Data Quality Enhanced Prediction Model for Massive Plant Data

Moon-Ghu Park [a*], Seong-Ki Kang [b], Hajin Shin [c]
[a]*Nuclear Engr. Sejong Univ., 209 Neungdong-ro, Seoul, Korea 05006*
[b]*Monitoring & Diagnosis, 1556 Duckyoungdaero, Suwon, Korea 16693*
[b]*Saint Paul Preparatory Seoul, 14-8, Seochojungang-ro 31-gil, Seoul, Korea 06593*

[*]*Corresponding author: mgpark@sejong.ac.kr*

1. Introduction

Recent extensive efforts for on-line monitoring implementation [1] insists that a big surprise in the modeling for predicting process variables was the extent of data quality problems in measurement data especially for data-driven modeling. Bad data for training will be learned as normal and can make significant degrade in prediction performance. For this reason, the quantity and quality of measurement data in modeling phase need special care. Bad quality data must be removed from training sets to the bad data considered as normal system behavior. The types of bad data are categorized as follows;

- Poor data acquisition
- Data lockup or frozen signals for extended periods.
- Missed data
- Unphysical abnormal fluctuations
- Loss of significant digits
- Random noise
- Unreasonable values
- Interpolation errors

This paper introduces an integrated signal pre-conditioning and model prediction mainly by kernel functions. The performance and benefits of the methods are demonstrated by a case study with measurement data from a power plant and its components transient data. The developed methods will be applied as a part of monitoring massive or big data platform where human experts cannot detect the fault behaviors due to too large size of the measurements.

2. Methods and Results

The main stream of recent technology in treatment of massive data in fault detection the invariant analysis [2]. The invariant means a mathematical function of the input-output data of $y = f(x;t)$. The function $f(x;t)$ is a data-driven model constructed under an intact condition [2]. The model produces reference signals to identify the system faults.

2.1 Noise filtering

The noisy measurements can give undesirable predictions. To filter out the measurement noise we introduce the bilateral kernel filter [3]. Measurement produces a set of random variables $\{t_i, y_i; i = 1, 2, ..., N\}$ on the interval $\{0 \le t_i \le T\}$. It is assumed that

$$y_i = y(t_i) + \varepsilon \tag{1}$$

where ε is a random noise variable with the mean equal to zero. The purpose of the bilateral kernel filter is to smooth out the small noise details and to preserve edge signals, no specific noise characteristic of ε is assumed. The kernel estimate of $y(t)$ at $t = \tau$ from this data is defined by

$$\hat{y}(\tau) = \frac{\sum_{i=1}^{N} y_i K(\tau - t_i)}{\sum_{i=1}^{N} K(\tau - t_i)} \tag{2}$$

The function K selected as the bilateral Gaussian function, i.e.,

$$\begin{aligned} K(t) &= K_D(\text{distance}) \times K_F(\text{feature}) \\ &= \exp(-D(t_i, t_q)^2 / \sigma_t^2) \times \exp(-D(y_i, y_q)^2 / \sigma_x^2) \end{aligned} \tag{3}$$

where σ_t^2, σ_x^2 are the variances for noise filtering and feature preservation, respectively. D is the Euclidean distance defined by

$$D(t_i, t_q) = \left\| t_i - t_q \right\| = \sqrt{(t_i - t_q)^2} \tag{4}$$

and t_q is the query. The bandwidth of the kernel σ^2 controls the width of measurements being spread around a query point. Figure 1 shows the noise filtering and edge preservation performance of the filter for a step signal.

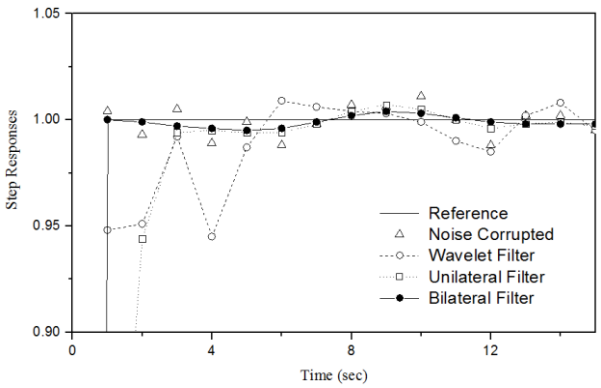

Figure 1. Bilateral filter performance for step signal

2.2 Kernel regression

After a preprocessing phase with the bilateral filter the data-driven model is given by the kernel regression with a same framework. The memory vectors of measurements used to develop the data-driven prediction model with p observations of n process variables is given by :

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p,1} & X_{p,2} & \cdots & X_{p,n} \end{bmatrix} \quad (5)$$

The weights are also generated by the Gaussian kernel :

$$\mathbf{w} = K_M(\mathbf{d}) = \exp(-\mathbf{d}^2 / \sigma^2) \quad (6)$$

where σ is the kernel bandwidth, w are the weights for the p memory vectors. Here the variables are grouped in to classes with large correlation coefficient. Hoeffding correlation function [4] is used to consider the nonlinearities between variables and optimize memory size. Figure 2 shows a typical grouping result grouped by prescribed cut-off value of the correlation.

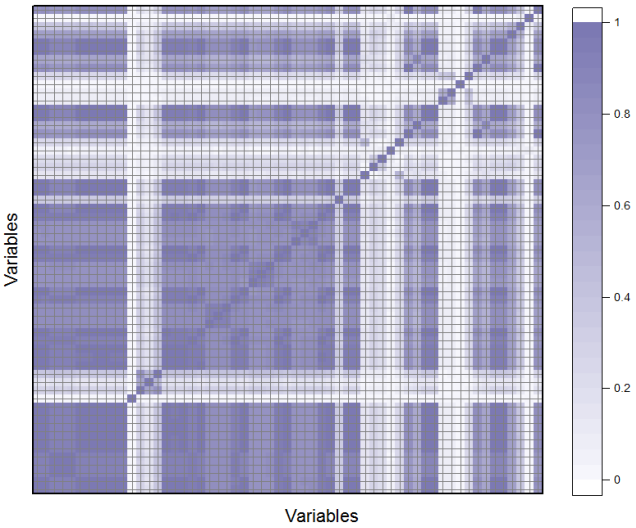

Figure 2. Group structure of correlation map

Figure 3 shows the prediction performance of the data-driven model which shows almost perfect predictions. Only 30% of the measurements are selected as the training set and the figure shows excellent prediction for entire data set.

3. Conclusions

This paper presented an integrated structure of supervisory system for monitoring the plants or sensors performance. The quality of the data-driven model is improved with a bilateral kernel filter for preprocessing of the noisy data. The prediction module is also based on kernel regression having the same basis with noise filter. The model structure is optimized by a grouping process with nonlinear Hoeffding correlation function. The framework is highly emphasized to be allied to massive data treatment due to its easy implementation and constructive structure via vector processing.

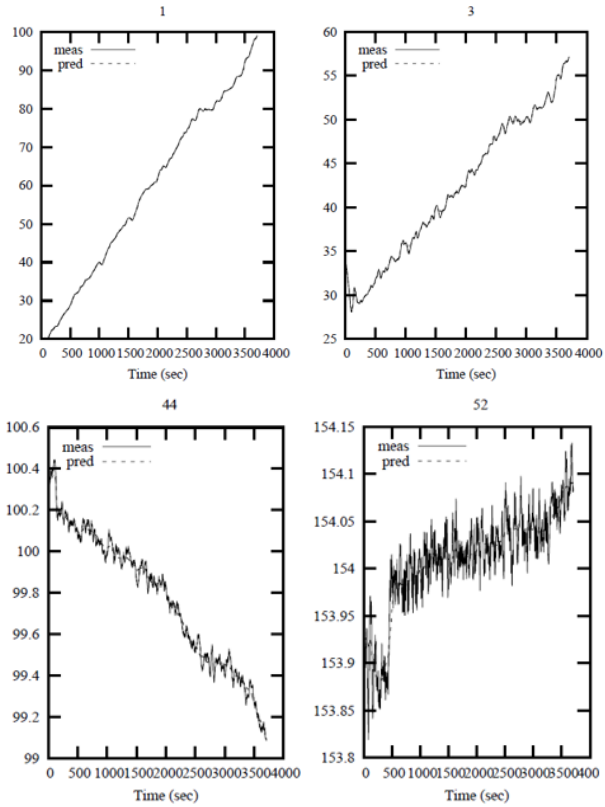

Figure 3. Prediction performance of data-driven model

REFERENCES

[1] Eddie Davis *et. al*., On-Line Monitoring at Nuclear Power Plants–Results From the EPRI On-Line Monitoring Implementation Project, 45th ISA POWID Symposium, Jun. 2002

[2] Fukushima K, Kato M, Hino I, Terasawa S, Yamamoto T, Ooishi T. Failure Sign Monitoring System for Large-scale Plants Applying System Invariant Analysis Technology (SIAT). NEC Technical Journal, Vol. 9(1), p.115, 2015.

[3] Park M, Shin H, Lee E. Kernel-Based Noise Filtering of Neutron Detector Signals. 2007; Nuclear Engineering and Technology, Vol. 39 (6), p. 1, 2007.

[4] Suzana Santos et. al. A comparative study of statistical methods used to identify dependencies between gene expression signals, Briefings in Bioinformatics, Vol. 15(6), p. 906, 2013.