Expert Systems with Applications 41 (2014) 1701–1707

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Gower distance-based multivariate control charts for a mixture of continuous and categorical variables

Gulanbaier Tuerhong, Seoung Bum Kim *

School of Industrial Management Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-701, South Korea

ARTICLE INFO

Keywords:
Gower distance
Multivariate control charts
Mixture data
Quality control
Statistical process control

ABSTRACT

Processes characterized by high dimensional and mixture data challenge traditional statistical process control charts. In this study, we propose a multivariate control chart based on the Gower distance that can handle a mixture of continuous and categorical data. An extensive simulation study was conducted to examine the properties of the proposed control chart under various scenarios and compared it with some existing multivariate control charts. The simulation results revealed that the proposed control chart outperformed the existing charts when the number of categorical variables increases. Furthermore, we demonstrated the applicability and effectiveness of the proposed control charts through a real case study.

1. Introduction

Statistical process control (SPC) tools are widely used in monitoring and improving output quality in the manufacturing and service industries (Woodall, 2000; Woodall & Montgomery, 1999). Control charts, which are based on solid statistical theory, are the most widely used tool in SPC (Montgomery, 2005). Their main purpose is to detect any assignable changes that affect output quality. Monitoring statistics and control limits are the two major components in construction of a control chart. Monitoring statistics, plotted on a control chart, can be established as a function of observations. Control limits are generally determined based on the probability distribution of the monitoring statistics with user-specified false alarm rates. Out-of-control signals for a monitored process are issued when the corresponding monitoring statistic exceeds (or falls below) the control limit.

Control charts can be divided into univariate and multivariate charts based on the number of quality characteristics that they monitor. Univariate charts monitor a single quality characteristic, and multivariate charts monitor a number of quality characteristics simultaneously. The most widely used multivariate control chart is a Hotelling's T^2 control chart. Its monitoring statistic is the distance between an observation and the scaled-mean, estimated from in-control observations. The control limit of a Hotelling's T^2 control chart is proportional to the percentile of the *F*-distribution, assuming that the data follow a multivariate normal distribution (Hotelling, 1947). The necessity of this distributional assumption has restricted the applicability of Hotelling's T^2 control charts to situations in which the data are nonnormally distributed.

To address this problem, many distribution-free control charts have been proposed (Bakir, 2006; Chakraborti, Van Der Laan, & Bakir, 2001; Liu, 1995; Liu, Singh, & Teng, 2004; Phaladiganon, Kim, Chen, Baek, & Park, 2011; Qiu, 2008; Qiu & Hawkins, 2001, 2003; Sukchotrat, Kim, & Tsung, 2009; Sun & Tsung, 2003; Tuerhong, Kim, Kang, & Cho, 2012; Yang, Lin, & Cheng, 2011). A comprehensive review of univariate distribution-free control charts can be found in Chakraborti et al. (2001). As for multivariate cases, Liu (1995) developed a multivariate nonparametric control chart that uses the concept of data depth. Moreover, to improve the location detection capability of the previous data depth-based chart, Liu et al. (2004) later proposed a nonparametric multivariate data depth moving average control charts. However, both of these data depth methods require a high computational load, which makes them less efficient for many modern processes that involve many quality characteristics (Ning & Tsung, 2012). Qiu and Hawkins have worked on developing distribution free rank-based multivariate cumulative sum procedures to handle nonnormal distributed process data (Qiu & Hawkins, 2001, 2003). However, their methods assume that the distribution of the in-control data is known. Recently, several other useful nonparametric multivariate control charts based on sign test have been proposed (Das, 2009; Zou & Tsung, 2011; Zou, Wang, & Tsung, 2012).

Further, some studies have been conducted to integrate data mining algorithms with control chart techniques. Sun and Tsung (2003) introduced a kernel-based multivariate control chart that uses support vector data description to handle nonnormally distributed processes. He and Wang (2007) presented a multivariate control chart based on a *k* nearest neighbor algorithm. In terms of low computational cost and better detection of out-of-control

* Corresponding author. Tel.: +82 2 3290 3397.
E-mail address: sbkim1@korea.ac.kr (S.B. Kim).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.068

signals, Cui, Li, and Wang (2008) proposed an improved version of kernel principal component analysis-based multivariate control charts. Sukchotrat et al. (2009) proposed a K^2 control chart based on a k nearest neighbor data description. Stefatos and Hamza (2009) proposed a multivariate control chart based on a robust covariance matrix and principal component analysis. Yu and Xi (2009) proposed an on-line monitoring approach based on a neural network ensemble technique. El-Midany, El-Baz, and Abd-Elwahed (2010) proposed a control scheme using artificial neural networks. Bush, Chongfuangprinya, Chen, Sukchotrat, and Kim (2010) developed a nonparametric multivariate control charts using a linkage ranking algorithm. Phaladiganon et al. (2011) proposed a bootstrap-based multivariate T^2 control chart for the situations in which the distribution of observed data is nonnormal or unknown. Kim, Jitpitaklert, Park, and Hwang (2012) proposed control charts for multivariate and autocorrelated processes that use various data mining algorithms. Verdier and Ferreira (2011) proposed an adaptive Mahalanobis distance-based multivariate control chart. Their approach showed good performance with data that have a local structure. Recently, Tuerhong, Kim, Kang, and Cho (2012) proposed a distribution-free multivariate control chart based on a hybrid novelty score.

All of the aforementioned approaches are designed for processes, characterized by continuous quality characteristics. However, in some modern industries the data contain both continuous and categorical variables. In service industries, for example, a credit card transaction dataset described in Prodromidis and Stolfo (1999) contain a mixture of 30 continuous and categorical variables, designed to detect fraudulent transaction. To the best of our knowledge, only a few efforts have been made to develop multivariate nonparametric control charts for mixture data. In one such effort, Hwang, Runger, and Eugene (2007) proposed a multivariate control chart using artificial contrast that converts the monitoring problem into a supervised classification problem. The basic idea behind their approach is to generate out-of-control data from a uniform distribution and create labels (classes) to build classification models. In another approach, Hu, Runger, and Eugene (2007) simulated artificial out-of-control data from a nonuniform distribution to detect the mean shifts in more specific directions. Hu and Runger (2010) proposed an exponentially weighted moving average version of the approach in Hwang et al. (2007) to improve detection capability. Deng, Runger, and Eugene (2012) proposed system monitoring with real-time contrasts. Unlike the recourse of the artificial contrasts embraced in the other approaches, Deng et al.' approach builds a new classifier for each new observation, and this enables its on-line monitoring capability. One advantage of these artificial contrast-based control charts (Hu & Runger, 2010; Hu et al., 2007; Hwang et al., 2007, Deng et al., 2010) is that they can treat mixture data. However, unlike conventional control charts, their construction relies on supervised classification methods that necessarily require out-of-control data as well as in-control data. Recently, Ning and Tsung (2012) proposed a density-based control chart that uses a local outlier factor and show that their approach can efficiently handle processes characterized by a mixture of continuous and categorical variables. However, the simulation study presented to demonstrate the usefulness of their proposed approach has limitations, especially with data that have a large number of categorical variables.

In the present study, we propose nonparametric multivariate control charts based on the Gower distance to handle a mixture of continuous and categorical data. In the proposed Gower distance-based control chart, the monitoring statistic is the value of the Gower distance, and the control limits can be calculated by a bootstrap percentile method.

The rest of the paper is organized as follows. In Section 2, we describe the proposed Gower distance-based multivariate control chart in terms of its monitoring statistics and control limits. Section 3 presents a simulation study that examines the performance of the proposed control chart and compared it with existing ones under various scenarios. In Section 4, we use real data to demonstrate the feasibility and effectiveness of the proposed control charts. Finally, Section 5 contains concluding remarks and topics for future study.

2. Proposed Gower distance-based multivariate control charts for mixture data

2.1. Gower's dissimilarity coefficient

Let q be the size of dimension and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{x}_{p+1}, \ldots, \mathbf{x}_q)$ be a mixture observation, characterized by p categorical variables and q–p continuous variables. Thus, the vector $\mathbf{x}$ can be rewritten as follows:

$$\mathbf{x} = (z_1, \ldots, z_p, c_1, \ldots, c_{q-p})^{\mathrm{T}} = (\mathbf{z}^{\mathrm{T}}, \mathbf{c}^{\mathrm{T}}) \qquad (1)$$

where $\mathbf{z}^{\mathrm{T}}$and $\mathbf{c}^{\mathrm{T}}$ represents the vector of the subset of $\mathbf{x}$ containing the p categorical variables and q–p continuous variables. Gower's dissimilarity coefficient is the weighted average of the distances calculated for each variable after scaling each variable to a $[0,1]$ scale. Gower's dissimilarity coefficient (Everitt, Landau, Leese, & Stahl, 2011) between the two mixture observations $\mathbf{x}_i = (\mathbf{z}_i^{\mathrm{T}}, \mathbf{c}_i^{\mathrm{T}})$ and $\mathbf{x}_j = (z_j^{\mathrm{T}}, c_j^{\mathrm{T}})$ can be calculated by the following equation:

$$D_{\mathbf{x}_i\mathbf{x}_j} = \frac{\sum_{r=1}^{p} w_{\mathbf{x}_i\mathbf{x}_j z_r} D_{\mathbf{x}_i\mathbf{x}_j z_r}}{\sum_{r=1}^{p} w_{\mathbf{x}_i\mathbf{x}_j z_r}} + \frac{\sum_{r=1}^{q-p} w_{\mathbf{x}_i\mathbf{x}_j c_r} D_{\mathbf{x}_i\mathbf{x}_j c_r}}{\sum_{r=1}^{q-p} w_{\mathbf{x}_i\mathbf{x}_j c_r}},$$

where $w_{\mathbf{x}_i\mathbf{x}_j z_r}$ and $w_{\mathbf{x}_i\mathbf{x}_j c_r}$ are, respectively, the weights for categorical variable z_r and continuous variable c_r. Note that each variable is equally weighted in this study. $D_{\mathbf{x}_i\mathbf{x}_j z_r}$ is the distance along a categorical variable z_r that can be obtained as follows:

$$D_{\mathbf{x}_i\mathbf{x}_j z_r} = \begin{cases} 0, & z_r^i = z_r^j \\ 1, & \text{otherwise}. \end{cases}$$

$D_{\mathbf{x}_i\mathbf{x}_j c_r}$ is the Manhattan distance (i.e., L1 norm) along a continuous variable c_r that can be computed as follows:

$$D_{\mathbf{x}_i\mathbf{x}_j c_r} = \frac{|c_r^i - c_r^j|}{\max(c_r) - \min(c_r)}. \qquad (2)$$

Although the Manhattan distance is used in the calculation of the original Gower's dissimilarity measure, other distance metrics can be used.

2.2. Monitoring statistics based on the Gower distance

Our proposed control charts use a monitoring statistic based on the Gower distance. First, we introduce some notations to describe the monitoring statistic. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set of training (in-control) mixture observations where $\mathbf{x}_i = (z_1^i, \ldots, z_p^i, c_1^i, \ldots, c_{q-p}^i)^{\mathrm{T}}$ containing p categorical and q–p continuous variables. Let $\mathbf{x}_{n+1} = (z_1^{n+1}, \ldots, z_p^{n+1}, c_1^{n+1}, \ldots, c_{q-p}^{n+1})^{\mathrm{T}}$ be a future observation ($\mathbf{x}_{n+1} \in \Re^q$).

2.2.1. Monitoring statistics based on global Gower distance

Global Gower distance of $\mathbf{x}_{n+1}$ is the average Gower distance between $\mathbf{x}_{n+1}$ and all the training observations in X and can be calculated from the following equation:

$$G(\mathbf{x}_{n+1}) = \frac{\sum_{i=1}^{n} \|\mathbf{x}_{n+1} - \mathbf{x}_i\|}{n} \qquad (3)$$