J Electr Eng Technol.2016; 11(4): 848-859
http://dx.doi.org/10.5370/JEET.2016.11.4.848

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

A Clustering-Based Fault Detection Method for Steam Boiler Tube in Thermal Power Plant

Jungwon Yu*, Jaeyel Jang, Jaeyeong Yoo***, June Ho Park* and Sungshin Kim†**

Abstract – System failures in thermal power plants (TPPs) can lead to serious losses because the equipment is operated under very high pressure and temperature. Therefore, it is indispensable for alarm systems to inform field workers in advance of any abnormal operating conditions in the equipment. In this paper, we propose a clustering-based fault detection method for steam boiler tubes in TPPs. For data clustering, *k*-means algorithm is employed and the number of clusters are systematically determined by slope statistic. In the clustering-based method, it is assumed that normal data samples are close to the centers of clusters and those of abnormal are far from the centers. After partitioning training samples collected from normal target systems, fault scores (FSs) are assigned to unseen samples according to the distances between the samples and their closest cluster centroids. Alarm signals are generated if the FSs exceed predefined threshold values. The validity of exponentially weighted moving average to reduce false alarms is also investigated. To verify the performance, the proposed method is applied to failure cases due to boiler tube leakage. The experiment results show that the proposed method can detect the abnormal conditions of the target system successfully.

Keywords: Thermal power plant, Boiler tube leakage, Fault detection, *k*-means clustering, Slope statistic

1. Introduction

The importance of condition monitoring and fault detection (FD) techniques has been growing for effective operation and performance improvement of various industrial processes such as aircraft, train, automobile, chemical factory and power plant. Fault is defined as an unpermitted deviation of at least one characteristic property or variable of a system from acceptable/usual/standard behavior [1]. Fault can result in system malfunctions and failure. In particular, system failures in thermal power plant (TPP) equipment with very high operating pressure and temperature can cause severe loss of life and materials. Monitoring and FD systems that can detect in advance the abnormal conditions of power plant units are essential for ensuring the safety, reliability and availability of power plants. The main objective of FD systems is to detect the abnormal operation conditions of power plants by analyzing complex and nonstationary behaviors of process parameters, and help field workers execute proper actions at the initiatory stage of faults.

Recently, as distributed control systems (DCSs) are built in power plants, massive operation data can be collected and managed efficiently. In DCS, historical operation data composed of various process variables is stored in discrete time intervals. The explosive growth of historical data has boosted efforts to extract useful knowledge from the data related to equipment health and maintenance information.

As described in Fig. 1, process monitoring procedures are basically performed in four steps [2]: FD, fault identification, fault diagnosis and system recovery. FD determines whether a fault has occurred. Fault identification confirms process variables in connection with the fault. Fault diagnosis identifies the type of fault. Finally, after removing the cause of the fault, the monitoring loop is closed. In this paper, the focus is on FD only.

The following summarizes several previous studies on condition monitoring and FD methods for TPPs using data mining techniques. Ajami and Daneshvar [2] used multivariate statistical signal processing techniques, such as principal component analysis and independent component analysis (ICA), for FD and diagnosis of TPP turbine systems. Hsu and Su [3] developed a method that combines

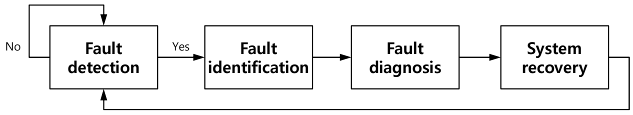


Fig. 1. General monitoring scheme for industrial processes [2]

† Corresponding Author: Dept. of Electrical and Computer Engineering, Pusan National University, Busan, Korea. (sskim@pusan.ac.kr)

* Dept. of Electrical and Computer Engineering, Pusan National University, Busan, Korea. ({garden0312, parkjh}@pusan.ac.kr)

** Technology & Information Department, Technical Solution Center, Korea East-West Power Co., Ltd., Dangjin, Korea. (zanggre250@ewp.co.kr)

*** CTO, XEONET Co., Ltd, Seongnam, Korea (joeyoo@xeonet.co.kr)

Received: March 23, 2016; Accepted: April 26, 2016

ICA and exponentially weighted moving average (EWMA) for early detection of TPP malfunctions at Taiwan Power Company. Cai et al. [4] introduced an on-line performance monitoring method for coal-fired power units, such as boilers and turbines, using support vector machine (SVM). Chen et al. [5] proposed a SVM-based method with dimension reduction schemes based on correlation analysis and decision tree to analyze turbine failures in thermal power facilities. Shashoa et al. [6] presented a data-driven FD and isolation approach for a steam separator at TEKO B1 Kostolac TPP using robust process identification procedures and Neyman-Pearson hypothesis test. Li et al. [7] presented a monitoring and fault diagnosis method for leak detection of feedwater heaters in coal-fired plants using group method of data handling based on ridge regression. Prasad et al. [8] proposed a performance monitoring strategy based on neural network and histogram plots to economize the operation of a 200 MW oil/gas-fired TPP. Guo et al. [9] reported a condition monitoring and FD method for tube-ball mills of coal-fired plants using a multi-segment mathematical model whose parameters are identified by genetic algorithms.

Although various statistical and machine learning techniques have been applied for condition monitoring and FD of TPP components, what seems to be lacking is attempts to use clustering-based FD methods for TPPs. In this paper, we propose a clustering-based FD method for tube leakage of steam boilers in TPPs. In classification-based approaches (e.g., SVM), labeled learning samples should be prepared to train binary classifiers. To prepare the labeled samples, experts determine whether arbitrary samples are normal or fault after checking historical operation data. When there are many monitored variables, this labeling procedure is a difficult and time-consuming process. On the other hand, clustering-based methods can find the hidden structure of unlabeled learning samples, and perform FD in unsupervised mode. Clustering-based methods that do not need pre-labeled samples have been widely applied to engineering fields, such as financial domain [10], network intrusion detection [11, 12], anomaly detection in surveillance videos [13] and steel industry [14]. In the proposed method, it is assumed that normal samples are close to cluster centroids and abnormal samples are far from the centroids. For data clustering, *k*-means algorithm with Euclidean distance is employed and slope statistic [15] proposed by Fujita is used to systematically determine the number of clusters. Slope statistic can handle situations when there is a dominant cluster in training samples, when the samples are not a mixture of Gaussian distributions, and when the dimensions of the samples are high.

After partitioning training samples gathered from normal target systems into several groups, fault scores (FSs) are assigned to unseen samples based on the distances between the samples and their closest centers. Using 95th, 97th and 99th percentiles, threshold values of FSs are calculated and alarm signals occur when the FSs of unseen samples are larger than the threshold values. The validity of EWMA for reducing false alarms is also investigated. In order to evaluate the performance, the proposed method is applied to collected dataset from the DCS of 200 MW TPP. The dataset corresponds to two failure cases due to boiler tube leakage. The simulation results show that the proposed method can detect the tube leakage in the early stages.

The remainder of this paper is organized as follows. Section 2 briefly summarizes *k*-means clustering algorithm and silhouette and slope statistics. In Section 3, FS assignments, their threshold settings and EWMA that consider the trends of FSs are explained. Section 4 describes the target system, a 200 MW coal-fired TPP, and the tube leakage in steam boiler. Section 5 shows the simulation results of the two failure cases and finally, Section 6 presents concluding remarks.

2. Data Clustering Algorithm

Data clustering techniques classify similar training samples into several groups or clusters and can find the hidden structure of unlabeled training samples.

2.1 *k*-Means Clustering Algorithm

The *k*-means algorithm [16, 17] partitions n given vectors $\mathbf{x}_j$, $j = 1,..., n$, into c groups (also called as clusters) G_i, $i = 1,..., c$, and finds cluster centers $\mathbf{c}_i$, $i = 1,..., c$, that minimize the objective function defined as follows:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\left(\sum_{k,\mathbf{x}_k \in G_i} \left\| \mathbf{x}_k - \mathbf{c}_i \right\|^2 \right), \tag{1}$$

where $\|\cdot\|$ denotes Euclidean distance and J_i is an objective function value of the ith cluster, which depends on its geometrical data structure and center position. The partitioned samples are described by c by n binary membership matrix U whose ith row and jth column, u_{ij}, is 1 if the jth sample, $\mathbf{x}_j$, belongs to the ith cluster, and 0 otherwise. Matrix U satisfies the following properties:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad \forall j = 1,...,n \tag{2}$$

and

$$\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} = n. \tag{3}$$

After cluster centers $\mathbf{c}_i$, $i = 1,..., c$, are fixed, u_{ij} is defined as

$$u_{ij} = \begin{cases} 1 & \text{if } \left\| \mathbf{x}_j - \mathbf{c}_i \right\|^2 \le \left\| \mathbf{x}_j - \mathbf{c}_k \right\|^2, \text{ for each } k \ne i, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

In other words, if the *i*th center is the closest center of the *j*th sample, the latter is included in the *i*th group. After determining u_{ij}, optimal centers $\mathbf{c}_i$ that minimize the objective function are calculated as

$$\mathbf{c}_i = \frac{1}{|G_i|} \sum_{k, \mathbf{x}_k \in G_i} \mathbf{x}_k, \tag{5}$$

where |·| denotes the size of a set. As explained above, in *k*-means algorithm, cluster centers $\mathbf{c}_i$ and membership matrix U are determined through iterative procedures (see [16] for more on this).

2.2 Silhouette statistic

The concept of silhouette [18] proposed by Rousseeuw is a useful tool for verifying how well the training samples are grouped. The silhouette plot not only provides validity to the clustering results but also outlines the target data structure. Silhouette statistic, an averaged value of each sample's silhouette value, can be employed to determine the proper number of clusters.

Suppose that n training samples $\mathbf{x}_j$, $j = 1,..., n$, are grouped into c clusters G_i, $i = 1,..., c$, and the *j*th sample belongs to the *i*th cluster. In order to calculate silhouette value $s(j)$ for the *j*th sample, let us define average dissimilarity $a(j)$ (i.e., inner dissimilarity) between the *j*th sample and all elements of the *i*th cluster, with the exception of the *j*th sample, as

$$a(j) = \frac{1}{|G_i|-1} \sum_{\mathbf{x} \in G_i} \|\mathbf{x}_j - \mathbf{x}\|^2, \tag{6}$$

where $\mathbf{x} \neq \mathbf{x}_j$. The average dissimilarity between the *j*th sample and all elements of the *k*th cluster, with the exception of the *i*th cluster, $d(\mathbf{x}_j, G_k)$, for $k = 1,..., c$, $k \neq i$, is also defined as

$$d(\mathbf{x}_j, G_k) = \frac{1}{|G_k|} \sum_{\mathbf{x} \in G_k} \|\mathbf{x}_j - \mathbf{x}\|^2, \quad \text{for each } k \neq i. \tag{7}$$

After calculating $d(\mathbf{x}_j, G_k)$, their minimum value (i.e., inter dissimilarity) is denoted by

$$b(j) = \min_k d(\mathbf{x}_j, G_k), \quad \text{for each } k \neq i, \tag{8}$$

where G_k, whose $d(\mathbf{x}_j, G_k)$ is minimum, is called as the second-best choice cluster. The silhouette value $s(j)$ of the *j*th sample is calculated as

$$s(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}}. \tag{9}$$

Silhouette values $s(j)$, $j = 1,..., n$, are in the range of [−1, 1] and they can be combined into a silhouette plot that graphically represents clustering results. Let us exemplify three extreme situations to deeply understand the meaning of a silhouette value. The first case is that where $s(j)$ is close to 1. This implies that inner dissimilarity is much smaller than inter dissimilarity, i.e., $a(j) \approx b(j)$. In this case, we can conclude that the *j*th sample is included in a proper cluster. In the second situation, $s(j)$ is approximately 0, i.e., $a(j) \approx b(j)$, and thus it is uncertain whether the *j*th sample should be assigned to the *i*th or second-best choice cluster. Lastly, the third case is the worst case, where $s(j)$ is close to −1. In this situation, it is valid to not classify the *j*th sample into the *i*th cluster but the second-best choice cluster.

The silhouette statistic used to determine the proper number of clusters is defined as

$$\bar{s}(c) = \frac{1}{n} \sum_{j=1}^{n} s(j), \quad c = 2, 3, ..., \tag{10}$$

After calculating the silhouette statistics for c = 2, 3,..., the number of clusters that maximizes them is finally selected.

2.3 Slope statistic

Slope statistic [15], proposed by Fujita et al., is based on the silhouette statistic described in the previous subsection. The basic idea of slope statistic is that the optimal cluster number has the maximum silhouette statistic and if the cluster number is larger than the optimum, the silhouette statistic decreases sharply. Based on this idea, slope statistic is defined as

$$\hat{s}(c) = -[\bar{s}(c+1) - \bar{s}(c)]\bar{s}(c)^p, \tag{11}$$

where p is a positive constant and controls the weight size for the two terms, $\bar{s}(c)$ and $\bar{s}(c+1) - \bar{s}(c)$. The reason for employing the silhouette approach in the construction of slope statistic is that the former considers both the inner and inter dissimilarity of each target sample. Using slope statistic, the proper number of clusters is determined as

$$c^* = \arg\max_c -[\bar{s}(c+1) - \bar{s}(c)]\bar{s}(c)^p, \quad c = 2, 3, ... \tag{12}$$

In this paper, the cluster number that maximizes slope statistic is determined for *k*-means clustering.

3. Clustering-Based Fault Detection

In the clustering-based FD method, after applying *k*-means algorithm to normal samples, unseen samples that do not match with the normal samples are regarded as fault samples. The advantages of clustering-based techniques are that FD can be performed in unsupervised mode and the computation time in the test phase is fairly short [19].

3.1 Fault score

In order to detect fault samples, FSs are assigned to unseen samples according to the distances between the samples and their closest cluster centers [20]. The following describes the procedures for imposing FS on a new sample $\mathbf{x}_{\text{new}}$. First, among c cluster centers, the nearest center $\mathbf{c}_k$ to $\mathbf{x}_{\text{new}}$ is found by

$$\mathbf{c}_k = \arg\min_{\mathbf{c}_i} \left\| \mathbf{c}_i - \mathbf{x}_{\text{new}} \right\|, \quad i = 1, ..., c. \tag{13}$$

Subsequently, the average distance l_k between $\mathbf{c}_k$ and the training samples included in G_k is calculated as

$$l_k = \frac{1}{\left| G_k \right|} \sum_{\mathbf{x} \in G_k} \left\| \mathbf{c}_k - \mathbf{x} \right\|. \tag{14}$$

Finally, the FS of $\mathbf{x}_{\text{new}}$ is defined as

$$FS_{\text{new}} = \frac{\left\| \mathbf{c}_k - \mathbf{x}_{\text{new}} \right\|}{l_k}. \tag{15}$$

FS measures the ratio of the dissimilarity between $\mathbf{x}_{\text{new}}$ and its nearest center $\mathbf{c}_k$ to the average distance l_k. The larger FS, the farther $\mathbf{x}_{\text{new}}$ is from its closest center. As explained in the next subsection, if FS exceeds the predefined threshold values, the corresponding samples are determined as fault samples and alarm signals are generated.

3.2 Threshold setup for fault score

This subsection provides the procedures to set up threshold value T for alarm signal generation. First, as presented in (14), mean distances l_k, k = 1,..., c, are calculated. Subsequently, FS is imposed on each training sample using (15). In other words, FSs, FS_j, j = 1,..., n, of the training samples are calculated by substituting the jth training sample $\mathbf{x}_j$ into (15) instead of $\mathbf{x}_{\text{new}}$. Finally, threshold values of FS are determined using FS_j.

In clustering-based FD, only the upper threshold value is considered because the possibility of fault increases when FS is large. Upper threshold $T = m + \zeta\sigma$ is generally used when FS_j follow Gaussian distribution, where m and σ are the mean and standard deviation of FS_j, respectively, and ζ is a positive integer [1]. The probability that FS of an arbitrary sample exceeds the upper threshold is equal to 0.07933, 0.011375 and 0.000675 for ζ = 1, 2, and 3, respectively. As shown in Section 5, FS_j follow a distribution where the right tail is longer than the left. In this paper, the 95th, 97th and 99th percentiles of FS_j are employed as threshold values for FD.

3.3 Exponentially weighted moving average

In the test phase, if an alarm signal is generated based only on the current FS, it is assumed that the current and previous FS are independent. In this case, the false alarm rate could increase because alarm signals occur regardless of the historical trend. In this paper, EWMA, which is widely used to smooth a time-series data, is employed to consider the trend of FSs. EWMA gives more weights to latest time-series and these weights decrease exponentially for older data. Using EWMA, the smoothed version of FSs at time t, $EWMA_w(t)$, is calculated as

$$\begin{aligned} &EWMA_w(1) = FS(1), \\ &EWMA_w(t) = \alpha FS(t) \\ &\qquad + (1-\alpha) EWMA_w(t-1), \quad \text{for } t > 1, \end{aligned} \tag{16}$$

where $FS(t)$ is FS at time t, α is a smoothing factor commonly calculated by $\alpha = \frac{2}{w+1}$ and w is window size. After calculating EWMA, an alarm signal is generated if it exceeds predefined threshold values. Alarm signal generation using EWMA could reduce false alarms because historical trends of FSs are considered.

3.4 Overview of the proposed approach

Table 1 summarizes the procedure for the proposed clustering-based FD approach. The procedure is divided into “training” and “test” phases. In the “training” phase, after determining the proper number of clusters, the training samples collected from the normal target system are partitioned using k-means algorithm. Subsequently, the FS of each training sample is calculated and three threshold

Table 1. Clustering-based FD algorithm

	Input: Multivariate training and test samples, X_{trn} and X_{test} $C_{\max}$ ← maximum number of clusters w ← window size of EWMA
“Training” phase	**for** c from 2 to $C_{\max}$ Partition X_{trn} into c groups using k-means algorithm Calculate the silhouette statistic using (10) **end** Calculate the slope statistic using (11) Determine the optimal number of cluster, c^*, using (12) Partition X_{trn} into c^* groups using k-means algorithm For each cluster, calculate the average distance, l_k, between $\mathbf{c}_k$ and $\mathbf{x} \in G_k$ using (14) Calculate FS_j, j = 1,..., $\lvert X_{\text{trn}} \rvert$, for each training samples using (15) Set the threshold values, $T_{95\text{th}}$, $T_{97\text{th}}$ and $T_{99\text{th}}$ using percentiles of FS_j

“Test” phase	**for** t from 1 to $\lvert X_{\text{test}} \rvert$
	    Find the closest center of tth test sample
	    Calculate fault score $FS(t)$ using (15)
	    Calculate $EWMA_w(t)$ using (16)
	    if $FS(t) \geq T_{95\text{th}}$ (or $EWMA_w(t) \geq T_{95\text{th}}$)
	        Generate ‘Caution’ alarm
	    else if $FS(t) \geq T_{97\text{th}}$ (or $EWMA_w(t) \geq T_{97\text{th}}$)
	        Generate ‘Alert’ alarm
	    else if $FS(t) \geq T_{99\text{th}}$ (or $EWMA_w(t) \geq T_{99\text{th}}$)
	        Generate ‘Critical’ alarm
	    end
	end

values (i.e., 95th, 97th and 99th percentiles) are set up. Depending on the strength of FS, three different alarm signals (i.e., “Caution”, “Alert” and “Critical”) are generated. In the “test” phase, after assigning FS to each test sample, alarm signals occur sequentially. If EWMA is not used, an alarm signal for each test sample is generated independently according to the strength of FS. If EWMA is employed, after calculating $EWMA_w(t)$ using (16), three different alarms are generated.

4. Description of Target System: 200 MW Coal-Fired Power Plant

The target system of this study is a 200 MW coal-fired TPP. Fig. 2 shows an example of the DCS screen in the plant. To verify the performance, the proposed method is applied to two failure cases collected from the target DCS.

4.1 Coal-fired thermal power plant

In the coal-fired power plant, after transforming feedwater into steam through the thermal energy produced from the combustion of bituminous coal, electricity is generated by driving the steam turbine and generator. Fig. 3 shows a simplified schematic diagram of the target TPP. The steam boiler raises steam by heating feedwater using thermal energy converted from fossil fuel. The steam boiler follows the thermodynamic steam cycle, i.e., Rankine cycle, which is a practical implementation of the ideal Carnot cycle [21]. Steam, an important medium for producing mechanical energy, can be generated from abundant water, does not react much with the materials of the power plant equipment and is stable at the required operation temperature in the power plant [22].

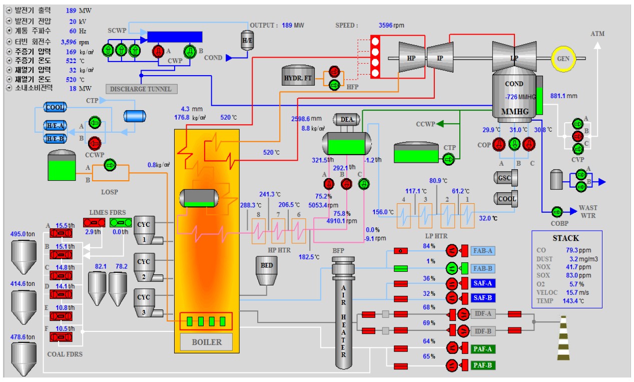
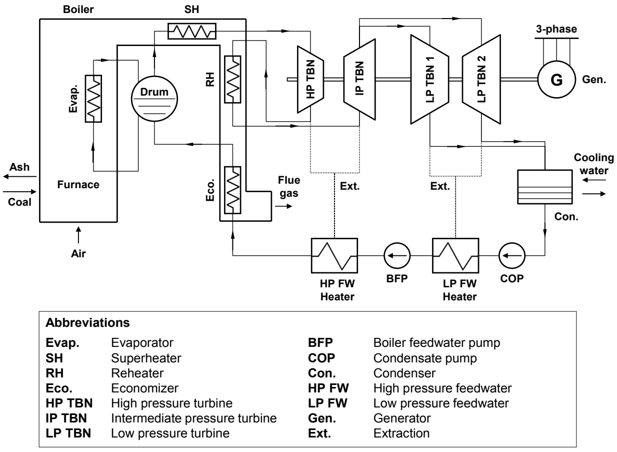


Fig. 2. Example of DCS screen in 200 MW coal-fired power plant



Abbreviations

Evap.	Evaporator	BFP	Boiler feedwater pump
SH	Superheater	COP	Condensate pump
RH	Reheater	Con.	Condenser
Eco.	Economizer	HP FW	High pressure feedwater
HP TBN	High pressure turbine	LP FW	Low pressure feedwater
IP TBN	Intermediate pressure turbine	Gen.	Generator
LP TBN	Low pressure turbine	Ext.	Extraction

Fig. 3. Simplified schematic diagram for target TPP

Bituminous coal pulverized in advance is transformed into thermal energy at the steam boiler furnace. Before flowing into the drum, feedwater is preheated by passing through a series of low- and high-pressure heaters and economizer. The heater and economizer raise feedwater by extraction steam from the turbine and high-temperature flue gas, respectively. These preheating steps improve the efficiency of the entire cycle. The drum supplies feedwater that will be converted to steam and temporarily stores the steam produced by the evaporator. The saturated steam by evaporator contains a small amount of moisture. A superheater converts the steam into the high-purity and high-pressure and temperature superheated steam that will be supplied to the turbine.

In the turbine, the superheated steam expands, turbine blades are rotated and thermal energy is transformed into mechanical energy. The rotating turbine blades drive the electric generator and three-phase electric power is generated. After performing mechanical works at the high-pressure turbine, the steam is reheated by a reheater and supplied to the intermediate-pressure turbine. The steam that exits from the low-pressure turbine is condensed into condensate water and stored at a condenser’s hotwell. The condensate water is boosted by a condensate pump and it passes through a low-pressure feedwater heater. Subsequently, the water is deaerated by a deaerator and boosted by the feedwater pump. The boosted water passes through a high-pressure heater and economizer and it is fed into the boiler again.

4.2 Boiler tube leakage

Failure from one or more tubes in the boiler can be detected by sound and either by an increase in the make-up water requirement (indicating a failure of the water-carrying tubes) or by an increased draft in the superheater or reheater areas (due to failure of the superheater or reheater tubes) [23]. The boiler tubes can be influenced by several damage processes such as inside scaling, waterside corrosion and cracking, fireside corrosion and/or erosion, stress rupture due to overheat and creep, vibration-induced and thermal fatigue cracking, and defective welds [24].

Tube leakage from a pin-hole might be tolerated because of an adequate margin of feedwater and the leakage can be corrected after suitable scheduled maintenance. However, if the boiler is continuously operated with the leakage, much of the pressurized fluid will eventually leak and cause severe damage to neighboring tubes. Tube leakage of boiler, superheater and reheater could result in a serious efficiency decline. In the short term, tube leakage of superheater and reheater is more serious than that of boiler. When severe tube leakage occurs, maintaining the boiler drum level properly is difficult. If leaking water is spilled onto the furnace, coal combustion is disturbed. In these cases, the plant should be shut down immediately.

In this paper, two unplanned shutdown cases due to boiler tube leakage are employed to demonstrate the validity of the proposed method.

5. Experiment Results

This section provides the results of applying the proposed clustering-based FD approach to the two unscheduled shutdown cases.

5.1 Data preparation

Table 2 lists the number of training and test samples and number of monitored variables for the two failure cases. In Table 2, each sample is recoded in 5-minute intervals and the training samples are gathered from a normally operating target system. After applying the “training” phase in Table 1 to the training samples, the performance of the proposed approach is evaluated by the test samples. Among hundreds of variables, 13 monitored variables are selected based on expert knowledge to detect boiler tube leakage at an early stage. The same variables are selected in Cases 1 and 2. Table 3 summarizes the 13 monitored variables selected by human experts.

Table 2. Summary from two unplanned shutdown cases due to boiler tube leakage

	No. of training samples	No. of test samples	No. of monitored variables
Case 1	2880	1165	13 continuous variables
Case 2	4320	1741	13 continuous variables

Table 3. Summary of monitored variables for boiler tube leakage in 200 MW TPP

Notation	Tag ID	Description	Unit
X_1	12DH-MW	Generator output	MW
X_2	2UL 10DP001 DXJ51	Steam flow	t/h
X_3	2MS PT 2 4 CXJ51	Main steam pressure	kg/cm^2
X_4	2MS 10EU001 DXJ51	Main steam temperature	°C
X_5	2CR PT 01A CXQ50	Reheater pressure	kg/cm^2
X_6	2RH 10DT001 DXJ51	Reheater temperature	°C
X_7	2FG 10DP001 DXJ51	Furnace pressure	kg/cm^2
X_8	2FW 10DL001 DXJ51	Drum level	m
X_9	2CM PT 04 CXQ50	Condenser pressure	kg/cm^2
X_{10}	2CM FT 01 CXQ50	Condenser make-up flow	t/h
X_{11}	2FW 10DF001 DXJ52	Feedwater flow	t/h
X_{12}	2FC 01DS001 DXJ51	Fuel supply	t/h
X_{13}	2AM 10DS001 DXJ52	Air supply	m^3/h

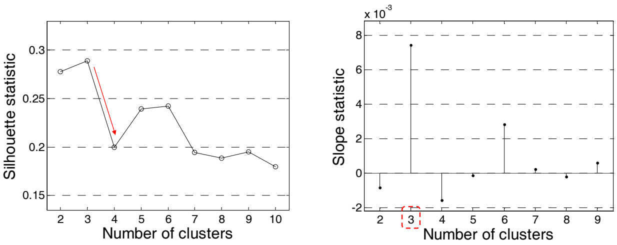
Before performing data clustering, *z*-score standardization is applied to each variable as

$$X_i^* = \frac{X - E[X_i]}{STD[X_i]}, \quad i = 1,\ldots,13, \tag{17}$$

where X and X^* are the original and standardized values, respectively, and $E[\cdot]$ and $STD[\cdot]$ are the expectation and standard deviation operators, respectively. After standardization, the mean and variance of each variable are equal to 0 and 1, respectively. One of the reasons for applying standardization is that the values of the mean and variance of each variable are different from each other.

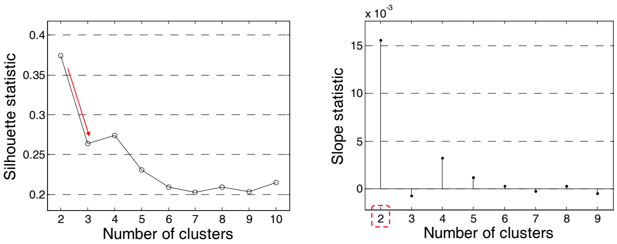
5.2 Determining proper number of clusters

In this study, as described in subsection 2.3, slope statistic is employed to determine the proper number of clusters. After partitioning the training samples and calculating the silhouette statistic with an increase in the number of clusters from $c = 2,\ldots, C_{\max}$, the slope statistic is computed, where $C_{\max}$ and positive constant p in (11) are set to 10 and 2, respectively. Figs. 4 and 5 show the plot of the number of clusters versus the silhouette and slope statistics in Cases 1 and 2, respectively. As shown in Figs. 4 and 5, the numbers of clusters that maximize the slope statistic in Cases 1 and 2 are 3 and 2, respectively. We can confirm that the slope statistic drops sharply in Cases 1 and 2 when the numbers of clusters increase from 3 to 4 and 2 to 3, respectively. In Cases 1 and 2, the training samples



(a) (b)

Fig. 4. Silhouette and slope statistics for Case 1: (a) Silhouette statistic; (b) Slope statistic



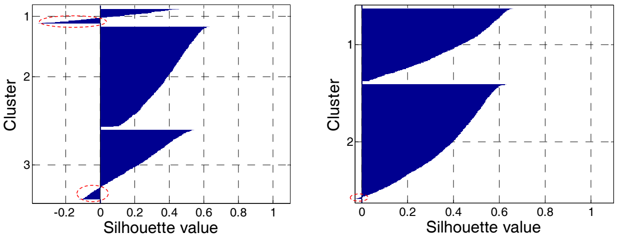
(a) (b)

Fig. 5. Silhouette and slope statistics for Case 2: (a) Silhouette statistic; (b) Slope statistic

are respectively grouped into 3 and 2 clusters using *k*-means algorithm. Fig. 6 shows the silhouette plots for the results of clustering in Cases 1 and 2. In Fig. 6, the samples whose silhouette values are negative are indicated by dotted red lines. It is appropriate for these samples to be classified into second-best choice clusters. Fig. 7 shows the results of applying *k*-means clustering to the training samples of Cases 1 and 2 in a three-dimensional space (i.e., X_5, X_{10} and X_{11}).

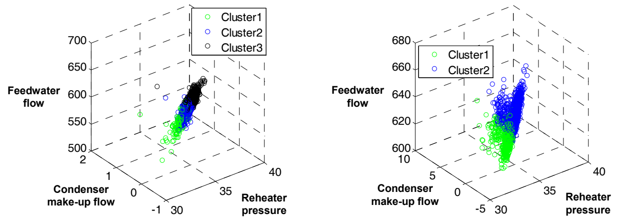
5.3 Threshold values of fault score

As described in subsection 3.2, after performing *k*-means clustering, the FS of each training sample is calculated and the threshold values are also determined. The distribution of the FSs is asymmetric, where the right tail is longer than the left. In this paper, 95th, 97th and 99th percentiles are employed for setting the threshold values. Fig. 8 shows the histograms of FSs of the training samples and their percentiles in Cases 1 and 2. In Fig. 8, solid red lines indicate nonparametric kernel smoothing of the histograms of FSs and vertical dotted yellow-green, blue and red lines correspond to the 95th, 97th and 99th percentiles of FSs, respectively. In Cases 1 and 2, the calculated 95th, 97th and 99th percentiles are 1.6986, 1.8499 and 2.2570, and 1.6499, 1.8821 and 2.2766, respectively. The reason for calculating three different percentiles is to generate diverse alarms based on the strength of FSs. For example, for an unseen sample, if its FS exceeds the 95th, 97th or 99th percentile, the “Caution”, “Alert” or “Critical” alarm occurs.



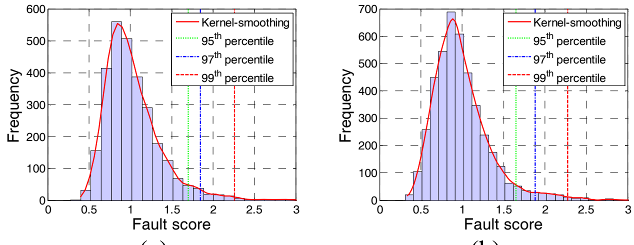
(a) (b)

Fig. 6. Silhouette plots: (a) Case 1; (b) Case 2



(a) (b)

Fig. 7. Results of applying *k*-means algorithm to training samples of: (a) Case 1; (b) Case 2



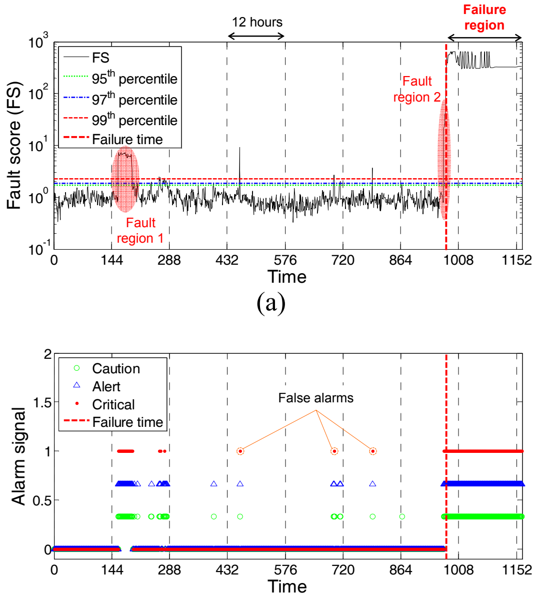
(a) (b)

Fig. 8. Histograms of FS and threshold values: (a) Case 1; (b) Case 2

5.4 Results of fault detection

In Case 1, EWMA is not employed because of its low false alarm rate. As explained in the previous subsection, after setting the threshold values, the FSs of unseen samples are calculated and alarm signals are generated when FSs exceed the threshold values. FS of a normal sample do not exceed threshold values. Fig. 9 shows the FSs of test samples in Case 1 and their alarm signals. In Fig. 9, unscheduled shutdown time due to boiler tube leakage is indicated by vertical solid dotted red lines. The horizontal dotted yellow-green, blue and red lines shown in Fig. 9 (a) represent the 95th, 97th and 99th percentiles, respectively, and “Caution”, “Alert” and “Critical” alarms are indicated by yellow-green circles, blue triangles and red points, respectively. In Fig. 9 (b), there are several improbable false alarms ignored in the real DCS. Fault regions where alarm signals occur intensively are indicated by shaded red regions and enlargements of such regions and their neighborhood are presented in Figs. 10 and 11.

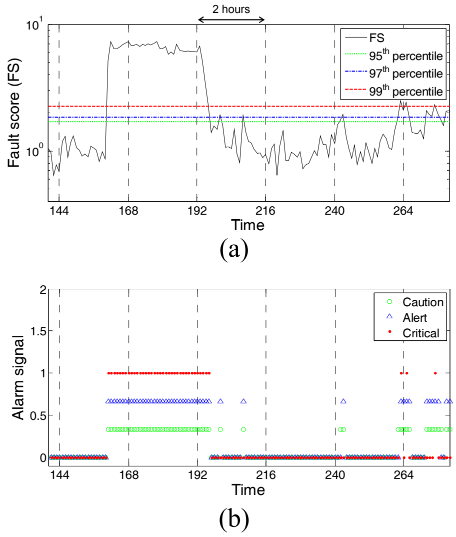
As shown in Figs. 9 (a) and 10, for a period that lasts approximately 3 hours, alarm signals occur intensively approximately 66 hours before the unplanned shutdown



(a)

(b)

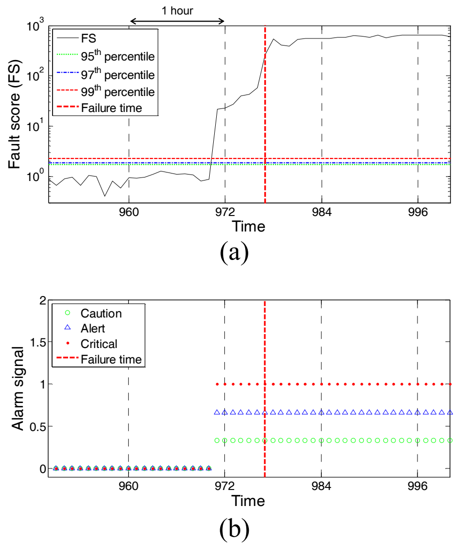
Fig. 9. FD results for Case 1: (a) FSs (b) alarm signals



(a)

(b)

Fig. 10. Enlargement of "fault region 1" in Fig. 9: (a) FSs; (b) alarm signals



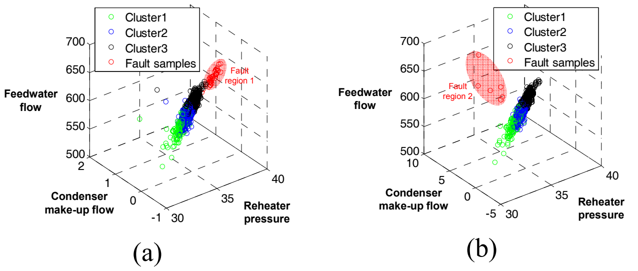
(a)

(b)

Fig. 11. Enlargement of "fault region 2" in Fig. 9: (a) FSs; (b) alarm signals

due to tube leakage. In Fig. 11, because of the dramatic increases of FSs, alarm signals are generated for 30 minutes immediately before the unscheduled shutdown. Fig. 12 shows fault samples that correspond to "Critical" alarms of the fault regions 1 and 2 in a three-dimensional space (i.e., X_5, X_{10} and X_{11}). As shown in Fig. 12, the behavior of the fault samples with "Critical" alarms is extremely inconsistent with that of the normal samples. In the fault region 1, intensive "Critical" alarms occur because of abnormal patterns of the reheater pressure and feedwater flow. The fault region 2 is an early warning region where the condenser make-up flow increases enormously.

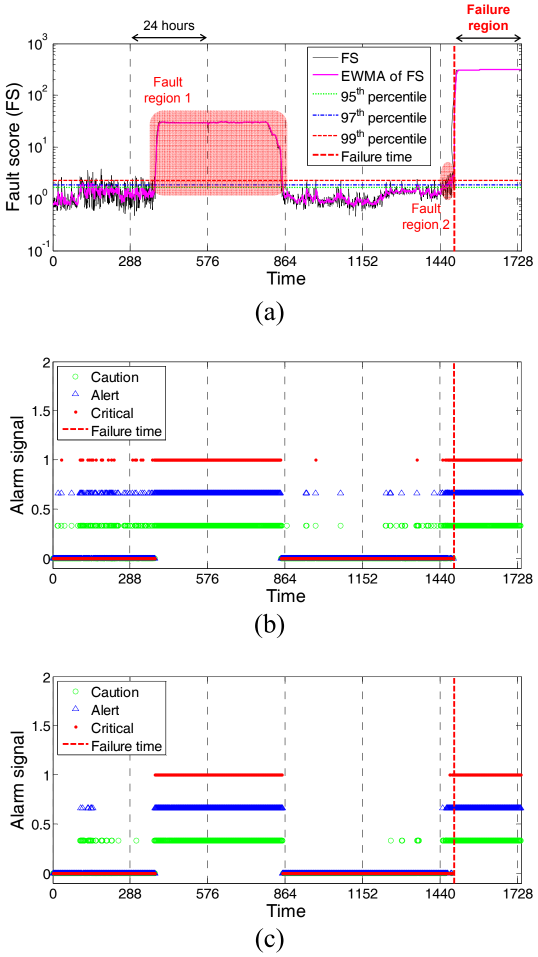
In Case 2, EWMA is employed for alarm signal generation. In EWMA, window size w in (16) is set to 6, i.e., the six most recent FSs from past to present are considered for calculating the present EWMA value for FD. Fig. 13 represents FSs and their EWMA values for the test samples in Case 2 and their alarm signals. In Fig. 13, vertical solid dotted red lines indicate the unplanned shutdown time caused by tube leakage. In Fig. 13 (a), the 95th, 97th and 99th percentiles are denoted by yellow-green, blue and red dotted horizontal lines, respectively, and EWMA values of FSs are indicated by a solid purple line. Figs. 13 (b) and (c) correspond to alarm signals without and with EWMA, respectively. Compared with Fig. 13 (b), numerous implausible false alarms are removed in



(a)

(b)

Fig. 12. Fault samples with "Critical" alarms in Case 1: (a) fault region 1; (b) fault region 2

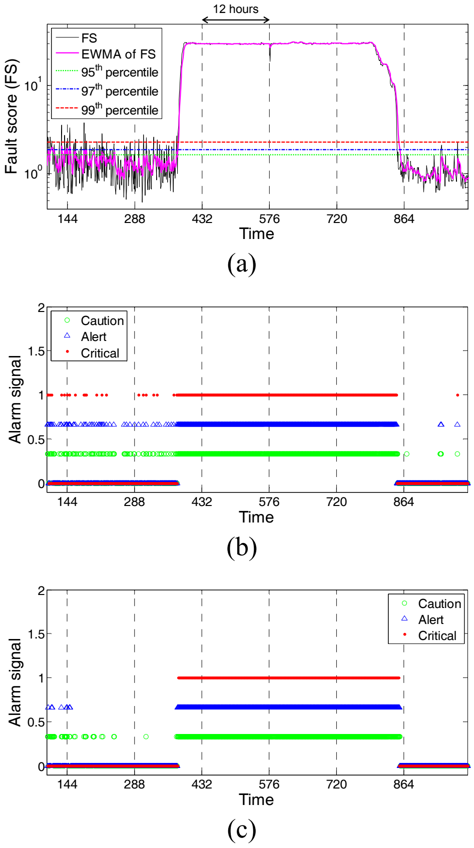


(a)

(b)

(c)

Fig. 13. FD results for Case 2: (a) FSs; (b) alarm signals without EWMA (c) alarm signals with EWMA



(a)

(b)

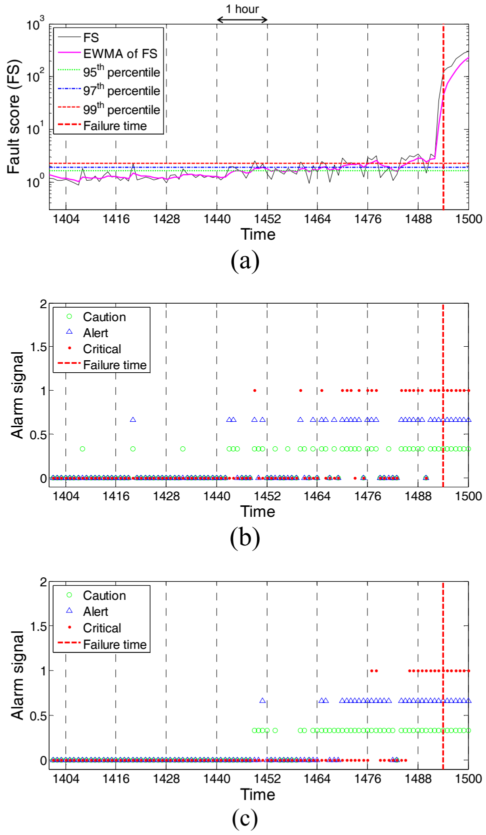
(c)

Fig. 14. Enlargement of "fault region 1" in Fig. 13: (a) FSs; (b) alarm signals without EWMA; (c) alarm signals with EWMA

Fig. 13 (c). The main reason is that the trend of FSs is considered in the EWMA-based FD. The shaded red regions in Fig. 13 (a) designate two fault regions where considerable alarm signals occur and magnification of the regions and their vicinity is illustrated in Figs. 14 and 15.

As illustrated in Figs. 13 (a) and 14, for a period that lasts approximately 40 hours, alarm signals occur intensively approximately 4 days before the unscheduled shutdown. In Fig. 15, "Caution" and "Critical" alarms occur considerably approximately 3 hours and 40 minutes immediately before the unplanned shutdown, respectively. Fig. 16 illustrates fault samples with "Critical" alarms of the fault regions in a three-dimensional space. As indicated in Fig. 16, the geometric patterns of the fault samples are completely dissimilar from those of normal samples. In the fault region 1, considerable "Critical" alarms occur because the reheater pressure and feedwater flow decline rapidly and the condenser make-up flow increases sharply. The fault region 2 corresponds to an early warning region where the condenser make-up flow increases gradually.

5.5 Performance evaluation and comparison

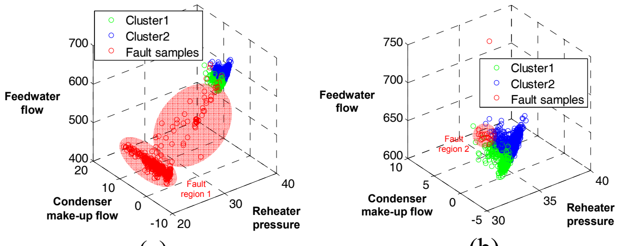


(a)

(b)

(c)

Fig. 15. Enlargement of "fault region 2" in Fig. 13: (a) FSs; (b) alarm signals without EWMA; (c) alarm signals with EWMA



(a)

(b)

Fig. 16. Fault samples with the "Critical" alarms in Case 2: (a) fault region 1; (b) fault region 2

In this subsection, we present the results of the performance comparison between the proposed method and PCA-based fault detection method using four evaluation measures. The closer the evaluation measures are to 1, the better the results are. The PCA-based method has been successfully applied to technical processes such as centrifugal chiller [25], thermal power plant [2], helical coil steam generator [26], continuously stirred tank reactor [27] and self-powered neutron detectors [28]. For the test samples, the four evaluation measures, i.e., accuracy (ACC), sensitivity (SEN), specificity (SPE), and precision (PRE), are calculated as follows [20]:

$$\mathrm{ACC} = \frac{TP + TN}{P + N}, \tag{18}$$

Table 4. Performance comparison for Case 1

	ACC	*SEN*	*SPE*	*PRE*
PCA	0.9805	1	0.9797	0.6885
Proposed method (without EWMA)	0.9928	1	0.9925	0.8571

Table 5. Performance comparison for Case 2

	ACC	*SEN*	*SPE*	*PRE*
PCA	0.9022	0.9979	0.857	0.7673
Proposed method (without EWMA)	0.9759	0.9979	0.9655	0.9318
Proposed method (with EWMA)	0.9933	0.9937	0.9931	0.9855

$$\text{SEN} = \frac{TP}{P}, \tag{19}$$

$$\text{SPE} = \frac{TN}{N}, \tag{20}$$

$$\text{PRE} = \frac{TP}{TP + FP}, \tag{21}$$

where

- P the number of fault samples;
- N the number of normal samples;
- TP the number of samples correctly detected as fault samples;
- TN the number of samples correctly determined as normal samples;
- FP the number of samples incorrectly detected as fault samples;
- FN the number of samples incorrectly determined as normal samples.

In PCA, the cumulative percent variance technique is used to decide the proper number of principal components, and Hotelling's T^2 statistic is employed for fault detection index. If the T^2 statistics of test samples are larger than or equal to T_α^2, where $\alpha = 1\%$, the samples are detected as fault samples. In the proposed method, the samples that satisfy the condition, $FS(t) \geq T_{99\text{th}}$ (or $EWMA_w(t) \geq T_{99\text{th}}$), are decided as fault samples. Tables 4 and 5 summarize the results of performance comparison in Cases 1 and 2, respectively. As listed in Tables 4 and 5, with the exception of sensitivity, the proposed method exhibits superior performance compare to PCA-based method.

6. Conclusion

In this paper, a clustering-based FD method was proposed for the steam boiler in a 200 MW TPP. Failure cases due to boiler tube leakage were collected from the target DCS and main monitored variables for leakage detection were selected based on expert empirical knowledge. In the proposed method, after applying *k*-means algorithm to training samples, FSs are assigned to test samples based on the distances between the samples and their closest cluster centers. To determine the proper number of clusters, slope statistic, an advanced version of silhouette statistic, is employed. The 95th, 97th and 99th percentiles for FSs of the training samples were used for threshold settings and three different alarm signals for unseen samples were generated according to the strength of their FSs. In a second failure case, EWMA was used to consider the trend of FSs.

The main advantages of the proposed method are summarized as follows. First, the proposed method did not require labeled training samples because unsupervised learning was employed. Second, the computation time in the test phase was fairly short because simply calculating the distance between unseen samples and their nearest cluster centers were required. In addition, more flexible FD was possible based on the strength of FSs because three different threshold values were set up using the 95th, 97th and 99th percentiles. Lastly, using EWMA to consider the trend of FSs, false alarms could be easily reduced.

To demonstrate the effectiveness, the proposed method was applied to collected failure cases. The experiment results showed that the proposed method can detect fault samples whose features are markedly different from those of normal samples. In addition, early detection of faults immediately before an unplanned shutdown was achieved successfully.

In this work, we only focus on FD that determines whether a fault has occurred. In future research, we will combine fault identification step with the proposed method to confirm monitored variables relevant to the fault.

Acknowledgements

This work was supported by the Energy Efficiency & Resources Core Technology Program of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea (No. 20151110200040)

References

[1] K. Patan, Artificial neural networks for the modelling and fault diagnosis of technical processes, Springer, 2008.

[2] A. Ajami and M. Daneshvar, “Data driven approach for fault detection and diagnosis of turbine in thermal power plant using Independent Component Analysis (ICA),” Int. J. of Elect. Power & Energy Syst., vol. 43, no. 1, pp. 728-735, Dec. 2012.

[3] C. C. Hsu and C. T. Su, “An adaptive forecast-based chart for non-Gaussian processes monitoring: with

application to equipment malfunctions detection in a thermal power plant," IEEE Trans. Control Syst. Technol., vol. 19, no. 5, pp. 1245-1250, Nov. 2010.

[4] J. Cai, X. Ma and Q. Li, "On-line monitoring the performance of coal-fired power unit: A method based on support vector machine," Appl. Thermal Eng., vol. 29, no. 11-12, pp. 2308-2319, Aug. 2009.

[5] K. Y. Chen, L. S. Chen, M. C. Chen and C. L. Lee, "Using SVM based method for equipment fault detection in a thermal power plant," Comput. in Ind., vol. 62, no. 1, pp. 42-50, Jan. 2011.

[6] N. A. A. Shashoa, G. Kvaščev, A. Marjanović and Ž. Djurović, "Sensor fault detection and isolation in a thermal power plant steam separator," Control Eng. Practice, vol. 21, no. 7, pp. 908-916, Jul. 2013.

[7] F. Li, B. R. Upadhyaya and L. A. Coffey, "Model-based monitoring and fault diagnosis of fossil power plant process units using group method of data handling," ISA Trans., vol. 48, no. 2, pp. 213-219, Apr. 2009.

[8] G. Prasad, E. Swidenbank and B. W. Hogg, "A novel performance monitoring strategy for economical thermal power plant operation," IEEE Trans. Energy Convers., vol. 14, no. 3, pp. 802-809, Sep. 1999.

[9] S. Guo, J. Wang, J. Wei and P. Zachariades, "A new model-based approach for power plant Tube-ball mill condition monitoring and fault detection," Energy Conversion and Manage., vol. 80, pp. 10-19, Apr. 2014.

[10] M. Ahmed, A. N. Mahmood and M. R. Islam "A survey of anomaly detection techniques in financial domain," Future Generation Comput. Syst., vol. 55, pp. 278-288, Feb. 2016.

[11] M. Ahmed, A. N. Mahmood and J. Hu, "A survey of network anomaly detection techniques," J. of Network and Comput. Applicat., vol. 60, pp. 19-31, Jan. 2016.

[12] K. A. P. Costa, L. A. M. Pereira, R. Y. M. Nakamura, C. R. Pereira, J. P. Papa and A. X. Falcão, "A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks," Inform. Sci., vol. 294, pp. 95-108, Feb. 2015.

[13] H. Li, A. Achim and D. Bull, "Unsupervised video anomaly detection using feature clustering," IET Signal Process., vol. 6, no. 5, pp. 521-533, Jul. 2012.

[14] J. Zhao, K. Liu, W. Wang and Y. Liu, "Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry," Inform. Sci., vol. 259, pp. 335-345, Feb. 2014.

[15] A. Fujita, D. Y. Takahashi and A. G. Patriota, "A non-parametric method to estimate the number of clusters," Computational Stat. & Data Anal., vol. 73, pp. 27-39, May 2014.

[16] J. S. R. Jang, C. T. Sun and E. Mizutani, Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence, Prentice Hall, 1997.

[17] K. B. Kim and D. H. Song, "Automatic Intelligent Asymmetry Detection Using Digital Infrared Imaging with K-Means Clustering," Int. J. of Fuzzy Logic and Intelligent Syst., vol. 15, no. 3, pp. 180-185, Sep. 2015.

[18] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," J. of Computational and Appl. Math., vol. 20, pp. 53-65, Nov. 1987.

[19] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: a survey," ACM Computing Surveys (CSUR), vol. 41, no. 3, Jul. 2009.

[20] J. Han, M. Kamber and J. Pei, Data mining: concepts and techniques, Elsevier, 2011.

[21] D. Flynn, Thermal power plant simulation and control, IET, 2003.

[22] A. K. Raja, Power plant engineering, New Age Int., 2006.

[23] D. Sarkar, Thermal power plant design and operation, Elsevier, 2015.

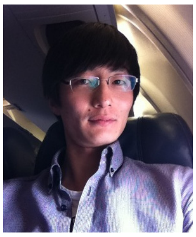
[24] J. E. Oakey, Power plant life management and performance improvement, Elsevier, 2011.

[25] S. Wang and J. Cui, "Sensor-fault detection, diagnosis and estimation for centrifugal chiller systems using principal-component analysis method," Appl. Energy, vol. 82, no. 3, pp. 197-213, Nov. 2005.

[26] K. Zhao and B. R. Upadhyaya, "Model based approach for fault detection and isolation of helical coil steam generator systems using principal component analysis," IEEE Trans. Nucl. Sci., vol. 53, no. 4, pp. 2343-2352, Aug. 2006.

[27] F. Harroua, M. N. Nounoua, H. N. Nounoub and M. Madakyaru, "Statistical fault detection using PCA-based GLR hypothesis testing," J. of Loss Prevention in the Process Ind., vol. 26, no. 1, pp. 129-139, Jan. 2013.

[28] X. Penga, Q. Lib and K. Wanga, "Fault detection and isolation for self powered neutron detectors based on Principal Component Analysis," Ann. of Nucl. Energy, vol. 85, pp. 213-219, Nov. 2015.



Jungwon Yu He received the B.S. and M.S. degrees from the Department of Electrical and Computer Engineering from Pusan National University (PNU), Busan, Korea, in 2012 and 2014, respectively, and is currently pursuing the Ph.D. degree in the Department of Electrical and Computer engineering at PNU. His research interests include time series analysis, data mining and fault detection and diagnosis, etc.

Jaeyel Jang He received the M.S. degrees from the Graduate School of Information Security from Korea University, Seoul, Korea, in 2013. He is currently a deputy general manager at the Technology & Information Department, Technical Solution Center, Korea East-West Power Co., Ltd. His research interests include multi-variable control, data mining and fault diagnosis, etc.

Jaeyeong Yoo He received the B.S. degrees in Electrical Engineering from Yonsei University, Korea, in 1982. He has been a senior researcher at LSIS Co., Ltd. from 1983 to 1991. He is currently a chief technology officer (CTO) at the XEONET Co., Ltd. His research interests include process controller design, fault diagnosis and prognosis, etc.

June Ho Park He received the B.S., M.S. and Ph.D. degrees from Seoul National University, Seoul, Korea in 1978, 1980, and 1987, respectively, all in electrical engineering. He is currently a Professor at the School of Electrical Engineering, Pusan National University, Busan, Korea. His research interests include intelligent systems applications to power systems. Dr. Park has been a member of the IEEE Power Engineering Society.

Sungshin Kim He received his B.S. and M.S. degrees in Electrical Engineering from Yonsei University, Korea, in 1984 and 1986, respectively, and his Ph.D. degree in Electrical Engineering from the Georgia Institute of Technology, USA, in 1996. He is currently a professor at the Electrical Engineering Department, Pusan National University. His research interests include fuzzy logic controls, neuro fuzzy systems, neural networks, robotics, signal analysis, and intelligent systems.