Prediction of Optimal Experimental Conditions for The Preparation of C4 Alkenes by Ethanol Coupling Based on Machine Learning Algorithm

Zihao Deng[1], Qiuliang Lin[1], Junjie Chen[1] and Shixian Zhang[1 *]

[1]*Jiangxi University of Science and Technology, Ganzhou City, Jiangxi Province, China*
Zihao Deng: 773791900@qq.com, Qiuliang Lin: 1808805412@qq.com, Junjie Chen: 756364097@qq.com
** Corresponding author: 249154340@qq.com*

Abstract:
The paper aims to combine computers with chemical experiments, using known experimental data to predict the optimal experimental conditions for the preparation of C4 alkenes by ethanol coupling. In this paper, the known data are predicted by using four machine learning algorithms of random forest, *XGBoost, SVR* and *LightGBM*, and by comparing the degree of fitting of the predicted value and the true value, the *XGBoost* algorithm is selected as the most suitable model to establish the algorithm and the degree of fit is 91.61%, and the best experimental condition prediction model based on the *XGBoost* algorithm is established. Temperature and catalyst combinations that yielded the best results were:400°C, 200mg 1wt%Co/SiO2-200mg HAP-ethanol concentration of 0.9ml/min. Temperatures under 350°C, and the optimal catalyst combination and temperature were: 325°C, 200mg 2wt%Co/SiO2- 200mg HAP-ethanol concentration of 1.68ml/min. Through machine learning algorithms, the most suitable ratio of chemical experiments is solved, it is to obtain the most efficient experimental ratio, reduce unnecessary experiments, save experiment time, and consume the least amount of chemical materials. Moreover, the model in this paper can also be used for designing auxiliary experiments in other chemical and physical fields with some applicability.

Keywords: *Machine learning, Process conditions, XGBoost algorithm, Computer-aided experimental design, Ethanol-coupled to prepare C4 alkenes.*

1 INTRODUCTION

1.1 The background of the question

With the sharp decline in fossil energy production and the increasing problem of environmental pollution, the traditional production method of using fossil energy as raw material is gradually being replaced. At present, the production and preparation of C4 alkenes using clean energy ethanol has been developed, but in the process of using ethanol for preparation, the use of different catalyst combinations and temperatures will lead to different effects of C4 alkene selectivity and C4 alkene yield. In practice, in order to obtain the optimal reaction conditions, it is often necessary to carry out multiple experiments, which will consume a lot of manpower, material resources and financial resources. At present, as computer technology continues to evolve, the use of machine learning to assist in deriving optimal response conditions may be a new trend.

1.2 Experimentation data illustration

The known experimental data used in this paper are derived from the data of Question B of the 2021 Higher Education Society Cup National College Students Mathematical Modeling Competition.

Definition of the term:

1. Selectivity: Refers to the proportion of all products in the preparation of C4 alkenes by ethanol coupling;
2. Cobalt load (CL): Refers to the ratio of cobalt to silicon dioxide by weight;
3. HAP(H): A catalyst carrier with the Chinese name hydroxyapatite;
4. Co/SiO2 and HAP charging ratio(C/H): Refers to the ratio of quality of Co/SiO2 and HAP;
5. Ethanol conversion: Refers to the one-way conversion rate of ethanol per unit time;

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 421-427, 2023.
https://doi.org/10.2991/978-94-6463-040-4_64

6. Temperature: Refers to the reaction temperature when ethanol-coupled to prepare C4 alkenes;
7. C4 alkene yield: he product of ethanol conversion and selectivity of C4 alkenes

1.3 Questions that need to be solved

The data processing of the known experimental data is obtained, the main factors affecting the preparation of C4 alkenes are obtained, and the appropriate mathematical model based on the machine learning algorithm is established in combination with the main factors obtained by the analysis, and the optimal experimental conditions for the preparation of C4 alkenes are predicted by computer programming. Through the predicted experimental conditions, speed up the process of experiments, reduce the number of experiments, reduce the cost of experiments, and use computers to assist in the design of experiments.

2 ANALYSIS OF THE QUESTION

2.1 Analysis of optimal experimental conditions predicting problems

For the optimal experimental condition prediction problem, the problem that needs to be solved is how to choose the appropriate catalyst combination and temperature, so that the yield of C4 alkene is the highest under the same experimental conditions, because this problem needs to predict the most suitable experimental conditions, and the known experimental data is small. Therefore, this paper uses four machine learning algorithms of random forest [5], *XGBoost* [1], *SVR* [4], and *LightGBM* [3] to predict the known experimental data, divide the data into training sets and test sets, use the training set data to train the models based on the four algorithms, bring the test set data into the models established by the four algorithms, and compare the fitting of the predicted values and actual values solved by the four algorithms. The algorithm with the highest degree of fitting is selected to solve the problem.

2.2 Assumptions of the model

1. It is assumed that the experimental data provided in this article is valid;
2. It is assumed that the temperature can be constant during the experimental process of this article;
3. Assume that only by-products from known data are considered in this paper.
4. Suppose this article only considers the study of the by-products given in the title.

3 SOLUTION OF THE PROBLEM

3.1 Description of experimental products

$$C_2H_5OH \xrightarrow{catalyst\&temperature} C_4H_8 + Byproduct$$

According to the above simplified formula of the chemical equation, it is simple to see the reaction of the experiment, and the statistical table of the specific by-products is as follows:

Table 1: By-product composition statistical table

Catalyst	Quartz sand or HAP
By-product	Ethylene
	Acetaldehyde
	Fatty alcohols with a carbon number of 4-12
	Methyl benzaldehyde and methylbenzyl alcohol
	Other products

According to the literature [2], it is known that when the acidity and alkalinity in the catalyst are different, the by-products generated will change.

3.2 Data processing and visualization

For the catalyst combination given by the known data, the above analysis is synthesized to show the following six types of conditions:

(1) The number of ethanol milliliters is different, and other conditions are the same

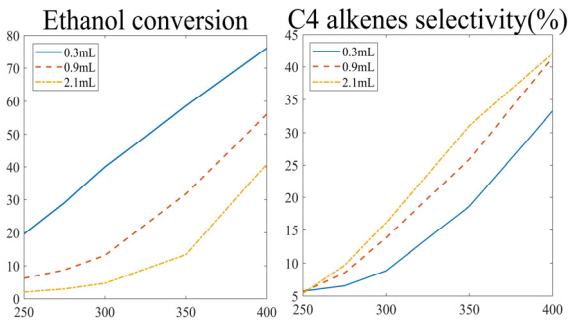

Figure 1: Change chart with ethanol milliliters

Other things being equal, the fewer millilitres of ethanol added per minute, the higher the ethanol conversion and the less selective the C4 alkene.

(2) The cobalt load is different, and other conditions are the same

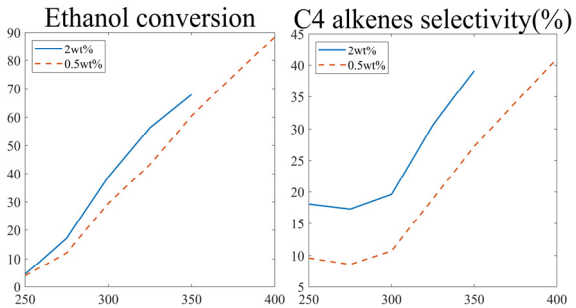

Figure 2: Change chart with cobalt load

The larger the cobalt loading, the higher the corresponding ethanol conversion and C4 alkene selectivity.

(3) The catalyst carriers are different and the other variables are the same

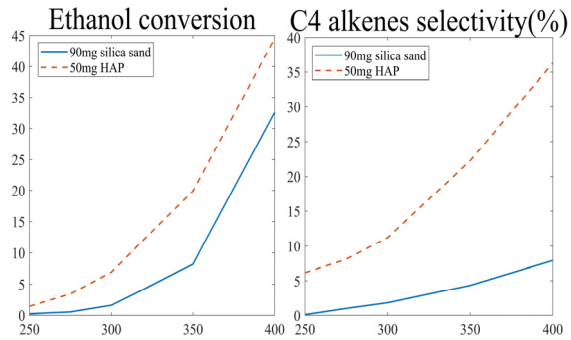

Figure 3: Change chart with catalyst carrier

Adding only HAP to the catalyst is better for the preparation of C4 alkenes than adding only quartz sand.

(4) Other variables remain the same, but the assembly method differs

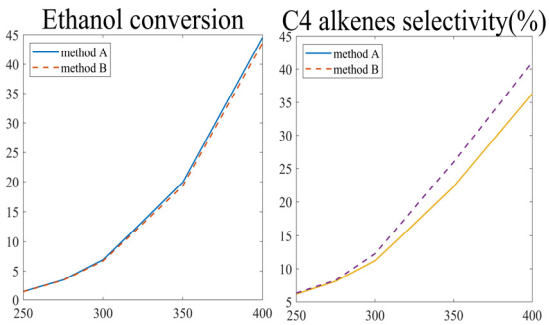

Figure 4: Change chart with assembly method

The effect of assembly method on ethanol conversion is less than the effect on the selectivity of C4 alkenes.

(5) Co/SiO2 and HAP charging ratios are different

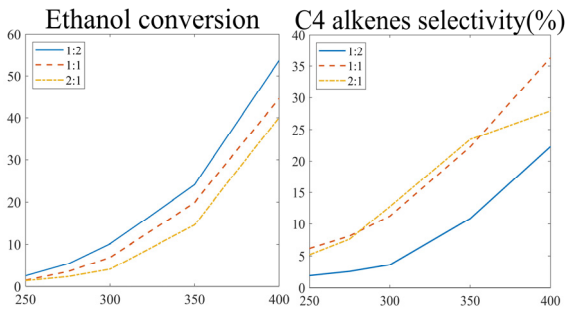

Figure 5: Change chart with charging ratios

To convert ethanol, the smaller the Co/SiO2 and HAP loading ratio, the higher the conversion rate.

(6) The change of ethanol conversion and C4 alkene selectivity over time at a given temperature

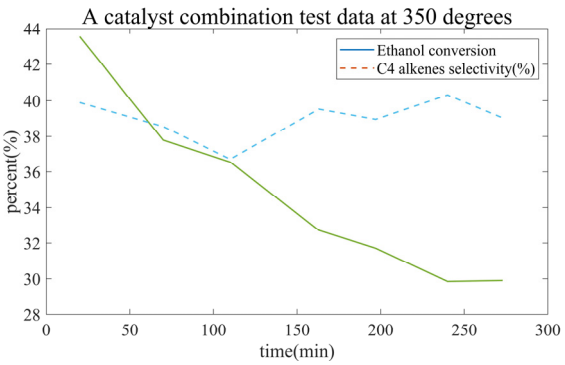

Figure 6: The temperature is constant 350°C

In the case of a combination of catalyst reactions at a given temperature, the longer the catalyst reacts in an ethanol atmosphere, the lower the conversion of ethanol, while the selectivity of C4 alkenes changes over time, reaching a maximum value at the right temperature.

Based on the presentation of the above six types of conditions and the analysis of the known data, the characteristic factors for the change of ethanol conversion and C4 alkene selectivity are summarized, namely the cobalt loading capacity in the catalyst combination, the Co/SiO2 and HAP charging ratios, the number of ethanol milliliters added per minute, whether quartz sand is added, the temperature, and the assembly method.

3.3 The selection of model methods

In this paper, the method of fitting prediction is used to expand the known small amount of experimental data, and then the optimal catalyst combination and temperature selection can make the solved results more accurate, and the specific method selection flow chart is as follows:

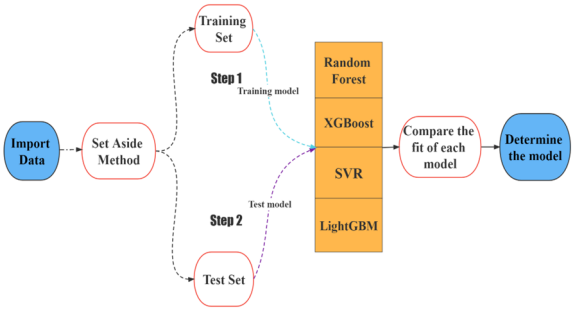

Figure 7: Model building method selection flowchart

The known data is divided into training sets and test sets, the training set data is substituted into the four models established according to the four machine learning algorithms to train the established models, and the Python programming solution is used to obtain the fit degree data obtained by the four models built by the four methods, and the fit degree data statistical table is as follows:

Table 2: Statistical table of fit degree data

Models	Degree of fit
SVR	0.7541
XGBoost	0.9161
random forest	0.8951
LightGBM	0.5705

It can be concluded that the model solved by *XGBoost* method has the highest degree of fit, which is 91.61%, so the *XGBoost* method is used to establish a prediction model for the optimal experimental conditions of ethanol-coupled preparation of C4 alkenes.

3.4 The optimal experimental condition prediction model is solved

According to the above analysis of the model method selection, it is concluded that the model established by the *XGBoost* method in this paper is the best effect on the problem solving, the first thing that needs to be done is to sort out the data required for the model to solve, because the *XGBoost* algorithm is one of the machine learning algorithms, there is *XGBoost*'s function library in Python software, and the establishment of the model for this article is mainly to combine the problem solving requirements with the *XGBoost* library for model establishment. The specific establishment situation is as follows.

The statistical chart of the specific prediction effect of the *XGBoost* method is as follows:

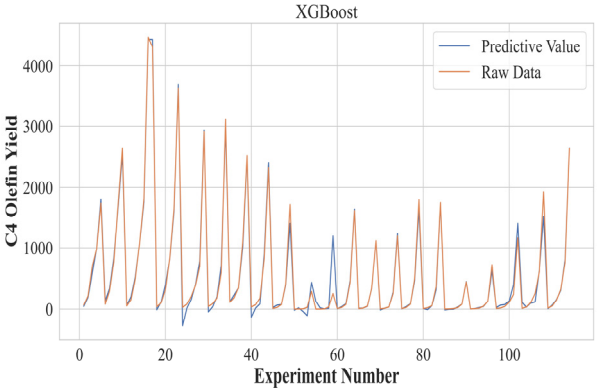

Figure 8: *XGBoost* method prediction plot

XGBoost is an additive model consisting of k base membrane vectors, assuming that the model of the tree we want to train for the t-iteration is $f_t(x_i)$, then there is:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{1}$$

The loss function of *XGBoost* is:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \tag{2}$$

where $\hat{y}_i$ is the predicted value, y_i is the true value and n is the sample size.

XGBoost's objective function:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \tag{3}$$

The second term in the above equation is added to the objective function to prevent the model from overfitting, and the first term represents the training error.

Apply Taylor's formula to approximate the loss function of the original *XGBoost*:

$$l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) = l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \tag{4}$$

The first derivative of the original loss function is g_i, while the second derivative is h_i, where the derivative is used to derive $\hat{y}_i^{(t-1)}$.

Take the squared loss function as an example:

$$l\left(y_i, \hat{y}_i^{(t-1)}\right) = l\left(y_i, \hat{y}_i^{(t-1)}\right)^2 \tag{5}$$

Then there is:

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} = -2\left(y_i - \hat{y}_i^{(t-1)}\right) \tag{6}$$

$$h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial\left(\hat{y}_i^{(t-1)}\right)^2} = 2 \quad (7)$$

Apply the above second-order expansion to the objective function of *XGBoost* to approximate the objective function:

$$Obj^{(t)} \approx \sum_{i=1}^{n} l\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) + cons\tan t \quad (8)$$

Apply Taylor's formula to approximate the loss function of the original *XGBoost*, and the final optimized objective function is:

$$Obj^{(t)} \approx \sum_{i=1}^{n} l\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \quad (9)$$

According to the addition model to obtain an overall model, through Python programming and calling the *XGBoost* library in Python, the data required for the solution in Annex I are sorted out, imported into the computer, and the degree of fitting of the prediction obtained by using the overall data is 97.46%, the prediction accuracy is high, and the optimal catalyst combination and temperature selection model based on the *XGBoost* algorithm can be obtained by testing it is reasonable and feasible.

The optimal experimental condition prediction model can obtain the highest experimental conditions of C4 alkene under different restrictions, and the statistical table of results is as follows:

Table 3: No temperature limit yield statistics table

T(°C)	C/H	CL	Etoh(ml)	C/S(mg)	H(mg)
400	1.0	1.0	0.9	200	200

Table 4: Statistical table of yield below 350°C

T(°C)	C/H	CL	Etoh(ml)	C/S(mg)	H(mg)
325	1.0	2.0	1.68	200	200

According to Table 3, the catalyst combination with the highest yield of C4 alkenes without temperature restrictions is 400 °C, the loading ratio of Co/SiO2 and HAP is 1:1, the Co loading amount is 1wt, the number of ethanol milliliters added dropwise per minute is 0.9ml, the mass of Co/SiO2 and HAP is 200mg, the ethanol conversion rate is 83.713%, the selectivity of C4 alkenes is 53.43%, and the C4 alkene yield is 44.72806%.

According to Table 4, the temperature limit under the condition of less than 350 °C makes the catalyst combination with the highest yield of C4 alkenes to the temperature condition is 325 °C, the Co/SiO2 and HAP charging ratio is 1:1, the Co load is 2wt, the number of ethanol milliliters added dropwise per minute is 1.68ml, the mass of Co/SiO2 and HAP is 200mg, the ethanol conversion is 56.382%, and the selectivity of C4 alkenes is 30.62%. C4 alkene yield was 17.26431%.

3.5 Optimize experimental group design

According to the prediction model of the optimal experimental conditions obtained from the previous solution, five sets of chemical experiments are added to the analysis of known data, and the chemical experiments are designed. The design of chemical experiments includes six elements: experimental purpose, experimental principle, experimental supplies and devices to be used, experimental operation and procedures, experimental results processing, and experimental procedures. The experimental design of this paper is to find out the optimal process conditions for C4 alkene yield under the experimental conditions not given in the known data, obtain the optimal optimized experimental group, and then obtain the auxiliary experimental design.

For the new experimental combination design, the method adopted in this paper is to use the optimal experimental condition prediction model combined with the known experimental data and the experimental situation solved by the analysis, the new impact factor summarized by the new impact factor situation to re-simulate the catalyst feature selection range, predict the new data, and obtain the five groups with the highest yield of C4 alkenes to design the experiment, the specific flow chart is:

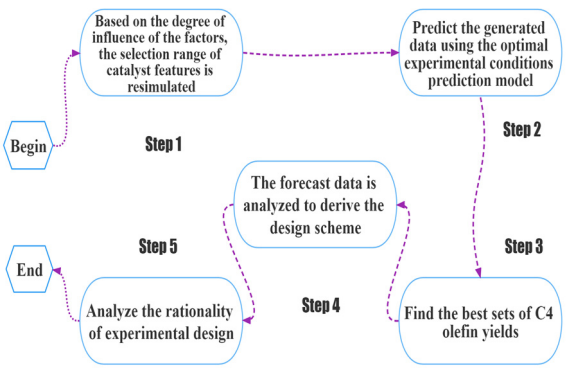

Figure 9: Optimize experimental group design flowcharts

The best sets of experiments with the best yield of C4 alkenes were obtained by Python programming, and the statistical table of the situation is as follows:

Table 5: Optimize the experimental group statistics table

T(°C)	C/H	CL	Etoh(ml)	C/S(mg)	H(mg)
375	1.0	1.0	0.9	150	150
375	1.0	1.0	0.9	200	200
400	1.0	1.0	0.9	150	150

400	1.0	1.0	0.9	200	200
425	1.0	1.0	0.9	150	150
425	1.0	1.0	0.9	200	200

Due to the small data in the experimental group, the model predicted the temperature and Co/Sio2 quality inaccurately, that is, the C4 alkene yield predicted by three different temperatures and different Co/Sio2 qualities was the same. Therefore, additional experiments are required to determine the optimal value of temperature and Co/Sio2 mass so that the C4 alkene yield can reach the maximum value. The experimental group of the third group in Table 5 is the same as the set of conditions in the known data, but the new data predicted by the optimal experimental condition prediction model that the C4 alkene yield is 42.804663% and the C4 alkene yield obtained by using the previously given data is 44.72806%, and there is an error in the two data. Therefore, the third set of data is used as the control group, and the remaining five groups are used as the experimental group for experimentation, that is, the five experiments are added to obtain the causes of true and accurate data analysis errors, the influence of errors is eliminated to determine the optimal value of temperature and Co/Sio2 quality, and the information obtained from the analysis of known data is analyzed according to the previous article, in order to make the experiment more accurate and reliable, the experiment needs to ensure that the reaction time is within the optimal time, and further explore the selection of catalyst combination and temperature.

4 CONCLUSIONS

The solution ideas and models established in this paper can be applied in physical or chemical experiments with relatively high cost, large risk coefficient or long test period, using a small amount of data to predict the experimental situation in other cases, deriving the experimental group that should be verified by the experiment, reducing the cost loss caused by repeated experiments, and at the same time being able to quickly find the most suitable experimental conditions.

REFERENCES

[1] Li F, Xu LJ, Zhu RB, et al. Demand prediction of shared bicycle borrowing based on XGBoost algorithm[J/OL]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition):1-10 [2021-08-04 17:52].

[2] Lu Shaopei. Preparation of butanol and C_4 alkenes by ethanol coupling [D]. Dalian University of Technology, 2018.

[3] Wang X, Liao B, Li M, Sun RNA. Fusion of LightGBM and SHAP for diabetes prediction and its characterization method [J/OL]. Small Microcomputer Systems: 1-11 [2021-09-06 18:16].

[4] ZHAO Rui, CHENG Xin, XU Xiaohui, SONG Tao, SUN Yuanlong. Early warning and control system of greenhouse diseases based on PSO-SVR model [J]. Journal of Jiangsu Agriculture, 2021, 37(04): 854-860.

[5] Zhao Y, Zhao Jiang Xuehui. Prediction of residual resistance of gliding boats based on random forest model [J/OL]. Journal of Huazhong University of Science and Technology (Natural Science Edition):1-6 [2021-09-10 10:29].