International Journal of Mechanical, Industrial and Aerospace Sciences
ISSN: 2517-9950
Vol:7, No:5, 2013

Automated Process Quality Monitoring with Prediction of Fault Condition Using Measurement Data

Hyun-Woo Cho

***Abstract*—**Detection of incipient abnormal events is important to improve safety and reliability of machine operations and reduce losses caused by failures. Improper set-ups or aligning of parts often leads to severe problems in many machines. The construction of prediction models for predicting faulty conditions is quite essential in making decisions on when to perform machine maintenance. This paper presents a multivariate calibration monitoring approach based on the statistical analysis of machine measurement data. The calibration model is used to predict two faulty conditions from historical reference data. This approach utilizes genetic algorithms (GA) based variable selection, and we evaluate the predictive performance of several prediction methods using real data. The results shows that the calibration model based on supervised probabilistic principal component analysis (SPPCA) yielded best performance in this work. By adopting a proper variable selection scheme in calibration models, the prediction performance can be improved by excluding non-informative variables from their model building steps.

***Keywords*—**Prediction, operation monitoring, on-line data, nonlinear statistical methods, empirical model.

I. Introduction

THE monitoring or detection of abnormal events of machines is an essential part of the operational tasks to produce high quality products consistently. To achieve this, one should build models that describe the nature or sources of variation. One option is to use mathematical models or knowledge-based models. Though they are potentially powerful methods, such models are time-consuming and practically difficult to develop [1]. An alternative approach is to construct empirical models based on historical data, which are readily available for most industrial processes.

Many multivariate statistical techniques had been developed and applied to fault detection and monitoring: principal component analysis (PCA), partial least squares (PLS), principal component regression (PCR), and so forth. In addition, there has been much interest in kernel-based nonlinear learning techniques such as support vector machines (SVMs) [2]. This facilitated the development of other kernel-based methods. These nonlinear kernel-based methods have been successfully applied to many problems. They have the common aspect that input data are mapped into a nonlinear space and then these mapped data are analyzed. Such a kernel trick has also been used to develop a nonlinear kernel PCA [3], kernel PLS [4] and kernel FDA [5].

Prediction of machine or parts operating conditions for a monitoring purpose can be formulated as a multivariate calibration problem. A major objective of this calibration model is to predict unusual conditions of machine from experimental or historical measurement data. In most cases, however, the high dimensionality and collinearity of measurement data makes it difficult to build a calibration model. Many applications in a variety of areas frequently utilized PLS-based multivariate calibration models. PLS is a dimension reduction technique that seeks to find a set of latent variables by maximizing the covariance of two variable blocks (i.e., predictor **X** and response **Y**).

Recently, new calibration methods have been developed such as orthogonal-PLS (O-PLS) and supervised probabilistic principal component analysis (SPPCA) [6], [7]. The predictor variable X, in general, contains unwanted variations that are unrelated (or orthogonal) to response variable Y. In such a case, the unwanted variation may degrade the predictive ability of a calibration model. The task of feature or variable selection is a quite important step in multivariate analysis because modern industrial processes are gathering high-dimensional data from automated sensor systems of plant in an on-line basis. The exclusion of unnecessary or non-relevant variables or noises in the data must produce better results even with simpler models. The idea of variable selection in calibration is to select a set of variables, in which prediction results are better than the results obtained using the full set of variables. The selection of variables for calibration can be considered as an optimization problem. In this respect, genetic algorithm (GA) is a very efficient technique for variable selection.

The objective of this study is to evaluate the predictive performance of several well established statistical prediction methods. These methods include PLS and some more recent methods such as SPPCA and KPLS, with and without GA-based variables selection scheme. Here, GA is used as an optimization tool to determine variables that maximize predictive abilities of calibration models. The adoption of feature selection in calibration framework helps to identify an optimal subset of original variables for calibration model building. For a performance comparison in this work the prediction performance of the several prediction schemes are tested and compared. This study uses real machine measurement data for faulty conditions. Part mismatch indicates one of the prevalent faulty events associated with machines. It may cause noise or vibration and sometimes lead to accelerated wear or

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).

malfunction.

The remainder of this paper is organized as follows. First, a brief review of PLS, KPLS, and SPPCA is given in section II. Then section III presents comparison results obtained from real machine data. Finally, concluding remarks are given.

II. Methodologies

A. PLS and KPLS

PLS was developed and adopted to model the relation between a predictor matrix **X** and a response matrix Y. It seeks to find a set of latent variables that maximizes the covariance between **X** (n×N) and **Y** (n×M). PLS decomposes **X** and **Y** into the form as follows [8]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (1)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \quad (2)$$

where **T** and **U** are (n×A) matrices of the extracted A score vectors, **P** (N×A) and **Q** (M×A) loading matrices, and **E** (n×N) and **F** (n×M) residual matrices. The PLS method based on nonlinear iterative partial least squares (NIPALS) algorithm searches for weight vectors w and c that maximizes the sample covariance between **t** and **u**.

By regressing **X** on **t** and **Y** on **u** after convergence, loading vectors **p** and **q** can be obtained and then PLS regression model can be expressed using regression coefficients **B** and residual matrix **G**:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{G} \quad (3)$$

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{C}^T \quad (4)$$

As shown in Fig. 1, the goal of PLS is to use the factors to predict the responses in the population. It can be achieved indirectly by extracting latent variables **T** and **U** from sampled factors and responses, respectively. The extracted factors **T** are used to predict **U**.

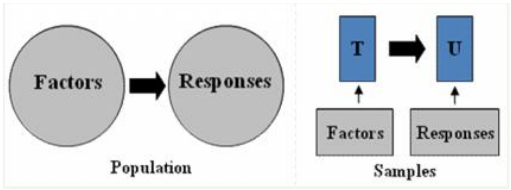

Fig. 1 A schematic diagram of partial least squares (PLS)

A nonlinear kernel version of PLS, called kernel PLS, differs from linear PLS in that original input data are first nonlinearly transformed into a nonlinear feature space via nonlinear mapping [4]. Then a linear PLS is performed in the feature space. KPLS can be easily obtained by using nonlinear kernel functions. KPLS requires only linear algebra and does not need to nonlinear optimization problems. This is the main advantage that KPLS has over other nonlinear PLS methods. KPLS also provides flexibility to model a wide range of nonlinearities by using various kernel functions. KPLS algorithm can be directly derived from linear PLS algorithm with some modifications.

B. SPPCA

While principal component analysis (PCA) originates from the analysis of data variances, probabilistic PCA, a latent variable model, defines a generative process for the data. Here, latent variables are conventionally assumed as a Gaussian distribution with zero mean and unit variance. The key concept of supervised probabilistic PCA considers that all the observations are conditionally independent given the latent variables. This indicates that the principal components in probabilistic PCA are the latent variables which best explain the data covariance. When supervised information is available, each object of input data can be associated with an output real values for regression task or binary values (e.g., 1 or -1) for classification.

The observed data (**x**, **y**) is generated from a latent variable model as:

$$\mathbf{x} = \mathbf{W}_x\mathbf{d} + \boldsymbol{\mu}_x + \boldsymbol{\varepsilon}_x$$
$$\mathbf{y} = \mathbf{W}_y\mathbf{d} + \boldsymbol{\mu}_y + \boldsymbol{\varepsilon}_y \quad (5)$$

The latent variable (**d**) and the error terms ($\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y$) are defined as isotropic Gaussians distribution:

$$\mathbf{d} \sim N(0,\mathbf{I}), \boldsymbol{\varepsilon}_x \sim N(0,\sigma_x^2\mathbf{I}), \boldsymbol{\varepsilon}_y \sim N(0,\sigma_y^2\mathbf{I}) \quad (6)$$

It was shown that the maximum likelihood estimate of $\mathbf{W}_x$ and $\mathbf{W}_y$ can be obtained (Yu et. al 2006). The projected latent variable $\mathbf{d}^*$ for centered new input $\mathbf{x}^*$ is given by

$$\mathbf{d}^* = \frac{1}{\sigma_x}\mathbf{R}^T(\mathbf{D}_p - \mathbf{I}_p)^{1/2}[\mathbf{U}_M^T\mathbf{U}_M + (\mathbf{D}_p - \mathbf{I}_p)^{-1}]^{-1}\mathbf{U}_M^T\mathbf{x}^* \quad (7)$$

III. Results and Performance

A. Data and Feature Selection

In this work data sets were collected using a motor attached through a coupling at two kinds of faulty conditions. The data recorder stored the input currents and voltages and the data sets for this analysis consist of 2,700 input variables of the frequency spectrum and the two response variables of the two faulty conditions. For the verification of a number of calibration models, a total of fifty observations were divided into three subsets so that each subset has almost equal observations. For each of subsets a calibration model is then constructed three times. That is, at each time we leave out one of the three subsets from training or model-building. As test data only the remaining subset is used so that we can obtain the prediction results using the different three test data sets. In order to do feature selection for the spectra data, genetic algorithm (GA) is implemented to select necessary variables for prediction. It should be noted that

many researchers have developed GA-based feature selection methods, each of them using a different GA structure. The algorithm [9] is used in this work because it was successfully applied to a large variety of data including spectral datasets. The feature selection in calibration is to select a subset of variables or features useful for building calibration model [10]. With feature selection prediction results of the prediction models are better than or comparable to the results obtained using the full set of variables. In this work, genetic algorithms are used as an optimization tool to select variables that maximize the predictive capabilities of calibration models for the two cases: case 1 for fault 1 condition and case 2 for fault 2 condition.

We performed comparative performance tests three times using the data obtained from the two cases of case 1 and case 2 fault conditions. The predictive performance of different calibration models was evaluated over the two faulty datasets. For comparison purposes, prediction errors for each of the 50 samples were calculated to evaluate the prediction results of calibration models. Here we used root mean squared error in prediction (i.e., RMSEP) for test datasets, which is the mean squared difference between the observed (true values) and the predicted (calibrated values).

B. Prediction Results

The prediction results of four calibration models are assessed using the data as stated earlier. The first model is a PLS model, which utilizes all the features available without performing GA-based feature selection (denoted as PLS). The second model is similar to the PLS model in that the linear technique of PLS is adopted, but it performed feature selection procedure based on genetic algorithms (denoted as GA-PLS). The remaining two models are as follows: nonlinear KPLS and SPPCA model with GA-based feature selection (denoted as GA-SPPCA).

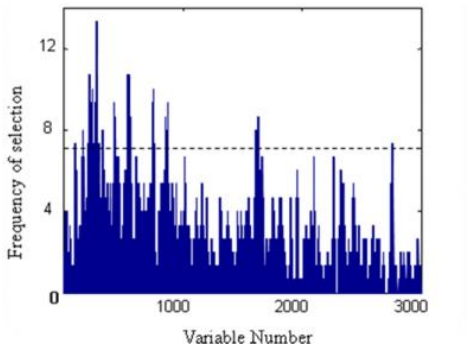

Fig. 2 Cumulative selection frequency for case 1

Prior to building the prediction models, GA was executed on each data set of the two faulty conditions with the optimal number of evaluations. This process was executed five times to verify the robustness of the predictive ability and effect of variable selection. It is due to the fact that the final solutions of different GA runs will not be exactly the same due to GA's stochastic nature. Common information from the results of five GA runs was extracted to select the most informative variables for prediction. This information is obtained from the frequency with which each variable is selected. Thus, we selected the variables that are consistently selected (i.e., variables having high selection frequency) in the five runs. Consequently, a total number of 173 (for case 1) and 213 (for case 2) variables were selected from the five selection frequency plots, one of which was shown in Fig. 2. Here, a dotted line of fig.2 indicates a cutoff value for variable selection. It is calculated by F-test, but it is out of scope of this paper [11], [12].

TABLE I
RESULTS FOR CASE 1

	RMSEP VALUE			
	PLS	GA-PLS	KPLS	GA-SPPCA
D1	1.60	1.25	1.31	1.01
D2	1.36	1.08	1.10	0.89
D3	1.54	1.18	1.22	0.96
Total	1.50	1.17	1.21	0.95

A prediction performance of the calibration models is evaluated over the datasets of the two different machine conditions. As stated before, a total of 50 observations were divided into three subsets. Then, calibration models are constructed three times, each time leaving out one of the three subsets from training or modeling data. RMSEP values for the different three test datasets were obtained for each of the four prediction models, based on which we evaluated the prediction results of the calibration models. Prediction results are summarized in Tables I (case 1) and II (case 2). Tables I and II showed root mean squared error in prediction (RMSEP) for each of the three subsets and compared the performance of the four calibration models. As listed in Table I, for example, the GA-SPPCA prediction model yielded the RMSEP value of 1.01 for subset 1, in which the training data for this case consist of subset 2 and subset 3.

TABLE II
RESULTS FOR CASE 2

	RMSEP VALUE			
	PLS	GA-PLS	KPLS	GA-SPPCA
D1	4.28	2.80	3.07	2.12
D2	3.82	2.03	2.34	1.85
D3	3.97	1.95	2.20	1.97
Total	4.02	2.26	2.54	1.98

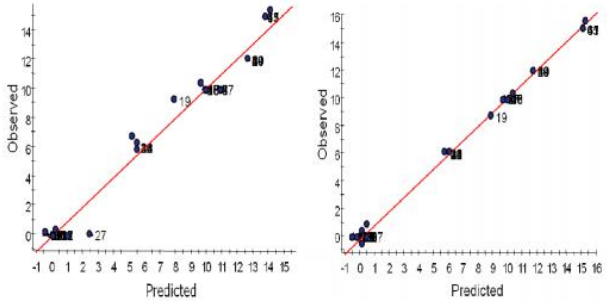

Fig. 3 Predicted vs. observed plots (a) PLS and (b) GA-SPPCA

For the case 1 of the faulty condition (Table I), the GA-SPPCA prediction model showed the best prediction performance in that it yielded the minimum average RMSEP value of 0.95. It is observed that the GA-SPPCA model has the minimum RMSEP over the entire datasets. Overall, GA-PLS, KPLS, and GA-SPPCA models showed a significantly better prediction performance for all data subsets than the linear PLS model. For example, for test data of subset 3 linear PLS calibration model yielded maximum RMSEP value while the performances of the other models are comparable.

RMSEP values for the three datasets of case 2 are listed in Table II based on the four calibration models. Similar to the faulty condition of case 1, the GA-SPPCA prediction model produced the minimum average RMSEP value of 1.98: GA-PLS with 2.26, KPLS with 2.54, and PLS with maximum of 4.02. It should be noted that in the subset 3 the RMSEP value of the GA-SPPCA (i.e., 1.97) is better than those of PLS (i.e., 3.97) and KPLS (i.e., 2.20), but lower than GA-PLS (i.e., 1.95). The effect of adopting GA-based feature selection in calibration can be seen by comparing the prediction results of calibration models. When compared to the PLS model, the GA-PLS prediction model produced better prediction performance: average RMSEP value from 2.54 to 2.26. Although not shown here, this observation is also valid for KPLS models of both case 1 and case 2. This is also the case for GA-SPPCA in that average RMSEP values of simple SPPCA models (without GA-based feature selection) for the two cases deteriorated slightly. Thus from the prediction results it can be stated that the models constructed based on selected features yielded better performance than the models without feature selection.

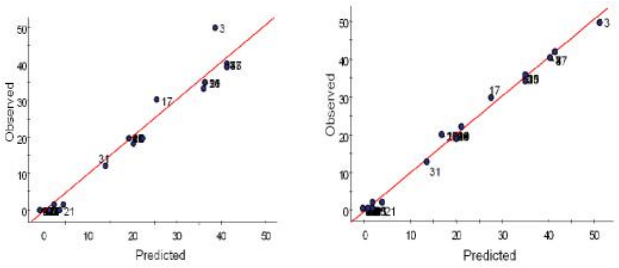

Fig. 4 Predicted vs. observed plots (a) PLS and (b) GA-SPPCA

For a visualization purpose of the prediction performance of the calibration models, predicted values were plotted against observed values as shown in Fig. 3 and 4. Fig. 3 shows the two plots (the PLS and GA-SPPCA models) for case 1 of the faulty condition. In such a plot the data should fall on the diagonal when calibration models predict the response values perfectly. In this respect, the GA-SPPCA models for case 1 and case 2 has a better predictive ability than PLS models. The GA-SPPCA models produced the predicted values close to the diagonal line. However, the predicted values of PLS are relatively different from the observed one.

IV. Conclusion

This study presented the use of a combination of multivariate calibration and GA-based feature selection as a method for analyzing high-dimensional machine measurement data. A GA-based feature selection was performed to identify key variables that retain most of information of the original data relevant to prediction of different fault conditions. The exclusion of non-informative variables made it possible to produce better prediction using fewer selected. The effectiveness of the presented prediction scheme was demonstrated using machine data in which the ultimate goal is to predict unusual events in wrong part setup or aligning. It turned out that the prediction results of the GA-SPPCA model improved significantly compared to other simple prediction models. The construction of appropriate prediction model for faulty condition monitoring helps us to make decisions on whether there is abnormal event occurred or not. In the near future, we will investigate how the proposed calibration scheme can be used together with other advanced data mining techniques for enhanced prediction of high-dimensional data.

References

[1] Y. S. Nga, R. Srinivasana, "An adjoined multi-model approach for monitoring batch and transient operations," *Computers and Chemical Engineering*, vol. 33, pp. 887–902, 2009.
[2] V. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, 1995, New York, NY.
[3] B. Schölkopf, A. J. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299-1319, 1998.
[4] R. Rosipal, and L. J. Trejo, "Kernel partial least squares regression in reproducing Kernel Hilbert space," *Journal of Machine Learning Research*, vol. 2, pp. 97–123, 2001.
[5] G. Baudat, and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, pp. 2385-2404, 2000.
[6] J. Trygg, and S. Wold, "Orthogonal projections to latent structures (O-PLS)," *Journal of Chemometrics*, vol. 16, pp. 19–128, 2002.
[7] S. Yu, K. Yu, V. Tresp, H. Kriegel, and M. Wu, "Supervised probabilistic principal component analysis. In: Proceedings of the 12th international conference on knowledge discovery and data mining (SIGKDD), pp 464–473, 2006.
[8] K. Kourti, "Application of latent variable methods to process control and multivariate statistical process control in industry," *International Journal of Adaptive Control and Signal Processing*, vol. 19, pp. 213–246, 2005.
[9] R. Leardi, and A. L. Gonzalez, "Genetic algorithms applied to feature selection in PLS regression: how and when to use them," *Chemometrics Intelligent Laboratory Systems*, vol. 41, pp. 195–207, 1998.
[10] A. Durand, O. Devos, C. Ruckebusch, and J. P. Huvenne, "Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared

quantitative analysis of cotton–viscose textiles," *Analytica Chimica Acta*, vol. 595, pp. 72–79, 2007.
[11] C. S.Soh, P. Raveendran, and R. Mukundan, "Mathematical models for prediction of active substance content in pharmaceutical tablets and moisture in wheat," *Chemometrics and Intelligent Laboratory Systems*, vol. 93, pp. 63–69, 2008.
[12] Y. Shao, and Y. He, "Nondestructive measurement of the internal quality of bayberry juice using Vis/NIR spectroscopy," *Journal of Food Engineering*, vol. 79, pp. 1015–1019, 2007.