生物质和煤共热解技术中特征选择和优化研究



答辩人: 陆子凯、狄诗琪 指导教师: 赵英姿

Team # 2024050727407, Problem B

2024年9月23日



- 4 ロ ト 4 昼 ト 4 差 ト - 差 - りへの

Team # 2024050727407, Problem B

致谢与答辩

<ロ > ← □ 答辩人: 陆子凯、狄诗琪

- 全球能源需求的增长与环境问题 的日益严重, 开发清洁、高效的可 再生能源成为关键。
- 生物质是一种丰富的可再生资源, 与煤共热解可利用两者的协同效 应, 提高产物质量和能源转化效 率。

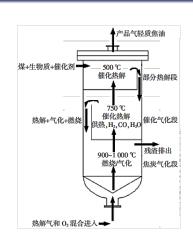


图 1: 生物质与煤共热解工 艺示意图[?]

 〈ロト 〈 □ ト 〈 豆 ト 〈 豆 ト 〈 豆 ト \ 豆 ト 〇 ○

 答辨人: 陆子凯、狄诗琪
 Team # 2024050727407, Problem B

研究目的与意义

研究目的

- 探究生物质与煤共热解过程 中影响产物组成和产率的关 键因素。
- 建立数学模型. 优化共热解 工艺参数,提高产物质量和 能源转化效率。

研究意义

- 为共热解技术的工业应用提 供理论指导, 推动技术进 步。
- 促进可再生能源的开发利 用,优化能源结构,减少环 境污染。

① 问题一:分析正己烷不溶物 (INS) 对各热解产物产率的影响。

- 问题一:分析正己烷不溶物(INS)对各热解产物产率的影 响。
- ② 问题二:研究 INS 及其与混合比例的交互效应对产物组成 的影响。

- 问题一:分析正己烷不溶物(INS)对各热解产物产率的影 响。
- □ 问题二:研究 INS 及其与混合比例的交互效应对产物组成 的影响。
- 3 问题三:确定最佳混合比例,提升产物利用率与能源转化效 率。

- 问题一:分析正己烷不溶物(INS)对各热解产物产率的影 响。
- □ 问题二:研究 INS 及其与混合比例的交互效应对产物组成 的影响。
- ③ 问题三:确定最佳混合比例,提升产物利用率与能源转化效 率。
- ④ 问题四:分析产物收率的实验值与理论值差异及其分布规 律。

- 问题一:分析正己烷不溶物(INS)对各热解产物产率的影 响。
- ② 问题二:研究 INS 及其与混合比例的交互效应对产物组成 的影响。
- ③ 问题三:确定最佳混合比例,提升产物利用率与能源转化效 率。
- △ 问题四:分析产物收率的实验值与理论值差异及其分布规 律。
- ⑤ 问题五:建立热解产物产率的预测模型,提高工艺控制精度。

问题分析

问题重述与分析

引言 问题 00000 000



图 2: 问题分析流程图

<ロ > ← □

数据预处理

附件数据缺失是影响分析质量的重要因素, 而直接忽视缺失值可 能导致结果偏差, 甚至使模型失效。针对附件一中缺失的正己烷

不溶物 (INS) 数据, 我们通过观察变量间的统计关系, 采用线性 插值法进行处理。正己烷不溶物与焦油比率数据波动较小(如表 ??),显示出潜在的线性关系,因此线性插值法能够较好地补足 缺失值,保持数据连续性。对于附件二中的实验数据缺失,考虑

到混合比例下数据可能呈现更复杂的趋势,我们选择了Lagrange 插值法。该方法通过构建插值多项式、能够更好地拟合非线性关 系,提高了缺失数据的估计精度。最后,我们对数据进行了重构转

换, 将混合比例从生物质/煤的比例形式转换为生物质/(生物质+ 煤)的小数格式,以提高数据处理的一致性与数值计算的便捷性。

在数据预处理中, INS 数据缺失的处理尤为关键。表??展示了正 己烷不溶物与其他变量间的化学比率统计数据。根据这些比率的 一致性, 我们选择了线性插值来填补缺失数据。

表 1: 化学比率的统计数据

化学比率	平均值	中位数	标准差
正己烷不溶物/焦油比率	0.283	0.280	0.106
正己烷可溶物/焦油产率比率	0.805	0.758	0.510
正己烷不溶物/水比率	0.313	0.284	0.186
焦渣产率/水产率比率	5.909	5.869	2.300

针对实验数据的缺失, 通过应用 Lagrange 插值对缺失实验数据进 行估计, 以确保插值结果不会偏离数据的总体趋势。

数据结构

下图??展示了数据预处理后中135个样本的部分变量和对应的数 值。

	配比 :	群品g ÷	焦油(Char)g ÷	水(Water)mL ÷	正己烷不溶物 (INS)g :	無油产率 :	水产率 :	焦渣产率 :	正己烷可溶物产率 :
	1.000000	10.5737	1.317900	0.580000	0.400732	8.124639	0.057944	0.759885	0.125236
	1.000000	10.2179	1.283200	0.590000	8.456600	0.125584	0.057900	0.758669	0.080900
	1.000000	10.3176	1.008171	0.943029	0.264736	0.097714	0.091400	0.736547	0.092842
	1.000000	10.1371	1.075369	0.926531	0.296578	0.106083	0.091400	0.728127	0.100176
	1.000000	9.2990	0.980300	0.850000	0.261862	0.105420	0.091408	0.726852	0.09112
138	8.166667	7.6822	0.880000	0.980800	0.177800	0.114551	0.117154	0.675796	0.091400
131	0.230769	7.1885	0.876892	0.880000	0.240083	0.121985	0.122500	0.647684	0.083348
132	0.230769	7.3468	0.912000	0.980800	0.215200	0.124136	0.122502	0.649085	0.094800
133	0.333333	8.5390	0.999800	1.380800	8.279642	8.117886	0.152243	0.612648	0.09435
134	0.333333	8.5047	0.975385	1.290000	0.232200	0.114688	0.152200	0.619763	0.087400

图 3: 数据结构

探索性数据分析

引言 问题重述与分析

数据预处理完成后,探索性数据分析(EDA)是进一步理解数据特征及其潜在规律的关键步骤。通过对描述性统计和相关性分析的深入探索,我们能够识别变量之间的关联性,并为后续模型构建提供可靠依据。

表 ?? 提供了部分变量的主要统计特征。其中,焦油产率的均值为 15.2,标准差为 2.5%。通过皮尔逊相关性分析,我们发现 INS 含量与焦油产率的相关性较强 (相关系数为 0.65),提示 INS 含量可能是影响焦油产率的重要因素。

表 2: 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
配比	135	1.00	0.048	0.295	0.301
焦油g	135	2.139	0.361	1.038	0.316
水 mL	135	1.91	0.58	1.037	0.295
INS g	135	0.935	0.025	0.304	0.158

4 ロ ト 4 回 ト 4 差 ト 4 差 ト 2 9 9 9 9 9

17 / 57

数据可视化分析

通过可视化分析,数据的分布和变量间关系得到了更为直观的展示。图??显示了焦油、水、INS等关键变量的箱线图,从中可以观察到变量的集中趋势及其潜在的异常点,各指标均不能通过正态性检验,所以在相关性分析前进行了 Z-score 标准化使数据具备标准正态分布的特征(均值为 0.标准差为 1)。

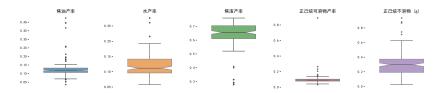


图 4: 关键变量箱线图

- 4 ロ ト 4 園 ト 4 恵 ト 4 恵 ト 9 年 9 9 0

答辩人: 陆子凯、狄诗琪

推理与总结

通过数据预处理与探索性分析填补了数据缺失, 且揭示了不同变 量之间的内在联系。探索性数据分析的初步结论为后续的模型构 建提供了重要的方向和依据:

- INS 含量与焦油产率显著相关、表明 INS 是影响焦油产率的 关键因素之一。
- 生物质混合比例增加时,焦油和水产率上升,而焦渣产率下 降。这一趋势反映了生物质高挥发分的特性。
- 数据中存在一定的线性与非线性关系, 提示后续分析需综合 考虑线性模型与非线性模型,以获得更精确的预测结果。

在接下来的研究中, 我们将基于这些分析结果, 进一步构建多元 回归模型与集成模型,尝试捕捉各变量间更为复杂的交互作用, 提升模型的预测精度与鲁棒性。

问题一: INS 对产率的影响分析

模型选择与理由

- 选择皮尔逊相关系数. 度量 INS 与产物产率之间的线性相关性。
- 适用于连续变量. 计算简便. 结果 易干解释。

皮尔逊相关系数公式

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

计算结果

变量关系	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

= 9000

结果解读

- INS 与焦油产率: INS 含量增 加有助于提高焦油产率。
- INS 与焦渣产率: INS 含量增 加会降低焦渣产率。
- INS 与水产率: 相关性较弱, 影响不明显。

结论

INS 在热解过程中促进了焦油的生 成,抑制了焦渣的形成,可能影响 了热解反应的化学路径。

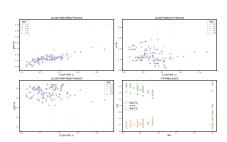


图 5: INS 与产物产率的关系

在散点图(图??)中呈现了生物质 混合比例与焦油产率、焦渣产率、 水产率的关系。随着生物质比例的 增加, 焦油和水的产率呈现上升趋 势, 而焦渣产率则下降。

<ロ > ← □

数据预处理与探索性分析 模型的建立与求解 致谢与答辩

问题二: INS 与混合比例的交互效应

模型选择与理由

- 选择 LightGBM 模型, 处理高维特征和复杂非线性关系。
- 引入交互特征 (INS × 配比), 捕捉两者的协同作用。

引言 问题重述与分析

- 基础特征: 配比、样品质 量、INS 含量等。
- 交互特征: INS × 配比。

模型训练与评估

- 使用预处理后的数据, 训练 LightGBM 模型。
- 通过交叉验证调整超参数. 提升模型性能。

结果解读

- 交互特征的重要性最高、表明 INS 与配比的协同效应显著影响焦油 产率。
- 基础特征的重要性次之, 仍对产 率有直接影响。

优化热解工艺时,应同时考虑 INS 含 量和混合比例. 充分利用协同效应, 提 高产物产率。

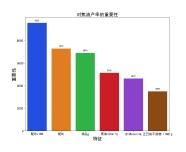


图 6: 焦油产率的特征重要 性排名

<ロ > ← □

问题分析

生物质-煤共热解过程涉及多个性能指标. 核心要点为:

- 多目标性: 需同时考虑产物利用率和能源转化效率
- 非线性关系: 各因素间存在复杂的交互作用
- 数据限制:实验样本量有限,难以应用传统大数据方法

理论依据

- ① 熵权法:基于信息论,能客观量化指标重要性
- ② 模糊综合评价:处理多指标决策问题的有效工具,将多目标问题简化为单目标优化
- ③ 多元多项式拟合: 捕捉变量间的非线性关系
- 4 粒子群优化 (PSO):解决高维非凸优化问题



29 / 57

熵权法核心推导

给定 m 个评价对象, n 个指标, 原始数据矩阵 $X = (x_{ii})_{m \times n}$:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}}, \quad E_j = -\frac{1}{\ln m} \sum_{i=1}^{m} p_{ij} \ln p_{ij}, \quad w_j = \frac{1 - E_j}{\sum_{k=1}^{n} (1 - E_k)}$$

计算得到的评价指标权重

指标	符号	权重 Wj
焦油产率	Y_{tar}	0.3572
焦渣产率	Y_{char}	0.2980
水产率	Y_{water}	0.1981
正己烷可溶物产率	Y_{INS}	0.1467

30 / 57

模糊综合评价实现目标融合

权重确定后, 模糊综合评价将多个目标融合为单一目标。其步骤 如下:

- 通过隶属函数将每个目标值标准化到区间 [0,1]。
- 对每个目标的隶属度进行加权求和:

$$S = W \cdot R = \sum_{j=1}^{m} w_j \cdot r_j$$

其中 w_i 为权重, r_i 为目标的隶属度, S 为综合得分。 综合得分 S 即为单一优化目标, 从而实现将多目标问题简化为 单目标优化。

Team # 2024050727407, Problem B

问题三: 熵权法-模糊综合评价模型

表 3: 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

32 / 57

问题三: 多元多项式拟合模型的构建与分析

模型构建理论基础

基于 Taylor 展开理论, 二阶多项式可以较好地逼近局部非线性关系:

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^T (\mathbf{x} - \mathbf{a}) + \frac{1}{2} (\mathbf{x} - \mathbf{a})^T H(\mathbf{a}) (\mathbf{x} - \mathbf{a})$$

其中, H(a) 为 Hessian 矩阵。

|拟合模型及其分析

$$f(x_1, \dots, x_5) = 0.119 + 0.176x_1 - 0.012x_2 - 0.120x_3 + 0.070x_4 - 0.071x_5 + (交互项)$$

模型解释:

- x1 (配比)和 x1 (水)对目标函数有正面影响
- x2 (样品)、x3 (焦油)和x5 (INS) 呈负面影响
- R-squared 为 0.861528,表明模型解释了 86.15% 的数据变异性

→ 4 = → 4 = → = ✓ 9 0 0

PSO 算法的理论基础

基于群体智能理论, 粒子的运动受三个因素影响:

- 惯性:保持当前运动趋势
- 认知:个体历史最优位置的吸引
- 社会: 群体最优位置的吸引

算法核心公式及其解释

速度更新:

$$v_{id}^{k+1} = \omega v_{id}^k + c_1 r_1 (p_{id}^k - x_{id}^k) + c_2 r_2 (p_{gd}^k - x_{id}^k)$$

位置更新:

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1}$$

其中, ω 为惯性权重, c_1 和 c_2 分别为个体和群体学习因子。

◆□ → ◆問 → ◆ = → ↑ □ ◆ ↑ Q ↑

问题三:结果分析与理论验证

优化结果及其物理意义

最佳混合比例: 28.44%

• 最大量化指标: 0.3511

• 物理解释: 在此比例下, 生物 质和煤的协同效应最强

与理论模型的一致性分析

- 结果落在 30% 临界值附近, 符 合文献 [?] 的预测
- 验证了低掺比下的正协同作用 假说,提供了定量优化结果

第36页

武汉科技大学 硕士学位论文

第四章 结论与展望

4.1 全文结论

 随着混合样中生物质掺混比例的加大,半焦产率不断降低,由66.23%下降到34.65%。 而气体产率由 7.78%上升到 19.52%, 焦油产率变化幅度同气体产率变化比较一致, 也呈现 上升趋势。由 8.74%上升到 18.58%。增大幅度较大。同时使油中直链偿烃。以及一些高限 加值的化合物如、卷、酚、苯、芴等含量不断增大。实现了低温煤焦油中高附加值化工产 品的富集。

 从 TG-DTG 曲线可以看出,随着温度升高,样品失重量逐渐增大;与煤单 相比,混合热解失重量及热解速率也陷之提高,混合物最终剩余物质量分数逐渐减小;随 着生物质掺湿比例增大,TG 曲线向低温区移动,煤热解起始温度逐渐降低,热解速率明 显增大,达到相同失重量所需的热解时间也缩短。

3) 生物质与煤共热解时,生物质的掺混比例低于 30%时,样品失重量比理论值大, 两者之间存在有利的协同作用。保伸了样品向热键方向讲行。且这些协同作用均在高温处 较为明显。当生物质排泄比例高于30%。样品失重量比理论值小、两者之间的协同作用越 来越弱、甚至可能抑制样品向热解方向的进行。

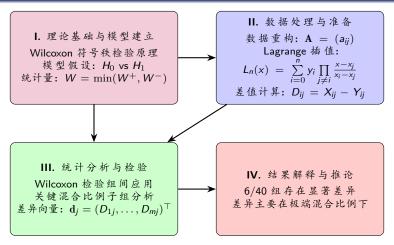
4)) 长镕煤、生物质及其不同配比的混合物构解动力学模型均符合一级动力学方程。 煤的平均活化修E,为46.43kJ/mol,指前因子为102.34 min⁻¹: 生物质的平均活化能E,=13.96 kJ/mol,指前因子为0.07min⁻¹:随着生物质掺湿比例的增大。混合后热解表观活化能和指 前因子明显减小、热解所需能量降低、系统反应速率加快。说明生物质的加入使混合热解 反应更加容易的进行。

图 7: 文献验证[]

4 D > 4 A > 4 B > 4 B >

答辩人: 陆子凯、狄诗琪

问题四: Wilcoxon 符号秩检验差异性分析模型



方法选择理由:

采用 Wilcoxon 符号秩检验, 适用于小样本和非正态分布的数据, 能够考虑差值的符号和 大小, 适合配对样本的比较, 对异常值不敏感, 提高结果的稳健性。

答辩人: 陆子凯、狄诗琪

Team # 2024050727407, Problem B

Wilcoxon 符号秩检验运行结果

通过 Wilcoxon 符号秩检验, 针对不同组别进行了组间和子组分析。以下展示了组间分析和各子组的检验结果。

图 8: 组间分析

图 9: CS/SM 组分析



Wilcoxon 符号秩检验运行结果

图 10: SD/SM 组分析

```
SD/HSSE

1 5001 - $Proped_pt_prope("15/M1")
2 5001-1500.cms("15/M1")
2 5001-1500.cms("15/M1")
2 5001-1500.cms("15/M1")
3 7 or claim 1 500.cdumm;
5 x > 5001.cms("15/M1", million(15/M1"))
5 x > 5001.cms("15/M1", million(15/M1"))
5 x > 5001.cms("15/M1", million(15/M1"))
6 x > 5001.cms("15/M1", million(15/M1"))
6 x > 5001.cms("15/M1", million(15/M1"))
7 x > 5001.cms("15/M1", million(15/M1"))
8 x * titlitto, powder 1 statisticon(0, y)
8 x * 15/M1" ("15/M1", "15/M1", "15/M
```

图 11: SD/HS 组分析

Wilcoxon 符号秩检验运行结果

```
MR96 "200/0" 延計組織程業: 統領值=1.0, p 值=0.0025
MRN '8/200' 延計社会保険・統令協+0.0, p 協+0.03125
MINERS '5/200' MANNINGER: MORES-1.0, p 88-0.0025
新味利 "29/100" 系数性経済業・統計値=1.0、p 値=0.0025
開味利 "29/100" 系数性経済業・統計値=1.0、p 値=0.0025
Mining '55/100' Ainthigam: (6) 2-1.0, p 2-0.0025
Minked 198/2001 ABSMMM Re (6)+ 10-1.0、p 10-0.0025
```

```
print(f"HM991 '(column)' EPHNAUM: NH-M-(statistic), p di-(p_value)")
#88N '100/0' 5696558#: GHS-0.0, p M-0.03125
新林列 15/1001 范的收拾结果: 经计值+3.0。 # 提+0.0525
BREST 10/1001 GENERALER: MHG-1.0, p G-0.0525
MRN 130/1001 SERVENIEN SCHIE-1.0. 8 M-0.0023
MEAN '50/100' SERVENIEN: (PHR-1.0. a M-0.002)
```

```
節除列 "200/0" 前別投放信用: 株計数-2.0, p 後-0.09375
#8850 "10/200" GONNAGER - MH E-2.0, p @+0.09375
#88N '20/200' SONESSIEM: MITE-2.0. 0 ($+0.0937)
搬车间 "32/202" 机砂板板模样: 核計畫=2.0, p 重=0.09375
#8956 "59/100" ESSEMENT: HHE-1.0, p $10.0025
```

图 12: GA/HN 组分 析

图 13: RH/HN 组分 图 14: RH/SM 组分 析

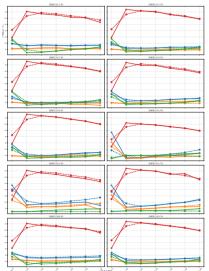
问题四: 差异显著的热解组合

差异显著的热解组合		共热解组合子组分析结果		
组合	物质类型	组合	关键混合比例	剔除后 p 值
CS/SM	HEX	CS/SM	0/100	0.03125
SD/SM	HEX	SD/SM	0/100	0.03125
SD/HS	HEX	SD/HS	0/100	0.03125
GA/HN	Water	GA/HN	0/100	0.03125
RH/HN	HEX	RH/HN	100/0	0.03125
RH/SM	HEX	RH/SM	0/100	0.03125

表 4: 差异显著的热解组合及子组分析结果

主要发现:在40组实验中,有6组存在显著差异,主要涉及HEX(正已烷)产物。关键混合比例集中在纯煤(0/100)和高比例生物质(100/0),表明在极端混合比例下,理论模型的预测准确性降低。非极性产物(HEX)的产率预测偏差较大,需进一步研究生物质与煤在极端比例下的相互作用机制,考虑改进理论模型,纳入更多影响因素。

不同生物质煤共热解与实验产物收率比较可视化





问题重述与分析 数据预处理与探索性分析 **模型的建立与求解** 参考文献 致谢与答辩

问题五: 热解产物产率预测模型 - 理论基础与方法选择

问题分析

生物质-煤共热解过程是一个复杂的热化学转化过程,其产物产率预测涉及多个挑战:

- 多因素影响:配比、样品质量、焦油量、水量等因素共同作用。
- 非线性关系:因素间可能存在复杂的非线性交互作用。
- 数据特征:实验数据可能存在噪声和不确定性。
- 预测精度要求: 需要在有限样本条件下实现高精度预测。

这些特点决定了我们需要选择能够适应复杂系统、捕捉非线性关系、并具有良好泛化能力的预测模型。

理论依据

基于问题特性,我们选择了三种不同类型的模型:

- 多元线性回归:适合捕捉线性关系,提供基础因素分析。
- 2 随机森林回归:非参数模型,适合处理非线性关系与复杂特征交互。
- ③ 高斯过程回归:非参数回归,在贝叶斯框架下优化选用,尝试量化不确定性并通过核函数捕捉复杂模式。

这种多模型策略允许我们从不同角度分析问题,比较模型性能,并获得更全面的预测结果。

答辩人: 陆子凯、狄诗琪

Team # 2024050727407, Problem B

问题五: 多元线性回归模型

模型原理

多元线性回归模型假设因变量(热解产物产率)与自变量(影响因素)之间存在线性关系:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

其中, y 为焦油产率, x_i 为影响因素, $β_i$ 为回归系数, ε 为随机误差项。模型估计采用最小二乘法, 最小化残差平方和:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2$$

通过求解正规方程 $(X^TX)\beta = X^Ty$, 我们可以得到参数估计 $\hat{\beta} = (X^TX)^{-1}X^Ty$ 。

- 4 ロ > 4 回 > 4 き > 4 き > - き - か Q G

致谢与答辩

问题五: 多元线性回归模型

参数估计结果与解释

拟合得到的模型为:

$$\hat{\mathbf{y}} = 0.0000 - 0.3108x_1 - 0.0431x_2 + 0.0194x_3 + 0.7672x_4$$

其中, x₁: 配比, x₂: 样品质量, x₃: 焦油量, x₄: 水量。 解释:

- 配比(x₁)和样品质量(x₂)对产率有负面影响,可能是由 于高配比和大样品量导致热解不充分。
- 焦油量(x₃)和水量(x₄)对产率有正面影响,特别是水量的影响最为显著(系数最大)。



问题五: 多元线性回归模型结果可视化

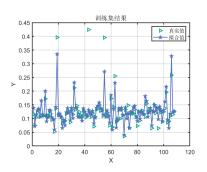


图 16: 训练集拟合结果

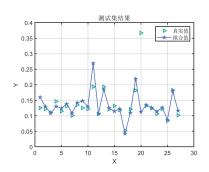


图 17: 测试集预测结果

从图中可以看出,模型在训练集上具有较好的拟合度,但在测试集上仍然存在一定的偏 差,表明线性模型可能无法完全捕捉系统中的非线性特征。

Team # 2024050727407, Problem B

随机森林是一种集成学习方法,通过构建多个决策树并结合它们的预测来进行回归:

- 基于 Bootstrap 采样构建多个决策树, 每棵树使用随机选择的特征子集。
- 决策树生长过程中,节点分裂基于最小化不纯度(如 Gini 指数或均方误差)。
- 最终预测诵过平均所有决策树的预测结果得到。

模型优势与局限性

优势:

- 随机森林能够有效处理非线性关系和特征交互。
- 不受多重共线性影响,适用于高维数据。
- 可以评估特征重要性, 具有良好的解释性。

局限性:

- 模型复杂度高、训练和预测时间可能较长。
- 难以捕捉超出训练数据范围的模式(外推能力有限)。
- 可能出现过拟合、尤其是当树的数量不足或深度过大时。



问题五:随机森林模型结果可视化

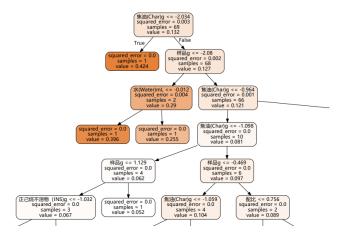


图 18: 随机森林中的一棵决策树部分可视化

4 D > 4 A > 4 B > 4 B >

问题五: 贝叶斯优化-高斯过程回归模型

模型原理

引言 问题重述与分析

高斯过程回归 (GPR) 是一种基于贝叶斯框架的非参数回归方法:

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

其中, m(x) 是均值函数, k(x,x') 是协方差函数 (即核函数), 通常采用 RBF 核:

$$k(x, x') = \sigma_f^2 \exp(-\frac{1}{2I^2}||x - x'||^2)$$

该模型能够量化预测的不确定性,并通过核函数捕捉复杂的非线性关系。

贝叶斯优化与超参数选择

- 贝叶斯优化可以帮助我们自动调整 GPR 的超参数 (如 σ_t^2 、I、 σ_n^2), 以提高模型 性能。
- 通过最大化边际似然(或最小化负对数边际似然)来选择最优的超参数组合。

问题五: 贝叶斯优化-高斯过程回归模型结果可视化

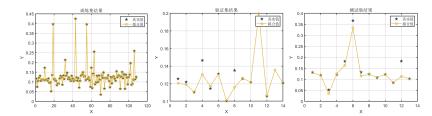


图 19: 高斯回归-贝叶斯优化模型结果

评估指标对比

模型	MAE	MSE	RMSE
线性回归	0.0215	0.0028	0.0527
随机森林	0.0201	0.0031	0.0561
高斯过程	0.0130	0.0005	0.0229

结果分析

- 高斯过程回归在所有指标上都优于其他两种方法,特别是在 MAE 和 RMSE 方面有显著改善。
- 随机森林在捕捉非线性特征方面表现不错,但在 MSE 和 RMSE 上的表现稍逊。
- 线性回归由于其线性假设,在捕捉复杂特征方面表现不佳。



答辩人: 陆子凯、狄诗琪

引言 问题重述与分析

未来工作方向

• 模型改进:考虑更多影响因素,设计多优化目标模型。

实际应用

工业实践:将模型应用于工业生产,验证其可行性和有效性。

参考文献 |

- [1] 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/i.issn.1671-5292.2018.02.001
- 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.
- 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, 2023, 56(05):193-202.
- Dehua Wang, Yang Zhang, and Yi Zhao. 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.
- 符学蔵. 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.
- 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, 2005, (12):3244-3251.
- 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, 2024. 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.

4 D > 4 A > 4 B > 4 B >

参考文献 ||

- [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, 2024. 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.

《四》《圖》《意》《意》

致谢与答辩

- 首先, 感谢指导老师对团队参加"数维杯"的支持。
- 特別感谢"数维杯"组委会提供的宝贵机会和展示平台。
- 最后, 感谢各位评委老师的聆听与指导。恭请评委老师提出 问题。

感谢您的聆听!

