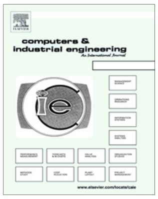
Computers & Industrial Engineering 85 (2015) 157–168

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie



A general framework for monitoring complex processes with both in-control and out-of-control information ☆

Chi Zhang [a], Fugee Tsung [a,*], Changliang Zou [b]

[a] *Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*
[b] *School of Mathematical Sciences, Nankai University, Tianjin 300071, China*

ARTICLE INFO

Article history:
Received 20 August 2014
Received in revised form 3 March 2015
Accepted 15 March 2015
Available online 20 March 2015

Keywords:
EWMA
Model-free procedure
Statistical process control
Support vector machine

ABSTRACT

Processes monitoring using multivariate quality variables remains an important and challenging problem in statistical process control (SPC). Although multivariate SPC has been extensively studied in the literature, the challenges associated with designing robust and flexible control schemes have yet to be adequately addressed. This paper develops a general monitoring framework for detecting location shifts in complex processes by employing data mining methods. The historical in-control (IC) and out-of-control (OC) data are combined to set up a support vector machine (SVM) model. The working status of the process is indicated by the probabilistic outputs of the SVM classifier and the multivariate exponentially weighted moving average strategy is applied to construct the control chart. A fast diagnostic procedure can be implemented as soon as the control chart gives an alarm. Our numerical studies show that the proposed control chart is able to deliver satisfactory IC and OC run-length performance regardless of the underlying distributions and data types. An example using real data from an industrial application demonstrates the effectiveness of the proposed method.

1. Introduction

Advanced sensing technologies have facilitated real-time data collection for process monitoring and fault diagnosis in increasingly complex industrial processes. A large amount of data and information related to quality measurements in complex processes has thus become available from a range of systems including natural (physical or chemical) and manmade (manufacturing or computer) systems. Although there is no explicit definition of complex processes, the concept of complexity can be divided into several categories, such as logistical complexity resulting from a high volume of tasks and product proliferation, and technological complexity, which is related to the inherent complexity of the system (Khurana, 1999). Complex processes often share some common features, including various data types, difficulty in modeling the data as well as large amount of data provided in the processes. Considering the complicated mechanism and high repairing cost of complex processes, statistical approaches that make use of the data and process information regarding control, monitoring and diagnosis would benefit industrial practice immensely.

Statistical process control (SPC) is a typical monitoring method and is widely used in manufacturing and service industries. The objective of SPC is to detect possible shifts in location or scale as soon as they occur, so that remedial actions can be taken without delay. A complex process usually involves multiple quality characteristics and hence it is common to monitor several quality variables simultaneously. Among others, the multivariate cumulative sum (MCUSUM, see Woodall & Ncube, 1985; Crosier, 1988), the multivariate exponential weighted moving average (MEWMA, see Lowry, Woodall, Champ, & Rigdon, 1992), the regression-adjusted control chart (Hawkins, 1991) and variable-selection-based control charts (Zou & Qiu, 2009; Wang & Jiang, 2009) have been developed. We recommend Woodall and Montgomery (2014) for an excellent literature review. In addition, most complex operations, such as semiconductor manufacturing and automotive body assembly, comprise multiple stages. Effort has also been devoted to the monitoring of multistage processes. Tsung, Li, and Jin (2008) and Shi and Zhou (2009) provide detailed reviews.

These aforementioned studies were mostly built based on certain assumptions, and it is assumed the collected data are continuous and they follow normal/multivariate normal distribution in previous works. However, as the complexity of modern industrial

☆ This manuscript was processed by Area Editor H.-S. Jacob Tsao.

* Corresponding author at: Room 5552, Academic Building, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Tel.: +852 2358 7092.

E-mail addresses: czhangaf@ust.hk (C. Zhang), season@ust.hk (F. Tsung), nk.chlzou@gmail.com (C. Zou).

http://dx.doi.org/10.1016/j.cie.2015.03.007
0360-8352/© 2015 Published by Elsevier Ltd.

processes grows, these assumptions may no longer hold and several challenges remain. The latter involves two major challenges, both of which come from the inherent features of complex process. First, data hybridity (multiple data types) prevents us from using the popular model-based methods. Due to intrinsic properties, it is not possible to obtain the exact continuous values for many quality characteristics. Categorical data and mixed-type data (including both continuous data and categorical data) have become increasingly common in complex processes. But it is generally difficult to set up an appropriate model to describe the relationship among categorical characteristics (Marcucci, 1985; Taleb, 2009; Li, Tsung, & Zou, 2014) as well as the relationship between categorical and numerical characteristics. And few literature has really addressed the mixed-type data monitoring problem.

The second challenge in the on-line monitoring of modern processes is the data distribution. As we have mentioned, normal/multivariate normal data are often assumed in control charts, which is not simply the case in many industrial applications. Furthermore, in modern processes, little prior system information is provided and the data distribution is often unknown. Without knowing the parametric form of the data distribution, it is quite difficult to describe process variables, let alone the correlation among variables. A motivating example is the application of the Hard Disk Drive Monitoring System (HDDMS). The HDDMS is a computerized system that monitors various attributes of hard disk drives, such as read error rate, spin-up time, and reallocated sector count. A scatter plot of two attributes—hard disk assembly temperature and current pending sector counts—is given in Fig. 1. It is easy to see that none of the commonly used distributions would fit this data. In such cases, the statistical properties of commonly used charts, which were designed with the normal distribution (or some known distribution/model) in mind, could potentially be severely affected. Considering data hybridity from the first challenge invalidates some traditional nonparametric method, model-free monitoring schemes are required to solve this problem.

Another feature of complex processes as mentioned is the large amount of data they generate. The HDDMS, for example, can produce more than 1 million pieces of IC data and 120,000 pieces of OC data a week. This OC data need not go to waste. In many industrial processes, the same faults may occur again and again. By accumulating and analyzing OC data, one may be able to extract common error patterns that could be used to predict the next shift.

To summarize, the challenges such as data hybridity and unknown distribution cause conventional monitoring methods to fail, and the availability of historical OC data motivates us to develop new monitoring schemes. In this paper, a general framework based on statistical learning techniques is proposed. A popular classification model—the support vector machine (SVM)—is employed and the multivariate EWMA technique is incorporated. The probabilistic output of the SVM model is used as the charting statistic. The proposed monitoring schemes has the following advantages: (i) It fully exploits historical OC information for detecting location shifts; (ii) our method well solves the challenges mentioned above, and it works robustly under multiple data types and unknown data distribution scenario; and (iii) the procedure for implementing the control chart is quite simple and straightforward, and little effort is required for model setup or parameter estimation.

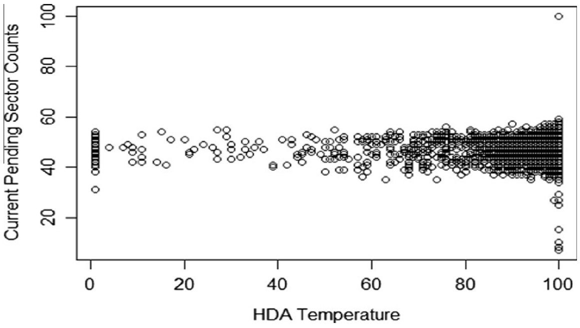


Fig. 1. Scatter plot of hard disk assembly temperature and current pending sector counts in the Hard Disk Drive Monitoring System.

The remainder of this paper is organized as follows: our proposed methodology is described in detail in Section 2; its numerical performance is thoroughly investigated in Section 3; the method is demonstrated in Section 4 using a real-data example from service processes; finally several remarks in Section 5 conclude the paper. Some technical details are provided in Appendix A.

2. A monitoring framework for complex processes

2.1. Problem description and literature review

In complex processes, we are provided with a large historical dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, in which $\boldsymbol{x}$ is often multivariate and it may contain both continuous and categorical variables. In the historical dataset, the IC and OC data may be mixed or well separated. If the data are mixed, Phase I methods may be applied to cluster them. Then to detect location shift, the on-line monitoring observations (supposing they start from $N+1$) are assumed to follow the change-point model:

$$\boldsymbol{x}_i \sim \begin{cases} F(\boldsymbol{x} - \boldsymbol{\mu}_0) & \text{for } i = N+1, \ldots, N+\tau, \\ F(\boldsymbol{x} - \boldsymbol{\mu}_1) & \text{for } i = N+\tau+1, \ldots \end{cases} \tag{1}$$

where $F(\cdot)$ is the unknown data distribution and $N+\tau$ is the change point. $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the process means of the IC and OC states, respectively, and $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$. In this paper, we also use $\text{sign}(\cdot)$ to denote the sign function, and $P(\cdot), \widehat{P}(\cdot)$ to denote the probability of certain event and its estimated value, respectively.

In practice, the real shift size $\boldsymbol{\mu}_1$ is not known in advance. As the shifts/faults in complex processes may occur repeatedly from practical experience, information about $\boldsymbol{\mu}_1$ could be retrieved by making use of the provided historical OC data. Assume that the historical data reveal K kinds of OC situations and the shift magnitudes equal to $\boldsymbol{\mu}_{h1}, \boldsymbol{\mu}_{h2}, \ldots, \boldsymbol{\mu}_{hK}$, then $\boldsymbol{\mu}_1$ should be close to one of those historical shifts. Using the K shifts reminds us of the directional charts, where the assumption lies on $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_{h1}$ or $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_{h2}$ or $\ldots\boldsymbol{\mu}_1 = \boldsymbol{\mu}_{hK}$. However, in our monitoring framework, the real shift $\boldsymbol{\mu}_1$ does not necessarily exist exactly in historical shifts. Instead, it should be similar to one of them, in terms of both magnitude and direction. Later we will show the case where the real shift the $\boldsymbol{\mu}_1$ varies in magnitude with historical shifts. In addition, the number of historical OC clusters K is often derived from Phase I, and the procedure will be described in later chapter.

Considering data hybridity and a lack of information about data distribution, data mining tools are utilized in our monitoring framework. Recently, control charts based on data mining tools have gained popularity as well, see the review paper by Hachicha and Ghorbel (2012). Hwang, Runger, and Tuv (2007), Hu and Runger (2010) and Deng, Runger, and Tuv (2012) developed the artificial contrast method to solve the problem, in which regularized least square classification (RLSC) and random forest (RF) are used as the classifiers. Guh and Shiue (2008) also applied decision tree for process control. On the other hand, as an efficient classification tool, the SVM model is widely used and some monitoring methods based on this model have been proposed. Sun and Tsung (2003) first developed the K-chart based on the support vector data description (SVDD), and some following works (Camci, Chinnam, & Ellis, 2008; Sukchotrat, Kim, & Tsung, 2010;