공학석사학위논문

Fault Diagnosis of an Industrial Plant Using Maintenance Record and Multivariate Analysis

실운영 플랜트의 정비기록부와 다변량 분석을 통한 이상감지 및 진단

2019 년 2 월

서울대학교 대학원

화학생물공학부

박 세 진

Abstract

Fault Diagnosis of an Industrial Plant Using Maintenance Record and Multivariate Analysis

Saejin Park

School of Chemical & Biological Engineering

The Graduate School

Seoul National University

Many algorithms have been introduced are for fault detection and diagnosis(FDD) over the years as FDD has been important in the chemical engineering industry. Recent improvements of computation power and advances in statistical techniques, data-driven method have been more popular and well-received approach for FDD. Actual operating process data sets are optimal for FDD algorithm validation but they are hard to acquire and most of FDD algorithm is tested on controlled simulation data for convenience. Preprocess is a crucial part to FDD result, but due to the scarcity of operating process data usage, there are no known specific steps to handle and preprocess actual operating process data.

Preprocess of actual operating data includes 2 parts: maintenance record and sensor raw data. Maintenance record entries are classified into 4 categories by analyzing the content of the entry and trait of maintenance.: corrective, preventive, predictive, periodic maintenance. Only 6 corrective maintenance record is used for FDD as they are the only type that will show fault attributes. Variables of sensor raw data have been reduced from 236 to 28 by analyzing the schematics and analyzing the data tendencies.

Dynamic principal component analysis (DPCA) and 1 class support vector machine (SVM) is used as a FDD algorithm for actual operating plant data. DPCA is used to reduce dimension of data and 1-class SVM is a useful tool to classify actual operating plant data as it only needs a single type of data class to construct a SVM structure. The conventional threshold of SVM classification score is zero, and a negative value is considered as a fault. Proposed new threshold in this study is difference of consecutive score that exceeds 130. If a difference score and following score is more than 130 than it can be classified as a fault. The result of proposed FDD with a new proposed threshold of 6 corrective maintenance record showed great detection accuracy by early detecting 5 fault scenarios.

With the proposed specific steps to preprocess operating process plant data sets and new SVM classification score threshold, accurate and early process detection/diagnosis is possible. Therefore, the proposed methods can help

optimal plant management by detecting a fault early to perform a predictive maintenance.

Keywords: Fault detection, Fault diagnosis, Machine learning, Multivariate analysis, Data-driven approach, Preprocess, 1 class-SVM

Student Number: 2017-29082

Contents

List of Figures

List of Tables

CHAPTER 1. Introduction

1.1. Research motivation

There has been an increasing demand for early and accurate fault detection and diagnosis (FDD) for modern large-scale, highly complex manufacturing facilities. Timely and optimal FDD is very valuable to fault management, whose objective is to increase the safety of plant operation, reduce manufacturing costs and minimize shut-down time. Due to its importance in industrial systems, many academic and industrial effort has been made over the years. Many fields such as pharmaceutical, environmental science have jumped in the FDD studies but process and manufacturing industries have been the field of FDD longer as the result of FDD immediately can influence the product quality, reduction of product rejection and satisfaction of safety and environmental regulations. Especially for chemical plants, which handles toxic, hazardous, flammable raw materials, the safety of the process is of an utmost importance[1].

A fault is defined as a state of one or more unexpected deviation from acceptable, normal operating condition.[2]. A fault may cause loss of required performance, possibly initiating a permanent interruption of a system's ability to perform, failure. It is a role of FDD to detect the fault before it aggravates into failure or malfunction[3].

Fault detection is determining if a fault has occurred and early detection of a fault may be able to give a warning on emerging problems to prevent serious

events with appropriate actions. Fault diagnosis is determining the cause of the fault. Some process faults may easily be detected and diagnosed with Shewhart charts, a control chart with upper and lower control limits that are typically at 3 standard deviations from mean. Shewhart chart is effective with univariate fault but most of the modern industrial systems consist of many complex units. Therefore, the Shewhart chart has limited detection ability against modern multivariable processes. In addition, many faults are not easy to identify as sensor values changes are hidden under noise and disturbances. Therefore, nowadays instead of simple control charts such as Shewhart chart, three methods are frequently used for FDD: model-based, knowledge-based, data-driven methods. In this study, data-driven methods will be used as the importance of data have emphasized over the past decade along with advancement in statistical studies as well as computation power.

Many FDD algorithms are tested with simulation data sets, which are easy to intentionally insert a type of fault of choosing at a designed time, therefore, making it a good data to test the FDD algorithms. Despite all the convenient usage of simulated data, no matter how much a simulated data set might resemble characteristics of actual data, it is still a simulated data not the actual data. It would be optimal choice to utilize actual real data from operating process plant but detailed operating real data sets are hard to acquire. Also unlike controlled simulate data where preprocess is unnecessary, actual operating plant data needs to be preprocessed with many considerations. Unfortunately, there are no known specific steps to handle and preprocess actual operating process data.

1.2. Research objectives

Most commonly used data-driven multivariate analysis method for FDD is Principal Component Analysis(PCA). PCA reduces the dimensionality of a large set of multivariable data set by transforming to a new set of variable, the principal components PCA is a tool that projects the reduced process data to latent space and can identify linearly related variables. PCA is used in FDD area with a threshold set by Hotelling T-squared or squared prediction error(SPE). PCA is a very effective tool for analyzing linear and static data but has limitation to describe dynamic and non-linear data. For dynamic and non-linear data, dynamic principal component (DPCA) is developed to capture the time-dependent correlation of data.

Support Vector Machine(SVM) is a statistical supervised learning technique that is used for classification and regression. SVM is widely used for its superior classification ability and with a proper kernel function, SVM is a powerful tool to classify non-linear data by constructing n-1 dimensional hyperplanes that can classify n-dimensional data with two different classes. The binary classifier is used for FDD by constructing SVM structure after first a characteristic of a faulty state is categorized. SVM structure will find the difference between the normal and faulty state in the new input data by analyzing the classifier scores of SVM model.

1.3. Description of the equipment used in this thesis

Fluid Catalytic Cracking (FCC), since the first industrial start-up in 1942, is still one of the most important conversion processes used in petroleum refineries. There are about 400 FCC unit in operation worldwide and continues and will play a major role as FCC process will be used for biofuels and for reduction of CO2 emissions[4]. The main objective of FCC unit is to convert high-boiling petroleum fraction gas oil and various heavy hydrocarbons to lighter and high-valued transportation fuels i.e. gasoline, jet fuel, and diesel. Main air blower (MAB) takes atmospheric air and delivers it to the regenerator for quick burn-off.

According to recent studies on the health of FCCU, major rotating equipment accounts for over 35% of all FCCU shutdown and air blower failures account for just under 50% of major rotating equipment failures. Considering the fact that scheduled shutdowns make up less than 10% of FCCU shut downs, it is safe to assume that MAB is one of the main cause of FCCU failures. That is why MAB data is used in this study for fault detection.

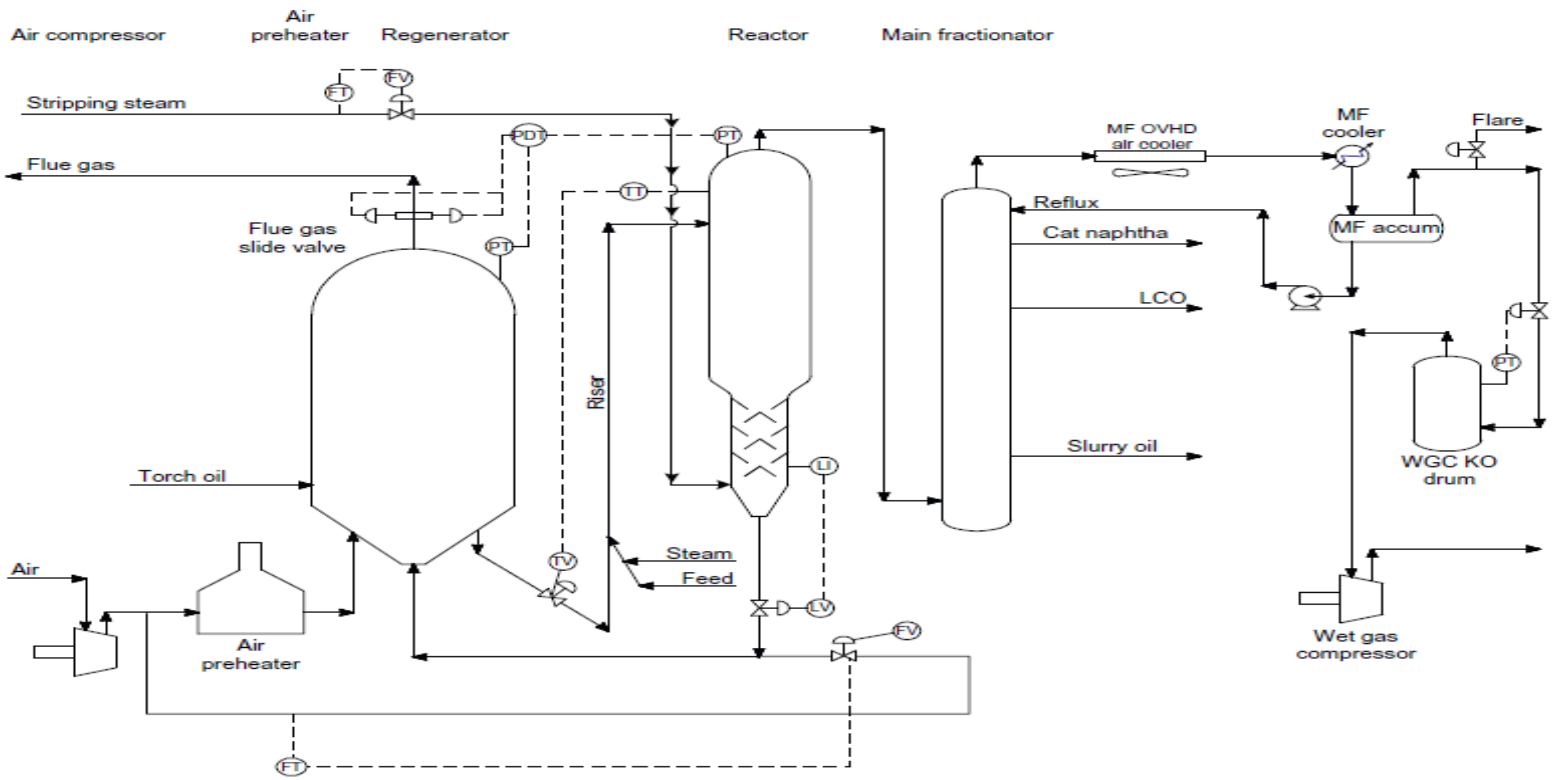

Figure 1.1 Schematic diagram of a typical FCCU process.

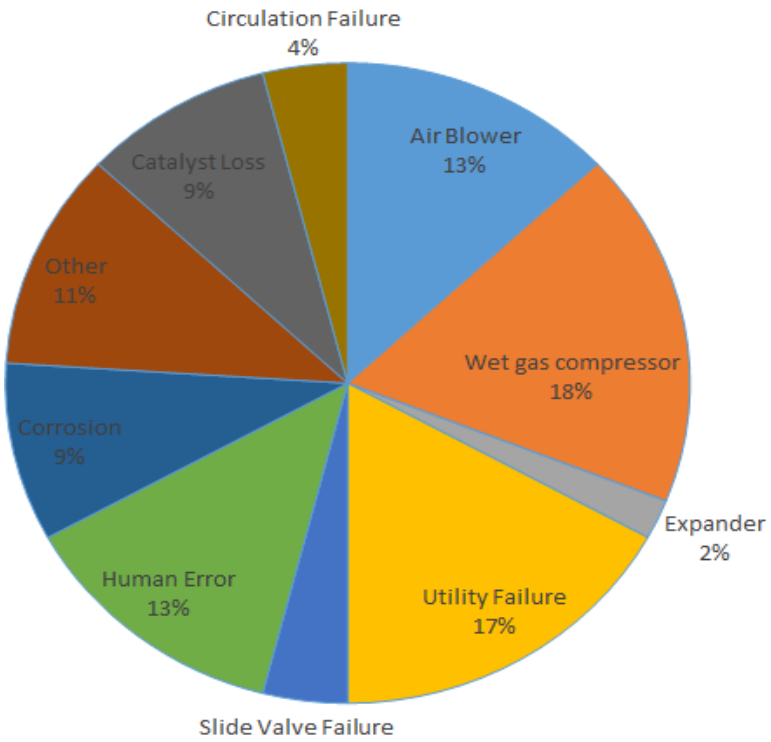

Figure 1.2 Cause of FCC unit outrage. Air blower, wet gas compressor and expander make up 33% of total outrage.

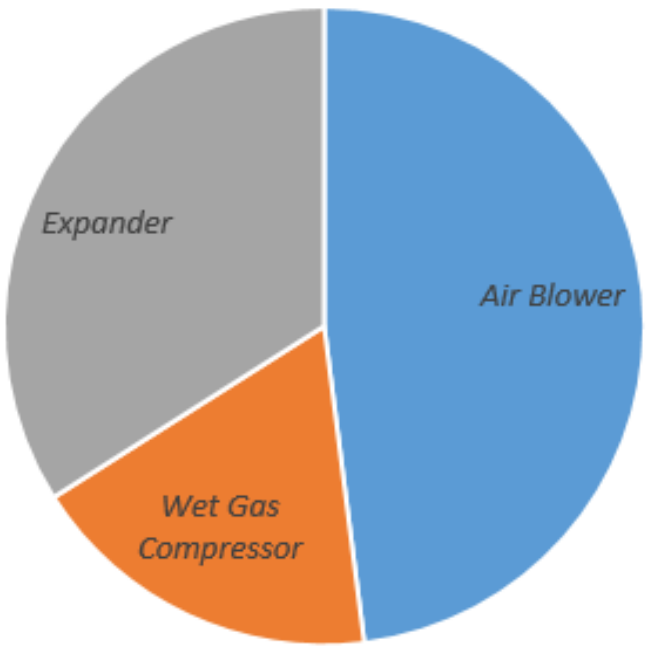

Figure 1.3 Breakdown of major rotating equipment failures. Air blower failures account for almost 50% of total failure

1.4. Outline of the thesis

Various methods have been developed for the purpose of accurate and early fault detection/diagnosis and most commonly used methods are (D)PCA and SVM. Those methods have proven to be very accurate with controlled simulated data. However, unlike controlled simulated data, actual operating plant data have to account for much more unpredictable variables such as climate and personnel mistakes and real data does not have categorized normal training data and abnormal test data sets for algorithm validation. It is a duty of a data analyst to classify the useable training and test data with meticulous preprocess procedure. To apply SVM for fault detection, 2 different class of data, normal and fault, is required as training data to classify the test data set. Since only maintenance records and raw data with tag list is provided without any classification labels, 1-class SVM is more suitable for the actual operating plant because it only requires normal data set for training. Also for FCC, which is a continuous process, a method that takes into account the serial correlations in the data, DPCA, is more suitable than PCA that takes only one time instant that is statistically independent to previous observations. DPCA has been proven to show higher accuracy for a process with long sampling times, i.e., 2 to 12 hours and in this study is used to reduce data dimension for SVM classification.

Most of developed FDD algorithms are tested with controlled simulation data that a researcher can easily control and manipulate. Advantages of simulation data are that many different types of fault can be injected at a designed time by an analyst. No matter how the accurate the simulation model may be, it is still

‘simulated’ data, which may not show the same accurate result when applied to a ‘real’ data. Conversely, ‘real’ data needs more polishing work than simulated data as unpredictable variables, noises may be hidden, and it’s a very hard task to acquire a data set with various types of faults to use for diagnostic classification.

1.5. Outline of the thesis

In this study, a detailed procedure of preprocessing industrial plant raw data, which consist of maintenance record and sensor data, and a FDD algorithm is introduced for early fault detection that will help determine the need for predictive maintenance. Unlike conventional preprocess procedure, that is usually scaling, eliminating missing data and etc., preprocessing raw industrial plant data requires two parts, maintenance record and raw sensor data. The steps to classify maintenance and find the relevant data from vast mass of raw data records to find adequate fault example is suggested. For FDD algorithm, 1-class SVM via DPCA with new threshold is used as FCC is a dynamic process and only unlabeled/unclassified data set is available.

CHAPTER 2. Methodology

2.1. Multivariate analysis methods

2.1.1. Principal component analysis

PCA is a useful dimensionality reduction technique method used to transform a set of correlated variables into a smaller set of new variables that are uncorrelated and express the data in such a way as to highlight their similarities and differences[5]. It determines a set of orthogonal vectors called loading vectors, ordered by the amount of variance explained in the loading vectors direction. Given a training set of n observations and m process variables, with mean zero and unit variance, stacked into a matrix X, where the matrices $\widehat{\mathbf{X}}$ and E represent the modeled and un-modeled variations of X. l represents the number of principal components and T and P are the score and loading matrices. Loading vectors P, are calculated by solving stationary point of the optimization problem using singular value decomposition (SVD) of the covariance matrix S. The loading vectors are the orthonormal column vectors in the matrix Y, and the variance of the training set projected along the ith column of Y is equal to σ^2. Solving (0.0) is equivalent to solving an eigenvalue decomposition of the sample covariance matrix S.

$$X = \widehat{\mathbf{X}} + E \tag{2.1}$$

,

$$\mathrm{T} = \mathrm{XP} \tag{2.2}$$

$$\hat{\mathrm{X}} = \mathrm{TP}^{\mathrm{T}} = \sum_{i=1}^{t} t_i P_i^T \tag{2.3}$$

$$\mathrm{E} = \mathrm{T_e P_e^T} = \sum_{i=l+1}^{t} t_i P_i^T \tag{2.4}$$

$$\mathrm{X} = \mathrm{TP}^{\mathrm{T}} + E = \sum_{i=1}^{a} t_i P_i^T + E = \hat{\mathrm{X}} + E \tag{2.5}$$

$$\mathrm{S} = \frac{1}{\mathrm{n}-1}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{V}\Lambda\mathrm{V}^{\mathrm{T}} \tag{2.6}$$

It is very important to choose the optimal number of principal components, a, because TP^{T} represents the principal sources of variability in the process, and E represents the variability corresponding to process noise.

For the FDD algorithm for this study, PCA scores of 1st and 2nd component were used in this study to construct SVM structure and as input for a classifier.

2.1.2. Hotelling's T-squared and squared prediction error

Hotelling's T-Squared and Q statistics, squared prediction error(SPE), is the multivariate analysis used to detect outlier in projected coordination of PCA[6]. The Hotelling's T-Squared statistics measure the variation within the PCA model for the lower-dimensional space for each new observation and can be calculated as below :

$$T^2 = x^T P(\Sigma_a)^{-2} P^T x \tag{2.7}$$

Where x is new observation and Σ_a contains the non-negative real eigenvalues corresponding to the number of principal components, a, and P contains the loading matrix of X. The graphical definition of Hotelling's T-squared value is Euclidean distance from the origin of PCs to the data point.

$$T^2 = x^T P \Lambda_\alpha^{-1} P^T x \tag{2.8}$$

The upper confidence limit of T^2 is calculated using the F-distribution :

$$T^2_{a,n,\alpha} = \frac{a(n-1)}{n-a} F_{(a,n-a,\alpha)} \tag{2.9}$$

where n is the number of samples in the data and α is the level of significance of F-distribution. A violation of the upper confidence limit is considered an outlier or a fault.

SPE's graphical definition could be explained as Euclidean distance from the PC's hyperplane to the data point. The magnitude of value of SPE expresses how far the data point is from the normal tendency of the other data. SPE can be calculated as below :

$$\mathrm{SPE} = x^{\mathrm{T}}\left(\mathrm{I} - \mathrm{PP}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathrm{I} - \mathrm{PP}^{\mathrm{T}}\right)x \tag{2.10}$$

where P contains loading vectors(orthogonal) of X and I is identity matrix. Upper limit of SPE is also calculated using approximate distribution.

$$\mathrm{SPE}_{\alpha} = \theta_1\left(\frac{\mathrm{h_0 C_\alpha}\sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 \mathrm{h_0}(\mathrm{n_0} - 1)}{\theta_1^2}\right)^{\frac{1}{\mathrm{h_0}}} \tag{2.11}$$

Where

$$\theta_1 = \sum_{\mathrm{j=a+1}}^{\mathrm{m}} \lambda_{\mathrm{j}}^{\mathrm{i}} \tag{2.12}$$

$$\mathrm{h_0} = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \tag{2.13}$$

C_α is the normal distribution value with the level of significance of F-distribution, α, and λ_{j} is the j-th largest eigenvalue of X. Any values that violate the threshold is considered as a fault.

Contribution plot is bar plot of the sum of the residuals that describes how much each variable contribute to the T-squared and SPE values at a specific observation.

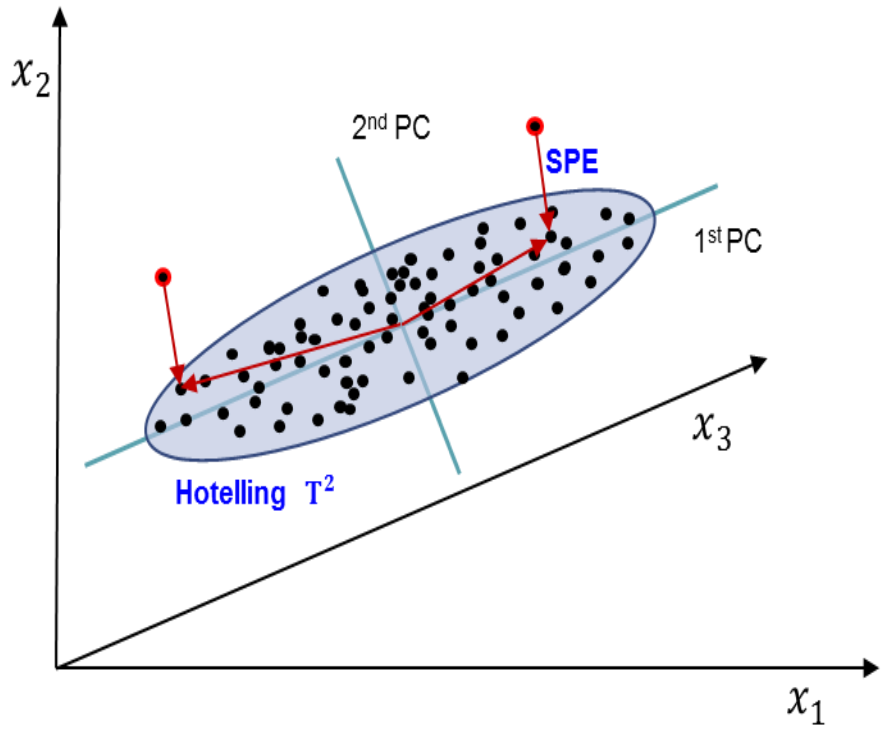

Figure 2.1 T- squared and SPE value described in three dimensional space. The distance between original observation to the PC plane is value of SPE and the distance from the origin of the PC plane to the projected observation point is the value of Hotelling T-squared.

2.1.3. DPCA

X is a set of data composed of n_t observations from p variables where $n_t \geq p$.

$$X = [X_1\ X_2\ \cdots\ X_P]_{(n_t \times p)} \tag{2.14}$$

PCA can be used to take into account the serial correlations by augmenting each observation with the previous w observations and stacking the data matrix.

$$\boldsymbol{X}_i^{\,w} = \begin{bmatrix} X_i(1) & X_i(2) & \cdots & X_i(w) \\ X_i(2) & X_i(3) & \cdots & X_i(w+1) \\ \vdots & \vdots & \ddots & \vdots \\ X_i(n_t - w + 1) & X_i(n_t - w + 2) & \cdots & X_i(n_t) \end{bmatrix}_{(n \times w)} \tag{2.15}$$

$$\boldsymbol{X}^w = \begin{bmatrix} \boldsymbol{X}_1^{\,w} & \boldsymbol{X}_2^{\,w} & \cdots & \boldsymbol{X}_p^{\,w} \end{bmatrix}_{(n \times m)} \tag{2.16}$$

$$n = n_t - w + 1, \qquad m = pw \tag{2.17}$$

w is the 'time lag shift', a trajectory matrix applied to stack up the data set for dynamic process[7]. Since variables in a multivariable analysis approach may have different range of values, it is convenient to perform a standardization of data with mean and standard deviation to obtain data matrix, Xs^w, with zero mean and unit variance. DPCA is applying PCA on Xs^w.

$$xs_{ij}^{\omega} = \frac{x_{ij}^{\omega} - \mu_j}{\sigma_j} \text{ for } i = 1, \ldots, n, \qquad j = 1, \ldots, m \tag{2.18}$$

For fault detection, the measures for each observation variable can be calculated by adding all the values of measures corresponding to previous w time lags.

2.2. Support Vector Machine

2.2.1. SVM

Support vector machines (SVM) is a supervised learning technique that was originally developed for classifying data from two different classes. The basic principle is illustrated in fig. that shows the classification of a series of points for two different classes of data, yellow triangles, blue squares.

The SVM classifier structure can be constructed by solving the optimization problem that minimizes the distance between hyperplane and data points[8]. In the equation n is the number of data points, α is the Lagrange multiplier, x is the data point and y_i are either 1 or -1.

$$\max \tilde{L}(\alpha) = \left[\sum_{i=1}^{n} \alpha_i - 0.5 \sum_{ij} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right] \tag{2.19}$$

K is a kernel function that is used to handle non-linear issues using linear classifier. Following qualities from Karush-Kuhn-Tucker (KKT) condition must be satisfied in the objective function.

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2.20}$$

LaGrange multiplier of hyperplane can be determined by solving objective function Equation 2.21. α parameter for the hyperplane can be calculated with

following equations where N_{sv} represents the number of all deciding boundary points or support vectors.

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{2.21}$$

$$b = \frac{1}{N_{sv}} \sum_{i=1}^{N_{SV}} (K(w, x_i) - y_i) \tag{2.22}$$

In order to construct SVM, optimal kernel function must be determined first. Some of the most commonly used kernel functions are linear, polynomial and radial basis. Linear kernel function is used if data points can be separated with linear hyperplanes. However, data from Process plant are non-linear data and radial basis function, or also called Gaussian kernel function is widely used as the nonlinear kernel function to construct SVM[9, 10]. Radial basis function is commonly used for non-linear data as it can separate data of different class in the form of hyper-sphere with nonlinear hyperplanes.

Linear $$K(x_j, x_k) = x_j^T x_k \tag{2.23}$$

Polynomial $$K(x_j, x_k) = (x_j^T x_k + C)^b \tag{2.24}$$

Radial basis function $$K(x_j, x_k) = \frac{exp(-\|x_j - x_k\|)^2}{\delta^2} \tag{2.25}$$

Sigmoid $$\mathrm{K}(x_\mathrm{j}, x_\mathrm{k}) = \tanh(k(x_\mathrm{j}, x_\mathrm{k}) + \upsilon) \quad (2.26)$$

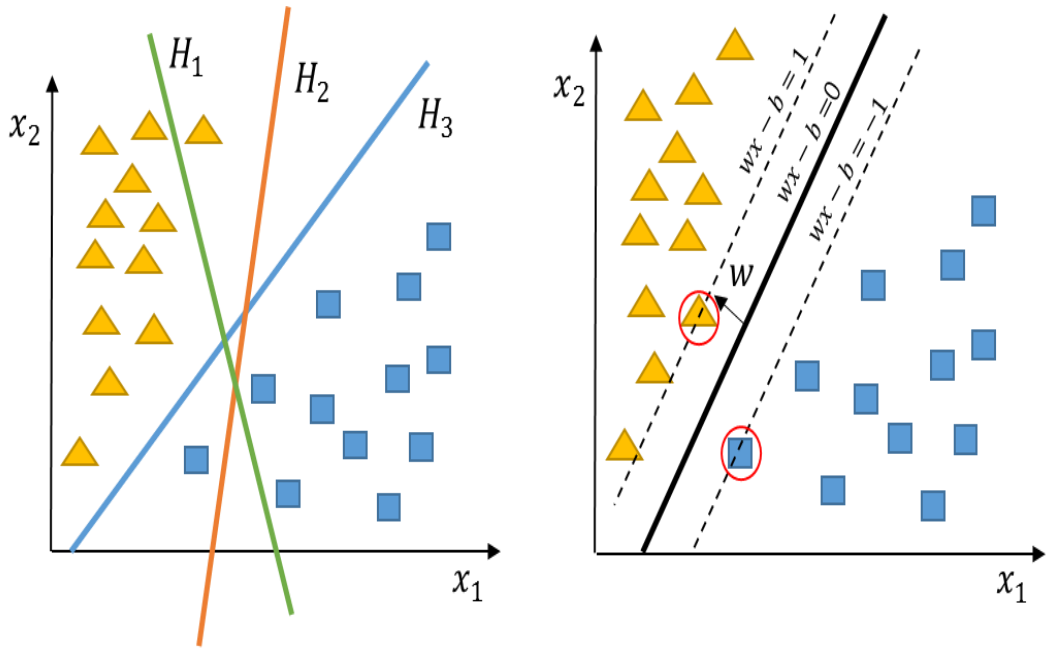

Figure 2.2 Example of deciding optimal support vector

2.2.2. 1-class support vector machine

1-class SVM is extended variant algorithm of SVM proposed by Scholkopf[11]. This divides possible faulty outlier data from normal data. Unlike traditional SVM, 1-class SVM only needs normal data to construct SVM structure. 1-class SVM computes the surface of a minimal hypersphere or margin support that includes sample normal training data. It also uses the kernel function similar to traditional SVM to map the non-linear data to a feature space[12]. Also, radial basis function is used as the data being tested is non-linear data and radial basis function is the optimal kernel function for non-linear data. 1-class SVM maximize the perpendicular distance from the origin, a test data from fault class. In summary, if a test data is outside the 'boundary' made up with single characterized data of a single class, it is classified under fault category. The modified optimization problem is as follows.

$$min_{w,\xi_i,b} \frac{1}{2}\|w\|^2 + \frac{1}{vm}\sum_{i=1}^{m}\xi_i + b$$
$$subject\ to\ w \cdot \phi(x_i) + b + \xi_i \geqslant 0, \quad \xi_i \geqslant 0, \quad i = 1 \ldots m$$

Score which expresses the distance from 'boundary' of normal samples, can be calculated using 'predict' function in MATLAB. If the value of score is greater than zero, it is inside the boundary which indicates it's normal and if the value is below zero, it indicates that it is outside the boundary thus classifying the test data as fault.

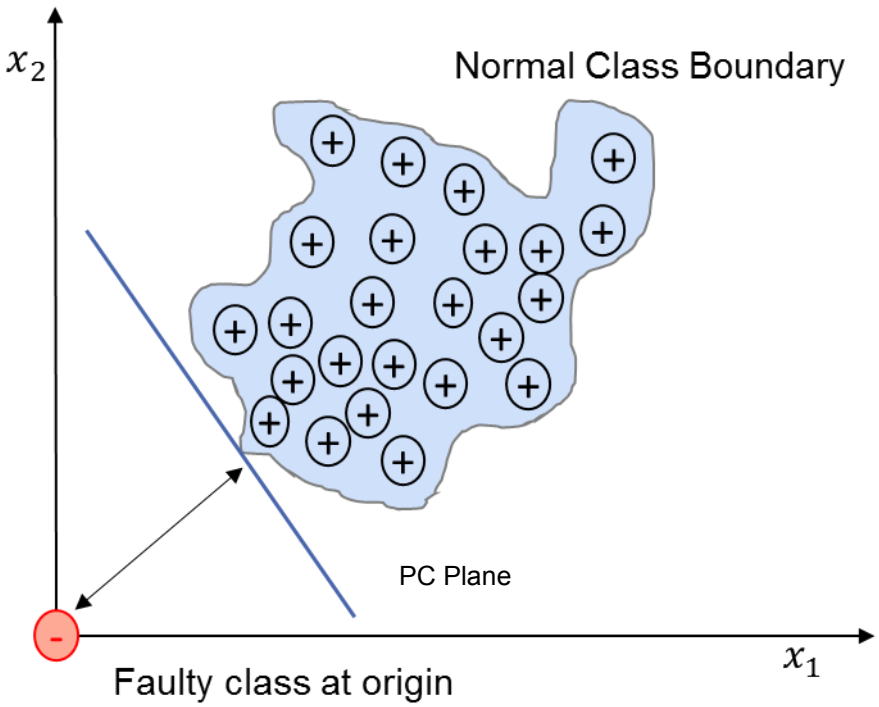

Figure 2.3 Graphical illustration of 1-class SVM

CHAPTER 3. Simulation

3.1. Process of pattern recognition

In accordance with the process of pattern recognition, which is shown in Figure 3.1, after acquiring the raw industrial plant data, preprocessing the raw data by eliminating irrelevant data and choosing the adequate case study to verify the algorithm[13]. Adequate case study can be determined by analyzing the maintenance record. After the preprocess, a dimension of chosen data will be reduced via PCA or Dynamic PCA. Utilizing the first and second score of PCA/DPCA, designated normal data set will construct SVM structure by 1-class SVM algorithm. Corresponding test data set will be PCA/DPCA score of chosen test data time span. SVM classification score can be calculated with given SVM structure and test data set.

SVM classification score is the signed distance from x to the decision boundary ranging from -∞ to +∞. A positive score class indicates that x is predicted to be in that class and a negative score indicates otherwise. In other words, if the score is a positive number, then it's normal and if the score is a negative number, then it's classified as a fault.

Figure 3.1 Process of pattern recognition

3.2. Preprocessing

Statistical analysis and machine learning algorithms learn from data. Therefore, the more disciplined the data is the more consistent and better result will be. Data preprocessing is a data mining technique that disciplines raw data into an understanding format. Data preprocessing is a proven method of resolving incomplete, inconsistent characteristics of real data[14].

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis

Machine learning algorithms learn from data. It is critical that you select the right optimal data for the problem. Even if a good data is acquired, data needs to be on a useful scale, format and include all the meaningful features.

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

3.2.1. Maintenance record

The first step of Pre-processing of maintenance record is categorizing the entries by type of maintenance. Typically, maintenance can be categorized into 4 categories, predictive maintenance, preventive maintenance, corrective maintenance, periodic maintenance.

Most of maintenance being performed are periodic maintenance and preventive maintenance. Periodic maintenance is the basic maintenance of equipment consists of elementary tasks. The maintenance interval is usually provided by the manufacturer to carry out. Preventive maintenance is maintenance that is regularly performed on a piece of equipment to reduce the risk of failures such as oil change, lubrication, and minor adjustments. Periodic maintenance and preventive is similar in a way that maintenance is done on the equipment while it's still on an operation and done to avert sudden failing. The main difference between periodic and preventive maintenance is that while periodic maintenance is based on time intervals, preventive maintenance is based on experience of past failures and weather changes.

Corrective Maintenance is any maintenance performed to restore the failed equipment to an operational condition. Unlike periodic and preventive maintenance, corrective maintenance is performed after a fault or failure has occurred to the equipment. Therefore, it is the type of maintenance an operator would want to avoid the most, as the damage has already been done in one way or another. That is why periodic and preventive maintenance is performed more

frequently despite the labor and part cost. Compared to even short process shut-down, opportunity cost of periodic and preventive maintenance is much cheaper.

The goal of predictive maintenance is to correctly predict failure of equipment and to prevent the failure by performing maintenance with necessary information and analysis. Fault detection is the key to a successful predictive maintenance.

At times depending on the format of the maintenance record, maintainer is required to classify the type of maintenance on the record. In this study case, there was no given classification by the maintainer. Therefore, intuition and knowledge of the process is needed to categorize the maintenance records. For example, charging of nitrogen is classified as periodic maintenance, as it is required by the operating manual on regular basis. On the hand, repair of air tube leak is classified as corrective maintenance, as it is actual correction of a failure of an equipment.

Time and Preventive maintenance was the majority of the maintenance record entries. Out of total of 63 maintenance record, 58 entries were periodic and preventive maintenance which made up for 90% and corrective maintenance record was only 5 entries. In an operating plant, time and preventive maintenance make up over 90% of the maintenance as shut-down due to a fault is critical to plant operation. Also there is a 24-hour crew on stand-by at an operating Plant to monitor any abnormalities before it escalates into a serious disastrous situation. Therefore, it only makes sense that there is a very few actual “serious” distinctive fault that requires corrective maintenance.

Although corrective maintenance is best suited data to test fault detection algorithm, number of data was too small. In addition to 5 corrective maintenance cases, 12 preventive maintenance cases selected for verification. Added 12 Preventive maintenance was chosen as they may now show distinctive fault characteristics as corrective maintenance but still may show more than periodic maintenance.

3.2.2. Raw data

The acquired raw data consists of total 2 years of observation in interval of minutes, 1440 data per day, with 238 sensor variables. Since processing 238 variable data with algorithm takes up too much time, it is necessary to reduce the number of variables into relevant variables. To reduce the number of variables, analysis of P&ID of the process is needed. Tag numbers and symbols on the P&ID schematics provide with the information of location, type of the instrument. By studying the P&ID, instruments that are part of a distributed control system(DCS) can be eliminated since they are display panels like a terminal screen that most likely has same values as actual sensors. Piping and connection symbols are also used to identify how the instruments in the process connect to each other and applied that information to eliminate data link and electrical panel related tag lists. After the process of eliminating irrelevant variables, 210 variables were deleted from the data and left with 28 variables.

3.3. Selecting optimal data set for validation

It is important to recognize that maintenance entry dates are inputted after maintenance action is completed, not when the maintenance action order has been issued. Also maintenance is performed by 24-hour crew, therefore the fault could have occurred during the night before the entry. That is the reason when determining dates for abnormal test data, day prior to the date of the entry must be included as well as day after to show change in the test data set.

Unlike controlled simulation data sets that have specified normal data set and abnormal data set, operating plant data does not have named normal and abnormal data sets. It is very important to choose the optimal set of normal and abnormal data for validating fault detection algorithm for even if the algorithm is flawless, data set decision could critically influence the result. During the analysis, a gradual increase in temperature in some variables was noticed. Although there was some gradual temperature increase over a span of several months, they were not caused by fault but part of normal pattern. That also means that the pattern of normal data changed as time passed. With change of the normal data patterns over time, 5,000 observations (about 3days of data) prior to day prior to entry was chosen to train.

Selected maintenance record will be the standard for data selection for optimal training and test data sets. PCA/DPCA is applied to reduce their dimensions and product of PCA and DPCA, first and second principal component score plot is used to construct SVM structure and calculate the SVM classification score.

Train Data	Test Data		
5000 observation	D-1	Date of Maintenance D+0	D+1

Table 3.1 Data selection criteria

3.4. Algorithm Validation

3.4.1. Tennessee Eastman Process

Tennessee Eastman Process(TEP) model that was first introduced by Downs and Vogel[15] is most widely used data set when validating FDD algorithms. Its popularity is because of the fact that it is modeled based on a real process and is very convenient to insert fault of user's choice at a time of one's choosing. Through numerous years of testing with TEP data sets, the minor glitches have been improved by engineers and are acknowledged by many as one of the an 'almost actual' data. Moreover, TEP has designated 'Normal Data' for training data set and 'Abnormal Data' for each fault scenarios to input as test data set. The types of fault provided with TEP data sets are listed in the table 3.2. For validation of the proposed algorithm, fault scenario 1&7 were chosen as the test data sets as they are known to have high detection rate than other scenarios. For this TEP data set, fault was inserted at 975 second which will be the standard for accuracy of detection time.

The result of FDD algorithm with step fault #1 & 7 data set as input is shown in figure 3.3 and figure 3.4. The conventional detection time, when a classification score goes below threshold '0', for fault #1 is 1276 second and 984 second for fault #7. Considering the fault was inserted at 975 second, the detection time for fault #1 was descent and detection time for fault #7 was very close to the time of insertion.

It was noticed from analyzing DPCA-SVM classification plot that normal scores before the classified fault data score are very apart than distance of

normal scores. Average of differences of normal consecutive classification scores was 1.03. But the score difference of observation when a fault is classified, 1276 sec, and 1 observation prior, 1275 sec was 136. Similar difference characteristics were observed for fault #7. Therefore, a new threshold of consecutive score value difference larger than 126 was established and the result of new threshold detection time was at least 5 observations faster than the conventional detection time. As the goal of fault detection is to detect early to prevent aggravation of fault, the new threshold is more suitable for FDD than the conventional threshold.

Scenario	Process variable	Type
1	A/C feed ratio, B composition constant (stream 4)	Step
2	B composition. A/C ratio constant (stream 4)	Step
3	D feed temperature (stream 2)	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser c4mting water inlet temperature	Step
6	A feed loss (stream 1)	Step
7	C header pressure loss-reduced availability (stream 4)	Step
8	A, B, C feed composition (stream 4)	Random variation
9	D feed temperature (stream 2)	Random variation
10	C feed temperature (stream 4)	Random variation
11	Reactor cooling water inlet temperature	Random variation
12	Condenser cooling water inlet temperature	Random variation
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16	Unknown	Unknown
17	Unknown	Unknown
18	Unknown	Unknown
19	Unknown	Unknown
20	Unknown	Unknown

Table 3.2 Detailed description of TEP faults.

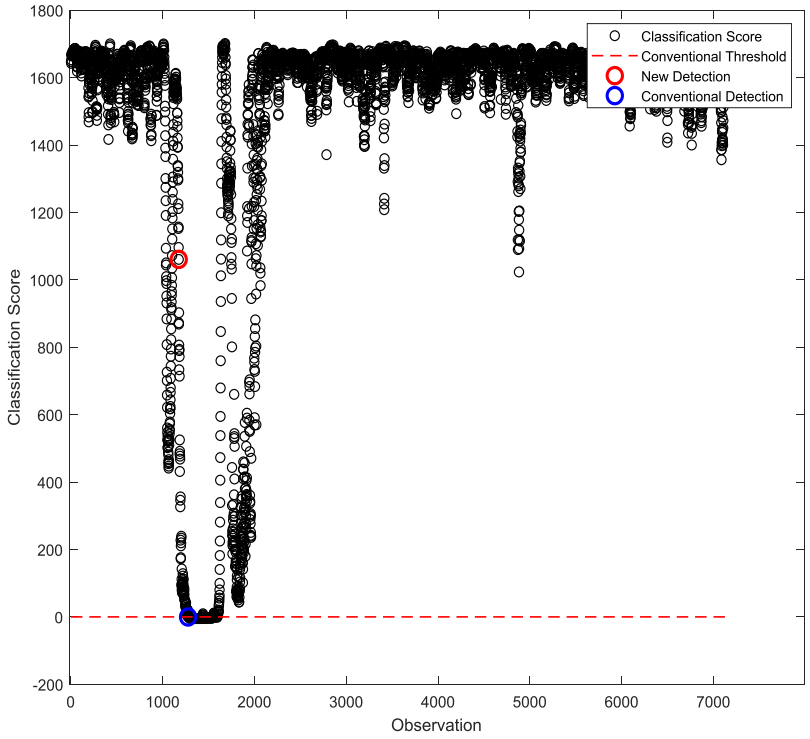

Figure 3.2 Fault detection result of TEP #1 fault with DPCA –SVM

(Conventional detection: 1276, New detection: 1174)

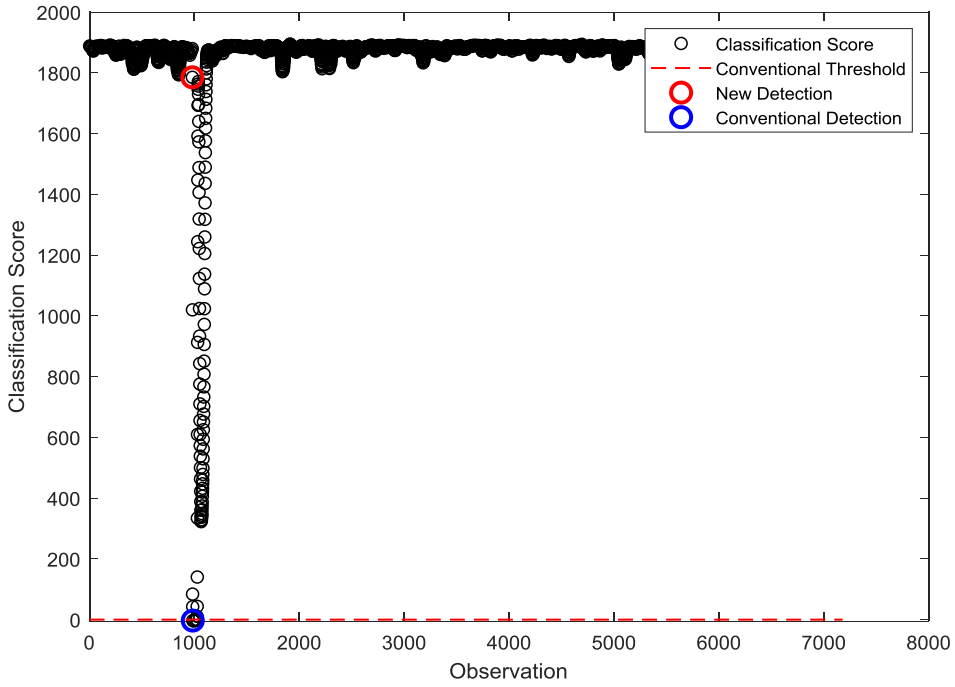

Figure 3.3 Fault detection result of TEP #1 fault with DPCA –SVM

(Conventional detection: 984, New detection: 979)

CHAPTER 4. Result

4.1. Fault detection and diagnosis

Fault detection and diagnosis using 1-class SVM via DPCA have proven accurate on TEP data. The proven algorithm was used to early fault detect MAB from FCCU data. To apply the algorithm, it is necessary to extract relevant data from a vast amount of data. The relevant data, 6 corrective maintenance entries which were classified from maintenance record entries and variable reduced data set are obtained by preprocessing procedure. 236 variables from raw data are reduced to 31 relevant variables and data of observation from maintenance entry is reduced via PCA and DPCA. 7 principal components are chosen as they explain 99.5% variance. 1st and 2nd scores from PCA/DPCA of training data set is used to construct the SVM classifier. Table 3.1 shows how training and data sets are selected based on the date of corrective maintenance entry. The standards are set due to completion of maintenance action and gradual increase of sensor data. Among the industrial plant data in the span of 2years, 6 corrective maintenance record was used for FDD as shown in Table 4.1.

	Date	Maintenance Entry	Category
1	Y1-Feb1	Intake Filter Calibration	corrective
2	Y1-June1	TBN NDE Hunting(TI253B)	corrective
3	Y1-June2	Set Point Error	corrective
4	Y1-Aug1	Repair Air Tube Leak	corrective
5	Y1-Aug2	Repair dryness of the sensor(TE257B)	corrective
6	Y2-Mar1	Inspect Cabinet Alarm RCP4	corrective

Table 4.1 Classified corrective maintenance record

4.1.1. Fault detection and diagnosis result

First tested data set is from entry Y1-June1 'TBN NDE hunting(T23B)' because it is fairly easy to notice the fault time from normalized data plot and also this is one of the 2 cases that specified fault tag list to validity diagnosis part as well. Figure 4.1 shows the plot of normalized data, PCA T^2 plot, PCA-SVM classification score plot and contribution plot. For detection with PCA T^2 plot, 95% confidence plot is used as the threshold and it showed many false alarms. PCA-SVM score plot showed fault scores below 0 with less false alarms. However, contribution plot with the time detected by PCA-SVM had inaccurate result. On the other hand, DPCA T^2 and DPCA-SVM showed much more accurate results compared to PCA T^2 and PCA-SVM score detection. Detection time of PCA SVM was 2237 min, and DPCA-SVM 2224 with threshold 0 and 2221 with new threshold of 100 consecutive score difference.

DPCA-SVM detection with new threshold was able to detect 16 min earlier than PCA-SVM and 3 min earlier than general DPCA-SVM threshold as shown in figure 4.2. 16 min and 3 min difference might not seem a big difference on paper, but to a 24 hour operating continuous process plant, it could save millions of dollars by preventing a critical stoppage of operation. Besides the early detection time, most distinct difference between PCA and DPCA is the accuracy of diagnosis via contribution plot. Precision of diagnosis via contribution plot heavily relies on the accuracy of detection time. Since FCC process is a dynamic process and DPCA is keen on dynamic process, the diagnosis via DPCA was in accordance with fault tag list specified on the maintenance record, T23B.

Second test data set, Y1-Aug2, is similar to first test data set with a specified fault tag list. Also as shown in the figure 4.3 and figure 4.4, the result accuracy was similar to first test data set, Y1-June1. PCA T^2 and DPCA T^2 plots showed great peak difference but it also showed numerous false alarms against 95% threshold. Fault detection time with PCA-SVM is 2319 min and 95 minutes earlier with DPCA 2224 with conventional threshold and 100 min earlier with new proposed threshold, 2219 min. Diagnosis result with PCA contribution was incorrect, but DPCA-SVM contribution plot result was consistent with the fault tag list identified on the maintenance record.

Third test data set, Y1-June2, only has maintenance entry without fault tag list mentioned. The normalize data plot shows characteristic of obvious step

fault and PCA T^2 plot also showed obvious peak at the time of the step fault. However high the peak value of PCA and DPCA T^2 may be, it still had numerous false alarms that surpassed 95% confidence limit. The normalized data showed 1 min of fault at 2319 and PCA T2 and PCA-SVM result showed exactly 1 min of fault at 2319 min. But DPCA showed more accurate, earlier and longer detection time than other results at 2221 min. Fourth test data, Y1-Feb1, also had similar result but PCA-SVM had earlier detection time by 2 min. as shown in figure 4.5 and 4.6.

The other 2 test data sets showed different from other data sets that showed step faults. Fault Y1-Aug1 did not have a clear cut fault results but the algorithm was able to accurately detect the fault as accuracy is determined by if the score recovers after the date of maintenance record. Even if the classification score of DPCA-SVM did not go below conventional threshold, the new threshold of score difference with 130 indicates fault which coincides with possible fault time window. However, all of fault Y2-Mach1 FDD result indicated that it had deviated after the intended time of fault occurrence, within 1440min. Due to the false alarm, this will be the only case of failed detection. However, newly proposed threshold showed only one other false alarm which is much less than other methods, proving that it's most reliable method of all.

Fault scenario	PCA	DPCA		Remark
	Conventional	Conventional	New Threshold	
Y1-Feb1	2052	1963	1954	New Threshold
Y1-June1	2237	2224	2221	New Threshold
Y1-June2	2320	2226	2221	New Threshold
Y1-Aug1	31~150	NA	1939	New Threshold
Y1-Aug2	2319	2224	2219	New Threshold
Y2-Mar1	83~800 5150~5439	170~500 5160~5300	637 / 5695	Detection Fail

Table 4.2 Fault detection time result for corrective maintenance

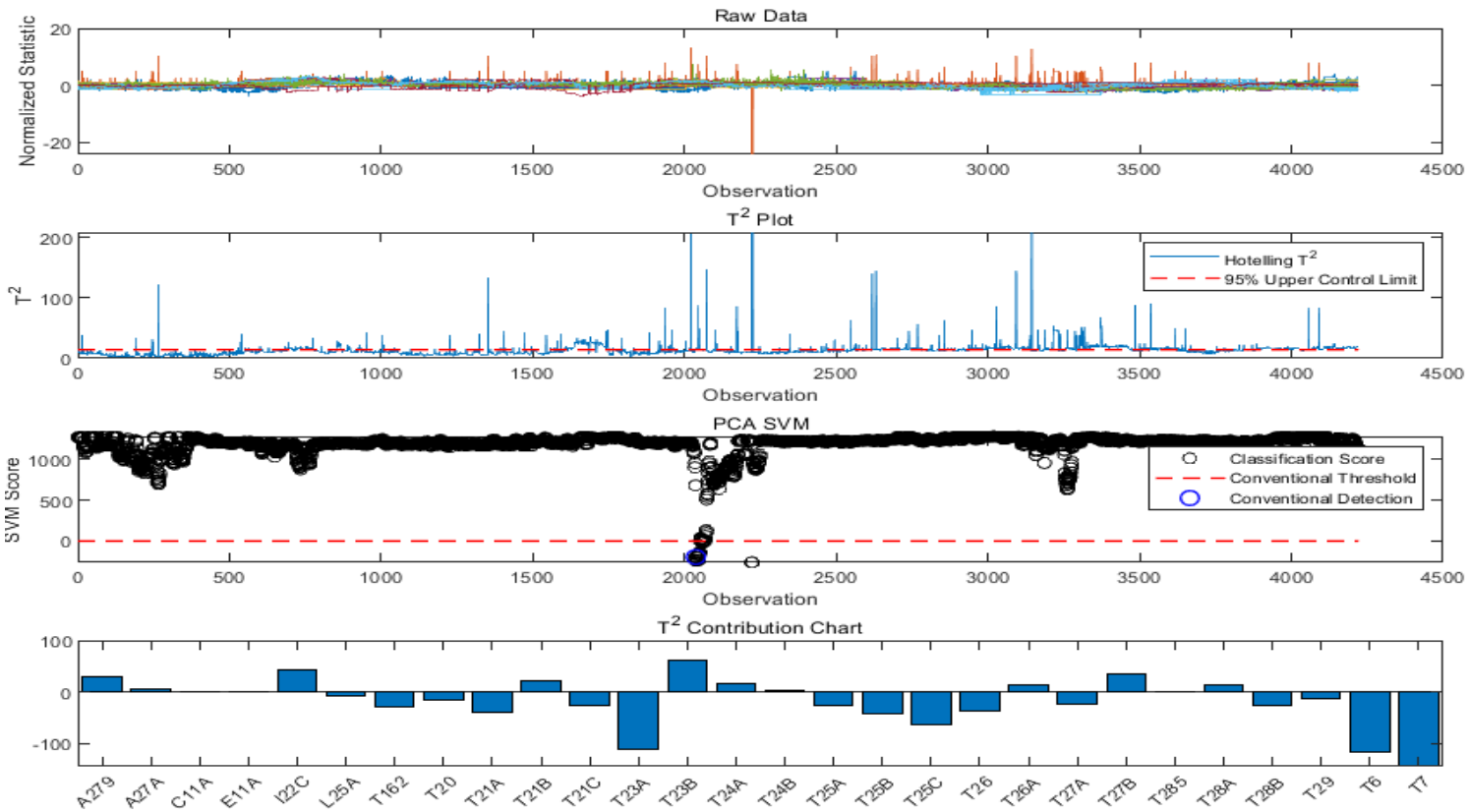

Figure 4.1 Fault detection and diagnosis result of Y1-June1 using PCA-SVM (Detection Time : 2237)

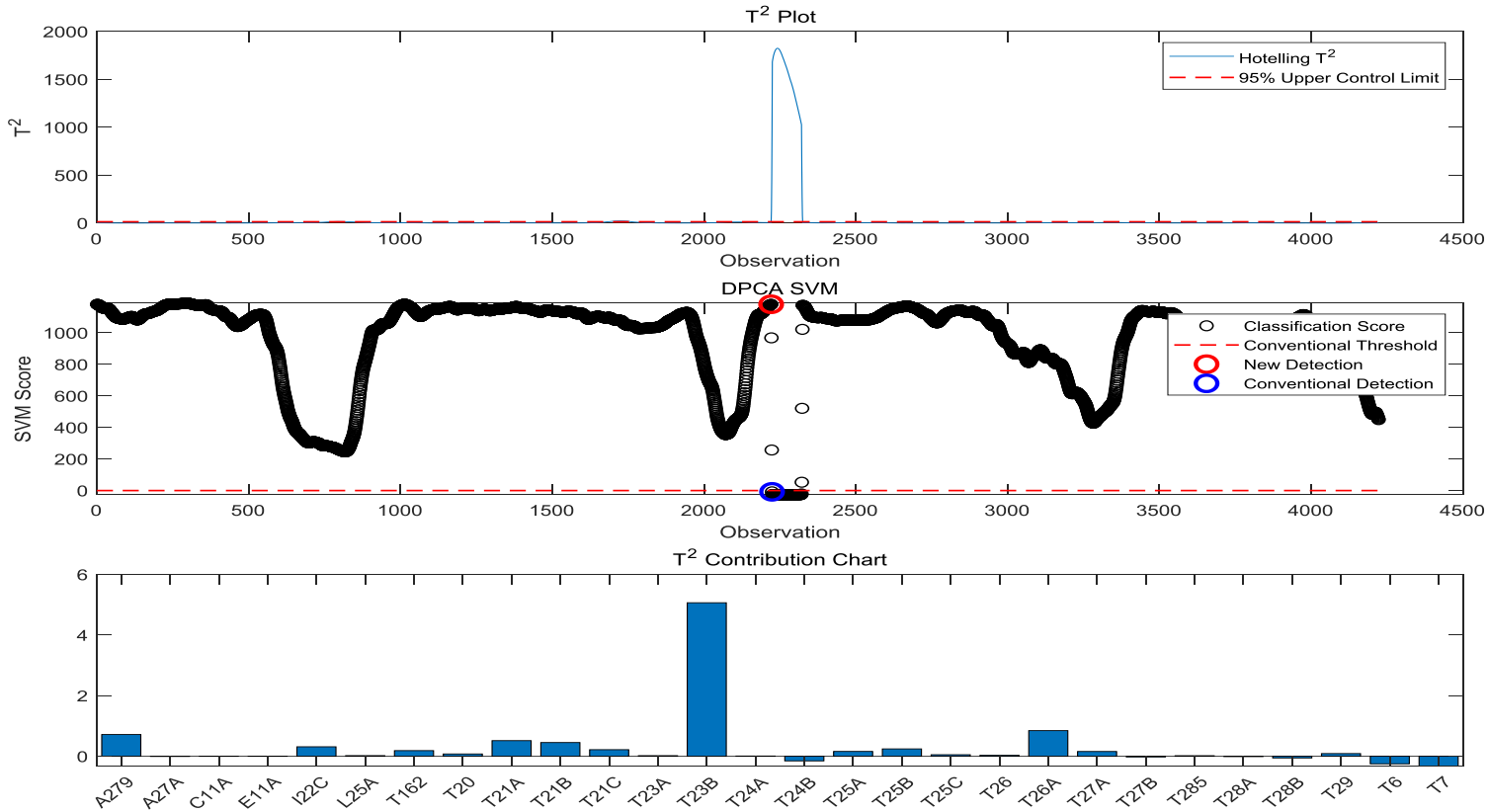

Figure 4.2 Fault detection and diagnosis result of Y1-June1 using DPCA-SVM (New detection time: 2221)

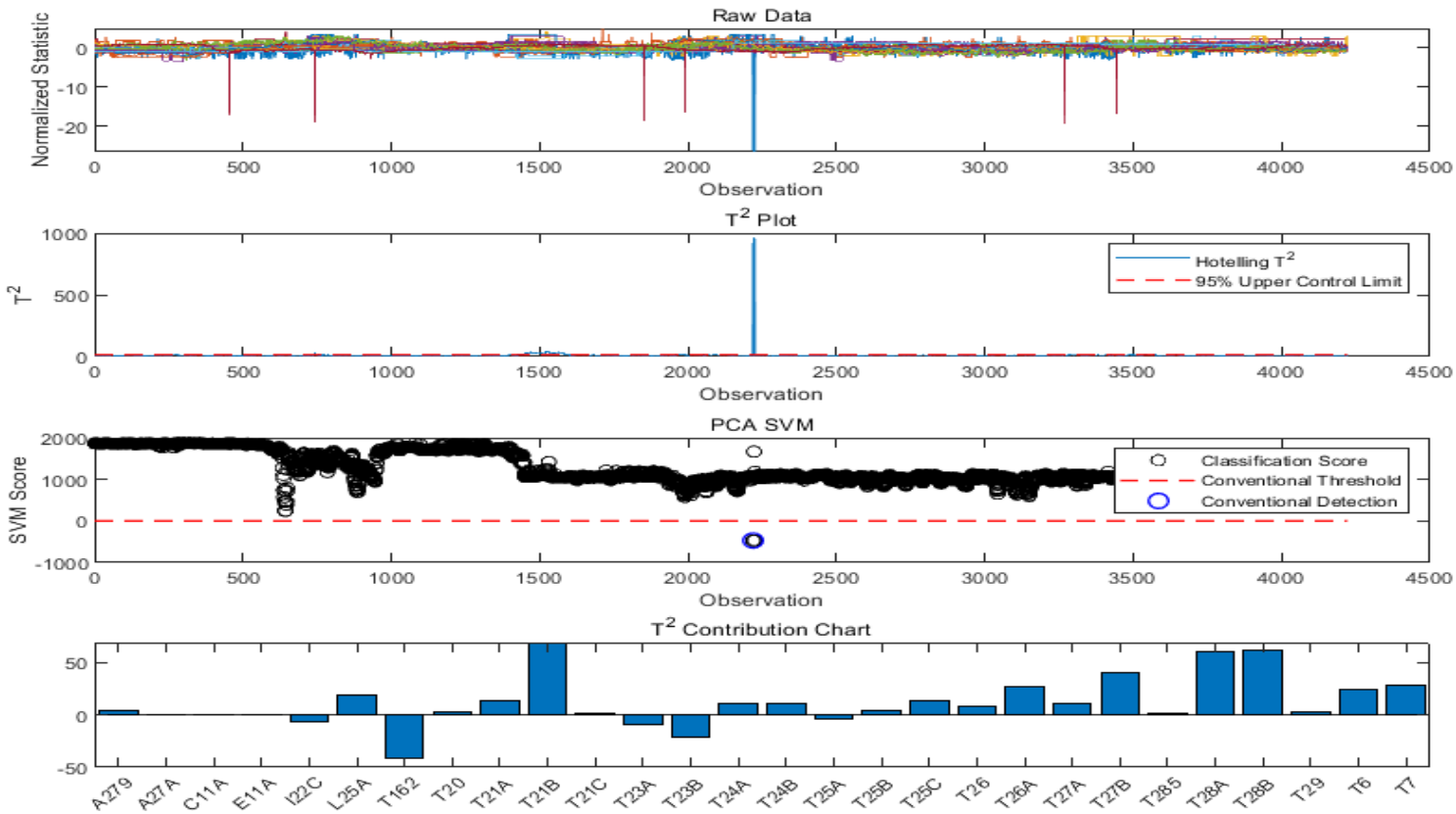

Figure 4.3 Fault detection and diagnosis result of Y1-Aug2 PCA-SVM(Detection time : 2319)

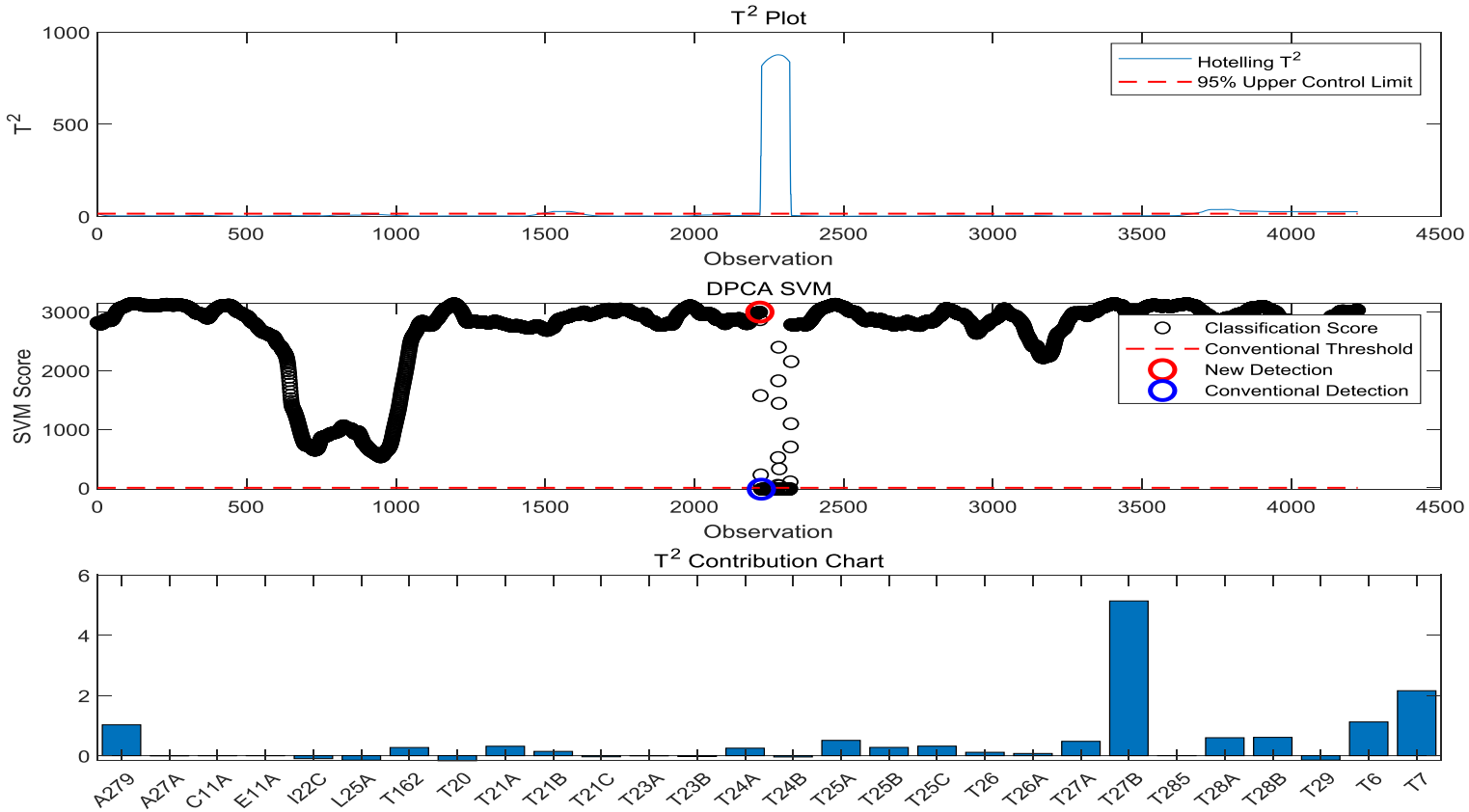

Figure 4.4 Fault detection and diagnosis result of Y1-Aug2 using DPCA-SVM(New detection time : 2219)

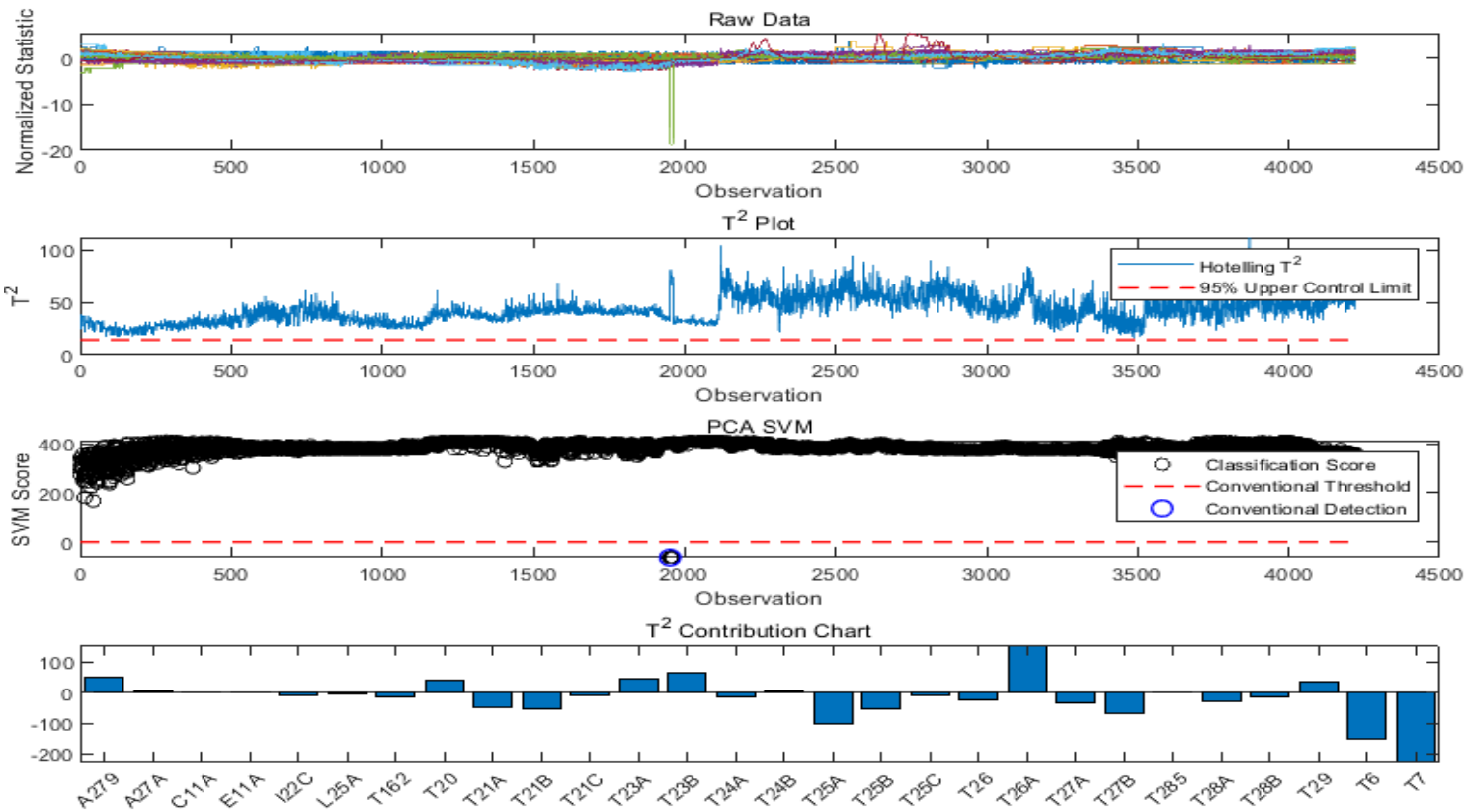

Figure 4.5 Fault detection and diagnosis result of Y1-Feb1 using PCA-SVM (Detection Time : 2152)

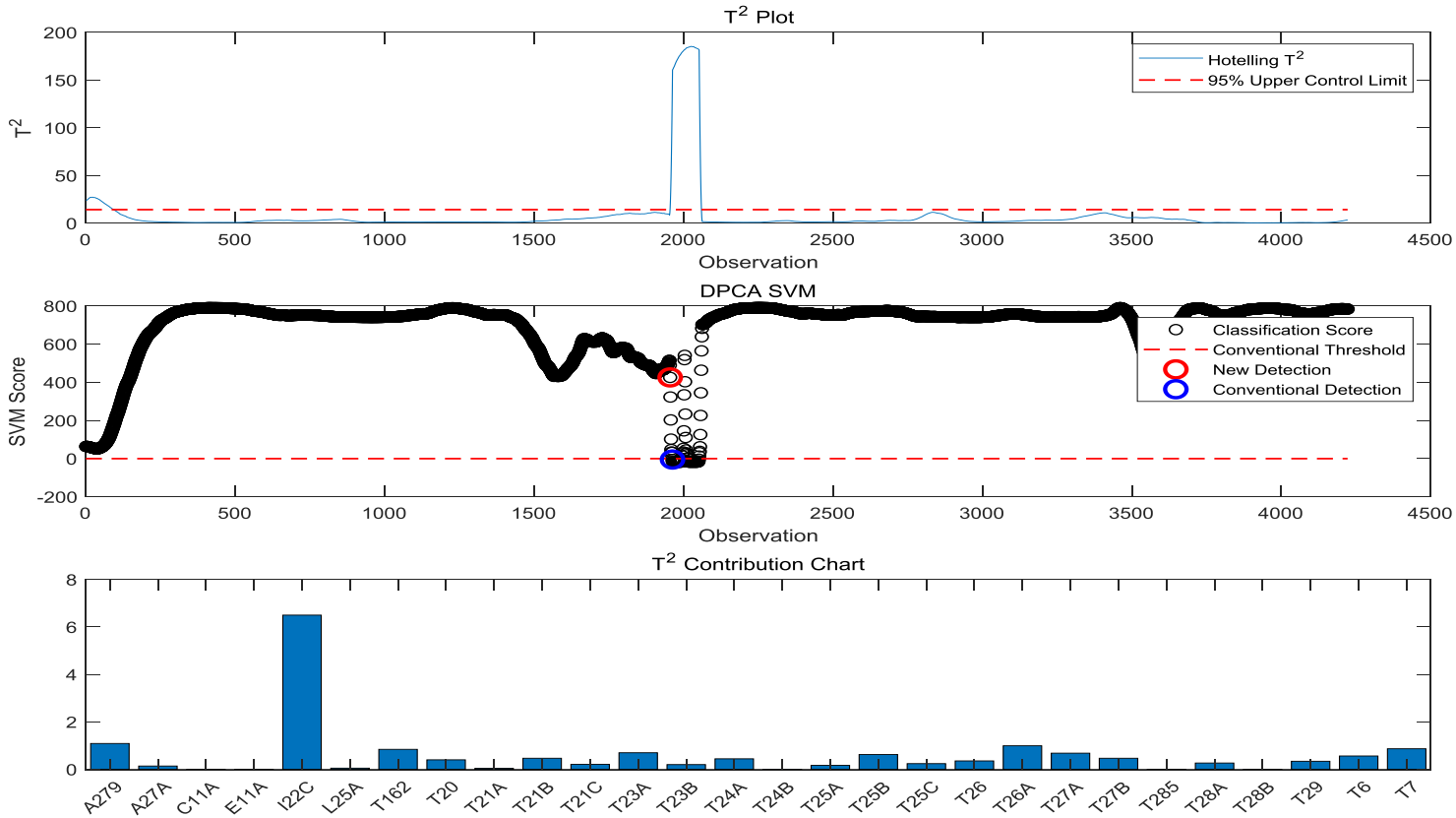

Figure 4.6 Fault detection and diagnosis result of Y1-Feb1 using DPCA-SVM (New detection time : 1955)

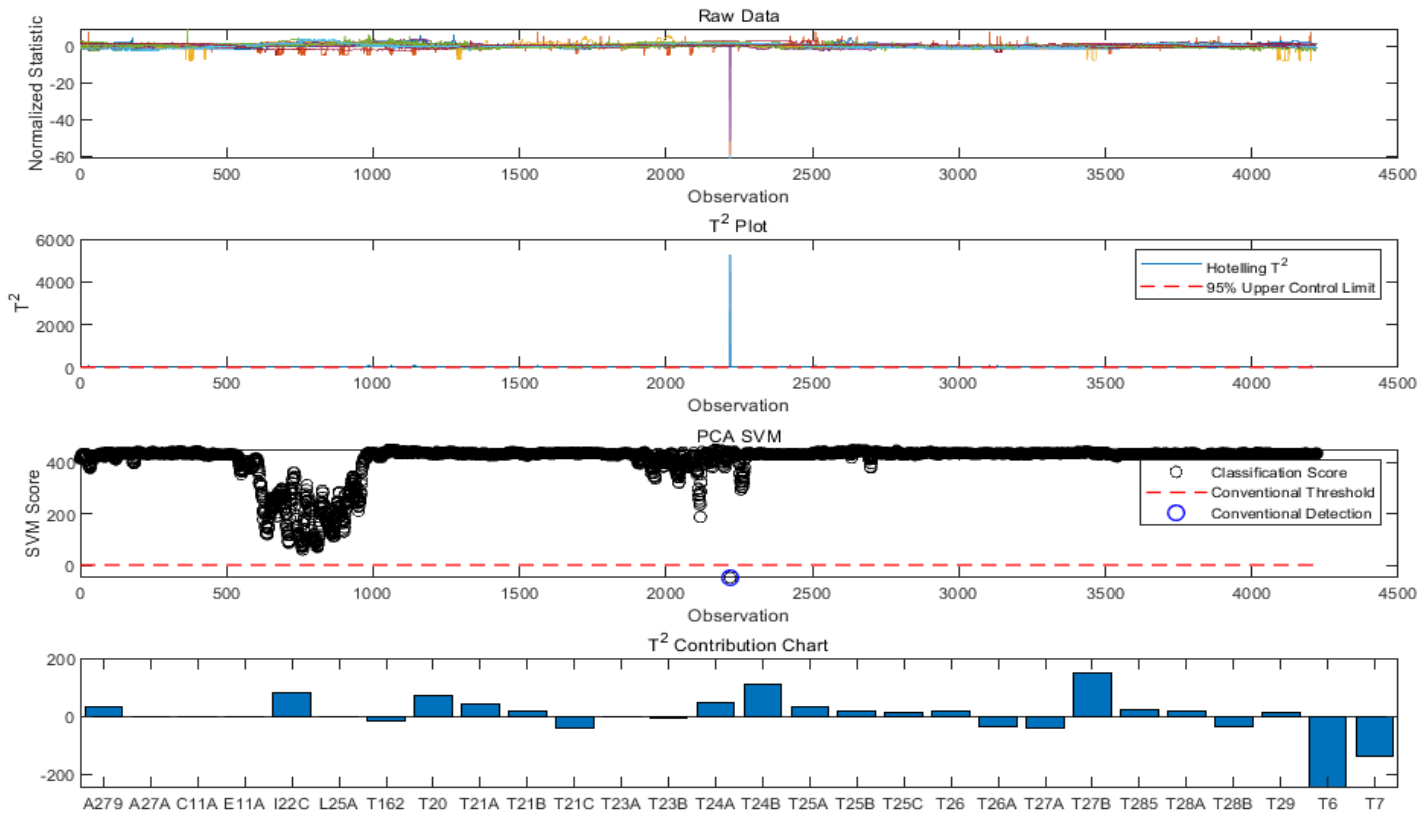

Figure 4.7 Fault detection and diagnosis result of Y1-June2 using PCA-SVM (Detection time : 2319)

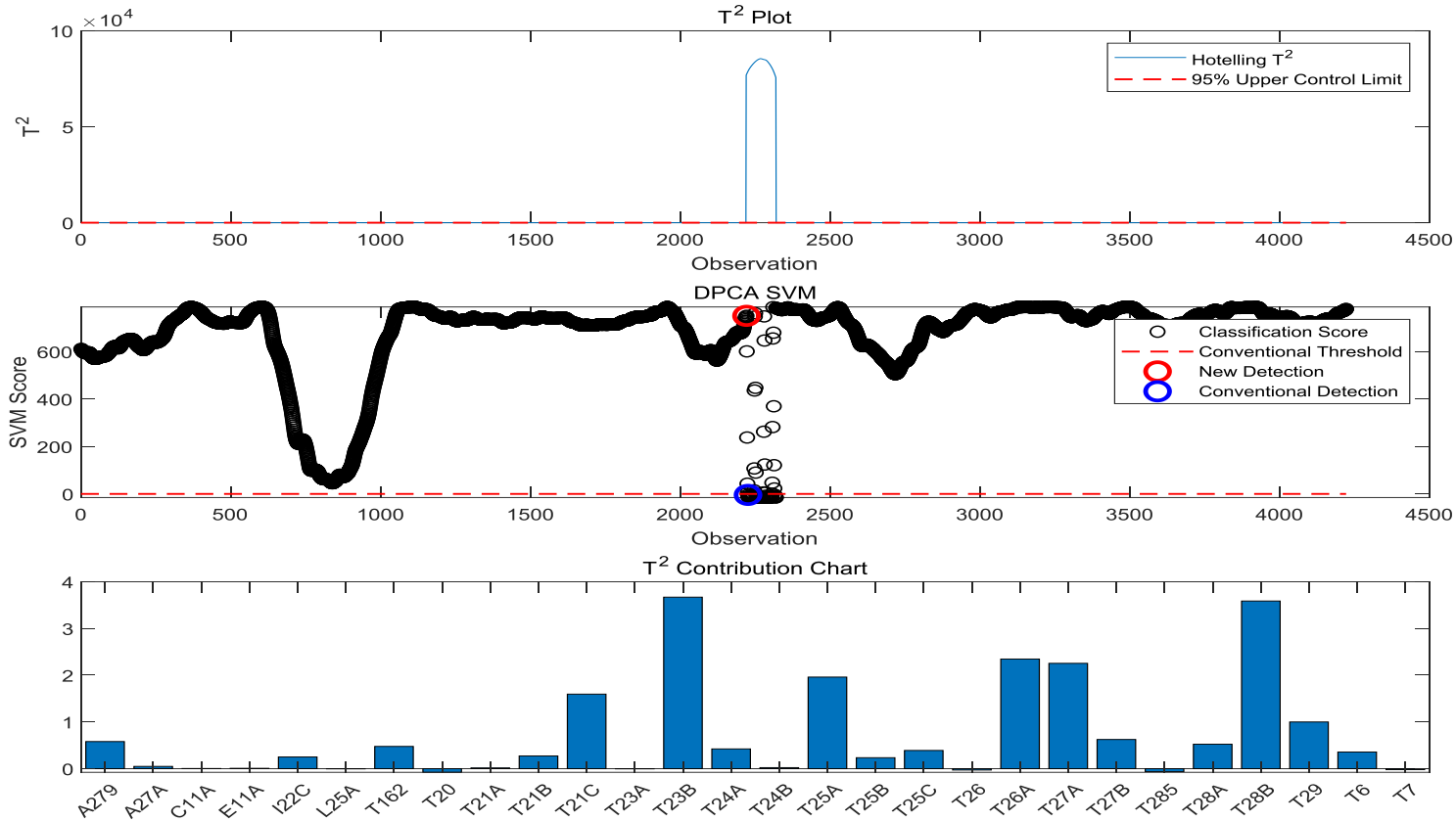

Figure 4.8 Fault detection and diagnosis result of Y1-June2 using DPCA-SVM (New detection time : 2221)

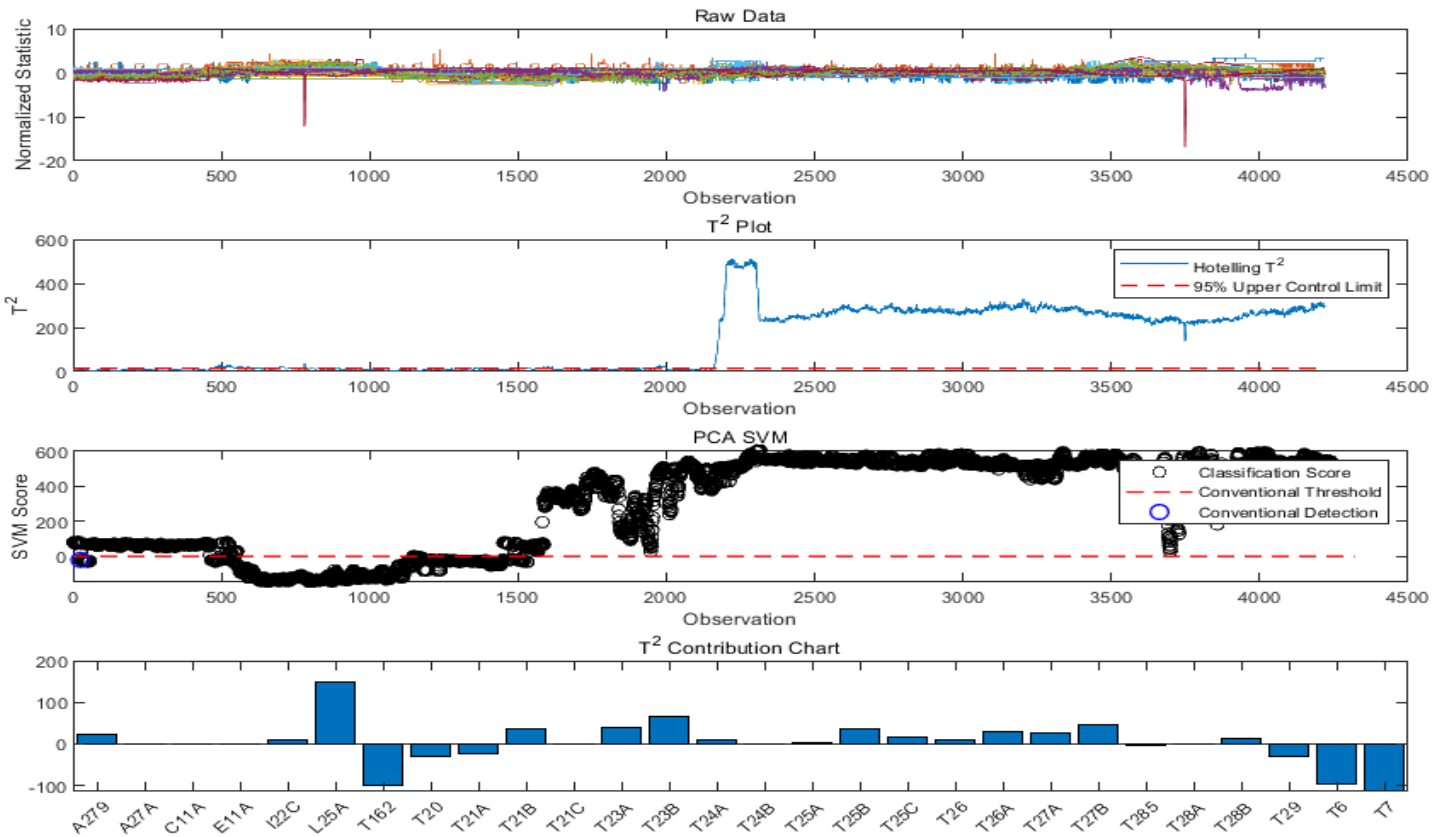

Figure 4.9 Fault detection and diagnosis result of Y1-Aug1 using PCA-SVM (Detection time : 25, 467)

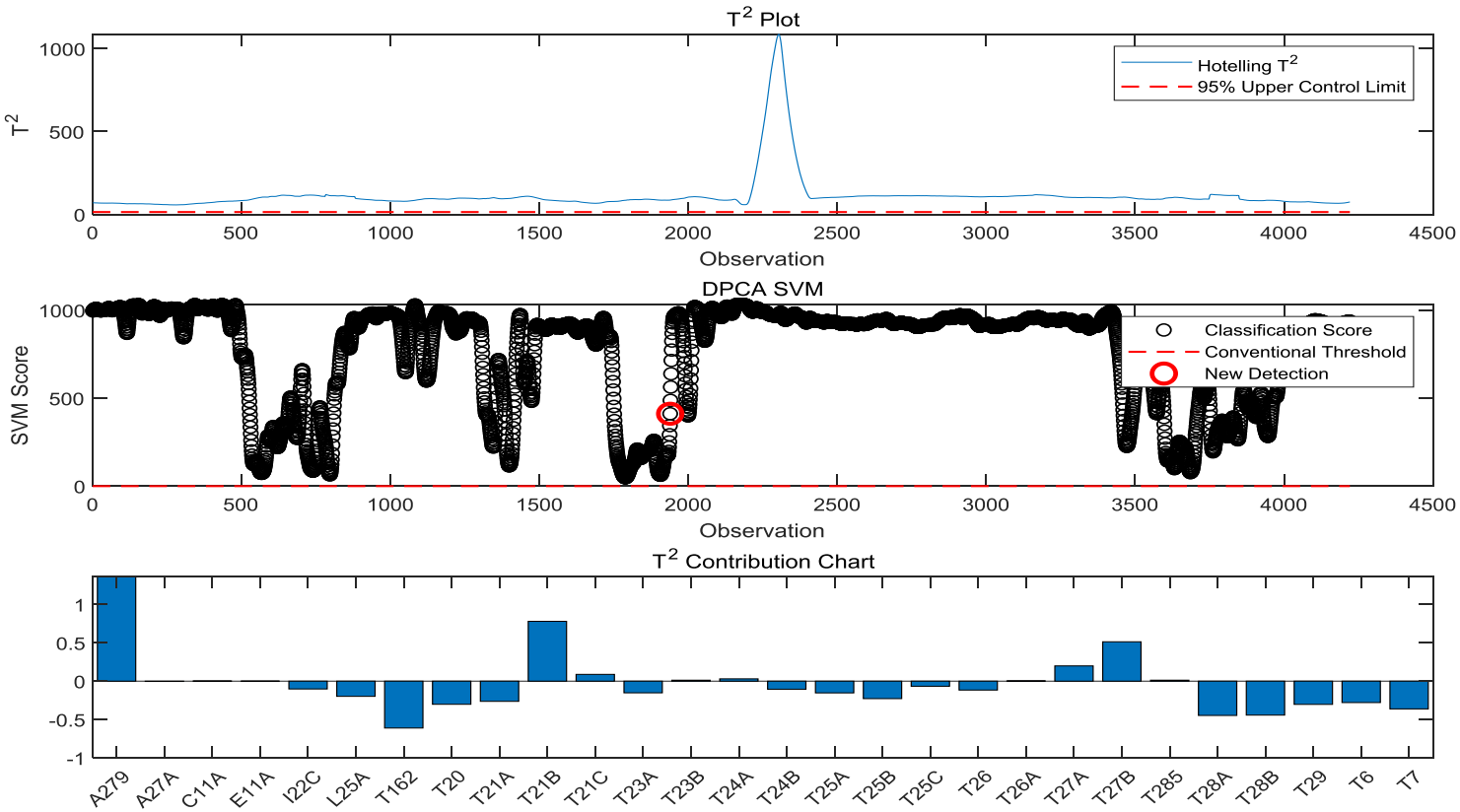

Figure 4.10 Fault detection and diagnosis result of Y1-Aug1 using DPCA-SVM (New detection : 1939)

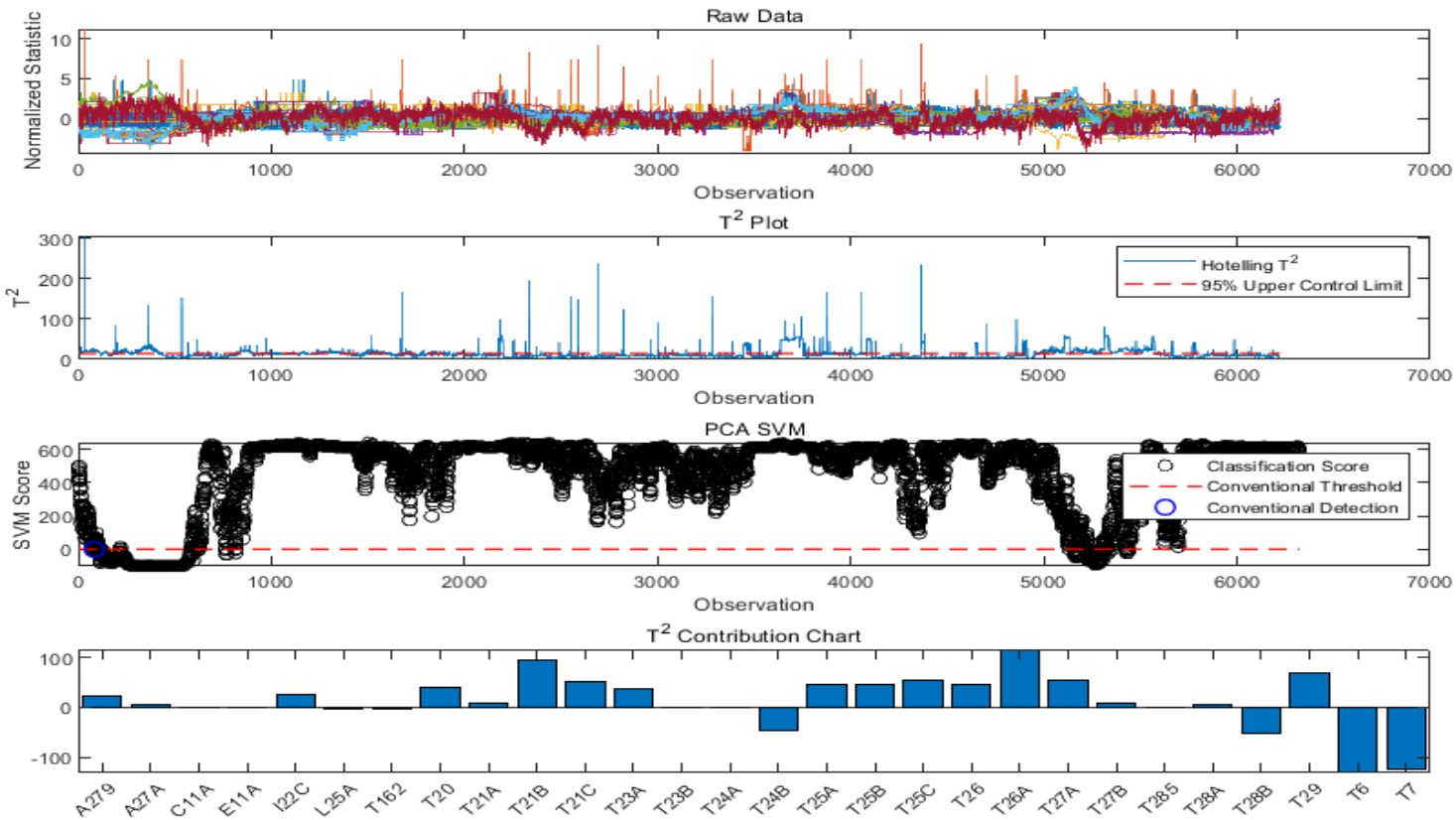

Figure 4.11 Fault detection and diagnosis result of Y2-March1 using PCA-SVM (Detection time : 83, 5200)

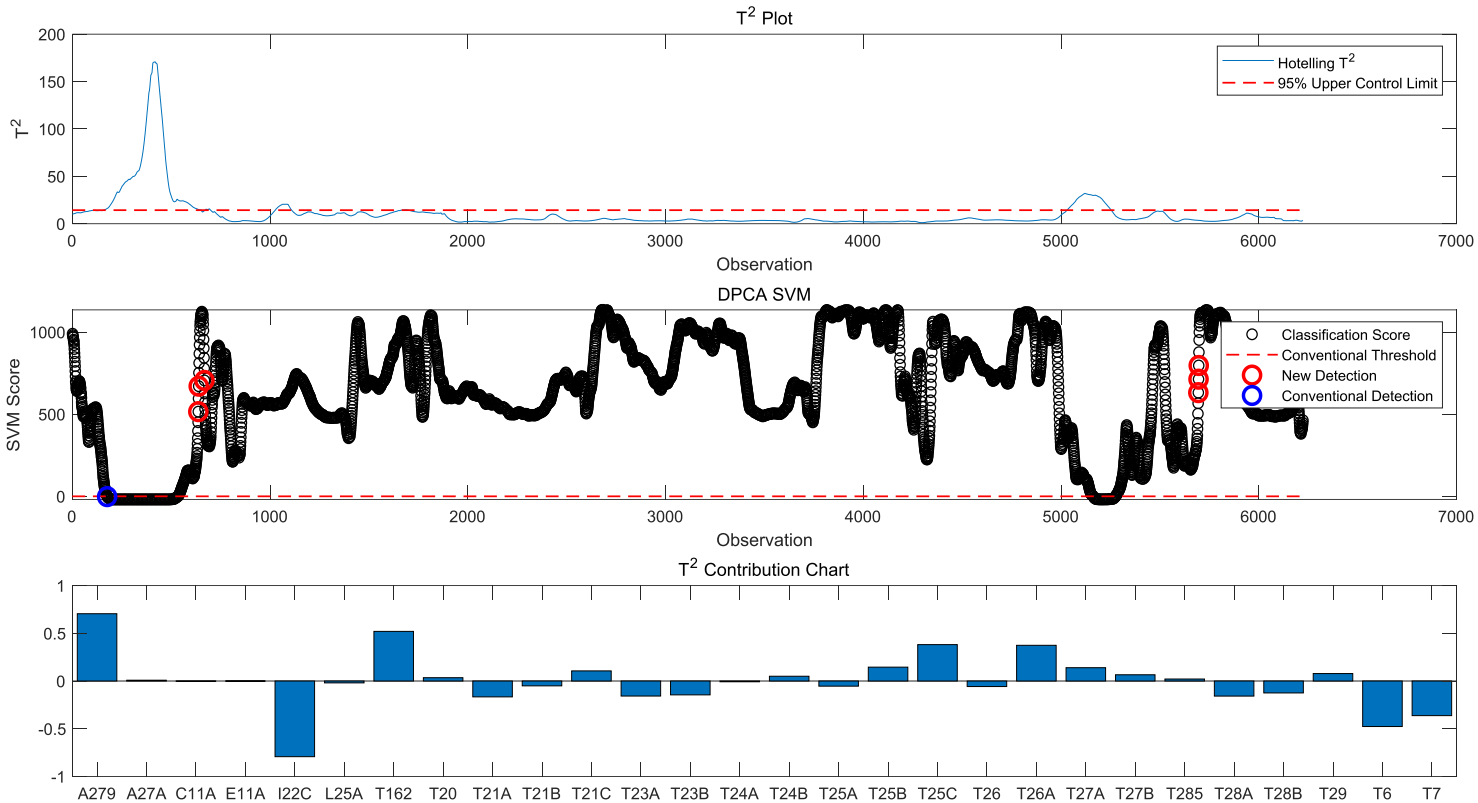

Figure 4.12 Fault detection and diagnosis result of Y2-March1 using DPCA-SVM (New detection time : 637, 5694)

CHAPTER 5. Conclusion

This work proposed detailed steps to preprocess operating process plant data and fault detect/diagnose using DPCA and 1-class SVM. The first step to a successful early fault detection is preprocessing of raw data. A general data preprocess is normalizing and eliminating outliers from raw data, but when handling data sets from an actual operating process plant, knowledge of the process and intuition is needed to eliminate the irrelevant variables. Selecting data sets to input FDD algorithm is done by analyzing and classifying provided maintenance record to single out corrective maintenance entries that shows characteristics of fault data. With dates of corrective maintenance entries as the set point, day before and after the fault occurrence are included as part of the test data set because the date of entry is inputted only after the completion of corrective action. 5000 data prior to day before the entry is selected as training data which is used to create the 1-class SVM normal boundary. Anything that is outside the SVM boundary is classified as fault and SVM classification score is used as the measurement to determine fault or normal.

Popular detection methods, PCA-T^2, PCA-SVM and DPCA-SVM were compared as means for FDD. The results of fault detection algorithm for 6 corrective maintenance entries showed that DPCA-SVM method was the most accurate method and newly proposed threshold, difference of consecutive score than exceed 130, showed much more early detection and diagnosis result than any other algorithm or methods in 5 scenarios out of 6. The corrective

maintenance #6, was considered as a failed detection because threshold was violated after the window of fault occurrence window but the margin of error was not a large as false alarm occurred only once. This may be due to a wrongful preprocessing or data selection of the data set.

This work has significance for providing robust FDD method with a new threshold does not require additional data fitting and also notably proposing detailed steps to preprocess actual operating process plant data. With the proposed FDD algorithm with a new threshold, it introduces more practical fault detection and diagnosis than other known algorithms with simulated data enabling to perform predictive maintenance for optimal plant management.

초록

이상 감지 및 진단 분야는 화학 공정 사업에서 매우 중요한 이슈로 부상되면서 관련된 여러 알고리즘이 개발되었다. 최근 컴퓨터 계산 능력 향상 및 새로운 통계 기법들의 개발로 인해 데이터 기반 접근법이 이상 감지 및 진단 분야에 많이 사용되고 있다. 이상 감지 및 진단을 위해 실제 운영 공정 데이터를 사용하는 것이 가장 이상적이지만, 실제 운영 데이터의 확보가 어렵고 자세한 데이터 선처리 방법이 알려진 바가 없어 대부분의 이상 감지 알고리즘은 통제 가능한 시뮬레이션 데이터로 검증이 수행되어 왔다. 이로 인해 이상 감지 및 진단에 매우 중요한 부분인 실제 운영 데이터에 대한 선처리에 대해 이번 연구에서 정립했다.

실제 운영 데이터에 대한 선처리는 크게 2 부분(정비 기록부, 센서 데이터)으로 구분된다. 정비기록부의 내용은 정비 행위 특성으로 예방정비, 기간별 정비, 시정 정비, 예측 정비 등 총 4 가지로 분류될 수 있다. 결함 특성이 드러나는 데이터인 시정 정비로 분류된 6 개의 기록이 이번 연구에서 사용되었다. 센서 데이터의 236 개의 변수는 개략도와 데이터의 성향 분석을 통해 28 개로 감소할 수 있다.

실제 운영 데이터의 이상 감지 및 진단을 위해 Dynamic Principal Component Analysis(DPCA)와 1-class Support Vector Machine(SVM) 기법을 사용했다. DPCA 는 데이터의 차원 축소를 위해 사용했고 1-class SVM 은 SVM 구축을 위해 1 종류의 데이터만 필요하기 때문에 운영 공정 데이터 분류에 적합하여 사용했다. 기존의 SVM score 분류 임계 값은 0 이고 score 값이 음수일 경우 결함으로 분류했다. 이 연구에서는 연속적인 SVM score 값의 차이가 130 일 경우를 결함으로 분류하는 새로운 임계 값을 제안했다. 선처리한 데이터를 활용하여 6 개의 시정 정비 기록부 내용에 대해 이상 감지 및 진단을 한 결과, 5 개의 시나리오에서 좋은 감지 및 진단 결과를 보였다.

이 연구에서 제안한 실제 운영 플랜트 데이터에 대한 선처리 세부 과정과 새로운 SVM score 임계 값으로 인해 이상 감지 및 진단을 정확하고 조기에 수행 가능하게 되었다. 따라서 제안한 방법들은 조기 결함 탐지/진단으로 예측 정비를 수행을 가능케 해서 최적의 플랜트 운영에 이바지할 수 있을 것이다.

주요어: 이상 감지, 이상 탐지, 기계학습, 다변량분석, 데이터 기반 접근법, 선처리, 1-class SVM

학 번: 2017-29082

References

[1] K. Mathioudakis, A. J. o. e. f. g. t. Stamatis, and power, "Compressor fault identification from overall performance data based on adaptive stage stacking," vol. 116, no. 1, pp. 156-164, 1994.

[2] R. Isermann, *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2006.

[3] P. M. J. E. J. o. c. Frank, "Analytical and qualitative model-based fault diagnosis–a survey and some new results," vol. 2, no. 1, pp. 6-28, 1996.

[4] R. Sadeghbeigi, *Fluid catalytic cracking handbook: An expert guide to the practical operation, design, and optimization of FCC units*. Elsevier, 2012.

[5] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*: Springer, 2011.

[6] J. F. MacGregor and T. J. C. E. P. Kourti, "Statistical process control of multivariate processes," vol. 3, no. 3, pp. 403-414, 1995.

[7] J. Mina and C. Verde, "Fault detection using dynamic principal component analysis by average estimation," in *Electrical and Electronics Engineering, 2005 2nd International Conference on*, 2005, pp. 374-377: IEEE.

[8] C. Jing and J. J. N. Hou, "SVM and PCA based fault classification approaches for complicated industrial process," vol. 167, pp. 636-642, 2015.

[9] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[10] B. Scholkopf *et al.*, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," vol. 45, no. 11, pp. 2758-2765, 1997.

[11] B. Schölkopf, A. J. Smola, and F. Bach, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[12] S. Mahadevan and S. L. J. J. o. p. c. Shah, "Fault detection and diagnosis in process data using one-class support vector machines," vol. 19, no. 10, pp. 1627-1639, 2009.

[13] N. M. J. J. o. e. i. Nasrabadi, "Pattern recognition and machine learning," vol. 16, no. 4, p. 049901, 2007.

[14] J. Gertler, *Fault detection and diagnosis*. Springer, 2013.

[15] J. J. Downs, E. F. C. Vogel, and c. engineering, "A plant-wide industrial process control problem," vol. 17, no. 3, pp. 245-255, 1993.