Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 432651, 10 pages
http://dx.doi.org/10.1155/2015/432651

Research Article

Fault Diagnosis in Condition of Sample Type Incompleteness Using Support Vector Data Description

Hui Yi,[1] Zehui Mao,[2,3] Bin Jiang,[2,3] Cuimei Bo,[1] Yufang Liu,[2,3] and Hui Luo[4]

[1]*College of Automation and Electrical Engineering, Nanjing Tech University, Nanjing 211816, China*
[2]*College of Automation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*
[3]*Jiangsu Key Laboratory of Internet of Things and Control Technologies, Nanjing 210016, China*
[4]*College of Engineering, Nanjing Agricultural University, Nanjing 210031, China*

Correspondence should be addressed to Bin Jiang; binjiang@nuaa.edu.cn

Received 8 October 2014; Revised 17 January 2015; Accepted 19 January 2015

Academic Editor: Gang Li

Faulty samples are much harder to acquire than normal samples, especially in complicated systems. This leads to incompleteness for training sample types and furthermore a decrease of diagnostic accuracy. In this paper, the relationship between sample-type incompleteness and the classifier-based diagnostic accuracy is discussed first. Then, a support vector data description-based approach, which has taken the effects of sample-type incompleteness into consideration, is proposed to refine the construction of fault regions and increase the diagnostic accuracy for the condition of incomplete sample types. The effectiveness of the proposed method was validated on both a Gaussian distributed dataset and a practical dataset. Satisfactory results have been obtained.

1. Introduction

Fault diagnosis has been the subject of great interest in the control research community in response to the increasing requirement of operating reliability and product quality [1–4]. During the last decades, fault diagnosis has been well studied and many useful approaches have been proposed, such as the parameter estimation [5] and state estimation [6]. Generally, these approaches could diagnose the faults by analyzing the residual between the real output and the model output, which are so-called model-based fault diagnoses. The core of the model-based diagnostic approach is to build a process model running parallel to the process [1].

Due to the ever-growing complexity of industrial systems, the modeling of a complex industrial process becomes very difficult and time-consuming. In some newly developed processes, such models are even unavailable. Consequently, data-driven approaches have been introduced to the field of fault diagnosis. This kind of approach does not need to build precise models for industrial processes. Instead, it utilizes the processing data, including offline and online data, to approximate the relationship between system inputs and the operating statuses. The data-driven approach is especially suitable for the fault detection (FD) and isolation for complex systems and hence has aroused much interest recently [4, 7–9].

One classical way for data-driven diagnostic approaches is to divide the data distributing space into several fault regions by employing particular classifiers [10]. Then, in each fault region, all data samples belong to the same operating status [11]. This classifier-based approach makes the diagnosis by locating the fault region that the testing sample falls into and avoids the requirement of system models and expert knowledge. During the last 10 years, numerous studies have been done concerning the issues of its diagnostic accuracy and decision speed. The reliability and feasibility of the classifier-based fault diagnosis approach have been significantly improved [12–14].

As has been pointed out by Russell et al. [7], the main drawback of data-driven approaches is that the diagnostic proficiency is highly dependent on the quantity and quality of the process data. When using the classifier-based

approaches for fault diagnosis, fine data samples are required to ensure the performance of the diagnosis. More specifically, according to the research of Hakkila et al. [15], the classification performance is especially sensitive to the completeness of the training samples. A complete set of training samples is essential for making the correct region divisions. This implies that when using classifier-based approaches, training samples for all types of faults should first be prepared.

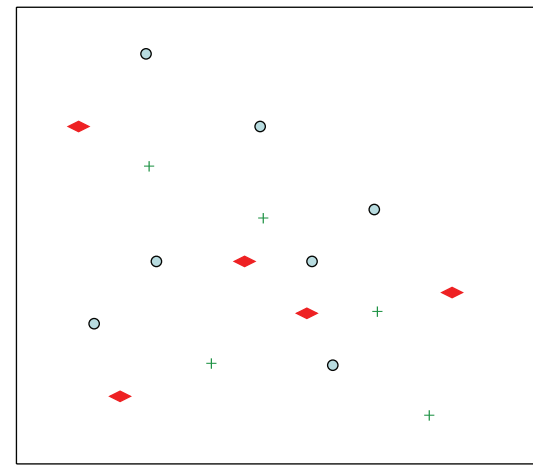
However, data-driven approaches show a significant difference in the difficulty of acquiring different faulty samples in real applications. In most cases, samples for common faults and for normal operating statuses could easily be obtained, whereas samples for rare faults and multifaults are seldom acquired. It is very hard to collect a complete training set which contains samples for all possible faults [14]. The problem of sample-type incompleteness has greatly reduced the feasibility of classifier-based approaches.

Consequently, the need to improve the diagnostic performance in the condition of incomplete faulty samples becomes an issue which greatly hampers the practical applications of data-driven fault diagnosis. However, little has been addressed in the literature on this issue. In this paper, a support vector data description (SVDD-) based approach has been proposed in an attempt to improve the diagnostic performance of the classifier-based approach in the condition of incomplete samples. It reduces the sensitivity of diagnostic accuracy towards sample completeness.

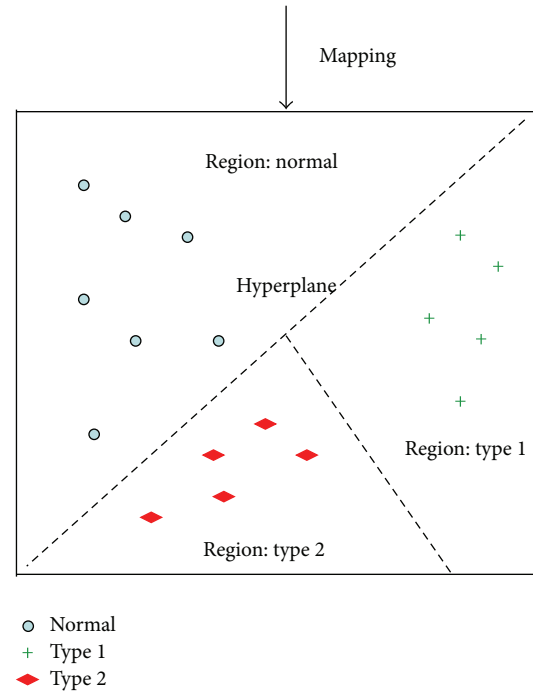
This paper is organized as follows. Section 2 illustrates how the sample-type incompleteness decreases the diagnosing accuracy. Section 3 introduces the classification mechanism of the SVDD. Section 4 presents the refined diagnostic algorithm using SVDD. In Section 5 the effectiveness of the proposed method is validated by experimental comparisons with conventional methods. Finally, conclusions and a discussion are given in Section 6.

2. Effects of the Sample-Type Incompleteness on the Classifier-Based Fault Diagnosis

Fault diagnosis includes FD and fault isolation (FI). FD attempts to judge whether the system is faulty or normal and then FI is employed to identify which fault has occurred in the system. In contrast to the conventional fault diagnosis where FD and FI are two different steps, the classifier-based diagnostic approaches complete both FD and FI in one step. Therefore, it has been widely used due to the high efficiency for fault diagnosis. The classifier-based approach uses particular classifiers to map the linear inseparable faulty samples into a high-dimensional space where different types of samples are linearly separable. Furthermore, a hyperplane is generated to divide the high-dimensional space into several regions, called fault regions, and each region consists of only one faulty-type sample, as shown in Figure 1. When testing samples are imported, they are first mapped into the same high-dimensional space. Then, by locating which fault region they have fallen into, the operating status can be identified, and the aim of FD and FI can be achieved [10, 11].



(a) Symptom dataset



(b) Fault regions

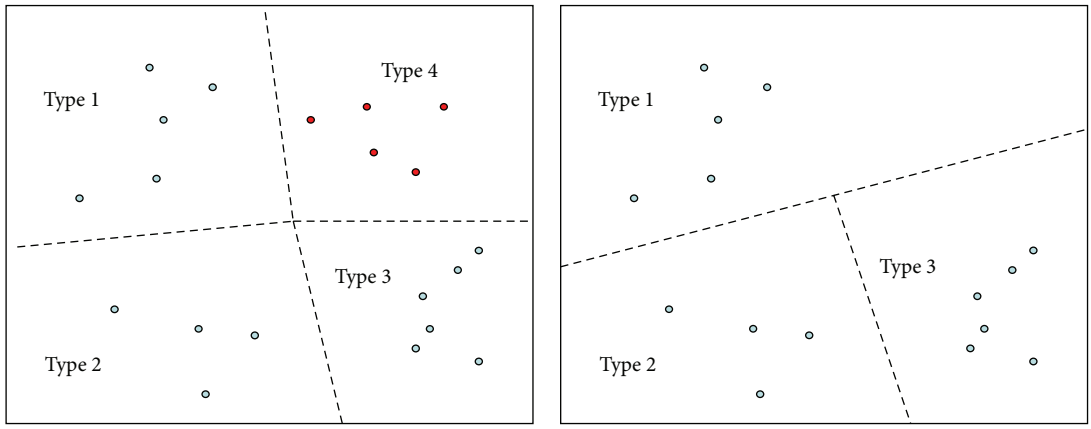
Figure 1: Fault regions made by classifiers.

Given a training set, D, which consists of process data for a normal status, N, and k types of faulty samples

$$D = \bigcup_{i=1}^{k} S_i \cup N, \tag{1}$$

where S_i is the sample set for the ith fault, classifiers like artificial neural networks (ANNs) and multiclass SVMs divide the sample distributing space into one "Normal" region and "k" faulty regions:

$$\text{Hyperplane}_{(w_0,b_0)} = \phi(D) = \phi\left(\bigcup_{i=1}^{k} S_i \cup N\right), \tag{2}$$



(a) Fault regions in condition of complete training set (b) Fault regions in condition of sample-type incompleteness

Figure 2: Fault regions of different sample-type completeness.

where ϕ is the mapping function for the given classifier and $H(x) = w_0 \cdot x + b_0$ is the hyperplane. The fault regions should be {"Normal", "Fault 1", ..., "Fault k"}.

When the testing samples, X_{Test}, are acquired, the fault can be isolated by locating which fault region they have fallen into by

$$F = \Psi\left\{\phi\left(D\right), X_{\text{Test}}\right\}, \tag{3}$$

where Ψ is the decision function and F is its decision value. Because the testing samples, X_{Test}, have already been obtained while diagnosing, X_{Test} can be considered constants in the above formula. Thus, the diagnostic performance mainly depends on three factors: ϕ, Ψ, and D. Here, the function, ϕ, is decided by the classification performance including algorithm selection and parameter setting, and the function, Ψ, is decided by the fault diagnostic flowchart. Thus the training set, D, can be regarded as the only variable in formula (3) that decides the diagnostic accuracy. In the condition of fault type incompleteness, for example, no process data has been acquired for some operating statuses; the variable, D, would be

$$D' = \bigcup_{i=1}^{k-m} S_i \cup N; \tag{4}$$

here m types of faulty samples are assumed to be missing. As the variable changes, a different hyperplane, $H'(x) = (W_0', x) + b_0' = \phi(D')$, would be generated, and furthermore the diagnosing result, $F' = \Psi\{\phi(D'), X_{\text{Test}}\}$, is likely to be changed. This implies that the incompleteness of fault types leads to a misclassification and a decrease of diagnostic accuracy.

Figure 2 is employed to illustrate how the sample-type incompleteness affects the diagnosis. In this figure, the complete training set is composed of 4 types of samples. Figure 2(a) shows the fault regions which are made in the condition of a complete training set. Then, taking the 4th faulty sample away, the new generated fault regions can be seen in Figure 2(b), where the region divisions are significantly different.

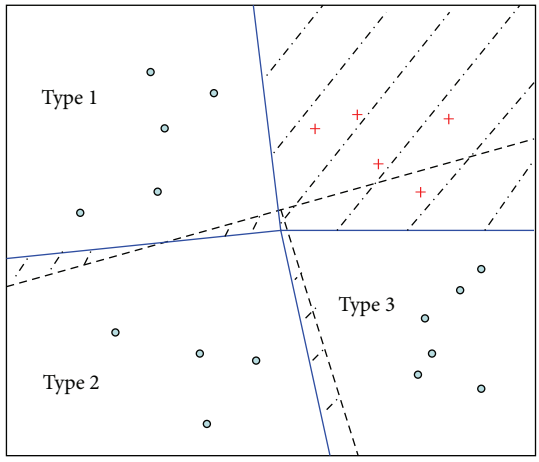


Figure 3: Misclassified fault regions brought on by sample-type incompleteness.

The shadow areas in Figure 3 denote the misclassified fault regions that were brought on by the incompleteness of the sample type. The testing samples which fall into these shadow areas will be diagnosed incorrectly.

3. Support Vector Data Description

SVDD is a kind of one-class classifiers which judges whether a testing sample belongs to the target sample type or not and requires only one type of samples while training [16]. It attempts to find a hypersphere with a minimum volume and contains all target samples, as shown in Figure 4. When the test sample has fallen into the hypersphere, it belongs to the same type as the target samples.

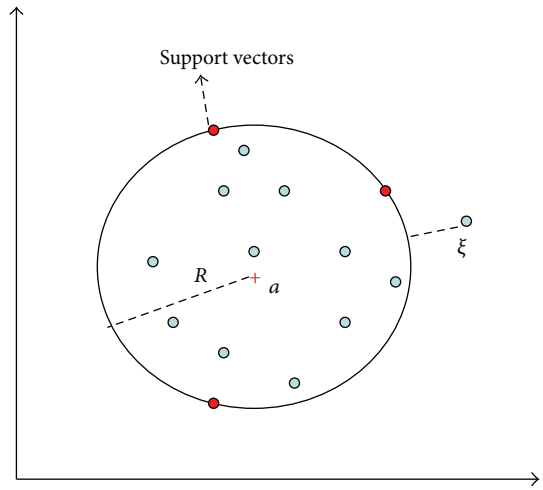


Figure 4: Region for target data made by SVDD.

Given the target data, $X = \{x_i\}_{i=1}^{n}$, where n is the number of training samples, the SVDD searches for the hypersphere, $\Omega = (a, R)$, by minimizing the radius:

$$\begin{aligned} \text{Min:} \quad & F(R,a) = R^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & (x_i - a)^T (x_i - a) \leqslant R^2 + \xi_i, \\ & \xi_i \geqslant 0, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{5}$$

where a denotes the centre of the sphere and R is its radius. C is the regular factor that gives the trade-off between the radius and the number of errors and ξ is the slack factor. The following Lagrangian was constructed to solve this problem:

$$\begin{aligned} & L(R, a, \alpha_i, \gamma_i, \xi_i) \\ & = R^2 + C\sum_i \xi_i \\ & \quad - \sum_i \alpha_i \left\{ R^2 + \xi_i - \left(\|x_i\|^2 - 2a \cdot x_i + \|a\|^2 \right) \right\} \\ & \quad - \sum_i \gamma_i \xi_i, \end{aligned} \tag{6}$$

where $\gamma_i \geqslant 0$ and $\alpha_i \geqslant 0$ are the Lagrange multipliers. After taking the partial derivative, the dual problem for formula (6) can be written as

$$\begin{aligned} \text{Max:} \quad & L = \sum_i \alpha_i (x_i \cdot x_j) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \\ \text{s.t.} \quad & 0 \leqslant \alpha_i \leqslant C \end{aligned} \tag{7}$$

Formula (7) is a standard quadratic programming problem. There are many well-studied methods to solve such problems, for example, the active set method [17]. As α can be obtained from formula (7), a and R in forumla (5) can also be deduced.

Given a testing sample, z, if

$$\begin{aligned} \|z - a\|^2 = {} & (z \cdot z) - 2\sum_i \alpha_i (z \cdot x_i) \\ & + \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \leqslant R^2, \end{aligned} \tag{8}$$

then z belongs to the same type as the target data, X.

Remark 1. Conventional classifiers such as Support Vector Machines (SVMs) require at least two types of samples while making the hyperplane. The output of these classifiers implies the type to which the testing sample is the most likely to belong. However, SVDD, which is specifically designed for one-class classification, requires only one type of training sample, that is, the target data. The aim of the SVDD classification is to judge whether the testing sample belongs to the type of target data.

4. Fault Diagnosis Using SVDD for the Condition of Sample-Type Incompleteness

As shown in formula (2), conventional classifiers take the whole training set, D, as the input variable for constructing the hyperplane. As D varies, the hyperplane varies, and furthermore the diagnostic result should be changed as in formula (3). This paper addresses a new framework for classifier-based fault diagnosis which is capable of implementing the type incomplete training set to construct a reasonable hyperplane and can gradually refine its diagnostic performance as more types of samples are added to the training set.

4.1. The Basic Idea of the Proposed Diagnosing Approach. Unlike traditional methods, the proposed approach does not construct all regions in a go. It constructs the regions step by step.

As shown in Figure 5, the SVDD-based approach firstly judges whether the testing sample belongs to "Known" types or "Unknown" types. Then, if the testing sample belongs to the "Known" types, the approach locates which region should the testing sample fall into.

Remark 2. "Known" types refer to types whose operating data samples have been acquired. "Unknown" types refer to types whose operating data samples are missing.

Figure 6 is implemented to illustrate how the SVDD-based approach constructs the fault regions.

The gray area in Figure 6 is the "Known" region and also refers to space where diagnosis could be made. From (a) to (d), we see the following.

(1) When new types of samples are imported, the SVDD-based approach just adds corresponding regions but does not reconstruct all regions. This means that if one fault region is constructed, the region will never change no matter how many more sample types are imported in the training.

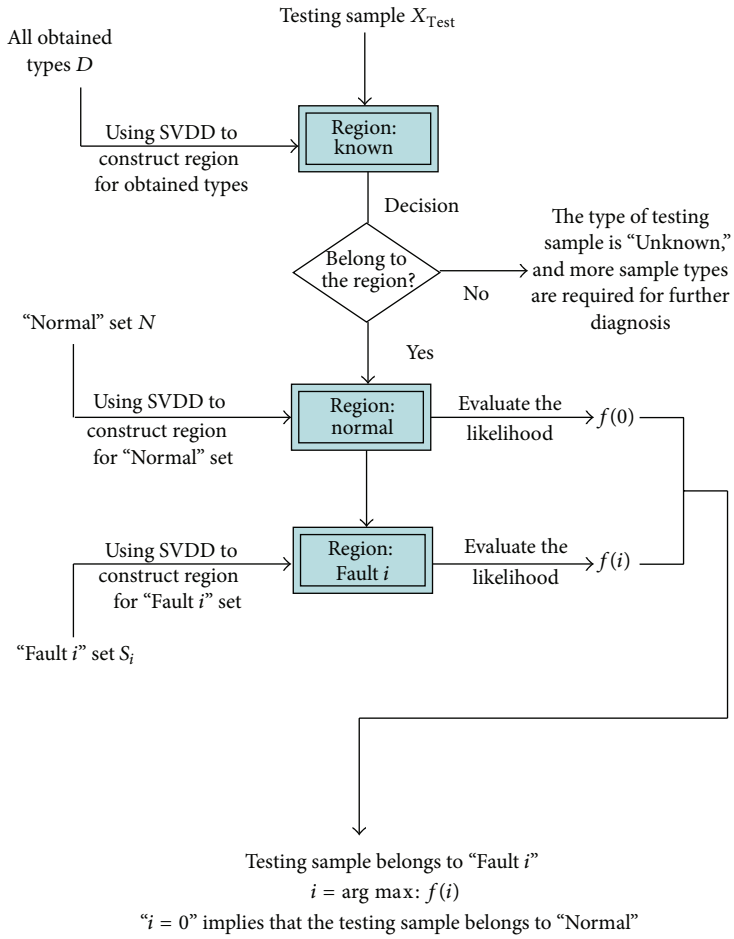


Figure 5: The basic idea of SVDD-based fault diagnosis.

(2) More sample types lead to a larger proportion of the gray area. And this indicates that the diagnostic ability could be improved as more sample types are acquired. Suppose a "Fault 2" sample was implemented as a testing sample. In (a) and (b), as the sample types for training are rare, the testing is classified to "Unknown" region. We can know that the testing sample belongs to neither "Normal" nor "Fault 1," but we do not know indeed which type it belongs to. However, in (c) and (d), as more sample types are imported for training and the fault regions are significantly refined, one can easily make an accurate diagnosis.

4.2. The Flowchart for the SVDD-Based Fault Diagnosis

4.2.1. The Definition of Evaluating Function f. Given a training set, $D = \bigcup_{i=1}^{k} S_i \cup N$, which consists of a normal sample set N and k types of faulty samples S_i. Using the SVDD approach, $k + 1$ hyperspheres, $\Omega_i = (a_i, R_i)$, $i = 0, 1, 2, \ldots, k$, can be obtained. Ω_0 denotes the hypersphere for normal set, and the other hyperspheres represent the faulty sets. These hyperspheres represent fault regions constructed by existing samples.

Definition 3. $f = R^2 - \|z - a\|^2$ is the evaluating function, which implies the likelihood that the testing sample belongs to the target set.

Here z is the testing sample; R and a are the radius and centre of the hypersphere, respectively. The centre, $a = \sum_{i=1}^{n} \alpha_i \cdot x_i$, and the radius can be obtained by calculating the distance between the centre and any relevance vector, x_{Sv} ($\alpha_{\text{Sv}} = 0$):

$$R^2 = \left\|x_{\text{Sv}} - a\right\|^2 = \left(x_{\text{Sv}} \cdot x_{\text{Sv}}\right) - 2\sum_{i=1}^{n} \alpha_i \left(x_i, x_{\text{Sv}}\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \left(x_i \cdot x_j\right). \tag{9}$$

Hence, the evaluating function can be rewritten as

$$f = R^2 - \|z - a\|^2 = \left(x_{\text{Sv}} \cdot x_{\text{Sv}}\right) - 2\sum_{i=1}^{n} \alpha_i \left(x_i \cdot x_{\text{Sv}}\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \left(x_i \cdot x_j\right)$$

Table 1: The relationship between F and fault type.

Fault type	Normal	Fault 1	⋯	Fault k	"Known" but hard to isolate
F value	0	1	⋯	k	$k+1$

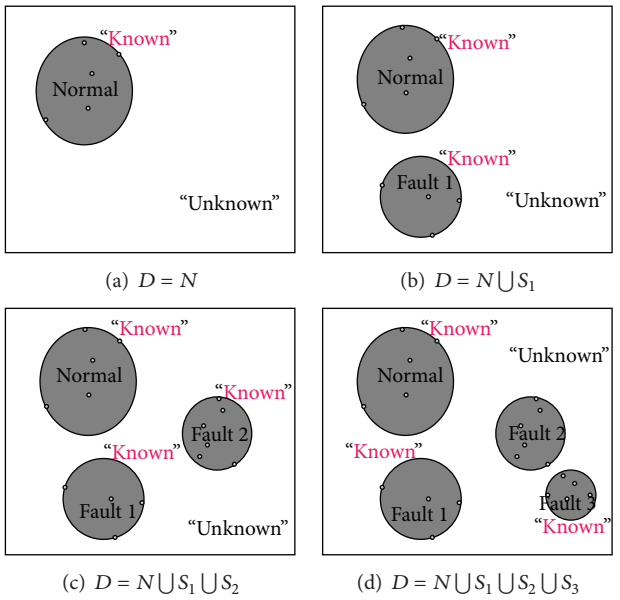


Figure 6: The process of fault region construction using SVDD.

$$
\begin{aligned}
&- \left\{ (z \cdot z) - 2\sum_{i=1}^{n} \alpha_i \left(z \cdot x_{\mathrm{Sv}}\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \left(x_i \cdot x_j\right) \right\} \\
&= \left(x_{\mathrm{Sv}} \cdot x_{\mathrm{Sv}}\right) - (z \cdot z) \\
&\quad - 2\left(\sum_{i=1}^{n} \alpha_i \left(x_i \cdot x_{\mathrm{Sv}}\right) - \sum_{i=1}^{n} \alpha_i \left(z \cdot x_{\mathrm{Sv}}\right) \right).
\end{aligned} \tag{10}
$$

A larger value of f implies a higher likelihood that the testing sample belongs to the same type as the target samples in the hypersphere. And we have

$$
\mathrm{dist} = \mathrm{sgn}\left(f\right) = \begin{cases} 1, & z \in \Omega \\ -1, & z \notin \Omega, \end{cases} \tag{11}
$$

where dist $= -1$ implies that the testing sample has not fallen into the hypersphere.

4.2.2. The Flowchart of the Proposed Approach. For a training set consisting of incomplete types, the SVDD approach has been introduced into the diagnosis [18–21]. The flowchart of the proposed approach is shown as follows.

Step 1. For the training set, $D = \bigcup_{i=1}^{k} S_i \cup N$, a hypersphere, $\Omega = (a, R)$, is constructed, which involves all existing samples while training. Given a testing sample, z, an index d is calculated:

$$
d = \mathrm{sgn}\left(R^2 - \|z - a\|^2\right). \tag{12}
$$

If $d = -1$, the testing sample belongs to "Unknown" types and requires more sample types to isolate the fault type. Otherwise, the testing sample belongs to the "Known" types, and Steps 2–4 are proposed for fault isolation.

Step 2. For the normal set, $N = S_0$, the training set can be written as $\bigcup_{i=0}^{k} S_i$. For S_i, the hypersphere, $\Omega_i = (a_i, R_i)$, is constructed by the SVDD approach.

Step 3. Given a testing sample, z, its evaluation function value can be calculated for all hyperspheres, $\Omega_i = (a_i, R_i)$, $i = 0, 1, 2, \ldots, k$.

$$
\begin{aligned}
f(i) &= R_i^2 - \left\|z - a_i\right\|^2 = \left(x_{\mathrm{Sv}} \cdot x_{\mathrm{Sv}}\right) - (z \cdot z) \\
&\quad - 2\left(\sum_{j=1}^{l} \alpha_j \left(x_j \cdot x_{\mathrm{Sv}}\right) - \sum_{j=1}^{l} \alpha_j \left(z \cdot x_{\mathrm{Sv}}\right) \right),
\end{aligned} \tag{13}
$$

where $x_{\mathrm{Sv}} \in S_i$, $S_i = \bigcup_{j=1}^{l} x_j$, α_j is the Lagrange multiplier for x_j, and l is the number of samples for S_i. We then get the decision value for diagnosis

$$
F = \begin{cases} \text{argmax: } f(i), & \exists f(i) \geqslant 0 \\ k+1, & \forall f(i) < 0. \end{cases} \tag{14}
$$

$F = k+1$ means that the testing sample has fallen into the "Known" region, but its likelihood for all types is very low. It is hard to decide which region should the testing sample belong to. In this paper, we just simply divide such sample to "Normal" set, with the consideration of reducing false alarms. The relationship between the F value and the fault type is shown in Table 1.

Step 4. According to the decision rules shown in Algorithm 1, the diagnostic decision can be calculated and the FI can be achieved. Here "decision = i" means the testing sample belongs to the ith fault type.

5. Experimental Validation

In this section, a group of Gaussian distributed synthetic datasets are firstly implemented to validate the feasibility and effectiveness of the proposed approach. Then, a real-world dataset for transformer faults is implemented. The performances of the SVDD-based approach are compared with the popular SVM-based approach [22–24] and some other classic approaches like Least Squares Support Vector Machine (LS-SVM) [25], learning vector quantization (LVQ) network [26], and random forests [27].

```
Switch  d
        Case  1
              if                    F = k + 1
                                    decision = 0
              else
                                    decision = F
              end
        Case  −1
                    Testing sample belongs to unknown types
end
```

Algorithm 1: Decision rules for the SVDD-based fault diagnosis.

5.1. The Synthetic Datasets. The training set consisted of three Gaussian distributed synthetic datasets, namely, Γ_1, Γ_2, and Γ_3, and each dataset represents one operating status, as shown in Figure 7.

5.1.1. One Type Is Missing. The given system yields three operating statuses: "Fault 1," "Fault 2," and "Normal," and their corresponding sample sets are "Γ_1," "Γ_2," and "Γ_3." The diagnostic performances of both SVM-based and SVDD-based approaches are first investigated in the condition with one sample type missing.

Suppose the sample set, Γ_3, is missing. We made the fault regions using the SVM-based approach and the proposed approach. As shown in Figure 8(a), the fault regions made by the SVM can efficiently identify the faulty types, Γ_1 and Γ_2; however, there is no region for Γ_3, and all samples belonging to Γ_3 will be misclassified into Γ_1 or Γ_2.

In contrast to the SVM-based approach, the proposed approach constructs an updatable framework for fault diagnosis and fully takes into consideration of the classification of unknown samples. As shown in Figure 8(b), where the fault regions for "Fault 1" and "Fault 2" are made, the SVDD-based approach also makes a region for unknown sample type. Therefore, the samples belonging to Γ_3 will be classified into the "Unknown" region instead of the "Fault 1" region or "Fault 2" region. This makes the diagnosis more reasonable.

100 samples were randomly selected as testing samples. We investigated the diagnostic accuracies of the two approaches for the condition of one sample type missing, which is shown in Table 2.

As shown in Table 2, the diagnostic accuracy of the SVM-based approach has been reduced from 100% to 64–71% when one of the sample types is missing. This implies that the sample-type incompleteness greatly affects the diagnostic accuracy. However, for the SVDD-based approach, the accuracy varies slightly when a sample type is missing. This demonstrates that the diagnostic accuracy of the SVDD-based approach is insensitive to the sample-type completeness. Moreover, when a sample type is missing, the accuracy of the SVM-based approach varies from 64% to 71% and the accuracy of the proposed approach varies from 86% to 89%. The proposed approach shows significant superiority to the classic SVM-based approach when dealing with incomplete sample types.

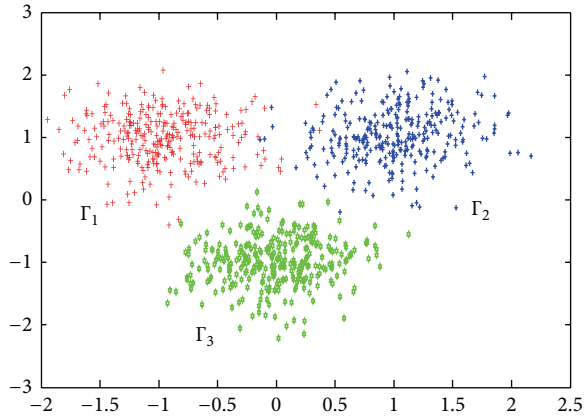
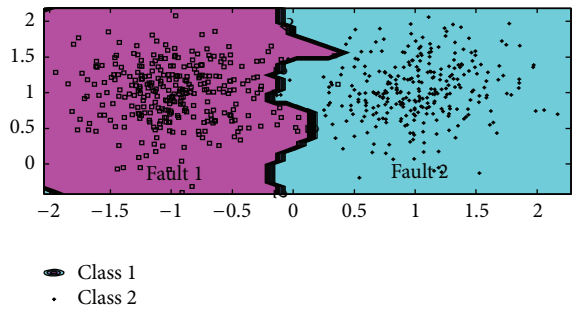
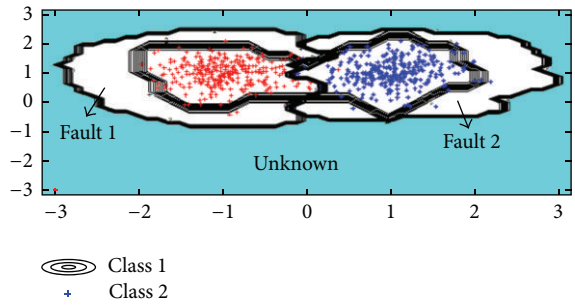


Figure 7: Three Gaussian distributed datasets.



(a) Fault regions made by SVM-based approach



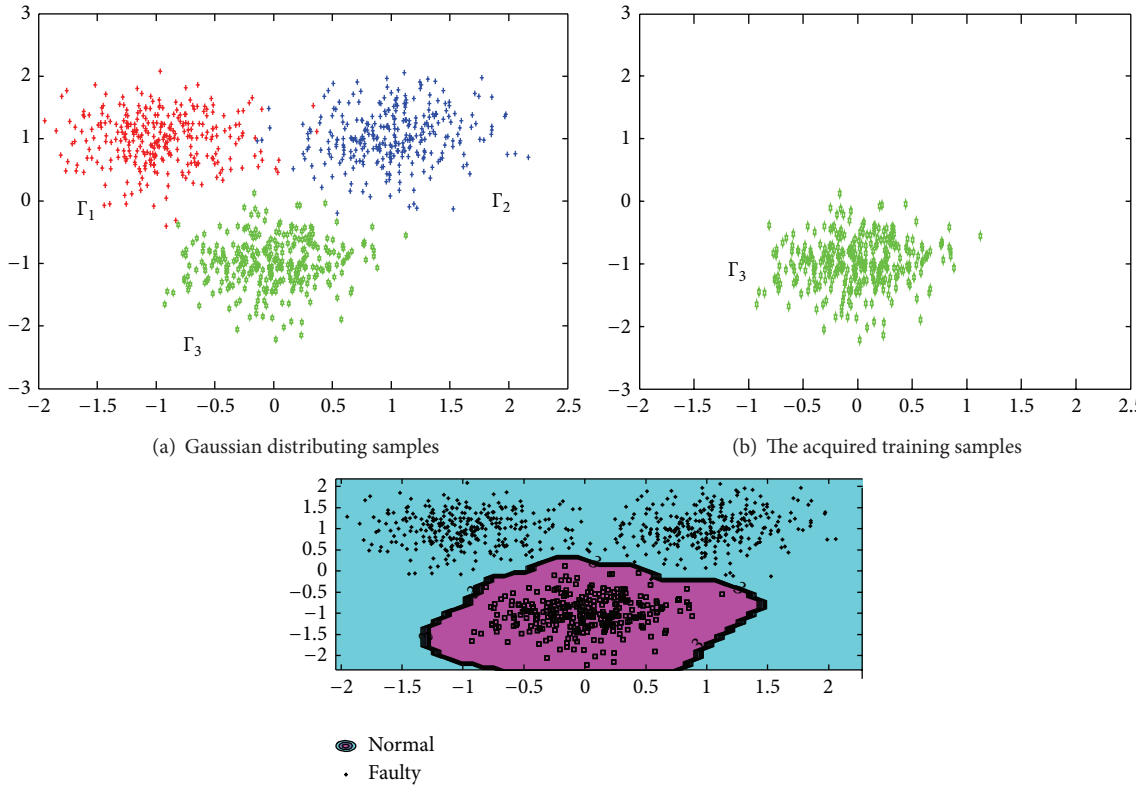
(b) Fault regions made by SVDD-based approach

Figure 8: Fault regions made by both SVM- and SVDD-based approaches (Γ_3 is missing).

Table 2: The diagnostic accuracy of the two approaches.

Missing type	Diagnostic accuracy of SVM-based approach	Diagnostic accuracy of SVDD-based approach
—	100%	89%
Γ_1	71%	87%
Γ_2	65%	86%
Γ_3	64%	89%

5.1.2. Fault Detection Using the SVDD-Based Approach. We suppose the system has acquired only the samples for a "Normal" status; that is, the training set is composed of Γ_3.



(a) Gaussian distributing samples

(b) The acquired training samples

(c) Fault detection using SVDD-based approach

Figure 9: FD using SVDD-based approach (only Γ_3 is acquired).

Because the conventional classifier-based approaches require at least two types of samples, the FD cannot be made. However, the SVDD-based approach solves the problem successfully. Using the proposed approach, the region for Γ_3 and the "Unknown" can be obtained. The region for Γ_3 is referred to as the "Normal" region and the "Unknown" region is thus referred to as the "Faulty" region. By locating which region the testing sample has fallen into, the FD is achieved.

As shown in Figure 9(c), the purple area is the "Normal" region and the blue area is the "Faulty" region. All samples belonging to Γ_3 fall into the "Normal" region and samples from Γ_1 and Γ_2 fall into the "Faulty" region. This demonstrates the feasibility of the proposed approach in FD.

5.2. The Real-World Dataset. The real-world dataset for a transformer fault diagnosis is obtained from the literature [28]. This dataset monitors the gas content dissolved in the insulating oil of the transformers, which has close relationships with the operating statuses. Here, 4 operating statuses are investigated, that is, "Normal" status, "High-energy discharge" fault, "Low-energy discharge" fault, and "Thermal heating" fault. The training set and testing set are prepared as shown in Table 3.

The radical basis function (RBF) kernel is selected for training the SVM and SVDD classifier and the parameter, σ, is set 0.5. In all of the experiments, the testing set is selected as $\{\Gamma_1', \Gamma_2', \Gamma_3', \Gamma_4'\}$ and the training set is selected, respectively; for example, when samples for "Low-energy discharge" fault are missing, the training set should be $\{\Gamma_1, \Gamma_2, \Gamma_4\}$.

Table 3: The dataset for the transformer fault diagnosis.

Faulty type	Training set	Testing set
Normal	$\Gamma_1 = \{x_i\}_{i=1}^{5}$	$\Gamma_1' = \{x_j\}_{j=1}^{4}$
Thermal heating	$\Gamma_2 = \{x_i\}_{i=1}^{25}$	$\Gamma_2' = \{x_j\}_{j=1}^{13}$
Low-energy discharge	$\Gamma_3 = \{x_i\}_{i=1}^{5}$	$\Gamma_3' = \{x_j\}_{j=1}^{6}$
High-energy discharge	$\Gamma_4 = \{x_i\}_{i=1}^{15}$	$\Gamma_4' = \{x_j\}_{j=1}^{2}$

In Figure 10, we investigated the diagnostic performances of the two approaches for the condition of each kind of sample-type incompleteness. The training sets selected are $\{\Gamma_1, \Gamma_2, \Gamma_3\}$, $\{\Gamma_1, \Gamma_2, \Gamma_4\}$, $\{\Gamma_1, \Gamma_3, \Gamma_4\}$, and $\{\Gamma_2, \Gamma_3, \Gamma_4\}$ for cases when one sample type is missing, and $\{\Gamma_1, \Gamma_2\}$, $\{\Gamma_1, \Gamma_3\}$, $\{\Gamma_1, \Gamma_4\}$, $\{\Gamma_2, \Gamma_3\}$, and $\{\Gamma_2, \Gamma_4\}$, $\{\Gamma_3, \Gamma_4\}$ for cases when two sample types are missing.

As shown in Figure 10, in all cases, the SVDD-based approach yields a better accuracy than the SVM-based approach. Moreover, for the SVM-based approach the accuracy is significantly decreased when more sample types are missing. The average accuracy for one sample type missing is

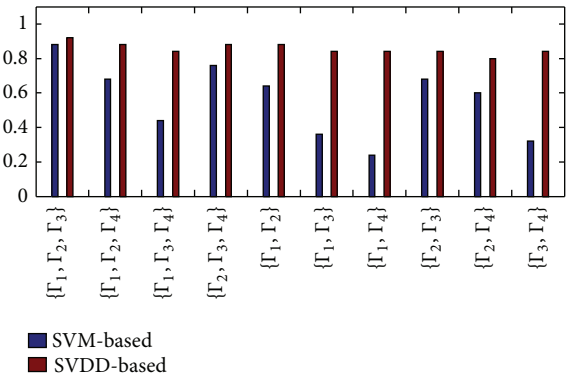


Figure 10: The comparison of diagnostic accuracy for the two methods for the condition of type incompleteness.

Table 4: Comparisons with classical methods.

Training set	Methods			
	SVDD	LS-SVM	LVQ network	Random forests
$\{\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4\}$	96%	100%	100%	100%
$\{\Gamma_1, \Gamma_2, \Gamma_3\}$	88%	84%	84%	80%
$\{\Gamma_1, \Gamma_2, \Gamma_4\}$	84%	68%	72%	68%
$\{\Gamma_1, \Gamma_3, \Gamma_4\}$	80%	56%	64%	56%
$\{\Gamma_2, \Gamma_3, \Gamma_4\}$	84%	68%	64%	64%
$\{\Gamma_1, \Gamma_2\}$	84%	64%	64%	68%
$\{\Gamma_1, \Gamma_3\}$	80%	48%	52%	56%
$\{\Gamma_1, \Gamma_4\}$	80%	40%	44%	44%
$\{\Gamma_2, \Gamma_3\}$	80%	60%	56%	60%
$\{\Gamma_2, \Gamma_4\}$	76%	56%	56%	60%
$\{\Gamma_3, \Gamma_4\}$	80%	48%	52%	52%

69% and when two sample types are missing the average accuracy falls to 47.33%. Conversely, the SVDD-based approach varies slightly while different portions of sample types are missing. The average accuracy for the proposed approach varies from 76% to 88%. This implies that the SVDD-based approach can efficiently improve the classifier-based diagnostic performance while the sample type is incomplete.

For a further comparison, classical methods such as Least Squares Support Vector Machine (LS-SVM) [26], learning vector quantization (LVQ) network [27], and random forests [28] are also introduced to solve the problem. For LS-SVM, the parameter $\gamma = 0.1$, and $\sigma = 0.5$; for LVQ network, the size of hidden layer is set 10, and the LVQ learning rate is set 0.01; for random forests, the parameter "mtry" is set 3, and the parameter "ntree" is set 500.

Training sets with different completeness are employed as the above experiment. The diagnostic performances for SVDD and these methods are investigated and shown in Table 4.

For most missing types, traditional methods have made an incorrect classification and lead to false alarms. But the SVDD-based approach just divides these samples into "Unknown" types. This result informs the users that more sample types are required for diagnosing the tested sample. The false alarms will greatly be reduced.

6. Conclusion and Discussion

The performance of the data-driven approach depends on the quality of the training set. However, a high quality training set is not easily acquired. Therefore, the remaining problem in both theoretic research and practical applications is how to make a reasonable fault diagnosis in the condition of a low quality training set.

In this paper, we have discussed the problem of sample-type incompleteness; that is, when some types of training samples are missing the diagnostic accuracy is commonly decreased. An SVDD-based diagnosis approach has been addressed for this issue. This approach provides a new framework for diagnosing the incomplete samples, which implements the one-class classifier, that is, SVDD, instead of conventional binary or multiclass classifiers and attempts to improve the diagnostic accuracy in this situation.

The effectiveness of the proposed approach has been validated by comparative experiments on both synthetic and practical training sets. The proposed approach has shown a significant superiority to the popular SVM-based fault diagnosis approach. This demonstrates that the SVDD-based approach is insensitive to the sample-type completeness and can efficiently increase the diagnostic accuracy when dealing with incomplete training sets.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work described in this paper is supported by grants from the Doctoral Fund of Ministry of Education of China (20113218110011 and 20113218120010), the National Science Foundation of China (61171191, 61203020, and 61401215), the Science Foundation of Jiangsu province (BK20140953), and the Science Foundation of Jiangsu High Schools (13KJB510013).

References

[1] S. X. Ding, *Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tool*, Springer, Berlin, Germany, 2008.

[2] J. Chen and R. J. Patton, *Robust Model-Based Fault Diagnosis for Dynamic Systems*, Kluwer Academic Publishers, Boston, Mass, USA, 1998.

[3] B. Jiang, Z. Mao, H. Yang et al., *Fault Diagnosis and Fault Accommodation for Control Systems*, National Defense Industry Press, Beijing, China, 2009.

[4] H. Wang, T.-Y. Chai, J.-L. Ding, and M. Brown, "Data driven fault diagnosis and fault tolerant control: some advances and possible new directions," *Acta Automatica Sinica*, vol. 35, no. 6, pp. 739–747, 2009.

[5] R. Isermman, "Experiences with process FD via parameter estimation," in *System Fault Diagnostics, Reliability & Related Knowledge-Based Approaches*, pp. 3–33, 1987.

[6] P. M. Frank, "Fault diagnosis in dynamic system via state estimation—a survey," in *System Fault Diagnostics, Reliability & Related Knowledge-Based Approaches*, pp. 35–98, 1987.

[7] E. Russell, L. H. Chiang, and R. D. Braatz, *Data-Driven Methods for FD and Diagnosis in Chemical Processes*, Springer, Berlin, Germany, 2000.

[8] S. M. Namburu, M. S. Azam, J. Luo, K. Choi, and K. R. Pattipati, "Data-driven modeling, fault diagnosis and optimal sensor selection for HVAC chillers," *IEEE Transactions on Automation Science and Engineering*, vol. 4, no. 3, pp. 469–473, 2007.

[9] D. Zhou, G. Li, and Y. Li, *Data-Driven Fault Diagnostic Techniques for Industrial Processes: Based on PCA and PLS*, Science Press, Beijing, China, 2011.

[10] R. Isermann, *Fault-Diagnosis Systems: An Introduction from FD to Fault Tolerance*, Springer, Berlin, Germany, 2006.

[11] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis part I: quantitative model-based methods," *Computers and Chemical Engineering*, vol. 27, no. 3, pp. 293–311, 2003.

[12] K. S. Gaeid and H. A. F. Mohamed, "Diagnosis and fault tolerant control of the induction motors techniques: a review," *Australian Journal of Basic and Applied Sciences*, vol. 4, no. 2, pp. 227–246, 2010.

[13] M. Catelani and A. Fort, "Fault diagnosis of electronic analog circuits using a radial basis function network classifier," *Measurement*, vol. 28, no. 3, pp. 147–158, 2000.

[14] S.-F. Yuan and F.-L. Chu, "Support vector machines-based fault diagnosis for turbo-pump rotor," *Mechanical Systems and Signal Processing*, vol. 20, no. 4, pp. 939–952, 2006.

[15] J. Hakkila, T. W. Giblin, R. J. Roiger, D. J. Haglin, W. S. Paciesas, and C. A. Meegan, "How sample completeness affects gamma-ray burst classification," *The Astrophysical Journal*, vol. 582, no. 1, pp. 320–329, 2003.

[16] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[17] J. R. Bunch and L. C. Kaufman, "A computational method for the indefinite quadratic programming problem," *Linear Algebra and Its Applications*, vol. 34, pp. 341–370, 1980.

[18] K. Y. Lee, D.-W. Kim, K. H. Lee, and D. Lee, "Density-induced support vector data description," *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 284–289, 2007.

[19] T. Mu and A. K. Nandi, "Multiclass classification based on extended support vector data description," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 5, pp. 1206–1216, 2009.

[20] H. Luo, Y. Wang, and J. Cui, "A SVDD approach of fuzzy classification for analog circuit fault diagnosis with FWT as preprocessor," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10554–10561, 2011.

[21] M.-Z. Tang, Y.-B. Wang, and C.-H. Yang, "Modified support vector data description for fault diagnosis," *Control and Decision*, vol. 26, no. 7, pp. 967–972, 2011.

[22] H. Yi, X.-F. Song, B. Jiang, and D.-C. Wang, "Support vector machine based on nodes refined decision directed acyclic graph and its application to fault diagnosis," *Acta Automatica Sinica*, vol. 36, no. 3, pp. 427–432, 2010.

[23] X. Song, S. K. Halgamuge, D. Chen, S. Hu, and B. Jiang, "The optimized support vector machine with correlative features for classification of natural spearmint essence," *International Journal of Innovative Computing, Information and Control*, vol. 6, no. 3, pp. 1089–1099, 2010.

[24] Y. Zhang, "Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM," *Chemical Engineering Science*, vol. 64, no. 5, pp. 801–811, 2009.

[25] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[26] P. Burrascano, "Learning vector quantization for the probabilistic neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 4, pp. 458–461, 1991.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] G. Lv, H. Cheng, H. Zhai, and L. Dong, "Fault diagnosis of power transformer based on multi-layer SVM classifier," *Electric Power Systems Research*, vol. 75, no. 1, pp. 9–15, 2005.