Statistics and Application 统计学与应用, 2025, 14(6), 239-252
Published Online June 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.146164

Hans汉斯

基于机器学习的正交设计应用分析

普文琪

广西师范大学数学与统计学院，广西 桂林

收稿日期：2025年5月27日；录用日期：2025年6月20日；发布日期：2025年6月30日

摘　要

本文探讨了化学实验数据在优化工艺条件方面的应用。首先对实验数据进行了预处理。随后，应用多种机器学习算法进行建模，并对各模型的性能进行了比较。结果表明，随机森林模型有较高的准确性。同时，本文还引入了正交试验设计方法。该方法通过优化实验组合减少试验次数，降低实验成本。在正交设计的基础上，确定了最优的工艺条件：Co/SiO_2与HAP的装料比为200:200，乙醇浓度为0.3 ml/min，反应温度为400℃。此外，通过贡献率分析，研究发现温度(因子C)对C4烯烃收率的影响最大，其贡献率高达76.41%。这一发现为进一步优化实验条件提供了重要参考。结合机器学习算法与正交试验设计，不仅能够有效减少试验次数，还能显著降低实验成本，为工业化学实验的优化提供了一种科学高效的途径，推动了机器学习算法在工业领域的应用，助力数字化转型。

关键词

试验设计，C4烯烃生产，机器学习，正交试验

Application Analysis of Orthogonal Design Based on Machine Learning

Wenqi Pu

School of Mathematics and Statistics, Guangxi Normal University, Guilin Guangxi

Received: May 27th, 2025; accepted: Jun. 20th, 2025; published: Jun. 30th, 2025

Abstract

This paper explores the application of chemical experimental data in optimizing process conditions. The experimental data is first preprocessed. Subsequently, several machine learning algorithms are applied to model the data, and the performance of each model is compared. The results show that the random forest model has high accuracy. Additionally, this paper introduces the orthogonal experimental design method. This method optimizes experimental combinations, reducing the number

文章引用: 普文琪. 基于机器学习的正交设计应用分析[J]. 统计学与应用, 2025, 14(6): 239-252.
DOI: 10.12677/sa.2025.146164

of trials and lowering experimental costs. Based on orthogonal design, the optimal process conditions are determined: a Co/SiO_2 to HAP loading ratio of 200:200, an ethanol concentration of 0.3 ml/min, and a reaction temperature of 400˚C. Furthermore, through contribution rate analysis, it is found that temperature (factor C) has the greatest impact on the C4 olefin yield, with a contribution rate of 76.41%. This discovery provides an important reference for further optimizing experimental conditions. The combination of machine learning algorithms and orthogonal experimental design not only effectively reduces the number of trials but also significantly lowers experimental costs, offering a scientifically efficient approach to optimizing industrial chemical experiments. This contributes to the application of machine learning algorithms in the industrial sector and aids in digital transformation.

Keywords

Experimental Design, C4 Olefin Production, Machine Learning, Orthogonal Experiment

1. 引言

1.1. 背景意义

C4 烯烃广泛应用于化工产品、医药生产及其他高技术领域，作为重要的化工原料，其生产过程的优化至关重要。C4 烯烃的制备通常以乙醇为原料，通过催化偶合反应合成。催化剂的选择及反应条件(如温度、乙醇浓度等)对 C4 烯烃的选择性和收率具有显著影响。因此，如何优化催化剂配比和反应条件，以提高 C4 烯烃的收率和选择性，成为了化学工程领域的一个研究热点。

乙醇催化偶合反应涉及复杂的化学反应机制，其中催化剂的负载量、装料比以及反应温度等因素均可能对最终产品的产量和质量产生影响。优化这些因素，不仅能提高 C4 烯烃的产量，还能够有效节约原料和降低生产成本，具有较大的经济和环境意义。

然而，传统的实验优化方法往往受到实验设计和数据处理能力的限制，无法全面考虑多个因素的交互作用。近年来，机器学习方法被广泛应用于化学反应优化过程中，尤其是在多因素优化问题上展现出了强大的优势。通过建立科学的预测模型，结合正交设计等实验优化方法，可以系统地探索和优化反应工艺，为生产过程的高效运行提供理论依据和技术支持。

1.2. 国内外研究

乙醇脱水制备 C4 烯烃的反应效率和选择性受催化剂性质和反应条件的显著影响。研究者们通过建立数学模型，分析了催化剂组合、反应温度、乙醇浓度等因素对乙醇转化率和 C4 烯烃选择性的影响。例如，Tang 等人提出了多元非线性回归模型，揭示了催化剂组合和温度对 C4 烯烃产率的影响规律，并提出了优化策略[1]。

随着数据驱动方法的发展，机器学习被广泛应用于催化反应的优化。研究者利用机器学习模型，如支持向量机回归(SVMR)、随机森林(RF)、反向传播神经网络(BPNN)等，分析了催化剂组合、温度等因素对乙醇转化和 C4 烯烃选择性的影响。例如，张新龙将随机森林模型与正交设计结合得出了化工实验的最佳参数组合[2]。Li 等人结合回归分析和机器学习算法，构建了 SVMR 模型，并采用遗传算法优化模型参

数，实现了对 C4 烯烃产率的预测和优化[3]。

正交实验设计是一种有效的实验优化方法，能够系统地研究多个因素对反应结果的影响。研究者们通过正交实验设计，分析了催化剂组成、反应温度、催化剂用量等因素对乙醇转化和 C4 烯烃选择性的影响。例如，Wang 等人采用正交实验设计，研究了不同催化剂组成和反应条件对 C4 烯烃产率的影响，为工业化生产提供了理论依据[4]。

在实际应用中，催化反应往往涉及多个目标的优化，如乙醇转化率、C4 烯烃选择性和产率等。研究者们采用多目标优化方法，如遗传算法、粒子群优化(PSO)等，综合考虑多个目标，实现催化反应的优化。例如，Yu 等人基于 BP 神经网络和遗传算法，优化了乙醇脱水反应的工艺条件，提高了 C4 烯烃的产率和选择性[5]。

2. 算法理论及试验设计

2.1. 多元线性回归

多元线性回归模型是用于分析多个自变量($x_1, x_2, \cdots, x_p$)与因变量(y)之间关系的统计方法。其模型形式如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

其中， β_0 称为截距， $\beta_1, \cdots, \beta_p$ 回归系数， ε 是误差项。在多元线性回归中，假定 y 与 $x_1, x_2, \cdots, x_p$ 之间存在线性关系，即当所有其他因素保持不变时，因变量的变化可以通过每个自变量的线性变化来表达。回归系数 $\beta_1, \cdots, \beta_p$ 的值揭示了每个自变量对因变量的影响程度和方向[6]。

2.2. 多项式回归模型

多项式回归模型是回归分析的一种扩展形式，旨在捕捉因变量与自变量之间的非线性关系，其模型形式可以表示为：

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_p x_i^p + \varepsilon \tag{2}$$

其中 $x_i, x_i^2, \cdots, x_i^p$ 是自变量的不同次幂，表明因变量和自变量之间的关系是非线性的[7]。

2.3. 决策树模型

决策树模型是一种常用于回归问题的机器学习模型。它通过一系列的决策规则对数据进行预测。CART 回归树用于预测连续型目标变量。与分类树不同，回归树的每个叶节点包含的是一个连续的预测值，而非类别标签。在每个节点，CART 回归模型根据特征的某个阈值将数据划分为两个子集，以最小化子集内目标变量的方差或平方误差。模型通过选择最优的特征和切分点，在每一步，将数据集划分为两个部分，然后递归地对每个子集进行相同的划分，直到满足停止条件为止。CART 回归模型在许多实际应用中非常有效，特别是在预测问题中，能够提供清晰的决策路径和解释性，是机器学习中常用的回归模型之一[8]。

2.4. 随机森林模型

随机森林(Random Forest)是一种集成学习方法，基于多个决策树构建模型。该方法由 Leo Breiman 于 2001 年提出[9]。随机森林通过训练多个决策树，并结合各个树的回归预测结果来得到最终的预测值。每棵树都是通过对数据的随机抽样和特征随机选择来训练的，最终的回归结果由所有树的预测值的平均值决定。

随机森林回归模型通过“自助抽样”从原始数据集 $D=\{(x_i,y_i)\}_{i=1}^{n}$ 中抽取多个子数据集，每个子数据集被用来训练一棵决策树。训练完成后，每棵树会对新数据 x_{new} 进行预测，得到回归值 $\hat{y}_i(x_{new})$。随机森林回归的最终预测结果是所有决策树预测值的平均值：

$$\hat{y}_{RE}=\frac{1}{T}\sum_{i=1}^{T}\hat{y}_i(x_{new}) \tag{3}$$

其中，T 是森林中树的数量，$\hat{y}_i(x_{new})$ 是第 i 棵树对输入 x_{new} 的预测值。

2.5. 正交试验设计

正交试验设计是一种实验设计方法，用于高效地研究多个因素对实验结果的影响。它通过合理安排不同因素的水平组合，以最少的实验次数获得多方面的信息，从而优化实验条件，节省时间和成本。正交试验设计广泛应用于工程、化学、工业制造等领域，特别适合于多因素和多水平的实验。

正交试验设计通过使用正交表安排实验，使得每个因子在试验中能够充分展示其对响应变量的影响，同时降低因子之间可能存在的交互作用影响。正交表提供了一个具有平衡性的实验设计，减少了重复实验，从而在较少的试验次数下获取足够的信息。

正交设计的应用步骤，首先确定实验中要研究的因子及其每个因子的水平。根据因子的个数及其水平数选择合适的正交表。通过正交表进行因子与水平的组合，确保试验的平衡性。安排实验，并记录每个实验组合的响应变量。最后根据实验数据，利用极差分析、方差分析等方法对结果进行分析，确定最佳因子水平组合[10]。

3. 数据处理

3.1. 数据准备

C4 烯烃在化工和医药产品的生产中有着广泛的应用，而乙醇则是其生产的主要原料。在制备过程中，催化剂的组合(包括 Co 负载量、Co/SiO_2 与 HAP 装料比、乙醇浓度)以及温度，都会对 C4 烯烃的选择性和收率产生影响[11]。

因此，研究催化剂组合、乙醇催化偶合制备 C4 烯烃的工艺条件，具有重要意义。本文的数据来自 2021 年全国数学建模竞赛 B 题的附件 1，具有较强的代表性，并在未来的工业应用中具有不可替代的作用和价值。相关数据见表 1。

Table 1. Partial data results
表 1. 部分数据结果

催化剂组合编号	催化剂组合	温度	乙醇转化率(%)	乙烯选择性(%)	C4 烯烃选择性(%)	乙醛选择性(%)
A1	200 mg 1 wt%Co/SiO_2-200 mg HAP-乙醇浓度 1.68 ml/min	250	2.07	1.17	34.05	2.41
		275	5.85	1.63	37.43	1.42
		300	14.97	3.02	46.94	4.71
		325	19.68	7.97	49.7	14.69
		350	36.80	12.46	47.21	18.66
A2	200 mg 2 wt%Co/SiO_2-200 mg HAP-乙醇浓度 1.68 ml/min	250	4.60	0.61	18.07	0.94
		275	17.20	0.51	17.28	1.43
		300	38.92	0.85	19.6	2.21
		325	56.38	1.43	30.62	3.79
		350	67.88	2.76	39.10	4.20

3.2. 数据清洗

数据清洗是数据分析过程中的重要步骤，其主要目的是提高数据质量，以确保分析结果的准确性和可靠性。数据清洗涉及识别和修正数据中的错误、不一致性和不完整性。

数据清洗过程中常见的问题包括：缺失值、重复数据、异常值、不一致的数据格式、错误的数据录入和字符数据清洗。缺失值需被识别并处理，以避免影响分析结果；重复数据需被识别并去除，以保证数据的唯一性；异常值则是偏离正常范围的数据点，可能源自录入错误；不一致的数据格式，如日期格式或单位，需要统一以确保准确性；错误的数据录入包括拼写或数值错误，需进行修正；字符数据可能包含多余空格、大小写不一致或无效字符，这些也需清理。有效的数据清洗有助于提高数据质量和分析结果的可靠性。

将不合理和不相关的数据删除后，将催化剂组合的结果分为两列，其中 Co/SiO_2 和 HAP 装料比表示 Co/SiO_2 与 HAP 的质量比，C4 烯烃收率通过乙醇转化率与 C4 烯烃选择性的乘积计算得出[12]。最终整理后的数据形成了表 2，处理后的数据总共有 108 条。

Table 2. Data cleaning partial results
表 2. 数据清洗部分结果

编号	Co/SiO_2 和 HAP 装料比	乙醇浓度 (ml/min)	温度	乙醇转化率 (%)	C4 烯烃选择性(%)	C4 烯烃收率 (%)
1	200:200	1.68	250	2.07	34.05	0.70
2	200:200	1.68	275	5.85	37.43	2.19
3	200:200	1.68	300	14.97	46.94	7.03
4	200:200	1.68	325	19.68	49.70	9.78
5	200:200	1.68	350	36.80	47.21	17.37
6	200:200	1.68	250	4.60	18.07	0.83
7	200:200	1.68	275	17.20	17.28	2.97
8	200:200	1.68	300	38.92	19.6	7.63
9	200:200	1.68	325	56.38	30.62	17.26
10	200:200	1.68	350	67.88	39.10	26.54

3.3. 数据相关性分析

在某些情况下，我们需要研究数据集中某些属性与指定属性之间的相关性。为此，通常可以采用统计学方法进行分析，其中皮尔逊相关系数是最常用的一种方法。皮尔逊相关系数用于评估数据指标之间的线性关系，其值范围从−1 到 1。值的绝对值越大，表示两个指标之间的相关性越强；绝对值越小，则表示相关性较弱。该系数特别适用于分析连续型数据变量之间的线性关系。然而，它只能捕捉线性相关性，对于存在非线性(如曲线)关系的指标，皮尔逊相关系数可能无法准确反映其相关性[13]。计算公式为：

$$\rho = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}} \tag{4}$$

通过皮尔逊相关系数，我们可以分析特征与因变量之间的线性关系。表 3 是数据结果。

Table 3. Pearson correlation results
表 3. 皮尔逊相关性结果

	Co/SiO_2 和 HAP 装料比	乙醇浓度(ml/min)	温度
person 相关性	0.306	−0.193	0.725

根据特征与因变量的皮尔逊相关系数分析，特征温度与因变量之间的线性关系最为显著，显示出较强的线性相关性。特征乙醇浓度、Co/SiO_2 和 HAP 装料比与因变量的线性关系较弱。

3.4. 模型评价指标

本文采用以下指标来评估模型预测的精确度：拟合优度、均方误差(MSE)，以及均方根误差(RMSE)。这些评价指标常用于回归模型及其他预测模型的精度测量。

3.4.1. 拟合优度

拟合优度用于检验模型对数据的拟合情况。在多元线性回归模型中，拟合优度通常通过样本决定系数 R^2 来衡量，其定义如下：

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \tag{5}$$

其中，SSE 是回归残差平方和，SST 是总平方和。样本决定系数的取值范围在[0, 1]之间。当 R^2 值接近 1 时，说明模型对数据的拟合能力较强；当 R^2 值接近 0 时，说明模型对数据的拟合能力较差。

3.4.2. 均方误差(MSE)

均方误差用于衡量预测值与真实值之间的平均平方差，其计算公式为：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{6}$$

其中，y_i 是真实值，$\hat{y}_i$ 是预测值，n 是样本总数。

3.4.3. 均方根误差(RMSE)

均方根误差是均方误差的平方根，用于衡量预测值与真实值之间的平均差异，其计算公式为：

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{7}$$

通过这些指标，我们可以有效评估模型的预测精度及其对数据的拟合程度。

4. 机器学习

4.1. 多元线性回归

为方便建立多元线性回归和多项式回归等模型，本文将 C4 烯烃收率表示为 y，Co/SiO_2 和 HAP 装料比表示为 x_1，乙醇浓度(ml/min)为 x_2，温度为 x_3。

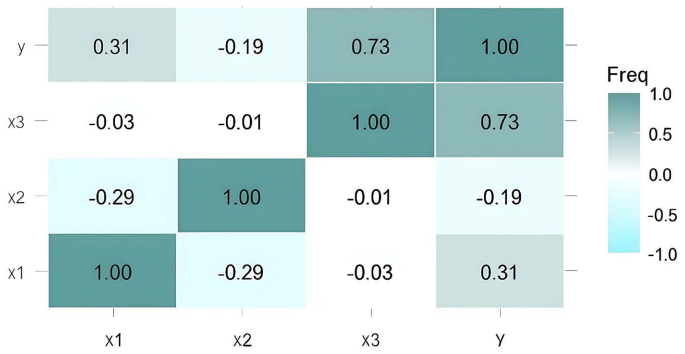

Figure 1. Heatmap
图 1. 热力图

从图 1 的结果中可以看到 x_1、x_2 和 y 的相关性较低，x_3 和 y 的相关性较高，因此温度和收率之间有一定的线性相关。

本文使用多元线性回归模型对数据集进行建模。直接应用多元线性回归后，得到回归系数表(表 4)。

Table 4. Regression coefficients table
表 4. 回归系数表

	截距	x_1	x_2	x_3
系数	−38.2007	1.1835	−1.6225	0.1273

由表 4 得到方程为：$y=-38.2007+1.1835x_1-1.6225x_2+0.1273x_3$。通过系数结果可以看到 x_1、x_3 的系数为正，即 y 与 x_1、x_3 成正相关，而 x_2 的系数为负，即 y 和 x_2 成负相关。适量提高 x_1、x_3，减低 x_2 即可提升 C4 烯烃收率。

4.2. 多项式回归

多项式回归是一种扩展线性回归的方法，用于捕捉自变量与因变量之间的非线性关系。通过代码运算得到二阶多项式回归的方程为：

$y=5.7683+28.3853x_1+5.0039x_1^2-8.3257x_2-4.6453x_2^2+67.1007x_3+25.0542x_3^2$。这个多项式回归模型通过引入因子的二次项，能够更好地捕捉因子与 C4 烯烃收率之间的非线性关系。

4.3. 决策树模型

决策树回归模型是一种基于树结构的预测模型，用于解决回归问题，即预测连续值输出。

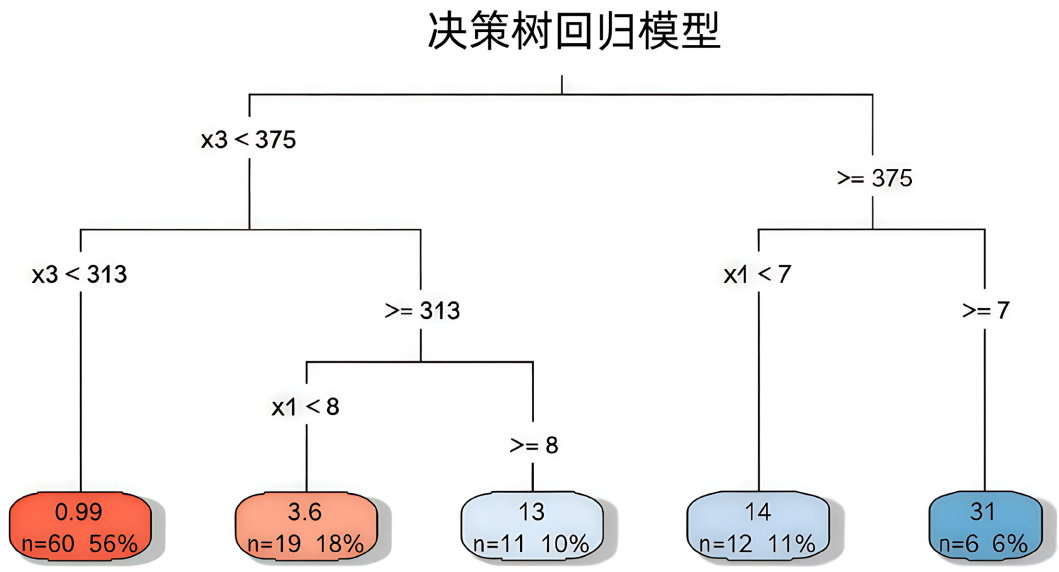

Figure 2. Decision tree regression model
图 2. 决策树回归模型

由图 2 可以看出来，模型表明特征 x_3 和 x_1 是影响目标变量的重要因素。不同的 x_3 和 x_1 值会显著影响预测结果，特别是当 x_3 增加到 375 或以上时，预测值明显增加。该模型为了解特征对目标变量的影响提供了直观的解释，有助于更好地理解和预测数据行为。

4.4. 随机森林模型

随机森林是一种强大的集成学习方法，通过构建多个决策树并集成其预测结果，能够显著提高模型的预测性能和稳健性。

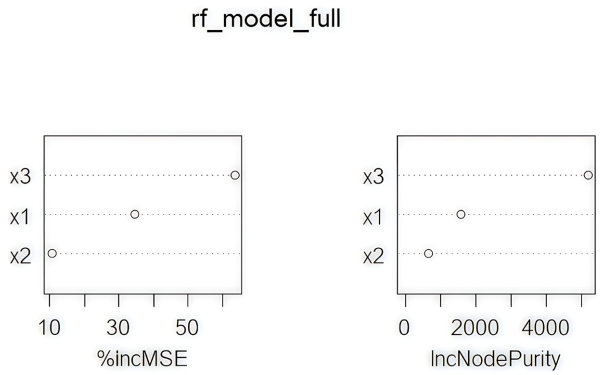

Figure 3. Random forest feature importance plot
图 3. 随机森林重要性变量图

%IncMSE (均方误差增加百分比)显示了每个特征对模型预测误差(均方误差)的影响。当某个特征被移除时，%IncMSE 值越大，说明该特征对模型的重要性越高。IncNodePurity (节点纯度增加值)显示了每个特征对节点纯度的贡献。节点纯度增加值越大，说明该特征在分裂节点时对减少数据的不纯度(如方差)贡献越大。由图 3 可以发现特征 x_3 是模型中最重要的特征，对预测准确性和节点纯度贡献最大。特征 x_1 次之，而特征 x_2 的重要性相对较低。在实际应用中，应重点关注和优化特征 x_3 和 x_1，以提高模型的预测性能。

4.5. 模型对比

通过上述四种模型的分析，根据实验预测结果，整理四种模型的评价指标如表 5 所示。

Table 5. Evaluation metrics for each model
表 5. 各个模型评价指标

	拟合优度	均方误差	均方根误差
多元线性回归	0.6426	27.5683	5.2505
多项式回归	0.7219	21.4558	4.6320
决策树	0.7922	16.0264	4.0033
随机森林	0.9187	6.2704	2.5040

由表 5 的结果可以看到随机森林模型的拟合优度最高，为 0.9187，表明其对数据的解释能力最强。随机森林模型的均方误差最低，为 6.2704，说明其预测误差最小。随机森林模型的均方根误差最低，为 2.5040，表明其预测精度最高。因此，我们选择随机森林模型进行建模，并结合正交设计实验，进一步优化工艺条件，以达到本文的最终目标。

5. 正交设计

5.1. 用正交表进行整体设计

在试验中考察三个三水平因子，该因子水平表如表 6 所示。

Table 6. Factor levels table
表 6. 因子水平表

因子	一水平	二水平	三水平
A：Co/SiO_2 和 HAP 装料比	75:75	100:100	200:200
B：乙醇浓度(ml/min)	0.3	0.9	1.68
C：温度(℃)	325	350	400

在有因子水平表后，需要进行表头设计。首先，根据试验中所考察的因子水平数，选择对应的正交表类型。然后，根据因子的数量，具体选定一张适合的表。在本例中选用 $L_9(3^4)$是合适的。确定正交表后，将因子分配到表格的列上，这一过程被称为表头设计。在忽略交互作用的情况下，因子可以自由分配到任意列，每个因子占一列[14]。用 $L_9(3^4)$安排试验，本文试验的表头设计如表 7 所示。

Table 7. Header design
表 7. 表头设计

表头设计	A	B	C	
列号	1	2	3	4

表 7 得到后，可以根据相应因子的水平写出试验计划。只需将正交表中各列的数字替换为因子的具体水平，未涉及因子的列则忽略不计。本文的试验计划如下：将第一列的 1、2、3 分别替换为 Co/SiO_2 和 HAP 装料比的三个水平 75:75、100:100、200:200；将第二列的 1、2、3 分别替换为乙醇浓度的三个水平 0.3、0.9、1.68；将第三列的 1、2、3 分别替换为温度的三个水平 325℃、350℃、400℃。这样可以得到具体的试验计划(见表 8)。例如，第一号试验的因子水平组合为：Co/SiO_2 和 HAP 装料比为 75:75，乙醇浓度为 0.3 ml/min，温度为 325℃，其他试验的因子组合则依此类推。

Table 8. Experimental design table
表 8. 试验计划表

Co/SiO_2 和 HAP 装料比	乙醇浓度	温度(℃)
(1) 75:75	(1) 0.3	(1) 325
(1) 75:75	(2) 0.9	(2) 350
(1) 75:75	(3) 1.68	(3) 400
(2) 100:100	(1) 0.3	(2) 350
(2) 100:100	(2) 0.9	(3) 400
(2) 100:100	(3) 1.68	(1) 325
(3) 300:300	(1) 0.3	(3) 400
(3) 300:300	(2) 0.9	(1) 325
(3) 300:300	(3) 1.68	(2) 350

5.2. 利用正交设计和随机森林优化工艺

使用随机森林模型进行对 C4 烯烃收率(%)进行预测，其中正交设计试验计划表的值为模型的输入变量，整理后得到直观分析计算表(表 9)。

Table 9. Visual analysis table
表 9. 直观分析表

表头设计	A	B	C	D	y
	1	2	3	4	
1	1	1	1	1	4.84
2	1	2	2	2	7.94
3	1	3	3	3	16.48
4	2	1	2	3	8.42
5	2	2	3	1	25.32
6	2	3	1	2	3.18
7	3	1	3	2	30.70
8	3	2	1	3	9.95
9	3	3	2	1	15.64
T1	29.28	43.97	17.98	45.81	
T2	36.93	43.22	32.02	41.84	
T3	56.29	35.31	72.51	34.86	
t1	9.76	14.66	5.99	15.27	
t2	12.31	14.41	10.67	13.95	
t3	18.76	11.77	24.17	11.62	
R	9.01	2.88	18.17	3.65	

5.3. 数据分析

5.3.1. 综合可比性

首先，我们需要确定各因子取何种水平组合最优。根据本文的试验结果，第 7 号试验的结果为 30.70，最高，因此可以认为其对应的水平组合 $A_3B_1C_3$ 是最佳的。然而，这一结论仅基于 9 个试验结果，是否适用于所有 27 个试验还需要进一步分析。为此，我们可以利用正交表的综合可比性进行深入分析。

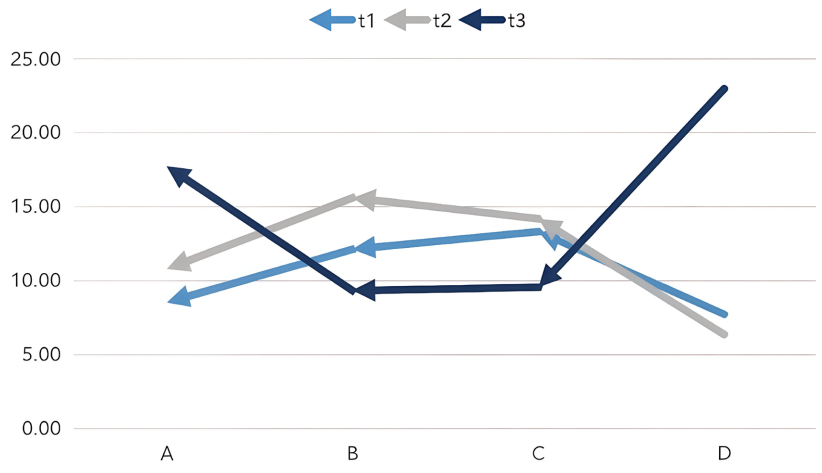

Figure 4. Mean and maximum values
图 4. 平均值最大值

由图 4 知，因子 A 取三水平，因子 B 取一水平，因子 C 取三水平。综上可知使指标达到最大的水平组合是 $A_3B_1C_3$，即当 Co/SiO_2 和 HAP 装料比为 200:200，乙醇浓度为 0.3 ml/min，温度为 400℃时，C4 烯烃的收率能够达到最大值。

5.3.2. 用极差分析各因子对指标的影响程度大小

各因子对指标影响程度的大小可以通过计算每个因子的极差来确定。极差是指最大值与最小值的差。如果极差较大，意味着调整该因子的水平会显著影响指标，因此该因子对指标的影响较大；反之，若极差较小，表示影响较小。反之，若极差较小，则表示该因子对指标的影响较小[15]。

在本文中各因子的极差分别为：$R_A = 18.76 - 9.76 = 9.01$，$R_B = 2.88$，$R_C = 18.17$。它们在表 9 的最后一行。从三个因子的极差分析来看，因子 C 对指标的影响最大，然后是因子 A，影响最小的是因子 B，即为 $R_C > R_A > R_B$。

5.3.3. 水平均值图

为了更直观地展示，可以将每个因子不同水平的均值绘制出来。本文的水平均值图如图 5~7 所示。从图中可以清楚地看出，最佳水平为 $A_3B_1C_3$，同时也能明显观察到各因子水平之间的差异。

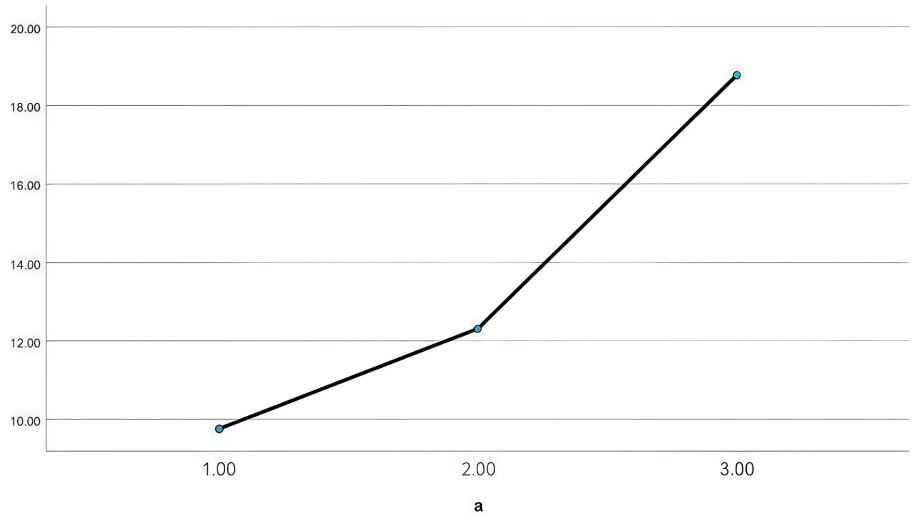

Figure 5. Mean values of A factor at different levels
图 5. A 因子的水平均值图

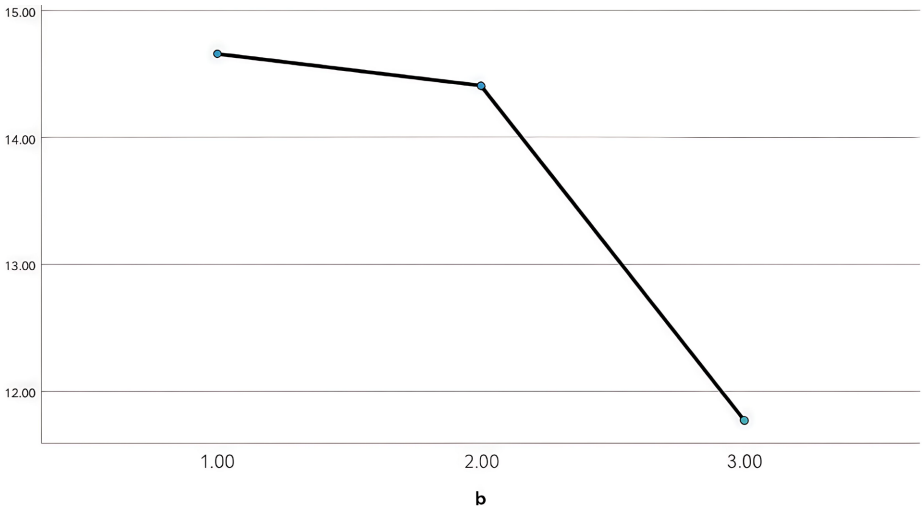

Figure 6. Mean values of B factor at different levels
图 6. B 因子的水平均值图

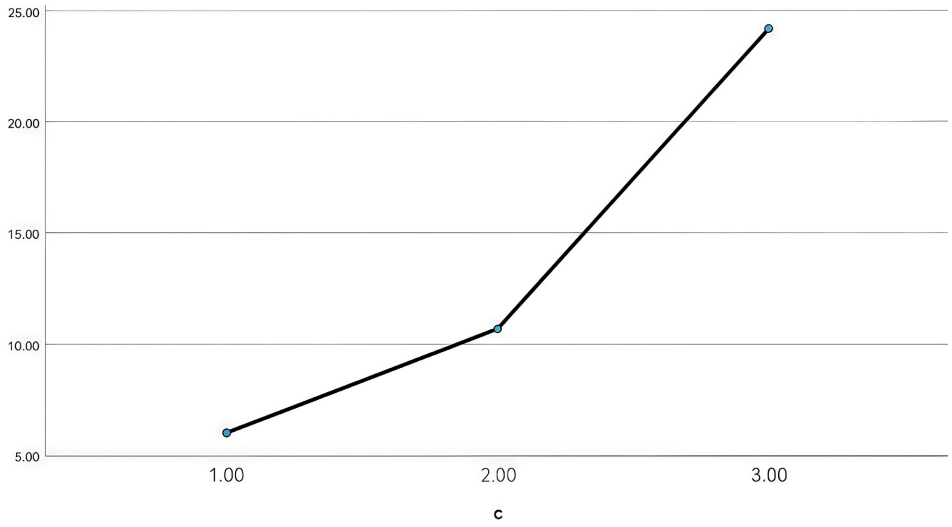

Figure 7. Mean values of C factor at different levels
图 7. C 因子的水平均值图

综上所述，最佳或满意的水平组合为 $A_3B_1C_3$，因为该方法是通过比较 27 个可能的水平组合得出的。因子的主次排序为因子 C、因子 A、因子 B，这也说明试验误差较小。

5.3.4. 数据的方差分析及最佳水平组合的估计

方差分析是一种统计方法，用于检测不同因素或因素水平对响应变量的影响是否显著。通过方差分析，可以确定哪些因素对实验结果有显著影响，从而为优化实验条件提供依据[16]。对于显著因子，应选择其最佳水平，因为水平变化会影响指标。对于不显著因子，可以自由选择水平，实操中会根据成本、操作便利等因素来决定水平[17]。

通过使用 SPSS 软件进行方差分析，结果如表 10 所示。$F_{0.9}(2,2)=9$，$F_{0.95}(2,2)=19$。从表中结果可以看出，因子 C 的 F_C 为 26.097，明显大于 19，因子 C 在显著性水平 0.05 下是显著的，而因子 A 和 B 不显著。由于因子 C 显著，因此应选择最佳水平 C_3，而因子 A 和 B 则可以选择任意水平。

Table 10. Analysis of variance (ANOVA) table
表 10. 方差分析表

源	III 类平方和	自由度	均方	F	显著性
修正模型	678.957[a]	6	113.159	11.053	0.085
截距	1667.443	1	1667.443	162.869	0.006
a	129.265	2	64.633	6.313	0.137
b	15.329	2	7.664	0.749	0.572
c	534.363	2	267.182	26.097	0.037
误差	20.476	2	10.238		
总计	2366.876	9			
修正后总计	699.433	8			

通过反差分析得到最佳水平组合为 ABC_3，点估计、区间估计都叙述该组合下指标均值 μ_3 的估计值，首先进行点估计：$\hat{\mu}=\bar{y}=13.61$，$\hat{c}_3=t_{33}-\bar{y}=24.17-13.61=10.56$，$\widehat{\mu_{..3}}=\hat{\mu}+\hat{c}_3=13.61+10.56=24.17$，从而 C_3 水平组合下指标的无偏估计为 $\widehat{\mu_{..3}}=24.17$。

接下来是区间估计：

$$\widehat{\mu_{\cdot\cdot 3}} = \hat{\mu} + \hat{c}_3 = t_{33} = \frac{1}{3}\left(y_3 + y_5 + y_7\right), var\left(\widehat{\mu_{\cdot\cdot 3}}\right) = \frac{1}{3}\sigma^2 = \frac{1}{n_e}\sigma^2, n_e = 3,$$

$$S_A = \frac{1}{3}\left(T_{11}^2 + T_{21}^2 + T_{31}^2\right) - \frac{1}{9}T^2 = 129.27, S_B = 15.33, S_C = 534.36, S_e = 20.48,$$

$$S_e' = S_e + S_A + S_B = 165.07,\ f_e' = f_e + f_A + f_B = 6, \hat{\sigma} = \sqrt{\frac{165.07}{6}} = 5.24, t_{0.975}(6) = 2.447.$$

所以 $\widehat{\mu_{\cdot\cdot 3}}$ 的 0.95 置信区间为： $24.17 \pm 5.24/\sqrt{3} = 24.17 \pm 3.02 = (21.15, 27.19)$。

5.3.5. 贡献率分析

当试验指标不符合正态分布时，方差分析的依据就不充分。在这种情况下，可以通过比较各因子的“贡献率”来评估因子的影响程度。

由于 S_{Foctor} 中除了因子的效应外，还包含误差，因此 $S_{Foctor} - f_{Foctor} \cdot MS_e$ 被称为因子的纯平方和。将因子的纯平方和与总平方和 S_T 的比值称为因子的贡献率[15]。例如，对于因子 A 来说，记其贡献率为 ρ_A，那么 $\rho_A = S_A / S_T = 129.26/699.43 = 18.48\%$。

类似可计算 ρ_B 与 ρ_C，以及纯误差平方和： $\rho_B = 2.19\%$， $\rho_C = 76.41\%$， $\rho_e = 2.93\%$。 $f_T * MS_e = S_e + f_A \cdot MS_A + f_B \cdot MS_B + f_C \cdot MS_C = 699.43$。

综合上述计算结果得到贡献率表格如表 11 所示。

Table 11. Contribution rate analysis table
表 11. 贡献率分析表

来源	平方和 S	自由度 f	纯平方和	贡献率(%)
因子 A	129.27	2	129.27	18.48
因子 B	15.33	2	15.33	2.19
因子 C	534.36	2	534.36	76.41
误差	20.48	2	20.48	2.93
T	699.43	8	699.43	

因子 C 的贡献率为 76.41%。这表明因子 C 在解释数据的变异性方面具有显著作用，是影响结果的主要因素。因子 A 的贡献率为 18.48%。因子 A 对数据变异也有一定的解释作用，但不如因子 C 显著。因子 B 的贡献率为 2.19%。因子 B 对数据变异的解释作用很小，几乎可以忽略不计，综上所述，主要应关注因子 C，其次是因子 A，而因子 B 的影响较小，误差也在可接受范围内。通过控制和优化因子 C，可以显著影响和改善结果。

6. 结论

本文通过多种机器学习算法和正交试验设计，对 C4 烯烃生产实验数据进行了深入分析。结果显示：

多种算法对比：随机森林算法在拟合优度、均方误差和均方根误差方面均优于多元线性回归、多项式回归和决策树模型。随机森林模型的拟合优度最高，均方误差和均方根误差最低，说明其预测性能最优。

正交试验设计：正交试验设计有效减少了试验次数，降低了成本。通过正交设计，本文确定了 C4 烯烃生产的最优工艺条件：Co/SiO_2 和 HAP 装料比为 200:200，乙醇浓度为 0.3 ml/min，温度为 400℃。

贡献率分析：因子 C (温度)对实验结果的贡献率最高，达到 76.41%，是影响实验结果的主要因素；

因子 A (Co/SiO_2 和 HAP 装料比)次之，贡献率为 18.48%；因子 B (乙醇浓度)影响最小，贡献率为 2.19%。

参考文献

[1] Tang, P., Li, H., Zhang, X., *et al*. (2023) A Mathematical Model of Catalyst Combination Design and Temperature Control in the Preparation of C_4 Olefins through Ethanol Coupling. *RSC Advances*, **13**, 10703-10714.

[2] 张新龙. 基于随机森林与正交设计的一类化工实验数据的统计分析[D]: [硕士学位论文]. 大连: 大连理工大学, 2022.

[3] Li, L., Hong, R., Yang, C. and Tao, J. (2022) Study on Influencing Factors of Ethanol Coupling to Prepare C_4 Olefins. *Highlights in Science, Engineering and Technology*, **17**, 176-184. https://doi.org/10.54097/hset.v17i.2595

[4] Wang, Q., Zhang, Y. and Liu, J. (2023) Process Schemes of Ethanol Coupling to C_4 Olefins Based on a Genetic Algorithm for Back Propagation Neural Network Optimization. *Heliyon*, **9**, e03589.

[5] Yu, J., Sun, W. and Qiu, P. (2023) Process Optimization of C_4 Olefins by Ethanol Coupling Based on BP Neural Network and Genetic Algorithm. *Highlights in Science, Engineering and Technology*, **41**, 258-264. https://doi.org/10.54097/hset.v41i.6825

[6] 刘严. 多元线性回归的数学模型[J]. 沈阳工程学院学报(自然科学版), 2005, 1(2): 128-129.

[7] 曹瑞, 周锋, 欧阳广帅, 等. 基于多项式回归的房价模型分析[J]. 科协论坛(下半月), 2010(10): 137-138.

[8] 陈峰. 基于 CART 算法的空气质量指数回归预测模型的学习[J]. 上饶师范学院学报, 2016, 36(6): 16-21.

[9] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[10] 王健. 基于正交试验的桩锚支护结构多参数优化设计[J]. 山西建筑, 2020, 46(21): 52-53+63.

[11] 崔恒建, 王冠鹏, 郑志伟. 乙醇偶合制备 C4 烯烃的统计分析与建模[J]. 数学建模及其应用, 2022, 11(1): 54-63.

[12] 王润墨, 沈桓羽, 陈静, 张芳芳, 孙凯, 韩永奇. 乙醇制备 C_(4)烯烃反应中催化条件的数理分析[J]. 齐鲁工业大学学报, 2022, 36(3): 73-80.

[13] 柯文俊, 王泊涵, 杜泽峰, 等. 一种无监督的软件复杂度度量与评估模型[J]. 高技术通讯, 2020, 30(4): 333-341.

[14] 余维新. 基于正交试验设计的建筑钢材质量优化研究[D]: [硕士学位论文]. 南京: 南京理工大学, 2011.

[15] 韩磊. 利用正交试验设计探究影响产品质量指标的因素[J]. 中国检验检测, 2021, 29(5): 25-28.

[16] 王波. 方差分析在企业人力结构化面试效果中的应用[J]. 中国商贸, 2014(18): 84-87.

[17] 陈敏. SPSS 在正交设计中的应用[J]. 科技风, 2020(26): 27-28.