Stepwise Grouping Algorithm for Statistical Learning Framework

Sang Ha An[a], Gyunyoung Heo[b*], Ho Joon Seo[c], Su Young Kim[c], Soon Heung Chang[a]
[a]*Korea Advanced Institute of Science and Technology, Yuseong-gu, Daejeon, 305-701, Korea*
[b] *Kyung Hee University, Yongin-si, Gyeonggi-do, 446-701, Korea*
[c] *BNF Technology Inc., Yuseong-gu, Daejeon, 305-500, Korea*
[*]*Corresponding author: gheo@khu.ac.kr*

1. Introduction

In general, condition-based maintenance means the maintenance program which introduces additional sensors like vibration, acoustic, ultrasonic and infrared sensor to monitor equipment condition. The methodology which uses only process variables already installed in the system to detect anomalies, efficiency degradation or malfunction of sensors also have been studied. Among them, commercialization potential of process/sensor anomalies detection system using empirical model is considered high because it does not have additional system requirements for complex systems [1-3]. However, some problems have been pointed out while these solutions having been introduced. For example, (1) missing important signals in variable selection and (2) variable duplicity problem for empirical model have been pointed out. This paper proposes complementary framework and detailed methodologies, and performs experimental validation by using a heat conduction experimental device which is available for flexible fault injections.

2. Methods and Results

In this section some of the techniques used for variable selection and the experiment results for validating those techniques are described.

2.1 Variable Grouping Methods

In a fault detection algorithm using empirical models, the variable selection is one of the most important parts because it has great effect on the overall accuracy of the model. Previous methods have normally used correlation coefficients for variable grouping. If the correlation coefficient of variables is higher than a certain value, the variables will be joined in a single group, and if it is lower than a set point, it will be discarded. Variables can be, therefore, missed if they show a low correlation coefficient for some reasons even though they have significant physical relationship.

Stepwise grouping method has been applied to solve this problem. The stepwise grouping procedure starts off by choosing an equation containing the single best X variable and then attempts to build up with subsequent additions of X's one at a time as long as these additions are worthwhile. The order of addition is determined by using the correlation coefficient to select which variable should enter next. The lowest RMSE (Root Mean Square of Error) is compared to a (selected or default) RMSE-to-enter value. After a variable has been added, the equation is examined to see if any variable should be deleted.

The basic procedure is as follows. First we select the Z most correlated with Y (suppose it is Z_1) and find the regression equation. We check if this variable is significant. If it is not, we quit and adopt the model $Y = \overline{Y}$ as best; otherwise we search for the second predictor variable to enter regression.

The Z_j with the highest such value (suppose this is Z_2) is now selected and a second regression equation is fitted. The overall regression is checked for significance, the improvement of RMSE is examined. The lower of these RMSE is then compared with an appropriate point, and the corresponding predictor variable is retained in the equation or rejected according to whether the test is significant or not significant. This testing of the "least useful predictor currently in the equation" is carried out at every stage of the stepwise procedure. A predictor that may have been the best entry candidates at an earlier stage may, at a later stage, be superfluous because of the relationships between it and other variables now in the regression [4].

2.2 Experimental Validation

A simple heat conduction experiment has been performed to validate the effectiveness of proposed method. An experimental device consists of a DC heater, a fan cooler and 6 thermocouples and they are shown in Figure 1.

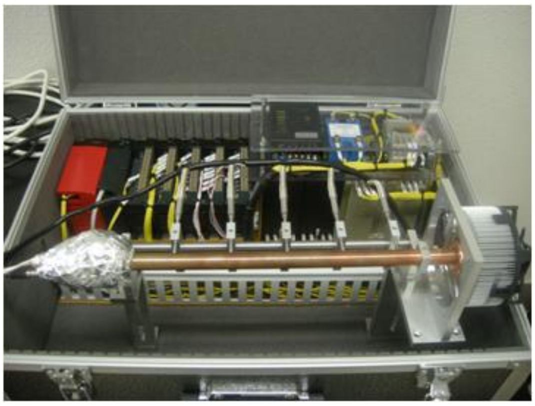

Figure 1. Device for heat conduction experiments

There are 9 variables available including a heater power, a cooler power, 6 temperatures on a copper rod, and atmospheric temperature. With these variables, the training was performed to check the model. It is obvious that the cooler and atmosphere temperature affects all of the temperatures in terms of physical relationship. But the method using only correlation coefficients made a group of 6 T/C signals only, while the stepwise method was able to detect the contribution of cooler and atmosphere and so it is composed of 6 T/Cs, cooler and atmosphere temperature. Root mean square of error resulting from two methods with training data is shown in Table 1. Training result using stepwise grouping is shown in Figure 2. Blue line in second row is actual value and green line is estimated value. And the red line in third row shows the residual.

Table 1. Comparison of RMSE by grouping methods

	RMSE
By Correlation Coefficient	0.3329
By Stepwise Grouping	0.2134

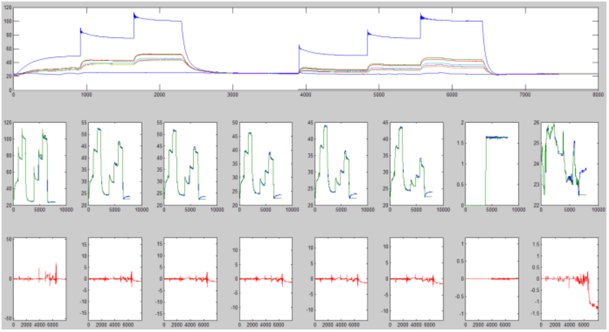

Figure 2. Training result using stepwise regression

3. Conclusions

This paper emphasizes that the pre-processing of data is carefully carried out especially for fault detection or diagnosis using empirical models while, only a modeling method is concerned. However, the most significant factor that affects the accuracy of model is how the training data is prepared and how a grouping is performed. Authors also suggest engineering judgment should be included even for empirical models.

REFERENCES

[1] L. Monostori, Computer-aided generation of monitoring strategy for complex machine tool monitoring systems, Measurement, Volume 8, Issue 3, July-September 1990, Pages 98-102.

[2] G. Heo, Condition monitoring using empirical models: technical review and prospects for nuclear applications, Nuclear Engineering and Technology, Volume 40, Issue 1, December 2007, Pages 98-102.

[3] S.H. An, G. Heo, H.J. Seo, S.Y. Kim, S.H. Chang, Statistical learning framework with adaptive retraining for condition-based maintenance, Transactions of the Korean Nuclear Society Spring Meeting, May 2009, Jeju, Pages 324.

[4] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, p.335, 1998.