Rolling Bearing Fault Diagnosis under Variable Working Conditions Based on Joint Distribution Adaptation and SVM

Ming Li[1], Zhao-Hui Sun[1], Weihui He[1,2], Siqi Qiu[1] and Bo Liu[1]
[1]:Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai 200240, China.
[2]:Xi'an Satellite Control Center, Xi'an, Shaanxi 710043, China.
nono-MAA@sjtu.edu.cn, zh.sun@sjtu.edu.cn, 18817558023@163.com, siqiqiu1988@163.com, lbws888@sjtu.edu.cn.

***Abstract*—The traditional fault diagnosis methods for rolling bearing usually require the test data and training data to follow the same distribution, which cannot be always meet in real-world scenarios, since the working condition of rolling bearing is often variable. Hence, to overcome the low performance of fault diagnosis traditional methods for different data distributions, a fault diagnosis approach based on transfer learning is proposed in this paper. And the main idea of our approach is to combine joint distribution adaptation and support vector machine to diagnose bearing faults under variable working conditions. In this research, kernel-JDA is used to reduce the difference between distributions of datasets taking both the marginal and conditional distributions into consideration, while the parameters of kernel-JDA are optimized to improve the performance. Besides, multi-features including time domain features and the relative wavelet packet energy are constructed at first to prepare for fault diagnosis. After mapping the multi-features through kernel-JDA, SVM is utilized to diagnose faults of rolling bearing under different working conditions. In addition, comparison experiments on vibration signal datasets of rolling bearings are carried out to verify the effectiveness and applicability of this approach for both the normal and small sizes of the sample sets.**

***Index Terms*—Fault diagnosis, Transfer learning, Joint distribution adaptation, Wavelet packet, Support vector machine.**

I. Introduction

The rolling bearing is a kind of important transmission components, which is widely used in wind turbines, helicopters, cranes, tanks and other important equipment. Under the poor working conditions, the rolling bearing is prone to failure, affecting the function of the whole machine. So, it is significant to study the bearing fault diagnosis method.

Over the past decades, bearing fault diagnosis based on the vibration signal is a kind of the most extensively studied methods. Then, many traditional machine learning algorithms have been successfully implemented in fault diagnosis, such as support vector machine (SVM), k-nearest neighbor (KNN), fuzzy inferences and etc. [1-4]. In addition, deep learning has also made great achievements in bearing fault diagnosis [5]. Jing et al. [6] have applied convolutional neural network (CNN) to directly learn features from vibration data and intelligently identify gearbox failures. Shao et al. [7] have proposed a bearing fault diagnosis method based on adaptive deep belief network (DBN). Yang et al. [8] have verified the effectiveness of long short-term memory (LSTM) in fault diagnosis of rotating machinery through a large number of comparative experiments. However, those studies are mainly focused on fault diagnosis for certain working conditions, where the classifier may not be effective in diagnose the data from other working conditions. Besides, the working conditions of the rolling bearings often change in actual projects. Hence, there have been researchers studied on fault diagnosis for various working conditions in past few years.

For example, Liu et al. [9] have utilized recurrent neural network (RNN) to diagnose rolling bearing fault under various conditions based on the features extracted with Hilbert-Huang transform and singular value decomposition. Fei et al. [10] has studied fault diagnosis of bearing under varying load conditions by utilizing adaptive feature selection method. However, the performance of the classifier in those studies can be degraded, when the characteristics of the training data and testing data are inconsistent [11]. To solve this problem, transfer learning is introduced into fault diagnosis in recent years.

The aim of transfer learning (TF) is to take advantage of experience learnt in one task and improve the performance in a similar but different task [12]. Recently, there are mainly 3 categories of methodologies that are widely studied and applied for transfer learning: instance based transfer learning, parameter/model based transfer learning and feature based transfer learning [11]. The methods of instance based TF include TrAdaBoost algorithm which increases the proportion of samples that can improve the accuracy of the target domain classification[13], the method which estimates distribution by kernel mean matching (KMM) and make the distribution of the target domain and source domain data closer through weight adjustment[14], and etc. Model based TF means the transfer learning by finding the parameters that can be shared by the model between the source and target domain. The representative studies on model based TF include the TransEMDT method designed by Zhao et al. [15], the methods introduced by Long et al. [16-18] to enhance the generalization ability of deep neural networks by introducing probability distribution and network fine-tuning, and etc. Feature based TF aims at narrowing the difference of data distribution between source domain and target domain [19-21]. The widely used methods include transfer component analysis (TCA) which

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

minimizes the difference between marginal distributions, joint distribution adaptation (JDA) which consider the difference between marginal distributions and that between conditional distributions at the same time, and etc [19-22]. Transfer learning offers a good idea for fault diagnosis under various situations, getting more universal diagnosis models.

As for the research on fault diagnosis of bearing based on transfer learning, there are a lot of studies shrinking the difference between marginal distributions [23-24]. Those studies have not considered adapting the conditional distributions explicitly, while there are situations that the difference between conditional distributions is much higher than that of marginal distributions, making a great effect on the performance of TF. Besides, parameter/model based TF has also been utilized for fault diagnosis. For example, Zhong [25] has utilized fine-tuning to realize transfer learning of CNN for fault diagnosis. Li et al. [26] have also demonstrated that CNN transfer learning could improve the accuracy of target domain fault diagnosis in different gearboxes and under different working conditions. Although those methods have worked well, they can only use a part of the source datasets whose distance from the target distribution is relatively small, leaving some source data wasted. Therefore, in order to make the best use of source data, JDA and SVM are combined to develop an effective method for rolling bearing fault diagnosis in various working conditions in this research.

In our study, to prepare for fault diagnosis, multi-features including time domain features and the relative wavelet packet energy are constructed, dealing with the non-stationary property and abrupt changes of the rolling bearing vibration signals. Besides, in order to improve the performance of JDA-SVM transfer learning, radial basis function (RBF) kernel is adopted in JDA and the parameters of RBF-JDA are optimized by grid search. Furthermore, the applicability of the proposed method in the situations with normal target sample size and small target sample size is validated through the experiments. The rest of the paper are organized as follows. In Section II, the principle of classifier based on optimized kernel-JDA-SVM is described. In Section III, the transfer learning fault diagnosis method applied for bearing in various working conditions is introduced. Experiment study and performance analysis using public datasets are carried out in Section IV. Finally, the conclusion is drawn in Section V.

II. THE PRINCIPLE OF CLASSIFIER BASED ON OPTIMIZED JDA-SVM

A. The Principle of JDA

JDA was first proposed by Long [22]. The main idea is to complete the transfer learning by reducing the distance of the joint probability distribution of the source and target domains. The joint probability distribution consists of two parts, marginal probability distribution and conditional probability distribution. The marginal probability distribution represents the overall difference between the target domain and the source domain, for example, the distance between classes in the two domains is different. The conditional probability distribution represents the difference between each class in the two domains, such as the difference in the distribution of similar samples in the two domains. JDA considers both the marginal probability distribution and the conditional probability distribution, further reducing the difference between the data distributions of the two domains.

In this research, the data in source domain is rich and the corresponding states of bearing are known, while the data in target domain has no labels or the amount of it is small. So joint distribution adaptation is utilized to make the great amount of source data which are collected in a known working condition more useful to train a effective classifier which can recognize the faults with target data in another working condition.

Given the labeled source data$\{(x_1, y_1), ..., (x_{n_s}), y_{n_s})\}$ and unlabeled target data $\{x_{n_s+1}, ..., x_n\}$, where $n = n_s + n_t$ is the total number of samples, n_s is the number of samples in source domain D_s and n_t that in target domain D_t. The implementation of joint distribution adaptation includes three steps:

1) STEP1: Marginal Distribution Adaptation (MDA) :

Utilize the maximum mean discrepancy(MMD) to compare the distributions of the source and target data. MMD of MDA is defined as MMD_0, which can be calculated by

$$\begin{aligned} MMD_0 &= ||\frac{1}{n_s}\sum\nolimits_{i=1}^{n_s} A^T x_i - \frac{1}{n_t}\sum\nolimits_{j=n_s+1}^{n_s+n_t} A^T x_j||^2 \\ &= tr(A^T X M_0 X^T A) \end{aligned} \tag{1}$$

where A is the transfer matrix, $X = \{x_1, ..., x_n\}$ and M_0 is the MMD matrix of MDA whose formula is

$$(M_0)_{ij} = \begin{cases} \frac{1}{n_s^2}, & x_i, x_j \in D_s \\ \frac{1}{n_t^2}, & x_i, x_j \in D_t \\ -\frac{1}{n_s n_t}, & otherwise \end{cases} \tag{2}$$

The difference between the marginal distributions of the source data and target data can be narrowed by minimize MMD in Eq.1.

2) STEP2: Conditional Distribution Adaptation (CDA) :

Considering the conditional distributions can be also different between D_s and D_t, conditional distribution adaptation is necessary for robust distribution adaptation. MMD can also be used to measure the difference of conditional distribution by computing the distance of each class. The MMD of CDA is defined as MMD_c, whose formula is

$$\begin{aligned} MMD_c &= ||\frac{1}{n_{s,c}}\sum_{x_i \in D_{s,c}} A^T x_i - \frac{1}{n_{t,c}}\sum_{x_j \in D_{t,c}} A^T x_j||^2 \\ &= tr(A^T X M_c X^T A)(c = 1, ..., C) \end{aligned} \tag{3}$$

where $n_{s,c}$ and $n_{t,c}$ are the numbers of the samples of class c, $D_{s,c} = \{x_i | x_i \in D_s \cap y_i = c\}$ is the sample sets of class c in the source domain, in which y_i is the known label of x_i. Similarly, $D_{t,c} = \{x_j | x_j \in D_t \cap y_j = c\}$ is the sample sets of class c in the target domain, in which however y_j is not

given. So, the pseudo label $\hat{y}(x_j)$ of x_j which is predicted by the classifier trained by the labeled source data is utilized to substitute y_j [22]. Thus, $D_{t,c} = \{x_j | x_j \in D_t \cap \hat{y}(x_j) = c\}$. M_c is the MMD matrix of CDA for class c, whose formula is

$$(M_c)_{ij} = \begin{cases} \frac{1}{n_{s,c}^2}, & x_i, x_j \in D_{s,c} \\ \frac{1}{n_{t,c}^2}, & x_i, x_j \in D_{t,c} \\ -\frac{1}{n_{s,c} n_{t,c}}, & \begin{cases} x_i \in D_{s,c}, x_j \in D_{t,c} \\ x_j \in D_{s,c}, x_i \in D_{t,c} \end{cases} \\ 0, & otherwise \end{cases} \tag{4}$$

Then sum up the $MMD_c s$ of all classes, getting the object function for conditional distribution adaption:

$$\sum_{c=1}^{C} tr(A^T X M_c X^T A) \tag{5}$$

The difference between the conditional distributions of the source data and target data can be narrowed by minimize Eq.5.

3) STEP3: Joint Distribution Adaptation: Add Eq.1 and Eq.5 together to get the general optimization goal:

$$min \sum_{c=0}^{C} tr(A^T X M_c X^T X) + \lambda ||A||_F^2 \tag{6}$$

where, $\lambda ||A||_F^2$ is the regulation term and λ is the regulation coefficient.

Besides, considering the variance of the data after the transformation should not be changed, a constraint term is added:

$$\begin{aligned} min \sum_{c=0}^{C} tr(A^T X M_c X^T A) + \lambda ||A||_F^2 \\ s.t. A^T X H X^T A = I \end{aligned} \tag{7}$$

where, $H = I - \frac{1}{n}\mathbf{1}$ is the centering matrix and I is the $n * n$ unit matrix. Then use eigen decomposition to solve it:

$$(X \sum_{c=0}^{C} M_c X^T + \lambda I) A = X H X^T A \phi \tag{8}$$

where, $\phi = diag(\phi_1, ..., \phi_k) \in R^{k*k}$ are the k largest eigenvalues. Thus, the transform matrix A can be get. For nonlinear problems, long et al. [22] adopt kernel mapping to JDA:

$$\begin{aligned} min \sum_{c=0}^{C} tr(A^T K M_c K^T A) + \lambda ||A||_F^2 \\ s.t. A^T K H K^T A = I \end{aligned} \tag{9}$$

where $K = \psi(X)^T \psi(X) \in R^{n*n}$ is the kernel matrix.

B. The principle of kernal-JDA-SVM

The main idea of JDA-SVM is to use JDA to make the sample distributions of source and target domains closer, and then the SVM classifier is trained using the mapped source samples to obtain better classification results for the target samples. The algorithm can be described as:

1) STEP1: Only adapt the marginal distributions and then utilize the mapped source samples to obtain the initial SVM classifier $M_{svm}^{(0)}$. Use this classifier $M_{svm}^{(0)}$ to obtain the classification results of the unlabeled target domain samples, and set them as the initial pseudo labels $\hat{y}^{(0)}(x_j)$.

2) STEP2: Iteration

a) Set $T_{max} = m$ for termination criterion, and the initial iteration $T = 1$.

b) Calculate the MMD matrix of MDA by Eq.2 and utilize the pseudo label to calculate the MMD matrices of CDA by Eq.4.

c) Add the MMD matrix of MDA and those of CDA together to get $\sum_{c=0}^{C} M_c$.

d) Solve the general optimization problem in Eq.9 of kernel-JDA by

$$(X \sum_{c=0}^{C} M_c X K^T + \lambda I) A = K H X^T A \phi \tag{10}$$

Then get the transform matrix A and the mapped sample matrix $Z = A^T K$.

e) Utilize the mapped sample sets of source domain Z_s to train an SVM classifier $M_{svm}^{(T)}$. In this research, the SVM with RBF kernel is used and its penalty term and the bandwidth of RBF are optimized by k-folds cross validation and grid search.

f) Use the classifier to get the predicted labels $\hat{y}^{(T)}(x_j)$ and the accuracy $Acc^{(T)}$ with the mapped sample sets Z_t of target domain.

g) Identify whether the iteration T meets the termination criterion $T_m = m$. If not, set $\hat{y}^{(T)}(x_j)$ as the new pseudo label and then return to *b)*. If so, break the loop.

From these steps, the predicted labels would be closer to the real labels, resulting a more accurate classifier for target domain.

C. The Optimized kernel-JDA-SVM Based on Grid Search

The general idea is to use grid search to optimize the regulation coefficient λ in Eq.9 and the parameter of the JDA kernel. RBF kernel is used as the JDA kernel in this research, whose formula is

$$K(x, y) = e^{-\gamma ||x-y||^2} \tag{11}$$

where γ is the term that needs to be optimized.

The algorithm can be described as:

1) STEP1: Determine the grid for parameters optimization. For example, set the grid of $log_2 \lambda$ as $[-5 : 1 : 5]$ and the grid of $log_2 \gamma$ as $[-5 : 0.5 : 5]$, where 1 and 0.5 are the intervals while $[-5, 5]$ is the range.

2) STEP2: Use each group of parameters for kernal-JDA-SVM training to obtain the target domain data classification accuracy rate $Acc_{\lambda,\gamma}$ of the classifier $M_{svm}^{\lambda,\gamma}$.

3) STEP3: Search for the maximum $Acc_{\lambda,\gamma}$. The corresponding parameters λ_b, γ_b of maximum $Acc_{\lambda,\gamma}$ are the optimized parameters. And the $M_{svm}^{\lambda_b,\gamma_b}$ is the optimized classifier for target samples.

Thus, the parameters of kernel-JDA-SVM are optimized automatically, improving the performance of kernel-JDA-SVM classifier for samples in target domain.

III. DESIGN OF TRANSFER LEARNING BASED FAULT DIAGNOSIS FOR BEARING IN VARIABLE WORKING CONDITIONS

The general framework of the transfer learning based fault diagnosis for bearing in variable working conditions is shown in Fig.1. It mainly contains two parts, construction of multi-features and transfer learning of fault classifier based on optimized kernel-JDA-SVM.

A. Construction of Multi-features

Multi-state vibration signals (including normal and multiple faults) of rolling bearing with various working conditions are pre-processed. Features are extracted from the time domain and wavelet packets to form the multi-features set.

The features extracted from time domain include root mean square (RMS), crest factor (CF), shape factor (SF), impulse factor (IF), margin factor (MF), kurtosis factor (KF), and skewness (SN). Given the vibration signal $x(i)$, then the formulas of the time domain features can be expressed as those Table I shows.

TABLE I
THE FORMULAS OF TIME DOMAIN FEATURES

Feature	Formula
RMS	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}x^2(i)}$
CF	$\frac{max\|x_i\|}{RMS}$
SF	$\frac{RMS}{\frac{1}{N}\sum_{i=1}^{N}\|x(i)\|}$
IF	$\frac{max\|x_i\|}{\frac{1}{N}\sum_{i=1}^{N}\|x(i)\|}$
MF	$\frac{max\|x_i\|}{(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\|x(i)\|})^2}$
KF	$\frac{\frac{1}{N}\sum_{i=1}^{N}(x(i)-\mu)^4}{S^4}$
SN	$\frac{1}{N}\sum_{i=1}^{N}(x(i)-\mu)^3$

$^{a\rightarrow}$. $\mu = \frac{1}{N}\sum_{i=1}^{N}x(i)$ is the mean and $S = \sqrt{1/(N-1)\sum_{i=1}^{N}(x(i)-\mu)^2}$ is the standard deviation of the signal.

In addition, wavelet packet analysis is utilized in this research to divide frequency bands and get the characteristic energy of the bands, improving the time-frequency resolution. Wavelet packet decomposition filters the analysis signal through a series of low-pass and high-pass filters, and decomposes the signal into independent frequency band sub-signals in the form of nodes of a binary tree [27]. After performing $l-layers$ wavelet packet decomposition on the original vibration signal, the decomposed signal can be expressed as

$$x(t)=\sum_{j=i}^{2^l}\sum_{k}x_l^j(k)u_l^j(k,t) \tag{12}$$

where $x_l^j(k)(k=1,2,...,N)$ is the wavelet packet coefficient of the node (l,j), $u_l^j(k,t)$ is the orthogonal wavelet basis. The energy at the wavelet packet node (l,j) is

$$E_l^j=\sum_{k=1}^{N}(x_l^j(k))^2 \tag{13}$$

The relative wavelet packet energy is

$$AE_l^j=\frac{E_l^j}{\sum_{j=1}^{2^l}E_l^j} \tag{14}$$

Then, 2^l relative wavelet packet energy can be obtained, expressed as $AE_l=(AE_l^1,AE_l^2,...,AE_l^{2^l})$. AE_l can be used to analyze the energy of each frequency band, which makes a great difference in analysis of non-stationary, nonlinear vibration, such as that of rolling bearing. So, select a suitable number of wavelet packet decomposition layers according to the natural frequency of the vibration acceleration sensor and the sampling frequency. In this study, 3 layers of decomposition are used to extract the features of the 8 components from the low frequency to the high frequency of the third layer.

Finally, 15 features are constructed from the vibration signals of bearings recorded as F.

$$F=(RMS,CF,SF,IF,MF,KF,SN,AE_3^1,AE_3^2,...,AE_3^8)^T \tag{15}$$

In this study, the labeled vibration signals of different bearing states in known working conditions are defined as the data in source domain D_s, and the vibration signals of different bearing states in other working conditions are defined as the data in target domain D_t. For n_s samples from D_s, the matrix of the multiple features is

$$M_{F_s}=(F_1,...,F_{n_s}) \tag{16}$$

Similarly, the matrix of the multiple features in D_t is

$$M_{F_t}=(F_1,...,F_{n_t}) \tag{17}$$

In addition, the multiple features in both domains are normalized to eliminate dimensional effects between features. The final matrices of the normalized multiple features are

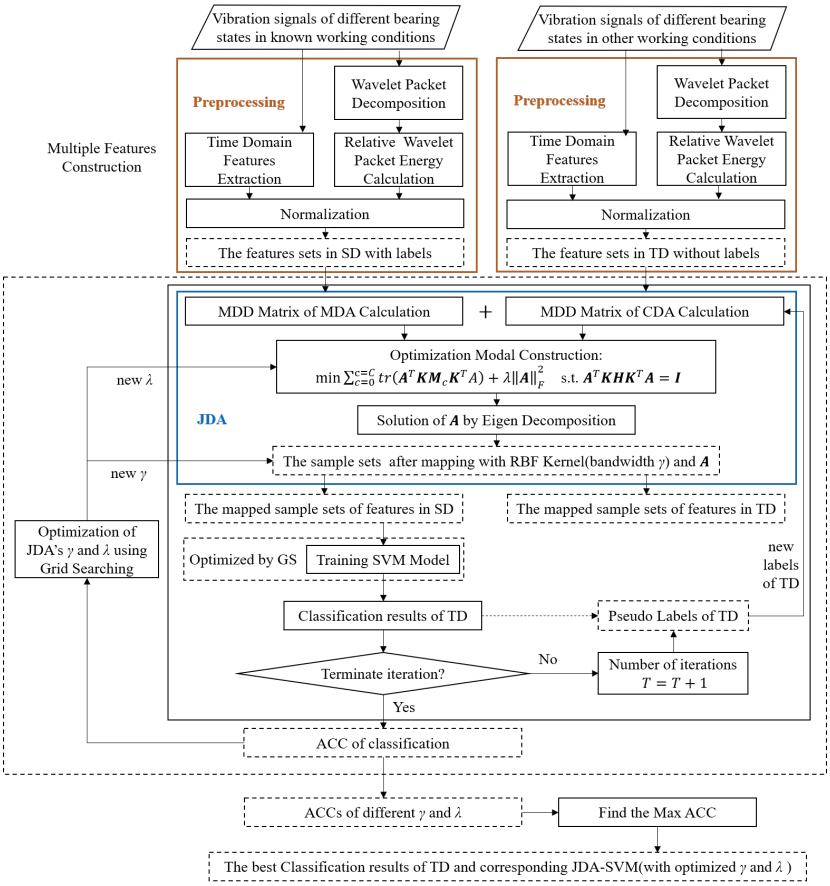


Fig. 1. The framework of fault diagnosis based on the optimized kernel-JDA-SVM for bearing in variable working conditions.

$$M'_{F_s} = (F'_1, ..., F'_{n_s}) \tag{18}$$

$$M'_{F_t} = (F'_1, ..., F'_{n_t}) \tag{19}$$

B. Transfer Learning of Fault Classifier based on Optimized kernel-JDA-SVM

Take the normalized multiple features obtained from Eq.18 and Eq.19 as the inputs of optimized kernel-JDA-SVM. Then utilize kernel-JDA to build the joint distribution MMD optimization model, and solve the transform matrix A. Then, use the mapped sample sets of the normalized multiple features with labels in source domain as the input to train SVM, while use grid search and k-folds cross validation to optimize the cost coefficient and the parameter of Gaussian kernel, getting a fault classifier. In addition, grid search is also used to optimize the values of parameters, the regulation coefficient λ and the RBF parameter γ, of kernel-JDA. Thus, different λs and γs in a certain range are tried in kernel-JDA-SVM to find the λ and γ with which the classifier is of the best performance on target domain. Finally, through the optimized kernel-JDA-SVM, the classifier with a relative high accuracy of fault diagnosis for bearings in unknown working conditions can be got.

IV. EXPERIMENT STUDY AND PERFORMANCE ANALYSIS

A. Datasets Description

In this research, the bearing data sets from Western Reserve University Bearing Data Center are selected to investigate the applicability of the presented approach [28]. The data was acquired from the bearing test system, which is shown in Fig. 2. The vibration signals were collected by a 16 channel DAT recorder, and the sampling frequencies contained 12 kHz and 48 kHz.

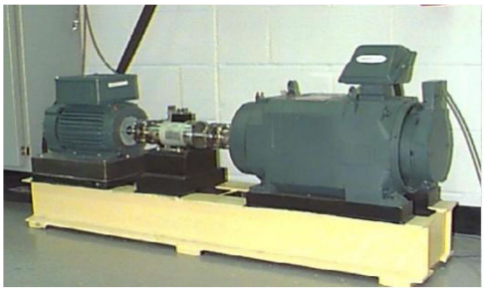


Fig. 2. The bearing test system at Western Reserve University.

The vibration signals of the rolling bearing at the drive end with 48 kHz sampling frequency are analyzed in this study.

3 different fault diameters and 4 different working conditions were carried out in the testing. Different test conditions are given in Table II. In this paper, the fault types contain ball defects of 3 diameters and inner race defects of 3 diameters. Thus, adding the normal state of bearing, there are 7 classes of the bearing status.

TABLE II
THE TEST CONDITIONS

Fault Diameter		0.007"	0.014"	0.021"
Working Conditions	**A**	0hp, 1797r/min		
	B	1hp, 1772r/min		
	C	2hp, 1750r/min		
	D	3hp, 1730r/min		

Comparison experiments are designed based on the sample sets of 7 types of state and under 4 kinds of working conditions. Fault diagnosis tests based on the optimized kernel-JDA-SVM and those based on direct SVM, which means training the SVM classifier with the unmapped target data, are carried out to demonstrates the validity of the presented approach. In this research, the data under the working condition C is used as the source samples, while the data under the working condition A\D is utilized as the target samples.

B. Preprocessing Results

In this research, 4096 vibration signal points are taken per sample. For each sample, elimination of mean value and other preprocessing are carried out before extracting the multiple features from the vibration signals. After extracting 15 features from each sample, the features from the C samples and those from A\D samples are normalized respectively. Fig.3 shows a part of normalized features extracted from the samples under the working condition C. It can be seen that the multi-features $F = (RMS, CF, SF, IF, MF, KF, SN, AE_3^1, AE_3^2, ..., AE_3^8)^T$ can reflect the 7 different states of rolling bearing effectively.

C. Performance Analysis of Optimized kernel-JDA-SVM Fault Diagnosis for Target Sample Sets with Normal Size

In the experiments, grid search and the 5-folds cross validation are used to optimized the parameters of SVM classifier. Other necessary parameters for JDA-SVM transfer learning are set as follows: **1)** The termination criterion for iterations of JDA $T = 15$; **2)** The dimension of mapped features $dim = 15$; **3)** The initial grid of optimized kernel-JDA-SVM $\{\lambda | 1 log_2 \lambda = [15 : 0.5 : 7]\}$ and $\{\gamma | log_2 \gamma = [-5 : 0.5 : 5]\}$.

Given the learning parameters, the direct SVM classifier, linear-JDA-SVM classifier and optimized kernel-JDA-SVM classifier are trained with C samples to identify different faults. Table II shows the diagnosis results. 'C2B' is not studied in this research because the distributions are almost the same between C and B and the accuracy of using the SVM trained the original C samples to identify B samples is 94.38 % which is good enough for fault diagnosis.

TABLE III
THE DIAGNOSIS ACCURACY OF EXPERIMENTS FOR TARGET SAMPLE SETS WITH NORMAL SIZE

Experiment	**Source Samples**	**Target Samples**	**Accuracy(%)**	
			SVM	**optimized kernel-JDA-SVM**
C2A	828	370	71.89	**91.35**
C2D	828	828	85.51	**96.26**

It can be seen from the table that the accuracy of optimized kernel-JDA-SVM is the highest. And the optimized kernel-JDA-SVM improves the ACCs of the SVM for 'C2A' and 'C2D' by more than 10%, making the fault diagnosis accuracy greater than 90%.

Results of grid search for optimized kernel-JDA-SVM are shown in Fig.4. The gird for 'C2D' remains the initial state while that for 'C2A' experiment has been fined to get the better results. The optimized parameters of kernel-JDA-SVM in 'C2A' experiment are $\gamma = 9$ and $\lambda = 32$. And the optimized parameters in 'C2D' experiment are $\gamma = 2^{-3.5}$ and $\lambda = 2^{-7}$. In addition, the MMD of marginal distribution and that of marginal distribution are calculated based on Eq.1 and Eq.5 to measure the difference between target distribution and source distribution. The total MMD is denoted as MMD_T, which is calculated by

$$MMD_T = \sum_{c=0}^{C} tr(A^T K M_c K^T A) \tag{20}$$

Table IV compares the MMD_T for original samples and mapped samples after using optimized kernel-JDA-SVM in both 'C2A' and 'C2D' experiments. The MMD_T are decreased after mapping based on optimized kernel-JDA. So, the SVM classifiers trained by mapped samples have higher accuracy than those trained by original samples.

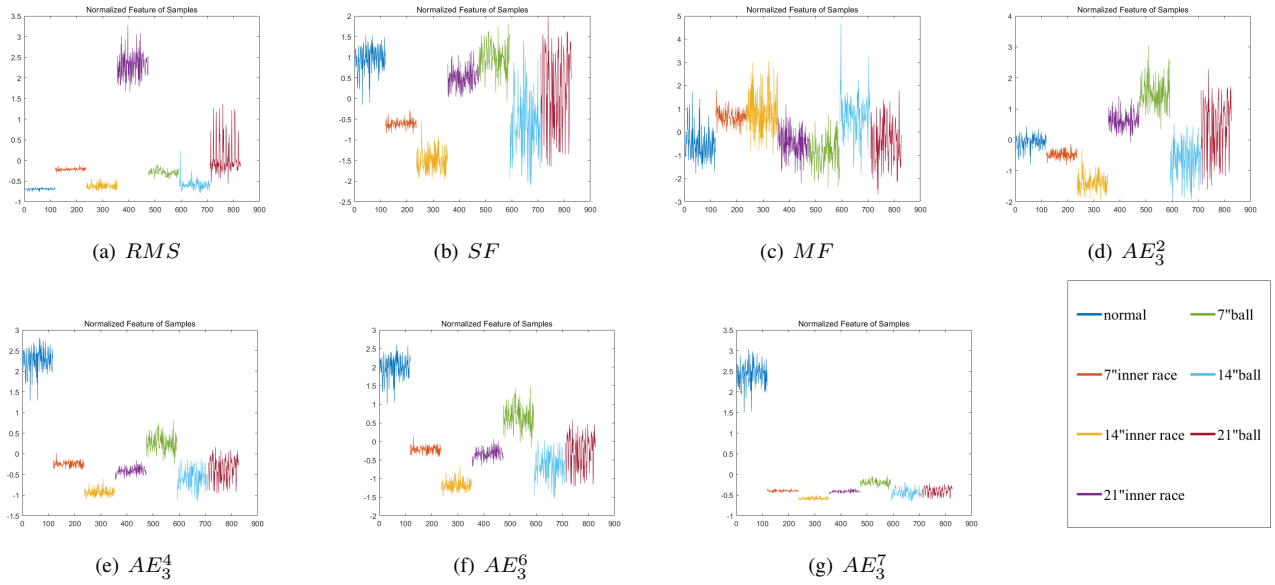
TABLE IV
THE MMD_T BETWEEN TARGET DATA AND SOURCE DATA BEFORE AND AFTER USING OPTIMIZED KERNEL-JDA-SVM

Experiment	$MMD_T(10^{-2})$	
	Before	*After*
C2A	205.7	**66.09**
C2D	45.85	**27.21**

In general, it can be concluded that optimized kernel-JDA-SVM can be effective in transfer learning fault diagnosis.

D. Performance Analysis of Optimized kernel-JDA-SVM Fault Diagnosis for Small Size of Target Samples

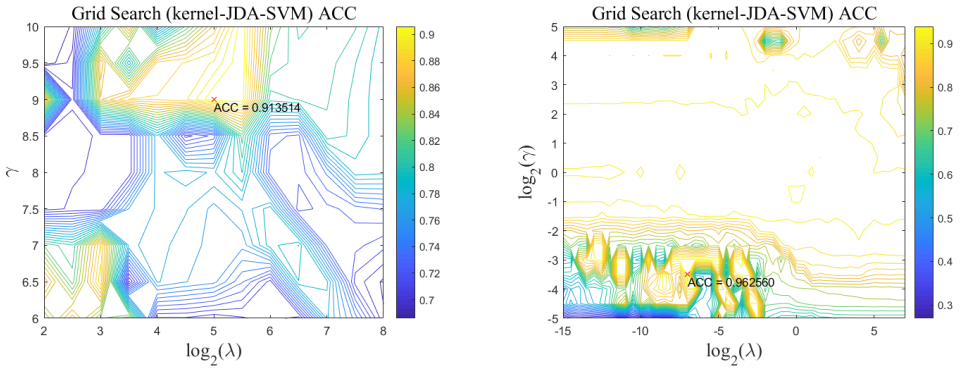
The experiments to validate the applicability of the presented method in small set of target samples are also conducted in this study. 77 independent samples under the D working



(a) RMS (b) SF (c) MF (d) AE_3^2

(e) AE_3^4 (f) AE_3^6 (g) AE_3^7

Fig. 3. A Part of Normalized features extracted from data under C working condition (The x-coordinate is the sequence number of the sample, while the y-coordinate is the normalized value of the feature.)



(a) C2A (b) C2D

Fig. 4. The gird search for optimized kernel-JDA-SVM (The red cross in the figure shows the optimal accuracy in the grid, and its corresponding coordinates are the optimized parameters.)

condition are taken as the target data at each time, while all the samples under C working condition are taken as the source data. Table V shows the results of fault diagnosis for small set of target samples based on the direct SVM and optimized kernel-JDA-SVM respectively.

It can be seen that the average accuracy of fault diagnosis for small target samples based on optimized kernel-JDA-SVM reaches 96%, which is higher than SVM without distribution adaptation. Especially for these with lower accuracy, the optimized kernel-JDA-SVM has increased the accuracy a lot. For example, the accuracy of fault diagnosis in 'C2D1' experiment is improved from 66.23% to 98.70%. For the situations where the distribution similarities between target and source domains are high enough to use SVM directly, the accuracies of the optimized kernel-JDA-SVM can still stay at a high level. Although the accuracies for 'C2D10' and 'C2D3' have been a little bit decreased by 1%, the accuracies of the optimized kernel-JDA-SVM are still larger than 95%. So, the optimized kernel-JDA-SVM is effective in gear fault diagnosis under the situation where the target samples are insufficient.

In general, it has been proved that optimized kernel-JDA-SVM is effective in bearing fault diagnosis under various working conditions, which make the source data useful for fault diagnosis in other working conditions with relative high accuracy.

V. CONCLUSION

In this study, a transfer learning based method has been designed for rolling bearing fault diagnosis under variable working conditions. The method includes multi-features extraction and classifier construction based on optimized kernel-JDA-SVM. The multi-features contain 7 kind of time domain features and 8 wavelet packet features. The parameters of kernel-JDA-SVM are optimized by grid search to get a better performance. Furthermore, the effectiveness of the proposed

TABLE V
THE DIAGNOSIS ACCURACY OF EXPERIMENTS FOR TARGET SAMPLE SETS WITH SMALL SIZE

Experiment	Source Samples	Target Samples	Accuracy(%) SVM	Accuracy(%) optimized kernel-JDA-SVM
C2D1	828	77	66.23	**98.70**
C2D2	828	77	92.21	**94.80**
C2D3	828	77	**97.40**	96.10
C2D4	828	77	92.21	**94.81**
C2D5	828	77	92.21	**94.81**
C2D6	828	77	89.61	**94.81**
C2D7	828	77	94.81	**96.10**
C2D8	828	77	96.10	**97.40**
C2D9	828	77	90.91	**98.70**
C2D10	828	77	**98.70**	97.40
AVG			**91.04**	**96.36**

approach is validated through groups of comparison experiments. In the experiments, above 90% accuracy is achieved in optimized kernel-JDA-SVM for fault diagnosis for target data under different working conditions. Meanwhile, for the situation when the size of target samples is relatively small, optimized kernel-JDA-SVM increase the accuracy from 66% to over 90% in the experiment.

For the future, the weight of marginal distribution difference and conditional distribution difference will be studied to adjust it automatically by evolutionary computation to improve the performance. In addition, the applicability of the method for different types of bearings will be also studied through experiments.

ACKNOWLEDGMENT

The author would like to thank SJTU Innovation Center of Producer Service Development, Shanghai Research Center for industrial Informatics, Shanghai Key Lab of Advanced Manufacturing Environment, National Natural Science Foundation of China (Grant No. 71632008) and Major Special Basic Research Projects for Aero engines and Gas turbines (Grant No. 2017-I-0007, Grant No.2017-I-0011) for the funding support to this research.

REFERENCES

[1] Wang, Y. S., et al. "An intelligent approach for engine fault diagnosis based on Hilbert–Huang transform and support vector machine." Applied acoustics 75 (2014): 1-9.

[2] Widodo, Achmad, and Bo-Suk Yang. "Support vector machine in machine condition monitoring and fault diagnosis." Mechanical systems and signal processing 21.6 (2007): 2560-2574.

[3] Pandya, D. H., S. H. Upadhyay, and Suraj Prakash Harsha. "Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN." Expert Systems with Applications 40.10 (2013): 4137-4145.

[4] Lei, Yaguo, et al. "Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs." Mechanical systems and signal processing 21.5 (2007): 2280-2294.

[5] Zhao, Rui, et al. "Deep learning and its applications to machine health monitoring." Mechanical Systems and Signal Processing 115 (2019): 213-237.

[6] Jing, Luyang, et al. "A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox." Measurement 111 (2017): 1-10.

[7] Shao, Haidong, et al. "Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet." ISA transactions 69 (2017): 187-201.

[8] Yang, Rui, et al. "Rotating machinery fault diagnosis using long-short-term memory recurrent neural network." IFAC-PapersOnLine 51.24 (2018): 228-232.

[9] Liu, Hongmei, Xuan Wang, and Chen Lu. "Rolling bearing fault diagnosis under variable conditions using Hilbert-Huang transform and singular value decomposition." Mathematical Problems in Engineering 2014 (2014).

[10] Fei, Sheng-wei. "Fault Diagnosis of Bearing Under Varying Load Conditions by Utilizing Composite Features Self-Adaptive Reduction-Based RVM Classifier." JOURNAL OF VIBRATION ENGINEERING & TECHNOLOGIES 5.3 (2017): 269-276.

[11] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.

[12] Taylor, Matthew E., and Peter Stone. "Transfer learning for reinforcement learning domains: A survey." Journal of Machine Learning Research 10.Jul (2009): 1633-1685.

[13] Dai, Wenyuan, et al. "Boosting for transfer learning." Proceedings of the 24th international conference on Machine learning. 2007.

[14] Huang, Jiayuan, et al. "Correcting sample selection bias by unlabeled data." Advances in neural information processing systems. 2007.

[15] Zhao, Zhongtang, et al. "Cross-people mobile-phone based activity recognition." Twenty-second international joint conference on artificial intelligence. 2011.

[16] Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." arXiv preprint arXiv:1502.02791 (2015).

[17] Long, Mingsheng, et al. "Deep learning of transferable representation for scalable domain adaptation." IEEE Transactions on Knowledge and Data Engineering 28.8 (2016): 2027-2040.

[18] Long, Mingsheng, et al. "Deep transfer learning with joint adaptation networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

[19] Pan, Sinno Jialin, et al. "Domain adaptation via transfer component analysis." IEEE Transactions on Neural Networks 22.2 (2010): 199-210.

[20] Long, Mingsheng, et al. "Transfer joint matching for unsupervised domain adaptation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[21] Duan, Lixin, Ivor W. Tsang, and Dong Xu. "Domain transfer multiple kernel learning." IEEE Transactions on Pattern Analysis and Machine Intelligence 34.3 (2012): 465-479.

[22] Long, Mingsheng, et al. "Transfer feature learning with joint distribution adaptation." Proceedings of the IEEE international conference on computer vision. 2013.

[23] Zhang, Wei, et al. "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals." Sensors 17.2 (2017): 425.

[24] Wen, Long, Liang Gao, and Xinyu Li. "A new deep transfer learning based on sparse auto-encoder for fault diagnosis." IEEE Transactions on Systems, Man, and Cybernetics: Systems 49.1 (2017): 136-144.

[25] Zhong, Shi-sheng, Song Fu, and Lin Lin. "A novel gas turbine fault diagnosis method based on transfer learning with CNN." Measurement 137 (2019): 435-453.

[26] Li, Xudong, et al. "Fault diagnostics between different type of components: A transfer learning approach." Applied Soft Computing 86 (2020): 105950.

[27] Ocak, Hasan, Kenneth A. Loparo, and Fred M. Discenzo. "Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: A method for bearing prognostics." Journal of sound and vibration 302.4-5 (2007): 951-961.

[28] Bearing Data Center, Case Western Reserve University. http://csegroups.case.edu/bearing datacenter/pages/download-data-file.