# Research on Big data Preprocessing Technology of Thermal System

Yi-Fan Zhao and Zhong-Guang Fu
*Key Laboratory of Condition Monitoring and Control for Power Plant Equipment, North China Electric Power University, Beijing, Changping district, China*
*E-mail: hbdl_zyf@126.com*

Fei Chen
*School of Control and Computer Engineering, North China Electric Power University Beijing, Changping district, China*

There is a big amount of overlap among the data of thermal system, which will seriously affect the accuracy and precision of the model. In order to improve the value density of big data of thermal system and improve the quality and efficiency of modeling, pretreatment and analysis have to be made. Aimed at the diagnosis and treatment of the multiple co-linearity among variables, three reduction models have been established, respectively by Pearson correlation coefficient diagnostic method, VIF auxiliary regression test and condition number diagnosis method. And the models have been validated by the data of a 600MW unit of a power plant. Through the analysis of the results, it is found that the condition number diagnosis method can effectively solve the problem of multiple co-linearity of big data of thermal system, realizing the pretreatment of big data of thermal system.

*Keywords*: Big data of thermal system; Pretreatment; Pearson correlation coefficient; VIF; Condition number

## 1. Introduction

The big data platform of thermodynamic system brings together the massive heat information collected by the power plant [1]. By analyzing the potential law in the data, it can predict the operation of power plant equipment and system, and help the staff to make right decisions in time [2], so as to improve the operational efficiency, achieving greater revenue.

But when analyzing the data, the quality of the data will directly affect the quality of the results [3]. The information overlapping of big data has a serious impact on the efficiency and quality of the modeling. Therefore, data

preprocessing is essential. Attribute reduction is a key link in the process of big data preprocessing [4]. By reducing the attributes of big data, it can greatly reduce the amount of data processing in the process of big data analysis, and improve the efficiency of data processing.

Aiming at the diagnosis and treatment of multiple co-linearity among variables, Pearson correlation coefficient diagnosis method, VIF auxiliary regression test method and condition number diagnosis method were used for modeling. And the effect of pretreatment had been verified by the data of a power plant 600MW unit. A method which is more suitable for the big data preprocessing of thermal power system would be selected after comprehensive comparison.

## 2. Multiple Co-Linearity

In the process of big data modeling, the selection of relevant variables is a key step in the process of big data modeling [5]. Accurate characteristic parameters can effectively guarantee the accuracy of the model, reduce the modeling time, and reduce the complexity of the model. A large number of operating parameters need to be monitored in power plants in order to ensure safety, but many of the parameters are not independent, among which exist a lot of information overlap.

In the modeling process, researchers tend to choose variables with experience and some existing theories. But because of the complexity of the problem itself, so many factors involved, and coupled with the limitations of the understanding level of the researchers, the approximate linear relationship between variables (i.e., multiple co-linearity) is difficult to avoid. Once the multiple co-linearity is confirmed, it will bring a series of design difficulties and design defects, and can also lead to the model structure unreasonable and design redundancy, resulting in a reduction in simulation accuracy and speed.

Multiple co-linearity means that the variables $X_1$,$X_2$, ⋯ ,$X_n$ are not independent of each other. In other words, the equation (1)has non-zero solution, i.e. $k_1, k_2, \cdots, k_n$ are not all zero [6].

$$X_1 k_1 + X_2 k_2 + \cdots + X_k k_n \approx 0 \tag{1}$$

If the the left of equation (1) is precisely equal to zero, there is a precise or complete linear relationship between $X_1$,$X_2$, ⋯,$X_n$. When it is confirmed that there is a multiple co-linear relationship between $X_1$,$X_2$, ⋯,$X_n$, then one or more of these variables can be expressed in linear representations of other variables. It shows that the correlation between this variable and other variables is very strong, and there is no need to exist as a separate explanatory variable in the model.

The existence of multiple co-linearity among the variables used for modeling will lead to the turbulence of model and distortion of coefficient.

## 3. Diagnosis and Treatment Methods

In order to eliminate the influence of multiple co-linearity when modeling, the most simple and effective method is identifying the information overlapping variables and exclude them out. With respect to the selection of variables, the following treatment methods are proposed on the basis of the improvement of several common diagnostic methods.

### 3.1. *Pearson correlation coefficient diagnostic method*

Over the years Pearson correlation coefficient which is used to measure the degree of linear correlation between variables, has been widely used in fields such as economics, medicine, statistics and so on [7]. A great deal of experience has been accumulated, and a set of relatively explicit quantitative relations to partition the degree of correlation is obtained.

Taking variables X and Y as an example, the Pearson correlation coefficient can be calculated by the following formula:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2-E^2(X)}\sqrt{(Y^2-E^2(Y)}} \quad (2)$$

When the unit of variable changes, correlation between variables will not be affected, this ensures the Pearson correlation coefficient unchangeable. So the correlation coefficient is also comparable when data unit is not at the same time.

The range of Pearson correlation coefficient is (-1, 1), and the absolute value is between 0 and 1 [8]. The absolute value of the correlation coefficient indicates the degree of correlation. The greater the value, the stronger the correlation degree is, and the smaller the value, the weaker the correlation degree is [9]. Experience shows that when the absolute value of the correlation coefficient is greater than 0.9, there may be a problem of multiple co-linearity among the variables [10]. So we can calculate the correlation coefficient matrix of the independent variables of all measured parameters of the power station, and retain variables of which the value is less than threshold 0.9, realizing the judgment and selection of co-linearity.

In order to facilitate the realization of automatic selection of variables, variables need to be sorted in the sense of data. Therefore, the input variables are sorted according to the eigenvalue at first.

Analyzed from mathematics, eigenvalue is obtained by decomposing the matrix in the linear space, and projecting on N standard orthogonal basis. The size of the eigenvalue represents the projection length of the matrix on each base.

The greater the eigenvalue, the greater the power of the matrix in the corresponding direction, and the more the information contained. In practical modeling applications, this feature of eigenvalue can be used for the sorting and filtering of variables. The greater the eigenvalue, the more information contained in the corresponding characteristic vector direction, and the greater contribution of the corresponding variable [11].

Calculating the Pearson correlation coefficient between any two variables after sorting the data, and retaining variables of which the value is less than threshold and with higher rank.

### 3.2. *VIF auxiliary regression test*

Sub-headings should be typeset in boldface italic and capitalize the first letter of the first word only. Section number to be in boldface roman. Experience shows that, when there is multiple co-linearity in the variables, the variance inflation factor (VIF) of the regression equation is greater than 100. According to this point, the specific co-linearity of the variable matrix can be judged by making auxiliary regression. Regarding each variable as the dependent variable, and the others as the independent variable in turn, and making regression analysis. Then calculate variance expansion factor for each regression equation [12].

The formula for calculating the variance inflation factor is

$$VIF_i = 1/(1 - R_i^2) \tag{3}$$

In the formula, $R_i^2$ represents coefficient of determination of the regression equation made by variable xi and the rest of the variables. The range of $R_i^2$ is [0, 1).

The change trend of $VIF_i$ and $R_i^2$ are the same, that is the larger the $R_i^2$, the larger the $VIF_i$. And at this time, the correlation between xi and other variables is relatively strong. Conversely, the smaller the $R_i^2$, the smaller the $VIF_i$. And at this time, the correlation between xi and other variables is relatively weak [6]. The value of $VIF_i$ is generally greater than or equal to 1. The greater the value of VIF is, the greater the possibility of multiple co-linearity between variables is. Experience shows that, generally when the VIF>100, it can be judged that there exists multiple co-linearity between the variables, and the data needs to be pretreated. Otherwise it will affect the regression estimate [13].

Take the historical operation data provided by the 600MW system of a power plant DCS system as an example to verify the pretreatment model. Regression analysis was performed regarding each variable as the dependent variable, and the others as the independent variable. Calculate variance expansion factor for each regression equation. Remove the variable

corresponding to the maximum VIF value [12]. Repeat the above steps for the remaining variables until the VIF values are all less than 100.

### 3.3. *Condition number diagnosis method*

When there is multiple co-linearity among variables, the variable matrix is ill-conditioned matrix. The more serious the multiple co-linearity is, the more ill-conditioned the matrix is. Condition number of matrix is commonly used to judge and measure the degree of ill condition for matrix [14].

Condition number equals the product of matrix norm and its inverse matrix norm, which represents the sensitivity to errors of matrix computation. Take linear regression as an example, the regression equation between the independent variable matrix $X$ and the dependent variable y is $y=AX$. When the condition number of $X$ is large, the small change of X will lead to large changes in the estimated value of $y$, which is not conducive to the estimation of the regression value; or slight changes in $y$, $A$ will have a large change, which is not conducive to the solution of the regression coefficient.

The calculation of condition number is defined as

$$cond(A) = \|A\| \cdot \|A^{-1}\| \tag{4}$$

From the analysis of linear algebra, the condition number of the matrix is always greater than 1. The condition number of orthogonal matrix is equal to 1, the condition number of singular matrix is infinite, and the condition number of the ill-conditioned matrix is relatively large. And this can be the basis for the diagnosis of multiple co-linearity between variables. Experience shows that when the condition number is greater than 30, matrix is ill-conditioned, and there is very likely to be a mutiple co-linear problem between variables. Therefore, all measurement parameters of the power station can be sorted according to the eigenvalue, and then the condition number of the variable matrix will be calculated. Variables of which the value is less than threshold 30 will be retained, realizing the judgment and selection of co-linearity. Specific steps are as follows:

(1)Sort the variables according to the eigenvalue; the sample matrix is obtained as $X = (X_1, X_2, \cdots, X_k)$.

(2)Calculate the condition number of matrix $[X_i \ \ X_j]$ ($i=1,2, \cdots k-1, j=i+1, i+2, \cdots k$). If the condition number is less than 30, then keep $X_j$; otherwise remove.

## 4. Example Verification and Methods Comparison

Affected by the coal market, the variety of the furnace coal of the power plant cannot always be the same as the designed series, and blended coals may also be

combusted. The frequent changes of the quality of coal combustion seriously affect the safety and economic operation of the power plant. The important evaluation index of coal quality is the calorific value of coal, so that the change of coal quality can be characterized by the change of net calorific value. Timely and accurate grasp of the coal heat has far-reaching significance for the operation optimization of the unit [15], for it provides online reference for performance calculation and combustion optimization. Therefore, online monitoring model of coal quality can be established by the preprocessed data for instance verification.

The historical data of 600MW unit of DCS system in a power plant was pretreated with the Pearson correlation coefficient diagnosis method, the VIF auxiliary regression test method and the condition number diagnosis method respectively. The original input data has 379 sets of variables, and the data dimension has been reduced to 133, 162 and 170 after screening.

According to genetic algorithm and BP neural network theory, online measurement parameters as input variables, and the low calorific value of coal as output variable, programming in MATLAB software based on the genetic algorithm optimized BP neural network to realize online monitoring of coal calorific value [16]. 4326 sets of data samples were obtained from the power plant DCS system, from which 4286 groups were randomly selected as training data for network training, and the remaining 40 sets of data as test data to test the fitting effect of the test network. Seperately establish models by data before and after pretreatment. The number of input variables were 379, 133, 162 and 170 respectively, and the number of output variables is 1. The GA-BP neural network model has been established to predict the coal heat. Part of the predicted results are shown in Table 1. The comparison result of measurement accuracy and model reduction is shown in Table 2.

Table 1 Forecast result of calorific value of coal

| No. | Coal Heat (MJ/kg) | Primitive Variable | | Pearson Correlation | | VIF Auxiliary Regression | | Condition Number Diagnosis | |
|---|---|---|---|---|---|---|---|---|---|
| | | Predicted Value (MJ/kg) | Relative Error (%) | Predicted Value (MJ/kg) | Relative Error (%) | Predicted Value (MJ/kg) | Relative Error (%) | Predicted Value (MJ/kg) | Relative Error (%) |
| 1 | 23.68 | 23.6441 | -0.1516 | 23.66253 | 0.07377 | 23.52182 | -0.6679 | 23.69827 | 0.07717 |
| 2 | 21.88 | 21.94447 | 0.29467 | 21.93833 | 0.26658 | 21.90138 | 0.09770 | 21.86509 | 0.06816 |
| 3 | 21.22 | 21.31079 | 0.42783 | 21.18963 | -0.1431 | 21.27453 | 0.25696 | 21.26261 | 0.20081 |
| 4 | 22.03 | 22.08679 | 0.25779 | 22.01064 | -0.0878 | 22.10173 | 0.32562 | 21.98482 | 0.20509 |
| 5 | 22.18 | 22.06288 | -0.5280 | 22.15296 | -0.1219 | 22.23909 | 0.26642 | 22.19183 | 0.05332 |
| 6 | 23.68 | 23.64187 | -0.1610 | 23.63473 | -0.1911 | 23.67192 | -0.0341 | 23.66313 | 0.07126 |
| 7 | 21.43 | 21.67495 | 0.74301 | 21.49031 | 0.28144 | 21.37914 | -0.2373 | 21.52815 | 0.45798 |
| 8 | 20.06 | 20.13997 | 0.39864 | 20.0908 | 0.15356 | 20.07783 | 0.08886 | 20.07784 | 0.08892 |
| 9 | 21.22 | 21.05465 | 0.77923 | 21.23572 | 0.07409 | 21.2444 | 0.11497 | 21.25859 | 0.18186 |
| 10 | 23.68 | 23.66514 | 0.06276 | 23.68179 | 0.00758 | 23.76208 | 0.34661 | 23.66115 | -0.0796 |

Table 2 Comparison of three kinds of model pretreatment

| | | Primitive Variable | Pearson Correlation | VIF Auxiliary Regression | Condition Number Diagnosis |
|---|---|---|---|---|---|
| Variable Dimension | | 379 | 133 | 162 | 170 |
| Prediction Accuracy (%) | Maximal Absolute Value of Relative Error | 0.9840 | 0.9418 | 0.9558 | 0.7155 |
| | Average Relative Error | 0.1693 | 0.1485 | 0.1412 | 0.1384 |

From table 1 and table 2, it can conclude that after comprehensive comparison, the treatment effect of condition number diagnosis method is better. The method has not only reduced the dimension of the variables, but also improved the quality and practicality of the data in big data paltform.

## 5. Conclusion

Thermal system big data provides massive information, but also inevitably there is a large amount of information overlap. To diagnose and deal with the multiple co-linearity of big data of thermal system, the data of the 600MW unit of a power plant was pretreated by Pearson correlation coefficient diagnosis method, VIF auxiliary regression test method and condition number diagnosis method respectively. And the model was verified by the online monitoring model of coal

calorific value. After comprehensive comparison, the application effect of the preprocessed data of condition number diagnosis method is the best, and the model is relatively simple. So we can use the condition number diagnosis method to reduce the data dimension, realizing the pretreatment of big data platform of thermal system, and increasing the value density of thermal system big data.

**References**

1. Gong Xiayi, Li Bohu, Chai Xudong. Summary of big data platform technology [J]. Journal of System Simulation, 2014, 26 (3): 489-496. (In Chinese)
2. Zhu Zhaoyang, Wang Jiye, Deng Chunyu. Research and design of big data platform for electric power system [J]. Electric Power Information and Communication Technology, 2015, 13 (6):1-2. (In Chinese)
3. Qian Jun. GIS spatial data processing and quality control system [D]. Tongji University, 2007:6-10. (In Chinese)
4. Song Yu, Jiao Pu, Li Gang. Characteristics of attribute reduction in big data preprocessing [J]. Computer Measurement and Control, 2015, 12:4191-4194. (In Chinese)
5. Wang Jianxing. Research on application of reverse modeling in modeling of complex thermal systems [D]. North China Electric Power University, 2012:14-16. (In Chinese)
6. Fan Jingcheng. Diagnosis and analysis of complex linear relationship under complex data [D]. Shandong University, 2008:4-8. (In Chinese)
7. Li Xiumin, Jiang Weihua. Correlation coefficient and correlation measure [J]. Practice and understanding of Mathematics, 2006, 12:188-192. (In Chinese)
8. Cui Xiujuan. Preliminary study on the national macro warning of safety accidents [D]. Capital University of Economics and Business, 2006:20-22. (In Chinese)
9. Hu Hui, Xu Haofeng, Bao Weihua. Ultrasonic echo location based on correlation algorithm [J]. Automation Instrument, 2015, 10:96-98+102. (In Chinese)
10. Xu Cunxing. Data mining and analysis of cash flow based on financial statements [J]. Journal of Shandong Agricultural University: Natural Science Edition, 2014 (4): 626-631. (In Chinese)
11. Chen Limin, Yang Jing, Zhang Jianpei. A big data Set spectrum clustering method based on accelerated iteration [J]. Computer Science, 2012, 05:172-176. (In Chinese)

12. Mao Lifan. Study on long term load forecasting technology in power network planning [D]. Hunan University, 2011:25-27. (In Chinese)
13. Yang Mei, Xiao Jing, Cai Hui. Multiple co-linearity and processing method in multivariate analysis [J]. China Health Statistics, 2012, 04:620-624. (In Chinese)
14. Sun Xiaofei. The condition number of two kinds of special matrix [D]. Taiyuan University of Technology, 2013:13-16. (In Chinese)
15. Jingyuan. Research on inverse modeling of coal calorific value of power plant boiler [D]. North China Electric Power University (Beijing), 2013:8-10. (In Chinese)
16. Liu Lin, Wang Peng, Zhai Yongjie, et al. Prediction of coal calorific value based on algorithm [J]. Thermal Power Generation, 2015 (2): 47-51. (In Chinese)