**Identifying the Source of Variance Shifts in Multivariate Statistical Process Control Using Ensemble Classifiers**

(Please note this sample is a reformatted version of a paper presented in ANQ Congress 2014.)

Hung-Ting Lee, Chuen-Sheng Cheng\*

Department of Industrial Engineering and Management, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li, Taoyuan, 320, Taiwan, R.O.C.

\*Corresponding Author: ieccheng@saturn.yzu.edu.tw, Tel.+886-3-4638899, Fax: +886-3-4638907

-----------------------------

The following example is for the case of multiple authors with different affiliations

Iuan-Yuan, Lu1, Chuen-Sheng Cheng2\*, Ching-Chow Yang3 and Pei-Hsun Tsai1

1Department of Business Management, National Sun Yat-Sen University, 70 Lianhai Road, Kaohsiung City 80424, Taiwan, R.O.C.

2Department of Industrial Engineering and Management, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li, Taoyuan, 32003, Taiwan, R.O.C.

3Department of Industrial and System Engineering, Chung Yuan Christian University, 200 Chung-Pei Road, Chung-Li, Taoyuan, 32023, Taiwan, R.O.C.

\*Corresponding Author: ieccheng@saturn.yzu.edu.tw, Tel.+886-3-4638899, Fax: +886-3-4638907

---------------------------------

***Abstract***

Statistical process control charts are important tools for monitoring process stability in manufacturing and service industries. Using multivariate control charts to monitor two or more related quality characteristic has become increasingly popular in today’s environment. Out-of-control signals in multivariate control charts may be caused by one or more variables or a set of variables. In the practice of quality control, in addition to the quick detection of process change, it is also critical to diagnose the change and identify which variable(s) has changed after an out-of-control signal occurred.

In this paper, we propose a decision tree (DT) based ensemble classifier to approach the problem of diagnosing out-of-control signal. The focus of this paper is on the variance shifts of a multivariate process. Simulation studies indicate that the proposed ensemble classification model is a successful method in identifying the source of variance change. A comparative study also reveals that DT using extracted features as input vector has slightly better classification performance than using raw data as input. The proposed method could be a useful diagnostic aid in monitoring shifts in process dispersion.

***Keywords****:* multivariate control chart, variance shift, ensemble classifications, decision tree, features.

**1. Introduction**

Statistical process control chart have been widely used to control various industrial processes. In many industrial applications, the quality of a product or process can be attributed to several correlated variables. In this situation, a multivariate chart is needed to determine whether the process is in control. Many types of multivariate control charts have been proposed to monitor process shifts in the mean or in the process dispersion. The Hotelling  control chart is commonly used to monitor the shifts in the mean of a multivariate process. Detecting shifts in the process variability is also of great importance in process control. In multivariate process control, process variability is characterized by the  covariance matrix , where  is the number of variables. The main diagonal elements of this matrix are the variances of the individual variables and the off-diagonal elements are the covariances. For detecting changes in the covariance matrix, a widely used approach is the generalized variance, , control chart (Alt, 1985). What follows is a brief introduction of this control chart, for further details the reader is referred to Montgomery (2013). Consider a multivariate process with  correlated quality characteristics of interest. It is also assumed that the quality distribution of the  quality characteristics is the  normal distribution with the in-control mean vector  and covariance matrix . The control limits of the control chart for  can be expressed as: ,,, where the coefficients  and  are given by,

 (1)

 (2)

The lower control limit is replaced with zero if the calculated value is negative. Usually  is unknown, it can be derived by a sample covariance matrix , based on a set of historical samples. In this situation, one would replace  by an unbiased estimator , where  is the determinant of the in-control covariance matrix. The chart statistic is , where  is the sample covariance matrix for sample . The process variability is monitored by plotting  on the chart. If a value of  exceeds the or , then the process variability is considered to be out of control indicating that there has been a change in . Surtihadi *et al.* (2004) considered various possible patterns of shifts in the covariance matrix and developed control chart to detect these process change. Costa and Machado (2009) proposed a control chart for detecting shifts in the covariance matrix of multivariate process. Machado *et al*. (2009) proposed a new statistic to control the covariance matrix of bivariate process. When an out-of-control event is identified as a signal in a control chart, a diagnostic procedure is needed for locating the specific variables that are the cause of the problem. Although control charts are efficient in detecting shift in the variance, they do not reveal which variable (or combination of variables) caused the signal. Due to this reason, identifying the responsible variable for the out-of-control process has become one of the most important topics in the field of multivariate process control. For monitoring process mean shift, the most effective method that addresses this problem is a decomposition of the control statistic (Runger *et al*., 1996). However, it is evident from the literature that little work has been directed towards the identification of the source of process variability.

Various soft computing techniques have recently been proposed to identify the responsible variables when the process is considered to be out of control (Low *et al*., 2003; Aparisi *et al*., 2006; Guh, 2007). To identify the source of variance shift, Cheng and Cheng (2008) considered two classifiers based on artificial neural networks (ANN) and support vector machine (SVM) to interpret the signals issued by a generalized variance chart. Cheng *et al*. (2011) proposed a model based on a least squares support vector machine (LS-SVM) to diagnose abnormalities in the bivariate processes of a covariance matrix.

Lately, ensemble classification techniques have received much attention and have demonstrated promising capabilities in improving classification accuracy over single classifiers (Hansen and Salamon, 1990; Kim and Kang, 2010; Yu *et al*., 2010). The ensemble techniques have been used in the area of multivariate process control to identify the source(s) of out-of-control signals. Recent examples reported in the literature include (Alfaro *et al*., 2009; Yu and Xi, 2009a, b; Yu *et al*., 2009, Cheng and Lee, 2014).

In this paper, we propose using the decision tree (DT) ensemble as a diagnostic aid to determine which variable(s) in the process that has changed after the chart triggered a signal. The present work considers the identification of the source(s) of the variance shifts. The proposed system includes a variance shift detector and a classifier. The traditional  chart is employed as a variance shift detector. An ensemble learning approach is used to construct the classifier. The ensemble classifier includes a number of decision tree-based models. The proposed system is depicted in Figure 1. The ensemble classifier, which determines the responsible variable(s) for the variance shift, starts after the  chart sends out an out-of-control signal. This study attempts to extend previous research in several directions. First, ensemble model is created by considering the statistical properties of dataset. Second, a set of features is proposed to further improve the classification accuracy.

We will illustrate the proposed approach using simulated multivariate data. This paper demonstrates that the decision tree ensemble can improve the classification accuracy greatly than using a single DT. The rest of this paper is organized as follows. In Section 2, we briefly describe the fundamental theory of decision tree. This section also provides the basic concept of ensemble based classification model. Section 3 presents the details of building the proposed ensemble classifier. Next, experimental results are given in Section 4. Finally, in Section 5, we summary the main conclusions from this study and provide some directions for future research.



Figure 1. The proposed diagnosis system.

**2. Theoretical Background**

**2.1 Decision Tree**

Decision tree (DT) is a non-parametric supervised learning method used for classification and regression. It is composed of root, internal and leaf nodes. Decision trees have attracted wide popularity due to their efficiency and simplicity in solving a wide range of problems in various fields.

There are several algorithms available to construct a decision tree. The research described in this paper considers the CART (Classification and Regression Trees) algorithm developed by Breiman *et al*. (1984). The goal of CART is to explain the variation of a single response (dependent) variable using a set of explanatory (independent) variables, via a binary partitioning procedure. Both the response and the explanatory variables can be either categorical or numerical. Depending on the data type of response variable, CART can create either classification tree or regression tree. A classification tree is produced when the response variable is categorical while a regression tree is obtained for a numerical response variable. This paper focuses on using CART to build a classification model.

…………

…………

**2.2 Ensemble Based Classification Model**

Ensemble learning is the process of creating multiple models and combining the results of individual models in order to solve a particular problem. This section focuses on classification related applications of ensemble learning. An ensemble classification model is defined as a set of basic classifiers whose individual outcomes are combined (or fused) in some manner to form a final decision. It has been shown that ensemble classification models outperform single classification models, whose performance is limited by the imperfection of feature extraction, learning algorithms, and the inadequacy of training data (Hansen and Salamon, 1990).

In general, an effective ensemble-based model consisting of diverse models with much disagreement is more likely to obtain a good generalization performance. In other words, the ensemble members in the ensemble model must possess different classification properties. Therefore, how to generate the diverse models is a crucial factor for constructing an effective ensemble classification system. There are various methods suggested for the creation of an ensemble of classifiers (Dietterich, 1997). Strategies that can be used to create an ensemble of classifiers include: (1) using different subset of training data with a single learning algorithm; (2) using different training parameters with a single learning algorithm; (3) using various types of learning algorithms. The above methods are usually referred to as diversity strategies.

…………

…………

**3. Proposed Decision Tree-Based Ensemble Classification Model**

The section describes the proposed ensemble classification model using different subsets of training data within a single training method. First, we describe the construction of a single classifier, including the training dataset, input vector and model training. Next, we present the development of ensemble classifiers.

**3.1 Building Single DT-Based Classification Model**

**3.1.1 Generation of Training Data Set**

Preparing an appropriate training dataset is a critical step in constructing an effective classifier. Due to extensive demands for training data, Monte Carlo simulation is the often-recommended method to generate training data. A-dimensional multivariate normal process was simulated by generating pseudo-random variates from a multivariate normal distribution whose mean was  and whose covariance matrix was . By manipulating the value of covariance matrix , an in-control or an out-of-control process can then be simulated. This paper considers the case of increasing in process dispersion, which corresponds to deterioration of process or product quality. We consider the patterns of shifts in the covariance matrix described in Surtihadi *et al*. (2004). Using the notation of Surtihadi *et al*. (2004), a description of the matrices follows.

Let  denote the known in-control covariance matrix. For simplicity, it is assumed that the covariance matrix  is known. In practice it will be necessary to collect sufficient data over a substantial amount of time when the process is in control to estimate .

 (3)

…………

…………

The above change patterns were considered in the work of Cheng and Cheng (2008). The present work considers the identification of the source(s) of the variance shifts after the control chart issues an out-of-control signal. Due to this reason, each selected sample data must exceed the UCL of the  chart. This study considered the case that process dispersion increases. Shifts were investigated for inflation factor  from 1.5 to 5.0 in steps of 0.25. In this study, the simulation was implemented using MATLAB. For convenience, all variables were scaled such that each process variable has mean zero and variance one. With this approach, the covariance matrix is in correlation form; that is, the main diagonal elements are all one and the off-diagonal elements are the pairwise correlation () between the process variables. Here we assume that the process mean vector remains constant as the process dispersion shifts. The current study considers the case of equi-correlation (pairs of variables are equi-correlated). The values of  were from 0.1 to 0.9 in steps of 0.2. It will also be assumed that the value of  is stable with respect to shifts in the process dispersion. Table 1 summarizes the shift patterns considered for each .

Table 1. Shift patterns.

|  |  |  |
| --- | --- | --- |
| Variable | Shift type | Class label |
|  | The 1st variable shifts | 1 |
|  | The 2nd variable shifts | 2 |
|  | All variables shift | 3 |
|  | The 1st variable shifts | 1 |
|  | The 2nd variable shifts | 2 |
|  | The 3rd variable shifts | 3 |
|  | All variables shift | 4 |

**3.1.2 Selection of Input Vector**

Determining the input feature vector to the classifier is an important task in building a classification model. In the present research, there are many different elements that could be used to construct an appropriate feature vector. The first feature vector (henceforth vector-I) consists of sample data observed in the  sample with . An alternative is to use computed sample statistics as the input components (hereafter called vector-II). The feature vector can be written as , where  is the standard deviation on the  variable in the  sample,  is the sample mean of the  variable, and  is the Mahalanobis distance obtained from the  sample data, with subscript  being the index of class . In the present research, Mahalanobis distance measures the distance between an observation and a given class (shift type). Obviously, this feature vector provides a compact and efficient representation of the input vector. A similar feature vector was proposed by Cheng and Lee (2014). Later in this paper, we will compare the performance of these two alternatives.

**3.1.3 Building Tree Models**

The tree models were built using MATLAB classregtree procedure. The class label was used as the response variable and input features described in Section 3.1.1 as explanatory variables. The “splitcriterion” was set to be Gini’s diversity index.

**3.2 Proposed DT Ensemble Model**

This section presents the specific design details of the ensemble model. Building ensemble classifier involves two parts: generation of classifiers and design of ensemble strategy. In this study, using different training datasets is selected as the diversity strategy to create ensemble members. This approach is referred to as data diversity. Under this approach, multiple classifiers are trained using different training subsets. This is in contrast to traditional approaches which try to train a classifier using a single training set (Aparisi *et al*., 2006; Cheng and Cheng, 2008). In previous studies, samples corresponding to various magnitudes of process changes must be considered simultaneously in a single training set. In other words, the training set involves a multiple-shift distribution. This approach complicates the training process.

…………

…………

After training, each classifier has generated its own result. The next step is to combine these classifiers into an aggregated classifier using an ensemble strategy. In the present study, a majority voting based ensemble strategy is adopted due to its ease of implementation. The decision of the ensemble classifier is obtained from the decision results of DTs via a majority voting based ensemble strategy. Figure 2 illustrates the implementation of the DT ensemble.



Figure 2. The structure of DT ensemble.

**4. Results and Discussion**

This section presents the comparisons of classification accuracy (%) of various classifiers, including single DT, bagging, AdaBoost, and the proposed DT ensemble. To demonstrate the power of the ensemble classifier and the benefits of proposed features, we used single DT method and vector-I as a baseline. The test data were generated in same way as training data. Shifts were investigated for  from 1.5 to 5.0 in steps of 0.25. The sample size considered in this study is . Bagging and AdaBoost ensemble models were created using “fit ensemble” function of MATLAB. For bagging ensemble model, the input argument “NLearn” was set to be 15. This setting was made so as to have a fair comparison between bagging and the proposed DT ensemble. For AdaBoost ensemble classifier, the input argument “Method” was set to be “AdaBoostM2”. Various values of “NLearn” ranging from 1 to 100 were tried. Only the best result was reported.

In each simulation run, 1000 test samples (five times the size of training samples) were created for each . For each , classification rate was approximated based on 15000 test samples for each variance shift pattern. Base on this simulation setting, each classifier has been performed 10 independent runs of simulation and the average performance is reported. Tables 2-3 summarize the overall performances of classification models in terms of the classification accuracy. The results are reported for 0.1, 0.3, 0.5, 0.7 and 0.9. It can be observed from Tables 2-3 that models built with vector-II perform better than those with vector-I. It can also be observed that DT ensemble uniformly performs better than single DT. Generally the proposed method has the best performance, followed by bagging and AdaBoost. Figure 3 presents the classification accuracy in the graphical form, as a function of . The results are reported for **** and 0.3, 0.5, and 0.7. It can be seen that DT ensemble consistently performs better than other classification models irrespective of the size of .

…………

…………

Table 2. Comparison of the classification accuracy ().

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | vector-I | | | |  | vector-II | | | |
| Single DT | AdaBoost | Bagging | DT ensemble |  | Single DT | AdaBoost | Bagging | DT ensemble |
| 0.1 | 76.66 | 79.96 | 80.89 | 84.44 |  | 80.30 | 80.82 | 85.56 | 90.83 |
| 0.3 | 77.88 | 81.57 | 82.15 | 85.71 |  | 80.95 | 81.56 | 85.77 | 90.88 |
| 0.5 | 79.80 | 83.42 | 84.34 | 87.19 |  | 83.29 | 84.38 | 87.60 | 91.19 |
| 0.7 | 84.37 | 85.88 | 87.91 | 89.13 |  | 86.54 | 87.82 | 90.45 | 92.90 |
| 0.9 | 90.61 | 91.60 | 93.36 | 94.12 |  | 93.32 | 94.73 | 95.73 | 96.42 |

Table 3. Comparison of the classification accuracy ().

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | vector-I | | | |  | vector-II | | | |
| Single DT | AdaBoost | Bagging | DT ensemble |  | Single DT | AdaBoost | Bagging | DT ensemble |
| 0.1 | 82.19 | 83.99 | 86.17 | 87.29 |  | 83.74 | 84.16 | 88.11 | 92.06 |
| 0.3 | 82.31 | 84.56 | 86.63 | 87.69 |  | 84.55 | 84.63 | 88.43 | 92.38 |
| 0.5 | 84.18 | 86.03 | 87.60 | 88.52 |  | 85.72 | 86.03 | 89.20 | 92.86 |
| 0.7 | 85.55 | 88.08 | 90.45 | 90.59 |  | 88.46 | 90.16 | 91.54 | 94.33 |
| 0.9 | 93.57 | 94.50 | 95.73 | 96.52 |  | 94.22 | 95.08 | 96.16 | 97.47 |

**5. Conclusions**

Identifying the variable(s) responsible for the variance shifts is an important topic in multivariate process control. In this paper, a DT-based ensemble classification model has been proposed for the interpretation of variance shift signal in multivariate processes. The focus is on using the data diversity strategy to create an effective ensemble classifier.

The results of extensive simulation studies indicate that the proposed DT-based ensemble classifier provides better performance than single classifier. The proposed DT ensemble model can also consistently outperform the other comparable models including bagging and boosting. Comparative studies also reveal that substantial improvement in classification accuracy can be obtained when DT was built using relevant features. It is believed that the proposed DT-based ensemble classification system can be used as a useful diagnostic tool in the interpretation of out-of-control signal in multivariate process control. The main contributions of this paper include a set of features useful for classification and a data diversity strategy based on shift size.

…………

…………

**References**

Alfaro, E., Alfaro, J. L., Gámez, M., and García, N. (2009), “A boosting approach for understanding out-of-control signals in multivariate control charts,” *International Journal of Production Research*, 47, 6821-6831.

Alt, F. B. (1985), “Multivariate quality control,” In *Encyclopedia of Statistical Science*, edited by Johnson, N. L., and Kotz, S., New York: John Wiley and Sons.

[Aparisi, F.](http://www.scopus.com/search/submit/author.url?author=Aparisi%2c+F.&origin=resultslist&authorId=6701528553&src=s), Avendaño G., and Sanz, J. (2006), “Techniques to interpret  control chart signals,” *IIE Transactions*, 38, 647-657.

Breiman, L. (1996), “Bagging predictors,” *Machine Learning*, 24(2), 123-140.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regressing Tree*, California: Wadsworth.

Cheng, C. S., and Cheng, H. P. (2008), “Identifying the source of variance shifts in the multivariate process using neural networks and support vector machines,” *Expert System with Applications*, 35, 198-206.

…………

…………

Kotsiantis, S. B., and Pintelas, P. E. (2004), “Combining bagging and boosting,” *International Journal of Computational Intelligence*, 1, 324-333.

…………

…………

Montgomery, D. C. (2013), *Introduction to Statistical Quality Control*, New York: Wiley & Sons.

Runger, G. C., Alt, F. B., and Montgomery, D. C (1996), “Contributors to a multivariate statistical process control signal,” *Communications in Statistics-Theory and Methods*, 25, 2203-2213.

Surtihadi, J., Raghavachari, M., and Runger, G. (2004), “Multivariate control charts for process dispersion,” *International Journal of Production Research*, 42, 2993-3009.

[Yu, J. B.](http://www.scopus.com/search/submit/author.url?author=Yu%2c+J.-b.&origin=resultslist&authorId=22942774900&src=s), and Xi, L. F. [(2009), “](http://www.scopus.com/search/submit/author.url?author=Xi%2c+L.-f.&origin=resultslist&authorId=7006667072&src=s)A neural network ensemble-based model for on-line monitoring and diagnosis of out-of-control signals in multivariate manufacturing processes,” *Expert Systems with Applications*, 36, 909-921.

Yu, J. B., and Xi, L. F. (2009a), “A hybrid learning-based model for on-line monitoring and diagnosis of out-of-control signals in multivariate manufacturing processes,” *International Journal of Production Research*, 47, 4077-4108.

Yu, J. B., Xi, L.F., and Zhou, X. J. (2009b), “Identifying source(s) of out-of-control signals in multivariate manufacturing processes using selective neural network ensemble,” *Engineering Applications of Artificial Intelligence*, 22, 141-152.

…………

…………

**Authors’ Biographical Notes**

Hung-Ting Lee *received the M.S. degree in Industrial Engineering and Management from National Formosa University in 2008. She is now a Ph.D. candidate in Industrial Engineering and Management, Yuan Ze University. Her primary research interests include multivariate statistical process control, profile monitoring and intelligent process control.*

Chuen-Sheng Cheng *is Professor and Chair in the Department of Industrial Engineering and Management at Yuan Ze University, Taiwan. He received his Ph.D. degree in industrial engineering from Arizona State University (U.S.A.) in 1989, an M.S. degree in industrial engineering from Texas Tech. University (U.S.A.) in 1985, and a B.S. degree in industrial engineering from National Tsinghua University (Taiwan) in 1980. His research interests include statistical quality control, intelligent process monitoring, and applications of artificial neural networks. He has authored several books in the areas of quality management and statistical quality control. Professor Cheng has extensive experience in training and consultancy in the areas of statistical quality control, quality management and Six Sigma. In 2000, Professor Cheng obtained the Individual Award of the 11th National Quality Awards of the Republic of China*. *Also, as the key facilitator, he led Yuan Ze University in winning the “Organization Award” of the 14th National Quality Award in 2003. Professor Cheng is the Editor-in-Chief of Journal of Quality, he is also a Fellow and a council member of the Chinese Society for Quality (CSQ).*