IOP Conf. Series: Materials Science and Engineering **339** (2018) 012039 doi:10.1088/1757-899X/339/1/012039

Fault detection of Tennessee Eastman process based on topological features and SVM

Huiyang Zhao [1,2,*] , Yanzhu Hu [1] , Xinbo Ai [1] , Yu Hu [1] and Zhen Meng [1]

[1] Beijing Key Laboratory of Work Safety Intelligent Monitoring, Beijing University of Posts and Telecommunications, Beijing 100876, China

[2] School of Information Engineering, Xuchang University, Xuchang 461000, China

Email: zhaohy@bupt.edu.cn

Abstract. Fault detection in industrial process is a popular research topic. Although the distributed control system(DCS) has been introduced to monitor the state of industrial process, it still cannot satisfy all the requirements for fault detection of all the industrial systems. In this paper, we proposed a novel method based on topological features and support vector machine(SVM), for fault detection of industrial process. The proposed method takes global information of measured variables into account by complex network model and predicts whether a system has generated some faults or not by SVM. The proposed method can be divided into four steps, i.e. network construction, network analysis, model training and model testing respectively. Finally, we apply the model to Tennessee Eastman process(TEP). The results show that this method works well and can be a useful supplement for fault detection of industrial process.

1. Introduction

The problem of fault detection and diagnosis is an essential part of industrial process. It is very important for production. The methods of fault detection can be broadly classified into two general categories: model-based methods and data-driven methods. Generally speaking, the model-based methods need a priori knowledge or a specific mathematic model. However, it is very difficult for some industrial systems. Different from the model-based methods, the data-driven methods are only dependent on the monitored process variables. Thus, the data-driven methods have been extensively studied and developed over the past few decades. We will review some data-driven methods as follows.

KNN(K-nearest neighbors) is a data-driven method used in fault detection. Xiong et al. [1] proposed an information fusion fault diagnosis method that is based on a static discounting factor and combines KNN with dimensionless indicators. Wang et al. [2] proposed a novel fault diagnosis method derived using KNN reconstruction on maximize reduce index (MRI) sensors. Tennessee Eastman (TE) process was provided to demonstrate that the proposed approach can identify the responsible variables for the multiple sensors fault.

Support vector machine (SVM)[3] is also an approach used in fault detection. Wu et al. [4] provided a combined measure of the original SVM and PCA (principle component analysis) to carry out the fault classification, and compared its result with what is based on SVM-RFE (Recursive Feature Elimination) method. Mahadevan and Shah [5] proposed a new approach for fault detection

and diagnosis based on One-Class Support Vector Machines. Gao and Hou [6] propose a multi-class support vector machine based on process supervision and fault diagnosis scheme to predict the status of the TE process.

Of course, there are many more data-driven methods for fault detection and we have not mentioned above, such as Decision tree classifier[7], Bayesian classifier [8], neural network[9], random forest[10], PLS (partial least squares)[11], and so on.

However, conventional data-based methods have two problems. One problem is dimension reduction for large-scale industrial processes and the other is that this kind of methods do not consider the relationships among different measured variables and the effect of single measured variables on the overall performance of the system [12].

To solve the problems described above, we proposed a novel method, integrated complex network theory with support vector machine and apply it to Tennessee Eastman process(TEP) which is a well-known typical industrial process.

2. Methods

The main principle of the proposed method is described in figure 1.

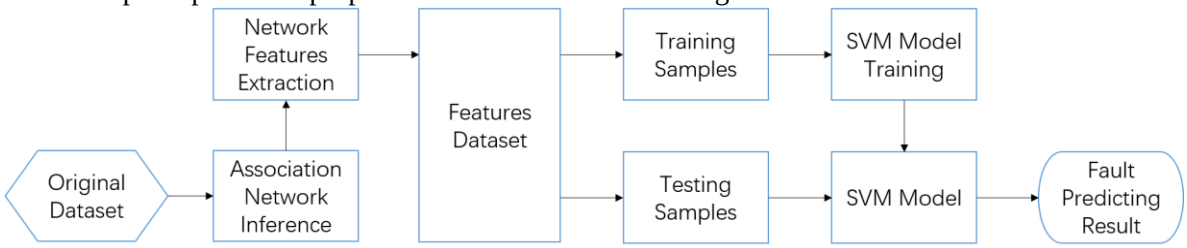

Figure 1. The main principle of the proposed method for fault detection.

From figure 1, we can see the detail of the proposed method. The start of the proposed method is original dataset, generally in form of multivariate time series. Then, the association network is inferred with SSPSTESGC [13, 14]. From the resulted network, some topological features are extracted and analyzed. Next, we will generate a features dataset which are composed of those topological features. This features dataset is considered as the sample of our selected classifier model which is SVM. Fault detection is a problem of classification. Thus, the features dataset is divided into training samples and testing samples. The training samples are used to train a SVM model. With the SVM model and testing samples, we can predict whether some faults have occurred in the observed system. The procedures of fault detection with the proposed method are described in detail as follows.

2.1. Network inference

Some essential methods for association network inference have been proposed for association networks inference, such as correlation [15], information theory [16], Granger causality (GC) [17], neural network [18], and so on.

We have proposed SSPSTESGC for association networks inference in [13, 14] and the good performance have been proved by experiments. Thus, it is selected as a part of the proposed method in this paper. The main principles of SSPSTESGC are the small-shuffled surrogate (SSS) method proposed in [19-21] and the partial symbolic transfer entropy(PSTE) shown in [22].

The primary task for inferring a network is how to define the relations between nodes. PSTE has been proved properly for this work. It is defined conditioning on the set of the remaining time series $z=\{v_3,v_4,\cdots,v_n\}$ and represented by Equation (1).

$$PSTE_{v_2\to v_1}=\sum p(\hat{v}_{1,t+\tau},\hat{v}_{1,t},\hat{v}_{2,t},\hat{z}_t)\log\frac{p(\hat{v}_{1,t+\tau}\mid\hat{v}_{1,t},\hat{v}_{2,t},\hat{z}_t)}{p(\hat{v}_{1,t+\tau}\mid\hat{v}_{1,t},\hat{z}_t)} \tag{1}$$

where the rank vector $\hat{z}_t$ is defined as the concatenation of the rank vectors for each of the embedding vectors of the time series in z. The partial symbolic transfer entropy can eliminate some of the indirect correlation and remain the pure or direct information flow between v_2 and v_1.

Partial Symbolic Transfer Entropy Spectrum(PSTES) is defined as follows:

The PSTES between time series Y and X is composed of their many partial symbolic transfer entropy curves drawn in a rectangular coordinate system. The horizontal axis represents different time delays and the vertical axis represents transfer entropy. One of the transfer entropy curves is resulted from original data and other curves are resulted from shuffled data.

2.2. Network analysis

Many topological measures have been proposed for complex network studying. Topological measures can be divided in two groups, i.e., measures at global network level and measures at local node level [23, 24], corresponding to the measurable element. Since the observed object in our study is the network as a whole, only those graph-level measures will be selected. In other words, node-level such as the degree of a certain node will not be taken into account. The topological features selected in the proposed method include average degree(AD), diameter(DIA), average path length(APL), density(DEN), clustering coefficient(CLU), degree centralization(DC), closeness centralization(CC), betweenness centralization(BC) and eigenvector centralization(EC).

2.3. Fault detection

After the step of network analysis, we will fit a model of fault detection by SVM. The topological features dataset is randomly split into training and testing groups. The ratio of training dataset is 70% and the ratio of testing dataset is 30%. When fitting an SVM model, a parameter should be specified, i.e. kernel function.

3. Results

3.1. The description of TEP

The TE process is a widely used realistic simulation program for chemical plants and has been widely accepted as a benchmark for control and monitoring studies [25, 26]. The process consists of five major transformation units: the reactor, the product condenser, the vapor-liquid separator, the recycle compressor and the product stripper.

In our experiment, 22 process measurements [27] are selected as original dataset. The TE process contains 21 preprogrammed faults[28] which are described. These faults are divided into six types, i.e. step, random variation, slow drift, sticking, unknown and constant position. 22 data sets (1 normal data set and 21 fault data sets) are generated from the TE process. With a sampling interval of 3 minutes, 960 observations are generated for each data set. All faults are introduced from the 161th observation.

3.2. Fault detection of TEP

3.2.1. Association network construction. In this section, the method SSPSTESGC [13, 14] will be used to infer the association network. The result of applying SSPSTESGC method is called PSTE spectrum between each pair of variables, such as two figures shown in figure 2 and figure 3. The value of horizontal axis is time delay and the value of vertical axis is PSTE. According the rule whether there is a relation from one variable to the other variable, we can conclude that *X7* is influenced by *X6* from figure 2 and *X2* is not influenced by *X1* from figure 3.

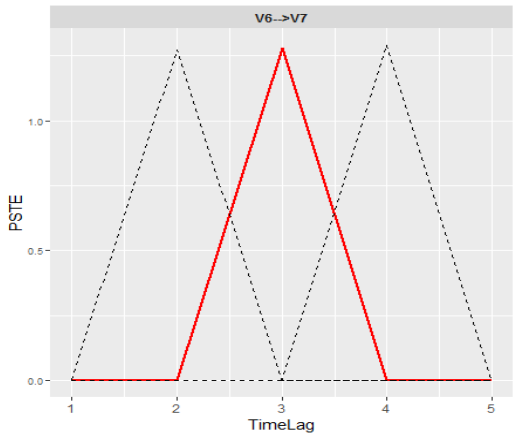

Figure 2. PSTE Spectrum from variable *V6* to *V7*.

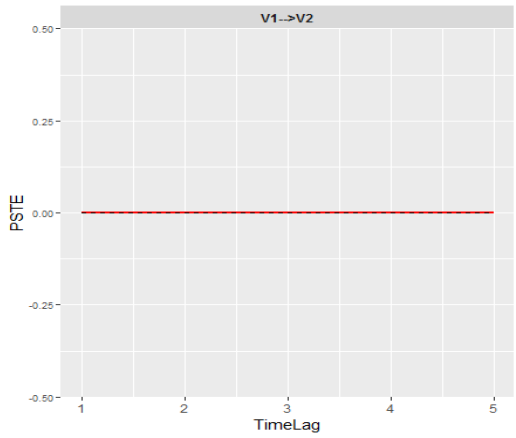

Figure 3. PSTE Spectrum from variable *V1* to *V2*.

According to the PSTE spectrums and the rule of network construction, the network at each time can be inferred. For example, the network with normal state is shown in figure 4 and the network with fault 1 is shown in figure 5. There are some differences between the two networks. For instance, the core vertex with maximal degree is *V6* in figure 4. In figure 5, the network changes with the occurrence of fault 1. The vertex with maximal degree changes to be *V11*. Additionally, there are some other differences between two networks. The degree of some vertex changes to be larger, such as *V7*, *V16* and *V22*. However, the degree of *V21* changes to be smaller. Moreover, some vertex changes from isolate vertex to normal vertex, such as *V2* and *V5*. However, some vertex changes from normal vertex to isolate vertex, such as *V1* and *V9*.

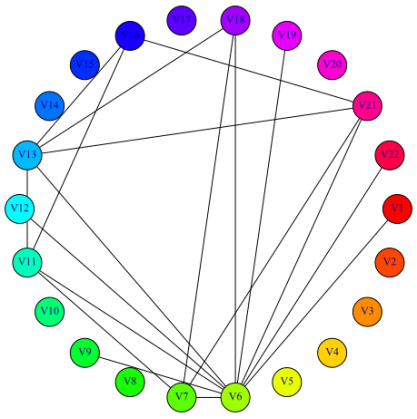

Figure 4. The network with normal state at time point 181.

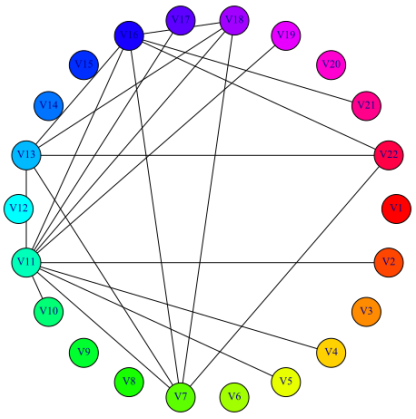

Figure 5. The network with fault 1 at time point 181.

3.2.2. Network analysis with complex network theory. Network features are calculated for each network and part of the results are shown in Table 1. The abbreviations SS represents the state of the system. This column has 22 different values, i.e. N (normal state), F1 (fault 1), F2 (fault 2) and so on.

Table 1. The topological features extracted from association networks.

ID	AD	DIA	APL	DEN	CLU	DC	CC	BC	EC	SS
1	1	2	1.56	0.02	0.29	0.12	0.02	0.06	0.9	N
2	1.27	2	1.48	0.03	0.44	0.14	0.02	0.05	0.89	N
3	1.09	2	1.54	0.03	0.38	0.12	0.02	0.05	0.89	N
4	1.73	4	2.18	0.04	0.12	0.23	0.05	0.21	0.87	F1
5	2.73	4	2.2	0.06	0.18	0.46	0.07	0.39	0.86	F1
6	2.36	4	2.07	0.06	0.43	0.29	0.06	0.26	0.82	F1
7	0.64	3	1.67	0.02	0	0.06	0.02	0.02	0.86	F2

8	0.64	5	2.29	0.02	0	0.06	0.02	0.04	0.88	F2
9	0.55	3	1.55	0.01	0	0.06	0.02	0.02	0.9	F2

3.2.3. Fault detection and evaluation. In this section, we will carry out fault detection by applying the SVM algorithm to the network features extracted above. With the fitted model, we can predict the system state is either normal or abnormal on testing dataset. In order to assess the performance of the proposed method, we introduce into three measures, i.e. precision, recall and accuracy.

The results of fault detection are shown in table 2. With the proposed method, precision is 0.84, the recall is 0.99 and the accuracy is 0.89. As a comparison, the performances of some other methods are also shown in table 2 As a whole, the proposed method is superior to the other five methods. Compared to precision, we pay more attention to the recall measure.

Table 2. The model assessment and comparison.

ID	Method	Precision	Recall	Accuracy
1	NFSVM	0.84	0.99	0.89
2	Naive Bayes	0.77	0.89	0.79
3	KNN	0.78	0.92	0.81
4	Decision Tree	0.79	0.95	0.84
5	Random Forest	0.81	0.96	0.86
6	Neural Network	0.79	0.96	0.85

4. Conclusions

In this paper, based on the complex network theory and SVM, we have proposed a novel method for fault detection of industrial systems. We applied the proposed method to TEP and the performance of proposed method is evaluated by three measures. As a result, the method makes a good performance on fault detection. However, there are still some topics that are worth studying in future.

References

[1] Xiong J, Zhang Q, Sun G, Zhu X, Liu M and Li Z 2016 An information fusion fault diagnosis method based on dimensionless indicators with static discounting factor and KNN *IEEE Sensors Journal* **16** 2060-9
[2] Wang G, Liu J and Li Y 2015 Fault diagnosis using kNN reconstruction on MRI variables *Journal of Chemometrics* **29** 399-410
[3] Cortes C and Vapnik V 1995 Support-vector networks *Machine Learning* **20** 273-97
[4] Wu F, Yin S and Karimi H R 2014 Fault detection and diagnosis in process data using support vector machines *Journal of Applied Mathematics* **2014**
[5] Mahadevan S and Shah S L 2009 Fault detection and diagnosis in process data using one-class support vector machines *Journal of process control* **19** 1627-39
[6] Gao X and Hou J 2016 An improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastman process *Neurocomputing* **174** 906-11
[7] Ye F, Zhang Z, Chakrabarty K and Gu X 2016 Adaptive board-level functional fault diagnosis using incremental decision trees *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **35** 323-36
[8] Duda R O, Hart P E and Stork D G 2012 *Pattern classification*: John Wiley & Sons)
[9] Eslamloueyan R 2011 Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee–Eastman process *Applied soft computing* **11** 1407-15
[10] Cerrada M, Zurita G, Cabrera D, Sánchez R-V, Artés M and Li C 2016 Fault diagnosis in spur gears based on genetic algorithm and random forest *Mechanical Systems and Signal Processing* **70** 87-103
[11] Khan A A, Moyne J R and Tilbury D M 2008 Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares *Journal of Process Control* **18** 961-74

[12] Cai E, liu D, Liang L and Xu G 2015 Monitoring of chemical industrial processes using integrated complex network theory with PCA *Chemometrics and Intelligent Laboratory Systems* **140** 22-35
[13] Hu Y, Zhao H and Ai X 2016 Inferring Weighted Directed Association Networks from Multivariate Time Series with the Small-Shuffle Symbolic Transfer Entropy Spectrum Method *Entropy* **18**
[14] Hu Y, Zhao H and Ai X 2016 Inferring Weighted Directed Association Network from Multivariate Time Series with a Synthetic Method of Partial Symbolic Transfer Entropy Spectrum and Granger Causality *PLOS ONE* **11** e0166084
[15] Guo X, Zhang Y, Hu W, Tan H and Wang X 2014 Inferring Nonlinear Gene Regulatory Networks from Gene Expression Data Based on Distance Correlation *PLoS ONE* **9** e87446
[16] Margolin A A, Wang K, Lim W K, Kustagi M, Nemenman I and Califano A 2006 Reverse engineering cellular networks *Nat. Protocols* **1** 662-71
[17] Schiatti L, Nollo G, Rossato G and Faes L 2015 Extended Granger causality: a new tool to identify the structure of physiological networks *Physiological Measurement* **36** 827
[18] Mahdevar G, Nowzaridalini A and Sadeghi M 2013 Inferring gene correlation networks from transcription factor binding sites *Genes & Genetic Systems* **88** 301-9
[19] Small M 2005 *Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance*
[20] Nakamura T, Hirata Y and Small M 2006 Testing for correlation structures in short-term variabilities with long-term trends of multivariate time series *Physical Review E* **74** 041114
[21] Nakamura T, Tanizawa T and Small M 2016 Constructing networks from a dynamical system perspective for multivariate nonlinear time series *Phys Rev E* **93** 032323
[22] Ai X 2014 Inferring a Drive-Response Network from Time Series of Topological Measures in Complex Networks with Transfer Entropy *Entropy* **16** 5753-76
[23] Sun X and Wandelt S 2014 Network similarity analysis of air navigation route systems *Transportation Research Part E: Logistics and Transportation Review* **70** 416-34
[24] Durek P and Walther D 2008 The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles *BMC Systems Biology* **2** 100
[25] Maurya M R, Rengaswamy R and Venkatasubramanian V 2007 Fault diagnosis using dynamic trend analysis: A review and recent developments *Engineering Applications of artificial intelligence* **20** 133-46
[26] Golshan M, Pishvaie M R and Boozarjomehry R B 2008 Stochastic and global real time optimization of Tennessee Eastman challenge problem *Engineering Applications of Artificial Intelligence* **21** 215-28
[27] Cai E, Liang L and Xu G 2015 Monitoring of chemical industrial processes using integrated complex network theory with PCA *Chemometrics and Intelligent Laboratory Systems* **140** 22-35
[28] Chiang L H, Russell E L and Braatz R D 2000 *Fault detection and diagnosis in industrial systems*: Springer Science & Business Media)

Acknowledgments
We thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This paper is supported by the National Natural Science Foundation of China (61503034), by the National Key Research and Development Program of China (2016YFC0701309-01), the National Natural Science Foundation of China (61627816), by the Key Scientific Research Project of Henan Province Universities (15B520031), and by the Xuchang Science and Technology Program (1502098). We thank all the fund organizations for offering us the studying conditions.