Journal of the Korea Academia-Industrial cooperation Society
Vol. 16, No. 5 pp. 3000-3005, 2015

http://dx.doi.org/10.5762/KAIS.2015.16.5.3000
ISSN 1975-4701 / eISSN 2288-4688

Diagnostic Classification Based on Nonlinear Representation and Filtering of Process Measurement Data

Hyun-Woo Cho[1*]

[1]**Department of Industrial and Management Engineering, Daegu University**

공정측정데이터의 비선형표현과 전처리를 활용한 분류기반 진단

조현우[1*]

[1]대구대학교 산업경영공학과

Abstract Reliable monitoring and diagnosis of industrial processes is quite important for in terms of quality and safety. The goal of fault diagnosis is to find process variables responsible for causing specific abnormalities of the process. This work presents a classification-based diagnostic scheme based on nonlinear representation of process data. The use of a nonlinear kernel technique is able to reduce the size of the data considered and provides efficient and reliable representation of the measurement data. As a filtering stage a preprocessing is performed to eliminate unwanted parts of the data with enhanced performance. The case study of an industrial batch process has shown that the performance of the scheme outperformed other methods. In addition, the use of a nonlinear representation technique and filtering improved the diagnosis performance in the case study.

요 약 신뢰할 수 있는 공정 감시와 진단은 생산 공정의 안전과 최종제품의 품질을 보장이라는 관점에서 중요하다. 공정 진단의 목적은 특정한 공정 이상의 원인을 밝혀내는 것이다. 본 연구에서는 분류기법에 기반한 공정진단 체계를 제시한다. 여기서는 공정데이터를 비선형 데이터 표현기법을 통해 변환함으로써 데이터의 크기를 줄이며 효율적인 데이터 표현이 가능하다. 추가적인 단계로서 공정 데이터의 전처리 과정을 통해 진단에 무관한 공정 패턴을 제거하고 진단 성능을 높이고자 한다. 진단 성능을 평가하기 위해 회분식 공정에 대한 사례연구를 수행한 결과 기존 선형 진단 방법론 및 전처리 과정이 없는 방법론에 비해 향상된 진단 결과를 얻을 수 있었다.

Key Words : Multivariate statistical methods, diagnosis, classification, noninear kernel, filtering

1. Introduction

As one of important elements for ensuing quality and safety of industrial processes, the task of fault diagnosis is crucial in on-going production runs. It seeks to find assignable causes of abnormal production conditions detected on-line. In process industry operating personnel often take remedial process inputs based on diagnostic decisions provided.[1] Various approaches have been developed including mathematical models and knowledge-based models, but they are not easy to develop and maintain especially when the process are complex or change frequently. [2] Due to the advances in sensing and data technology multivariate statistical diagnosis approaches have become popular using automated on-line data collection. These methods are considered to be easy to implement and computationally efficient.

This research was supported by the Daegu University Research Grant, 2011.
*Corresponding Author : Hyun-Woo Cho (Daegu Univ.)
Tel: +82-2-850-6540 email: hwcho@daegu.ac.kr
Received January 21, 2015 Revised May 5, 2015
Accepted May 7, 2015 Published May 31, 2015

The identification of causes of faults can be treated in the statistical formulation of diagnosis as classification problems.[3] Breiman *et al*. (1984) developed a recursive partitioning methodology for classification problems called classification and regression tree.[4] The classification and regression tree classifier is a tree constructed by recursively partitioning the predictor space, which is based on training data sets. It identifies important independent variables when there are many potential considered. Its main advantage is that the resulting classification model can be easily interpreted.[5]

In classification domains, dimension reduction is quite useful because the dimension of data is normally high when compared with small model data or samples. Many nonlinear versions of popular linear statistical approaches have been developed and extensively utilized in practical classification and dimension reduction problems including kernel PCA (KPCA), kernel PLS (KPLS), kernel FDA (KFDA), and so on.[6]

This work presents a classification-based diagnostic scheme based on nonlinear representation of raw process measurement data. The use of a nonlinear kernel technique in the diagnosis of possible faults in processes helps us to decrease the size of the data considered. In addition, it provides efficient and reliable representation of the measurement data that is suitable for differentiating different fault cases or classes. As a filtering stage, in this work, a multivariate preprocessing is performed to eliminate unwanted parts of the data, which is expected to improve the performance of empirical diagnosis models. The performance of the proposed diagnostic scheme is demonstrated using a case study of an industrial batch process. Compared to continuous processes commonly encountered, batch processes are widespread in the production of high value added products such as pharmaceutical and semi-conductor industries. However, batch processes are difficult to control because they have finite duration time and nonlinear characteristics of measurement data.

This paper is organized as follows. First, a brief review of related and proposed methods used in this work is provided, and then a case study is conducted to demonstrate the presented diagnosis scheme. Finally, concluding remarks and future research issues are given.

2. Method

As a statistical technique a classification tree construction process in a classification problems recursively partitions variable space based on training measurements in which groups are known. The class assigned to each terminal node minimizes the estimated cost of misclassification[4]:

$$r(t) = \min_i \sum_j c(i|j)\, p(j|t) \tag{1}$$

where $c(i \mid j)$ represents the cost misclassifying a class j as a class i and the estimated probability of the class j in node t.

Among several measures of impurity of the node developed the follow index is the most commonly used function of node impurity

$$i(t) = \sum_{i \neq j} p(i|t)\, p(j|t) = 1 - \sum_j p^2(j|t). \tag{2}$$

The goodness of a split can be checked by the deviance reduction related to the split, which is given by

$$d(t) = -2 \sum_j n_{tj} \log p(j|t) \tag{3}$$

where n_{tj} represents the frequency of class j in node t. Kernel Fisher's discriminant analysis (KFDA) is the nonlinear kernel version of linear FDA. Kernel FDA performs LFDA in the feature space F, and as a result kernel FDA produces a set of nonlinear discriminant vectors. The discriminant weight vector is determined by maximizing between-class scatter matrix $\mathbf{S}_b^{\Phi}$ while minimizing total scatter matrix $\mathbf{S}_t^{\Phi}$, which are defined in F as follows:

$$\mathbf{S}_b^{\Phi} = \frac{1}{M}\sum_{i=1}^{C} c_i (\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi})(\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi})^{\mathrm{T}} \quad (4)$$

$$\mathbf{S}_t^{\Phi} = \frac{1}{M}\sum_{i=1}^{M} (\Phi(\mathbf{x}_i) - \mathbf{m}^{\Phi})(\Phi(\mathbf{x}_i) - \mathbf{m}^{\Phi})^{\mathrm{T}} \quad (5)$$

This can be done by maximizing the Fisher criterion[6]:

$$J^{\Phi}(\boldsymbol{\psi}) = \frac{\boldsymbol{\psi}^{\mathrm{T}} \mathbf{S}_b^{\Phi} \boldsymbol{\psi}}{\boldsymbol{\psi}^{\mathrm{T}} \mathbf{S}_t^{\Phi} \boldsymbol{\psi}} \quad (6)$$

The optimal discriminant vectors can be obtained by solving the eigenvalue problem $\mathbf{S}_b^{\Phi}\boldsymbol{\psi} = \lambda \mathbf{S}_t^{\Phi}\boldsymbol{\psi}$ instead of equation (6). They are actually the eigenvectors of $\mathbf{S}_b^{\Phi}\boldsymbol{\psi} = \lambda \mathbf{S}_t^{\Phi}\boldsymbol{\psi}$. There exist coefficients b_i such that

$$\boldsymbol{\psi} = \sum_{k=1}^{M} b_k \Phi(\mathbf{x}_k) = \mathbf{H}\boldsymbol{\alpha} \quad (7)$$

where $\mathbf{H} = [\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_M)]$ and $\boldsymbol{\alpha} = (b_1, \ldots, b_M)^{\mathrm{T}}$.

By substituting equation (7) into equation (6) and following equation is obtained:

$$J^{\mathrm{K}}(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{KWK})\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{KK})\boldsymbol{\alpha}} . \quad (8)$$

Then, equation (8) can be converted to

$$J(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_t\boldsymbol{\beta}}, \quad (9)$$

where $\mathbf{S}_b = \Lambda^{1/2}\mathbf{P}^{\mathrm{T}}\mathbf{W}\mathbf{P}\Lambda^{1/2}$ and $\mathbf{S}_t = \Lambda$. Finally the optimal discriminant vectors are given by

$$\boldsymbol{\Psi}_j = \mathbf{H}\boldsymbol{\alpha}_j = \mathbf{H}\mathbf{P}\Lambda^{-1/2}\boldsymbol{\beta}_j, \ \mathrm{j}=1, \cdots, \mathrm{d}. \quad (10)$$

Overall diagnostic scheme of this work is shown in Fig. 1, in which detailed steps of the scheme are also provided. An orthogonal filtering of raw measurement data is first done in order to get rid of noise parts of raw measurement data. This step is also necessary because ultimate classification performance in diagnosis can be improved by filtering the noise unrelated to meaningful patterns before nonlinear kernel modeling. In this work one of orthogonal filtering techniques is used to remove the unwanted portions of measurement variation orthogonal to class membership from process variables. To this end, calculation of the first score vector of process variables is required, and then these scores are orthogonalized followed by obtaining weight vector. Finally it needs to update score vectors, and whole processes are repeated until score values have converged. For the next component a loading vector is generated giving residual. The next components can be calculated in such a way.

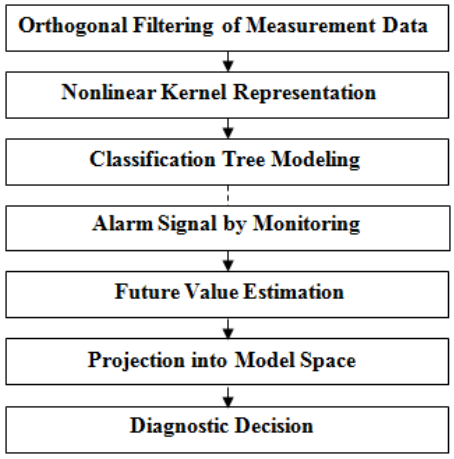

Fig. 1. Overall scheme

As shown in Fig. 1, the next step is the extraction of nonlinear features of various measurement data, which is executed by performing nonlinear kernel representation of the filtered data. For a classification training purpose the classification tree model is built based on the resulting kernel score values. The nonlinear modeling of the measurement data is to find out the directions where different data class or groups for diagnosis are separated well. The within-group-scatter matrix ($\mathbf{S}_w$) and the between-group-scatter matrix ($\mathbf{S}_b$) are given by:

$$\mathbf{S}_w = \sum_{p=1}^{P}\sum_{\mathbf{z}_i \in g_p} (\mathbf{z}_i - \bar{\mathbf{z}}_p)(\mathbf{z}_i - \bar{\mathbf{z}}_p)^T \quad (11)$$

$$\mathbf{S}_b = \sum_{p=1}^{P} n_p (\bar{\mathbf{z}}_p - \bar{\mathbf{z}})(\bar{\mathbf{z}}_p - \bar{\mathbf{z}})^T \quad (12)$$

The model dimension can be determined by finding alpha values that minimize the following criterion.[7] Based on nonlinear score values obtained the next step is to build a classification tree. It is necessary to split each variable at all its possible split points where a parent node is divided into two child nodes. Then it

selects the variables and split point with the highest impurity reduction. Dividing the parent node into the two child nodes is repeated until the tree has maximum size. In a pruning stage cross validation is performed in order to get the optimal tree size.

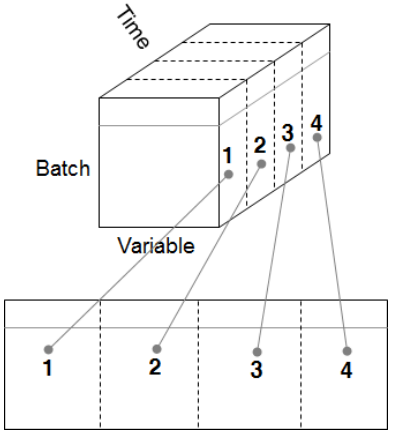

Fig. 2. Characteristics of batch data

The on-line diagnosis module of the diagnostic scheme is initiated by an out-of-control signals determined by on-line fault detection or monitoring systems implemented. In doing on-line diagnostic decision using test data, however, it is necessary to estimate future observations of the current measurement data collected on-line. It is because that the current new batch run is not complete until the end of its operation (Fig 2). The incomplete parts of the measurement data should be estimated to make the data matrix full.

Once the measurement data become full with future observations estimated on-line, an on-line KFDA score vector $\mathbf{s}_{\text{new}}(k^*)$ for current measurement data of the test run can be obtained by projecting the observation onto discriminant vectors: $\mathbf{s}_{new}(k^*) = \mathbf{\Psi}^{\mathrm{T}} \Phi(\mathbf{x}_{new}(k^*))$. Then the proposed scheme classifies complete test measurement data into four fault classes by calculating the distance between on-line scores and each of mean score vectors of the fault classes. Finally, the fault class with the minimum distance is selected as the assignable cause of the fault at that time.

3. Case study: diagnosis results

The performance of the proposed diagnostic scheme is demonstrated in this section using a case study of an industrial batch process. This industrial batch process is a polyvinyl chloride (PVC) straight resin polymerization, which consists reactor, condenser, agitator, and cooling jacket. Here eleven process variables are automatically measured on-line, which are used to measure various physical states such as temperature, pressure, flow rate, agitator speed, and so on. As shown in Fig. 3, the loading vectors of two variables are displayed in order to represent of relative importance of original variables in new variables. Due to the characteristics of polymerization reactor temperature needs to be managed carefully. Training data sets for off-line model building consist of 75 abnormal batch operations in four fault groups. In addition, a total of eight abnormal batches (i.e., two batches for each fault group) are used as test data sets, none of which are included in the training data sets. Hereafter they are denoted as Run1 through Run8 in this work. For example, “Run1” and “Run2” have same fault membership class.

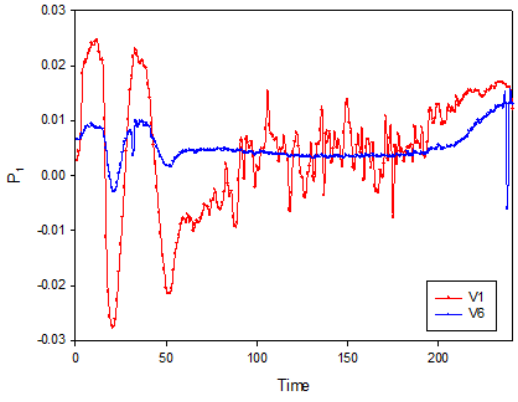

Fig. 3. Plot of two variables’ loading vectors

An empirical classification model for separating four classes or groups was constructed using the training measurement data. During the refinement stage of the model diagnosis results for the test data sets are tested

using various kernel functions of nonlinear representation. It turned out that the second-order polynomial kernel is better to represent the data nonlinearity rather than other kernels of Gaussian or sigmoid. In terms of handling future observation for test data fault library-based estimation approach was adopted in order to estimate the future observations of test batches.

Table 1 shows the diagnosis results for the eight test batches for the case study, in which classification success percentage values were calculated during faulty operation runs of test batches after out-of-control signals. It means, for example, that the value of 0.9 in this table indicates 90% successful diagnostic decision made. To compare the diagnosis performance of the method with those of other variants three diagnostic schemes are evaluated. The results obtained from the proposed scheme are listed along with two similar methods. That is, the first method is the proposed scheme with linear representation technique used instead of the nonlinear technique (denoted as "Linear" in the table). The second method is exactly the same as the proposed one, but it did not use filtering or preprocessing of measurement data at all.

Table 1. Diagnosis results for case study

Run	Linear	No Filtering	Proposed
1	0.92	0.94	0.95
2	0.89	0.93	0.95
3	0.86	0.92	0.97
4	0.89	0.90	0.93
5	0.78	0.89	0.91
6	0.80	0.94	0.95
7	0.86	0.93	0.94
8	0.85	0.90	0.89
avg.	0.86	0.92	0.94

As shown in Table 1 the proposed diagnostic scheme (denoted as "Proposed" in the table) produced the best diagnosis performance in terms of average diagnosis success rate (i.e., 0.94). On the other hand, lower average values of 0.86 and 0.92 are obtained from "Linear" and "No Filtering" methods respectively. Some performance improvement is achieved in terms of average daignosis values. Moreover, the proposed scheme yielded the highest diagnosis success rates for all test batches except Run8 (0.90 of "No Filtering" vs. 0.89 of "Proposed"). Thus it can be said that the proposed diagnostic scheme outperforms the linear and no filtering variants. It should be also noted that the performance of the proposed scheme without filtering (i.e., "No Filtering") is better than that of the proposed scheme with linear representation technique used instead of nonlinear one (i.e., "Linear"). The effect of using linear or nonlinear representation techniques is much more important than the effect of filtering in diagnosis. It is mainly because the nonlinear patterns in measurement data cannot be represented well by linear techniques.

4. Conclusion

This work presented a nonlinear representation-based diagnostic scheme based on classification tree. It also includes noise filtering as a pre-treatment of raw measurement data. It has been demonstrated using a dataset of a batch process that the proposed diagnostic scheme outperforms other similar diagnosis schemes. It turned out that the use of a nonlinear representation technique in the case study improved diagnosis performance due to the better representation of nonlinear patterns of the measurement data. In addition, the use of noise filtering and preprocessing of the data was shown to be quite effective in increasing diagnosis success rates in most test runs. Such characteristics of the proposed scheme would be useful in diagnosing various fault data of complex industrial processes. As one of future research issues, continuous updates of training data for empirical diagnosis models is quite essential in improving the performance dramatically. Also small sample size issues related to the training batch runs and future value estimation need further researches in terms of reliability of diagnosis and computational speed.

References

[1] S. J. Qin, "Survey on data-driven industrial process monitoring and diagnosis", Annual Reviews in Control, 36, pp. 220-234, 2012.
DOI: http://dx.doi.org/10.1016/j.arcontrol.2012.09.004

[2] S. Bersimis, S. Psarakis, and J. Panaretos, "Multivariate statistical process control charts: an overview", Qual. & Reliability. Engineering International, 23 (5), pp. 517–543, 2007.
DOI: http://dx.doi.org/10.1002/qre.829

[3] S. X. Ding, "Data-driven design of monitoring and diagnosis systems for dynamic processes: A review of subspace technique based schemes and some recent results", Journal of Process Control, 24, pp. 431–449, 2014.
DOI: http://dx.doi.org/10.1016/j.jprocont.2013.08.011

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression tree", Wadsworth, Monterey, CA, USA, 1984.

[5] P. S. Gromski, Y Xu, K. A. Hollywood, M. L. Turner, and R. Goodacre, “The influence of scaling metabolomics data on model classification accuracy”, Metabolomics, 11, pp 684-695, 2015.
DOI: http://dx.doi.org/10.1007/s11306-014-0738-7

[6] G. Baudat and F. Anouar, Generalized discriminant analysis using a kernel approach. Neural Computation, 12, pp. 2385–2404, 2000.
DOI: http://dx.doi.org/10.1162/089976600300014980

[7] 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, 50, pp. 243-252, 2000.
DOI: http://dx.doi.org/10.1016/S0169-7439(99)00061-1

Hyun-Woo Cho [Regular member]

- Aug. 2003 : POSTECH., Industrial Eng., PhD
- Aug. 2003 ~ Aug. 2007 : GIT/UT, Research Associate
- Sep. 2007 ~ Feb. 2011 : SEC, Senior Engineer
- Mar. 2011 ~ Current : Daegu Univ., Dept. of Industrial. & Management Eng., Professor

<Research Interests>
Intelligent Process Monitoring, Data Mining