Energy and Buildings 112 (2016) 270–278

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

A statistical training data cleaning strategy for the PCA-based chiller sensor fault detection, diagnosis and data reconstruction method

Yunpeng Hu[a,b,c], Huanxin Chen[a,*], Guannan Li[a], Haorong Li[d], Rongji Xu[e], Jiong Li[c]

[a] Department of Refrigeration and Cryogenic Engineering, School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan 430074, PR China
[b] Wuhan Business University, Wuhan 430056, PR China
[c] State Key Laboratory of Compressor Technology, Hefei 230031, PR China
[d] Department of Architectural Engineering, University of Nebraska-Lincoln, PKI Room245 1110S, 67th Street, Omaha, NE 68182, United States
[e] Beijing University of Civil Engineering and Architecture, Beijing 100044, PR China

ARTICLE INFO

Article history:
Received 17 May 2015
Received in revised form 7 November 2015
Accepted 26 November 2015
Available online 2 December 2015

Keywords:
Chiller sensor fault
Training data cleaning
Euclidean distance
Principal component analysis
Outlier detection

ABSTRACT

This paper presents a statistical training data cleaning strategy for PCA-based chiller sensor Fault Detection, Diagnosis and Data Reconstruction method. Finding and removing outliers from the original training data set, the training data quality can be improved by the presented data-cleaning strategy. This can enhance the efficiency of the fault detection and increase the accuracy of the data reconstruction. Outliers cannot be easily found in the original data set used for training the PCA model. These outliers would severely affect the projection directions of PCA's two orthogonal subspaces (PC subspace and Residual subspace). Therefore, the threshold of *Q*-statistic is changed by the unexpected projection subspaces so that the detection efficiency of the sensor fault is decreased. The Euclidean distance was employed as an index to detect outliers from the original training data. In order to achieve optimal training data for the sensor FDDR, the *z*-scores of each sample's Euclidean Distance were employed as the key to remove the outliers. A field measured data set of a screw water-cooled chiller was used to validate the presented strategy. Results demonstrate that the quality of the training data is optimized and sensor fault detection efficiency, as well as the reconstruction data accuracy, is improved when compared to the normal PCA method.

1. Introduction

Reliable, accurate and real-time measurement from sensors in the control and monitor systems is critical to maintain and optimize the operation conditions of HVAC&R systems. However, sensor faults are inevitable. HVAC&R system is a thermodynamic system and the measurements of sensors are not isolated. Therefore, data-driven methods for sensor fault detection, diagnosis and data reconstruction are very useful in decades. Due to measurement noises and system dynamics, outliers in the original training data set are inevitable. Especially when the online data analysis would be further applied to HVAC&R systems, the original data will be huge and dirty more and more. Training data is the critical condition and the necessary condition for the performance of the data-driven methods. Therefore, the data cleaning method will be fundamental for the data-driven model.

Sensor faults lead to a deviation on the system control, and further impact the operational state, such as lower indoor air quality [1], and higher energy consumption [1–3]. Simulation results showed that the mixed air temperature sensor fault would increase the annual energy of an air handling system up to 30%, or the cooling coil discharge air temperature sensor fault would cause up to 150% energy consumption [2]. The increment of energy consumption is 14% with the bias fault of the return air flow rate sensor [4]. Also, the energy consumption can be up to 195% if the outdoor air flow rate sensor was completely failure [5]. Therefore, researches on the sensor fault detection, diagnosis and data reconstruction (FDDR) in HVAC&R system are of great importance and have been paid more attention to in the recent years.

The data derived from the sensors in the HVAC&R system are highly coupled according to the energy balance principle, the flow-pressure balance principle and other basic principles. Due to the disturbance of the indoor and outdoor ambient conditions, HVAC&R system is normally operated under a wide range. Thus, the sensor fault cannot be directly detected by its own historical data. One common technique used on the sensor FDDR is the

[*] Corresponding author. Tel.: +86 2787558330.
E-mail address: chenhuanxin@tsinghua.org.cn (H. Chen).

http://dx.doi.org/10.1016/j.enbuild.2015.11.066
0378-7788/© 2015 Elsevier B.V. All rights reserved.

Nomenclature

T_{chws}	chilled-water supply temperature (°C)
T_{chwr}	chilled-water return temperature (°C)
M_{chw}	chilled-water flow rate (m^3/h)
T_{cws}	condenser-water supply temperature (°C)
T_{cwr}	condenser-water return temperature (°C)
M_{cw}	condenser-water flow rate (m^3/h)
M_{ref}	mass flow rate of refrigerant (kg/s)
W_{comp}	chiller electrical-power input (W)
$\boldsymbol{X}$	original matrix
$\boldsymbol{X}^0$	normalized original matrix
$\boldsymbol{R}$	covariance matrix
$\boldsymbol{U}$	eigen vector matrix
VE	variance explained
CV	cumulative contribution of variance
FDD	fault detection and diagnosis
FDDR	fault detection, diagnosis and reconstruction
HVAC&R	heating, ventilating, air-conditioning and refrigeration
PC	principal component
PCA	principal component analysis
SPCA	principal component analysis with a statistical data-cleaning
P	PC subspace projection matrix
$\tilde{P}$	residual subspace projection matrix
Q_α	threshold of the Q-statistic
D	Euclidean distance
$\vec{x}$	a sample
$\hat{\vec{x}}$	estimate of a sample
$\vec{e}$	residual of a sample
$\vec{x}_{rc}$	reconstruction of a sample
Greek letters	
μ	mean
σ	standard deviation
$\lambda_1,\ldots,\lambda_n$	eigenvalues

multi-dimensional data analysis. Due to the rapid development of auto control and communication technology in the recent years, it becomes more flexible and convenient to derive the original operation data from the Building Management System. Therefore, various multi-dimensional data-driven methods have been introduced into the FDDR of HVAC&R system.

Many data analysis methods were proposed to develop the model for sensor FDDR in recent years, such as the neural network [6], the wavelet neural network [7], the feature selection [8], the support vector machine [9], the principal component analysis [10], the fisher discriminant analysis [11], the expert system [12,13], the fractal correlation dimension [14], the cumulative sum [15], etc. Principal Component Analysis (PCA) method [16,17] is one of the multivariate statistical analysis methods. After projected by PCA, the original sample can be represented by less principal components without losing the feature of the system. Instead of analyzing all the isolated variables, the PCA method focuses on the collection of the different measuring samples as a whole. Therefore, the PCA sensor FDD method has been introduced into HVAC systems, subsystems, and components in the last decades. Many researchers applied the PCA method to the chiller sensor FDD, like Wang and Cui [18,19], Xu et al. [20], Chen and Lan [21], Hu et al. [22] and so on.

A common flowchart of a multi-dimensional data-driven sensor FDDR strategy is illustrated in Fig. 1. The training data is chosen carefully from the original operational data due to the inevitability of noise or outlier. Therefore, the data cleaning process is extremely important to enhance the reliability of the training data set. Due to measurement noises and system dynamics, a filter based on Hotelling T^2 was used to eliminate the outlier from the measurements of an AHU in an existed commercial building [4]. A fuzzy Kohonen clustering network algorithm was used to clean the real data of a data mining model for electric load forecasting [23]. With the rapid development of real-time data analysis, the data cleaning of real measurements will be of great important for real-time online calculation. Sometimes data cleaning is included in the step of data preprocessing. Learning from the training data by different kinds of data-driven methods, a routine operation model is established and the fault boundary is identified. The new sample is transferred, analyzed by the algorithm driven from the training data. Then the analysis result of a new sample should be compared with the fault boundary to detect whether there is on the fault condition. If the analysis result is outside of the fault boundary, someone sensor in the model would be faulty one. Then the faulty sensor should be diagnosed and the erroneous measurement data should be reconstructed.

Most of these investigations were focused on the study of applying different data-driven methods into the sensor FDDR of HVAC&R system. In the entire procedures of any data-driven method, the analysis results would be highly dependent on the quality of the training data. Therefore, the quality of the training data is fundamental to the data-driven methods. The quality of a large real world data set depends on a number of issues, such as sampling interval, sampling location, and so on. Among them, the source of the data is the crucial factor [24] when we apply the FDDR methods to the real HVAC&R system. However, rare work was reported on how to clean training data to enhance FDDR results in detail. In this study, a training data cleaning strategy based on the statistical distance, the Euclidean distance, has been proposed to find out outliers in the original measurement data. The data quality of the training data for PCA-based sensor FDDR method was improved after removing outliers. The presented strategy is named as SPCA for short when compared to the normal PCA method, shortened as PCA. The results of SPCA are validated by the field tested data. The SPCA method can improve the efficiency of the sensor fault detection in the water-cooled chiller and can increase the accuracy of the fault data reconstruction.

2. Principal of SPCA method

2.1. Principal component analysis and its projection

In PCA method, the original data matrix $X^0 \in R^{m\times n}$ usually consists of m samples (rows) and n process variables (columns) obtained from the field measurements. Due to different engineering units and different orders of magnitude, $\boldsymbol{X}^0$ should be normalized to a normalized matrix $\boldsymbol{X}$ with zero mean and unit variance. This step makes the covariance matrix $\boldsymbol{R}$ the same as the correlation coefficient matrix. The covariance matrix $\boldsymbol{R}$ of $\boldsymbol{X}$ can be defined as

$$R \approx \frac{X^T X}{n-1} \tag{1}$$

The eigenvalues, $\lambda_1,\ldots,\lambda_n$, ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$) are obtained after eigenvalue decomposition of $\boldsymbol{R}$. The eigenvector matrix $\boldsymbol{U}$ is the sequence of the corresponding eigenvectors of the eigenvalues in turn.

Therefore, PCA's two orthogonal projection subspaces are obtained. The first subspace that captures the process systematic variations is PC subspace. Its projection matrix is $\boldsymbol{P}$ that contains the former k columns of $\boldsymbol{U}$. The other subspace that captures random noise, error or useless information is Residual subspace. Its projection matrix is $\tilde{P}$ that contains the last n–k columns of $\boldsymbol{U}$. If the

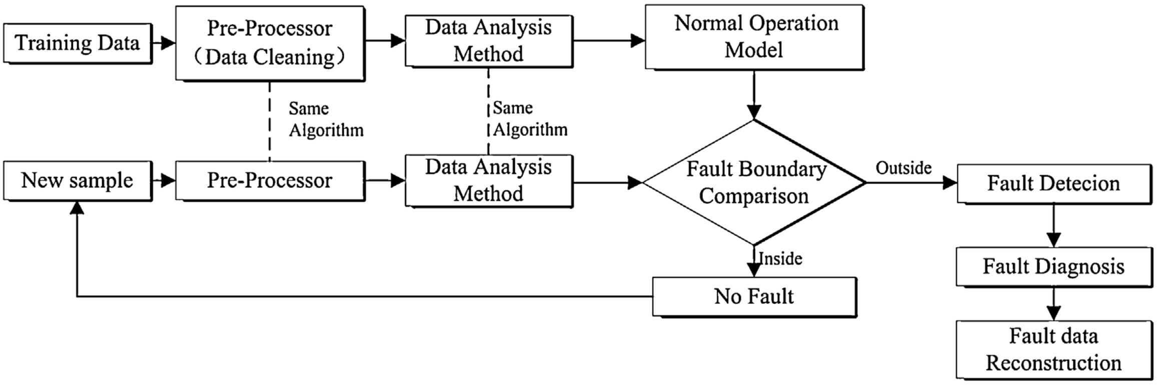

Fig. 1. Flowchart of a data-based sensor fault detection, diagnosis and reconstruction.

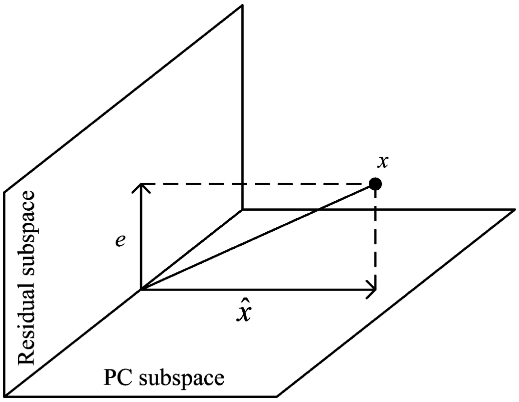

Fig. 2. Projection of a sample *x* by PCA method.

cumulative contribution of variance of former *k* PCs is just greater than a certain percentage number, e.g. 85% or 90%, the number of principal components is *k*. In this paper, 85% is used as the criterion of the CV to determine the number of principal components.

A sample, namely a vector $\vec{x}$ in the mathematics, can be projected into PC and Residual subspace, i.e. estimation $\hat{\vec{x}}$ and residual $\vec{e}$. The projection relationship can be illustrated in Fig. 2. $\vec{x}$ can be defined as

$$\vec{x} = \hat{\vec{x}} + \vec{e} \tag{2}$$

The estimate $\hat{\vec{x}}$, which is the projection of $\vec{x}$ into the PC subspace, can be defined as

$$\hat{\vec{x}} = \vec{x}PP^T \tag{3}$$

The residual $\vec{e}$, which is the projection of $\vec{x}$ into the residual subspace, can be defined as

$$\vec{e} = \vec{x} - \hat{\vec{x}} = \vec{x}\left(\tilde{P}\tilde{P}^T\right) = \vec{x}\left(I - PP^T\right) \tag{4}$$

From Eqs. (3) and (4), obviously, the values of $\hat{\vec{x}}$ and $\vec{e}$ heavily depend on the projection relationship of PC subspace and Residual subspace. It is illustrated in Fig. 3 that the values of $\hat{\vec{x}}$ and $\vec{e}$ can be absolutely different, even if a same vector $\boldsymbol{A}$ is projected into the different projection relationships, *x*–*o*–*y* and *x*′–*o*–*y*′.

The existence of outliers in the training data set can strongly affect the data covariance structure, and, therefore, the projection relationship of the PC subspace and Residual subspace. The fault boundary is broader than the expected. As a result, the sensitivity of fault detection would be reduced. Therefore, obtaining a good pair of PC subspace and Residual subspace from the original training matrix is a primary and significant step in the PCA method.

2.2. Euclidean distance outlier detection

In this paper, a statistical training data cleaning strategy, Euclidean distance-based outlier detection, is employed to remove the outliers from the original training data. From a statistical point, the multivariate data comparison problem can be transformed into a univariate data comparison problem by using different distance values. The Euclidean distance means a straight line distance between two points. For a vector $\vec{A}$ in the certain space, the distance between $\vec{A}$ and the origin of coordinates can be defined as

$$D\left(\vec{A}\right) = \sqrt{\sum_{i=1}^{n}(a_i)^2} \tag{5}$$

where a_i is the coordinate of $\vec{A}$ in the *i*th dimension.

Considering the PCA normalization training data $\boldsymbol{X}$ is a $m \times n$ multivariate matrix. The Euclidean distance D_i of *i*th row, namely *i*th sample, can be defined as

$$D_i = \sqrt{\sum_{j=1}^{n}\left(x_{i,j}\right)^2} \tag{6}$$

where $x_{i,j}$ is the *j*th variable of *i*th sample in the $\boldsymbol{X}$. The mean of the Euclidean distances of all samples, μ_D, can be defined as

$$\mu_D = \frac{1}{m}\sum_{i=1}^{m} D_i \tag{7}$$

The standard deviation of the Euclidean distances of all samples, σ_D, can be defined as

$$\sigma_D = \sqrt{\frac{1}{(m-1)}\sum_{i=1}^{m}(D_i - \mu_D)} \tag{8}$$

The Euclidean distances of all samples can be expressed as a univariate data sequence, $\{D_1, D_2, \ldots, D_m\}$. The problem of multivariate analysis has been transformed into a problem of univariate analysis. In this paper, the *z*-score is used to identify outliers. The *z*-score of the Euclidean distances of *i*th sample, z_i, is defined as

$$z_i = \frac{\left|D_i - \mu_D\right|}{\sigma_D} \tag{9}$$

Based on Chebyshev's theorem [25], there are about 68.27% of the values lying within one standard deviation of the mean, about 95.45% of the values within two standard deviations of the mean and about 99.73% of the values within three standard deviations of the mean. The *z*-scores of the three cases above are 1, 2 and 3, respectively. In this paper, if the *z*-score of the Euclidean distance

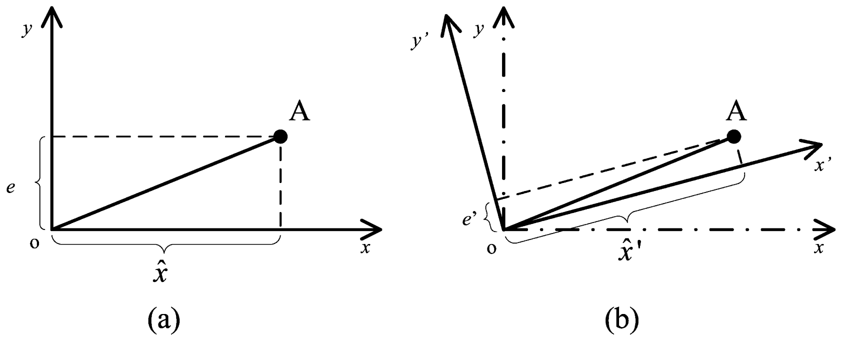

Fig. 3. Comparison of a same vector in the different projection relationships.

of the *i*th sample is greater than 2, this sample will be identified as an outlier and removed from the training data.

2.3. PCA-based fault detection and data reconstruction

The *Q*-statistic of $\vec{x}$ is determined by analyzing the residual. The *Q*-statistic can be expressed as

$$Q = \left\|\vec{e}\right\|^2 < Q_\alpha \tag{10}$$

The *Q*-statistic translates the multivariate residual $\vec{e}$ into a univariate datum, so the thresholds Q_α of the *Q*-statistic can be used as the fault boundary to detect the fault conditions. Q_α can be defined as

$$Q_\alpha = \theta_1 \left[\frac{c_a \sqrt{2\theta_2 h_0^2}}{\theta_2} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{11}$$

where $\theta_i = \sum_{j=k+1}^{n} \lambda_j^i, \quad i = 1, 2, 3, h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$ and c_α is the normal deviate corresponding to the upper $(1-\alpha)$ percentile.

Due to the continuity of monitor process, only one new sample or small sample data set in the process cannot indicate the sensor fault caused by random error or accidental error. In order to compare the sensitivity of different methods, the statistical result of fault detection, the detection efficiency, is defined as

$$\eta = \frac{\sum \mathrm{num}_{\mathrm{fault}}}{\sum \mathrm{num}_{\mathrm{new}}} \tag{12}$$

where $\sum \mathrm{num}_{\mathrm{new}}$ is the total number of the new samples to be detected, $\sum \mathrm{num}_{\mathrm{fault}}$ is the number of samples under fault conditions, i.e. the corresponding *Q*-statistic values are greater than Q_α. The greater the detection efficiency is, the more sensitivity of the FDD method is. If the fault detection efficiency is greater than 50%, the sensor fault would be alarmed.

Considering $\mathrm{d}Q_{\mathrm{rc}}/\mathrm{d}\tilde{f}_i = 0$, the reconstruction data, presented by Dunia and Qin [26], is defined as

$$\vec{x}_{\mathrm{rc}} = \left(I - \Xi_i \Xi_i^+\right) \vec{x} \tag{13}$$

where $\tilde{f}_i$ is an estimate of the fault magnitude f_i which measures the displacement in the direction Ξ_i, and $()^+$ is the Moore–Penrose pseudo inverse.

3. SPCA sensor fault detection strategy for water-cooled chiller

Considering the basic heat balance analysis of the chiller in the past works, there are eight important sensors in the chiller system and its control system [22]. Therefore, sensors included in the PCA model should be at least shown as

$$X = \begin{bmatrix} T_{\mathrm{chws}} & T_{\mathrm{chwr}} & M_{\mathrm{chw}} & T_{\mathrm{cws}} & T_{\mathrm{cwr}} & M_{\mathrm{cw}} & W_{\mathrm{comp}} & M_{\mathrm{ref}} \end{bmatrix} \tag{14}$$

where M_{ref} refers the sensor related to the capacity control that indicates the mass flow rate of refrigerant.

The structure of SPCA sensor FDD strategy is shown in Fig. 4. The process of the SPCA method is described as follows:

(1) Calculate each row's Euclidean distance of $\boldsymbol{X}$ after normalizing the original matrix $\boldsymbol{X}^0$, then calculate the mean and standard variance of the Euclidean distance vector.
(2) Calculate the *z*-score of each row's Euclidean distance and compare it with the *z*-score threshold. If the *i*th row's *z*-score is greater than the *z*-score threshold, this sample is an outlier. Obtain new original data set $\boldsymbol{X}^0$ by deleting all the outliers from the original data set.
(3) Normalize the new training data set to matrix $\boldsymbol{X}$ with zero mean and unit variance.
(4) Obtain covariance matrix *R* by Eq. (1) and then decompose $\boldsymbol{R}$ to obtain eigenvalues and eigenvectors.
(5) Get *k*, the number of PCs and the residual subspace project matrix $\boldsymbol{P}$.
(6) Calculate Q_α of the training model.
(7) Input a new sample x_{new}, calculate and compare its *Q*-statistic to Q_α, then decide whether there is on the sensor fault conditions.

4. Validation

A field data set was employed to validate and evaluate the SPCA method in practice. The details of a real screw chiller were described in the reference [22].

4.1. An example of error sensor data

The sensor fault cannot be found out easily by observing its own historical data isolated. Chinese standards recommend designing chiller operational conditions at chilled water supply temperatures of 7 °C and chilled water return temperatures of 12 °C. In fact, the chilled water supply temperatures can be changing around 7 °C irregularly, as well as the chilled water return temperatures around 12 °C, at the real operational conditions, because of the irregular change of the external disturbances such as outdoor weather, indoor occupancy, and so on. The different faults of chilled water return temperature, which occurred on day 18, with bias fault level of −0.5 °C and 1.0 °C, are shown in Figs. 5 and 6, respectively. Due to the measured data of T_{chwr} are around 12 °C, the sensor fault cannot be detected easily only by observing the historical data of T_{chwr}. Therefore, the multi-dimensional data-driven sensor FDDR strategy is favorable to detect the faulty sensor.

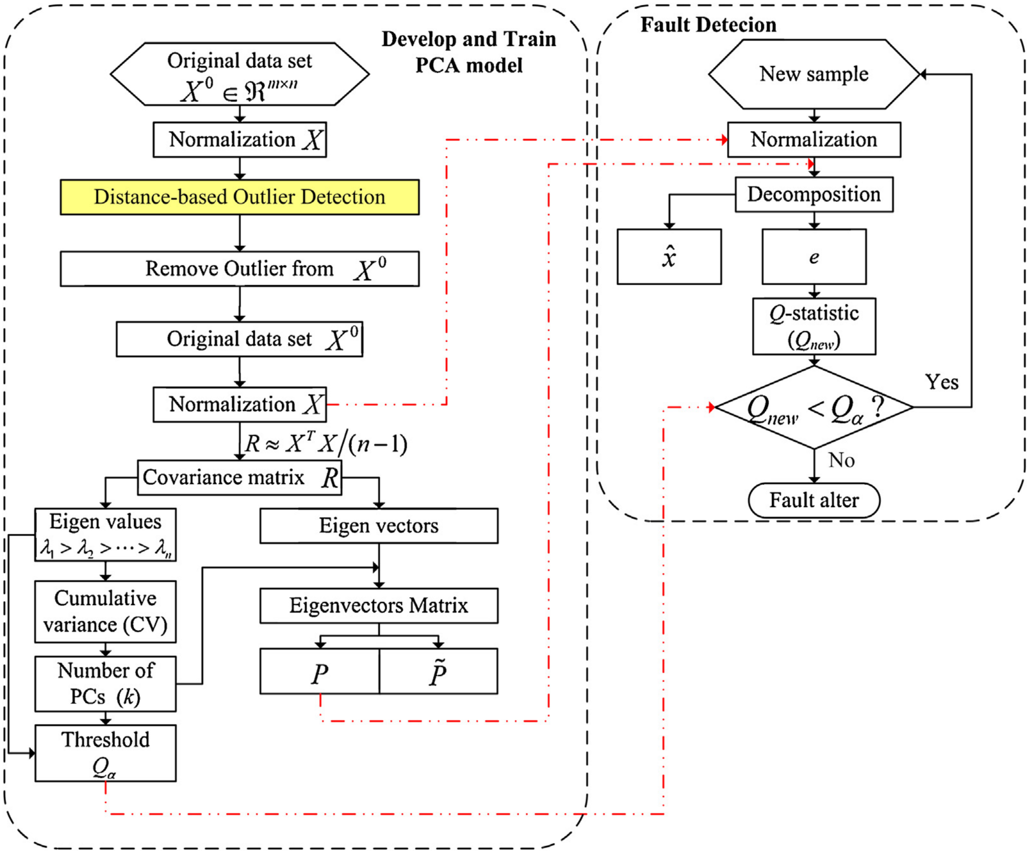

Fig. 4. Flowchart of fault detection of SPCA.

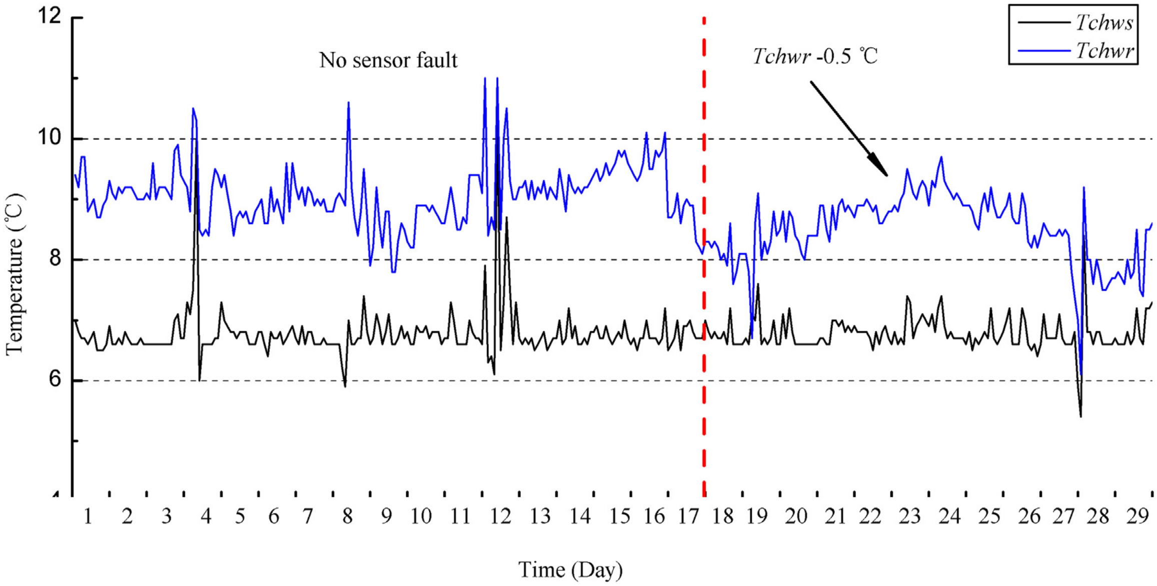

Fig. 5. The historical data of chilled water return temperature at −0.5 °C bias fault level.

4.2. Training model comparison

The z-score plot of the normalized matrix in the process of outlier detection is shown in Fig. 7. There are 4 samples, 2% of total original data, of which corresponding Euclidean distance's z-scores are greater than 2. Four outliers' z-scores are 7.86, 2.37, 8.90, and 4.46, respectively. After removing outliers, there are 196 remained samples within the 200 original samples.

Euclidean Distances' mean and standard deviation of two methods is shown in Table 1. Euclidean Distances' standard deviation of the PCA model is greater than that of SPCA method. Training data after outlier detection is more compact and clustered, when compared to the training data before outlier detection. The training model after outlier detection is closer to the normal condition. The SPCA method is more sensible than PCA.

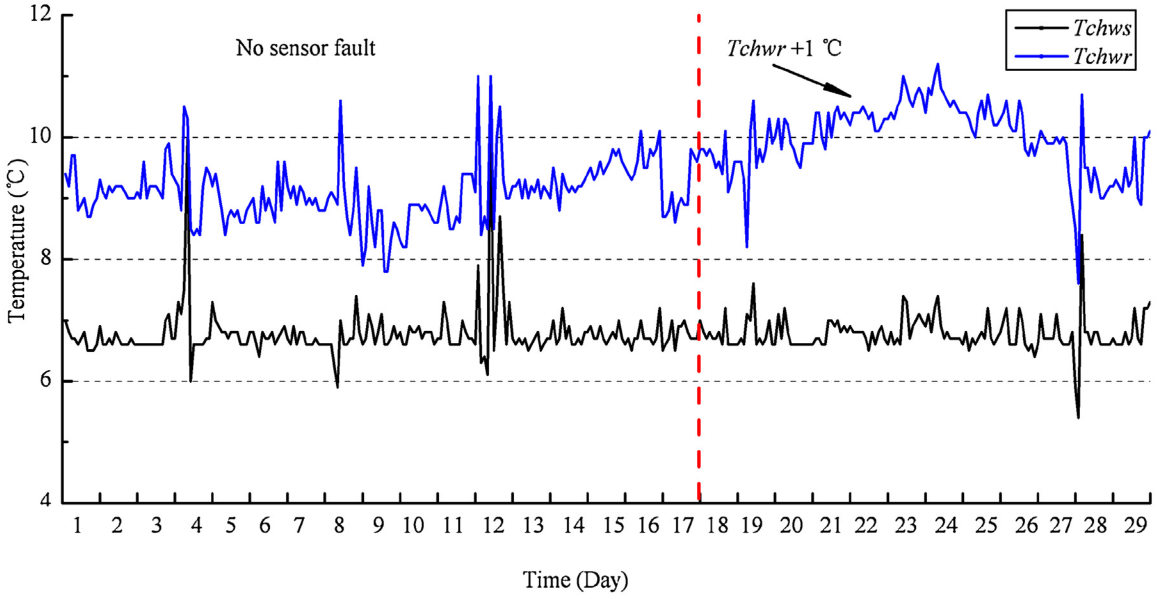

Fig. 6. The historical data of chilled water return temperature at +1.0 °C bias fault level.

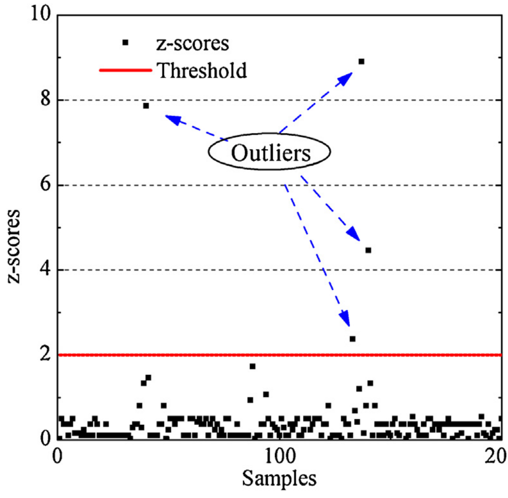

Fig. 7. The *z*-score plot of the original training data using a field data.

Table 1
The Euclidean distances' mean and standard deviation of the two methods.

	PCA method	SPCA method
Mean	0.4899	0.3850
Standard deviation	0.8711	0.7165

The eigenvalues of all PC, the variance explained, and the corresponding CV explained of two methods has been presented as shown in Table 2. They are only mathematical results of the data training. From the results shown in Table 2, PCs numbers of two methods are both four. The CV of PCA and the CV of SPCA are 89.32% and 88.37%, respectively.

4.3. Fault detection efficiency comparison

Different fault levels were introduced to the chilled-water supply temperature sensor. Introduced fault levels were increased by an increment of 0.5 °C from −2.0 °C to +2.0 °C. The detection efficiencies of two methods are shown in Table 3.

Table 2
The eigenvalues and VE, CV of the two methods.

The *i*th PC	PCA method			SPCA method		
	Eigenvalue	VE (%)	CV (%)	Eigenvalue	VE (%)	CV (%)
1	3.2186	40.23	40.23	3.2821	41.03	41.03
2	1.8589	23.24	63.47	1.6682	20.85	61.88
3	1.1101	13.88	77.35	1.1670	14.59	76.47
4	0.9582	11.98	89.32	0.9523	11.90	88.37
5	0.4775	5.97	95.29	0.5288	6.61	94.98
6	0.2097	2.62	97.91	0.2363	2.95	97.93
7	0.1048	1.31	99.22	0.1087	1.36	99.29
8	0.0622	0.78	100.00	0.0566	0.71	100.00

Table 3
The chilled-water supply temperature sensor fault detection efficiencies of the two methods.

Introduced fault level (°C)	T_{chws}	
	PCA	SPCA
+2.0	100.00	100.00
+1.5	100.00	100.00
+1.0	100.00	100.00
+0.5	69.59	97.30
−0.5	4.73	10.81
−1.0	8.11	75.68
−1.5	55.41	99.32
−2.0	97.97	99.32

By PCA method, when there is a positive introduced fault level, sensor fault detection efficiency is nearly no less than 70%. But when there is a negative introduced sensor fault level, the fault detection efficiency is very low. As shown in Table 3, if the fault levels are greater than or equal to 1 °C or at −2.0 °C, the sensor fault detection efficiency is greater than 97.97%. PCA method works very well. At +0.5 °C fault level, PCA method works a little worse and the detection efficiency is decreased to nearly 70%. However, when the introduced fault level is −1.5 °C, detection efficiency is only 55.41%. Fault detection efficiency is almost zero, when fault levels are from 0 °C to −1.0 °C.

After removing the outliers from the original training data, the sensor fault detection efficiency of the SPCA method has been promoted, except for the introduced fault level −0.5 °C. Fig. 8 shows

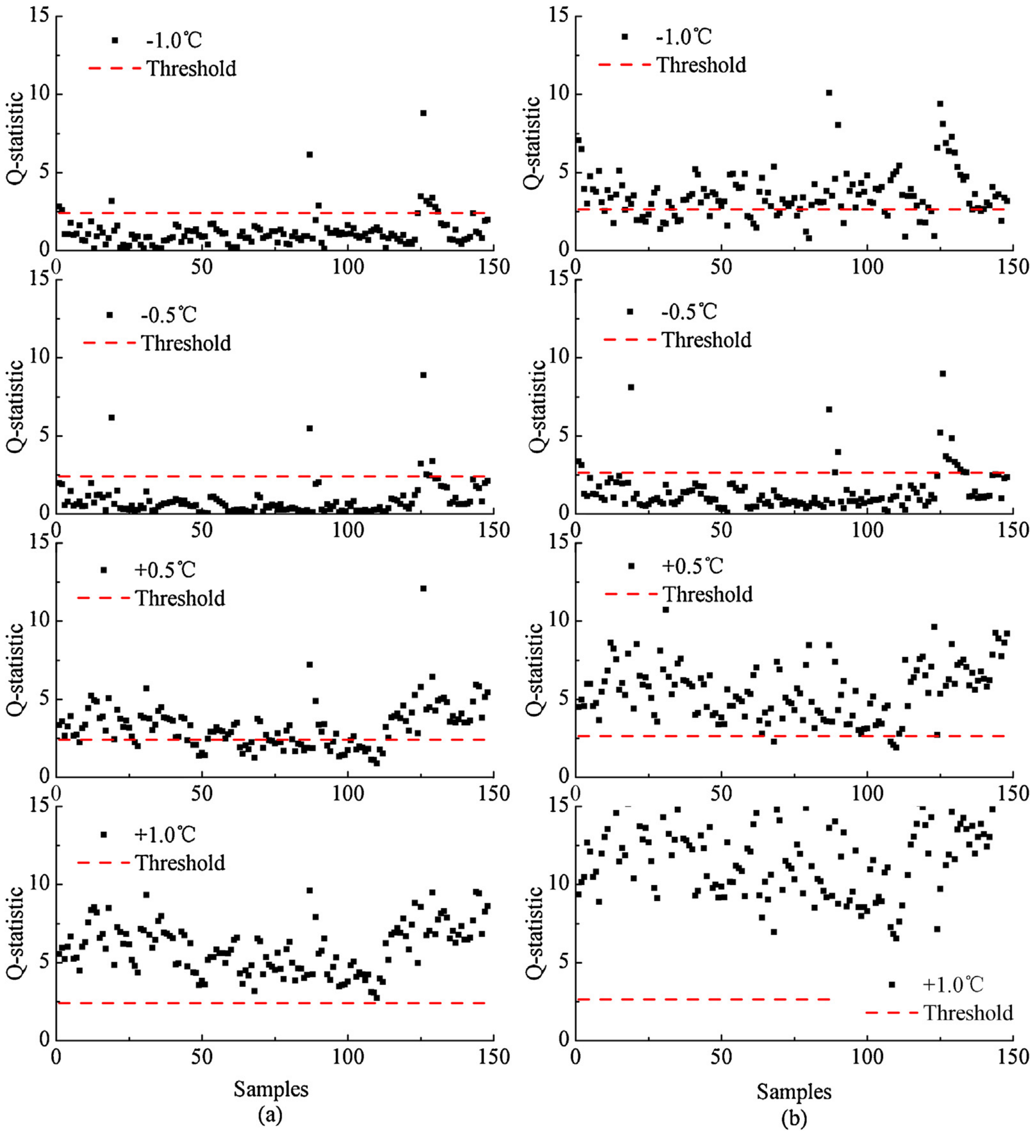

Fig. 8. *Q*-statistic plots of T_{chws} with different fault levels by the two methods at the fault levels from −1 °C to +1 °C (a) of the PCA method, (b) of the SPCA method.

the differences of *Q*-statistic plots of T_{chws} sensor with different fault levels by PCA method and SPCA method. As mentioned in Table 3, detection efficiency of PCA is 100% at a fault level greater than +1.0 °C. Fig. 8 just shows the *Q*-statistic plots at fault level from −1.0 °C to +1.0 °C with a 0.5 °C step. It can be observed clearly that at the same fault level, the detection efficiency of SPCA method is higher than that of PCA method.

All detection efficiencies of two PCA method are compared as shown in Fig. 9, with fault levels from −2 °C to +2 °C. At −0.5 °C fault level, detection efficiency of PCA method is almost below 10%, but that of SPCA method is greater than 77.7%. At −1.0 °C fault level, the detection efficiency of SPCA method is nearly 100%. On the contrary, detection efficiency of PCA is only 8.11%. The detection efficiency of SPCA has been promoted significantly and performed much better than PCA at low-level sensor fault.

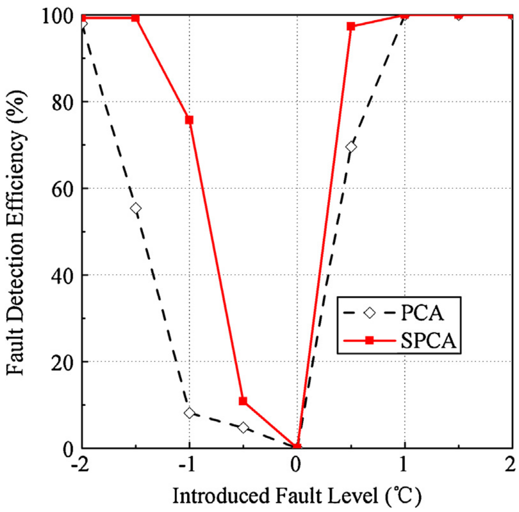

Fig. 9. The T_{chws} sensor fault detection efficiencies of the two methods at the fault levels from −2 °C to +2 °C.

4.4. Fault reconstruction comparison

The error between the reconstruction data and the original data refers to the performance of sensor fault reconstruction. T_{chws} sensor fault reconstruction results of PCA and SPCA are shown in Fig. 10. The mean of T_{chws} reconstruction error by the PCA is 0.57 °C, while

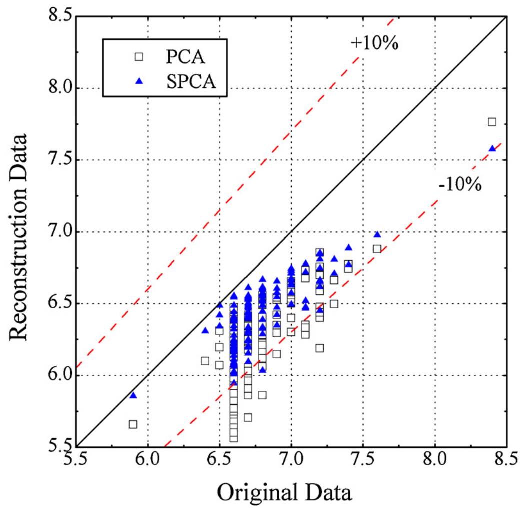

Fig. 10. T_{chws} sensor fault reconstruction results comparison between PCA and SPCA.

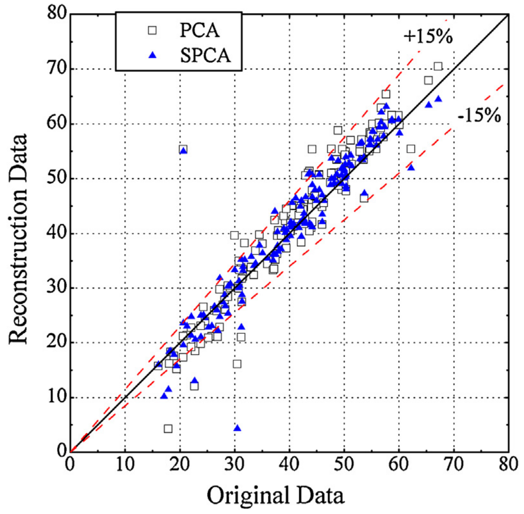

Fig. 11. M_{ref} sensor fault reconstruction results comparison between PCA and SPCA.

that by SPCA is 0.37 °C. The stand deviation of T_{chws} reconstruction error by PCA and SPCA are 0.24 °C and 0.19, respectively. It's found that T_{chws} reconstruction results of SPCA are better than that of PCA. M_{ref} sensor reconstruction results of two methods are shown in Fig. 11. The mean of M_{ref} reconstruction error by the PCA and by SPCA are −0.88 and −0.62, respectively. The stand deviation of M_{ref} reconstruction error by PCA, which is 5.10, is greater than that by SPCA, which is 4.65.

5. Conclusion

In general, the PCA-based sensor FDDR method worked well in applications for HVAC&R system. However, it is found that the sensor fault is not completed detected at low-fault levels for a field measured data set. Training data quality should be one of the major reasons for this incomplete detection. Outlier detection is critical for enhancing the quality of training data for PCA-based sensor FDDR method. The existence of outliers in the training data set can significantly alter the projection structure of the PC subspace and Residual subspace, then to affect the efficiency of fault detection and to decrease the accuracy of data reconstruction.

In this study, a statistical data-cleaning method is presented to remove outliers for obtaining high-quality training data. In SPCA method, the *z*-score of the Euclidean Distance was employed as the criterion to remove the outliers in the original data. SPCA method was validated by the field data of a screw chiller system. With outlier detection, the two orthogonal partition subspaces have been well developed, and the sensor fault detection efficiencies and the reconstruction error of the prevented SPCA method can be improved significantly. The results also show that the training data cleaning process is useful for the fault reconstruction. When the original data, including training data and the tested data, are the online data for a long period of operation, the data cleaning process will be of great important. Based on the present work, investigations on how to improve the quality of field data for chiller sensor FDDR will be further studied.

Acknowledgements

The research work presented in this paper is supported by National Natural Science Foundation of China (Project 51576074 and Project 51328602), and is funded by Beijing Key Lab of Heating, Gas Supply, Ventilating and Air Conditioning Engineering (Project NR2016K02), China, and is supported by state key laboratory of compressor technology, China.

References

[1] S.H. Lee, F.W.H. Yik, A study on the energy penalty of various air-side system faults in buildings, Energy Build. 42 (1) (2010) 2–10.
[2] J.Y. Kao, E.T. Pierce, Sensor errors: their effects on building energy consumption, ASHRAE J. 25 (12) (1983) 42–45.
[3] S.H. Yoon, W.V. Payne, P.A. Domanski, Residential heat pump heating performance with single faults imposed, Appl. Therm. Eng. 31 (5) (2011) 765–771.
[4] S.W. Wang, F. Xiao, AHU sensor fault diagnosis using principal component analysis method, Energy Build. 36 (2) (2004) 147–160.
[5] S.W. Wang, Y.M. Chen, Fault-tolerant control for outdoor ventilation air flow rate in buildings based on neural network, Build. Environ. 37 (7) (2002) 691–704.
[6] W. Lee, J.M. House, N. Kyong, Subsystem level fault diagnosis of a building's air-handling unit using general regression neural networks, Appl. Energy 77 (2) (2004) 153–170.
[7] Z.M. Du, X.Q. Jin, Y.Y. Yang, Wavelet neural network-based fault diagnosis in air-handling units, HVAC&R Res. 14 (6) (2008) 959–973.
[8] H. Han, et al., Important sensors for chiller fault detection and diagnosis (FDD) from the perspective of feature selection and machine learning, Int. J. Refrig. 34 (2) (2011) 586–599.
[9] H. Han, et al., Study on a hybrid SVM model for chiller FDD applications, Appl. Therm. Eng. 31 (4) (2011) 582–592.
[10] F. Xiao, S.W. Wang, J.P. Zhang, A diagnostic tool for online sensor health monitoring in air-conditioning systems, Autom. Constr. 15 (4) (2006) 489–503.
[11] Z.M. Du, X.Q. Jin, L.Z. Wu, PCA-FDA-based fault diagnosis for sensors in VAV systems, HVAC&R Res. 13 (2) (2007) 349–367.
[12] N. Kocyigit, M.O. Isikan, AHUX: a knowledge-based expert system to teach troubleshooting of a typical air handling unit leaded by the symptom of the system failure, Energy Educ. Sci. Technol., A: Energy Sci. Res. 30 (1) (2012) 577–590.
[13] F. Xiao, C.Y. Zheng, S.W. Wang, A fault detection and diagnosis strategy with enhanced sensitivity for centrifugal chillers, Appl. Therm. Eng. 31 (17–18) (2011) 3963–3970.
[14] X.B. Yang, et al., A novel model-based fault detection method for temperature sensor using fractal correlation dimension, Build. Environ. 46 (4) (2011) 970–979.
[15] H.T. Wang, et al., A robust fault detection and diagnosis strategy for pressure-independent VAV terminals of real office buildings, Energy Build. 43 (7) (2011) 1774–1783.
[16] W. Härdle, L. Simar, Applied Multivariate Statistical Analysis, second ed., Springer Berlin Heidelberg, New York, NY, 2007.
[17] J.E. Jackson, A User's Guide to Principal Components, first ed., John Wiley & Sons, Inc, New York, NY, 1991.
[18] S.W. Wang, J.T. Cui, A robust fault detection and diagnosis strategy for centrifugal chillers, HVAC&R Res. 12 (3) (2006) 407–428.
[19] S.W. Wang, J.T. Cui, Sensor-fault detection, diagnosis and estimation for centrifugal chiller systems using principal-component analysis method, Appl. Energy 82 (3) (2005) 197–213.
[20] X.H. Xu, F. Xiao, S.W. Wang, Enhanced chiller sensor fault detection, diagnosis and estimation using wavelet analysis and principal component analysis methods, Appl. Therm. Eng. 28 (2–3) (2008) 226–237.

[21] Y.M. Chen, L.L. Lan, A fault detection technique for air-source heat pump water chiller/heaters, Energy Build. 41 (8) (2009) 881–887.

[22] Y.P. Hu, et al., Chiller sensor fault detection using a self-adaptive principal component analysis method, Energy Build. 54 (2012) 252–258.

[23] X. Zhang, C. Sun, Dynamic intelligent cleaning model of dirty electric load data, Energy Convers. Manage. 49 (4) (2008) 564–569.

[24] O. Maimon, L. Rokach, Data Mining and Knowledge Discovery Handbook, second ed., Springer, New York, Dordrecht, Heidelberg, London, 2010.

[25] V. Barnett, T. Lewis, Outliers in Statistical Data, John Wiley and Sons, Chichester, West Sussex, England, 1994.

[26] R. Dunia, S.J. Qin, Joint diagnosis of process and sensor faults using principal component analysis, Control Eng. Pract. 6 (4) (1998) 457–469.