World Academy of Science, Engineering and Technology
International Journal of Chemical and Molecular Engineering
Vol:6, No:12, 2012

Diagnosis of Multivariate Process via Nonlinear Kernel Method Combined with Qualitative Representation of Fault Patterns

Hyun-Woo Cho

***Abstract*—**The fault detection and diagnosis of complicated production processes is one of essential tasks needed to run the process safely with good final product quality. Unexpected events occurred in the process may have a serious impact on the process. In this work, triangular representation of process measurement data obtained in an on-line basis is evaluated using simulation process. The effect of using linear and nonlinear reduced spaces is also tested. Their diagnosis performance was demonstrated using multivariate fault data. It has shown that the nonlinear technique based diagnosis method produced more reliable results and outperforms linear method. The use of appropriate reduced space yielded better diagnosis performance. The presented diagnosis framework is different from existing ones in that it attempts to extract the fault pattern in the reduced space, not in the original process variable space. The use of reduced model space helps to mitigate the sensitivity of the fault pattern to noise.

***Keywords*—**Real-time Fault diagnosis, triangular representation of patterns in reduced spaces, Nonlinear kernel technique, multivariate statistical modeling.

I. Introduction

THE impact of abnormal process operations is enormous both on safety and cost. To ensure safety and stability, it is necessary to continuously monitor the process operations, detect and diagnose process abnormalities, and take appropriate remedial actions. The fault diagnosis is to identify an assignable cause of the detected abnormal events, which helps operators to ensure productivity and final product quality [1][2]. On the other hand, the availability of on-line process data in most industrial processes has motivated the study of data-driven diagnosis methods. For the fault diagnosis, many multivariate statistical techniques have been developed: principal component analysis, partial least squares, and Fisher discriminant analysis [3]-[6]. These multivariate statistical techniques have been adopted frequently because of the sensors and data measurement technology. In addition, some researchers have also developed various techniques like wavelet transforms and multi-scale PCA [7][8].

There has been much interest in nonlinear kernel-based statistical learning methods such as support vector machines [9]. They have the common characteristics that input data are mapped into a nonlinear space and then these mapped data are analysed. The use of such a kernel trick enables us to develop various kernel methods including, kernel PCA, kernel PLS and kernel FDA [10][11]. In this work, a nonllinear kernel-based fault diagnosis is presented. To represent qualitative fault pattern in reduced spaces triangular representation method is combined with kernel PCA with an emphasis on improving on-line diagnosis performance. Among many fault diagnosis approaches, multivariate statistical methods depend on process measurement data to build empirical diagnosis models. Due to the nature, they are easy to construct, computationally efficient and relatively robust to noise [2]. In this area, various techniques have been utilized such as contribution plots, wavelet transforms, neural networks , multi-scale PCA and discriminant analysis [2][9]. To evaluate the diagnosis performance, the diagnosis performance using linear and nonlinear kernel combined with triangular representation of process data is demonstrated using multivariate simultation data of Tenneessee Eastman process.

Hyun-Woo Cho is with the Department of Industrial and Management Engineering, Daegu University, 712-714 Kyungsan, Republic of Korea (phone: +82-53-850-6547; fax:+82-53-850-6549; e-mail: hwcho@daegu.ac.kr)

II. Methodologies

A. Linear and Nonlinear Methods

Principal component analysis (PCA) decomposes a large number of correlated original variables into an uncorrelated set of principal components. Informative and relevant information of raw data can be summarized in scores. When the original variables are highly correlated, several PCs are sufficient to explain the major behavior of the data. The remaining ones explain the noise of the data, and noise-filtering is done by excluding them from further analysis [1]. To derive nonlinear kernel version of PCA, called kernel PCA (KPCA), it is necessary to solve the eigenvalue problem $\lambda \mathbf{v} = \mathbf{C}^{\mathrm{F}}\mathbf{v}$. Here, $\mathbf{C}^{\mathrm{F}}$ is the M sample estimate of the covariance matrix in the feature space F:

$$\mathbf{C}^{F} = \frac{1}{M}\sum_{j=1}^{M} \Phi(\mathbf{x}_j)\Phi(\mathbf{x}_j)^{\mathrm{T}} \tag{1}$$

where $\Phi(\cdot)$ is a nonlinear function. The eigenvalue equation can be written as

$$\lambda\langle \Phi(\mathbf{x}_k), \mathbf{v}\rangle = \langle \Phi(\mathbf{x}_k), \mathbf{C}^{F}\mathbf{v}\rangle \tag{2}$$

and there exists coefficients α_i, $i = 1, \ldots, M$, such that

$$\mathbf{v} = \sum_{j=1}^{M} \alpha_j \Phi(\mathbf{x}_j) \cdot \quad (3)$$

Combining (1), (2) and (3) yields (Schölkopf et al. 1998)

$$\lambda \sum_{j=1}^{M} \alpha_j \langle \Phi(\mathbf{x}_k), \Phi(\mathbf{x}_j) \rangle = \frac{1}{M} \sum_{j=1}^{M} \alpha_j \left\langle \Phi(\mathbf{x}_k), \sum_{i=1}^{M} \Phi(\mathbf{x}_i) \right\rangle \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \cdot \quad (4)$$

The principal components for a test vector x, are calculated by:

$$t_k = \langle \mathbf{v}_k, \Phi(\mathbf{x}) \rangle = \sum_{j=1}^{M} \alpha_j^k \langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}) \rangle = \sum_{j=1}^{M} \alpha_j^k \mathrm{k}(\mathbf{x}_j, \mathbf{x}) \cdot \quad (5)$$

B. Triangular Representation of Process Trends

A triangular representation method was proposed by [12] to extract process features in a systematic manner. The qualitative state of $\mathbf{x}(t)$, QS($\mathbf{x}$,t), is defined with the triplet as QS($\mathbf{x}$,t)= $< [\mathrm{x(t)}], [\partial \mathrm{x(t)}], [\partial\partial \mathrm{x(t)}] >$. If QS($\mathbf{x}$,$t$) remains constant, it means the uniform pattern or trend during that time interval. Basically, there are seven basic triangular components, which are determined by the first and second derivatives, as shown in Table I. For example, a triangular component with positive $[\partial \mathrm{x}]$ and $[\partial\partial \mathrm{x}]$ shows the pattern of concave upward and monotonic increase. For the representation of process data, triangular components serve as the geometric primitives so that any trend can be represented by a series of triangular components. Thus, such triangular representation of a process trend helps us to model the important features. In terms of fault diagnosis, it is actually the fingerprint of a fault that may appear in different magnitude or time duration.

TABLE I
SEVEN BASIC TRIANGULAR COMPONENTS

True Cause	COMPONENT		
	$[\partial \mathrm{x}]$	$[\partial\partial x]$	Description
1	0	0	Constant
2	+	0	Linear Increase
3	-	0	Linear Decrease
4	+	+	Concave Upward/Monotonic Increase
5	+	-	Concave Downward/Monotonic Increase
6	-	-	Concave Downward/Monotonic Decrease
7	-	+	Concave Upward/Monotonic Decrease

First, when a fault is detected, the fault data are projected onto a reduced space of PCA or KPCA to obtain score values. For this purpose, the measurement values of the process variables after the detection of the fault are recorded. By extracting the fault pattern via the triangular representation method, we can compare the extracted fault pattern with the existing fault patterns. More specifically, fault pattern vectors $\mathbf{y}(\mathrm{j})$ can be accumulated over time, which is represented as $\mathbf{y}(\mathrm{j})=[(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_j)]^{\mathrm{T}}$. Here, $\boldsymbol{y}_{\mathrm{i}}$ is a (7×1) fault element vector at the ith sequence with $\mathbf{y}_{\mathrm{i}} = [(\mathrm{y}_{1\mathrm{i}}, \mathrm{y}_{2\mathrm{i}}, \cdots, \mathrm{y}_{6\mathrm{i}}, \mathrm{y}_{7\mathrm{i}})]^{\mathrm{T}}$, in which each element of $\boldsymbol{y}_{\mathrm{i}}$ is either zero or one indicating the presence (absence) of triangular components. As shown in Fig. 1, for example, suppose the fault patterns observed are 2-4-7 for sequence 1, 2, and 3. Then the pattern vector at the first sequence $\mathbf{y}_1 = [0100000]^T$, at the second sequence $\mathbf{y}_2 = [0001000]^T$ and at the third sequence $\mathbf{y}_3 = [0000001]^T$. Thus in this case, we can obtain $\boldsymbol{y}(3)$ as $[0100000\ \ 0001000\ 0000001]^T$.

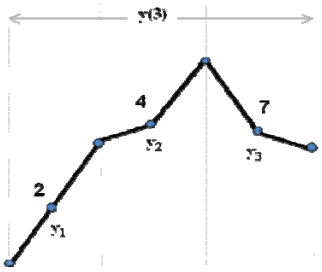

Fig. 1 An example of triangular representation

To perform pattern matching for diagnosis, the similarity is calculated using the distance between the two pattern vectors $\mathbf{y}^{\mathrm{r}}(\mathrm{j})$ and fault library vector $\mathbf{z}_{\mathrm{k}}^{\mathrm{r}}(\mathrm{j})$ in the rth reduced dimension, which is given by $\mathrm{D}_{\mathrm{k}}^{\mathrm{r}}(\mathrm{j}) = \left\| \mathbf{z}_{\mathrm{k}}^{\mathrm{r}}(\mathrm{j}) - \mathbf{y}^{\mathrm{r}}(\mathrm{j}) \right\|$. For the kth fault in the library, the similarity index in the rth PC is given by

$$S_k^r(j) = \frac{\frac{1}{D_k^r(j)}}{\sum_{m=1}^{K} \frac{1}{D_m^r(j)}}. \quad (6)$$

Here, when $\mathrm{D}_{\mathrm{k}}^{\mathrm{r}}(\mathrm{j})=0$ $\mathbf{y}^r(j)$ is identical to $\mathbf{z}_k^r(j)$. Such a case is likely to happen at the beginning sequence of a fault.

III. RESULTS AND PERFORMANCE COMPARISON

The diagnosis performance based on the triangular representation of process data is demonstrated here. This work utilizes simulated data from the Tennessee Eastman process, which is a common test bed for continuous processes [13]. This process has various equipments including reactor, condenser, and compressor. As shown in Fig. 2, it consists of five major units: a reactor, a product condenser, a separator, a recycle compressor, and a product stripper. This process also produces two products G and H from four reactants A, C, D, and E. Also there are an inert B and a byproduct F. A total of 53 process variables are measured on-line. The gaseous reactants are fed to the reactor, where the liquid products G and H are formed. The reactions in the reactor are as follows:

A+C+D→ G, A+C+E→ H, and A+E→ F+3D→ 2F.

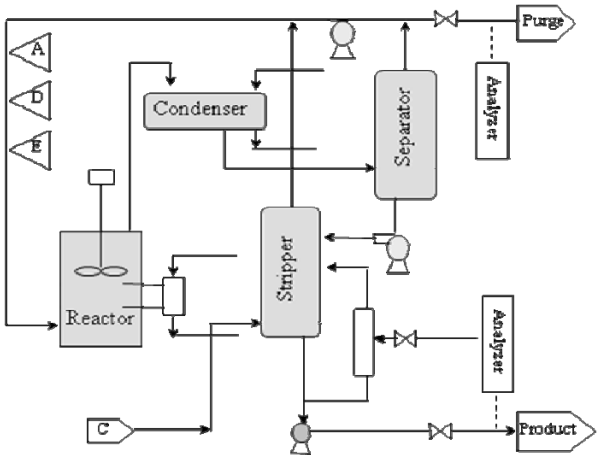

Fig. 2 A schematic for TE process

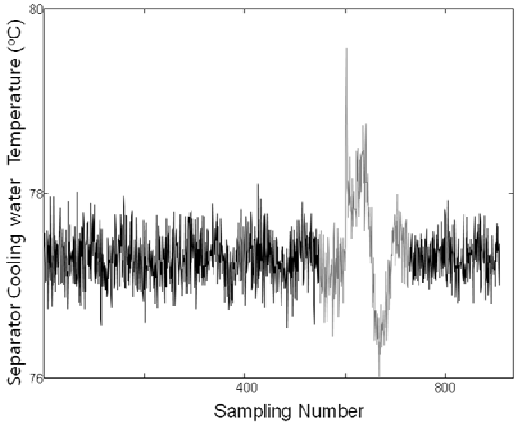

Fig. 4 A plot for separator cooling water temperature

Seven different process faults are investigated to test the diagnosis performance. In this work they are referred to as FT1 through FT7. These seven faults represent bias changes in the process. For example, a step change in the condenser cooling water inlet temperature, as shown in Fig. 3 and Fig. 4, results in some fluctuation in separator temperature and separator cooling water temperature accordingly. We need to find the on-line fault pattern at a given sequence. If the fault pattern in the first PC at the first sequence is constant, $\boldsymbol{y}_1$ is represented by $\mathbf{y}_1 = [1000000]^T$. In the same manner, the fault patterns for the next sequences can be determined. Similarly, the on-line fault pattern in the first PC can be formulated. The next step is to compare the on-line pattern vector with the existing fault patterns, i.e., $\mathbf{z}_k^r(j)$. As mentioned before, the distance $D_k^r(j)$ between $\mathbf{y}^r(j)$ and $\mathbf{z}_k^r(j)$ and the similarity index $S_k^r(j)$ are obtained. Then $S_k(j)$ is calculated for each cause candidate, and the cause candidate with the highest value of $S_k(j)$ is selected as the assignable cause.

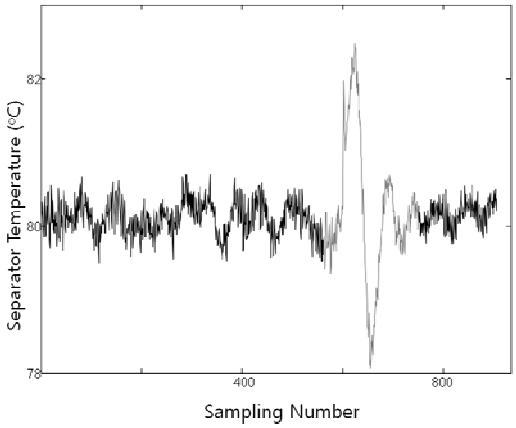

Fig. 3 A plot for separator temperature

Table II shows the diagnosis results for the TE process based on the triangular representation of process fault data combined with the linear technique. The overall similarity indices for the seven faults at the fifth sequence are displayed. As an example, when the true cause of the fault is FT2, the overall similarity index values $S_k(j)$ for each of cause candidates are 0.13, 0.30, 0.18, 0.12, 0.09, 0.10, and 0.08. It means that the overall similarity index for the true cause (i.e., 0.30) is higher than those of other cause candidates. For a clear comparison, the highest value is highlighted by a bold style in each row of the table. In terms of diagnosis performance, Table II yielded incorrect diagnosis decisions for the two fault cases, i.e., FT1 and FT3. Unlike other fault cases, these two cases did not produce the highest index values for the true causes of the cases. In case of FT1, for example, the true cause FT1 has the similarity index value of 0.18 which is lower than 0.20 of FT7. This is also the case for FT3 where the true cause FT3 possesses lower similarity index value of 0.21 than 0.22 of FT1. In summary, the linear technique based diagnosis method shows a limited diagnosis result in this case study.

TABLE II
DIAGNOSIS RESULTS OF LINEAR METHOD

True Cause	SIMILARITY INDEX						
	FT1	FT2	FT3	FT4	FT5	FT6	FT7
1	0.18	0.12	0.10	0.09	0.14	0.17	0.20
2	0.13	0.30	0.18	0.12	0.09	0.01	0.08
3	0.22	0.09	0.21	0.11	0.11	0.10	0.16
4	0.04	0.08	0.30	0.34	0.10	0.09	0.05
5	0.05	0.07	0.20	0.22	0.24	0.10	0.12
6	0.05	0.04	0.03	0.07	0.06	0.68	0.07
7	0.10	0.05	0.12	0.03	0.15	0.09	0.46

Based on the nonlinear technique diagnosis results for the TE process were also evaluated as shown in Table III. Similar to Table II, the highest value of the similarity index for fault candidates is highlighted by a bold style in each row of Table III. The major difference between the diagnosis results of linear and nonlinear techniques used can be seen by comparing the case of FT1. In case of FT1, the use of the linear technique selected FT7 as the highest index value of 0.20 though the true

cause is FT1. On the other hand, the use of the nonlinear technique yielded the right cause (FT1) with the index value of 0.32 whilst FT7 has the index value of 0.17. It should be also noted that the use of nonlinear technique increased the index values for the right cause candidate in the other cases except the case of FT3. The fault pattern of FT3 still selected incorrect cause candidate of FT1 as the highest index value, which is similar to Table II. It turned out that the nonlinear technique based diagnosis framework outperformed the linear one for this case study.

TABLE III
DIAGNOSIS RESULTS OF NONLINEAR METHOD

True Cause	SIMILARITY INDEX						
	FT1	FT2	FT3	FT4	FT5	FT6	FT7
1	0.32	0.10	0.08	0.07	0.12	0.14	0.17
2	0.10	0.37	0.15	0.11	0.11	0.08	0.08
3	0.19	0.12	0.18	0.13	0.12	0.08	0.18
4	0.02	0.10	0.23	0.41	0.09	0.09	0.06
5	0.06	0.09	0.14	0.19	0.41	0.05	0.06
6	0.02	0.03	0.02	0.05	0.05	0.77	0.06
7	0.07	0.05	0.08	0.03	0.10	0.06	0.61

IV. CONCLUSION

The use of a triangular representation of fault pattern in reduced spaces is presented to make a diagnostic decision using on-line multivariate process data. Using simulated data it was demonstrated that the nonlinear kernel method outperforms the linear method in terms of diagnosis performance. It is easy to implement because the historical data are already stored in a database, and thus can offer one of solutions for the on-line diagnosis of complicated industrial processes. Nonetheless, the improvement and stabilization of the diagnosis performance at the beginning of a fault need to be considered. It is because reliable diagnosis results at the beginning are meaningful to operators who need to take control actions. In this respect, the time delay in detect a fault can result in incorrect diagnostic decisions at all.

REFERENCES

[1] F. Akbaryan, and P. R. Bishnoi, "Fault diagnosis of multivariate systems using pattern recognition and multisensor data analysis technique," *Computers and Chemical Engineering*, vol. 25, pp. 1313-1339, 2001.

[2] A. K. S. Jardine, D. Lin, D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, pp. 1483–1510, 2006.

[3] V. A. Sotiris, P. W. Tse, and M. G. Pecht, "Anomaly detection through a bayesian support vector machine," *IEEE Transactions on Reliability,* vol. 59, pp. 277-286 , 2010.

[4] R. Lombardo, J.-F. Durand, A. P. Leone, "Multivariate additive PLS spline boosting in agro-chemistry studies," *Current Analytical Chemistry*, vol. 8, pp. 236-253, 2012.

[5] L. H. Chiang, E. L. Russell, and R. D. Braatz, "Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, pp. 243-252, 2000.

[6] J. C. Wong, K. A. Mcdonald, and A. Palazoglu, "Classification of abnormal plant operation using multiple process variable trends," *Journal of Process Control*, vol. 11, pp. 409-418. 2001.

[7] A. Bakhtazad, A. Palazoglu, and J. A. Romagnoli, "Detection and classification of abnormal process situations using multidimensional wavelet domain hidden markov trees," *Computers and Chemical Engineering*, vol. 24, pp. 769-775. 2000.

[8] M. Misra, S. J. Qin, H. Yue, and C. Ling, "Multivariate process monitoring and fault identification using multi-scale PCA," *Computers and Chemical Engineering*, vol. 26, pp. 1281-1293, 2002.

[9] P. K. Kankar, S. C. Sharma, and S. P. Harsha, "Faultdiagnosis of ball bearings using machine learning methods," *Expert Systems with Applications*, vol. 38, pp. 1876–1886, 2011.

[10] L. Dobos, and J. Abonyi, "On-line detection of homogeneous operation ranges by dynamic principal component analysis based time-series segmentation," *Chemical Engineering Science*, vol. 75, pp. 96–105, 2012.

[11] J. McBain, and M. Timusk, " Feature extraction for novelty detection as applied to fault detection in machinery," *Pattern Recognition Letters*, vol. 32, pp. 1054–1061, 2011.

[12] J. T.-Y. Cheung, and G. Stephanopoulos, "Representation of process trends-part I. a formal representation framework," *Computers and Chemical Engineering*, vol. 14, pp. 495-510, 1990.

[13] J. J. Downs, and E. F. Vogel, "A plant-wide industrial process problem," *Computers and Chemical Engineering*, vol. 7, pp. 245-255, 1993.