A root-cause analysis method for fault diagnosis in condenser

Banglei Han1, ChaoZhang*, Chu Xie1

1College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China

*Corresponding author’s e-mail: skdzhangchao@sdust.edu.cn

Abstract. Condenser is an important part of thermal power units. When the condenser fails, the quick determination of the root-cause fault can effectively avoid major safety accidents in thermal power units. Most existing methods are based on the assumption of stationary relationships among the variables. However, the relationships among the variables are non-stationary in the real industrial process. These methods are not suitable for real industries. According to the process knowledge, the condenser is a typical system of non-stationary. Aiming at the non-stationary relationships, a root-cause analysis method based on multiple regression models is proposed in this paper. Firstly, the bottom-up piecewise linear representation method is applied to yield qualitative trends of historical data for each variable. Secondly, qualitative trend combinations of multiple variables are obtained based on amplitudes for each variable. Finally, the contributions of the related variables in qualitative trend combinations to the main variable are calculated based on multiple regression models, and the root-cause variable can be determined according to the size of the contribution. An industrial example is provided to illustrate the effectiveness of the proposed method.

1. Introduction

Condenser is an important part of thermal power units. When the condenser fails, the quick determination of the root-cause fault can effectively avoid major safety accidents in thermal power units. Therefore, the root-cause analysis of the condenser plays a very important role in ensuring the safe operation of the thermal power units.

In recent years, research on root-cause analysis for fault diagnosis has received extensive attention from the industry and academia. Tao et al. and Zhang et al. put forward a method of main component analysis model fault root analysis[1-2].Folmer et al. clustered frequently occurring historical alarm data and matched the pattern of online alarm data with the alarm patterns frequently occurring in the history, so as to help operators conduct root cause analysis[3].Gonzalez et al. proposed an offline root cause analysis method based on machine learning technology, which was strengthened by summarizing, graphic operation and visualization, thus helping operators identify the root-cause of faults[4].Wang et al. proposed an alarm root-analysis method based on Bayesian network for a special type of alarm variable in thermal power units[5].All these methods above require complete process knowledge, mechanism information, learning algorithm and data mining. In addition to mechanism knowledge or learning class algorithms, the method based on causal topological analysis has also been researched. Cauvin, Chiang, and Bauer et al. proposed a qualitative model based on causality to establish an abnormal propagation path to determine the root-cause variable [6-8]. Bauer and Hu et al. proposed a data-driven method based on historical process data to determine the direction of disturbance propagation by using transfer entropy [9-10]. Yuan et al. used Granger causality to diagnose the oscillation source and propagation under the hypothesis of stationary time series.
determine the root cause of alarms[11]. Wang et al. put forward process based on Granger causality method, and clarified the interaction between similar alarm variables. Combining similarity analysis of alarm data with causal analysis of process data, the corresponding alarm and its evolution path can be identified effectively [12]. Wang et al. proposed a root-cause analysis method based on a time causal graph, which eliminated the unknown variables in the equation through the forward and backward propagation methods of causal path, so as to determine the root variables of faults [13].

Since the existing root-cause analysis methods are mostly based on assumption that the relationships between variables are stationary sequences, so only one root variable can be analyzed. However, in the actual industrial process, the relationships between the main variable and each related variable are non-stationary, which make the root-cause time-varying. According to the process knowledge, the condenser is a typical system which the relationships among the process variables are time-varying. It is difficult to obtain the correct root-cause variable by existing methods, so this paper proposed a root-cause analysis method based on multiple regression models. The method can divide a non-stationary sequence into many stationary sequences. And the contribution of each related variable to the main variable is calculated by a multiple regression model when the main variable is abnormal. The root-cause variable for fault is determined according to the size of the contribution. Themethod proposed in this paper can obtain the correct root-cause analysis results.

2. Proposed method

The proposed method consists of three parts: (i) piece-wise linear representations are used to obtain the trends of historical time sequences, (ii) threshold of variables’ changes are used to obtain the qualitative trends and statistics of their combinations, (iii) root-cause analysis of multiple regression models are used to determine the root-cause variable by the contributions. The detail steps are as follows:

Firstly, in order to extract the trends of historical data, piece-wise linear representations are used to divide the time series \( \{x(t)\}_{t=1}^{N} \) into segments \( \{x(t)\}_{k=n_{k-1}}^{n_k-1} \), \( \{x(t)\}_{k=n_{K-1}}^{n_K-1} \) .... Each segment is represented by a straight line. \( n_k (k \in [1, K]) \) is the time index of the first sample in the \( K \)-th segment. It can be approximated as a linear regression model,

\[
x(t) = a_k + b_k t + e(t),
\]

where \( e(t) \) is assumed to be Gaussian white zero mean and variance \( \sigma^2 \). The unknown parameters \( a_k \) and \( b_k \) can be estimated from the least square method. So, the \( K \)-th segment \( \{x(t)\}_{k=n_{K-1}}^{n_K-1} \) can be estimated as

\[
\hat{x}(t) = \hat{a}_k + \hat{b}_k t.
\]

The bottom-up algorithm is used to divide times series [14]. Its main idea is to merge data segments constantly merged from more to less. The specified steps of this method are: (i) the time series is fitted at every two points, and the number of segments is \( N/2 \), (ii) the fitting error of each segment and the next segment is calculated, and the data segment with the smallest fitting error is merged, (iii) the above calculation is repeated until the number of segments of the data segment is equal to a given number \( K \), the number \( K \) of data segments can be determined by a method by Wang et al. [15]. In the multivariable system, the starting and ending positions of each segmentation result of the main variable are time intervals of other related variables, and straight line segment fitting is performed to realize a piecewise linear representation of multi-variables.

Secondly, since the trends of historical data are represented by piecewise linear representations, the amplitude changes of each variable are obtained. \( A_k \) is defined as the threshold of significant changes by Wang et al.[16]. The amplitude change of \( K \)-th data segment is represented as \( A_k \), if \( |A_k| \geq A_0 \), the variable changes significantly. The threshold \( A_0 \) is calculated as
where $R_0^2$ is the fitting, $\hat{\sigma}^2$ is the estimated value of $\sigma^2$

$$\hat{\sigma}_c^2 = \frac{1}{K} \sum_{k=1}^{K} \left[ \frac{1}{t_{k+1} - t_k} \sum_{t_i=t_k}^{t_{k+1}-1} (x(t) - \hat{x}(t))^2 \right].$$

Determining the trend of the $K$-th data segment based on $A_0$ as

$$z_c(k) = \begin{cases} 1, & A_k > A_0, \\ 0, & -A_0 < A_k < A_0, \\ -1, & A_k < -A_0. \end{cases}$$

The trend combinations are obtained of the $K$-th data segment for the $m+1$ variables of multivariable system

$$z = [z_{i_1}, z_{i_2}, ..., z_{i_m}, z_{i_{m+1}}].$$

A qualitative trend combination of all historical data is obtained $Z = [z_1, z_2, ..., z_4]$.

Thirdly, multiple regression models are reestablished for the main variable with the related variables of the amplitude changes significantly as input. The contributions of the related variables to the main variable are calculated, and the root-cause variable is determined by the contribution.

It is assumed that the number of the related variables is $m$. The $K$-th data segment is $y(t), x_1(t), x_2(t), ..., x_m(t)_{t_{k-1}}^{t_k}$, where $y(t)$ is the main variable, and $x_1, x_2, ..., x_m$ are the related variables. Then, the multiple regression model of main variable about related variables is calculated as

$$y_k = \beta_0 + \beta_{1,k} x_1 + \beta_{1,k} x_2 + ... + \beta_{m,k} x_m,$$

where $\beta_{i,k}, i \in [0, ..., m]$, are regression coefficients, obtained by the standard linear least-squares algorithm. The physical unit of each variable is different. So it is necessary to standardize the variable $x_i$ as

$$x'_i = \frac{x_{i,k} - \bar{x}_{i,k}}{A_{0,i}},$$

where $x'_{i,k}$ is the standardized data of $x_i$ in the $K$-th data segment, $\bar{x}_{i,k}$ is the mean of $x_i$ in the $K$-th data segment.

In a multivariable system, when the main variable $y$ changes abnormally in the $K$-th data segment, the contribution ratio of each related variable to the main variable is different. An index $\gamma_{i,k}$ is defined for the changes of $y$ as the contribution ratio of $x_i$

$$\gamma_{i,k} = \frac{\hat{\beta}_{i,k} A_{0,y}}{A_{0,x_i}} \Delta x_{i,k},$$

where $\Delta x_{i,k}$ is the amount of change of each related variable $\Delta x_{i,k} := x_{i,k,t_{k+1}} - x_{i,k,t_k}$. The derivation process of (9) is as follows:

In the $K$-th data segment, after normalizing the data, a multiple regression model is established according to (7). Since historical data of all variables are presented piecewise linear models, there is only one characteristic trend for each variable in $K$-th segment. The difference between the start and end points of each variable represents the changes of this variable $\Delta y_k := y_{k,t_{k+1}} - y_{k,t_k}$. So, the starting and ending point values of each related variable are brought into the multiple regression model, and obtained the estimated value of multiple regression model of these two points, $y'_{i,k}$ and $y'_{i,t_{k+1}}$

$$y'_{i,k} = \beta_{0,k} + \beta_{1,k} x'_{1,k} + \beta_{2,k} x'_{2,k} + ... + \beta_{m,k} x'_{m,k},$$

$$y'_{i,t_{k+1}} = \beta_{0,k} + \beta_{1,k} x'_{1,t_{k+1}} + \beta_{2,k} x'_{2,t_{k+1}} + ... + \beta_{m,k} x'_{m,t_{k+1}}.$$
\[ y'_{t,k} = \beta_0 + \beta_1 x_{t,k} + \beta_{2,k} x'_{t,k} + \ldots + \beta_{m,k} x'_{t,m,k} \]  
\[ (11) \]

According to (8), on the type equations right end expands to

\[ y'_{t,k} = \beta_0 + \beta_1 x_{t,k} + \beta_{2,k} x'_{t,k} + \ldots + \beta_{m,k} x'_{t,m,k} \]  
\[ (12) \]

\[ y'_{t,m,k} = \beta_0 + \beta_1 x_{t,m,k} + \beta_{2,k} x'_{t,m,k} + \ldots + \beta_{m,k} x'_{t,m,m,k} \]  
\[ (13) \]

The normalized amplitude change of the main variable \( y \) of \( K \)-th data segment is obtained \( \Delta y'_{k} \), which is calculated as

\[ \Delta y'_{k} = y'_{t,k} - y'_{t} = \beta_1 x_{t,0} + \beta_2 x_{t,1} + \ldots + \beta_{m} x_{t,m} \]  
\[ (14) \]

It is assumed that the multiple regression model has a good fitness. There is \( \Delta y'_{k} \) is equal to \( \frac{\Delta y_{k}}{A_{0,y}} \), so it can be represented as

\[ \Delta y_{k} = \frac{\beta_1}{A_{0,y}} \Delta x_{1,k} + \frac{\beta_2}{A_{0,y}} \Delta x_{2,k} + \ldots + \frac{\beta_{m}}{A_{0,y}} \Delta x_{m,k} \]  
\[ (15) \]

Finally, \( y'_{k} \) is obtained of \( x_{k} \), and the root-cause variable is determined as

\[ y_{k} > y_{0} \]  
\[ (16) \]

where \( y_{0} \) is selected by users.

3. Industrial example

This section presents an industrial example to illustrate the effectiveness of the proposed method. The example is from a large-scale 330MW thermal power units located at Shandong Province in China.

According to process knowledge, the condenser vacuum is a key observation variable for thermal power units. If the condenser vacuum is too high or too low, it will cause a major safety accident. The variables that affect condenser vacuum include \( x_{1} \) as the condenser inlet temperature with the unit \( ^{\circ}C \), \( x_{2} \) as the condensate flow with the unit \( t/h \) and \( x_{3} \) as the circulating water inlet temperature with the unit \( ^{\circ}C \). The abnormally threshold of condenser vacuum is set at a value of -90.7 Kpa. When the value overpasses-90.7 Kpa, and it is necessary to root-cause analysis for fault diagnosis.

Firstly, historical data are obtained from June to August in 2018 of the thermal power units with the sampling period of 1 sec. For each variable, the bottom-up algorithm is used here to obtain the piece-wise linear representations, which are shown in Fig 1 (red dash).

Secondly, for \( A_{0,y} \) of \( y \), \( x_{1} \), \( x_{2} \) and \( x_{3} \) are calculated by (3), which are 0.224 Kpa, 0.486 \( ^{\circ}C \), 46.901 \( t/h \), 0.152 \( ^{\circ}C \), respectively. The trends of each variable are determined by (5), and the root-cause combinations are obtained of history data. The statistic results of trend combinations are shown in Table 1.

| y | x1 | x2 | x3 | Total |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 134 |
| 1 | 0 | 1 | 0 | 39 |
| 1 | 1 | 1 | 0 | 124 |
| 1 | 1 | 0 | 1 | 139 |
| 1 | 1 | 1 | -1 | 24 |
| 1 | 1 | 1 | 1 | 55 |
Finally, for each variable, standardized data are calculated from (8), and $\gamma_i$ of each variable is calculated by (9), $\gamma_0$ is set 30% here. Fig 1 shows the results of root-cause analysis, which included different root-cause variables (yellow background). The root-cause variable are $x_1$ and $x_3$ with the trends combination [1, 1, 0, 0] and [1, 0, 0, 1] respectively, while the fitting of multiple regression model is also shown in the figure. The explanatory degree is the product of the percentage of the total contribution of each related variable and the model fit.

![Graph showing root-cause analysis results](image)

**Fig 1** The results of root-cause analysis: (a) with the trends [1, 1, 0, 0], (b) with the trends [1, 0, 0, 1].

Table 2 shows all root-cause analysis of historical data. In the trend combination of [1, 1, -1, 1], the root-cause variable is $x_1$, accounting for 75%, the root-cause variable is $x_2$, accounting for 9%, the root-cause variable is $x_3$, accounting for 16%.

| Trend combinations | $x_1$ (%) | $x_2$ (%) | $x_3$ (%) |
|--------------------|-----------|-----------|-----------|
| [1,1,0,0]          | 100       | 0         | 0         |
| [1,0,0,1]          | 0         | 0         | 100       |
| [1,1,1,0]          | 100       | 0         | 0         |
| [1,1,0,1]          | 92.8      | 0         | 7.2       |
| [1,1,1,-1]         | 75        | 9         | 16        |
| [1,1,1,1]          | 100       | 0         | 0         |

**4. Conclusion**

The existing root-cause analysis methods are difficult to obtain the correct root-cause analysis results when the relationships between the main variable and the related variables are non-stationary. This paper proposed a root-cause analysis method based on multiple regression models. The proposed method was able to obtain the correct root-cause variables when the relationships among variables were non-stationary. The proposed method was composed by three major steps: (i) the piece-wise linear representations were used to obtain qualitative trends of all variables, (ii) the amount of each...
variable’s contribution to the main variable was calculated by the multiple regression model, (iii) the root cause-variable was determined by the size of the contribution. An industrial example was provided to illustrate the effectiveness of the proposed method.

References
[1] Tao, S., Li, C. (2012) Fault detection strategy based on difference of score reconstruction associated with main component analysis. Control Theory & Applications, 29(02):178-180 (in Chinese)
[2] Zhang, C., Guo, Q., Li, Y., et al. (2019) Chemical process fault root diagnosis based on the sequence of variable abnormal. Computers and Applied Chemistry, 36(05):774-782 (in Chinese)
[3] Folmer, J., Pantforder D., Vogel-Heuser, B. (2011) An analytical alarm flood reduction to reduce operator’s workload. In: Human-Computer Interaction. Orlando. pp. 297-306.
[4] Gonzalez, J.M.N., Jimenez, J.A., Lopez, J.C.D., et al. (2017) Root cause analysis of network failures using machine learning and summarization techniques. IEEE Communications Magazine, 55(9): 126-131.
[5] Wang, J., Yang, Z., Su, J. (2018) Root cause analysis of occurring alarms in thermal power plants based on Bayesian networks. International Journal of Electrical Power & Energy Systems, 10(3): 67-74.
[6] Cauvin, S., Cordier, M.O., Dousson, C., et al. (1998) Monitoring and alarm interpretation in industrial environments. AI Communications, 11(34):139-173.
[7] Chiang, L.H., Braatz, R.D. (2003) Process monitoring using causal map and multivariate statistics: fault detection and identification. Chemometrics and Intelligent Laboratory Systems, 65(2): 159-178.
[8] Bauer, M., Thornhill, N.F. (2008) A practical method for identifying the propagation path of plant-wide disturbances. Journal of Process Control, 18(7-8): 707-719.
[9] Bauer, M., Cox, J.W., Caveness, M.H., et al. (2006) Finding the direction of disturbance propagation in a chemical process using transfer entropy. IEEE Transactions on Control Systems Technology, 15(1): 12-21.
[10] Hu, W., Wang, J., Chen, T., et al. (2017) Cause effect analysis of industrial alarm variables using transfer entropies. Control Engineering Practice, 65: 205-214.
[11] Yuan, T., Qin, S.J. (2014) Root cause diagnosis of plant-wide oscillations using granger causality. Journal of Process Control, 24(2): 450-459.
[12] Wang, J., Li, H., Huang, J., et al. (2015) A data similarity based analysis to consequential alarms of industrial processes. Journal of Loss Prevention in the Process Industries, 35: 29-34.
[13] Wang, J., Qian, Z., Chen, M., et al. (2017) Research on fault diagnosis of bond graph model based on temporal causal graph. Control Engineering of China, 24(12): 2443-2449. (in Chinese)
[14] Zhou, B., Ye, H. (2016) A study of polynomial fit-based methods for qualitative trend analysis. Journal of Process Control, 37:21-33.
[15] Wang, J., Yu, Y., Chen, K., (2019) Determining the number of segments for piece-wise linear representation of discretetime signals. Computers and Chemical Engineering, 120: 46-53.
[16] Wang, J., Pang, X., Gao, S., et al. (2019) Assessment of automatic generation control performance of power generation units based on amplitude changes. International Journal of Electrical Power & Energy Systems, 108:19-30.