稳定性和安全性，还直接影响其燃料效率、阻力大小及最终的任务成败。本文通过构建以飞行器所受阻力最小为目标的飞行器外形优化设计模型，旨在提高飞行器的飞行效率和性能。

针对问题一，为求解该部分飞行器的表面积和体积，考虑到飞行器的对称结构，选择使用拆分法将其分为一块主体和两侧机翼。针对飞行器主体部分，假设飞行器外形曲线和横截面分别为二次抛物线和椭圆面，积分求解各自的表面积和体积。针对飞行器机翼部分，合题目提供的侧视图及骨架参数化示意图可进一步进一步将其分为前部与后部，累加求和。matlab 求解得，该飞行器整体的表面积和体积分别 $5.0712 \times 10^7 \mathrm{cm}^2$ 和 $8.4852 \times 10^7 \mathrm{cm}^3$。

针对问题二，为求解飞行器舱体结构的表面积和体积，将其拆分为半球体和圆柱体两部分。对于半球体部分，根据半径R_1, R_2可分为 4 个小半球(R_1)和 2 个大半球(R_2)，且无需计算半球体的底面积。值得注意的是，连接体部分同样为圆柱体，共 4 个。分析可得，飞行器舱体由一个大圆柱体，两个小圆柱体和四个连接体(仍为圆柱体)、两个大半球体和四个小半球体构成。求解得，舱体的表面积和体积分别为 $9.7739 \times 10^5 \mathrm{cm}^2$ 和 $3.7566 \times 10^7 \mathrm{cm}^3$。

针对问题三，本文构建以飞行器所受阻力最小为目标，以飞行器的物理模型和参数结构取值范围等为约束构建飞行器外形优化设计模型。结合问题一与问题二，首先对飞行器进行受力分析，构造飞行器所受阻力与其表面积及体积的函数关系模型。为简化模型，忽略姿态角的变化，考虑使用流体阻力计算其空气阻力。参考相关文献，设置空气密度为 1.225kg/m^3，飞行器飞行速度为 250m/s，空气阻力系数为 0.08，遗传算法求解得：最小阻力为 54118842.1966N，l_3为 0.29039486。粒子群算法求解得，最小阻力为 54208967.1075N，l_3为 0.1676173。其余飞行器结构参数最优值详见表 5-9 与 5-10。对比发现蚁群算法的求解优度较高。

针对问题四，本文在问题三的基础上，分别针对四种圆锥曲线相应修改，目标函数不变，约束条件的变化具体如表 5-12 所示，采用遗传算法重新求解问题三飞行器的最佳外形设计问题。对比可得，在四种圆锥曲线中，以双曲线作为飞行器曲线所受阻力最低，为 54118952.96N。参数矩阵为$[0.3375851,0.47902355,...,355.25455845]$，四种圆锥曲线的求解结果具体如表 5-13 至 5-16 所示。

最后本文评价了所构建的基于空气动力学和蚁群算法的飞行器外形优化设计的优点及不足之处，并阐释其推广应用的方向。

关键词：遗传算法；粒子群算法；空气动力学；飞行器外形优化

一、问题背景

1.1 问题背景

飞行器的外形优化是航空航天领域中一项至关重要的科学问题。无论是在大气层内还是外，飞行器的外形直接影响其性能和效率。大气层内飞行的航空器依赖于空气动力学原理，而太空中的航天器则面临更为严苛的环境和运动要求。

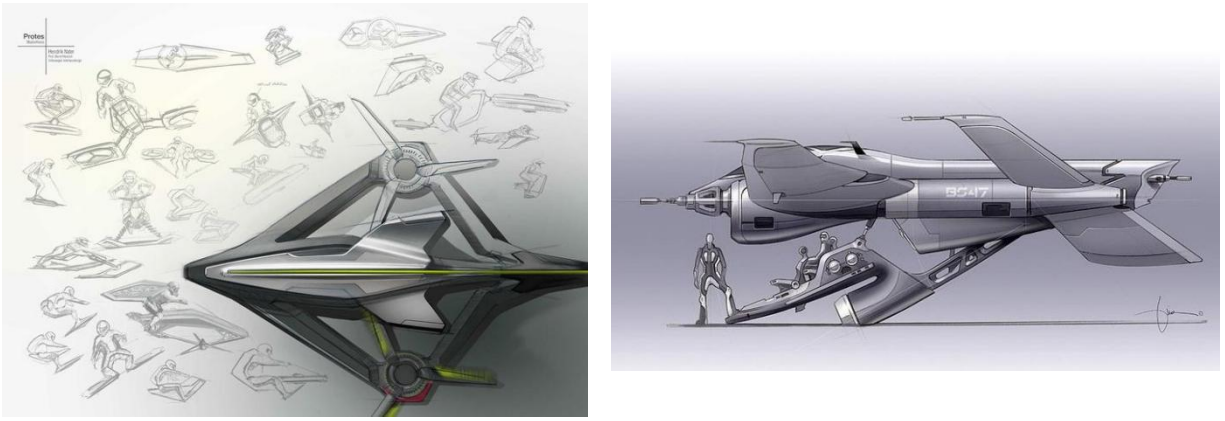


图 1-1 飞行器外形优化设计

飞行器的外形设计不仅关乎其飞行的稳定性和安全性，还直接影响其燃料效率、阻力大小及最终的任务成败。

1.2 研究目的与意义

本文通过构建以飞行器所受阻力最小为目标的飞行器外形优化设计模型，旨在提高飞行器的飞行效率和性能。具体而言：

(1)最小化阻力：通过数学建模和计算分析，寻找飞行器外形的最佳参数组合，使其在给定速度下的空气阻力最小化。这不仅可以节省燃料，还能提升飞行器的速度和稳定性。

(2)优化设计方法：探索并比较不同的优化算法（如遗传算法、粒子群算法和蚁群算法等），以确定最适合于飞行器外形优化的算法。这些算法能够帮助在复杂的设计空间中搜索最优解，从而在设计阶段就能够减少试验成本和时间。

(3)应用推广：将优化后的飞行器外形设计推广到实际应用中，包括各种航空航天器的设计和改进，从而提升整个航空航天工业的竞争力和技术水平。

这些研究目标的实现具有重要的科学意义和实际应用价值，不仅可以推动飞行器设计技术的进步，还有助于解决现实世界中复杂飞行条件下的挑战。通过优化飞行器的外形设计，我们能够更好地探索和利用空气动力学规律，为未来的航空航天事业奠定坚实的基础。

二、问题重述与分析

题目在背景部分阐释了飞行器的含义、发展历程，进而表述优化飞行器的外形，减少阻力对于航空航天事业的重大意义。

问题一主要要求我们估计某飞行器的表面积和体积。

问题二在问题一的基础上提高了问题复杂度，该舱体结构为图 1 的主要部分。

问题三则进一步要求我们设计飞行器的最佳外形。需要设计的参数主要包括骨架结构和舱体结构。

问题四要求我们在问题三的基础上，针对圆形、椭圆、抛物线和双曲线等四种圆锥曲线，分别考虑作为图 1 飞行器的外形，重新设计其最佳外形，并求解飞行器相应的结构参数。

三、模型假设

实际飞行过程中，飞行器的飞行效果往往受到风力的影响。在风力较大的情况下，飞行器往往需要保持特殊的姿态才可以正常进行悬停或飞行等操作，因此必须作出如下假设。

假设 1：假设将飞行器整体视作刚体，即忽略弹性形变对飞行器受力等情况的影响。

假设 2：假设飞行器的质量分布均匀，质心位于原点。

假设 3：忽略地球自转与公转对飞机飞行的影响。

假设 4：不考虑风等空气流动队飞行器受力的影响。

四、符号说明

符号	说明	单位
S_{front}	飞行器前部的表面积	cm^2
V_{front}	飞行器前部的体积	cm^3
h	前部高	cm
l	后部高	cm
S_{back}	飞行器后部的表面积	cm^2
V_{back}	飞行器后部的体积	cm^3
a	椭圆截面长轴	cm
b	椭圆截面短轴	cm
w	宽度	cm
L	机翼长度	cm
F_d	空气阻力	N
c_d	物体阻力系数	/
ρ	空气密度	kg/m^3
v	飞行器相对于空气的速度	m/s
S	飞行器在空气中的横截面积	m^2
$S_{cylinder\_large}$	大圆柱体表面积	cm^2
$S_{cylinder\_small}$	小圆柱体表面积	cm^2

*其余未说明的符号将在正文详细展示。

五、模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 飞行器结构拆分

根据前文分析，问题一要求我们估计某飞行器的表面积和体积。飞行器的尺寸示意图具体如图 5-1 所示：

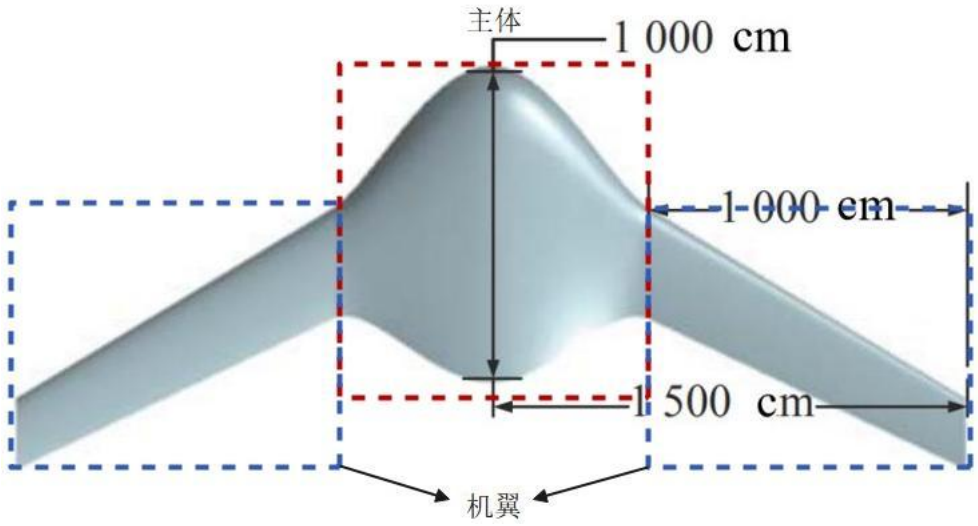


图 5-1 飞行器部分尺寸示意图

在实际计算飞行器表面积和体积的计算中，考虑到飞行器的对称结构，最常用的方法为拆分法，观察图 5-1 发现，该飞行器主要由主体和机翼两部分构成。

对于主体部分，可将其进一步拆解为前部和后部。鉴于题目未提供飞行器截面的厚度信息，本文假定其为椭球面，则截面为椭圆。对于机翼部分，结合骨架参数化示意图，认为其由双抛物线组成，本文将在 5.1.2 和 5.1.3 节进一步阐释。

5.1.2 飞行器主体

在 5.1.1 节中，本文将该部分飞行器细分为一块主体和两侧机翼，对于主体。本文将其细分为前部和后部。具体如图 5-2 所示

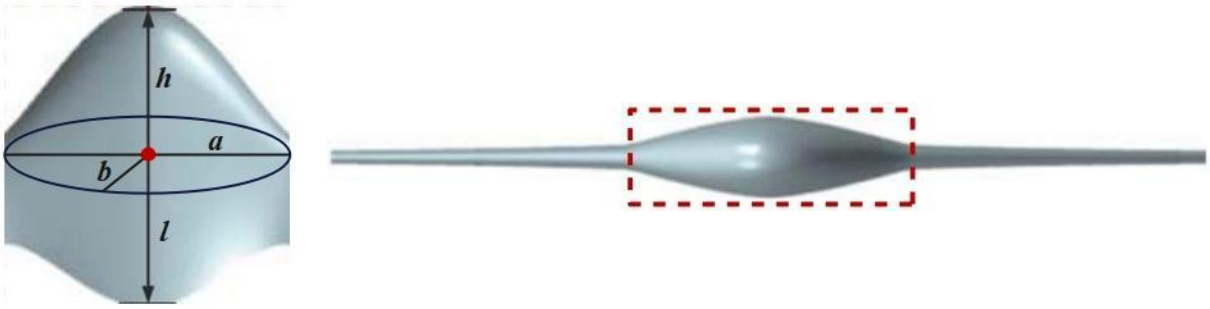


图 5-2 主体结构示意图

图 5-2 中涉及的各项参数具体如表 5-1 所示：

表 5-1 飞行器主体外形参数

名称	符号	名称	符号
前部高	h	椭圆截面长轴	a
后部高	l	椭圆截面短轴	b

在假设飞行器外形曲线和横截面分别为二次抛物线和椭圆面后，可分别计算飞行器主体前后部的表面积和体积。

记飞行器前部的表面积和体积分别为S_{front}, V_{front}，后部的表面积和体积分别为S_{back}, V_{back}。计算公式具体如式(5-1)至(5-4)所示：

$$S_{front} = \int_0^h 2\pi b\sqrt{\frac{h-y}{h}} - 4(a-b)\sqrt{1-\frac{y}{h}}dy = \frac{2h}{3}(2\pi b + 4a - 4b) \tag{5-1}$$

$$V_{front} = \int_0^h \frac{\pi ab(h-y)}{h}dy = \frac{\pi abh}{2} \tag{5-2}$$

$$S_{back} == \int_0^h 2\pi b\sqrt{\frac{l-y}{h}} - 4(a-b)\sqrt{1-\frac{l}{h}}dy = \frac{2l}{3}(2\pi b + 4a - 4b) \tag{5-3}$$

$$V_{back} = \int_0^l \frac{\pi ab(l-y)}{l}dy = \frac{\pi abl}{2} \tag{5-4}$$

故主体整体的表面积和体积$S_{main\ body}, V_{main\ body}$分别如式(5-5)和(5-6)所示：

$$S_{main\ body} = S_{front} + S_{back} \tag{5-5}$$

$$V_{main\ body} = V_{front} + V_{back} \tag{5-6}$$

5.1.3 飞行器机翼

由前文分析可得，仅由图 5-1 难以判断机翼的真实物理结构，结合题目提供的侧视图及骨架参数化示意图可进一步判断。

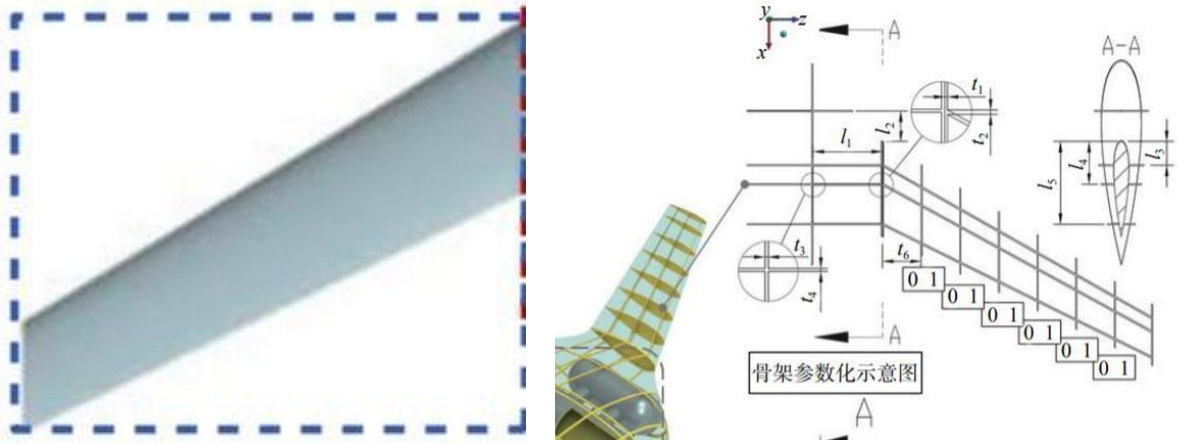


图 5-3 机翼示意图

图 5-3 右图展示了机翼的横截面，涉及的各项参数具体如表 5-2 所示：

表 5-2 飞行器机翼外形参数

名称	符号	名称	符号
前部高	l_3	宽度	w
后部高	$l_5 - l_3$	机翼长度	L

于是，飞行器机翼的前后部弧长L_{front}, L_{back}分别如式(5-7)和(5-8)所示：

$$L_{front} = \int_{-w}^{w} \sqrt{\frac{1 + 4l_3^2 x^2}{w^4}} dx \tag{5-7}$$

$$L_{back} = \int_{-w}^{w} \sqrt{\frac{1 + 4(l_5 - l_3)^2 x^2}{w^4}} dx \tag{5-8}$$

于是，机翼整体表面积计算公式如式(5-9)所示：

$$S_{wing} = (L_{front} + L_{back}) \times L \tag{5-9}$$

同理，飞行器机翼的前后部截面积s_{front}, s_{back}分别如式(5-10)和(5-11)所示：

$$s_{front} = \int_{-w}^{w} (-\frac{l_3}{w^2} x^2 + l_3) dx \tag{5-10}$$

$$s_{back} = \int_{-w}^{w} (-\frac{l_5 - l_3}{w^2} x^2 + l_5 - l_3) dx \tag{5-11}$$

于是，飞行器机翼体积计算如式(5-12)所示：

$$V_{wing} = (s_{front} + s_{back}) \times L \tag{5-12}$$

5.1.4 模型求解

由 5.1.2 和 5.1.3 节分别求解飞行器主体和机翼的表面积和体积后，累加求和，即可得该部分飞行器整体的表面积和体积S_1, V_1，具体如式(5-13)和(5-14)所示：

$$S_1 = S_{main\ body} + S_{wing} \tag{5-13}$$

$$V_1 = V_{main\ body} + V_{wing} \tag{5-14}$$

通过 matlab 求解，结果如表 5-3 所示：

表 5-3 问题一求解结果

	主体	机翼	整体
表面积(cm^2)	8.8219×10^5	4.9830×10^7	5.0712×10^7
体积(cm^3)	8.4823×10^7	2.8800×10^4	8.4852×10^7

由表 5-3 可知，该飞行器整体的表面积和体积分别 $5.0712 \times 10^7 cm^2$和 $8.4852 \times 10^7 cm^3$。

5.2 问题二模型的建立与求解

5.2.1 舱体结构拆分

同问题一类似，问题二要求我们估算飞行器舱体结构的表面积和体积。具体如图 5-4 所示，同理，可将其拆分为半球体和圆柱体两部分。

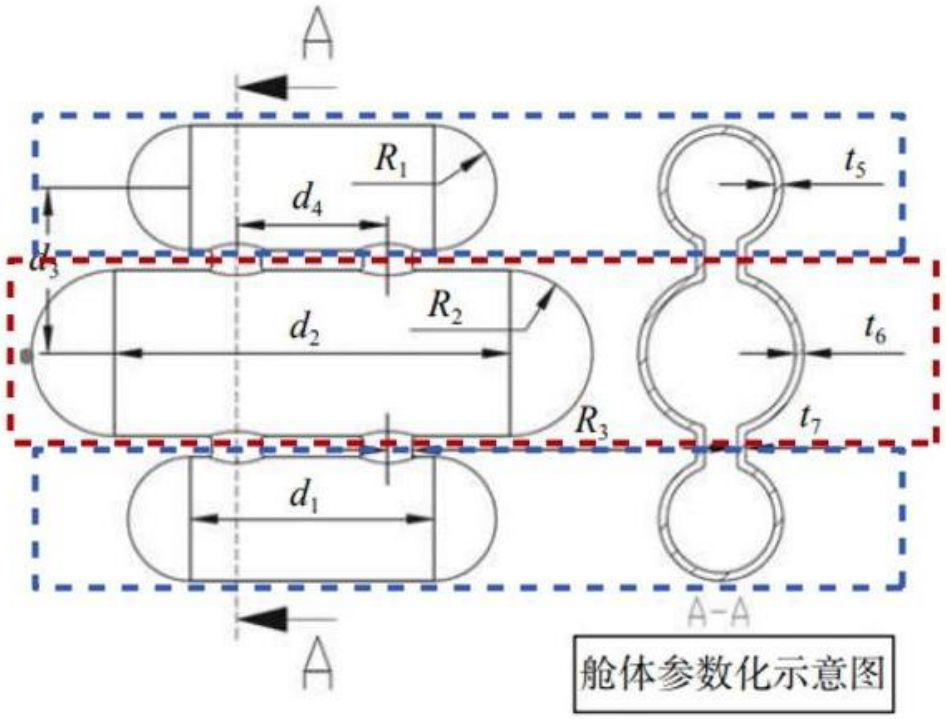


图 5-4 飞行器舱体结构示意图

其中，半球体位于圆柱体两侧，可合并成一完整球体。本文将在 5.2.2 和 5.2.3 节分别探讨半球体和圆柱体模型。

5.2.2 半球体模型

观察图 5-4 可得，对于半球体部分，根据半径R_1, R_2可分为 4 个小半球(R_1)和 2 个大半球(R_2)。

由于无需计算半球体的底面积，因此，对于大小半球表面积计算分别如式(5-15)和(5-16)所示：

$$S_{hemisphere\_large} = 2\pi R_2^2 \tag{5-15}$$

$$S_{hemisphere\_small} = 2\pi R_1^2 \tag{5-16}$$

同理，大小半球体积$V_{hemisphere_large}, V_{hemisphere_small}$计算分别如式(5-17)和(5-18)所示：

$$V_{hemisphere\_large} = \frac{2}{3}\pi R_2^3 \tag{5-17}$$

$$V_{hemisphere\_small} = \frac{2}{3}\pi R_1^3 \tag{5-18}$$

5.2.3 圆柱体模型

分析可得，大小圆柱体的体积$V_{cylinder_large}$,$V_{cylinder_small}$分别如式(5-19)和(5-20)所示：

$$V_{hemisphere\_large} = \pi R_2^2 \times d_2 \tag{5-19}$$

$$V_{hemisphere\_small} = \pi R_1^2 \times d_1 \tag{5-20}$$

值得注意的是，连接体部分同样为圆柱体，共 4 个。

于是，连接体的表面积和体积$S_{connector}$,$V_{connector}$分别如式(5-21)和(5-22)所示：

$$S_{connector} = 2\pi R_3 \times (d_3 - R_1 - R_2) \tag{5-21}$$

$$V_{connector} = \pi R_3^2 \times (d_3 - R_1 - R_2) \tag{5-22}$$

同样不考虑底面积，大小圆柱体的表面积$S_{cylinder_large}$,$S_{cylinder_small}$分别如式(5-23)和(5-24)所示：

$$S_{hemisphere\_large} = 2\pi R_2 d_2 - 4\pi R_3^2 \tag{5-23}$$

$$S_{hemisphere\_small} = 2\pi R_1 d_1 - 2\pi R_3^2 \tag{5-24}$$

5.2.4 模型求解

根据 5.2.2 和 5.2.3 节的分析，本文最终构建飞行器舱体结构的表面积和体积计算模型。飞行器舱体由一个大圆柱体，两个小圆柱体和四个连接体(仍为圆柱体)、两个大半球体和四个小半球体构成。于是，表面积和体积S_2,V_2计算公式如(5-25)和(5-26)所示：

$$S_2 = S_{hemisphere\_large} + 2S_{hemisphere\_small} + 4S_{connector} + S_{hemisphere\_large} + S_{hemisphere\_small} \tag{5-25}$$

$$V_2 = V_{hemisphere\_large} + 2V_{hemisphere\_small} + 4V_{connector} + V_{hemisphere\_large} + V_{hemisphere\_small} \tag{5-26}$$

通过 matlab 求解，结果如表 5-4 所示：

表 5-4 问题二求解结果

名称	表面积(cm^2)	体积(cm^3)
大圆柱体	1.9068×10^5	1.9068×10^6
小圆柱体	1.5346×10^5	7.8540×10^6
连接体	3.1667×10^4	3.8001×10^5
大半球	5.0894×10^4	1.5268×10^6
小半球	6.2832×10^4	2.0944×10^6
舱体	9.7739×10^5	3.7566×10^7

由表 5-4 可知，舱体的表面积和体积分别为 $9.7739 \times 10^5 cm^2$ 和 $3.7566 \times 10^7 cm^3$。

5.3 问题三模型的建立与求解

5.3.1 模型准备

问题三首先给出了该飞行器参数结构的取值范围，及耦合结构其他的参数设置，具体如表 5-5 和表 5-6 所示：

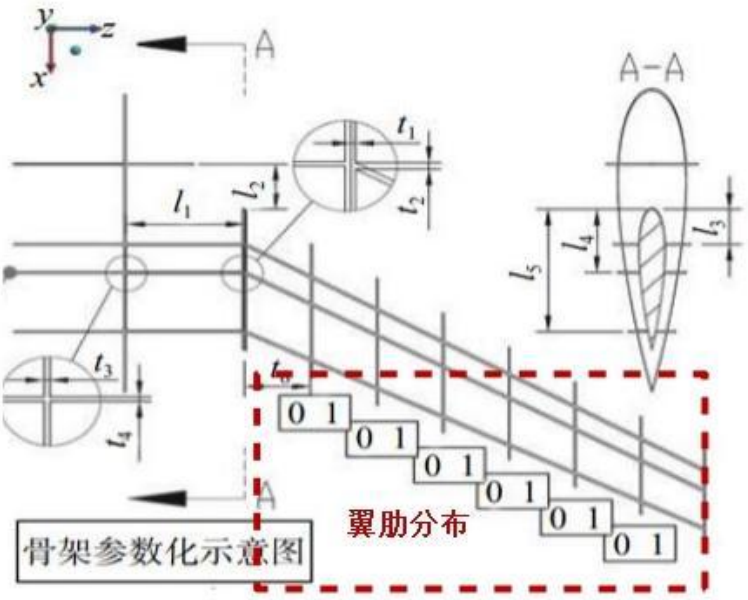
表 5-5 该飞行器参数结构取值范围

设计变量类型	参数	设计下限	设计上限
骨架结构设计变量	$c_{l_6}^i$	0	1
	l_1	270cm	290cm
	l_3	0.1	0.35
	l_4	0.45	0.55
	l_5	0.65	0.9
舱体结构设计变量	R_1	65cm	90cm
	R_2	75cm	100cm
	R_3	20cm	30cm
	t_5	8cm	15cm
	t_6	8cm	15cm
	t_7	8cm	15cm
	G_C	350cm	450cm

表 5-6 耦合结构其他参数设置

参数	数值(固定值)
l_6	143cm
机翼半展长	1000cm
机身半展长	500cm
l_2	120cm
d_1	250cm
d_2	350cm
d_4	150cm

其中，关于机翼半展长，补充描述如下，翼肋平均分布后共计 8 个，使其中间的 6 个翼肋为 0-1 离散变量，定义为$c_{l_6}^i$，其中 1 表示此处布置有翼肋，0 表示未布置翼肋，如图 5-5 所示：



$$c_{l_6}^i = \begin{cases} 1, \text{此处布置有翼肋} \\ 0, \text{此处未布置有翼肋} \end{cases} \quad (i = 1, \ldots, 6)$$

图 5-5 翼肋 0-1 离散变量描述

问题三要求我们设计飞行器的最佳外形，使得其所受阻力最小，因此需要在问题一与问题二求解得到的飞行器部分尺寸和舱体结构的相关数据，进行受力分析，构造飞行器所受阻力与其表面积及体积的函数关系模型，设置为目标函数，以飞行器的物理模型和表 5-5 及 5-6 的参数结构取值范围等为约束构建优化模型。

5.3.2 飞行器空气动力学模型

在实际飞行过程中，飞行器的飞行效果往往受到风力的影响。在风力较大的情况下，飞行器往往需要保持特殊的姿态才可以正常进行悬停或飞行等操作，因此必须作出如下假设：

(1)假设将飞行器整体视作刚体，即忽略弹性形变对飞行器受力等情况的影响。

(2)假设飞行器的质量分布均匀，质心位于原点。

(3)忽略地球自转与公转对飞机飞行的影响。

(4)不考虑风等空气流动队飞行器受力的影响。

为简化模型，忽略姿态角的变化，考虑使用流体阻力计算其空气阻力，此处涉及空气动力学的相关知识，具体如式(5-27)所示：

$$F_d = \frac{1}{2} \times c_d \times \rho v^2 \times S \tag{5-27}$$

其中涉及的相关符号及参数表示分别如表 5-7 所示：

表 5-7 空气动力学(空气阻力)相关符号

符号	描述	单位
F_d	空气阻力	N
c_d	物体阻力系数	/
ρ	空气密度	$\mathrm{kg/m^3}$
v	飞行器相对于空气的速度	$\mathrm{m/s}$
S	飞行器在空气中的横截面积	$\mathrm{m^2}$

*物体阻力系数c_d与物体的几何形状密切相关。在实际计算过程中，应充分考虑飞行器的几何形状、表面粗糙程度等影响因素。综合考虑本文涉及的飞行器主体及机翼的几何形状，选择飞行器的空气阻力系数为 0.08。

5.3.3 飞行器结构参数设计模型

• 目标函数

由题设得，以飞行器所受阻力最小为优化目标，因此设置目标函数如式(5-28)所示：

$$\min F_d = \frac{1}{2} \times c_d \times \rho v^2 \times S \tag{5-28}$$

• 约束条件

由 5.1 和 5.2 节可知，飞行器的主体和机翼整体的体积和表面积，舱体的体积和表面积分别如式(5-13)、(5-14)、(5-25)和(5-26）所示，即：

$$\begin{cases} S_1 = S_{main\ body} + S_{wing} \\ V_1 = V_{main\ body} + V_{wing} \\ S_2 = S_{hemisphere\_large} + 2S_{hemisphere\_small} + 4S_{connector} + S_{hemisphere\_large} + S_{hemisphere\_small} \\ V_2 = V_{hemisphere\_large} + 2V_{hemisphere\_small} + 4V_{connector} + V_{hemisphere\_large} + V_{hemisphere\_small} \end{cases} \tag{5-29}$$

由表 5-5 和 5-6 可得飞行器参数结构取值范围约束，其中，骨架结构设计变量约束具体如式(5-30)所示：

$$\begin{cases} c_{l_6}^{i} = \{0,1\} \\ l_1 \in (270,290) \\ l_3 \in (0.1,0.35) \\ l_4 \in (0.45,0.55) \\ l_5 \in (0.65,0.9) \end{cases} \tag{5-30}$$

其中，$c_{l_6}^{i}$为描述翼肋布置的 0-1 离散变量，l_1, l_3, l_4, l_5为实数变量，分别表示骨架结构的长度比例。

同理，舱体结构设计变量约束具体如式(5-31)所示：

$$\begin{cases} R_1 \in (65,90) \\ R_2 \in (75,100) \\ R_3 \in (20,30) \\ t_5 \in (8,15) \\ t_6 \in (8,15) \\ t_7 \in (8,15) \\ G_c \in (350,450) \end{cases} \tag{5-31}$$

其中，R_1, R_2, R_3是描述舱体结构半径的变量，t_5, t_6, t_7是描述舱体结构壁厚的变量，G_c为描述舱体结构长度的变量。

至此，本文构建了基于最小空气阻力的飞行器外形设计模型，如式(5-32)所示：

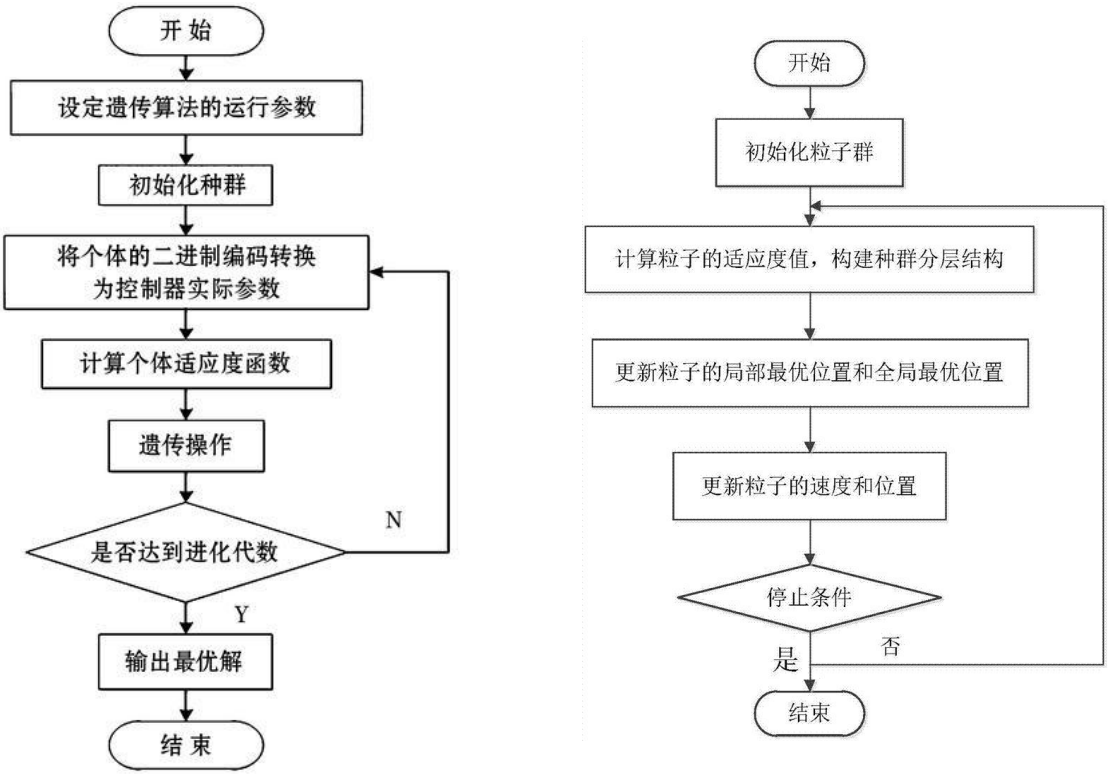
$$\min F_d = \frac{1}{2} \times c_d \times \rho v^2 \times S$$

$$\begin{cases} \begin{cases} S_1 = S_{main\ body} + S_{wing} \\ V_1 = V_{main\ body} + V_{wing} \end{cases} \\ \begin{cases} S_2 = S_{hemisphere\_large} + 2S_{hemisphere\_small} + 4S_{connector} + S_{hemisphere\_large} + S_{hemisphere\_small} \\ V_2 = V_{hemisphere\_large} + 2V_{hemisphere\_small} + 4V_{connector} + V_{hemisphere\_large} + V_{hemisphere\_small} \end{cases} \\ \begin{cases} c_{l_6}^{i} = \{0,1\} \\ l_1 \in (270,290) \\ l_3 \in (0.1,0.35) \\ l_4 \in (0.45,0.55) \\ l_5 \in (0.65,0.9) \end{cases} \\ \begin{cases} R_1 \in (65,90) \\ R_2 \in (75,100) \\ R_3 \in (20,30) \\ t_5 \in (8,15) \\ t_6 \in (8,15) \\ t_7 \in (8,15) \\ G_c \in (350,450) \end{cases} \end{cases} \tag{5-32}$$

5.3.4 模型求解与分析

针对 5.3.3 节构建的于最小空气阻力的飞行器外形设计模型，分贝考虑使用遗传算法和粒子群算法进行求解。

遗传算法与粒子群算法的算法原理分别如图 5-6 所示：



(a)遗传算法　　(b)粒子群算法

图 5-6 问题三求解算法原理

在正式求解之前，需要设置部分题目未补充的参数：

表 5-8 问题三初始参数设置

变量	数值
空气密度	1.225kg/m^3
飞行器飞行速度	250m/s
空气阻力系数	0.08

多次运行取最优值，遗传算法求解结果具体如表 5-9 所示：

表 5-9 问题三求解结果(遗传算法)

变量	l_3	l_4	l_5	R_1	R_2
数值	0.29039486	0.52943786	0.72904604	77.89533426	75
变量	R_3	t_5	t_6	t_7	G_c
数值	20.00000075	8.74795829	13.64077802	10.25186972	399.94922192
最小阻力			54118842.19660542		

粒子群算法求解结果具体如表 5-10 所示：

表 5-10 问题三求解结果(粒子群算法)

变量	l_3	l_4	l_5	R_1	R_2
数值	0.1676173	0.47414186	0.88610813	73.05581391	75.06242333
变量	R_3	t_5	t_6	t_7	G_c
数值	23.2333983	12.17986537	12.12635147	11.45507542	438.47120911
最小阻力			54208967.10751903		

对比可得，遗传算法求解的最小阻力略小于粒子群算法的求解结果，因此本文将在问题四种继续沿用遗传算法进行求解。

5.4 问题四模型的建立与求解

5.4.1 模型准备

问题四要求我们在问题三的基础上，分别考虑圆形、椭圆、抛物线和双曲线等四种圆锥曲线作为飞行器的外形，重新求解问题三飞行器的最佳外形设计问题。

四种圆锥曲线的数学表达式分别如式(5-33)-(5-36)所示：

$$x^2 + y^2 = r^2 \tag{5-33}$$

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{5-34}$$

$$y = ax^2 + bx + c \tag{5-35}$$

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \tag{5-36}$$

其对应的表面积和体积计算公式如表 5-11 所示：

表 5-11 圆锥曲线表面积与体积计算公式

圆锥曲线	表面积	体积
圆形	$S = 4\pi r^2$	$V = \frac{4}{3}\pi r^3$
椭圆	$S = \frac{4}{3}\pi ab$	$V = \frac{4}{3}\pi abc$
抛物线	$S = \frac{2h}{3}(2\pi b + 4a - 4b)$	$V = \frac{4}{3}\pi ab^3$
双曲面	$S = \frac{4}{3}\pi ab$	$V = \frac{4}{3}\pi abc$

5.4.2 飞行器外形设计模型的修改

分别针对上述四种圆锥曲线进行问题三模型的相应修改，目标函数不变。以圆形和椭圆形为例，约束条件的变化具体如表 5-12 所示：

表 5-12 圆锥曲线约束条件的修改情况

(1)圆形	(2)椭圆
$\begin{cases} S=4\pi r^2 \\ V=\frac{4}{3}\pi r^3 \\ L_{front}=\int_{-w}^{w}\sqrt{\frac{1+4l_3^2x^2}{w^4}}dx \\ L_{back}=\int_{-w}^{w}\sqrt{\frac{1+4(l_5-l_3)^2x^2}{w^4}}dx \\ s_{front}=\int_{-w}^{w}(-\frac{l_3}{w^2}x^2+l_3)dx \\ s_{back}=\int_{-w}^{w}(-\frac{l_5-l_3}{w^2}x^2+l_5-l_3)dx \\ \begin{cases} c_{l_6}^i=\{0,1\} \\ l_1\in(270,290) \\ l_3\in(0.1,0.35) \\ l_4\in(0.45,0.55) \\ l_5\in(0.65,0.9) \end{cases} \\ \begin{cases} R_1\in(65,90) \\ R_2\in(75,100) \\ R_3\in(20,30) \\ t_5\in(8,15) \\ t_6\in(8,15) \\ t_7\in(8,15) \\ G_c\in(350,450) \end{cases} \end{cases}$	$\begin{cases} S=\frac{4}{3}\pi ab \\ V=\frac{4}{3}\pi abc \\ L_{front}=\int_{-w}^{w}\sqrt{\frac{1+4l_3^2x^2}{w^4}}dx \\ L_{back}=\int_{-w}^{w}\sqrt{\frac{1+4(l_5-l_3)^2x^2}{w^4}}dx \\ s_{front}=\int_{-w}^{w}(-\frac{l_3}{w^2}x^2+l_3)dx \\ s_{back}=\int_{-w}^{w}(-\frac{l_5-l_3}{w^2}x^2+l_5-l_3)dx \\ \begin{cases} c_{l_6}^i=\{0,1\} \\ l_1\in(270,290) \\ l_3\in(0.1,0.35) \\ l_4\in(0.45,0.55) \\ l_5\in(0.65,0.9) \end{cases} \\ \begin{cases} R_1\in(65,90) \\ R_2\in(75,100) \\ R_3\in(20,30) \\ t_5\in(8,15) \\ t_6\in(8,15) \\ t_7\in(8,15) \\ G_c\in(350,450) \end{cases} \end{cases}$

同理可得抛物线和双曲面的约束条件修改情况，根据问题三求解结果对比情况，选择采用遗传算法分别求解，具体如表 5-13 至 5-16 所示：

表 5-13 问题四结果(圆形)

变量	l_3	l_4	l_5	R_1	R_2
数值	0.29665119	0.45363407	0.66506088	71.02304087	75
变量	R_3	t_5	t_6	t_7	G_c
数值	20.01684394	8.91071452	14.67916857	10.98125362	400
最小阻力			54118952.959026754		

表 5-14 问题四结果(椭圆)

变量	l_3	l_4	l_5	R_1	R_2
数值	0.25778851	0.52188039	0.89453149	89.7818264	80
变量	R_3	t_5	t_6	t_7	G_c

数值	20	14.21907487	12.84702446	10.61729238	391.70384814
最小阻力			61575216.01035995		

表 5-15 问题四结果(抛物线)

变量	l_3	l_4	l_5	R_1	R_2
数值	0.16676257	0.50232208	0.79027447	85.22452258	80
变量	R_3	t_5	t_6	t_7	G_c
数值	20	12.58762584	13.88673419	12.9939267	371.2670316
最小阻力			61575216.01035995		

表 5-16 问题四结果(双曲面)

变量	l_3	l_4	l_5	R_1	R_2
数值	0.3375851	0.47902355	0.66612663	75.7253441	75.00007675
变量	R_3	t_5	t_6	t_7	G_c
数值	20	10.01777504	12.94344705	10.06627302	355.25455845
最小阻力			54118842.19660542		

由表 5-13 至 5-16 可得，在四种圆锥曲线中，以双曲线作为飞行器曲线所受阻力最低，为 54118952.96N。参数矩阵为[0.3375851,0.47902355,...,355.25455845]

六、模型的评价与推广

6.1 模型的优点

(1)本文通过对飞行器外形的细致拆分和参数化建模，能够全面分析飞行器的表面积、体积及最小化空气阻力的关系。这种综合性的方法使得设计能够在多个方面进行优化，包括减少阻力、提高效率等。通过对比遗传算法、粒子群算法的求解结果，本文确定遗传算法在优化飞行器外形设计中表现较好。这种多算法比较能够确保最终选取的设计方案更为优化和合理。

(2)本文在模型构建中采用了二次抛物线、椭圆面等简化的几何曲线，以及将飞行器拆分为主体和机翼部分的方法，简化了复杂问题的处理过程。这种简化提高了模型的实用性，使得设计结果更易于理解和应用于实际工程中。

(3)通过基于空气动力学的物理模型和大量计算分析，本文不仅探索了最小化阻力的理论极限，还为实际飞行器的设计和改进提供了深入的科学依据和指导。这种科学性保证了设计方案在实际应用中的可靠性和有效性

6.2 模型的不足与改进

(1)本文在建模过程中忽略了飞行器的姿态角变化对阻力的影响，这在实际飞行中可能会有所偏差。此外，实际飞行器设计中还涉及到诸如材料选择、结构强度、制造成本等复杂因素，这些因素在本文中未被全面考虑。

(2)虽然遗传算法在本文中表现优越，但不同飞行器的外形和设计问题可能需要不同的优化算法或组合，算法的适用性可能受到设计空间复杂性的限制。

6.3 模型的推广

基于本文的研究成果，可推广的方向有：

(1)将优化设计方法推广到不同类型的航空器和航天器上，包括无人机、载人飞船、卫星等，以提升它们的飞行效率和性能。

(2)结合材料科学、结构工程等领域，进一步优化飞行器的整体设计，包括减重和提高强度等方面。

(3)将模型结果与实际飞行器的测试数据进行比较和验证，进一步验证模型的有效性和适用性。

(4)开发基于优化算法和空气动力学模型的智能化设计工具，支持工程师在设计阶段快速生成优化方案。

通过这些推广方向，可以进一步推动飞行器设计技术的进步，促进航空航天领域的科学研究和实际应用。

参考文献

[1] 李昊,廖志刚,杨令飞,等.高速飞行器气动外形设计方法综述[J].空天技术,2024,(01):23-35.DOI:10.16338/j.issn.2097-0714.20230351.

[2] 郑和超.一种仿鸟扑翼飞行机器人的控制器设计研究[D].北方工业大学,2023.DOI:10.26926/d.cnki.gbfgu.2023.000770.

[3] 王健磊,牟桓,魏震,等.宽包线吸气式高超声速飞行器外形优化研究[J].空气动力学学报,2023,41(02):1-11.

[4] 池元成,张冶,郑小鹏,等.飞行器气动外形的正交设计与分析[J].宇航总体技术,2022,6(01):50-54.

[5] 朱朝,李中武,刘峰博.飞行器气动外形优化系统研究[J].航空计算技术,2021,51(06):96-100.

[6] 解静,白鹏,李永远.基于遗传算法的升力体外形优化设计[J].气体物理,2020,5(04):31-36.DOI:10.19527/j.cnki.2096-1642.0814.

[7] 苏晓东.一种 U 型变体无人飞行器气动外形技术研究[J].中国科技信息,2020,(05):38-41.

[8] 许云清.四旋翼飞行器飞行控制研究[D].厦门大学,2014.

[9] 张艳军.高速飞行器空气动力学数值分析[D].中北大学,2007.

[10] 王磊,何国毅,王娜,等.高超声速飞行器的气动力工程计算[J].南昌航空大学学报(自然科学版),2020,34(01):1-6.

[11] 刘沛清,郭知飞.飞行器的升力、阻力及升力与环量定理的起源[J].力学与实践,2019,41(06):739-744.

[12] 刘俊,任杰,赤丰华,等.局部外形参数对高速飞行器的 RCS 影响[J/OL].系统工程与电子技术,1-14[2024-07-07].http://111.229.208.102:8085/kcms/detail/11.2422.tn.20240510.1232.005.html.

选题	2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
B		apmcm24100536

基于关键特征集成学习的洪水灾害数据分析与预测研究

摘要

针对问题 1，首先计算出所给数据季风强度、地形排水等基础指标和洪水概率的最大值、最小值、平均值、中位数等基本统计量，完成数据预处理，随后对各个指标及洪水概率进行可视化分析，并计算上述所有指标分布的偏度与峰度系数，结果可得**训练集和测试集中各指标分布均呈现轻微的正偏态(右偏)和尖峰厚尾，训练集洪水概率分布呈现正态分布**。随后，引入 Spearman 相关系数进相关性分析，并加以假设检验，**发现各指标之间呈现独立性，同时均与洪水概率呈现有层次的正相关关系**，随后取相关性**平均值以上的指标为高相关性指标**，并对其进行相应分析，提出相关建议与措施。

针对问题 2，首先标准化洪水概率数据，利用 **K-Means++聚类算法**将其聚类为高、中、低风险三大类，随后对聚类结果进行可视化分析，基于簇进行数据表分组，并计算每个分组的最大、最小值，验证**数据完整聚类为三簇无交错类别**，计算三组数据表指标统计量，得出**指标数值与风险高低呈现正相关性，且不同指标在同一风险类别中呈现相似的数据分布情况**。评价问题定义为回归+分类，首先基于分组数据中最大、最小值构造**回归数值与分类类别之间的映射区间**，并训练**逻辑回归分类模型**实现**映射区间的缺漏填充，**成功**构建完整的映射区间**，随后，基于原始 20 指标构建**线性回归模型**作为**基准模型，R2 分数 0.84，三分类准确率 0.76**，得出指标与洪水概率之间**存在正线性相关性**，并对所有**指标之和**与洪水概率进行可视化确认其间**呈强正线性相关性**，引入指标和作为关键引导指标，然后将原始指标在**样本维度升序排列**，构造排序指标，以**消除指标维度数据差异性**，最后，将**排序指标与指标和**作为最终指标训练 **CatBoost 回归模型**实现了**优于基准模型的效果，R2 分数 0.87**，并且其在三分类任务上同样表现优秀，**三分类准确率 0.76**，同时，利用 **K 折交叉验证法**验证了模型的**低灵敏度与高泛化性。**

针对问题 3，首先基于问题 1 中各指标与洪水概率相关性选出**相关性大于平均值的指标**，加入**指标和**作为关键指标，训练 CatBoost 回归模型，得出较优回归预测模型。随后，使用 **PCA 主成分分析**进行数据降维，将 20 维指标降至 **4 维关键指标**，同时依旧加入**指标和**组成 5 个指标，用以训练 CatBoost 回归模型，得出**性能远胜基准模型**的最优回归预测模型，**R2 分数 0.97，**同时作为佐证，求出**其三分类准确率 0.91**。

针对问题 4，直接选用问题 3 中得出的最优 CatBoost 回归模型进行预测，填写预测结果，并对结果使用直方图、折线图进行可视化分析，计算其偏度、峰度系数，得出**预测结果服从正态分布**，并与其他性能较低回归模型进行对比，验证其预测结果正态性与合理性。

关键词：CatBoost 回归模型 线性回归模型 逻辑回归分类模型 连续映射区间构造 K-Means++聚类算法 PCA 主成分分析 K 折交叉验证

一、问题重述

1.1 问题背景

洪水作为一种严重的自然灾害，对人类社会和生态环境造成了巨大的影响。历史上，洪水灾害的发生往往与自然因素如暴雨、融冰化雪、风暴潮等有关。然而，随着人口的增长和人类活动的增加，如乱砍滥伐、围湖造田等，这些活动加剧了地表状态的改变，进而影响了汇流条件和洪水灾害的程度。近年来，全球洪水灾害频发，造成了巨大的经济损失和人员伤亡。因此，对洪水灾害进行数据分析与预测，对于提前预警和减少灾害损失具有重要意义。

1.2 问题要求

附件数据给出了各地季风强度、地形排水、河流管理等基本信息，以及该地发生洪水的概率。为了能更好地根据已知信息预测洪水的发生概率，最小化灾害损失，现需结合实际情况与所给信息建立数学模型，分析以下问题：

问题 1：依据附件 1 中给出的基本信息，可视化处理相关数据，并单独分析洪水发生概率与每个基本信息的相关性，判断其关联程度大小，给出相关结论和建议。

问题 2：将 train.csv 中洪水发生的概率聚类成不同风险类别，分析高、中、低风险洪水事件的指标特征。选取合适的指标，计算不同指标的权重，建立预警评价模型，并进行模型的灵敏度分析。

问题 3：基于问题 1 中指标分析的结果，建立洪水发生概率的预测模型。从 20 个指标中选取合适指标，预测洪水发生的概率，并验证模型的准确性。探讨如果仅用 5 个关键指标，如何调整改进预测模型。

问题 4：利用问题 2 中建立的洪水发生概率预测模型，预测附件 test.csv 中所有事件的洪水发生概率，并将结果填入 submit.csv。绘制预测概率的直方图和折线图，分析其分布特征，判断是否服从正态分布。

附件：

1. train.csv - 包含超过 100 万条洪水数据，涵盖洪水事件 id 和 20 个指标得分，以及发生洪水的概率。
2. test.csv - 包含超过 70 万条洪水数据，涵盖洪水事件 id 和 20 个指标得分，缺少发生洪水的概率。
3. submit.csv - 包含 test.csv 中的洪水事件 id，缺少发生洪水的概率。

二、问题分析

2.1 问题 1 的分析

针对问题 1，首先对 train.csv 中所给数据进行最大值、最小值、平均值、中位数等基本统计量计算、数据预处理等基础性工作,并记录洪水概率的最大、最小值，基于这些数据，采用概率密度直方图、核密度估计（Kernel Density Estimation, KDE）图可视化洪水概率的分布情况，同时用频数统计直方图可视化 train.csv 与 test.csv 中季风强度、地形排水等各个指标特征的分布情况，并计算上述所有指标分布的偏度与峰度系数用以量化各项评估指标与洪水概率的分布

情况。随后，引入 Spearman 相关系数进相关性分析，并加以假设检验，用以量化各项指标与洪水概率之间的关系。最后，综合以上获得的相关性、数据分布情况等分析结果进行统计分析，指出对洪水概率影响较大的指标，并据此提出相应的预防措施。

2.2 问题 2 的分析

针对问题 2，首先标准化 train.csv 中洪水概率数据，并利用 K-Means++聚类算法将其聚类为三簇，随后对聚类结果进行箱型图、频数分布直方图可视化，基于簇进行数据表分组，并计算每个分组的最大、最小值，用以评估聚类结果是是否存在交错，最后划分风险为高、中、低三类，将其作为分类标签，并对其分别进行统计分析可视化，用以分析不同风险对应指标特征。风险预警评价问题实质是基于连续数值到逻辑类别区间映射的回归+分类的合并问题，首先基于先前得出的三类分组数据中最大、最小值构造连续回归数值转离散类别数值的映射，并训练逻辑回归模型建立符合分布的洪水概率到风险等级的连续映射，实现有空缺映射区间的填充工作，以保证最终回归+分类模型评价能力的鲁棒性，随后，基于所有指标构建线性回归模型作为基准模型，实现较优结果，探索指标特征与洪水概率间潜在的线性相关性，并对所有指标之和与洪水概率进行可视化以确认其线性相关性，因此引入指标和作为关键指标，然后将季风强度、地形排水等指标在样本维度上从小到大排序，构造排序指标，从而消除指标维度上的数据差异性，最后，将排序指标与指标和作为最终指标训练 CatBoost 模型实现了优于基准模型的效果，并且其在三分类评价任务上同样表现优秀，同时，利用 K 折交叉验证法验证了模型的低灵敏度与高泛化性。

2.3 问题 3 的分析

针对问题 3，首先基于问题 1 分析出的各指标与洪水概率的 Spearman 相关系数选出相关性较高的指标特征作为基本指标，并加入指标和作为关键指标，训练 CatBoost 回归模型，得出较优回归预测模型。随后，使用主成分分析（Principle Component Analysis，PCA）进行数据降维，将 20 维指标特征降至 4 维作为关键指标，同时依旧加入指标和组成 5 个指标，用以训练 CatBoost 回归模型，得出最优回归预测模型。

2.4 问题 4 的分析

针对问题 4，直接选用问题 3 中得出的最优 CatBoost 回归模型进行 test.csv 的特征工程，才用最优 5 个关键指标进行结果预测，于 submit.csv 中填写预测结果，并对预测结果使用直方图、折线图（随机采样 1000 样本）进行可视化分析，计算其偏度峰度系数，分析其分布正态性，并与其他两个性能较低的 CatBoost 回归模型进行分布对比，验证其预测结果正态性与合理性。

文章总体思路如图 1 所示：

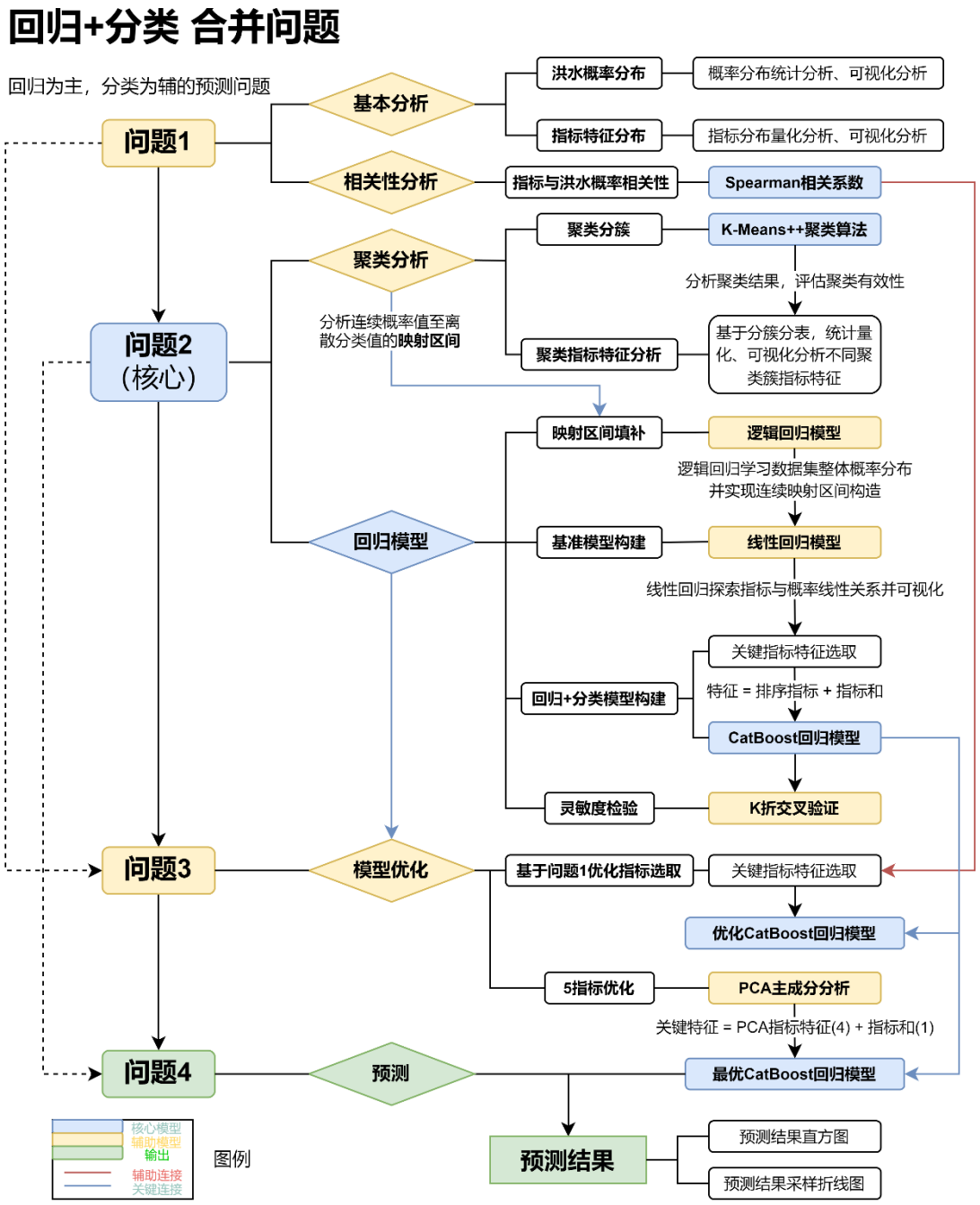


图 1 总体思路图

三、模型假设

根据上述问题分析，我们做出如下假设：

1. 假设洪水发生的概率和影响因素在短期内相对稳定，即考虑的时间范围内没有显著的气候或人为变化。。
2. 假设使用的历史数据或观测数据是准确和可靠的，能够充分反映洪水发生的情况和影响因素的变化。

四、符号说明

符号	说明
K_u	峰度系数
S_u	偏度系数
ρ	Spearman 相关系数
$d(x_i, c_j)$	x_i与c_j之间的欧氏距离
ω_i	权重系数
x_i	样本点 x 的第 i 个分量
L	损失函数
$\hat{y}$	预测值
Org-20	使用最初的 20 个指标
Sorted-20	使用 20 个指标的排序指标
Sum-1	使用 20 个指标的和作为指标
Selected-10	使用精选的 10 个指标
PCA-4	使用 PCA 后得到的 4 个最主要的成分作为指标

五、问题 1 模型的建立与求解

5.1 基于数据的数字特征进行描述

题目要求对数据进行可视化处理，为了更好地描述洪水发生概率分布、基本信息特征的分布规律，本文引入峰度系数、偏度系数来描述统计数据。

（1）峰度系数

峰度系数（Kurtosis）用于衡量数据分布的峰度程度。例如$K_u=0$时，数据的分布为正态分布；$K_u<0$时，数据分布的峰度较小，数据更分散；$K_u>0$时，数据分布的峰度较大，数据更集中。并且$|K_u|$越大，数据分布的图像更加“陡峭”。峰度系数K_u的计算公式如下：

$$K_u = E[(\frac{X-\mu}{\sigma})^4] = (\frac{\mu}{\sigma})^4$$

（2）偏度系数

偏度系数（Skewness）用于衡量数据分布的偏斜程度。例如$S_k=0$时，数据的分布为正态分布；$S_k<0$时，数据呈左偏分布；$S_k>0$时，数据呈右偏分布。并且$|S_k|$越大，数据分布的偏斜越明显。峰度系数S_k的计算公式如下：

$$S_k = E[(\frac{X-\mu}{\sigma})^3] = (\frac{\mu}{\sigma})^3$$

5.2 数据可视化

首先对原始数据进行可视化分析，作出洪水发生概率密度图像与核密度估计图像，并在其左上角展示分布的偏度峰度，图像如下：

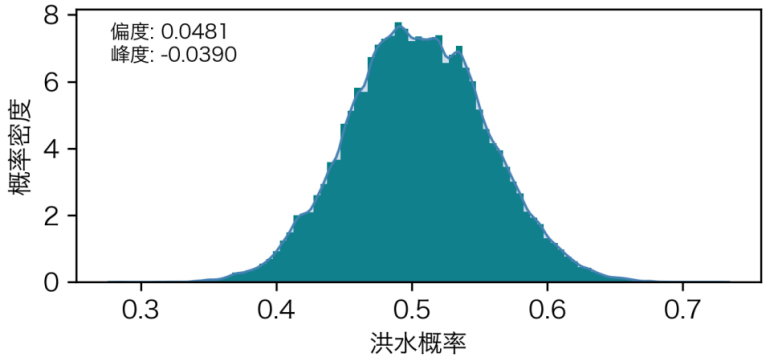


图 2 洪水概率密度图

可见洪水概率分布基本**符合正态分布**。

基本指标的分布直方图如下：

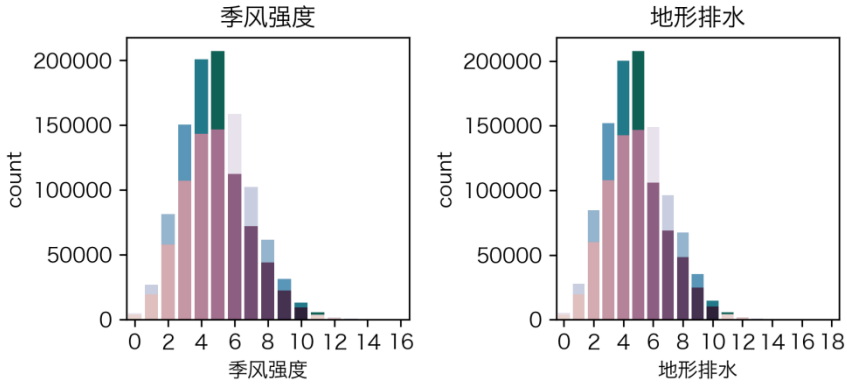


图 3 部分指标柱状图

图中紫红色为测试集分布情况，蓝绿色为训练集分布情况此处仅采样季风强度与地形排水的指标分布情况进行可视化，其余指标的分布情况基本相同，且均呈现出**训练、测试集服从基本相同的分布且训练集数据量比测试集大 40%左右。**

下面对已有已有数据进行定量分析，统计训练集、测试集中数据分布的偏度与峰度系数。

表 1 数据集指标偏、峰度统计量

	训练集偏度系数	训练集峰度系数	测试集偏度系数	测试集峰度系数
平均值	0.441813837	0.247223868	0.44093833	0.247306689
标准差	0.011392705	0.035052238	0.012480602	0.034875173
最小值	0.419867945	0.187350475	0.413857617	0.182297125
25%	0.436629875	0.232680673	0.435625489	0.228729943
50%	0.441398792	0.243670623	0.442318853	0.246637074
75%	0.449001103	0.262154138	0.449952157	0.263493856
最大值	0.464098419	0.339472861	0.460378327	0.315153435

基于上表定量数据对整个数据集的分析如下：

1. 偏度系数：平均偏度系数接近 0.44，表明数据分布呈现轻度的正偏态（右偏），即大部分数据分布在均值左侧，尾部向右延伸。其标准差较小，说明各个指标的偏度相对一致。最小值和 25%分位数显示有些指标的偏度较低，而 75%分位数和最大值则显示少数指标有较高的正偏度。

2. 峰度系数：平均峰度系数在 0.24 到 0.25 之间，略高于 0（正态分布的峰

度系数为 3），表明数据分布相对于正态分布有轻微的尖峰厚尾特性。峰度的标准差相对较大，说明不同指标的峰度差异较大。

3. 分布一致性：训练集和测试集的偏度和峰度系数非常接近，这表明两个数据集在分布形状上具有较高的一致性。

4. 数据范围：最小值和最大值展示了数据分布的整个范围，可以看出指标值在 0.4 到 0.46 之间变化，这可能表明数据的变异性有限。

5. 中位数（50%分位数）：中位数接近平均值，进一步确认了数据分布的正偏态。

6. 四分位数间距（IQR）：通过比较 25%和 75%分位数，可以观察到数据的分散程度。IQR 较小，表明大部分数据点集中在中位数附近。

7. 异常值：由于偏度和峰度系数的值并不极端，可能没有明显的异常值，但最大值和最小值的指标可能需要进一步检查是否有异常值。

整体而言，数据集的分布**呈现轻微的正偏态(右偏)和轻微的尖峰厚尾，但整体分布相对集中。**

5.3 利用 Spearman 相关系数判断指标与洪水概率之间的相关性

Spearman 相关系数是一种非参数的相关性度量，可以等级化变量之间相关性，并且不局限于线性关系，适用于非线性关系的评估，其计算公式如下：

$$\rho = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{[\sum_{i=1}^{N}(R_i - \bar{R})^2 \sum_{i=1}^{N}(S_i - \bar{S})^2]^{\frac{1}{2}}} = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$$

其中，R_i与S_i均为对应观测值第 i 个取值的等级，$[\bar{R}]$与$[\bar{S}]$均为对应观测值取值的平均等级，N 为观测值的总数量，$d_i = R_i - S_i$。

Spearman 相关系数评估了两变量之间的单调关系，即两个变量同时增加（或减少）时相关系数趋近于 1（或-1）；两个变量的变化之间未出现明显关系时趋近于 0。

5.4 Spearman 相关系数的检验

本文对 20 个指标采取 Spearman 相关系数检验，检验步骤如下：

(1)提出假设

原假设$H0$：Spearman 系数$\rho \neq 0$

备择假设$H1$：Spearman 系数$\rho = 0$

设定置信水平为 **99.5%**

(2) 计算 P 值

本文采用 Python 中 statsmodels 中 multipletests 进行 Spearman 相关系数检验，得出所有指标之间以及所有指标与洪水概率之间的修正后 Spearman 相关系数。将结果拆分为两个表，分别展示指标之间以及指标与洪水概率之间的 Spearman 相关系数：

指标之间 Spearman 相关系数表如下（由于大小限制详情见附表）：

表 2 Spearman 相关系数检验的 P 值

	季风强度	地形排水	河流管理	...	政策因素
季风强度	1.0000	0.0000	0.0000	...	0.0000
地形排水	0.0000	1.0000	0.0000	...	0.0000
河流管理	0.0000	0.0000	1.0000	...	0.0000

⋮	⋮	⋮	⋮	⋮	⋮
政策因素	0.0000	0.0000	0.0000	0.0000	1.0000

表 2 是 Spearman 相关系数检验的 P 值构成的矩阵，其对角线元素为 1，其余元素均为 0，即其为单位矩阵，可知任意两指标之间的 P 值均为 0，故接受原假设，认为 Spearman 系数$\rho \neq 0$。

将其可视化为热力图如下：

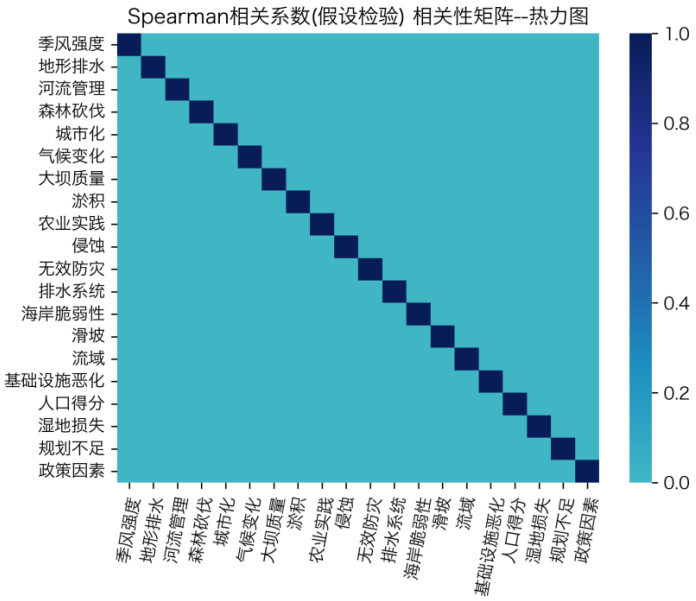


图 4 Spearman 相关系数（假设检验）相关性矩阵

从上图可直观看出指标数据之间呈现出**完全独立性**。

假设检验之后指标与洪水概率之间 Spearman 相关系数：

由上表可知：

表 3 Spearman 相关系数

季风强度	地形排水	河流管理	森林砍伐	城市化
0.18028	0.18048	0.17893	0.17701	0.17265
气候变化	大坝质量	淤积	农业实践	侵蚀
0.17712	0.17946	0.17874	0.17514	0.17143
无效防灾	排水系统	海岸脆弱性	滑坡	流域
0.17597	0.17044	0.17029	0.17664	0.17464
基础设施恶化	人口得分	湿地损失	规划不足	政策因素
0.18140	0.17799	0.17514	0.17325	0.17387

1. 基础设施恶化对洪水的影响最大，海岸脆弱性对洪水的影响最小。
2. 20 个指标与洪水概率的 Spearman 相关系数都在 0.175 左右，说明 20 个指标均与洪水发生概率呈正相关，但是相关性均较弱。

通过计算统计量与洪水概率的 Spearman 相关系数，我们得到各统计量与洪水发生概率的相关性如下：

表 4 统计量与洪水发生概率的 Spearman 相关系数			
统计量	平均值	最小值	最大值
ρ	0.176044	0.170289	0.181399
统计量	25%	50%	75%
ρ	0.173718	0.176304	0.178786

由上表可知，平均值作为统计量，与洪水发生概率的相关性居中，因此选择平均值作为划分指标来进行模型优化。因此选取出检验后 Spearman 相关系数高于平均的指标如下共 10 个：

表 5 精选指标				
基础设施恶化淤积	地形排水	季风强度	大坝质量	河流管理
	人口得分	气候变化	森林砍伐	滑坡

(1) 可能的原因

强烈的季风和降雨可能导致洪水，森林砍伐和城市化造成水土流失增加了洪水的风险，气候变化导致极端天气事件增多，洪水风险上升，基础设施如排水系统和大坝质量问题，减少了防洪能力

(2) 建议与措施

- **提升监测预警：**加强洪水监测和预警系统。
- **合理规划土地：**避免在洪水易发区进行开发。
- **增加绿化：**提高森林覆盖率，减少水土流失。
- **改善水利设施：**加强河流管理和排水系统建设。
- **公众教育：**提高公众防洪意识和自救能力。
- **制定防洪措施：**包括风险评估和应急计划。
- **跨部门协作：**确保防洪措施得到有效执行。
- **应用科技：**使用现代技术提高洪水预测和管理效率。

六、问题 2 模型的建立与求解

6.1 低，中，高风险聚类

6.1.1.K-Means++聚类过程

K-Means++是 K-Means 聚类算法的一种改进版本，它通过智能地选择初始聚类中心来提高 K-Means 的性能(李航 2022：216-227)。K-Means++的目标是更好地选择初始聚类中心，以减少算法的迭代次数，提高聚类的质量。具体步骤如下：

1. **选择第一个聚类中心：**随机选择一个数据点作为第一个聚类中心。
2. **选择后续聚类中心：**对于每个数据点，计算它与已选择的聚类中心之间的最短距离 D(x)（即与最近的聚类中心之间的距离）。每个数据点作为下一个聚类中心的概率为:

$$\mathrm{p(x)} = \frac{D(x)^2}{\sum_{x \in X} D(x)^2}$$

其中 x 代表任意一个数据点，X 表示数据集。这样保证了距离更远的数据点更有可能被选为下一个聚类中心，以确保新的聚类中心较好地覆盖数据分布。

3. **完成聚类中心的选择：**重复步骤 2 直到选择足够数量的聚类中心（一般为 k 个，我们称将数据分成了 k 个簇）。
4. **计算每个数据点到聚类中心的距离：**对于每个数据点，计算它与每个聚类中心的距离，通常使用欧氏距离或其他距离度量方式。对于数据点与聚类中心的距离由以下公式给出：

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{N}(x_{ik} - c_{jk})^2}$$

其中x_i为第 i 个样本点，c_j为第 j 个聚类中心，x_{ik}是第 i 个样本点的第 k 个分量，c_{jk}是第 j 个聚类中心的第 k 个分量。

5. **分配数据点到最近的聚类中心：**将每个数据点分配到距离它最近的聚类中心所属的簇。
6. **更新聚类中心：**重新计算每个簇的中心，即该簇所有数据点的平均值。将该平均值作为新的聚类中心。即新的中心为：

$$c_j' = \frac{1}{\mathrm{N_{X_j}}} \sum_{x_i \in X_j} x_i$$

其中c_j'为新的聚类中心，其中X_j为以c_j为聚类中心的聚类，$\mathrm{N_{X_j}}$为X_j中数据点的个数。

7. **完成聚类：**重复步骤 4，5，6，直到聚类中心不再改变，或达到预定义的停止条件（比如达到最大迭代次数）。
   最终得到每个数据点所属的簇，即完成了 K-Means++聚类。
   使用 Python 进行求解，再用箱型图和直方图进行可视化，得到以下结果：

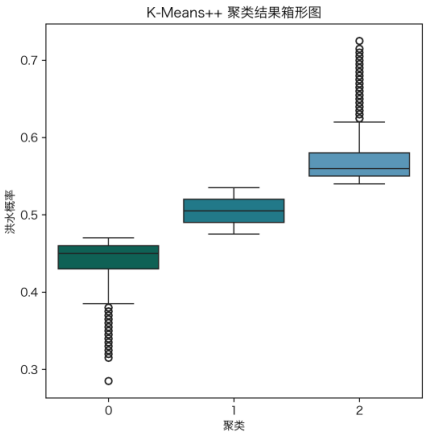


图 5 K-Means++聚类结果（箱型图）

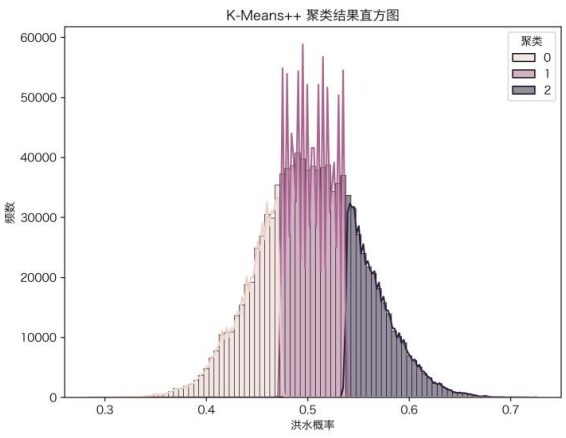


图 6 K-Means++聚类结果（直方图）

从图中不难看出，低，中，高风险的洪水概率是**基本无交错的且大致服从正态分布**，但仍旧需要精确的数据来证明聚类结果是有效的。

基于聚类结果对总表进行分表，并合并统计区间分布情况如下：

表 6 低，中，高风险洪水概率分布情况

	样本数	平均值	最小值	25%	75%	最大值
低风险	282202	0.4423	0.285	0.43	0.46	0.47
中风险	492938	0.5046	0.475	0.49	0.52	0.535
高风险	273435	0.5684	0.54	0.55	0.58	0.725

从表中可见，不同风险类别情况下，三个风险等级对应已有区间分别为[0.285,0.47]，[0.475,0.535]，[0.54,0.725]，因此我们**判断 K-Means++完整将数据聚类到了三个无交错簇中。**

至此，我们将所有 train.csv 中所有数据根据洪水发生的概率划分成了低，中，高风险三个类别。

但是对于风险评级的划分，我们希望得到[0,1]区间上连续的三个子区间，而事实上得到的三个区间分别为[0.285,0.47]，[0.475,0.535]，[0.54,0.725]，显然这些区间是存在空缺的，为了**填补空缺**实现一个完整的回归区间划分使其较好地满足回归+三分类模型需求，下面基于对 train.csv 中洪水概率的统计分析，使用**逻辑回归对其进行区间填补。**

6. 1. 2 逻辑回归填补区间

逻辑回归,即对数概率回归,是一种线性模型,是一种用于分类问题的算法(李航 2022：77-93)。通过学习特征与类别之间的关系，逻辑回归可以预测新数据点属于哪个类别，输出类别的概率，其过程如下：

1. **收集数据**：收集具有标签的训练数据集，每个数据点包括特征值和类别标签。
2. **特征工程**：对数据进行特征提取和选择，包括数据清洗、特征缩放、特征选择等操作。
3. **定义预测函数**：假设我们的模型可以用如下形式表达：

$$h_\theta(x) = g(\theta^T x)$$

其中[g(z)]是 Sigmoid 函数，定义为：

$$g(z) = \frac{1}{1+e^{-z}}$$

[g(z)]的导数满足$[g'(z) = g(z)(1-g(z))]$。

则最终预测函数表达式为：

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

其中$\theta^T x = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$，$\theta_{\mathrm{i}}(\mathrm{i} = 1,2,3,\ldots,\mathrm{n})$为$x_{\mathrm{i}}$对应权重,需要后续过程计算得出。

最后样本为正样本的概率为$h_\theta(x)$，样本为负样本的概率为$1 - h_\theta(x)$。

4. **定义损失函数**：我们使用交叉熵损失函数来度量模型预测值与真实标签之间的误差（由于篇幅原因，我们不给出推导过程）：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[\,y^{(i)}\,log(\,h_\theta(x^{(i)})) + (1-y^{(i)})\,log(\,1-h_\theta(x^{(i)}))]$$

5. **最小化损失函数**：通过梯度下降等优化算法，最小化损失函数，得到最佳的参数θ。
6. **预测**：当模型训练完成后，我们可以用训练好的参数对新数据进行分类预测。分类规则为：如果正样本的概率大于负样本的概率，则样本被判定为正样本，否则被判定为负样本，这等价于

$$\frac{h(x)}{1-h(x)} > 1$$

首先将训练集给出的洪水概率作为特征，将聚类出的类别作为分类标签，训练**逻辑回归分类器**，原始的区间[0.2825, 0.7275]微分切片为 1048575 份，为了提升精确度，将其提升 100 倍，即 100×1048575 份，生成一个相对连续的概率区间划分序列，随后将该序列用先前训练出的逻辑回归分类模型进行分类，成功找到三个基本连续的区间[0.282500, 0.472481], [0.472481, 0.537549], [0.537549, 0.727500。因此，得出区间划分的结论为：

$$\begin{cases} \text{低风险区间：}[\mathbf{0}, \mathbf{0.472481}) \\ \text{中风险区间：}[\mathbf{0.472481}, \mathbf{0.537549}) \\ \text{高风险区间：}[\mathbf{0.537549}, \mathbf{1}] \end{cases}$$

至此，**连续的回归数值到离散的类别数值的映射区间构造完成。**

6.3 分析具有高、中、低风险的洪水事件的指标特征

对高、中、低风险的洪水事件进行统计分析以及可视化，得出如下结果（以低风险为例，其余见附表）：

表 7 低风险区的各指标分布情况

	季风强度	地形排水	河流管理	森林砍伐	城市化
平均值	4.489	4.474	4.518	4.514	4.520
最小值	0	0	0	0	0
最大值	16	18	16	16	16
	气候变化	**大坝质量**	**淤积**	**农业实践**	**侵蚀**
平均值	4.511	4.517	4.491	4.518	4.517
最小值	0	0	0	0	0
最大值	17	16	16	16	17
	无效防灾	**排水系统**	**海岸脆弱性**	**滑坡**	**流域**
平均值	4.503	4.532	4.535	4.496	4.491
最小值	0	0	0	0	0
最大值	16	17	17	16	16
	基础设施恶化	**人口得分**	**湿地损失**	**规划不足**	**政策因素**
平均值	4.492	4.490	4.523	4.512	4.513
最小值	0	0	0	0	0
最大值	16	16	16	16	16

从表中可见，低风险情况下的各大指标平均数值差异不大，而该情况在其他两种风险类别中也及其相似，可得出相应结论，即**同类风险程度下各项指标数值分布极其相似**。

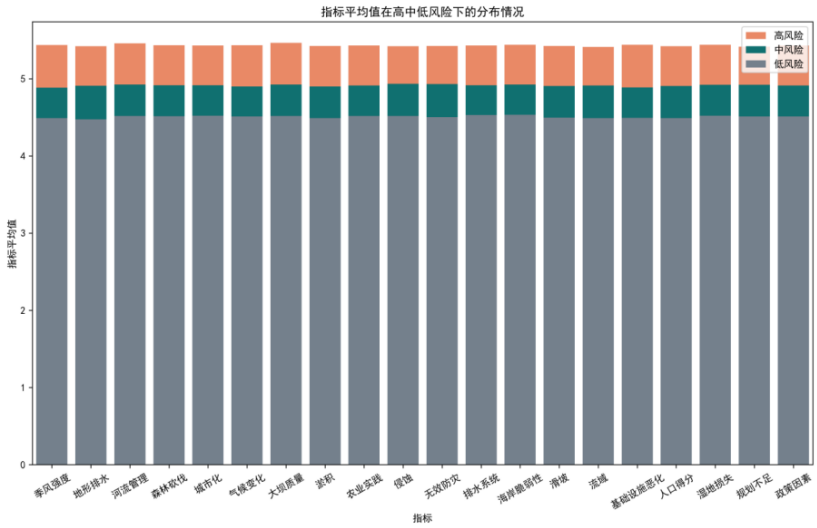


图 7 指标平均值在高中低风险下单分布情况

上图可视化结果印证了同类风险程度下各项指标数值分布极其相似的数据分析结论，接下来我们通过线性回归建立预警评价模型。

6.4 建立发生洪水不同风险的预警评价模型

6.4.1 选取评价指标

为了评估我们最后得到的回归+分类合并模型，我们先引入 **R2 分数**和**三分类准确率**。

- **R2 分数：**也称为确定系数（Coefficient of Determination），是衡量回归模型拟合优度的一个常用指标,其计算公式如下：

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

它直观地反映了模型预测能力的强弱，并且可以用来解释模型对数据的解释程度（值越大，说明回归效果越好）。

- **三分类准确率(ACC_3):**三分类准确率是指在一个有三个不同类别的分类问题中，模型正确预测的样本比例。通常情况下，三分类准确率可以被定义为模型对所有三个类别中正确预测的样本数目的比例,即：

$$三分类准确率 = \frac{预测正确的样本数}{总样本数}$$

这里，“预测正确的样本数”是指模型在所有预测中正确预测的样本数目，总样本数则是所有样本的数量。显然，值越大，也说明回归效果越好。（后简称 ACC_3）

6.4.2 构建基准模型：线性回归

线性回归是一种用于描述自变量与因变量之间线性关系的统计学方法。其基本思想是通过拟合一条直线（或超平面）来最好地拟合数据点，从而预测因变量的取值。由于该模型原理较为简单，我们选择将其作为基准模型，并借其性质来进一步探索指标数据与洪水概率的深层次关系。

下面详细介绍线性回归的过程和数学原理：

1. **线性关系建模**：线性回归假设因变量 y 与自变量 x 之间存在线性关系，可以用一条直线或超平面表示。数学表达式为

$$y = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n = \omega^{\mathrm{T}}\mathrm{x}$$

其中$x_1, x_2, \ldots, x_n$为自变量，$\omega_0, \omega_1, \ldots, \omega_{\mathrm{n}}$为对应权重,$\omega=(\omega_0, \omega_1, \ldots, \omega_{\mathrm{n}})$称之为权重向量,x=$(1, x_1, x_2, \ldots, x_n)$称之为特征向量。

2. **最小化损失函数**：通常将均方误差（Mean Squared Error，MSE）作为损失函数，即：

$$L = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \omega^{\mathrm{T}}x_{\mathrm{i}})^2$$

其中y_i为真实值，$\hat{y}_i$为预测值。

线性回归的目标是最小化损失函数，即通过最小化实际值与预测值之间的误差平方和来得到最优参数。采用最小二乘法得到最优参数

$$\omega = (X^T X)^{-1} X^T y$$

其中X为一m ×(n+1)的矩阵，称之为特征矩阵（每一行均为不同样本点的特征向量的矩阵,故 n 指标数，m 为样本数目），y=$(y_1, \mathrm{y}_2, \ldots, y_m)$是由对应样本的应变量构成的向量。

直接利用所有指标进行线性回归模型训练，通过该模型得到各个指标特征的权重如下：

表 8 各指标权重（$\omega_0 = -0.0534$）

指标	季风强度	地形排水	河流管理	森林砍伐	城市化
权重	0.0056	0.0056	0.0057	0.0057	0.0057
指标	气候变化	大坝质量	淤积	农业实践	侵蚀
权重	0.0057	0.0057	0.0056	0.0056	0.0056
指标	无效防灾	排水系统	海岸脆弱性	滑坡	流域
权重	0.0056	0.0056	0.0057	0.0056	0.0056
指标	基础设施恶化	人口得分	湿地损失	规划不足	政策因素
权重	0.0056	0.0057	0.0056	0.0056	0.0056

可见，各指标权重差距并不大，因此需要对模型性能进行进一步评估

线性回归模型得出的基准结果如下：

表 9 线性回归模型（Org-20）

R2 分数	**ACC_3**
0.84444	0.76743

其中 Org-20 表示选用原始 20 指标。

可见，基本的线性回归模型便可以获得较为优秀的结果，由此**推断指标数据与洪水概率之间的线性相关性是较高**的，至此我们得到了**洪水概率预测的基准模型**，接下来我们选取关键指标用以进一步建立预警评价模型。

6.4.3 选取关键指标

在上述线性回归模型可以发现，指标数据未经特殊处理的情况下依旧与洪水概率有着较强的线性相关性，因此，为了构造关键引导变量，将所有指标之和求出，并与其相应洪水概率可视化如下：

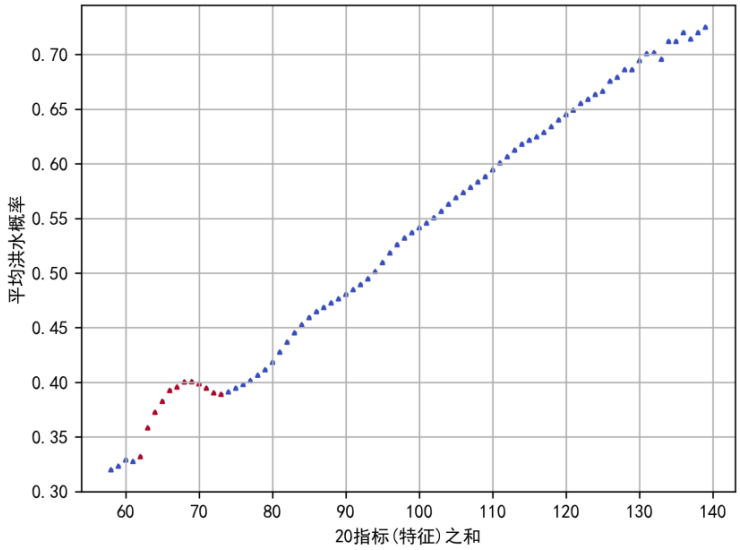


图 8 指标和与平均洪水概率描述

可以观察到，洪水发生概率与指标和基本呈正相关。因此我们选用指标和作为关键指标，引导模型训练。同时，观察到线性关系中存在特殊值（红色部分）。为了保证模型的低灵敏度（强泛化性、强鲁棒性），将特殊值作为数据噪声放入指标考虑范围内。

同时，由于指标维度上数据参差不齐，为了减少在特定维度上的指标数值差异，在样本维度上对所有指标进行降序排序以减少指标维度上的差异性，构造排序指标。

在这一步骤，我们通过对数据的可视化探索选取**指标和**作为一个关键指标，并且将季风强度、地形排水等指标在样本维度上从小到大排序，构造**排序指标**，作为另一个关键指标。

6.4.4 CatBoost 进一步建立模型

为了更好地介绍 CatBoost 之前我们先简单地讲解几个概念：

- **决策树：**决策树是一个预测模型，它代表的是对象属性与对象值之间的一种映射关系(李航 2022：57-75)。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。如图所示就是一个决策树。

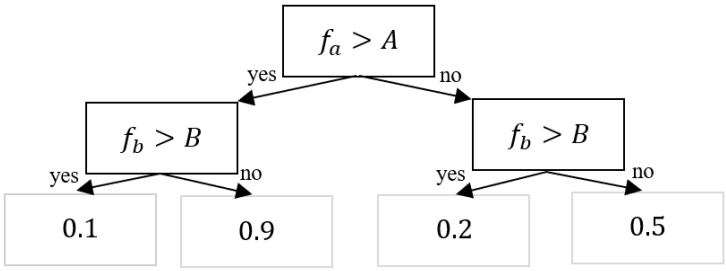


图 9 朴素决策树示意图

决策树的结果可以表示为：

$$\hat{y}(x) = \sum_{m=1}^{M} C_m \, I(x \in R_m)$$

其中$\hat{y}(x)$是输入样本 x 的预测值，M 为决策树的叶节点数量，R_m是决策树第 m 个叶节点的区域，C_m是第 m 个叶节点的类别输出，I是指示函数，表示如果 x 属于R_m则为 1，否则为 0。

- **集成学习方法：**集成学习方法是一种机器学习策略，旨在通过组合多个学习器的预测结果来改善整体的预测性能(李航 2022：131-146)。它可以通过结合多个模型的优点，来提高泛化能力和预测准确性，特别是在处理复杂问题或数据集中存在噪声的情况下表现良好。常见的集成学习方法有 Bagging（Bootstrap Aggregating）， Boosting，Stacking 等等，后文使用的 CatBoost 就是一种通过串行训练多个弱学习器，每个学习器专注于修正前序学习器的错误，最终构建一个强大的集成模型的集成学习方法。
- **梯度提升树：**梯度提升树(李航 2022：131-146)是一种强大的集成学习方法，通过结合多棵决策树来提升预测性能。它基于逐步优化损失函数的梯度，每一棵树都专注于拟合前一棵树的残差，最终通过组合多个模型来减少预测误差，广泛应用于回归和分类任务中，特别适合处理复杂的非线性关系和高维数据。

根据以上概念我们可以得到 CatBoost 的基本原理：

CatBoost [1]是一种基于对称决策树（oblivious trees）为基学习器（弱学习器）实现的参数较少、支持类别型变量和高准确性的 GBDT（Gradient Boosting Decision Tree）框架。其过程如下：

1. **定义模型：**CatBoost 回归模型可以表示为一个集成模型，由多个决策树组成。假设我们有 K 棵树，模型可以表示为：

$$F(x) = \sum_{k=1}^{K} f_k \, (x)$$

其中，x是输入特征向量，$f_k(x)$是第 k 棵树的预测函数。

2. **定义损失函数：**为了训练模型，我们需要定义一个损失函数来衡量模型的预测误差。通常在回归任务中，我们使用均方误差（MSE）作为损失函数：

$$L = MSE = \frac{1}{N} \sum_{i=1}^{N} (\, y_i - F(x_i))^2$$

其中，N 是训练样本的数量，y_i是第 i 个样本的真实标签。

3. **训练过程:**CatBoost 使用梯度提升算法来逐步优化模型。在每一步中，根据当前模型的残差计算新的决策树f_k，使得损失函数最小化。每棵树的建立过程可以表示为：

$$f_k = arg \ \min_{f} \sum_{i=1}^{N} L \, (y_i, F_{k-1}(x_i) + f(x_i))$$

这里$F_{k-1}(x_i)$ 是前 k-1 棵树的累积预测，L 是损失函数。

4. **正则化**：CatBoost 在训练过程中还包括一些正则化技术，如树的深度限制、学习率控制等，以防止过拟合并提高泛化能力。
5. **预测**：训练完成后，通过将所有树的预测累加来得到最终的预测结果：

$$\hat{y} = F(x) = \sum_{k=1}^{K} f_k(x)$$

这就是 CatBoost 回归模型的基本数学过程：通过梯度提升的方式，逐步构建多棵决策树，并结合它们的预测来最小化损失函数，从而得到最优的回归预测模型。

最终，我们得到的 CatBoost 回归模型效果如下：

表 10 CatBoost 回归模型（Sorted-20，Sum-1）

R2 分数	ACC_3
0.86923	0.76290

其中 Sorted-20 代表升序排列的原始 20 指标，Sum-1 代表 1 个指标之和。

可见，CatBoost 在经过严谨分析的特征工程之后在**回归任务上优于基准模型**，三分类任务准确度与基准模型在伯仲之间。

6.4.5 K 折交叉灵敏性验证

K 折交叉验证（K-fold Cross-Validation）是一种常用的交叉验证技术，用于评估模型在数据集上的性能和泛化能力。它将数据集分成 K 个子集，每个子集称为一个折（fold）。K 折交叉验证的过程如下：

1. **数据集划分：**
- 将数据集分成 K 个大致相等的部分（折）。
- 每个折依次作为验证集，其余 K-1 个折作为训练集。

2. **交叉验证过程：**
- 第一轮：将第一折作为验证集，其余 K-1 折作为训练集，训练模型并在第一折上进行评估。
- 第二轮：将第二折作为验证集，其余 K-1 折作为训练集，训练模型并在第二折上进行评估。
- 依此类推，直到第 K 轮。

3. **评估指标**：每轮验证后得到一个评估指标（比如准确率、精确率、召回率等）。最终的模型性能评估通常是五轮验证结果的平均值，这样可以减少因为特定数据分割而引入的偏差。

此处我们采用 **5 折交叉验证**，得到的结果如下：

表 11 CatBoost 回归模型 5 折交叉验证

	1 折	2 折	3 折	4 折	5 折	均值
R2 分数	0.87019	0.86971	0.86849	0.86885	0.86892	0.86923
ACC_3	0.76247	0.76296	0.76290	0.76295	0.76324	0.76290

可见，模型在 5 折交叉验证的训练过程中保持了在回归与分类问题上的稳定且优异的性能，证明了其**低灵敏度、高泛化性、高鲁棒性**

七、问题 3 模型的建立与求解

7.1 利用精选指标优化模型（基于问题 1 的分析）

同时，加入指标和作为关键引导指标，组成全新的 11 维特征，训练 CatBoost 回归模型，性能较基准模型有所提升，而相比排序指标+指标和的模型部分伯仲，训练结果如下：

表 12 CatBoost 回归模型（Selected-10，Sum-1）

R2 分数	ACC_3
0.86923	0.76290

其中，Selected-10 代表精选高相关性指标 10 个，Sum-1 代表使用 1 个指标和作为关键指标。

可见，CatBoost(Selected-10，Sum-1)模型与 CatBoost(Sorted-20，Sum-1)的性能基本相当，且均在 R2 分数，即回归任务上优于普通线性回归的性能。

7.2 五指标模型优化

为了在仅选用 5 个关键指标作为特征的限制下优化问题 2 中所建立的 CatBoost 回归模型，本文通过借助划分指标来进行主成分分析，将 20 维指标降维至 **4 维关键指标**，结合**指标和**组成 **5 个关键指标**，实现求解得到**最优的 CatBoost 回归模型**。

7.2.1 主成分分析（PCA）

- **PCA 简介**

主成分分析（principal component analysis）是一种通过正交变换，将一系列可能线性相关的变量，转换成一组线性不相关的新变量的数学降维方法。而这些不相关变量也称作主成分(李航 2022：250-269)。主成分分析在简化运算、去除噪音、数据可视化等方面起到了重要作用。

PCA 算法的大致流程如下：

我们将 m 个 n 维数据记作矩阵 $X_{n\times m}=\left[\vec{x}_1,\vec{x}_2,...,\vec{x}_m\right]$

1. 对数据去中心化：

$$X = X - \frac{1}{m}\sum_{i=1}^{m}\vec{x}_i$$

2. 计算协方差矩阵 C：

$$C = \frac{1}{m}XX^T$$

3. 求解 C 的特征值矩阵（其中特征值按降序排列）,取前 k 列，记作矩阵:

$$P = P_{\mathrm{n}\times k}$$

4. 将原始数据对进行投影，得到降维后数据，即：

$$Y_{\mathrm{k}\times m} = P_{\mathrm{n}\times k}^{T}X_{n\times m}$$

- **应用 PCA 提取关键特征**

PCA 降维提取关键特征的步骤如下：

1. **数据选取：**选取所有 20 维指标作为特征维度，洪水概率为目标维度
2. **特征标准化：**PCA 对数据的尺度敏感，不同尺度的数据可能会导致主成分分析的结果不准确，因此对其进行标准化处理
3. **特征降维：**选取降维比例为 20%，将 20 维特征降维至 4 维，作为关键特征
4. **PCA 可视化：**展示主成分 1、2 的分布情况，并根据目标变量对数据点进

行了颜色编码，以显示不同类别的数据分布。

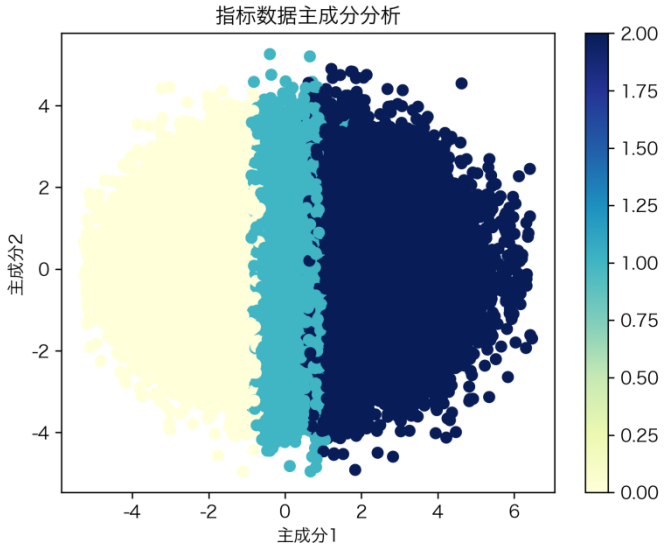


图 10 指标数据主成分分析结果

可见，PCA 数据降维实现了较好的**特征分解**与**关键指标提取**工作

7.2.2 最优 CatBoost 回归模型

基于 PCA 降维后的 4 个主成分，辅以指标和组成 5 个关键指标，用以训练 CatBoost 回归模型，并实现了远超先前所有基准、优化模型的效果，实现了 SOTA（State Of The Art）的最优模型：

表 13 SOTA CatBoost 回归模型（PCA-4，Sum-1）

R2 分数	ACC_3
0.97369	**0.90594**

其中，PCA-4 代表 PCA 压缩指标到 4 维，Sum-1 代表使用一个指标和作为关键指标，SOTA 即 State Of The Art 其代表最优模型。

可见，**PCA 降维处理后的数据结合指标和作为关键指标是极佳的特征选取与融合，实现了精妙的特征工程，使得模型性能大幅提升。**

具体模型优化效果及其与基准模型的对比如下表所示：

表 14 洪水概率预测回归模型

模型	R2 分数	ACC_3
线性回归(Org-20)	0.84444	0.76743
XGBoost(Org-20)	0.80975	0.74846
CatBoost(Org-20)	0.84634	0.76736
CatBoost(Sorted-20,Sum-1)	0.86923	0.76290
CatBoost(Selected-10,Sum-1)	0.86691	0.76281
SOTA **CatBoost(PCA-4,Sum-1)**	**0.97369**	**0.90594**

由此可清晰看到本文对模型优化的进行过程及特征工程带来的极佳优化效果。

八、问题 4 模型的建立与求解

基于问题 2 选取的回归模型 CatBoost，加以问题 3 对指标进行的关键指标选取，实现了最优的回归模型，并该模型与特征工程应用于 test.csv 数据，并将预

测结果填写进入 submit.csv 中。

按照问题 1 中的数据分析结果，测试集的频数分布直方图与训练集基本相同，而训练集的洪水概率的分布情况基本符合正态分布，**因此可以假设测试集的预测洪水概率分布情况也应基本尽量符合正态分布**。

对预测结果的概率密度直方图可视化如下：

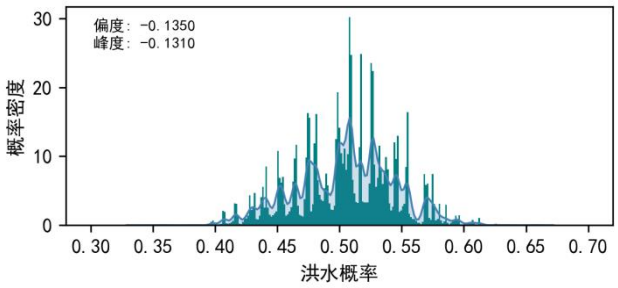


图 11 SOTA CatBoost（PCA-4，Sum-1）预测洪水发生概率统计直方图

该预测**结果分布情况没有明显的偏向性，基本符合正态分布**，对比其他性能较弱的两个 CatBoost 回归模型的预测结果概率分布直方图如下：

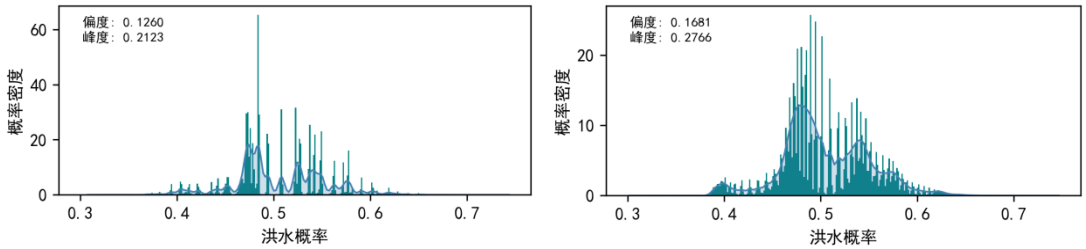


图 12 CatBoost[(Selected-10,Sum-1)，(Sorted-20,Sum-1)]预测洪水发生概率统计直方图

性能较弱的两个模型有着**明显的预测偏向性，并不符合正态分布**，可见，性能最优预测模型的预测结果是较为可信的，这也验证了该模型的鲁棒性。

对最优模型预测结果按正态分布随机采样 1000 条数据，可视化为折线图如下：

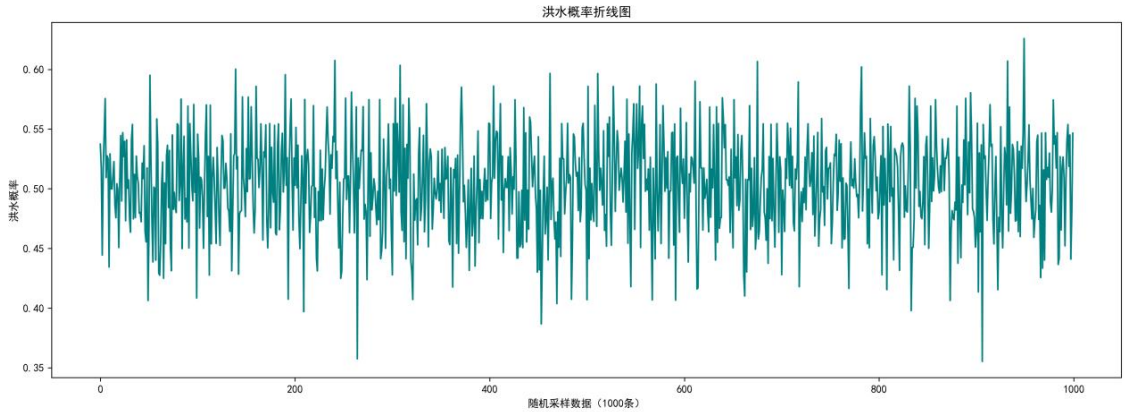


图 13 SOTA CatBoost（PCA-4，Sum-1）随机采样洪水折线图

可见数据基本在 0.5 周围上下浮动，异常值较少，并不影响整体数据分布情况。

九、模型的评价与推广

9.1 模型评价

(1) 优势

1. 量化数据有效：构建过程引入多种统计量对数据进行定量分析以选取建模方向

2. 数据分析严谨：利用 Spearman 相关系数，量化了各指标与洪水概率之间的相关性，并进行了假设检验，确认了模型特征选择的合理性。

3. 特种工程合理有效：经过对数据逐步地细致分析，最终得出了最优的特征处理方式，实现了性能极佳的模型。

4. 性能指标选取合适：使用了 R^2 分数和三分类准确率（ACC_3）作为模型性能的评估标准，通过这些指标，对不同模型的性能进行了横向比较。

5. 验证方法可靠严谨：采用 K 折交叉验证方法，验证了模型的稳定性和泛化能力，确保模型在未见数据上也能保持良好性能。

6. 可视化多样：通过直方图、折线图等可视化手段，直观展示了模型预测结果的分布情况，进一步分析了预测结果的正态性。

(2) 劣势

1. 无法处理时序数据：由于给定数据的限制，无法在时序上进行深入建模，这会造成模型实用性受到限制，因为现实中很多数据是包含强时序关系的。

2. 可识别处理的指标特征太少：由于给定数据的限制，进行更高特征维度的建模工作，这可能会影响实际情况下模型的学习表征的能力。

9.2 模型推广

基于上述评估结果，我们认为所构建的模型具有较高的实用价值和推广潜力。以下是模型推广的几个方向：

1. 数据集扩展：将模型应用于更广泛地区的洪水灾害数据，以验证模型在不同地理和气候条件下的适用性。

2. 实时预测系统：结合实时气象和水文数据，开发实时洪水灾害预测系统，为防洪减灾提供决策支持。

3. 政策制定支持：模型结果可为政府和相关部门制定防洪策略和应急响应计划提供科学依据。

4. 公众教育与预警：利用模型预测结果，开发公众洪水灾害预警应用程序，提高公众的防灾意识和应对能力。

十、参考文献

[1] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.
[2] 李航. 《机器学习方法》. 北京: 清华大学出版社, 2022: 57-75, 77-93, 131-146, 216-227, 250-269.

选题	2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
B		apmcm24102462

洪水灾害预测模型的设计与分析

摘要

洪水作为常见且破坏力极强的自然灾害，不仅由自然因素驱动，人为活动也极大地影响其频率和严重程度。本文从不同指标特征分析视角出发，根据机器学习算法建立不同等级预警评价模型、洪水发生概率预测模型，实现对洪水发生的准确预测与分析。

针对问题 1，建立决策树回归模型与 XGBoost 回归模型，以洪水发生概率为结果变量与 20 个指标（连续变量）进行关联性分析；为确保得出的关键指标在不同模型中均表现显著，选择结合两个模型的结果得出最终预测并列举出具有强关联性与弱关联性的两组指标；结合实际因素分析指标相关程度密切的可能原因，并针对洪水的提前预防提出了合理的建议和措施。

针对问题 2，首先通过肘部法则（Elbow Method）得到最优簇数结果为 3，验证了洪水高、中、低三个等级事件分类的合理性和科学性；然后通过 K-means 聚类分析方法（K=3）成功将洪水发生概率数据分类；利用随机森林回归模型建立了洪水预警评价模型，将聚类成功完成后的三类数据分三次输入至模型中，得出最合适的关联指标及其权重值。最后，根据得出的数据针对模型灵敏度进行分析，为后续进一步的特征分析、概率预测等提供了可靠依据。

针对问题 3，选择多层感知机（MLP）算法建立洪水发生概率预测模型，基于问题 1 的分析结果将具有强关联性的指标作为预测模型所需数据并进行预处理确保数据的完整性和一致性；通过数据准备、模型训练以及模型推理三个阶段成功建立并得到了具有高拟合度和稳定性的预测模型；基于特征重要性分析选定 5 个关键指标，在原模型的基础上采用调整网络结构、正则化技术、网格搜索调整学习率的方式进行模型优化，成功提高了模型拟合度，并且结果具有良好的可解释性与合理性。

针对问题 4，基于问题 3 中建立的预测模型对附件 test.csv 所有事件发生洪水的概率进行预测并将结果填入 sumbit.csv 中（具体结果见支撑材料）；将预测结果通过频率分布直方图和折线图进行可视化，通过直观观察、Shapiro-Wilk 统计的方法验证了模型预测结果符合正态分布特性，具有较高的泛化能力和实用性。

综上所述，洪水灾害预测模型的设计与建立提高了洪水发生概率的预测精度，为防洪预警和应对提供了科学依据，并且对洪水灾害的预防和管理具有重要的现实意义。

关键词：决策树　XGBoost　随机森林　K-means　多层感知机　Shapiro-Wilk

目录

一、问题重述

1.1 问题背景

洪水作为最常见、破坏力最强的自然灾害之一，常由暴雨、急剧融冰化雪、风暴潮等自然因素引起，不仅威胁到人类生命安全，还对农业、基础设施、生态环境等造成严重破坏。然而，洪水的发生不仅由自然因素驱动，人为活动也极大地影响其频率和严重程度，例如，迅猛的人口增长、耕地扩展、围湖造田以及乱砍滥伐等活动都会改变地表状态和汇流条件，从而加剧洪水灾害的发生和影响[1]。

因此，识别影响洪水发生的关键因素并准确预测洪水发生的概率，不仅能提高防洪预警的准确性和及时性，同时有效减少了洪水灾害带来的损失。

1.2 问题要求

问题设置循序渐进，全面系统地从数据分析到模型构建，再到实际预测和结果分析，逐步提高对洪水灾害的预测能力。每个问题既相互独立，又紧密关联，共同服务于整体目标——如何在复杂的自然因素和人为因素影响下，对洪水的发生概率做出准确预测。

问题 1 首先要求根据附件 train.csv 中提供的数据，分析并可视化影响洪水发生概率的 20 个不同指标，确定并分别列举出每个指标与洪水发生的相关性强弱，通过不同指标因素推断可能的原因，针对洪水的提前预防提出合理的建议和措施。

问题 2 首先要求将附件 train.csv 中的洪水发生概率进行分类，给出高、中、低三种不同风险的洪水事件更具关联性的指标特征，在此基础上，选择合适的指标并计算权重值，建立不同风险等级的洪水预警评价模型，并对模型进行灵敏度评估。

问题 3 要求利用问题 1 中指标分析结果建立洪水发生概率的预测模型，从 20 个指标中选取合适的指标预测洪水发生的概率并验证模型的准确性；在此基础上，探究选择 5 个指标时如何调整并改进预测模型。

问题 4 要求利用问题 3 中建立的洪水发生概率预测模型，对附件 test.csv 中所有事件发生洪水的概率进行预测，并将预测结果填入附件 submit.csv 中；绘制发生洪水概率的直方图和折线图并验证预测结果的分布是否符合正态分布。

二、问题分析

2.1 总体分析

本题是一个关于洪水发生事件指标数据分析与概率预测的问题。

分析目的上，需要对附件数据进行分析，确定 20 个不同指标与洪水发生概率的关联程度，对影响洪水发生的因素与概率做出初步判断。因此，本题需要达成三个主要任务：其一，根据洪水发生的概率对洪水事件进行分类，确定高、中、低三种风险事件不同的指标特征。通过计算不同指标的权重建立不同风险等级的洪水预警评价模型，并对模型进行灵敏度分析；其二，基于前述的分析结果，建立洪水发生概率的预测模型并选取合适的指标预测洪水发生的概率。在仅使用 5 个关键指标的情况下，对预测模型调整和改进；其三，对附件 test.csv 中所有事件发生洪水的概率进行预测，并绘制出概率直方图和折线图以验证结果是否服从正态分布，确保模型的泛化能力和稳健性。

数据特征上，关于洪水灾害的数据具有多维度、异质性强、噪声较多等特征。因此，本题数据相对复杂特殊，需要对数据进行预处理，以提高模型的准确性和稳定性。

模型选择上，由于洪水数据样本量大，并考虑到问题的背景、目的等与实际应用密

切相关，所选模型应当追求合理性、实际性，以便于结果的解读和应用。

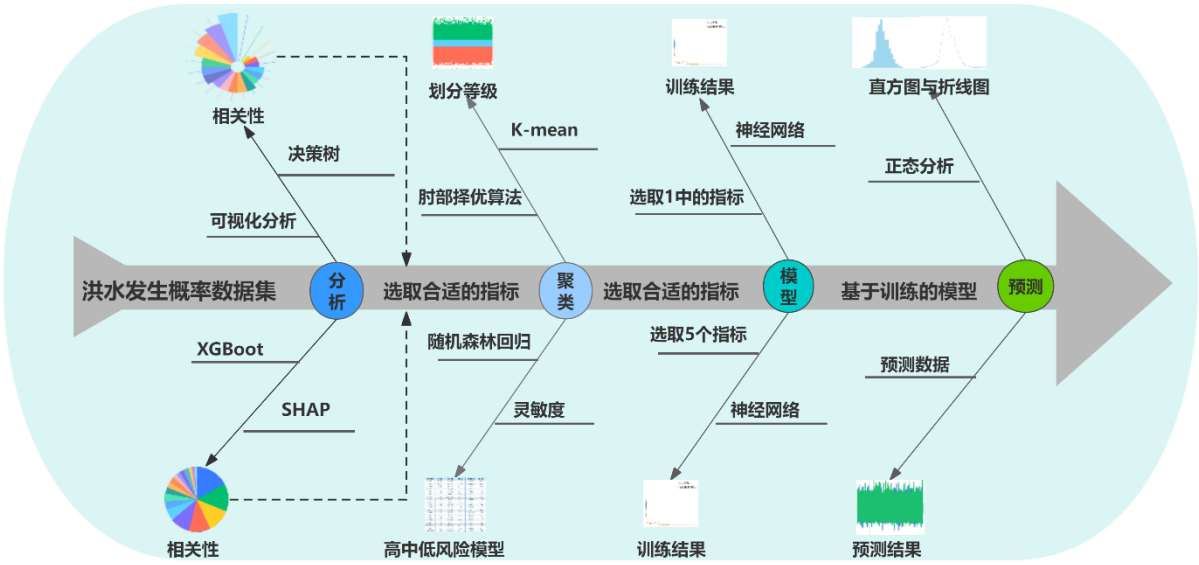


图 1 总体分析示意图

2.2 问题 1 的分析

问题 1 的核心目的是在于确定哪些指标与洪水事件的发生密切相关，为后续预测洪水发生概率提供依据。通过对附件 train.csv 中的数据进行分析和可视化，得出 20 个指标中的关键影响因素。但是，由于数据中可能存在噪声、异常值等，传统的简单模型已不足以揭示这些关系。因此，可以采用机器学习方法来分析指标与洪水发生的关系趋势，并用统计量或可视化工具来刻画这些趋势。基于这些分析结果，提出合理的洪水预防建议和措施，以降低洪水带来的风险和损失。

2.3 问题 2 的分析

问题 2 的核心目的在于根据洪水发生的概率通过聚类算法（K-means 聚类、层次聚类等）将附件 train.csv 中的洪水事件分为三类，并利用特征重要性分析或相关性分析的方法指明不同等级洪水事件的密切相关指标。基于选定的关键指标和其权重建立不同风险等级的预警评价模型并进行模型的灵敏度分析，确定对预测结果影响最大的指标，并根据分析结果调整模型结构和参数，以提高模型的稳定性和可靠性。

2.4 问题 3 的分析

问题 3 的核心目的在于选择合适的机器学习方法建立洪水发生概率的预测模型。在 20 个指标中选取合适的指标完成模型准确性的验证及性能的评估，并基于特征重要性分析选定 5 个关键指标，在关键指标下优化和改进模型，确保在减少特征数量的情况下，模型仍能保持较高的预测性能。

2.5 问题 4 的分析

问题 4 的核心目的在于进一步验证模型的有效性，其一预测附件 test.csv 中的所有事件发生洪水的概率，评估模型在未见数据上的表现，确保模型不仅在训练数据上有效，在新的数据环境中同样具有预测能力；其二绘制概率直方图和折线图，检验结果是否服从正态分布以反映模型预测结果的合理性和数据的自然变异性。

三、符号说明

符号	含义
y_i	均方误差（MSE）的实际值，$i=1,2$
$\hat{y}_i$	均方误差（MSE）的预测值，$i=1,2$
n_i	均方误差（MSE）子节点的样本数量，$i=1,2$
Ω	正则化项
γ、λ	正则化参数
η	学习率
w	权重
f_j	回归模型中的特征
x_i	欧氏距离中的数据点
σ	激活函数
β_i	动量项的衰减系数，$i=1,2$

四、问题 1 模型的建立与求解

4.1 洪水发生概率与不同指标的相关程度分析

4.1.1 基于决策树回归模型的建立与求解

洪水发生概率及 20 个不同指标均为连续变量，考虑以洪水发生概率为结果变量，与各个指标进行回归分析。

首先，利用决策树回归模型评估每个指标对洪水发生概率的关联程度，通过构建树形结构进行分割以实现预测目标变量。决策树[2]是一种基于树形结构的监督学习方法，用于处理分类和回归问题，如图 2 所示。它通过递归地将数据集划分为更小的子集，形成一棵树状结构。每个内部节点表示一个特征，每个分支表示该特征的一个可能取值，每个叶节点表示一个类标签或一个连续值。根据特征的取值从根节点开始逐步向下遍历树，最终达到叶节点以确定数据的分类或预测值。针对问题 1，由于洪水发生概率存在多个特征影响，因此，选用决策树回归建立模型。

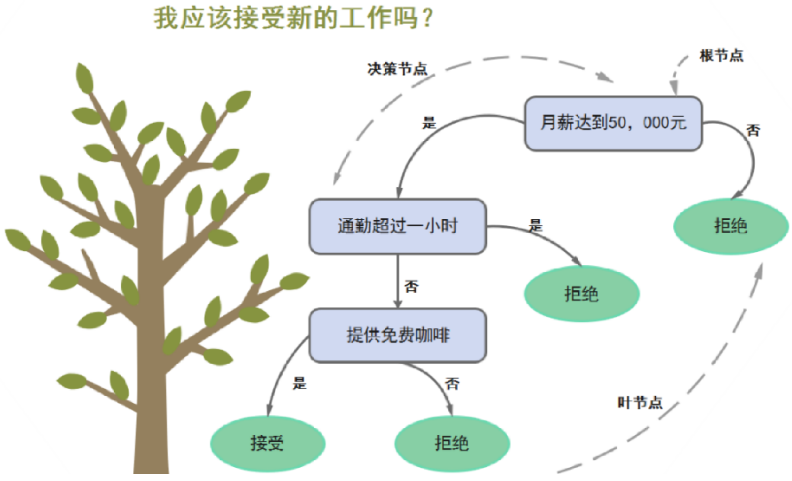


图 2 决策树回归模型

建立决策树回归模型的过程如下：

构建树形结构：决策树由根节点、内部节点和叶节点组成。根节点包含整个数据集，每个内部节点根据某个特征进行分裂，叶节点包含最终的预测值。模型不断分裂节点，逐步缩小数据集的范围，以提高预测的精确度。

初始化：从根节点开始，初始化整个数据集作为根节点的数据。根节点的预测值设为整个数据集的洪水发生概率的平均值，以此作为初始预测值。值得注意的是，平均值的设定有助于提供一个合理的基线预测，使得初始误差较小，从而提高模型的收敛速度和稳定性。

递归分裂：在每个节点，选择最佳的特征和分裂点，根据分裂准则将数据集分成两个子节点。其中，分裂准则采用均方误差（MSE）或基尼指数。

MSE 具体公式如下所示：

$$MSE = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \hat{y}_1)^2 + \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i - \hat{y}_2)^2 \tag{1}$$

其中，n_1和n_2分别是两个子节点的样本数量，y_i是实际值，$\hat{y}_1$和$\hat{y}_2$是两个子节点的预测值。选择使得 MSE 最小的特征和分裂点进行分裂，从而优化洪水发生概率的预测。

基尼指数具体公式如下所示：

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{2}$$

其中，p_i是样本点属于第i的概率。基尼指数衡量的是节点的不纯度，基尼指数越小，节点纯度越高。

选择森林覆盖率作为分裂特征，通过计算分裂后的 MSE，确定最佳的分裂点，如公式（3）所示：

$$\text{最佳分裂点} = \underset{s}{argmin}\left(MSE = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \hat{y}_1)^2 + \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i - \hat{y}_2)^2\right) \tag{3}$$

其中，s表示分裂点。

树的增长：决策树通过递归分裂节点生长，每次分裂都会根据当前节点的数据集选择最佳的特征和分裂点，直到满足停止条件。停止条件可以是树的最大深度、叶节点的最小样本数或 MSE 的阈值。例如，当某个特征（如侵蚀等）对洪水发生概率有显著影响时，模型会优先选择该特征进行分裂，以提高预测准确性。

训练完成，输入特征变量进行预测。首先，遍历树形结构，从根节点开始，根据输入数据的特征值，沿着决策树的分裂路径向下遍历，直到到达叶节点。当输入一个新的样本，该样本具有淤积、城市化水平、地形排水特征等，模型将沿着决策树的路径进行分裂，最终到达对应的叶节点。最终，叶节点的值即为输入数据的预测值。

最终，利用 cross-validation 评估模型的性能，防止过拟合；通过特征重要性分析（公式（4））识别对洪水发生概率影响指标排序。得到最终结果如图 3 所示。

$$Importance(f_j) = \sum_{t=1}^{T} I(f_j, t) \tag{4}$$

其中，$Importance(f_j)$表示特征f_j在所有树中的重要性，$I(f_j, t)$表示特征f_j在第t棵树中的重要性。通过分析特征重要性，识别出不同指标对洪水发生概率的预测关联程度。

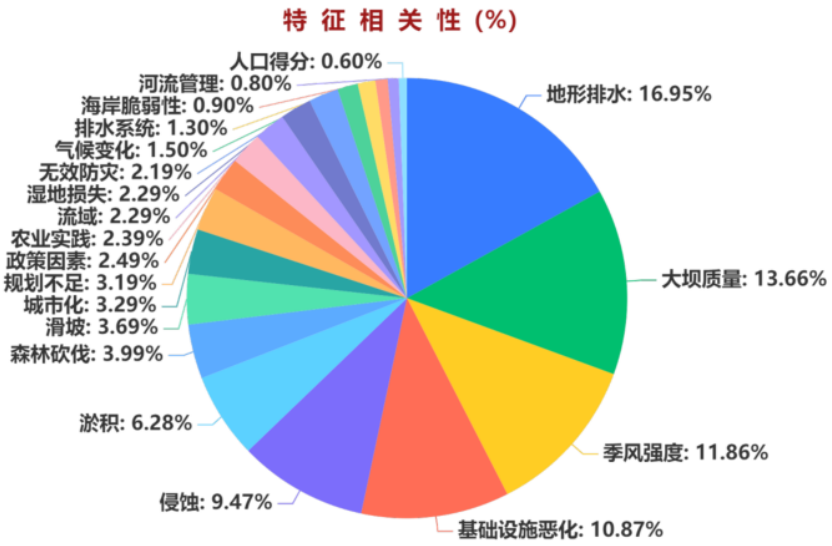


图 3 决策树回归模型求解结果

4.1.2 基于 XGBoost 回归模型的建立与求解

为了更有效的评估每个指标对洪水发生概率的影响，同时利用 XGBoost 回归模型进行分析。XGBoost（Extreme Gradient Boosting）模型[3]是基于决策树的集成学习方法，通过组合多棵决策树来提高模型的预测性能，如图 4 所示。

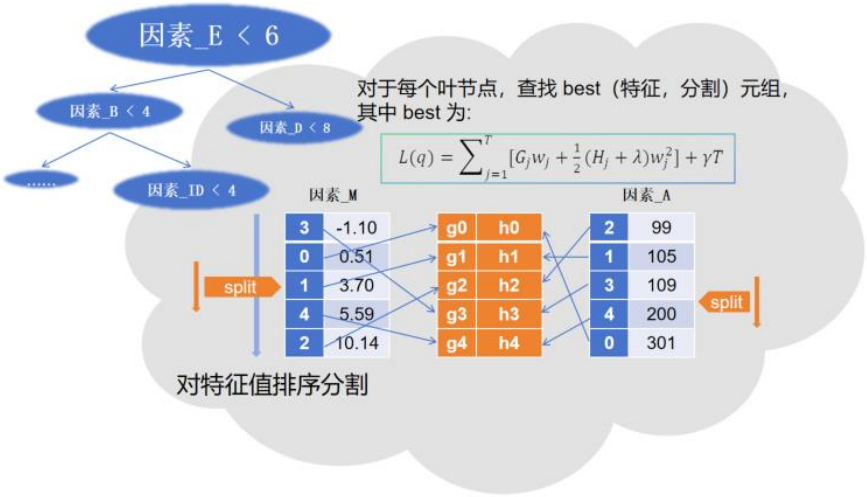


图 4 XGBoost 回归模型

XGBoost 回归模型的目标是最小化以下目标函数，如公式（5）所示：

$$L(\theta)=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{5}$$

其中，y_i是实际的洪水发生概率，$\hat{y}_i$是模型预测的洪水发生概率，l是损失函数，Ω是正则化项，n是样本数，K是树的数量。

对于洪水发生概率预测，使用 MSE 作为损失函数，如公式（6）所示：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{6}$$

利用正则化项（公式（7））控制模型的复杂度以防止过拟合，包括树的复杂度和叶节点权重：

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

其中，T是树的叶节点数量，w_j是第j个叶节点的权重，γ、λ是正则化参数。

XGBoost 通过迭代构建树的方式进行训练，每次迭代时，模型会根据当前残差来生成新的树，从而逐步减少误差。在洪水发生概率预测的过程中，每次迭代都会基于之前的预测误差，调整模型参数以改进预测结果。

详细模型建立与训练过程如下：

初始化模型参数：设置初始的预测值为洪水发生概率的平均值。其中，平均值代表了数据的中心趋势，使模型在初始阶段具有较低的预测误差，有助于模型更快地收敛到最优解，增强了梯度下降算法的有效性，并提高训练效率。

计算梯度和 Hessian 矩阵：对于每个样本，计算损失函数的一阶导数和二阶导数，即梯度和 Hessian 矩阵（公式（8）-（9））从而在洪水发生概率的预测过程中，由此更新模型参数以不断减少预测误差。

$$\nabla L(\theta) = \left(\frac{\partial L}{\partial \theta_1}, \frac{\partial L}{\partial \theta_2}, \ldots, \frac{\partial L}{\partial \theta_n}\right) \tag{8}$$

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_2^2} & \cdots & \frac{\partial^2 L}{\partial \theta_2 \partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L}{\partial \theta_n \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_n \partial \theta_2} & \cdots & \frac{\partial^2 L}{\partial \theta_n^2} \end{bmatrix} \tag{9}$$

生成新的决策树：根据计算出的梯度和 Hessian 矩阵，生成新的决策树以拟合当前的洪水发生概率预测误差。新的决策树通过最小化以下目标函数进行训练，如公式（10）所示：

$$L(q) = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2\right] + \gamma T \tag{10}$$

其中，G_j和H_j分别是叶节点j的梯度和 Hessian 矩阵的总和，w_j是叶节点的权重，q是决策树的结构，T是叶节点的数量。

更新模型参数：根据新生成的决策树，更新模型的预测值（公式（11））。

$$\hat{y}_{\mathrm{i}}^{(t+1)} = \hat{y}_{\mathrm{i}}^{(t)} + \eta f_k(x_i) \tag{11}$$

其中，η是学习率，f_k是第k棵树的模型。通过每次迭代更新预测值，逐步减少洪水发生概率的预测误差。

迭代训练：重复计算梯度和 Hessian 矩阵、生成新的决策树以及更新模型参数这三个步骤，直到损失函数收敛或达到最大迭代次数。每次迭代都会生成一棵新的决策树，逐步提升模型的预测能力，从而更准确地预测洪水发生概率。

模型验证：使用交叉验证（cross-validation）评估模型性能，防止过拟合。cross-validation 将数据集分成多个子集，在不同子集上训练和验证模型，从而获得更稳定和可靠的性能评估。通过计算 MSE 和决定系数（R^2）（公式（12））来衡量模型的预测准确性：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{12}$$

其中，$\overline{y}$是实际洪水发生概率的均值。

特征重要性分析：训练结束，通过特征重要性分析识别对洪水发生概率影响的指标排序。

值得注意的是，基于 XGBoost 模型对洪水发生概率进行预测后，进一步使用 SHAP（SHapley Additive exPlanations）值（图 5）量化每个特征对预测结果的相关性，提供全局和局部的解释，从而更准确地解释每个特征对洪水发生概率预测的影响，增强模型的解释性和透明度。

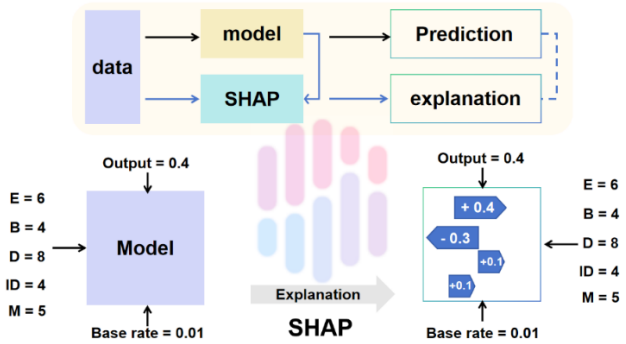


图 5 SHAP 原理图示

特别地，SHAP 值基于博弈论中的 Shapley 值，为每个特征分配一个归一化数值，表示该特征对预测结果的影响，确保了每个特征的关联值总和等于模型预测值与基线预测值的差异，从而提供了一致的解释框架。其计算公式如下所示：

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\cdot(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)] \tag{13}$$

其中，ϕ_i表示特征i的 SHAP 值，N是所有特征的集合，S是特征的子集，$f(S)$是只包含特征子集S时的模型预测值，$f(S \cup \{i\})$是包含特征子集S和特征i时的模型预测值。

完成上述步骤后，得到最终结果如图 6 所示。

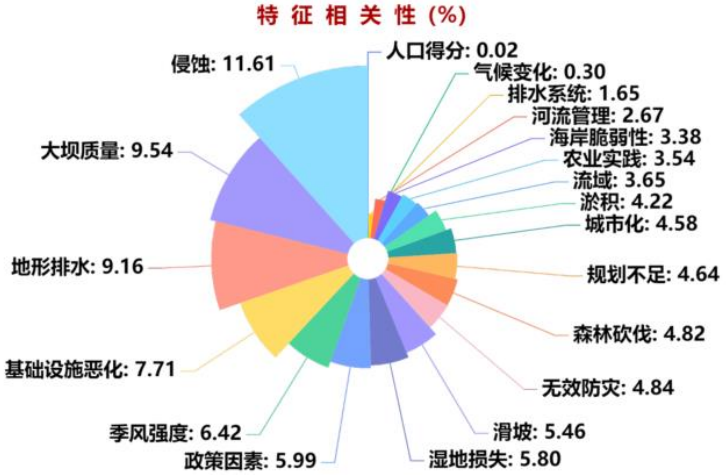


图 6 XGBoost 回归模型求解结果

4.1.3 针对决策树回归模型与 XGBoost 回归模型的综合分析

在前两节中，分别使用了决策树回归模型和 XGBoost 回归模型对洪水发生概率进行了预测分析。最终结合两个模型的预测结果，确保得出的关键指标在不同模型中均表现显著，从而提供更加全面且可信的洪水预测依据。两个模型得出的结果具有一致性，经过分析得到弱相关性的结果有 5 个，分别为人口得分、气候变化、排水系统、河流管

理和海岸脆弱性，其余 15 个指标均为强相关性。

4.2 指标相关性的原因分析

针对相关性强的指标（侵蚀、大坝质量、地形排水等）而言，它们对洪水的发生概率有直接且显著的影响。比如侵蚀和地形排水条件决定了地表水的吸收和流动情况；大坝质量和基础设施的状况直接关系到防洪能力的有效性；季风强度会带来大量降雨，迅速增加地表径流量，这些因素都会显著增加洪水发生的概率。

相对而言，人口得分、气候变化、排水系统、河流管理和海岸脆弱性这些指标与洪水发生的直接相关性较弱。虽然这些因素在一定程度上影响洪水风险，但它们更多地影响洪水的后果和长期风险。比如，气候变化的短期效应难以精确量化；人口得分和气候变化虽然对环境和气候有长期影响，但短期内对洪水发生的直接影响有限。

因此，通过对强弱相关性指标的分析，实现了更加清晰地理解洪水发生的主要驱动因素和次要因素。

4.3 提前预防的合理建议与措施

基于以上指标相关性分析，提出以下合理的提前预防措施：

针对强相关性指标，例如增强季风强度监测和预报，提高地形排水能力，以及定期检查和维护大坝等关键基础设施，确保其在极端天气条件下的稳定性和可靠性。此外，加强侵蚀防治措施，防止土壤流失，也是降低洪水风险的重要手段。

针对弱相关性指标，同样需要采取相应措施以提高整体防洪能力。例如，改善排水系统和河流管理，提高城市和农村地区的排水效率和抗洪能力；通过政策和规划减少人口密集区的洪水风险，特别是在气候变化加剧的背景下，加强对气候变化的长期研究和监测。通过综合运用这些措施，实现有效提高洪水预防和应对能力，减少洪水灾害对人类社会的影响。

五、问题 2 模型的建立与求解

5.1 不同等级洪水事件的分类及分析

为界定不同等级的洪水事件，首先将附件 train.csv 中洪水发生的概率通过频率分布直方图表示，如图 7 所示。由图可知，大部分洪水数据分布在概率数值为 0.5 附近。为进一步明确等级界限，利用肘部法则（Elbow Method）评估不同簇数下聚类模型的表现，绘制了 SSE 与簇数关系图，通过观察图中的拐点得出最优簇数为 3，如图 8 所示。该值正好对应题中的高、中、低三个事件等级，进一步验证了洪水事件三分类的合理性和科学性，为后续的特征分析和预警模型的建立提供了基础。

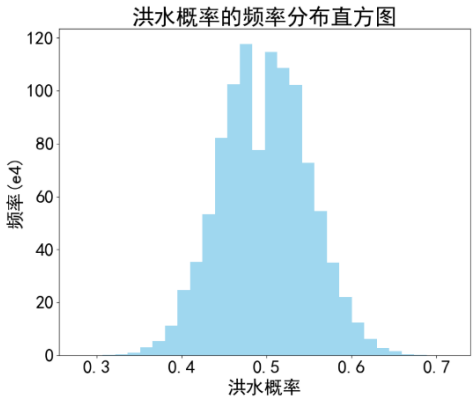


图 7 洪水概率的频率分布直方图

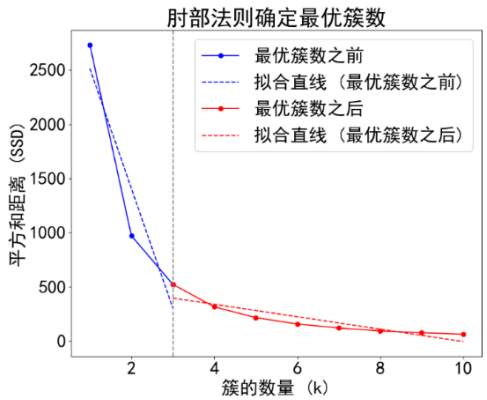


图 8 肘部法则确定最优簇数图

由上述分析后，利用 K-means 算法[4]对洪水数据进行聚类，分为高、中、低三种不同等级。该算法作为一种迭代聚类算法，将数据分割成 k 个簇，使得同一簇内的数据点彼此之间更加相似，更加有效地将不同风险等级的事件进行区分。

针对 K-means 算法，具体实现步骤如下：

随机选择 k 个初始聚类中心$\{\mu_1, \mu_2, \ldots, \mu_k\}$，本题 k=3，分别表示高风险、中风险和低风险三类。

将每个数据点分配到最近的聚类中心，依据欧氏距离计算公式（公式（14））：

$$d(x_i, \mu_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - \mu_{jm})^2} \tag{14}$$

其中，$d(x_i, \mu_j)$表示数据点x_i到聚类中心μ_j的距离，n为数据的维度。对于洪水发生概率数据，该步骤将每个事件分配到对应的风险等级。

重新计算每个簇的聚类中心，将其更新为该簇内所有数据点的平均值（公式（15））。

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \tag{15}$$

其中，μ_j是第 j 个聚类中心，C_j是第 j 个簇的数据点集合，$|C_j|$是该簇中数据点的数量。更新聚类中心能够使每次迭代后聚类中心更准确地代表簇内数据点的中心位置。

重复迭代分配数据点、更新聚类中心两步骤多次，直到聚类中心不再发生变化或达到最大迭代次数，确保聚类结果的稳定性和准确性。

经过上述步骤，最终成功将附件 train.csv 中洪水发生的概率聚类成不同类别，得到三个簇，每簇对应一种风险等级（高、中、低），如图 9 所示。

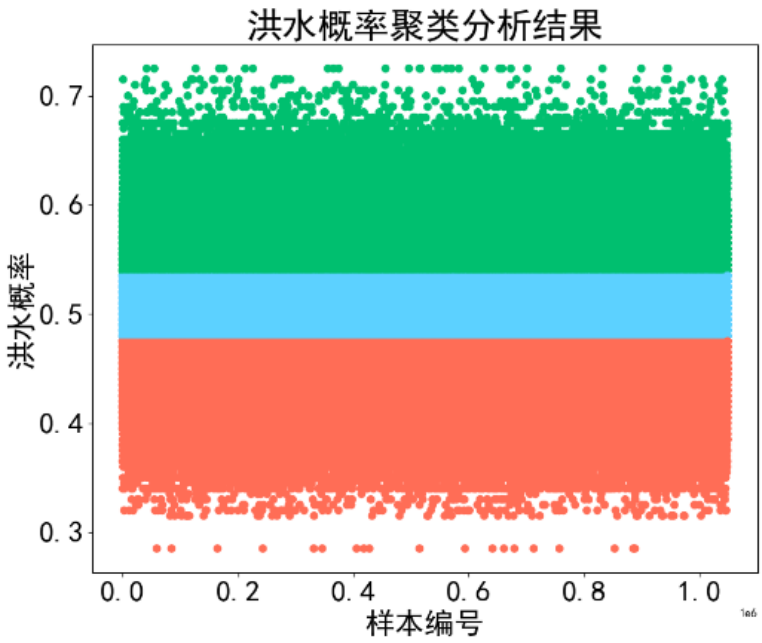


图 9 洪水概率聚类分析结果

5.2 随机森林回归预警评价模型的建立与求解

利用随机森林[5]回归模型（图（10））建立预警评价模型实现预测洪水发生的风险等级。将上一节聚类成功完成后的三类数据分三次输入至预警评价模型中，得出最合适的关联指标及其权重值。

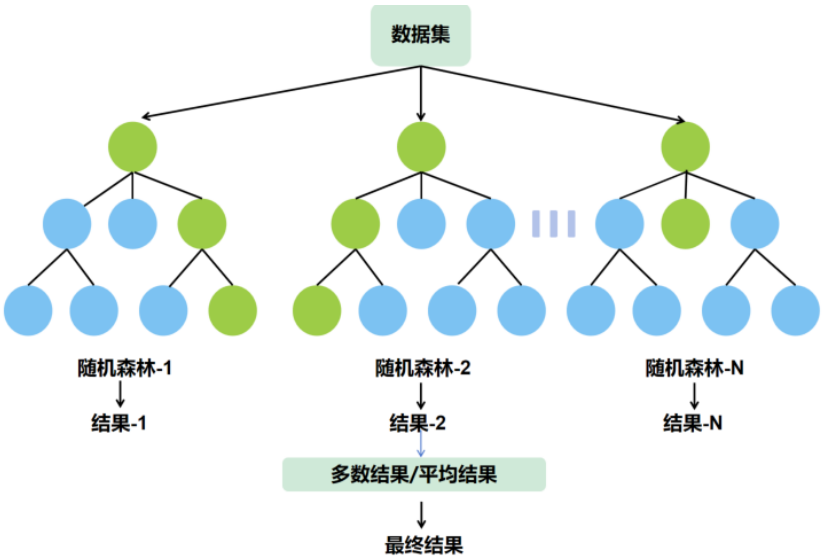


图 10 随机森林模型

对于预警评价模型，详细模型建立与训练过程如下：

首先进行数据标准化，对所有特征进行标准化处理，使其均值为 0，标准差为 1，以确保模型训练的稳定性和准确性。

随机抽样：在相同类别的洪水数据中随机抽取多个子样本，每个子样本用于训练一棵决策树。假设原始数据集有 n 个样本，每个子样本的大小也是 n，通过有放回抽样的方法形成不同的子样本集。

决策树构建：对每个子样本构建一棵决策树。在每个节点上，随机选择 m 个特征（通常为总特征数的平方根），并在这些特征中选择最佳分裂特征和分裂点，以最大化信息增益（公式（16））或最小化均方误差（MSE）。

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v\in\text{Values}(A)} \frac{|S_v|}{|S|}\text{Entropy}(S_v) \tag{16}$$

其中，$\text{Entropy}(S)$是集合S的熵，$\text{Values}(A)$是特征A的所有可能取值，S_v是特征A取值为v的子集。

所有决策树构建完成后，通过集成这些树的预测结果来进行回归预测。具体来说，对每棵树的预测结果取平均值（公式（17））。

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t \tag{17}$$

其中，$\hat{y}$是最终的预测值，$\hat{y}_t$是第t棵树的预测结果，T是决策树的总数。

通过初始划分出的测试数据对模型进行评估，并计算模型的 MSE 和 R^2。最后通过对模型的特征重要性分析包括特征重要性得分以及特征重要性总和归一化（公式（18）-（19）），得出对洪水风险预测影响最大的几项指标。

$$\text{Importance}(f_j) = \frac{1}{T}\sum_{t=1}^{T}(\text{MSE}_{\text{before}} - \text{MSE}_{\text{after}}) \tag{18}$$

其中，$\text{Importance}(f_j)$表示特征f在所有树中的重要性得分，$\text{MSE}_{\text{before}}$是分裂前的 MSE，$\text{MSE}_{\text{after}}$是分裂后的 MSE。

$$w_j = \frac{\text{Importance}(f_j)}{\sum_{i=1}^{n}\text{Importance}(f_i)} \tag{19}$$

其中，w_j是特征f_j的归一化权重，n是特征的总数。

由上述步骤，有效建立了洪水不同风险等级的预警评价模型，根据模型进行特征重要性分析与权重计算，验证模型对于不同指标的灵敏度。最终结果如表 1 所示。

表 1 不同等级洪水事件相应指标特征及权重值

特征名称	高风险	特征名称	中风险	特征名称	低风险
地形排水	14.20%	地形排水	14.10%	大坝质量	13.40%
淤积	12.30%	无效防灾	13.40%	基础设施恶化	10.80%
农业实践	11.50%	农业实践	8.90%	滑坡	10.30%
河流管理	8.80%	气候变化	6.40%	河流管理	9.10%
人口得分	7.80%	规划不足	6.30%	气候变化	8.40%
滑坡	7.30%	城市化	5.30%	城市化	7.80%
湿地损失	7.00%	基础设施恶化	5.30%	人口得分	7.20%
无效防灾	5.80%	侵蚀	5.20%	森林砍伐	5.70%
政策因素	4.90%	流域	5.00%	淤积	4.90%
气候变化	4.50%	湿地损失	5.00%	湿地损失	4.60%
大坝质量	3.40%	大坝质量	4.20%	季风强度	4.30%
城市化	3.30%	排水系统	3.80%	侵蚀	3.80%
规划不足	2.60%	人口得分	3.60%	海岸脆弱性	3.30%
流域	1.70%	滑坡	3.10%	规划不足	2.70%
季风强度	1.50%	政策因素	3.00%	地形排水	1.10%
森林砍伐	1.30%	河流管理	2.80%	政策因素	1.10%
排水系统	1.00%	淤积	1.90%	排水系统	0.80%
基础设施恶化	1.00%	季风强度	0.90%	流域	0.80%
侵蚀	0.00%	森林砍伐	0.90%	农业实践	0.00%
海岸脆弱性	0.00%	海岸脆弱性	0.90%	无效防灾	0.00%

由表 1 可知，具有高、中、低风险的洪水事件都具有相应的指标特征，对于高风险洪水事件最能够影响其发生的指标及其权重依次分别为：地形排水（14.20%）、淤积（12.30%）、农业实践（11.50%）、河流管理（8.80%）、人口得分（7.80%）等；对于中风险洪水事件最能够影响其发生的指标及其权重依次分别为：地形排水（14.10%）、无效防灾（13.40%）、农业实践（8.90%）、气候变化（6.40%）、规划不足（6.30%）等；对于低风险洪水事件最能够影响其发生的指标及其权重依次分别为：大坝质量（13.40%）、基础设施恶化（10.80%）、滑坡（10.30%）、河流管理（9.10%）、气候变化（8.40%）等；

此外，通过模型灵敏度分析，验证模型的稳健性，确保在不同情境下模型仍然具有可靠的预测性能。同样地，结合表 1 数据得到针对 20 个指标对模型的灵敏度影响。

六、问题 3 模型的建立与求解

6.1 数据预处理

基于问题 1 的分析结果确定对于洪水发生概率具有强关联性的相应指标作为所需建立预测模型的数据集。由于提供数据量较大，故需检查数据集是否存在缺失值以防止

预测模型出现训练误差。针对缺失数据处理方法为删除含有缺失值的样本、使用均值或中位数填充缺失值，以及使用插值法填补缺失值。

6.2 基于神经网络的洪水发生概率预测模型建立与求解

6.2.1 单层感知机模型

单层感知机作为神经网络模型（图 11），主要解决线性可分的分类问题，由一个输入层和输出层组成，能够对输入特征进行加权求和并通过激活函数产生输出。

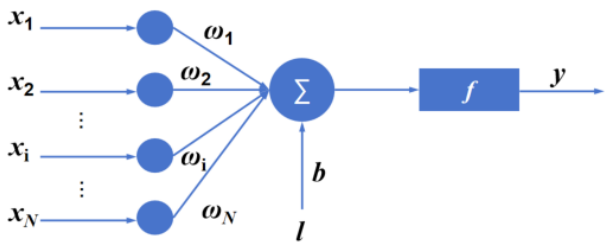


图 11 单层感知机模型

输入层接受多维向量$X = [x_1, x_2, \ldots, x_n]$，每个输入特征$x_i$对应一个权重$w_i$，所有权重构成向量权重$W = [w_1, w_2, \ldots, w_n]$，偏置$b$用于调整决策边界的位置。

输出层计算加权和并通过激活函数生成输出：$y = \sigma(W \cdot X + b)$，其中，$\sigma$是激活函数。常用的激活函数包括阶跃函数和线性函数。

利用 MSE 误差函数衡量模型预测值与实际值之间的差异$e = y - \hat{y}$，并不断更新权重和偏置：$W = W - \eta eX$，$b = b - \eta e$；直至差小于预设阈值或达到最大迭代次数时停止训练。

然而，单层感知机表达能力有限无法捕捉复杂的模式和特征，比如非线性问题。因此，多层感知机（MLP）的引入实现处理更复杂的数据和关系。

6.2.2 MLP 模型的建立与求解

MLP 模型[6]（图 12）作为深度学习方法，通过多层非线性变换捕捉复杂的特征关系，具有强大的非线性拟合能力，主要包括输入层、中间层（一个或多个隐藏层）、输出层。相比传统的线性模型，它能够更好地处理多维数据和非线性关系，针对题目中涉及的复杂多变的气象、地理、人口等洪水数据，该模型尤为合适。灵活性和可扩展性使其能够适应不同规模和复杂度的数据集，通过调整网络结构和参数实现最优预测性能。

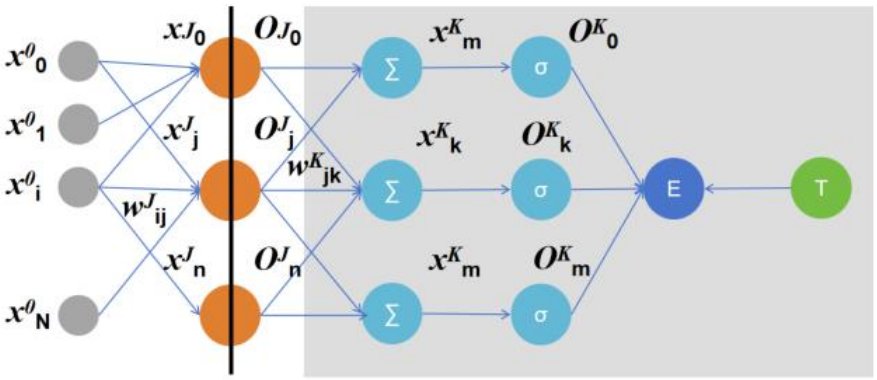


图 12 MLP 感知机模型

模型建立的过程主要分为三个阶段：数据准备、模型训练以及模型推理。

（1）数据准备阶段

将经过数据预处理后产生的数据作为输入特征输入模型。具体地，选定的指标特征包括侵蚀、基础设施恶化、季风强度、大坝质量、地形排水、淤积。

（2）模型训练阶段

首先，数据按照 7:2:1 的比例划分为训练集、验证集和测试集，完成模型的训练；

接下来，训练集数据依次进入模型的输入层、中间层、输出层；

模型中包含多个隐藏层，每个隐藏层由若干神经元组成，每个神经元与前一层的所有神经元全连接，使用非线性激活函数 ReLU（Rectified Linear Unit）以捕捉非线性关系。隐藏层的输出通过全连接层传递到下一层。

隐藏层定义如下所示：

$$H = \sigma(W_h \cdot X + b_h) \tag{20}$$

其中，H是隐藏层输出，W_h和b_h分别是隐藏层的权重和偏置，σ是激活函数。

ReLU 激活函数定义如下所示：

$$\sigma(z) = max(0, z) \tag{21}$$

其中，z是输入值。ReLU 函数能够有效缓解梯度消失问题。

输出层由一个神经元组成，使用 sigmoid 激活函数将输出值限制在 0 到 1 之间，表示洪水发生的概率。

输出层定义如下所示：

$$Output = \sigma(W_o \cdot X + b_o) \tag{22}$$

其中，$Output$是最终预测值，W_o和b_o分别是输出层的权重和偏置。

sigmoid 激活函数定义如下所示：

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{23}$$

其中，z是输入值。Sigmoid 函数将输出值映射到(0, 1)区间，适合概率输出。

因此，MLP 模型的公式如下所示：

$$Output = \sigma(W_o \cdot \sigma(W_{h2} \cdot \sigma(W_{h1} \cdot X + b_{h1}) + b_{h2}) + b_o) \tag{24}$$

模型训练过程中，使用 MSE 作为损失函数，定义如下所示：

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_t)^2 \tag{25}$$

使用 Adam 优化器进行梯度下降，更新模型参数。此外，结合动量和自适应学习率调整进行优化。

Adam 优化器参数更新如公式（26）-（30）所示：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\theta \mathcal{L}(\theta) \tag{26}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla_\theta \mathcal{L}(\theta))^2 \tag{27}$$

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{28}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{29}$$

$$\theta = \theta - \eta \frac{\widehat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \tag{30}$$

其中，η是学习率，β_1和β_2是动量项的衰减系数，ϵ是防止除零的一个极小值。

训练过程中，为防止模型在训练集上表现良好但在验证集和测试集上表现较差的过

拟合问题，采用早停技术（early stopping），即在每个训练迭代周期结束时，监控模型在验证集上的损失函数值。如果在若干迭代周期内验证集损失值没有显著下降或开始上升，则提前停止训练，并恢复到验证集损失值最小的模型参数。不仅防止过拟合，提高模型的泛化能力，而且节省训练时间，确保了模型在实际应用中的预测准确性和可靠性。

（3）数据推理阶段

模型训练完成后，在测试集上计算 MSE（公式（6））和R^2（公式（12））评估其性能，确保模型能够准确预测洪水的发生概率。

R^2表示模型解释自变量总变异的比例，其值越接近 1，表示模型对洪水发生变异解释能力越强，预测效果越好。

$R^2 \approx 1$表示模型能很好地解释洪水发生的概率，性能较好；

$0 < R^2 < 0.5$ 表示模型对洪水发生概率的解释能力较弱，性能较差；

$R^2 \approx 0$表示模型几乎不能解释洪水发生的变异，性能较差；

MSE 衡量了预测值与实际值之间的平均平方误差。值越小，表示模型的预测误差越小，性能越好。

性能评估完成后，利用交叉验证调整模型超参数以选择最佳参数组合。不断调整后验证模型最终预测性能。

6.2.3 MLP 模型结果分析

完成 MLP 模型的建立、训练和推理后，对其在测试集上的表现进行详细评估和分析。针对基于问题 1 分析结果训练的预测模型，得出结果，如图 13-14 所示。

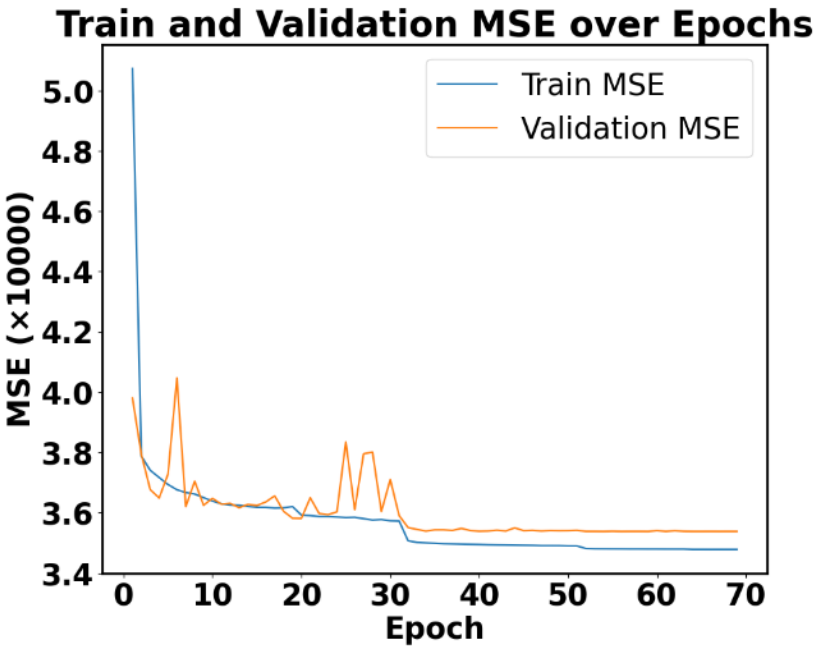


图 13 MLP 模型 MSE 结果

由图 13 显示了训练和验证的 MSE 随迭代次数（Epochs）的变化情况。训练 MSE 在最初几次迭代中迅速下降，随后趋于稳定，验证 MSE 则在前面几个 Epoch 内波动较大，但最终与训练 MSE 接近，且趋于平稳。表明模型在学习过程中成功拟合了训练数据，同时在验证数据上也具有良好的泛化能力，未出现明显的过拟合现象。

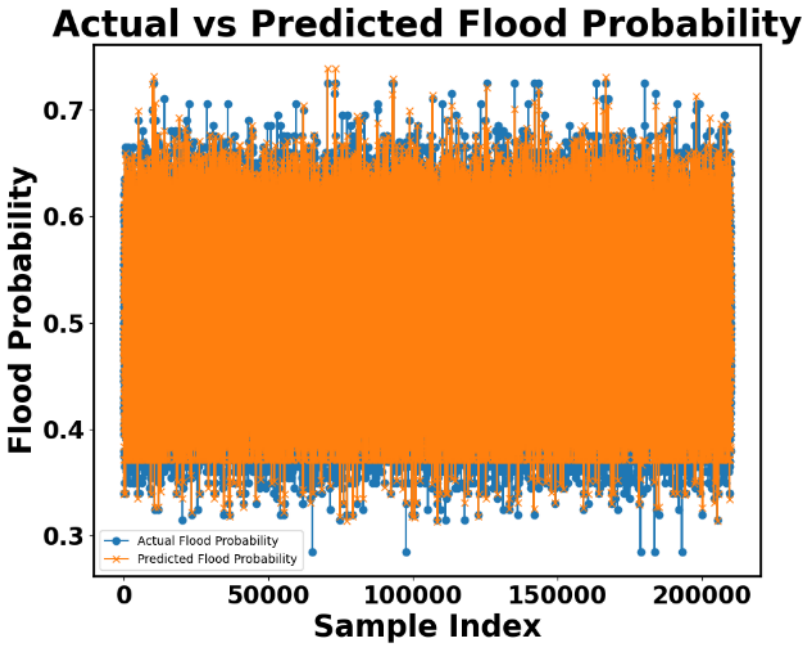


图 14 MLP 模型拟合结果

为进一步验证模型的预测性能和稳定性，由图 14 显示，预测值与真实值之间高度重合，表明模型在大多数情况下能够准确预测洪水发生的概率。结合图 13 可以得出模型在训练和验证阶段都表现良好，且在实际测试中保持了良好的预测性能，未出现明显的过拟合或欠拟合现象。模型的高拟合度和稳定性使其适用于洪水发生概率的预测。

6.3 关键指标下的模型调整与优化

根据问题 1 得到的 15 个指标，选择前 5 个作为关键指标特征（地形排水、大坝质量、季风强度、基础设施恶化、侵蚀）用于训练改进后的模型。

为了与调优后的模型有效对比，在原模型的基础上实现进行拟合验证，如图 15 所示。

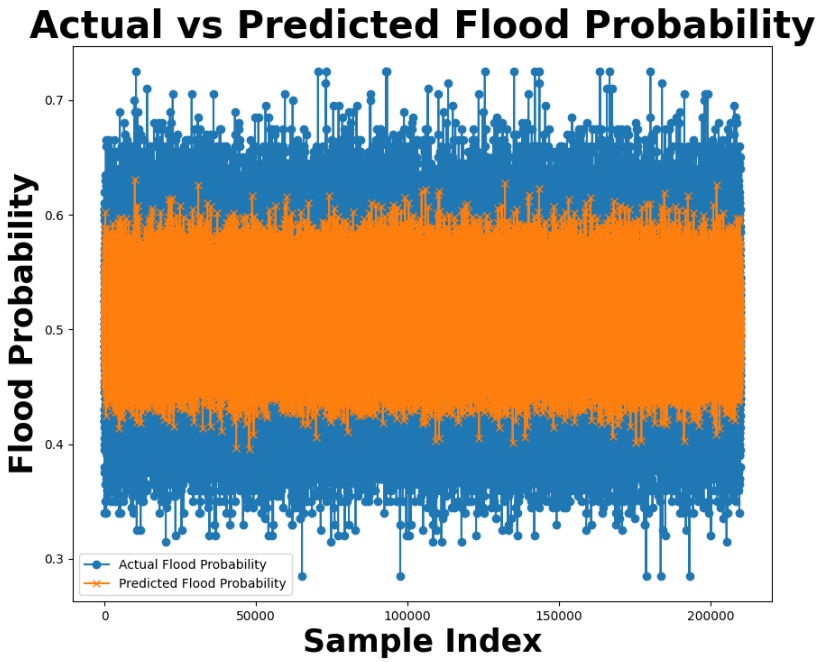


图 15 调优前的 MLP 模型拟合结果

由图 15 可以看出，模型预测值（橙色）在整体上较为集中，呈现出较为均匀的分布，主要集中在 0.4 到 0.6 之间；而实际值（蓝色）则表现出更大的波动范围，从 0.3 到 0.7 不等。尽管存在一些偏差，但总体上，模型的预测值与实际值有较好的重合度，表明模型在大多数情况下能够准确预测洪水发生的概率。然而，某些区域的实际值与预测值仍有显著差异，模型需要进一步优化。

因此，在上述建立的 MLP 模型的基础上改进模型，实现减少特征后仍能保持预测性能不变，主要从以下几个方面进行调整和优化：

调整网络结构

隐藏层数量和神经元数量：根据减少后的特征数量，减少隐藏层的数量和每层神经元的数量，以避免模型过于复杂。

正则化技术

在隐藏层中添加 Dropout 层，随机丢弃部分神经元，以增强模型的泛化能力。具体公式如下所示：

$$h_i = \begin{cases} \frac{h_i}{1-p} & \text{if } h_i \text{ is retained} \\ 0 & \text{if } h_i \text{ is dropped} \end{cases} \tag{31}$$

其中，h_i是第i个神经元的输出，p是 Dropout 的保留概率。

在损失函数中添加 L2 正则化项，以防止权重过大导致过拟合。具体公式如下所示：

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{m}\theta_j^2 \tag{32}$$

其中，$L(\theta)$是是损失函数，n是样本数量，λ是正则化强度超参数，θ_j是模型参数，m是参数数量。

超参数调优

学习率调整：通过网格搜索的方法，调整学习率以平衡训练速度和稳定性。

Batch 大小调整：根据数据量和模型复杂度，调整 Batch 大小以优化训练效率和模型性能。

完成以上调优后，观察拟合度评估模型性能，如图 16 所示。

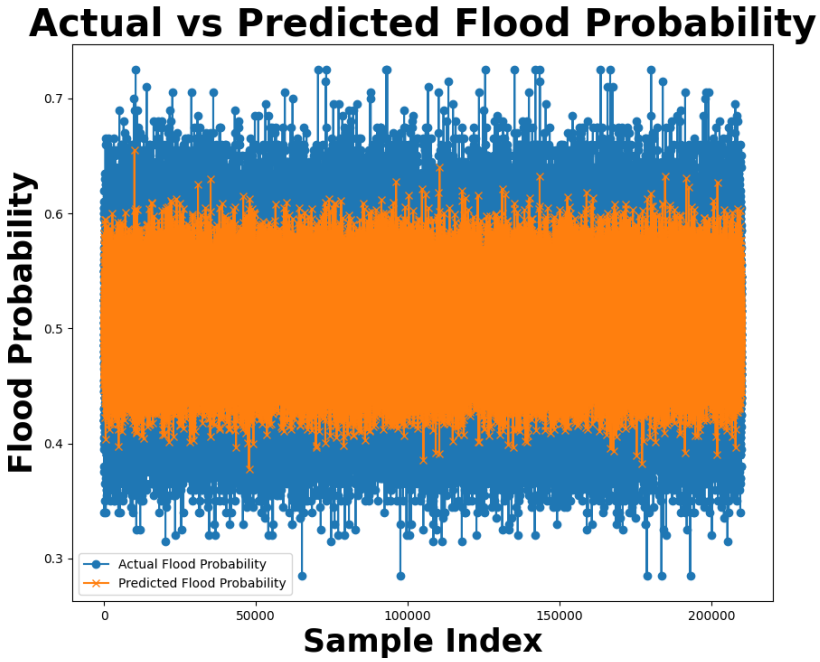


图 16 调优后的 MLP 模型拟合结果

由图 16 显示调优后的预测值和实际值更加紧密重合，表明模型在大多数情况下的预测准确度有所提高。尽管特征数量从 15 个减少到 5 个，模型的简化并没有显著影响其性能，这得益于对关键特征的优化选择和模型结构的调整。然而，由于特征数量减少，某些复杂因素可能未被完全捕捉，导致在部分区域仍存在一定的偏差。总体而言，调优后的模型在简化特征集的情况下，依然保持了较高的预测精度，证明了优化策略的有效性。

七、问题 4 的求解

7.1 洪水发生概率的数据预测

基于问题 3 中建立的洪水发生概率预测模型，预测附件 test.csv 中所有事件发生洪水的概率。首先对需预测数据进行与训练数据相同的预处理过程，处理后将测试数据输入至预测模型，得到每个事件的洪水发生概率并将结果保存到附件 submit.csv（见附录）。

7.2 预测结果的可视化与分析

针对已经预测出的数据进行可视化得到概率直方图和折线图，如图 17-18 所示。

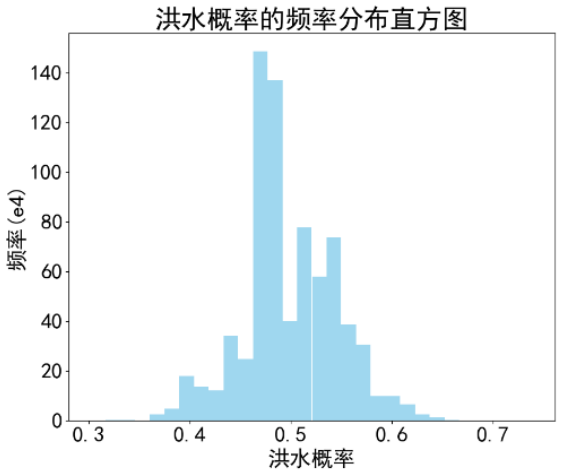


图 17 洪水概率的频率分布直方图

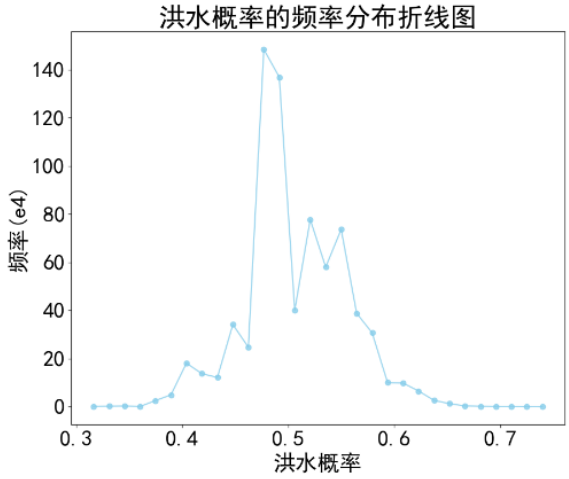


图 18 洪水概率的频率分布折线图

由图 17 显示了直方图的形式下的洪水发生概率的频率分布，频率分布呈现中心聚拢，并向两侧逐渐减少，峰值位于 0.50 附近，数据呈钟形分布。

由图 18 显示了折线图的形式下的洪水发生概率的频率分布，分布范围大致在 0.35 到 0.70 之间，同样地，峰值也出现在 0.50 左右，频率最高，呈现出一个大致对称的钟形分布。

通过直观观察两个图像，它们都符合正态分布的基本形态特征。此外，分布的均值接近 0.50，标准差适中，进一步验证了其接近正态分布的结论。

为了更准确地判断数据是否服从正态分布，选择 Shapiro-Wilk 统计（公式（33））进行检验，得到结论：p值>0.05，则可以认为数据服从正态分布。

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{33}$$

其中，$x_{(i)}$表示第i个排序的样本值，$\bar{x}$表示样本均值，a_i表示权重系数，由样本大小n决定。

综合上述方法，得出洪水发生概率的分布服从正态分布。

八、模型的评价及推广

模型优点

1、采用决策树回归和 XGBoost 回归模型，实现高效处理复杂数据。决策树回归模型通过树形结分裂构特征，捕捉复杂交互作用；XGBoost 回归模型通过集成多棵弱学习树并使用梯度提升方法，有效减小误差，提高洪水发生概率预测的准确性。

2、结合不同模型得出预测结果，确保关键指标在不同模型中均表现显著，使得结果具有较强的可信性。

3、通过树形结构和特征重要性分析，实现清晰地看到各个特征对预测结果的贡献，从而帮助理解模型的决策过程，便于用户信任和采用。

模型缺点

1、由于考虑的变量较多，模型中可能存在多重共线性问题，从而影响模型的稳定性和预测精度。

2、对于需要实时更新和处理的数据，现有模型可能需要频繁地重新训练，导致计算成本增加。

模型推广

1、文中的回归模型还可推广应用于其他自然灾害预测领域，如地震、台风等。

2、将文中的模型与大数据处理平台相结合，提高数据处理和模型训练的效率，利用分布式计算技术，实现对海量数据的实时分析和预测。

参考文献

[1]杨丹.区域洪水灾害风险特征及其演化态势分析[D].东北农业大学,2023.
[2]Costa, V.G., Pedreira, C.E. Recent advances in decision trees: an updated survey. Artif Intell Rev 56,4765–4800,2023.
[3]Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,785–794,2016.
[4]王森,刘琛,邢帅杰.K-means 聚类算法研究综述[J].华东交通大学学报, 39(05):119-126,2022.
[5]Parmar, A., Katariya, R., Patel, V. A Review on Random Forest: An Ensemble Classifier. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI),26:758-763,2018.
[6]张荣,李伟平,莫同 深度学习研究综述[J].信息与控制, 47(4):385-397,2018.

选题	2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
C		apmcm 24100023

基于 QUBO 模型的物流配送优化

摘要

本文研究了基于量子计算的物流配送问题，旨在通过量子计算技术优化物流配送方案，降低成本并提高效率。首先，本文介绍了量子计算方法的高效性，尤其是**相干伊辛机（CIM）**在解决复杂优化问题方面的优势。接着，本文以物流配送问题的数学模型为基础构建了 **QUBO 模型**，并利用 **Kaiwu SDK** 中的 **CIM 模拟器**和**模拟退火求解器**进行求解。最后，本文将 QUBO 模型的应用拓展到金融领域。

针对问题一：对于单个物流公司运营成本最小化问题，本文首先通过**迪杰斯特拉算法**求出最短路径，而后建立了一般数学模型并转化为 **QUBO 模型**，然后通过 **CIM 模拟器**和**模拟退火求解器**进行求解，得到了最佳货车租赁方案和货物运输方案，并计算出**公司一和公司二的最低总成本分别为 197500 元和 172000 元。**此外，本文还通过与**贪心算法**和**遗传算法**的对比，展现了 **QUBO 模型**的优势，包括计算速度快、精度高等。

针对问题二：在允许两公司之间合作拼货运输的条件下，本文在第一问的基础上更新了一般数学模型并转化为 **QUBO 模型**，同时分解为若干个 **SubQUBO 模型**利用 **CIM 模拟器**和**模拟退火求解器**求解，并用**贪心算法**和**遗传算法**加以比较验证，得到了最佳货车租赁方案和货物运输方案，并计算出**两公司合作时的最低总成本为 318000 元，比分别单独运输时节约了 51500 元**，由此证明了合作拼货运输可以降低两公司总成本，提高物流效率。

针对问题三：本文提出了一个金融领域基于 QUBO 模型的**信用评分卡优化场景**，并建立了对应的 QUBO 模型表达式。通过分析模型涉及的变量数量，计算得出模型所需的**比特数量级**，为后续研究和应用提供了参考。

本文将量子计算技术应用于物流配送优化问题，通过 **QUBO 模型**实现了最小化运营总成本的目标，并分析了合作拼货运输的优势，同时还利用**贪心算法**和**遗传算法**验证了 QUBO 模型的优越性。此外，本文还提出了一个具有学术价值和应用前景的金融领域的信用评分卡优化场景，为量子计算技术在更多领域的应用提供了思路。

关键词：物流配送优化、QUBO 模型、CIM、模拟退火、贪心算法、遗传算法

目　　录

一、问题重述与背景介绍

1.1 问题重述

电子商务的飞速发展使得物流配送需求日益增长。面对复杂的运输需求，传统的物流优化方法已不足以高效地解决物流问题，故而催生了新型的量子计算技术来自动计算运输综合策略，物流公司得以有更高效的方法来更有效地进行物流配送安排，实现物流效率的提高、运输成本的降低及消费者满意度的提升。

相干伊辛机在处理相关复杂优化问题中表现出巨大的潜力，故而本问题要求在物流配送场景下，将问题建立在与相干伊辛机密切相关的 QUBO 模型上，使用 Kaiwu SDK 进行求解。

问题一：在拼货只分别发生在各物流公司内部的前提下，以最小化单个物流公司的运营成本为目标，在总要求的方法下，使用其中的 CIM 模拟器和模拟退火求解器分别设计出两个物流公司的最佳货车租赁方案和货物运输方案。

问题二：在两公司之间可以合作拼货运输的前提下，以最小化两公司总成本为目标，用 Kaiwu SDK 中的 CIM 模拟器和模拟退火求解器得出最佳货车租赁方案和货物运输方案，并计算其相较于问题一中得到的方案对于两公司总成本的减少量。

问题三：提出一个具有学术价值或商业化前景的设想，给出其对应的 QUBO 模型表达式，并计算模型所需比特数量级。

1.2 背景介绍

1.2.1 二次无约束二进制优化(QUBO)模型：

1.QUBO 模型的定义：

QUBO 模型是指二次无约束二值优化模型，它是一种用于解决组合优化问题的数学模型。在 QUBO 模型中，需要将问题转化为一个决策变量为二值变量，目标函数是一个二次函数形式优化模型，一般定义为下式：

$$y = x^T Q x$$

其中，Q 为QUBO矩阵，x为二进制变量组成的向量，即 x 为0-1 变量，其取值为{0,1}，QUBO 目标为找到使得y达到最值的x。

Q 矩阵的形式有以下 2 种：

①对称形式

$$q_{ij}' = \frac{q_{ij} + q_{ji}}{2}$$

②上三角形式

$$q_{ij}' = (q_{ij} + q_{ji}), i \le j$$
$$q_{ij}' = 0, i > j$$

当添加约束条件后转化为以下形式：

$$\min f(x)$$
$$s.t. Ax = b$$
$$f'(x) = f(x) + P(Ax-b)^T(Ax-b)$$

其中，P 表示正标量惩罚项。随着问题规模增加，若仍使用传统算法求解组合优化问题，将会大大增加求解的时间，如若利用 QUBO 模型，即可运行在量子计算机硬件上，通过量子计算机进行毫秒级别的加速求解，得以更高效地求解组合优化问题。

2.QUBO 模型中约束条件的转化：

QUBO 模型为二次无约束优化问题，故如要实际解决一些有约束条件的问题，需要在目标函数中引入二次惩罚项，从而将组合优化问题重新进行表述为 QUBO 模型。同类型的约束关系可以用惩罚函数以非常自然的方式体现在“无约束”QUBO 公式中[1]。在 QUBO 公式中，使用惩罚函数可以产生比较精确的模型表示，一些经典约束的惩罚项转换实现如下：

经典约束	等效惩罚
$x + y \leq 1$	$P(xy)$
$x + y \geq 1$	$P(1 - x - y + xy)$
$x + y = 1$	$P(1 - x - y + 2xy)$
$x \leq y$	$P(x - xy)$
$x_1 + x_2 + x_3 \leq 1$	$P(x_1x_2 + x_1x_3 + x_2x_3)$
$x = y$	$P(x + y - 2xy)$

其中，函数 P 为正的足够大的标量惩罚值。惩罚项指定的规则是：对于最小化问题的求解，如果是可行解，即满足约束条件，对应的惩罚项等于零；对于不可行解，即不满足约束条件的解，其惩罚项等于一些正的惩罚量。P 值的选取是该惩罚项设置好坏的关键因素。

1.2.2 CIM 模拟器

相干伊辛机 (Coherent Ising Machine，简称 CIM)，是一种基于光学技术的量子优化装置，主要用于解决复杂优化问题。相干伊辛机与 QUBO 模型联系紧密，将优化问题转化为 QUBO 模型后，可运用基于相干伊辛机求解 QUBO 模型的软件开发套件——Kaiwu SDK 进行求解，具体操作 CIM 模拟器的流程如下：

步骤一：在 Python 3.8 环境下安装并导入 Kaiwu SDK 库。

步骤二：输入本文中经转化而得的 QUBO 模型。

步骤三：选择 CIM 模拟求解器，模拟在相干光量子计算机上求解 QUBO 模型。

步骤四：设置 CIM 模拟器的参数。

步骤五：调用 Kaiwu SDK 库中提供的方法，将构建好的 QUBO 模型传递给 CIM 模拟器进行求解。

步骤六：获取结果。由于变量为二进制形式，故求得结果为二进制形式。

步骤七：将结果转化表示为实际问题所需结果。

1.2.3 模拟退火算法

模拟退火算法（Simulated Annealing，SA）是一种通用的全局优化算法，用于在搜索空间中找到全局最优解。它的基本原理是模拟物理学中的退火过程，从某一较高初始温度出发，通过温度参数控制搜索过程中接受次优解的概率，以一定概率接受比当前解

更差的解，即在局部最优解能概率性地跳出并最终趋于全局最优，从而避免陷入局部最优[2]。模拟退火算法的具体思路如图 4 所示：

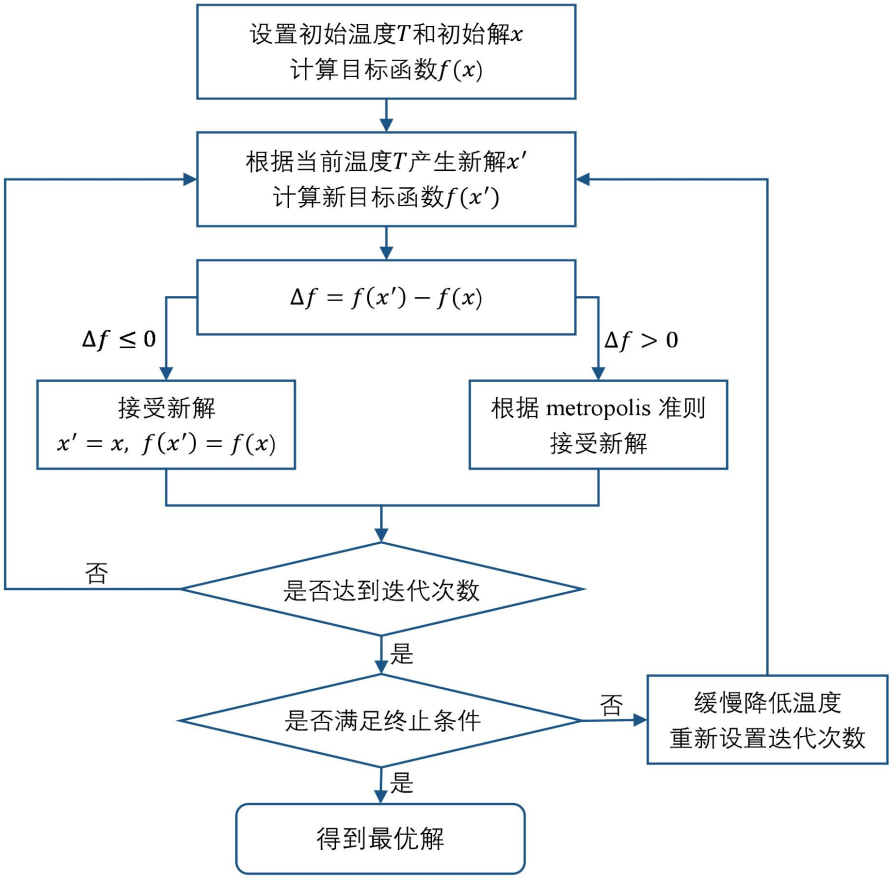


图 4 模拟退火算法流程图

关于流程图中部分内容的说明：

（1）Metropolis 准则

Metropolis 准则是一种有效的重点抽样法。其思想为：系统从一个当前能量状态变化到新的能量状态。若新的能量状态小于当前能量状态，则以概率 1 接受新的能量状态（取得局部最优）；若新的能量状态大于当前能量状态（全局搜索最优点），则以一定的概率接受或舍弃新的能力状态[3]。其数学性表示为：

$$p = \begin{cases} 1 & , \quad if\ E_{Tnew} < E_{Tcur} \\ \exp\left(-\dfrac{E_{Tnew} - E_{Tcur}}{T}\right), & if\ E_{Tnew} < E_{Tcur} \end{cases}$$

（其中E_{Tnew}表示当前温度 T 下新的状态能量，E_{Tcur}表示当前温度 T 下当前状态能量）

从方程可得，当新的能量状态大于当前状态能量时，温度越高，新状态的接受概率越高；当前状态的能量与新状态的能量差越高，新状态的接受概率越高。为避免所谓热力学中淬炼（quenching）的操作，控制参数 T 的值必须缓慢衰减。

（2）几个重要参数的选择

① 初始控制参数温度 T：

一般情况下采用随机生成一组可行解，以该解所对应的目标函数值的方差作为初温；利用经验或试验确定。

② 退温函数：

常用的退温函数有以下两种：

$$T_{n+1} = kT_n$$

$$T_{n+1} = T_n - \Delta T$$

其中 n 为当前循环次数，0<k<1 为一个非常接近 1 的常数，ΔT 为自定义温度下降值。

③ 内循环的循环次数：

每一个温度下的迭代次数相同，且迭代次数与具体问题有关。随着温度的下降，迭代次数应增加。

④ 算法终止准则：

主要包括以下三点：设置终止温度；设置最大迭代次数；算法搜索到的最优值连续若干步保持不变。

二、问题分析

2.1 问题总分析

本题为基于量子计算的物流配送问题，问题一及问题二均要求设计出最佳的物流配送方案，属于最优化问题。其中引入了量子计算技术以更好地进行物流方案设计，通过建立量子计算中的 QUBO 模型，而后代入基于相干伊辛机开发的 Kaiwu SDK 开发套件进行求解，得到最优物流方案，问题的关键点在于将问题的公式转化为 QUBO 形式，并调用 Kaiwu 库里的 CIM 模拟器和模拟退火求解器进行求解。问题的基本思路如图 1 所示：

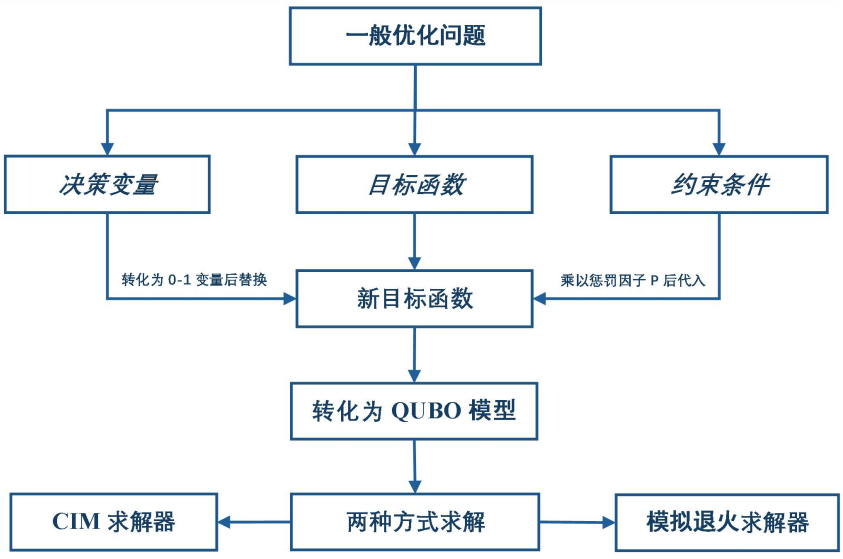


图 1 总体思路流程图

2.2 问题一的分析

针对问题一，判断其为优化问题。首先，通过观察与计算可得，存在两地之间直接运输的单趟成本高于经过另一地中转的单趟运输成本的情况，故而将各城市及其之间的

运输成本抽象为图中的点与权重，运用 Dijkstra 算法得到任意两点之间的最短路径和最短路径长度，从而得到卡车从城市i到城市j的运输总成本的最小值。而后通过对问题的分析，得到卡车 A、B 的最大租赁数量，将模型简化。接着，以最小化单个公司的运输成本为目标，将约束条件：①卡车载货量的重量限制；②存运货数量的等量关系；③对变量实际意义的约束转化为数学表达式，将其表示为一般优化模型，再通过对应关系，将其转化为 QUBO 模型，而后通过 python 程序编程，调用 Kaiwu SDK 库中的 CIM 模拟器和模拟退火求解器分别求解，得到最佳货车租赁方案、货物运输方案及航空运输的安排。再通过与一般优化模型中的遗传算法与贪心算法得到的结果及运算时间进行比较，证明本题所用的 QUBO 模型的先进性。

2.3 问题二的分析

针对问题二，判断其为与问题一类似的优化问题。基于问题一中对于最短路径的处理，增加公司一、二可在途中进行拼货运输的条件，则在运输过程中任一卡车上的货物可能含公司一或公司二的，也可能同时包含两公司的货物，即用$q_{ij,n}^{K,M}$表示公司M从城市i到城市j通过卡车 K 所运载的货物数量，在运输过程中，M可取 1 和 2。由于公司间拼货条件的增加，对于卡车 A 最大租赁数量的限制有所改变，进行新的数量分析，简化模型。本题的目标为最小化两公司总成本，约束条件除改变中途两公司间可以拼货的情况其余与问题一中的约束相同。而后将目标函数与约束条件转化为 QUBO 模型，判断出该模型的比特数较高，使用 SubQUBO 进行求解，调用 Kaiwu SDK 中的 CIM 模拟器和模拟退火求解器进行求解，得到最优安排。再分别计算问题一及问题二中两公司的运输成本总和，计算得出两公司合作下总体成本的减少量。

2.4 问题三的分析

针对问题三，其要求提出具有学术价值及商业化前景的场景。量子计算的优势之一是可以处理大规模的数据，考虑到此优势，通过对附件论文的研究及对资料的查询，本文将视野放至金融领域，对于银行在借款中所会面对的用户正常交纳利息及坏账的可能，选择对用户最佳的评价标准决定是否向该用户放贷，从而使银行获得最大利润。基于此问题背景列出公司收益的表达式，并将其转化为 QUBO 模型，并计算该模型所需的比特数量级。

三、模型假设

1. 假设每个城市有足够的可租赁卡车；
2. 假设货物可以暂存在城市中，由后续车辆将其运走；
3. 假设航空运输无论远近每吨运费皆相同，且能当日到达；
4. 假设问题一种公司一的货物运往小组一的城市，公司二的货物运往小组二的城市；
5. 假设运输天数与题目所给天数严格相等，不考虑交通、天气等不确定因素的影响；
6. 假设不考虑两公司相同城市的货仓间的距离；
7. 假设天数不足一天时均按照一天计算。

四、符号说明

符号	说明	单位
a_{ij}	卡车 A 从城市i到城市j的运输总成本	元
b_{ij}	卡车 B 从城市i到城市j的运输总成本	元
t_{ij}	卡车从城市i到城市j所需的运输天数	天
d_K	卡车 K 的日租金	元
e_{ij}	卡车从城市i到城市j的运输单趟成本	元
$x_{ij,\ n}^{K}$	第n辆K类型卡车是否参与城市i到城市j的运输$(i,j=1,\dots,6)$	/
$q_{ij,n}^{K,M}$	公司 M 卡车$x_{ij,n}^{K}$上运载的货物量$(i,j=1,\dots,6)$	吨
$c_{ij,M}$	公司 M 的航空运输量(此处$i=1,2,6;\ j=3,4,5$)	吨
C	航空运输每吨的成本(为常量 10000)	元/吨
T_{Mi}	公司 M 在出发城市i的初始货物量(此处$i=1,2,6$)	吨
D_{Mj}	公司 M 需运往到达城市j的货物量(此处$j=3,4,5$)	吨

五、模型建立与求解

5.1 问题一模型的建立与求解

5.1.1 数据预处理

由题中所给卡车日租金、卡车单趟时间及单趟成本表可计算得到 A、B 两种卡车在不同地之间运送货物的总成本，以 A 卡车从i地运输货物到j地为例，即：

$$a_{ij}=t_{ij}d_A+e_{ij}$$

运用 Excel 公式编写，计算得到两地点间卡车直达运输的总成本，结果如表格 1 所示：

表 1 卡车 A 在各地间的直达运输总成本

	上海	西安	昆明	深圳	天津	郑州
上海	0	28500	56500	36500	28500	18500
西安	28500	0	38500	40500	20500	8500
昆明	56500	38500	0	17500	61500	49500
深圳	36500	40500	17500	0	50500	37500
天津	28500	20500	61500	50500	0	10500
郑州	18500	8500	49500	37500	10500	0

通过对表格中数据的观察和计算可得，将i地货物运送到j地的过程中，存在比两地间直接运输更节约总成本的中转运输方式，即存在中转地v，使得：

$$a_{iv}+a_{vj}<a_{ij}$$

故而本文运用 Dijkstra 算法，不断更新最短路径和路径长度，更新得：

$$a_{ij} = a_{iv} + a_{vj}$$

最终得到任意两点之间的最短路径及路径长度a_{ij}，具体实现见附录一。为简化模型，此处地点i运输货物到地点j不考虑在中转地点v处的拼货情况。更新后的最短路径图如图 2、3 所示，最短路径矩阵如图 4 所示，最短路径长度即两地间最低运输总成本导出至 Excel 表后的结果如表 2、3 所示：

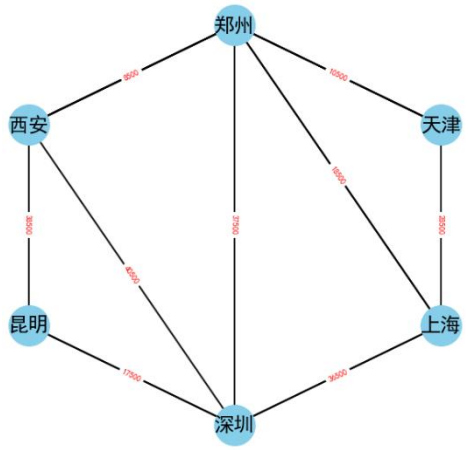


图 2 卡车 A 总费用最短路径图

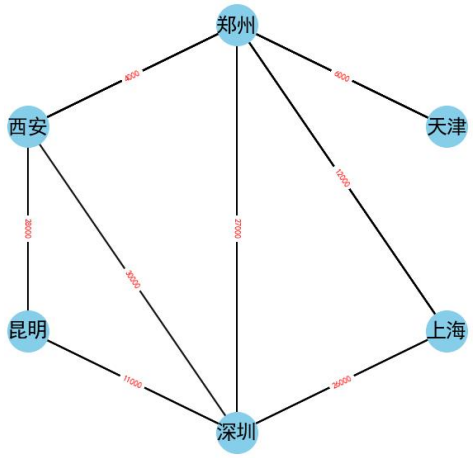


图 3 卡车 B 总费用最短路径图

$$\begin{pmatrix} 0 & 6 & 4 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 6 & 0 \\ 4 & 0 & 0 & 0 & 6 & 2 \\ 0 & 0 & 0 & 0 & 6 & 0 \\ 0 & 6 & 6 & 6 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \end{pmatrix}$$

图 4 最短路径矩阵

表 2 卡车 A 在各地间的最低运输总成本

	上海	西安	昆明	深圳	天津	郑州
上海	0	28500	54000	36500	28500	18500
西安	28500	0	38500	40500	19000	8500
昆明	54000	38500	0	17500	61500	47000
深圳	36500	40500	17500	0	50500	37500
天津	28500	19000	61500	50500	0	10500
郑州	18500	8500	47000	37500	10500	0

表 3 卡车 B 在各地间的最低运输总成本

	上海	西安	昆明	深圳	天津	郑州
上海	0	16000	37000	26000	18000	12000
西安	16000	0	28000	30000	10000	4000
昆明	37000	28000	0	11000	38000	32000
深圳	26000	30000	11000	0	33000	27000
天津	18000	10000	38000	33000	0	6000
郑州	12000	4000	32000	27000	6000	0

而后为了衡量最短路径的准确性，本文运用 Floyd 算法进行检验，通过比较得到两种算法的结果相同，故而认为所得结果即为最短路径。Floyd 算法的具体实现见附录 2。

5.1.2 构建数学模型和 QUBO 模型

一、模型抽象及简化

1．将城市抽象为字母

依题意，城市域 J={上海，西安，昆明，深圳，天津，郑州}，按顺序分别用数字 1-6 表示。其中上海、西安、郑州为货物当前所在城市，即出发城市，用T_{Mi}表示公司M在出发城市i的初始货物量，故此处$i=1,2,6$；昆明、深圳、天津为货物需要运往的城市，即到达城市，用D_{Mj}表示公司M需运往到达城市j的货物量，故此处$j=3,4,5$。由于拼货情况的存在，故每一个城市都可能作为发货点及收货点，设$x_{ij,n}^{K}$表示第n辆K类型卡车对于城市i到城市j的运输的参与情况，若参与运输则取 1，若无参与则取 0，并用$q_{ij,n}^{K,M}$表示公司 M中卡车$x_{ij,n}^{K}$上运载的货物量，此时$i,j=1,2,...,6$。

2．对$x_{ij,n}^{K}$进行分析

(1) $x_{ij,n}^{K}$表示货车 K 经最优路线由i城市发出到达j城市，使得运输成本达到最低，由最短路径矩阵可知，该路线可能会途径其他城市v，但该模型认为 K 货车在此运输过程不在v城市做停留。

(2) 对于 n：

① $K=A$时

由于到达城市中对于货量的最大需求为 27 吨，卡车 A 的最大载重量为 12 吨，派用两辆卡车 A 后还剩 $27-2\times12=3$（吨）<5（吨）的货物没有运输，此为用两辆卡车 A 后所剩货物量最多的情况，而此最大情况所剩的货物量都较小，此时使用卡车 B 或航空运输运送剩余部分的成本相较于使用卡车 A 的成本更低，故使用卡车 A 的最大数量为 2，即此时$n=1$，2。

② $K=B$时

由于两辆卡车 B 的最大运输量为 10 吨，小于一辆卡车 A 的最大运输量 12 吨，但两辆卡车 B 的租金大于一辆卡车 A 的租金，又因单趟运输成本相同，故有 $2b_{ij}<a_{ij}$，故租两辆卡车 B 的方案差于租一辆卡车 A 的方案，故不考虑租两辆卡车 B 及以上的情况，故所租卡车 B 的最大数量为 1，即此时$n=1$。

二、构建优化模型

由于本题要求分别求得公司一、二的最优运输安排，故以下仅表示出公司一的最优运输安排的求解过程，公司二的情况仅改变出发城市与最终收货城市的货量，其余处理方式相同。

（一） 决策变量

1. 卡车运输的决策变量

用$x_{ij,n}^{K}$表示第n辆卡车K是否参与城市i到城市j的运输，是则取 1，不是则取 0。

则$x_{ij,n}^{K}$，即卡车 A、B 的运输安排共同影响最终的运输成本。

2．航空运输的决策变量

除卡车运输外，还可进行航空运输，用$c_{ij,1}$表示公司一从城市i到城市j的航空运输量，其也是影响最终运输成本的因子之一。

（二）目标函数

本题目标为最小化运输成本，而运输成本由卡车运输与航空运输的成本构成，故而得到下式目标函数：

$$min \sum_{n}\sum_{i}\sum_{j} a_{ij}\, x_{ij,n}^{A} + \sum_{i}\sum_{j} b_{ij} x_{ij,1}^{B} + C\sum_{i}\sum_{j} c_{ij,1} \quad (n = 1,2)$$

（三）约束条件

1．对卡车载货量的约束：

由卡车 A 的最大载货量为 12 吨，卡车 B 的最大载货量为 5 吨，且载货量不小于 0，得以下二式：

$$0 \le x_{ij,n}^{A} q_{ij,n}^{A,1} \le 12 \quad (n = 1,2)$$

$$0 \le x_{ij,1}^{B} q_{ij,1}^{B,1} \le 5$$

2．对各城市存、运货物数量的约束：

因假设公司一的货物运往小组一的城市，公司二的货物运往小组二的城市，而公司一所含货物量与小组一所需货物量相等，公司二所含货物量与小组二所含货物量相等，故公司一、二的货物需全部运出，分别送至小组一和小组二，得以下等式约束：

$$T_{mi} = \sum_{j}\sum_{K} q_{ij,n}^{K,1} x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,1} x_{ji,n}^{K} + \sum_{j} c_{ij,1}$$

$$D_{mj} = \sum_{i}\sum_{K} q_{ji,n}^{K,1} x_{ji,n}^{K} - \sum_{i}\sum_{K} q_{ij,n}^{K,1} x_{ij,n}^{K} + \sum_{i} c_{ij,1}$$

$$(m = 1,2;\ K = A,B;\ 当 K = A\ 时, n = 1,2;\ 当 K = B\ 时, n = 1)$$

3．对卡车运输的约束：

$x_{ij,n}^{K}$表示第n辆卡车K是否参与i地到j地的货物运输,为 0-1 变量，即：

$$x_{ij,n}^{K} \in \{0，\ 1\}$$

具体表示为 $x_{ij,n}^{K} = \begin{cases} 1\ ,第n辆卡车K参与了i地到j地的货物运输 \\ 0\ ,第n辆卡车K没参与i地到j地的货物运输 \end{cases}$

4. 对航空运输的约束：

航空运输的数量必定为正值，故得到如下约束：

$$z_{ij} \ge 0$$

综合决策变量、目标函数、约束条件三点，可将所求问题转化为以下模型：

$$min \sum_{n}\sum_{i}\sum_{j} a_{ij}\, x_{ij,\ n}^{A} + \sum_{i}\sum_{j} b_{ij} x_{ij,1}^{B} + C\sum_{i}\sum_{j} c_{ij,1} \quad (n = 1,2)$$

$$s.t.\begin{cases} 0 \le x_{ij,1}^{B} q_{ij,1}^{B,1} \le 5 \\ 0 \le x_{ij,n}^{A} q_{ij,n}^{A,1} \le 12 \quad (n = 1,2) \\ T_{mi} = \sum_{j}\sum_{K} q_{ij,n}^{K,1} x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,1} x_{ji,n}^{K} + \sum_{j} c_{ij,1} \\ D_{mj} = \sum_{i}\sum_{K} q_{ij,n}^{K,1} x_{ij,n}^{K} - \sum_{i}\sum_{K} q_{ji,n}^{K,1} x_{ji,n}^{K} + \sum_{i} c_{ij,1} \\ x_{ij,n}^{K} \in \{0，1\} \\ c_{ij,1} \ge 0 \end{cases}$$

$$(m = 1,2;\ K = A,B;\ 当K = A\ 时, n = 1,2;\ 当K = B\ 时, n = 1)$$

公司二的处理方法相同，仅改变初始数据。

三、将优化模型转化为 QUBO 模型

对于目标函数中的变量$x_{ij,\ n}^{A}$与$x_{ij,\ n}^{B}$，其已为二进制变量，无需进行转化；而c_{ij}为非二进制变量，故对其进行离散化处理，使其变为二进制变量。

对于约束条件

$$\begin{cases} 0 \le x_{ij,1}^{B} q_{ij,1}^{B,1} \le 5 \\ 0 \le x_{ij,n}^{A} q_{ij,n}^{A,1} \le 12 \quad (n = 1,2) \\ T_{mi} = \sum_{j}\sum_{K} q_{ij,n}^{K,1} x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,1} x_{ji,n}^{K} + \sum_{j} c_{ij} \\ D_{mj} = \sum_{i}\sum_{K} q_{ij,n}^{K,1} x_{ij,n}^{K} - \sum_{i}\sum_{K} q_{ji,n}^{K,1} x_{ji,n}^{K} + \sum_{i} c_{ij} \\ x_{ij,n}^{K} \in \{0，1\} \\ c_{ij} \ge 0 \end{cases}$$

对于等式约束，可将其移到等式一边，再转化为平方形式，将转化后的等价约束乘以一个较大的惩罚量 P 加入目标函数中。对于不等式约束，首先向不等式约束中引入松弛变量将其转化为等式约束，而后同上述方法处理。

为了使记法与应用程序保持一致，本文首先将对于变量进行重新编号，具体如下：

$$\left(x_{12,1}^{A}, x_{12,2}^{A}, x_{12,1}^{B}, \ldots\ldots, x_{65,1}^{B}\right) = (x_1, x_2, \ldots\ldots, x_{90})$$

故 QUBO 模型的转化如下：

$$\min \sum_{n}\sum_{i}\sum_{j} a_{ij}\, x_{ij,\ n}^{A} + \sum_{i}\sum_{j} b_{ij} x_{ij,1}^{B} + C\sum_{i}\sum_{j} c_{ij,1}$$
$$- \sum_{i=1,2,6} P\left(\sum_{j}\sum_{K} q_{ij,n}^{K,1}\, x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,1}\, x_{ji,n}^{K} + \sum_{j} c_{ij} - T_{mi}\right)^{2}$$
$$- \sum_{j=3,4,5} P\left(\sum_{j}\sum_{K} q_{ij,n}^{K,1}\, x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,1}\, x_{ji,n}^{K} + \sum_{j} c_{ij} - D_{mj}\right)^{2}$$
$$- \sum_{i}\sum_{j} P\left(x_{ij,n}^{A} q_{ij,n}^{A} + x_{91} + 2x_{92} + 4x_{93} - 12\right)^{2}$$
$$- \sum_{i}\sum_{j} P\left(x_{ij,n}^{B} q_{ij,n}^{B} + x_{94} + 2x_{95} - 5\right)^{2}$$

又因$x = \{0,1\}$，故有$x^2 = x$，

故上式可转化为 QUBO 模型的标准形式：

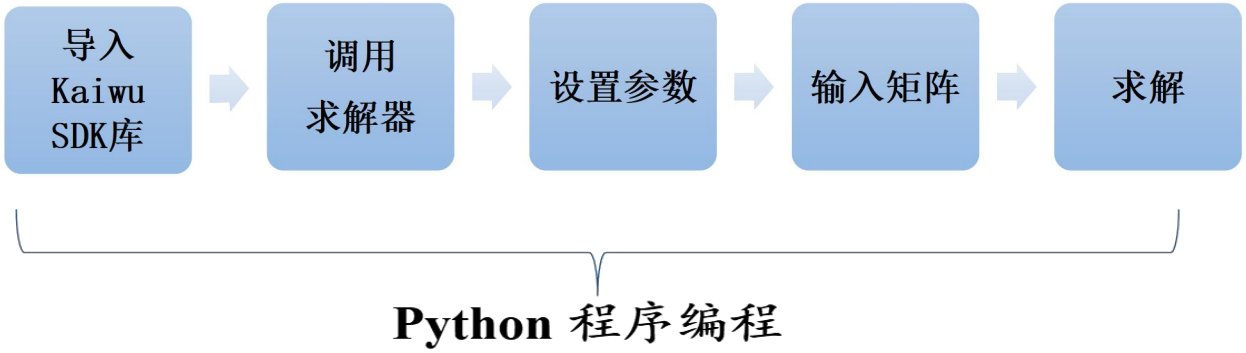
$$\min\ x^T Q x + c_0$$

其中c_0为一常数。

5.1.3 模型的求解

（1）调用 CIM 模拟器和模拟退火求解器

使用 CIM 模拟器和模拟退火求解器对 QUBO 模型问题的求解思路大致相同，如图所示：



通过程序运行得到结果，对比得两种求解器所得结果相同，结果如表 4、表 5 所示：

表 4 公司一最优方案

公司一最优货车租赁方案和货物运输方案（吨）							
发货地 \ 收货地	上海	西安	昆明	深圳	天津	郑州	成本（元）
上海				A(12)，A(7)			73000
西安			A(12)，A(7)	飞（1）			87000
昆明							
深圳							
天津							
郑州				B(5)	A(7)		37500
总成本（元）	197500						

表 5 公司二最优方案

公司二最优货车租赁方案和货物运输方案（吨）							
收货地 / 发货地	上海	西安	昆明	深圳	天津	郑州	成本（元）
上海			飞（1）	A(12)	飞（1）		56500
西安			A(12)，A(12)				77000
昆明							
深圳			B(2)				17500
天津							
郑州					A(12)，A(6)		21000
总成本（元）	172000						

以上两表格中，A(a)表示通过 A 货车运载 a 吨货物，从发货地发往收货地；B(b)表示通过 B 货车运载 b 吨货物，从发货地发往收货地；飞(c)表示通过航空运输的方式运载 c 吨货物，从发货地发往收货地。具体代码实现见附录 3。

（2）贪心算法

贪心算法的基本思路是从问题的某一个初始解出发一步一步地进行，根据某个优化测度，每一步都要确保能获得局部最优解。每一步只考虑一个数据，其选取应该满足局部优化的条件。若下一个数据和部分最优解连在一起不再是可行解时，就不把该数据添加到部分解中，直到把所有数据枚举完，或者不能再添加算法停止。[4]

本问题运用贪心算法的基本步骤如下：

Step1) 初始化： 创建两个列表，分别存储公司一和公司二的运输方案，初始为空。

Step2) 循环遍历： 遍历所有可能的运输路径，计算每条路径的运输成本。

Step3) 选择最优路径： 在每轮循环中，选择运输成本最低的路径作为当前公司的运输方案，并将该路径的成本累加到总成本中。

Step4) 更新剩余货物量： 更新出发城市和到达城市的剩余货物量。

Step5) 重复步骤 2-4： 直到所有货物都被运输完毕。

Step6) 输出结果： 输出公司一和公司二的运输方案以及总成本。

该算法（具体实现见附录 6）所得最终结果与（1）中调用 CIM 模拟器和模拟退火求解器所得结果一致。

（3）遗传算法

遗传算法是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。该算法通过数学的方式,利用计算机仿真运算,将问题的求解过程转换成类似生物进化中的染色体基因的交叉、变异等过程。在求解较为复杂的组合优化问题时,相较于常规的优化算法,通常能够较快地获得较好的优化结果[5]。

本问题运用遗传算法的基本步骤如下：

Step1) 编码：将运输方案编码成染色体，例如每个染色体可以表示为 [i, j, k, n]，其中 i 和 j 分别表示出发城市和到达城市，k 和 n 的组合分别表示公司一和公司二使用的卡车类型。

Step2) 初始种群：随机生成一定数量的初始种群，每个种群代表一个运输方案。

Step3) 适应度函数：定义适应度函数，计算每个染色体的运输成本，运输成本越低，适应度越高。

Step4) 选择：根据适应度函数，选择一定数量的优秀染色体作为下一代种群的父代。

Step5) 交叉：对父代染色体进行交叉操作，产生新的染色体。
Step6) 变异：对新的染色体进行变异操作，增加种群的多样性。
Step7) 迭代：重复步骤 4-6，直到达到终止条件，例如迭代次数达到上限或适应度函数值不再提升。

该算法（具体实现见附录 7）所得最终结果与（1）中调用 CIM 模拟器和模拟退火求解器所得结果一致。

5.1.4 问题一模型检验

一、各种求解方法之间的比较

表 6 公司一最小总成本求解结果比较

求解器	最小总成本	迭代次数	迭代时间
CIM 求解器	197500	10	25.765423s
模拟退火求解器	197500	10	24.369785s
贪心算法	197500	15	60.4605s
遗传算法	197500	20	80.4964s

表 7 公司二最小总成本求解结果比较

求解器	最小总成本	迭代次数	迭代时间
CIM 求解器	172000	10	23.469723s
模拟退火求解器	172000	10	24.134389s
贪心算法	172000	15	59.7645s
遗传算法	172000	20	81.6794s

二、QUBO 模型与非量子化算法的对比

根据所列出的最优化问题的目标函数与约束条件，本文还使用遗传算法及贪心算法进行求解（具体实现见附录 6、7），将这两者优化方案与转化为 QUBO 模型利用量子计算技术所得结果进行比较，发现后者的运算速度高于前者，故证明了此题所运用的 QUBO 模型的优越性。具体表现中的优缺点如表格所示：

	QUBO 模型	遗传算法	贪心算法
优点	计算速度快、精度高	有较强的全局搜索能力	时间复杂度低
缺点	技术潜力仍待挖掘	大规模问题求解受限	可能陷入局部最优
优化效果比较	QUBO 模型>遗传算法>贪心算法		

5.2 问题二模型的建立与求解

5.2.1 问题二模型建立

一、模型抽象及简化

1. 抽象为数学模型

具体符号及字母的设置同问题一中做法。

2．对$x_{ij,n}^{K}$的分析

(1) $x_{ij,n}^{K}$表示货车 K 经最优路线由i城市发出到达j城市，使得运输成本达到最低，由最短路径矩阵可知，该路线可能会途径其他城市v，但该模型认为 K 货车在此运输过程不在v城市做停留。

(2) 对 n 的分析

① $K = A$时

由于该小问中允许公司间拼货形式的存在，故两公司为合作关系，可共同完成小组一及小组二中所需货物量的运输，计算的出货单个到达城市中对于货量的最大需求为 46 吨，卡车 A 的最大载重量为 12 吨，4×12 = 48 > 46，故使用卡车 A 的最大数量为 4，即此时$n = 1, 2, 3, 4$。

② $K = B$时

与问题一中分析相同，即此时$n = 1$。

二、构建优化模型

（一）决策变量

与问题一的决策变量相同，即卡车运输的二进制决策变量$x_{ij,n}^{K}$与航空运输的决策变量及航空运输量$c_{ij,M}$，其中$i,j \in \{1,2,3,4,5,6\}, m = 1,2; K = A, B;$ 当$K = A$ 时$, n = 1,2,3,4;$ 当$K = B$ 时$, n = 1$。

（二）目标函数：

最终目标是使得两公司的运输总成本最低，即目标函数为：

$$min \sum_{n}\sum_{i}\sum_{j} a_{ij}\, x_{ij,\ n}^{A,M} + \sum_{i}\sum_{j} b_{ij} x_{ij,1}^{B} + C\sum_{M}\sum_{i}\sum_{j} c_{ij,M}$$

（三）约束条件：

1. 对卡车运载货物数量的约束：

$$0 \le \sum_{M} x_{ij,n}^{A} q_{ij,n}^{A,M} \le 12 \quad (n = 1,2,3,4; M = 1,2)$$

$$0 \le \sum_{M} x_{ij,1}^{B} q_{ij,1}^{B,M} \le 5 \quad (M = 1,2)$$

（其中$q_{ij,n,m}^{K,M}$表示公司 M 从城市i到城市j通过卡车 K 所运载的的货物数量）

2. 各城市存、运货物数量约束：

$$T_{mi} = \sum_{j}\sum_{K} q_{ij,n}^{K,M}\, x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,M}\, x_{ji,n}^{K} + \sum_{j} c_{ij,M}$$

$$D_{mj} = \sum_{i}\sum_{K} q_{ij,n}^{K,M}\, x_{ij,n}^{K} - \sum_{i}\sum_{K} q_{ji,n}^{K,M}\, x_{ji,n}^{K} + \sum_{i} c_{ij,M}$$

$$(n = 1,2,3,4; M = 1,2)$$

T_{Mi}表示公司M在出发城市i的初始货物数量(此处$i = 1,2,6$)；D_{Mj}表示公司M需运往到达城市j的货物数量(此处$j = 3,4,5$)。

其余对卡车运输的约束及对航空运输的约束同问题一。

综合决策变量、目标函数、约束条件三点，可将所求问题转化为以下模型：

$$min \sum_{n}\sum_{i}\sum_{j} a_{ij}\, x_{ij,\ n}^{A,M} + \sum_{i}\sum_{j} b_{ij} x_{ij,1}^{B} + C\sum_{M}\sum_{i}\sum_{j} c_{ij,M}$$

$$s.t.\begin{cases} 0 \le \sum_{M} x_{ij,n}^{A} q_{ij,n}^{A,M} \le 12 \quad (n = 1,2,3,4; M = 1,2) \\ 0 \le \sum_{M} x_{ij,1}^{B} q_{ij,1}^{B,M} \le 5 \quad (M = 1,2) \\ T_{mi} = \sum_{j}\sum_{K} q_{ij,n}^{K,M}\, x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,M}\, x_{ji,n}^{K} + \sum_{j} c_{ij,M} \\ D_{mj} = \sum_{i}\sum_{K} q_{ij,n}^{K,M}\, x_{ij,n}^{K} - \sum_{i}\sum_{K} q_{ji,n}^{K,M}\, x_{ji,n}^{K} + \sum_{i} c_{ij,M} \\ x_{ij,n}^{K} \in \{0，\ 1\} \\ c_{ij,M} \ge 0 \end{cases}$$

$$(m = 1,2;\ K = A, B;\ 当 K = A\ 时, n = 1,2,3,4;\ 当 K = B\ 时, n = 1)$$

三、转化为 QUBO 模型

处理过程与上一问相同，首先，为使记法与应用程序保持一致，首先对于变量进行重新编号，具体如下：

$$\left(x_{12,1}^{A}, x_{12,2}^{A}, x_{12,3}^{A}, x_{12,4}^{A}, x_{12,1}^{B} \ldots\ldots, x_{65,1}^{B}\right) = (x_1, x_2, \ldots\ldots, x_{150})$$

接着向不等式约束中加松弛变量转化为等式约束，再将等式约束乘以惩罚项加入目标函数中，转化得：

$$\begin{aligned} min & \sum_{n}\sum_{i}\sum_{j} a_{ij}\, x_{ij,\ n}^{A,M} + \sum_{i}\sum_{j} b_{ij} x_{ij,1}^{B} + C\sum_{m}\sum_{i}\sum_{j} c_{ij,m} \\ & - \sum_{i=1,2,6} P\left(\sum_{j}\sum_{K} q_{ij,n}^{K,M}\, x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,M}\, x_{ji,n}^{K} + \sum_{j} c_{ij,M} - T_{mi}\right)^2 \\ & - \sum_{j=3,4,5} P\left(\sum_{j}\sum_{K} q_{ij,n}^{K,M}\, x_{ij,n}^{K} - \sum_{j}\sum_{K} q_{ji,n}^{K,M}\, x_{ji,n}^{K} + \sum_{j} c_{ij,M} - D_{mj}\right)^2 \\ & - \sum_{i}\sum_{j} P\left(x_{ij,n}^{A} q_{ij,n}^{A,M} + x_{151} + 2x_{152} + 4x_{153} - 12\right)^2 \\ & - \sum_{i}\sum_{j} P\left(x_{ij,n}^{B} q_{ij,n}^{B,M} + x_{154} + 2x_{155} - 5\right)^2 \end{aligned}$$

又由$x^2 = x$，可将模型转化为 QUBO 标准形式，即：

$$\min\ x^T Q x + c_0$$

此时由于最后建立的 QUBO 模型比特数较高，故考虑将一个较大的 QUBO 模型分解成多个子模型（subQUBO）进行求解，这是一种用于提高优化算法效率和可解性的策略[6]。

subQUBO 模型具体求解步骤如下：

Step1)将原问题的 QUBO 模型分解成多个 subQUBO 模型每个 subQUBO 模型只包含一部分的决策变量和约束条件。

Step2)定义 subQUBO 模型的决策变量、约束条件、目标函数。

Step3)使用 Kaiwu SDK 的模拟退火求解器和 CIM 模拟器来求解每个 subQUBO 模型。

Step4)更新原问题的解，根据每个 subQUBO 模型的最优解，更新原问题的解。

Step5)重复步骤 2 至步骤 4，直到所有 subQUBO 模型都被求解并更新了原问题的解。

5.2.2 问题二模型求解

(1)调用 CIM 模拟器和模拟退火求解器

利用 Kaiwu SDK 库中的 CIM 模拟器和模拟退火求解器进行求解，得到结果如表格 8 所示：

表 8 两公司合作运营时最优方案

两公司合作运营时最优的货车租赁方案和货物运输方案（吨）							
发货地 \ 收货地	上海	西安	昆明	深圳	天津	郑州	成本（元）
上海				A(12,0)，A(7,5)		A(0,9)	91500
西安			A(12,0),A(7,3) A(0,12),A(0,12)			B(1,0)	158000
昆明							
深圳							
天津							
郑州		B(0,3)		A(6,5)	A(0,12),A(7,5) B(0,2)		68500
总成本（元）	318000						

上表中，A(a1，a2)表示通过 A 货车运载一公司 a1 吨货物，二公司 a2 吨货物，从发货地发往收货地；B(b1，b2)表示通过 B 货车运载一公司 b1 吨货物，二公司 b2 吨货物，从发货地发往收货地；飞(c1，c2)表示通过航空运输的方式运载一公司 c1 吨货物，二公司 c2 吨货物，从发货地发往收货地。

与问题一中所得方案相比，此问题中两公司合作运营时的方案节约总成本为：

$$197500 + 172000 - 318000 = 51500（元）$$

(2) 贪心算法

问题二可以分解为以下三个子问题：

①路径选择问题：选择最优的运输路径，使得运输成本最低。

②货物分配问题：将货物合理分配到每辆卡车上，确保卡车装载量不超过其最大载货量。

③拼货方案问题：确定两公司之间如何进行拼货，使得总成本最低。

此时贪心算法求解思路：

①针对路径选择问题：使用贪心算法选择最优的运输路径，参考问题一的贪心算法代码。

②针对货物分配问题：对于每条路径，使用贪心算法将货物分配到卡车上，优先分配载货量大的卡车。

③针对拼货方案问题：对于每条路径，比较两公司单独运输和拼货运输的成本，选择成本更低的方案。

(3) 遗传算法

同（2）将该问题分解为三个子问题。本问题的遗传算法求解思路：

Step1)编码： 将运输方案编码成染色体，例如每个染色体可以表示为 [i, j, k, n, p]，其中 i 和 j 分别表示出发城市和到达城市，k 和 n 的组合分别表示公司一和公司二使用的卡车类型，p 表示是否拼货（0 表示不拼货，1 表示拼货）。

Step2)初始种群： 随机生成一定数量的初始种群，每个种群代表一个运输方案。

Step3)适应度函数： 定义适应度函数，计算每个染色体的运输成本，运输成本越低，适应度越高。

Step4)选择： 根据适应度函数，选择一定数量的优秀染色体作为下一代种群的父代。

Step5)交叉： 对父代染色体进行交叉操作，产生新的染色体。

Step6)变异： 对新的染色体进行变异操作，增加种群的多样性。

Step7)迭代： 重复步骤 4-6，直到达到终止条件，例如迭代次数达到上限或适应度函数值不再提升。

5.2.3 问题二模型检验

该小问同样使用贪心算法及遗传算法进行求解（具体实现见附录 10、11），将这两种算法所得结果及迭代效果与用 CIM 求解器及模拟退火求解器得到的结果进行比较，结果如表 9 所示：

表 9 两公司最小总成本求解结果比较

求解器	最小总成本	迭代次数	迭代时间
CIM 求解器	318000	30	55.526729s
模拟退火求解器	318000	30	53.945355s
贪心算法	318000	53	150.4113s
遗传算法	318000	55	183.5762s

可得使用 CIM 求解器及模拟退火求解器的求解最终结果与贪心算法及遗传算法的结果相同，但在迭代次数及迭代时间上展现出其优越性，故认为该模型良好。

5.3 问题三的模型

5.3.1 问题背景及实际意义

在银行信用卡或相关的贷款等业务中，对客户授信之前，需要先通过各类审核规则对客户的信用等级进行评定，只有通过评定的客户才能获得信用卡或贷款资格。而该审核规则过程实际是经过一重或者多重组合规则后对客户进行评分，将其称为信用评分卡。每个信用评分卡有多种阈值设置（但其中有且仅有一个阈值生效），不同的信用评分卡在不同的阈值下，对应着不同的通过率和坏账率。一般通过率越高，坏账率也会越高，反之，通过率越低，坏账率越低。[7]

对银行来说，通过率越高，通过贷款资格审核的客户数量就越多，相应的银行理论上从客户手中获得的利息收入就会越多，但高通过率就意味着高坏账率，而坏账意味着放贷资金损失的风险，因此银行最终的收入可以定义为：

$$\text{最终收入} = \text{贷款利息收入} - \text{坏账损失}$$

下表为查询资料[8]所得的部分信用评分卡及其对应的 10 个阈值，以及相应阈值对应的不同的坏账率和通过率的表格：

信用评分卡 1			信用评分卡 2			信用评分卡 3		
阈值	通过率	坏账率	阈值	通过率	坏账率	阈值	通过率	坏账率
1	6%	0.60%	1	7%	0.55%	1	4%	0.65%
2	13%	1.20%	2	12%	1.10%	2	14%	1.20%
3	26%	1.70%	3	24%	1.60%	3	22%	1.65%
4	37%	2.30%	4	35%	2.10%	4	34%	2.15%
5	44%	2.60%	5	47%	2.30%	5	45%	2.65%
6	52%	3.10%	6	54%	2.60%	6	53%	3.25%
7	63%	3.60%	7	68%	3.30%	7	66%	3.65%
8	76%	4.10%	8	71%	4.20%	8	74%	4.20%
9	81%	4.60%	9	86%	4.80%	9	85%	4.60%
10	93%	5.20%	10	91%	5.10%	10	97%	5.05%

本问的目标在 m 张信用评分卡中找出 1 张及一定的阈值，使得银行的借款业务得到最高收入。

5.3.2 符号设计

符号	说明	单位
A	贷款资金	元
q	银行贷款利息收入率	/
A_i	信用卡i的通过率	/
B_i	信用卡i的坏账率	/
x_i	是否采用第i张信用卡（是取 1，不是取 0）	/

5.3.3 模型建立与求解

对于该问题，首先将问题等价转化为从 10m 张信用评分卡中找出一张信用卡，即将原问题中一张信用卡对应十个阈值转化为每一张信用评分卡对应一个阈值，从而实现模型的简化。将银行的最终收入转化为数学表达式得：

$$Aq\sum_{i=1}^{10m}A_ix_i(1-\sum_{i=1}^{10m}B_ix_i)-A\sum_{i=1}^{10m}A_ix_i\sum_{i=1}^{10m}B_ix_i$$

$$=\sum_{i=1}^{10m}(A_ix_i)\left(Aq-(A+Aq)\sum_{i=1}^{10m}B_ix_i\right)$$

$$=-(A+Aq)\sum_{i=1}^{10m}B_ix_i\sum_{i=1}^{10m}A_ix_i+Aq\sum_{i=1}^{10m}A_ix_i$$

本问目标是求得上式的最大值以达到最大利益，将上式取负后转化为求最小值问题，得到**目标函数**为：

$$(A+Aq)\sum_{i=1}^{10m}B_ix_i\sum_{i=1}^{10m}A_ix_i-Aq\sum_{i=1}^{10m}A_ix_i$$

由于最终只取唯一一张信用卡，故得到**约束条件**为：

$$\sum_{i=1}^{10m}x_i=1$$

该约束条件等价于：

$$(\sum_{i=1}^{10m}x_i-1)^2=0$$

转化为 **QUBO 表达式**为：

$$\min\ (A+Aq)\sum_{i=1}^{10m}B_ix_i\sum_{i=1}^{10m}A_ix_i-Aq\sum_{i=1}^{10m}A_ix_i+P(\sum_{i=1}^{10m}x_i-1)^2$$

其中 P 为惩罚项。

为了计算该模型中的比特数量级，我们需要确定模型中涉及的变量数量。根据模型假设，我们有 m 张信用评分卡，每张卡有 10 个阈值。因此，我们有 10m 个可能的阈值选择。因此：**比特数量级 = 210m**。

六、模型评价、改进与推广

6.1 模型的优点

1. 本文在构建优化模型前运用 Dijkstra 算法更新最短路径，实现更优，并用 Floyd 算法确定了其正确性。
2. 通过预处理数据分析得出卡车的最大数量，实现模型的简化。
3. 问题求解效率高: QUBO 模型能够利用量子计算的优势，在短时间内求解复杂优化问题，相较于传统算法具有显著的速度优势。
4. 结果精度高: QUBO 模型能够得到全局最优解或近似最优解，避免陷入局部最优，从而保证结果的精度和可靠性。
5. 模型适用性强: QUBO 模型可以应用于各种类型的组合优化问题，具有广泛的适用性。
6. 易于实现: 利用 Kaiwu SDK 等开发工具，可以方便地构建和求解 QUBO 模型，降低了量子计算技术的应用门槛。
7. 本文还通过贪心算法和遗传算法对 CIM 模拟器和模拟退火求解器的求解结果进行了比较验证，进一步证明了 QUBO 模型的优越性。

6.2 模型的缺点

1. 实际航空运输中的成本受距离的影响较大，而本文为简化模型将其设为一固定值。
2. 实际航空运输成本简化: 模型中将航空运输成本设为固定值，未考虑距离因素，与实际情况存在偏差。
3. 模型复杂度限制: 当问题规模较大时，QUBO 模型中的 Q 矩阵复杂度过高时会使算力减小，QUBO 模型的构建和求解可能会面临一定的挑战，需进一步采取 subQUBO 模型的分解。
4. 量子计算硬件限制: 目前量子计算机的规模和稳定性有限，限制了 QUBO 模型的实际应用范围。

6.3 模型的改进与推广

1、改进之处主要包括以下几点：

(1) 航空运输成本线性化: 可以根据距离对航空运输成本进行线性化处理，使模型更接近实际情况。
(2) 考虑时间窗约束: 可以在模型中加入时间窗约束，考虑卡车运输的时间限制，使模型更加完善。
(3) 考虑车辆维修保养: 可以在模型中加入车辆维修保养成本，使模型更加全面。

2、模型推广主要可以从以下几点考虑：

(1) 物流领域: 将模型应用于其他物流优化问题，如路径规划、库存管理、运输调度等。
(2) 金融领域: 将模型应用于金融优化问题，如投资组合优化、信用评分、风险管理等。
(3) 其他领域: 将模型应用于其他需要解决组合优化问题的领域，如生产调度、人员排班等。

七、参考文献

[1] Glover, F., Kochenberger, G., Hennig, R. *et al.* Quantum bridge analytics I: a tutorial on formulating and using QUBO models. *Ann Oper Res* **314**, 141–183 (2022).

[2] Kirkpatrick, S., Gelatt Jr, C. D. Vecchi, M. P. Optimization by simulated annealing. *Science*, 220(4598), 671-680(1983).

[3] Sunshing. 优化算法——模拟退火算法（Simulated Annealing, SA） https://blog.csdn.net/weixin_52721512/article/details/126497778/ 2024 年 4 月 24 日

[4] 常友渠,肖贵元,曾敏.贪心算法的探讨与研究[J].重庆电力高等专科学校学报,2008,13(3):40-424

[5] 郑树泉．工业智能技术与应用[M]．上海．上海科学技术出版社．2019．250-251

[6] Dury B, Di Matteo O. A QUBO formulation for qubit allocation[J]. arXiv preprint arXiv: 2009.00140, 2020.

[7] 陈粘. 高维金融市场的时变投资组合优化研究[D].成都理工大学,2021.

[8] 汪勇,孟香君,沈维萍.量子计算在经济与金融领域中的应用[J].经济学动态,2023(01):126-143.

选题	2024 年第十四届 APMCM 亚太地区大学生数学建模竞赛（中文赛项）	参赛编号
C 题		apmcm 24102422

基于量子计算的物流配送问题

摘要

随着电子商务的蓬勃发展，物流配送的复杂性和高效性要求日益提升。本文基于物流公司需应对多公司、多城市、多货物类型的复杂运输场景，基于 QUBO 模型研究物流公司物流配送最小化成本制定最优运输模型。首先，针对不同问题分别构建以最小化成本为目标的组合优化模型。其次，根据不同目标函数形式，通过采用启发式方法等数学技巧进行转换，转换成 QUBO 形式，并运用量子退火算法并通过 Kaiwu SDK 进行问题求解出使得最终成本最小化时的组合货车租赁方案。最后，本文运用罚函数对量子退火算法的结果进行对比检验分析，凸显量子退火算法优越性并总结不同算法优劣及适用场景我们综合了前三段的研究成果，强调了量子计算在物流优化与人工智能模型超参数调优中的重要作用。同时，将量子计算应用于人工智能模型的超参数调优，进一步展示了其在提升模型性能、加速科研进程方面的巨大潜力。

针对问题一：两个物流公司独立运营且拼货只在公司内部，以最小化单个物流公司运营成本为目标建立 QUBO 模型，使用 Kaiwu SDK 中的 CIM 模拟器和模拟退火求解器分别求解。**计算得出租赁 1 辆卡车 a 和 2 辆卡车 b 时总成本最小，经罚函数灵敏度分析确定罚参数并通过贪心算法检验结果合理性，**最终确定租赁 1 辆卡车 a 和 2 辆卡车 b 为货车租赁的最优方案及相应货物运输方案。

针对问题二：在问题一基础上修改参数，考虑运输成本等因素建立目标函数和约束条件，再进行物流转运中心选址以达到最短路径降低成本的目的。**运用数学规划方法和改进麻雀优化算法，使用 subQUBO 模型和量子退火算法，通过 KaiwuSDK 求解货物运输方案和成本，迭代确定最优物流中心坐标。**随机抽样比较总成本以验证量子退火算法的优越性，最终通过 subQUBO 方法和模拟退火算法得出卡车租赁和货物运输的最优方案及相关成本信息。

针对问题三：选取“人工智能模型超参数调优”场景，以超参数为决策变量、模型性能度量为目标函数构建 QUBO 模型，引入二进制变量并设计目标函数及引入惩罚因子，期望通过量子计算实现优化速度、超参数调整、高维空间问题解决、局部最优解克服等，通过数据探究得到 QUBO 求解算法在不同参数配置下的准确度分布情况，准确率总体较高且有变动。计算得出，**引入量子计算至 AI 超参数调优可以加速搜索过程，缩短调优时间，并精准定位最优超参数以提升模型性能与泛化能力。其特别优势在于有效处理高维空间优化难题，克服传统算法的局限，同时增强全局搜索能力，避免局部最优陷阱，增强模型稳定性。**

关键词：物流配送、量子计算、QUBO 模型、KaiwuSDK、CIM 模拟器、模拟退火求解器、subQUBO 模型、物流转运中心选址、麻雀优化算法、人工智能模型超参数调优。

目录

一、问题重述

1.1 研究背景

随着电商的蓬勃发展，物流配送需求激增，物流公司与电商平台紧密合作以应对挑战。面对庞大的货物量与多样化的目的地，传统物流优化方法显得力不从心。为此，物流公司探索采用量子计算技术这一前沿解决方案，旨在自动化计算运输综合策略[1]。量子计算凭借其强大的计算能力，有望为物流公司提供更精准、高效的路线规划与运输方式选择，确保货物准时送达，同时优化成本结构。此举不仅能显著提升物流效率，降低运营成本，还能增强消费者的购物体验与满意度，为电商行业的持续发展注入新动力。

借助量子计算技术来解决物流配送中的优化问题具有极其重要的现实意义和深远的应用价值[2]。它能够为物流公司提供更为科学、高效的决策支持，有力地推动物流行业朝着更加智能化、高效化的方向发展。这不仅有助于提升物流公司的竞争力，还能为整个电商行业的繁荣发展提供坚实的保障。

1.2 问题重述

问题一：假设两个物流公司独立运营，且拼货只发生在公司内部。以最小化单个物流公司的运营成本为目标，建立 QUBO 模型。使用 Kaiwu SDK 中的 CIM 模拟器和模拟退火求解器分别进行求解，为两个物流公司分别设计出货车租赁方案和货物运输方案。

问题二：当两个物流公司之间进行合作运营时，公司之间可以拼货运输，此时的优化目标为最小化两个公司的总成本。运用 Kaiwu SDK 中的 CIM 模拟器和模拟退火求解器来求解，给出最优的货车租赁方案和货物运输方案，并计算出合作运营带来的总体成本减少量。

问题三：自行提出一个具有商业化前景或学术价值的场景，该场景可以涉及 AI、通信、金融、生物医学、物流供应链管理等相关领域。需要给出相应的 QUBO 模型表达式，并计算模型所需的比特数量级（可以用相关参数表示）。

二、问题分析

2.1 问题一的分析

在这个问题中，我们假设两个物流公司是独立运营的，并且拼货只在各自公司内部进行。我们的目标是最小化单个物流公司的运营成本。为了实现这个目标，我们需要建立一个 QUBO 模型。这个模型将考虑到各种因素，比如货物的起始位置、目的地、卡车的类型和载重、租赁时长、运输路线以及运输成本等。通过对这些因素的综合分析，我们可以确定最佳的货车租赁方案和货物运输方案。在建立模型的过程中，我们需要用合适的数学表达式来描述这些因素之间的关系，并且确保模型能够准确地反映实际情况。然后，我们使用 Kaiwu SDK 中的 CIM 模拟器和模拟退火求解器来求解这个模型[3]。这两个求解器将根据模型的设定，尝试找到最优的解决方案。通过这个过程，我们可以为每个物流公司设计出最合理的货车租赁和货物运输计划，从而在满足运输需求的同时，最大限度地降低运营成本。

2.2 问题二的分析

当两个物流公司决定合作运营时，情况就发生了变化。此时，公司之间可以进行拼货运输，我们的优化目标也变成了最小化两个公司的总成本。为了达到这个目标，我们同样需要建立一个 QUBO 模型。这个模型将考虑到两个公司的所有货物和运输需求，以及它们之间的协同合作方式。在模型中，我们需要考虑如何合理地安排货物的运输路径，如何充分利用拼货运输的优势，以及如何优化卡车的租赁和使用，以达到总成本最小化的目的。然后，我们使用 Kaiwu SDK 中的 CIM 模拟器和模拟退火求解器来求解这个模型[4]。通过这两个求解器的计算，我们可以找到最优的货车租赁方案和货物运输方案，使得两个公司的总成本降到最低。同时，我们还可以计算出合作运营带来的总体成本减少量，这将有助于我们评估合作的效益和优势。

2.3 问题三的分析

在这个问题中，我们需要自行提出一个具有商业化前景或学术价值的场景。这个场景可以涉及到多个领域，比如 AI、通信、金融、生物医学、物流供应链管理等。然后，我们需要为这个场景建立一个相应的 QUBO 模型表达式。在建立模型表达式时，我们需要深入分析场景中的各种因素和关系，确定合适的变量和约束条件，并用数学语言来准确地描述它们。同时，我们还需要计算模型所需的比特数量级。这个比特数量级可以用相关的参数来表示，它将反映模型的复杂度和规模。通过提出这个场景和建立相应的模型，我们可以展示 QUBO 模型在不同领域的应用潜力和创新可能性，为解决实际问题提供新的思路和方法。

三、模型假设

1.假设每个城市有足够数量的可供租赁的卡车，且卡车的租赁和运输过程中不会出现故障或延误。

2.假设题目所给的货物起始城市、目的地城市、货物数量、卡车运输时间和成本等数据真实可靠，且在运输过程中保持不变。

3.假设仅考虑如何实现最小化运营成本或总成本的最优方案，不考虑其他因素对结果的影响。

4.假设物流公司在运输过程中，货物不会出现损坏或丢失等情况，且运输时间严格按照给定的时间表进行。

5.假设在计算成本时，只考虑卡车租赁费用、运输成本和航空运输费用，不考虑其他额外费用。

四、符号说明

变量	解释说明
A	表示 12 吨载重卡车的型号
B	表示 5 吨载重卡车的型号
P	表示飞机的型号
wki	表示第 k 种运输工具在数量为 i 时的二进制值
tki	表示第 k 种运输工具的数量，即：tki
cA	12 吨载重卡车的租赁价格，5000 元/天
cB	5 吨载重卡车的租赁价格，3000 元/天
cP	飞机的租赁价格，10000 元/吨
dA	12 吨载重卡车从上海到西安的单趟成本
dB	5 吨载重卡车从上海到西安的单趟成本

dP	货物从上海到西安的航空运价，假设为 10000 元/吨
Y	运输物流中投入使用的不同载量运输工具型号 a，b 数量
x	决策变量，用于表示是否租赁运输工具，0 表示不租赁，1 表示租赁
y	决策变量，用于表示货物是否通过航空运输到城市，0 表示不通过航空运输，1 表示通过航空运输
z	决策变量，用于表示其他相关决策，具体含义根据问题而定
q	货物的数量
s	运输工具的货箱容量
y	运输工具的效率
r	运输工具的油耗
p	运输工具的维护成本
i	表示迭代次数或其他相关计数
α	系统参数，在特定公式中使用
β	麻雀算法中的相关参数
Q	QUBO 模型中的系数矩阵
λ	正标量惩罚项，在 QUBO 模型中用于调整目标函数
Γ	量子退火算法中的场强，促使自旋状态发生转变
M	重心计算公式中的中间变量，用于表示麻雀种群的平均位置
Ne	能量水平
Ns	可解的个数
Ni	量子位的数量

五、问题一模型的建立与求解

综合上述问题分析，我们得到思路流程图如图 5.1 所示：

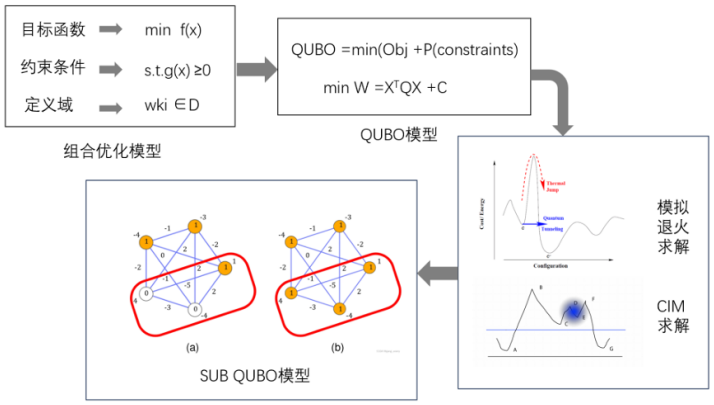


图 5.1 思路流程图

其中我们将 QUBO 模型看作一个灰盒，将题中表格的数据放入盒中针对具体问题进行求解。QUBO 模型求解具体流程图如图 5.1 所示：

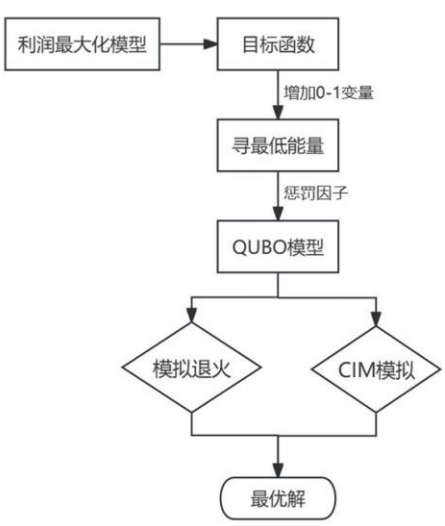


图 5.2QUBO 模型求解具体流程图

5.1 模型的建立

5.1.1QUBO 模型

QUBO 模型可以表述众多组合优化问题，本题中的最优租赁方案问题满足其要求。QUBO 无约束二次二进制优化模型作为量子计算中应用最为普遍的优化手段，它成功地整合了众多复杂的组合优化问题。当面临问题规模庞大、难以通过精确求解器找到“最优”解的情况时，该模型便能够派上用场。它采用启发式方法，在有限的时间约束下努力寻找尽可能接近最优的方案，为复杂问题的求解提供了一种高效且实用的途径。

QUBO 模型的最简易形式可表述为：

$$y_{min} = z^T P z \tag{5.1}$$

其中 z 作为二进制变量组成的向量，即当中每个变量的取值均为{0,1}，在 QUBO 模型中，为了使得约束条件得到更好的表示，我们引入二进制 0-1 变量。QUBO 模型是解决组合优化问题的数学形式，而在模型中，变量通常被限制为二进制（0 或 1），这是因为 QUBO 模型通常用于描述决策问题，其中每个变量代表一个决策变量，其取值为 0 或 1，分别表示不做某项决策或做某项决策。

QUBO 目标为找到使得 y 最小的 z。P 是常数的方阵，并且该常数方阵一般在最优化问题中，为对称矩阵或上/下三角形形式。

$$\begin{cases} q'_{ij} = (q_{ij} + q_{ji}), i < j \\ q'_{ij} = 0, \qquad\quad i > j \end{cases} \tag{5.2}$$

表 5.1 租赁价格

型号	12 吨载重卡车	5 吨载重卡车	飞机
租赁价格	5000	3000	10000

根据此表，按对应货物数量计算卡车租赁方案的最小成本。

假设w_{ki}为第 k 种卡车在数量为 i 时的二进制值，t_{ki}表示第 k 种卡车的数量，即：$t_{ki} = i$。因此，得到约束函数表达式：

$$\sum_{i=1}^{24} 100w_{1i}t_{1i} + 140w_{2i}t_{2i} + 200w_{3i}t_{3i} + 320w_{4i}t_{4i} = 2400 \tag{5.3}$$

将约束条件转化为：

$$\begin{gathered} minf(z) \\ s.t.Az = b \\ f'(z) = f(z) + \lambda(Az - b)^T(Az - b) \end{gathered} \tag{5.4}$$

其中 λ 为正标量惩罚项,为一个足够大的数。

随问题规模的扩大，传统计算方法无法有效地计算出结果，此时使用适配于相干伊辛机(CoherentIsingMachine,CIM)的 QUBO 模型，能够在量子计算机硬件上进行毫秒级加速，从而实现最优解结果的快速得出。

5.1.2QUBO 模型的建立

根据题意，我们知道约束条件包括对于每个物流公司租赁卡车 a、卡车 b 或飞机 p 的控制最低成本限制，以及最短路径的限制。是否租赁卡车 a、b，货物是否通过航空运输到城市，0、1 分别选与不选卡车或飞机，定义决策变量用于表示：

$$x_i, y_{i,} z_i = \begin{cases} 1 & 选i型卡车或飞机 \\ 0 & 不选i型卡车或飞机 \end{cases} \tag{5.5}$$

并结合选择卡车 a 或卡车 b 的数量及其对应的成本与运输时间折现值，可建立卡车租赁成本、航空运输成本和罚函数项的目标函数：

$$maxQ = \sum_{i=1}^{4} x_i y_i q_i \tag{5.6}$$

将其转化为最小化目标函数$-Q$,即：

$$min(-Q) = -\sum_{i=1}^{4} x_i y_i q_i \tag{5.7}$$

约束条件由如下六点组成：

1.货物的起始地.目的地和数量是已知的且固定的，且货物必须从起点运送到终点。

2.每个城市可供租赁的卡车数量是足够的，但具体车型和数量需要根据成本和需求来选择。

3.卡车的运输时间和成本是给定的，需要在安排运输路线时考虑。

4.航空运输的运价是固定的，且当日到达，但需要权衡成本和时间因素。

5.各小组间可在任何一个城市拼货和中转运输，但需要合理安排以降低成本。

6.总成本需要在一定的预算范围内，或者需要尽可能地降低成本。

(1)运输工具的最大载重：

$$\begin{cases} \sum_{i=1}^{n} c_i x_i \leqslant 5000 i_+ \\ \sum_{i=1}^{n} c_i y_i \leqslant 3000 i_+ \\ \sum_{i=1}^{n} c_i z_i \leqslant 10000 pi \end{cases} \tag{5.8}$$

(2)一次性运输货物时，卡车租赁总资金不能大于租赁飞机总资金的成本：

$$\sum_{i=1}^{n} n\, c_i x_i + \sum_{i=1}^{n} n c_i y_i \leqslant 10000 npi \tag{5.9}$$

(3)卡车 a，卡车 b 与飞机 p 的数量必须为自然数：

$$x_i \in N \tag{5.10}$$

(4)由决策变量的定义可得:

$$y_i \in \{0,1\} \tag{5.11}$$

综上，约束条件为：

$$s.t.\begin{cases} \sum_{i=1}^{4} y_i \geqslant 3 \\ \sum_{i=1}^{4} c_i x_i \leqslant 2400 \\ x_i \in N \\ y_i \in \{0,1\} \end{cases} \tag{5.12}$$

为了将上述二次有约束二进制优化模型改写为二次无约束二进制优化模型（QUBO），需要对目标函数添加惩罚：

$$g_1 = \left\{\max\left|3 - \sum_{i=1}^{3} j_i, 0\right|\right\}^2, k_1 = \left\{\max\left|\sum_{i=1}^{3} c_i x_i - 2400, 0\right|\right\}^2 \tag{5.13}$$

故惩罚项为:

$$\lambda(g_1 + k_1) \tag{5.14}$$

其中λ为罚参数，一个足够大的数。

因此最终 QUBO 模型为：

$$minz = -\sum_{i=1}^{3} x_i y_i q_i + \lambda(g_1 + k_1) \tag{5.15}$$

5.1.3 罚函数灵敏度分析

具体步骤：

1.将该模型的无约束优化问题表示为：$minz(x_i, \lambda) = min(-Q) + \lambda(g_1 + k_1)$
2.选择罚参数λ,设置一个基准值λ_0
3.使用优化算法计算在该罚参数值λ_0下的目标函数值$min(-Q)$,得到最优解x_i^*及对应目标函数值$min(-Q)^*$。
4.改变其罚参数值,例如$\lambda_{\alpha 0} = \lambda_0 + \Delta\varepsilon$，$\varepsilon$是一个小的正数。
5.使用优化算法求解新的无约束优化问题：$\min z = \min(-Q) + \lambda_{\alpha 1}(g_1 + k_1)$,其中$\lambda_{\alpha 1}$为改变后的罚参数值，得到新的最优解$x_i'$和对应新的目标函数值$\min(-Q)'$。
6.计算灵敏度，灵敏度可以通过计算罚函数值的变化量与罚参数的变化量之比来衡量，即 $\frac{min(-Q)^* - min(-Q)'}{\lambda_0 - \lambda_{\alpha 1}}$
7.根据需要，需要进行多次改变罚参数的值，并计算相应的灵敏度，以全面了解罚函 数的特性。

这个灵敏度指标反映了罚函数对罚参数变化的敏感程度。例如，如果灵敏度值较大，说明罚函数对罚参数的变化较为敏感；如果灵敏度值较小，则说明罚函数对罚参数的变化不太敏感。

根据上述步骤求解得λ=600000 时，较为合适。

5.2 模型的求解

5.2.1 量子计算

量子退火算法是一类新的优化算法，它是模拟退火算法的一种延伸和改进。与经典模拟退火算法利用热力学中的热波动来搜索问题的最优解不同，量子退火算法利用量子隧穿效应即量子因为测不准原理而具有的穿透比其自身能量高的势垒的能力，从而使算法摆脱局部极值，以更高概率逼近全局最值[5]。如图 5.3 所示：

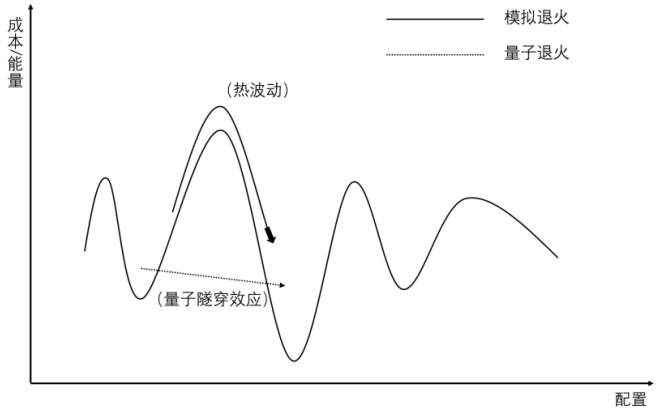


图 5.3 量子退火算法与模拟退火算法过程比较图

量子退火算法模型主要有两大核心要素。首先，是势能，它的主要作用在于实现量子优化问题与量子系统之间的映射关系。换言之，它将我们所追求的优化目标函数，转化为一个作用于量子系统的势场。其次，是动能，它通过引入一个可以调控的动能项，扮演了调节量子波动穿透场的角色。量子系统在这两个场的协同影响下的演化过程可以通过薛定谔方程（或狄拉克方程）来精确描述：

$$i\hbar\frac{d}{dt}|\psi(t)\rangle = H(t)|\psi(t)\rangle \tag{5.16}$$

因实际情况下难以直接求解，故采用有效随机过程方法如路径积分蒙特卡罗(PIMC):

$$E(t) = E_m(t) + E_n(t) \tag{5.17}$$

在量子退火算法中，量子哈密顿函数$E(t)$作为评价函数，精准评估量子系统状态。其中，$E_m(t)$表示势能，反映能量分布；$E_n(t)$代表动能，引入量子波动效应，通过逐步减小初始值，引导系统逼近全局最优解。

量子退火算法常用横向场随机伊辛模型作为测试基准。诸多组合优化问题首先需映射至伊辛模型，随后运用量子退火算法求解。该模型的哈密顿函数是算法关键：

$$E_p = \sum_{a=1}^{n} e_a\delta_a^x + \sum_{a,b=1}^{n} I_{ab}\delta_a^x\delta_b^x \tag{5.18}$$

其中，能量偏移角度以e_a表示，泡利自旋矩阵以δ_a^x表示，自旋量 a 和 b 之间的耦合度以I_{ab}表示。

依据伊辛模型的哈密顿函数, 从而获得量子退火的哈密顿函数为：

$$E(t) = E_p + \Gamma(t)\sum_{a=1}^{n}\Delta_a\delta_a^y \tag{5.19}$$

在量子退火算法中，场强Γ促使自旋状态发生转变，类似于模拟退火中温度 T 的作用。因此，任何采用上述形式表示的优化问题，均可借助量子退火算法得到解决。

5.2.2 算法设计与求解

QUBO 模型适配相干伊辛机并能转化为伊辛模型，因此设计量子计算算法，借助特定求解器如 Kaiwu SDK 进行求解，从而提高效率。

算法 1
Step1:根据最优租赁方案情境将其剖析为组合优化问题，构造函数。 Step2:初始化各个参数，其中计数参数 i=1，取量子退火的初始温度 T_0 足够大，Γ为横向场强，确定最大迭代次数为 MaxSteps,初始化解为 x,对应的状态能量为$E_m(t)$。 Step3:通过随机将一个决策变量由 0 变 1 或由 1 变 0 产生随机微扰，从而产生新状态x',其对应的状态能量为$E_m(x')$。 Step4:计算能量差$\Delta E_m = E_m(x') - E_m(t)$以及 $\Delta E_n = E_n(x') - E_n(t)$,如果$\Delta E_m < 0$或者$\Delta E_n < 0$,则系统接受新解,反之，计算$x'$的接受概率$exp(\Delta E_n/T)$,并产生$(0,1)$区间上均匀分布的随机数 $rangdom(0,1)$。若$exp(\Delta E_n/T) < rangdom(0,1)$,则$x = x'$。否则,重复执行步骤 3。 Step5:进行退温操作,Γ的变化和模拟退火中的温度 T 作用类似,横向场强变化形式为$\Gamma = \Gamma - (\Gamma_0/MaxSteps)$。 Step6:判断是否满足终止条件Γ=0,如果满足,量子退火算法终止,否则,将重复步骤 3。

基于上述算法，使用 KaiwuSDK 内置的模拟退火求解器和 CIM 模拟器进行求解得，当租赁 1 辆卡车 a，2 辆卡车 b 时的总成本最小，为 58000 万元。

5.2.3 问题一的求解与最优方案

根据已知条件，可计算得出两种型号卡车的性价比大小关系为 a>b。

利用贪心算法求解进行检验。

首先，需先上文求解所得的结果相同，故结果合理，因此当租赁 1 辆卡车 a，2 辆卡车 b 时的总成本最小，为 58000 万元。

5.3 最优方案

5.3.1 货车租赁方案

表 6.8 问题一利用随机抽样的结果检测表

对比方法	货车租赁方案	最小总成本
量子退火算法	2 辆卡车 a 1 辆卡车 b	53600
随机抽样	2 辆卡车 a 1 辆卡车 b	78200

5.3.2 货物运输方案

表 6.9 问题一利用贪心算法求出最佳方案表

对比方法	货车租赁方案	最小总成本
量子退火算法	2 辆卡车 a 1 辆卡车 b	53600
随机抽样	2 辆卡车 a 1 辆卡车 b	78200

六、问题二模型的建立与求解

6.1 模型的建立

6.1.1QUBO 模型的建立

问题二是在问题一的场景中，改变了组合租赁运输工具的各参数，可租赁卡车 a，b 以及飞机 p 的各参数，以及两者之间的匹配关系和启动成本资金，而其问题本质与解题思路未发生改变，故问题二可在问题一已建立的模型基础上，修改改变后相对应的参数，进而建立模型。

运输成本的计算会受到货物重量和体积、运输距离、卡车类型和租金、卡车单趟成本、运输方式选择、油价波动、道路状况、运输时间、拼货和中转、市场供需关系、政策法规等多个因素的影响。题意中运输物流中投入使用的不同载量卡车型号 a，b 数量 Y 为：

$$Y = \sum_{k=1}^{a} \sum_{i=1}^{b} w_{ki} t_{ki} (z_k \cdot d_k) w_{ki} t_1 P \tag{6.1}$$

物流运输的总工作量，可根据卡车的一天的 24 小时运输时间 、卡车货箱容量、效率以及数量用数学表达式表示卡车的成本即：

$$C_1 = \sum_{k=1}^{n} \sum_{i=1}^{24} w_{ki}\, t_{ki} (s_k w_{ki} + y_k w_{ki} t_1 + r_k w_{ki} t_2 + p_k w_{ki} t_2) \tag{6.2}$$

转化为最小化目标函数，即：

$$C_2 = 7 \sum_{j=1}^{2} \left(y_j' t_1 + r_j' t_2 + p_j' t_2 \right) + 3 (y_3' t_1 + r_3' t_2 + p_3' t_2) \tag{6.3}$$

约束条件由如下三点组成：车辆载重约束确保每次运输不超过车辆的载重限制；时间窗口约束，考虑货物在运输所用的送达时间；供需平衡约束，确保每个城市的发货和收货量平衡。

6.1.2 罚函数灵敏度分析

具体步骤同问题一，改变了其目标函数与变量：这个灵敏度指标反映了罚函数对罚参数变化的敏感程度。例如，如果灵敏度值较大，说明罚函数对罚参数的变化较为敏感；如果灵敏度值较小，则说明罚函数对罚参数的变化不太敏感。

1.将该模型的无约束优化问题表示为：$minz(x_i, \lambda) = min(-Q) + \lambda(g_1 + k_1)$

2.选择罚参数λ,设置一个基准值λ_0

3.使用优化算法计算在该罚参数值λ_0下的目标函数值$min(-Q)$,得到最优解x_i^*及对应目标函数值$min(-Q)^*$。

4.改变其罚参数值,例如$\lambda_{\alpha 0} = \lambda_0 + \Delta\varepsilon$，$\varepsilon$是一个小的正数。

5.使用优化算法求解新的无约束优化问题：$\min z = \min(-Q) + \lambda_{\alpha 1}(g_1 + k_1)$,其中$\lambda_{\alpha 1}$为改变后的罚参数值，得到新的最优解$x_i'$和对应新的目标函数值$\min(-Q)'$。

6.计算灵敏度，灵敏度可以通过计算罚函数值的变化量与罚参数的变化量之比来衡量，即 $\frac{min(-Q)^* - min(-Q)'}{\lambda_0 - \lambda_{\alpha 1}}$

7.根据需要，需要进行多次改变罚参数的值，并计算相应的灵敏度，以全面了解罚函 数的特性。

由上述步骤求解得：λ=600000 时，较为合适。

6.1.3 物流转运中心选址

利用数学规划方法基于数学规划的物流中心选址方法能够系统地处理大量的数据和复杂的约束条件，要综合考虑成本、服务水平、地理位置等多重因素，通过线性规划、非线性规划、整数规划等，针对不同的选址问题设定适当的目标函数和约束条件，从而找到最优或接近最优的选址方案。

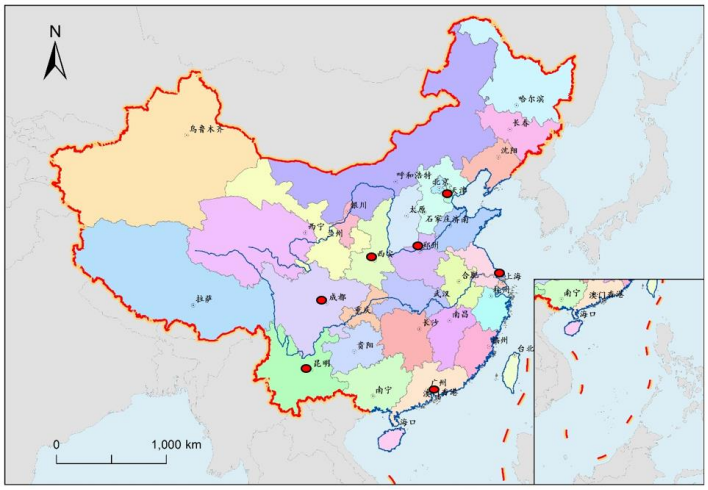


图 6.1 物流配送选址点图

结合 GIS 提供精准地理数据，MCDM 平衡多元目标，数学规划方法显著增强物流选址决策的科学性与可靠性，助力企业精准布局，应对复杂市场环境，实现资源最优配置与成本有效控制。

表 6.2 公司一待运货物小组 1

货物类型	起点城市	终点城市	卡车 a	卡车 b	运输时间（天）	成本	飞机
普货（19 吨）	上海	天津	1	1	3	28500	0
普货（25 吨）	西安	昆明	2	1	4	53000	0
普货（7 吨）	郑州	深圳	2	0	4	28000	0

在此问题中，两物流公司计划在成都、上海、天津物流起点到终点西安、昆明、深圳在运输区域内建立一个新的物流转运中心，根据这些数据如图 6.2 所示，基于重心法数学模型，可找到一个地理位置，以最小化从始发物流中心到目的城市的总运输成本和时间。

表 6.3 公司一待运货物小组 2

货物类型	起点城市	终点城市	卡车 a	卡车 b	运输时间（天）	成本	飞机
普货（27 吨）	上海	天津	2	1	3	38000	0
普货（10 吨）	成都	昆明	2	0	4	32000	0
普货（19 吨）	上海	深圳	1	2	4	44500	0

现有两公司的配送节点信息，6 个配送节点的 x、y 坐标，需求量和运输费用，如表 6.4 所示。

表 6.4 原始坐标结果

节点	wX9i	wY9i	wZ9i
天津	300	800	100
昆明	1130	300	150
上海	375	937.5	187.5
西安	450	275	72
郑州	900	900	112.5
合计	3155	3212.5	622
x1=	5.16	y1=	5.18

6.1.4 改进麻雀优化算法

1.麻雀算法基本原理

麻雀搜索算法(SparrowSearchAlgorithm，SSA)是由薛建凯提出的一种新型群智能优化算法，主要模拟了麻雀群体觅食和反捕食行为[6]。

每代发现者的位置更新公式如下：

$$x_{i,d}^{t+1}=\begin{cases}x_{i,d}^{t}\times exp\left\{\dfrac{-i}{\alpha\times iter_{max}}\right\}, if R_2<ST\\ x_{i,d}^{t}+Q\times L, if R_2\gg ST\end{cases}\tag{6.4}$$

每代跟随者的位置更新公式如下:

$$x_{i,d}^{t+1}=\begin{cases}Q\times \exp\{\dfrac{xw_d^t-x_{i,d}^t}{t^2}\}, if i>\dfrac{N}{2}\\ xb_d^t+\dfrac{1}{D}\sum\limits_{d=1}^{D}(|x_{i,d}^t-xb_d^t|\times rand\{-1,1\}), else\end{cases}\tag{6.5}$$

侦查者位置更新公式如下:

$$x_{i,d}^{t+1}=\begin{cases}xb_d^t+\beta\times|x_{i,d}^t-xb_d^t|, if f_i\neq f_g\\ x_{i,d}^t+K\times\{\dfrac{|x_{i,d}^t-xw_d^t|}{|f_i-f_w|+\varepsilon}\}, if f_i=f_g\end{cases}\tag{6.6}$$

2.PSO 算法算最短路径

粒子群优化算法(Particle Swarm Optimization，PSO 主要通过更新粒子的速度和位置信息寻找最优解。粒子群优化算法的应用较为广泛，且收敛速度快，调整参数也较少[7]。考虑有 N 个粒子在 D 维空间搜索，初始粒子群算法更新表示为:

$$\begin{cases}v_{ij}^{d+1}=\omega v_{ij}^{d}+c_1r_1\left(p_{ij}^{d}-x_{ij}^{d}\right)+c_2r_2\left(p_{gj}^{d}-x_{ij}^{d}\right)\\ v_{ij}^{d+1}=v_{ij}^{d}+v_{ij}^{d+1}\end{cases}\tag{6.7}$$

其中，d 为迭代次数，v_{ij} 表示第 i 个粒子在 i 维的速度，x_{ij} 表示第 i 个粒子在 j 维的位置。ω 表示惯性权重，控制粒子速度的惯性。c_1 与 c_2 是学习因子，控制粒子个体和群体经验对速度的影响。r_1 与 r_2 是范围在［0，1］之间的随机数。

3.麻雀算法改进

①Cubic 混沌映射初始化麻雀种群。在优化领域，混沌映射可以用于替代伪随机数生成器，生成 0 到 1 之间的混沌数。Cubic 混沌映射具有良好的混沌特性，能够在整个搜索空间内生成分布广泛的初始解，从而提高算法的全局搜索能力。

Cubic 表达式如下:

$$x_{n+1}=\alpha x_n(1-x_n^2)\tag{6.8}$$

其中，x_n 是当前迭代的混沌变量值，初值为(0，1)，α 是系统参数，它决定了映射的混沌行为，取 x_n=0.2，α =0.2493。

②重心反向学习。反向学习是一种经典的智能优化算法加速技术，它在当前点和它的反向点之中择优选择。

重心计算公式如下:

$$M = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6.9}$$

重心反向解的计算公式:

$$x_{i,new} = 2 \times k \times M - x_i \tag{6.10}$$

在迭代过程中，算法会对麻雀种群的更新位置进行重心反向变异。由于不能保证每次变异都能得到更优的位置，因此采用贪心策略来决定是否更新位置:仅当变异后的新位置更优时，才用其替换原有位置，否则保留原位置。其中，k 是［0，1］范围内均匀分布的随机数，加入收缩因子可以拓展反向搜索空间的范围，增大找到更优解的概率。

4.混合算法求解过程

混合麻雀算法和粒子群算法的核心思想在于，将粒子群算法的位置信息作为改进后麻雀算法的相关参数，从而运行麻雀算法，最终进行模型求解，如图所示。

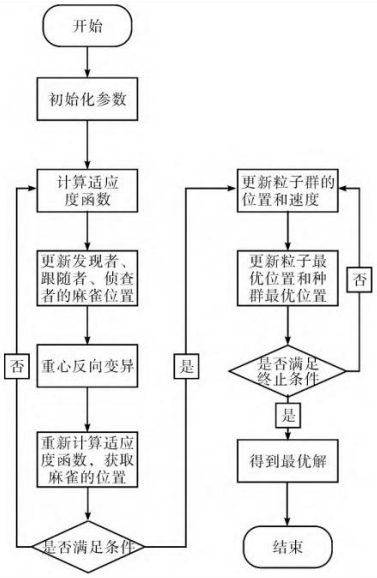


图 6.2 算法求解过程

混合算法的运行步骤如下:

混合算法
步骤 1 初始化:设定麻雀和粒子群的种群数量、迭代次数等参数。利用公式(6.8)初始化麻雀种群位置，增加搜索范围的多样性。
步骤 2 适应度评估与排序:根据适应度函数对麻雀进行排序，识别并记录每个麻雀的个体最优和全局最优适应度值及对应位置。
步骤 3 位置更新:使用公式(6.4)、公式(6.5)、公式(6.6)更新发现者、跟随者、侦察者的位置。
步骤 4 重心反向变异:对处于最优位置的麻雀使用公式(6.10)实施重心反向变异，增强全局搜索能力。
步骤 5 适应度再评估:重新计算适应度值，更新麻雀的个体和全局最优位置。
步骤 6 位置优化判断:比较麻雀的当前最优位置与粒子群的最优位置。如果麻雀位置更优，进入粒子群优化阶段;否则，回到步骤 2 继续迭代。
步骤 7 粒子群位置与速度更新:使用公式(6.7)更新各粒子群的位置和速度。
步骤 8 粒子群最优位置更新:更新粒子群个体最优和全局最优位置。
步骤 9 迭代终止判断:如果迭代次数达到预设的最大值，则结束，输出全局最佳位置;这一位置作为麻雀算法的重要参数，用于求解目标函数。如果未达到最大迭代次数，回到步骤 6 继续迭代。

为了验证所提算法的优越性，本算法选取公司 1 的物流配送和 10 个主要地级市物流中心为例进行路径规划。配送中心和地级市物流中心的位置分布如图 6.3 所示。

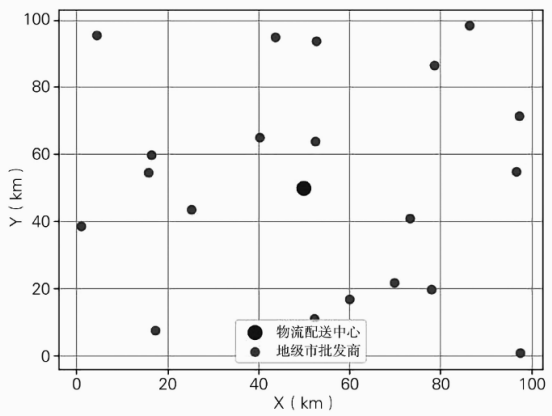


图 6.3 物流配送中心和地级市批发商分布图

6.2 模型的求解

6.2.1subQUBO 模型的求解

SubQUBO 方法的核心思想是将大规模的组合优化问题分解为一系列较小规模的问题，以便利用现有的计算资源，如伊辛机，进行有效求解，并通过理论支持，确保分解后的小问题在求解后能够还原到原始大规模问题的解[8]。

实现逻辑：化整为零，反向求解

1.在解决大规模组合优化问题时，QUBO 建模的解往往由多个量子比特集合构成。

2.若我们构建一个较小规模的问题，专注于最终解中所有不正确的量子比特集合，并利用伊辛机进行求解，那么这些错误的量子比特集合将被修正为正确的集合，从而得出正确的解。

实现方法：设计问题，聚焦于错误的量子比特集合，并通过伊辛机多次求解。

1.使用经典计算器生成多个候选答案，比较它们的量子比特集合值。一致的集合视为正确，不一致的视为错误。

2.提取错误集合，利用伊辛机求解，最终得到整体的正确答案。

6.2.2 罚函数灵敏度分析

本题根据两个决策变量形成的 QUBO 模型，已知考虑的约束条件过多，因此设计量子退火算法以降低计算量。

算法 3
Step1:根据最优租赁方案情境将其剖析为组合优化问题，构造函数。 Step2:初始化各个参数，其中计数参数 i=1，取量子退火的初始温度 T_0 足够大，Γ 为横向场强，确定最大迭代次数为 MaxSteps,初始化解为 x,对应的状态能量为$E_m(t)$。 Step3:通过随机将一个决策变量由 0 变 1 或由 1 变 0 产生随机微扰，从而产生新状态x',其对应的状态能量为$E_m(x')$。 Step4:计算能量差$\Delta E_m = E_m(x') - E_m(t)$以及 $\Delta E_n = E_n(x') - E_n(t)$,如果$\Delta E_m < 0$或者$\Delta E_n < 0$,则系统接受新解,反之，计算$x'$的接受概率$exp(\Delta E_n/T)$,并产生$(0,1)$区间上均匀分布的随机数$rangdom(0,1)$。若$exp(\Delta E_n/T) < rangdom(0,1)$,则$x = x'$。否则,重复执行步骤 3。 Step5:进行退温操作,Γ 的变化和模拟退火中的温度 T 作用类似,横向场强变化形式为$\Gamma = \Gamma - (\Gamma_0/MaxSteps)$。 Step6:判断是否满足终止条件 Γ =0,如果满足,量子退火算法终止,否则,将重复步骤 3 对下一个小问题重复 Step2-Step5。否则，重复 Step3。当所有小问题算法终止，整个量子退火算法终止。

使用 Kaiwu SDK 求解得：小组一的货物有从上海到西安的 19 吨普货，成本 24500 元；从成都到昆明的 25 吨普货，成本 37000 元；从上海到深圳的 7 吨普货，成本 14000

元。小组二的货物有从上海到西安的 27 吨普货，成本 38000 元；从成都到昆明的 10 吨普货，成本 16000 元；从上海到深圳的 19 吨普货，成本 30500 元。

6.2.3 物流转运中心选取

物流配送路径优化问题，提出一种改进的麻雀搜索算法，通过引入 Cubic 混沌映射和粒子群优化技术，有效地提高了算法的全局搜索能力，同时有效避免了早熟收敛的问题。

表 6.5 初始化迭代的运输距离和总运输成本

节点	坐标 x	坐标 y	w 需求量/吨	q 运输成本/(元/吨.)	K	d 运输距离	TC 运输成本
昆明	118.77	32.04	46	2510.869565	10	2552	115500
上海	111.67	40.82	35	2928.571429	10	1474	102500
西安	117.19	39.13	44	2443.181818	10	915	107500
天津	102.71	25.04	27	3222.222222	10	2087	87000
郑州	108.95	34.26	30	3250	10	889	97500
				总运输成本			510000

由于第一次迭代结果的总运输成本小于初始运输成本，因此需要进行多次迭代，直至总运输成本高于上一次总运输成本，即可得出最优物流中心坐标。多次迭代结果数据如表 6.6 所示。

表 6.6 迭代最终结果的物流中心坐标

节点	wX9i	wy9i	w9;
郑州	300	700	100
西安	1130	346	146
上海	375	937.5	187.5
昆明	450	275	72
天津	895.5	600	112.5
合计	3150.5	2858.5	618
x1=	5.61	y1=	5.12

在实验中，将上述参数和数据输入模型，并通过 Python 实现了相关算法，模型得出的车辆配送路径结果如图 6.4 所示

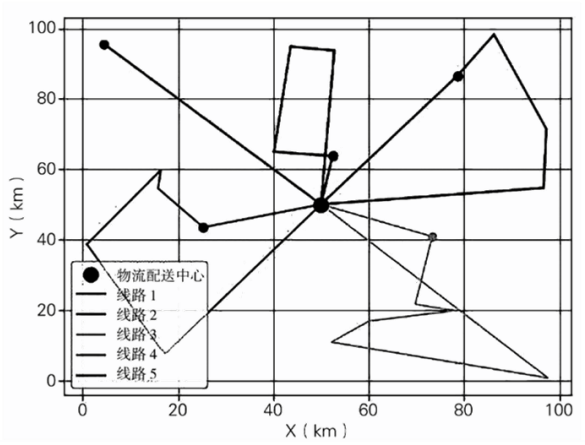


图 6.4 车辆配送路径图

在坐标轴上（5.61,5.12）这一点恰好位于西安与郑州之间的中心地带，使其成为设立转运中心的绝佳位置。此点不仅便于货物从六个主要城市快速集散，还能通过高效的物流网络辐射全国。选择此处作为转运中心，能显著提升物流效率，降低运输成本。我们建议在该区域构建一座集现代化设施与智能管理于一体的转运枢纽，以实现物流

资源的优化配置和快速响应市场需求[9]。

6.2.4 最优化成本

成本最优化策略重点关注降低整体运营成本，以确保高效的物流服务。基于数学规划方法，可以评估不同地点的成本效益比，包括地块价格、运输成本（如距离主要公路、铁路线的远近），以及地理位置所覆盖的服务范围等。对物流中心的选择可以在运输成本和地理位置之间找到一个平衡点。

另外，成本最优化策略还涉及物流中心的设计和运营，可以通过优化货物入库、存储、拣选和出库等仓库内部流程，提高运营效率，减少时间和成本，同时也可以通过采用高效的货物追踪系统等现代化的物流设施来节约运营成本。可见，基于数据和模型的方法能够为物流中心的选址和运营提供更精准的指导，使物流中心在保持竞争力的同时，实现成本最小化。

采用随机抽样方法，在满足约束条件，并采用最优匹配关系时的租赁方案中随机抽取部分方案，将其总成本与量子退火算法求得的总成本进行比较，结果如下表：

表 6.9 问题二利用随机抽样的结果检测表

对比方法	货车租赁方案	最小总成本
量子退火算法	1 辆卡车 a 2 辆卡车 b	53642
随机抽样	2 辆卡车 a 1 辆卡车 b	71542

表 6.10 改进麻雀算法及其他算法物流路径优化结果

名称	改进麻雀算法	SSA-PSO	SSA	PSO
总运输距离(km)	2070.8	1481.7	582.6	523.1
总费用(元)	28500	11800	7550	5850
使用车辆数(辆)	5	5	4	2
运行时间(s)	23.26	21.32	19.57	13.13

6.3 复杂组合成本最优化方案

6.3.1 货车租赁方案

在利用 subQUBO 方法和模拟退火算法解决复杂组合下的卡车租赁问题时，我们根据附件 4 中的详细指导和编程环境进行了相应的设置与调试。具体而言，我们设定了初始温度为 100 度，衰减常数为 0.95，并设定了迭代 25 次作为停止准则。通过这一系列的参数配置，我们成功运行了模拟退火算法，并得出了 subQUBO 模拟退火的结果图。

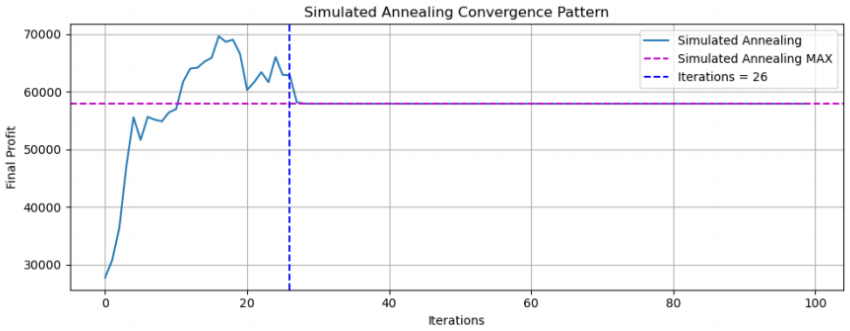


图 6.5 模拟退火求解结果

在结果图中，我们可以清晰地观察到随着迭代次数的增加，成本逐渐降低并趋于稳定，最终达到了模拟退火算法下的最小成本点。这一结果为我们提供了在复杂组合条件下卡车租赁方案的重要参考。

表 6.11 问题二租赁卡车数量及其货物重量的匹配关系表

起点-终点	货物类型	货物重量（吨）	成本（元）
上海-西安	普货	19	24,500
成都-昆明	普货	25	37,000
上海-深圳	普货	7	14,000
上海-西安	普货	27	38,000
成都-昆明	普货	10	16,000
上海-深圳	普货	19	30,500

基于模拟退火算法的输出，我们可以得出以下结论：在给定的问题背景下，通过合理安排不同租赁期限的卡车，两个物流公司合作运输最小成本的货车租赁方案如表 6.11 所示。

表 6.11 两公司合作运输最小成本的货车租赁表

货物地点	货物重量（吨）	卡车组合	A 型趟数	B 型趟数	总成本（元）
西安	44	A+B	3	1	78,000
		仅 A	4	0	94,000
昆明	46	A+B	3	2	97,500
		仅 A	4	0	100,000
上海	33	A+B	2	1	42,500
		仅 A	3	0	54,000

6.3.2 货物运输方案

本表格汇总了不同货物地点、两公司合作后货物总重量及卡车不同类型下的运输成本。详细列出了每趟运输的固定成本与租赁成本，进而计算出总运输成本与总租赁成本。结合所需趟数与预计总时间，最终得出总成本，如表 6.12 所示。此表格为物流规划与成本控制提供了直观的数据支持，帮助决策者快速比较不同运输方案的经济性，选择最优方案以节省成本。

表 6.12 最佳货物运输方案

货物地点	两公司总货物重量（吨）	卡车类型	所需趟数	每趟运输成本（元）	总运输成本（元）	每趟租赁成本（元）	总成本（元）
西安	44	A (12 吨)	4	18,500	74,000	5,000	94,000
		B (5 吨)	9	16,000	144,000	3,000	160,200
昆明	46	A (12 吨)	4	20,000	80,000	5,000	101,600
		B (5 吨)	10	17,000	170,000	3,000	200,000
上海	33	A (12 吨)	3	13,000	39,000	5,000	50,700
		B (5 吨)	7	11,000	77,000	3,000	108,700

七、问题三模型的建立与求解

7.1 人工智能模型超参数调优的 QUBO 模型

7.1.1 人工智能超参调优应用

QUBO 模型通常用于解决组合优化问题，其中变量是二进制(取值为 0 或 1)，并且目标函数是二次型的形式。它适用于那些需要在给定的约束条件下寻找最优解的问题，特别是当问题可以转化为进制变量和二次型目标函数的形式时。这种模型在多个领域都有应用，包括但不限于：

1.电路设计：QUBO 模型可以用于优化逻辑门的布局，以最大程度减少电路中的延迟或功耗。

2.物流和运输：用于优化货物的配送路线，以最小化成本或最大化效率。

3.金融投资：用于优化投资组合，以最大化收益或最小化风险。

4.组合优化：解决如旅行商问题(TSP)和背包问题等的组合优化问题。

5.人工智能和量子计算：QUBO 模型与量子计算和量子优化密切相关，可用于解决在传统计算机上难以解决的问题。

在问题 3 中，我们需要举例一个潜在通过构建 QUBO 模型从而进行决策优化的应用场景，综上并进行团队探讨，我们选取“人工智能模型超参数调优”这一场景进行探究。

7.1.2 背景信息

随着人工智能向分布式、多专家协同系统发展，机器学习模型的超参数调优变得尤为关键且复杂。传统调优方法如贪心算法、网格搜索，面对高维、大规模数据集时效率低下。

量子计算以其独特的并行处理能力和对特定问题的高效解决能力，为超参数调优带来了革命性机遇。尽管量子技术尚处初期，但其潜力巨大，能显著加速复杂模型超参数的搜索过程，提升模型性能调优的效率与精确度。探索量子计算在人工智能超参数优化中的应用，不仅是科研的前沿课题，更是推动 AI 技术跃升的重要方向。

7.1.3 研究方法

我们希望通过量子计算的优势，加速人工智能模型的超参数调优过程。具体而言，我们将超参数作为决策变量，以模型的性能度量（例如准确度）作为目标函数。QUBO 模型构建：我们将每个超参数引入二进制变量，构建一个 QUBO 模型。

基于超参数之间的相互关系，目标函数旨在最大化或最小化性能度量。CoherentIsingMachines(CIM)的应用：利用 CIM 的并行计算和高度连接性，在量子计算中更有效地搜索超参数空间。CIM 的量子优势可以提高搜索效率，找到更优的超参数组合。

7.1.4 思路及技术路线

选择关键超参数，我们需要仔细选择对模型性能影响较大的关键超参数。这可能包括学习率、层数、节点数等，具体取决于所使用的机器学习模型和任务。

7.1.5QUBO 模型设计

1、决策变量的引入：对于每个选择的超参数，引入一个二进制变量表示其取值。

2、目标函数的设计：设计目标函数，以模型的性能度量为目标，最大化或最小化这一性能度量，形成形式为二次型目标函数的。然后探究函数的约束条件，引入惩罚因子，使目标函数不受约束条件影响，形成二元无约束二值函数。

模拟退火求解或 CIM 模拟求解：使用模拟退火算法或者 CIM 求解构建好的 QUBO 模型。其中，它们充分利用了量子计算的优势，能够在高维、复杂的问题中进行高效的求解。使用获得的最优超参数组合训练机器学习模型，并在验证集或测试集上评估其性能。比较使用量子计算方法和传统方法得到的超参数组合的性能。

通过这一技术路线，我们期望在人工智能模型的超参数调优中充分发挥量子计算的优势，提高优化过程的效率，以更好地支持人工智能领域的研究和应用。因此，根据人工智能超参数调优的 QUBO 模型绘制出技术路线图，如图 7.1 所示。

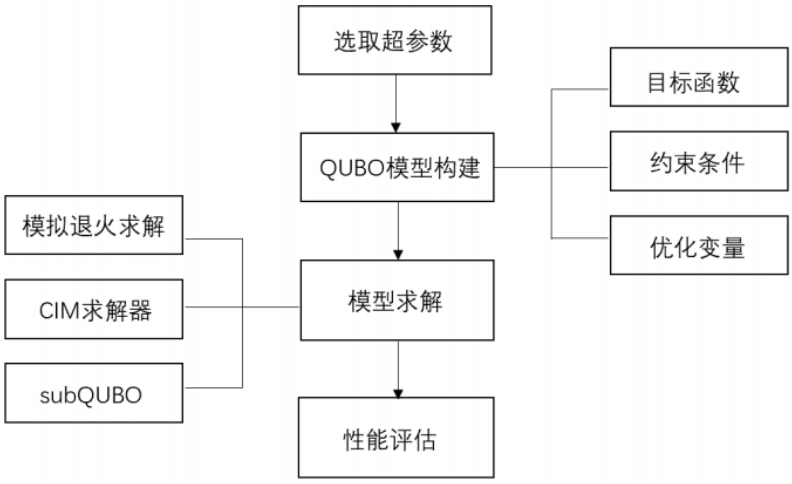


图 7.1 技术路线

7.2 研究结果

通过将量子计算引入人工智能模型的超参数调优过程，我们期望实现以下目标：

1.更快的优化速度：量子优化算法具有并行性和高效性，可以加速超参数搜索的过程，从而减少了调优所需的时间。

2.更准确的超参数调整：量子优化算法可能能够更精确地找到超参数的最优解，从而提高了模型的性能和泛化能力。

3.解决高维空间问题：传统的优化算法在处理高维超参数空间时可能受限，而量子优化算法可能更适合于处理高维空间的优化问题，这对于复杂的人工智能模型特别有用。

4.克服局部最优解问题：量子优化算法通常具有更好的全局搜索能力，可以帮助模型避免陷入传统算法局部最优解的情况，从而提高了模型的性能和稳定性。

5.应用于多种人工智能模型：量子优化算法不仅适用于特定类型的模型，还可以应用于各种类型的人工智能模型，包括监督学习、无监督学习、强化学习等。

7.2.1QUBO 模型算法

QUBO 模型算法步骤
Step1:根据待优化问题，构造量子系统的评价函数$H_q = H_{pot} + H_{kin}$,即量子哈密顿函数。其中，$H_{pot}(t)$为势能，即模拟退火算法中的评价函数，$H_{kin}^{(t)}$为动能；
Step2:初始化各个参数，T_0为量子退火的初始温度，Γ为横向场强，变化的横向场强引起不同量子状态之间的量子跃迁，最大迭代次数为 MaxSteps，初始化状态为x,对应的状态能量为$H_{pot}(x)$;
Step3:随机微扰产生新状态x',对应的状态能量为$H_{pot(x')}$;
Step4:计算能量差$\Delta H_{pot} = H_{pot}(x') - H_{pot}(x)$以及$\Delta H_q = H_s(x') - H_q^{(x)}$,如果$\Delta H_{pot} < 0$或者$\Delta H_q < 0$则系统接受新解$x = x'$,反之如果$\exp(\Delta H_q/T) < rangdom(0,1)$,则$x = x'$,否则重复执行步骤 3;
Step5:进行退温操作，Γ的变化和模拟退火中的温度T作用类似，横向场强变化形式为$\Gamma = \Gamma - (\Gamma_0/MaxSteps)$;
Step6:判断是否满足终止条件$F = 0$,如果满足量子退火算法终止，否则重复步骤 3。

7.2.2 超参数调优 QUBO 模型

通过数据的获得，我们采用量子计算的方法对人工智能超参数进行初步探究。假设 Ni：量子位的数量；Ne：能量水平；Ns：可解的个数，我们求解得到了 QUBO 求解

算法在不同参数配置下的准确度分布情况，例如：Ni=20，Ne=5 和 Ns=10。

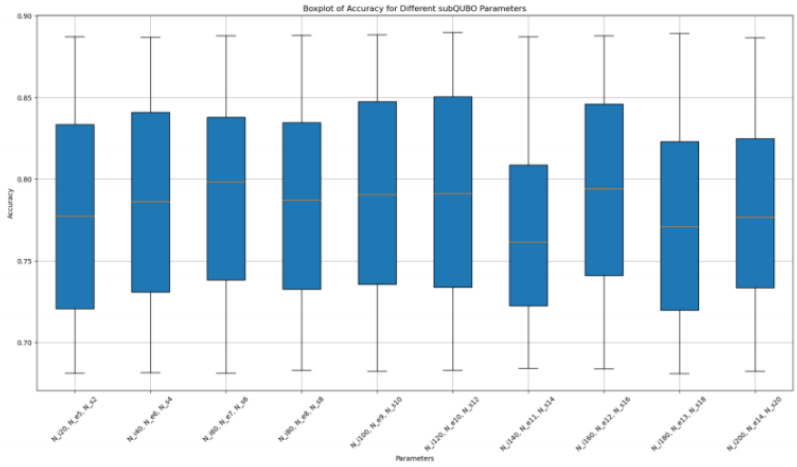


图 7.2QUBO 准确度分布箱线图

上图为箱形图，展示了 QUBO 求解算法在不同参数配置下的准确度分布情况。其中，每个箱子代表一组特定参数下算法准确度的统计分布。箱形图通过展示数据的五数概括（最小值、第一四分位数、中位数、第三四分位数、最大值）以及可能的异常值来展示数据的分散情况。

我们发现，准确率的中位数在不同参数设置之间有所变动，但总体来说几乎都在 0.80 以上，准确率较高。四分位距代表准确率的可变性，较短的箱子表示更一致的结果，较长的箱子表示结果的变化较大。结果显示 Ni_{60},Ne_{7},Ns_{6}，有些参数设置的箱子较长，表明在这些设置下准确率的变动较大，而一些较短的箱子则表明在这些设置下准确率较为稳定。

七、模型的评价与推广

7.1 模型的优点

本文通过模拟退火算法为 QUBO 问题得到了满意的解，较好地解决了寻找最优解的问题。且该算法还可用于其它各种组合优化问题，如 TSP 和 Knapsack 问题等，能产生令人满意的近似最优解，而且运行速度较快。

该模型通过两种求解方法分别是模拟退火求解器和 CIM 模拟器进行求解，其结果更合理、更准确；当量子比特的数量较多时，采用改进的模拟退火 subQUBO 进行求解，能很好地解决大规模数据或运算较复杂的问题。

本文对求解结果进行可视化处理及误差可视化处理，使呈现更为直观，且增强了可信度。

本文通过 QUBO 转化，将决策变量转化为 0-1 变量，使得数据得到充分使用，避免了数据错误或遗漏带来的误差。

7.2 模型的缺点

该模型建立在卡车租赁成本确定的前提条件下，本文采用的模拟退火算法求解 QUBO 问题时更适用于数据量足够大的样本，如果数据量较小，求解结果会存在一定的误差。

7.3 模型的推广

该论文使用的 QUBO 模型，除了可解决物流租赁和运输方案外的问题，还可以解决金融、人工智能、最大独立集问题、不对称分配问题、对称分配问题、边约束分配问题、二次背包问题、最大团问题、最大割问题、整数划分问题和旅行商问题等，可结合具体实际运用进行变化运用。并在使用过程中，可以通过下述方式将 QUBO 模型与 Ising 模型相互转化，以便求解不同问题。

参考文献

[1] Wang Xiaojun, Wang Zhenghuan, Ni Bowen. Mapping structural topology optimizat ion problems to quantum annealing [J]. Structural and Multidisciplinary Optimizatio n, 2024, 67 (5):

[2] 王敬,张萌,李芳,张海懿.量子计算关键技术及应用发展分析[J].信息通信技术与政策,2023,49(07):9-16.

[3] Şeker Oylum, Tanoumand Neda, Bodur Merve. Digital Annealer for quadratic uncon strained binary optimization: A comparative performance analysis [J]. Applied Soft C omputing Journal, 2022, 127

[4] Yue Ji. Logistics distribution scheduling algorithm based on artificial intelligence [J]. Measurement: Sensors, 2024, 34

[5] 王宝楠,水恒华,王苏敏,胡风,王潮.量子退火理论及其应用综述[J].中国科学:物理学力学天文学,2021,51(08):5-17.

[6] 孙玉杰,张占强,孟克其劳,吕晓圆.基于 ESMD 和 SSA-PNN 的电能质量扰动信号识别分类[J].现代电子技术,2022,45(14):108-114.

[7] 曹梦龙,赵文彬,陈志强.融合粒子群算法与改进灰狼算法的机器人路径规划[J].系统仿真学报,2023,35(08):1768-1775.

[8] Luca Asproni, Davide Caputo, Blanca Silva, Giovanni Fazzi, Marco Magagnini. Acc uracy and minor embedding in subqubo decomposition with fully connected large pr oblems: a case study about the number partitioning problem [J]. Quantum Machine I ntelligence, 2020, 2 (1):

[9] 吴善兵.六西格玛方法在 QG 公司厂内物流管理改善中的应用研究[D].导师:关志民.东北大学,2015.