World Academy of Science, Engineering and Technology
International Journal of Industrial and Manufacturing Engineering
Vol:6, No:12, 2012

Nonparametric Control Chart using Density Weighted Support Vector Data Description

Myungraee Cha, Jun Seok Kim, Seung Hwan Park, and Jun-Geol Baek

Abstract—In manufacturing industries, development of measurement leads to increase the number of monitoring variables and eventually the importance of multivariate control comes to the fore. Statistical process control (SPC) is one of the most widely used as multivariate control chart. Nevertheless, SPC is restricted to apply in processes because its assumption of data as following specific distribution. Unfortunately, process data are composed by the mixture of several processes and it is hard to estimate as one certain distribution. To alternative conventional SPC, therefore, nonparametric control chart come into the picture because of the strength of nonparametric control chart, the absence of parameter estimation. SVDD based control chart is one of the nonparametric control charts having the advantage of flexible control boundary. However,basic concept of SVDD has been an oversight to the important of data characteristic, density distribution. Therefore, we proposed DW-SVDD (Density Weighted SVDD) to cover up the weakness of conventional SVDD. DW-SVDD makes a new attempt to consider dense of data as introducing the notion of density Weight. We extend as control chart using new proposed SVDD and a simulation study of various distributional data is conducted to demonstrate the improvement of performance.

Keywords—Density estimation, Multivariate control chart, One-class classification, Support vector data description (SVDD)

I. Introduction

ADVANCEMENT in scientific technology leads to the development of collecting information so that it could be possible to manufacture more technology-intensive products. However, as the amount of data increase, monitoring a number of variables simultaneously has been a trouble in production management. To improve performance of quality control in process, the univariate control chart has been extended to the multivariate control chart. As representative, Statistical Process Control (SPC) is commonly used multivariate control chart method. SPC has the advantage of cost saving because of its fast construction of chart. However, there is the weakness of SPC. Usually SPC demand data distributional assumption. As a typical example of SPC, Hotelling T^2 chart is one of the typical multivariate control charts in statistical process control (SPC) [1]. Hotelling suggest T^2 as the monitoring statistic and integrate variables as a statistic. As monitoring one statistic, more less computational complexity makes T^2 chart have merit of classifying in-control data and out-of-control data fast. Nevertheless, the application of T^2 is restricted because of the assumption that process data is following a multivariate normal distribution. In case of non-normal process data, the performance of the control chart to monitoring variable will decrease [2].

Unfortunately, most of modern industries manufacture a complex product throughout various stages of processes. That is why most data are not only non-normal but also the mixture of data having different distributional parameters respectively [3]. Thus, parametric control charts requiring distributional assumption is restricted their applications in processes so that the demand of nonparametric control chart has risen as an alternative of parametric control chart [4].

One-class classification (OCC) based control chart is one of the nonparametric control charts. One-class classification algorithm aims to classify one class from all other class and when all other class are recognized as outlier, it called as outlier detection method. In order to distinguish one class from all others, OCC method learns using only one target class like Phase I in control chart. For such a reason, Studies about OCC based control chart have been conducted.

Especially, in OCC method, control chart based on SVDD algorithm is widely used for their flexibility [5]. SVDD algorithm is basically one of the one-class classification methods. The objective of SVDD is to find optimal sphere classifying whether data is normal or outlier. As control chart, SVDD has strong point that it can construct flexible in-control region no matter how target data distributed. As mentioned before, process data are the mixture of processes and its distributional formation is not deterministic. Therefore, SVDD which can develops a flexible control boundary is effective as nonparametric control chart.

Although SVDD has strengths to reflect target data characteristics, the limitation of applied characteristic is exist. For calculating abnormality, SVDD considers only a distance between sphere and data not distribution of data [6]. When SVDD decides optimal sphere using support vectors without considering density distribution, there is great likelihood to ignore a dense area. Hence, the efficiency of control chart based on SVDD will be decreased.

To improve SVDD performance, we proposed new SVDD reflected density distribution. By introducing density weight into learning, we look forward to find optimal sphere and reduce error rate as control chart.

The structure of this paper is organized as follows. In section II, we briefly mention about conventional SVDD and next section III will be introduction of new SVDD suggested by this research. Moreover, in section IV, we present how new SVDD works as control chart and in order to demonstrate

Myungraee Cha, Jun Seok Kim and Seung Hwan Park are with the School of Industrial Management Engineering, Korea University, Seoul 136-713, Korea, e-mail: (myungraeecha@gmail.com; bliths@gmail.com; udongpang@korea.ac.kr)

Jun-Geol Baek is an associate professor in the School of Industrial Management Engineering, Korea University, Seoul, 136-713, Korea(Corresponding author to provide phone +82-2-3290-3396; fax:82-2-929-5888; e-mail:jungeol@korea.ac.kr).

improvement of conventional SVDD, we experiment various type of data distribution in section V. Final comments are presented in section VI.

II. Support Vector Data Description (SVDD)

Assume that a data set contains l one target class data objects $\{x_i, i = 1, \ldots, l\}$. Objective of SVDD is to find an optimal sphere with minimum volume containing all possible target data. Objective function expressed as follows

$$\min\ R^2 + C\sum_i^l \xi_i$$
$$s.t. \quad \|\mathbf{x}_i - \mathbf{a}\|^2 \le R^2 + \xi_i, \quad \xi_i \ge 0 \quad \forall_i \tag{1}$$

The objective of SVDD is composed by relationship between volume of sphere and the number of data in sphere. SVDD Boundary attempts to decrease volume of sphere to minimize objective function. However, when the sphere is small, the number of slack variables leaving out of sphere is increasing. Therefore, objective function is also influenced by the slack variables and this relationship is key factor to determine volume of sphere. To adjust the effect of slack variable, we could set the penalty C in (1). When penalty is relatively high value, out-of-sphere data is critical to objective function and to minimize function, it makes to cover all of data in boundary to get rid of the influence of out-of-sphere data. Thus, penalty C play a key role in adjusting the volume of sphere[7]. To solve maximize problem, Lagrange multiplier method is used and Lagrange function is

$$L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) = R^2 + C\sum_i^l \xi_i - \sum_i^l \alpha_i \left\{ R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2 \right\} - \sum_i^l \gamma_i \xi_i \tag{2}$$

With constraint Lagrange multiplier $\alpha_i \ge 0, \gamma_i \ge 0$, Lagrange function constructs to dual problem with new constraints.

$$L = \sum_i \alpha_i K(\mathbf{x}_i \cdot \mathbf{x}_i) - \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j)$$
$$s.t. \quad 0 \le \alpha_i \le C \qquad \sum_i \alpha_i = 1 \tag{3}$$

Using kernel trick, data are mapping to high-dimensional space. There are various kernel tricks such as Linear, Polynomial, RBF (Radial basis function) and usually RBF are used [8]. Boundary of SVDD will be determined by optimal value of Lagrange multiplier with solving (3) and which data having optimal Lagrange multiplier value are called as support vectors.

$$K(\mathbf{x}_i \cdot \mathbf{x}_j) = \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle = exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sigma^2} \right) \tag{4}$$

III. Density Weighted SVDD (DW-SVDD)

As we mentioned in section II, in SVDD algorithm, the criteria of abnormality is setting by distance from center of sphere to data. Only taking distance between two points into account as abnormal criteria is possible to be hesitated the consideration about density. To improve the performance of ordinary SVDD algorithm, therefore, we suggest considering density distribution of data set. To applying the concept of our idea, we introduce the notion of density weight. Density weight is expressed using k nearest neighbor (k-NN) algorithm [9]. k-NN helps to get a distance between x and kth nearest neighbor from x and it expressed as $d\left(x_i,\, x_i^k\right)$ which x_i^k represents the kth nearest neighbor from x_i. When data is located in comparatively dense area, local density is smaller than low dense area. After collecting the information about local density, we designate the density weight as

$$\rho(x_i) = 1 - \frac{d\left(x_i, x_i^k\right)}{\max_{i \in l}\, d\left(x_i, x_i^k\right)} \tag{5}$$

The distance of each data indicates local density of area data exist in real space not feature space. There is other way to consider data density in feature space. However, there might be the possibility of distortion caused by mapping data into feature space and the reflection of density in kernel space is might be a lack skill at reflecting real data characteristic of density. To apply the effect of density weight on SVDD, it replaces established penalty denoted by C in previous section. Reformed objective function of New SVDD is

$$\min\ R^2 + C\sum_i^l \rho(x_i)\, \xi_i$$
$$s.t. \quad \|\mathbf{x}_i - \mathbf{a}\|^2 \le R^2 + \xi_i, \quad \xi_i \ge 0 \quad \forall_i \tag{6}$$

In the same tale with conventional SVDD, Lagrange multiplier method is used and new formulated Lagrange function is

$$L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) = R^2 + C\sum_i^l \rho(x_i)\, \xi_i - \sum_i^l \alpha_i \left\{ R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2 \right\} - \sum_i^l \gamma_i \xi_i \tag{7}$$

By partial differentiation is 0, new equation and constraints set down as follows.

$$L = \sum_i \alpha_i K(\mathbf{x}_i \cdot \mathbf{x}_i) - \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j)$$
$$s.t. \quad 0 \le \alpha_i \le \rho(x_i)\, C \qquad \sum_i \alpha_i = 1 \tag{8}$$

As result, constraint is change by the effect of density weight. In step of determining support vector, density weight will be influence on the choice of giving constraints. As smaller the density weights are, the effect of slack variable presenting distance from center of sphere to data is reduced. As a result of density weight, each of data has different penalty depending on density unlike conventinal SVDD which have

the same penalty to all of data. Hence, by applying density weights in algorithm, we can get normal space boundary more influenced by density distribution.

IV. Constructing control chart

On account of the advantage of nonparametric control chart, recently several studies have implemented about control chart based on SVDD. Sun and Tsung suggested K chart which consisted of statistic K, distance from center of optimal sphere to data and control limit as radius of sphere [10]. Robust K chart using normalized kernel distance is proposed by Kumar et al. [11].

However, control chart using SVDD has weakness for constructing control limit. D^2 chart compensates the statistical weakness of K chart by using Bootstrap method [12]. They expected to build a nonparametric control limit using bootstrap method which widely used to estimate statistical variable when population distribution is hard to know [13]. Mentioned in previous section, the learning algorithm is important to set up a control chart and to improve performance of control chart based on SVDD, we proposed density-weighted SVDD (DW-SVDD). Based on DW-SVDD, we construct control chart. Statistic is equal to K chart in (9) and it represents the distance between the center of optimal sphere and new data $\mathbf{z}$.

$$\|\mathbf{z}-\mathbf{a}\|^2 = (\mathbf{z}\cdot\mathbf{z}) - 2\sum_i \alpha_i (\mathbf{z}\cdot\mathbf{x}_i) + \sum_{i,j}\alpha_i\alpha_j (\mathbf{x}_i\cdot\mathbf{x}_j) \quad (9)$$

In addition, control limit is derived from R^2 indicating radius of optimal sphere and calculated by the distance between support vectors[7].

$$R^2 = (\mathbf{x}_k\cdot\mathbf{x}_k) - 2\sum_i \alpha_i (\mathbf{x}_i\cdot\mathbf{x}_k) + \sum_{i,j}\alpha_i\alpha_j (\mathbf{x}_i\cdot\mathbf{x}_j) \quad (10)$$

V. Experiment

A. Simulation Setup

We conduct to simulation study to investigate the improvement and analyze the adjustment of new SVDD in control chart. Bivariate normal, banana-shaped and gamma distribution is generated for testing the performance. We set the parameter of each distribution as Table I. To test the performance of chart respectively, we generate out-of-control data which chart should detect. out-of-control data so-called as fault data are defined as data generated in mean shifted process and the shifted mean size of fault data is based on λ indicating the magnitude of non-centrality. There are three type of λ as λ_1, λ_2, λ_3. In case of bivariate normal distribution, shifted mean as fault is $\mu_1 = \mu_0 + \delta$ and the relationship between δ and λ is introduced by Kim [12].

TABLE I
Parameter of simulating data set

Distribution	Parameter	Parameter value
Gamma	Shape, Scale	[1,1]
Bivariate normal	μ_0, Σ_0	[0,0], $\begin{bmatrix} 1 & .35 \\ .35 & 1 \end{bmatrix}$
Banana-shaped	Standard deviation	1

Simulation scenarios are composed by 200 in-control data as train data set for construction of control chart (Phase I) and one thousand test data containing 900 in-control data and 100 out-of-control data for monitoring the performance of chart (Phase II).

For SVDD algorithm, the parameter of penalty C in formulation of sphere and variance σ in RBF kernel trick are important factors to the performance. To compare error rate of each chart, we use 5 cross-validation method to optimize the value of parameter C and σ for each chart in range of from 2^{-8} to 2^8 respectively.

B. Performance Measurement

TABLE II
Definition of terminology

		Actual value	
		Normal	Fault
Prediction outcome	Normal	True positive(TP)	False Positive(FP)
	Fault	False Negative(FN)	True Negative(TN)

In order to evaluate each control chart performance, we select true positive rate (TPR), true negative rate (TNR) and accuracy [14] as performance measurement. As you see in Table II, the measurement is derived following as

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP+FN}} \quad (11)$$

$$\text{True Negative Rate (TNR)} = \frac{\text{TN}}{\text{FP+TN}} \quad (12)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{(TP+FN)+(FP+TN)}} \quad (13)$$

Considering three terms simultaneously is helpful to compare the performance of control chart. If fault detection accuracy is only performance measurement, the control boundary will be made as smaller as they could in order to raise fault detection rate giving up the target accuracy. Therefore, to prevent the situation, we decide to handling conflicting rate, TPR and TNR. Also higher score of accuracy means optimal solution of control chart because it is the point that TPR and TNR both are highest.

C. Comparison Results

Afterward the construction of control chart, we compare the performance with conventional SVDD [5] and another SVDD introduced by Liu et al.[15]. They suggest the different notion of penalty and proved the effect of proposed SVDD as novelty detection algorithm using UCI repository data. To compare the effectiveness of each SVDD as control chart, we also construct a control chart. Statistic and control limit of all control charts using the performance comparison are under the

TABLE III
EXPERIMENTAL RESULT

		T^2			SVDD			Liu et al.[15]			Density-Weighted SVDD		
		TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy
Gamma	λ_1	0.7878	0.8308	0.8093	0.9067	0.8933	0.9000	0.9023	0.8698	0.8860	0.9111	0.9040	0.9075
	λ_2	0.8698	0.9463	0.9081	0.9796	0.9810	0.9803	0.9574	0.9840	0.9707	0.9843	0.9837	0.9840
	λ_3	0.9450	0.9688	0.9569	0.9800	0.9998	0.9899	0.9779	0.9916	0.9848	0.9797	0.9991	0.9894
Banana-shaped	λ_1	0.5830	0.4822	0.5326	0.7220	0.4934	0.6077	0.7178	0.4934	0.6056	0.7161	0.4999	0.6080
	λ_2	0.8067	0.3683	0.5875	0.8717	0.5854	0.7285	0.7043	0.7412	0.7227	0.8663	0.6591	0.7627
	λ_3	0.8439	0.4388	0.6414	0.8689	0.7388	0.8039	0.8745	0.7410	0.8078	0.8642	0.7541	0.8091
Normal	λ_1	0.7518	0.5474	0.6496	0.7224	0.5211	0.6217	0.7300	0.5032	0.6166	0.7498	0.5095	0.6297
	λ_2	0.8384	0.8769	0.8577	0.8554	0.8118	0.8336	0.7024	0.8606	0.7815	0.9369	0.7644	0.8507
	λ_3	0.9661	0.9627	0.9644	0.9700	0.9775	0.9738	0.9536	0.9875	0.9706	0.9750	0.9935	0.9842

same conditions. Also, we compare with Hotelling T^2 chart, most commonly used multivariate control chart.

As you see in Table III, density-weighted SVDD has good performance for data following gamma distribution having small size of λ. In case of shift size is λ_1, accuracy of T^2 chart is 80.93%, Liu et al. suggesting SVDD has 88.6% accuracy. Next higher accuracy is SVDD and DW-SVDD having 90% and 90.75% respectively. In median shift in gamma distribution, the ranking is represented as T^2, SVDD proposed Liu, SVDD and DW-SVDD having 90.81%, 97.07%, 98.03%, 98.40%. However, in case of large shift, λ_3 in gamma distribution dataset, conventional SVDD has highly efficient. The reason is that when DW-SVDD implements the learning with data having differential density distribution, DW-SVDD tends to construct boundary more intensively into dense area. It is natural effect of density weight and it causes target accuracy lower. Therefore, DW-SVDD might be more sensitive to density distribution of data than standard SVDD and SVDD suggested by Liu et al. So, DW-SVDD is more proper to treat or monitoring dataset which needs to handle more carefully because it is more customized to detect small shift. Generally fault data with small shift is hard to be detected. Therefore, tt is meaningful improvement that DW-SVDD is strong at detecting smaller shift data than conventional SVDD.

Also, in Banana-shaped distribution, Proposed SVDD is outstanding in performance without reference to how much data are shifted. In smaller fault shift λ_1,accuracy is 60.8% in DW-SVDD. λ_2 and λ_3 are 76.27%, 80.91% respectively and this accuracy is the highest score comparing with other algorithms.

As against other experimental scenario, in normal distribution, T^2 chart works well more than other algorithms in small shift and medium shift. The reason is that T^2 chart has the advantage of controlling data as following normal distribution. In case of λ_2, the highest accuracy is 85.77% in T^2 and the next is DW-SVDD as 85.07%. Comparing with conventional SVDD which has 83.36%, it is enough to insist that for normal distributional data DW-SVDD is useful more than SVDD based control chart. Even in case of large mean shift case, DW-SVDD has the best performance having 98.42% accuracy much higher than T^2. It is meaningful to us that SVDD we proposed have efficient as control chart more than control chart based on SVDD and SVDD proposed by Liu et al.

Therefore, we conclude that DW-SVDD is competitive nonparametric control chart not only in non-normal data but normal distributional data. Introducing the notion of density weight in SVDD learning step, DW-SVDD is improved the performance of SVDD as control chart. DW-SVDD is efficient to increase the true negative rate as keeping the true positive rate high which is in inverse proportion to true negative rate.

VI. CONCLUSION

In this paper, we introduce the notion of density weight to reflect the density data description in SVDD. Density weight is composed of local density rate from k-NN method. Using distance between data and kth nearest neighbor of data, we estimate local density of data and ratio of maximum distance of kth nearest neighbor to distance of each data from kth neighbor define density weight. As applying the notion in conventional SVDD, we proposed a new SVDD method named DW-SVDD. As well as proposing new SVDD, we also contribute to monitoring field introducing control chart based on DW-SVDD. As control chart, weakness of SVDD is stand out. Conventional SVDD tends to decide the data as outliers only considering how far the data are. However, the inclination of SVDD is restrict its application and capacity as control chart. To improve performance of control chart based on SVDD, we proposed control chart based DW-SVDD and we demonstrate that the reflection of density has a good

influence on performance of control chart regardless of data description. In addition to prove the performance of control chart based on DW-SVDD, we are planning to verify the performance of DW-SVDD as outlier detection method or one-class classification method using data from real research data. Furthermore, when using k-NN method as estimator of local density, the value of k has effect on the performance of classification. To optimize the value of k, various optimization method can be used such as Tabu search, Genetic algorithm and so on. In the future research, therefore, application of optimization of parameter and variables will be translated into action.

ACKNOWLEDGMENT

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the IT R&D Infrastructure Program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-(B1100-1101-0002)).

REFERENCES

[1] H. Hotelling, *Multivariate Quality Control*. McGraw-Hill, 1947.
[2] D. C. Montgomery, *Introduction to statistical quality control*. New York Singapore : John Wiley, 1985.
[3] S. T. Bakir, "Distribution-free quality control charts based on signed-rank-like statistics," *Communications in Statistics - Theory and Methods*, vol. 35, no. 4, pp. 743–757, 2006.
[4] S. Bersimis, S. Psarakis, and J. Panaretos, "Multivariate statistical process control charts: An overview," *Quality and Reliability Engineering International*, vol. 23, no. 5, pp. 517–543, 2007.
[5] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, no. 11-13, pp. 1191–1199, 1999.
[6] K. Lee, D. W. Kim, K. H. Lee, and D. Lee, "Density-induced support vector data description," *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 284–289, 2007.
[7] D. J. Tax and R. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.
[8] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge University Press, 2004.
[9] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *Sigmod Record*, vol. 29, no. 2, pp. 93–104, 2000.
[10] R. Sun and F. Tsung, "A kernel-distance-based multivariate control chart using support vector methods," *International Journal of Production Research*, vol. 41, no. 13, pp. 2975–2989, 2003.
[11] S. Kumar, A. K. Choudhary, M. Kumar, R. Shankar, and M. K. Tiwari, "Kernel distance-based robust support vector methods and its application in developing a robust k-chart," *International Journal of Production Research*, vol. 44, no. 1, pp. 77–96, 2006.
[12] T. Sukchotrat, S. B. Kim, and F. Tsung, "One-class classification-based control charts for multivariate process monitoring," *IIE Transactions*, vol. 42, no. 2, pp. 107–120, 2010.
[13] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.
[14] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, ser. Springer Series in Statistics. New York, NY: Springer-Verlag New York, 2009.
[15] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "Svdd-based outlier detection on uncertain data," *Knowledge and Information Systems*, pp. 1–22, 2012.