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

**A R T I C L E  I N F O**

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

**A B S T R A C T**

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.

© 2013 Elsevier Ltd. All rights reserved.

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; Sukhotrat, 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 processes.

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) contains a mixture of 30 continuous and categorical variables. In service industries, for example, in some modern industries the data contain both continuous and categorical variables. Let \( x \) be a mixture observation, characterized by \( p \) categorical variables and \( q-p \) continuous variables. Thus, the vector \( x \) can be rewritten as follows:

\[
\begin{align*}
x &= (z_1, z_p, c_1, \ldots, c_{q-p})^T
\end{align*}
\]

where \( z \) and \( c \) represents the vector of the subset of \( x \) 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 \( x_1 = (z_{11}, z_{1p}, c_{11}, \ldots, c_{1q-p}) \) and \( x_2 = (z_{21}, z_{2p}, c_{21}, \ldots, c_{2q-p}) \) can be calculated by the following equation:

\[
D_{x_1x_2} = \frac{\sum_{j=1}^{p} w_{x_1j} D_{x_1j} + \sum_{j=p+1}^{q} w_{x_2j} D_{x_2j}}{\sum_{j=1}^{p} w_{x_1j} + \sum_{j=p+1}^{q} w_{x_2j}},
\]

where \( w_{x_1j} \) and \( w_{x_2j} \) are, respectively, the weights for categorical variable \( z_j \) and continuous variable \( c_j \). Note that each variable is equally weighted in this study. \( D_{x_1x_2} \) is the distance along a categorical variable \( z_j \) that can be obtained as follows:

\[
D_{x_1x_2} = \begin{cases} 
0, & z_{1j} = z_{2j} \\
1, & \text{otherwise}
\end{cases}
\]

\( D_{x_1x_2} \) is the Manhattan distance (i.e., L1 norm) along a continuous variable \( c_j \) that can be computed as follows:

\[
D_{x_1x_2} = \frac{|c_{1j} - c_{2j}|}{\max(c_i) - \min(c_i)}.
\]

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 \( X = \{x_1, x_2, \ldots, x_n\} \) be the set of training (in-control) mixture observations where \( x_1 = (z_{11}, z_{1p}, c_{11}, \ldots, c_{1q-p})^T \) containing \( p \) categorical and \( q-p \) continuous variables. Let \( x_{n+1} = (z_{n+11}, z_{n+1p}, c_{n+11}, \ldots, c_{n+1q-p})^T \) be a future observation \( x_{n+1} \in \mathbb{R}^q \).

2.2.1 Monitoring statistics based on global Gower distance

Global Gower distance of \( x_{n+1} \) is the average Gower distance between \( x_{n+1} \) and all the training observations in \( X \) and can be calculated from the following equation:

\[
G(x_{n+1}) = \frac{\sum_{i=1}^{n} ||x_{n+1} - x_i||}{n}
\]
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات