Applied Soft Computing 13 (2013) 1520–1527

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Change point determination for a multivariate process using a two-stage hybrid scheme

Yuehjen E. Shao, Chia-Ding Hou *

Department of Statistics and Information Science, Fu Jen Catholic University, New Taipei City, Taiwan

ARTICLE INFO

Article history:
Received 23 December 2011
Received in revised form 31 January 2012
Accepted 3 February 2012
Available online 28 February 2012

Keywords:
Hybrid
Logistic regression
Support vector machine
Multivariate adaptive regression splines
Change point
Multivariate process

ABSTRACT

Effective identification of the change point of a multivariate process is an important research issue since it is associated with the determination of assignable causes which may seriously affect the underlying process. Most existing studies either use the maximum likelihood estimator (MLE) method or the machine learning (ML) method to estimate or identify the change point of a process. Typically, the MLE method may be criticized for its assumption that the process distribution is known, and the ML method may have the deficiency of using a large number of input variables in the modeling procedure. Diverging from existing approaches, this study proposes an integrated hybrid scheme to mitigate the difficulties of the MLE and ML methods. The proposed scheme includes four components: the logistic regression (LR) model, the multivariate adaptive regression splines (MARS) model, the support vector machine (SVM) classifier and the change point identification strategy. It performs three tasks in order to effectively identify the change point in a multivariate process. The initial task is to use the LR and MARS models to reduce and refine the whole set of input or explanatory variables. The remaining variables are then served as input variables to the SVM in the second task. The last task is to integrate use of the SVM outputs with our proposed identification strategy to determine the change point in a multivariate process. Experimental simulation results reveal that the proposed hybrid scheme is able to effectively identify the change point and outperform the typical statistical process control (SPC) chart alone and the single stage SVM methods.

1. Introduction

Process improvement can be successfully achieved by quickly detecting disturbances and accurately identifying their root causes as early as possible. Due to their easy use and effective performances, statistical process control (SPC) charts have been widely used in practice. One of the main characteristics of the SPC charts is that the out-of-control signal would be triggered when disturbances occur in the process. As a consequence, the associated disturbances should be quickly recognized and rectified. The underlying process can then be improved in time.

However, the recognition of the process disturbance may be a time-consuming task. This problem can be mitigated by accurately determining the change point since it typically carries the most related information about the underlying disturbances. Therefore, effective determination of the change point is an important research issue for the process industry.

In recent years, there have been many studies done on change point determination [1–15]. Most of these studies are concerned with the estimation of the maximum likelihood estimator (MLE) or the use of machine learning (ML) method. They proposed the MLE approach to estimate the change point when the process mean or variance has been shifted [1,2]. The MLE approach with the use of EWMA and Cusum control charts was also studied to determine the change point for a normal process [3]. The MLEs of the change point for a Uniform [4] and a Gamma processes [5–7] were also derived. The statistical properties for the MLE were also discussed [5–8]. The MLE of a change point for a multivariate process was also discussed [9]; however, the performance of this MLE is not stable when the number of quality variables is large (say, greater than 5). In addition, the MLE of a change point can be estimated for an attribute process. MLE for the Binomial, Poisson and Geometric processes have been discussed [10–12]. However, those studies have the same assumption in which the underlying process distribution must be known in advance. When the process distribution is not known, which is commonly occurred in the real applications, the MLE cannot be derived. Also, the misuse of the MLE method would usually cause the over-estimation problem [13].

Due to the fact that the process distribution may not be known, some studies focused on the ML method to identify the change point [13,14]. However, the input variables in those studies are considered to be very few. Actually, very few or no research works

* Corresponding author.
E-mail address: stat0002@mail.fju.edu.tw (C.-D. Hou).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2012.02.008

have discussed the problems on how to manage the considerable number of input variables when determining the change point. One of the major problems, when ML methods with many input variables are involved, is that a long training process time would have occurred [16,17]. As a consequence, both of the MLE and ML methods have their own deficiencies. While the MLE method is criticized for its strict assumption that the underlying process distribution must be known, the processing time of ML methods can be lengthy when many input variables are used. Those difficulties are specially pronounced in a multivariate process. A multivariate process can have 2, 3 or even more than 5 quality characteristics. The mathematical derivation of the MLE would be much more difficult than the case of a univariate process. The recognition of underlying process distribution would also be much harder when the numbers of quality variables increased. Additionally, when the number of the quality variables is large (e.g., 10), it may result in the problem of creating a large number of input variables for the ML methods. Therefore, this study is motivated to solve the difficulties where the underlying process distribution is unknown and there is a large number of input variables.

Since a multivariate process would most likely possess the above difficulties, this study considers the change point determination in a multivariate normal process with 10 quality variables. The identification of a step change point in a multivariate process mean vector has been addressed [15]. In their study [15], it is found that fewer explanatory or input variables are needed in the case of a change for the multivariate process mean vector. For example, we may have 10 explanatory variables in the case of mean change when the number of quality characteristic is 10 in a multivariate process. Accordingly, the identification of a step change point in a process mean vector is not the main concern in the present study. In contrast to the typical assumption of process mean shifts, this study assumes another practical and even more serious problem, the process variance shifts. In this study, we have 56 input variables to be considered. It is not necessary to use all the 56 variables as the inputs for a ML method. Consequently, this study integrates the two-stage hybrid models with an effective identification strategy to determine the change point of a multivariate process. To achieve the highest quality solution for a particular manufacturing process with smaller computational time, hybrid evolutionary systems or evolutionary computation algorithms are commonly used for optimizing the manufacturing process [18,19]. Although many studies have been conducted using hybrid modeling for SPC applications [20,21], there are very few studies addressing hybrid modeling in change point determination. In this study, the proposed two-stage hybrid models include the logistic regression (LR) and support vector machine (SVM), referred to as the LR-SVM, and the multivariate adaptive regression splines and support vector machine, referred to as the MARS-SVM.

When the process shifts have occurred, we assume that the multivariate process covariance matrix changes from Σ_0 to Σ_1. Because a large number of input variables for the ML methods is not appropriate, the proposed two techniques, LR and MARS, can be used to finely select fewer but more important explanatory variables. This is the purpose of the first stage of building a hybrid model. The selected variables are then served as inputs for the proposed SVM classifier. The SVM modeling is the second stage of the hybrid model. The main function of the SVM is to classify the multivariate process as in-control (i.e., the SVM output is represented by "0") or out-of-control (i.e., the SVM output is represented by "1"). After performing the SVM classification, a useful identification strategy is integrated to determine the change point.

The SVM is able to lead to great potential and superior performance in practical applications. This is largely due to the structure of risk minimization principles in SVM, which has greater generalization ability and is superior to the empirical risk minimization principle as adopted in neural networks [22–24]. Accordingly, due to the usefulness of the generalization power with a unique and global optimal solution, this study addresses the SVM.

The existing ML methods may be capable of identifying the change point. However, the existing ML methods have not been used in determining the change point of a multivariate process which has a very large number of input variables. Using the proposed mechanisms, we are able to effectively detect the change point of a multivariate process. The superiority of the proposed approaches is also demonstrated with the use of simulated experiments. The structure of this study is organized as follows. The following section states the structure of a multivariate process and the variance shifts. Section 3 addresses the components of the two-stage hybrid models and the identification strategy. Section 4 describes the experimental simulations. The performances for the typical and the proposed approaches are demonstrated and discussed. The final section concludes this study.

2. Structures of the process and the variance shift

The multi-stage hybrid procedure is commonly used in various fields such as medical area [17], credit risk modeling [25], statistical inference [26,27] and in manufacturing process [20,21]. In this study, we aim to develop a hybrid scheme to effectively determine the change point of a multivariate process.

This study assumes that the multivariate normal process is initially in control and the sample observations come from a pair of known mean vector $\underset{\sim}{\mu}$ and covariance matrix Σ_0. After an unknown time $\tau + 1$, this study assumes that the process covariance matrix changes from Σ_0 to Σ_1. This study also assumes that once the covariance matrix Σ_0 changes, it remains at the new level of Σ_1 until the special causes of a disturbance have been identified and removed. Let

$$\underset{\sim}{X}_{i,j} = [X_{i,j,1},\ X_{i,j,2}, \ldots,\ X_{i,j,p}]' \tag{1}$$

be a $p \times 1$ vector which represents the p characteristics on the jth observation in subgroup i with multivariate normal distribution function $N(.,.)$. It follows that

$$\begin{array}{c}
\underset{\sim}{X}_{1,1}, \ldots, \underset{\sim}{X}_{1,n} \overset{\text{iid}}{\sim} N(\underset{\sim}{\mu}, \Sigma_0);\\
\underset{\sim}{X}_{2,1}, \ldots, \underset{\sim}{X}_{2,n} \overset{\text{iid}}{\sim} N(\underset{\sim}{\mu}, \Sigma_0);\\
\cdot\\
\cdot\\
\cdot\\
\underset{\sim}{X}_{\tau,1}, \ldots, \underset{\sim}{X}_{\tau,n} \overset{\text{iid}}{\sim} N(\underset{\sim}{\mu}, \Sigma_0);\\
\underset{\sim}{X}_{\tau+1,1}, \ldots, \underset{\sim}{X}_{\tau+1,n} \overset{\text{iid}}{\sim} N(\underset{\sim}{\mu}, \Sigma_1);\\
\cdot\\
\cdot\\
\cdot\\
\underset{\sim}{X}_{T,1}, \ldots, \underset{\sim}{X}_{T,n} \overset{\text{iid}}{\sim} N(\underset{\sim}{\mu}, \Sigma_1),
\end{array} \tag{2}$$

where the notation "$\overset{\text{iid}}{\sim}$" stands for "independent and identically distributed", n is the sample size, $\tau + 1$ is the change point, T is the signal time that a subgroup covariance matrix exceeds an $|S|$ control