Journal of the Korea Academia-Industrial cooperation Society
Vol. 20, No. 8 pp. 216-223, 2019

https://doi.org/10.5762/KAIS.2019.20.8.216
ISSN 1975-4701 / eISSN 2288-4688

Improved Fault Detection Based on One-Class Classification and Feature Selection

Hyun-Woo Cho
Department of Materials-Energy Science and Engineering, Daegu University

단일 클래스 분류와 특징 선택에 기반한 향상된 이상 감지

조현우
대구대학교 신소재에너지공학과

Abstract Fault detection during production processes is one of the required operational tasks to run production processes both safely and consistently. Unexpected operational events or undetected process faults can have a serious impact on the production systems and subsequently on the final products' quality. In addition, such situations may lead to malfunctions or breakdowns of production processes. To reliably detect such abnormalities, a new one-class classification-based detection scheme has recently been developed The proposed method consists of four steps:1) noise filtering, 2) feature selection, 3) nonlinear representation and 4) outlier detection. The performance of the proposed scheme was demonstrated using the multivariate data obtained from a simulation process. The results have shown that the proposed method produced reliable monitoring results and outperforms any existing methods with an average improvement of 25.4%. The use of proper feature selection in the proposed framework yielded better detection performance.

요 약 생산 공정에서 발생되는 공정 이상을 적시에 감지하는 것은 생산 공정의 안전하고 일관된 조업 및 운영에 필수적인 요소 중 하나로서 반드시 필요하다. 예측되지 못하거나 적절하게 감지되지 못한 공정 이상은 전체 생산 공정과 공정에서 생산되는 최종 제품의 품질에 심각한 영향을 줄 수 있기 때문이다. 또한 이러한 상황은 공정 기능 불량과 고장으로 이어지게 된다. 이러한 공정 이상을 신뢰성 있게 적시에 검출하기 위해 본 연구에서는 새로운 단일 클래스 분류에 기반한 공정 이상 감지 기법을 제안한다. 본 연구의 제안된 방법은 잡음 필터링, 특징 선택, 비선형 표현 및 특이치 검출의 네 단계로 구성된다. 본 연구에서는 시뮬레이션 공정의 측정치를 활용하여 제안된 방법의 성능을 평가하였다. 그 결과 제안된 공정 이상 탐지 기법이 신뢰할 수 있는 모니터링 결과를 산출하였으며 기존 비교 대상 방법들보다 평균 25.4% 향상된 성능을 보여 주었다. 또한 적합한 특징 선택을 통하여 보다 향상된 이상 감지 성능을 얻을 수 있었다.

Keywords : Detection, One-Class Classification, Feature Selection, Nonlinear Representation, Filtering

1. Introduction

Unexpected events such as breakdowns and malfunctions in a production system have a critical impact on process operation and the quality of final products. Detection of faults or outlier is one of the operational tasks needed to maintain a process safely. For this purpose, there has been also much interest in nonlinear statistical methods such as support vector

This research was supported by the Daegu University Research Grant, 2015.
*Corresponding Author : Hyun-Woo Cho(Daegu Univ.)
email: hwcho@daegu.ac.kr
Received April 22, 2019 Revised July 25, 2019
Accepted August 2, 2019 Published August 31, 2019

machines (SVM)[1]. Similar to SVM, other kernel-based nonlinear techniques have been also developed: kernel partial least squares (KPLS), kernel principal component analysis (KPCA) and kernel Fisher discriminant analysis (KFDA)]. They have been applied to many practical issues of classification, detection, prediction, and so on[2-4].

Recently, one-class classification approaches to fault detection have been studied, which differ from conventional classification based approach in the way how a one-class classifier is trained[5]: it is trained by normal data or target data only and never considers abnormal or fault data. Such a characteristic is quite useful because in most cases one of classes (i.e., fault data) is under-sampled relatively. Measurements on normal operating conditions of a process are very cheap and easy to obtain. On the contrary, it is very expensive and time-consuming, though not impossible, to obtain measurements on all faulty situations. Support vector data description (SVDD), as one of one-class classification techniques, provides a compact description of target data[5]. SVDD seeks to represent original data in a spherical minimal-volume domain enclosing target points of the datasets.

This work develops a new SVDD-based one-class classification method for fault detection. It also include additional steps of noise filtering, feature selection, and nonlinear representation. The noise filtering is to remove from target data unwanted variation or noises of data. Then, feature selection step is performed to select important variables. The exclusion of redundant variables from original data may yield better results with simpler models. Nonlinear techniques also can be used to extract nonlinear patterns of data. Compared to linear techniques, nonlinear one can provide an efficient lower-dimensional representation of data.

The first objective of this paper is to compare the proposed method and existing detection methods. Based on simulation data of a test process detection results of the proposed method are compared with those of four frequently used methods. The second objective is to evaluate the advantage or importance of feature selection. The proposed method is tested by using different feature selection methods.

2. Methods

Fig. 1 shows an overall picture of the proposed method. It includes four steps, namely, noise filtering, feature selection, nonlinear representation, and fault detection. An orthogonal filter-based preprocessing is first performed. It can remove unwanted variation of data. Then, feature selection step is performed to select important variables that contribute to the separation between normal and abnormal data. The next step is to extract nonlinear patterns of data using one of nonlinear representation techniques. They provide an efficient representation of original. Finally, SVDD-based detection model is constructed to detect a fault. Such an empirical model is obtained by determining optimal decision boundary, against which future operations can be referenced or monitored.

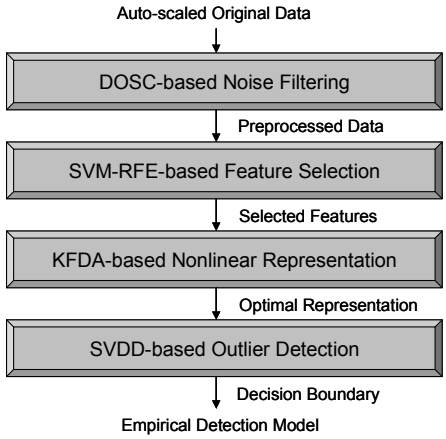

Fig. 1. Overall framework

As one of pre-processing methods, orthogonal filtering's goal is to get rid of systematic variation from independent variables **X** that is unrelated to response variables **Y**[6]. **X**'s largest variation having no correlation with **Y** must be removed from **X**. In this work OSC is modified so that it functions as noise filtering for a detection of outliers. The **Y** matrix includes information about class memberships so that each column of the **Y** has binary values of zero or one. The first task determines the first principal component scores of the data. These scores are orthogonalized to **Y** yielding correction vectors and PLS weight vectors. Then scores are updated by processing the correction vectors, which is also orthogonalized to **Y**. Entire steps are repeated until scores have converged, and correction vectors are located towards the vectors orthogonal to **Y**. As a result, a loading vector can be calculated producing residuals, and in such a way the next components can be obtained.

Support vector machines (SVM) feature selection method combined with a recursive feature elimination (RFE) procedure performs sensitivity analysis for an appropriately defined cost function[7]. In the linear kernel case, for example, let us define a cost function $J=(1/2)\parallel w \parallel$. Then the least sensitive feature with the minimum weight is eliminated first. This eliminated feature becomes ranking *n*. Then the SVM classification model is re-trained without the eliminated feature. The next step is to remove the feature having minimal magnitude of weights. The eliminated feature becomes ranking *n*-1 at this time. By repeating this process until no feature is left, one can rank all the features. Given training instances $X_{all}=[x_1,...,x_l]$ with class labels $y=[y_1,...,y_l]^T$, initialize the subset of features $s=[1,2,...,n]$ and r=[]. For general kernel cases, let us define a cost function

$$J=(1/2)\alpha^T H\alpha-\alpha^T e \quad (1)$$

where e is an l dimensional vector of ones, $H_{hk}=y_h y_k K(x_h,x_k)$, and α is a Lagrange multiplier. It is assumed that there are no changes in α in order to get the change in J caused by removed feature *i*.

$$H(-i)_{hk}=y_h y_k K(x_h(-i),x_k(-i)) \quad (2)$$

where (-*i*) represents that the feature *i* has been removed. The sensitivity function is given by

$$\begin{aligned} DJ(i) &= J-J(-i) \\ &= (1/2)\alpha^T H\alpha-(1/2)\alpha^T H(-i)\alpha \end{aligned} \quad (3)$$

The SVM-RFE algorithm for general kernels can be implemented by repeating (i) through (v) until *s* becomes an empty array as follows:

(i) Construct new instances $X=X_{all}(:,s)$

(ii) Train SVM(X,y) to obtain α

(iii) Compute the ranking criterion

$$DJ(i)=(1/2)\alpha^T H\alpha-(1/2)\alpha^T H(-i)\alpha$$

(iv) Find the feature *f* such that

$$f=\arg\min_i DJ(i)$$

(v) Update *r* and remove the feature *f* from *s*: $r=[s(f),r], s=s-s(f)$.

When a linear kernel is used, the same procedure is repeated except (iii) (where one computes a gradient $w=\nabla g(x)=\sum_{i\in SV}\alpha_i y_i x_i$).

The objective of nonlinear kernel discriminant analysis is to obtain certain directions, along which hidden groups of data are separated as clearly as possible. These directions can be obtained by maximizing between-class scatter S_b^Φ while minimizing total scatter S_t^Φ. Similar to linear FDA, it is done by maximizing the Fisher criterion[4]:

$$J^\Phi(\Psi)=\frac{\Psi^T S_b^\Phi \Psi}{\Psi^T S_t^\Phi \Psi}, \Psi\neq 0. \quad (4)$$

The optimal discriminant vectors are described as a linear combination of the data in feature space. Thus there exist coefficients b_i such that

$$\Psi=\sum_k^M b_k \Phi(x_k)=H\alpha \quad (5)$$

where $H=[\Phi(x_1),\dots,\Phi(x_M)$ and $\alpha=(b_1,...,b_M)^T$. Let the elements of *K*, a kernel matrix, be given

as $[\tilde{K}]_{ij} = < \Phi(x_i), \Phi(x_j) > = k(x_i, x_j)$ in which $< x, j >$ indicates the dot product and $k(x_i, x_j)$ is a kernel function. A kernel function of $k(x_i, x_j)$ is used to avoid executing the nonlinear mapping $\Phi(x)$ and dot products in the feature space[4]. In summary, KFDA first performs a nonlinear mapping $\Phi(x)$ to project an input vector to a feature space. Then linear FDA is performed in feature space, giving a lower dimensional KFDA space based representation. In this work, instead of KPCA, KFDA is employed for representing nonlinear data because it provides better discrimination between normal and abnormal data.

SVDD is one of one-class classification methods. One-class classification methods can be divided into density estimation, reconstruction, and boundary methods. As one of boundary methods, SVDD seeks to envelop data within a feature space with the volume as small as possible. When a suitable kernel is introduced, this model can be more powerful and may give reliable results. Let us consider a SVDD model with a hyper-sphere boundary around data. Here the sphere is characterized by center μ and the radius R. The problem of SVDD is to determine μ and R that has minimal-volume hyper-sphere containing all samples x_i, i=1, 2$\cdots I$. Here the error function is $F(R, \mu) = R^2$ with $\| x_i - \mu \|^2 \leq R^2$. When there are some abnormal samples, a large sphere can be obtained but it will not represent the data well. Slack variables $\xi (\geq 0)$ are introduced to allow for some samples outside the sphere.

The distance between x_I and μ is not smaller than the R^2 so that larger distances are penalized. The minimization problem is expressed in the following:

$$F(R, \mu) = R^2 + C \sum^{I} \xi_i \tag{6}$$

where the parameter C indicates trade-off between the sphere's volume and the number of samples outside it. This should be minimized under the following constraints[5]:

$$\| x_i - \mu \|^2 \leq R^2 + \xi_i, \xi_i \geq 0 \tag{7}$$

By incorporating Eq. 7 into Eq. 6, one can construct the Lagrangian function:

$$L(R, \mu, \alpha_i, \beta_i, \xi_i) = R^2 + C \sum^{I} \xi_i - \sum^{I} \alpha_i [R^2 + \xi_i - (\| x_i \|^2 - 2\mu x_i + \| \mu \|^2)] - \sum^{I} \beta_i \xi_i \tag{8}$$

with the Lagrange multipliers $\alpha_i \geq 0$ and $\beta_i \geq 0$. Here, L is minimized with respect to R, μ, ξ_i and maximized with respect to α_i and β_i. Equation (8) can be expressed as

$$L = \sum^{I} \alpha_i (x_i \cdot x_j) - \sum^{I} \alpha_i \alpha_j (x_i \cdot x_j) \tag{9}$$

with $0 \leq \alpha_i \leq C$. Maximization of Eq.9 yields a set of α_i. When a sample x_i satisfies the inequality $\| x_i - \mu \|^2 \leq R^2 + \xi_i$, the constraint is satisfied and αi = 0. Thus only samples xi with $\alpha_i > 0$ are needed in SVDD, which are called support vectors. By using kernel trick the SVDD problem of Eq. 9 can be expressed as[5]:

$$L = \sum^{I} \alpha_i K(x_i, x_j) - \sum^{I} \alpha_i \alpha_j K(x_i, x_j) \tag{10}$$

with constraints $0 \leq \alpha_i \leq C$ and $\sum \alpha_i = 1$. The use of different kernel functions result in different boundaries in SVDD.

3. Results

The performance of the proposed scheme is demonstrated using simulated data. A test process is a simulation of an actual industrial process which has been widely used for optimization strategies, monitoring, and diagnosis[8]. It has 22 continuous process variables, 12 manipulated variables, and 19 composition data. Training data and test data sets for each of faults include Gaussian noise. A variable plot for the fault F3 are displayed in Fig. 2, where Y axis represents measurement values.

Fluctuations of measurement variables can be seen after around 480 sampling time. Table 1 lists a total of 10 faults of the process considered in the case study. In this work the proposed framework was implemented in Matlab environment.

Table 1. List of process faults of case study

Fault	Description
F1	A/C feed ratio, B composition constant
F2	B composition, A/C ratio constant
F3	D feed temperature
F4	Reactor cooling water inlet temperature
F8	A/B/C feed composition
F9	D feed temperature
F10	C feed temperature
F13	Reaction kinetics
F14	Reactor cooling water valve
F21	Fixed valve position

We first performed orthogonal filtering on training data and then test data. It is interesting to find the effect of using the orthogonal filtering and its number of components in the proposed framework. For this comparison purpose, three regression models are constructed using dependent variables as class memberships. These models are classical PLS algorithms applied to classification problems, which is often called discriminant PLS (DPLS).

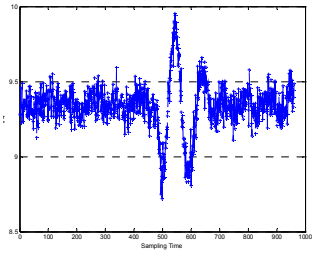

Fig. 2. Variable plot for F3

Table 2 summarizes the modeling results of the three DPLS models. The three DPLS models are built based on training data to predict class memberships of data points: a DPLS model is constructed without direct orthogonal signal correction (DOSC0) and the others by doing orthogonal filtering with 1 (DOSC1) or 2 (DOSC2) components retained. Here, There $R^2\mathbf{X}$ (cumulative sum of the squares of $\mathbf{X}$ explained) and $R^2\mathbf{Y}$ (cumulative sum of the squares of $\mathbf{Y}$ explained) represent measures of model's ability to fit data, and Q^2 indicates its discriminating power, i.e., cumulative fraction of $\mathbf{Y}$'s total variance predicted by extracted components.

For example, the DOSC0 model used 61.5% of $\mathbf{X}$ to explain 37.0% of $\mathbf{Y}$ with Q^2 31.4%. As shown in Table 2, the use of the orthogonal filtering produced better discriminating power than DPLS without it: Q^2 values 0.841 (DOSC1) and 0.829 (DOSC2) vs. 0.314 (DOSC0). It means that the use of orthogonal filtering produces good separation between different groups of data, resulting in better discrimination performance. Thus the main advantage of using orthogonal filtering in the proposed framework is to improve the separability of different faults. It is made possible by removing the irrelevant variation that is not related to the separation. The number of components retained in orthogonal filtering should also be determined. In this work, one component of orthogonal filtering (DOSC1) is chosen because its Q^2 value is higher than that of DOSC2 (i.e., 0.841 vs. 0.829). In addition, SVM-RFE feature selection algorithm was applied to orthogonally filtered training data of the test process. The feature selection results (i.e., ranking of variables) would be different depending on which kernel function to use.

Table 2. DPLS results

	No. of DOSC Components		
	DOSC0	DOSC1	DOSC2
R2X	0.615	0.407	0.450
R2Y	0.370	0.852	0.873
Q2	0.314	0.841	0.829

Thus it is necessary to determine the optimal kernel function to be used in SVM-RFE feature selection. For this purpose, we tested 4 different

kernels: linear, polynomial, sigmoid, and RBF kernel functions. The feature selection results based on them are displayed in Figure 3. A detection success rate (i.e., proportion of the observations correctly detected) is plotted for each of the four kernels with respect to the number of features selected. As a result, RBF kernel was selected for the test process because it has the maximum rate of 0.86 using only 25 selected features.

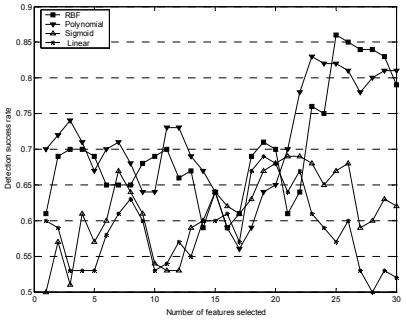

Fig. 3. SVM-RFE feature selection results

The second maximum is obtained from polynomial (i.e., 0.83) with 23 selected features. On the contrary, the use of linear or sigmoid kernels degrades the detection performance significantly.

As shown in Table 3, detection results of the proposed method were obtained and compared with those of two PCA-based methods. The first is a linear PCA providing lower-dimensional representation of original data. The second method is an extended version of traditional PCA, called dynamic PCA, which takes into account the serial correlations. For each of the two methods, Hotelling's T^2 and Q statistics are used to detect a fault.

Table 3. Detection results

Fault	detection success rate		
	PCA	DPCA	Proposed
F1	0.92/1.00	0.91/0.98	1.00
F2	0.45/0.41	0.68/0.85	0.98
F3	0.13/0.18	0.31/0.23	0.35
F4	0.89/0.85	0.91/0.99	0.99
F8	0.95/0.96	0.97/0.98	0.75
F9	0.18/0.29	0.26/0.34	0.63
F10	0.37/0.43	0.51/0.59	0.81
F13	0.89/0.88	0.90/0.97	0.98
F14	0.87/1.00	0.92/1.00	1.00
F21	0.29/0.42	0.43/0.70	0.83
Average	0.59/0.64	0.68/0.76	0.83

Table 3 shows the detection results for the test data, in which we underlined the maximum detection rates for each of the faults. For example, the fault case of F1 has the maximum rate of 1.00, which was obtained from Q statistic-based PCA method and the proposed method. The proposed method, as shown in Table 3, showed the best detection performance in that it yielded the highest detection rates for all fault cases. It is also observed that the maximum rates in some fault cases were achieved by two or three methods. The fault case of F4 has the maximum rate of 0.99 obtained from Q statistic-based DPCA method and the proposed method. This is also the case in F14, in which three methods yielded the maximum rates of 1.00.

It is observed that the proposed method outperformed other detection methods. The average detection rate of the proposed method (i.e., 0.83), in addition, is better than those of the other methods: 0.59, 0.64, 0.68, and 0.76. Table 3 also showed that the dynamic PCA method is better than the linear PCA method. It may be due to the fact that DPCA takes into account the serial correlations. The linear PCA method assumes that the data at one time is independent to the data at past time instances. But it is not the case in real data. From Table 3 we can find that the proposed method improved the detection performance of this case study significantly.

Table 4. Performance comparison

Fault	detection success rate		
	Without	GA	SVM-RFE
F1	0.88	1.00	1.00
F2	0.84	0.95	0.98
F3	0.21	0.28	0.35
F4	0.79	0.94	0.99
F8	0.58	0.70	0.75
F9	0.51	0.60	0.63
F10	0.61	0.75	0.81
F13	0.81	0.92	0.98
F14	0.85	0.99	1.00
F21	0.63	0.78	0.83
Average	0.67	0.79	0.83

Additionally, the proposed method is evaluated by using different feature selection methods, including no feature selection at all. Genetic algorithm (GA) based feature selection method is considered because it has been used successfully in solving many problems. This work utilized the GA algorithm developed by Leardi and Gonzalez[9] in order to do feature selection for the data. The detection performance of the proposed framework is re-examined with a slight modification using the same test process. The only change was made in the feature selection step of the proposed framework.

The results of the performance comparison are summarized in Table 4. For an easy comparison, the last column of Table 3 is reproduced in the last column of Table 4. The result shows that the use of SVM-RFE feature selection yielded the best performance in all fault cases. Equal maximum detection rates are reported from F1. In terms of average detection success rate, in addition, it outperforms the others: 0.83 vs. 0.79 (GA feature selection) and 0.67 (without feature selection). It is notable that the detection performance decreases significantly when no feature selection is used.

4. Conclusion

In this work, a new one-class classification detection method has been proposed and evaluated in terms of detection performance and feature selection. The proposed method consists of four steps: DOSC noise filtering, SVM-RFE feature selection, KFDA nonlinear representation, and SVDD based detection. The DOSC noise filtering removed unwanted variation of training data. As given in the case study the proposed method yielded good detection results and outperforms linear and dynamic PCA methods: 25.4% average performance improvement as shown in Table 3. In addition, this work illustrated the importance of feature selection. The use of feature selection (SVM-RFE and GA) in the proposed method yielded better detection performance than the method without feature selection.

References

[1] V. N. Vapnik, “The nature of statistical learning theory“, Springer, USA, 1999.

[2] B. Schölkopf, A. J. Smola, K. Müller, “Nonlinear component analysis as a kernel eigenvalue problem”, *Neural Computation*, Vol.10, pp.1299-1319, 1998. DOI: https://doi.org/10.1162/089976698300017467

[3] R. Rosipal, L. J. Trejo, “Kernel partial least squares regression in reproducing Kernel Hilbert space”, *Journal of Machine Learning Research*, Vol.2, pp.97-123, 2001.

[4] G. Baudat, F. Anouar, “Generalized discriminant analysis using a kernel approach”, *Neural Computation*, Vol.12, pp.2385-2404, 2000. DOI: https://doi.org/10.1162/089976600300014980

[5] D. M. J. Tax and R. P. W. Duin, “Support vector data description”, *Machine Learning*, Vol.54, pp.45-66, 2004. DOI: https://doi.org/10.1023/B:MACH.0000008084.60811.49

[6] S. Wold, H. Antti, F. Lindgren, J. Öhman, “Orthogonal signal correction of near-infrared spectra”, *Chemometrics and Intelligent Laboratory Systems*, Vol.44, pp.175-185, 1998. DOI: https://doi.org/10.1016/S0169-7439(98)00109-9

[7] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, “Gene selection for cancer classification using support vector machines”, *Machine Learning*, Vol.46, pp.389-422, 2002. DOI: https://doi.org/10.1023/A:1012487302797

[8] J. J. Downs, E. F. Vogel, "A plant-wide industrial process problem", Computers and Chemical Engineering, vol.7, pp.245-255, 1993.
DOI: https://doi.org/10.1016/0098-1354(93)80018-I

[9] R. Leardi. A. L. Gonzalez, "Genetic algorithms applied to feature selection in PLS regression: how and when to use them", Chemometrics and Intelligent Laboratory Systems, Vol.41, pp.195-207, 1998.
DOI: https://doi.org/10.1016/S0169-7439(98)00051-3

Hyun-Woo Cho [Regular member]

- Aug. 2003 : POSTECH., Industrial Eng., PhD
- Aug. 2003 ~ Aug. 2007 : GIT/UT, Research Associate
- Sep. 2007 ~ Feb. 2011 : SEC, Senior Engineer
- Mar. 2011 ~ Current : Daegu Univ., Professor

〈Research Interests〉
Process Monitoring, Data Mining, Energy