IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 10, 2018, accepted November 4, 2018, date of publication November 12, 2018,
date of current version December 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2880770

A New Transfer Learning Method and Its Application on Rotating Machine Fault Diagnosis Under Variant Working Conditions

WEIWEI QIAN, SHUNMING LI, AND JINRUI WANG
College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Corresponding author: Weiwei Qian (qianweiwei33@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51675262, in part by the Project of National Key Research and Development Plan of China "New Energy-Saving Environmental Protection Agricultural Engine Development" under Grant 2016YFD0700800, and in part by the Advance Research Field Fund Project of China under Grant 6140210020102.

ABSTRACT Effective data-driven rotating machine fault diagnosis has recently been a research topic in the diagnosis and health management of machinery systems owing to the benefits, including safety guarantee, labor saving, and reliability improvement. However, in vast real-world applications, the classifier trained on one dataset will be extended to datasets under variant working conditions. Meanwhile, the deviation between datasets can be triggered easily by rotating speed oscillation and load variation, and it will highly degenerate the performance of machine learning-based fault diagnosis methods. Hence, a novel dataset distribution discrepancy measuring algorithm called high-order Kullback–Leibler (HKL) divergence is proposed. Based on HKL divergence and transfer learning, a new fault diagnosis network which is robust to working condition variation is constructed in this paper. In feature extraction, sparse filtering with HKL divergence is proposed to learn sharing and discriminative features of the source and target domains. In feature classification, HKL divergence is introduced into softmax regression to link the domain adaptation with health conditions. Its effectiveness is verified by experiments on a rolling bearing dataset and a gearbox dataset, which include 18 transfer learning cases. Furthermore, the asymmetrical performance phenomenon found in experiments is also analyzed.

INDEX TERMS Adaptive signal processing, artificial neural network, fault diagnosis, softmax regression, sparse filtering, transfer learning.

I. INTRODUCTION

Traditional vibration signal-based fault diagnosis methods can give quite comprehensive interpretation of vibration signals and mine the intrinsic information embedded in them [1]. The feature extraction methods utilized by them are always based on signal processing methods. These methods include time-domain analysis [2], frequency transform [3], high resolution time-frequency analysis [4], wavelet transform [5], and envelope demodulation algorithms [6]. Well developed and diverse as traditional fault diagnosis methods are, the feature extraction procedure is always time consuming and labor intensive [7]. In contrast, intelligent fault diagnosis methods are always free of manpower and fast. Meanwhile, intelligent fault diagnosis methods are desirable for the complex systems because locating signal symptoms or establishing explicit system models is challenging for them. The most striking characteristic of intelligent fault diagnosis methods is the feature extraction part of them, which widely takes advantage of artificial neural networks [8]. Lei *et al.* [9] introduced sparse filtering (SF) [10] into rotating machine fault diagnosis. Jia *et al.* [11] proposed a stacked autoencoders (SAE) based DNN for roller bearing and planetary gearbox fault diagnosis. Liu *et al.* [12] introduced CNN into fault diagnosis to deal with 1 dimension vibration signals directly. Zhang *et al.* [13] constructed a stacked denoising autoencoders based DNN to diagnosis fault signals with lower signal to noise ratio (SNR).

Although significant successes have already been achieved in the field of intelligent fault diagnosis [9]–[16], most of these algorithms perform well under a universal assumption: the training and testing datasets have the same domain distribution [17]. Frustratingly, in many real-world application scenarios, it does not hold and the performance will drop

remarkably when distributions of datasets for model training and model application differ [17], [18]. Specially for rotating machine fault diagnosis, vibration signals collected under variant working conditions will differ much in their distributions. Meanwhile, the diversity between dataset distributions is common and occurs frequently which can be triggered easily by working condition variation such as rotating speed oscillation and load variation. This phenomenon is widely called domain shift [19], and it will challenge the effectiveness of the most machine learning-based approaches in fault diagnosis setting in real-world scenarios. Therefore, it is significant and more practical to take consideration of these factors. Universally, the training dataset is called source domain and the dataset for model application is called target domain [20]. Aiming at reducing domain shift, domain adaptation [17] is developed and it can be categorized into transfer learning. The main idea of domain adaptation [18] is adapting the model from the auxiliary source domain to the target domain. It means constructing a prediction model with good generalization ability for the target domain utilizing little or no label information of the target domain. Plenty of application cases in natural language processing, object recognition, image classification [21]–[26] show that the domain adaptation is truly beneficial and promising. Mainly, the forms of information transfer are categorized into four general transfer categories [20] and two widely used categories are: (1) instance transfer and (2) subspace transfer. The former utilizes the source instances which are reweighed according to the sharing information in the target domain as inputs, such as [27] and [28]. The latter aims at mapping the two domains into a sharing feature subspace. Following the idea of learning the sharing feature subspace, plenty of algorithms which realize it by minimizing the domain distribution discrepancy have been proposed [17], [22]. These distribution discrepancy evaluation methods include maximum mean discrepancy (MMD) [29], proxy A-distance [30], and central moment discrepancy (CMD) [22], Kullback-Leibler (KL) divergence [31] *et al.* They are always embedded into original feature extraction methods to encourage the networks mapping the samples from different domains into a sharing feature subspace and retaining the discriminative feature extraction capability of the original network. To achieve this target, evaluation terms such as KL divergence, MMD are always introduced into the objective function to measure the distribution discrepancy of features extracted from source and target domains [17], [21]. Despite the fact that they are all effective in a certain degree, they each have some inherent deficiencies. KL divergence matches the first-order moments and computes fast but takes no consideration of the higher-order moments. Widely used MMD can match distributions better but requires computationally expensive distance and kernel matrix computation.

Recently, for fault diagnosis under variant working conditions, some people used domain adaptation in artificial neural networks to tackle the domain discrepancy between source and target datasets collected from variant working conditions, and obtained quite good results [17], [18]. The domain adaptation approaches they used are both MMD-based methods. Inspired by KL divergence and CMD [22], [31], we propose matching the high-order moments of the domain-specific distributions explicitly via the proposed high-order KL (HKL) divergence. Then, a three-stage intelligent fault diagnosis method with domain adaptation ability is constructed based on HKL divergence for fault diagnosis under variant working conditions. The main contributions of the paper are described as follows.

1) HKL divergence is developed to align the high-order moments of domain distributions. Compared with MMD-based approaches, the proposed method is more effective and computes faster.
2) Two transfer learning methods are developed based on HKL divergence. Sparse filtering with HKL divergence (SF-HKL) is constructed to learn both discriminative and sharing features of the source and target domains. Meanwhile, softmax regression with HKL divergence (SOF-HKL) is developed to take consideration of the labels in domain adaptation, and it is validated that the performance can be further improved by this operation.
3) A three-stage intelligent machine fault diagnosis method is constructed. Firstly, samples are obtained from raw signals in an overlapping way and then transformed into frequency spectra in the preprocessing. Secondly, sharing features are extracted in the network trained by SF-HKL. Finally, features are fed into SOF-HKL to identify the health conditions. The constructed fault diagnosis network is capable of adapting the network to work well on datasets collected from variant working conditions. It fits the real-world applications better, and can handle rotating speed oscillation and load variation in a certain degree. Furthermore, parameter sensitivity is also investigated for the future application.
4) The performance asymmetrical phenomenon [18] which occurs when vibration signals of two working conditions serve as source and target domains alternatively is pointed out and investigated.

The rest of the paper is organized as follows. Section II describes the preliminary knowledge briefly. Section III details the developed method. In Section IV and V, the experimental diagnosis cases are studied using the proposed method. The asymmetrical performance phenomenon is investigated in Section VI. Finally, in Section VII, the conclusions are presented.

II. RELATED KNOWLEDGE

A. PROBLEM DEFINITION

For clarity, the frequently used notations are summarized in Table 1. The fault diagnosis under variant working conditions refers to the problem that using the labeled source dataset $\boldsymbol{Z}_s$ and the unlabeled target dataset $\boldsymbol{Z}_t$ to predict the labels of the samples in the target dataset. It assumes that $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ are collected under different working conditions.

TABLE 1. Notation and description.

Notation	Description	Notation	Description
$\boldsymbol{Z}_s$, $\boldsymbol{Z}_t$	Source/target dataset	$\boldsymbol{f}_s$, $\boldsymbol{f}_t$	Source/target feature matrix
$\boldsymbol{z}_s$, $\boldsymbol{z}_t$	Source/target sample	m_s, m_t	Source/target dataset sample number
$\boldsymbol{x}_{sn}$, $\boldsymbol{x}_{tn}$	The n th order moment of source/target dataset	k_s, k_t	Source/target training sample number
N_{in}, N_{out}	Input/output feature dimension	y_s^i	Label of the i th sample in source dataset

B. KULLBACK-LEIBLER(KL) DIVERGENCE

KL divergence [21] is an asymmetrical measurement of the discrepancy between two probability distributions, and it is also called the relative entropy. Supposing there are two probability distributions $\boldsymbol{P} \in R^{k\times 1}$ and $\boldsymbol{Q} \in R^{k\times 1}$, the KL divergence of $\boldsymbol{Q}$ from $\boldsymbol{P}$ is defined in (1), where $\boldsymbol{P}(i)$ and $\boldsymbol{Q}(i)$ are elements of the ith dimension in $\boldsymbol{P}$ and $\boldsymbol{Q}$ separately.

$$D_{KL}(\boldsymbol{P} \parallel \boldsymbol{Q}) = \sum_{i=1}^{k} \boldsymbol{P}(i) \ln(\frac{\boldsymbol{P}(i)}{\boldsymbol{Q}(i)}) \tag{1}$$

Since it is significant to evaluate and optimize the two distributions equally, the symmetrical form of KL divergence is employed to measure the divergence of two distributions here, as shown in (2).

$$KL(\boldsymbol{P}, \boldsymbol{Q}) = D_{KL}(\boldsymbol{Q} \parallel \boldsymbol{P}) + D_{KL}(\boldsymbol{P} \parallel \boldsymbol{Q}). \tag{2}$$

The similarity between the two distributions increases with the decrease of KL divergence value, which accounts for its ability of evaluating the domain divergence. As it presents, it measures the mean values of $\boldsymbol{P}$ and $\boldsymbol{Q}$ namely the first-order moments of them, so it can be employed to align the two domain distributions in the latent space. However, it should be noted that the higher-order moments of distributions are not considered in KL divergence.

C. SPARSE FILTERING (SF)

The fundamental idea of SF focuses on constraining the sparsity of output features instead of explicitly modeling the input distribution. SF aims at obtaining sparse features that satisfy three principles [9], [10]. The architecture of SF is shown in Fig. 1. $\boldsymbol{z}^i \in \Re^{N_{in}\times 1}$ is the ith sample, which is composed of N_{in} data points, k is the number of samples. The training dataset is used to train SF and obtain an optimized weight matrix $\boldsymbol{W} \in \Re^{N_{out}\times N_{in}}$, where N_{out} is the dimension of output feature vectors. In the training of SF, the following four steps should be conducted sequentially in each iteration, as shown in (3) to (6).

Firstly, the element-wise soft absolute function namely $\sigma(x) = \sqrt{10^{-8}+x^2}$ is used as activation function 1, as depicted in (3). $\boldsymbol{W}_j$ is the jth row of $\boldsymbol{W}$; f_j^i is the ith column and jth row element of feature matrix $\boldsymbol{f} \in \Re^{N_{out}\times k}$.

$$f_j^i = \sqrt{10^{-8} + \left(\boldsymbol{W}_j \boldsymbol{z}^i\right)^2} \approx \left|\boldsymbol{W}_j \boldsymbol{z}^i\right|. \tag{3}$$

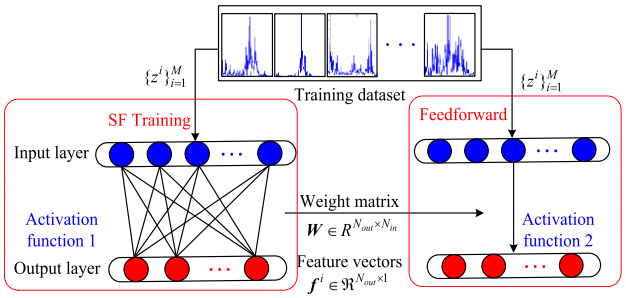


FIGURE 1. Architecture of the SF network.

Next, by means of dividing each kind of feature by its L2 norm [9], [10] across all samples, each kind of feature $\boldsymbol{f}_j$ is normalized to be equally active.

$$\tilde{\boldsymbol{f}}_j = \boldsymbol{f}_j / \left\|\boldsymbol{f}_j\right\|_2 \quad (j = 1, 2 \ldots N_{out}). \tag{4}$$

And then, each column of $\tilde{\boldsymbol{f}}$, namely $\tilde{\boldsymbol{f}}^i$ is normalized through L2 norm in (5) so feature vectors of all samples are equally active.

$$\hat{\boldsymbol{f}}^i = \tilde{\boldsymbol{f}}^i / \left\|\tilde{\boldsymbol{f}}^i\right\|_2 \quad (i = 1, 2 \ldots k). \tag{5}$$

At last, the normalized features are optimized for sparsity by L1 norm. The objective function of SF is depicted in (6). The derivation can be derived through the four steps above correspondingly.

$$L_{SF}(\hat{\boldsymbol{f}}) = \sum_{i=1}^{k} \left\|\hat{\boldsymbol{f}}^i\right\|_1. \tag{6}$$

After SF training, the sample $\boldsymbol{x}^i \in \Re^{N_{in}\times 1}$ is mapped into a feature vector $\boldsymbol{f}^i \in \Re^{N_{out}\times 1}$ in the feed forward process. The mapping can be realized through the optimized $\boldsymbol{W}$ and activation function 2 namely $\sigma(x) = \log\left(1+x^2\right)$. It is an element-wise function and the final feature can be calculated as follows.

$$f_j^i = \log(1 + (\boldsymbol{W}_j \boldsymbol{z}^i)^2). \tag{7}$$

D. SOFTMAX REGRESSION

Softmax regression [11] is often implemented after feature extraction for multiclass classification and performs well. Therefore, we utilize softmax regression as the classifier in this paper. Supposing the input feature matrix $\boldsymbol{f} = \{\boldsymbol{f}^i\}_{i=1}^k$ has label set $\boldsymbol{Y} = \{y^i\}_{i=1}^k$, where $y^i \in \{1, 2, \ldots, R\}$, and k is the number of samples. For each column vector $\boldsymbol{f}^i$, the softmax regression tries to calculate the probability, namely $p(y^i = r|\boldsymbol{f}^i)$ for each label in the output layer. Then, softmax regression can be trained by minimizing the objective function as follows.

$$L_{SOF}(\boldsymbol{f}, \boldsymbol{Y}) = -\frac{1}{K}\left[\sum_{m=1}^{k}\sum_{r=1}^{R} 1\{y_m = r\} \log \frac{e^{\boldsymbol{W}_2^r \boldsymbol{f}^m}}{\sum_{l=1}^{R} e^{\boldsymbol{W}_2^r \boldsymbol{f}^m}}\right] + \lambda_2 \sum_{i=1}^{R}\sum_{j=1}^{N_{out}} (\boldsymbol{W}_2^{i,j})^2. \tag{8}$$

where $1\{\cdot\}$ represents the indicator function, which outputs 0 when the condition is false, and 1 otherwise; $\boldsymbol{W}_2$ is the parameter matrix of softmax regression and $\boldsymbol{W}_2^r$ is the rth row of $\boldsymbol{W}_2$. The second term is the weight decay term, and $\boldsymbol{W}_2^{i,j}$ is the ith row and jth column of $\boldsymbol{W}_2$, where N_{out} is the input dimension of the last layer. As far as we know, few literatures shed light on softmax regression when performing domain adaptation. In this paper, we fuse the HKL divergence into softmax regression explicitly to link the sample labels with the domain adaptation process.

III. PROPOSED METHOD

In this section, HKL divergence which attempts to align high-order moments of distributions is firstly developed. Then, two transfer learning methods namely SF-HKL and SOF-HKL are proposed. Based on them, a robust fault diagnosis network is constructed for fault diagnosis under variant working conditions.

A. HKL DIVERGENCE

KL divergence can align the first-order moments of distributions explicitly and computes fast, but it takes no consideration of higher-order moments. Hence, HKL divergence is developed to align the high-order moments of distributions. Supposing the source and target input matrices are $\boldsymbol{Z}_s = \{\boldsymbol{z}_s^i\}_{i=1}^{k_s}$ and $\boldsymbol{Z}_t = \{\boldsymbol{z}_t^j\}_{j=1}^{k_t}$ separately, where $\boldsymbol{z}_s^i \in \Re^{N_{out}\times 1}$ is the ith column instance vector of $\boldsymbol{Z}_s$ and $\boldsymbol{z}_t^i \in \Re^{N_{out}\times 1}$ is the jth column instance vector of $\boldsymbol{Z}_t$. HKL divergence is obtained by applying equations from (9) to (15) on $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ sequentially, where the square means the element-wise square operation; L_1 and L_n measure the first and nth order moment discrepancy between $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ separately, and $n = 2, 3\ldots$.

Equation (9) and Equation (10) calculate the mean values of each dimension in $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ respectively, namely the first-order moments of distributions. $z_s^{i,j}$ and $z_t^{i,j}$ are the jth elements of $\boldsymbol{z}_s^i$ and $\boldsymbol{z}_t^i$; $\boldsymbol{x}_{s1}^i$ and $\boldsymbol{x}_{t1}^i$ are the ith dimension elements of $\boldsymbol{x}_{s1}$ and $\boldsymbol{x}_{t1}$ separately.

$$\boldsymbol{x}_{s1}^i = \frac{1}{k_s}\sum\nolimits_{j=1}^{k_s} \boldsymbol{z}_s^{i,j}. \tag{9}$$

$$\boldsymbol{x}_{t1}^i = \frac{1}{k_t}\sum\nolimits_{j=1}^{k_t} \boldsymbol{z}_t^{i,j}. \tag{10}$$

Equation (11) and Equation (12) calculate the nth order moments of each dimension in $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ respectively. $\boldsymbol{x}_{s1}^i$ and $\boldsymbol{x}_{t1}^i$ are the nth order moments of the ith dimension elements in $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ separately.

$$\boldsymbol{x}_{sn}^i = \left(\frac{1}{k_s}\sum\nolimits_{j=1}^{k_s}\left|\boldsymbol{z}_s^{i,j} - \boldsymbol{x}_{s1}^i\right|^n\right)^{\frac{1}{n}}. \tag{11}$$

$$\boldsymbol{x}_{tn}^i = \left(\frac{1}{k_t}\sum\nolimits_{j=1}^{k_t}\left|\boldsymbol{z}_t^{i,j} - \boldsymbol{x}_{t1}^i\right|^n\right)^{\frac{1}{n}}. \tag{12}$$

Equation (13) and Equation (14) calculate the first and nth order moment discrepancy of the two domains separately.

$$L_1 = \sum_{i=1}^{N_{out}} \boldsymbol{x}_{s1}^i \log\left(\frac{\boldsymbol{x}_{s1}^i}{\boldsymbol{x}_{t1}^i}\right) + \boldsymbol{x}_{t1}^i \log\left(\frac{\boldsymbol{x}_{t1}^i}{\boldsymbol{x}_{s1}^i}\right). \tag{13}$$

$$L_n = \sum_{i=1}^{N_{out}} \boldsymbol{x}_{sn}^i \log\left(\frac{\boldsymbol{x}_{sn}^i}{\boldsymbol{x}_{tn}^i}\right) + \boldsymbol{x}_{tn}^i \log\left(\frac{\boldsymbol{x}_{tn}^i}{\boldsymbol{x}_{sn}^i}\right). \tag{14}$$

Equation (15) calculates the overall distribution discrepancy of domains. This term can be fused into the original objective function of networks and then involved in the training of the networks. Although the effectiveness of HKL divergence will increase with n, the computing cost will increase accordingly too, so we make a tradeoff and set n to 2 in the following.

$$L_{HKL}(\boldsymbol{Z}_s, \boldsymbol{Z}_t) = L_1 + \sum_{n=2}^{\infty} L_n. \tag{15}$$

B. FAULT DIAGNOSIS NETWORK CONSTRUCTION

Mainly, the proposed fault diagnosis network is composed of three parts: (1) preprocessing; (2) feature extraction and (3) feature classification. In feature extraction, SF-HKL is constructed to extract sharing features. In feature classification, SOF-HKL is constructed to align domain adaptation further by linking the domain adaptation with the source sample labels. The proposed method is concise and computes fast, and the flowchart of the method is displayed in Fig. 2.

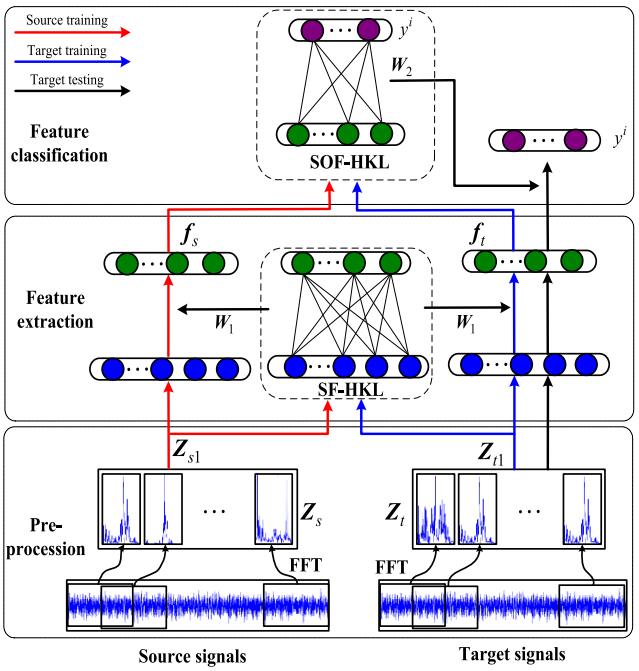


FIGURE 2. The flowchart of the proposed method.

1) PREPROCESSING

In rotating machine fault diagnosis, the source and target domain samples are always acquired from different working

conditions, and the domain shift always arises from load variation or rotating speed oscillation. To utilize the limited data better, data augmentation [32] is adopted through applying the overlapping sampling strategy on the raw signals. Then, FFT is conducted on signal samples to acquire the frequency coefficients correspondingly. Finally, the first-half frequency coefficients are normalized and used as the final inputs of the model. Supposing $\boldsymbol{Z}_s = \{z^i, y^i\}_{i=1}^{m_s}$ represents the source dataset and $\boldsymbol{Z}_t = \{z^i\}_{i=1}^{m_t}$ represents target dataset, where each sample z^i contains N_{in} frequency coefficients; $y^i \in \{1, 2, \ldots, R\}$ is the health condition label of source sample z^i. The training dataset contains a number of samples selected from each health condition in $\boldsymbol{Z}_s$, which consist $\boldsymbol{Z}_{s1} = \{z^i\}_{i=1}^{k_s}$ and the samples selected from each health condition in $\boldsymbol{Z}_t$, which consist $\boldsymbol{Z}_{t1} = \{z^i\}_{i=1}^{k_t}$. And the whole unlabeled target dataset $\boldsymbol{Z}_t$ is for testing.

2) FEATURE EXTRACTION

Feature extraction refers to the network between input layer and latent feature layer, and we construct SF-HKL as the feature extraction network. As SF is an unsupervised feature extraction method, $\boldsymbol{Z}_{s1}$ and $\boldsymbol{Z}_{t1}$ compose the training dataset for SF without the involvement of labels. The HKL divergence term namely $L_{HKL}(\hat{\boldsymbol{f}}_s, \hat{\boldsymbol{f}}_t)$ is fused into the original SF objective function $L_{SF}(\hat{\boldsymbol{f}})$ to reduce the distribution discrepancy of features extracted from source and target domains, and the final objective function of SF-HKL is shown in (16).

$$L_1 = L_{SF}(\hat{\boldsymbol{f}}) + \lambda_1 L_{HKL}(\hat{\boldsymbol{f}}_s, \hat{\boldsymbol{f}}_t). \tag{16}$$

where $\hat{\boldsymbol{f}}$ is the combination of $\left\{\hat{\boldsymbol{f}}_s, \hat{\boldsymbol{f}}_t\right\}$; $\hat{\boldsymbol{f}}_s = \left\{\hat{\boldsymbol{f}}_s^i\right\}_{i=1}^{k_s}$ and $\hat{\boldsymbol{f}}_t = \left\{\hat{\boldsymbol{f}}_t^j\right\}_{j=1}^{k_t}$ are the normalized feature matrix of $\boldsymbol{Z}_{s1}$ and $\boldsymbol{Z}_{t1}$ respectively, and the normalization process is the same as the one described above; $\hat{\boldsymbol{f}}_t^i$ and $\hat{\boldsymbol{f}}_t^j$ are the normalized column vectors of the ith and the jth samples in $\boldsymbol{Z}_{s1}$ and $\boldsymbol{Z}_{t1}$ respectively; λ_1 makes a tradeoff between the effects of $L_{SF}(\hat{\boldsymbol{f}})$ and $L_{HKL}(\hat{\boldsymbol{f}}_s, \hat{\boldsymbol{f}}_t)$. In this way, the network can extract both discriminative and sharing features by minimizing the new objective function iteratively.

Finally, with the optimized weight matrix $\boldsymbol{W}_1 \in \Re^{N_{out} \times N_{in}}$ and (7), we can map $\boldsymbol{Z}_{s1}$ and $\boldsymbol{Z}_{t1}$ to feature matrix $\boldsymbol{f}_s \in \Re^{N_{out} \times k_s}, \boldsymbol{f}_t \in \Re^{N_{out} \times k_t}$ separately.

3) FEATURE CLASSIFICATION

Feature classification refers to the network between the latent feature layer and output layer. In domain adaptation, domain distributions are always aligned in unsupervised feature extraction but it can not be linked to the labels. Therefore, SOF-HKL is developed to solve the domain shift problem further. The input matrices of HKL divergence term can be calculated through (17) and (18), where $|\cdot|$ is the element-wise absolute function. Supposing the weight matrix of softmax regression is $\boldsymbol{W}_2 \in \Re^{C \times N_{out}}$, the optimized weight matrix $\boldsymbol{W}_2$ can be obtained by minimizing (19), where the inputs of SOF-HKL are $\boldsymbol{f}_s$ and $\boldsymbol{f}_t$; $\boldsymbol{Y}_s = \{y^i\}_{i=1}^{k_s}$ is the label set of $\boldsymbol{f}_s$; λ_3 is the parameter which makes a tradeoff between $L_{SOF}(\boldsymbol{f}_s, \boldsymbol{Y}_s)$ and $L_{HKL}(\boldsymbol{F}_s, \boldsymbol{F}_t)$.

$$\boldsymbol{F}_s = \left|\boldsymbol{f}_s\right|. \tag{17}$$

$$\boldsymbol{F}_t = \left|\boldsymbol{f}_t\right|. \tag{18}$$

$$L_2 = L_{SOF}(\boldsymbol{f}_s, \boldsymbol{Y}_s) + \lambda_3 L_{HKL}(\boldsymbol{F}_s, \boldsymbol{F}_t). \tag{19}$$

Finally, the trained fault diagnosis network can be applied directly to the testing dataset $\boldsymbol{Z}_t$ and the corresponding health conditions of target domain samples can be obtained.

IV. CASE STUDY I: FAULT DIAGNOSIS OF MOTOR BEARING UTILIZING THE PROPOSED METHOD

A. DATA DESCRIPTION

The dataset provided by Case Western Reserve University [33] is adopted to validate the effectiveness of the proposed method. The signals were all acquired from the drive end of the motor through a tri-axial accelerator under 4 health conditions: 1) normal condition (NO); 2) outer race fault (OF); 3) inner race fault (IF) and 4) roller fault (RF). Vibration signals of three different severity levels (0.18, 0.36 and 0.54 mm) were collected from health condition OF, IF and RF separately. Additionally, the signals were all collected under 4 different loads (0, 1, 2 and 3 hp) and the sampling frequency was 12 kHz. For convenience, the outer race fault of severity level 0.18 mm is denoted as OF18, and other health conditions are also denoted in this way.

As the distributions of domains vary with the working conditions, the samples are separated into corresponding domains according to their working conditions. The components of each domain are presented in Table 2. Each domain contains 10000 samples, which are obtained evenly from ten health conditions, and each sample contains 1200 data points. Then time-domain samples are preprocessed into frequency-domain samples. Some preprocessed samples are shown in Fig. 3 and the details of the preprocessing are as follows [17].

TABLE 2. Details of each domain in the bearing dataset.

Domain	Load (hp)	Rotating speed (r/min)	Number of samples	Number of health conditions
A	0	1797	10000	10
B	1	1772	10000	10
C	2	1750	10000	10
D	3	1730	10000	10

1) FFT is applied to each time-domain sample to acquire the corresponding frequency coefficients.
2) The single-sided frequency coefficients are used as inputs for the model. So the dimension of the input is 600.
3) The single-sided frequency coefficients are normalized to bigger ones [17], and we set the upper bound to 20 here.

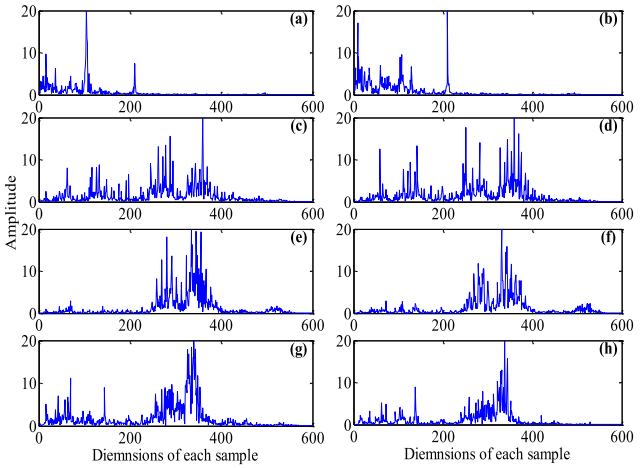


FIGURE 3. The preprocessed samples from domains A and D: (a) a sample of NO in A; (b) a sample of NO in D; (c) a sample of IF18 in A; (d) a sample of IF18 in D; (e) a sample of OF18 in A; (f) a sample of OF18 in D; (g) a sample of BF18 in A and (h) a sample of BF18 in D.

To denote the transfer learning process, we use denotation 'a−b' represents that the source domain contains samples of all health conditions under load 'a' and the target domain contains samples of all health conditions under load 'b'.

B. NETWORK PARAMETER SELECTION AND SENSITIVITY ANALYSIS

There are 3 tunable parameters in the constructed network: λ_1, λ_3 and N_{out}. The sensitivities of the performance to these parameters are investigated along with the tuning. In initialization, λ_1, λ_3 and N_{out} are set to 1000, 0.01 and 600 respectively. The numbers of training samples from source and target domains namely k_s and k_t are set to 3000 and 1000, and the samples are selected from each health condition randomly and evenly. As 12 transfer learning cases exist, we will demonstrate the ones which have the most different working conditions, namely case: A-D and D-A. However, the final parameter selection is also validated on the other transfer learning cases. Meanwhile, the number of iterations is also tuned and is set to 100 for there exists the degeneration phenomenon in SF [34].

First of all, N_{out} is investigated. This parameter is critical for it is directly linked with the computing cost, including the training time and the storage cost. For both cases, it is displayed in Fig. 4 that the accuracy varies more than 10% and the training time increases almost linearly with the rise of N_{out} within the variation range of N_{out}. And the performances of both cases increase little when N_{out} is bigger than 500, so we make a tradeoff and set N_{out} to 500.

Secondly, λ_1 and λ_3 are investigated together, for they both weigh the effects of the original terms and the HKL divergence terms in the objective functions, and influence the performance of the network together. As shown in Fig. 5, it presents in both cases that the version without HKL divergence namely when $\lambda_1 = 0$, $\lambda_3 = 0$ has lower testing accuracies than the network with properly tuned parameters.

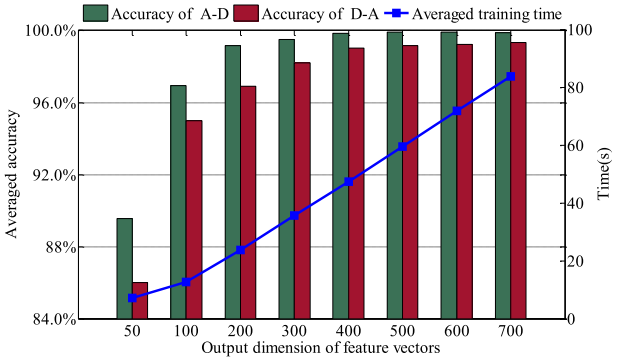


FIGURE 4. Diagnosis results of the proposed method with various output dimensions of feature vectors.

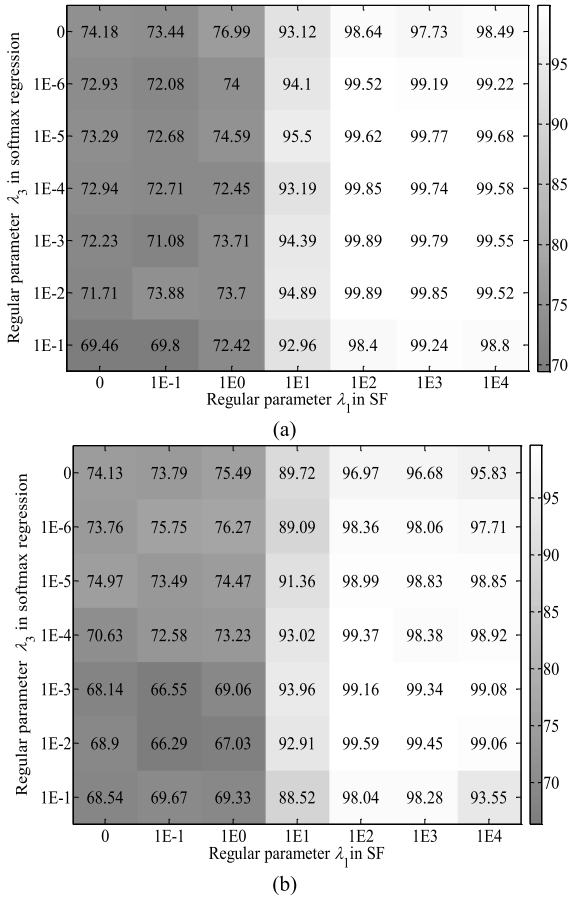


FIGURE 5. Testing results with various regular parameters λ_1 and λ_3 in cases: (a) case A-D and (b) case D-A.

The performance peaks when λ_1 is around 1E3 and λ_3 is around 1E-2. Meanwhile, it is also presented in Fig. 5 that the performance is more sensitive to λ_1 for the accuracies vary more than 30% in the variation range of λ_1 and vary less than 10% in the variation range of λ_3. Finally, we set λ_1 to 1E3 and λ_3 to 1E-2.

As shown above, the most sensitive parameter is λ_1 in the variation range of each parameter. Therefore, more attention should be paid to the selection of λ_1, and the other parameters can be set in more flexible ranges without increasing the risk of affecting the diagnosis performance a lot.

C. DIAGNOSIS RESULTS OF THE CONSTRUCTED NETWORK

In this section, the performance of the proposed method is investigated using all the 12 transfer learning cases, as shown in Table 3. The parameters and the proportion of samples from source and target domains namely k_s/k_t are tuned in each case for SF. For convenience, the value of k_s/k_t is denoted as μ. It should be pointed out that the target domain samples also participate in the unsupervised feature extraction part of SF, because unsupervised feature extraction can get some domain-adaptation ability when samples from both domains are adopted. Furthermore, we denote the proposed method with a fixed μ in all cases as Method 1, and denote the one with a tuned μ for each case as Method 2.

TABLE 3. Diagnosis results of all transfer learning cases (%).

Target Domains		Source Domains			
		A	B	C	D
A	SF	0	90.01 ± 1.47	70.08 ± 1.35	89.87 ± 1.95
	Method 1	0	99.73 ± 0.28	82.83 ± 4.18	99.39 ± 0.25
	Method 2	0	99.86 ± 0.13	88.50 ± 1.45	100
B	SF	93.98 ± 2.17	0	90.63 ± 3.98	76.48 ± 4.78
	Method 1	99.47 ± 0.60	0	98.68 ± 0.92	90.76 ± 4.12
	Method 2	99.80 ± 0.16	0	99.23 ± 0.19	95.17 ± 4.01
C	SF	73.42 ± 1.79	78.86 ± 1.22	0	72.22 ± 2.04
	Method 1	87.00 ± 3.83	95.58 ± 2.01	0	92.95 ± 4.12
	Method 2	87.54 ± 3.71	99.59 ± 0.20	0	97.81 ± 2.26
D	SF	85.59 ± 2.25	78.00 ± 2.60	71.23 ± 1.96	0
	Method 1	99.66 ± 0.21	95.27 ± 3.51	96.54 ± 3.41	0
	Method 2	99.70 ± 0.19	95.50 ± 2.41	98.16 ± 2.51	0

The format of the result is: averaged accuracy ± standard deviation.

Firstly, the situation with a fixed μ in the proposed method is investigated, and the numbers of samples from source and target domains are still 3000 and 1000 respectively. Generally, it is shown in Table 2 that the testing accuracies of A-D and D-A are beyond 99.5% and the testing accuracies of most cases are beyond 95%. By comparison, it shows that the proposed method outperforms SF in all cases. In detail, the lowest improvement of testing accuracy is 5.49% and the averaged improvement is 13.96%. It demonstrates that the developed method is quite effective in performing domain adaptation.

Then, the diagnosis results of the constructed network with a tuned μ for each case are investigated, as shown in Table 3. Generally, the performance of all transfer learning cases can be further improved by tuning μ. The biggest improvement of the diagnosis accuracy is 5.67%, which occurs in case C-A. The averaged improvement is 1.96%. It indicates the value of μ effects the performance in a certain degree. Meanwhile, the tuning of μ in cases A-D and D-A is presented in Fig. 6(a) and Fig. 6(b) separately. It shows that the optimized values of μ are different in the two cases.

To further show the results, the confusion matrices of the pair of cases B-D and D-B which have the biggest testing

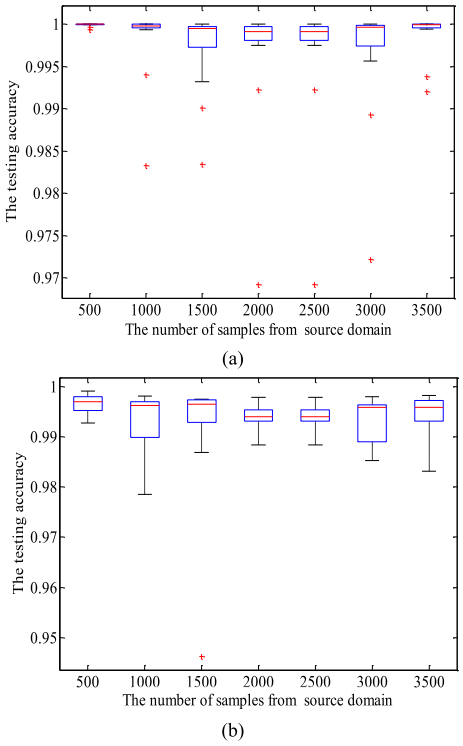


FIGURE 6. The tuning of μ for some cases: (a) case A-D and (b) case D-A.

accuracy discrepancy are shown in Fig. 7, and they are the averaged results of 20 trials. The rows represent the actual health condition types, and the columns represent the predicted health condition types. It presents that samples of categories IF18, BF18 and BF54 are always misclassified in both cases, and BF18 and BF54 are always misclassified as each other mainly because they are similar and are just distinct in fault severity.

D. PERFORMANCE COMPARISONS WITH OTHER ALGORITHMS

To better present the superiority of the proposed method, we compare it with the related works and some domain-adaptation methods in Table 4, which adopt the same motor bearing dataset. In Method 1, the prediction results of SF are displayed for comparison. In Method 2, SF-HKL is employed in the feature extraction. It shows that the averaged testing accuracies of all cases can be improved by more than 9% compared with Method 1, which validates the effectiveness of SF-HKL. In Method 3, SOF-HKL is used in feature classification. It presents that the testing accuracy can be improved in all cases compared with Method 1, which validates the effectiveness of the proposed SOF-HKL. In Method 4, KL divergence is embedded into SF and softmax regression together, and the constructed network is denoted as SF+SOF+KL. In Method 5, SF+SOF+HKL is constructed by embedding HKL divergence into SF and softmax regression together. The comparison shows that SF+SOF+HKL outperforms SF+SOF+KL by 4.98% higher averaged testing accuracy. In Method 6, we fuse the widely used MMD into

TABLE 4. Diagnosis result comparisons with other methods in the bearing dataset.

Method	Description	A-B (%)	B-A (%)	A-C (%)	C-A (%)	A-D (%)	D-A (%)	B-C (%)	C-B (%)	B-D (%)	D-B (%)	C-D (%)	D-C (%)	Mean (%)	Time (S)
1	SF	93.98	90.01	73.42	70.08	85.59	89.87	78.86	90.63	78	76.48	71.23	72.22	**80.87**	**57.87**
2	SF+HKL	97.55	97.75	81.38	79.07	98.18	96.48	94.03	93.27	86.61	86.07	87.96	87.55	**90.49**	**65.26**
3	SOF+HKL	96.7	92.45	79.85	82.31	92.02	90.51	93.6	91.37	90.3	92.97	90.6	93.52	**90.52**	**64.56**
4	SF+SOF+KL	96.05	99.71	80.62	83.16	99.21	97.29	97.9	96.76	88.51	85.24	88.24	88.46	**91.76**	**70.46**
5	SF+SOF+HKL	99.8	99.86	87.54	88.5	99.7	100	99.59	99.23	95.5	95.17	98.16	97.81	**96.74**	**72.89**
6	SF+SOF+MMD	96.4	98.71	82.1	83.16	98.43	98.04	96.89	95.88	87.47	86.46	93.54	92.35	**92.45**	**77.82**
7	SF+CORAL	95.24	90.99	76.88	76.42	89.81	92.18	82.86	78.38	79.94	78.13	80.76	84.9	**83.87**	**60.45**
8	SAE	89.47	93.25	72	76.9	79	77.88	81	92.13	81.03	82.02	86.37	86.64	**83.14**	**91.52**
9	TCA	95.36	95.55	76.28	72.44	92.94	92.37	96.23	92.55	79.91	81.82	95.36	95.36	**88.85**	**82.47**
10	TJM	97.81	97.96	95.4	74.63	94.98	98.34	97.92	95.64	86	85.71	98	93.14	**92.96**	**312.01**
11	SVM	56.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	59.6	N/A	**57.90**	**N/A**
12	DBN	86.7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	88.2	N/A	**87.45**	**N/A**
13	ANN	67.3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	68.1	N/A	**67.70**	**N/A**

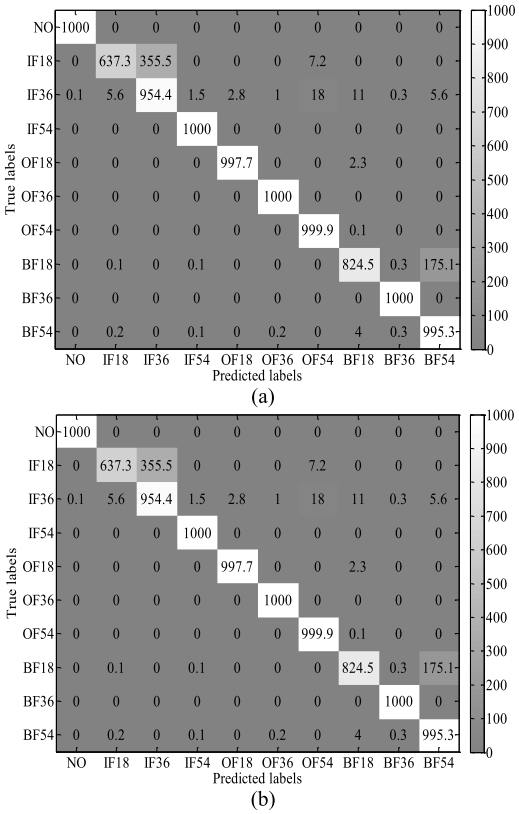


FIGURE 7. Confusion matrices of transfer learning cases: (a) case B-D and (b) case D-B.

SF and softmax regression and denote it as SF+SOF+MMD. By comparison, it shows that SF+SOF+MMD is effective in improving the diagnosis accuracy compared with Method 1, but the effectiveness is inferior to SF+SOF+HKL. In Method 7, we construct a network where a domain adaptation strategy called correlation alignment (CORAL) [35] is employed to do domain adaptation on the samples before feature extraction, the result shows that it can improve the averaged prediction accuracy by 3%, but still inferior to SF+SOF+KL or SF+SOF+HKL. In Method 8, a 5-layer deep fault diagnosis network is constructed and denoted as SAE, which is firstly pre-trained by AE and then fine tuned by global fine tuning. The results show that it outperforms SF for having deeper network structure and is inferior to SF+HKL for having no domain adaptation. In Method 9, we construct a fault diagnosis network based on a domain adaptation method called transfer component analysis (TCA) [36]. It shows that TCA outperforms SF in all cases mainly for having domain adaptation, and is inferior to SF+SOF+MMD mainly for taking no consideration of label information. In Method 10, we construct a network based on a domain adaptation method called transfer joint matching (TJM) [37], which conducts domain adaptation by both instance transfer and subspace transfer. By comparison, it shows that TJM outperforms SF+KL mainly for having instance transfer operation, and is inferior to SF+SOF+MMD mainly for taking no consideration of labels in the domain adaptation. In Method 11, 12 and 13, support vector machine (SVM), deep belief network (DBN) and artificial neural network (ANN) [38] are selected as contrast, and the details can be found in [18]. The averaged testing accuracies of them are 87.45%, 57.90% and 67.70% respectively, and it shows that their performances are obviously inferior to SF+SOF+KL and SF+SOF+HKL. As complexity is also an important performance index of machine learning-based methods, we also present the comparisons of training time among these transfer learning methods in Table 4. It shows that SF+SOF+HKL needs a little more time in network training when compared with SF+SOF+KL, and it needs less training time when compared with SF+SOF+MMD. Meanwhile, it also presents that SF+CORAL needs the least time for training among all the transfer learning methods for it only needs to reweight the training samples in the preprocessing, which is not quite time-consuming.

V. CASE STUDY II: FAULT DIAGNOSIS OF GEARBOX UTILIZING THE PROPOSED METHOD

A. DATA DESCRIPTION

A gearbox vibration signal dataset [39] is utilized to verify the effectiveness of the proposed method, and the test bench is shown in Fig. 8. There are six health conditions of gears: 1) normal condition (NC); 2) pinion gear worn (PW); 3) wheel gear pitting (WP); 4) wheel gear broken-tooth (WB);

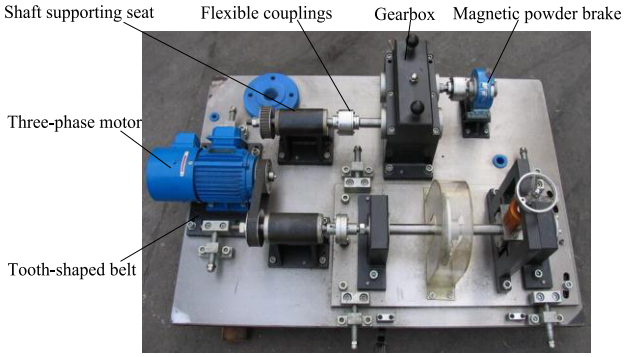


FIGURE 8. The test bench of the gearbox.

5) wheel gear pitting and pinion gear worn (WPPW) and 6) wheel gear broken-tooth and pinion gear worn (WBPW), and each health condition has 3 loads. The sampling frequency was 5120 Hz and 500 samples are acquired from each health condition under one load. The details of the domains under different working conditions are given in Table 5, and there are 3000 samples in each domain. The dimensions of the time-domain samples and the corresponding frequency-domain samples are 1500 and 750 separately.

TABLE 5. Details of each domain in the gearbox dataset.

Domain	Load (hp)	Rotating speed (r/min)					
		NC	PW	WP	WB	WPPW	WBPW
A	1	880	880	878	877	877	881
B	2	820	834	840	825	825	830
C	3	852	850	860	849	849	854

B. DIAGNOSIS RESULTS AND COMPARISONS

The diagnosis results of the proposed method and comparisons with other methods are shown in Table 6. The descriptions of the methods are the same as the ones shown in Section V, and all the parameters of them are tuned. As there are six possible transfer learning cases, the results of them are all presented in Table 6. Generally, the testing accuracies of the proposed method in most cases are beyond 90%. In detail, compared with SF, the lowest improvement is 7.22% and the averaged improvement is 17.90%. Meanwhile, it shows that the proposed method outperforms the others in all cases, which verifies that the proposed method is fairly robust to working condition variation. Furthermore, the diagnosis results are consistent with the results shown in the bearing case.

TABLE 6. Diagnosis result comparisons with other methods in the gearbox dataset.

Description	A-B (%)	B-A (%)	A-C (%)	C-A (%)	B-C (%)	C-B (%)
SF	78.66	74.93	72.77	62.59	72.26	91.83
SF+HKL	96.84	92.23	89.60	81.93	91.83	98.06
SOF+HKL	92.36	92.53	91.14	77.13	89.43	96.84
SF+SOF+KL	97.40	93.03	88.34	81.03	91.14	98.31
SF+SOF+HKL	98.31	94.49	92.53	83.09	92.97	99.05
SF+SOF+MMD	95.18	91.87	92.40	81.60	89.03	98.53
SF+CORAL	92.53	93.17	89.23	78.77	87.84	93.39
SAE	95.18	84.34	82.62	82.03	81.03	93.56
TCA	94.96	90.31	84.65	80.85	73.92	96.77
TJM	97.50	92.56	86.50	81.80	83.33	97.46

VI. ASYMMETRICAL PERFORMANCE PHENOMENON

It is observed that the performance of case 'a-b' is different from case 'b-a' [18] when the value of μ is fixed in both SF+SOF+HKL and SF, and we call it asymmetrical performance phenomenon here, which appears in all pairs of transfer learning cases. The testing accuracy discrepancies of some pairs of transfer learning cases are shown in Fig. 9. There are mainly two reasons accounting for the phenomenon.

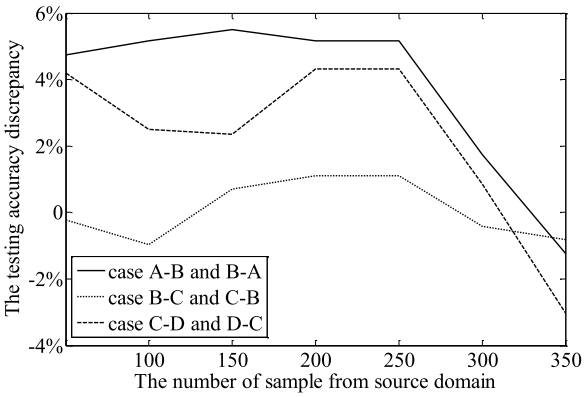


FIGURE 9. Diagnosis result discrepancy of some pairs of cases with various k_S/k_t values.

1) It could be noted that as SF is an unsupervised feature extraction method, the proportion of the numbers of samples from the two domains, namely μ can cause the phenomenon, as shown in Fig. 9. Meanwhile, for HKL divergence, it is extended from the symmetrical version of KL divergence, so it will not cause this phenomenon when μ is set to 1.
2) Although domain adaptation by sharing feature subspace learning can extract the sharing features of both domains, there is no guarantee that the conditional distributions [20] of the sharing features are same for source and target domains due to no labels participate in feature extraction. So the sharing features may have different distributions in the two domains. Supposing the distributions of the ith dimension features of all samples in source and target domains are X1 and X2 respectively, and the distributions of labels are Y for both domains. It is unnecessary that P(Y|X1) is identical to P(Y|X2) and that accounts for the asymmetrical performance phenomenon in a certain degree. It can also be verified by the situation that although relieved by using SOF-HKL, the asymmetrical phenomenon still exists when the value of μ is tuned.

VII. CONCLUSION

In real-world fault diagnosis environments, working condition variation caused by rotating speed oscillation and load variation is a common setting, and it can lead to the diversity between source and target dataset distributions easily. Hence, HKL divergence is proposed to adapt domain distributions further by aligning the high-order moments of distributions in source and target domains. Meanwhile, HKL divergence is embedded into SF and softmax regression, and a robust fault diagnosis network based on them is constructed in this paper. Experiments on a widely-used bearing dataset and a gearbox dataset validate the superiority of the proposed method in eliminating the effect of the working condition variation. The results show that HKL divergence can do domain adaptation further than KL divergence by aligning the high-order moments of distributions. Furthermore, the effectiveness of employing HKL divergence in SF or softmax regression for domain adaptation is also validated separately by experiments.

The asymmetrical performance phenomenon is pointed out and analyzed, which occurs when two datasets from different working conditions serve as source or target domain alternatively. Two main reasons account for the phenomenon are given. (1) The numbers of training samples from source and target domains are unequal; (2) There are no labels participating in feature extraction, so the extracted features of source and target domains may have different conditional distributions. Softmax regression with HKL divergence can relieve this situation but the effectiveness is limited. We will investigate more powerful network structure based on HKL divergence in our future work.

REFERENCES

[1] M. Kang, J. Kim, J.-M. Kim, A. C. C. Tan, E. Y. Kim, and B.-K. Choi, "Reliable fault diagnosis for low-speed bearings using individually trained support vector machines with kernel discriminative feature analysis," *IEEE Trans. Power Electron.*, vol. 30, no. 5, pp. 2786–2797, May 2015.

[2] J. Yin, W. Wang, Z. Man, and S. Khoo, "Statistical modeling of gear vibration signals and its application to detecting and diagnosing gear faults," *Inf. Sci.*, vol. 259, no. 3, pp. 295–303, 2014.

[3] W. Li, Z. Zhu, F. Jiang, G. Zhou, and G. Chen, "Fault diagnosis of rotating machinery with a novel statistical feature extraction and evaluation method," *Mech. Syst. Signal Process.*, vols. 50–51, pp. 414–426, Jan. 2015.

[4] Z. Feng, M. Liang, and F. Chu, "Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples," *Mech. Syst. Signal Process.*, vol. 38, no. 1, pp. 165–205, 2013.

[5] C. Shen, D. Wang, F. Kong, and P. W. Tse, "Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier," *Measurement*, vol. 46, no. 4, pp. 1551–1564, 2013.

[6] A. B. Ming, W. Zhang, Z. Y. Qin, and F. L. Chu, "Envelope calculation of the multi-component signal and its application to the deterministic component cancellation in bearing fault diagnosis," *Mech. Syst. Signal Process.*, vols. 50–51, pp. 70–100, Jan. 2015.

[7] Z. 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, Jan. 2015.

[8] K. Worden, W. J. Staszewski, and J. J. Hensman, "Natural computing for mechanical systems research: A tutorial overview," *Mech. Syst. Signal Process.*, vol. 25, no. 1, pp. 4–111, 2011.

[9] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.

[10] J. Ngiam, Z. Chen, S. A. Bhaskar, P. W. Koh, and A. Y. Ng, "Sparse filtering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1125–1133.

[11] F. Jia, Y. G. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 303–315, May 2016.

[12] R. Liu, G. Meng, B. Yang, C. Sun, and X. Chen, "Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine," *IEEE Trans. Ind. Inf.*, vol. 13, no. 3, pp. 1310–1320, Jun. 2017.

[13] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, 2017.

[14] X. Jiang, S. Li, and Q. Wang, "A study on defect identification of planetary gearbox under large speed oscillation," *Math. Problems Eng.*, vol. 2016, Dec. 2016, Art. no. 5289698, doi: 10.1155/2016/5289698.

[15] M. Wang, H.-X. Li, X. Chen, and Y. Chen, "Deep learning-based model reduction for distributed parameter systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 12, pp. 1664–1674, Dec. 2016.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[17] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.

[18] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2017.2754287.

[19] V. Kalogeiton, V. Ferrari, and C. Schmid, "Analysing domain shift factors between videos and images for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2327–2334, Nov. 2016.

[20] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, 2016.

[21] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Proc. 24th Int. Conf. Artif. Intell.* New York, NY, USA: AAAI Press, 2015, pp. 4119–4125, doi: 10.1145/3108257.

[22] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," in *Proc. Int. Conf. Represent. Learn.*, 2017.

[23] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 24, no. 4, Nov. 2011, pp. 999–1006.

[24] T. Turki, Z. Wei, and J. T. L. Wang, "Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients," *IEEE Access*, vol. 5, pp. 7381–7393, 2017.

[25] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[26] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowl.-Based Syst.*, vol. 80, pp. 14–23, May 2015.

[27] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[28] M. Chen, K. Q. Weinberger, and J. C. Blitzer, "Co-training for domain adaptation," in *Proc. NIPS*, 2011, pp. 2456–2464.

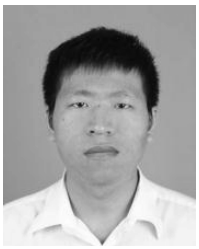
[29] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[30] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, vol. 6314, 2010, pp. 213–226.

[31] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[32] I. Goodfellow, Y. Bengio, A. Courville, and F. Bach, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[33] *Bearing Data Center, Case Western Reserve University*. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file
[34] J. Lederer and S. Guadarrama, "Compute less to get more: Using ORC to improve sparse filtering," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3797–3803.
[35] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2015, pp. 2058–2065.
[36] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[37] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1410–1417.
[38] H. Shao, H. Jiang, X. Zhang, and M. Niu, "Rolling bearing fault diagnosis using an optimization deep belief network," *Meas. Sci. Technol.*, vol. 26, no. 11, p. 115002, 2015.
[39] X. X. Jiang, S. M. Li, and Y. Wang, "A novel method for self-adaptive feature extraction using scaling crossover characteristics of signals and combining with LS-SVM for multi-fault diagnosis of gearbox," *J. Vibroeng.*, vol. 17, pp. 1861–1878, 2015.

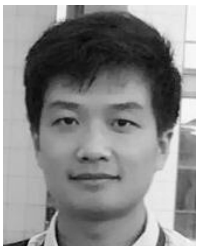


WEIWEI QIAN received the B.S. degree from the Jiangsu University of Science and Technology, Zhenjiang, China, in 2012 and 2016, respectively.

He is currently pursuing the Ph.D. degree with the College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include rotating machinery fault diagnosis, mechanical signal and information processing, and machine learning.

SHUNMING LI received the Ph.D. degree in mechanics from Xi'an Jiaotong University, China, in 1988.

He is currently a Professor with the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include noise and vibration analysis and control, signal processing, machine fault diagnosis, sensing and measurement technology, and intelligent vehicles.



JINRUI WANG received the B.S. and M.S. degrees from the Shandong University of Science and Technology, Tsingdao, China, in 2013 and 2015, respectively.

He is currently pursuing the Ph.D. degree with the College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include rotating machinery fault diagnosis, and mechanical signal and information processing.