Preliminarily Realize the Remote Diagnosis of the Running State of the Power Station Unit with Principal Component Analysis Based on Java Program

Hui Zhang $^{1,2}$ and Xiaochuan Niu $^3$

$^1$ School of Electrical and Power Engineering, China University of Mining and Technology, Xuzhou 221116, China;  
$^2$ Xuhai College, China University of Mining and Technology, Xuzhou 221008, China  
$^3$ Guohua Power Generation Co., Ltd, Xuzhou 221135, China  

Correspondence: cumthui@126.com

Abstract: For realizing the remote diagnosis of the running state of the power station unit preliminarily, the article accomplishes relevant calculation module for implementing the PCA program, and sets up the data interface of power plant data and PCA module by using the Java program, which completes the calculation and analysis of power plant data in a timely and efficient manner, removes the complexity of manual processing in large dimension data filed and realizes the real-time processing of operation parameters of power plant. Furthermore, the portability of program modules is enhanced and the work efficiency is greatly improved by making the program module executable, which provides a certain reference basis for the monitoring of operation of power plant unit.

1. Introduction

Due to the complexity of the boiler system, the parameters are highly coupled, and the relationship between the cause of the fault and the symptom of the fault is complicated [1]. At the same time, with the aging of equipment and changes in the external environment, the possibility of faults is increasing [2]. It is to monitor and evaluate the operating status of the unit in real time, and eliminate the influence of subjective factors on the overall operation status of the unit, to maintain the relationship between the original state parameters to ensure the reliability and safety of the thermal power generation process and improve the quality of service, it is urgent to realize timely detection and effective diagnosis of process faults [3] [4] [5].

There is a strong multi-dimensional and nonlinear coupling relationship between the variables affecting the running state of the unit. The changes of these variables bring great difficulties to the optimal operation of the unit [6] [7]. Therefore, it is desirable to be able to select a small number of important variables to include indicator information for evaluating unit status [8] [9]. Principal component analysis method statistically analyzes the relationship between various parameters on the basis of retaining the relationship between the original variables to the greatest extent, and integrates the original variables into a few new variables that are not related to each other [10] [11] [12].

Principal component analysis has been widely used in various fields such as signal processing and statistics [13] [14], and can also be used to extract the main information in power plant measurement data [15], while the traditional PCA method is currently only used to process smaller dimensional data. While the actual process of the data sample dimensions are hundreds of thousands, resulting in greater
processing difficulty. At the same time, the traditional PCA method does not perform data screening detection on the input sample data, and uniformly processes some operating factors and other factors that have less influence on unit efficiency in the actual operation process, which seriously affects the efficiency of data processing and the accuracy of final evaluation.

In order to meet the needs of modern intelligent power plant construction, and to analyze and process the operation data of the power plant more quickly and efficiently, and evaluate the operation status of the power plant in real time, this paper makes full use of computer means to write a big data processing module for implementing PCA method in Java language. The small-dimensional sample data is selected, and the PCA method is used to comprehensively evaluate the operating conditions of the same unit with the same output, the same unit with different output, and the same unit with different output, and compare with the actual operating status and related literature. The evaluation results are consistent, but the calculation time is only 1/10000 of the traditional calculation, which greatly improves the efficiency and accuracy of the calculation. Finally, based on the large-scale running samples provided by a power plant (a total of 300 samples, each sample data dimension is 10000×302), the paper uses the written Java program module to calculate and obtain the best operation of the unit in the past time period and the best operating parameters for each device.

2. Principle of Principal Component Analysis

2.1 Definition of Principal Component Analysis
Principal Component Analysis (PCA), first proposed by Hotelling in 1933, is a multivariate statistical analysis method that converts multiple indicators into fewer comprehensive indicators[16].

2.2 Steps of Principal Component Analysis

2.2.1 Standardization of Sample Indicators
The original data selects P samples of the unit operating under stable conditions, each containing m indicators $X_1, X_2, \ldots, X_m$. In order to eliminate the influence of the difference between the dimension and the order of magnitude, the raw data needs to be standardized first [17][18].

The formula for standardization of sample indicators is as follows:

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{\text{var}(x_j)}} \quad i = 1,2, \ldots, p, j = 1,2, \ldots, m$$

Here, $x_{ij}^*$ is the $j$th parameter value in the $i$th sample; $\bar{x}_j$ is the mean value of $x_j$; $\sqrt{\text{var}(x_j)}$ is the covariance of $x_j$.

2.2.2 Calculate the Correlation Matrix of the Sample

$$R = (r_{ij})_{m \times m} = X'X$$

Here, $r_{ij}$ is the simple coefficient of indicator $i$ and indicator $j$.

2.2.3 Calculate the Eigenvalues and Unit Eigenvectors of the Correlation Matrix
(1) Calculate the Eigenvalue

Let $|R - \lambda_1 I| = 0, |R - \lambda_2 I| = 0, \ldots, |R - \lambda_m I| = 0$ (3)

Solve the equation and get m non-negative eigenvalues, in fact, they are the variances of the principal components $Y_1, Y_2, \ldots, Y_m$ respectively.

(2) Calculate Eigenvalue Contribution Rate and Cumulative Contribution Rate

The variance contribution rate $\gamma_k$ of the $k$th principal component indicates that the sample contains information as a weight of all samples information content.

$$\gamma_k = \frac{\lambda_k}{\sum_{i=1}^{m} \lambda_i}$$

The cumulative contribution rate of the principal components $Y_1, Y_2, \ldots, Y_n$ is as follows:
\[
\sum_{i=1}^{n} \lambda_i / \sum_{i=1}^{m} \lambda_i
\]

The principal components can be represented by the eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_m \) and arranged according to the size of the eigenvalues. In practical applications, \( n \) principal components (\( n < m \)) are generally selected, and the cumulative contribution rate is greater than 85%.

(3) Calculate the Eigenvectors of the Standard Matrix

\[
\mathbf{u} = \begin{bmatrix}
\mathbf{u}_{11} \\
\mathbf{u}_{21} \\
\vdots \\
\mathbf{u}_{m1}
\end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix}
\mathbf{u}_{12} \\
\mathbf{u}_{22} \\
\vdots \\
\mathbf{u}_{m2}
\end{bmatrix}, \ldots, \quad \mathbf{u} = \begin{bmatrix}
\mathbf{u}_{1m} \\
\mathbf{u}_{2m} \\
\vdots \\
\mathbf{u}_{mm}
\end{bmatrix}
\]

(6)

2.2.4 Principal Component Function

\[
\begin{align*}
Y_1 &= \mathbf{u}_{11} x_1 + \mathbf{u}_{12} x_2 + \cdots + \mathbf{u}_{1m} x_m \\
Y_2 &= \mathbf{u}_{21} x_1 + \mathbf{u}_{22} x_2 + \cdots + \mathbf{u}_{2m} x_m \\
&\vdots \\
Y_m &= \mathbf{u}_{m1} x_1 + \mathbf{u}_{m2} x_2 + \cdots + \mathbf{u}_{mm} x_m
\end{align*}
\]

(7)

2.2.5 Comprehensive Evaluation Function

\[
F = a_1 Y_1 + a_2 Y_2 + \cdots + a_m Y_m
\]

(8)

The corresponding principal component normalization coefficient is selected according to the principle of contribution rate of eigenvalues, and it is brought into the comprehensive evaluation function to obtain the comprehensive score of each sample and the criteria are sorted to find the most reasonable operating conditions in the corresponding time.

3. Application

3.1 Boiler System Description

This paper takes the boiler of a domestic power plant as the research object. All the analysis and diagnosis are based on this boiler, but the diagnosis principle and analysis method are not limited to the boiler, and can be applied to other boilers.

The boiler belongs to subcritical parameters, intermediate reheating, coal burning, natural circulation boiler, and the boiler layout is \( \pi \) type. As shown in Figure 1, the design parameters are shown in Table 1.

The super-heater system is in the order of steam flow: ceiling pipes, wall pipes, low temperature super-heater, front screen super-heater, rear screen super-heater and high temperature super-heater; the super-heater system is in the order of flue gas flow: front screen super-heater, rear screen super-heater, high temperature super-heater and low temperature super-heater.
Figure 1. Schematic diagram of boiler structure

The re-heater system is in the order of steam flow: low temperature wall re-heater, medium temperature re-heater and high temperature re-heater; the re-heater system is in the order of flue gas flow: wall re-heater, medium temperature re-heater and high temperature re-heater.

During the actual operation of the unit, parameters of boiler steam side and flue gas side such as super-heater temperature and pressure, re-heater temperature and pressure, feed water temperature, feed water flow rate, primary and secondary air volume, primary and secondary air temperature, re-heater water spray, super-heater water spray and auxiliary equipment such as coal mill, flue gas regulating baffle and high-pressure heater have an important influence on the coal consumption of the unit, that is, the economics of the unit.

3.2 Program Function

The PCA data interface program reads data from the plant's DCS or SIS platform, according to the text type, select the corresponding data processing module, matrix the data text, and use the matrix processing module to eliminate the static variables or quasi-static variables in the running process of the unit according to the corresponding algorithm, thereby constructing a new matrix and completing the new matrix and the related operations of data processing, result analysis and storage are completed, and the optimal operating condition nodes and corresponding parameters of the unit are obtained.

3.3 Program model description

This paper implements the traditional PCA method by programming, and uses the Java language to write the main program control module, the data reading module and the data processing and analysis module and the result storage module for different data types such as CSV and Excel. The specific module architecture is shown in Figure 2.

Before Principal Component Analysis, because there are large differences in the magnitude and dimension of each variable, it is first necessary to normalize the data, then calculate the eigenvalues and unit eigenvectors of the correlation matrix, perform principal component analysis on the data of each unit, and then calculate the corresponding weight of each indicator in each principal component and the cumulative contribution rate of each principal component, the final principal component is obtained according to the relevant threshold criteria, and a comprehensive evaluation function is established to realize real-time data reception and analysis. The specific process is shown in Table 1.
1) Program Module

![Program module flow chart]

Figure 2. Program module flow chart

2) Specific Program

| Item       | Content                                                                 |
|------------|--------------------------------------------------------------------------|
| Input      | SIS/DCS transfer CSV/Excel real-time data set                            |
| Output     | Comprehensive evaluation results of operating conditions objlist        |

function PrincipalComponentAnalysis (Data) {
    // Segmentation of data according to different text input formats
    pcanalysis.readatarr(D);
    // Calculate the matrix dimensions and build an array list
    Linedimension; // Horizontal dimension
    Rowdimension; // Vertical dimension
dataarray; // Array list
    computestandardddmatrix(dataarray, rowdimension, linedimension, firstArray);// data transmission
    // Divide the transmission data matrix according to specific requirements and calculate the mean
Parameters($x_i$)=average;
// Calculating the variance set
\[ x_{ij} = \text{var}(x_j) \]

Variance.length= rowdimension;
// Storage variance
Variance($x_i$)=ave;
// Matrix normalization, calculate correlation coefficient, judgment of applicability of PCA method
if(Math.abs(variances[j]) == 0) {
    coefficient[i][j] = 0;
} else {
    // Storage standardization matrix
    \[ x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\text{var}(x_j)}} \quad i = 1, 2, ..., n, j = 1, 2, ..., m \]
    coefficient[i][j] = (dataarray[i][j] - parameters[j]) / Math.sqrt(variances[j]);
}

This paper is based on the operating data of a 200 MW unit provided by a certain electric power institute during a continuous period of time, the sample interval is one second, and each sample data contains 302 operating parameters. Randomly select consecutive data samples corresponding to 600 seconds (arbitrary continuous time) to form a matrix of 600×302, as shown in Table 2. Principal component analysis method is used to eliminate redundant parameter information of the unit, and low-dimensional variable information is used to effectively represent high-dimensional data information.

Table 2. Operating parameters of the unit in random continuous time

| time/s | Pressure of economizer inlet flue gas of A side (kpa) | Wall temperature of low temperature super-heater 1(℃) | Air preheater secondary air inlet air temperature of A side (℃) | Main steam flow of Unit 4 (t/h) |
|--------|-----------------------------------------------------|-----------------------------------------------------|-----------------------------------------------------|-----------------------------|
| 1      | -0.08523                                            | 396.6904                                            | 11.9848                                             | 523.0215                    |
| 2      | -0.08513                                            | 396.6914                                            | 11.9847                                             | 522.6436                    |
| 3      | -0.08502                                            | 396.6923                                            | 11.9846                                             | 522.2658                    |
| …      | …                                                   | …                                                   | …                                                   | …                           |
| 600    | -0.09840                                            | 398.2255                                            | 11.9378                                             | 521.2247                    |

Table 1b

| ALGO item content | Input SIS/DCS transfer CSV/Excel real-time data set | Data output Comprehensive evaluation results of operating conditions objlist |
|-------------------|-----------------------------------------------------|--------------------------------------------------------------------------|
| ALGO              |                                                     |                                                                           |
|                   | //Calculate the transposed matrix
|                   | transposedmatrix[i][j]=coefficient[i][j];          |                                                                           |
|                   | //Computational correlation matrix                  |                                                                           |
|                   | Xmultiplematrix[i][k]=transposedmatrix[i][j]*coefficient[j][k]; |                                                                           |
|                   | //Calculate eigenvalue                              |                                                                           |
|                   | Matrix(copyvalue).eig().getD();                    |                                                                           |
|                   | Matrix(copyvalue).eig().getV();                    |                                                                           |
|                   | //Obtain parameter weight                          |                                                                           |
|                   | lambdaSum+=Avalue.eig().getD().get(i, i);         |                                                                           |
Calculate eigenvalue contribution rate and cumulative contribution rate

\[
\alpha_k = \lambda_k / \left( \sum_{i=1}^{m} \lambda_i \right) \sum_{i=1}^{n} \lambda_i / \left( \sum_{i=1}^{m} \lambda_i \right)
\]

if (threshold>0.85) {
    if (savey.length==1) {
        "there are"+savey.length+" main elements ";
        @SuppressWarnings("rawtypes")
        map.put(savey[0][i], i);
        list.add((savey[0][i]));
        "the"+(int) map.get(savey[0][savey[0].length-1])+" row of boiler parameters represents the best operating conditions"
    } else if (savey.length==2) {
        "there are"+savey.length+" main elements ";
        @SuppressWarnings("rawtypes")
        map.put(savey[0][i], i);
        list.add((savey[0][i]));
        "the"+(int) map.get(savey[0][savey[0].length-1])+" row of boiler parameters represents the best operating conditions"
        }
    }
}

and comprehensive evaluation is carried out to calculate the optimal operating conditions in the above time period [19][20].

Table 1 shows some important parameters related to the overall economic characteristics of the unit during operation, so that a matrix can be established.

\[
\mathbf{X} = (x_1, x_2, \ldots, x_m)
\]  

Here, \(x_1\) is the pressure of economizer inlet flue gas of A side; \(x_1\) is the wall temperature of low temperature super-heater; \(\ldots, x_m\) is the main steam flow.

The program module is executable, and the standardized matrix is obtained by analyzing the data, as follows:

\[
\mathbf{X} = \begin{bmatrix}
2.569 & 1.633 & -1.634 & -1.633 & -1.633 & -1.633 \\
1.970 & 1.429 & -1.427 & \cdots & -1.429 & -1.429 & -1.429 \\
1.371 & 1.225 & -1.224 & \cdots & -1.225 & -1.225 & -1.225 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
-0.663 & -1.224 & 1.223 & 1.224 & 1.224 & 1.225 \\
-0.689 & -1.429 & 1.431 & \cdots & 1.428 & 1.429 & 1.429 \\
-0.716 & -1.633 & 1.633 & 1.634 & 1.633 & 1.633 \\
\end{bmatrix}
\]  

Calculating the eigenvalue, contribution rate and cumulative contribution rate of each principal component, as shown in Table 3.

| principal components | Eigenvales \(\lambda\) | Contribution rate \(\alpha_k/%\) | Cumulative contribution rate \(\sum \alpha_k/%\) |
|----------------------|--------------------------|-------------------------------|------------------------------------------|
| \(y_1\)              | 190.123                  | 0.932                         | 0.932                                    |
| \(y_2\)              | 8.487                    | 0.042                         | 0.974                                    |
| \(y_3\)              | 5.207                    | 0.023                         | 0.997                                    |
| \(\ldots\)           | \(\ldots\)              | \(\ldots\)                   | \(\ldots\)                               |
| \(y_{600}\)          | 0.0018                   | 0.000012                      | 1.00                                     |

Calculate the unit eigenvector of the sample correlation matrix \(\mathbf{X}\) and the corresponding weight of
As can be seen from Table 3, the comprehensive evaluation value of the first principal component is 93.2%, the comprehensive operation characteristics of the unit are fully expressed, indicating that the first principal component has covered enough information of the original data. Therefore, it is selected as the main component, and other principal components are negligible.

Comprehensive evaluation function:

\[ F = a_1 y_1 = 1.901 \]  

According to the comprehensive evaluation function, the running data corresponding to all time points can be calculated to obtain the comprehensive sorting of the corresponding operating parameters, and then the optimal operating conditions and corresponding parameter values in the randomly selected time period are determined, as shown in Table 5.

### Table 5. Comprehensive ordering of parameters

| time/s | 1   | 2   | ……  | 17  | ……  | 600  |
|--------|-----|-----|------|-----|------|------|
| F value| 0.159 | -0.295 | ……  | 1.132 | ……  | -0.643 |
| Sort   | 26  | 125 | ……  | 1   | ……  | 327  |

As shown in Table 5, the calculation of the randomly selected 600 consecutive time nodes shows that the instantaneous working condition corresponding to the 17th time node makes the equipment have the best economy during this period, and can be used as a reference standard for the unit to operate in real time. And adjust the unit system equipment according to the above node parameters.

### 4.Conclusions

Based on the original principal component analysis method, this paper realizes the real-time processing of data through Java language programming, which eliminates the complexity of manual processing under large-dimensional data field. At the same time, the program automatically completes the real-time analysis of similar data in large-scale data transmission, which greatly improved the work efficiency and won valuable time for the power plant to adjust the operating conditions in time.

In addition, by writing interface programs for different types of data such as CSV and Excel, the program adaptability is expanded, and the program modules are executable, which enhances the portability of the program modules.

By analyzing the operation data of a 200 MW unit provided by a certain electric power institute in a continuous time period, the operating condition corresponding to the 17th time point can be used as the best reference operation value in the current time period, and combine multiple indicators into one new indicator with little loss of parameter information, and key factors are extracted and used for the state evaluation of the unit, which greatly eliminates the influence of subjective factors on the operating state of the unit and provides a basis for on-line monitoring of the unit status.

### Acknowledgments

The authors are very grateful to Dr. Ma GH for his help in programming and scheme design.

### References

[1] Wang HJ. (2008) Research on real-time fault diagnosis of power station boiler. [Doctoral
thesis]. North China Electric Power University. Beijing. pp. 1-8.

[2] Kim, H., Na, MG., Heo, G. (2014) Application of monitoring, diagnosis, and prognosis in thermal performance analysis for nuclear power plants. Nuclear Engineering and Technology., 46:737-752.

[3] Chandrasekharan, S., Panda, RC., Swaminathan, BN. (2014) Modeling, identification, and control of coal-fired thermal power plants. Reviews in Chemical Engineering., 30: 217-232.

[4] Aurora, C. , Magni, L., Scattolini, R., Colombo, P., Pretolani, F., Villa, G. (2004) Predictive control of thermal power plants. Int. J. Robust Nonlinear Control., 14:415–433.

[5] Nikula, RP., Ruusunen, M., Leiviskä, K. (2016) Data-driven framework for boiler performance monitoring. Applied Energy., 183: 1374-1388.

[6] Tang, ZH., Zhang, ZJ. (2019) The multi-objective optimization of combustion system operations based on deep data-driven models. Energy., 182: 37-47.

[7] Chandrasekharan, S., Panda, RC., Swaminathan, BN., Panda, A. (2018) Operational control of an integrated drum boiler of a coal fired thermal power plant., Energy., 159: 977-987.

[8] Yan CL. (1998) Some questions about the comprehensive evaluation by principal component analysis. Mathematical Statistics and Management., 3: 22-25.

[9] Fu ZG.,Wang LP.,Ge ZH. (2008) Comprehensive evaluation of the operating status of power plant units by principal component analysis.Power Engineering., 28: 0548-0551.

[10] Ma R., Kang R., Luo B., et al. (2013) Identification method of energy consumption characteristics of thermal power units based on improved principal component analysis. Power Grid Technology., 37: 1196-1202.

[11] Yuan XW. (2012) Analysis of the application of principal component analysis method in fault detection of thermal power plant production process. Academic Discussion., 8: 40.

[12] He XQ. (2007) Modern statistical analysis methods and applications. China Renmin University Press., Beijing. pp. 358-376.

[13] Pyun, H., Kim, K., Ha, D., Lee, CJ., Lee, WB. (2020) Root causality analysis at early abnormal stage using principal component analysis and multivariate Granger causality. Process Safety and Environmental Protection., 135: 113-125.

[14] Si FQ.,Hong J.,Xu GZ. (2002) Data inspection PCA method based on vector projection. Chinese Society for Electrical Engineering., 22: 157-160.

[15] Zhang XT. (2001) Data Mining and Statistics. China Statistics., 10: 9-12.

[16] Ye SF. (2001) Improvement on the comprehensive evaluation of principal component analysis. Mathematical Statistics and Management., 20: 52-55.

[17] Yu JH., Yang WQ. (2005) Multivariate statistical analysis and application EM3. Zhongshan University Press., Guangzhou. pp. 189-206.

[18] Wang DY.,Shuang GC. (1998) Mathematical model for comprehensive evaluation of unit operation economy. East China Electric Power., 7: 7-9.

[19] Wang JR.,Li PK.,Zheng HW. (2004) Application of Principal Component Analysis Method in Process Control of Thermal Power Plants.Journal of Instrumentation., 81: 1028-1029.

[20] Li XX.,Fan ZY.,Wang HJ. (2016) Application of Principal Component Analysis in Optimal Operation of Thermal Power Units. Power plant auxiliary machine.,37:33-36.