Correlation Analysis Neural Network: C4-alkene Yield Prediction and Ethanol CR Evaluation

Zhihui Sun[1, *], Jiayu Li[1, a], Fan Liu[2, b], and Qian Xu[1, c]

[1] School of Mechatronics and Information Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China

[2] School of Science, China University of Mining and Technology-Beijing, Beijing 100083, China

[*]Corresponding author e-mail: 18832725855@163.com, [a] 1599964805@qq.com, [b]1599964805@qq.com, [c] xuqian_work@163.com

Abstract

C4-alkene is a type of polymeric material with a huge production capacity and a wide range of applications in the chemical industry. With the rising global demand for alkene, the existing industrial sources of C4-alkene are mostly from catalytic cracking at refineries and extraction from ethylene cracking reaction products, both of which rely on fossil fuels, therefore the production of C4-alkene from ethanol has a bright future. As a result, it is critical to investigate the process of catalytic coupling of ethanol to produce C4-alkene through the design of catalyst combinations. BP neural network is a multilayer feedforward neural network that constantly adjusts the parameters to converge to a reduced mean square error training model by backward transfer and error correction. In this paper, linear regression analysis was used to analyze the correlation and importance of each catalyst component, and then a multilayer feedback network model with different catalysts and temperatures was trained to obtain the optimized catalyst combination for producing the highest possible yield of C4-alkene.

Keywords

Correlation Analysis; Control Variates; Linear Regression; BP Neural Network.

1. Introduction

With the shortage of fossil energy production and the aggravation of environmental impact, the development of new energy sources has become increasingly urgent. Ethanol molecules can be prepared by fermentation, and the preparation of C4-alkene from ethanol has great application prospects. As a result, it is critical to investigate the process of catalytic coupling of ethanol to produce C4-alkene through the design of catalyst combinations. The control variate approach is extensively used in classical chemical analysis to examine the reaction ratios of chemical components, but it takes a large number of long-term trials to produce very reliable results. It is easy to generate experimental analysis errors in this scenario if the chemical analysis is performed with a large number of reaction variables.

BP neural network is a multilayer feedforward neural network, which is one of the most widely used neural network models that constantly adjusts the parameters to converge to a reduced mean square error training model via backward transfer and error correction. In this paper, a BP neural network is proposed for the predictive analysis model of C4-alkene yield and ethanol conversion rate under different catalyst combinations and temperatures. Firstly, the correlation analysis and significance analysis of each catalyst component were performed using linear regression analysis, based on the experimental results of the influence of different catalyst combinations and temperatures on ethanol conversion and C4-alkene selectivity. Then the multilayer feedback network model with different catalysts and temperatures was trained based on the significance analysis, and the results and data were integrated and input into the neural network. Finally the analysis and conversion prediction of chemical components were performed with the good fit of the neural network to the data to obtain the optimized catalyst combination.

2. Model Selection

2.1 Research Approach

2.1.1 Model Overview

The whole model is divided into two main steps: principal component analysis and neural network prediction, the flow chart of the model is as follows:

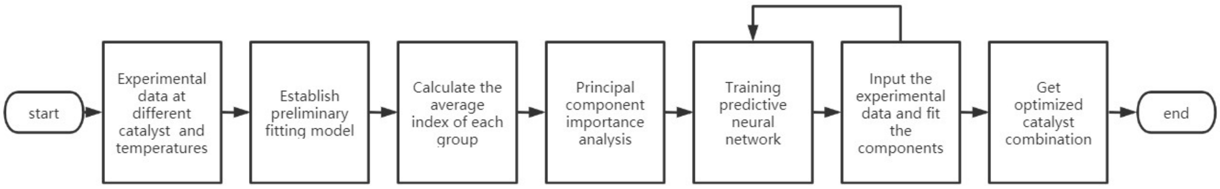
start
Experimental data at different catalyst and temperatures
Establish preliminary fitting model
Calculate the average index of each group
Principal component importance analysis
Training predictive neural network
Input the experimental data and fit the components
Get optimized catalyst combination
end

Figure 1. Flow chart of the model

We will evaluate each variable in the first step, since it is obviously unreasonable and very difficult to perform a direct evaluation analysis among many chemical variables, we will first perform an importance and correlation analysis, using the principal component analysis concept to rank the importance of each chemical variable, and this importance ranking will have an impact on the decision making in the second step.

In the second step, we'll use the results from the previous section to constrain the variables to make the neural network's input as concise and meaningful as possible, so that the neural network can better fit the relationship between the chemical variables and the reaction output value, and in the final output stage, we'll add a decision range to limit the size of the output to prevent too abnormal data.

2.1.2 First Step Model: Principal Component Analysis

First, correlation analysis was performed to establish a preliminary fitting model by scatter plot and two fitting methods, namely the effect of temperature on ethanol conversion and C4-alkene selectivity.

Based on the data analysis of the influence of temperature on ethanol conversion and selectivity of C4-alkene respectively. It was hypothesized that an exponential fit was performed between temperature and ethanol conversion in the given temperature interval, with all other factors held constant and unaffected by the results, and that least squares was used for the linear fit, converting the nonlinear to linear via equation transformation, and a quadratic linear regression was performed between temperature and C4-alkene selectivity.

The influence of each component on the results is then investigated based on the average indicators (R^2-statistic, F-value, error variance, p-value) for each group of catalysts that are obtained from the preliminary fitted model. All of the average R^2 are very close to 1, indicating a good fit of the fitted curve. However, for some samples where the data is already clear, R^2 can be a good indication of the correlation between y_i and $\hat{y}$, whereas because the data obtained from the test is subject to some chance, it cannot show that y and x must have some linear relationship.

Second, the variation pattern of the effect of individual independent variables on the dependent variable is derived by using the control variables method. The model is approximated as linear, and stepwise regression analysis is used to derive the weight and the significant effects of the single components, as well as the degree of importance and optimal value of each independent variable.

Although the stepwise regression analysis is linear, the significance analysis and R-squared in it still help us a lot in deciding the degree of significance, so we ignore the effect brought by nonlinearity in this paper and focus on the significance judgment.

2.1.3 Second Step Model: Neural Network Prediction

In previous chemical analyses, we commonly used the controlled variable method to analyze the reaction ratios of chemical components, but this takes a large number of long-term experiments to produce reliable results. On this premise, it is quite easy to generate experimental analysis errors if the chemical analysis is performed in the situation of a particularly large number of reaction variables.

Based on the study of [1][2], we propose a neural network for decision-making based on importance analysis to provide an alternative analytical idea for chemical analysis. In the study of [1], the analysis of propylene ontogeny productivity was performed by using neural networks with fuzzy calculations, while [2] used neural networks to analyze the distillation ratio control, both of which give us suggestions for the analysis of ethanol conversion and alkene selectivity. We combined the correlation analysis of linear regression and integrated the results with the data into a neural network, and used the good fit of the neural network to the data for chemical composition analysis and conversion prediction.

In contrast to traditional chemical analysis, neural network analysis offers good properties in both linear and nonlinear fitting. Another point is that neural networks can process data faster and learn correlation features that are difficult to see in the data. Although this information is hidden in neuronal calculations, the degree of chemical variable analysis by neural networks can be clearly seen through certain methods, such as visualization.

2.2 Formulas and Symbols Description

2.2.1 Exponential Fitting of Temperature to Ethanol Conversion

Let the dependent variable be y (y = Y1, Y2) and the independent variable be x (here represents X1, i.e. temperature), observe all the scatter plots from the experimental data, and conclude that the exponential fit is better, so we can set y=ae^{bx}, and take the logarithm on both sides at the same time to obtain the formula as follows:

$$ln\, y = ln\, a + bx \tag{1}$$

2.2.2 Quadratic Polynomial Fit of Temperature to C4-alkene Selectivity

$$y = ax^2 + bx + c \tag{2}$$

3. Analysis of Results

3.1 Data Processing and Environment

3.1.1 Data Processing

Based on the data in Annex I, conclusions and conjectures were generated by categorizing the applicable data through treatment comparisons. The degree of importance of each component was then evaluated independently by analyzing the degree of significance of a single component on the target component index through stepwise regression analysis, based on the preceding conclusions and hypotheses.

When dealing with a specific problem, some anomalous data are excluded to simplify the model and reduce the error in the results.

3.2 Analysis of Results

3.2.1 Analysis of Ethanol Conversion and C4-alkene Selectivity by Model

Table 1. Three Scheme comparing

Index Index influence and feeding method		Average R^2	Avetage F	Mean error variance	Average p
Effect of temperature on ethanol conversion	A	0.95	185.6980	0.0646	0.0032
	B	0.9394	145.7048	0.0352	0.0036
Temperature dependence on C4-alkene selectivity	A	0.9688	70.3947	8.0995	0.0233
	B	0.9873	332.4239	1.7030	0.0026

3.2.2 Model Evaluation of the Importance of Different Catalyst Combinations and Temperatures on Ethanol Conversion and C4-alkene Selectivity

As shown in the figure, the stepwise regression of temperature on C4-alkene selectivity is shown in Fig.2, while the stepwise regression of temperature on ethanol conversion is shown in Fig.3.

Table 2. Parameters that mainly affect the conversion of ethanol

Model	R	R^2	Adjusted R^2	Errors in standard estimates	Variation of R^2	Variation of F
1	$.700^{a}$	.490	.485	9.70861	.490	107.517
2	$.800^{b}$	.641	.634	8.18339	.151	46.640
3	$.837^{c}$	.700	.692	7.51056	.059	21.779

a) predictive variables:(constant), temperature.

b) predictive variables:(constant), temperature,HAP content.

c) predictive variables:(constant), temperature,HAP content,Co load capacity.

Table 3. Parameters that mainly affect the conversion of ethanol

Model	R	R^2	Adjusted R^2	Errors in standard estimates	Variation of R^2	Variation of F
1	$.771^{a}$	.594	.590	14.61559866	.594	163.738
2	$.872^{b}$	.760	.756	11.28755604	.166	76.781
3	$.892^{c}$	.795	.789	10.47584727	.035	18.868

a) predictive variables:(constant), temperature.

b) predictive variables:(constant), temperature,HAP content.

c) predictive variables:(constant), temperature,HAP content, ethanol concentration.

The data were analyzed as follows:

Table 4. Parameters that mainly affect alkene selectivity

Average index Index influence and feeding method		Fitting model	Mean error variance
Effect of temperature on ethanol conversion	A	Index model	0.0646
		Cubic polynomial fitting	1.428193077
		Power function fitting	2.2503
	B	Index model	0.0352
		Cubic polynomial fitting	0.514398571
		Power function fitting	0.986285714
Temperature dependence on C4-alkene selectivity	A	Quadratic polynomial fitting	8.0995
		Cubic polynomial fitting	0.619958462
		Power function fitting	2.186377
	B	Quadratic polynomial fitting	1.7030
		Cubic polynomial fitting	0.655492857
		Power function fitting	1.338457143

The results obtained from the model: the amount of change in R-squared and F owing to temperature added at the start was the greatest among the three groups of models, indicating that the effect of temperature on both was the greatest among all factors by significance analysis. The degree of variation in the R-squared was also shown to be more significant when HAP content was introduced in addition to temperature. The experimentally obtained 0.151 and 0.166 were both substantially higher than the widely used significance test of 0.05, indicating that HAP had a second only to temperature effect on the dependent variable.

After temperature and HAP, Fig.2 adds the index of Co loading. Since 0.059 is slightly greater than 0.05, the influence of Co loading on ethanol conversion is judged to be minimal and much smaller than that of HAP content and temperature. Similarly, ethanol concentration has almost no effect on alkene conversion, which also illustrates the accuracy of the assumptions established by the control variable method from the side.

3.2.3 Model Prediction of the Optimized Combination of Catalysts

The following outcomes are predicted by the model.

Table 5. Table of best conversion predictions for ethanol

	The first set of optimal solutions	The second group of optimal solutions	The third group of optimal solutions	The fourth group of optimal solutions
Co/SiO_2 content(mg)	68.58229225	61.31148249	71.28460064	67.05945846
Co load capacity(%)	1.785719026	1.51408165	2.155466206	1.818422294
HAP content(mg)	70.86293981	83.16767242	79.2355876	77.75539995
Co/SiO_2 / (Co/SiO_2 + HAP)	0.491822425	0.424362134	0.473588304	0.463257621
Ethanol concentration(ml/min)	0.591654308	0.463323732	0.583831297	0.546269779
Temperature (℃)	431.9532654	413.9116491	429.2134644	425.0261263
Ethanol conversion	87.20329455	81.65019725	80.48942622	83.11430601

Table 6. Table of best selectivity predictions for alkene

	The first set of optimal solutions	The second group of optimal solutions	The third group of optimal solutions	The fourth group of optimal solutions	Average optimal solution
Co/SiO_2 content(mg)	123.4955476	80.70971064	121.0561864	94.62644049	104.9719713
Co load capacity(%)	2.198092404	2.17631856	1.560115338	1.829653705	1.941045002
HAP content(mg)	82.46635041	86.21694797	87.55960754	84.80824237	85.26278707
Co/SiO_2 / (Co/ SiO_2 + HAP)	0.599603853	0.483504021	0.580282941	0.527358697	0.547687378
Ethanol concentration(ml/min)	1.140263523	1.309152478	1.221115144	1.119048205	1.197394837
Temperature (℃)	439.5260536	449.1293432	448.3447394	406.0456486	435.7614462
Ethanol conversion	45.6323585	41.7T060189	50.35490532	37.05793299	43.70394967

3.2.4 Final Prediction Results of the Model

In terms of the optimal solutions for ethanol conversion and C4-alkene selectivity, Tables 3 and 4 show that the values of the four primary constrained independent variables are essentially the same. Furthermore, the values of the optimal solutions for the four main parameters of the two dependent variables are approximately equal in the equation C4-alkene yield = ethanol conversion * C4 selectivity, so this paper concludes that optimal solutions will exist when the four parameters are in the range interval corresponding to as in Table 4. And because there is no existing catalyst combination that meets the aforementioned criteria, this paper will be compared to the findings of other researchers' experiments.

Table 7. Optimal range of values

parameter	Temperature (℃)	HAP content(mg)	Co/SiO_2 loading ratio	Co / (Co + HAP)(%)
section	415~425	75~90	about 2	About 0.5

Compare with the experimental results obtained by Shaopei Lü [3], whose experimental results [3] are shown below:

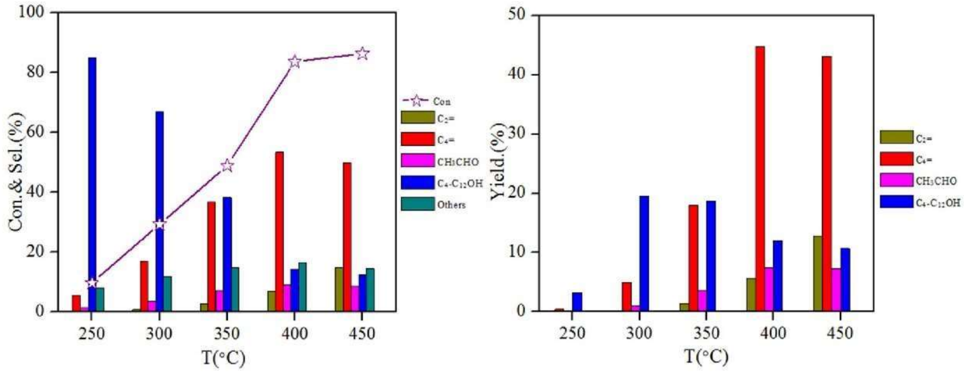

Figure 2. Histogram of catalyst impact (a,b)

It is clear to see that 400 is the optimal solution among the temperatures it contains, but 450 also performs well. He also concludes in his analysis of the paper that the catalyst functions best when the Co/SiO2 content to HAP ratio is 1:1. Combining the predictions of this paper with the analysis of the experimental results, i.e. the catalyst combinations already listed in Annex 1, we can exclude the effect of HAP mass. Since the data on HAP content = 75 mg is insufficient, but Co content/(Co content + HAP content) indirectly contains HAP information, HAP content will be excluded and the other three main parameters will be compared. When the temperature approached 400°C, we found that group A4 had the highest catalytic efficiency and good chemical reaction. Our chosen group A4 ranked third-best C4-alkene yield in Excel when compared to other data, and the difference with the previous data was tiny, so we assumed the analysis was more accurate.

When the temperature is set to 350°C:

As in the preceding sub-question, based on the correlation study of Figs. 4.a and 4.b, we believe that the higher the temperature possible, the better. As a result, we only consider the case in which the temperature rises to 350°C while the other three main parameters remain constant. Because of the small amount of data about HAP=75mg among the known catalysts, we'll exclude it in this article and focus on the remaining main parameters as in the prior question. After excluding the catalyst parameters at temperatures greater than 350°C, we find that the A2 catalyst produces the first C4-alkene yield among all catalysts at 350°C, which is a very accurate prediction.

4. Conclusion

The improved BP neural network presented in this research is highly accurate and can explain most of the data well, as well as perform some predictive analyses of the effect of several unknown catalyst groups and temperature on the dependent variable. In comparison to the classic neural network algorithm, the model's stability is improved, and the algorithm in this paper combines a stepwise regression algorithm, which employs the data gained via stepwise regression and control variables to constrain the results generated by the neural network, improving the accuracy of the algorithm.

This prediction model still has some defects, such as the aforementioned optimal prediction results and optimal parameter ranges can be learned from the data, so the next step of the model should be to strengthen its learning ability so that it can be spontaneously filtered instead of artificially limiting the outcomes. In addition, only four independent variables were considered when the original BP neural network was used for predictive analysis based on conditional decision making: Co load, HAP mass, ethanol concentration, and temperature. Two additional independent variables were added as part of the improvement: Co mass and Co mass/(Co mass + HAP mass). The addition of parameters can also help enhance the model's accuracy while using the same quantity of data.

References

[1] Ezzatzatzadegan L . Neuro fuzzy modeling of propylene polymerization[J]. Tp Chemical Technology, 2011.

[2] J, Fernandez, de, et al. Dual composition control and soft estimation for a pilot distillation column using a neurogenetic design - ScienceDirect[J]. Computers & Chemical Engineering, 2012, 40(1):157-170.

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