Incremental Weighted Support Vector Data Description Method for Incipient Fault Detection of Rolling Bearing

Chao Yang
School of Computer and Information Engineering
Henan Normal University
Xinxiang, CHINA, 453007
Email: yc_mat@qq.com

Wentao Mao*
School of Computer and Information Engineering
Henan Normal University
Xinxiang, CHINA, 453007
Email: maowt@htu.edu.cn

Yamin Liu
School of Computer and Information Engineering
Henan Normal University
Xinxiang, CHINA, 453007
Email: ym2352@qq.com

Siyu Tian
School of Computer and Information Engineering
Henan Normal University
Xinxiang, CHINA, 453007
Email: tianxiaosi77@qq.com

***Abstract*—Incipient fault detection is a key technical link in the field of rolling bearing prognostic and health management. The traditional incipient fault detection models are generally built on offline data and unable to update timely for matching the online data of target bearing. Moreover, when using the anomaly detection algorithm represented by support vector data description (SVDD) for incipient fault detection, it is easy to cause high false alarm rate due to slight and anomalous fluctuation of online data. To solve the above problems, an incremental weighted support vector data description (IW-SVDD) is proposed for incipient fault detection of rolling bearing. First, we train an initial SVDD detection model based on a small amount of online data that exists at initial stage, and use this model to pre-detect the sequentially-arrived online data. Second, in order to adapt the detection model to the anomalous fluctuations of online data, we design a strategy to determine the sample state. This strategy divides the fluctuation of online data into four states: abnormal appears, abnormal appears in succession, abnormal disappears and abnormal re-appears. Then we assign proper weights on the corresponding samples in each state. Finally, we update training set repeatedly by replacing the earliest samples in the training set with the same amount of samples which violates KKT condition in pre-detection results. In this way, the detection model is re-trained in incremental mode. Experiment results on IEEE PHM Challenge 2012 show that the proposed IWSVDD model can effectively reduce false alarm rate with ensuring detection accuracy.**

Keywords-Incipient fault detection; Anomaly detection; SVD-D; False alarm; Incremental learning

I. INTRODUCTION

As an important component of mechanical equipment, rolling bearings are easily damaged under complex working condition. Once accidental failure of bearing occurs, a huge damage even casualty will be caused. Therefore, incipient fault detection of bearing is extremely important. Accurate and reliable incipient fault detection becomes a key technical link in prognostic and health management (PHM) [1]. How to improve the accuracy and real-time performance of incipient bearing fault detection has clear academic value and application requirement.

Since vibration signal is able to directly reflect the working condition of rolling bearing, it is widely studied in existing incipient fault detection methods. The traditional incipient fault detection methods, i.e. signal analysis-based methods, mainly work on signal analysis. For example, [2] utilized approximate entropy of different frequency components at the time of incipient fault occurrence. [3] filtered noise and extracted incipient fault features by means of wavelet decomposition. However, this kind of methods is prone to produce false alarm due to the anomalous fluctuation of normal state data. In recent years, with the rapid development of machine learning techniques, more and more attentions have been paid on data-driven intelligent method. This kind of method usually contains two steps: 1) feature extraction and 2) model construction. Traditional hand-crafted features are statistical features in time/frequency/time-frequency domains from raw signal. We can also adopt deep features extracted from long-short time memory (LSTM) [4, 5], sample enhancement network [6], and deep auto-encoder (AE) [7] etc. Detection model construction is then conducted using one-class classification algorithm based on normal state data of offline bearing or the earliest data of target bearing. However, this kind of methods usually is unable to update the model in real time in online detection scenario, and would not avoid the negative effect of anomalous fluctuation of normal state data.

As a typical one-class classification algorithm, support vector data description (SVDD) [8] only needs normal data to train an anomaly detection model, then SVDD is often used for incipient fault detection. Current research focus of incipient fault detection is to improve SVDD model theoretically according to actual requirements. From this aspect, Cha et al. [9] utilized K-nearest neighbor to conduct classification, and took the distance between target samples and class-center as the weight of slack variable in SVDD model. However, this method didn't consider the characteristic of online anomaly detection. The model couldn't adapt to the anomalous fluctuation of online data and is easy to cause false alarms. Yin et al. [10] proposed an online anomaly detection method based on incremental SVDD and extreme learning machine (ELM) [11] with incremental output structure. This

978-1-7281-0860-5/19/$31.00 ©2019 IEEE

method uses incremental SVDD to find new fault pattern, and incrementally adjusts the output neurons of ELM. But this method doesn't consider the significance of different samples and the impact of anomalous data fluctuation in model update.

To sum up, due to anomalous data fluctuation in normal state caused by noise and working condition, the normal data is easily misjudged as anomaly, i.e. false alarm. Although some works [9, 10] have solved the above problems to some extent by using weighted and incremental learning pattern, it still ignores the false alarm problem in the scenario of online detection. To adapt detection model to anomalous fluctuation of online data, this paper develops SVDD model by integrating weighted learning mode and incremental learning mode together. In this way, the detection model is not only dynamically updated using online data, but also can be adjusted according to the importance of different samples. By intuitive understanding, the above mechanism can calculate the weight of different samples dynamically based on the detection result of online data batch, so that the hyper-sphere radius of SVDD model would be expanded or shrinked to adapt the fluctuation of online data and then avoid the negative influence of data fluctuation.

Based on the above analysis, this paper proposes an incremental weighted support vector data description (IW-SVDD) algorithm for incipient fault detection of rolling bearing. Specifically, we adopt a widely-used deep learning technique, i.e. deep AE, to extract deep feature of online data. In order to adjust sample weight according to the detection results of online data, this paper divides data fluctuation into four interference scenarios: anomaly appears, anomaly appears in succession, anomaly disappears and anomaly re-appears. By assigning corresponding weights to the samples which violates KKT condition , while replacing the earliest samples in the training set by these samples, an incremental weighted learning model is conducted. Experiment results on IEEE PHM Challenge 2012 dataset show the proposed method can effectively reduce false alarm rate with ensuring an acceptable accuracy.

The main contribution of this paper can be summarized as follows:

1) This paper proposes a new SVDD algorithm for incipient fault detection. This algorithm integrates weighted strategy into dynamic model update process for online data, and dynamically calculates sample weights according to different interference scenarios. Then the dynamic model update and optimization can be realized.

2) This paper proposes an online detection strategy for bearing incipient fault. This strategy is based on weighted and incremental learning modes, and can effectively reduce false alarm rate for incipient fault while maintaining detection accuracy. What's more, this strategy can be generalized well to other traditional anomaly detection algorithms.

II. BACKGROUND

A. Stacked Auto-Encoder

The sketch map of Stacked Auto-Encoder is shown in Fig.1. SAE is stacked by multiple Auto-Encoders (AEs). Auto-Encoder is a three layers neural network for unsupervised feature extraction. Given training sample $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T \in \Re^N$, the encoder transforms x into a code representation $\mathbf{h} = [h_1, h_2, \cdots, h_M]^T \in \Re^M$ using the function $\mathbf{h} = f(\mathbf{Wx} + \mathbf{b})$ where W is input weight matrix, b is bias vector, and f is activation function. The output of hidden layer is reconstructed through the decoder $\mathbf{z} = g(\mathbf{W'h} + \mathbf{b'})$ where $\mathbf{W'}$ is output weight matrix. The goal of network optimization is to make the reconstruction error between $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T \in \Re^N$ and $\mathbf{z} = [z_1, z_2, \cdots, z_N]^T \in \Re^N$ as small as possible. The output of the hidden layer can be regarded as a feature representation of input vector when the reconstruction error is small enough.

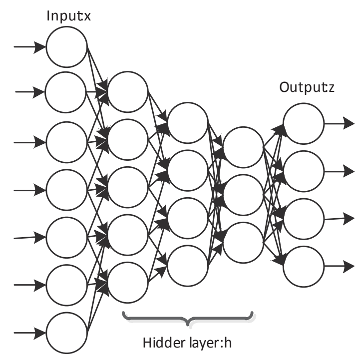


Fig. 1: Stacked Auto-Encoder

SAE is trained by minimizing the loss function: $L(\mathbf{x}, \mathbf{z}) = -\sum_{i=1}^{N} [x_i \log(z_i) + (1 - x_i) \log(1 - z_i)]$. Gradient descent method is used to optimize the weight matrix and bias matrix.

B. Support Vector Data Description

SVDD is one-class classification method which seeks a hyper-sphere with smallest radius to contain all or most of training samples in feature space. A sample which falls into the hyper-sphere is regarded as a normal sample, otherwise the sample is considered as an anomaly.

Given a training set $\{x_i | x_i \in R^d, i = 1, 2, \cdots, n\}$. The objective function of SVDD is [8]:

$$\begin{cases} \min\limits_{R,a,\xi} & R^2 + C \sum\limits_{i=1}^{l} \xi_i \\ s.t. & \|x_i - c\|^2 \leq R^2 + \xi_i \;\; i = 1, ..., l \\ & \xi_i \geq 0, i = 1, ..., l \end{cases} \tag{1}$$

where R is the radius of hyper-sphere, the parameter C controls the trade-off between hyper-sphere volume and empirical error, and ξ_i is slack variable, C is center of the hyper-sphere. Equation (1) can be solved by using Lagrange multiplier method. Please refer to [8] for more details.

III. INCREMENTAL WEIGHTED SUPPORT VECTOR DATA DESCRIPTION

Affected by noise and complex working condition, anomalous fluctuation inevitably occurs in normal state of target bearing. Then detection model could misclassify a normal sample as anomaly. To avoid this phenomenon, it is better to assign different weights on the samples with anomalous fluctuation according to bias level. After updating the training set and retraining SVDD, the detection model would then be adapted to the fluctuation and the false alarm rate would be reduced as well.

In original SVDD, only a same weight with value 1 is assigned to the slack variable of each sample, which leads to the following fact: although some samples are more abnormal, the punishment to them is same to the samples with small deviation. If a sample is given a small weight on slack variable, its contribution to the model will become small. As a result, the SVDD model will be inclined to set this sample as non-support vector. In other words, the model will attempt to keep this sample within the hyper-sphere. Based on this understanding, by combining weighting strategy with incremental updating strategy, this section proposes an incremental weighted SVDD.

A. Weighted SVDD

As there are many studies on weighted SVDD, this section first gives a brief introduction. The objective function and constraints are as follows[9]:

$$\begin{cases} \min\limits_{R,a,\xi} & R^2 + C\sum\limits_{i=1}^{l} w_i\xi_i \\ s.t. & \|x_i - c\|^2 \le R^2 + \xi_i \\ & \xi_i \ge 0, i = 1, ..., l \end{cases} \tag{2}$$

where w_i is the weight on slack variable. Using Lagrange method to optimize Equation (2), we have:

$$L = R^2 + C\sum_i w_i\xi_i - \sum_i \alpha_i\{R^2 + \xi_i - \|x_i - c\|^2\} - \sum_i \gamma_i\xi_i \tag{3}$$

Setting partial derivatives to zero gives the constraints, we have

$$\begin{cases} \frac{\partial \mathrm{L}}{\partial R} = 0: & \sum\limits_i \alpha_i = 1 \\ \frac{\partial \mathrm{L}}{\partial c} = 0: & c = \sum\limits_i \alpha_i x_i \\ \frac{\partial \mathrm{L}}{\partial \xi_i} = 0: & C\omega_i - \alpha_i - \gamma_i = 0 \end{cases} \tag{4}$$

Re-substituting Equations (4) into Equation (3), we have:

$$\begin{cases} L = \sum\limits_i \alpha_i K(x_i \cdot x_i) - \sum\limits_{i,j} \alpha_i\alpha_j K(x_i \cdot x_j) \\ s.t. \quad 0 \le \alpha_i \le Cw_i \quad i = 1, 2, ..., l \\ \qquad \sum\limits_i \alpha_i = 1 \end{cases} \tag{5}$$

We use RBF kernel function in our SVDD model.

According to Equations (5), we find the weight will change the position of hyper-sphere center and the range of Lagrange multiplier as well. Thus the detection results will be adjusted accordingly. For IW-SVDD, its KKT condition can be defined as:

$$\begin{cases} \alpha_i \ge 0, \gamma_i \ge 0, \\ \|x_i - a\|^2 - R^2 - \xi_i \le 0 \\ \alpha_i(\|x_i - a\|^2 - R^2 - \xi_i) \le 0 \\ \xi_i \ge 0, \lambda_i\xi_i = 0 \end{cases} \tag{6}$$

To classify an object z, the distance to the hyper-sphere center has to be calculated as follows:

$$\|z - c\|^2 = K(z \cdot z) - 2\sum_i \alpha_i K(z \cdot x_i) + \sum_{i,j} \alpha_i\alpha_j K(x_i \cdot x_j) \tag{7}$$

B. Sample state determination

Since the rolling bearing is easily affected by noise and operating environment, the data in normal state have inevitable fluctuation. However, if we directly use offline data or some online data at initial stage to train a detection model, it is hard to adapt such anomalous fluctuation and then very easy to cause false alarm. Considering the characteristics of anomalous fluctuation in normal data, this paper divides the data fluctuation into four scenarios and designs a sample state determination (SSD) strategy. This strategy can calculate different weight values according to corresponding states, so that the detection model can be dynamically adjusted to reduce false alarm. Also, SSD determines when the model needs to be updated.

Please note the four scenarios of data fluctuation are inspired by [12]. In [12], online data are divided into four operating states, and a score can be calculated and accumulate. Alarm will be triggered when this score exceeds a pre-fixed threshold. These four states are displayed in Fig. 2. State A indicates the beginning of anomaly appearance; State B represents the situation that anomalies appear in succession. State C is the stage in which anomalies disappear while their previous contexts are abnormal. State D represents the stage in which anomalies re-appear.

In this paper, we integrate the alarm stategy into detection model, rather than an independent alarm strategy which links to detection process like [12]. To determine when the model needs to be updated, this paper sums all weights in the four states to a score. The model continues to be updated until the score reaches a threshold. It has been verified by many experiments that the threshold set to 1 will get best detection result.

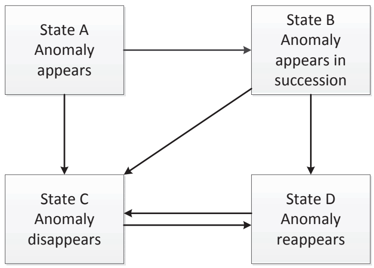


Fig. 2: Sample state transition

The weight of sample in State A can be calculated as follows:

$$WeightA = sigmoid(1 + \exp^{RMS_i - RMS_{average}}) \tag{8}$$

In Equation (8), factors that mainly correlate with *WeightA* include: the *RMS* of current sample (RMS_i) and the average *RMS* of all samples in original training set ($RMS_{average}$). The term 1 in Equation (10) represents the anomaly sample itself. Since *RMS* reflects the level of a signal energy, the operator *exp* can represent the degree of abnormality.

State B: Anomaly appears in succession. The weight of sample in *State B* can be calculated as follows:

$$WeightB = sigmoid(n_{th} + \exp^{RMS_i - RMS_{average}}) \tag{9}$$

where the term n_{th} denotes the number of anomalies that appear in succession. The larger this number is, the higher the importance of the sample is.

State C: Anomaly disappears. The weight of sample in *State C* can be calculated as follows:

$$WeightC = -sigmoid(n_{error}/\Delta) \tag{10}$$

where n_{error} is the number of monitored samples since the last anomaly occurred. n_{error} is 1 when the anomaly first disappears. Δ is regulatory factor. In this paper, Δ is set to 0.2. The samples in *State C* are only used to determine when the model needs to be updated and do not participate in sample weighting, because this paper mainly considers the false alarm problem.

State D: Anomaly re-appears. There are two kinds of weight calculation rules in *State D*, as follows. It may be a real fault sample when its RMS is greater than $RMS_{average}$. Therefore, its weight should be set large enough so that the model can detect it as an anomaly. Otherwise, its weight should be set smaller to ensure that the model can set the sample as non-support vector.

$$f(RMS_i) = \begin{cases} RMS_i & RMS_i > RMS_{ave} \\ RMS_i - RMS_{average} & RMS_i \leq RMS_{ave} \end{cases} \tag{11}$$

C. Algorithm description

In IW-SVDD, if the pre-detection results indicate some online samples violate KKT condition, SSD will be aroused to calculate the weights for these samples. Then replace the earliest samples in the training set by these samples. After updating the training set, the detection model should be retrained. In this way, the support vectors in online data that violate KKT condition can be converted to non-support vectors. In other words, the normal samples that are misclassified due to anomalous fluctuation will be contained within the hypersphere after adjusting the detect model by SSD.

The detailed steps of IW-SVDD are shown as follows:

The flowchart of IW-SVDD is shown in Fig.3

Algorithm1. IW-SVDD with SSD

Input:AE features of raw vibration signal. The sequence data of the first batch is noted as T_0.
Output: The output Y
Step 1: Train IW-SVDD model with existing online data by Equation (3) with all weight set 1. Then, use Equation (9) for pre-detection and obtain the detection result Y_0. Initialize the score to 0
Step 2: If (score>threshold): Output Y_0
else:
(1) Calculate the weight for each sample according to Equations (10)-(13), and update the score.
(2) Select the sample set R_0 in T_0 that violates KKT condition, and replace the same amount earliest samples in training set with R_0. Denote the updated training set by S_1.
(3) Retrain IW-SVDD using Equation (2) and obtain the updated model. Set the new detection result as Y_1.
Step 3: Go to Step 2 if new online data batch arrives, else set Y_1 as a final output Y.

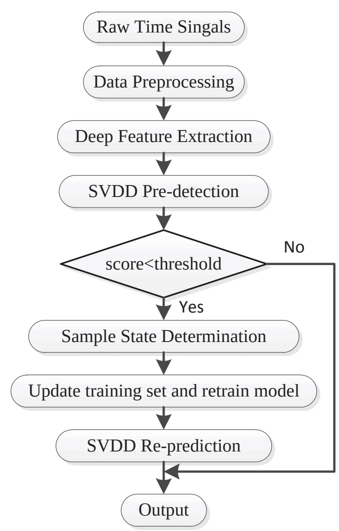


Fig. 3: Flowchart of the proposed method

IV. Experiments

A. Data description

The dataset used in this paper comes from the open IEEE PHM 2012 Data Challenge(downloaded by http://csegroups.case.edu/bearingdatacenter/pages/download-data-file). Collected from the test platform named PRONOSTIA (shown in Fig. 4), this dataset provides a

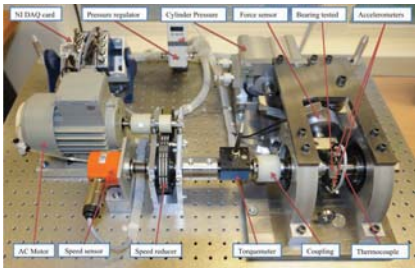


Fig. 4: PRONOSTIA test platform [13]

series of run-to-failure vibration signal by conducting an accelerated degradation test.

B. Data preprocessing

In this paper, Hilbert-Huang Transform(HHT) [14] which is suitable for non-stationary signal is used to preprocess raw vibration signal as input of AE. Compared with raw vibration signal, it can be observed from the obtained marginal spectrum that fault changes obviously in high frequency part. Therefore, by using HHT to preprocess the raw signal through time-frequency domain transformation, the trend information of degradation process can be revealed in more detail, which benefits for the further deep feature extraction.

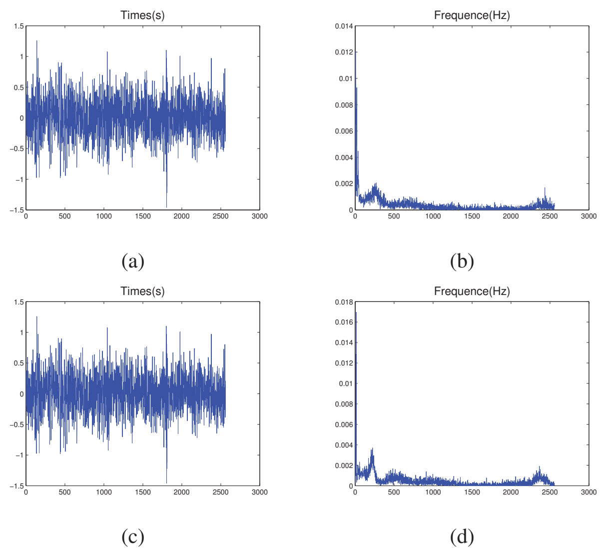


Fig. 5: Comparison of vibration signal in time domain and HHT marginal spectrum. Here the subfigure (a) and (c) are the raw signal of bearing 2 and 4 under the first working condition, (b) and (d) are the corresponding HHT marginal spectrum.

C. Deep feature representation

In this section, we utilize deep AE to extract deep features of the HHT marginal spectrum extracted from the original signal. For the structure of AE network, three hidden layers are set with size [500, 200, 50], and the learning rate is set to 0.001. After 1000 iterations, the training loss reaches around 0.003. The deep features are finally obtained by taking the whole sample set as input. In order to observe the feature distribution in different feature spaces, PCA and T-SNE method are used to reduce the dimension of the obtained features. Fig. 6 shows their distribution characteristic. Obviously, the features of normal samples and abnormal samples are still not completely separable even though extracted by deep AE, which will cause false alarm in online detection. The experimental environment is: CPU i7-4790, Memory: 8G, Windows 7 operation system, Python 3.5 with Tensorflow 1.2.0, MATLAB R2014a.

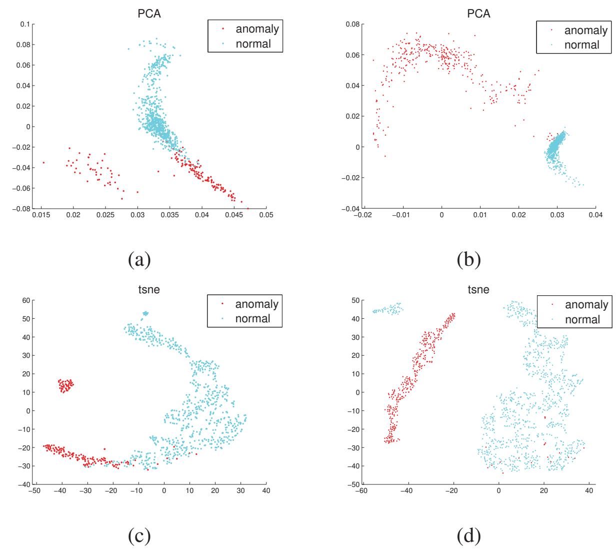


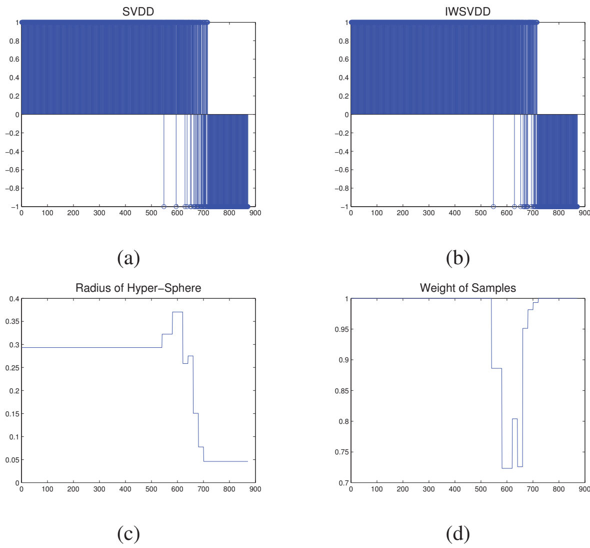
Fig. 6: Distribution of deep features whose dimension is reduced by PCA and T-SNE. Here the subfigure (a) and (b) are for the bearing 2, (c) and (d) are for the bearing 4.

D. Experimental results

To establish the initial SVDD, the first 500 samples of the bearings 2 and 4 are used for feature extraction by deep AE. To simulate the online scenario, 20 samples are selected sequentially as a data batch. The detection results are listed in Fig. 7 and Fig. 8.

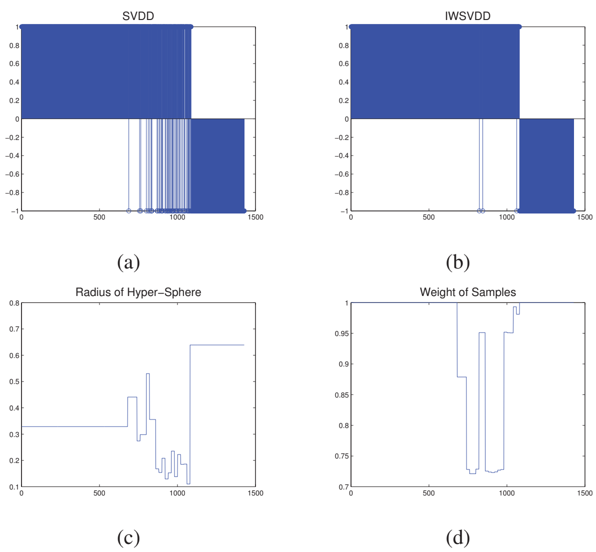
The experimental results in Fig. 7 and Fig. 8 both show that the radius of hyper-sphere will change in line with the trend of sample weight, which also make the change of detection results. As shown in Fig. 7(c), once the weight decreases, the radius of the hyper-sphere increases, and vice versa. This is because that a sample's contribution to the model will be enhanced when its weight become larger, which indicates that this sample will be placed on or outside the hyper-sphere. If the weight is set smaller, the sample has less contribution and then tends to be pushed into the hyper-sphere (equal to have a larger radius). Consequently, the fluctuation of normal state data can be adapted effectively by setting weights in SVDD model, and SSD strategy works well for calculating the proper sample weights.

Furthermore, to provide a comprehensive comparison, the number of abnormal samples before the fault state, i.e., false alarm number, is also counted. The fault state is considered



(a) (b) (c) (d)

Fig. 7: Comparative detection results of IW-SVDD and SVDD for the bearing 2. Here the subfigures (a) and (b) are from SVDD and IW-SVDD respectively, the subfigures (c) and (d) are the radius of hyper-sphere and sample weight respectively.



(a) (b) (c) (d)

Fig. 8: Comparative detection results of IW-SVDD and SVDD for the bearing 2. Here the subfigures (a) and (b) are from SVDD and IW-SVDD respectively, the subfigures (c) and (d) are the radius of hyper-sphere and sample weight respectively.

from the point after which only abnormal samples are detected. This point is called here the *alarm time point*. The comparative results are listed in TABLE. 1. It can be observed that the proposed method can effectively reduce the false alarm rate with ensuring the accuracy of detection results.

TABLE I: Comparative results of SVDD and IW-SVDD

Data	Bearing 2	Bearing 4
False alarm number of SVDD	27	53
False alarm number of IW-SVDD	17	3
Alarm time point of SVDD	717	1088
Alarm time point of IW-SVDD	717	1081

V. Conclusions

This paper presents an new detection method for bearings early fault detection. This method integrates incremental learning strategy and weighted learning strategy into SVDD to improve its robust performance for online detection. The most vital innovation is proposing a sample state determination strategy to calculate effective weights for online data. The experimental results show the proposed method can keep the detection results unchanged or even slightly better, with much less false alarm rate.

VI. Acknowledgement

This work was supported by the National Natural Science Foundation of China(No. U1704158, 11702087), and China Postdoctoral Science Foundation Special Support(No. 2016T90944)

References

[1] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, “Machinery health prognostics: A systematic review from data acquisition to rul prediction,” *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, 2018.

[2] R. Yan and R. Gao, “Machine health diagnosis based on approximate entropy,” in *Proceedings of the 21st IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No. 04CH37510)*, vol. 3. IEEE, 2004, pp. 2054–2059.

[3] H. Qiu, J. Lee, J. Lin, and G. Yu, “Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics,” *Journal of sound and vibration*, vol. 289, no. 4-5, pp. 1066–1090, 2006.

[4] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, “Long short term memory networks for anomaly detection in time series,” in *Proceedings*. Presses universitaires de Louvain, 2015, p. 89.

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, “Adversarially learned one-class classifier for novelty detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3379–3388.

[7] W.-b. Huang, F.-c. Sun *et al.*, “Building feature space of extreme learning machine with sparse denoising stacked-autoencoder,” *Neurocomputing*, vol. 174, pp. 60–71, 2016.

[8] D. M. Tax and R. P. Duin, “Support vector data description,” *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[9] M. Cha, J. S. Kim, and J.-G. Baek, “Density weighted support vector data description,” *Expert Systems with Applications*, vol. 41, no. 7, pp. 3343–3350, 2014.

[10] G. Yin, Y.-T. Zhang, Z.-N. Li, G.-Q. Ren, and H.-B. Fan, “Online fault diagnosis method based on incremental support vector data description and extreme learning machine with incremental output structure,” *Neurocomputing*, vol. 128, pp. 224–231, 2014.

[11] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: theory and applications,” *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[12] W. Lu, Y. Li, Y. Cheng, D. Meng, B. Liang, and P. Zhou, “Early fault detection approach with deep architectures,” *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 7, pp. 1679–1689, 2018.

[13] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello, N. Zerhouni, and C. Varnier, “Pronostia: An experimental platform for bearings accelerated degradation tests.” in *IEEE International Conference on Prognostics and Health Management, PHM’12*. IEEE Catalog Number: CPF12PHM-CDR, 2012, pp. 1–8.

[14] J. Liu, S. Ji, and J. Ye, “Multi-task feature learning via efficient l 2, 1-norm minimization,” in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 339–348.