Asian Business Research; Vol. 2, No. 2, 2017
ISSN 2424-8479 E-ISSN 2424-8983
Published by July Press

Diagnosis of Out-of-control Signals in Multivariate Statistical Process Control Based on Bagging and Decision Tree

Jing Jiang[1] & Hua-Ming Song[2]

[1] Quality and Technology Supervision Bureau of Jiangsu Province, Nanjing, China

[2] Department of Management Science and Engineering, Nanjing University of Science & Technology, Nanjing, China

Correspondence: Hua-ming Song, Department of Management Science and Engineering, Nanjing University of Science & Technology, Nanjing 210094, China. Tel: 86-135-1250-0816.

Received: April 13, 2017 Accepted: May 2, 2017 Online Published: May 16, 2017

doi:10.20849/abr.v2i2.147 URL: https://doi.org/10.20849/abr.v2i2.147

Abstract

In this paper, we propose an ensemble method based on bagging and decision tree to resolve the problem of diagnosing out-of-control signals in multivariate statistical process control. To classify the out-of-control signals, we obtain a series of classifiers through ensemble learning on decision tree. Then we will integrate the classification results of multiple classifiers to determine the final classification. The experimental results show that our method could improve the accuracy of classification and is superior to other methods in terms of diagnosing out-of-control signals in multivariate statistical process control.

Keywords: out-of-control signals, multivariate statistical process control, bagging, decision tree

1. Introduction

In the practice of multivariate statistical process control, there are two basic issues in front of managers. One is how to detect and signal if the process changes into out-of-control state; another is how to interpret and justify the reasons when the abnormal state occurred, that is, which variate or variates combination should be response the abnormal state. The first issue is almost addressed through a powerful tool, Hotelling's T^2 control chart. In fact, Hotelling's T^2 control chart is a natural extension from a single variate statistical process control to multivariate statistical process control. However, the second issue, Diagnosis of out-of-control signals, is more complex to be resolved, this results to many explorations both in industry and academic field.

Diagnosis of out-of-control signals in multivariate statistical process control (MSPC) has received tremendous attention in the field of multivariate quality control, and many scholars have conducted abundant research in this area. With the rapid development of computer technology, many artificial intelligence algorithms in machine learning, such as neural networks (NN), support vector machines (SVM) and decision trees, etc., have found broad application to multivariate process control. Cook et al. (2001) and Low et al. (2003) apply back propagation (BP) to construct monitoring scheme for detecting variance shifts in multivariate process. Wang and Chen (2002) propose a neural-fuzzy model based on BP to detect mean shifts and classify shift magnitudes. Hwarng (2004) presents a neural network model based on BP algorithm to detect process mean shift and provide information about the shift magnitude. Chen and Wang (2004) develop an artificial neural network-based model to not only identify the characteristic or group of characteristics that shift but also classify the magnitude of shifts when the multivariate χ^2 chart signals that mean shifts have occurred. Niaki and Abbasi (2005) also propose an artificial neural network-based model to diagnose faults in out-of-control conditions with the use of Hotelling's T^2 control chart, and they show that the multilayer perceptron (MLP) network with error back propagation performs better than multivariate Shewhart chart methods (MSCH). Zhou et al. (2014) design a comprehensive fault detection and identification procedure based on principal component analysis (PCA), clustering of K-means and machine learning of BP in multiple processes and its effectiveness is finally verified with experimental data.

Cheng and Cheng (2008) apply NN and SVM to build shift classifiers for identifying the source of variance shifts in the multivariate process and simulation studies show that the NN-based classifier and SVM-based classifier have similar performance. Salehi et al. (2011, 2012) present a modular model based on SVM and NN for on-line analysis of out-of-control signals (mean shift/ mean and variance shifts) in multivariate manufacturing processes. Shao et al. (2012) propose a hybrid scheme which is composed of independent component analysis (ICA) and support vector machine (SVM) to determine the fault quality variables in a multivariate process. All of their experimental results indicate the effectiveness of their hybrid methods.

Guh and Shiue (2008) use decision tree learning-based model to detecting mean shifts in a bivariate process, and experimental results show that the learning speed of their model is much faster than that of a neural network-based model. Later, motivated by their research, He et al. (2013) propose improved decision tree based model for bivariate process monitoring and fault identification,

In this paper, we propose an ensemble learning method based on bagging and decision tree to diagnose the out-of-control signals in MSPC. The performance of our method can be demonstrated by experimental results and comparisons with competing methods, such as MLP and MSCH (see Niaki and Abbasi (2005)), CART (see Alfaro et al. (2009b)) and SAMME (Stagewise Additive Modeling Using a Multi-class Exponential Loss Function) which is proposed by Alfaro et al. (2009a) to diagnose out-of-control signals.

2. Classifier Ensemble Algorithm Based on Bagging and Decision Tree

With the wide application of machine learning in production and research, ensemble learning, which is one of the hottest directions of research in machine learning, has been gradually used to diagnose the out-of-control signals in. And the diagnostic problem is a classification problem in nature. For classification problem, ensemble methods are learning algorithms that construct a series of single classifiers whose individual results are combined in some way to classify new examples. Dietterich (1997) showed experimental evidence has proved that ensembles of classifiers are often much more accurate than the single classifier, and then explained why ensemble methods work and gave the necessary and sufficient condition in his research (see Dietterich, 2002). There are lots of ensemble learning algorithms, such as bagging, boosting, arcing, random forest etc., while there exist a variety of base classifier: Bayes, NN, SVM, decision trees and so on. In this paper, an ensemble method based on bagging and decision tree is proposed.

Bagging, the acronym of bootstrap aggregating, was firstly proposed by Breiman (1996). And then bagging has found an increasingly wide utilization and showed excellent performance in various fields, as an effective method of turning the weak learning algorithm into a strong one. The key factor to make sure bagging works well is the instability of learning algorithms. The *instability* here means learning algorithms are sensitive to the training set, i.e. small changes in training set can lead to great changes in the predicted results. And decision tree belongs to the unstable algorithms (see Beriman, 1996). Decision tree includes two varieties, classification tree and regression tree. Diagnosing the out-of-control signals in multivariate statistical process control is a classification problem, so classification tree is chosen as the weak classifier in our method. The construction of an optimal tree consists of two steps: the growth of tree in which we use Gini index to measure the impurity of one node and the pruning of tree in which we select cost-complexity function as the pruning criterion. And the ensemble method based on bagging and decision tree is introduced particularly below.

First of all, we get a series of new training sets from the initial training set $S=\left\{\left(\boldsymbol{x}_n, y_n\right), n=1,2,\cdots,N\right\}$, where y_n is the label. Then we train the decision tree algorithm based on each new training set to generate a series of tree classifiers. At last, we use the tree classifiers to classify the testing set and the final classification result is decided by typically simple majority vote. The procedure is as follows:

Input: a training set of n labeled samples $S=\left\{\left(\boldsymbol{x}_n, y_n\right), n=1,2,\cdots,N\right\}$, label $y_n \in Y=\left\{1,2,\cdots,k\right\}$, a testing set $\boldsymbol{X}$, times of training T.

```
Process:
(1) for t=1, 2, ···, $T$ do
```

(2) Acquire new training set S_t from the initial training set S by bootstrapping;

(3) Get a weak classifier C_t after training the decision tree algorithm with the training set S_t;

(4) Use classifier C_t to classify the test set $\boldsymbol{X}$ and get the classification results R_t ;

(5) end for

(6) Determine the final classification result by voting: $H(\mathbf{X}) = argmax_{y \in Y} \sum_{t=1}^{T} \mathrm{I}(R_t(\boldsymbol{x}) = y)$.

Output: the classification result of testing set ***X.***

3. Experimental Results and Comparisons

Simulated experiments would be carried out with the application of software MatlabR2013b. In the practice of multivariate quality control, HotellingT^2 control chart is applied to detect the out-of-control signals and then we use the ensemble method of bagging-trees to diagnose the out-of-control signals (i.e. identify which variable or variables have changed). So firstly, we generate the training sets and testing sets through HotellingT^2 control chart, and then we do the data preprocessing--standardization. Finally, we take advantage of ensemble algorithm of bagging-trees which is trained based on training data to classify the testing data.

In order to prove our ensemble method performs well in diagnosing out-of-control signals in multivariate statistical process control, three benchmark examples that include two, three and four variables respectively from Alfaro et al. (2009a) are adopted. Besides, for facilitating comparisons between our method and previous ones proposed by Niaki and Abassi (2005) and Alfaro et al (2009a, 2009b), not only the quantity, mean and covariance matrix but also the shift values of mean vector adopted when generating data are as same as theirs. However, randomness of data generation leads to the difference.

Generation and standardization of data sets. Monte Carlo simulation is applied to generate the data sets for training and testing, and HotellingT^2 control chart is used to monitor the process to generate the out-of-control data sets. In HotellingT^2 control chart, the shift values in mean vector are greater than or equal to 1.50 times the standard deviation in general. And the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ could usually be evaluated from a large quantity of observations (i.e. $\boldsymbol{\mu}$ and Σ are often known) so that T^2 statistics is replaced by χ^2 . When the distance between the in-control data and out-of-control data which is measured by Mahalanobis distance is greater than the control limit (calculated from $\chi^2_{1\text{-}\alpha,\ p}$, where $\alpha = 0.05$ is the significance level and p is the number of variables), the control chart will give an out-of-control alarm. Furthermore, we standardize the data using zero-mean normalization in order to eliminate the influence exercised by different dimension and variables' degree of variation.

Data experiment. The simulation experiment of three cases that include two, three and four variables is described in detail as following.

The example of two variables: The first example is derived from a lumber manufacturing process in which two quality variables (stiffness (x_1) and bending strength (x_2)) that have an effect on grade of lumber need to be detected. The mean vector and covariance matrix obtained from a large number of historical data are given as follow:

$$\boldsymbol{\mu} = \begin{bmatrix} 265 \\ 470 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 100 & 66 \\ 66 & 121 \end{bmatrix}$$

In this case ($p = 2$), there are total three ($2^p - 1 = 3$) possible classes to be identified. The shift value of mean vector is set to 1.5σ when we generate the training set. Therefore, the process mean vectors of above mentioned three out-of-control classes are $(265+1.5\times\sqrt{100},470)^{\mathrm{T}}$, $(265,470+1.5\times\sqrt{121})^{\mathrm{T}}$ and $(265+1.5\times\sqrt{100},470+1.5\times\sqrt{121})^{\mathrm{T}}$. When we generate the testing sets, the shift values of mean vector are set to 2.0σ, 2.5σ and 3.0σ respectively. 500 shift samples are generated for each out-of-control class so that there are 1500 shift samples in both the training and testing sets. The classification results can be seen in Table 1. The vector $(1,0)^{\mathrm{T}}$ indicates x_1 only is out of control, while vector $(0,1)^{\mathrm{T}}$ indicates only x_2 is out of control, and vector $(1,1)^{\mathrm{T}}$ indicates that both variable x_1 and variable x_2 are out of control.

Table 1. The classification results of two-variable example

Shift		Predicted class		
	Class	$(1,0)^{\mathrm{T}}$	$(0,1)^{\mathrm{T}}$	$(1,1)^{\mathrm{T}}$
2.0σ	$(1,0)^{\mathrm{T}}$	447	0	53
	$(0,1)^{\mathrm{T}}$	0	453	47
	$(1,1)^{\mathrm{T}}$	53	40	407
2.5σ	$(1,0)^{\mathrm{T}}$	471	0	29
	$(0,1)^{\mathrm{T}}$	0	470	30
	$(1,1)^{\mathrm{T}}$	62	45	393
3.0σ	$(1,0)^{\mathrm{T}}$	478	0	22
	$(0,1)^{\mathrm{T}}$	0	485	15
	$(1,1)^{\mathrm{T}}$	43	32	425

The example of three variables: This example is connected with detergent-making company, and there are three variables to be monitored: color (x_1), free oil percentage (x_2), and acidity percentage (x_3). When the process is in control, the estimated mean vector and covariance matrix are:

$$\mu=\begin{bmatrix}67.5\\12.0\\97.5\end{bmatrix},\quad \Sigma=\begin{bmatrix}0.68 & 0.36 & -0.07\\0.36 & 1.00 & -0.12\\-0.07 & -0.12 & 0.03\end{bmatrix}$$

In this case ($p=3$), there are total seven ($2^p-1=7$) possible classes. When we generate the training set, the shift value of mean vector is set to 3.0σ. We consider the mean vector shifts 2.0σ, 3.0σ and 4.0σ respectively when generating testing sets. For every out-of-control class, one hundred shift samples are generated. Therefore, there are total 700 samples in each data set. The classification results for various shifts values are summarized in Table 2.

The example of four variables: The example of four variables is relevant to ballistic missile testing, and the mean vector and covariance matrix are given as follow:

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 102.74 & 88.34 & 67.03 & 54.06 \\ 88.34 & 142.74 & 86.55 & 80.02 \\ 67.03 & 86.55 & 64.57 & 69.42 \\ 54.06 & 80.02 & 69.42 & 99.06 \end{bmatrix}$$

In this case, there will be total $2^p - 1 = 15$ ($p = 4$) possible classes to be identified. We consider the process mean vector shifts 3.0σ when generating the training set. The shift values of mean vector are set to 2.0σ , 2.5σ and 3.0σ respectively when generating testing sets. Similar to the two-variable example, we generate 500 shift samples for every out-of-control class. Then there will be 7500 shift samples in each training set or testing set. The classification results for various shifts values are summarized in Table 2.

Comparison of results with previous studies. We compare our ensemble method of bagging-trees with SAMME (Alfaro et al., 2009a), CART (Alfaro et al. 2009b), MLP and MSCH (Niaki and Abassi, 2005), and the misclassification rates of the methods can be seen in Table 2.

Table 2. The misclassification rates of methods [%]

Methods	Two variables			Three variables			Four variables			Total Average
	2.0σ	2.5σ	3.0σ	2.0σ	3.0σ	4.0σ	2.0σ	2.5σ	3.0σ	
Bagging-trees	12.87	11.07	7.47	30.71*	10.86	3.14*	8.28*	2.91*	1.13*	9.83*
SAMME	7.20	3.33*	2.87*	46.00	2.86*	14.43	11.85	9.39	5.73	11.50
CART	9.07	6.53	5.47	49.86	12.71	18.71	15.28	14.08	11.19	15.88
MLP	13.93	10.73	8.73	50.00	62.14	45.29	33.03	24.87	18.69	29.71
MSCH	4.60*	10.93	10.93	n.a.[d]	n.a.[d]	n.a.[d]	71.15	66.68	68.53	33.80[e]

The values with label '' are the optimal results of all methods.*

The values with label 'd' indicate no answer given by Alfaro et al. (2009a).

The value with label 'e' is average value of only six cases (not including the case of three variables).

According to Table 2, in the case of two variables, the classification results of bagging-trees method are worse than that of SAMME and CART. However, bagging-trees significantly outperforms the other methods in the cases of three and four variables, in addition to the case of three variables with shift value 3.0σ .The misclassification rates of bagging-trees in cases of three variables with shift value 4.0σ , four variables with shift value 2.5σ and 3.0σ are all below 5%, which are quite better than those of other methods. It is apparent that our method works especially well in the cases of more than two variables. Furthermore, the total average misclassification rate of bagging-trees is 9.83% which is lower than that of all other methods. The comparison shows that our bagging-trees method proposed in this paper as an effective tool for diagnosing out-of-control signals in multivariate statistical process is superior to previous methods.

4. Conclusions and Future Work

In this paper, the classifier ensemble based on bagging and decision tree is proposed to diagnose out-of-control signals in multivariate statistical process control. When classifying the out-of-control data, we can obtain a series of classifiers by training decision tree with use of the ensemble learning method--bagging algorithm, and then we'll integrate the results of multiple classifiers to determine the final categories. Simulation experiments show our method is useful for interpreting out-of-control signals. The comparisons with previous research results prove that our method could get substantial increase in accuracy and works better than other methods, such as SAMME, CART, MLP, MSCH etc. on the whole.

However, some aspects have not been covered in this paper, such as determining shift values in mean vector, selecting times of training and using bagging-trees to diagnose other types (downward shift, trend and cycle etc.) of out-of-control data. It is worthwhile studying these issues in the future research.

References

Alfaro, E., Alfaro, J. L., Gámez, M., & García, N. (2009a). A boosting approach for under-standing out-of-control signals in multivariate control charts. *Int. J. Prod. Res., 47*, 6821-6834. https://doi.org/10.1080/00207540802474003

Alfaro, E., Alfaro, J. L., Gámez, M., & García, N. (2009b). Árboles de clasificación para el análisis de gráficos de control multivariantes. *Rev. Mate. Teor. Aplic., 16,* 30-42.

Breiman, L. (1996). Bagging Predictors. *Mach. Learn., 24*, 123-140. https://doi.org/10.1007/BF00058655

Chen, L. H., & Wang, T. Y. (2004). Artificial neural networks to classify mean shifts from multivariate T^2 chart signals. *Comput. Ind. Eng. 47*, 195-205. https://doi.org/10.1016/j.cie.2004.07.002

Cheng, C. S., & Cheng, H. P. (2008). Identifying the source of variance shifts in the multivariate process using neural networks and support vector machines. *Expert Syst. Appl., 35*, 198-206. https://doi.org/10.1016/j.eswa.2007.06.002

Cook, D. F., Zobel, C. W., & Nottingham, Q. J. (2001). Utilization of neural networks for the recognition of variance shifts in correlated manufacturing process parameters. *Int. J. Prod. Res.*, *39,* 3881-3887. https://doi.org/10.1080/00207540110071750

Dietterich, T. G. (1997). Machine-learning research. *Ai Mag., 18*, 97-136.

Dietterich, T. G. (2002). Ensemble learning. In M.A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks* (2nd ed., pp. 405-408). Mit Pr., Cambridge.

Guh, R. S., & Shiue, Y. R. (2008). An effective application of decision tree learning for on-line detection of mean shifts in multivariate control charts. *Comput. Ind. Eng., 55*, 475-493. https://doi.org/10.1016/j.cie.2008.01.013

He, S. G., He, Z., & Wang G. A. (2013). Online monitoring and fault identification of mean shifts in bivariate processes using decision tree learning techniques. *J. Intell. Manuf., 24*, 25-34. https://doi.org/10.1007/s10845-011-0533-5

Hwarng, H. B. (2004). Detecting process mean shift in the presence of autocorrelation: a neural-network based monitoring scheme. *Int. J. Prod. Res., 42*, 573-595. https://doi.org/10.1080/0020754032000123614

Low, C., Hsu, C. M., & Yu, F. J. (2003). Analysis of variations in a multi-variate process using neural networks. *Int. J. Adv. Manuf. Tech., 22,* 911-921. https://doi.org/10.1007/s00170-003-1631-0

Niaki, S. T. A., & Abbasi, B. (2005). Fault diagnosis in multivariate control charts using artificial neural networks. *Qual. Reliab. Eng. Int., 21*, 825-840. https://doi.org/10.1002/qre.689

Salehi, M., Bahreininejad, A., & Nakhai, I. (2011). On-line analysis of out-of-control signals in multivariate manufacturing processes using a hybrid learning-based model. *Neurocomputing, 74*, 2083-2095. https://doi.org/10.1016/j.neucom.2010.12.020

Salehi, M., Kazemzadeh, R. B., & Salmasnia, A. (2012). On line detection of mean and variance shift using neural networks and support vector machine in multivariate processes. *Appl. Soft Comput., 12*, 2973-2984. https://doi.org/10.1016/j.asoc.2012.04.024

Shao, Y. E., Lu, C. J., & Wang, Y.C. (2012). A hybrid ICA-SVM approach for determining the quality variables at fault in a multivariate process. *Math. Probl. Eng.*, 1-12. https://doi.org/10.1155/2012/284910

Wang, T. Y., & Chen, L.H. (2002). Mean shifts detection and classification in multivariate process: a neural-fuzzy approach. *J. Intell. Manuf., 13,* 211-221. https://doi.org/10.1023/A:1015738906895

Zhou, J., Guo, A., Celler, B., & Su, S. (2014). Fault detection and identification spanning multiple processes by integrating PCA with neural network. *Appl. Soft Comput., 14*, 4-11. https://doi.org/10.1016/j.asoc.2013.09.024

Copyrights