Competitive Adaptive Reweighted Sampling Method for Fault Detection

Wang Xuan¹,², Yanxue Wang¹,²,*

¹School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044
²Beijing Key Laboratory of Performance Guarantee on Urban Rail Transit Vehicles, Beijing University of Civil Engineering and Architecture, Beijing 100044
*Corresponding author: bemail: wangyanxue@bucea.edu.cn

Abstract: According to Darwin's theory of evolution, a variable combination with the strongest correlation between selection process variables and quality indicators is developed. This is called competitive adaptive reweighted sampling (CARS). In this article, the absolute value of the partial least squares regression coefficient is used to assess the importance of each variable. Next, we select variables based on the regression coefficients of variables and quality indicators, including forced variable selection based on exponential decreasing function (EDF) and competitive variable selection based on adaptive reweighted sampling (ARS). Finally, we use cross-validation (CV) to select the subset with the lowest root mean square error of CV (RMSECV). The Tennessee-Eastman (TE) process is used to evaluate the performance of the proposed fault detection method. The results show that CARS can find the best combination of certain key variables, improving the fault detection capability of quality indicators.

1. Introduction
A variety of fault occur in industrial production, and fault related to product quality have always been one of the most concerned issues. Therefore, quality-related fault detection has become a hot topic in recent years. Because data-driven fault diagnosis[1-2] is easy to implement and requires fewer system models, it has attracted great attention in the field of process monitoring. Partial Least Squares (PLS)[3] is a typical multivariate statistical analysis method used for fault detection. Principal Component Analysis (PCA)[4] is another most commonly used method in process monitoring. However, Li et al.[5] pointed out that the standard PLS performs tilt decomposition on the process variable space, and its residual subspace may still contain important process variables related to the output. Therefore, for quality-related fault detection, the statistical data and fault detection provided by PLS are problematic. In order to overcome the shortcomings of standard PLS, Zhou et al.[6] proposed a PLS-based post-processing method, namely Total Projection to Latent Structures (T-PLS). In this way, by designing appropriate statistics in the four subspaces, the quality-related faults can be classified. Later, Yin et al.[7] proposed another method, called Modified Partial Least Squares (M-PLS), which decomposes the process space X into two orthogonal subspaces. Compared with T-PLS, M-PLS achieves orthogonal decomposition in the process space. By comparing the advantages of the two methods, Qin et al.[8] proposed a Concurrent Partial Least Squares (C-PLS), which is said to have more advantages than the former.
But our further research found that the choice of variables in the traditional method is determined according to the correlation between each variable and the quality index. Generally, the correlation of a single variable to quality indicators is generally low. Under normal circumstances, industrial process data has high collinearity problems. Selecting variables based on the correlation between a single variable and quality indicators is obviously not the best method, which will have a negative impact on the detection ability of the model. From an optimization perspective, variable selection can be seen as a process to maximize the detection performance of the calibration model. Therefore, it is natural to adopt an optimization algorithm, which can find the best combination of variables. The rest of this article is arranged as follows. The second part reviews CARS theory. The third part uses CARS algorithm for fault analysis and fault detection. The fourth part draws a conclusion.

2. CARS theory

2.1. PLS and weights of variables

In partial least squares regression, there is only one output variable, and Y is a vector at this time. Partial least squares method modeling is to decompose X and Y as follows

\[ X = TP^T + E \]  
\[ Y = UQ^T + F \]  

In the formula, T and U are score matrices. P and Q are load matrices corresponding to X and Y respectively. E and F are residual matrices.

PLS is a widely used process. Establish a linear relationship between X and y based on the latent variables. Supposed the score matrix is represented by T, it is a linear combination of X and W. c is the least squares regression coefficient of y to T. Therefore, we have the following formula

\[ T = XW \]  
\[ Y = Tc + e = WXc + e = Xb + e \]  

Where e is the prediction error, and \( b = Wc = [b_1 \ b_2 \ \cdots \ b_p] \) is \( p \)-dimensional coefficient vector. The absolute value of the i-th element in b represents the contribution of the variable to the quality index. therefore, it is natural to say that when \( |b_i| \) is larger, the i-th variable is more important. In order to evaluate the importance of each variable, we define the normalized weight as

\[ W_i = \frac{|b_i|}{\sum_i |b_i|} \quad i = 1,2,3,\ldots,p \]  

Therefore, we manually set the weight of the variable eliminated by CARS to 0. We ensure that the weight vector \( W \) is \( p \)-dimensional.

2.2. Exponentially decreasing function

Assuming that the entire process data contains \( p \) variables, the variable selection of CARS involves two steps. The first step is to remove variables with relatively small regression coefficients by using EDF. The second step uses EDF to remove other variables. Define the ratio of EDF to calculate the variables you want to keep.

\[ r_i = ae^{-ki} \]  

Where a and k are two constants. In the first sampling, all \( p \) variables are used for modeling, which means \( r_1 = 1 \). In the \( N \) th sampling, only two variables are kept, so we get \( r_N = \frac{2}{p} \). Under these two conditions, a and k can be calculated as

\[ a = \left( \frac{2}{p} \right)^{1/(N-1)} \]  
\[ k = \frac{\ln(\frac{p}{2})}{N-1} \]  

Where \( \ln \) represents the natural logarithm.
Figure 1: Graphical illustration of the exponentially decreasing function. In the first stage, the number of the variables are reduced fast while in the second stage, it decreases very slowly which realizes a refined selection.

Figure 1 illustrates an example of EDF. It can be clearly seen that the process of variables reduction can be roughly divided into two stages. In the first stage, variables are quickly eliminated, performing a ‘fast selection’. While in the second stage, variables are reduced in a very gentle way, called a ‘refined selection’. Due to the characteristics of EDF, it can effectively remove variables with no or little information in the process variables. This is why we choose EDF.

Assuming that the MC sampling times of CARS is set to $N$, CARS will select $N$ variable subsets in turn. In short, in each sampling, CARS requires four consecutive steps: (1) Monte Carlo is used for the sampling of the model. (2) Using EDF to perform forced variable selection. (3) Using ARS to realize the choice of competitive variables. (4) Cross validation[9] is used to evaluate the subset.

3. Tennessee eastman example

The Tennessee-Eastman Process (TE) [10] developed by Eastman can be used to evaluate process control and inspection methods. Before fault application, we select 8 types of quality-related fault IDVs (1, 2, 5, 6, 8, 10, 12, 13), and 5 types of quality-unrelated fault IDVs (3, 4, 9, 11, 15).

Figure 2: TE process data at fault ID1, the increase in the number of samples (plot above), 10-fold the RMSECV values (plot center) and regression coefficients of each variables (plot below) as the sampling runs increase. The line (marked with an asterisk) represents the best point at which the 10-fold the RMSECV value reaches the lowest value.
As can be seen in Figure 2, the RMSECV value first drops rapidly from sampling 1-3, which should be attributed to the elimination of non-informative variables. Then the gentle change from sampling 4-6, 8-46, corresponding to the stage where there is no significant change in sampling variables. Finally, eliminating the key variables P1, P2 and resulting in loss of information, the RMSECV value rises rapidly at L2, L3.

It should be noted that the plot (below) of Figure 2 shows the coefficient path of each variable. Each line of plot (below) records the coefficient of each variable in different sampling runs. Therefore, a subset of variables and regression coefficients can be extracted from each sampling run. The best subset with the lowest RMSECV value can be marked with an asterisk vertical line. More interestingly, because the coefficient of the variable (indicated by P1) just dropped to zero, the RMSECV value jumped to a higher level at the sampling point (indicated by the dashed line L2). When the coefficient of the other variable P2 drops to zero, the dotted line marked L3 also jumps sharply. Such observations show that there are key variables. Without these variables, the performance of the model will drop sharply. Therefore, they are called key variables. All in all, the simulation study shows that CARS is an effective variable selection method.

Two commonly used indicators, fault detection rates (FDRs) and fault alarm rates (FARs), are used for performance evaluation. Therefore, FDR and FAR can be defined as

\[
FDR = \frac{N_{\text{nea}}}{N_{\text{tfa}}} \quad (9)
\]

\[
FAR = \frac{N_{\text{faa}}}{N_{\text{tfa}}} \quad (10)
\]

Table 1: FDRs of quality-related faults of the TE process(%)

| Fault ID | PLS  | CARS-PLS |
|----------|------|----------|
| IDV(1)   | 99.75| 99.75    |
| IDV(2)   | 98.50| 99.00    |
| IDV(5)   | 29.84| 30.21    |
| IDV(6)   | 99.38| 99.88    |
| IDV(8)   | 97.00| 98.88    |
| IDV(10)  | 84.77| 91.01    |
| IDV(12)  | 99.63| 99.38    |
| IDV(13)  | 95.01| 95.13    |

Table 2: FARs of quality-unrelated faults of the TE process(%)

| Fault ID | PLS  | CARS-PLS |
|----------|------|----------|
| IDV(3)   | 8.49 | 8.86     |
| IDV(4)   | 84.14| 6.62     |
| IDV(9)   | 7.37 | 9.49     |
| IDV(11)  | 66.92| 12.98    |
| IDV(15)  | 14.86| 12.48    |

Based on Table 1, 2, CARS-PLS has a slightly higher detection capability than PLS for 9 quality-related fault. Among the 5 types of fault that are not related to quality, CARS-PLS showed a lower false alarm rate. Among the IDV3, IDV9, and IDV15, the difference in FARs between CARS-PLS and PLS was small. However, Among the IDV4 and IDV11, the FARs of CARS-PLS is far lower than that of PLS. In these two types of faults, the FARs of PLS reaches more than 66%, and PLS basically fails. This is because CARS uses the optimal combination of variables to model. It can avoid the influence of highly linearly related variables on the model, improving the anti-interference ability of the model, and making the model more robust. In summary, for the TE process, the CARS method shows a lower FARs for quality-independent faults.
4. Conclusion
This paper proposes a competitive adaptive reweighted sampling method, which combines PLS technology to select key variables. In each sampling process, exponential decreasing function and adaptive weighted weighted sampling control variable number. By applying it to industrial data sets, it shows that CARS is a method to eliminate uninformed variables and select variables to establish a high-performance model. Using a combination of some key variables can achieve better fault detection. Highly collinear variables may reduce the stability of the calibration model.

Although in this work, the selection of variables is carried out through CARS combined with PLS. It should be pointed out that it can also be used in combination with other modeling methods. Our future work focuses on the study of CARS for minor faults and the application of CARS in other fields.

Acknowledgments
The financial sponsorship from the project of National Natural Science Foundation of China (51875032, 51475098 and 61463010) and Guangxi Natural Science Foundation (2016GXNSFFA380008) are gratefully acknowledged.

References
[1] Ding, S.X., Zhang, P., Jeinsch, T.(2011) A survey of the application of basic data-driven and model-based methods in process monitoring and fault diagnosis. IFAC Proceedings. 44: 12380-12388.
[2] Liu, Q., Wang, Y.X., Xu, Y.G. (2020) Synchrosqueezing extracting transform and its application in bearing fault diagnosis under non-stationary conditions. Measurement, 108569.
[3] Dayal, B.S., MacGregor, J.F.(1997) Improved PLS algorithm. Journal of Chemometrics. 11: 73-85.
[4] Wang, G., Luo, H., Peng, K.(2016) Quality-related fault detection using linear and nonlinear principal component regression. Journal of the Franklin Institute. 353: 2159-2177.
[5] Li, G., Qin, S.J., Zhou, D.H.(2010) Geometric properties of partial least squares for process monitoring. Automatica. 46: 204-210.
[6] Zhou, D.H., Li, G., Qin, S.J.(2010) Total projection to latent structures for process monitoring. AIChE Journal. 56: 168-178.
[7] Yin, S., Ding, S.X., Zhang, P.(2011) Study on modifications of PLS approach for process monitoring. IFAC Proceedings. 44: 12389-12394.
[8] Qin, S.J., Zheng, Y.Y.(2012) Concurrent projection to latent structures for output-relevant and input-relevant fault monitoring. In: Conference on Decision and Control. New York. pp. 10-13.
[9] Wold, S., Sjöström, M., Eriksson, L.(2001) PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems. 58: 109-130.
[10] Vogel, D.(1997) A plant-wide industrial process control problem. Computers and Chemical Engineering. 17: 245-255.