Received November 28, 2019, accepted December 16, 2019, date of publication December 23, 2019,
date of current version January 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961711

Adaptive Cross-Domain Feature Extraction Method and Its Application on Machinery Intelligent Fault Diagnosis Under Different Working Conditions

ZENGHUI AN[1], SHUNMING LI[1], XINGXING JIANG[2], YU XIN[1], AND JINRUI WANG[3]
[1]College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2]School of Rail Transportation, Soochow University, Suzhou 215123, China
[3]College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266510, China

Corresponding author: Zenghui An (me_anzenghui@163.com)

This work was supported in part by the Major National Science and Technology under Project 2017-IV-0008-0045, in part by the National Natural Science Foundation of China under Grant 51675262 and Grant 51975276, in part by the Advance Research Field Fund Project of China under Grant 61400040304, and in part by the Key Laboratory of Ministry of Industry and Information Technology under Grant KL2019N001.

ABSTRACT Transfer learning based models have been employed for intelligent fault diagnosis under different working conditions. However, an actual and important problem is neglected in existing intelligent fault diagnosis methods, which is that target domain mechanical fault datasets are always highly imbalanced with abundant normal condition mechanical samples but a paucity of samples from rare fault conditions. To solve this actual problem, this paper proposed a novel adaptive cross-domain feature extraction (ACFE) method which can automatically extract similar features between different feature spaces. ACFE wants to obtain more information for determining the category of each training target sample, so it avoids the distribution adaptation and is suitable for the imbalanced problem. Specifically, high dimensional distance through kernel method is employed for ensuring the strong identify ability first. Then, for automatically extracting cross-domain feature, we calculate the posterior probability of category of each target domain training sample based on high dimensional distance, and employ entropy loss to capture the cross-domain information. Besides, we propose the guide loss to avoid the features of a category overall falling into false category caused by imbalanced dataset. Based on ACFE, the intelligent fault diagnosis method for dealing with the imbalanced target dataset is described. To verify the effectiveness, we carry out two specially designed experiments, and the results shows that, comparing with related method, the proposed method achieve a superior performance.

INDEX TERMS Transfer learning, imbalanced dataset, information entropy, intelligent fault diagnosis.

I. INTRODUCTION

In modern industries, rotating machinery plays a crucial role in the fields of aviation, machine tool, automobile and so on. For the important effect, its operating conditions will directly influence the performance of the whole mechanical equipment [1], [2]. But unfortunately, it is prone to failures due to their harsh working environment such as humidity, high temperature, and variable load, resulting in a catastrophic failure of the rotating machinery [3], [4]. Meanwhile, rotating machinery in modern industry becomes more sophisticated than ever before [5]. So in order to comprehensively check the health condition of the machinery, a large amount of signals are obtained after the long-time monitoring, which also brings great difficulties to mechanical fault diagnosis. Therefore, various intelligent algorithms have been proposed for fault diagnosis, such as Artificial Neural Networks (ANN) [6], [7], Autoencoders [8], [9], Restricted Boltzmann Machine (RBM) [10], Convolutional Neural Networks (CNN) [11], [12], Sparse Filtering [13], [14] and k-Nearest Neighbor [15].

But, there is a problem which plagues the mentioned methods. As we all know, intelligent method trained by training dataset can successfully diagnose the testing dataset.

The associate editor coordinating the review of this manuscript and approving it for publication was Youqing Wang.

It should be noticed that training dataset and testing dataset are under the same working condition. In the field of machine learning, this can be regarded as that the training dataset and target dataset follow the same distribution or belong to the same domain. But, in practice, the working condition may change, which lead to the model trained by the original training dataset cannot diagnose the samples under the new working condition, i.e., the changing working conditions makes the distribution different and lead to the training dataset and testing dataset belong to different domains [16]. To solve this problem, transfer learning based intelligent fault diagnosis method have been proposed. Lu *et al.* [17] combined Autoencoder with domain adaptation and proposed Deep neural network for domain Adaptation in Fault Diagnosis (DAFD), and verified the proposed method using gearbox datasets under different working conditions. Wen *et al.* [18] employed Sparse Auto-Encoder as the basic framework and add the domain adaptation loss. The ability to diagnose the bearing dataset under different loads was verified. An *et al.* [19] generalized the deep neural network using multiple kernel method and improved the accuracy of bearing fault diagnosis under different working conditions. They both get the satisfying testing accuracies under their experiment condition.

But an actual and important problem is neglected in the above method based on transfer learning. The unlabeled data, which is under a different working condition from the labeled training data and is used to train the model, is imbalanced in basically [20]. Mechanical fault datasets, similar to medical datasets, genomics and financial datasets [21], are also very limited since the vast majority of samples are normal samples. When the machine is employed under another working condition, the samples collected in a relatively short time must be imbalanced. For data driven based intelligent methods, how to deal with imbalanced datasets, is also an important problem, and have drawn some attention [22]. For transfer learning based methods, to solve the problem of imbalanced data of target domain is a more meaningful task, because the condition will lead to change of distribution and further result in the negative transfer, which troubles the existing methods.

In this paper, we proposed an adaptive cross-domain feature extraction (ACFE) method. ACFE has strong clustering ability. The aim of ACFE is to seek certainty of event that which category does each target sample belong to. So, it avoids dealing with skewing distribution adaptation caused by imbalanced dataset and can automatically find and capture the similar cross-domain feature. Then we apply it in intelligent fault diagnosis, and study the advantage of ACFE. The main contributions of our work can be summarized as follows.

(1) For guaranteeing the more stable and classable features and solving the unboundedness of distance in Euclidean space, we use the high dimensional distance between vectors based on kernel method for clustering.

(2) For automatically extracting cross-domain feature, we calculate the posterior probability of category of each target domain training sample based on high dimensional distance and employ entropy loss to capture the information.

(3) The guide loss is proposed to avoid the features of a category overall falling into false category caused by imbalanced dataset

This paper is organized as follows. In Section II, several theory backgrounds are described. The framework of ACFE and the intelligent fault diagnosis method are detailed in Section III. In Section IV, the diagnosis case of bearing dataset under variable load is studied to test the effectiveness of ACFE. Section V investigates the diagnosis case under different rotational speeds and loads. Finally, main conclusions are given in Section VI.

II. PRELIMINARIES

A. TRANSFER LEARNING AND MOTIVATION OF ACFE

For clearly describing ACFE, we first introduce transfer learning [23], [24]. As we all know, traditional intelligent method trained by training dataset can successfully diagnose the testing dataset. It should be noticed that training dataset and testing dataset are under the same working condition. However, transfer learning based method can deal with the problem that the working conditions of training dataset and testing dataset are different. The requirement of transfer learning based intelligent fault diagnosis method are labeled source domain ($\mathcal{D}^s$) dataset and unlabeled target domain ($\mathcal{D}^t$) dataset (in this paper, domain represents the working condition).

Traditionally, transfer learning is realized by distribution adaptation or called domain adaptation. Because of mechanical imbalanced dataset, distribution adaptation is powerless. As shown in Figure 1, the three histograms above the line mean the feature distributions of different categories and the feature distributions of different categories are different. The Pie charts mean the proportions of different categories. Under this condition, the category ratio directly determines the feature space distribution. So the feature space distributions of different category ratios (the two histograms below the line) are distinguishing obviously. In general, source domain samples are abundant. So we usually select the balanced dataset for training to avoid the gradient skewing. However, the labels of target domain samples are unknown. The unknown category ratio will lead to skewing feature space distribution. Therefore, transfer learning based on distribution adaptation cannot deal with imbalanced dataset.

Above all, it is necessary to develop a new method which can avoid distribution adaptation for intelligent fault diagnosis method under different working condition. For transfer learning, the important precondition is that source domain and target domain are related, which means that if the clustering result of source domain samples is very good, the feature of target domain sample should be close to the correct source domain feature. Therefore, the model should have strong clustering ability and can catch the connection between source domain and target domain, which is the main motivation of our ACFE.

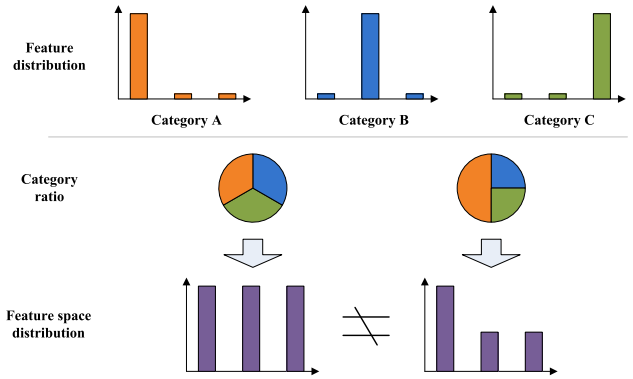


FIGURE 1. The different feature space distribution caused by different category ratio.

B. INFORMATION AMOUNT AND ENTROPY

In the field of information, information is used to dispel uncertain factors [25]. This means that if the probability of the described thing is quite high, the information it provided is less. As a measurement of information, the information amount is a function that depends on the probability [26]. Information theory provides us that the mathematic essence of the information amount is the equation as $-\log(p_i)$, where p_i is the probability of an event [27]. The information amount can describe the information when the event occurs. If we want estimate the indeterminacy of event, the information entropy is provided, which is the mathematic expectation of the amount of information and is defined as follows [28]:

$$H(\boldsymbol{p}) = -\sum_{p_i \in \boldsymbol{p}} p_i \log(p_i) \tag{1}$$

where, $\boldsymbol{p}=\{p_i\}_{i=1}^{n_p}$ is the probability of all the conditions, and $H(\boldsymbol{p})$ is the information entropy of event. Higher $H(\boldsymbol{p})$ means that the indeterminacy of event is also high. Otherwise, the event is more certain.

C. HIGHER DIMENSIONAL DISTANCE BETWEEN TWO VECTORS

In Euclidean space, distance between two vectors is defined as the form of *Frobenius norm*,

$$d(\boldsymbol{v}_1, \boldsymbol{v}_2) = \|\boldsymbol{v}_1 - \boldsymbol{v}_2\| \tag{2}$$

In the higher dimensional space, the difference and similarity can be found more easily. For example, Support Vector Machine (SVM) is based on this theory. In this paper, for improving the stability of learned feature and learning efficiency we adopt multiple kernel method to calculate the higher dimensional distance which is as follows,

$$\begin{aligned} d_{MK}(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \sum_{\varphi \in \Psi} \|\varphi(\boldsymbol{v}_1) - \varphi(\boldsymbol{v}_2)\|^2 \\ &= \sum_{k \in K} [k(\boldsymbol{v}_1, \boldsymbol{v}_1) + k(\boldsymbol{v}_2, \boldsymbol{v}_2) - 2k(\boldsymbol{v}_1, \boldsymbol{v}_2)] \end{aligned} \tag{3}$$

where, $\varphi(\bullet)$ is referred to higher dimensional map. Ψ is the set of maps. $k(\bullet,\bullet)$ is a kernel function, which could compute the inner product in a higher dimensional space, i.e., $k(\boldsymbol{x},\boldsymbol{y})=<\varphi(\boldsymbol{x}), \varphi(\boldsymbol{y})>$. K is the set of kernel functions [29]. So d_{MK} can measure the distance between two vectors under multiple higher dimensional map with the help of kernel functions.

III. ACFE AND PROPOSED FRAMEWORK

A. ADAPTIVE CROSS-DOMAIN FEATURE EXTRACTION METHOD

ACFE is constructed by defining the objective function include clustering loss, entropy loss, guide loss and regularization loss as shown in Figure 2. Firstly, feature extraction process is abstracted into Φ which represents the neural network transformation. We can adjust the structure of Φ to actual case. According to the actual condition, we have labeled source dataset $\{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{n_s}$ and unlabeled target dataset $\{\boldsymbol{x}_i^t\}_{i=1}^{n_s}$, where target dataset is imbalanced in general.

1) CLUSTERING LOSS

The source domain sample $\boldsymbol{x}^s$ is mapped to feature space and is changed to activation vector $\boldsymbol{v}^s = \Phi(\boldsymbol{x}^s; \theta)$, where, θ represents weight of the neural network transformation Φ. With the help of labels, we want to make the activation vectors under the same health condition close together and disperse the activation vectors under different health conditions. Therefore, we calculate the distances between two different activation vectors first. We deal with the vectors in the higher dimensional space so as to enhance the clustering ability. Higher dimensional distance can guarantee the more stable and classable features. Besides, if we directly calculate the distances using the form of *Frobenius norm*, the amplitude of final clustering loss must be quite large because of the unboundedness of distance. This will lead to the failure of training. Therefore, we calculate the distances after mapping the activation vector to higher dimensional space. The kernel method is employed. So the higher dimensional distance is defined as follows:

$$d_{ij} = d_{MK}\left(\boldsymbol{v}_i^s, \boldsymbol{v}_j^s\right) \tag{4}$$

where, $i, j = 1, 2, \ldots, n_s$ and n^s is the number of source domain samples. We want to reduce d_{ij} when the two activation vectors have the same label and increase d_{ij}, otherwise. So we construct the $(n^s \times n^s)$ similarity matrix $\boldsymbol{A}$ with the help of source domain provided labels, where $a_{ij} = 1$ if $\boldsymbol{v}_i^s$ and $\boldsymbol{v}_j^s$ belong to the same category and 0, otherwise. Therefore, the clustering loss L_c is defined as follows:

$$L_c\left(\theta; \boldsymbol{x}^s, y^s\right) = \sum_{a_{ij} \in A} a_{ij} d_{ij} - \left(1 - a_{ij}\right) d_{ij} \tag{5}$$

where, y^s is the label of $\boldsymbol{x}^s$. Minimizing the clustering loss can achieve the desired effect that features belonging to the same category are close together and far to otherwise in the higher dimensional space.

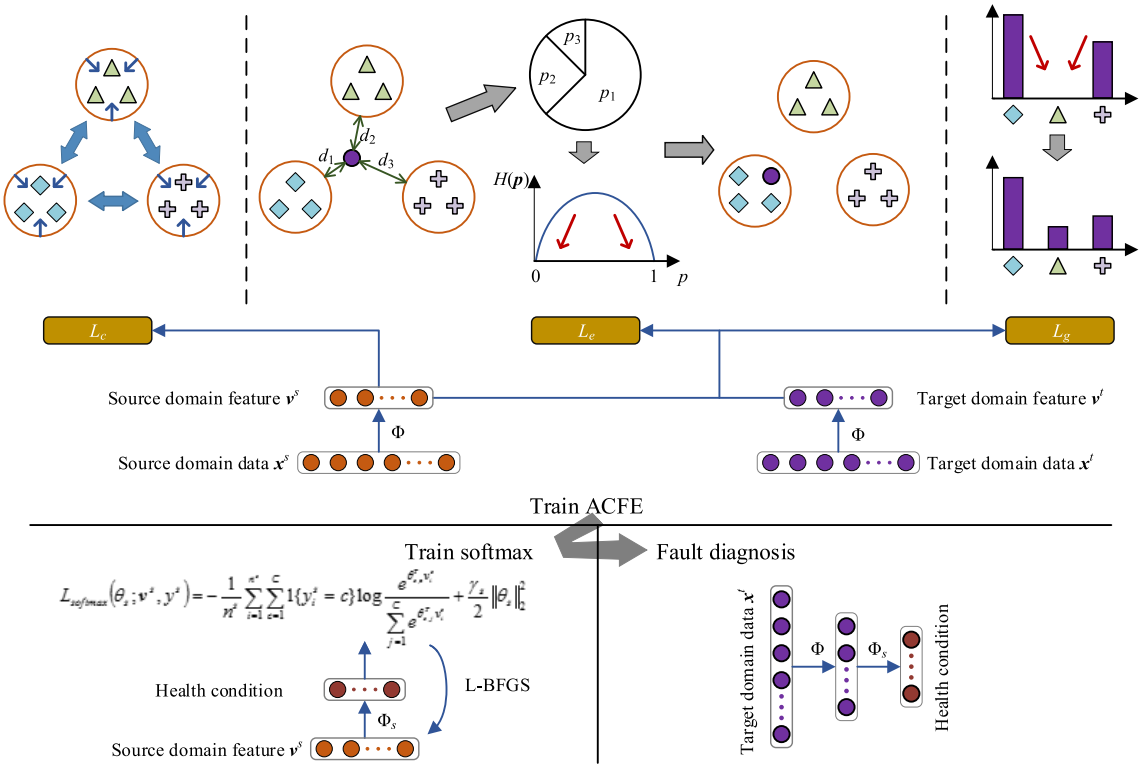


FIGURE 2. The illustration of proposed ACFE and intelligent fault diagnosis method.

2) ENTROPY LOSS

After training with clustering loss, the distances between activation vectors of source domain under different health conditions are very large. The target domain activation vector $\boldsymbol{v}_i^t = \Phi(\boldsymbol{x}_i^t; \theta)$ should be close to some $\boldsymbol{v}^s$ belonging to same category, because the target domain is related to the sources domain. Therefore, if the distance between a target domain activation vector and a source domain activation vector is small, the probability that the two samples belong to the same category is large. Besides, "which category the target sample $\boldsymbol{x}_i^t$ belonging to" can be regarded as an event. So we can obtain the information entropy of the event. The information entropy can measure the indeterminacy of the event. However, we want to make the event more certain, which can be realized by minimizing the information entropy. This is the motivation of entropy loss and the key ability of the entropy loss is to capture the similarities automatically and to further learn them.

For matching with the clustering loss in the higher dimensional space, we also employ the higher dimensional distance. The higher dimensional distances $de_{ij} = d_{MK}(\boldsymbol{v}_i^t, \boldsymbol{v}_j^s)$ are calculated and $\boldsymbol{D}_e = \{de_{ij}\} \in \boldsymbol{R}^{n^t \times n^s}$, where, $i = 1, 2, \ldots, n^t$, $j = 1, 2, \ldots, n^s$ and n^t is the number of target domain samples. With the help of source domain labels, we can calculate the distances between activation vectors $\boldsymbol{v}^t$ and the source category c in the feature space as follows:

$$dc_{ic} = \sum_{y_j^s = c} de_{ij} \tag{6}$$

where, y_j^s is the label of activation vectors $\boldsymbol{v}_j^s$, and $c = 1, 2, \ldots, C$ is the number of category. According to $\boldsymbol{D}_c = \{dc_{ic}\} \in \boldsymbol{R}^{n^t \times C}$, we can obtain the posterior probability of category of each target domain sample. If the distance dc_{ic} is small, the likelihood that the sample $\boldsymbol{x}_i^t$ belongs to the category c is large. Therefore, the posterior probability is

$$p_{ic} = \text{soft max}\,(dc_{ic}) = \frac{e^{-dc_{ic}}}{\sum_{c=1}^{C} e^{-dc_{ic}}} \tag{7}$$

When the target domain activation vectors is only similar to one category and dissimilar to all the other categories, the probability vector $\boldsymbol{p}_i = [p_{i1}, \ldots, p_{iC}]$ tends to be a one-hot vector [30]. A one-hot vector can be viewed as a low information entropy realization of $\boldsymbol{p}_i$. So we employed the information entropy to capture the similar information. The entropy loss is given by,

$$L_e\left(\theta; \boldsymbol{x}^s, y^s, \boldsymbol{x}^t\right) = \sum_{i=1}^{n^t} H\left(\boldsymbol{p}_i\right) \tag{8}$$

As we can see from the curve of information entropy in Figure 2, minimizing the information entropy means that the probability is trended to 0 and 1. Because of the competitive relation caused by softmax activation function, there will be only one condition becoming high probability. So minimizing the entropy loss gives us probability vectors $\boldsymbol{p}_i$ that tend to be one-hot vectors, i.e., the target domain vectors are similar to source domain vectors from any one category only.

3) GUIDE LOSS

Training with imbalanced dataset will lead to the skewing of weight, i.e., even though the samples are representative, the learned feature will be drowning because of the small amounts. Besides, as is well-known, some features belonging to different categories are sometimes similar. When the map Φ is employed to deal with target data, the skewing of features maybe increase the similarity. Then, using transfer learning method to learn the cross-domain features will result in negative transfer. These reasons maybe lead to the condition shown in the top-right corner of Figure 2, which is that samples, belonging to some categories with less target training samples, are all misclassified as other categories.

So we propose the guide loss to solve the problem. If the negative transfer exists, the two similar categories must be regarded as one category, which leads to $\mathrm{Y}^s \neq \mathrm{Y}^t$, where, Y^s and Y^t represent the source label space and target label space respectively. We should guide the training process to the right transfer. We consider that although the target data is imbalanced, the label spaces Y^s and Y^t are same, which means that there should be some target samples belonging to every category. Therefore, the final distribution of target labels must be a dense vector instead of a sparse vector, which is the motivation of guide loss.

According to detail in section III.A.II, $\boldsymbol{P}=\{\boldsymbol{p}_i\}_{i=1}^{n^t}$, in fact, can be regarded as label space Y^t (we call it target domain pseudo label space in this paper). We reduce the dimension of $\boldsymbol{P}$ to obtain the target domain pseudo label distribution $\boldsymbol{P}_c \in \boldsymbol{R}^C$ by averaging all $\boldsymbol{p}_i$, i.e.,

$$\boldsymbol{P}_c = \sum_{i=1}^{n^t} \boldsymbol{p}_i \tag{9}$$

We first normalize $\boldsymbol{P}_c$ though $l2$-normalization,

$$\boldsymbol{P}_{cn} = \frac{\boldsymbol{P}_c}{\|\boldsymbol{P}_c\|} \tag{10}$$

If $\boldsymbol{P}_{cn}$ is a sparse vector, this means the negative transfer is reinforced, i.e., the target samples are identified as several limited categories. This goes against the facts. We should make $\boldsymbol{P}_{cn}$ a dense vector. Therefore, the guide loss is proposed as follows:

$$L_g\left(\theta; \boldsymbol{x}^s, y^s, \boldsymbol{x}^t\right) = -\left|\boldsymbol{P}_{cn}\right| \tag{11}$$

where, $|\bullet|$ is $l1$-norm. Minimizing the guide loss gives us a dense vector $\boldsymbol{P}_{cn}$, i.e., all categories have samples.

IV. REGULARIZATION LOSS

Appropriate regularization method is necessary for solve the problem of overfitting. For ACFE, the simple $l2$-regularization loss, $L_r(\theta) = ||\theta||^2$, is employed to penalize θ.

Combining those four optimization objects, the final optimization object can be written as,

$$L = L_c + \alpha L_e + \beta L_g + \gamma L_r \tag{12}$$

where the hyper-parameters α, β and γ determine how strong the optimization objects is.

Based on the equation (12) and SGD algorithm, the parameters θ are updated as follows,

$$\theta \leftarrow \theta - \varepsilon\left(\frac{\partial L_c}{\partial \theta} + \alpha\frac{\partial L_e}{\partial \theta} + \beta\frac{\partial L_g}{\partial \theta} + \gamma\frac{\partial L_r}{\partial \theta}\right) \tag{13}$$

where, ε is the learning rate.

When the training process is completed, the map Φ is able to automatically extract the cross-domain classable features. Then, the feature can be used to further process.

A. PROPOSED FRAMEWORK

This section details the intelligent fault diagnosis method based on the proposed ACFE. The method includes two stage, adoptive cross-domain feature extracting and fault diagnosis. Our method is designed to utilize the labeled samples under A working condition and imbalanced unlabeled samples under B working condition to obtain the transferable feature extraction model, which can also be applied to diagnose fault samples under B working condition.

1) ADOPTIVE CROSS-DOMAIN FEATURE EXTRACTION

First, the form of map Φ should be designed. We adopt Fast Fourier Transform (FFT) to time domain signals, and the spectrum is used as the input $\boldsymbol{x}$ for Φ. A three layers neural network is adopted as Φ. Every layer includes weight matrix $\boldsymbol{W}_i$ and biases vector $\boldsymbol{b}_i$, where $i = 1, 2, 3$, i.e. $\theta = \{\boldsymbol{W}_i, \boldsymbol{b}\}_{i=1}^3$. Besides, activation functions of every layer are Rectified Linear Unit (ReLU).

Then, ACFE can be adopted with map of three layers neural network. According to Ref. [29], the Gaussian radial basis function (RBF), i.e., $k_G(x_1, x_2) = \exp(||x_1 - x_2||^2/2s^2)$, has been proven usefulness in practice [31], where s is the standard deviation. Therefore, the proposed method uses G different RBF and sums them to calculate the distances for ensuring the features stabilized in different infinite-dimensional spaces. We use 5 RBF whose mid-value is m and times between two s, where, m is the mean of $||x_1 - x_2||$ of the training data.

In the beginning of optimization, the clustering capability of model is quite poor. So we want to train the clustering loss. When the model has the ability of classification, we should pay more attention to learn the cross domain feature. Therefore, α and β should increase with the training epoch. We set it as follows.

$$\alpha, \beta = \frac{2}{1 + e^{-10q}} - 1 \tag{14}$$

where q is linearly changing from 0 to 1 when we train the model.

2) FAULT DIAGNOSIS BASED ON CROSS-DOMAIN FEATURE

After cross-domain feature extracting, source features and target features belong to the same feature space. Therefore, we can train the classifier only with the labeled source

features $\{\boldsymbol{v}_i^s, y_i^s\}_{i=1}^{n^s}$. For diagnosing fault, softmax regression [32] is employed. In this paper, it is called map Φ_s. Because the features output from ACFE include much classable information, the main effect of softmax is actually to show the result clearly. So one layer softmax regression is adopted and the loss is defined as,

$$L_{softmax}\left(\theta_s; \boldsymbol{v}^s, y^s\right) = -\frac{1}{n^s}\sum_{i=1}^{n^s}\sum_{c=1}^{C} 1\{y_i^s = c\}\log\frac{e^{\theta_{s,c}^T \boldsymbol{v}_i^s}}{\sum_{j=1}^{C} e^{\theta_{s,j}^T \boldsymbol{v}_i^s}} + \frac{\gamma_s}{2}\|\theta_s\|^2 \quad (15)$$

where, γ_s is the hyper-parameter and $1\{\cdot\}$ denotes the indicator function, θ_s is the weight matrix of softmax regression. As mentioned above, ACFE has strong clustering ability. Therefore, training softmax regression should be quite easy and cost a little time. But if we also use SGD algorithm to optimize softmax regression loss, it is hard to select learning rate, because large learning rate will result in low accuracy and small learning rate will spend more time. Therefore, we adopt L-BFGS optimization algorithm for its ability of determining learning rate adaptively.

After two learning stages, we use target test samples to verify the proposed method. For each test sample, we transform it to the label space by the two trained map Φ and Φ_s. Then the health conditions of test samples are decided.

V. CASE STUDY I: FAULT DIAGNOSIS OF ROLLING BEARING

A. DATA DESCRIPTION

The validity of the method was verified by the experimental data of motor bearings provided by Case Western Reserve University Lab [33]. The experimental bench includes an induction motor, a torque sensor and a load motor. Test bearings are installed at the driving end of the induction motor. The load motor provides 0 hp, 1 hp, 2 hp and 3 hp load for the induction motor. In order to simulate the common faults of bearings, electrical discharge machining is used to process a single fault into the test bearings, including bearings with fault in inner ring(IF), rolling element (RF) and outer ring (OF) and with fault diameters of 0.1778 mm (7 mils), 0.3556 mm (14 mils) and 0.5334 mm (21 mils). The accelerometer is placed near the drive end of the motor housing and acquires vibration data at a sampling frequency of 12 kHz.

Before training the model, the data preprocessing procedure is completed. 1200 data points of time-domain signals are selected for FFT to get a sample, and the size of shift is 600. For each load, we obtained 20000 samples. For simulating the scenario of ACFE, four settings are designed.

(1) 1800 source domain labeled and balanced samples with load A hp consisting of ten health conditions are collected as the source training dataset.

(2) 1800 target domain unlabeled and imbalanced samples with load B hp consisting of ten health conditions are collected as the target training dataset. (3) The source testing dataset is consisted of 5000 balanced samples randomly selected from the rest source domain samples with load A hp.

(4) The target testing dataset is consisted of 5000 balanced samples randomly selected from the rest target domain samples with load B hp.

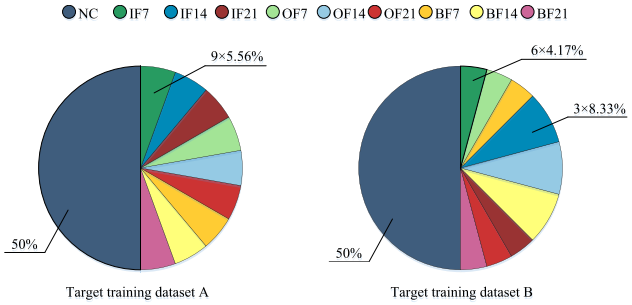


FIGURE 3. The imbalanced form of target training dataset.

We design two imbalanced forms of target training dataset. We consider that most samples are with normal condition. Besides, the collected and saved fault samples are usually with the medium level faults, because the incipient faults are difficult to detect, meanwhile, people do not allow the machine run with the major fault in general. The ratios of every health condition are shown in Figure 3. We can see that most samples are with normal condition and the samples with fault conditions are scarce.

B. SETUP OF PROPOSED MODEL

Each layer size of ACFE is [600, 400, 200, 100]. The learning rate is 0.001 and training step is set as 200. Every training batch contains 400 labeled data samples from the source domain and 400 unlabeled imbalanced data samples from target domain. The penalty parameters γ and γ_s are both 0.01.

In order to verify the effectiveness of the proposed entropy loss and guide loss, two contrastive forms of the ACFE are investigated for proving their effectiveness. ACFE without entropy loss and guide loss (ACFE-EG), ACFE only without guide loss (ACFE-G) are studied. The results are shown in Table 1 and Table 2. The results (r_1, r_2, r_3) denote the average accuracy, mid-value accuracy and standard deviation, respectively. It should be noticed that we experiment 10 times for each transfer condition and the highest accuracy and lowest accuracy are discarded to reduce the effects of the randomness. Besides, all the training accuracies are 100% so we do not show in the tables. For ACFE, we randomly selected an experiment and give out the changes of losses in Figure 4

In order to compare with the existing successful fault diagnosis method based on transfer learning, the results of generalization of deep neural network (GDN) reported in Ref. [19] are also shown in Table 1 and Table 2. The results (r_1, r_2) denote the average accuracy and standard deviation, respectively. It should be noticed that the target domain training datasets of GDN are balanced.

TABLE 1. Full transfer accuracy table training with target dataset A.

Source domain	Method	Target domain			
		Load 0	Load 1	Load 2	Load 3
Load 0	GDN	-	100.00%, 0.00%	100.00%, 0.00%	99.75%, 0.25%
	ACFE-EG		74.30%, 74.79%, 1.79%	74.32%, 74.59%, 1.84%	71.68%, 71.64%, 1.99%
	ACFE-G		99.30%, 99.35%, 0.37%	93.94%, 99.10%, 7.98%	93.81%, 98.50%, 8.25%
	ACFE		**99.35%,** 99.33%, 0.37%	**99.48%, 99.43%, 0.27%**	**99.09%, 98.89%, 0.63%**
Load 1	GDN	100.00%, 0.00%	-	100.00%, 0.00%	99.87%, 0.06%
	ACFE-EG	75.24%, 75.26%, 1.75%		72.47%, 72.53%, 1.77%	72.28%, 72.18%, 2.75%
	ACFE-G	99.01%, 99.00%, 0.33%		98.82%, 98.72%, 0.54%	93.71%, 99.16%, 7.86%
	ACFE	**99.51%, 99.66%, 0.41%**		**99.48%, 99.51%, 0.31%**	**99.32%, 99.34%, 0.46%**
Load 2	GDN	100%, 0.00%	100.00%, 0.00%	-	100.00%, 0.00%
	ACFE-EG	71.69%, 71.64%, 1.74%	73.69%, 73.49%, 1.26%		72.71%, 73.93%, 2.84%
	ACFE-G	93.82%, 98.16%, 8.07%	98.70%, 98.84%, 0.70%		94.56%, 99.27%, 7.16%
	ACFE	**99.51%, 99.55%, 0.38%**	**99.57%, 99.60%, 0.24%**		**99.73%, 99.79%, 0.29%**
Load 3	GDN	95.86%, 1.63%	96.95%, 1.02%	100.00%, 0.00%	-
	ACFE-EG	70.08%, 70.10%, 2.33%	74.62%, 74.99%, 1.23%	73.38%, 73.47%, 1.97%	
	ACFE-G	91.24%, 94.94%, 9.24%	93.70%, 97.96%, 7.18%	99.54%, 99.59%, 0.28%	
	ACFE	**99.08%, 98.97%, 0.34%**	**98.89%, 98.77%, 0.60%**	**99.63%, 99.62%, 0.25%**	

TABLE 2. Full transfer accuracy table training with target dataset B.

Source domain	Method	Target domain			
		Load 0	Load 1	Load 2	Load 3
Load 0	GDN	-	100.00%, 0.00%	100.00%, 0.00%	99.75%, 0.25%
	ACFE-EG		74.36%, 73.40%, 1.63%	74.09%, 74.30%, 2.06%	72.40%, 72.84%, 1.21%
	ACFE-G		99.27%, 99.31%, 0.30%	92.97%, 95.54%, 8.18%	93.10%, 96.02%, 7.69%
	ACFE		99.61%, 99.67%, 0.30%	99.19%, 99.11%, 0.37%	98.75%, 98.84%, 0.54%
Load 1	GDN	100.00%, 0.00%	-	100.00%, 0.00%	99.87%, 0.06%
	ACFE-EG	74.47%, 74.35%, 1.44%		72.64%, 72.80%, 1.28%	73.22%, 73.10%, 0.73%
	ACFE-G	99.07%, 98.82%, 0.53%		99.17%, 99.17%, 0.52%	92.91%, 96.89%, 8.07%
	ACFE	99.56%, 99.56%, 0.38%		99.61%, 99.64%, 0.30%	99.44%, 99.51%, 0.54%
Load 2	GDN	100%, 0.00%	100.00%, 0.00%	-	100.00%, 0.00%
	ACFE-EG	70.58%, 71.49%, 1.66%	73.03%, 72.05%, 1.95%		73.47%, 74.41%, 2.45%
	ACFE-G	91.94%, 96.03%, 9.32%	99.08%, 99.16%, 0.61%		92.93%, 95.23%, 7.62%
	ACFE	99.29%, 99.35%, 0.47%	99.58%, 99.59%, 0.29%		99.59%, 99.61%, 0.16%
Load 3	GDN	95.86%, 1.63%	96.95%, 1.02%	100.00%, 0.00%	-
	ACFE-EG	70.95%, 70.76%, 1.32%	73.87%, 73.31%, 2.41%	72.57%, 73.63%, 2.34%	
	ACFE-G	92.36%, 95.28%, 9.19%	92.41%, 94.85%, 7.91%	99.45%, 99.42%, 0.21%	
	ACFE	**98.45%, 98.42%, 0.76%**	**98.81%, 98.78%, 0.62%**	99.56%, 99.55%, 0.34%	

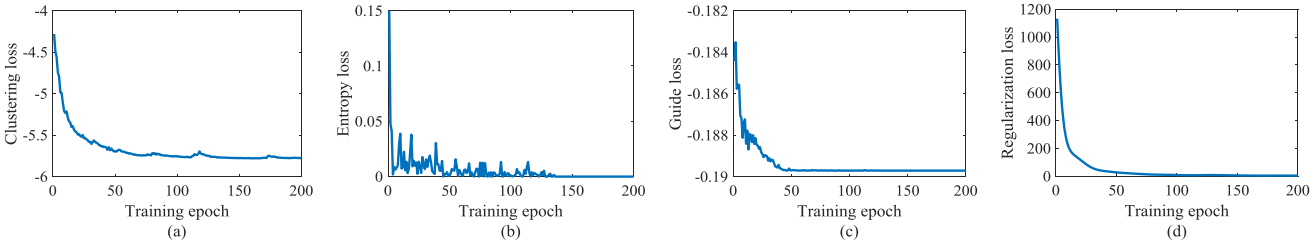


FIGURE 4. The changes of (a) clustering loss, (b) entropy loss, (c) guide loss and (d) regularization loss with the training epoch.

C. PROPOSED FRAMEWORKTESTING RESULTS AND DISCUSSION OF PROPOSED METHOD

As we can see in Table 1 and Table 2, if we only train the model with clustering loss and regularization loss (ACFE-EG), i.e. we don't adapt transfer learning, the target diagnosis accuracies of all transfer experiments are in the low accuracy. But the results of ACFE-E are worse than ACFE-EG. Because guide loss match with entropy loss for

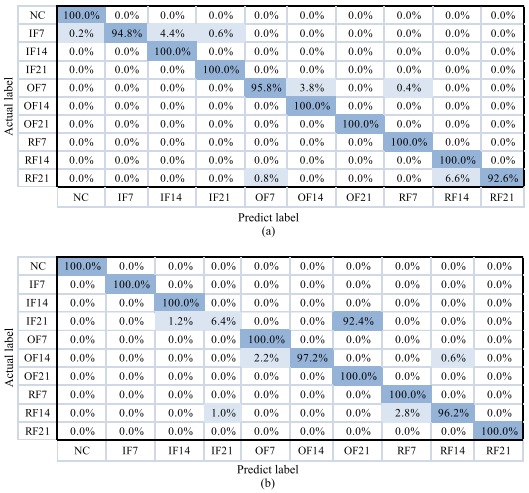


FIGURE 5. Confusion matrixes of (a) ACFE and (b) ACFE-G in the case of 3-0.

correcting the negative transfer. When we train ACFE-G, the diagnosis results are increasing. But as we can see from the results with underline, the low testing accuracy and high standard deviation mean that samples belonging to one category are integrally misclassified as another category. By comparing the results of ACFE-G and ACFE, we can draw a conclusion that it is necessary to employ the guide loss. Guide loss can effectively reduce the negative transfer. It is obvious that the accuracy of ACFE is the highest and the standard deviation is minimum when diagnose the two different degrees of imbalance datasets. This means that ACFE have the advantage for handling the imbalance of datasets.

Comparing with the existing successful method GDN, the performance of ACFE is satisfactory. ACFE trained with imbalanced target dataset obtains comparative testing accuracies as GDN trained with balanced target dataset. The results show that ACFE can improve the prediction accuracy, especially in 3-0 and 3-1, in which ACFE improves 8.68% and 6.88%. The standard deviations of ACFE have shown good results, and they are less than 1% in all cases. This means that the proposed ACFE method gets enormously successful. Besides, as we can see in Figure 4, all the losses decline obviously when the training epoch goes higher, which means that the proposed losses are trained effectively.

For further verifying the effectiveness of guide loss and showing more details about the diagnosis information, the confusion matrixes of ACFE and ACFE-G of dataset A in the case of 3-0 are presented in Figure 5. The testing accuracy of ACFE-G is 89.98% and the testing accuracy of ACFE is 98.32%. As shown in Figure 5(b), ACFE-G misclassifies nearly all the target testing samples of IF21 as OF21, which means that the ACFE-G is caught in negative transfer. After adding the guide loss, the negative transfer has been overcome, and there are a few samples misclassified as other similar faults as shown in Figure 5(a). This further verifies that guide loss can effectively control the label distribution to reduce the negative transfer.

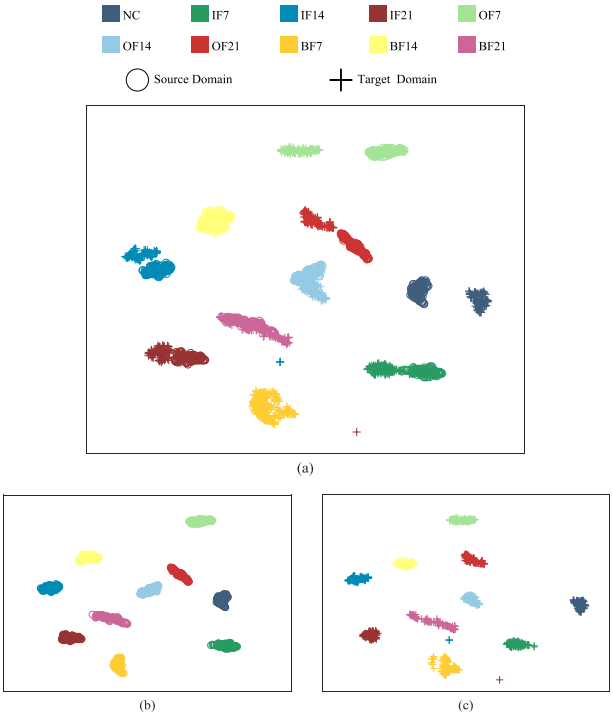


FIGURE 6. The t-SNE visualization of features: (a) All features, (b) Source domain features and (c) Target domain features.

To study these activation vectors extracted by ACFE obviously, we use t-distributed stochastic neighbor embedding (t-SNE) [32], [34]. We choose one result of transfer fault diagnosis experiment 3-0 of dataset A to study and plot in Figure 6. As we can see from Figure 6(a), features with the same health condition are grouped closer together even though they are under different working conditions. Figure 6(b) and (c) show that the distributions of source domain features and target domain features are basically consistent, which means that our ACFE achieves the goal of distribution adaptation but avoids employing distribution adaptation method.

As we can see from the above work, there is only two penalty parameters γ and γ_s which is uncertain. We investigate the impact of different parameters under the transfer condition 3-0 of dataset A. Figure 7 and Figure 8 shows the results. ACFE has strong robustness for selecting the penalty parameters γ and γ_s. All the results have little difference except γ and γ_s equal 0. Considering that the standard deviation is lower and the testing accuracies are a little higher than most others, we choose 1E-3 as the penalty parameters γ and γ_s.

In practice, selecting of kernel function is difficult. Therefore, we investigate the impact of different kernel function and select several kernel function for studying.

1) SIGMOID KERNEL (SK)

$$k(\boldsymbol{x}, \boldsymbol{y}) = \tanh(a\langle \boldsymbol{x}, \boldsymbol{y}\rangle + b) \tag{16}$$

The multiple kernel parameter a is set as [0.25, 0.5, 1, 2, 4], and b is 0.

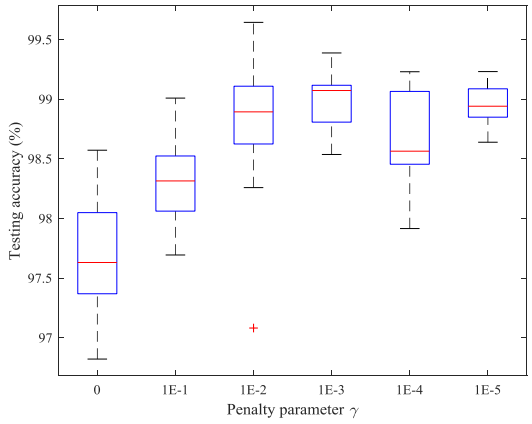


FIGURE 7. Boxplot using various γ.

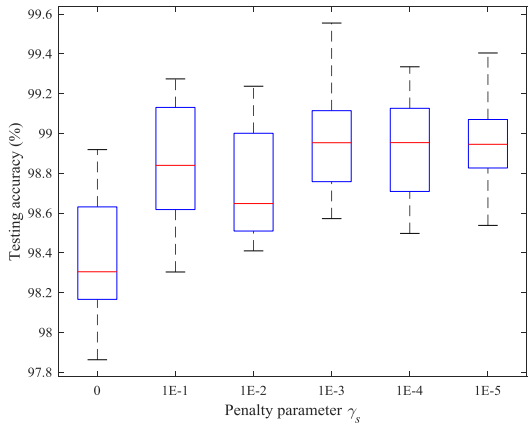


FIGURE 8. Boxplot using various γ_S.

2) RATIONAL QUADRATIC KERNEL (RQK)

$$k(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{\|\boldsymbol{x} - \boldsymbol{y}\|^2 + b} \tag{17}$$

We use this multiple kernel whose max-value is the mean of $||\boldsymbol{x}-\boldsymbol{y}||$ on the training data and decrease to half. For example, if the mean of $||\boldsymbol{x}-\boldsymbol{y}||$ is 1, the final b is set as [0.5, 0.625, 0.75, 0.875, 1].

3) CAUCHY KERNEL (CK)

$$k(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{s} + 1} \tag{18}$$

For Cauchy Kernel, we use the method of selecting parameter of RBF.

4) LOG KERNEL (LK)

$$k(\boldsymbol{x}, \boldsymbol{y}) = -\log\left(1 + \|\boldsymbol{x} - \boldsymbol{y}\|^a\right) \tag{19}$$

The multiple kernel parameter a is set as [0.5, 1, 1.5, 2, 2.5].

Considering the accuracies of transfer condition 0-3 and 3-0 are a litter lower than others, we investigate the impact of different kernel function under these condition. The results are shown in Figure 9. The results of RQK are lower than others and the accuracies are all below 98.5%. The performances of CK and LK are better. The accuracies are about 98.5%. For the two transfer condition (0-3 and 3-0) the results of same kernel function are different. But RBF obtain the highest accuracies and lower standard deviation, which means RBF is more suitable for our ACFE.

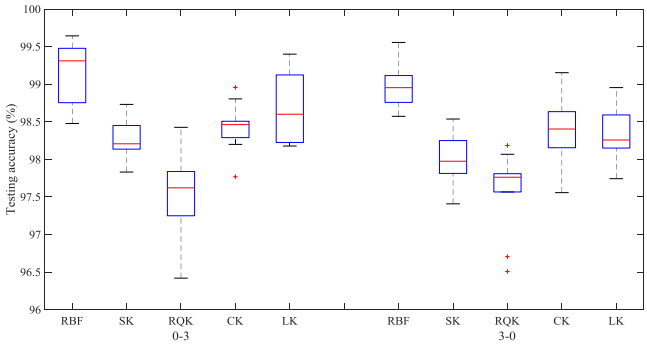


FIGURE 9. Boxplot using various kernel function.

D. COMPARING WITH RELATED WORK

To show the effectiveness of ACFE, we compare it with the methods in related work such as Transfer Component Analysis (TCA) [35], Joint Distribution Adaptation (JDA) [36], Balanced Distribution Adaptation (BDA) [37] and GDN [19] using the same rolling bearing dataset in section IV.A. The comparisons under transfer condition 3-0 of dataset A are displayed in Table 3.

For all the related works, each parameter is selected by random search in a large parameter range, and the suitable parameters are used for the final models. TCA, JDA and BDA both employ RBF kernel and one layer softmax regression for classification which is same as ACFE. For TCA, the best testing accuracies are only about 20% with 0.5 standard deviation. JDA obtains 82.04%±5.48% which is quite higher than TCA. For BDA, the accuracies of 83.67%±4.88% are obtained. For GDN, we employ the set of Ref. [19] and obtain the accuracies of 74.43%±3.46%. It should be noticed that, for the imbalanced target dataset domain adaptation, marginal distribution adaptation is difficult to achieve good results and joint distribution adaptation is suitable relatively. Compared with the methods above, the proposed ACFE is trained by imbalanced target dataset and obtains a higher and more stable accuracy.

VI. CASE STUDY II FAULT DIAGNOSIS UNDER DIFFERENT SPEED AND LOAD

A. DATA DESCRIPTION

A bearing fault dataset under time-varying rotational speed and loads is used to verify the proposed model. The bearing test bench is shown in Figure 10, which consists of a diesel engine, 5 bearing seats, 3 couplings, a breake disc and so on. The brake disc can provide artificial variable loads.

TABLE 3. Classification comparison of the rolling bearing dataset.

Method	Source testing dataset accuracy	Target testing dataset accuracy
TCA	99.27%, 99.21%, 0.3%	20.03%, 21.14%, 6.75%
JDA	99.39%, 99.41%, 0.25%	82.04%, 81.32%, 5.48%
BDA	99.29%, 99.26%, 0.39%	83.67%, 84.73%, 4.88%
GDN	100%, 100%, 0%	74.43%, 74.37%, 3.46%
ACFE	100%, 100%, 0%	99.08%, 99.07%, 0.27%

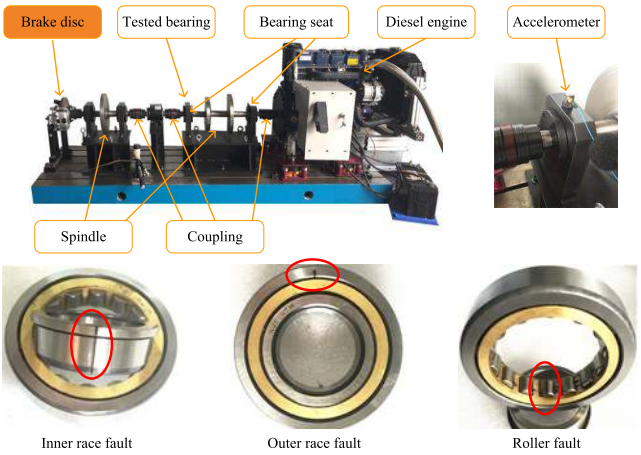


FIGURE 10. The arrangement of bearing test bench and bearing with fault.

TABLE 4. The description of the datasets.

Health condition	Number of target training samples	Number of source training samples
NC	1500	600
IF	300	600
RF	300	600
OF	300	600

The rotational speed range of diesel engine is 850rpm∼2600rpm. The datasets are collected at the tested bearing seats. The health condition includes four types: normal condition (NC), inner race fault (IF), roller fault (RF), outer race fault (OF). The sampling frequency used in the experiment is 12.8 kHz.

In this section, we test our ACFE on the transfer ability of rotating speed and load. Samples are selected under the rotating speed S1 (800rpm) and S2 (1000rpm), where, S1 is under a constant load by brake disc. 1200 data points of time-domain signals are selected for FFT to get a sample, and the size of shift is 600. As displayed in Table 4, the imbalance degree target datasets are collected.

B. TESTING RESULTS

In this section, we directly adopt the set of the rolling bearing case in section IV to test the robustness of ACFE. For other related works, each parameter is selected by random search in a large parameter range, and the suitable parameters are used for the final models. The methods based on the related works are detailed in section IV.D. The results are shown in Table 5.

TABLE 5. Target testing dataset accuracy.

Method	S1-S2	S2-S1
TCA	27.54%±7.03%	32.08%±6.75%
JDA	65.07%±4.93%	72.67%±4.90%
BDA	67.94%±5.01%	73.02%±4.77%
GDN	59.86%±4.25%	61.53%±5.86%
ACFE	**91.72%±0.57%**	**92.04%±0.63%**

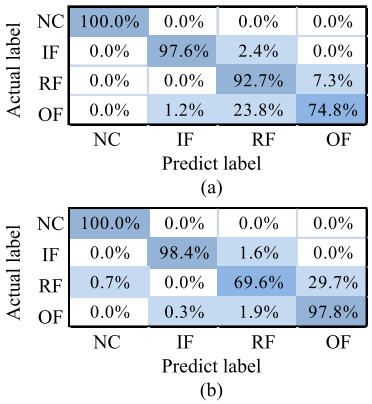


FIGURE 11. Confusion matrix of the ACFE prediction results in the case of (a) S1-S2 and (b) S2-S1.

As can be seen from the table, the performances of the five methods under the condition of S2-S1 are a little better than the performances under the condition of S1-S2. When the target training datasets are imbalanced, accuracies of methods based on marginal distribution adaptation (TCA and GDN) are lower than others. According to our investigations, when the standard deviation is 0.1 and $\gamma_s = 0.00001$, TCA obtains 27.54%±7.03% accuracies in the case of S1-S2 and 32.08%±6.75% accuracies in the case of S2-S1. For GDN, the 59.86%±4.25% classification accuracies of S1-S2 and 61.53%±5.86% classification accuracies of S2-S1 are achieved. When the standard deviation is 1 and $\gamma_s = 0.001$, JDA obtains 65.07%±4.93% accuracies of S1-S2 and 72.67%±4.90% accuracies of S2-S1. BDA achieves 67.94%±5.01% accuracies in the case of S1-S2 and 73.02%±4.77% accuracies in the case of S2-S1 when the three parameters are 0.9, 1 and 0.0001, respectively. Compared with the methods above, the proposed ACFE is more suitable for imbalanced datasets and obtains higher accuracies, which means that our ACFE is adapted for fault diagnosis under different speeds and loads successfully.

The diagnostic accuracy confusion matrixes of the ACFE method for the transfer fault diagnosis experiments S1-S2 and S2-S1 are obtained in Figure 11. The accuracy of S1-S2 is 91.28% and the accuracy of S2-S1 is 91.45%. As can be seen from Figure 11, the roller fault samples and outer race fault samples are more difficult to classify. For example,

ACFE misclassifies 23.8% testing samples of outer race fault samples as roller fault samples in the case of S1-S2. For S2-S1, ACFE misclassifies 29.7% testing samples of roller fault as outer race fault samples. The reason may be that the feature of them is similar, which makes it more difficult to distinguish them.

VII. CONCLUSION

The problem of imbalanced dataset is actual and important. In this paper, we proposed an adaptive cross-domain feature extraction method, which has strong abilities of clustering and automatically extracting cross-domain feature. This paper studied the effect of proposed entropy loss and guide loss based on the calculated posterior probability of category. The results proved that they are effective and necessary. However, in the process of study, we find that, in rare circumstances, ACFE also has the performance of negative transfer, even though the guide loss is employed. Besides, target domain training dataset sometimes lacks several categories, i.e. the label spaces of source domain and target domain are different. How to diagnose the fault under this condition is a more difficult task. The authors would investigate these topics in the future study.

REFERENCES

[1] L. Wang, Z. Liu, Q. Miao, and X. Zhang, "Time-frequency analysis based on ensemble local mean decomposition and fast kurtogram for rotating machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 103, pp. 60–75, Mar. 2018.

[2] C. Zhao and F. Gao, "Fault subspace selection approach combined with analysis of relative changes for reconstruction modeling and multifault diagnosis," *IEEE Trans. Control Syst. Technol.*, vol. 24, no. 3, pp. 928–939, May 2016.

[3] L. Dong, S. Liu, and H. Zhang, "A method of anomaly detection and fault diagnosis with online adaptive learning under small training samples," *Pattern Recognit.*, vol. 64, pp. 374–385, Apr. 2017.

[4] S. Haidong, J. Hongkai, L. Xingqiu, and W. Shuaipeng, "Intelligent fault diagnosis of rolling bearing using deep wavelet auto-encoder with extreme learning machine," *Knowl.-Based Syst.*, vol. 140, pp. 1–14, Jan. 2018.

[5] Y. Wang, S. Yabin, and H. Biao, "Survey on the theoretical research and engineering applications of multivariate statistics process monitoring algorithms: 2008–2017," *Can. J. Chem. Eng.*, vol. 96, no. 10, pp. 167–179, 2018.

[6] B. A. Paya, I. Esat, and M. Badi, "Artificial neural networks based fault diagnosis of rotating machinery using wavelet transforms as a preprocessor," *Mech. Syst. Signal Process.*, vol. 11, no. 5, pp. 751–765, 1997.

[7] J. Rafiee, F. Arvani, A. Harifi, and M. H. Sadeghi, "Intelligent condition monitoring of a gearbox using artificial neural network," *Mech. Syst. Signal Process.*, vol. 21, no. 4, pp. 1746–1754, May 2007.

[8] 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.

[9] H. Shao, H. Jiang, Y. Lin, and X. Li, "A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders," *Mech. Syst. Signal Process.*, vol. 102, pp. 278–297, Mar. 2018.

[10] L. Liao, W. Jin, and R. Pavel, "Enhanced restricted boltzmann machine with prognosability regularization for prognostics and health assessment," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7076–7083, Nov. 2016.

[11] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, Sep. 2016.

[12] S. Wang, J. Xiang, Y. Zhong, and Y. Zhou, "Convolutional neural network-based hidden Markov models for rolling element bearing fault identification," *Knowl.-Based Syst.*, vol. 144, pp. 65–76, Mar. 2018.

[13] 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.

[14] Z. An, S. Li, W. Qian, Q. Wu, and J. Wang, "An intelligent fault diagnosis approach considering the elimination of the weight matrix multi-correlation," *APPL SCI-BASEL*, vol. 8, no. 6, p. 906, 2018.

[15] Y. Lei and M. J. Zuo, "Gear crack level identification based on weighted K nearest neighbor classification algorithm," *Mech. Syst. Signal Process.*, vol. 23, no. 5, pp. 1535–1547, 2009.

[16] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, Sep. 2018.

[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.*, vol. 49, no. 1, pp. 136–144, Jan. 2017.

[19] Z. An, S. Li, J. Wang, Y. Xin, and K. Xu, "Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method," *Neurocomputing*, vol. 352, pp. 42–53, Aug. 2019.

[20] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 281–288, Feb. 2009.

[21] H. Salehinejad and S. Rahnamayan, "Customer shopping pattern prediction: A recurrent neural network approach," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Athens, Greece, Dec. 2016, pp. 1–6.

[22] F. Jia, Y. Lei, N. Lu, and S. Xing, "Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization," *Mech. Syst. Signal Process.*, vol. 110, pp. 349–367, Sep. 2018.

[23] Z. Ding, N. M. Nasrabadi, and Y. Fu, "Semi-supervised deep domain adaptation via coupled neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5214–5224, Nov. 2018.

[24] M. Yang, W. Tu, Z. Zhao, X. Chen, J. Zhu, and Q. Qu, "Personalized response generation by Dual-learning based domain adaptation," *Neural Netw.*, vol. 103, pp. 72–82, Jul. 2018.

[25] C. E. Shannon, W. Weaver, and N. Wiener, "The mathematical theory of communication," *Phys. Today*, vol. 3, no. 9, pp. 31–32, 1950.

[26] S. P. Strong, R. Koberle, R. R. de R. van Steveninck, and W. Bialek, "Entropy and information in neural spike trains," *Phys. Rev. Lett.*, vol. 80, no. 1, pp. 197–200, 1998.

[27] J. A. Núñez, P. M. Cincotta, and F. C. Wachlin, "Information entropy," *Celestial Mech., Dyn. Astron.*, vol. 64, nos. 1–2, pp. 43–53, 1996.

[28] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, 2003.

[29] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, and N. Zhang, "Deep domain confusion: Maximizing for domain invariance," Dec. 2014, *arXiv:1412.3474*. [Online]. Available: https://arxiv.org/abs/1412.3474

[30] H. Venkateswara, J. Eusebio, S. Panchanathan, and S. Chakraborty, "Deep hashing network for unsupervised domain adaptation," Jun. 2017, *arXiv:1706.07522*. [Online]. Available: https://arxiv.org/abs/1706.07522

[31] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, Mar. 2012.

[32] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.

[33] K. Loparo. *Case Western Reserve University Bearing Data Center*. Accessed: Jul. 15, 2013. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages/12k-drive-end-bearing-fault-data.

[34] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 2605, pp. 2579–2605, Nov. 2008.

[35] 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.

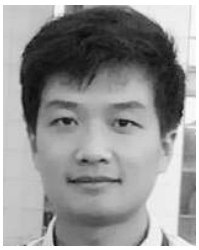
[36] M. Long, J. Wang, J. Sun, P. S. Yu, and G. Ding, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, vol. 2013, pp. 2200–2207.
[37] J. Wang, Y. Chen, W. Feng, Z. Shen, and S. Hao, "Balanced distribution adaptation for transfer learning," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, New Orleans, LA, USA, Nov. 2017, pp. 1129–1134.

ZENGHUI AN received the B.S. and M.S. degrees from the University of Jinan, Jinan, China, in 2013 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 (NUAA), Nanjing, China. His current research interests include mechanical fault diagnosis and deep 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 (NUAA), 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.

XINGXING JIANG received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. He is currently a Postdoctoral Researcher with Soochow University, Suzhou, China. His current research interests include rotating machine fault diagnosis, and mechanical signal and information processing.

YU XIN is currently pursuing the Ph.D. degree in aerospace propulsion theory and engineering with the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include fault diagnosis of rotor machine and signal processing.



JINRUI WANG received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. He is currently a Postdoctoral Researcher with the Shandong University of Science and Technology, Qingdao, China. His current research interests include rotating machine fault diagnosis, and mechanical signal and information processing.