Key Variable Screening Method for Complex Industry System Based on Multi-source Monitoring Data

Wei Wang, Juntai Xie*, Zhiyong Gao and Xiangpeng Chen
State Key Laboratory for Manufacturing Systems Engineering, Xi’an Jiaotong University, Xi’an 710049, China
*Email: xiejuntai@xjtu.edu.cn

Abstract. Aiming at the problem that the value density of variables in the state-aware network of complex industry system is low, which leads to poor timeliness of state evaluation, a key variable screening method for state-aware network is proposed. Firstly, a causal network model that can accurately reflect the interaction mode between the monitoring variables of the system is established. Secondly, each node of the causal network can be ranked by the LeaderRank algorithm and the variable set can be divided into multiple sets of variables. Finally, each variable set can be used to evaluate the performance state of the system, and the effectiveness index of variable screening is constructed to evaluate the accuracy and timeliness of the evaluation results of each variable set, then the key variables can be obtained. The Tennessee Eastman (TE) process data is used to test the proposed method, the result shows that the key variables obtained by the method can effectively improve the evaluation efficiency of the performance state.

1. Introduction
The process industry occupies an important position in the national economy. With the continuous improvement of demand, the scale of production equipment is getting larger and larger [1]. In order to effectively evaluate the performance state of the equipment, thousands of digital instruments and sensors to monitor the performance state of the system are installed, forming a state-aware network that can effectively perceive the performance state [2]. The characteristics of multi-point and large data volume result in the value density of variables in the state-aware network is low. How to screen out the key variables that can accurately reflect the performance state become necessary.

In order to solve the problem, Gao [3] established an information flow model according to the system structure. Through this model, the key information nodes of the system are obtained based on the PageRank algorithm. Liu [4] proposed a multi-attribute decision making method based on the TOPSIS to identify the key nodes in the complex water conservancy project. Liu [5] determines the vulnerable nodes according to the global active and reactive power effects of nodes in the power grid. These studies put forward different evaluation methods for node importance. However, it only identifies the key variables of the physical topology, and these methods cannot accurately reflect the information interaction mode between the monitoring variables of the system. It cannot be effectively applied to the key variables screening of complex industry system [6] with unclear correlation between nodes.

In view of the performance state is evaluated by calculating the coupling relationship between all the variables of system in the existing performance state evaluation technology [7][8], which leads to the problems that the evaluation process requires a large amount of computing resources and the existing key variables screening methods for complex systems are difficult to identify the key variables that can accurately reflect the system performance state based on the monitoring data [9], a key variables
screening method for the state-aware network is proposed. Based on the monitoring data of the system, a causal network model that reflects the information interaction mode of the monitoring variables of the system is constructed. The variables of the causal network are ranked and divided into multiple sets of variables, each set of variables is used to evaluate the performance state, and the evaluation result of all variables is used as a benchmark to construct the effectiveness index of variable screening $MCI$ to evaluate the accuracy and timeliness of each variable set. Then the key variables can be obtained and the effective screening of the key variables in the state-aware network can be realized.

2. The key variable screening methods
There are usually multiple monitor variables that are not strongly related to the performance state in the process industry production system. If we cannot effectively eliminate these variables when evaluating the system state, it will cause longer time be consumed. Based on this, the key variable screening method is proposed, as shown in Figure 1.

2.1. Monitoring data acquisition and pre-processing
The working conditions of the process industry fluctuate sharply, this will cause the monitoring data to be disturbed by noise. In addition, the dimensions of each variable are different. In order to prevent these problems from affecting the effectiveness of the analysis results, wavelet noise reduction and min-max standardization methods are used before analysis. The min-max standardization method normalizes the data by transforming each variable time series $x_1, x_2, \cdots, x_n$ as follows:
2.2. Analysis and modelling of causality between variables

Using Copula-Granger causality analysis method [10] to calculate the causality value of variables. In order to determine the threshold value of the causality between each variable, this paper constructs a Gaussian noise sequence. Using the causality value between each variable and Gaussian noise as the threshold value, and getting the causal matrix between the monitoring variables, as shown below:

\[
C = \begin{bmatrix}
0 & c_{12} & \cdots & c_{1n} \\
c_{21} & 0 & \cdots & c_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
c_{n1} & c_{n2} & \cdots & 0
\end{bmatrix}
\]

(2)

In the formula, \(c_{ij}\) represents the causality value between variable \(i\) and \(j\).

2.3. Partition of variable set in the state-aware network

Few special nodes play a role in the causal network. Therefore, this paper uses the Leader Rank algorithm [11] to rank the variables of the causal network. According to the ranking results, multiple sets of variables are obtained at a certain proportion. Then the multiple sets of variables are used to evaluate the multiple fault states of the system. The specific steps are as follows:

1. Selecting a set of historical monitoring data \(X^n_1, X^n_2, \ldots, X^n_l\) corresponding to the variable sets \(n_1, n_2, \ldots, n_k\), where \(k\) is the number of nodes in the variable set, \(l\) is the length of the monitoring data.
2. Dividing the monitoring data with a length of \(l\) by the sliding window method, the width of the window is \(T\), and the sliding step is \(SST\), the obtained \(n\) sets of monitoring data samples are as follows:

\[
n = \text{floor} \left( \frac{l - T}{SST} \right)
\]

(3)

3. Obtaining the monitoring data of each variable, and calculating the causality value of the monitoring data in each window, and obtaining the dynamic causal network that can reflect the evolution mode of information interaction among the variables in real time. The network structure entropy of the causal network in the window are used to evaluate the performance state.

\[
\hat{E} = -\frac{2}{2 \ln N - \ln 4(N - 1)} \sum_{i=1}^{N} P(k_i) \cdot \ln P(k_i) - \ln 4(N - 1)
\]

(4)

In the formula, \(P(k_i)\) is the intensity distribution of the nodes, \(k_i\) is the node intensity of node \(i\), and \(N\) is the number of variables in the causal network.

2.4. The key variable screening in the state-aware network

This paper calculates the absolute percentage error, contour similarity and the calculation time of the evaluation results between each set of variables and all variables on the system performance state, and uses the entropy weight method to fuse each indicator to obtain the effectiveness index of variable screening \(MCI\) to evaluate the accuracy and timeliness of each variable set. So as to screen out the key variables that can effectively characterize the performance state. The specific steps are as follows:

1. Calculating the average absolute percentage error \(MAPE\) of each variable set:

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y(i) - x(i)}{x(i)} \right|
\]

(5)

In the formula, \(x_i\) and \(y_i\) are the \(i\)-th evaluation result of each variable set \(X\) and all variables \(Y\).
2. Respectively normalizing the evaluation results \(X\) and \(Y\) to eliminate the influence of numerical differences on the similarity judgment results.
3. Using the Pearson correlation coefficient to calculate the contour similarity \(CS\) of the evaluation results \(X\) and \(Y\), as follows:
\[ CS = \frac{\sum_{i=1}^{N}(x(i) - \bar{x})(y(i) - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x(i) - \bar{x})^2 \sum_{i=1}^{N}(y(i) - \bar{y})^2}} \]  

(6)

In the formula, \( \bar{x} \) and \( \bar{y} \) are the averages of all samples of \( X \) and \( Y \) respectively.

(4) Calculating the root mean square of the contour similarity \( CS \) and the absolute percentage error \( MAPE \), and obtaining the accuracy evaluation index \( Era \):

\[ Era = \sqrt{\frac{(1 - CS)^2 + MAPE^2}{2}} \]  

(7)

(5) Obtaining the unit window evaluation time-consuming index \( Ute \) according to the calculation time of the single sliding window.

(6) Calculating the information entropy of the evaluation indicators \( Era \) and \( Ute \):

\[ E_j = -\ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \]  

(8)

Among them, \( p_{ij} = Y_{ij}/\sum_{i=1}^{n} Y_{ij} \), \( j \) represents the \( j \)-th evaluation index, \( i \) represents the \( i \)-th dimension feature of each evaluation index, \( Y_{ij} = [era_1, ute_1; erase_2, ute_2; \ldots; era_k, ute_k] \). \( E_j \) is the information entropy of each evaluation index.

(7) According to the information entropy \( E_j \) of each indicator, calculating its weight:

\[ W_j = \frac{1 - E_j}{k - \sum E_j} \]  

(9)

(8) Normalizing the data of the \( Era = \{era_1, erase_2, \ldots era_k\} \) and \( Ute = \{ute_1, ute_2, \ldots ute_k\} \) to obtain the normalized evaluation indicators \( Era' \) and \( Ute' \).

(9) Through the index \( Era' \), \( Ute' \) and its corresponding weights \( W_1 \), \( W_2 \), the effectiveness index of variable screening \( MCI \) is obtained:

\[ MCI = Era' \cdot W_1 + Ute' \cdot W_2 \]  

(10)

The set of key variables with the smallest value of the \( MCI \) is the key variables in the state-aware network of complex industry system.

3. Instance verification

3.1. Monitoring data acquisition

As the classic data of chemical process, TE process data set has been widely used in the fields of fault diagnosis and state evaluation in chemical process. This paper simulates the monitoring data of 22 variables in TE process under normal performance state and various fault states.

3.2. Causality modelling and the ranking of variables

In order to avoid the identification of false causality, through experiments, the 1.5 times causality value between each monitored variable and Gaussian noise is selected as the threshold value to eliminate false causality between variables. Taking each monitoring variable as the node, and the causality between the variables as the edge weight, a causal network model reflecting the mode of information interaction between the variables is constructed.

This paper mainly ranks each variable by extracting the characteristics that can reflect the interactive behaviour of the variable in the network. The Leader Rank algorithm is used to calculate the LR value of each variable in the network, and the result is shown in Figure 2.

3.3. The screening of key variables

Obtaining the monitoring data of the 22 variables about the TE process, and using the wavelet noise reduction and min-max standardization methods to deal with the data. The sliding window method is
used to divide the monitoring data of each variable, the width of the window $T$ is 500, the sliding step $SST$ is 50, and 278 sets of monitoring data samples are obtained.

Using the Copula-Granger causality analysis method to calculate the causality value of the monitoring data in each window, a dynamic causal network that can reflect the evolution mode of information interaction between variables in real time is obtained, and the network structure entropy of the causal network in each window are extracted, which is the evaluation result of the system performance state.

Calculating the average absolute percentage error $MAPE$ and the contour similarity $CS$ of the evaluation results between each variable set and all variables, and obtaining the index $Era$ that can effectively characterize the accuracy of the evaluation results. The index $Ute$ is obtained according to the evaluation time of each unit window. The effectiveness index of variable screening $MCI$ is calculated with the index $Era$ and $Ute$ by the entropy method.

The three indicators $Era$, $Ute$, and $MCI$, which reflect the effectiveness of the evaluation results of each variable set, are normalized, and the chart of three indicators change trend as shown in Figure 3.

Figure 2. The chart of LR value and ranking of each variable

Figure 3. The chart of main indicators change trend

It can be seen from Figure 3 that with the expansion of the ratio, the index $MCI$ decreases first and then increases. The extreme value is reached when the ratio is 50%.

(a) The causality heat map among key variables   (b) The causality heat map among all variables

Figure 4. The causality chart of all and 50% key variables

Comparing Figure 4 (a) and (b), we can find that the causality between the key variables is more significant than all variables, it indicates that this method has screened out most of the variables with insignificant impact on the performance state, and can effectively improve the evaluation efficiency.
4. Conclusion

Aiming at the problem that the value density of some variables in state-aware network of complex industry system in the process industry is low, which leads to the poor timeliness of system state evaluation, this paper proposes a key variable screening method for state-aware network driven by monitoring data. Compared with the traditional key variables screening method of complex systems, this method can use the monitoring data to find out the mode of information interaction between variables, through the construction of the causal network model between the variables, the ranking of variables, the variable set division and the validity evaluation of variable realize the key variables screening in the state-aware network of complex industry system.

Using TE process data to verify, the results show that the key variables screened by the method in this paper can effectively characterize the performance state, and improve the evaluation efficiency.

Acknowledgment

This research was funded in part by the National Key R&D program of China, Grant number 2019YFF0217500, in part by the open research project fund of the key laboratory of quality infrastructure effectiveness research, Grant number KF20180301, and in part by the science and technology plan project of Xi'an, Grant number 2019521415KYPT005JC007. The authors would also like to thank the Institute of New Energy Equipment and Quality Engineering of Xi'an Jiaotong University.

References

[1] Qian Y, Yang S Y, Jia X P, Li X X and Li H C. 2013 J. Life cycle assessment and sustainability of energy and chemical processes. CIESC, 64 133-147.
[2] Xu G N, Gao Z Y, Liang Y J, Gao J M, Liu Q Q and Cheng Y H. 2020 J. Verification and restoration method of abnormal monitoring data by compressive sensing for process industry. Journal of Xi'an Jiaotong University, 54 59-70.
[3] Gao Z Y, Guo C and Gao J M. 2016 J. Determination method for key points of information quality control in process industry system. Computer Integrated Manufacturing Systems. 22 1323-28.
[4] Liu Z H, Yu H and Yang F T. 2014 J. Evaluate the node importance for water network based on complex network theory. Sci Sin Tech. 44 1280-94.
[5] Liu X L, Mao Y, Liang S, Li Y X, Fan X, Deng H C and Peng W Q. 2016 J. Identification of vulnerable lines in power grid based on comprehensive betweenness index. Power System Protection and Control, 044 116-121.
[6] Wijayasekara D, Linda O, Manic M and Rieger C. 2014 J. FN-DFE: fuzzy-neural data fusion engine for enhanced resilient state-awareness of hybrid energy systems. IEEE Transactions on Cybernetics. 44 2016-75.
[7] Liang Y J, Gao Z Y, Gao J M, Wang R X, Liu Q Q and Cheng Y H. 2020 J. A new method for multivariable nonlinear coupling relations analysis in complex electromechanical system. Applied Soft Computing. 94.
[8] Xie J T, Gao J M, Gao Z Y, Wang R X and Wang Z. 2018 J. Application Research of Multivariate Linkage Fluctuation Analysis on Condition Evaluation in Process Industry. Science China-Technological sciences. 61. 397-407.
[9] Zheng R C, Gu J, Jin Z J, Peng H Q and Cai L. 2020 J. Research on Short-term Load Forecasting Variable Selection Based on Fusion of Data Driven Method and Forecast Error Driven Method. Proceedings of the CSEE. 40. 487-500.
[10] Hu M and Liang H. 2014 J. A copula approach to assessing Granger causality. Neuro Image. 100. 125-134.
[11] Gen X Q, She Q S, Zhang Q Z and Luo Z Z. 2018 J. Multivariate Causality Analysis Method of Motor Imagery EEG Signals Based on Copula. Space Medicine & Medical Engineering. 031, 49-56.