Quantitative analysis of a process industry’s operating status of based on DCS data set

Qian Chen¹ and Kai Sun²

¹School of Mechanical and Electrical Engineering, Xi’an University of Architecture and Technology, Xi’an 710055, China
²Institute of Microelectronic of Chinese Academy of Science, Beijing 100029, China

E-mail: clara_sun96@163.com

Abstract. The process industry represented by energy chemical enterprise is a typical distributed and complex electromechanical system. Its distributed control system (DCS) records overall operating status information of a process industry. It is a challenge to quantitative operating status through analyzing DCS data set. This paper proposed a novel method to calculate fault score of system’s operating status by classifying its DCS data set and forming a classifier matrix. Firstly, define a classifier matrix depends on the baseline of the DCS data. Secondly, calculate the classifier matrix to obtain the fault score. The higher fault score means the lower security of the system. Last but not least, a period of the system operating status trend could be observed by plotting the fault score. A case study of a real chemical plant is introduced to verify the validation of the method.

1. Introduction

The process industry, such as chemical industry, nuclear system, electrical power system etc., plays an important role in the Chinese economy. It is a typical complex electromechanical system that consists of a collection of electromechanical devices coupled and operated with each other. Failure of any internal component may cause dysfunction of the whole system. As the increase of the DCS data set, it is difficult to establish mathematical expressions because of the huge, nonlinear, and coupling of DCS data. Therefore, it is significant to quantify the high-dimensional data and master the overall operation state of the system to prevent the occurrence of safety accidents.

The DCS data set is a multivariate time series, which compose thousands of various time series, reflecting the operating status of the whole system. Conventionally, data-driven multivariate statistical analysis [1] is widely used to analyze DCS data set. including Principal Component Analysis (PCA), Kernel Function Principal Component Analysis (KPCA), Independent Component Analysis (ICS), Partial Least Squares Algorithm (PLS). The partial least squares (PLS) [2-3] and principal component...
analysis (PCA) [4] are the representative techniques of Measurement System Analysis (MSA). Both methods are challenged by the process of non-Gaussian distribution and nonlinear multivariate time series. Independent component analysis (ICA) focuses on dealing with non-Gaussian characteristics [5-7]. Neural-network-related PCA (NN-PCA) [8-9], kernel principal component analysis (KPCA) [10-12] and kernel independent component analysis (KICA) [13-14] are developed to solve nonlinear problem. SVM [15-16] establishes a robust non-linear model to offer accurate quantitative predictions in classification. The main idea of these methods is to reduce the dimension of data by using multivariable mapping, and to filter some "important" data and omit other data. Because of the complex nonlinear coupling relationship between the variables in the huge DCS data set, the traditional methods tend to lose part of "useful" data, which leads to the inaccuracy grasp the system operation status. Therefore, it is a challenge to find a way to quantitative analysis of all high-dimensional data and system operation status.

Sun Kai, et al. [17] proposed that the whole DCS data set be mapped into a group of black and white digital images distinguish the normal status and abnormal status of the system with two colors. This method initially realizes the quantification of system operation status, but the classification of operation status is simple, the classification of system fault anomaly degree is not detailed enough, and the grasp of the quantification operation status trend of the system is not accurate enough, and the operation status trend has some offset from the real working condition. In order to solve these problems, this paper subdivides the normal state and abnormal state of the system, and proposed the definition of classification matrix to quantitatively grade the monitor data according to the 0-N deviation degree. The fault score of the system is calculated to accurately realize the quantitative analysis of the operation status of the system.

The outline of this paper is as follows. Section 2 defines the DCS data matrix, and extract perfect value range of each sensor in DCS data as operation status baseline. Section 3 introduces data classifiers and classification matrices to achieve data abnormal degree. Section 4 defines the system fault score and proposes a method to quantify the system operation status. Section 5 illustrates the process of the proposed method through an actual case of chemical enterprise, which verifies effectiveness of the work in this paper. This paper is closed with a conclusion in section 6.

2. DCS Baseline

DCS data set can be regarded as a multivariable time series set, including $n$ monitor variables. Each variable intercept $m$ sample values in a monitor interval. Sampling monitor variables are arranged as row vectors and sampling time series as column vectors. A $m \times n$ two-dimensional data matrix can be constructed.

**Definition 1** Data Matrix $X$ is a matrix of monitor data set, as equation (1) shows. Each row of the data matrix $X$ represents overall sensors’ monitor data in a special time period.

$$X(m, n) = \begin{bmatrix}
    x_{11} & x_{12} & \ldots & x_{1n} \\
    x_{21} & x_{22} & \ldots & x_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{m1} & x_{m2} & \ldots & x_{mn}
\end{bmatrix}^{\text{Sensors}}_{\text{Time}} \quad (1)
$$

In equation (1) $x_{ij} \in X, 1 \leq i \leq m, 1 \leq j \leq n, i, j \in N$, where $i$ represents the special time point and $j$ marked the sensor number of the DCS. The row vector $\begin{bmatrix} x_{i1} & x_{i2} & \ldots & x_{in} \end{bmatrix}$ of data matrix $X$ represents the dynamic characteristics of the system at a specific time point. The one-dimensional column vector $\begin{bmatrix} x_{ij} & x_{2j} & \ldots & x_{3j} \end{bmatrix}^T$ of data matrix $X$ represents the one-dimensional time series of system monitor variables.

DCS data baseline defines mean and as important attributes. Each sensor in the system has mean
value and standard deviation as the evaluation standard of monitor data. The benchmark of mean value and variance of the system has been determined at the beginning of DCS design, which is the inherent property of DCS. The data baseline of each variable can be obtained directly from DCS.

**Definition 2**  
Per_value is $1 \times n$ vector composed with $N$ parameters’ benchmark, as equation (2) shows.

$$\text{Per\_value} = [\mu_1 \ \mu_2 \ \ldots \ \mu_j \ \ldots \ \mu_N]_{1 \times N}$$  

**Definition 3**  
Std is a standard deviation vector of overall sensors, as equation (3) shows.

$$\text{Std} = [\sigma_1 \ \sigma_2 \ \ldots \ \sigma_j \ \ldots \ \sigma_N]_{1 \times N}$$  

**Definition 4**  
Standard Region is a benchmark range for the classification of DCS monitor data anomalies. When the monitor data of variable $j$ is within the standard region, it means that the system is in normal operation status, as equation (4) shows.

$$\mu_j - \sigma_j \leq x_j \leq \mu_j + \sigma_j$$  

In equation (4), $x_j \in X$. The more the monitor data of variable $j$ deviates from the standard region, the higher the anomaly degree of the system will be.

### 3. Classifier Matrix

The monitor data of sensor $j$ during the sampling period represents the operation status of the system. By defining the concept of Data Classifier, the monitor data of sensor $j$ can be classified according to the degree of deviation from Standard Region $\mu_j \pm \sigma_j$, and the purpose of quantitative grade can be achieved.

**Definition 5**  
Data classifier (status Region) is a matrix composed of standard region value and abnormal deviation region value of each variable, as equation (5) shows.

$$\text{status\_Region} = \begin{bmatrix} \mu_1 + h\sigma_1 & \mu_2 + h\sigma_2 & \ldots & \mu_j + h\sigma_j & \ldots & \mu_N + h\sigma_N \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_1 + \sigma_1 & \mu_2 + \sigma_2 & \ldots & \mu_j + \sigma_j & \ldots & \mu_N + \sigma_N \\ \mu_1 - \sigma_1 & \mu_2 - \sigma_2 & \ldots & \mu_j - \sigma_j & \ldots & \mu_N - \sigma_N \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mu_1 - h\sigma_1 & \mu_2 - h\sigma_2 & \ldots & \mu_j - h\sigma_j & \ldots & \mu_N - h\sigma_N \end{bmatrix}_{2h \times N}$$  

The abnormal degree of DCS data can be separated into $h$ degree. For each variable parameter of the system, the quality grade interval between perfect and fault is $h$. Figure 1 shows the arrangement of the quality grade interval parameter $j$. 
The main feature of the data classifier is to classify the floating-point data which is difficult to identify directly, and to replace the floating-point value with a simple integer value 0-N to achieve simple and meaningful data, as Figure 2 shows.

**Definition 6**

Classifier Matrix \((T)\) is a matrix that obtained by classifying and compressing DCS data set by data classifier, as equation (6) shows.
According to the three Sigma quality control principle, when the fault deviation degree in the classification matrix is 0~2, the system is in the no-fault operation status. Once the fault deviation level is greater than 2, the system has different degrees of fault. The greater the fault deviation, the higher the fault degree.

4. System Fault score
Each element in the classifier matrix $T$ represent the abnormal degree of the variable at a special time point. That is, element equal with zero represents the best operation status. When all the elements of $T$ are zero, which means the idealism status, the cumulative sum of is zero. Vice versa, the summary equals with $mn \times n \times N$ means the worst status. Under the actual operation condition, the fault deviation degree of all variables is added to obtain the overall operation fault degree of the system in a certain period of time.

**Definition 7** Fault Score is obtained by accumulating all the fault deviation degree in the classification matrix, and calculate the total fault deviation degree of the system within a period of time, as equation (7) shows.

$$\text{Fault score } = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2, \quad x_{ij} \in T$$

Fault score represents the overall operation status of the system. According to the three Sigma quality control principle, when the system fault score is within the range of 0~$m \times n \times 2$ that the system is in a safe operation status. Once the system score exceeds this upper limit range, the higher the system fault score, the worse operating state of the whole system.

5. Use Case
The air compressor unit of a chemical plant is a distributed electromechanical complex system, and its equipment connection is shown in Figure 3. The compressor unit is composed of steam turbine, air compressor, supercharger gearbox and auxiliary equipment. All equipment of the system is connected by connecting pipes, and the equipment is coordinated with each other to achieve the function of air compression. The air compressor unit of the chemical plant includes 250 sensors in total, and there are various of sensors, such as temperature, flow, pressure, speed, power etc. These sensors upload the monitor data to the DCS in real time.
The DCS data set of the compressor unit in the second half of 2012 and the first half of 2013 were analyzed. According to definition 2–6, the daily monitor data in these two time periods are classified and compressed according to the deviation degree of fault. According to definition 7, elements of classification matrix are accumulated to get system fault score. Plot the daily fault scores of after half of 2012 and first half of 2013 were, shown in Figure 4 and Figure 5.

Figure 3. Air compressor equipment connection diagram

The DCS data set of the compressor unit in the second half of 2012 and the first half of 2013 were analyzed. According to definition 2–6, the daily monitor data in these two time periods are classified and compressed according to the deviation degree of fault. According to definition 7, elements of classification matrix are accumulated to get system fault score. Plot the daily fault scores of after half of 2012 and first half of 2013 were, shown in Figure 4 and Figure 5.

Figure 4. Air Compressor health status of in the second half of 2012

Figure 4 shows the overall health of the system tends to deteriorate from July 15 to December 22 of 2012. On Oct. 27, the fault score of the system was at the maximum value, which was about three times higher than before, and the degree of system abnormality was greatly increased. Although the fault score showed a downward trend after October 27, the overall fault score was still not within the safe range. And on November 22, the system operation state deteriorated again. According to the trend shown in Figure 4, it can be used as the basis for the early warning of overhaul. At this time, the whole system should be shut down for overhaul. Actually, it has been overhauled on Dec.23.
Figure 5. Air Compressor health status in the first half of 2013

After the shutdown and overhaul at the end of 2012, the system restarted on Feb 13, 2013. It’s can be seen from Figure 5 that the system fault score on seven peak days is higher than that on adjacent days. From February 14 to 16, 2013, the system was in the worst operation status. However, since the system was in started state during that period, it was a normal phenomenon that the system was unstable. As time goes by, the overall operation status of the system is gradually improved, and the operation status of the system tends to be stable. Hence, the fault score decreases gradually except several tiny peak after April 4, 2012.

6. Conclusion
This paper meets the challenge with how to quantify a complex electromechanical system based on its DCS data set. By proposing a concept of classifier matrix and fault score, the operating status of system which is recorded by thousands of variables every minute can be described by one numerical score. Plotting the score for a period of time, the working condition trend can be showed directly, which could be the judging basis for the system overhaul. The use case of an air compressor from a real chemical plant validate the effectiveness of the method.

Acknowledgments
This paper is supported by National Nature Science Foundation of China (grant no. 51705393).

References
[1] Baris Gun Sürmeli and M. Borahan Tümer 2020 Multivariate Time Series Clustering and its Application in Industrial Systems J. Cybernetics and Systems 51(3)
[2] Paul Nomikos and John F. MacGregor 1995 Multivariate SPC charts for monitoring batch processes J. Technometrics.37(1) 41–59
[3] Kumar K. and Mishra AK. 2015 Application of Partial Least Square (PLS) Analysis on Fluorescence Data of 8-Anilinonaphthalene-1-Sulfonic Acid, a Polarity Dye, for Monitoring Water Adulteration in Ethanol Fuel J. Constr.Eng.Prac. 25(4) 1055–61
[4] Wold S., Esbensen K. and Geladi P. 1987 Principal component analysis J. Chemometrics and Intelligent Laboratory Systems 2 37–52
[5] Lee J. M., Yoo C. K. and Lee I. B. Statistical process monitoring with independent component analysis J. Journal of Process Control 14 467–85

[6] Lee J. M., Qin S. J. and Lee I. B. 2006 Fault detection and diagnosis of multivariate process based on modified independent component analysis J. AIChE Journal 52 3501–14

[7] Wang J., Zhang Y., Cao H. and Zhu W. 2012 Dimension reduction method of independent component analysis for process monitoring based on minimum mean square error J. Journal of Process Control 22 477–87

[8] Chen J. H. and Liao C. M. 2002 Dynamic process fault monitoring based on neural network and PCA J. Journal of Process Control 12 277–89

[9] Dong D., McAvoy and T. J. 1996 Nonlinear principal component analysis—based on principal curves and neural networks J. Computers and Chemical Engineering 20 65–78

[10] Alcala C.F. and Qin S. J. 2010 Reconstruction-based contribution for process monitoring with kernel principal component analysis J. Industrial and Engineering Chemistry Research 49 7849–57

[11] Ge Z. Q., Yang C. J. and Song Z. H. 2009 Improved kernel PCA-based monitoring approach for nonlinear processes J. Chemical Engineering Science 64 2245–55

[12] Lee J. M., Yoo C. K., Choi S. W., Vanrolleghem P. A. and Lee I. B. 2004 Nonlinear process monitoring using kernel principal component analysis J. Chemical Engineering Science 59 223–34

[13] Lee J. M., Qin S. J. and Lee I. B. 2007 Fault detection of non-linear processes using kernel independent component analysis J. Canadian Journal of Chemical Engineering 85 526–36

[14] Zhang Y. W. 2008 Fault detection and diagnosis of nonlinear processes using improved kernel independent component analysis (KICA) and support vector machine (SVM) J. Industrial and Engineering Chemistry Research 47 6961–71

[15] Jemwa G. T. and Aldrich C. 2005 Improving process operations using support vector machines and decision trees J. AIChE Journal 51 526–43

[16] Zeng G. M., Li X. D., Jiang R., Li J. B. and Huang G. H. 2006 Fault diagnosis of WWTP based on improved support vector machine J. Environmental Engineering Science 23 1044–54

[17] Sun Kai, Gao Jianmin, Gao Zhiyong, Jiang Hongquan and Gao Xu 2015 Plant-wide quantitative assessment of a process industry system’s operating state based on color-spectrum J. Elsevier Ltd 60-61