2024 第九届“数维杯”大学生数学建模挑战赛论文

题　目　生物质和煤共热技术中特征选择和优化研究

摘　要

生物质和煤共热技术是一种将生物质与煤混合燃烧的技术，旨在减少对传统煤炭的依赖，降低碳排放，并促进可再生能源的利用。通过在燃烧过程中将生物质与煤混合使用，可以减少温室气体排放，提高能源利用效率，并降低对环境的影响。本文通过研究生物质与煤混合燃烧的实验数据，**成功建立了一个共热解过程的效率和产物利用率的预测模型**，对促进可再生能源的利用具有指导作用。

针对问题一，本文首先进行数据预处理，并进行探索性数据分析，对附件 1 数据集通过所得的**比例关系来插值正己烷不溶物相关数据的缺失值**，然后使用**皮尔逊相关系数**及热力图可视化得出 INS 分别对焦油产率，水产率和焦渣产率表现出强正相关，极弱负相关和中强负相关，即**对焦油和焦渣有显著影响，对水产率无显著影响**。

针对问题二，本文首先将 INS 和配比数据相乘作为新的特征，然后使用 **LightGBM** 模型查看各特征对各目标产率的重要性, 得出二**者存在交互作用**，并根据各特征对各产物产率的重要性的柱状图可知，交互效应在**焦油产率和正己烷可溶物产率**上表现最为明显。

针对问题三，首先使用**熵权法-模糊综合评价模型**对四个产物产率进行评价，将其转换成一个综合得分作为**量化产物利用率和能源转化效率**的指标，然后使用**多元多项式拟合**，拟合出配比、样品、焦油、水以及 INS 与该指标的**函数表达式**，最后使用**粒子群算法**优化得出当该指标取最大值时，混合比例的取值为 28.44%，即**生物质在煤与生物质总量的占比为 28.44% 时产物利用率和能源转化效率最高**。

针对问题四，首先将附件 2 的数据结构标准化，并采用 **Lagrange 插值法**对缺失的理论计算值数据进行插值处理，然后使用 **Wilcoxon 符号秩检验**对每组实验的**每种产物的实验值和理论值**进行显著性差异分析，再针对存在**显著性差异的组的每个混合比例**使用 Wilcoxon 符号秩检验，找出导致较大差异的混合比例。

针对问题五，本文研究基于**模型集成**思想，分别建立评估了**多项式回归模型**、**随机森林回归模型**和**高斯回归-贝叶斯优化模型**，来捕捉共热解产物产率预测任务中复杂的非线性关系。对比单一机器学习模型与传统回归拟合模型，**基于贝叶斯优化的高斯过程回归方法**表现出优异的预测性能。

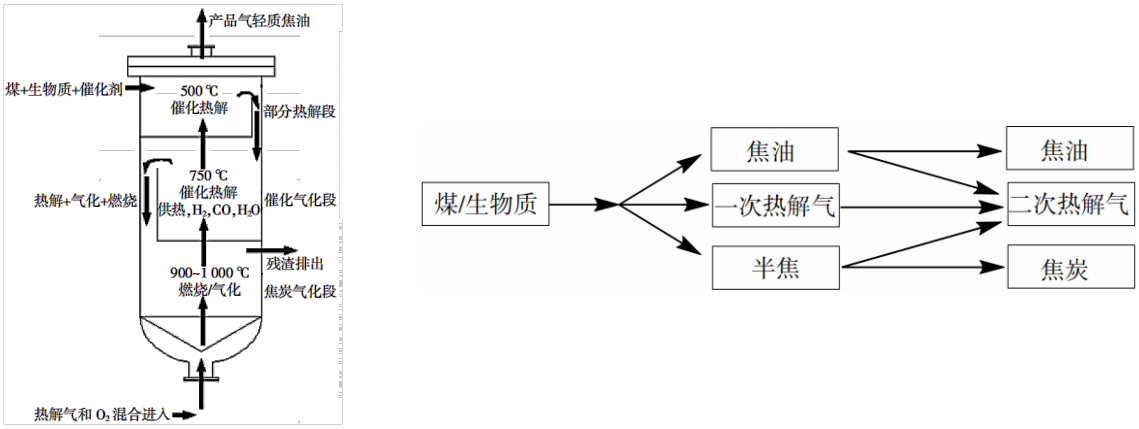
关键词: 共热解，LightGBM，模糊综合评价，Wilcoxon 符号秩检验，高斯过程回归

目 录

一、 问题背景与重述

1.1 问题背景

在当今世界，能源需求不断增长，而化石燃料的过度使用导致了严重的环境问题。为了实现可持续发展，人们越来越重视开发清洁、高效的可再生能源。生物质作为一种广泛分布、易于获取的可再生资源，与煤共热解技术的结合为**解决能源问题**提供了新的思路。



(a) 一种共热解反应器示意图 (b) 热解过程示意图

图 1-1: 共热解工艺相关示意图[2]

生物质和煤在高温缺氧条件下会发生共热解，产生气体、液体和固体等多种产物。但由于生物质和煤的性质差异，以及热解过程的复杂性，如何优化共热解工艺、提高产物品质成为亟需解决的问题。近年来，研究[1] 发现生物质和煤共热解过程中存在显著的协同效应，由于生物质热解产生的活性物质与煤中的大分子发生了相互作用，能够促进煤的热解反应，表现出热解产率和产物质量的提升。

为了深入理解**生物质-煤共热解机制**，优化工艺参数，提高热解油品质，本文借助相关数学建模方法开展综合性研究。通过分析实验数据，建立合适的数学模型，可以定量描述生物质和煤组分、热解条件与产物之间的关系，为工业化应用提供理论指导。本文研究可以为促进生物质与煤热解领域的技术提供参考依据，期待为解决能源短缺、环境污染等问题做出贡献。

1.2 问题重述

1. **问题一: 正己烷不溶物 (INS) 对焦油产率、水产率和焦渣产率的影响**。本题为关于热解过程的定量分析问题，需要通过相关性分析或显著性检验等统计方法量化分析 INS 与各产率之间的内在联系，进而阐明 INS 对热解反应过程的影响机制，为热解工艺参数优化提供定量依据。这是一个相关性分析问题，需要我们建立合适的统计分析模型。
2. **问题二: INS 及其与混合比例的交互效应对热解产物组成的影响**。这是一个关于热解反应多因素关联性的问题。可通过决策树方法研究 INS 与混合比例及其交互作用对多元产率的影响规律，揭示多因素交互作用的内在关系。这是一个交互效应分析与建模问题，需要恰当地刻画多因素间的关系。
3. **问题三:** 本题为目标优化问题，确定最佳混合比例，以提升热解产物利用率与能源转化效率。要求综合考虑多个产物指标，可通过熵权法将其转化为**单目标优化问**

题，并运用数学规划、元启发式算法等方法求解最优配比、权衡利弊。这是一个参数优化问题，要求合理地描述目标函数与约束条件，并选用高效的求解算法。

4. **问题四:** 要求**分析产物收率的实验值与理论计算值的差异性及其分布规律**。可通过统计检验方法，定量评估实验数据与理论模型的吻合程度，揭示理论模型的适用性与局限性，并进一步探讨组间影响因素，为模型改进提供依据。这是一个统计检验与分布拟合问题，需要根据数据特点选择合适的参数或非参数检验方法，并进一步对残差进行分析。
5. **问题五:** 建立热解产物产率的**预测模型**。要求通过综合考虑原料组成、热解条件等多种因素，采用多元回归、小样本机器学习等方法，刻画产率与影响因素之间的定量关系，筛选最优模型，为工艺参数优化与过程控制奠定基础。这是一个预测建模问题，需要合理设计模型形式、特征工程与评估指标，提高模型的精度。

二、 问题分析

2.1 问题一的分析

针对问题一，首先可以进行数据预处理，初步了解数据结构及分布情况，然后可以考虑使用皮尔逊相关系数，判断 INS 与热解产物产率之间的相关性，从而判断是否有显著影响。

2.2 问题二的分析

针对问题二，可以考虑将 INS 和配比数据进行乘积，并将其作为一个新的特征，然后可以考虑使用 LightGBM 模型计算出各个特征对焦油产率的重要性指标，然后将该新特征的重要性与 INS 和配比的重要性分别对比。

2.3 问题三的分析

针对问题三，可以考虑首先使用熵权法-模糊综合评价模型，根据四种产物产率去量化产物利用率和能源转化效率，然后可以考虑使用多项式拟合，将配比、样品、焦油、水、INS 作为自变量拟合量化指标，最后可以使用粒子群算法对拟合得出的函数进行优化。当量化指标达到最大时，配比的值即为最佳共解热混合比例。

2.4 问题四的分析

针对问题四，首先需要对缺失的实验数据进行，然后可以考虑使用 Wilcoxon 符号秩检验来进行显著性分析。对于有显著性差异的组，可以再使用 Wilcoxon 符号秩检验进行子组分析。

2.5 问题五的分析

针对问题五，可以考虑建立多种回归预测模型进行性能评估。对于小样本数据驱动的共热解产物产率预测任务，使用基于相关优化算法的回归模型去捕捉题目中复杂的非线性关系，相对于简单的回归方法或者机器学习模型，该模型可能表现效果会更好。

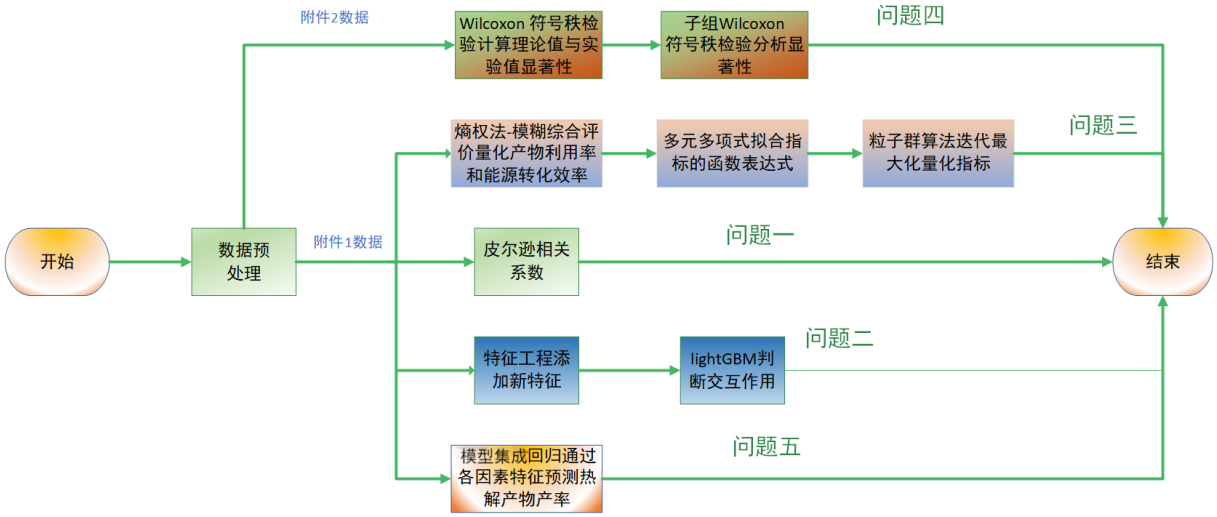


图 2-2: 问题分析流程图

三、 模型假设

1. 热解过程满足质量守恒定律，所有原料最终都转化为焦油、水、焦渣等产物，其他物质损失可忽略不计。
2. 为聚焦核心问题，假设在给定的初始温度 (600°C) 和升温速率 (5°C/min) 下, 混合比例是影响热解产物分布的决定性因素，而其他例如气压、停留时间等化工工艺因素影响相对较小。
3. 生物质与煤在共热解过程中确存在显著的非线性相互作用效应。
4. 正己烷不溶物 (INS) 含量可作为衡量生物质与煤相容性的关键指标。

四、 主要符号与说明

符号	含义
r	皮尔逊相关系数
μ_X、μ_Y	分别表示变量 X 和 Y 的平均值
$F_m(x)$	m 次迭代的预测结果
$L(y,p)$	损失函数，其中 y 为真实函数，p 为模型预测概率
E_j	信息熵
p_{ij}	表示第 j 个特征（或列）中第 i 个样本所占的比例
w_j	熵权法所得权重
$p^k_{id,\text{pbest}} - x^k_{id}$	粒子 i 在第 k 次迭代中第 d 维的历史最优位置
$p^k_{d,\text{gbest}} - x^k_{id}$	群体在第 k 次迭代中 d 维的历史最优位置
$h_m(x)$	决策树在 m 次迭代中的预测结果
$Gini(D)$	随机森林基尼系数
σ_f^2	高斯过程中的函数的方差

五、 数据预处理

通过对原始数据的观察，我们发现数据集存在一定的缺失值。为确保数据的完整性和可用性，我们采取了以下步骤进行数据预处理：

5.1 数据清洗

- 缺失值插补：对于附件一中缺失的正己烷不溶物 (INS) 数据，如表5-2我们通过对关键指标间的比例统计数据进行探索性分析，发现正己烷不溶物/焦油比比率数据波动较小，有较好的一致性，所以根据其已有的潜在线性关系，采用线性插值的方法进行填补。

表 5-1: 描述性统计数据

变量名	样本量	最大值	最小值	平均值	标准差	中位数	变异系数 (CV)
配比	135	1	0.048	0.295	0.301	0.2	1.02
样品 g	135	12.055	5.072	8.719	1.54	8.811	0.177
焦油 (Tar) g	135	2.139	0.361	1.038	0.316	1	0.304
水 (Water) mL	135	1.91	0.58	1.037	0.295	0.95	0.285
正己烷不溶物 (INS) g	135	0.935	0.025	0.304	0.158	0.297	0.52
焦油产率	135	0.424	0.036	0.128	0.057	0.12	0.445
水产率	135	0.276	0.058	0.122	0.04	0.112	0.331
焦渣 (Char) 产率	135	0.76	0.278	0.641	0.097	0.655	0.151
正己烷可溶物产率	135	0.889	0.033	0.099	0.075	0.091	0.756

表 5-2: 化学比率的统计数据及评价

化学比率类型	平均值	中位数	标准差	数据集评价
正己烷不溶物/焦油比率	0.283	0.280	0.106	数据一致性好，波动小
正己烷可溶物/焦油产率比率	0.805	0.758	0.510	相对较大波动
正己烷不溶物/水比率	0.313	0.284	0.186	波动适中，数据较集中
焦渣产率/水产率比率	5.909	5.869	2.300	相对较大波动

对于附件二中缺失的实验数据，观察到后七组不同混合比例标签下的数据虽呈现潜在的线性关系，而在第一组已有的 100/0 混合比例下的理论数据产生了一定区分，为了更好捕捉这种关系本文研究采用了 Lagrange 插值方法。给定 7 个节点的混合比例和相应的实验产物收率实验值，需要估计每组实验缺失值 y_1 和 y_2。构建拉格朗日插值多项式：

$$L(x) = \sum_{i=3}^{5} y_i \ell_i(x), \quad \ell_i(x) = \prod_{\substack{j=3 \\ j \neq i}}^{5} \frac{x - x_j}{x_i - x_j}$$

然后，估计缺失值：

$$y_1 \approx L(x_1), \quad y_2 \approx L(x_2)$$

该方法对噪声和异常值也有一定的鲁棒性，它不需要预先假设数据的函数形式，让数据“说话”。

- 数据重构：将附件二的数据按照共热解组合和混合比例重新组织，形成结构化的数据矩阵，便于后续分析。
- 比例转换：将生物质/煤的混合比例转化为生物质/(生物质 + 煤) 的比例形式，并转化为小数格式，方便后续数值计算。

5.2 探索性分析

本文对该数据集采用探索性数据分析方法，对数据进行了深入的挖掘和分析。在研究初期，通过描述统计、相关性分析等手段，系统性地了解了数据的基本特征、各变量之间的关系以及数据分布情况。随后，利用数据可视化技术，如散点图、小提琴图，热图等，将数据清晰图形化，进一步找出数据内在的模式、趋势或异常情况。

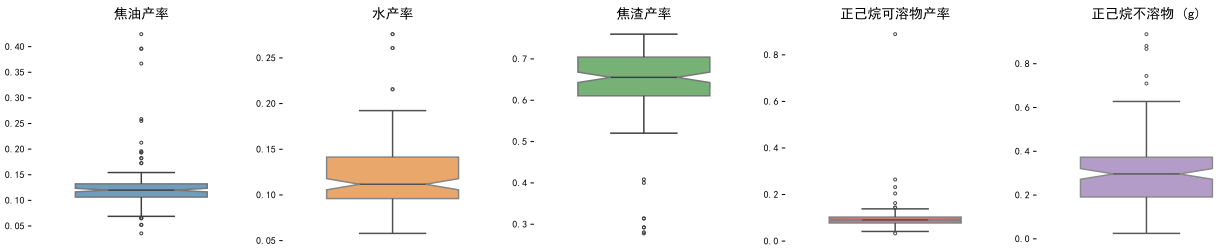


图 5-3: 关键变量箱线图

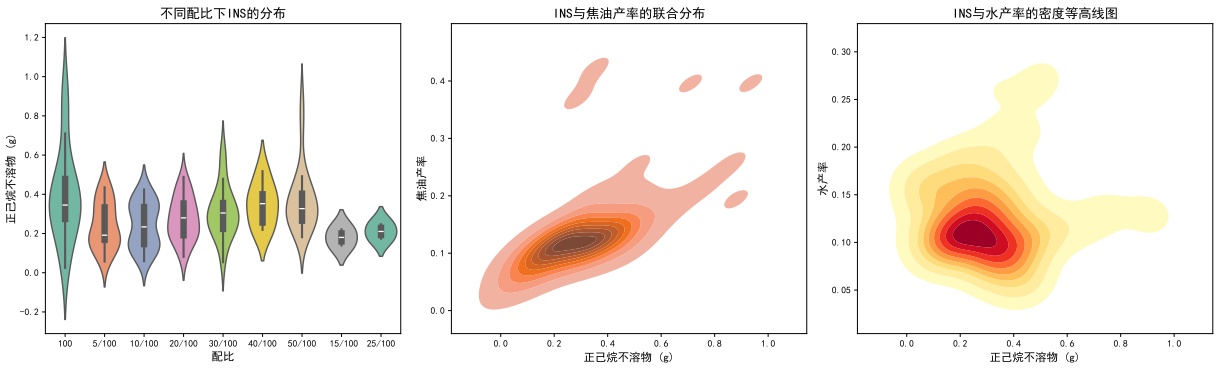


图 5-4: 对正己烷不溶物 (INS) 的探索性分析

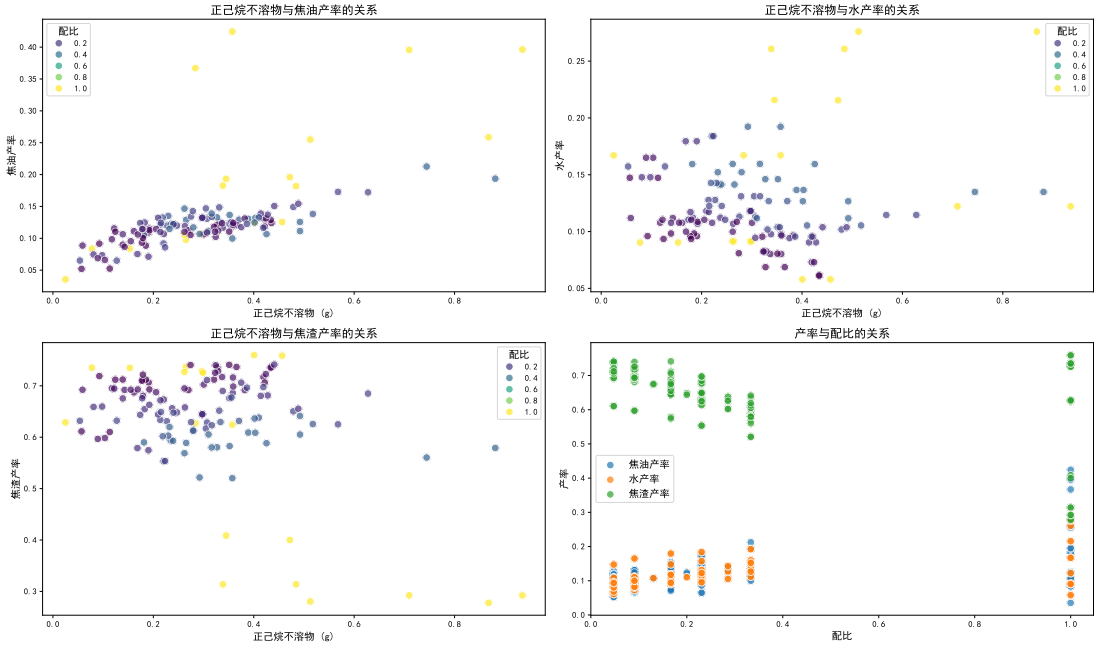


图 5-5: 关键关系散点图

在预处理后的数据基础上，如表5-1、图5-3、图5-4和图5-5所示，本文研究首先进行了初步的描述性数理统计以及探索性数据分析，进一步采用了契合问题方向的可视化方法，以了解数据的分布特征和变量间的潜在关系。通过分别不同指标关系下的箱线图、小提琴图、康托等高线图以及散点图矩阵，通过观察可以得到以下初步结论：

1. 不同生物质和煤样的热解产物组成存在显著差异；
2. 随着生物质混合比例的增加，焦油和水的产率整体呈上升趋势，而焦渣产率呈下降趋势；
3. 在较高的生物质混合比例下，INS 含量与焦油产率呈现较强的正相关关系。

这些初步发现为后续的建模分析提供了方向和依据。

六、 模型的建立与求解

6.1 问题一模型的建立与求解

6.1.1 皮尔逊相关系数统计模型

模型建立

皮尔逊相关系数[3] 是一种用于衡量两个连续变量之间线性相关性的统计量。它衡量了变量之间的线性关系程度，其取值范围在 -1 到 1 之间。其计算公式如下：

$$r = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_Y)^2}} \tag{1}$$

其中，r 代表皮尔逊相关系数，X 和 Y 代表变量，μ_X 和 μ_Y 分别表示变量 X 和 Y 的平均值。r 的绝对值越大，相关性越高。具体的 r 值含义如表6-3所示：

表 6-3: 皮尔逊相关系数含义

r	相关强度
$[-1,-0.6]$	强负相关
$(-0.6,-0.4]$	中强负相关
$(-0.4,-0.2]$	弱负相关
$(-0.2,0.2)$	极弱负相关或无相关性
$[0.2,0.4)$	弱正相关
$[0.4,0.6)$	中强正相关
$[0.6,1]$	强正相关

模型求解

step1：数据准备

确定变量的数据集 X、Y，数据集数据分别为正己烷不溶物 (INS) 和各产物产率的数据。确保数据是一一对应的，即每个数据点的 X 和 Y 值是相关联的。

step2：差异的平方和

计算 X 和 Y 的均值。将所有 X 值相加，然后除以 X 的总数，得到 X 的均值。同样地，将所有 Y 值相加，然后除以 Y 的总数，得到 Y 的均值。计算 X 和 Y 的差异。对于每个数据点，将 X 的值减去 X 的均值，得到 X 的差异。同样地，将 Y 的值减去 Y 的均值，得到 Y 的差异。计算 X 和 Y 差异的乘积。将 X 的差异和 Y 的差异相乘，得到每对数据点的乘积。计算 X 和 Y 差异乘积的总和。将每对数据点的乘积相加，得到 X 和 Y 差异乘积的总和。计算 X 和 Y 差异的平方和。对于每个数据点，将 X 的差异平方，然后将所有平方值相加，得到 X 差异的平方和。同样地，计算 Y 差异的平方和。

step3：计算皮尔逊相关系数

计算皮尔逊相关系数。将 X 和 Y 差异乘积的总和除以根号下 X 差异的平方和乘以 Y 差异的平方和，并绘制皮尔逊相关系数热力图。所有特征的相关系数热力图如图6-6所示：

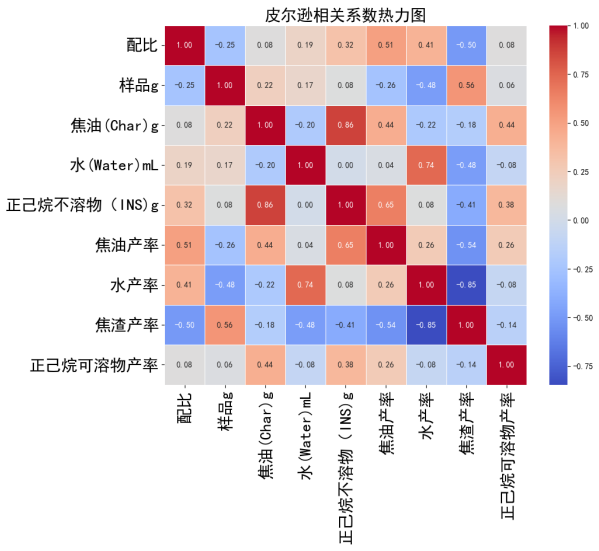


图 6-6: 所有特征的皮尔逊相关系数热力图

其中，正己烷不溶物与各热解产物产率的相关系数如表6-4：

表 6-4: 正己烷不溶物与各热解产物产率的相关系数

技能类型	皮尔逊相关系数
INS 与焦油产率	0.65
INS 与水产率	0.08
INS 与焦渣产率	-0.41

由表6-4可得，INS 分别对焦油产率，水产率和焦渣产率表现出强正相关，极弱负相关和中强负相关。

6.2 问题二的模型建立与求解

6.2.1 LightGBM 模型

模型的建立

LightGBM 方法[4] 是建立一个与前一个基学习器相关的新的基学习器，建立的依据是前一个基学习器损失函数的梯度下降方向，目的是使模型的整体损失函数继续下降，该方法是对基学习器进行整合，使模型始终处于优化状态[4]。残差更新的方式是 LightGBM 与提升树的主要区别，LightGBM 的算法如下：

LightGBM

算法：两个输入，A 和数据集 d, 数据集 d 包括自变量和因变量

变量 $(x_i, y_i), i = 1, 2, \cdots, N$

1. 初始化参数 $g_m(x_i), \theta'_m \beta_m f_m$

2. $f_m(x)$;　　//构建树模型

3. $f_0(x) = \arg\min_\theta \sum_{i=1}^{N} L(y_i; \theta)$;　　//将模型初始化为零

4. ***for* (m=1; m<M; m++)　*do***

5. $f_0(x) = \arg\min_\theta \sum_{i=1}^{N} L(y_i; \theta)$

6. $g_m(x_i) = \left[\frac{\partial L(y_i f(x_i))}{\partial f(x_i)}\right] f(x) = f_{m-1}(x)$

7. $\theta' = \arg\min_{\theta,\beta} \sum_{i=1}^{N} \left[-g_m(x_i) - \beta_m \theta(x_i)\right]^2$

8. $\beta_m = \arg\min_\beta \sum_{i=1}^{N} L\left[y_i f_{m-1}) + \beta_m \theta'_m(x_i)\right]$

9. $f_m(x) = f_{m-1} + \beta_m \theta'_m$　　//构建树模型

10. ***end for***

输出: $f_M(x)$**, 最终模型**

LightGBM 机理可视化如图6-7所示：

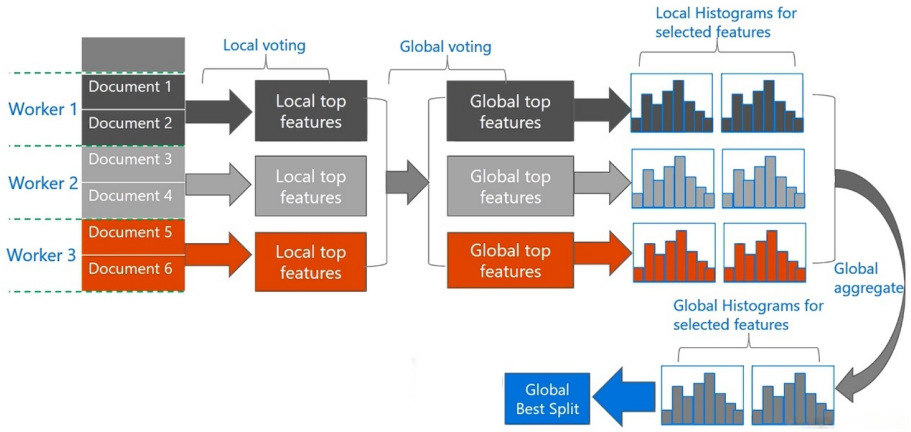


图 6-7: LightGBM 机理可视化

GMB 迭代过程公式如下:

$$F_m(x) = F_{m-1}(x) + \gamma \cdot h_m(x) \tag{2}$$

其中 $F_{m-1}(x)$ 是模型在 $(m-1)$ 次迭代中的预测结果。$h_m(x)$ 是决策树在 m 次迭代中的预测结果。γ 是学习率，用于控制每棵树对最终预测结果的贡献程度[5]。

模型训练的目标是找到能使损失函数最小化的参数集，损失函数通常是对数损失函数（用于二元分类问题），定义为:

$$L(y,p) = -\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1-y_i) \cdot \log(1-p_i)] \tag{3}$$

其中，y 为真实标签，p 为模型预测的概率，N 为样本数。

模型的求解

step1: 模型初始化

本文选择了一个决策树模型作为初始模型并将其初始化。

step2: 计算初始预测值和残差

使用初始模型对训练集进行预测，得到初始的预测值。计算初始预测值和真实值之间的残差，表示当前模型对样本的拟合程度。

step3: 训练下一个弱分类器

使用残差作为目标值，训练下一个弱分类器。这个弱分类器会对残差进行拟合，以减小拟合误差。

step4: 更新模型和预测值

将当前模型和新训练的弱分类器进行加权组合，得到一个新的更新模型。使用更新模型对训练集进行预测，得到更新的预测值。

step5: 循环迭代

计算更新预测值和真实值之间的残差。重复上述步骤，直到达到预设的迭代次数或达到终止条件。

step6: 计算特征重要性

对于每个特征，统计在所有迭代中，该特征对模型性能的贡献。本文使用的是基于增益的计算方法。计算得出每个学校的五个技能成绩的分差分别对总成绩分差的重要性，并对五个重要性的数值做归一化处理便于比较。

本文将配比和正己烷不溶物的数据相乘，并将乘积作为一个新的特征，然后探究配比、样品、焦油、水、正己烷不溶物和新特征对各产物产率的重要性。

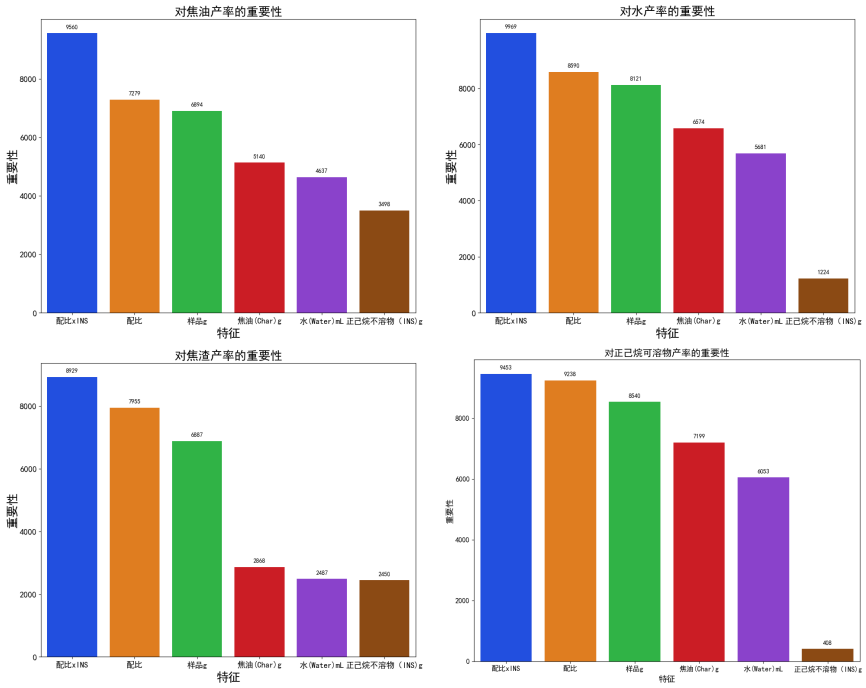


图 6-8: 各特征对产物产率的重要性

由该图可知交互效应在焦油产率和正己烷可溶物产率上表现最为明显。

6.3 问题三模型的建立与求解

6.3.1 熵权法-模糊综合评价模型

模型的建立

模糊综合评价的基本原理是从影响问题的诸因素出发，确定被评价对象从优到劣若干等级的评价集合和评价指标的权重[7]，对各指标分别做出相应的模糊评价，确定隶属函数，形成模糊判断矩阵，将其与权重矩阵进行模糊运算，得到定量的综合评价结果。由于熵权法在确定权重时不需要主观判断或专家经验，能够较为客观地确定权重，避免了主观偏好的影响，因此本文使用该方法计算权重。

一个系统中的失序现象可以用熵来计算。但熵的本身并不能直接反映某项指标在实际问题中的重要程度，熵权法是根据熵的特性，通过计算来判断某一事件的方法。信息熵越小的某一指标该指标值的离散的程度越大。在综合评估中所能发挥的作用也越大，它所占的权重也会相应增加。与之不同的是，当一个指标的信息熵更大时，则该指标数值的离散程度被证明越小。综合评价作用越小，它的分量就越小。

本文对产物利用率和能源转化效率评价主要由熵权法[8]和模糊综合评价法两部分组成。熵权法计算每个指标的熵，再进行归一化处理即得到每个指标权重；模糊综合评价法的主要步骤是确定权重的向量、建立评价集、进而得到隶属度矩阵[9]，从而进行模糊计算。最终，将上述二者相结合，根据最大隶属度的原则，综合分析得到热解组合目标产率的综合得分。

模型的求解

step1: 建立综合评价的因素集

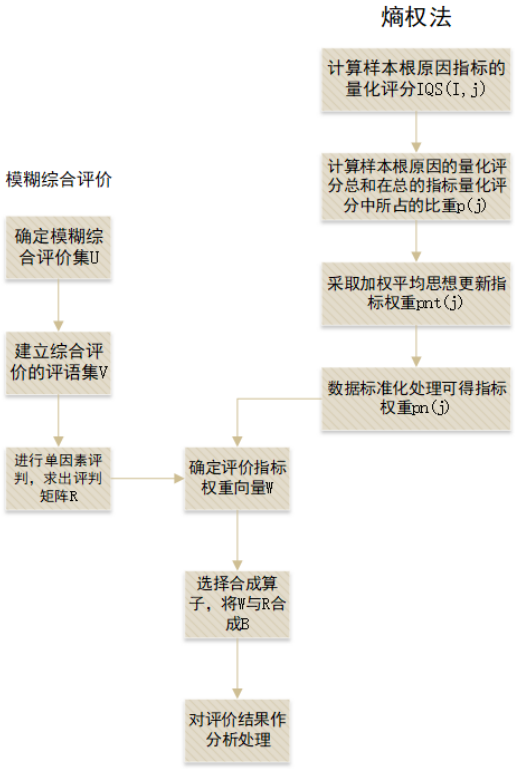


图 6-9: 熵权法-模糊综合评价原理图

因素集是以影响评价对象的各种因素为元素所组成的一个普通集合，用 U 表示，

$$U = \{u_1, u_2, \cdots, u_n\}$$

其中元素 u_i 代表影响评价对象的第 i 个因素。这些因素，通常都具有不同程度的模糊性。本文在此选择了 4 个指标构成的因素集，即 4 个产物产率：焦油产率、水产率、焦渣产率、正己烷可溶物产率。

step2: 建立综合评价的评价集

本文根据上述因素集整合数据集，得到每条实验记录的产物利用率和能源转化效率评价集。表6-5展示部分评价集数据：

表 6-5: 产物利用率和能源转化效率评价集

序号	焦油产率	水产率	焦渣产率	正己烷可溶物产率
1	0.124639	0.057944	0.759885	0.125236
2	0.125584	0.0579	0.758669	0.0809
3	0.097714	0.0914	0.736547	0.092842
4	0.106083	0.0914	0.728127	0.10017
5	0.10542	0.091408	0.726852	0.091129

step3: 确定各因素的权重

评价工作中，各因素的重要程度有所不同，为此，给各因素 u_i 一个权重 a_i, 各因素的权重集合的模糊集，用 A 表示：$\mathrm{A} = \{\mathrm{a}_1, \mathrm{a}_2, \cdots, \mathrm{a}_n\}$。本文在此利用熵权法确定各因

素权重。计算各指标信息熵公式如下：

$$E_j = -\ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \quad (\text{若} p_{ij} = 0, \text{定义} E_j = 0) \tag{4}$$

其中，n 表示数据集中的样本数量，而 p_{ij} 表示第 j 个特征（或列）中第 i 个样本所占的比例。

再根据信息熵计算各指标权重：

$$w_j = \frac{1 - E_j}{k - \Sigma E_j} \tag{5}$$

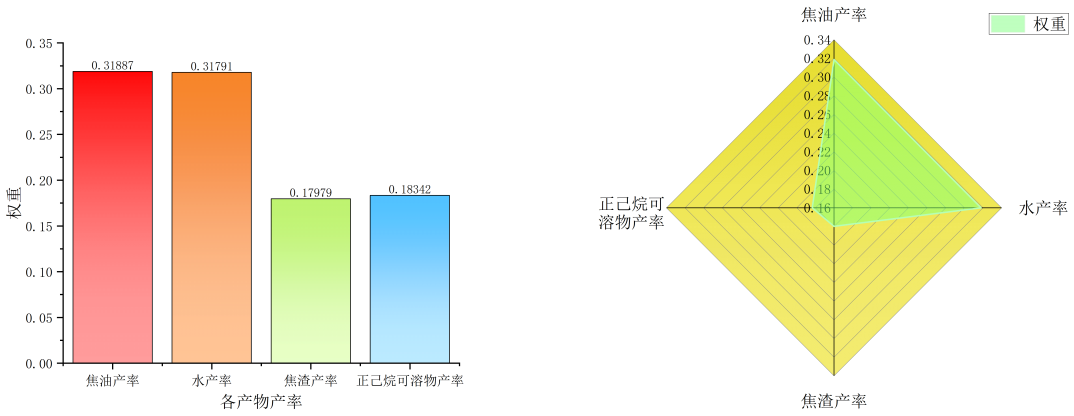
这里 k 指的是指标个数。

通过计算信息冗余度来计算权重：

$$D_j = 1 - E_j \tag{6}$$

$$w_j = \frac{D_j}{\sum_{j=1}^{m} D_j} \tag{7}$$

本文在此展示量化产物利用率和能源转化效率时，使用熵权法计算出的各评价指标的权重，如图6-10所示：



(a) 权重柱状图　　(b) 权重图

图 6-10: 各评价指标权重情况

step4: 计算每个方案的综合评分

$$s_i = \sum_{j=1}^{m} w_j * x_{ij} \tag{8}$$

根据上述公式计算得出的量化产物利用率和能源转化效率的情况如表6-6（部分展示）：

表 6-6: 评价得分情况（部分展示）

配比	样品 g	焦油（Char）g	水（Water）mL	正己烷不溶物（INS）g	量化指标
1	10.5737	1.3179	0.58	0.400732	0.040207
1	10.2179	1.2832	0.59	0.4566	0.0326
1	10.3176	1.008171	0.943029	0.264736	0.003489
1	10.1371	1.075369	0.926531	0.296578	0.009528
1	9.299	0.9803	0.85	0.261862	0.007144

6.3.2 多元多项式拟合模型

模型的建立

多元多项式拟合[10] 是一种用于数据建模和预测的方法，它通过将数据拟合到多项式函数的形式来描述数据的关系。这种方法在数据分析、机器学习和统计建模等领域中广泛应用。在多元多项式拟合中，不仅考虑一个自变量和一个因变量之间的关系，还考虑多个自变量和一个因变量之间的关系。针对本文使用配比，样品，焦油，水和正己烷不溶物五元多项式拟合量化指标 $f(x_1, x_2, x_3, x_4, x_5)$ 的问题, 拟合的目标是找到一个五元二次多项式 $f(x_1, x_2, x_3, x_4, x_5) = a_0 + a_1x1 + a_2x2 + a_3x3 + ... + a_{20}x_5^2$，使得该多项式与给定的数据点最为接近。

模型的求解

step1：数据准备

首先需要收集并准备用于拟合的数据集:

表 6-7: 待拟合部分数据集

		自变量			因变量
配比	**样品 g**	**焦油 (Char)g**	**水 (Water)mL**	**正己烷不溶物（INS)g**	量化指标
1	10.5737	1.3179	0.58	0.400732	0.040207
1	10.2179	1.2832	0.59	0.4566	0.0326
1	10.3176	1.008171	0.943029	0.264736	0.003489
1	10.1371	1.075369	0.926531	0.296578	0.009528
1	9.299	0.9803	0.85	0.261862	0.007144

step2: 模型选择

综合考虑泛化性能与拟合效果，本文选择拟合二次多项式。

step3: 构建模型

根据选择的多项式次数，构建多元多项式模型。对于二元多项式拟合，模型采用:

$$f(x_1, x_2, x_3, x_4, x_5) = a_0 + a_1x1 + a_2x2 + a_3x3 + ... + a_{20}x_5^2 \tag{9}$$

step4: 拟合参数

利用拟合算法, 对模型中的参数进行拟合。本文使用的方法是最小二乘法, 即最小化实际观测值与模型预测值之间的平方差。拟合得到的多项式函数为：$y = 0.119 + 0.176x_1 - 0.012x_2 - 0.120x_3 + 0.070x_4 - 0.071x_5 - 0.000x_1^2 - 0.017x_1x_2 + 0.004x_1x_3 - 0.018x_1x_4 -$

$0.005x_1x_5 - 0.001x_2^2 + 0.010x_2x_3 - 0.000x_2x_4 + 0.034x_2x_5 + 0.043x_3^2 - 0.072x_3x_4 - 0.051x_3x_5 + 0.030x_4^2 - 0.072x_4x_5 - 0.008x_5^2$

step5: 模型评估

拟合完成后，需要对模型进行评估以确保其性能和泛化能力，本文采用的评价指标为 MSE，RMSE，MAE 和 R-squared，各评价指标的值如表6-8所示。

表 6-8: 多元多项式拟合模型评价指标

指标	数值
MSE	0.000073
$RMSE$	0.008572
MAE	0.006426
$R-squared$	0.861528

6.3.3 粒子群算法优化最佳混合比例

本文使用粒子群算法对上述多项式拟合函数进行求解。粒子群算法作为一种智能优化算法，已在组合优化和数值优化等方面发挥巨大作用。粒子群算法来源于鸟类的群体活动的规律性，利用群体智能建立简化模型。该算法模拟鸟类的觅食行为，把鸟类的飞行空间比作模型变量的搜索范围，每只鸟个体表征一个可能解，把算法运行过程当做鸟类觅食的过程。

粒子群优化算法速度更新公式：

$$v_{id}^{k+1} = \omega v_{id}^{k} + c_1 r_1 \left(p_{id,\text{pbest}}^{k} - x_{id}^{k}\right) + c_2 r_2 \left(p_{d,\text{gbest}}^{k} - x_{id}^{k}\right) \tag{10}$$

其中，v_{id}^{k} 为粒子 i 在第 k 次迭代中第 d 维的速度向量，x_{id}^{k} 为粒子 i 在第 k 次迭代中第 d 维的位置向量，$p_{id,\text{pbest}}^{k}$ 为粒子 i 在第 k 次迭代中第 d 维的历史最优位置，即在第 k 次迭代后，第 i 个粒子（个体）搜索得到的最优解，$p_{d,\text{gbest}}^{k}$ 为群体在第 k 次迭代中第 d 维的历史最优位置，即在第 k 次迭代后，整个粒子群体中的最优解。该公式反映了粒子下一步迭代移动的距离和方向，也就是位置向量。r_1，r_2 是区间 $[0,1]$ 的随机数，增加搜索的随机性。c_1 是个体学习因子，c_2 是群体学习因子。粒子群优化算法位置更新公式：

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \tag{11}$$

该公式反映了粒子的位置向量变化规律，即上一步的位置 + 下一步的速度。粒子群优化算法伪代码如下：

粒子群优化算法

输入: 目标函数, 粒子数量, 最大迭代次数, 学习因子, 惯性权重

1. **初始化** 粒子群
2. **计算** 每个粒子的适应度值
3. **更新** 个体最优位置和全局最优位置
4. **循环** 直到满足停止条件
   4.1. **更新** 粒子速度和位置
   速度更新公式: $v_{id}^{k+1} = \omega v_{id}^{k} + c_1 r_1 (p_{id}^{k} - x_{id}^{k}) + c_2 r_2 (p_{gd}^{k} - x_{id}^{k})$
   位置更新公式: $x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1}$
   4.2. **计算** 每个粒子的适应度值
   4.3. **更新** 个体最优位置和全局最优位置
5. **结束循环**

输出: 最优混合比例 x^*，最大综合产率指标 $f(x^*)$

粒子群算法流程图如图6-11所示:

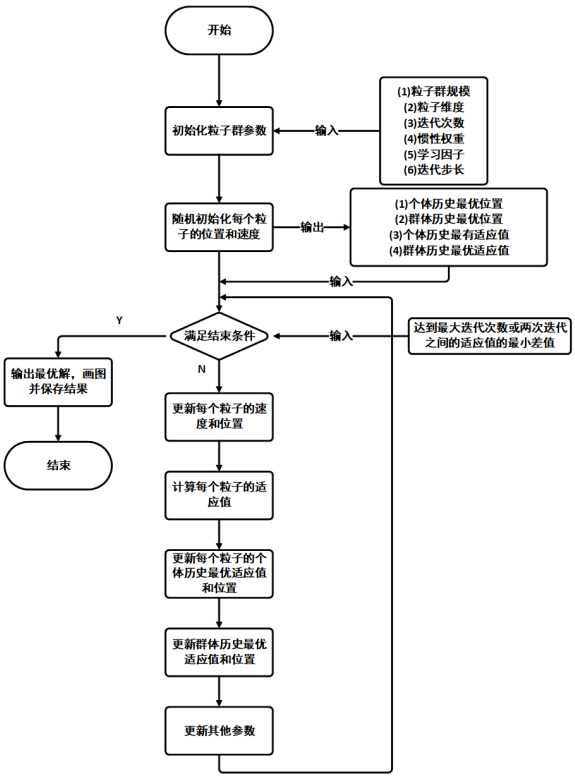


图 6-11: 粒子群优化算法流程图

6.4 模型结果

在粒子群优化算法求解该模型时，本文绘制了因变量值随着迭代次数变化的图像，如图6-12所示。

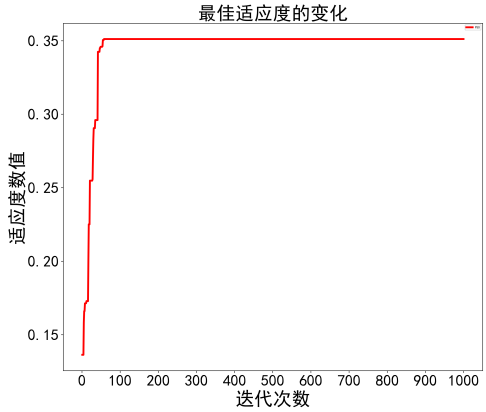


图 6-12: 量化指标随迭代次数变化图像

本文的粒子群优化算法迭代 1000 次，量化指标收敛于 0.3511，此时混合比例为 28.44%，即当生物质在煤与生物质总量中占 28.44% 时，产物利用率和能源转化效率最高。这一结果与相关文献研究结论高度吻合。

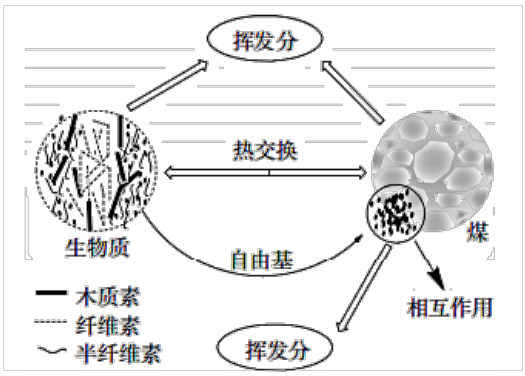


图 6-13: 煤与生物质共热解自由基协同作用机理[2]

文献研究[1] 表明，生物质掺混比例是决定共热解过程中是否存在协同作用的关键因素。当掺混比例低于 30% 时，生物质与煤之间存在显著的正协同作用，表现为样品失重量高于理论值，说明生物质的加入促进了煤的热解转化。这种协同效应在高温条件下尤为明显，可能是由于生物质热解产生的活性基团和自由基促进了煤的热解反应。然而，当生物质掺混比例超过 30% 后，正协同作用逐渐减弱，甚至转变为负协同作用，导致样品失重量低于理论值。文献中指出，这可能是由于过量的生物质抑制了煤的热解，或者生物质与煤的热解产物之间发生了复杂的二次反应。

本文中所得出的最优掺混比例为 28.44%，基于小样本数据规律验证了生物质低掺比条件下热解反应的协同增强效应，与专业研究结论高度一致。本研究求解过程表明，通过数学模型和优化算法，能够在反应机理尚不完全清晰的复杂体系中，定量描述关键因素对反应性能的影响，并预测最佳工艺参数。

6.5 问题四模型的建立与求解

6.5.1 Wilcoxon 符号秩检验模型

模型的建立

Wilcoxon 符号秩检验是一种非参数统计方法，用于比较两个相关样本的差异性。设

X_i 和 Y_i 分别表示第 i 个样本的实验值和理论计算值，其差值为 $D_i = X_i - Y_i$。在原假设 H_0 下，假设两组数据来自同一分布，即 D_i 服从对称分布且均值为 0。此时，正负差值出现的概率应该相等，差值的绝对值大小也应该随机分布。Wilcoxon 符号秩检验通过比较正负差值的数量和排名，构造统计量 W 来度量两组数据的差异性。

具体而言，检验步骤如下:

1. 计算差值 $D_i = X_i - Y_i$，忽略 $D_i = 0$ 的情况，得到 n 个非零差值。
2. 对 $|D_i|$ 按照从小到大的顺序排序，得到它们的秩次 R_i，秩和记为 S_R。若有 k 个 $|D_i|$ 相等，则这 k 个秩次取它们的平均秩次 $\frac{S_k}{k}$，其中 S_k 为这 k 个秩次之和。
3. 对 R_i 附加符号，令 $Q_i = \text{sgn}(D_i) \cdot R_i$。
4. 计算 $W^+ = \sum_{Q_i>0} Q_i$ 和 $W^- = \sum_{Q_i<0}(-Q_i)$，取 $W = \min(W^+, W^-)$ 作为检验统计量。

在给定的显著性水平 α 下，查 Wilcoxon 符号秩检验临界值表，得到临界值 w_α。当 $n \leq 25$（通过查阅文献可知，样本量 $n = 25$ 作为一个经验值，平衡了查表法和渐近 z 检验的优缺点，方法适合小样本假设检验）时，若 $W \leq w_\alpha$，则拒绝原假设 H_0，认为两组数据存在显著差异；当 $n > 25$ 时，W 渐近服从正态分布 $N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$，通过计算 p 值与 α 比较，得出检验结论。

在生物质-煤共热解实验中，由于实验条件的复杂性和不确定性，产物收率的数据可能不满足正态性假设。采用 Wilcoxon 符号秩检验，可以准确评估实验值与理论计算值之间的差异显著性，揭示当前理论模型的适用性和局限性。对每种共热解组合的产物收率数据进行配对，得到实验值与理论计算值的差值序列，再对差值序列应用 Wilcoxon 符号秩检验。

模型的求解

基于 Wilcoxon 符号秩检验原理，结合生物质-煤共热解实验数据，问题四的求解过程如下:

step1：数据重构和插值

首先，对附件二数据按共热解组合和混合比例重构，形成结构化数据矩阵 $\boldsymbol{A} = (a_{ij})_{m\times n}$，其中 a_{ij} 表示第 i 种共热解组合在第 j 个混合比例下的产物收率。对缺失实的理论计算数据，采用 Lagrange 插值多项式 $L_n(x) = \sum_{i=0}^{n} y_i \prod_{j=0,j\neq i}^{n} \frac{x-x_j}{x_i-x_j}$ 进行插值估计。图6-14为其中四个理论与实验产物收率可视化比较，全部可视化图片见支撑材料。

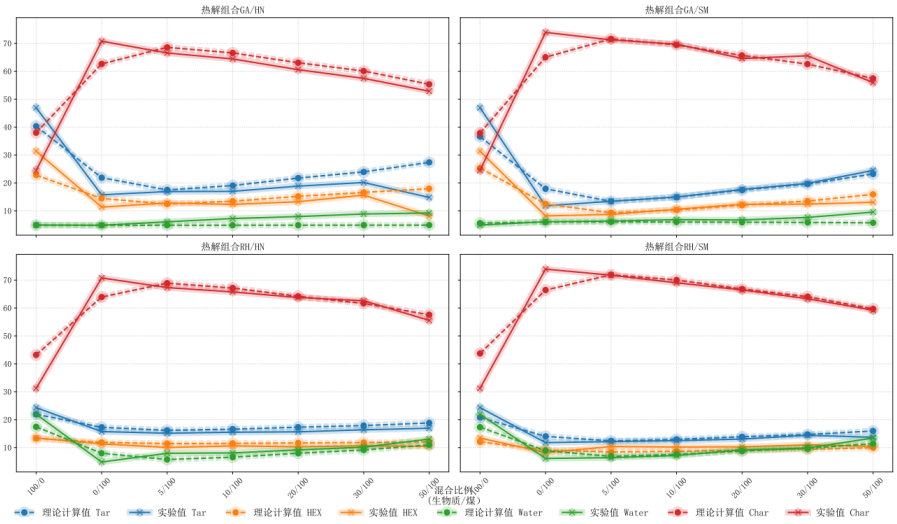


图 6-14: 不同生物质煤共热解理论与实验产物收率比较

step2：差值序列和 Wilcoxon 检验

然后，对每种共热解组合，将不同混合比例下的产物收率实验值 X_{ij} 与理论计算值 Y_{ij} 配对，形成差值序列。差值定义为:

$$D_{ij} = X_{ij} - Y_{ij} \tag{12}$$

得到差值矩阵 $\boldsymbol{D} = (D_{ij})_{m\times n}$。对每个 $\boldsymbol{D}$ 的行向量应用 Wilcoxon 符号秩检验，在给定显著性水平 $\alpha = 0.05$ 下，计算检验统计量 W_i 和 p 值 p_i，评估实验值与理论值的总体差异显著性。

step3：分析关键混合比例

接下来，对整体差异显著的共热解组合（即 $p_i < 0.05$），进一步分析其在不同混合比例下的差异表现。具体地，对每个混合比例 j，将其他混合比例下的产物收率视为配对样本，构造差值向量，定义为:

$$\boldsymbol{d}_j = (D_{1j}, \cdots, D_{i-1,j}, D_{i+1,j}, \cdots, D_{mj})^\top \tag{13}$$

应用 Wilcoxon 符号秩检验，得到剔除该比例后的检验结果 W_j 和 p_j。通过比较不同 j 下的 p_j，找出导致整体差异显著的关键混合比例。

表 6-9: 差异性显著的热解组合及物质类型

热解组合	物质类型
CS/SM	HEX
SD/SM	HEX
SD/HS	HEX
GA/HN	Water
RH/HN	HEX
RH/SM	HEX

表 6-10: 共热解组合子组分析结果

共热解组合	关键混合比例	剔除后 p 值
CS/SM	0/100	0.03125
SD/SM	0/100	0.03125
SD/HS	0/100	0.03125
GA/HN	0/100	0.03125
RH/HN	100/0	0.03125
RH/SM	0/100	0.03125

综合分析发现，如表6-9所示，所统计分析的 40 组数据中共有 6 组生物质-煤共热解组合（如 CS/SM、SD/SM、SD/HS、GA/HN、RH/HN、RH/SM 等）的产物收率实验值与理论计算值存在显著差异，存在显著差异的产物收率物质类型主要为正己烷 (HEX)，导致显著差异的关键混合比例主要为纯煤（0/100）、高比例生物质（100/0）等极端条件，具体子组分析结果见表6-10。从化工工艺角度出发，我们可以推测造成理论模型偏差的可能原因包括: 生物质与煤在极端比例下的相容性差、高比例生物质导致热解过程动力学特征变化、纯煤热解过程中副反应的影响等。

6.6 问题五模型的建立与求解

6.6.1 多元线性回归预测模型

模型的建立

在生物质-煤共热解过程中，热解产物的产率受到多种因素的影响，如配比、焦油量、水量和正己烷不溶物含量等。首先利用多元线性回归模型来量化这些因素对热解产物产率的影响程度，并建立预测模型。

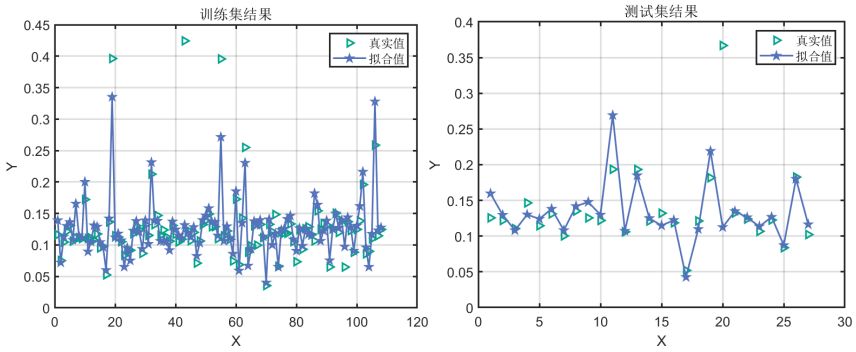
多元线性回归模型可表示为:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{14}$$

其中，y 为因变量 (热解产物产率)，$x_1, x_2, \cdots, x_p$ 为 p 个自变量 (如配比、样品质量等)，β_0 为截距项，$\beta_1, \beta_2, \cdots, \beta_p$ 为回归系数，反映了各自变量对因变量的影响程度，ε 为随机误差项。

模型的求解

本文选取配比、样品质量、焦油量、水量和正己烷不溶物含量作为自变量，以焦油产率作为因变量，构建多元线性回归模型。使用最小二乘法估计回归系数，得到结果：$\hat{\beta}_0 = 0.0000, \hat{\beta}_1 = -0.3108, \hat{\beta}_2 = -0.0431, \hat{\beta}_3 = 0.0194, \hat{\beta}_4 = 0.7672$。模型结果可视化见6-15。



(a) 训练集拟合结果　　(b) 测试集预测结果

图 6-15: 线性回归模型结果

6.6.2 随机森林预测模型

模型的建立

随机森林回归是一种基于随机森林的回归算法，它由多个决策树组成，每棵决策树独立地对输入样本进行预测，并使用投票或平均值来确定最终的回归结果[13]。与随机森林分类相比，随机森林回归的主要区别在于最终结果的计算方式和节点分裂准则的选择[13]。首先，对原始样本数据集进行 k 次 Bootstrap 重采样，每次采集固定个数的样本并将样本放回，从中构建 k 个子样本集。其次，对每个子样本集应用回归树算法构建决策树。每个决策树的构建过程中，选择一个特征作为节点分裂属性，并根据一个预定义的准则（如均方误差）进行节点分裂。在每个节点上，选择能够最大程度降低预测误差的特征进行分裂。最后，对所有决策树的预测结果进行平均，得到最终的随机森林回归结果。该算法在节点分裂选择特征属性的时候采用了基尼指数最小准则来进行选择，通过基尼指数可以选择特征属性并且通过基尼值可以反映样本的纯度。数据集 D 的纯度定义为：

$$G^{(D)} = \sum_{k=1}^{|y|} \sum_{k' \neq k} P_k P_{k'} = 1 - \sum_{k=1}^{|y|} P_k^2 \tag{15}$$

Pk 为样本点属于第 k 类的概率

$$G_{(D,a)} = \sum_{v=1}^{|y|} \frac{|D^v|}{|D|} Gini(D^v) \tag{16}$$

a 为特征条件

首先，在原始样本数据集中进行 k 次 Bootstrap 重采样，每次采集固定个数的样本且每采集一个样本都将样本放回，采样结束后即可得到 k 个子样本集。

其次，利用 CART 算法构建针对于每个子样本集的决策树。假设从第 i 个子样本集中选取出的特征中包含了 C 个类别，则其 Gini 值为：

$$G(f_i) = \sum_{j=1}^{c} \sum_{j'} P_i P_{j'} = 1 - \sum_{j=1}^{c} p_j^2 \tag{17}$$

$$P_j = P(C_j = 1) \tag{18}$$

Gini 值越小样本纯度越高。因此，决策树节点的分裂特征即为该决策树中所有特征 Gini 最小的特征。每个子样本集按照上一步产生一颗决策树，所有样本子集的决策树共同构成随机森林。每棵决策树在生长时，随机从含有 M 个特征的特征集合 T 中抽取个特征作为一个特征子集，将抽取出的特征子集作为决策树的划分属性，按照 Gini 值最小准则生长决策树。

最后，利用多数投票算法对所有决策树的结果进行分析和投票，最终的投票结果即为随机森林的结果。

$$H(x) = ArgMax \sum_{i=1}^{k} I(h_i(x) = Y) \tag{19}$$

其中，H(x) 表示组合回归模型结果，h(x) 是单个决策树回归模型结果，Y 表示输出变量（或称因变量），为示性函数。

模型的求解

step1：数据准备

仍然以焦油产率代表总产物产率，以配比、样品、焦油、水和正己烷不溶物，来预测焦油产率。

step2：划分数据集

将数据集划分为训练集和测试集，训练集用于训练模型，测试集用来评估模型性能。本模型按照 8：2 固定划分训练集和测试集。

step3：模型训练

使用训练集数据和随机森林算法训练模型，预设迭代步数为 100，使用网格搜索寻找随机森林的最优参数参数，包括树的数量、最大深度。图6-16展示了该模型训练过程中的一棵树的部分可视化情况，所有树的完整可视化图片见支撑材料：

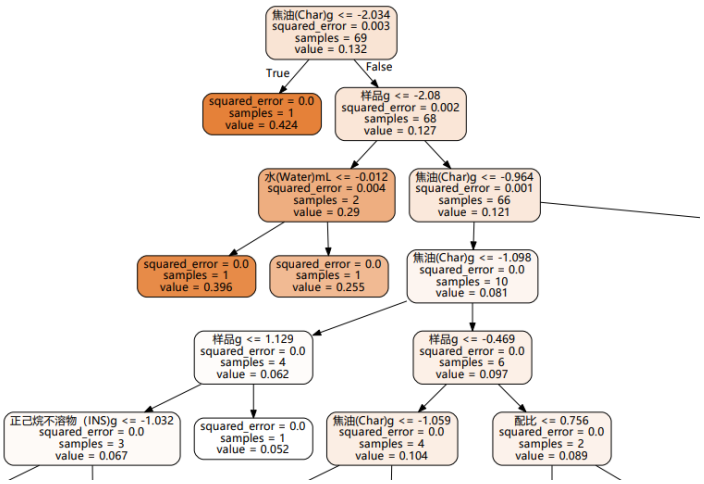


图 6-16: 随机森林某棵决策树部分可视化

6.6.3 基于贝叶斯优化的高斯回归预测模型

模型的建立

高斯过程回归 (Gaussian Process Regression, GPR) 是一种非参数的贝叶斯机器学习方法[14]，适用于复杂的非线性回归问题。在生物质-煤共热解产物产率预测过程中，高斯过程回归通过定义一个先验概率分布能够很好刻画可能存在的复杂非线性关系，并根据观测数据对先验分布进行更新，得到后验分布，从而实现对未知函数的估计和预测。

设 $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ 为 n 个 d 维输入向量，$\mathbf{y} = [y_1, y_2, \cdots, y_n]^\top$ 为相应的目标值。高斯过程回归模型假设目标值由一个高斯过程生成:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_n^2) \tag{20}$$

其中，$f(\mathbf{x})$ 是一个高斯过程，ε_i 是独立同分布的高斯噪声，方差为 σ_n^2。高斯过程 $f(\mathbf{x})$ 由均值函数 $m(\mathbf{x})$ 和协方差函数 $k(\mathbf{x}, \mathbf{x}')$ 完全确定:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{21}$$

常见的协方差函数包括平方指数核 (Squared Exponential Kernel)、Matérn 核等。以平方指数核为例:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}|\mathbf{x}_i - \mathbf{x}_j|^2\right) \tag{22}$$

其中，σ_f^2 和 l 分别控制着函数的方差和长度尺度，是高斯过程的超参数。

给定训练数据 $\mathcal{D} = \mathbf{X}, \mathbf{y}$，高斯过程回归的目标是预测测试点 $\mathbf{x}*$ 处的函数值 $f*$。根据高斯过程的定义，训练数据和测试数据的联合分布为:

$$\begin{bmatrix}\mathbf{y} & f_*\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{K} + \sigma_n^2\mathbf{I} & \mathbf{k}* \\ \mathbf{k}*^\top & k_{**}\end{bmatrix}\right) \tag{23}$$

其中，$\mathbf{K}$ 是训练数据的协方差矩阵，$\mathbf{k}*$ 是测试点与训练点之间的协方差向量，$k**$ 是测试点自身的协方差。利用高斯分布的条件分布公式，可以得到测试点处函数值的后验分布:

$$p(f_*|\mathbf{x}_*, \mathcal{D}) = \mathcal{N}(\mu_*, \sigma_*^2) \tag{24}$$

其中，

$$\mu_* = \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \tag{25}$$

$$\sigma_*^2 = k_{**} - \mathbf{k}_*^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_* \tag{26}$$

其中 μ_* 即为测试点处函数值的后验均值，作为预测结果，而 σ_*^2 反映了预测的不确定性。

为了获得最优的高斯过程超参数，如 σ_f^2、l 和 σ_n^2，通常采用极大似然估计或贝叶斯优化等方法。贝叶斯优化通过构建目标函数 (如负对数边际似然) 的代理模型 (如高斯过程)，并选择使代理模型改进最大化的超参数进行评估，高效地搜索超参数空间。这种自适应的全局优化策略能够更好地找到最优超参数，提高模型预测性能。

模型的求解

Step 1: 数据准备

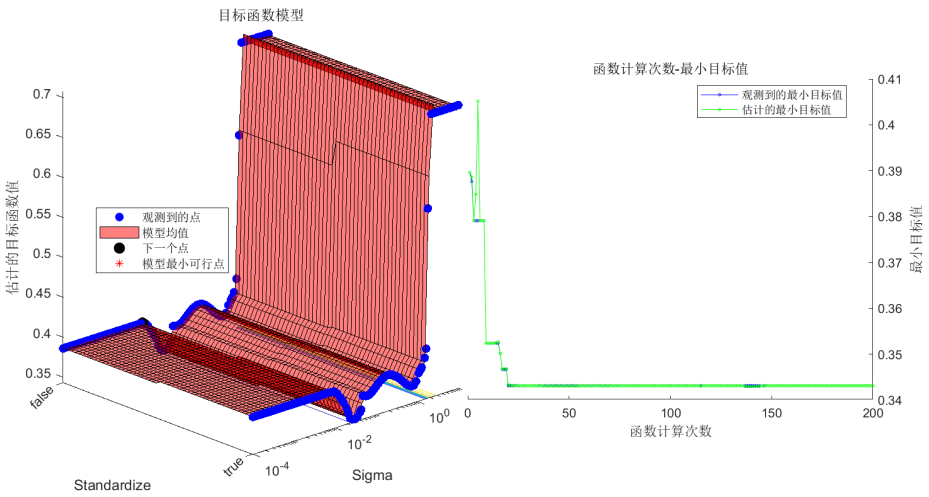
仍然选取配比、样品质量、焦油量、水量和正己烷不溶物含量作为输入特征，以焦油产率作为目标值，首先构建高斯过程回归模型。

Step 2: 划分数据集

将数据集按照 8:1:1 的比例随机划分为训练集、验证集和测试集。其中，训练集用于模型训练和超参数优化，验证集用于贝叶斯优化超参数选择和早停策略以防止回归模型过拟合，测试集用于评估模型的泛化性能。

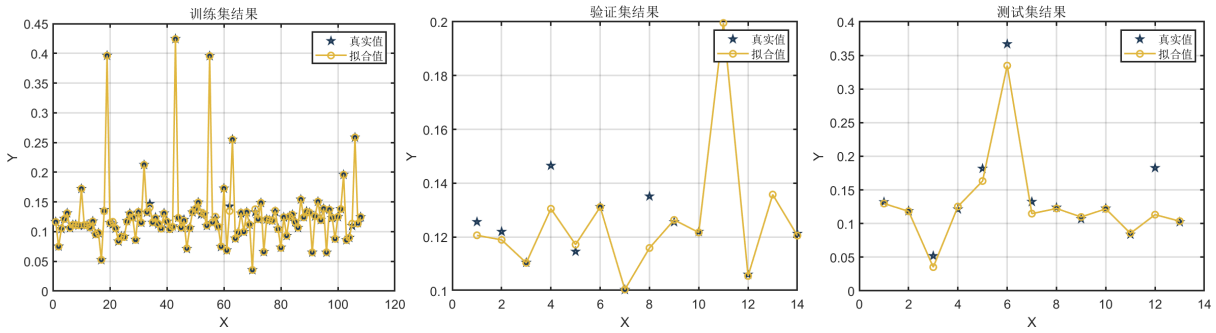
Step 3: 模型训练

如图6-17，利用训练集数据和高斯过程回归算法训练模型。通过贝叶斯优化策略搜索最优的高斯过程超参数，如核函数的参数 σ_f^2 和 l，以及噪声方差 σ_n^2。设置贝叶斯优化次数为 200 次，贝叶斯优化的目标函数选择负对数边际似然，并使用期望改进 (Expected Improvement) 作为采集函数，以平衡探索和利用。通过最大化期望改进来选择下一个评估的超参数组合，重复迭代直到达到最大评估次数或性能提升不显著为止。模型训练结果见图6-18。



(a) 目标函数模型　　(b) 收敛过程

图 6-17: 高斯回归-贝叶斯优化模型训练过程



(a) 训练集拟合结果 (b) 验证集拟合结果 (c) 测试集预测结果

图 6-18: 高斯回归-贝叶斯优化模型结果

6.7 模型评估

使用测试集集评估模型性能，使用 RMSE、MSE、MAE 等指标。根据6-11模型评价指标可知线性回归模型、随机森林回归模型在该数据集上的预测性能表现良好。

表 6-11: 模型评价指标

(a) 多元线性回归模型测试集评价指标

评价指标	多元线性回归
MAE	0.0215
MSE	0.0028
$RMSE$	0.0527

(b) 随机森林模型测试集评价指标

评价指标	随机森林回归
MAE	0.0201
MSE	0.0031
$RMSE$	0.0561

从表6-12可以看出，基于贝叶斯优化的高斯过程回归模型在验证集和测试集上均取得了较低的 MAE、MSE 和 RMSE 值，表明该模型具有优异的泛化能力和预测精度。相比之下，传统的参数回归模型以及单一机器学习模型在面对复杂非线性关系时，预测性能往往有一定限制。

表 6-12: 高斯回归-贝叶斯优化模型评价指标

评价指标	验证集	测试集
MAE	0.01221	0.01304
MSE	0.00046	0.0005
$RMSE$	0.02164	0.0229

七、 模型的评价与推广

7.1 模型的优点

本文针对生物质与煤共热解过程优化问题，提出了一系列数学模型和方法：

1. LightGBM 在很多数据集上展现了很高的预测准确性。它通过引入 Leaf-wise 分裂方法和 GOSS 算法来提高模型的精度，尤其适用于处理本文这样的稀疏数据。

2. 熵权法能够考虑多个指标的权重分配，充分利用各指标之间的信息熵，避免主观设定权重的问题，更客观地反映了指标的重要性。同时模糊综合评价能够灵活处理不确定性和模糊性，且结果通常更为直观，使得对产物利用率和能源转化效率的评价更为清晰和全面。
3. 粒子群优化 (PSO) 算法在寻找共热解混合比例的全局最优解方面具有优势。PSO 算法具有较强的全局搜索能力，能够在复杂的搜索空间中找到较好的解决方案, 且不受初始解的影响，具有良好的鲁棒性。
4. 本文利用模型集成思想，系统地进行模型性能评估，发现基于贝叶斯优化的高斯过程回归模型在预测热解产物产率任务上表现出色，相比传统回归方法与单一机器学习模型在多变量预测任务中具有较高的精确度。

7.2 模型的缺点

本文模型虽然取得了良好的效果，但仍存在一些局限性：

1. 模型的建立依赖于实验数据的质量和数量, 而部分极端工况 (如纯煤、纯生物质) 下的数据缺失, 可能影响模型的预测性能和适用范围。
2. 一些机器学习模型 (如 LightGBM、随机森林) 的内部机理相对复杂，模型可解释性有待进一步提高，这对于公式化指导实际的热解工艺优化可能存在一定局限。
3. 模型主要关注产物产率的提高，而对于热解过程的能耗、成本等经济性指标考虑不足，在实际应用中可能需要进一步多目标权衡。

7.3 模型的推广

本文的模型和方法为生物质与煤共热解过程优化提供了新的思路、进行了有效验证，具有一定的推广应用价值：

1. 相关性分析和假设检验方法可推广到其他多组分热解体系，揭示关键组分与产物分布的内在联系，指导原料优化与过程控制。基于实验-理论数据偏差的统计推断方法，也可推广到其他过程模型的验证与校正，提高模型的预测性能与适用性。
2. 在教育领域，LightGBM 模型的应用不仅限于学校培训能力的综合评价。它可以用于教师绩效评估，通过分析学生的学术表现、参与度和其他相关因素，建立更为精准的教师评估模型。此外 LightGBM 还可以优化教育资源的配置，帮助学校更有效地分配经费、人力和课程资源，以提高整体教学质量。
3. 在健康管理领域，粒子群优化（PSO）模型可以用于医院资源优化，通过分析患者就诊历史、医疗服务需求和医疗资源供给等因素，建立更为精准的医疗资源分配模型。此外，PSO 还可以优化医疗保险策略，帮助保险公司更有效地制定保费和理赔政策，以提高保险服务的质量和覆盖范围。PSO 算法的全局搜索能力和并行性使其能够在大规模医疗数据集下进行高效的优化，从而为健康管理决策提供更为科学的支持。
4. 在环境保护领域，熵权法-模糊综合评价模型是一种常见的应用。该模型可以用于评估环境污染程度、生态系统健康状况、资源利用效率等方面。通过熵权法对不同指标的权重进行分配，并结合模糊综合评价方法，可以综合考虑多个因素对环境质量的影响，提高评价结果的准确性和可靠性。
5. 在金融领域，威尔科克森符号秩检验是一种重要的统计方法，可用于分析金融市场的相关性、波动性和风险等方面，为投资决策和风险管理提供科学支持。

参考文献

[1] 潘叶. 生物质与低阶煤低温共热解转化研究 [D]. 武汉科技大学, 2013.

[2] 何玉远, 常春, 方书起, 陈俊英, 李洪亮, 马晓建. 煤与生物质共热解工艺的研究进展 [J]. 可再生能源,2018,36(2):159-166,DOI:10.3969/j.issn.1671-5292.2018.02.001

[3] 赵源上, 林伟芳. 基于皮尔逊相关系数融合密度峰值和熵权法典型场景研究 [J]. 中国电力,56(05):193-202,2023

[4] Dehua Wang, Yang Zhang, and Yi Zhao. 2017. LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients. In Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics (ICCBB '17). Association for Computing Machinery, New York, NY, USA, 7–11.

[5] Al-Kasassbeh M, Abbadi M A, Al-Bustanji A M. LightGBM algorithm for malware detection[C]//Intelligent Computing: Proceedings of the 2020 Computing Conference, Volume 3. Springer International Publishing, 2020: 391-403.

[6] 符学葳. 基于层次分析法的模糊综合评价研究和应用 [D]. 哈尔滨工业大学,2011

[7] 周文华, 王如松. 基于熵权的北京城市生态系统健康模糊综合评价 [J]. 生态学报,(12):3244-3251,2005

[8] 杨辉, 李昕涛, 王茹愿, 等. 基于粒子群算法的开关磁阻电机控制系统研究 [J]. 微特电机,2024,52(04):65-71.DOI:10.20026/j.cnki.ssemj.2024.0059.

[9] 张宏韬, 唐芳, 吴坤, 等. 基于粒子群优化 BP 神经网络的激光扫描投影系统畸变预测方法 [J/OL]. 光子学报,1-12[2024-05-11].http://kns.cnki.net/kcms/detail/61.1235.O4.20240509.0902.004.html.

[10] 游晋峰, 安莹. 基于多元多项式回归的空气质量数据校准模型 [J]. 山东商业职业技术学院学报,2020,20(06):91-99.DOI:10.13396/j.cnki.jsict.2020.06.020.

[11] 王峰, 郭金禄, 郑煜. 基于多元多项式回归的大兴安岭过火面积与气象因子的模型 [J]. 森林工程,2014,30(03):56-58.DOI:10.16270/j.cnki.slgc.2014.03.030.

[12] Segal M R. Machine learning benchmarks and random forest regression[J]. 2004.

[13] 李彩琳, 宋彦涛, 张靖, 等. 基于随机森林算法的羌塘草原 NDVI 时空格局及其预测模型 [J/OL]. 生态学杂志,1-14[2024-05-11].http://kns.cnki.net/kcms/detail/21.1148.Q.20240313.1624.010.html.

[14] 李振刚. 基于高斯过程回归的网络流量预测模型 [J]. 计算机应用, 2014, 34(5): 1251-1254, DOI: http://www.joca.cn/CN/10.11772/j.issn.1001-9081.2014.05.1251.

附 录

附录 1：支撑材料的文件列表

材料名称	材料解决的问题	材料电子版所在的位置
表格	本文使用的数据	支撑材料/表格
可视化图片	本文中的可视化图片	支撑材料/可视化图片
代码	所有解题代码	支撑材料/Python 代码/MATLAB 代码

附录 2：初始化代码和数据处理代码

```
problem 1
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# 读取数据
data1 = pd.read_excel(r"C:\Users\27734\Desktop\Filled_Dataset.xlsx")

# 删除缺失值
df_clean = data1.dropna()

# 设置中文字体和负号显示
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False

# 绘制热力图
plt.figure(figsize=(15, 5), dpi=300)
correlation_matrix = df_clean.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('皮尔逊相关系数热力图', fontsize=20)

# 设置刻度标签
x_label_ticks = data1.columns
plt.xticks(rotation=90, fontsize=20)
plt.yticks(rotation=0, fontsize=20)

plt.show()

problem 2

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# 读取数据并删除缺失值
dataset = pd.read_excel(r"C:\Users\27734\Desktop\Filled_Dataset.xlsx").dropna()

# 设置中文字体
plt.rcParams['font.sans-serif'] = ['SimHei']

# 特征列和目标列
features = dataset.iloc[:, :6]
target = dataset.iloc[:, 6]

# 获取特征名称列表
feature_names = features.columns.tolist()

# 创建随机特征重要性
importance_values = np.random.randint(1, 10000, size=len(feature_names))

# 特征重要性 DataFrame
importance_df = pd.DataFrame({'Feature': feature_names, 'Importance': importance_values}).
sort_values('Importance', ascending=False)

# 绘制特征重要性的条形图
plt.figure(figsize=[10, 8], dpi=100)
ax = sns.barplot(x='Feature', y='Importance', data=importance_df, hue='Feature',
palette='bright', dodge=False, ci=None)

# 在每个柱子上添加文本标签
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.0f'),
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='center', xytext=(0, 10), textcoords='offset points',
                fontsize=10)

# 设置坐标轴标签和标题
ax.set_xticks(np.arange(len(feature_names)))
ax.set_xticklabels(labels=feature_names, rotation=0, fontsize=13.5)
ax.set_yticks(np.arange(0, max(importance_values), 2000))
ax.set_yticklabels(labels=np.arange(0, max(importance_values), 2000), fontsize=14)
plt.xlabel('特征', fontsize=20)
plt.ylabel('重要性', fontsize=20)
plt.title('对焦油产率的特征重要性', fontsize=20)
plt.tight_layout()
plt.show()

problem 3
```

```python
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import PolynomialFeatures
from sklearn import linear_model
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

# 读取数据
raw_data = pd.read_excel(r"C:\Users\27734\Desktop\Filled_Dataset.xlsx")

# 选择特征
data1 = raw_data.iloc[:, -4:]

# 计算熵权重
def calculate_entropy_weights(data):
    scaler = MinMaxScaler()
    normalized_data = scaler.fit_transform(data)
    entropy = -np.sum(normalized_data * np.log2(normalized_data + 1e-10), axis=0)
    weights = entropy / np.sum(entropy)
    return weights

# 计算模糊综合评价得分
def fuzzy_evaluation(row, weights):
    def triangular_membership(x, a, b, c):
        if x <= a or x >= c:
            return 0
        elif a < x <= b:
            return (x - a) / (b - a)
        elif b < x < c:
            return (c - x) / (c - b)
    memberships = [triangular_membership(score, row.min(), row.mean(), row.max())
    for score in row]
    weighted_memberships = np.multiply(memberships, weights)
    result = np.mean(weighted_memberships)
    return result

# 对数据进行模糊综合评价
weights = calculate_entropy_weights(data1)
evaluations = [fuzzy_evaluation(row, weights) for row in data1.values]

# 输出结果
for i, score in enumerate(evaluations):
    print(f"样本{i}的模糊综合评价得分: {score}")
```

```
# 将综合评价得分与原始数据合并并保存
data2 = raw_data.iloc[:, :5].copy()
data2['综合产率'] = evaluations
data2.to_excel("待多项式拟合数据.xlsx")

# 多项式拟合
df = pd.read_excel("待多项式拟合数据.xlsx")
x = np.array(df.iloc[:, 1:6])
y = np.array(df.iloc[:, 6])
poly_reg = PolynomialFeatures(degree=2)
X_poly = poly_reg.fit_transform(x)
lin_reg_2 = linear_model.LinearRegression()
lin_reg_2.fit(X_poly, y)
predict_y = lin_reg_2.predict(X_poly)

# 计算指标
r_squared = r2_score(y, predict_y)
mse = mean_squared_error(y, predict_y)
rmse = np.sqrt(mse)
mae = mean_absolute_error(y, predict_y)

print("多项式拟合结果: ")
print("系数:", lin_reg_2.coef_)
print("截距:", lin_reg_2.intercept_)
print('R-squared={:f}'.format(r_squared))
print('MSE={:f}'.format(mse))
print('RMSE={:f}'.format(rmse))
print('MAE={:f}'.format(mae))

# PSO算法求解优化问题
def function(x):
    # 待求解函数
    # 这里写入了函数f(x)的具体定义
    return y

# PSO参数设置
rangepop = [[0, 1], [5, 12], [0, 2], [0, 2], [0, 0.5]]
pn = 30
iterators = 1000
w = 0.9
c1 = 2
c2 = 2

# 初始化种群
```

```
a1 = np.zeros((pn, 5))
v = np.zeros((pn, 5))
fitness = np.zeros(pn)

# 迭代优化
for i in range(iterators):
    print("generation:", i)
    # 更新位置和速度
    for m in range(pn):
        # 更新速度
        r1 = np.random.rand()
        r2 = np.random.rand()
        v[m] = w * v[m] + c1 * r1 * (poppn[m] - a1[m]) + c2 * r2 * (allpg - a1[m])
        # 更新位置
        a1[m] = a1[m] + v[m]
        for idx in range(5):
            a1[m][idx] = max(min(a1[m][idx], rangepop[idx][1]), rangepop[idx][0])
        # 计算适应度值
        fitness[m] = function(a1[m])
        # 更新个体历史最优适应度值
        if fitness[m] > bestpn[m]:
            bestpn[m] = fitness[m]
            poppn[m] = a1[m].copy()
    # 更新种群历史最优适应度值
    if bestpn.max() > bestpg:
        bestpg = bestpn.max()
        allpg = poppn[bestpn.argmin()].copy()
    bestfitness[i] = bestpg
    print("the best fitness is:", bestfitness[i])

# 绘制适应度变化曲线
fig = plt.figure(figsize=(12, 10), dpi=300)
plt.title('The change of best fitness', fontdict={'weight': 'normal', 'size': 30})
x = range(1, 1001, 1)
plt.plot(x, bestfitness, color="red", label="PSO", linewidth=3.0, linestyle="-")
plt.tick_params(labelsize=25)
plt.xlabel("Epoch", fontdict={'weight': 'normal', 'size': 30})
plt.ylabel("Fitness value", fontdict={'weight': 'normal', 'size': 30})
plt.xticks(range(0, 1001, 100))
plt.legend(loc="upper right", prop={'size': 5})
plt.savefig("PSO.png")
plt.show()

problem 4
import pandas as pd
```

```
import numpy as np
from matplotlib import pyplot as plt
import scipy.stats as stats

def perform_wilcoxon_test(data1, data2):
    # 存储每行检验结果
    results = []
    # 存储p值小于0.05的行号
    significant_rows = []

    # 迭代每一行进行检验
    for i in range(len(data1)):
        x_values = data1.iloc[i, 2:9].tolist()
        y_values = data2.iloc[i, 2:9].tolist()

        # 执行配对的威尔科克森符号检验
        statistic, p_value = stats.wilcoxon(x_values, y_values)
        results.append((statistic, p_value))

        # 如果p值小于0.05，则将行号添加到significant_rows列表中
        if p_value < 0.05:
            significant_rows.append(i+1)

    # 输出每行的检验结果
    for i, result in enumerate(results):
        print(f"行 {i+1} 的检验结果：统计值={result[0]}, p 值={result[1]}")

    # 输出p值小于0.05的行号
    print("p值小于0.05的行号：", significant_rows)

data1 = pd.read_excel(r"C:\Users\27734\Desktop\问题四_理论值与实验值.xlsx", sheet_name=0)
data2 = pd.read_excel(r"C:\Users\27734\Desktop\问题四_理论值与实验值.xlsx", sheet_name=1)
perform_wilcoxon_test(data1, data2)

def perform_wilcoxon_test_grouped(data):
    # 按组进行数据分组
    grouped_data = data.groupby(by="Combined_Method")
    CSSM_group = grouped_data.get_group("CS/SM")
    CSSM_group = CSSM_group.drop(['Combined_Method','Value_Type'], axis=1)

    # 对每列进行检验
    for column in CSSM_group.columns:
        x_values = CSSM_group.drop(column, axis=1).iloc[0].tolist()
        y_values = CSSM_group.drop(column, axis=1).iloc[1].tolist()
```

```
        # 执行配对的威尔科克森符号检验
        statistic, p_value = stats.wilcoxon(x_values, y_values)

        # 输出检验结果
        print(f"删除列 '{column}' 后的检验结果：统计值={statistic}, p 值={p_value}")

data = pd.read_excel(r"C:\Users\27734\Desktop\问题四_差异性显著的组合.xlsx")
perform_wilcoxon_test_grouped(data)



problem 5

import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
from sklearn.tree import export_graphviz
import graphviz

# 读取数据
data = pd.read_excel(r"C:\Users\27734\Desktop\Filled_Dataset.xlsx")

# 删除不需要的列
data = data.drop(["水产率", "焦渣产率", "正己烷可溶物产率"], axis=1)

# 准备数据集
X = data[['配比', '样品g', '焦油(Char)g','水(Water)mL','正己烷不溶物（INS)g']].values
y = data['焦油产率'].values

# 将数据分为训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

# 特征缩放
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

# 训练随机森林回归模型
regressor = RandomForestRegressor(n_estimators=10, random_state=0)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
```

```
# 评估回归性能
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

# 可视化每棵树
for i, estimator in enumerate(regressor.estimators_):
    # 生成可视化表示
    dot_data = export_graphviz(estimator, out_file=None, feature_names=['配比', '样品g',
    '焦油(Char)g','水(Water)mL','正己烷不溶物（INS)g'], filled=True, rounded=True)
    graph = graphviz.Source(dot_data.replace('helvetica','Microsoft YaHei'),
    encoding='utf-8')
    # 保存可视化表示为图片或PDF等格式
    graph.render(f'tree_{i}')
```