Article:
Liu, Z.-H., Lu, B.-L., Wei, H.-L. orcid.org/0000-0002-4704-7346 et al. (2 more authors) (2019) Fault diagnosis for electromechanical drivetrains using a joint distribution optimal deep domain adaptation approach. IEEE Sensors Journal. ISSN 1530-437X

https://doi.org/10.1109/jsen.2019.2939360

Paper accepted for publication by IEEE Sensors Journal (30-08-2019)

Fault Diagnosis for Electromechanical Drivetrains Using a Joint Distribution Optimal Deep Domain Adaptation Approach

Zhao-Hua Liu, Member, IEEE, Bi-Liang Lu, Hua-Liang Wei, Xiao-Hua Li, and Lei Chen

Abstract: Robust and reliable drivetrain is important for preventing electromechanical (e.g., wind turbine) downtime. In recent years, advanced machine learning (ML) techniques including deep learning have been introduced to improve fault diagnosis performance for electromechanical systems. However, electromechanical systems (e.g., wind turbine) operate in varying working conditions, meaning that the distribution of the test data (in the target domain) is different from the training data used for model training, and the diagnosis performance of an ML method may become downgraded for practical applications. This paper proposes a joint distribution optimal deep domain adaptation approach (called JDDA) based auto-encoder deep classifier for fault diagnosis of electromechanical drivetrains under the varying working conditions. First, the representative features are extracted by the deep auto-encoder. Then, the joint distribution adaptation is used to implement the domain adaptation, so the classifier trained with the source domain features can be used to classify the target domain data. Lastly, the classification performance of the proposed JDDA is tested using two test-rig datasets, compared with three traditional machine learning methods and two domain adaptation approaches. Experimental results show that the JDDA can achieve better performance compared with the reference machine learning, deep learning and domain adaptation approaches.

Index Terms—fault diagnosis, electromechanical drivetrain, deep neural network, deep learning, domain adaptation (DA), joint distribution optimal, auto-encoder(AE), machine learning, artificial intelligence, bearing, gearboxes, wind turbine, varying working conditions.

I. INTRODUCTION[1]

Electromechanical systems (e.g., wind turbine) play an important role in industrial systems [1], [2]. However, electromechanical drivetrains are typically exposed to invariable and harsh environments, and usually suffer from high failure rate. It is prone to failure due to the some severe operating environment and the wide range of load fluctuations. Drivetrain failures can cause serious damage to the whole equipment, so it is necessary to discover potential faults in drivetrain system as early as possible. Usually, the failures of bearing, gearboxes, and other drivetrain components can usually result in long downtime, thus can cause considerable drivetrain maintenance costs [3]. Therefore, fault diagnosis for bearing and gearboxes components in drivetrain is one of the most important parts in the condition monitoring systems. In fact, in order to monitor the bearing and gearboxes conditions, many useful fault diagnosis methods were proposed, such as wavelet transforms [4], time-frequency manifold [5], and Morphological Hilbert-Huang (MH) technique [6]. The implementation of the fault diagnosis process using these methods usually need expert manual intervention. However, methods with expert manual intervention usually cannot provide easily-understood diagnosis results, therefore, a more convenient algorithm is necessary to solve fault diagnosis problems.

In recent years, machine learning and deep learning technologies have attracted the attention of many researchers to deal with the fault classification problems [7]. Commonly used machine learning (ML) methods include logistic regression (LR) [8], Naive Bayes classifier [9], SVM [10] and neural networks [11]. Although these ML methods can work well for most fault diagnosis problems without expert manual intervention as required by traditional approaches, they need sufficient prior knowledge and large labeled training samples. However, it is difficult to collect sufficient labeled data and then train a reliable diagnosis model in most engineering scenarios.

There is a similar situation in the deep learning networks (DLN) [12], such as recurrent neural networks (RNN) [13], convolutional neural network (CNN) [14] and deep brief networks (DBN) [15]. More specifically, Sun et al. [16] proposed an intelligent bearing fault diagnosis method, combining compressed data acquisition and deep learning, and including a sparse auto-encoder (SAE) as the DLN infrastructure. Of course, due to the structural advantages of DLN, these methods not only automatically extract the features, but also the extracted features can represent the original data well. So the classifier trained with those features would normally have an outstanding performance. However, the good performance of DLN is based on a hypothetical condition – the training data and test data have the same distribution.

Inevitably, electromechanical systems usually work in varying conditions due to the changeable working conditions, environmental noise and product quality etc., so data sets from a same process may have different distributions in practical application. Moreover, the lack of labeled training

[1]Manuscript received January 24, 2019; revised April 02, 2019, and July 03, 2019; accepted August 30, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61972443, Grant 61573299, and Grant 61503134, and the Hunan Provincial Hu-Xiang Young Talents Project of China under Grant 2018RS3095, Hunan Provincial Natural Science Foundation of China under Grant 2018JJ2134.

Z.-H. Liu, B.-L Lu, X.-H Li and L.Chen are with the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, 411201, China (e-mail:zhaohualiu2009@hotmail.com, 1197393632@qq.com, lixiaohua_0227@163.com, chenlei@hnust.edu.cn).

H.-L. Wei is with the Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield S1 3JD, U.K(w.hualiang@sheffield.ac.uk).

data makes it necessary to use historical labeled data (source domain) to achieve correct label prediction of new data (target domain), and such knowledge transfer can cause a distribution discrepancy between training data (source domain) and testing samples (target domain), which leads to significant diagnosis performance deterioration. In short, it is a challenging task for existing ML and DLN based fault diagnosis methods: How to overcome the data disparity issue (e.g. between training and test data) to achieve good performance for fault diagnosis? In other words, how to achieve a good diagnosis model with insufficient labeled training data for engineering problems?

Recently, a useful method based on domain adaptation (DA) [17] has been introduced into the ML and DLN. DA strategies can be roughly summarized into two categories: instance reweighting and feature extraction [18], [19]—the former re-adjusts the training set on the basis of the common knowledge contained in the test set, and then further analyzes the reweighted training set, while the latter aims to detects a shared subspace and draws the distance between training data (source domain) and testing samples (target domain). The DA approach considered in this study belongs to the latter one. In fact the training data space and their distribution used for building the fault diagnosis model is defined as the source domain, the test data space and their distribution for model application is defined as the target domain, and the problem that we are eager to solve is the cross-domain learning problem[17]-[19].

Due to the existence of cross-domain learning problem, the performance of the ML and DLN based classifiers may not be satisfactory, meaning that the classifier trained with the source domain may not work well for classification in the target domain. In many real applications, it is very expensive or difficult, if not possible; to collect test data that has the same distribution as the training set. To overcome such an issue, Lu et al. [20] proposed a novel deep model in which the DLN was combined with domain adaptation. Wen et al. [21] proposed a new deep transfer learning approach based on Sparse Auto-Encoder (SAE) in which the DLN utilizes the maximum mean discrepancy (MMD) measure [22] to minimize the distance between two distributions, so the classification of the target domain can make use of the common knowledge in the source domain. However, these two methods just optimize the difference of the marginal distribution, in other words, they only use a part of distribution information based on the original data.

In this paper, we put forward a different model from previous studies in which the advantage of DLN can be further exploited, and the DA [23] technique can be used to reveal and take advantage of data distribution information. Combining the DA strategies and the DLN, a novel approach using joint distribution adaptation for a deep learning model (JDDA), in which representative features of source domain and target domain are extracted by the DLN, and then, by making use DA, the distance between the two domains is narrowed, so the model trained in the source domain can be directly applied to the classification of the target domain through the classification hyper-plane. As far as we know, this is a novel work in the literature on such a method based on DA to solve fault diagnosis problems with large complicated fault data. The main contributions of the paper are summarized as follows:

1) A novel deep domain adaptation learning architecture based on DA techniques is proposed for fault diagnosis under variable working conditions. The proposed JDDA integrates deep learning, DA and machine learning in one model, where representative features can be easily extracted by the DLN. Afterwards, the representative features distance of source domain and target domain is decreased through the DA algorithm. Therefore, a classifier trained with the representative features of source domain can be used for fault prediction of the target domain.

2) A domain-adaptation fault diagnosis algorithm for electromechanical systems (wind turbine) based on joint distribution optimization is proposed and its performance is tested on actual device data. Besides, the joint distribution adaptation algorithm has been improved in the terms of labeled samples acquisition.

3) Empirical analysis is performed on the hyper-parameters of the domain-adaptation algorithm, aiming to facilitate the determination of important parameters of DA.

II. RELATED WORK

This section describes the related work on fault diagnosis of electromechanical drivetrains and DA.

A. Fault Diagnosis of electromechanical drivetrains

The failures of bearing, gearboxes, and other drivetrain components can usually result in long downtime, thus can cause considerable drivetrain maintenance costs. Similar approaches have been reported in the literature, for example, Zheng et al. [24] proposed a fault diagnosis method based on support vector machine for a rolling element bearing system fault detection; Li et al. [25] proposed a fault diagnosis method for motor rolling bearing using a neural network classifier. However, these methods need to manually determine a representative feature of the original data, but it often is very difficult or impossible to properly define the most representative feature when data is big. LeCun, Bengio and Hinton [12] proposed the concept of DLN, introducing the pre-training skill and minimizing the network's reconstruction error, so it can get a good representation of the original data automatically. Recently, some fault diagnosis methods [26] established on deep learning technology has been proposed. Comparing with traditional machine learning methods, the classification accuracy of these deep learning approaches has been significantly improved.

B. Domain Adaptation

Domain adaptation (DA), as a transfer learning method, utilizes a different but related source domain to solve the problem of the target domain. It is inspired by the idea that people can apply past experiences to new things. In the early stages, most of the domain adaptation strategies are coupled with machine learning [27]-[30]. But gradually DA has been integrated with deep learning in many applications, and an excellent achievement has been achieved in the field of computer vision [31], [32].

Broadly speaking, there are two types of feature extraction methods with DA: transfer subspace learning [33]-[35] and

transfer classifier induction [36]-[38], the DA is the part of the former one, and the DA has several ways to get it.

1) Marginal distribution adaptation, which aims to make the distance between the source domain and the target domain as close as possible, and a way to get closer is to minimize the predefined distance, e.g., maximum mean discrepancy (MMD), Kullback–Leibler divergence or Bregman divergence.

2) Conditional distribution adaptation, which estimates the effect of DA by shortening the distance between conditional distributions, and the detailed information can be found in [19].

3) Joint distribution adaptation (JDA), which combines the two methods mentioned above [39], [40]. Normally, the smaller the distance between two domains, the more robust of the JDA embedded models. More information about JDA is given in Section III-2.

Deep learning with DA opens a new door for the problem of electromechanical device fault diagnosis. The deep learning method usually needs enough training data, so it would be challenging and difficult for fault diagnosis using machine learning methods if there is only a small data set. But now the issue can be alleviated by means of DA. The DA can use similar but not identical source domain to solve classification problems in target domain. Pan et al. [28] proposed a Transfer Component Analysis (TCA) technique to map features (data) in two domains (source and target domains) to a Reproducing Kernel Hilbert Space (RKHS) using Maximum Mean Discrepancy (MMD). Long et al. [27] devised a transfer learning framework, called Adaptation Regularization Based Transfer Learning Framework (ARTL), by incorporating MMD into the machine learning. The major difference between our proposed model and these methods discussed above is that a deep learning scheme (i.e. deep learning network) is introduced in our model and used to extract features automatically, the transfer learning scheme (i.e. joint distribution adaptation) is embedded in the DLN, and in this way a robust and high-performance model can be acquired.

III. THE PROPOSED JDDA FOR FAULT DIAGNOSIS OF ELECTROMECHANICAL DRIVETRAINS

This section introduces a joint distribution optimal deep domain adaptation approach (JDDA), and the JDDA framework is showed in Fig.1.

1. Problem Definition

Definition 1 (Domain). Given a sample set X and the feature space χ, where $X = \{x_1, x_2, x_3, ..., x_n\} \in \chi$, let P(X) be a marginal probability distribution. A domain, designated by $\mathcal{D}$, is a set that consists of the feature space χ and a marginal probability distribution P(X). Note that two domains $\mathcal{D}_S$ and $\mathcal{D}_T$ are said to be different if their feature space χ or the marginal distribution P(X) are different, or simply, $\chi_S \neq \chi_T$ or $P_S(X) \neq P_T(X)$. In this paper, the source domain is defined as the training data space and their distribution which used for building the fault diagnosis model. Correspondingly, the target domain is composed of the test data space and the associated distribution, where the diagnosis model is applied to.

Definition 2 (Task). A task $\mathcal{T}$ is made up of the label spaces Y and the conditional probability distribution $P(Y|X)$. Consider two different tasks $\mathcal{T}_S$ and $\mathcal{T}_T$, which possess different characteristics in two different domains $\mathcal{D}_S$ and $\mathcal{D}_T$. Two tasks $\mathcal{T}_S$ and $\mathcal{T}_T$ are said to be different, if the associated label spaces Y or the conditional distributions $P(Y|X)$ are different.

In this paper, to bridge the gap between the source domain and the target domain, it is assumed that the label spaces are the same, but the conditional distributions $P(Y|X)$ are different. In other words, the types of failures between the source domain and the target domain are overlapping, but the probability of failure occurrence is different in the changeable engineering environments. In short, it is assumed that $\mathcal{Y}_S = \mathcal{Y}_T$, $P(Y_S|X_S) \neq P(Y_t|X_t)$.

Definition 3 (Motivation). Given the source domain containing n samples, that is, $\mathcal{D}_S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ (n is large enough), and a target domain containing m samples, that is, $\mathcal{D}_T = \{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2}), x_{n+3}, ..., x_{n+m}\}$ (m is much smaller than n, i.e., m<<n) . The objective is to find a conversion function F , such that the labeled data of two domains have the following property: $P_S(F(X)) = P_T(F(X))$, $P(Y_S|F(X_S)) = P(Y_T|F(X_T))$. In this paper, the former corresponds to the maximum mean discrepancy, while the latter uses the conditional distribution adaptation to match the difference. Therefore, a classification hyper-plane co-trained by the source domain $\mathcal{D}_S$ and the smaller labeled data set of the target domain $\mathcal{D}_T$ can be used to predict the unlabeled samples in $\mathcal{D}_T$.

2. Deep Learning Network

In this paper, taking into account the good feature extraction performance of auto-encoder (AE), it is introduced for feature extraction, and it is the basis of the proposed JDDA. The key idea of AE is that the reconstruction of the original data in one layer [21], to achieve such a goal, AE is designed to consist of two parts: coding part and decoding part. On the one hand, the function of the coding part is to extract feature of the original data, and the procedure that extracts feature of the original data X can be defined as $h = f(Z_C)$, $Z_C = \theta$. On the other hand, the decoding part is to restore the extracted features back to a set of data that possesses the same latitude with the original data, the process of decoding part can be defined as $\hat{X} = f(Z_d)$, $Z_d = \theta'(h)$,and $f(.) = 1/1 + \exp(.)$, which are parameterized respectively as follows: $\theta = \omega_C X + b_C$, $\theta' = \omega_d h + b_d$, where $\theta \triangleq \{\omega_C, b_C\}$ and $\theta' \triangleq \{\omega_d, b_d\}$ are the weight and bias matrixes of the encoder

and decoder, respectively, so the basic structural loss function is defined as:

$$\min_{\omega,b} \frac{1}{2m}\sum_{i=1}^{m}\left\|(\hat{X}_i - X_i)\right\|_F^2 \qquad (1)$$

$$\text{s.t.}\quad Z_C = \theta,\ h = f(Z_C),\ Z_d = \theta',\ \hat{X} = f(Z_d).$$

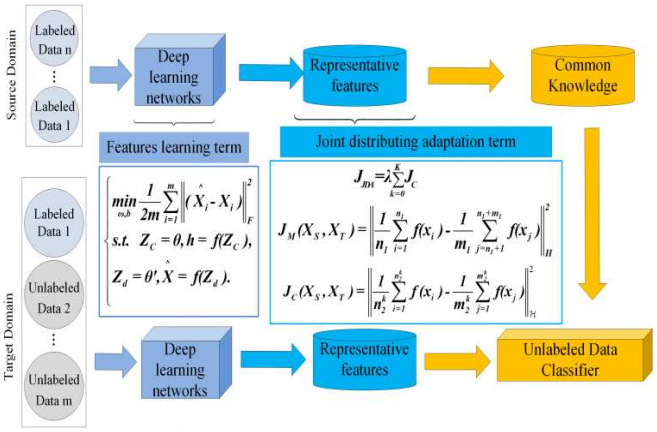


Fig. 1. The proposed fault diagnosis framework based on joint distribution adaptation.

The symbol $\|\ \|_F$ represents a symbolic notation of the Frobenius norm, $\hat{X}_i$ and X_i individually represent the single sample from the feature of decoding part and the sample set X.

3. Joint Distribution Optimal Deep Domain Adaptation Architecture for Fault Diagnosis

The JDA is designed to find the best path to minimize the distance between two domains, integrating marginal distribution and conditional distribution is performed through the JDA term function in which the distance between the marginal distributions and the distance between the conditional distributions could be shorten in $\mathcal{H}$, and this can be formulated as follows:

$$\min_{F} \left\|\mathbb{E}_{P(X_S,Y_S)}[F(X_S,Y_S)] - \mathbb{E}_{P(X_T,Y_T)}[F(X_T,Y_T)\right\|^2$$
$$\approx \left\|\mathbb{E}_{P(X_S)}[F(X_S)] - \mathbb{E}_{P(X_T)}[F(X_T)]\right\|^2 \qquad (2)$$
$$+\left\|\mathbb{E}_{P(Y_S|X_S)}[Y_S|F(X_S)] - \mathbb{E}_{P(Y_T|X_T)}[Y_T|F(X_T)]\right\|^2$$

where $\mathbb{E}$ represents mathematical expectation.

JDA utilizes the convenience of the Maximum Mean Discrepancy (MMD), the discrepancy of two distributions is aligned by the MMD term in which the sample mean from the two domains is subtracted in the reproducing kernel Hilbert space (RKHS), and the calculation form can be written as:

$$\mathrm{MMD}(X_S, X_T) = \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\phi(x_i) - \frac{1}{m_1}\sum_{j=n_1+1}^{n_1+m_1}\phi(x_j)\right\|_{\mathcal{H}} \qquad (3)$$

where n_1 is the number of samples in the source domain and m_1 is the number of samples in the target domain. $\phi: X \to \mathcal{H}$ is the mapping function of the original feature space mapped to RKHS. In this study we fine-tune the MMD form to be:

$$J_M(X_S, X_T) = \mathrm{MMD}^2(X_S, X_T) = \left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\phi(x_i) - \frac{1}{m_1}\sum_{j=n_1+1}^{n_1+m_1}\phi(x_j)\right\|_{\mathcal{H}}^2 \qquad (4)$$
$$= \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}k(x_i^s, x_j^s) + \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_1}k(x_i^t, x_j^t) - \frac{2}{n_1 m_1}\sum_{i=1}^{n_1}\sum_{j=1}^{m_1}k(x_i^s, x_j^t)$$

where subscript M indicates that this objective function to be optimized is the marginal distribution, and $k(x,y) = \exp(\|x-y\|^2/2\sigma^2)$. By minimizing (4), the marginal distributions between domains are drawn close， and this only takes advantage of the differences in the marginal distributions. In order to get a high classification accuracy model, the discrepancy between the conditional distributions

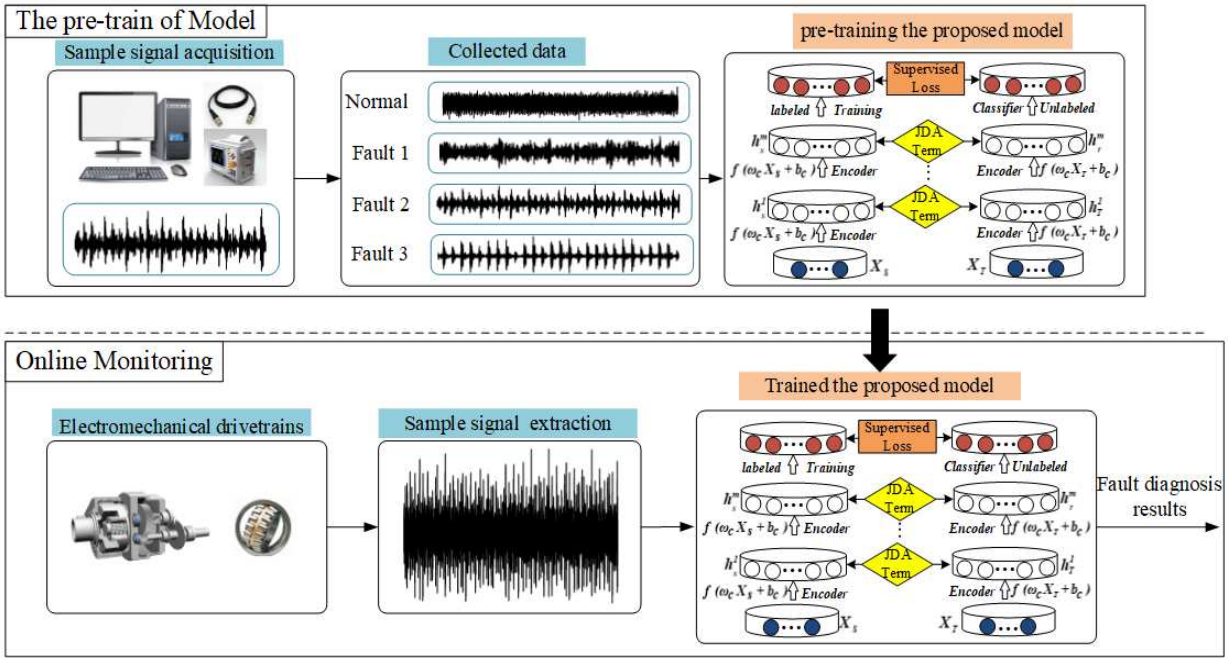


Fig. 2. The fault diagnosis algorithm for electromechanical drivetrain based on the JDDA.

$P(Y_S|X_S) \neq P(Y_t|X_t)$ of the two domains $\mathcal{D}_S$ and $\mathcal{D}_T$ is another optimization objective. Long et al. [19] proposed the transfer feature learning with joint distribution adaptation (JDA) in which the representation feature is designed to optimize this discrepancy by using pseudo target labels, and the pseudo target labels are predicted by the supervised classification hyper-plane (SVM) trained on the source domain labeled data. So the difference between the two distributions of features in $\mathcal{D}_S$ and $\mathcal{D}_T$ can be reduced as much as possible under knowing the pseudo target labels $P(Y_S|F(X_S)) \approx P(Y_t|F(X_t))$. In particular, we can calculate the distance of the average for normal type of samples by applying the real label directly by assuming that $P(Y_S|F(X_S)) = P(Y_t|F(X_t))$, so that a high classification accuracy model can be achieved. The specific details are computed as follows:

$$\begin{aligned} J_C(X_S, X_T) &= \left\| \frac{1}{n_2^k} \sum_{i=1}^{n_2^k} \phi(x_i) - \frac{1}{m_2^k} \sum_{j=1}^{m_2^k} \phi(x_j) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n_2^k} \sum_{i=1}^{n_2^k} \sum_{j=1}^{n_2^k} k(x_i^s, x_j^s) + \frac{1}{m_2^k} \sum_{i=1}^{m_2^k} \sum_{j=1}^{m_2^k} k(x_i^t, x_j^t) - \frac{2}{n_2^k m_2^k} \sum_{i=1}^{n_2^k} \sum_{j=1}^{m_2^k} k(x_i^s, x_j^t) \end{aligned} \tag{5}$$

where $\phi: X \to \mathcal{H}$ is the mapping function of the original feature space mapped to RKHS; and $k(x, y) = \exp(\|x - y\|^2 / 2\sigma^2)$. $n_2^k = \{ x_i \in \mathcal{D}_S \wedge y(x_i) = k \}$ is the set of samples belonging to class k in the target domain that fall into the common subset, and $m_2^k = \{ x_j \in \mathcal{D}_t \wedge y(x_j) = k \}$ is the set of samples belonging to class k in the target domain that fall into the common subset.

By minimizing (5), the conditional distributions between domains are drawn close. In fact, combining (4) and (5), it leads to the following JDA optimization problem as follows:

$$\begin{aligned} J_{JDA} &= \lambda J_M + \mu \sum_{k=1}^{K} J_C \\ &= \lambda \left(\sum_{k=0}^{K} J_C \right) \end{aligned} \tag{6}$$

Note that the MMD can be viewed as a special case of JDA with $k = 0$, we can simultaneously adapt both the marginal distributions and conditional distributions between domains.

4. The Proposed JDDA for Drivetrain Fault Diagnosis

This section proposes a fault diagnosis framework, which is shown in Fig.2. Taking into account the existing computing power, the JDDA is designed to include only one hidden layer in this study, and the main data processing is as below. Data from the source and target domains go through the AE network to get associate features, so all data is pulled into a same feature space (RKHS), then, the key step of the JDDA—JDA term is used as a tool to narrow the features distance of both domains in the RKHS, the SVM classification hyper-plane trained by labeled data of the source domain can be applied to the classification of the unlabeled data of the target domain. According to the structure of the JDDA, the cost function of the JDDA can be expressed as in equation (7).

$$\begin{aligned} &\min_{\omega, b} \ \frac{1}{2m} \sum_{i=1}^{m} \left\| (\hat{X}_i - X_i) \right\|_F^2 + \lambda \left(\sum_{k=0}^{K} J_C \right) \\ &\text{s.t.} \ \ Z_C = \theta, \ h = f(Z_C), \ Z_d = \theta', \ \hat{X} = f(Z_d) \end{aligned} \tag{7}$$

where λ is the trade-off parameter of the JDA term, the MMD can be viewed as a special case of JDA with $k = 0$. There are a total of K samples in the common subset that belongs to both of the two domains.

This cost function contains the two parts mentioned in III-2, namely, deep learning network and joint distribution based deep domain adaptation architecture. The former comes from the direct loss of reconstruction error, and another is used to reduce the distance between two different domains in the same feature space. Another implementation aspect of the JDDA is the learning algorithm. The decoding layer is not shown in Fig. 2, because we use the features in the previous layer directly as the input to the next layer. In fact, the decoding layer still exists. The process of the learning algorithm is summarized below.

Learning Algorithm: A Joint Distribution Optimal Deep Domain Adaptation Method for Fault Diagnosis

Begin:

Step 1: Randomly initialize the parameter of AE network

Building a basic AE network structure, and the corresponding parameters ω_C , ω_d , b_C , b_d needs to initialize by following the structural loss function in (1) with the labeled data.

Step 2: Pre-train the AE network

The raw data is made up of the unlabeled data from $\mathcal{D}_S$ and $\mathcal{D}_T$, and it will be used to pre-train the JDDA, the iteration process of solving the parameters can be written as:

$$\omega_C = \omega_C - \alpha_1 \frac{\partial J_{cost}}{\partial \omega_C} \tag{8}$$

$$\omega_d = \omega_d - \alpha_2 \frac{\partial J_{cost}}{\partial \omega_d} \tag{9}$$

$$b_C = b_C - \alpha_3 \frac{\partial J_{cost}}{\partial b_C} \tag{10}$$

$$b_d = b_d - \alpha_4 \frac{\partial J_{cost}}{\partial b_d} \tag{11}$$

where $\alpha_i \ (i = 1, 2, 3, 4)$ is the learning rate.

Step 3: Establishing the JDDA network

Those parameters ω_C , ω_d , b_C , b_d are used to build the JDDA based on the AE structure. Meanwhile, $J_M + J_C$ has been inserted into the loss of AE, and the final loss function (6) is optimized by retraining the labeled data from the two domain.

Step 4: Training joint distribution adaptation classifier

At this step, the classification hyper-plane of SVM is only determined by features $\phi(x)_S$ from $\mathcal{D}_S$. Due to the role of JDA term, the $\mathcal{D}_T$ features $\phi(x)_T$ can be directly separated by trained classifier.

Step 5: Output classification results

The classification results of the $\mathcal{D}_T$ features $\phi(x)_S$ are generated from the SVM.

End

Ⅳ. EXPERIMENT TEST

We consider two test-rig systems, which are shown in Fig. 3. Two simulation datasets were used to test the performance of the proposed method. The two experiment datasets were acquired in two places: ball bearing test data from Case Western Reserve University Bearing Data Center (CWRU) [41], and a gearbox fault data from the prognostics and health management society (PHM Society) [42].

(a) Bearing test rig [41] (b) Gearbox test rig [42]

Fig.3. Experiment setup for drivetrain fault diagnosis

A. Data Description

1) Ball Bearing Test Data: experiments data was from the single-point drive end of the bearing in which the accelerometer was used to get the normal and fault data, and the fault data contain defects in the inner race (IN), the outer race (OU) and the ball (BA). Of course, the mentioned three kinds of fault data have four fault diameters (0.007, 0.014, 0.021, and 0.028), respectively. In addition, the motor load was set in four stages (0, 1, 2 and 3hp), and the sampling frequency was 12 kHz. In this part, we used data selected from the four to create six DA condition (0-1hp, 0-2hp, 0-3hp, 1-2hp, 1-3hp, 2-3hp) to verify the performance of the model. Taking 0-3hp as an example, the form of the problem definition in section III-1 can be specifically designed as:

a) **Source Domain:** The source domain contains normal and defect data from a 0hp motor load, in this paper, the fault diameters are selected as 0.007 and 0.014.So

$$\Omega_S = \left\{\text{normal}^0, \text{IN}^0_{0.007,0.014}, \text{OU}^0_{0.007,0.014}, \text{BA}^0_{0.007,0.014}\right\}.$$

b) **Target Domain:** Similar to the source domain, the target domain contains a lot of normal and defect data from 3hp motor loads, but a different place is that there is only the normal data for labeled samples, so the available target domain $\Omega_T = \left\{\text{normal}^3\right\}$.

c) **Task:** The task is categorizing the unlabeled data in the target domain into

$$\left\{\text{normal}^3, \text{IN}^3_{0.007,0.014}, \text{OU}^3_{0.007,0.014}, \text{BA}^3_{0.007,0.014}\right\}.$$

2) Gearbox fault dataset: This is a compound fault data. In order to maximize the use of this data, the tachometer information of helical is chosen with the accelerometers mounted on both the input and output shaft retaining plates, three kinds of data– normal gear, chipped gear (CG) and broken gear (BG) are included in tachometer information, and this fault data is collected under both high and low load conditions, in addition, five different (30Hz, 35Hz, 40Hz, 45Hz, 50Hz) shaft speed have also been set. To the convenience of experimental data recording, several abbreviations are used to represent specific data, for example, 45L, meaning that the load condition for this data is low, and its shaft speed is 45Hz.For the sake of creating a TL situation, the problem definition in the part III-1 can be specifically designed as:

a) **Source Domain:** In gearbox fault dataset, the source domain is composed of the three kinds of data of 45L and the normal data of the target area. $\Omega_S = \left\{\text{normal}, \text{CG}, \text{BG}\right\}$.

b) **Target Domain:** As same as source domain, the source domain contains five different shaft speed data with high load condition (30H, 35H, 40H, 45H, 50H), but the only data that can be labeled is normal data. $\Omega_T = \left\{\text{normal}\right\}$.

c) **Task:** The task of this part is deal with the unlabeled data in the target domain, this unlabeled data which can be classified into $\left\{\text{normal}, \text{CG}, \text{BG}\right\}$.

3) Data preprocessing: the Ball Bearing Test Data needs to be pre-processed. Firstly, a total of 1200 data points (samples) were chosen from the two domains, with 80% overlap. Then, due to the presence of noise, Fast Fourier Transform (FFT) is used to reduce effect of noise on model training. Next, the value of the data after FFT is magnified 10 times, because it becomes too small. For Gearbox Fault Dataset, in order to remove noise interference, we assign a value of 0 to the value less than zero in the original data.

B. Experimental Results

1) Methods Used

For comparison purpose, the JDDA, along with several state-of-the-art machine learning and TL methods, are used: including SVM [10], [43] Logistic Regression (LR) [8], Back-propagation neural network (BP) [11], Transfer Component Analysis (TCA) [28], deep model based domain adaptation for fault diagnosis (DAFD) [20], and without the JDA term of the JDDA that we proposed (JDDA-R). The first 3 approaches are traditional machine learning methods, the fourth and fifth approaches are DA methods which have been successively applied to fault diagnosis. The last one is a comparison method of JDDA and it is also a deep learning method (AE).

2) Experimental Details

For the first three methods, the source domain data is used to train the associated machine learning models. After the training is completed, unlabeled data from the target domain is used to test the classification accuracy of the model, but the training data includes not only label data from the source area but also normal label data from the target domain. Data normalization is performed for the fourth method.

TABLE Ⅰ
CLASSIFICATION ACCURACY FOR THE DRIVETRAIN BEARING DATA

Without distribution adaptation technology							
Trial number	1	2	3	4	5	6	
Methods	0-1HP	0-2HP	0-3HP	1-2HP	1-3HP	2-3HP	avg.
LR	88.8%	74.9%	79.4%	75.0%	72.5%	77.8%	78.1%
svm	93.9%	87.7%	83.1%	74.7%	77.9%	97.8%	85.9%
BP	74.8%	72.1%	73.7%	65.7%	89.2%	84.6%	76.7%
JDDA-R	78.6%	74.1%	80.1%	74.9%	80.2%	75.6%	77.3%
With distribution adaptation technology							
TCA	97.8%	75.0%	86.9%	80.1%	**99.7%**	80.4%	86.7%
DAFD	96.7%	92.3%	93.6%	86.4%	93.2%	92.5%	92.5%
JDDA	**99.6%**	**98.6%**	**99.6%**	**97.4%**	97.8%	**100.0%**	**98.8%**

In terms of parameter adjustment, an empirical search approach is used to find the optimal parameters for the six comparative methods. For SVM, the LIBSVM package is used for classification [43], the kernel function is set to Gaussian kernel, and the value of the trade-off parameter is set to 1.5. For LR, the trade-off parameter is selected from $\{0.002, 0.02, 0.2, 2, 20\}$. For BP, the number of hidden layer is set to 2, the number of hidden neurons of each layer is 1000, and the learning rate is set to 0.1. For JDDA-R, it means that the parameter λ is set to 0, so only auto-encoder is used for feature extraction without domain adaptation term. For TCA, the kernel type is selected as Radical Basis Function (RBF), and the optimized subspaces for the processed features can choose from $8,16,32,64,128$ .For DAFD, as it uses of a back propagation algorithm, the reconstruction error is gradually reduced, and three main adjustable parameters are: $\lambda = 1, \mu = 1000, \text{and } \sigma = 0.001$, more details can be found in [20].

For all the DA methods considered in this study, the method designed for the SVM method can be used to find the associated optimal model parameters. For JDDA, the number of hidden layer is set to one and the numbers of hidden units are set to 1000. For the convenience of the experiments, the value of the model's regularization parameter is set to two.

At last, the classification accuracy of each method is defined as:

$$\text{accuracy}(C\%) = \frac{\text{label}(x) = k \cap \text{predict}(x) = k}{x_n} \qquad (12)$$

where x_n is the total number of test samples, and k is the true label value that a classifier correctly identified.

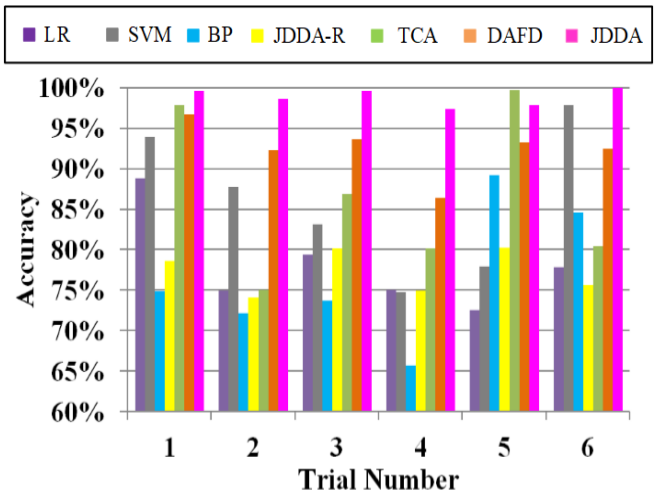


Fig.4. Fault diagnosis accuracy of each method on drivetrain bearing data.

3) Results of Bearing Case Study

As shown in TABLE Ⅰ and Fig. 4, for the methods without distribution adaptation, it is generally lower than the method that with distribution adaptation. In the trial number 3, the classification accuracy of JDDA-R is 16.5% lower than the novel JDDA. For the methods with distribution adaptation, the classification accuracy is also lower than that of JDDA, for example, the accuracy of DAFD is 6% lower than that of JDDA. It is worth mentioning that the classification accuracy of BP is the lowest among all methods. This may be explained from two aspects. First, we can only empirically find the best hidden layer and learning rate for BP, it uses a semi-supervised approach, and the experiment data contains part of the data from the target area, which affects the classification accuracy of the BP network.

TABLE Ⅱ
CLASSIFICATION ACCURACY FOR THE GEARBOX DATASET

Methods	45L-30H	45L-35H	45L-40H	45L-45H	45L-50H	avg.
Without distribution adaptation technology						
LR	50.0%	48.5%	49.8%	52.3%	51.6%	50.4%
svm	50.0%	50.0%	49.5%	50.4%	50.4%	50.1%
BP	46.9%	33.9%	38.1%	28.9%	34.0%	36.4%
JDDA-R	48.5%	46.0%	48.2%	62.3%	53.6%	51.7%
With distribution adaptation technology						
TCA	50.8%	52.5%	60.3%	60.8%	60.1%	56.9%
DAFD	53.8%	54.6%	57.4%	72.5%	65.1%	60.7%
JDDA	**57.9%**	**61.6%**	**69.5%**	**80.3%**	**70.4%**	**67.9%**

4) Results of Gearbox Case Study

The results of the five different TL circumstances are listed in Table Ⅲ, where 45L-30H means that the $\mathcal{D}_S$ is the data from 45L, and the $\mathcal{D}_T$ is the data from 30H. Although the average accuracy of the model in all methods is high, the accuracy of the model is even lower than the supervised learning algorithm in some papers. The main reasons are as follows, unlike this supervised learning algorithm, the experimental data of the JDDA is performed under the condition that the train data and the test data are subject to different distributions.

5) Results Summary

As we expect, the excellent results of novel model has achieved when it is used to deal with the classification problem in the test of two actual data, the proposed method can indeed improve the classification accuracy of TL situation.

C Results Analysis

In order to further explore why the performance of JDDA is good, the t-distributed stochastic neighbor embedding (t-SNE), as a dimension reduction visualization method, is used to reduce the dimension of features involved in both JDDA-R and JDDA. The reason for choosing t-SNE is that high dimensional data can be well visualized at low-dimensional space, as shown in Fig. 5(a)–(d), where the normal features of the two models are clearly observed and each fault feature is rendered in a two-dimensional map after dimensionality reduction. More details about the t-SNE application can be found in [44].

The visualization of the JDDA features is used for reference purpose. For example, in Fig. 5(a), the distance between two domains features of the JDDA(green and blue cross marker) is smaller than the distance between $\mathcal{D}_S$ and $\mathcal{D}_T$ features of the JDDA-R (green and blue point marker). This characteristic proves that our model can make the distance between $\mathcal{D}_S$ and $\mathcal{D}_T$ closer in the Reproducing Kernel Hilbert Space (RKHS), so a high-performance SVM classifier is available by training with the labeled features of the JDDA.

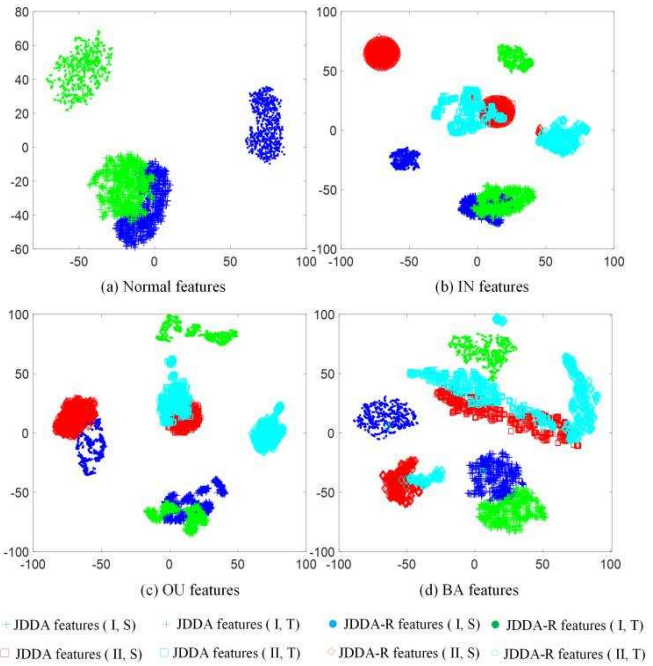


Fig.5. The features of JDDA and JDDA-R are displayed by t-sne in a reduced-dimensional dimension. The bracketed symbol S represents the source domain D_S and symbol T represents the target domain D_T, in particular, the number Ⅰ represents the fault diameter value is 0.007, the number Ⅱ means the fault diameter value is 0.014.

D Empirical Analysis of Parameters

In this section, the effect of the trade-off parameters of the JDA term on the accuracy of model classification for the CWRU Bearing Data is analyzed. As the mentioned above, the classification accuracy is a standard measure to evaluate a classifier's performance. Let λ be a trade-off parameter, we use $\log_{10}\lambda$ as the abscissa to show the effect of the change of λ on the classification accuracy. As shown in Fig.6, in phase 1, $\log_{10}\lambda \in [-2, 0.4]$. With the increasing of λ, the performance of the JDDA becomes better and better. In phase 2, $\log_{10}\lambda \in [0.4, 3.5]$. The JDDA maintains a good performance for the test data, which means that the JDDA has a robust classification effect. In phase 3, $\log_{10}\lambda \in [3.5, 4]$. The classification accuracy of JDDA drops rapidly.

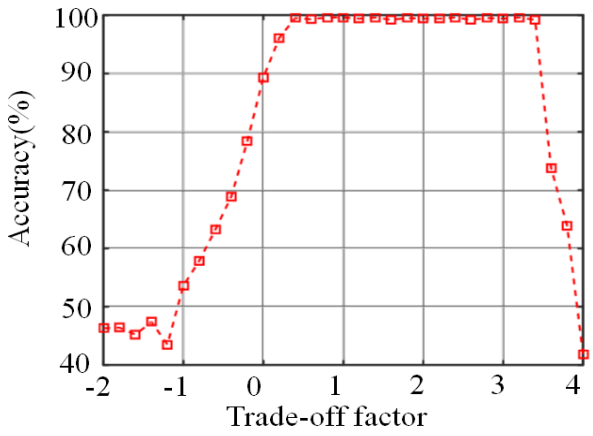


Fig. 6. The trade-off parameter (λ) influence for the JDDA.

V. CONCLUSION

A novel deep domain adaptation learning architecture, combined with deep learning model and joint distribution adaptation (JDA), is proposed for fault diagnosis of electromechanical drivetrains system. The performance of the proposed JDDA method is tested using simulation datasets for bearing and gearbox, and compared with other five state-of-the-art methods. The main contribution of this paper is that a novel method of domain adaptation has been explored in which the distance of the representative features of the source domain and target domain is reduced through the RKHS, and the JDDA can not only be applied to fault diagnosis of variable working conditions, but also to other fields. An explanation of the better performance of JDDA is presented using the t-SNE. The analysis of the impact of the trade-off parameter on the classification accuracy of the JDDA provides some useful information for further development and improvement of the JDDA.

In future work, we would consider the following two topics. Firstly, we would apply the method to more real scenario datasets to further test its performance, and then applied to real electromechanical drivetrains problem solution. So, it can reduce the downtime of electromechanical drivetrains (wind turbines, high-speed railway, etc.), save maintenance costs, increase power generation rate and economic benefits. Secondly, the distributed parallelism approach will be further explored and enhanced to improve the real-time performance.

We will carry out the proposed method to the reality wind farm in our future work.

REFERENCES

[1] R. Yan, R. Gao, and X. Chen, “Wavelets for fault diagnosis of rotary machines: a review with applications”, Signal Processing, Vol. 96, Part A, pp. 1-15, March 2014.

[2] Y. Lei, J. Lin, M. Zuo, Z. He, “Condition monitoring and fault diagnosis of planetary gearboxes: A review”, Measurement, Vol. 48, no.2, pp. 292-305, Feb. 2014.

[3] Z. Wang, Q. Zhang, J. Xiong, M. Xiao, G. Sun, and J. He, “Fault diagnosis of a rolling bearing using wavelet packet denoising and random forests,” IEEE Sensors Journal., vol. 17, no. 17, pp. 5581–5588, Sep. 2017

[4] R. Yan, R. X. Gao, and X. Chen, “Wavelets for fault diagnosis of rotary machines: A review with applications,” Signal Process., vol. 96,pp. 1–15, Mar. 2014.

[5] J. Wang and Q. He, “Wavelet packet envelope manifold for fault diagnosis of rolling element bearings,” IEEE Trans. Instrum. Meas.,vol. 65, no. 11, pp. 2515–2526, Nov. 2016.

[6] S. Osman and W. Wang, “A morphological hilbert-huang transform technique for bearing fault detection,” IEEE Trans.Instrum. Meas., vol. 65, no. 11, pp. 2646-2656, Nov. 2016.

[7]Z.W. Gao, C. Cecati and S. X. Ding,“A survey of fault diagnosis and fault-tolerant techniques—part ii: fault diagnosis with knowledge-based and hybrid/active approaches,”IEEE Trans. Ind. Electron., vol. 62, no. 6,pp. 3768-3774, Jun. 2015.

[8] M.K.Bodla,S.M. Malik, M.T. Rasheed, M.Numan, M.Z. Ali, J.B.Brima, “Logistic regression and feature extraction based fault diagnosis of main bearing of wind turbines,” in Proc. IEEE Ind. Electron., Appl. Conf., pp. 1628-1633, June. 2016.

[9]R.K. Sharma, V. Sugumaran, H. Kumar, M. Amarnath, “A comparative study of naive Bayes classifier and Bayes net classifier for fault diagnosis of roller bearing using sound signal,” Int. J. Decision Support Systems., vol. 1, No. 1, pp. 1-115, Jun. 2015.

[10]Y. Yang, D. Yu, and J. Cheng, “A fault diagnosis approach for roller bearing based on IMF envelope spectrum and SVM,” Measurement.,vol. 40, no. 9–10, pp. 943-950,Nov. 2007.

[11]Y. Shatnawi and M. Al-Khassaweneh,“Fault diagnosis in internal combustion engines using extension neural network,”IEEE Trans. Ind. Electron.,vol. 61, no. 3, pp. 1434-1443, March. 2014.

[12]Y. LeCun, Y. Bengio and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[13]T. D.Bruin, K.Verbert, R.Babuška, “Railway track circuit fault diagnosis using recurrent neural networks,”IEEE Trans. Neural Netw. Learn.Syst., vol.28,no.3,pp.523-533,March.2017.

[14] H. Jiang, F. Wang, H. Shao, H. Zhang, “Rolling bearing fault identification using multilayer deep learning convolutional neural network,” Journal of Vibroengineering., vol. 19, no.1, pp. 138-149, Feb. 2017.

[15] Z.Y. Chen, and W. Li, “Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network,” IEEE Trans. Instrum. Meas., vol. 66, no.7, pp. 1693-1702, July. 2017.

[16] J.D. Sun, C. Yan and J. Wen, "Intelligent bearing fault diagnosis method combining compressed data acquisition and deep learning," IEEE Trans. Instrum. Meas.,vol. 67, no. 1, pp. 185 – 195, Jan. 2018.

[17]S.J. Pan and Q. Yang,“A survey on transfer learning,”IEEE Trans. Knowl.Data Eng.,vol. 22.no, 10, pp. 1345-1359, Oct. 2010.

[18]W. Dai, Q. Yang, G.-R. Xue and Y. Yu, “Boosting for transfer learning,” in Proc. 24th ICML, Corvallis, OR, USA, 2007.

[19] M.S. Long,J.M. Wang,G.G. Ding,J.G. Sun andP.S. Yu, “Transfer feature learning with joint distribution adaptation,” in Proc.IEEE Int. Conf.Comput. Vis., pp. 2200-2207, Dec. 2013.

[20]W.N. Lu,B. Liang, Y. Cheng, D.S.Meng, J. Yang and T. Zhang,“Deepmodel based domain adaptation for fault diagnosis,”IEEE Trans. Ind. Electron.,vol. 64, no. 3, pp.2296 – 2305, March. 2017.

[21]L. Wen,L.Gaoand X.Y. Li,“A new deep transfer learning based on sparse auto-encoder for fault diagnosis.”IEEE Trans. Syst., Man, Cybern, Syst., doi: 10.1109/TSMC.2017.2754287.

[22]A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J.Smola, “A kernel method for the two-sample problem,” in Proc.NIPS, 2006.

[23]L. Shao , F. Zhu, and X. Li,“Transfer learning for visual categorization: a survey.”IEEE Trans. Neural Netw. Learn.Syst.,vol. 26, no.5, pp. 1019-1034, May. 2015.

[24] H. Zheng and L. Zhou, “Rolling element bearing fault diagnosis based on support vector machine,” in Proc.IEEE Int. Conf. Consumer Electron. Comm. Net., pp. 544-547, April 2012.

[25]B.Li, M.Y. Chow, Y. Tipsuwanand J.C. Hung, “Neural-network-based motor rolling bearing fault diagnosis,” IEEE Trans.Ind.Electron., vol.47, no.5, pp. 1060-1069, Oct. 2000.

[26]H. Miao, and D. He, “Deep learning based approach for bearing fault diagnosis,” IEEE Trans. Ind. Appl., vol.53, no.3, pp. 3057-3065, May-June. 2017.

[27] M.S. Long, J.M. Wang, G.G. Ding, S.J. Pan andP.S. Yu, “Adaptation regularization: a general framework for transfer learning,”IEEE Trans. Knowl. Data Eng., vol.26, no.5, pp. 1076-1089, May. 2014.

[28] S.J. Pan,I.W. Tsang, J.T.Kwok andQ. Yang, "Domain Adaptation via Transfer Component Analysis," IEEE Trans. Neural Netw.,vol. 22, no.2, pp.199-210, Feb. 2011.

[29] R. Zhang,H.Y. Tao,L.F. Wu,Y. Guan, "Transfer learning with neural networks for bearing fault diagnosis in changing working conditions," IEEE Access., vol. 5. pp. 14347-14357, Jun. 2017.

[30] L.X. Duan,I.W. Tsang and D. Xu, "Domain transfer multiple kernel learning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no.3, pp. 465-479, March. 2012.

[31]N. Dethlefs,"Domain transfer for deep natural language generation from abstract meaning representations,"IEEE Comput. Intell. Mag.,vol. 12, no.3, pp. 18-28, Aug. 2017.

[32]M.S. Long, Y. Cao, J M. Wang, M.I. Jordan,"Learning transferable features with deep adaptation networks,"in Proc.32nd Int. Conf. Mach. Learn.,pp. 97-105, July.2015.

[33]S. J. Pan, X. Ni, J.-T. Sun, Q. Yang and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proc.19th Int. Conf. WWW, Raleigh., NC, USA, 2010.

[34]X.X. Shi, Q. Liu,W. Fan andP.S. Yu,"Transfer across completely different feature spaces via spectral embedding," IEEE Trans. Knowl. Data Eng., vol. 25, no.4, pp. 906-918, April. 2013.

[35]Y. Xu, X.Z. Fang,J. Wu,X.L. Li andD. Zhang,"Discriminative transfer subspace learning via low-rank and sparse representation,"IEEE Trans. Image Process.,vol. 25, no.2, pp. 850-863, Feb. 2016.

[36]J.W. Tao, F. L. Chung and S. Wang,"On minimum distribution discrepancy support vector machine for domain adaptation,"PatternRecognLett.,vol. 45, no.11, pp. 3962-3984, Nov. 2012.

[37]L.X. Duan, I. W. Tsang and D. Xu,"Domain transfer multiple kernel learning,"IEEE Trans. Pattern Anal. Mach. Intell.,vol. 34, no.3, pp. 465-479, March. 2012.

[38]Q.H. Liu,X.J. Liao,H.L. Carin,J.R. Stack andL.Carin"Semisupervised multitask learning,"IEEE Trans. Pattern Anal. Mach. Intell.,vol. 31, no.6, pp. 1074-1086, June. 2009.

[39]Y-T. Hsieh, S-Y. Tao, Y-H.H. Tsai, Y-R.Yeh, Y-C.F. Wang,"Recognizing heterogeneous cross-domain data via generalized joint distribution adaptation,"in Proc.IEEE Int. Conf. Multimedia.Expo.,pp. 1-6, July 2016.

[40] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptationregularization:A general framework for transfer learning," IEEE Trans. Knowl.Data Eng., vol. 26, no. 5, pp. 1076–1089, May 2014.

[41]K. Loparo, Case western reserve university bearing data center. (2013). [Online]. Available: http://csegroups. case.edu/ bearing data center/ pages/ 12k-drive-end-bearing-fault-data.

[42][Online]. Available: https://www.phmsociety.org/competition/ PHM/ 09/ apparatus.

[43] C.W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. Neural Netw., vol. 13, no. 12, pp. 415-425, Mar. 2002.

[44]L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, no. 2579/2605, p. 85, Nov. 2008.

Zhao-Hua Liu (M'16**)** He received M.S. degree in computer science and engineering, and the PhD. degree in automatic control and engineering from the Hunan University, China, in 2010 and2012, respectively. He worked as a visiting researcher in the Department of Automatic Control and Systems Engineering at the University of Sheffield, United Kingdom, from 2015 to 2016.

He is currently an Associate Professor of Control and Systems Engineering with the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, China. His current research interests include artificial intelligence algorithm design and applications, condition monitoring and fault diagnosis for electric power equipment, and intelligent control of wind power system. Dr. Liu was a selected as Hu-Xiang Excellent Young Talents of Hunan Province of China in 2018.

Dr. Liu has published a monograph in the field of Biological immune system inspired hybrid intelligent algorithm and its applications, and published more than 30 research papers in refereed journals and conferences, including IEEE TRANSACTIONS/JOURNAL/MAGAZINE .He is a regular reviewer for several international journals and conferences.

Bi-Liang Lu He received B. Eng. degree in Electrical engineering and automation from the Hunan university of science and technology, Xiang tan, China, in 2017. He is currently pursuing the M.S. degree in the automatic control and engineering, at Hunan University of Science and Technology, Xiangtan, China.

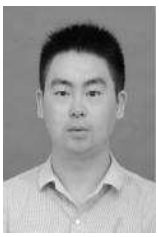
His current research interests include deep learning algorithm design, and condition monitoring and fault diagnosis for electric power equipment.

Hua-Liang Wei received the Ph.D. degree in automatic control and engineering from the Department of Automatic Control and Systems Engineering, the University of Sheffield, UK, in 2004.

He is currently a Senior Lecturer and Head of the Dynamic Modelling, Data Mining and Decision Making (DM3) laboratory, the Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, U.K. His current research interests include system identification and data analytics for complex systems, data driven modeling and data mining, NARMAX methodology and its applications, machine learning and neural networks algorithm design, and parameter estimation.

Xiao-Hua Li received the B.Eng. degree in computer science and engineering from the Hunan University of Science and Engineering, Yongzhou, China, in 2007 and the M.Sc. degree in computer science fromHunan University, Changsha, China, in 2010.

She is currently a Lecturer of computer science with the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, China. Her research interest includes evolutionary computation and machine learning algorithm design and applications.



Lei Chen received M.Sc. degree in computer science and engineering, and the PhD. degree in automatic control and electrical engineering from the Hunan University, China, in 2012 and 2017, respectively.

He is currently a Lecturer with the School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, China. His current research interests include deep learning, cloud computing, and big data analysis.