

乙醇耦合制备 C4 烯烃效率的研究

摘要

C4 烯烃在工业生产和医疗领域有着广泛的应用，在制备 C4 烯烃的过程中，温度以及催化剂组合对 C4 烯烃的选择性和 C4 烯烃收率有着重要的影响，研究最为有效的催化剂组合对生产有着极为重要的意义。本文通过对数据进行定性分析，并且建立回归与预测模型，给出最优的制备条件。

针对问题一，需要研究温度、C4 烯烃的选择性与乙醇转化率的相关关系。对于每种催化剂组合，我们首先对三者分别进行夏皮洛-威尔克检验 (Shapiro-wilk)，得出数据满足正态分布的条件，进而可以使用皮尔逊相关系数对三者的相关性进行分析，求出相关系数矩阵，并对相关系数的显著性进行假设检验，得出乙醇转化率、温度和 C4 烯烃的选择性两两之间有显著的相关性。在 350℃、给定的催化剂组合的条件下，我们对时间与乙醇转化率和时间与 C4 烯烃选择性分别进行曲线拟合，根据 SSE 与 R-square 选取最优的拟合曲线，并结合拟合结果进行了定性分析。

针对问题二，我们首先通过 Python 中的 pandas 数据分析包对附件一中的数据进行处理，提取了温度、Co 负载量、乙醇浓度、HAP 质量、Co/SiO₂ 和 HAP 装料比这五个定量变量，以及装料方式这个定性变量，并且使用 Stata 软件来对其进行描述性统计。接着，对乙醇转化率与 C4 烯烃选择性分别进行 OLS 标准化回归，并对回归结果进行显著性 *t* 检验，得到各因素对因变量影响的显著程度。最后，我们对扰动项的异方差以及多重共线性进行检验，发现存在异方差，根据统计学领域的专家 Stock and Watson (2011) 推荐，我们使用 OLS 与稳健的标准误对模型进行修正。

针对问题三，在问题二拟合结果的基础上，我们发现拟合优度(R-square)并不理想，这对于预测结果来说会产生较大的误差，为了得到更准确的预测结果，我们使用 BP 神经网络。根据回归中自变量的显著性检验，我们选取了对结果产生较大影响的五个自变量作为输入，加入隐藏层后，对 C4 烯烃收率进行输出，整体的拟合优度为 0.986，预测结果非常理想。我们给定自变量的范围，使用网格搜索法，得到大量的输出结果，对数据结合实践进行筛选后，给出了当温度为 400℃，Co 负载量为 1wt%，乙醇浓度 0.9 ml/min，Co/SiO₂ 和 HAP 装料比为 1: 1，HAP 质量为 240mg 时，C4 烯烃收率为 45.42% 的结论。在温度小于 350℃ 的限定下，同理，我们给出了最优解的实验条件。

针对问题四，我们在 BP 神经网络模型预测结果的基础上，结合分位数的概念，为了得到尽可能高的 C4 烯烃收率，而且考虑到预测偏差、大数据量的影响，我们将 C4 烯烃收率大于 35% 的数据分为 5 组，每一组计算出上四分位数，以及距离其上四分位数最近的烯烃收率，该烯烃收率可描述该范围内烯烃收率的大体趋势。增加这五组实验，可以检验模型的预测准确性，且 C4 烯烃收率也大幅度提高。

关键词：皮尔逊相关系数、OLS 与稳健的标准误、假设检验、BP 神经网络

一、问题重述

1.1 问题背景

随着人类活动导致化石原料日益匮乏，人们发现通过乙醇制备的 C4 烯烃等产品，为医药、化工领域生产寻找合适的能源补足和替代提供了选择^[1]。一种通过大量科学实验得到的 Co/SiO₂-HAP 催化剂组合参与乙醇偶合的一系列化学变化后效果显著。它与 Co 负载量、Co/SiO₂ 与 HAP 装料比、乙醇浓度有关，并同温度一起影响 C4 烯烃的选择性和 C4 烯烃收率。因此探究这种催化的工艺尤为关键。

1.2 目标任务

C4 烯烃的主要原料是乙醇，其制备过程受温度与催化剂组合的影响。在一系列实验的基础上，探寻这种催化工艺的最佳条件。

问题一需要对于附件一中的每种催化剂组合，研究温度对乙醇转化率、C4 烯烃选择性的影响，并且在温度为 350 度且给定的催化剂组合的实验条件下，分析不同时间的测试结果。

问题二需要我们研究不同温度、不同催化剂组合对乙醇转化率以及 C4 烯烃选择性的影响。

问题三需要我们探索在相同实验条件下，使得 C4 烯烃收率尽可能高的温度与催化剂组合。在温度低于 350 度的情况下，给出合适的催化剂组合与温度，得到尽可能高的 C4 烯烃收率。

问题四要我们增加并且设计五次重要的实验，探索对于制备 C4 烯烃的工艺条件具有重要意义实验。

二、问题分析

2.1 对问题一的分析

第一问要解决两个问题。首先是根据附件一，探究不同催化剂组合环境下温度与乙醇转化率及 C4 烯烃选择性的相关性及其具体影响。对每组催化剂组合分别考虑，由于样本量较少，本文使用夏皮洛-威尔克检验的 p 值，若经过显著性检验，可采用皮尔逊相关系数来判断上述变量的关系；反之则采用斯皮尔曼系数。考察显著性后，可以对问题以定性表述。

其次要对附件二的数据进行处理，要得到制备反应中不同时间乙醇转化率及 C4 烯烃选择性的规律，需要对数据进行曲线拟合，以观察变量之间的函数关系，方便定性描述特征。

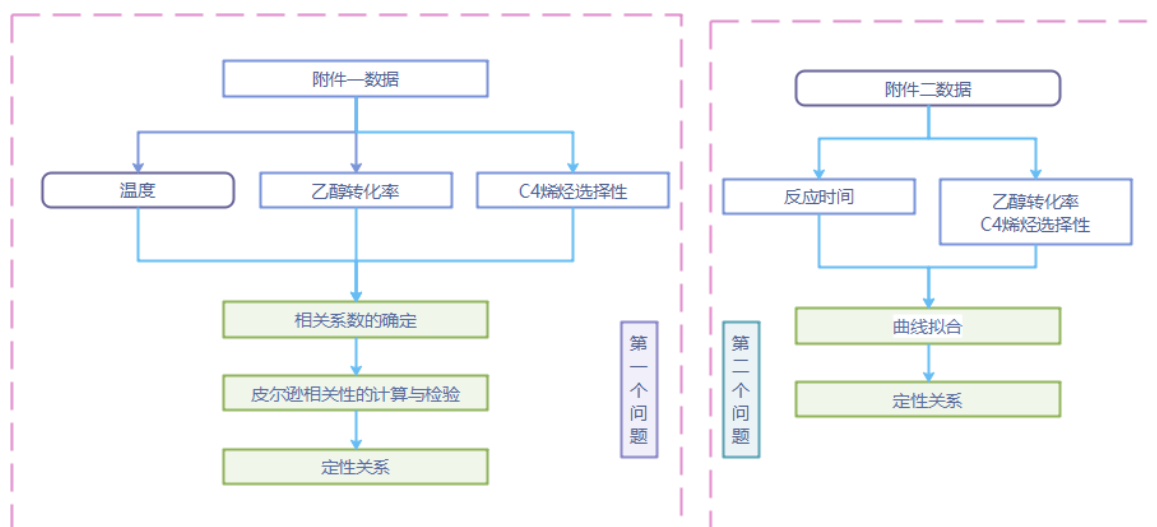


图 1 问题一分析框图

2.2 对问题二的分析

针对第二问，为研究温度与催化剂组合对两个求解量如何影响，首先对附件一的第一列催化剂组合数据进行处理。将不同的催化剂组合拆分为具体的 Co 负载量，Co/SiO₂ 质量和 HAP 质量以及两者的比值，乙醇浓度。考虑温度的协同影响，选择合适的自变量分别对乙醇转化率和 C4 烯烃选择性进行 OLS 标准化回归。检验各自变量的显著性，并对两个回归的异方差与多重共线性问题分别修正，得出更为准确的结论。

2.3 对问题三的分析

针对第三问，由烯烃收率的定义可知，通过回归结果可以得到烯烃收率最大的目标函数，但观察函数发现在约束边界处取得极值、且回归的拟合优度较小，因此用 BP 神经网络对样本数据集进行训练和验证划分，用网格搜索法求得最大的烯烃收率。

2.4 对问题四的分析

第四问需要给出五组实验设计，为检验乙醇偶合制备 C4 烯烃催化条件模型的合理性，在第三问 BP 神经网络模型求解的基础上，利用四分位数将解区间划分为五种不同组合，以对模型的合理性进行检验。

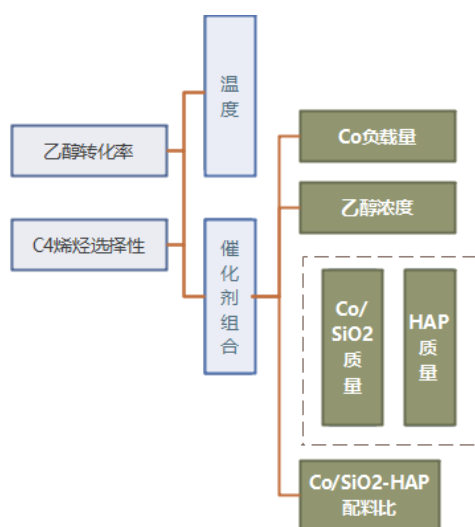


图 2 问题二分析框图

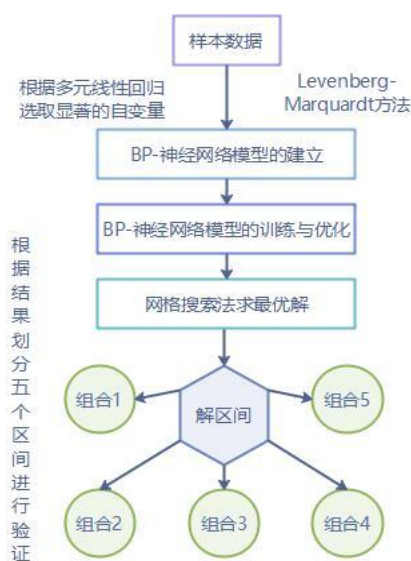


图 3 问题三、四分析框图

三、模型假设

- 1.该催化反应具有较高的灵敏度和连续性。
- 2.题目中所给的数据均是可重复得出的，不考虑偶然因素对实验的影响。
- 3.控制好适宜的催化反应条件，可以达到与附件所给数据匹配的环境精度。
- 4.只考虑温度、催化剂条件等题目所给条件对 C4 烯烃收率的影响。

四、符号说明

符号	说明	单位
p	显著性水平	-
t	t 检验值	-
SSE	和方差	-
$R-square$	判定系数	-
$Coef.$	解释变量系数	-
β	回归系数	-
$_cons$	回归方程中的截距项	-
VIF	方差膨胀因子	-

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 相关系数的确定

相关分析是以相关关系为对象，研究两个或两个以上随机变量之间线性依存关系的紧密程度。通常用相关系数表示，多元相关时用复相关系数表示^[2]。度量两个变量间的线性关系时，通常采用皮尔逊（Pearson）相关系数或斯皮尔曼（Spearman）相关系数进行分析。但是，只有当数据满足连续、正态分布且符合线性关系时，才可以使用皮尔逊（Pearson）相关系数来描述变量间的相关性；而斯皮尔曼（Spearman）相关系数却没有使用条件限制。

对于每种催化剂组合，样本量在 n 位于 5~8 的范围内。因此，首先使用小样本的夏皮洛-威尔克检验（Shapiro-wilk）， H_0 :该随机变量服从正态分布， H_1 :该随机变量不服从正态分布。将 p 值与 0.01 比较，小于 0.01 时则可以在 99%的置信水平下拒绝原假设，否则不能拒绝原假设。

每种催化剂组合的样本是否满足正态分布条件。我们使用 SPSS 软件对每组进行了检验，我们选取四组重点进行说明。首先，可视化它们之间的相关关系。下图为催化剂组合 A1,A2,B1,B2 的乙醇转化率、C4 烯烃的选择性与温度的散点图。

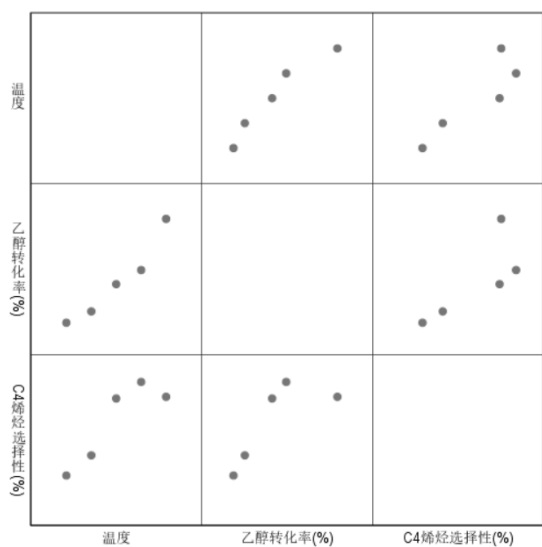


图 4 催化剂组合 A1 的散点图

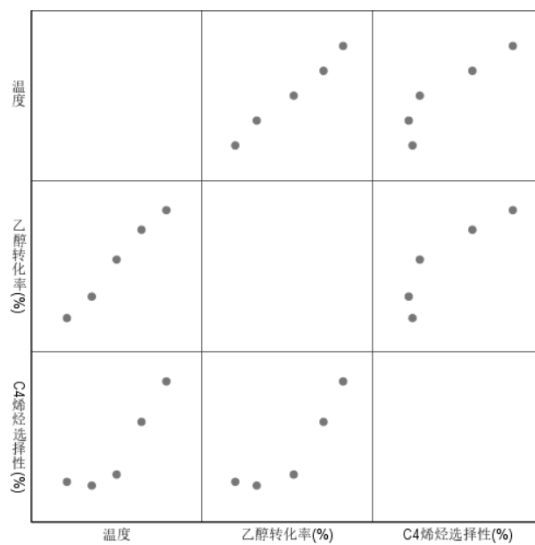


图 5 催化剂组合 A2 的散点图

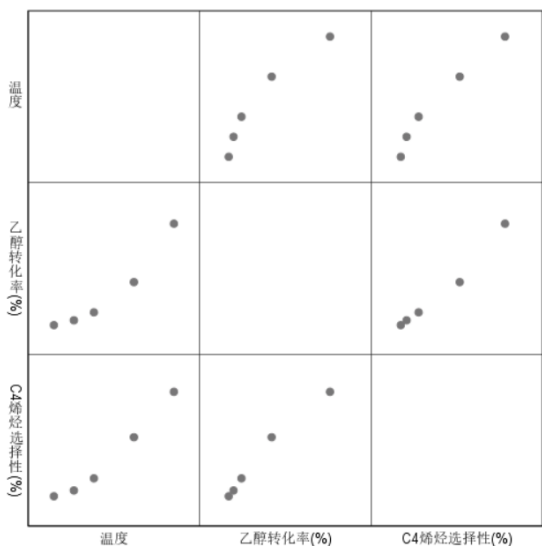


图 6 催化剂组合 B1 的散点图

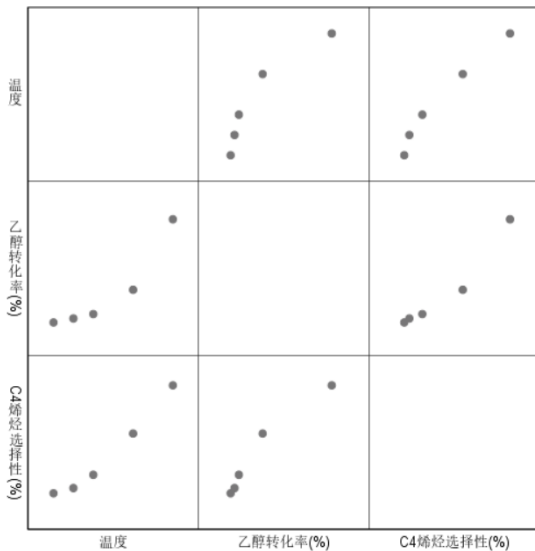


图 7 催化剂组合 B2 的散点图

对于每种催化剂组合，进行夏皮洛-威尔克检验（Shapiro-wilk），其中催化剂组合 A1,A2,B1,B2 的检验结果如下表所示：

表 1 夏皮洛-威尔克正态分布显著性检验表

催化剂组合	指标	自由度	显著性 p
A1	温度	5	0.967
	乙醇转化率	5	0.783
	C4 烯烃的选择性	5	0.157
A2	温度	5	0.967
	乙醇转化率	5	0.652
	C4 烯烃的选择性	5	0.231

B1	温度	5	0.787
	乙醇转化率	5	0.045
	C4 烯烃的选择性	5	0.275
B2	温度	5	0.787
	乙醇转化率	5	0.140
	C4 烯烃的选择性	5	0.285

由上表的显著性水平均大于 0.01, 故在 99%的置信水平下, 接受原假设, 该样本服从正态分布, 可以使用皮尔逊 (Pearson) 相关系数。

5.1.2 皮尔逊相关性的计算与检验

在 5.1.1 中得样本服从正态分布, 对此, 我们计算出每种催化剂组合中乙醇转化率、C4 烯烃的选择性与温度的相关系数 (结果见附录 2), 对其显著性进行检验, 皮尔逊相关系数 r , 检验它是否显著地异于 0, 设定原假设: $H_0: r = 0$, 备择假设 $H_1: r \neq 0$ 。构造统计量 $t = r\sqrt{\frac{n-2}{1-r^2}}$, 可以证明 t 是服从自由度为 $n-2$ 的 t 分布^[3]。

画出该分布的概率密度函数 pdf , 并且计算出相应的 p 值, 若 $p < 0.05$, 说明在 95% 的置信水平上拒绝原假设。

画出该分布的概率密度函数 pdf , 并且计算出相应的 p 值, 若 $p < 0.05$, 说明在 95% 的置信水平上拒绝原假设。

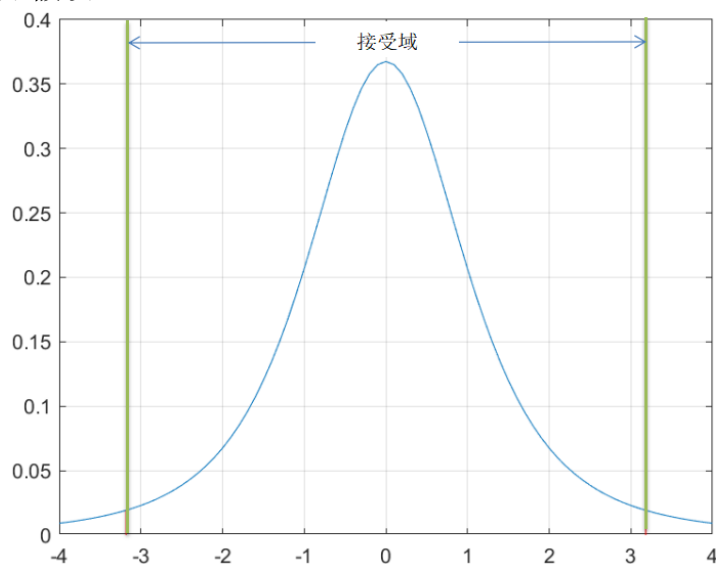


图 8 t 分布表

其中接收检验的 A1,A2,B1,B2 中样本量 $n=5$, 下面分别给出 A1,A2,B1,B2 四种催化剂组合的相关系数矩阵并且对相关系数的显著性进行检验。

A1 的相关系数矩阵

$$\begin{pmatrix} 1.0000 & 0.9655 & 0.8871 \\ 0.9655 & 1.0000 & 0.7712 \\ 0.8871 & 0.7712 & 1.0000 \end{pmatrix}$$

A2 的相关系数矩阵

$$\begin{pmatrix} 1.0000 & 0.9950 & 0.9143 \\ 0.9950 & 1.0000 & 0.9029 \\ 0.9143 & 0.9029 & 1.0000 \end{pmatrix}$$

B1 的相关系数矩阵

$$\begin{pmatrix} 1.0000 & 0.9621 & 0.9858 \\ 0.9621 & 1.0000 & 0.9898 \\ 0.9858 & 0.9898 & 1.0000 \end{pmatrix}$$

B2 的相关系数矩阵

$$\begin{pmatrix} 1.0000 & 0.9293 & 0.9848 \\ 0.9293 & 1.0000 & 0.9704 \\ 0.9848 & 0.9704 & 1.0000 \end{pmatrix}$$

表 2 A1,A2,B1,B2 组合 p 检验值表

组合	温度与乙醇转化率 p 检验值	温度与 C4 烯烃的 选择性 p 检验值	乙醇转化率与 C4 烯 烃的选择性 p 检验值
A1	0.0077	0.0448	0.1268
A2	0.0004	0.0297	0.0358
B1	0.0088	0.0020	0.0012
B2	0.0223	0.0022	0.0061

由 p 值检验的结果可知, 除 A1 的乙醇转化率与 C4 烯烃的选择性检验值外(实验数据可能存在误差), 其它的 p 值均小于 0.05, 故在 95% 的置信水平上拒绝原假设, 即皮尔逊相关系数显著地异于 0。

根据查阅资料得到相关系数评价指标:

表 3 相关系数评价指标表

相关性	负	正
无相关性	-0.09 to 0.0	0.0 to 0.09
弱相关性	-0.3 to -0.1	0.1 to 0.3
中相关性	-0.5 to -0.3	0.3 to 0.5
强相关性	-0.9 to -0.5	0.5 to 0.9
极强相关性	-1.0 to -0.9	0.9 to 1

从计算得出结果可知, 每种催化剂组合中乙醇转化率、C4 烯烃的选择性与温度两两之间具有强相关性, 并且乙醇转化率与温度, C4 烯烃的选择性与温度之间的相关性更为显著。

5.1.3 曲线拟合

附件 2 中给出了温度为 350 度时且给定的催化剂组合在一次实验不同时间的测试结果, 选取时间作为横坐标, 乙醇转化率与 C4 烯烃选择性作为纵坐标进行画出散点图, 如下图所示, 从图中我们可以定性得到: 乙醇转化率随着时间的推移, 不断下降, 且下降的速度越来越慢, C4 烯烃选择性随着时间的推移先降低后上升。

我们使用 matlab 曲线拟合工具箱对两条曲线分别进行拟合, 考虑曲线的复杂度与拟合效果(主要分析不同拟合曲线的 SSE、R-square), 最后选出最合适函数进行拟合。拟合结果为:

乙醇转化率(y_1)与时间(x)的关系: 选择二次函数进行拟合

$$y_1 = 0.0001772x^2 - 0.1051x + 45.25$$

C4 烯烃选择性(y_2)与时间(x)的关系: 选择 Fourier 函数进行拟合

$$y_2 = 38.78 + 1.253\cos(0.03x) + 0.265\sin(0.03x) - 0.095\cos(0.06x) + 0.292\sin(0.06x)$$

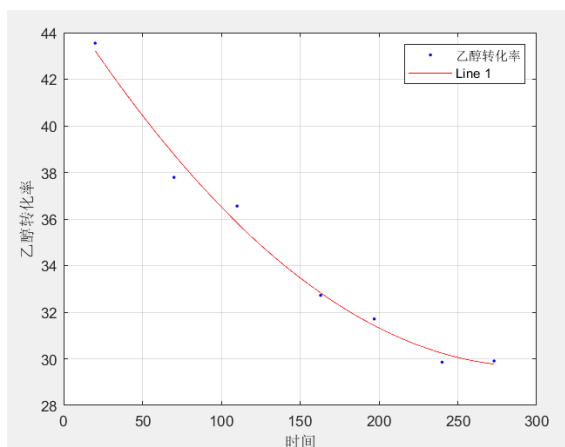


图 9 时间与乙醇转化率关系图

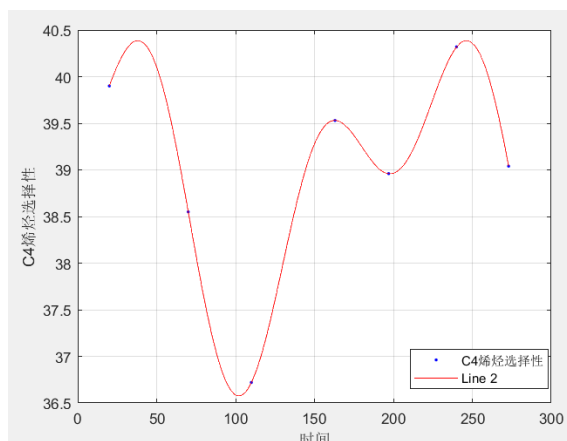


图 10 时间与 C4 烯烃选择性关系图

表 4 时间与乙醇转化率的拟合效果评价表

函数类型	Degree	SSE	R-square
Exponential	-	1.733	0.9885
Fourier	-	0.934	0.9938
Polynomial	1	10.090	0.9329
Polynomial	2	1.835	0.9878
Polynomial	3	1.775	0.9882

表 5 时间与 C4 烯烃选择性的拟合效果评价表

函数类型	Degree	SSE	R-square
Exponential	-	4.43	0.4622
Fourier	-	2.228e-05	1
Polynomial	1	7.855	0.04642
Polynomial	2	6.256	0.2405
Polynomial	3	2.697	0.6726

5.1.4 模型的结论与分析

对附件 1 给出的 14 种 A 装料方式和 7 种 B 装料方式的催化剂组合分别进行研究，得到乙醇转化率，C4 烯烃选择性与温度之间的相关性显著的结论。在一定的温度区间内，温度越高，乙醇转化率和 C4 烯烃选择性升高。相比而言，C4 烯烃选择性到达平稳的温度更低，且在温度开区间存在最优值。

对附件 2 的数据进行拟合处理可知：乙醇转化率与时间近似呈现二次减函数关系，C4 烯烃选择性与时间多项式和指数拟合结果不理想，采用傅里叶级数展开处理。但得到 SSE 接近 0，函数关系仍然不明显，这可能是数据量过少，噪声点过多导致的。

5.2 问题二模型的建立于求解

5.2.1 数据预处理与描述性统计

通过对题目和数据的分析得知，不同催化剂组合与温度的选取，对乙醇转化率以及

C4 烯烃选择性大小有着重要的影响。本文首先通过 Python 中的 pandas 数据分析包对附件一中的数据进行处理，对每一种催化剂组合提取并且增加 Co 负载量、乙醇浓度、HAP 质量、Co/SiO₂ 质量、Co/SiO₂ 和 HAP 装料比、C4 烯烃收率、装料方式，其中装料方式为分类变量，0 表示使用装料方式 I 进行实验，1 表示使用装料方式 II 进行实验，其余变量均为定量变量。对于 A11 这组数据，因为无 HAP，为特殊数据，我们对其舍去。

之后，本文根据各变量之间的独立性以及各变量对因变量的影响重要程度，提取了温度、Co 负载量、乙醇浓度、HAP 质量、Co/SiO₂ 和 HAP 装料比这五个定量变量，以及装料方式这个定性变量，并且使用 Stata 软件来对其进行描述性统计，结果如下表：

表 6 定量变量的描述性统计表

Variable	Obs	Mean	Std	Min	Max
温度/℃	109	315.596	52.240	250	450
Co 负载量/wt%	109	1.440	1.183	0.5	5
乙醇浓度/ml/min	109	1.474	0.528	0.3	2.1
HAP 质量/mg	109	99.633	70.970	10.0	200
Co/SiO ₂ 和 HAP 装料比	109	1.024	0.246	0.49	2.03

表 7 定性变量的描述性统计表

装料方式	Freq	Percent
0	69	63.30
1	40	36.70
Total	109	100.00

5.2.2 多元线性回归模型的建立

回归分析是对具有相关关系的变量之间的数量变化规律进行测定，研究某一随机变量（因变量）与其他一个或几个普通变量（自变量）之间的数量变动关系，并据此对因变量进行估计和预测的分析方法^[4]。

根据 5.1.1 的数据分析，选取温度、Co 负载量、乙醇浓度、HAP 质量、Co/SiO₂ 和 HAP 装料比、装料方式六项指标，对乙醇转化率与 C4 烯烃选择性分别进行回归分析。

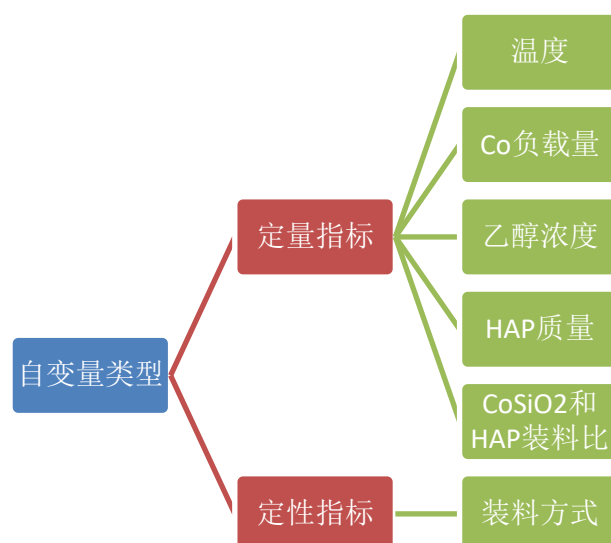


图 11 自变量类型情况介绍

使用 OLS 进行标准化回归：

$$y_i = \beta_0 + \delta_0 scheme + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \mu \quad (1)$$

其中 $X = \{x_1, x_2, x_3, x_4, x_5\} = \{\text{温度}, \text{Co 负载量}, \text{乙醇浓度}, \text{HAP 质量}, \text{Co/SiO}_2 \text{ 和 HAP 装料比}\}$ 。 $y = \{y_1, y_2\} = \{\text{乙醇转化率}, \text{C4 烯烃选择性}\}$, $scheme$ 为装料方式, μ 为扰动项。

使用 Stata 求得乙醇转化率的 OLS 标准化回归的系数, 并且对六个自变量的回归系数进行联合显著性检验, 对于因变量是乙醇转化率的回归函数, $p = 0.00 < 0.01$, 因此在 99% 的置信区间下, 拒绝原假设, 可认为 δ_0 与 $\beta_i (i = 1, 2, 3, 4, 5)$ 不全为 0。对于因变量是 C4 烯烃选择性的回归函数, $p = 0.00 < 0.01$, 在 99% 的置信区间下, 也可以拒绝原假设, 可认为 δ_0 与 $\beta_i (i = 1, 2, 3, 4, 5)$ 不全为 0。

表 8 乙醇转化率的 OLS 标准化回归结果

	温度	Co 负载量	乙醇浓度	HAP 质量	装料比	装料方式 1
P> t 	0.000	0.705	0.000	0.000	0.493	0.047
β	0.772	-0.018	-0.184	0.293	-0.031	0.104

对所有的回归系数进行显著性 t 检验, 其中温度、乙醇浓度、HAP 质量的 p 值小于 0.01, 故在 99% 的置信区间下, 三者的回归系数显著的异于 0, 且由标准化回归系数 β 可知, 温度、HAP 质量与乙醇转化率呈正相关, 乙醇浓度与其呈负相关, 且温度影响最大, 其余因素则对乙醇转化率的影响不显著。

使用 Stata 软件分析 C4 烯烃选择性与自变量之间的关系, 得到的结果如下表:

表 9 C4 烯烃选择性的 OLS 标准化回归结果

	温度	Co 负载量	乙醇浓度	HAP 质量	装料比	装料方式 1
P> t 	0.000	0.000	0.028	0.000	0.173	0.064
β	0.726	-0.305	0.122	0.407	0.714	0.112

对所有的回归系数进行显著性 t 检验, 其中在 99% 的置信区间下, 温度、Co 负载量、HAP 质量的回归系数显著的异于 0, 其中温度、HAP 质量与 C4 烯烃选择性呈正相关, 与 Co 负载量呈负相关, 且温度对 C4 烯烃选择性影响最为显著。

5.2.3 多元线性回归模型的检验与改进

多元线性回归模型的拟合结果往往会受到扰动项的异方差以及多重共线性的影响。对此, 我们对 5.2.2 中的 OLS 标准化回归进行检验。首先, 分别做出乙醇转化率与 C4 烯烃选择性的残差与拟合值的散点图进行大致观察。从图中可以看出, 随着拟合值的增大, 两个图的残差均呈扩散状, 且越来越大, 故扰动项可能存在异方差。

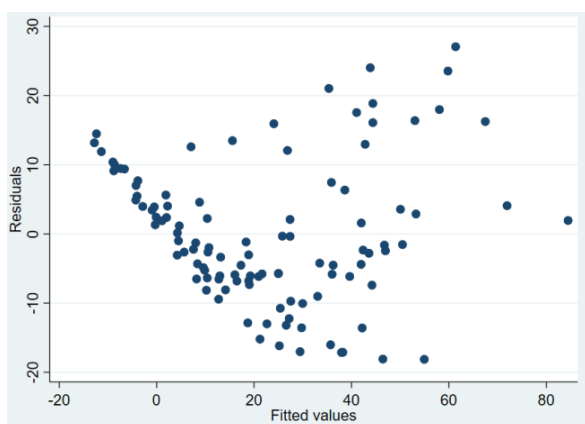


图 12 乙醇转化率残差与拟合值散点图

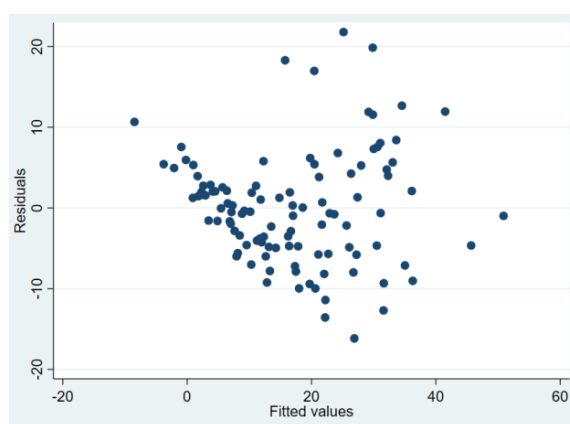


图 13 C4 烯烃选择性残差与拟合值散点图

对扰动项进行异方差检验,使用怀特检验(White test), p 值分别为 0.0005 与 0.0003, 均小于 0.01, 故在 99%的置信水平下拒绝原假设, 可认为扰动项存在异方差。扰动项存在异方差, 会导致 OLS 估计量不再是最优线性无偏估计量。通过查阅资料可知, 对于异方差, 通常有两种处理方法, 一种是仍然进行 OLS 回归, 但使用稳健标准误, 另一种是使用广义最小二乘法 GLS。统计学领域的专家 Stock and Watson (2011)推荐, 在大多数情况下应该使用 OLS 且引入稳健标准误来处理异方差。使用 Stata 软件, 求得改进后的回归结果如下表:

表 10 乙醇转化率 OLS 标准化结合标准误的回归结果

	Coef.	t	P> t
温度	0.340	16.42	0.000
Co 负载量	-0.351	-0.34	0.737
乙醇浓度	-8.033	-3.08	0.003
HAP 质量	0.095	4.99	0.000
装料比	-2.886	-1.04	0.300
装料方式 1	4.957	2.45	0.016
_cons	-81.98156	-8.93	

表 11 C4 烯烃选择性 OLS 标准化结合标准误的回归结果

	Coef.	t	P> t
温度	0.188	14.71	0.000
Co 负载量	-3.480	-4.71	0.000
乙醇浓度	3.134	2.16	0.033
HAP 质量	0.077	5.45	0.000
装料比	3.918	1.68	0.096
装料方式 1	3.130	2.36	0.020
_cons	-55.539	-11.34	

由表格可得温度、乙醇浓度、HAP 质量对乙醇转化率影响最为显著。乙醇转化率的多元线性回归模型为:

$$y_1 = 0.340x_1 - 8.033x_3 + 0.095x_4 - 81.982 \quad (2)$$

温度、Co 负载量、HAP 质量对 C4 烯烃选择性影响最为显著。C4 烯烃选择性的多元线性回归模型为：

$$y_1 = 0.188x_1 - 3.480x_2 + 0.775x_4 - 55.539 \quad (3)$$

其中, $X = \{x_1, x_2, x_3, x_4, x_5\} = \{\text{温度}, \text{Co 负载量}, \text{乙醇浓度}, \text{HAP 质量}, \text{Co/SiO}_2 \text{ 和 HAP 装料比}\}$

最后计算出方差膨胀因子 VIF ，假设有 k 个自变量，那么第 m 个自变量的 VIF_m 为：

$$VIF_m = \frac{1}{1 - R_{1-k|m}^2} \quad (4)$$

$R_{1-k|m}^2$ 是将第 m 个自变量作为因变量，对其余 $k-1$ 个回归得到的拟合优度。 VIF_m 越大，说明第 m 个变量和其他变量的相关性越大。 k 个自变量求得的中最大者记为 VIF ，若 $VIF > 10$ ，认为该回归方程存在严重的多重共线性。

根据两个回归方程的回归结果，计算出二者的 VIF 均为 1.20，故可认为回归方程不存在严重的多重共线性。

5.2.4 模型的结论与分析

本文以乙醇转化率和 C4 烯烃选择性为回归目标，通过对 5 类定量回归系数和 1 类定性回归系数联合进行显著性 t 检验并求得了标准化回归系数 β 。

结合 OLS 标准化检验，观察得到的多元线性回归结果可知，温度、HAP 质量与乙醇转化率呈正相关，乙醇浓度与其呈负相关，且温度影响最为显著；温度、HAP 质量与 C4 烯烃选择性呈正相关，与 Co 负载量呈负相关，温度对 C4 烯烃选择性影响最大。处理异方差后的获得的回归方程经过膨胀因子的检验，不具有多重共线性，与实际数据符合良好。

5.3 问题三模型的建立与求解

5.3.1 BP 神经网络模型的建立

在 5.2.3 中分别建立了乙醇转化率与 C4 烯烃选择性的多元线性回归模型，但二者的拟合优度(R-square)仅为 0.804 与 0.737，我们加入平方项进行拟合，拟合优度仍然提升不明显，各自变量和因变量之间的关系并非简单的线性组合，故上述模型仅可用来描述各自变量对因变量的影响程度。对于数据的分析和预测我们采用 BP 神经网络模型。

在回归模型中，我们对于各自变量地重要程度进行了考察。温度、Co 负载量、乙醇浓度、HAP 质量、Co/SiO₂ 和 HAP 装料比对乙醇转化率与 C4 烯烃选择性来说最为重要，进而影响 C4 烯烃收率。因此，我们采用这五个因素作为输入层，C4 烯烃收率作为输出层，对 BP 神经网络进行训练。通过参数调节，我们选取的隐藏层神经元个数为 10，使用 Levenberg-Marquardt 方法进行训练，该算法能借由执行时修改参数达到结合高斯-牛顿算法以及梯度下降法的优点，并对两者之不足作改善^[5]。

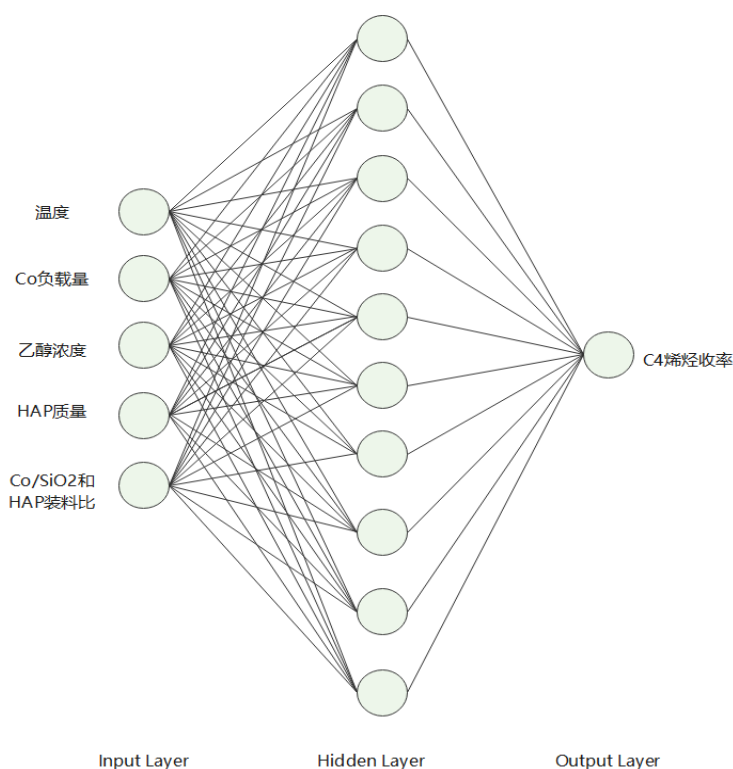


图 14 神经网络模型的示意图

5.3.2 BP 神经网络模型的训练与优化

对样本数据进行划分，划分为训练集（Training set）、验证集（Validation set）和测试集（Testing set）。其中训练集用作模型拟合的数据样本，验证集用于调整模型的超参数和用于对模型的能力进行初步评估，测试集用于评估模型最终模型的泛化能力。通过评估验证集的均方差，找出最优的迭代次数，在迭代次数为 24 次时，对模型的评估达到最优。此时，训练集、验证集、测试集的拟合优度分别为 0.998，0.9887，0.909，模型整体的拟合优度为 0.986，极大的优于多元线性回归模型的拟合优度。

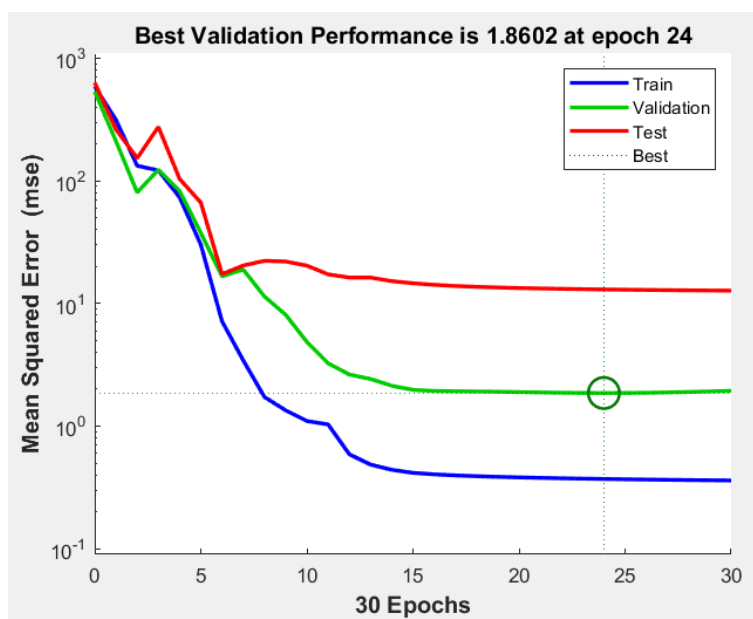


图 15 迭代次数与数据均方差关系图

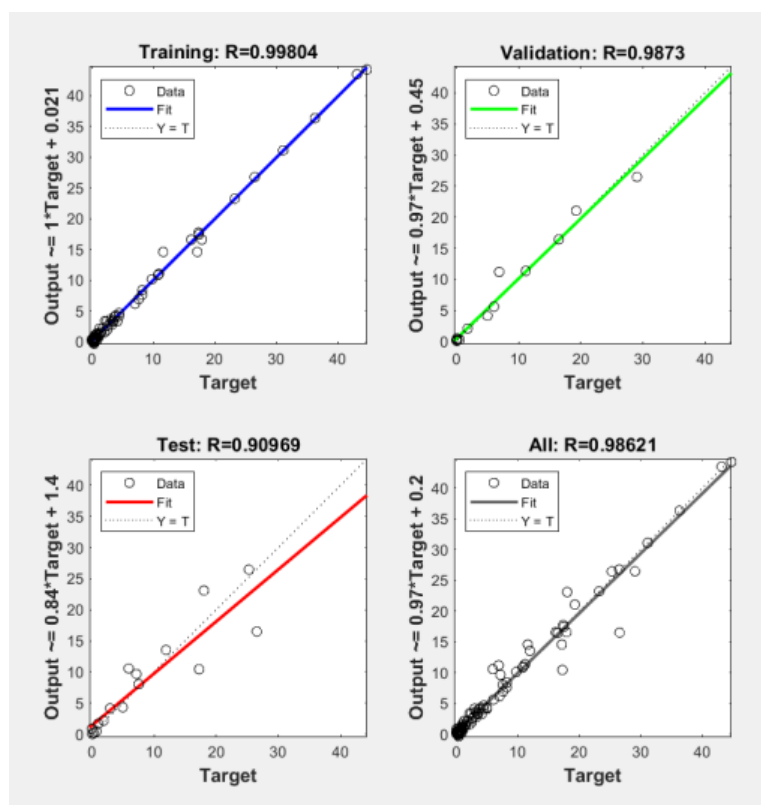


图 16 神经网络预测结果评价图

5.3.3 网格搜索法求最优解

通过 5.2.1 中的定量数据的描述性统计表，可以得到输入变量的大致范围。选取温度的变化范围为 250~450，每隔 25 取一组值，Co 负载量范围 0.5~5，每隔 0.5 取一组值，乙醇浓度 0.3~2.1，每隔 0.2 取一组值，HAP 质量变化范围为 10~240，每隔 30 进行取值，Co/SiO₂ 和 HAP 装料比范围为 0.7~2.1 每隔 0.3 取一组数据。再将这些数据使用 pandas 进行组合，每个数据组合导入 5.3.2 中建立的神经网络模型中，预测 C4 烯烃选择性(%)。剔除异常值后并且对预测值进行降序排列，结果的箱型图如下(具体数据见附件 ans3.xlsx)。

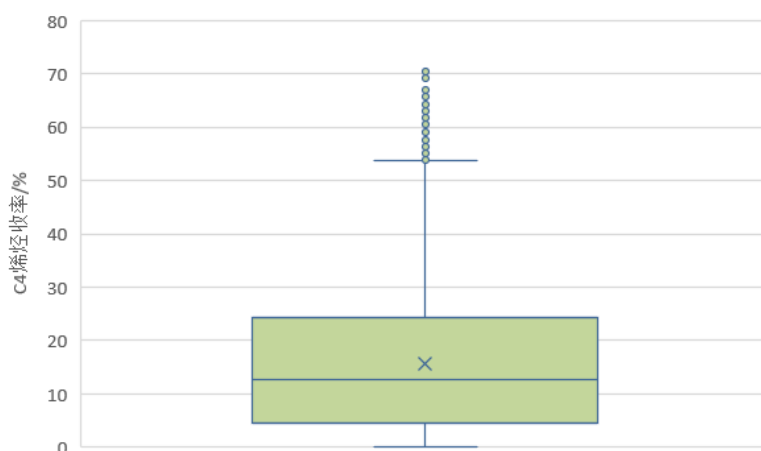


图 17 预测结果的箱型图

5.3.4 模型的结论与分析

根据 BP 神经网络预测得到的数据，我们对其进行分析，在附件 1 中给出的实验的

基础上进行分析，附件 1 中的各组实验最高温度大多为 400 摄氏度，Co 负载量仅包括 0.5，1，2，5 几种情况，乙醇浓度包括 0.3，0.9，1.68，2.1 几种情况，CoSiO₂ 和 HAP 装料比均为 1。根据预测数据与常见的实验条件，我们可以得到在实验条件允许的前提下，最优的三组解如下表所示：

表 12 实际条件下，三组最优解表

温度 /℃	Co 负载量 /wt%	乙醇浓度 /ml/min	HAP 质量 /mg	Co/SiO ₂ 和 HAP 装料比值	C4 烯烃收率 的预测值
400	0.5	0.9	240	1	45.53%
400	1	0.9	240	1	45.42%
400	1	0.9	240	1	45.37%

当温度条件限定为小于 350 度的情况下，我们对得到的数据再次进行分析，可以得到在实验条件允许的前提下最优的三组解，如下表所示：

表 13 温度限制下，三组最优解表

温度 /℃	Co 负载量 /wt%	乙醇浓度 /ml/min	HAP 质量 /mg	Co/SiO ₂ 和 HAP 装料比值	C4 烯烃收率 的预测值
350	0.5	1.7	240	1	30.25%
350	1	1.7	240	1	29.37%
350	1.5	1.7	240	1	27.32%

本文根据神经网络求解结果与附件一部分数据，对催化条件探究模型输入层变量的影响做以定性的解释：

考虑到温度对 C4 烯烃收率的影响，随着反应温度的升高，乙醇转化率提高。高碳醇的选择性逐渐减小，C4 烯烃选择性随之先升高后降低。在 400℃时最适合目标产物的转化。温度小于等于 350℃，产物主要是高碳醇，反应温度大于 350℃主要是 C4 烯烃。

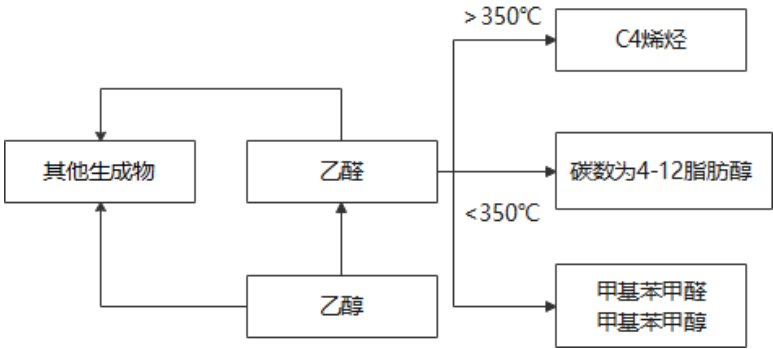


图 18 乙醇偶合过程中产物可能的关系

对 Co 负载量的探究表明，Co 质量的影响较弱，催化反应的主要承担者是 SiO₂，Co 负载量为 0.5,1 时，Co 在催化剂表面难以被还原，当负载量提升至 2,5 时，Co 在催化剂表面聚集。而单质钴可以促进乙醇脱氢生成乙醛。

装料比为 1:1 的 C4 烯烃选择性最佳。随着装料比的增大，C4 烯烃的选择性呈现先上升后下降的趋势。高碳醇选择性增加，其他生成物基本保持稳定。HAP 含量增加，有利于烯烃的转化，当增加到一定程度时，烯烃会进一步耦合生成高碳醇。

5. 4 问题四模型的建立与求解

根据 BP 神经网络的预测结果，我们认为 C4 烯烃收率大于 35%为较好的实验条件。

因 BP 神经网络预测与真实结果存在偏差，而且预测的数据量较大，故我们选取四分位数对预测结果进行描述。四分位数的主要概念可以理解为将数据从小到大排序后进行四等分，位于等分点上的数值则可以分别确定为该位置上的分位数^[6]。

为了得到尽可能高地得到 C4 烯烃收率，而且考虑到预测偏差、大数据量地影响，我们将大于 35%C4 烯烃收率的数据分为 5 组，每一组计算出上四分位数，以及距离其上四分位数最近的烯烃收率，该烯烃收率可描述该范围内烯烃收率的大体趋势。使用 SPSS 软件进行计算，结果如下。

表 14 预测数据的分位数表

\ 组数	百分位数		
	25	50	75
1	61.06	62.18	64.20
2	51.43	53.56	55.79
3	46.11	46.99	48.24
4	41.11	42.22	43.30
5	35.98	37.10	38.38

其中 1 组代表 C4 烯烃收率大于 60%，2 组代表 C4 烯烃收率位于 50%~60%之间，3 组代表 C4 烯烃收率位于 45%~50%之间，4 组代表 C4 烯烃收率位于 40%~45%之间，5 组代表 C4 烯烃收率在 35%~40%之间。其对应的变量取值如下表：

表 15 增加的五组实验表

温度 /℃	Co 负载量 /wt%	乙醇浓度 /ml/min	HAP 质量 /mg	Co/SiO ₂ 和 HAP 装料比值	C4 烯烃收率的预测值
450	0.5	0.3	130	0.7	64.16%
450	5	1.1	130	1.9	55.79%
450	5	0.3	70	1.3	48.24%
400	2	1.5	240	1.6	43.30%
250	3.5	2.1	240	1.3	38.38%

六、结果的分析与检验

6.1 多元线性回归置信区间的分析

在不同的置信区间下，多元线性回归方程的自变量的显著性不同。通常选取 90%，95%或 99%的置信区间。对于 5.2.3 求得的回归结果进行分析，在 90%的置信区间下，乙醇转化率的回归方程中，温度、乙醇浓度、HAP 质量、装料方式的影响显著，C4 烯烃选择性的回归方程中，温度、Co 负载量、乙醇浓度、HAP 质量、装料方式的影响显著。在 95%的置信区间下，二者的回归方程中，各因素的显著性与在 90%的置信区间下相同。在 99%的置信区间下，温度、乙醇浓度、HAP 质量对乙醇转化率影响显著，温度、Co 负载量、HAP 质量对 C4 烯烃选择性影响最为显著。考虑到精度以及模型复杂度的影响，本文采取在 99%的置信区间下进行回归分析。

6.2 神经网络相关参数的灵敏度分析

本文对模型三用到的神经网络的参数进行三种方法的验证，得到如下结果：

方法一：采用莱文贝格—马夸特方法（Levenberg-marquardt），
 方法二：采用贝叶斯正则化，通过修改损失函数防止过拟合。
 方法三：采用优共轭梯度法（Scaled Conjugate Gradient）^[7]。
 并且选择不同的隐藏层中神经元的个数，以确定最优的训练结果。

表 16 不同参数下的训练结果表

神经元个数	方法	训练集的均方误差
8	一	17.94
	二	7.52
	三	18.84
10	一	5.10
	二	0.44
	三	28.9
13	一	5.18
	二	0.39
	三	43.10

在验证集对模型表现的最优验证，其结果的大小反映了神经网络的预测性能，训练集的均方误差越小，拟合效果越好。综合比较神经元个数在 8、10、13 时三种方法的最优验证。观察得到，方法二可能会引起过拟合等一些问题，预测结果不理想。方法三的拟合效果差。由此可知，当采用莱文贝格—马夸特方法对 10 个神经元进行预测时，表现最优验证的值为 5.10，拟合效果良好。

七、模型的优缺点分析

7.1 多元线性回归模型的优点

多元线性回归模型，对于分析多个自变量对因变量的影响有着重要的作用，通过多元线性回归分析以及对自变量的显著性检验，可以得出对因变量影响最为重要的因素，为神经网络的输入参数进行更好的配置。

7.2 最适催化条件探究模型的优点

1. 自学习和自适应能力：训练 BP 神经网络时，能够通过学习自动提取输出、输出数据间的合理规则，并自适应的将学习内容记忆于网络的权值中。

2 非线性映射能力：BP 神经网络，在本质上完成了一个从输入到输出的映射过程，数学理论证明三层的神经网络就能够以任意精度逼近任何非线性连续函数。因此适于探寻诸如乙醇偶合制备 C4 烯烃的催化条件内部机制这类复杂的作用。

3. 泛化能力：BP 神经网络具有将其学习成果应用于新知识的能力。在设计模式识别分类器时，即要正确分类网络所需对象，又要关心它在经过训练后，可否对没有见过的模式或有部分噪声污染的模式，进行正确分类。

4. 容错能力：BP 神经网络在它局部或部分神经元受破坏后对全局的训练结果造成的影响很小，即使系统在受到局部损伤时可以维持正常运作。

7.3 模型的缺点与改进

由于数据量较少以及自变量的个数较少，对于预测结果的准确度有着一定的影响。

在样本量固定的情况下,可以考虑从数据中挖掘对结果有影响的信息,增加自变量的个数,来提高拟合以及预测的精确度。

八、模型的改进与推广

本文所给出的基于多元线性规划与 BP-神经网络的探究最适乙醇偶合制备 C4 烯烃催化条件模型,能够很好的反应 C4 烯烃收率与所提供的显著因变量之间的关系。考虑到实际中,同组变量可能要进行多次重复实验。本模型能够在人力、物力有限的制备实验后,仅得到多组变量较少的实验数据下进行回归预测,有着较为合理的预估。

由于附录中的实验数据量较少,多元线性回归与多元非线性二次回归的结果对目标烯烃收率的误差较大。无法通过乙醇转化率和 C4 烯烃选择性两个拟合函数相乘得到 C4 烯烃收率的表达式。可以考虑建立与化学动力学有关的指数模型,对本例样本或增加一定补充实验后的大样本进行预测。如此得到的方程可用粒子群算法进行多元非线性规划求解。将得到的结果与本 BP-神经网络模型做以比对,从而更好地解释乙醇偶合制备 C4 烯烃过程中因变量的影响。

九、参考文献

- [1] 张俊旺,汪洋,杨利敏, et al. 固定床反应器中乙醇制混合烃工艺过程考察[J].化学反应工程与工艺, 2016, 32(01): 15-21.
- [2] 盛骤、谢式千、潘承毅.概率论与数理统计(第四版)[M].北京:高等教育出版社.2008
- [3] Davies H M . A Course in Theoretical Statistics. By Rahman N. A. Pp. xv, 542. 4.20. 1968. (Griffin)[J]. Mathematical Gazette, 1971, 55(394):xv-458.
- [4] 盛骤、谢式千、潘承毅.概率论与数理统计(第四版)[M].北京:高等教育出版社.2008
- [5] 习国泰.改进 Levenberg-Marquardt 算法的复杂度分析[D].上海交通大学, 2012.
- [6] 温有栋.基层统计分析中四分位数计算方法的思考[J].统计与管理.2018, (3):55-58.
- [7] 百度百科.共轭梯度法.<https://baike.baidu.com/item/共轭梯度法/7139204?fr=aladdin>. 2021.1.25

十、附录

附录 1
支撑材料文件列表
<p>支撑材料分为两类</p> <p>第一类为文章中所使用的代码，在附录中进行了说明，且附在了支撑材料的“代码.docx”中。第一问中所有组合的相关系数求解代码为“相关系数求解.rar”。神经网络相关代码在“神经网络代码”文件夹中，包括训练的网络(net.mat)，训练数据的输入值(three.mat)，训练数据的输出值(y_predict.mat)，以及代码的调用函数(函数.mat)，代码的主函数。(主函数.mat)。</p> <p>第二类为表格数据</p> <p>A1.xlsx、A2.xlsx、B1.xlsx、B2.xlsx 为文章中举例的四种催化剂组合的温度，乙醇转化率，C4 烯烃选择性。</p> <p>pre1.xlsx 为所有催化剂组合的温度，乙醇转化率，C4 烯烃选择性。</p> <p>pre2.xlsx、data.xlsx 为使用 python 中的 pandas 数据处理包预处理数据所得到的结果。</p> <p>pre3.xlsx 为自变量与 C4 烯烃收率的数据表。</p> <p>data3.xlsx 为使用 python 中的 pandas 数据处理包生成的各种自变量组合的情况，来作为神经网络的输入数据进行预测。</p> <p>附件 1，附件 2 为题目中所给的表格。</p>

附录 2			
各种组合皮尔逊相关系数(剔除 A11)			
组合	温度与乙醇转化率 皮尔逊相关系数	温度与 C4 烯烃的选择性 皮尔逊相关系数	乙醇转化率与 C4 烯烃的 选择性皮尔逊相关系数
A1	0.9655	0.8871	0.7712
A2	0.9950	0.9143	0.9029
A3	0.9820	0.9554	0.9770
A4	0.9975	0.9578	0.9691
A5	0.9345	0.9696	0.9862
A6	0.9836	0.8854	0.8990
A7	0.9994	0.9682	0.9600
A8	0.9771	0.9916	0.9959
A9	0.9210	0.9974	0.9150
A10	0.9229	0.8615	0.9878
A12	0.9631	0.9832	0.9944
A13	0.9363	0.9883	0.8932
A14	0.9639	0.9592	0.9975

B1	0.9621	0.9858	0.9898
B2	0.9293	0.9848	0.9704
B3	0.8901	0.9710	0.9439
B4	0.8993	0.8950	0.9663
B5	0.9124	0.9776	0.9720
B6	0.9399	0.9821	0.9328
B7	0.9361	0.9944	0.9599

附录 3

附件 1 数据预处理代码

```
import pandas as pd
import numpy as np

data=pd.read_excel(r'C:\Users\DELL\Desktop\附件 1.xlsx')
data['Co 负载量']=data['催化剂组合'].dropna().map(lambda
x:x.split('mg')[1].split('wt')[0]).astype('double')
data['Co 负载量']=data['Co 负载量'].fillna(method='pad')
data['催化剂组合编号']=data['催化剂组合编号'].fillna(method='pad')
data['乙醇浓度']=data['催化剂组合'].dropna().map(lambda x:x.split('乙醇浓度
')[1].split('ml')[0]).astype('double')
data['乙醇浓度']=data['乙醇浓度'].fillna(method='pad')
data=data.drop(index=[54,55,56,57,58])
data['催化剂组合'].dropna()
data['Co/SiO2 质量']=data['催化剂组合'].dropna().map(lambda
x:x.split('mg')[0]).astype('double')
data['HAP 质量']=data['催化剂组合'].dropna().map(lambda
x:x.split('Co/SiO2-')[1].split('mg')[0]).astype('double')
data['Co/SiO2 质量']=data['Co/SiO2 质量'].fillna(method='pad')
data['HAP 质量']=data['HAP 质量'].fillna(method='pad')
data['Co/SiO2 和 HAP 装料比']=round(data['催化剂组合'].dropna().map(lambda
x:x.split('mg')[0]).astype('double')/data['催化剂组合
'].dropna().map(lambda
x:x.split('Co/SiO2-')[1].split('mg')[0]).astype('double'),2)
data['Co/SiO2 和 HAP 装料比']=data['Co/SiO2 和 HAP 装料比
'].fillna(method='pad')
data['C4 烯烃收率(%)']=round(data['乙醇转化率
(%)'].astype('double')*data['C4 烯烃选择性(%)'].astype('double')/100,2)
data['装料方式']=data['催化剂组合编号'].str.contains('B').astype('int')
data.to_excel(r'C:\Users\DELL\Desktop\data.xlsx')
```

附录 4

第一问求解 R 值和 p 值的代码

```
clear;clc
load 'A1.mat' %以 A1 组催化剂加载为例，当文件名有空格时，加引号
%% 计算各列之间的相关系数以及 p 值
[R,P] = corrcoef(X) %用自带计算 p 值的函数，注意双侧检验的 p 值要乘以 2
%% 进行假设检验，绘 t 分布图
x = -4:0.1:4;
y = tpdf(x,3); %求 t 分布的概率密度值 5 是自由度
figure(1)
plot(x,y,'-')
grid on
hold on % 保留原图，方便继续操作
% tinv 函数返回 t 分布的双尾反函数
tinv(0.975,3) % 结果为 3.1824, 0.975 为置信区间
% 该函数为累积密度函数 cdf 的反函数
plot([-3.1824,-3.1824],[0,tpdf(-3.1824,3)],'r-')
plot([3.1824,3.1824],[0,tpdf(3.1824,3)],'r-')
```

附录 5

第一问 SPSS 作四种催化剂组合的散点图

```
GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\DELL\Desktop\A1.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 1 WINDOW=FRONT.
GRAPH
  /SCATTERPLOT(MATRIX)=温度 乙醇转化率 C4 烯烃选择性
  /MISSING=LISTWISE.

GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\DELL\Desktop\A2.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
```

```

/READNAMES=ON
/DATATYPEMIN PERCENTAGE=95.0
/HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 2 WINDOW=FRONT.
GRAPH
  /SCATTERPLOT (MATRIX)=温度 乙醇转化率 C4 烯烃选择性
  /MISSING=LISTWISE.

GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\DELL\Desktop\B1.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 3 WINDOW=FRONT.
GRAPH
  /SCATTERPLOT (MATRIX)=温度 乙醇转化率 C4 烯烃选择性
  /MISSING=LISTWISE.

GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\DELL\Desktop\B2.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 4 WINDOW=FRONT.
GRAPH
  /SCATTERPLOT (MATRIX)=温度 乙醇转化率 C4 烯烃选择性
  /MISSING=LISTWISE.

```

附录 6

第一问 SPSS 进行夏皮洛-威尔克检验检验

```

GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\DELL\Desktop\A1.xlsx'

```

```

/SHEET=name 'Sheet1'
/CELLRANGE=FULL
/READNAMES=ON
/DATATYPEMIN PERCENTAGE=95.0
/HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 5 WINDOW=FRONT.
EXAMINE VARIABLES=温度 乙醇转化率 C4 烯烃选择性
/PLOT BOXPLOT STEMLEAF NPLOT
/COMPARE GROUPS
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.

GET DATA
/TYPE=XLSX
/FILE='C:\Users\DELL\Desktop\A2.xlsx'
/SHEET=name 'Sheet1'
/CELLRANGE=FULL
/READNAMES=ON
/DATATYPEMIN PERCENTAGE=95.0
/HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 6 WINDOW=FRONT.
EXAMINE VARIABLES=温度 乙醇转化率 C4 烯烃选择性
/PLOT BOXPLOT STEMLEAF NPLOT
/COMPARE GROUPS
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.

GET DATA
/TYPE=XLSX
/FILE='C:\Users\DELL\Desktop\B1.xlsx'
/SHEET=name 'Sheet1'
/CELLRANGE=FULL
/READNAMES=ON
/DATATYPEMIN PERCENTAGE=95.0
/HIDDEN IGNORE=YES.
EXECUTE.

```

```

DATASET NAME 数据集 7 WINDOW=FRONT.
EXAMINE VARIABLES=温度 乙醇转化率 C4 烯烃选择性
  /PLOT BOXPLOT STEMLEAF NPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.

GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\DELL\Desktop\B2.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.

EXECUTE.

DATASET NAME 数据集 8 WINDOW=FRONT.
EXAMINE VARIABLES=温度 乙醇转化率 C4 烯烃选择性
  /PLOT BOXPLOT STEMLEAF NPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.

```

附录 7

第一问曲线拟合代码

```

%% 导入数据
[~, ~, raw] = xlsread('C:\Users\DELL\Desktop\附件 2.xlsx', '稳定性测试', 'A4:H10');
data = reshape([raw{:}], size(raw));
Untitled = table;
Untitled.VarName1 = data(:,1);
Untitled.VarName2 = data(:,2);
Untitled.VarName3 = data(:,3);
Untitled.VarName4 = data(:,4);
Untitled.VarName5 = data(:,5);
Untitled.VarName6 = data(:,6);
Untitled.VarName7 = data(:,7);

```



```

Untitled.VarName8 = data(:,8);
clearvars data raw;
%拟合曲线 1
x=table2array(Untitled(:,1));
y1=table2array(Untitled(:,2));
y2=table2array(Untitled(:,4));
[xData, yData] = prepareCurveData( x, y1 );
ft = fittype( 'poly2' );
[fitresult, gof] = fit( xData, yData, ft );
figure( 'Name', 'Line 1' );
h = plot( fitresult, xData, yData );
legend( h, '乙醇转化率', 'Line 1', 'Location', 'NorthEast' );
xlabel 时间
ylabel 乙醇转化率
grid on
%拟合曲线 2
[xData, yData] = prepareCurveData( x, y2 );
ft = fittype( 'fourier2' );
opts = fitoptions( 'Method', 'NonlinearLeastSquares' );
opts.Display = 'Off';
opts.StartPoint = [0 0 0 0 0 0.0248347245343067];
[fitresult, gof] = fit( xData, yData, ft, opts );
figure( 'Name', 'Line 2' );
h = plot( fitresult, xData, yData );
legend( h, 'C4 烯烃选择性', 'Line 2', 'Location', 'NorthEast' );
xlabel 时间
ylabel C4 烯烃选择性
grid on

```

附录 8

第二问 OLS 回归代码与 OLS 结合稳健的标准误代码

```

import excel "C:\Users\DELL\Desktop\pre2.xlsx", sheet("Sheet1") firstrow
sum 温度 Co 负载量 乙醇浓度 HAP 质量 CoSiO2 和 HAP 装料比 乙醇转化率 C4 烯烃选择性
tab 装料方式,gen(A)
reg 乙醇转化率 温度 Co 负载量 乙醇浓度 HAP 质量 CoSiO2 和 HAP 装料比 A1 A2,b
estat imtest,white
reg 乙醇转化率 温度 Co 负载量 乙醇浓度 HAP 质量 CoSiO2 和 HAP 装料比 A1 A2,r
estat vif
reg C4 烯烃选择性 温度 Co 负载量 乙醇浓度 HAP 质量 CoSiO2 和 HAP 装料比 A1 A2,b
estat imtest,white
reg C4 烯烃选择性 温度 Co 负载量 乙醇浓度 HAP 质量 CoSiO2 和 HAP 装料比 A1 A2,r
estat vif

```

附录 9

第三问生成神经网络的预测数据代码

```
import pandas as pd
import numpy as np
x1=[250,275,300,325,350,375,400,425,450]
x2=[0.5,1,1.5,2,2.5,3,3.5,4,4.5,5]
x3=[0.3,0.5,0.7,0.9,1.1,1.3,1.5,1.7,1.9,2.1]
x4=[10,40,70,100,130,160,190,210,240]
x5=[0.7,1,1.3,1.6,1.9,2.1]
import itertools
for item in itertools.product(x1,x2,x3,x4,x5):
    l.append(item)
data3=pd.DataFrame(l)
data3.to_excel(r'C:\Users\DELL\Desktop\data3.xlsx')
```

附录 10

第三问 matlab 神经网络预测与训练代码

```
%主函数
load three.mat
[r,c]=size(three)
predict_y = zeros(r,1); % 初始化 predict_y
for i = 1: r
    result = sim(net, three(i,:));
    predict_y(i) = result;
end

function [Y,Xf,Af] = myNeuralNetworkFunction(X,~,~)
y1_step1.ymin = -1;
y1_step1.gain = 0.0447227191413238;
y1_step1.xoffset = 0.01;
isCellX = iscell(X);
if ~isCellX
    X = {X};
end
% Dimensions
TS = size(X,2); % timesteps
if ~isempty(X)
    Q = size(X{1},1);
else
```

```

    Q = 0;
end
% Allocate Outputs
Y = cell(1,TS);
% Time loop
for ts=1:TS
    % Input 1
    X{1,ts} = X{1,ts}';
    Xp1 = mapminmax_apply(X{1,ts},x1_step1);
    % Layer 1
    a1 = tansig_apply(repmat(b1,1,Q) + IW1_1*Xp1);
    % Layer 2
    a2 = repmat(b2,1,Q) + LW2_1*a1;
    % Output 1
    Y{1,ts} = mapminmax_reverse(a2,y1_step1);
    Y{1,ts} = Y{1,ts}';
end
% Final Delay States
Xf = cell(1,0);
Af = cell(2,0);

% Format Output Arguments
if ~isCellX
    Y = cell2mat(Y);
end
end
% Map Minimum and Maximum Input Processing Function
function y = mapminmax_apply(x,settings)
y = bsxfun(@minus,x,settings.xoffset);
y = bsxfun(@times,y,settings.gain);
y = bsxfun(@plus,y,settings.ymin);
end
% Sigmoid Symmetric Transfer Function
function a = tansig_apply(n,~)
a = 2 ./ (1 + exp(-2*n)) - 1;
end
% Map Minimum and Maximum Output Reverse-Processing Function
function x = mapminmax_reverse(y,settings)
x = bsxfun(@minus,y,settings.ymin);
x = bsxfun(@rdivide,x,settings.gain);
x = bsxfun(@plus,x,settings.xoffset);
end

```

附录 11

第四问 SPSS 求分位数代码

```
GET DATA
  /TYPE=XLSX
  /FILE=&apos;C:\Users\DELL\Desktop\类别 1.xlsx&apos;
  /SHEET=name &apos;Sheet1&apos;
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 1 WINDOW=FRONT.
FREQUENCIES VARIABLES=C4 烯烃收率
  /NTILES=4
  /ORDER=ANALYSIS.
GET DATA
  /TYPE=XLSX
  /FILE=&apos;C:\Users\DELL\Desktop\类别 2.xlsx&apos;
  /SHEET=name &apos;Sheet1&apos;
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 1 WINDOW=FRONT.
FREQUENCIES VARIABLES=C4 烯烃收率
  /NTILES=4
  /ORDER=ANALYSIS.
GET DATA
  /TYPE=XLSX
  /FILE=&apos;C:\Users\DELL\Desktop\类别 3.xlsx&apos;
  /SHEET=name &apos;Sheet1&apos;
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 2 WINDOW=FRONT.
FREQUENCIES VARIABLES=C4 烯烃收率
```

```

/NTILES=4
/ORDER=ANALYSIS.

GET DATA
  /TYPE=XLSX
  /FILE=&apos;C:\Users\DELL\Desktop\类别 4.xlsx&apos;
  /SHEET=name &apos;Sheet1&apos;
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 3 WINDOW=FRONT.
FREQUENCIES VARIABLES=C4 烯烃收率
  /NTILES=4
  /ORDER=ANALYSIS.

GET DATA
  /TYPE=XLSX
  /FILE=&apos;C:\Users\DELL\Desktop\类别 5.xlsx&apos;
  /SHEET=name &apos;Sheet1&apos;
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 5 WINDOW=FRONT.
FREQUENCIES VARIABLES=C4 烯烃收率
  /NTILES=4
  /ORDER=ANALYSIS.

```