International Journal of
Applied Mathematics,
Electronics and Computers
ISSN: 2147-8228 http://ijamec.atscience.org

Advanced Technology and Science

Original Research Paper

SVDD Based Data-Driven Fault Detection

Yusuf SEVIM

Accepted 3rd September 2016

Abstract: Conventional data driven process monitoring algorithms are limited to Gaussian process data for principal component analysis (PCA) algorithm and non-Gaussian process data for independent component analysis (ICA) algorithm. This paper provides a comparison study between the conventional data driven methods and support vector data description (SVDD) algorithm for fault detection (FD). Different from the traditional methods, SVDD algorithm has no Gausssian assumption. Thus the distribution of process data is not important for SVDD method. In order to compare their FD performances of the proposed methods from the application viewpoint, Tennessee Eastman (TE) benchmark process is utilized to compare the results of all the discussed methods. Simulation results on TE process show that ICA and SVDD methods perform better for false faults than the PCA method.

Keywords: Process Monitoring, Fault Detection, Support Vector Data Description, Independent Component Analysis, Principal Component Analysis, Statistical Process Control.

1. Introduction

In statistical process control (SPC) systems there exist many variables need to be monitored. These measurements provide useful information about the status of the systems. So applying the univariate SPC methods on that systems may produce false alarms. Using multivariate SPC (MSPC) methods can overcome that problem. MSPC has several advantages over univariates, for example, showing relationships between variables which cannot be detected with univariate statistics, and help to understand the interaction between variables. PCA and ICA algorithms have been widely used as a multivariate statistic intended to find latent variables in the FD field [1-6].

PCA is a well-known algorithm, and depends only on the second order statistics which means that the latent variables capture the most variance of the source signals. Unlike PCA, the goal of ICA algorithms is to minimize the statistical dependence between the basis vectors, and there is no closed form expression for ICA algorithms.

Another efficient FD algorithm is SVDD algorithm proposed by Tax and Duin [7]. SVDD algorithm aims at finding spherically shaped boundary around a data set. In this boundary, a hypersphere enclosing most of the data set belonging to the class of interest and rejecting the outliers. SVDD is a new method in the FD area, but it has been used in a wide range of FD applications [8-10]. In this paper, it is employed the one-class SVDD to find the separating boundary the normal data set and faulty data set. In this respect, SVDD is implemented and compared with standard MSPC methods for FD.

The remainder of this paper is organized as follows. Section II describes PCA and ICA algorithms. How to obtain SVDD algorithm is given in section III. Section IV gives monitoring results of the application to TE process. Finally, section V provides a concluding summary of this paper.

In statistical process control (SPC) systems there exist many variables need to be monitored. These measurements provide useful information about the status of the systems. So applying the univariate SPC methods on that systems may produce false alarms. Using multivariate SPC (MSPC) methods can overcome that problem. MSPC has several advantages over univariates, for example, showing relationships between variables which cannot be detected with univariate statistics, and help to understand the interaction between variables. PCA and ICA algorithms have been widely used as a multivariate statistic intended to find latent variables in the FD field [1-6].

PCA is a well-known algorithm, and depends only on the second order statistics which means that the latent variables capture the most variance of the source signals. Unlike PCA, the goal of ICA algorithms is to minimize the statistical dependence between the basis vectors, and there is no closed form expression for ICA algorithms.

Another efficient FD algorithm is SVDD algorithm proposed by Tax and Duin [7]. SVDD algorithm aims at finding spherically shaped boundary around a data set. In this boundary, a hypersphere enclosing most of the data set belonging to the class of interest and rejecting the outliers. SVDD is a new method in the FD area, but it has been used in a wide range of FD applications [8-10]. In this paper, it is employed the one-class SVDD to find the separating boundary the normal data set and faulty data set. In this respect, SVDD is implemented and compared with standard MSPC methods for FD.

The remainder of this paper is organized as follows. Section II describes PCA and ICA algorithms. How to obtain SVDD algorithm is given in section III. Section IV gives monitoring results of the application to TE process. Finally, section V provides a concluding summary of this paper.

2. Process Monitoring Based on PCA and ICA

The basic model considered in PCA and ICA is

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \quad (1)$$

where $\boldsymbol{s} = [s_1, \ldots, s_d]^{\boldsymbol{T}}$ is a vector of unknown source signals which are independent, and $\boldsymbol{x}$ is obsevation vector which is mixtures of the source signals via unknown mixing matrix $\boldsymbol{A}$. The objective of PCA is to make variables uncorrelated through orthogonal rotation. The orthogonal rotation matrix is obtained by using eigenvalue decomposition on covariance matrix

Department of Electrical and Electronics Engineering, KAradeniz Technical Universiy, Campus, 61080, Trabzon/Turkey
Corresponding Author: Email: ysevim@ktu.edu.tr
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}}{m-1} = \boldsymbol{V\Lambda V^{T}} \tag{2}$$

where $\boldsymbol{\Lambda}$ is the diagonal eigenvalue matrix with its diagonal elements in decreasing order, $\boldsymbol{V}$ is orthogonal eigenvector matrix . The loading matrix $\boldsymbol{P}$ is formed by a first vector which possesses acceptable percent of data variance. The transformation of $\boldsymbol{X}$ matrix is called as the score matrix and calculated as follows

$$\boldsymbol{T} = \boldsymbol{XP} \tag{3}$$

The transformation into the original space is achieved by using (3)

$$\widehat{\boldsymbol{X}} = \boldsymbol{TP^{T}} \tag{4}$$

The residual matrix $\boldsymbol{E}$ is calculated as

$$\boldsymbol{E} = \boldsymbol{X} - \widehat{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{TP^{T}} \tag{5}$$

PCA algorithm divides data space into two parts. First part is determined by a first major component and has the greatest data variance. The other part is defined by a small percentage of data variance and shows noise. T^2 statistic can be used to measure the variation of PCA model for a loading vector.

$$T^2 = \boldsymbol{x^{T}V\Lambda_{a}^{-1}V^{T}x} \tag{6}$$

where $\boldsymbol{\Lambda_a}$ is the first a rows and columns of $\boldsymbol{\Lambda}$. In PCA the $\boldsymbol{X}$ data is assumed to follow a multivariate normal distribution so T^2 follows an F distribution with degrees of freedom A and $A - M$, and confidence limits can be calculated as follows

$$\mathrm{T}_{\alpha}^{2} = \frac{A(M^2-1)}{M(M-A)} F_{A,M-A,\alpha} \tag{7}$$

Squared Prediction Error (SPE) statistic measures the average size of the residuals corresponding to the lowest $d - a$ eigenvalues

$$SPE = \boldsymbol{x}(\boldsymbol{I} - \boldsymbol{PP^{T}})\boldsymbol{x^{T}} \tag{8}$$

The confidence limit of SPE statistic is defined as

$$SPE_{\alpha} = \theta_1 \left(\frac{h_o c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_o (h_o - 1)}{{\theta_1}^2} \right)^{\frac{1}{h_o}} \tag{9}$$

with

$$\theta_i = \sum_{j=a+1}^{m} \lambda_j^i \tag{10}$$

$$h_o = 1 - \frac{2\theta_1\theta_3}{3{\theta_2}^2} \tag{11}$$

where c_α is the value of normal distribution with α level of significance.

ICA tries to estimate source signals without knowing $\boldsymbol{A}$ and $\boldsymbol{s}$. In order to achieve that, Hyvarinen introduced a fast-fixed point algorithm (FastICA) [11]. In FastICA, Negentropy method is used to measure statistical independency. The negentropy is defined as $\boldsymbol{J(y) = H(y_{gauss}) - H(y)}$, which can be approximated by

$$\boldsymbol{J(y)} = \left[\boldsymbol{E\{G(y)\} - E\{G(y_{gauss})\}}\right]^{\boldsymbol{2}} \tag{12}$$

where $\boldsymbol{G(y)}$ is the nonquadratic function [12]. The FastICA algorithm is stated as follows

1. Choose an initial weight vector $\boldsymbol{w}$ of unit norm.
2. Update $\boldsymbol{w} \leftarrow \boldsymbol{E\{zg(w^{T}z)\} - E\{\acute{g}(w^{T}z)\}w},$ where $\boldsymbol{g(y) = tanh(a_1 y)}$
3. Normalize $\boldsymbol{w} \leftarrow {}^{\boldsymbol{w}}\!/_{\|\boldsymbol{w}\|}$.
4. If not converged, go back to step 2.

Separating matrix $\boldsymbol{W}$ is obtained by assembling all the vector $\boldsymbol{w}$, and the demixing sources are calculated as

$$\hat{\boldsymbol{s}} = \boldsymbol{Wx} \tag{13}$$

After obtaining $\boldsymbol{W}$, it is divided into two parts, dominant part ($\boldsymbol{W_d}$), excluded part ($\boldsymbol{W_e}$), and using these parts three statistics are calculated as follows

$$I^2 = \widehat{\boldsymbol{S}}_{\boldsymbol{d}}^{\boldsymbol{T}}\widehat{\boldsymbol{S}}_{\boldsymbol{d}} = \boldsymbol{x^{T}{W_d}^{T}W_d x} \tag{14}$$

$$I_e^2 = \widehat{\boldsymbol{S}}_{\boldsymbol{e}}^{\boldsymbol{T}}\widehat{\boldsymbol{S}}_{\boldsymbol{e}} = \boldsymbol{x^{T}{W_e}^{T}W_e x} \tag{15}$$

$$SPE = (\boldsymbol{x} - \hat{\boldsymbol{x}})^{\boldsymbol{T}}(\boldsymbol{x} - \hat{\boldsymbol{x}}) \tag{16}$$

where $\hat{\boldsymbol{x}} = \left(\boldsymbol{\Lambda^{-0.5}V^{T}}\right)^{-1}\boldsymbol{B_d W_d x}$. In ICA, the latent variables are assumed to be non-Gaussian distributed, hence the confidence limits of tree statistics are calculated by kernel density estimation method [13].

3. Process Monitoring Based on SVDD

The SVDD is a new method in the FD that constructs a univariate monitoring statistics for variables [14]. The aim of SVDD is to transform original variables into a high dimensional feature space by using kernel. The transformed variables fall into a minimal sphere of radius R [7]:

$$\min_{R,\boldsymbol{a},\varepsilon} R^2 + C\sum_{i=1}^{n} \varepsilon_i \ s.t. \|\phi(\hat{\boldsymbol{s}}_{\boldsymbol{i}}) - \boldsymbol{a}\|^2 \le R^2 + \varepsilon_i \tag{17}$$
$$\varepsilon_i \ge 0 \quad \forall 1 \le i \le n$$

where $\boldsymbol{a}$ is the center of hypersphere, C is the trade off between the volume of the hypersphere and the number of transformed samples outside the hypersphere, and ε_i is slack variable. The $\boldsymbol{a}$ and R are obtained as follows

$$\boldsymbol{a} = \sum_{i=1}^{n} \alpha_i \phi(\hat{\boldsymbol{s}}_{\boldsymbol{i}}) \ , \tag{18}$$

$$R = \sqrt{K(\hat{\boldsymbol{s}}_{\boldsymbol{p}}, \hat{\boldsymbol{s}}_{\boldsymbol{p}}) - 2\sum_{i=1}^{n} \alpha_i K(\hat{\boldsymbol{s}}_{\boldsymbol{p}}, \hat{\boldsymbol{s}}_{\boldsymbol{i}}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K(\hat{\boldsymbol{s}}_{\boldsymbol{i}}, \hat{\boldsymbol{s}}_{\boldsymbol{j}})} \tag{19}$$

where the sample points $\hat{\boldsymbol{s}}_{\boldsymbol{p}}$ is support vector. Using (17), the univariate monitoring statistic is calculated as

$$d^2(\phi(\hat{\boldsymbol{s}}_{\boldsymbol{i}})) = \|\phi(\hat{\boldsymbol{s}}_{\boldsymbol{i}}) - \boldsymbol{a}\|^2 \tag{20}$$

If the square distance $d^2(\phi(\hat{\boldsymbol{s}}_i)) = \|\phi(\hat{\boldsymbol{s}}_i) - \boldsymbol{a}\|^2 \le R^2$, sample is accepted as a normal sample. The decision based on SVDD can be precisely described as

$$d^2(\phi(\hat{\boldsymbol{s}}_i)) = K(\hat{\boldsymbol{s}},\hat{\boldsymbol{s}}) - 2\sum_{i=1}^{n}\alpha_i K(\hat{\boldsymbol{s}}_i,\hat{\boldsymbol{s}}) + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\hat{\boldsymbol{s}}_i,\hat{\boldsymbol{s}}_j) \le R^2 \tag{21}$$

4. Comparision Study Based on TE

4.1. Applications for Determination of Disease Risk

In this section PCA, ICA and SVDD methods will be applied to TE process for a comparison study. TE simulator was developed by Downs and Vogel which produces two products from four reactants [15]. It has five major units, i.e. reactor, condenser, compressor, separator, and stripper. There are 41 measured variables and 12 manipulated variables. In TE process, 20 process faults and an additional valve faults were defined by Downs and Vogel [15]. The sample length of the training data set is 500 under normal operating conditions. Each testing data set for one fault mode consist of 960 samples, and fault was introduced at sample 160 for each data set. All the data were autoscaled prior to application of the algorithms. To obtain better effect of simulation, 52 variables are selected for analysis. The confidence limit of kernel density estimation is selected as 99%. Fault detection rate (FDR) and false alarm rate (FAR) are considered to evaluate FD performance [16,17]. FDR and FAR rates are defined as

$$FDR = \text{No. of samples } (J > J_{th} | f \neq 0) \tag{22}$$

$$FAR = \text{No. of samples } (J > J_{th} | f = 0) \tag{23}$$

If one of the test statistics exceeds threshold, a fault can be detected successfully. The FD performances of the methods are investigated for 11 typical fault modes in TE process, and the results are calculated and tabled in Table I. Also, in Table 1 the fault types and fault modes can be seen. As it can be seen in Table 1, FDRs are almost close each other. For fault 15 all algorithms show poor FD performance. FARs for fault 6 is very low for FastICA and SVDD in Table II.

According to FARs given in Table II, PCA has higher FARs than FastICA and SVDD, which means PCA produces more false alarms than the others. For fault 12 they have almost the same FARs, but for fault 8 only FastICA approach gives the best FAR.

In this paper, PCA, ICA and SVDD methods have been introduced, which was based on linear (PCA, ICA) and nonlinear process monitoring techniques (SVDD). Then, all the methods were implemented on TE process to compare the sensivity of the algorithms quantified by calculating the FARS.

Table 1. Description of process faults, and FDRs of algorithms

Fault Mode	**ALGO-RITHMS**	**Fault Detection Rate (FDR)**				**Fault Type**
		SVDD	**SPE**	**T^2-I**	**Ie**	
1	**PCA**		800	796		Step
	FastICA		797	796	797	
	SVDD	798				
2	**PCA**		798	790		Step
	FastICA		789	747	789	
	SVDD	789				
4	**PCA**		800	675		Step
	FastICA		755	21	724	
	SVDD	781				
6	**PCA**		800	796		Step
	FastICA		800	800	800	
	SVDD	800				
7	**PCA**		800	800		Step
	FastICA		800	326	800	
	SVDD	800				
8	**PCA**		787	785		Random Variation
	FastICA		791	725	795	
	SVDD	796				
12	**PCA**		790	794		Random Variation
	FastICA		800	790	797	
	SVDD	798				
13	**PCA**		772	766		Slow Drift
	FastICA		767	747	765	
	SVDD	769				
14	**PCA**		798	800		Sticking
	FastICA		800	722	800	
	SVDD	800				
15	**PCA**		247	176		Sticking
	FastICA		185	7	180	
	SVDD	242				
18	**PCA**		745	728		Unknown
	FastICA		736	713	735	
	SVDD	742				

Table 2. Description of FARs

Fault Mode	ALGO-RITHMS	False Alarm Rate (FAR)			
		SVDD	SPE	T^2-I	Ie
1	PCA		29	15	
	FastICA		3	0	3
	SVDD	7			
2	PCA		33	11	
	FastICA		2	0	3
	SVDD	4			
4	PCA		31	16	
	FastICA		5	1	3
	SVDD	6			
6	PCA		25	3	
	FastICA		0	0	0
	SVDD	1			
7	PCA		26	5	
	FastICA		0	0	2
	SVDD	6			
8	PCA		33	11	
	FastICA		5	2	10
	SVDD	27			
12	PCA		32	12	
	FastICA		29	0	30
	SVDD	36			
13	PCA		30	7	
	FastICA		3	0	3
	SVDD	3			
14	PCA		34	15	
	FastICA		3	1	2
	SVDD	7			
15	PCA		29	10	
	FastICA		1	0	0
	SVDD	3			
18	PCA		39	13	
	FastICA		1	0	1
	SVDD	8			

5. Conclusion

According to Table I, all the tested methods give similar FDRs but different FARs (Table II). However FARs of ICA and SVDD have almost the same, and are lower than PCA. Therefore, ICA and SVDD algorithms can be preferred to PCA to obtain lower FARs, and may provide the process operator with more reliable information. The drawback of the ICA algorithm is that using the kernel density estimation method is computationally expensive. Unlike the ICA, SVDD does not suffer from a high computational load, so using SVDD is more appropriate for FD applications.

References

[1] Shams M. B., Budman H. M., and Duever T. A., Fault detection, identification and diagnosis using CUSUM based PCA, *Chemical Engineering Science*, *66*(20), 4488-4498, 2011.

[2] Villegas T., Fuente M. J., and Rodríguez M., Principal component analysis for fault detection and diagnosis. experience with a pilot plant, In *CIMMACS'10 Proceedings of the 9th WSEAS international conference on computational intelligence, man-machine systems and cybernetics* , 2010, December, pp. 147-152.

[3] Alkaya A., and Eker İ., Variance sensitive adaptive threshold-based PCA method for fault detection with experimental application, *ISA transactions*, *50*(2), 287-302, 2011.

[4] Kano M., Tanaka S., Hasebe S., Hashimoto I., and Ohno H., Monitoring independent components for fault detection, *AIChE Journal*, *49*(4), 969-976, 2003.

[5] Yoo C. K., Lee J. M., Vanrolleghem P. A., and Lee I. B., On-line monitoring of batch processes using multiway independent component analysis, *Chemometrics and Intelligent Laboratory Systems*, *71*(2), 151-163, 2004.

[6] Lee J. M., Yoo C., and Lee I. B., Statistical monitoring of dynamic processes based on dynamic independent component analysis, *Chemical engineering science*, *59*(14), 2995-3006, 2004.

[7] Tax D. M., and Duin R. P., Support vector data description, *Machine learning*, *54*(1), 45-66, 2004.

[8] Ge Z., Xie L., Kruger U. , Lamont L., Song Z., and Wang S., Sensor fault identification and isolation for multivariate non-Gaussian processes, *Journal of Process Control*, *19*(10), 1707-1715, 2009.

[9] Xie L., and Kruger U., Statistical processes monitoring based on improved ICA and SVDD, In *International Conference on Intelligent Computing*, 2006, August, pp. 1247-1256, Springer Berlin Heidelberg.

[10] Wang D., Peter W. T., Guo W., and Miao Q., Support vector data description for fusion of multiple health indicators for enhancing gearbox fault diagnosis and prognosis, *Measurement Science and Technology*, *22*(2), 025102, 2010.

[11] Hyvärinen A., and Oja E., A fast fixed-point algorithm for independent component analysis, *Neural computation*, *9*(7), 1483-1492, 1997.

[12] Hyvärinen A., and Oja E., Independent component analysis: algorithms and applications, *Neural networks*, *13*(4), 411-430, 2000.

[13] Silverman B. W., *Density estimation for statistics and data analysis* (Vol. 26). CRC press, 1986.

[14] Liu X., Xie L., Kruger U., Littler T., and Wang S., Statistical-based monitoring of multivariate non-Gaussian systems, *AIChE journal*, *54*(9), 2379-2391, 2008.

[15] Downs J. J., and Vogel E. F., A plant-wide industrial process control problem, *Computers & chemical engineering*, *17*(3), 245-255, 1993.

[16] Chiang L. H., Braatz R. D., and , Russell E. L., *Fault detection and diagnosis in industrial systems*, Springer Science & Business Media, 2001.

[17] Lee J. M., Qin S. J., and Lee I. B., Fault detection and diagnosis based on modified independent component analysis, *AIChE journal*, *52*(10), 3501-3514, 2006.