Proceedings of the 2nd International Conference on Green Energy, Environment and Sustainable Development (GEESD2021), D. Dobrotă and C. Cheng (Eds.)

doi:10.3233/ATDE210339

Key Feature Selecting in the Clean Oil Refinery Process Based on a Two-Stage Data Mining Framework

X LIN[a], J Y MA[b†], F X LIN[a], D Q WANG[c] and X S XIAO[d,1]
[a] *School of Management, Zhejiang University, Hangzhou 310058, China*
[b] *College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*
[c] *College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China*
[d] *Bigdata Technology Institute, Guizhou Light Industry Polytechnic College, Guiyang 550025, China*

Abstract. Maintaining the ratio of octane number and reducing the proportion of harmful substances in the heavy oil fluid catalytic cracking process meets both environmental and economic benefits. Through collecting tremendous processing data by digital hardware, gasoline refiners are still hard to do well data analytical work in production process control due to the large scale of ambiguous intermediate operating variables. This paper proposes a two-stage data mining framework integrates the strengths of Ridge regression and Person correlation analysis to extract a scale limited group of key features. Different with traditional recursive feature elimination methods, we pay more attention to the correlation analysis between every couple of features in the result. Two stop criterions guarantee to fulfil refining standards and limit the computational work in finite steps. A real word case study contains 325 samples, 13 quality indicators and 354 operating variables which testifies the validity and practicality of our algorithm. The result shows only 13 features (operating variables) are significant to the rationality of process design and the improvement of process control.

Keywords. Clean oil refinery, two-stage data mining framework, key features

1. Introduction

Since the introduction of the Paris Climate Agreement and other environment protection documents, governments join hands with related guilds to state some occupation standards, for cutting emission in some industrial processes. To reduce city emission generated by petrol cars and gasoline vehicles, fuel admission standards were updated stricter year by year: China national standard of clean gasoline had reduced sulphur index nearly 15 times in its version III than version IV since 2010 to 2019 [1]; Europe standards also stipulated a low-level pollution index in a long term [2].

[1] Corresponding Author, X S XIAO, Bigdata Technology Institute, Guizhou Light Industry Polytechnic College, Guiyang 550025, China; Email: shiauxsh@126.com.
† Co-first Author

While paying attention to these environmental constraints, the maintenance of economic benefit lies also important. Therefore, a gasoline purity measurement indicator, namely the octane number (RON) [3], should be well-keeping in the heavy oil fluid catalytic cracking process (HO-FCC). The higher RON in a gasoline sample is, the greater dielectric constant is, indeed the higher combustion and energy utilization efficiency is and the lower carbon emission is [4]. However, some environmentally friendly processes such as part of detail steps in desulfurization would not totally positive to the RON keeping. Fortunately, with the development of digital industrial monitoring technology, oil refining enterprises have the ability to record and store tremendous dynamic changing data of all the intermediate operating variables during the HO-FCC process. Thus, it is possible to design a data mining framework to extract limited representative, independent, economic positive and environmentally friendly key features in a large indicator pool. This leads to a foundational work for finding, predicting and evaluating HO-FCC refined control solutions.

This paper proposes a two-stage data mining framework, which integrates ridge regression recursive feature elimination method (Ridge-RFE) [5] and Pearson correlation analysis method (PCC) [6], to better avoid correlation and strengthen representation between each selected key indicator in its solution. The rest of this paper is organized as follows: Section 2 reviews the related literature. Section 3 describes the feature selection problem. Section 4 designs two-stage solving framework. The numerical results of a case study using a real-word data from an oil refiner in China are reported in Section 5. Conclusions and future research directions are outlined in Section 6.

2. Literature Review

In order to do this interdisciplinary research work, we should review the literatures in two parts as follow: (1) HO-FCC refined process control in chemical engineering; (2) Feature selection in data science.

2.1. HO-FCC Refined Process Control

Early HO-FCC process control researches focused more on a posterior analysis of gasoline samples in using various analytical techniques, such as fluorescent indicator adsorption (FIA) [7], Near infra-red (NIR) [8], Nuclear magnetic resonance (NMR) [9], e.g., to obtain gasoline compositional data and evaluate samples' quality. Due to the requirement of fast process design, predicting samples' quality in new process based on historical posterior sample data has become frontiers recently. Guan et al. [4] used dielectric spectroscopy (DES) in association with partial least squares (PLS) multivariate calibration method to predict the RON of clean gasoline samples. Nikolaou et al. [10] proposed a non-linear calculation method to predict the RON based on gas chromatographic data. Wang et al. [11] presented an adaptive sample space expansion approach (ASSEA) to improve the prediction accuracy in real-time gasoline quality measurement. However, limited works joint analysed the tremendous intermediate operating variables in HO-FCC process. This paper selects key features in this variables' pool by data mining and makes preparation to develop a data-driven algorithm for gasoline quality prediction.

2.2. Feature Selection

For data-driven models, feature selection method can process and respond to high-dimensional data quickly, reduce the time of data calculation and improve the performance of the prediction model [12]. The most practical method is recursive feature elimination (RFE) [13], which selects the optimal features by some regression techniques: Gauraha [14] established a high-dimensional linear regression model with LASSO to quickly reduce the dimension of variables Kibria [15] testified the superiority for ridge regression with other regression methods in multi-collinearity problems. Chen et al. [16] introduced the support vector regression (SVR) in RFE when the indicators were powerful nonlinear to labels. However, these researches unconsidered the correlation analysis among all the selected features. This paper integrated Pearson correlation analysis (PCA) and design a reasonable stop criterion to avoid selecting similar features.

3. Key Feature Selecting Problem in HO-FCC

Intermediate operating variable that oils refinery is able to identify and record in HO-FCC is denoted as i such as ratio of hydrogen to oil, reducer pressure, etc., and the variables' pool is denoted as set I , $i \in I$. Gasoline quality indicator is denoted as j such as RON, sulphur content, etc., and the indicators' pool is denoted as set J , $j \in J$. Gasoline sample in real time sampling is denoted as k , $k \in K$. The database record could be symbolized as follow:

- Process record x_{ik} : value of intermediate operating variable i in gasoline sample k ;
- Quality record y_{jk} : value of quality indicator j in gasoline sample k .

Figure 1 shows the key feature selecting process in HO-FCC. The scientific problem is to design an efficient framework to extract a group of key features in fulfillment with both the appropriate group size and the correlation constraint between each feature in it.

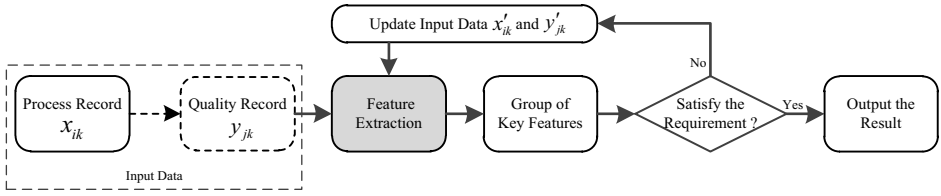

Figure 1. The key feature selecting process in HO-FCC.

4. Two-Stage Data Mining Framework

To fulfil the requirement in section 3, we propose a two-stage feature selection algorithm in figure 2. The importance of each feature is output by Ridge-REF and the correlation analysis result is output by PCA. This paper finds a specific stop criterion to extract feature in both independent and representative Some details steps are listed as follow:

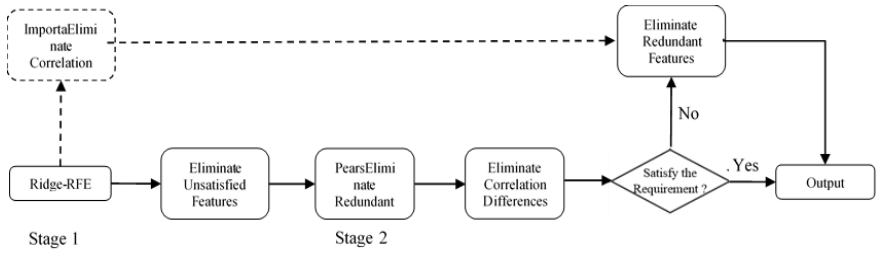

Figure 2. The two-stage feature culling algorithm.

4.1. Data Preprocessing

- For the loci containing only part of the time point, if the incomplete data is too large to be supplemented, such loci should be deleted;
- Delete the sites with empty data in the sample;
- For some sites with null data, the null value is replaced by the average data of two hours before and after the null value;
- In the data of a certain column of operating variables, if the data exceeding the operating range exceeds 10% of the total data of the column, the operating variable will be considered abnormal and the operating variable will be eliminated;
- If there is still any sample data beyond the operating range in all the remaining sample data, the sample will be considered abnormal and the sample data in this row will be excluded.

4.2. First Stage: Ridge Regression in Recursive Feature Elimination

The recursive feature elimination of the main idea is through repeated building ridge regression model of training for many times, each time after training according to the characteristics of the lowest weight coefficient to remove weight, then redo the above process, the remaining features until the traverse all features, has been removed in the process of the whole order is characteristic of sorting.

The stability of RFE largely depends on the underlying model during iteration. RFE based on linear regression has not been regularized, and the model is unstable at this time. The so-called regularization refers to the addition of a constraint term after the regression model cost function. The sum of squares of all parameters of the constraint term adopted by Ridge regression is L2 norm. The regression model based on Ridge regularization is stable, and Ridge regression is more suitable for the situation that there is multi-collinearity between independent variables or the number of independent variables is more than the sample size. The details of ridge-RFE algorithm are shown in table 1.

Table 1. The ridge-RFE model algorithm introduction.

Input:	Training variables $\mathbf{X_0} = [\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_k, \ldots \mathbf{x}_l]^T$
	Category labels $\mathbf{y} = [y_1, y_2, \ldots y_k, \ldots y_l]^T$
Initialization:	Character subset $\mathbf{s} = [1, 2, \ldots n]$
	Feature ranking list $\mathbf{q} = [\ ]$
Start:	Limit the training variable to the required feature set $\mathbf{X} = \mathbf{X}_0(:, \mathbf{s})$
	Use the ridge to return to training $\alpha = Ridge - train(\mathbf{X}, \mathbf{y})$
	Calculate the weight of each vector dimension $w = \sum_k \alpha_k \mathbf{y}_k \mathbf{x}_k$
	Calculate sorting index $c_i = (w_i)^2$
	Find the feature with the lowest ranking value $f = \arg min(\mathbf{c})$
	Update the feature sorting list $\mathbf{q} = [\mathbf{s}(f), \mathbf{q}]$
	Removes the characteristics of sorted values $\mathbf{s} = \mathbf{s}(1 : f-1, f+1 : length(s))$
Output:	Feature ordering list $\mathbf{q}$

4.3. Second Stage: Pearson Correlation Analysis

Pearson correlation analysis can well measure the linear correlation between two variables. The calculation formula of correlation coefficient is:

$$r_{i,j} = \frac{\sum_{k=1}^{n}(x_{i,k} - \frac{1}{n}\sum_{k=1}^{n} x_{i,k})(x_{j,k} - \frac{1}{n}\sum_{k=1}^{n} x_{j,k})}{\sqrt{\sum_{k=1}^{n}(x_{i,k} - \frac{1}{n}\sum_{k=1}^{n} x_{i,k})^2}\sqrt{\sum_{k=1}^{n}(x_{j,k} - \frac{1}{n}\sum_{k=1}^{n} x_{j,k})^2}}, \forall i, j \in I$$

where, $r_{i,j}$ represents the correlation coefficient between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Generally, correlation can be divided into three levels: the low linear correlation when $|r| < 0.4$; the significance correlation when $0.4 \le |r| < 0.7$; the high linear correlation when $0.7 \le |r| < 1$.

4.4. Stop Criterion

- Criterion 1: all PCA values are smaller than a threshold: $r_{i,i'} < \varepsilon, i \ne i', \forall i \in I, \forall i' \in I$;
- Criterion 2: the number of key features should less than $0.1|I|$.

We delete related feature when it breaks criterion 1 and then select no more than first $0.1|I|$ features in the rest feature group according to the Ridge-RFE ranking results.

5. Case Study

5.1. Data

This paper uses Python to conduct simulation experiment on the model, among which 325 samples are collected from Gaoqiao Petrochemical of Sinopec Time database (Honeywell PHD) and LIMS experimental database. Each sample includes 13 quality indicators and 354 operating variables.

5.2. Solution

In the part of data processing, we delete 20 samples and 4 operating variables. We use 11 quality indicators in three major types (raw materials, raw adsorbents and regenerated adsorbents) in modelling due to their inherent properties of gasoline refining process and decisive effect on RON. Figure 3 shows the 21 features PCA result after Ridge-RFE solving and figure 4 shows the 13 features PCA result when eliminating criterion 1 broken items. Note, red block represents positive correlation and blue represents negative correlation; The darker colour is, the greater correlation is. Compared with these two figures, we find that the colour of whole map is significantly lighter. This phenomenon can well testify the validity in extracting both independent and representative features of our algorithm. As $13 < 0.1 \times 354$, criterion 2 is not triggered. Thus, the 13 features become the final result and are listed in table 2.

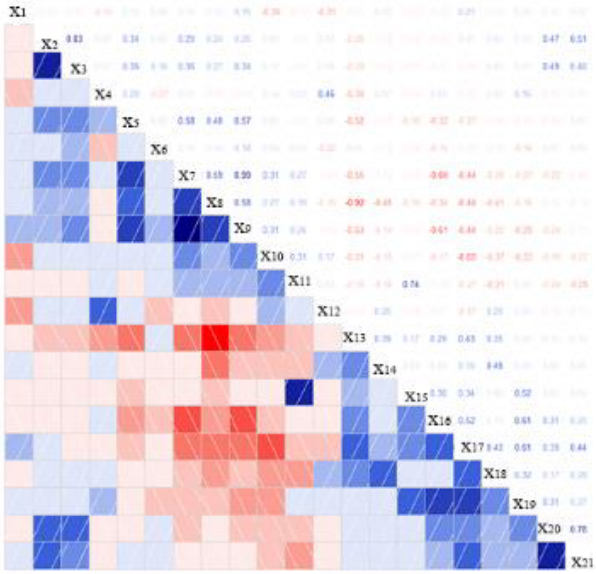

Figure 3. 21 features obtained by Ridge-RFE.

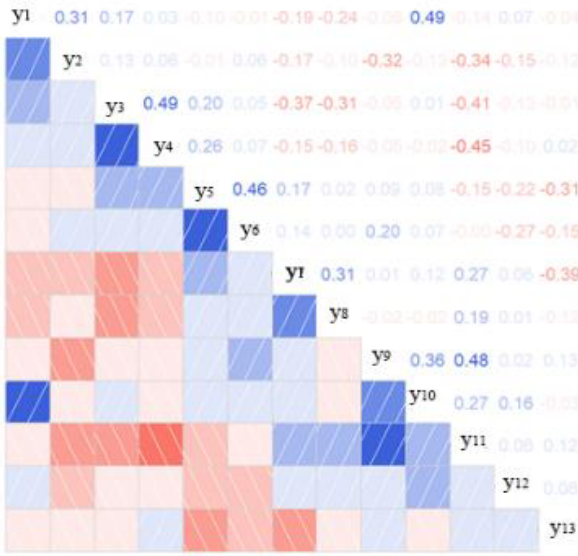

Figure 4. 13 features obtained by Pearson correlation analysis.

Table 2. The remaining major variables after Pearson correlation analysis.

Sequence number	Column number	Serial number	Name
1	FI	S-ZORB.PT_2905.DACA	D-109 pressure
2	AW	S-ZORB.PT_9403.PV	Nitrogen inlet pressure
3	AE	S-ZORB.PC_5101.PV	Stabilize tower top pressure
4	FQ	S-ZORB.PT_2502.DACA	D-107 bottom pressure
5	HS	S-ZORB.PT_6003.DACA	Blower inlet pressure
6	C	Raw material properties	Sulphur content, μg/g
7	D		Octane number, RON
8	E		Saturated hydrocarbon, v% (alkane + cycloalkane)
9	F		Aromatic hydrocarbons, v%
10	G		"Bromine value, gBr/100g"
11	H		"Density (20°C), kg/m³"
12	M	Preparation of adsorption properties	Coke, wt%
13	N		S, wt%

6. Conclusion

In HO-FCC process, reducing harmful substances such as sulfur compounds while maintaining RON levels is a challenge work in process design and control. Digital production facilities provide tremendous data to engineering but large scale of intermediate operating variables makes data analytical difficult. However, traditional RFE method in feature engineering could well reduce feature scale but ignore the correlation between each couple of features. This would influence both the forecast precision of product quality and control precision of production process. This paper proposes a two-stage data mining framework which integrated the strengths of Ridge

regression and Person correlation analysis to solve this problem. Case study testifies the validity and practicality of it.

Future works will consider the relationship between feature engineering and quality forecast learning. We hope this work can provide reference for the operation management in oil refining enterprises related to economic benefit planning and production emission control.

Acknowledgments

This research is supported in part by the Key R&D Program of Zhejiang Province (No. 2021C01104), the National Training Program of Innovation and Entrepreneurship for Undergraduates of China (No. 202010335011) and the Provincial Fundamental Researches Plan of Guizhou (No. [2019]20013).

References

[1] Mao J, Zou Y, Zhang Z, et al. 2017 A brief analysis of the revision of GB 17930-2016 "Motor Gasoline" standard *Journal of China Standardization* **261** 21-24.
[2] Group REM 2016 Euro-5 gasoline *Russian Energy Monthly.*
[3] David S 2002 Octane and the environment *Science of the Total Environment* **299** 37-56.
[4] Guan L, Feng X L, Li Z C, et al. 2009 Determination of octane numbers for clean gasoline using dielectric spectroscopy *Fuel* **88** 1453-1459.
[5] Li F and Yang Y 2005 Analysis of recursive feature elimination methods *International ACM SIGIR Conference on Research & Development in Information Retrieval* pp 633-634.
[6] Coefficient P 1996 Pearson's correlation coefficient *New Zealand Medical Journal* **109** 38.
[7] ASTM D 1319-Hydrocarbon types in liquid petroleum products by fluorescent indicator adsorption.
[8] Rohrback B 1991 Computer-assisted rating of gasoline octane *Trends Annual Chemistry* **10.**
[9] Meusinger R and Moros R 1995 Application of genetic algorithms and neural networks in the analysis of multi-component mixtures using NMR spectroscopy.
[10] Nikolaou N, Papadopoulos C, Gaglias I, et al. 2004 A new non-linear calculation method of isomerisation gasoline research octane number based on gas chromatographic data *Fuel* **45** 247.
[11] Wang K, He K, Du W, et al. 2021 Novel adaptive sample space expansion approach of NIR model for in-situ measurement of gasoline octane number in online gasoline blending processes *Chemical Engineering Science* 242.
[12] Chandrashekar G and Sahin F 2014 A survey on feature selection methods *Computers & Electrical Engineering* **40** 16-28.
[13] Guyon I and Elisseeff A 2003 An introduction to variable and feature selection *Journal of machine learning research* **3** 1157-1182.
[14] Gauraha N 2018 Introduction to the LASSO *Resonance* **23** 439-464.
[15] Kibria B and Banik S 2016 Some ridge regression estimators and their performances *Journal of Modern Applied Statistical Methods* **15** 206-238.
[16] Chen Y, Xu P, Chu Y, Li W, Wu Y, Ni L and Wang K 2017 Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings *Applied Energy* **195** 659-670.