International Journal of Mechanical, Industrial and Aerospace Sciences
ISSN: 2517-9950
Vol:7, No:11, 2013

Empirical Process Monitoring Via Chemometric Analysis of Partially Unbalanced Data

Hyun-Woo Cho

***Abstract*—**Real-time or in-line process monitoring frameworks are designed to give early warnings for a fault along with meaningful identification of its assignable causes. In artificial intelligence and machine learning fields of pattern recognition various promising approaches have been proposed such as kernel-based nonlinear machine learning techniques. This work presents a kernel-based empirical monitoring scheme for batch type production processes with small sample size problem of partially unbalanced data. Measurement data of normal operations are easy to collect whilst special events or faults data are difficult to collect. In such situations, noise filtering techniques can be helpful in enhancing process monitoring performance. Furthermore, preprocessing of raw process data is used to get rid of unwanted variation of data. The performance of the monitoring scheme was demonstrated using three-dimensional batch data. The results showed that the monitoring performance was improved significantly in terms of detection success rate of process fault.

***Keywords*—**Process Monitoring, kernel methods, multivariate filtering, data-driven techniques, quality improvement.

I. Introduction

THE process monitoring methods including fault detection and fault identification have been extensively studied as one of essential topics of statistical process control. This is largely due to the fact that most of manufacturing or production processes are susceptible to unexpected abnormalities such as process faults, breakdowns and malfunctions [1]. Unfortunately, these special or uncommon events are apt to give a negative impact on final product quality. A fault generally indicates an abnormal process event. Thus the main goal of process monitoring is to detect the occurrence of a fault [2].

Batch type process monitoring is quite difficult to implement or maintain because these processes have challenging issues like nonlinear process behavior, finite duration of operation time, etc. A batch process operation includes a set of tasks like charging ingredients, processing them under controlled conditions, and discharging final product [3]. Recent reports showed that the development of monitoring schemes concentrated on the application of multivariate statistical methods due to the availability of real-time and historical process data. Machine learning techniques have been extensively used in practical monitoring problems including nonlinear kernel techniques based on Fisher discriminant analysis, principal component analysis, and partial least squares [4]-[6]. Empirical modeling techniques have been widely used because of widespread sensor and data measurement technology in production processes.

Automated data collection has become popular in most of industrial processes with the help of the advances in sensing and data measurement technology. Thus the availability of large historical and/or real-time dataset has motivated the statistical approaches to the on-line monitoring and fault diagnosis. Various multivariate statistical techniques have been employed including principal component analysis (PCA), partial least squares (PLS), and Fisher discriminant analysis (FDA). Theses multivariate statistical techniques, in general, are considered to be easy to implement, computationally efficient, and relatively robust to noise.

When monitoring a process using process measurement data, data sets for specific classes may be under-sampled or not enough to build empirical monitoring models. Frequently, measurement data of normal operations without faults are easy to gather, but the measurement data of out-of-control states or faults are expensive to collect. Such an unbalanced measurement data can be covered by adopting the appropriate method which depends mainly on normal data. Support vector data description (SVDD) is quite helpful in describing samples of high density areas of normal operating conditions. This is able to adapt to the real shape of normal samples and seeks to find flexible boundary with a minimum volume by introducing kernel functions.

The idea of SVDD is to generate a boundary around data samples with a volume as small as possible. The purpose of SVDD is to represent given set of data in a unique minimal volume spherical domain enclosing most of the samples on interest. The compact representation of the sample data is given as a hyper-sphere with minimal volume containing most of the sample data in a high-dimensional feature space based on kernel functions. A one-class classification problem can be also solved by estimating a probability density of normal data. However, such an approach needs a large number of samples and is not robust to normal data that may contain only limited area of normal operation data. Thus it focuses on describing samples of high density areas of normal operating conditions. It also rejects samples of low density areas though they are actually normal. The advantage of SVDD is that it can adapt to the real shape of samples and find flexible boundary with a minimum volume by introducing kernel functions.

This work presents the utilization of an empirical model-based quality monitoring approach to batch processes

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).

with unbalanced measurement data. The monitoring scheme is combined with nonlinear representation of raw process data. In addition, preprocessing or filtering schemes are implemented for better monitoring performance. In this work the nonlinear representation of finite and three-way batch data is applied. Prior to building empirical monitoring models, filtering of the data is performed to trim the irrelevant information of the process data. In this work, several monitoring schemes are evaluated, in which two filtering techniques are also tested. Due to the characteristics of batch data, the selection of estimation approaches for future observation is also discussed with the comparison of monitoring performance. The performance of the proposed process monitoring schemes is demonstrated using batch process data. It is organized as follows: an introduction of multivariate statistical techniques followed by monitoring performance comparison using batch process data. Finally, concluding remarks are given.

II. Methodologies

A. Linear Projections

A linear version of principal component representation, i.e., principal component analysis (PCA), is used to decompose correlated original variables into an uncorrelated set of linear principal components. In most cases, only several components are enough to explain the data variability. It seeks to decompose the data matrix **X** into the sum of the outer products of score vectors (**t**) and loading vectors (**p**) plus a residual matrix (**E**):

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r^{\mathrm{T}} + \mathbf{E} \tag{1}$$

On the other hand, the goal of linear Fisher discriminant analysis (FDA) is to find certain directions in original variables, along which hidden groups are discriminated as clearly as possible [8]. As an extension of linear FDA, nonlinear kernel FDA (KFDA) executes linear FDA in the feature space F. As a result, the discriminant weight vector is determined by maximizing between-class scatter matrix while minimizing total scatter matrix, which are defined in feature space:

$$\mathbf{S}_b^{\Phi} = \frac{1}{M}\sum_{i=1}^{C} c_i (\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi})(\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi})^{\mathrm{T}} \tag{2}$$

$$\mathbf{S}_b^{\Phi} = \frac{1}{M}\sum_{i=1}^{C} c_i (\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi})(\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi})^{\mathrm{T}} \tag{3}$$

by maximizing the Fisher criterion

$$J^{\Phi}(\boldsymbol{\psi}) = \frac{\boldsymbol{\psi}^{\mathrm{T}} \mathbf{S}_b^{\Phi} \boldsymbol{\psi}}{\boldsymbol{\psi}^{\mathrm{T}} \mathbf{S}_t^{\Phi} \boldsymbol{\psi}} \tag{4}$$

the optimal discriminant vectors can be obtained. As one of filtering techniques, orthogonal signal correction (OSC) is a PLS-based solution which removes unwanted variation. In this work, the OSC method is used so that a coding is introduced where each column in Y matrix contains information about class memberships of samples. The binary **Y** matrix has a structure where each row sums to unity. The first step of an OSC is to calculate the first PC score vector, and actual correction vector is produced:

$$\mathbf{t}^* = \{\mathbf{I} - \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\}\mathbf{t}. \tag{5}$$

Then PLS weight vector **w** is computed such that **Xw=t***, which is followed by the calculation of a new score vector **t=Xw**. Finally, a loading vector **p** is computed and the correction term $\mathbf{tp}^{\mathrm{T}}$ is subtracted from **X**, giving the residual. The next components can be calculated in a similar way. An alternative approach is discriminant partial least squares (PLS), which is the classical PLS algorithm applied to classification problems. One common way to use PLS in classification problems is to introduce a coding in which each column in **Y** contains information about the class memberships of samples [7].

B. Nonlinear Mappings

Support vector machine (SVM) is a training algorithm to learn classification and regression rules of patterns from raw data. It is basically a linear method that is nonlinearly mapped from the input data space. In a real computation, input data are first mapped into high dimensional feature space. As shown in Fig. 1, in the feature space optimal decision function is obtained having a maximum margin, in which the decision function satisfies inequality constraints

$$y_i(\mathbf{w}\Phi(\mathbf{x}_i) + b) - 1 \geq 0 \ \forall_i \tag{6}$$

Based on the optimal decision function nonseparable problems are solved by Lagrangian as follows:

$$L = 1/2\|\mathbf{w}\|^2 + C\sum\xi_i - \sum\alpha_i[y_i(\mathbf{w}\Phi(\mathbf{x}_i) + b) - 1 + \xi_i] - \sum\mu_i\xi_i \tag{7}$$

Rather than such a quadratic problem, a dual problem is handled because it's easy to solve:

$$L_d = \sum\alpha_i - 1/2\sum\alpha_i\alpha_j y_i y_j \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) \tag{8}$$

Training SVM is to find α_i, b, and support vectors with given kernel function parameters and C.

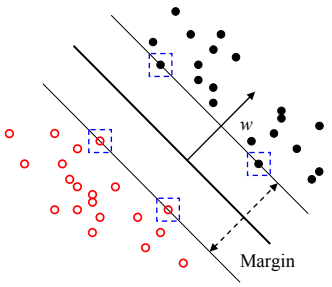

Fig. 1 A simple representation of SVM

Support vector data description (SVDD) is a one class classification method to envelop samples or objects within a high dimensional space with the volume as small as possible. It is necessary to find **μ** and R that has the minimum volume of hyper-sphere containing all samples. In the end the minimization problem can be denoted as follows:

$$F(R,\boldsymbol{\mu}) = R^2 + C\sum_{i=1}^{I}\xi_i \tag{9}$$

Here, the parameter C represents the trade-off between the volume of the sphere and the number of samples outside it, and thus it should be minimized with the following constraints:

$$\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \le R^2 + \xi_i, \xi_i \ge 0 \tag{10}$$

The Lagrangian function can be given by incorporating these equations:

$$L(R,\boldsymbol{\mu},\alpha_i,\beta_i,\xi_i) = R^2 + C\sum_{i=1}^{I}\xi_i - \sum_{i=1}^{I}\alpha_i\{R^2 + \xi_i - (\|\mathbf{x}_i\|^2 - 2\boldsymbol{\mu}\cdot\mathbf{x}_i + \|\boldsymbol{\mu}\|^2)\} - \sum_{i=1}^{I}\beta_i\xi_i \tag{11}$$

To obtain more flexible boundaries, inner products of samples are replaced by a kernel function so that the SVDD problem can be formulated as follows:

$$L = \sum_{i=1}^{I}\alpha_i K(\mathbf{x}_i,\mathbf{x}_i) - \sum_{i,j=1}^{I}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) \tag{12}$$

III. Comparison of Results

This part demonstrates the monitoring performance of the proposed scheme which utilizes nonlinear kernel method combined with preprocessing techniques for three dimensional process data. The test process is a polyvinyl chloride batch process. Here a straight resin polymerization process is initiated by vinyl chloride monomer. This process contains a polymerization reactor, reflux condenser, agitator, and cooling jacket. Eleven process variables are automatically measured on-line. A total of 170 batches are used in building nonlinear kernel monitoring models. In order to overcome the limitation of three way characteristics of batch process data, as stated before, future observations of a new batch, i.e., unmeasured data parts of current batch operation, should be estimated appropriately as shown in Fig. 2. It is due to the fact that a new or current batch operation is not complete until the end of its operation.

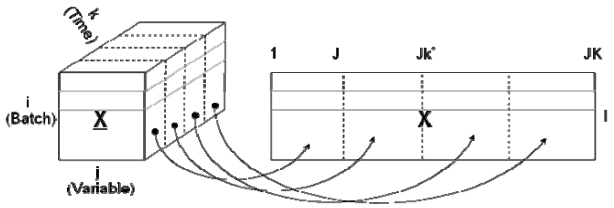

Fig. 2 Batch data and future values

TABLE I
Results Based On Library Values

	Monitoring success rate (%)			
	Model1	Model2	Model3	Model4
Ft1	84	89	93	95
Ft2	89	90	92	95
Ft3	84	89	93	94
Ft4	83	86	92	97
Ft5	87	91	92	93
Ft6	85	89	93	96
Ft7l	78	83	84	89

A filtering or preprocessing of process data is performed prior to main model building in order to obtain monitoring results using several monitoring schemes, To evaluate difference the monitoring performance based on several multivariate projection methods, the two representation techniques were applied. In addition, two filtering methods for the test process are tested. That is, discriminant partial least squares and orthogonal signal correction are considered. The selection of a kernel function in implementing monitoring models using different representation and preprocessing techniques was evaluated with the test of various kernel functions. In this work, second order polynomial kernel was chosen to capture nonlinearity of the data. The monitoring results for the seven test batches of Fault1 (denoted as Ft1) through Fault7 (Ft7) are summarized in Table I. As shown in Table I monitoring accuracy values, i.e., % of detection success rate of faults, are listed to evaluate the monitoring performance of four monitoring schemes denoted as Model1 through Model4. Here Monitoring accuracy is defined as the proportion of the observations correctly detected.

As shown in Table I, Model 1 denotes the monitoring scheme of using DPLS, KPCA, and SVDD. In addition, Model2 indicates the Model1 with the use of OSC instead of DPLS. Similar to the relationship between Model1 and Model2, on the other hand, Model3 differs from Model1 in that it utilizes KFDA rather than KPCA. The only difference of Model4 is the use of OSC instead of DPLS of Model3. As reported in Table I, the Model4 monitoring scheme showed the best monitoring performance in that it yielded the highest

monitoring accuracy for all test batches. The Model4 monitoring method (with KFDA and OSC) produced the best monitoring performance, i.e., average monitoring success rate (%) of 94.1 in terms of the average monitoring accuracy over the test batches.

Meanwhile, it is observed that the average monitoring success rate values of Model1, Model2, and Model3 are 84.3, 88.1, and 91.3, respectively. It should be noted that the overall monitoring performance of using KFDA, i.e., Model3 and Model4, outperforms those of using KPCA, i.e., Model1 and Model2, irrespective of preprocessing techniques used. Consequently, using KFDA monitoring methods has significantly improved the monitoring performance for this test process. It may be meaningful to operating personnel, who have to take remedial actions using the monitoring results.

As executed in Table I, monitoring results are obtained using the same monitoring models such as Model1 through Model4 monitoring schemes, though not shown here. On the other hand, the difference between Table I and these results is which estimation approaches to use in the monitoring model to estimate future observations of current operation of batch process.

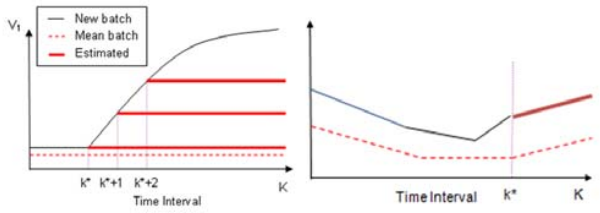

Fig. 3 Estimation of future values using current values

Specifically, results of Table I were obtained using the method of fault library approach [9]. On the other hand, PCA projection-based estimation method was used whilst current deviation method was applied to produce the monitoring results. For more information about the current deviation approach this is illustrated as shown in Fig. 3. For these results, overall observations of monitoring results are quite similar to Table I in that the Model4 produced the best monitoring accuracy in all the faults tested. For example of projection-based estimation method, the Model4 yielded maximum success rates. As analyzed in Table I, in case of comparing the effect of different estimation methods, fault library method [9] of Table I outperformed performance of the other two cases. In particular, the difference between PCA-based estimation method and current deviation method can be seen by comparing these tables. On the other hand, the use of KFDA improved monitoring performance significantly when compared to using KPCA. As shown in Fig. 4, relative performance of the overall results can be easily distinguished graphically.

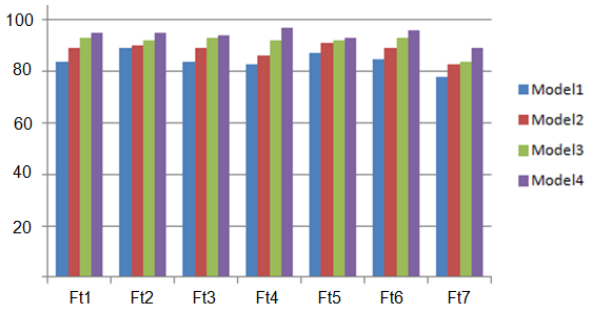

Fig. 4 Graphical comparison of monitoring results

IV. Conclusion

This work proposes the use of nonlinear kernel representation of unbalanced measurement data combined with the filtering techniques to provide reliable empirical model-based quality monitoring performance for batch processes. In this work a case study on the batch process has been executed using different monitoring schemes. Resultantly it has shown that the use of appropriate technique like discriminant analysis combined with preprocessing method of orthogonal signal correction produced reliable monitoring results on the test process. In particular, monitoring results showed that among different monitoring models tested the Model4 outperforms the other three monitoring schemes of Model1, Model2, and Model3.

In terms of future observations handling, the three future estimation methods were also tested. The use of the discriminant analysis technique was shown to represent monitoring pattern in the test process data. The unbalanced data problem of batch processes was solved by the use of the support vector data description technique. Here, it helps to define a control region or boundary around sample data with a volume as small as possible. Taking into accounts of frequent use of batch processes, kernel-based nonlinear technique is quite helpful to make monitoring decision in an on-line basis. Though not shown here, empirical model based monitoring schemes can be implemented, maintained, and updated efficiently. On the other hand, monitoring performance of the empirical models is subject to the quality of historical batch data. In this case, it would help to gather as many batch data as possible, but this inevitably results in a computational problem.

References

[1] S. J. Qin, "Statistical process monitoring: basics and beyond," *Journal of Chemometrics*, vol. 17, pp. 480–502, 2003.

[2] S. Bersimis, S. Psarakis, J. Panaretos, "Multivariate statistical process control charts: an overview," *Quality and Reliability Engineering International*, vol. 23, pp. 517–543, 2007.

[3] X. Meng, A. J. Morris, E. B. Martin, "On-line monitoring of batch processes using PARAFAC representation," *Journal of Chemometrics*, vol. 17, pp. 65-81, 2003.

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

[5] 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.

[6] L. H. Chiang, E. L. Russell, R. D. Braatz, "Fault monitoring 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.
[7] J. A. Westerhuis, S. de Jong, A. K. Smilde, "Direct orthogonal signal correction," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, pp. 13-25, 2001.
[8] G. Baudat, and F. Anouar, " Generalized discriminant analysis using a kernel approach," Neural Computation, vol. 12, pp. 2385-2404, 2000.
[9] H.-W. Cho, K. J. Kim, "A method for predicting future observations in the monitoring of a batch process," *Journal of Quality Technology*, vol. 35, pp. 59-69, 2003.