

Study on Influencing Factors of Ethanol Coupling to Prepare C4 Olefins

Longlong Li^{1,#}, Rui Hong^{2,#}, Chenxia Yang^{3,#}, Jiaqi Tao^{4,#}

¹Department of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu, 730070

²Shen School of Economics, Northwest Normal University, Lanzhou, Gansu, 730070

³College of Economics, Northwest Normal University University, Lanzhou, Gansu, 730070

⁴Department of Chemistry and Chemical Engineering, Northwest Normal University, Lanzhou, Gansu

2624127071@qq.com

[#]These authors contributed equally.

Abstract. In this paper, we study the data analysis and calculation of the factors influencing the preparation of C4 olefins by ethanol coupling. Specifically, by combining each catalyst, the relationship between temperature and ethanol conversion and C4 olefin selectivity was analyzed by regression analysis and correlation analysis, respectively. In addition, the grey relational degree analysis was used to solve the grey relational degree between the influencing factors of the experiment. The nonlinear correspondence between ethanol conversion and C4 olefin selectivity with different catalyst combinations and temperatures was evaluated by means of random forest model and BP neural network model. Finally, we build a support vector machine regression (SVMR) model and optimize the cost and gamma parameters of this model using a genetic algorithm.

Keywords: Regression analysis, machine learning algorithm, support vector machine regression (SVMR), C4 olefin, random forest model, BP neural network model.

1. Introduction

The traditional production methods of C4 olefins all rely on fossil energy, but with the development of society, the shortage of fossil energy and the environmental problems brought about by its conversion application become more and more obvious. It is necessary to find a green, sustainable and alternative new energy source. To prepare chemical industrial products such as C4 alkenes, ethanol has attracted the attention of researchers due to its wide source, easy conversion, and low pollution. During the preparation of C4 olefins, catalyst combination and temperature will have an impact on C4 olefins and yields.

As chemical raw materials, C4 olefins have high chemical research and pharmaceutical production value. Its traditional preparation method has problems such as shortage of raw materials and environmental pollution, while ethanol, as a clean and produced chemical reagent, can produce C4 olefins [1]. In this reaction, the combination of various catalysts (including: Co loading, Co/SiO₂ and HAP charging ratio, ethanol concentration) and temperature have an effect on the selectivity and yield of C4 olefins.

In the chemical experiments for the preparation of C4 olefins by ethanol coupling, the catalyst combination and temperature will have an impact on the selectivity of C4 olefins and the yield of C4 olefins. In this paper, by analyzing the relationship between catalyst combination, temperature and ethanol conversion and olefin selectivity, the maximum yield of C4 olefins under different conditions was obtained by analyzing the experimental data. In this paper, the relationship between temperature and ethanol conversion is firstly analyzed by regression, and the relationship between temperature and C4 olefin selectivity is analyzed. Secondly, the experimental data were analyzed with the help of machine learning algorithm to explore the effects of different catalyst combinations and temperatures on the ethanol conversion and C4 olefin selectivity. In addition, using machine learning algorithm, a

support vector machine regression (SVMR) model is established, and the cost and gamma parameters of the model are optimized using genetic algorithm. Finally, we investigated whether the high temperature would affect the activity of the catalyst and reduce the conversion of ethanol and the selectivity of C4 olefins and the effect of alkaline environment on product distribution.

2. Univariate Linear Regression Model

In this paper, Stata software was used to perform regression analysis on the relationship between temperature and ethanol conversion under different catalyst combinations, and the relationship between temperature and C4 olefin selectivity was analyzed.

For the study of the relationship between C4 olefin selectivity and temperature, this paper firstly carries out the correlation analysis of the corresponding data. The correlation coefficient is a quantitative indicator that describes the closeness of the correlation between two variables [2], and its calculation formula is:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (1)$$

where r represents the correlation coefficient, x and y represent the variables, respectively, $\bar{x}$ and $\bar{y}$ represent the mean values of the variables, respectively.

In the case of a certain combination of catalysts, the coefficient of determination of the linear regression model is greater than that of the nonlinear regression model, indicating that for this group of data, the linear fitting is better than the nonlinear fitting. Therefore, in view of the research problem of the relationship between temperature and ethanol conversion rate, the linear regression model is used, which can be expressed as:

$$y = ax + b + \mu \quad (2)$$

The P test of the linear regression model shows that its P value is always less than 10%, so the linear model is significant at the 10% level [3], and passed the significance test.

The curves of ethanol conversion and C4 olefin selectivity as a function of temperature were plotted using matlab software, as shown in Fig.1.

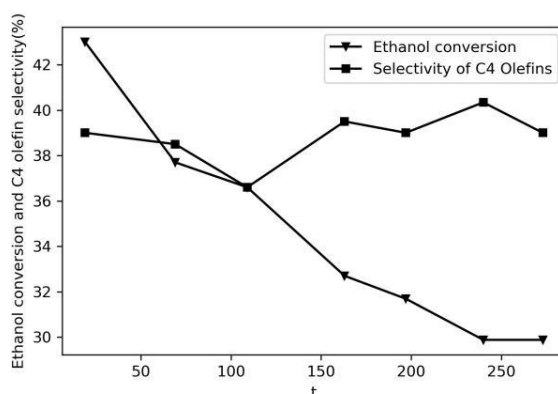


Figure 1. Ethanol conversion and C4 olefin selectivity as a function of temperature

It can be seen from Fig. 1 that with the progress of the reaction, the conversion rate of ethanol shows a downward trend, while the selectivity of C4 olefins has no fixed trend of change, generally showing a trend of first decreasing and then increasing.

We analyze the correlation between conversion rate and C4 olefins selectivity and time. The correlation coefficient between time and conversion rate is -0.9658795, which proves that there is a negative correlation between time and conversion rate of ethanol. However, the correlation coefficient between time and C4 olefin selectivity is 0.21544993, which cannot prove the linear correlation between the two factors due to its small correlation coefficient.

3. Machine Learning Algorithms

3.1. Grey Relational Degree

"Grey system" refers to a system in which some information is clear and some information is unknown, and the functional relationship of various factors in the system cannot be clearly defined. Therefore, in this paper, in the construction of the coupling preparation system of C4 olefins, the grey correlation comprehensive evaluation method is used to analyze the influence of different catalyst combinations and temperatures in the reaction process to highlight the strength of the influence factors of various countries in the process of ethanol coupling to prepare C4 olefins. Provide reference for the production of chemical products and pharmaceuticals.

The comprehensive evaluation method of grey correlation degree describes the relative change of each index factor in an information matrix, that is, the stronger the correlation, the greater the correlation degree. The comprehensive evaluation method based on grey relational degree is divided into the following steps [4]:

Step1: Determine the analysis sequence

Subsequences is the factors affecting system behavior, which can be expressed as:

$$C_i(j) = (C_i(1), C_i(2), \dots, C_i(m)) \quad (3)$$

where $i = 1, 2, 3, \dots, n$ is the number of evaluation modules, $j = 1, 2, 3, \dots, m$ is the number of indicators in each evaluation module, in this part, $n = 4$, $m = 109$.

Parent sequence reflects the behavioral characteristics of the system, which can be expressed as:

$$C_0(j) = (C_0(1), C_0(2), \dots, C_0(m)) \quad (4)$$

Step2: Dimensionless generalization of variables

Because the data of each factor column in the system may be different due to the dimension, it is inconvenient to compare or it is difficult to get the correct conclusion when comparing. Therefore, in the analysis of grey relational degree, it is generally necessary to carry out dimensionless processing of the data. Dimensionless can eliminate the difference in dimension and order of magnitude between indicators and ensure the comparability of indicators. It mainly includes two methods: initial value processing and mean value processing. Since the data involved in the analysis of this question are all greater than 0, this paper uses the mean processing method, and its expression is:

$$C'_i = \frac{C_i}{\text{mean}(C_i)} \quad (5)$$

The grey correlation coefficient is:

$$\xi_i(j) = \frac{\delta_{\min}|x'_0(j) - x'_i(j)| + \rho \delta_{\max}|x'_0(j) - x'_i(j)|}{|x'_0(j) - x'_i(j)| - \rho \delta_{\max}|x'_0(j) - x'_i(j)|} \quad (6)$$

where $\rho = 0.5$ is the resolution coefficient.

In the gray correlation degree evaluation system, the greater the correlation degree, the better the performance of the corresponding evaluation module, which is recorded as:

$$\gamma_i = \frac{1}{n} \sum_{j=1}^m \xi_i(j) \omega_j \quad (7)$$

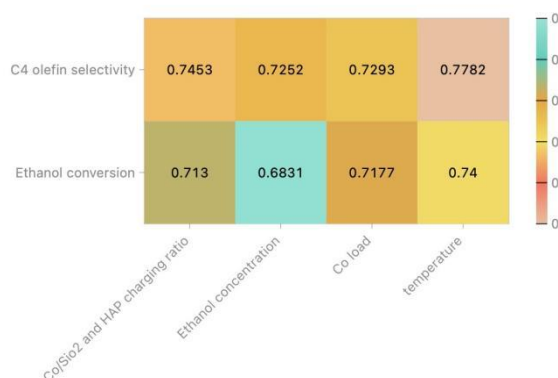


Figure 2. Grey correlation heat map

From Fig.2, it can be seen more intuitively that the area where (2, 3) is located is red, indicating that there is a strong correlation between these variables. In addition, other indicators and modules also have high correlations.

3.2. Random Forest Model

Random forest is an ensemble learning model based on the idea of bagging. It completes the learning task by constructing multiple learners [5]. The principle of the Bagging algorithm of this model is similar to voting to train a weak learner with one training set at a time. After n times of random sampling with replacement, n weak learners are trained according to different training sets. For classification problems, according to the voting of all weak learners, the "minority obeys the majority" principle is used to make the final prediction result. For regression problems, take the average of all learners as the final result. The statistical gain of the Gini coefficient is employed in decision trees [6]. The Gini coefficient indicates the degree of sample difference in the data set. The larger the Gini coefficient, the more diverse the data set is, that is, there are various classification results, and the more diverse the current characteristics of the data set, the more impure, that is, it may be a Multi-classification problem [7], which can be expressed as:

$$Gini(D) = 1 - \sum_{k=1}^{|y|} p_k^2 \quad (8)$$

Generally, the feature with the smallest Gini coefficient is selected as the initial root node. The specific formula is as follows:

$$Gini_{index}(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} Gini(D^v) \quad (9)$$

The gain of the Gini coefficient is:

$$Gini(D, a) = Gini(D) - Gini_{index}(D, a) \quad (10)$$

The random forest binary classification evaluation index is as follows:

The accuracy can be expressed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

The precision can be expressed as:

$$precision = \frac{TP}{TP+FP} \quad (12)$$

The recall can be expressed as:

$$recall = \frac{TP}{TP+FN} \quad (13)$$

The F1-score can be expressed as:

$$F_{\beta} = \frac{(1+\beta^2)*precision*recall}{\beta^2*precision+recall} \quad (14)$$

The 109 data sets were divided into training sets and validation sets, with the training set accounting for 0.9 and the validation set accounting for 0.1, and K-fold cross-validation was performed. Using Matlab to train the model, predict the validation set, and calculate the F1 value and AUC value. The calculation results are shown in Table 1.

From table 1, it can be seen that the random forest model is suitable for both ethanol conversion rate and C4 olefin selectivity.

Table 1. Calculation results

Model Name	Ethanol conversion		C4olefins selectivity	
	F1 value	AUC value	F1 value	AUC value
random forest model	0.9590	0.9580	0.9181	0.9600

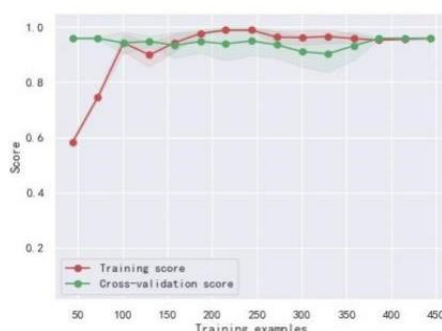


Figure 3. The training curve of ethanol conversion rate

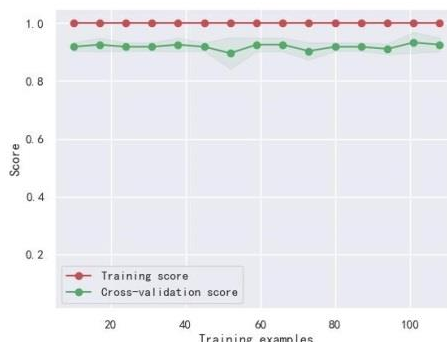


Figure 4. C4 olefin training curve

The ethanol conversion rate curve is shown in Fig. 3, and the C4 olefin selectivity curve is shown in Fig. 4. As can be seen from the Fig. 3 and Fig. 4, the number of decision trees and the maximum depth of decision trees are first determined by grid search, and then the minimum number of samples required for splitting and the leaf nodes carry the least number of samples, so it is quite intuitive to see that the performance of the random forest model is better, which has a great advantage over the linear model. And when running the random forest, it is found that the code runs faster. This is because the random forest is easy to be parallelized. When the random forest is trained, the trees are independent of each other.

3.3. Experimental Mining Model of Single-layer Perceptron

It can be seen from the literature that the algorithm principle of the single-layer perceptron network is the gradient descent operation, and the gradient search technology is used to minimize the mean square error of the actual output value of the network and the expected output value. It is especially suitable for solving problems with complex internal mechanisms [8]. The normalization method is the min-max normalization method:

$$x^* = \frac{x - \min}{\max - \min} \quad (15)$$

where μ is the mean of all sample data and σ is the standard deviation of all sample data.

The optimal number of neurons in the hidden layer is [9]:

$$N_h = \frac{N_s}{(a * (N_i + N_o))} \quad (16)$$

Among them, N_i is the number of input layer nodes, N_o is the number of output layer nodes, N_h is the number of hidden layer nodes, N_s is the number of training set samples, a is a constant between 2 and 10.

The constructed single-layer perceptron network model is shown in Fig. 5 [10].

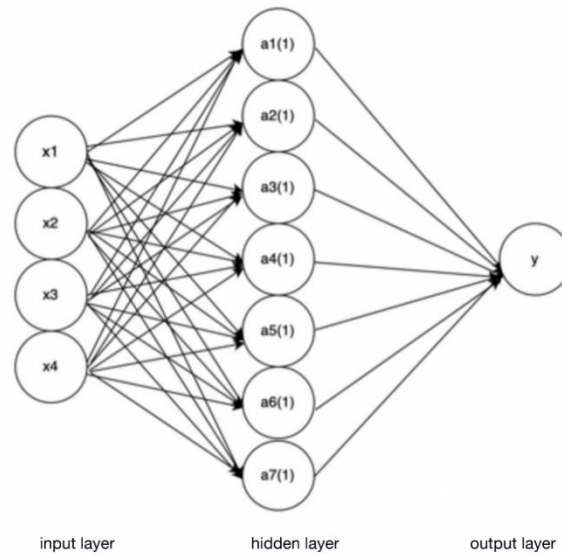


Figure 5. Structure diagram of single hidden layer neural network

The loss function uses the cross entropy loss:

$$J = -\frac{1}{m} \sum_{i=0}^m (y^{(i)} \log(a^{[2](i)}) + (1 - y^{(i)}) \log(1 - a^{[2](i)})) \quad (17)$$

The parameter update algorithm is the gradient descent method:

$$\theta = \theta - \alpha \frac{\partial J}{\partial \theta} \quad (18)$$

The training mode selected in this paper is L2 regularization. The regularization parameter is equivalent to adding the absolute value of the weight parameter directly to the parameter based on the original loss function, which is expressed as:

$$L = E_{in} + \lambda \sum_j |W_j| \quad (19)$$

After normalization, Matlab neural network toolbox is called to build a BP neural network model with 4 neurons in the input layer, 7 neurons in the hidden layer and one neuron in the output layer, which is shown in Fig. 6.

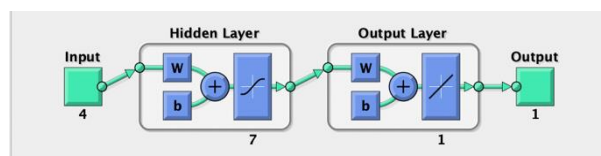


Figure 6. BP neural network model

Using the model predicted value obtained by training to compare with the real value, and calculate the error, and the obtained results are shown in Fig. 7 and Fig.8.

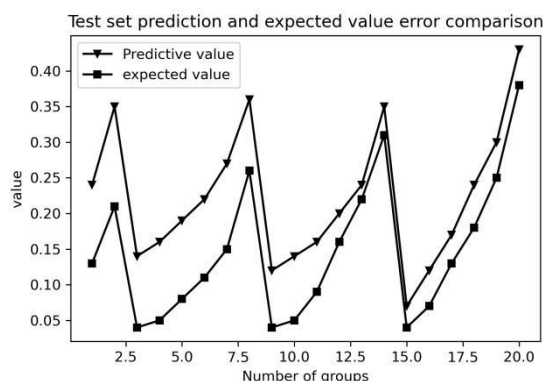


Figure 7. Ethanol conversion rate

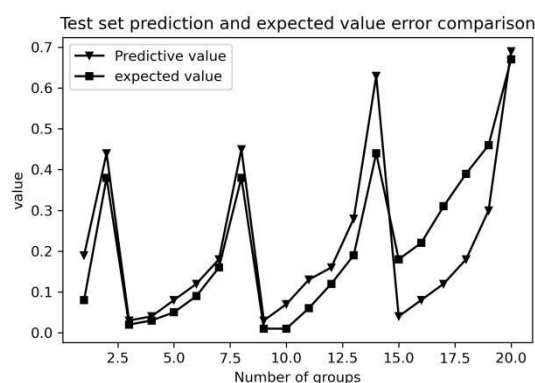


Figure 8. C4 olefin selectivity

3.4. SVMR Model

SVMR is a type of penalized learning that is applied in the field of functional regression. Its purpose is to simulate the regression relationship between the input x and the result y . To solve the nonlinear regression problem, the kernel mapping method is introduced, the variable x is mapped to a high-dimensional nonlinear space with the transformation function Φ , and the kernel function $K(x_i, x_j) = x_i^T K x_j = \Phi(x_i) * \Phi(x_j)$ is introduced to avoid the same feature. The inner product $\Phi(x_i) * \Phi(x_j)$ is calculated in space, and the nonlinear regression expression is finally obtained as:

$$y = f(x, a_i^+, a_i^-) = \sum_{i=1}^l (a_i^+ - a_i^-) K(x_i, x) + b \quad (20)$$

The Gaussian kernel function is expressed as:

$$k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \quad (21)$$

Fitting a nonlinear functional model through the SVMR model:

$$y = f(x1, x2, x3, x4, x5, x6) \quad (22)$$

Finding a set of $(x1, x2, x3, x4, x5, x6)$ to make the yield of C4 olefins as large as possible, then construct the optimal solution function:

$$MAX f(x1, x2, x3, x4, x4, x6) \quad (23)$$

The maximum value of each column of data is recorded as max_i , and the minimum value is recorded as min_i , then the constraints are:

$$Min_i \leq x1 \leq max_i \quad i = (1, 2, 3, 4, 5, 6) \quad (24)$$

The idea of controlling variables refers to a scientific experimental idea of controlling all variables that can cause changes in dependent variables except independent variables, and then clarifying the causal relationship [11].

First of all, the factors that affect the experimental results include catalyst combination and temperature, analyze the experimental conditions and results given in each group, and combine the control The idea of variables, comparing the existing experimental groups, the questions that can be explored are:

Comparing the experimental results of B2 and B7 catalyst combinations, it can be concluded that at the same temperature, the ethanol concentration is negatively correlated with ethanol conversion and C4 olefin selectivity.

Choosing A4, A6 or A9, A10 two groups of catalysts to compare the experimental results, it can be concluded that at the same temperature, the Co loading is negatively correlated with ethanol conversion and C4 olefin selectivity.

Comparing the experimental results of A13 and A14 catalyst combinations, it can be concluded that at the same temperature, the Co/SiO₂ and HAP charging ratios are negatively correlated with ethanol conversion and C4 olefin selectivity.

In the reaction process of preparing C4 olefins from ethanol, the catalyst plays a crucial role, and the temperature will have a certain influence on the effect of the catalyst, which in turn affects the selectivity of the target product. From the experimental data provided, it can be seen that under certain conditions of catalyst combination, with the increase of temperature, the conversion rate of ethanol and the selectivity of C4 olefins increase, but the data of A3 catalyst combination is analyzed, and it is found that at the temperature from 400 °C When the °C rises to 450 °C, the conversion rate of ethanol increases slowly, and the selectivity of C4 olefins decreases. Based on this, three groups of catalyst combinations A3 with the same temperature but at 500 °C, 550 °C and 600 °C are set up. Experiments were carried out, and the experimental results were combined with the A3 catalyst combination and the results at the corresponding temperature to explore whether too high temperature would affect the activity of the catalyst and reduce the conversion rate of ethanol and the selectivity of C4 olefins.

Ethanol is usually converted to hydrocarbon mixtures including ethylene, C₃-C₄ light olefins and C₅+ long-chain hydrocarbons through oligomerization-cracking reactions over zeolite catalysts. The product distribution is related to the reaction conditions (such as temperature), the acidity and alkalinity of the catalyst surface, and the water content of the raw materials. Co is dissolved in water to make the solution alkaline. The A11 catalyst combination at 400 °C is compared with the designed experiment to explore whether the different catalyst support materials will affect the experimental results. At the same time, the effect of alkaline environment on product distribution was investigated.

4. Conclusion

In this paper, the data analysis and calculation of the influencing factors of the preparation of C4 olefins by ethanol coupling method are carried out. The relationship between ethanol conversion, C4 olefin selectivity and temperature was studied. The effects of different catalyst combinations and different temperatures on ethanol conversion and C4 olefin selectivity were explored. However, using machine learning algorithm, a SVMR model was established, and the cost and gamma parameters of the model were optimized by genetic algorithm. Finally, the influence of catalyst combination and temperature on experimental data was explored.

References

- [1] L. S. Pei, Preparation of butanol and C₄ alkenes by ethanol coupling, Dalian University of Technology, 2018.
- [2] Z. M. Lin, Applied statistics. Second Edition, Tsinghua University Press, 2015.

- [3] P. S. Chu, *Econometrics* (Third Edition), Renmin University of China Press, 2009.
- [4] L. Tao, W. S. Yu, Research on my country's electricity market transaction evaluation system based on gray correlation degree and fuzzy comprehensive evaluation method, *Industrial Technology and Economics*, 2018, 37(09): 130-137.
- [5] G. Xuan, Research on movie box office prediction based on random forest, China University of Petroleum (Beijing), 2018.
- [6] Y. D. Ju, Y. Jing, Z. X. Juan, Feature selection algorithm based on random forest, *Journal of Jilin University (Engineering Edition)*, 2014 (01); 142-146.
- [7] L. X. Hai, Application of random forest model in classification and regression analysis, *Journal of Applied Entomology (Insect Knowledge)*, 2013, 50(004): 001190-1197.
- [8] L. Y. Sheng, X. Xin, et al., *Data Mining Principles and Application of Algorithms*, Beijing: Machinery Industry Press, 2014.
- [9] Y. J. Qi, L. Jia, S. S. Qian, Influencing factor analysis model and empirical research on the impact factor of sci-tech journals based on BP neural network, *China Market*, 2008(26): 103-104.
- [10] L. Ning, Analysis of influencing factors of the impact factors of management journals based on BP neural network, *Journal of Shaanxi Institute of Education*, 2007(03):83-86.
- [11] L. G. Zhou, The idea of osmotic control variables to cultivate chemical experimental ability, *New Education*, 2021(14):77-78.