SAX Breakpoints for Random Forest Based Real-Time Contrast Control Chart

In-Seok Lee, Jun-Geol Baek*

Department of Industrial Management Engineering

Korea University

Seoul, Republic of Korea

E-mail: dark7710@korea.ac.kr, jungeol@korea.ac.kr (*corresponding author)

Abstract

In the manufacturing process, process monitoring is very important. Real-time contrast (RTC) control chart outperforms existing monitoring methods. However, the performance of RTC control chart depends on the classifier. The existing RTC charts use random forest (RF), support vector machine (SVM), or kernel linear discriminant analysis (KLDA) as a classifier. RF classifier can find cause of faults but the performance is lower than others. Therefore, we suggest the data representation method to improve the RF based RTC control chart. Symbolic aggregate approximation (SAX) is famous method to improve the performance of classification and clustering. We convert the input data by using SAX. We change the parameters of SAX such as alphabet size and breakpoints to improve the performance. Experiment shows that represented data is efficient method to improve the performance of RTC control chart.

Introduction

In the process monitoring, RTC control chart outperforms existing method. The RTC control chart learns a classifier when new observations are generated and calculates the statistics by the learned results. The traditional RTC control chart uses random forest as a classifier. This classifier provides the variable importance that is the degree of the effect for a variable in the process. The RTC control chart can analyze the cause of faults by using the variable importance (Deng et al, 2012).

However, the random forest classifier distinguishes observations by using the decision boundary of each variable and yields the class as an output. Because the output is a categorical value, it is difficult for the classifier to reflect the distance between the decision boundary and the observation. Thus, the monitoring statistics have discrete values, which can degrade the performance (Wei et al., 2016).

To resolve the problem, a distance-based RTC control chart is suggested. These charts are used to calculate the statistics through the distance between decision boundaries and the observation using SVM and KLDA as the classifier (Wei et al., 2016). The control charts using these classifiers are less effective in process management because these charts could not find the cause of a fault.

Therefore, we suggest the RF based RTC control chart with SAX. SAX is famous method to improve the performance of the classification and clustering. We convert the input data by using SAX. To improve the performance, we change the parameters: alphabet size and breakpoints. Experiment shows that our method improves the performance of RTC control chart.

RTC Control Chart

RTC control chart sets reference and contrast to classify each other. The reference and contrast indicate observations in the normal and abnormal state, respectively. In Phase I, the reference data S_0 consist of the data in the normal state and the reference data size is denoted by N_0.

Let x_t be the observation at time t and N_W be the moving window size. The contrast data $S_w(t)$ changes by moving the window whenever a new observation occurs at time t, i.e., $S_w(t) = \{x_{t-N_w+1}, \cdots, x_{t-1}, x_t\}$.

The classifier builds a classification boundary at each time t. S_0 and $S_w(t)$ are labeled Class 0 and Class 1. In the normal state, S_0 and $S_w(t)$ are similar to each other. Therefore, the classifier has difficulty distinguishing between S_0 and $S_w(t)$. On the other hand, if the process is out of control, then S_0 and $S_w(t)$ are relatively easy to classify. $\hat{p}_k(x_i|t)$ is a predicted probability of the arbitrary observation x_i with class k $(k = 0, 1)$ at time t. We compute the average of S_0 to reflect the classification results (Deng et al, 2012). The monitoring statistics is as follows:

$$p(S_0, t) = \frac{\sum_{x_i \in S_0} \hat{p}_0(x_i|t)}{N_0} \quad (1)$$

Proposed Method

We convert raw data by using SAX. SAX is composed of piecewise aggregate approximation (PAA) and string-based algorithms. PAA divides a sequence X of length k (k-dimensional), $X = \{x_1, \dots, x_k\}$, into p-dimensional equally sized segments, $\bar{X} = \{\bar{x}_1, \dots, \bar{x}_p\}$, where $k \geq p$. The ith segment $\bar{x}_i$ is described as follows:

$$\bar{x}_i = \frac{p}{k} \sum_{j=k/p(i-1)+1}^{(k/p)i} x_j. \quad (2)$$

PAA represents the dimension or length of the data and the value for each segment as its average (Lin et al., 2007). After PAA transformation, SAX represents each segment of $\bar{X}$ as part of an alphabet by the decision boundary. The decision boundary determines two parameters: the alphabet size and breakpoints. According to the number of breakpoints τ, alphabet size is determined as $\tau + 1$. Generally, the determination of breakpoints is to divide the equally sized areas under the Gaussian curve (Lin et al., 2007). To improve the performance of RTC control chart, we use SAX to represent the data. We set the breakpoints by using the standard Z score. We allocate the topmost breakpoint on the standard Z score. Except the topmost breakpoint, the location of breakpoints follows Gaussian curve.

Experiment

We generate the multivariate normal distributions with zero mean and identity covariance matrix. The data has two different dimensions ($d = 10,\ 100$) and one variable shifts by two different sizes ($\lambda = \sqrt{5}, \sqrt{10}$). The shift sizes are computed by the non-centrality parameter λ described in equation (3).

$$\lambda = \sqrt{\delta^T \Sigma_X^{-1} \delta}, \quad (3)$$

To compare the performance, we use the same parameters of RTC including RF as Deng et al. (2012). To apply real-time detection, we set $k = p$ in order to prevent time information loss. In process monitoring, the control limit (CL) is a statistical criterion to classify between the normal and abnormal states. The CL is set through the statistics of an in-control process after setting the allowable limit of the Type-I error. When the monitoring statistics is greater than the CL, the process state changes from normal to abnormal. The number of observations is the run length RL_w until the first detecting points. We use the average run length (ARL) as the performance measurement as follows:

$$ARL = \frac{1}{R} \sum_{w=1}^{R} RL_w, \quad (4)$$

where R indicates the number of repetitions. We set the CL by the normal state ARL (ARL_0) that is equal to the prespecified level as $ARL_0 \cong 200$ and use the first abnormal state run length ARL_1 to compare the performance. We perform the experiment about alphabet size from 3 to 9 when the breakpoints follow the Gaussian curve and we decide the alphabet size 9 as the shortest ARL_1. Table 1 is the result of experiment. In the Table 1, each method indicates original RTC control chart, RTC control chart with Gaussian curve based SAX, and proposed methods. We repeated the experiment 100 times. The standard Z scores of proposed methods are 1.96 (SAX 1.96), 2.58 (SAX 2.58). Final method (SAX 1.6) changes second breakpoint is 1.6 and other breakpoints is the same 1.96 (SAX 1.96). We mark the performance better than RTC chart with cross.

Table 1. The comparative results of experiment (ARL_1).

Method	*d = 10*		*d = 100*	
	$\sqrt{5}$	$\sqrt{10}$	$\sqrt{5}$	$\sqrt{10}$
RTC	7.53	6.47	9.45	7.93
Gaussian	158.9	157.0	152.1	147.3
SAX 1.96	8.35	7.50	**9.40**$^+$	8.20
SAX 2.58	7.85	6.70	10.90	7.85$^+$
SAX 1.6	**7.40**$^+$	**6.20**$^+$	10.50	**7.70**$^+$

Conclusion

We proposed the RTC control chart by using SAX. The experiment shows that SAX representation method is more suitable than other methods and our method improves the performance of RTC control chart. The performance of RTC chart depend on the location of breakpoints. We confirm the SAX is possible to improve the RTC chart and necessity for adaptive breakpoints to reflect data information.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2016R1A2B4013678). This work was also supported by the BK21 Plus (Big Data in Manufacturing and Logistics Systems, Korea University) and by the Samsung Electronics Co., Ltd.

References

Deng, H.; Runger, G.; and Tuv, E. 2012. System monitoring with real-time contrasts. *Journal of Quality Technology*, 44(1), 9–27.

Lin, J.; Keogh, E.; Wei, L.; and Lonardi, S. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2), 107–144.

Wei, Q.; Huang; W.; Jiang, W.; and Zhao, W. 2016. Real-time process monitoring using kernel distances. *International Journal of Production Research*, 54(21), 6563–6578.