Expert Systems with Applications 39 (2012) 2073–2081

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Data mining model-based control charts for multivariate and autocorrelated processes

Seoung Bum Kim [a], Weerawat Jitpitaklert [b], Sun-Kyoung Park [c], Seung-June Hwang [d,*]

[a] School of Industrial Management Engineering, Korea University, Seoul, Republic of Korea
[b] Department of Industrial and Manufacturing Systems Engineering, University of Texas at Arlington, Arlington, TX, USA
[c] School of Business Administration, Hanyang Cyber University, Seoul, Korea, Republic of Korea
[d] Collage of Business & Economics, Hanyang University ERICA Campus, Ansan, Republic of Korea

ARTICLE INFO

Keywords:
Autocorrelated process
Multivariate process
Model-based control chart
Statistical process control
Data mining

ABSTRACT

Process monitoring and diagnosis have been widely recognized as important and critical tools in system monitoring for detection of abnormal behavior and quality improvement. Although traditional statistical process control (SPC) tools are effective in simple manufacturing processes that generate a small volume of independent data, these tools are not capable of handling the large streams of multivariate and autocorrelated data found in modern systems. As the limitations of SPC methodology become increasingly obvious in the face of ever more complex processes, data mining algorithms, because of their proven capabilities to effectively analyze and manage large amounts of data, have the potential to resolve the challenging problems that are stretching SPC to its limits. In the present study we attempted to integrate state-of-the-art data mining algorithms with SPC techniques to achieve efficient monitoring in multivariate and autocorrelated processes. The data mining algorithms include artificial neural networks, support vector regression, and multivariate adaptive regression splines. The residuals of data mining models were utilized to construct multivariate cumulative sum control charts to monitor the process mean. Simulation results from various scenarios indicated that data mining model-based control charts performs better than traditional time-series model-based control charts.

1. Introduction

One of the key management systems in organizations is planning for quality. Organizations consider planning for quality as a part of their strategic planning. Without careful strategic planning for quality, organizations could lose large amounts of money, market share, time, and effort (Montgomery, 2001). Therefore, business/manufacturers should focus on planning for quality as a way to develop a competitive edge in the market. Quality control and improvement include a set of activities implemented to achieve product and service specifications. SPC methodologies have frequently been used to avoid poor quality. A control chart is an important tool used in SPC to monitor the performance of a process over time to keep the process within control limits. Control charts are based on solid statistical theory and provide a comprehensive graphical display that can be readily configured by the users with minimal assistance. A typical control chart comprises the monitoring statistics and the control limits. When the monitoring statistics exceed (or fall below) the control limits, an alarm is generated so that the process can be investigated before defective units are produced.

Univariate control charts were devised to monitor the quality of a single process characteristic. However, modern processes often involve a large number of highly correlated process characteristics. Although univariate control charts can be applied to each individual characteristic, this technique may lead to unsatisfactory results when multivariate problems are involved. Moreover, high-throughput technologies in modern industries are capable of generating data for short intervals that in their brevity leads to an autocorrelation problem. Traditional multivariate control charts were developed and came into use to solve these problems. However, they have become less and less capable of handling the large streams of complex and auto/cross-correlated data found in modern manufacturing and service systems.

Hotelling's T^2 chart is the most widely used multivariate control chart because it can simultaneously and efficiently monitor multiple correlated process characteristics. The main assumptions of T^2 control charts are the normality and independency of observed process data. That is, successive multivariate observations are assumed to be independent, identically, and normally distributed over time. Some other types of multivariate control charts include the multivariate cumulative sum (MCUSUM) control chart (Crosier,

* Corresponding author.
E-mail address: sjh@hanyang.ac.kr (S.-J. Hwang).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.010

1988; Healy, 1987; Pignatiello & Runger, 1990; Woodall & Ncube, 1985) and the multivariate exponentially weighted moving average (MEWMA) control chart (Lowry, Woodall, Champ, & Rigdon, 1992). Both were devised for increased sensitivity to detect small process shifts. Although the MCUSUM and MEWMA charts are known to be relatively robust, compared with Hotelling's T^2 control chart, for nonnormal and autocorrelated data, failure to use multivariate control charts carefully with autocorrelated data may result in deterioration of monitoring performance (Alwan, 1992; Montgomery & Mastrangelo, 1991). Increased rates of false alarms are one possible result of such deterioration.

Model-based control charts that yield the residuals – the difference between the actual values and the fitted values from the models used – have been the traditional way to address autocorrelation problems in process monitoring. Model-based control charts have been effectively used in monitoring multistage processes in which the output process variable(s) of interest are related to the input process variables from the previous and current stages (Loredo, Jearkpaporn, & Borror, 2002). A regression adjustment control chart, developed by Hawkins (1991), monitors the residuals from the process variable of interest when that variable is regressed on all the others. A regression adjustment control chart is especially useful when a process variable of interest exhibits autocorrelation because the residuals from the regression model are typically uncorrelated. However, its parametric assumption of an error term in linear regression analysis limits its applicability for handling nonnormal process data. A number of other model-based control charts are available (Alwan L.C. & Roberts, 1988; Jiang, Tsui, & Woodall, 2000; Montgomery & Mastrangelo, 1991; Runger & Willemain, 1995; Zhang, 1998).

Alwan L.C. and Roberts (1988) proposed a two-step approach containing two control charts, one called a common-cause chart and the other, a special-cause chart. The approach works well in detecting large process mean shifts. Montgomery and Mastrangelo (1991) proposed the EWMA center line control chart. Their approach works well if the observations are positively autocorrelated and if the process mean does not drift too rapidly. Runger and Willemain (1995) proposed the unweighted batch means (UBM) chart. This approach monitors the average value of observations and does not use a residual-based control chart. Zhang (1998) proposed an exponentially weighted moving average for stationary process (EWMAST) chart to deal with a stationary autocorrelated process. The chart works well when the process has low positive autocorrelation and small mean shifts. Jiang et al. (2000) proposed a charting technique based on autoregressive moving average statistics, the ARMA chart. All of the methods discussed above, however, deal with the occurrence of autocorrelation in univariate processes. They do not address autocorrelation in multivariate processes.

As the limitations of SPC methodology become increasingly obvious in the face of evermore complex manufacturing processes, data mining algorithms, because of their proven capabilities to effectively analyze and manage large amounts of data, have the potential to resolve the problems that are stretching SPC to its limits. Despite their great potential, however, few efforts have been made to integrate data mining algorithms with SPC. Arkat, Niaki, and Abbasi (2007) used artificial neural networks (ANNs) to build a model and construct a MCUSUM chart using the residuals for multivariate and autoregressive processes. They compared the average run length (ARL) performance of the three methods: autocorrelated charts, time-series-based residuals charts, and ANN-based residuals charts and concluded that ANN-based residuals charts outperformed the other two charts for small mean shifts in processes. ARL is the average number of observation required for the chart to detect a change (Woodall & Montgomery, 1999). In-control ARL (ARL_0) and out-of-control ARL (ARL_1) were, respectively, calculated under in-control and out-of-control processes. Issam and Mohamad (2008) used support vector regression (SVR) to construct the residuals-based MCUSUM control chart. They calculated the residuals from one-step-ahead prediction. That is, current observations are used as input to forecast future observations. They concluded that SVR-based residuals charts performed better than time-series-based residuals charts and ANN-based residuals charts when small mean shifts were involved. This idea is interesting, but their main conclusion was derived based on limited simulation scenarios. Their studied did not investigate the different degrees of autocorrelation. Thus, their methods need to be justified much more thoroughly via simulation under various scenarios.

Our proposed approach differs from Issam and Mohamad (2008) in how it finds the residuals. To illustrate, for a process with three variables; x_1, x_2, and x_3, we use x_1 and x_2 as inputs to create a model that predicts x_3. The residuals of this model are obtained for monitoring x_3. We apply the same procedure to the other variables until we get the residuals from all variables. The assumption behind our proposed approach to obtain the residuals is that degrees of autocorrelation of individual process variables are not significantly different. This is a reasonable assumption because the process variables from an equipment may have similar degrees of autocorrelation. In the present study, we conducted a simulation study under various scenarios including multiple dimensions and different degree of autocorrelation.

The focus of the present study is the development of the new process monitoring methodology that can effectively deals with complex multivariate autocorrelated processes. Specifically, we use such state-of-the-art data mining models as multivariate adaptive regression splines (MARS), ANNs, and SVR. Multivariate control charts will then be used to monitor the residuals of the output variables from these data mining models.

The rest of this paper is organized as follows. In Section 2, we briefly explain the data mining models used for the model-based control charts. Section 3 illustrates the simulation study and performance comparisons among various data mining model-based control charts based on ARL measures. Section 4 presents our concluding remarks.

2. Data mining algorithms and MCUSUM Chart

2.1. Multiple linear regression

Multiple linear regression (MLR) is a parametric approach that renders a linear equation to examine the relation of the mean response to multiple predictor variables. The coefficient of each predictor variable in the linear equation is estimated by a least squares estimation technique that minimizes the summation of the squared deviation between the actual and fitted values. MLR models have been widely used for prediction problems because of their simplicity. However, MLR models may lead to inefficient and unsatisfactory conclusions when the relationship between the response and predictor variables is nonlinear. Moreover, a parametric assumption of error term in MLR often restricts its applicability to many complicated multivariate data.

2.2. Time-series regression

Although linear regression models are easy to implement, they do not account for the autocorrelation structure of the process. The time-series regression procedure consists of two steps. In the first step, an ordinary least square regression procedure is implemented to fit the model. Next, the autocorrelation function and the partial autocorrelation function of the residuals are employed to determine the appropriate autoregressive and moving average time-series model (Box, Jenkins, & Reinsel, 2008). In the second step, the