Mechanical Systems and Signal Processing 186 (2023) 109884

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Deep discriminative transfer learning network for cross-machine fault diagnosis

Quan Qian [a,b], Yi Qin [a,b,*], Jun Luo [a,b,*], Yi Wang [a,b], Fei Wu [a,b]

[a] State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400044, People's Republic of China
[b] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, People's Republic of China

ARTICLE INFO

Communicated by Yaguo Lei

Keywords:
Discriminative feature learning
Joint domain adaptation
Distribution alignment
Classification loss
Fault transfer diagnosis

ABSTRACT

Many domain adaptation methods have been presented to deal with the distribution alignment and knowledge transfer between the target domain and the source domain. However, most of them only pay attention to marginal distribution alignment while neglecting the discriminative feature learning in two domains. Thus, they still cannot satisfy the diagnosis requirement in some cases. To enhance the distribution alignment and match the marginal distributions as well as conditional distributions of two domains, an improved joint distribution adaptation (IJDA) mechanism is proposed. In IJDA, to enhance domain confusion, maximum mean discrepancy and CORrelation Alignment (CORAL) are combined as a new distribution discrepancy metric. Furthermore, an improved conditional distribution alignment mechanism is constructed. To contribute to feature learning and learn more separable features, a new I-Softmax loss that can be optimized like the original Softmax loss and possesses a stronger classification ability is proposed. With the IJDA mechanism and I-Softmax loss, the deep discriminative transfer learning network (DDTLN) is built to implement fault transfer diagnosis. Under the unlabeled target-domain samples, the experimental results on six cross-machine diagnostic tasks verify that the proposed DDTLN has a higher performance of transfer fault diagnosis than other typical domain adaptation methods.

1. Introduction

Rotating machines are widely applied to the manufacturing industry, energy supply, rail transportation, and aerospace industry. However, they usually work in complex and harsh environments. Unexpected faults may occur with the long-time operation, which may result in economic losses and serious casualties [1]. Thus, precisely identifying their health condition is vital for ensuring a safe operating environment and high production efficiency. Due to the rapid development of industrial big data and measurement technology, cutting-edge fault diagnosis and prognosis algorithms have attracted the attention of many researchers [2,3]. To automatically mine the fault-relation features and achieve an end-to-end diagnosis scheme, the deep learning (DL)-based diagnosis methods have become a research hotspot in recent five years. More importantly, compared with the traditional machine learning-based methods and signal analysis-based methods, DL-based diagnosis methods can greatly reduce the interference of human experience.

Generally, the appealing performance and successful application of deep learning rely on a significant amount of annotated training

* Corresponding authors at: State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400044, People's Republic of China.
E-mail addresses: qy_808@cqu.edu.cn (Y. Qin), qy_808@cqu.edu.cnluoj, un@cqu.edu.cn (J. Luo).

https://doi.org/10.1016/j.ymssp.2022.109884
Received 30 June 2022; Received in revised form 12 September 2022; Accepted 15 October 2022
Available online 25 October 2022

data. It is extremely difficult to obtain enough labeled samples in the actual industry, which also means the robustness and generalization ability of the deep learning model cannot be effectively guaranteed. The high-quality labeled data is still rare as the industrial supply end lacks the related expertise experience. On the other hand, these DL-based diagnosis models generally demand that the training dataset and the test dataset satisfy the same probability distribution. Unfortunately, this precondition is very hard to meet in actual applications. For example, due to working loads, transmission path, noise interference, fault degree, and even complex mechanical structure, the gathered monitoring data from rotating machines will unavoidably result in prominent distribution discrepancy. Therefore, advanced diagnosis methods are urgently required under the limitation of small and unlabeled samples.

To deal with the above-stated issues on DL, transfer learning (TL) derived from computer vision provides a surprising but feasible way to diagnose the fault of rotating machines with the unlabeled target-domain dataset. By reducing the distribution discrepancy between two domains, the TL can share the knowledge obtained from the labeled source-domain dataset to the small labeled target-domain dataset or the unlabeled target-domain dataset. Domain adaptation (DA) mitigates the two-domain gap and learns the domain-invariant features. As such, it plays an integral role in the TL. Mainstream deep DA mechanisms can be divided into adversarial mechanism-based [4,5] and statistic metric-based [6–9] mechanisms. For example, deep domain confusion (DDC) [6] and deep adaptation network (DAN) [7] were proposed to execute cross-domain image classified tasks with the maximum mean discrepancy (MMD) distance metric. Deep correlation alignment (DCORAL) [8] also obtains a better consequence than typical covariance methods. Motivated by the generative adversarial network (GAN) [10], Ganin et al. [4] presented a domain discriminator to discriminate the source domain and target domain. Then, domain confusion was achieved by adversarial learning between the feature extractor and the domain discriminator. In the area of fault transfer diagnosis, Long et al. [9] employed a three-layer sparse auto-encoder network and MMD metric to perform fault diagnosis on the case western reserve university (CWRU) bearing dataset. To further boost the ability of domain confusion, some methods were proposed to improve the accuracy of transfer diagnosis under different loads by combining the adversarial mechanism and distance metric [11]. Aiming at various types of transfer tasks, DA-based methods can be divided into partial domain adaptation [12], close-set domain adaptation [13], open-set domain adaptation [14], universal domain adaptation [15], multi-to-one domain adaptation in the source and target domains [16], and one-to-multi domain adaptation in the source and target domains [17]. For instance, to execute the partial transfer diagnosis of bearings and gears, a new weight selection adversarial network was proposed by Li et al. [12], which constructed an auxiliary discriminator to obtain the instance weights of source-domain samples and target-domain samples. Zhang et al. [15] built a deep hybrid weighted DA mechanism to diagnose bearing faults, where the prior relationship between the source-domain label space and target-domain label space was unknown. A multisource refined transfer learning network was presented by Chai et al. [16] to obtain the shared classes corresponding to the target domain from multi-source domains by the weight selection mechanism, which broke the assumption that the label space of each source domain was equal to the target domain.

Despite the above DA-based methods making significant achievements in various fields and transfer tasks, they neglect the following two key factors. First, they only focus on the marginal distribution alignment (MDA) between the target domain and the source domain while neglecting the conditional distribution alignment (CDA) on the same class of two domains. To enhance DA ability, joint distribution adaptation, including the MDA and CDA, was proposed by Long et al. [18]. However, the conditional probability distribution is approximatively replaced by the class conditional probability distribution [18], and it affects the performance of domain confusion to a certain degree. Secondly, the goal of classified transfer tasks is to obtain discriminative yet domain-invariant features. However, almost all DA models mainly consider domain-invariant feature learning while simultaneously overlooking discriminative feature learning. In Fig. 1, the probability density functions (PDF) of four healthy conditions on CWRU bearing dataset is plotted. It can be observed that PDFs are jumbly due to noise interference and other factors, which is not conducive to executing fault transfer diagnosis. Therefore, a more separable feature learning mechanism (discriminative feature learning), i.e., the smaller inner-class distance and the larger inter-class distance, is required in DA.

In discriminative feature learning, the related works can be divided into two respects: the designed loss function [19–21] and the network structure [22]. For instance, Liu [19,20] et al. proposed the L-Softmax and A-Softmax to adjust the desired margin by mapping

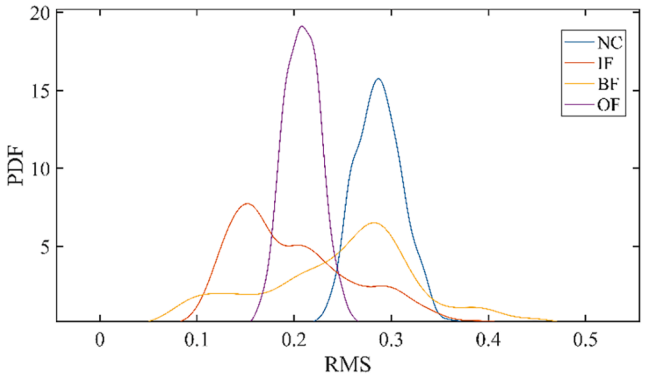

Fig. 1. PDF of four healthy conditions in the CWRU bearing dataset, and the x-axis denotes the root mean square (RMS) of vibrational signals listed in Table 2, and the y-axis denotes the PDF.

the original Euclidean space of features to the angular space. However, optimization is extremely difficult due to the nonmonotonicity of the cosine function. A new network structure including two classifiers was designed by Wu et al. [22] to obtain a better discriminating performance via a maximum classifier discrepancy (MCD) adversarial mechanism. Unfortunately, the adversarial mechanism will lead to an unstable task score [23].

In order to better demonstrate the relationship between DA mechanism and discriminative feature learning, a binary diagnostic task is taken as an instance in Fig. 2. The following consensuses about DA should be firstly emphasized: (1) As the source domain has the prior knowledge of label information, thus the target-domain features aim to align the distribution of source domain during DA. (2) The distribution distance between two domains with the same category is minimum, which means the target-domain features will give priority to align the distribution of same category in source domain during DA. As shown in Fig. 2 (a), the universal DA mechanism always adopts the original Softmax loss whose decision margin illustrated in Fig. 2(b) is equal to zero for obtaining the separable features, which can only guarantee the separability of labeled source-domain features. Due to the distribution discrepancy between source domain and target domain, some target-domain features will be inevitably misclassified by the decision boundary of source domain. Even after the domain confusion, these misclassified target-domain features may be still not correctly classified as they are far away from the decision boundary. For the discriminative DA mechanism, as there is a large decision margin between two decision boundaries, the misclassified target-domain features are still closer to the same source-domain category. Thus they can be reclassified correctly via the domain confusion, as illustrated in Fig. 2(b).

By the above analysis and discussion, there are key issues in the current fault transfer diagnosis: (1) The universal joint distribution adaptation mechanism cannot better achieve the domain confusion because of its approximation. (2) The existing DA diagnostic methods neglect the discriminative feature learning. (3) The current discriminative feature learning methods have the difficulty of optimization or instability. Aiming to solve these issues, the deep discriminative transfer learning network (DDTLN) is proposed based on a convolutional neural network (CNN). DDTLN is mainly composed of an improved joint distribution adaptation (IJDA) and improved Softmax (I-Softmax) loss. In IJDA, CORAL and MMD are combined as a new distribution discrepancy metric (DDM) to enhance domain confusion. Moreover, an improved CDA mechanism is built to achieve the domain confusion to a larger degree. For obtaining higher diagnosis accuracy and learning more separable features, the I-Softmax is proposed. The main contributions of this paper are listed as follows:

(1) Considering the proximity of the existing CDA mechanism, a novel CDA mechanism is built for better aligning the real conditional probability distributions of two domains. Then the IJDA mechanism is proposed using the MDA and the proposed CDA.
(2) To better measure the distribution distance from two aspects of mean and covariance, an improved metric combining MMD and CORAL is designed to further reduce the distribution discrepancy.
(3) For learning more separable fault features, a new I-Softmax loss with a flexible margin is proposed to bring the transfer framework a more excellent diagnostic capacity in cross-machine transfer diagnosis tasks.

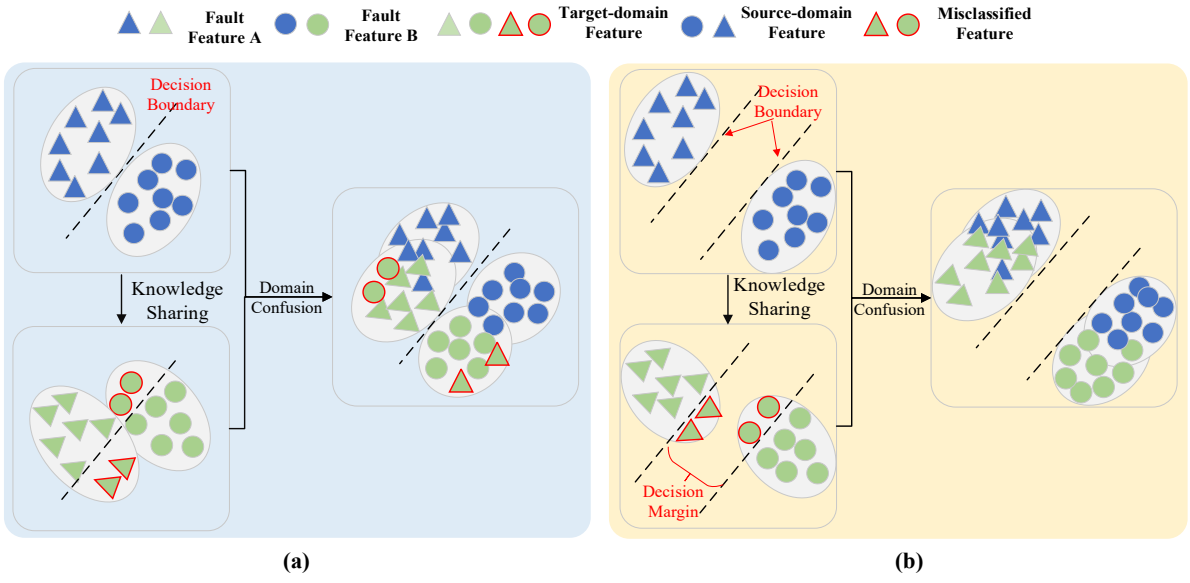

Fig. 2. Different DA mechanisms: (a) universal DA mechanism, (b) discriminative DA mechanism, where the decision margin denotes the distance between two decision boundaries.

2. Preliminaries

2.1. Problem definition

(1) **Domain:** Given the marginal probability distribution $P(\mathbf{X})$ and the n-dimension feature space $\mathscr{X} \in \mathrm{R}^{\mathrm{n}}$, the domain can be represented as $\mathscr{D} = \{\mathscr{X}\mathrm{hP(X)}\}$, where $\mathbf{X} \in \mathscr{X}$. The labeled source domain and the unlabeled target domain are denoted by $\mathscr{D}_{\mathrm{S}} = \{(\mathbf{X}_{\mathrm{S}}^1, \mathbf{Y}_{\mathrm{S}}^1), (\mathbf{X}_{\mathrm{S}}^2, \mathbf{Y}_{\mathrm{S}}^2), \cdots (\mathbf{X}_{\mathrm{S}}^{\mathrm{ns}}, \mathbf{Y}_{\mathrm{S}}^{\mathrm{ns}})\}$ and $\mathscr{D}_{\mathrm{T}} = \{(\mathbf{X}_{\mathrm{T}}^1), (\mathbf{X}_{\mathrm{T}}^2), \cdots (\mathbf{X}_{\mathrm{T}}^{\mathrm{nT}})\}$ respectively, where the $\boldsymbol{Y}_S^i$ represents the corresponding true-ground label of the ith feature sample $\boldsymbol{X}_S^i$.

(2) **Task:** The task $\mathscr{T}$ is defined as $\mathscr{T} \in (\mathscr{Y}, \mathrm{f}(\mathbf{X}))$, where $f(\cdot)$ denotes the learnable C-cardinality classifier and $\mathscr{Y}$ represents the label space of the features $\mathbf{X}$. In the unsupervised DA tasks, the unlabeled target-domain task $\mathscr{T}_{\mathrm{T}}$ is executed by the knowledge learned via the labeled source-domain task $\mathscr{T}_{\mathrm{S}}$. Assuming $f(\mathbf{X})$ is subjected to $P(\mathbf{Y}|\mathbf{X})$, $P(f(\mathbf{X})) = P(\mathbf{Y}|\mathbf{X})$ can be seen as the conditional probability distribution.

(3) **Joint distribution adaptation:** According to [18], joint distribution adaptation is defined as the combination of MDA (i.e., $P(\mathbf{X}_S) = P(\mathbf{X}_T)$) and CDA (i.e., $P(\mathbf{Y}_S|\mathbf{X}_S) = P(\mathbf{Y}_T|\mathbf{X}_T)$). Its optimization goal learning a feature representation T in which distribution discrepancy of marginal distribution and conditional distribution in two domains can be significantly reduced. Under the assumption that the class conditional probability distribution ($P(\mathbf{Y}|\mathbf{X})$) is equal to the conditional probability distribution ($P(\mathbf{X}|\mathbf{Y})$), the objective function L_{JDA} of joint distribution adaptation is defined as:

$$\begin{aligned} L_{JDA} &= \left\| \mathrm{E}_{\mathrm{P}(\mathbf{X}_{\mathrm{S}})}[\mathrm{T}(\mathbf{X}_{\mathrm{S}})] - \mathrm{E}_{\mathrm{P}(\mathbf{X}_{\mathrm{T}})}[\mathrm{T}(\mathbf{X}_{\mathrm{T}})] \right\|^2 \\ &+ \sum_{c=1}^{C} \left\| \mathrm{E}_{\mathrm{P}(\mathbf{X}_{\mathrm{S}}|\mathbf{Y}_{\mathrm{S}}=\mathrm{c})}[\mathrm{T}(\mathbf{X}_{\mathrm{S}})|\mathbf{Y}_{\mathrm{S}} = \mathrm{c}] - \mathrm{E}_{\mathrm{P}(\mathbf{X}_{\mathrm{T}}|\mathbf{Y}_{\mathrm{T}}=\mathrm{c})}[\mathrm{T}(\mathbf{X}_{\mathrm{T}})|\mathbf{Y}_{\mathrm{T}} = \mathrm{c}] \right\|^2 \end{aligned} \tag{1}$$

where $P(\mathbf{Y}_T|\mathbf{X}_T)$ is estimated by the knowledge from the labeled source domain, i.e., by the pseudo label. Parameter C denotes the number of categories.

2.2. Maximum mean discrepancy and correlation alignment

Maximum mean discrepancy is the most frequently used distribution distance metric in transfer learning tasks. It is defined as:

$$MMD(\mathbf{X}_S, \mathbf{X}_T) = \left\| \frac{1}{n_S} \sum_{\mathbf{X}_S \in \mathscr{D}_{\mathrm{S}}} \Phi(\mathbf{X}_S) - \frac{1}{n_T} \sum_{\mathbf{X}_T \in \mathscr{D}_{\mathrm{T}}} \Phi(\mathbf{X}_T) \right\|_H^2 \tag{2}$$

where n_S and n_T respectively represent the mini-batch size of source domain samples and target domain features, $\|h\|_H$ denotes the reproducing kernel Hilbert space (RKHS), and $\Phi(\cdot)$ denotes the mapping function in RKHS. For simplicity purposes, the explicitly

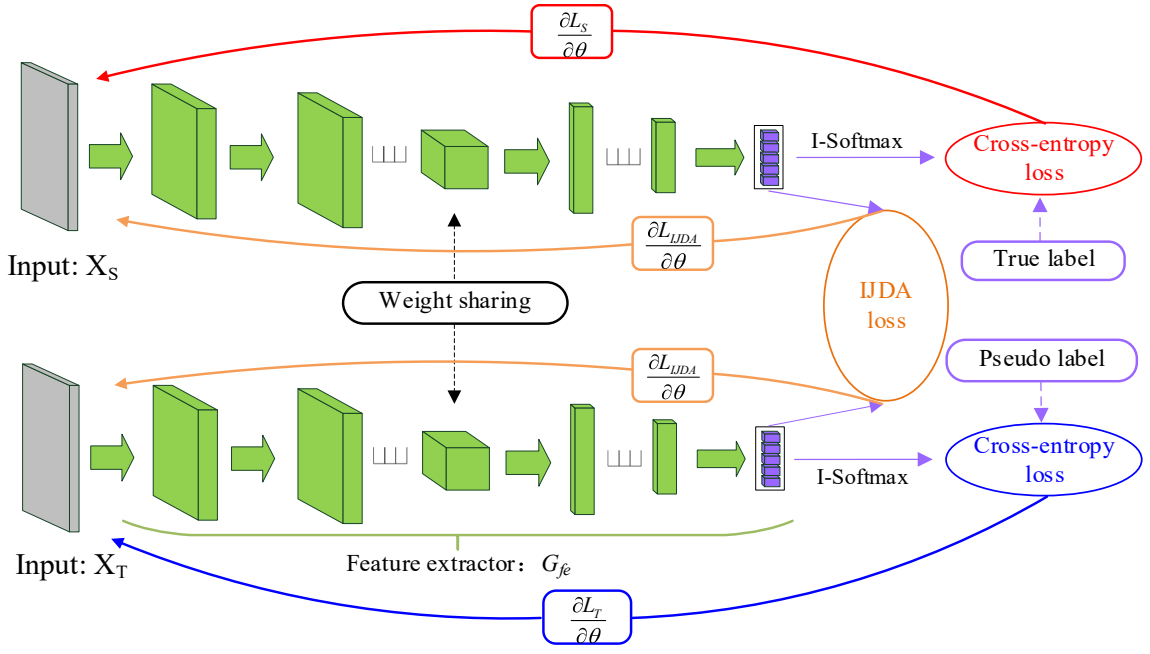

Fig. 3. The network structure of DDTLN; the right arrow and the left arrow respectively denote the forward propagation and the back-propagation.

mapping function is replaced by the kernel function in practical application, i.e., kernel trick. The common kernel function includes the Gaussian kernel, polynomial kernel, Laplace kernel, and Sigmoid kernel.

Just as the MMD aligns the mean of source-domain and the target-domain in RKHS, the correlation alignment aims to match the two-order covariance statistics:

$$CORAL(\mathbf{X}_S, \mathbf{X}_T) = \frac{1}{4d^2}\|\boldsymbol{Cov}_S - \boldsymbol{Cov}_T\|_{\mathrm{F}}^2 \tag{3}$$

where Cov_S and Cov_T denote the covariance metrics that can be computed as:

$$\begin{cases} Cov_S = \dfrac{1}{n_S - 1}\left(D_S^T D_S - \dfrac{1}{n_S}\left(I^T D_S\right)^T \left(I^T D_S\right)\right) \\ Cov_T = \dfrac{1}{n_T - 1}\left(D_S^T D_S - \dfrac{1}{n_T}\left(I^T D_T\right)^T \left(I^T D_T\right)\right) \end{cases} \tag{4}$$

where **I** is a row vector with all elements equal to 1.

3. The proposed transfer learning model

3.1. The DDTLN framework

The structure of the proposed DDTLN is plotted in Fig. 3. CNN is employed as the feature extractor due to its strong feature learning ability. The network parameters are listed in Table 1. It can be seen that the table includes five "Cov1D" blocks, one global average pooling (GAP) layer, and two fully connected (FC) layers. Every "Cov1D" block is composed of the Convolution layer, batch normalization (BN) layer, and max-pooling layer. The GAP and BN can accelerate network convergence and mitigate the overfitting phenomenon.

3.2. Improved joint distribution adaptation

To overcome the negative effect of the CDA approximation in Eq. (1), the improved CDA mechanism is proposed to align the conditional probability distributions in two domains. Using the Bayes theorem, the conditional probability distribution can be transformed as the form of class conditional probability distribution, which is expressed as:

$$P(Y = c|X) = \frac{P(Y = c) \cdot P(X|Y = c)}{P(X)} \tag{5}$$

Table 1
Detailed parameters of DDTLN.

Type of layer		Receptive field size/stride/number of channels	Input size	Output size
Conv1D-1	Convolution	64/16/32	(3072,1)	(192,32)
	BN	/	(192,32)	(192,32)
	Max-pooling	2/2/32	(192,32)	(96,32)
Conv1D-2	Convolution	3/1/64	(96,32)	(96,64)
	BN	/	(96,64)	(96,64)
	Max-pooling	2/2/64	(96,64)	(48,64)
Conv1D-3	Convolution	3/1/128	(48,64)	(48,128)
	BN	/	(48,128)	(48,128)
	Max-pooling	2/2/128	(48,128)	(24,128)
Conv1D-4	Convolution	3/1/256	(24,128)	(24,256)
	BN	/	(24,256)	(24,256)
	Max-pooling	2/2/256	(24,256)	(12,256)
Conv1D-5	Convolution	3/1/512	(12,256)	(12,512)
	BN	/	(12,512)	(12,512)
	Max-pooling	2/2/512	(12,512)	(6,512)
	GAP	/	(6,512)	512
FC1	1/*/512	512	512	
FC2	1/*/4	512	4	

where the class conditional probability distribution $P(\mathbf{X}|\mathbf{Y}=\mathrm{c})$ can be computed according to [18]. $P(\mathbf{Y}=\mathrm{c})$ denotes the class prior distribution that can be computed as:

$$\begin{cases} P(Y_S = c) = n_S^c / \sum_{i=1}^{C} n_S^i \\ P(Y_T = c) = n_T^c / \sum_{i=1}^{C} n_T^i \end{cases} \tag{6}$$

where C denotes the category number, while n_S^c and n_T^c respectively represent the batch size of the cth class in the entire source-domain batch size and the target-domain batch size:

$$\begin{cases} \sum_{c=1}^{C} n_S^c = n_s \\ \sum_{c=1}^{C} n_T^c = n_T \end{cases} \tag{7}$$

The goal of MDA is to align the marginal probability distribution, i.e.,$P(\mathbf{X}_S) = P(\mathbf{X}_T)$.Using Eq. (5), the improved CDA mechanism is formulated as:

$$L_{CDA} = \sum_{c=1}^{C} \left\| \mathrm{E}_{P(X_S|Y_S=c)}[T(X_S)|Y_S = c]P(Y_S = c) - \mathrm{E}_{P(X_T|Y_T=c)}[T(X_T)|Y_T = c]P(Y_T = c) \right\|^2 \tag{8}$$

Then, with the improved CDA mechanism and MDA mechanism, the final IJDA mechanism can be defined as:

$$\begin{aligned} L_{IJDA} &= \left\| \mathrm{E}_{P(X_S)}[T(X_S)] - \mathrm{E}_{P(X_T)}[T(X_T)] \right\|^2 \\ &+ \sum_{c=1}^{C} \left\| \mathrm{E}_{P(X_S|Y_S=c)}[T(X_S)|Y_S = c]P(Y_S = c) - \mathrm{E}_{P(X_T|Y_T=c)}[T(X_T)|Y_T = c]P(Y_T = c) \right\|^2 \end{aligned} \tag{9}$$

After defining the IJDA mechanism, we need to find a distribution distance metric to evaluate the marginal distribution discrepancy and conditional distribution discrepancy in Eq. (9). Due to a significant amount of random noise, the collected vibration signals of rotating machines are approximatively subjected to Gaussian distribution that includes two estimated parameters (mean and variance). Therefore, in order to better achieve the IJDA mechanism while further boosting the ability of domain confusion, CORAL and MMD distribution discrepancy metrics are combined as a new metric named$DDM(\mathbf{A},\mathbf{B})$:

$$DDM(\mathbf{A},\mathbf{B}) = MMD(\mathbf{A},\mathbf{B}) + CORAL(\mathbf{A},\mathbf{B}) \tag{10}$$

Taking the designed DDM metric into the IJDA mechanism, the final IJDA loss function can be rewritten as:

$$\begin{aligned} L_{IJDA} &= DDM(T(X_S), T(X_T)) \\ &+ \sum_{c=1}^{C} DDM[P(Y_S = c)\cdot(T(X_S)|Y_S = c),\ P(Y_T = c)\cdot(T(X_T)|Y_T = c)] \end{aligned} \tag{11}$$

where the feature representation T refers to the feature extractor in Fig. 3.

3.3. I-Softmax loss

For the multi-classified tasks, the Softmax function is widely employed in the neural network due to its probabilistic interpretation and simplicity. However, in some cases, it still cannot meet the requirements of intra-class compactness and inter-class separability. Thus, a new I-Softmax loss is designed to learn more separable features and improve the score in transfer tasks, which is defined as follows:

$$L_y = \begin{cases} -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{e^{F^i(c)/m-k}}{e^{F^i(c)/m-k} + \sum_{j\neq c} e^{F^i(j)}}\right), F^i(c) \rangle 0 \\ -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{e^{mF^i(c)-k}}{e^{mF^i(c)-k} + \sum_{j\neq c} e^{F^i(j)}}\right), F^i(c) \leqslant 0 \end{cases} \tag{12}$$

where $\mathbf{F}^i$ represents the feature vector outputted by the feature extractor, i.e.,$F^i = G_{fe}(Xi\Theta_{fe})$. Parameters $F^i(c)$ and $\mathbf{F}^i(j)$ denote the cth element corresponding to the label index of Xi and other elements respectively. n denotes the number of the feature vector, and $k \geqslant 0$ and $m \geqslant 1$ are defined as the hyper-parameters to control the decision boundary. If $m = 1$ and$k = 0$, the I-Softmax loss will be equal to

the original Softmax loss.

To clearly state the I-Softmax, it is interpreted from the angular perspective. As shown in Fig. 4, the original Softmax can be factorized into the angular and amplitude components via the cosine similarity:

$$F^i(c) = W_c^T Z^i = \|W_c\|\|Z^i\|\cos(\theta_c) \tag{13}$$

where $\mathbf{Z}^i$ denotes the feature vector outputted by the penultimate FC layer, W_c denotes the cth weight vector between the last two FC layers, and $\theta_c \in [-\pi, \pi]$ denotes the angular between W_c and z.

Binary classification is taken as an instance to elaborate the I-Softmax. Assuming that a sample-feature $\mathbf{Z}$ is obtained from class 1, the original Softmax loss must satisfy the following inequality for classifying the $\mathbf{Z}$ into class 1:

$$W_1^T Z > W_2^T Z \tag{14}$$

However, the original Softmax still cannot hold rigorous decision marginal, thereby influencing its compactness and separability. To make I-Softmax obtain a more proper decision boundary and achieve the correct classification, the following expression has to be satisfied:

$$\begin{cases} W_1^T Z > W_1^T Z/m - k > W_2^T Z, \ W_1^T Z > 0 \\ W_1^T Z > W_1^T Zhm - k > W_2^T Z, \ W_1^T Z \leqslant 0 \end{cases} \tag{15}$$

Obviously, the proposed I-Softmax loss holds more rigorous decision criteria for class 1 than the original Softmax loss. In essence, the intra-class compactness and inter-class separability of the I-Softmax function are also achieved by controlling the θ_c according to [24]. Taking $\mathbf{W}_c^{\mathrm{T}}\mathbf{Z} > \mathbf{0}$ as an example, we can prove that:

$$\begin{aligned} &\|Wc\|\|Z^i\|_2\cos\left(\widehat{\theta}\right) = \|Wc\|\|Z^i\|\cos(\theta_c)/m - k \\ &\Rightarrow k = \|Wc\|\|Z^i\|\left[\cos(\theta_c)/m - \cos\left(\widehat{\theta}\right)\right] \geqslant 0 \\ &\qquad\Rightarrow \widehat{\theta} \geqslant \theta_c \end{aligned} \tag{16}$$

where $\widehat{\theta}$ denotes the equivalent marginal angular. It immediately follows from Eq. (16) that the impact of k and m equivalently act on augmenting the angularθ_c.

Let us consider the case of $\|\mathbf{W}_1\| = \|\mathbf{W}_2\|$ and$\mathbf{W}_1^{\mathrm{T}}\mathbf{Z} > \mathbf{0}$, as shown in Fig. 5. Given a sample feature $\mathbf{Z}$ from class 1, the original Softmax loss to force $\mathbf{W}_1^{\mathrm{T}}\mathbf{Z} > \mathbf{W}_2^{\mathrm{T}}\mathrm{Z}$ ($\theta_1 < \theta_2$) can be classified. However, the obtained decision boundary is not rigorous enough to learn separable features. For the I-Softmax loss, it strictly demands $\mathbf{W}_1^{\mathrm{T}}\mathbf{Z}/\mathrm{m} - \mathrm{k} > \mathbf{W}_2^{\mathrm{T}}\mathrm{Z}$ ($\theta_1 << \theta_2$). Therefore, it makes a more discriminative decision marginal.

According to [25], feature normalization will encourage more separable feature learning. Therefore, we employ L_2-norm to normalize the weight vector:

$$\widehat{W}_c = \frac{W_c}{\|W_c\|_2} \tag{17}$$

Finally, the I-Softmax loss can be redefined as follows:

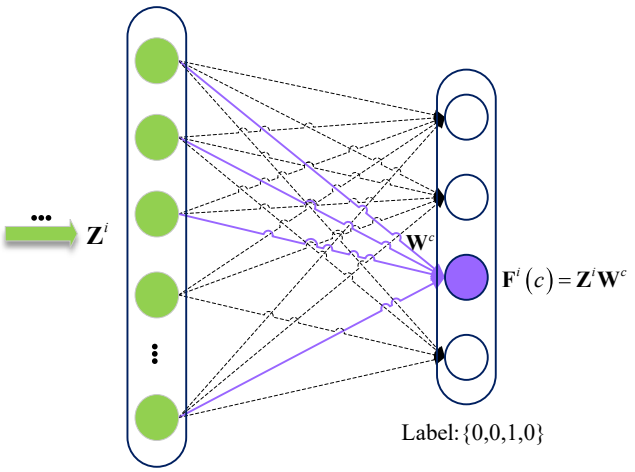

Fig. 4. The principle diagram of Eq. (13).

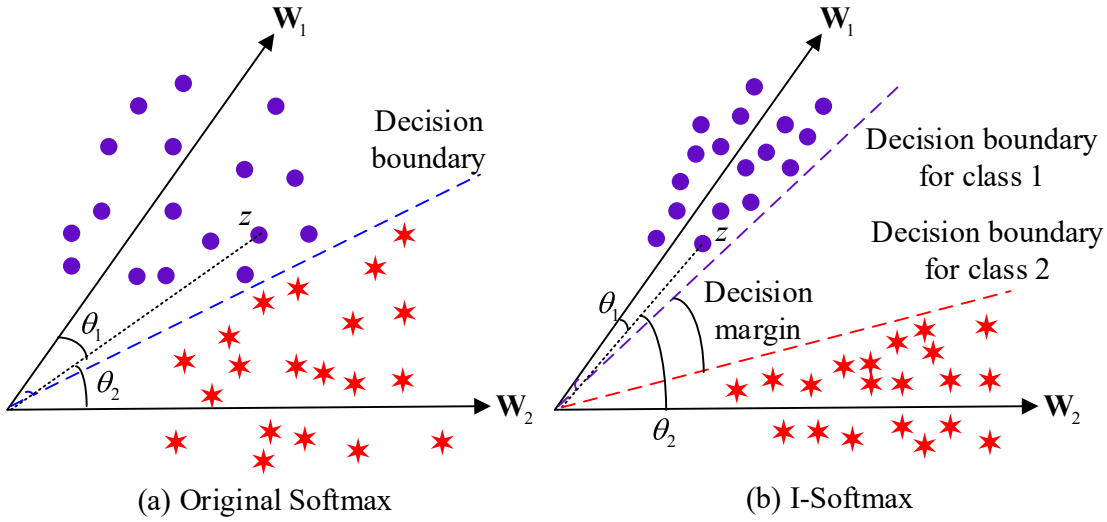

Fig. 5. Interpretation of I-Softmax.

$$L_y = \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{\widehat{W}_c^T Z^i/m-k}}{e^{\widehat{W}_c^T Z^i/m-k}+\sum_{j\neq c} e^{\widehat{W}_j^T Z^i}}\right), \widehat{W}_c^T Z^i > 0 \\ -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{m\widehat{W}_c^T Z^i-k}}{e^{m\widehat{W}_c^T Z^i-k}+\sum_{j\neq c} e^{\widehat{W}_j^T Z^i}}\right), \widehat{W}_c^T Z^i \leqslant 0 \end{cases} \tag{18}$$

3.4. Optimization objective

The proposed DDTLN model includes two optimization objectives: mining the IJDA loss with unsupervised training and mining the classified I-Softmax loss with supervised training.

(1) **IJDA loss.**

In the proposed IJDA loss, an improved joint domain adaptation mechanism is provided. Moreover, the MMD and CORAL are combined to achieve domain confusion from the normal distribution characteristic of signals. IJDA loss equation is shown in Eq. (10), where label information of target-domain samples is approximately obtained by the pseudo label. After optimizing the DDTLN by the IJDA loss, the obtained features will be domain-invariant. Otherwise, DDTLN can be directly optimized by the gradient back-propagation and chain rule. Finally, the IJDA loss gradient corresponding to the network parameters Θ_{fe} is expressed as:

$$\nabla\Theta_{fe} = \frac{\partial L_{IJDA}}{\partial\Theta_{fe}} = (\nabla L_{CORAL})^T\frac{\partial F}{\partial\Theta_{fe}} + (\nabla L_{MMD})^T\frac{\partial\Phi(F)}{\partial\Theta_{fe}} \tag{19}$$

$$\begin{cases} \nabla L_{CORAL} = \nabla CORAL(F_S, F_T) + \nabla CORAL\left(\frac{n_S^c}{n_S}(F_S|Y_S = c), \frac{n_T^c}{n_T}(F_T|Y_T = c)\right) \\ \nabla L_{MMD} = \nabla MMD(F_S, F_T) + \nabla MMD\left(\frac{n_S^c}{n_S}(F_S|Y_S = c), \frac{n_T^c}{n_T}(F_T|Y_T = c)\right) \end{cases} \tag{20}$$

Taking the $\nabla CORAL(\mathbf{F}_S, \mathbf{F}_T)$ and $\nabla MMD(\mathbf{F}_S, \mathbf{F}_T)$ as examples, the detailed equations are computed as follows:

$$\nabla MMD(\mathbf{F}_S, \mathbf{F}_T) = \begin{cases} \frac{2}{n_S}\left(\frac{1}{n_S}\sum_{\mathbf{F}_S\in\mathscr{D}_S}\Phi(\mathbf{F}_S) - \frac{1}{n_T}\sum_{\mathbf{F}_T\in\mathscr{D}_T}\Phi(\mathbf{F}_T)\right), \mathbf{F}\in D_S \\ \frac{2}{n_T}\left(\frac{1}{n_T}\sum_{\mathbf{F}_T\in\mathscr{D}_T}\Phi(\mathbf{F}_T) - \frac{1}{n_S}\sum_{\mathbf{F}_S\in\mathscr{D}_S}\Phi(\mathbf{F}_S)\right), \mathbf{F}\in D_T \end{cases} \tag{21}$$

$$\nabla CORAL(F_S, F_T) = \begin{cases} \frac{1}{d^2(n_S-1)}\left(D_S^T - \frac{1}{n_S}\left(\left(I^T Ds\right)^T I^T\right)^T (Cov_S - Cov_T)\right), & F \in D_S \\ \frac{1}{d^2(n_T-1)}\left(D_T^T - \frac{1}{n_S}\left(\left(I^T DT\right)^T I^T\right)^T (Cov_S - Cov_T)\right), & F \in D_T \end{cases} \tag{22}$$

where d represents the dimension of the feature vector **F**.

(2) **I-Softmax loss**.

Different from the original Softmax loss, I-Softmax loss can separate and compact learned features. This is more helpful for obtaining higher accuracy than the original Softmax loss on multi-classified tasks. According to Eq. (18), the original Softmax is only a particular case of the I-Softmax. Given that the vector **Z** is outputted by the I-Softmax function and its one-hot label vector **Y**, the gradient of I-Softmax loss is computed as follows:

$$\nabla\Theta_{fe} = \left(\nabla L_y\right)^T \frac{\partial F}{\partial \Theta_{fe}} \tag{23}$$

where ∇L_y is defined as:

$$\nabla L_y = \begin{cases} (Z-Y)/m, F^i(c) \rangle 0 \\ (Z-Y)\cdot m, F^i(c) \leqslant 0 \end{cases} \tag{24}$$

It can be seen that the gradient of I-Softmax loss with respect to the network parameters Θ_{fe} is approximatively equal to the original Softmax loss. In other words, the I-Softmax loss can easily optimize network parameters by backpropagation and it does not require additional training tricks as found in [19,20].

(3) **Global loss**.

Generally, the classified cross-entropy loss is applied to the labeled source domain for learning discriminative features. To learn more separable features in the TL tasks, the I-Softmax loss is applied to target-domain samples by the pseudo label. Hence, the entire classified loss is defined as:

$$L_W = L_S + \gamma L_T \tag{25}$$

where L_S and L_T respectively represent the source-domain I-Softmax loss and the target-domain I-Softmax loss. The parameter γ is the trade-off parameter.

By integrating the proposed IJDA loss and the I-Softmax loss, the entire objective function is defined as:

$$L_{all} = L_W + \lambda L_{IJDA} \tag{26}$$

where λ denotes the trade-off parameter. Then, the root mean square prop (RMSProp) optimizer is utilized to update the trainable parameters of DDTLN:

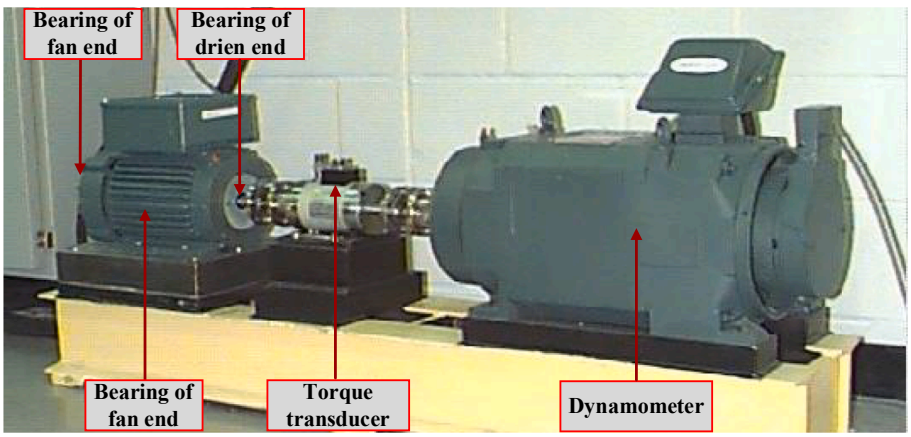

Fig. 6. The test rig of CWRU bearing datasets.

$$\Theta_{fe} \leftarrow \Theta_{fe} - \varepsilon\left(\frac{\partial L_S}{\partial \Theta_{fe}} + \gamma \frac{\partial L_T}{\partial \Theta_{fe}} + \lambda \frac{\partial L_{IJDA}}{\partial \Theta_{fe}}\right) \tag{27}$$

where ε denotes the learning rate. Ultimately, the DDTLN will obtain domain-invariant and more separable features.

4. Experimental study on cross-machine diagnosis

4.1. Dataset description

To verify the effectiveness of the proposed DDTLN, the cross-machine diagnostic tasks based on three bearing datasets are implemented in this section. The detailed information of three datasets is introduced as follows.

(1) **CWRU:** The CWRU datasets, collected by the Case Western Reserve University [26], is widely regarded as the standard benchmark dataset in bearing diagnostic cases. Its experimental platform, which includes a driven motor, a loading motor, a torque transducer, a dynamometer, and several testing bearings, is shown in Fig. 6. A total of four loads are simulated: 0 hp, 1 hp, 2 hp, and 3 hp. During bearing testing, raw vibration signals of several fault types are collected, including normal condition (NC), inner race fault (IF), ball fault (BF), and outer race fault (OF). The sampling frequency of acceleration sensors is set as 12000 Hz.

(2) **RTS:** RTS bearing dataset is built according to the RTS rotor dynamics test rig, which is a custom-made experimental platform. The fault types of RTS datasets are similarly composed of NC, IF, BF, and OF. The structure of the test rig shown in Fig. 7 is composed of a servo motor, a coupling, the bearings, two rotors, and a sensor. Raw vibration signals are collected by the CMS Wireless sensor placed on the right bearing pedestal. Several loads are simulated to gather sufficient raw vibration signals, which include 0 kN, 1 kN, 2 kN, and 3 kN. The sampling frequency is set as 8 000 Hz. The input speed of the bearing is 1 000 r/min, 2 000 r/min, and 3 000 r/min.

(3) **SWJTU:** The SWJTU bearing datasets are collected by the Southwest Jiaotong University [27]. As shown in Fig. 8, the test rig of SWJTU datasets consists of a three-phase motor, two bearings, an accelerometer, and a loading system. The fault types are also the same as the CWRU and RTS bearing datasets. The testbed can also collect raw signals under different loads. The sampling frequency of the accelerometer is 10 000 Hz. The input speed is set as 896 r/min.

4.2. Fault diagnostic tasks and implementation details

The sample number of each category in the source domain and target domain is 1000, thus source-domain and target-domain respectively have 4000 samples. The training dataset includes source-domain samples and target-domain samples, while the testing dataset only includes target-domain samples. Considering that the fault samples are rare in practice, the sliding sampling technique is employed to divide the raw data for augmenting the fault samples, and there are the overlapping points between two neighboring samples. Besides, each sample has 3072 data points so as to obtain enough fault information. To reduce the extra calculation and the influence of expertise, this paper directly uses the raw vibration samples as the input of the fault diagnosis model.

By employing the aforementioned three bearing datasets, six cross-machine transfer tasks are built to verify the effectiveness of DDTLN: A → B, B → A, A → C, C → A, B → C, and C → B. It must be noted that the six cross-machine transfer tasks comprehensively include the load and speed transfer, as listed in Table 2. By taking the A → B as an instance, the "A" and "B" respectively represent the labeled source domain and the unlabeled target domain. Detailed information of three datasets is listed in Table 2. All parameters of these datasets are mutually different apart for the healthy condition. This indicates that six transfer tasks represent a challenge when

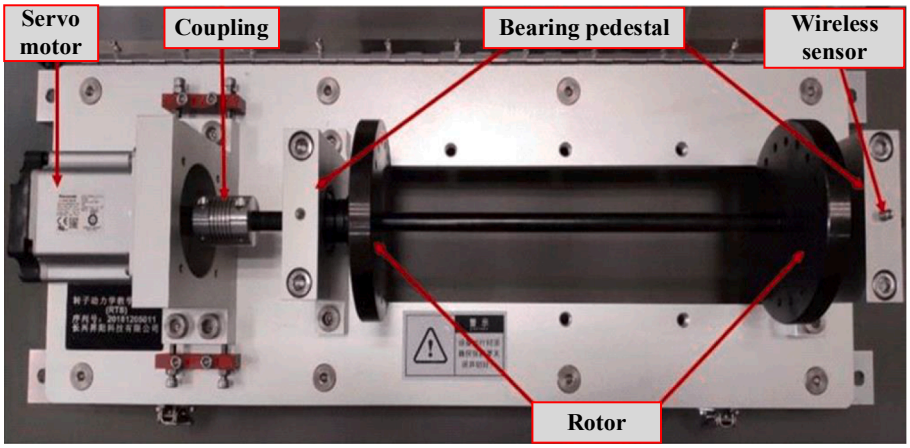

Fig. 7. The test rig of RTS bearing datasets.

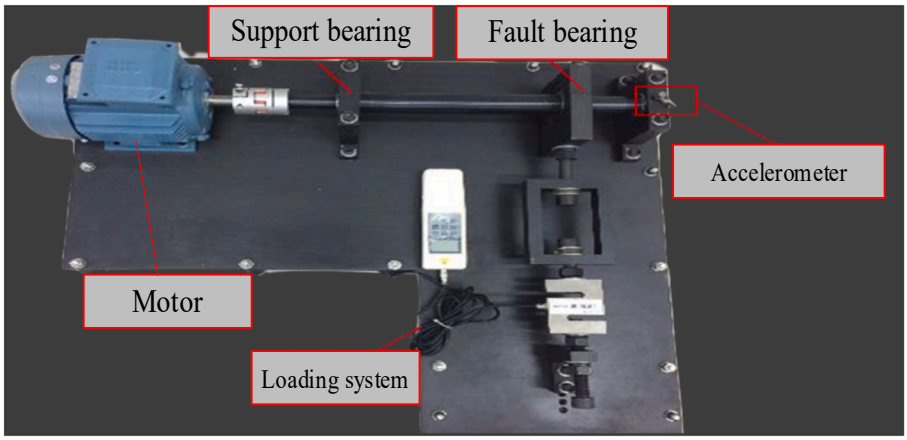

Fig. 8. The test rig of SWJTU bearing datasets.

Table 2
Detailed information of three datasets.

Name	Bearing	Healthy	Speed	Load	Fault size
A	CWRU bearing of drive end	NC	1750 r/min	2 hp	/
		IF	1750 r/min	2 hp	0.007 in.
		OF	1750 r/min	2 hp	0.007 in.
		BF	1750 r/min	2 hp	0.007 in.
B	RTS bearing	NC	1000 r/min	44 N	/
		IF	1000 r/min	44 N	0.5 mm
		OF	1000 r/min	44 N	0.5 mm
		BF	1000 r/min	44 N	0.5 mm
C	SWJTU dataset	NC	896 r/min	2 kN	/
		IF	896 r/min	2 kN	0.3 mm
		OF	896 r/min	2 kN	0.3 mm
		BF	896 r/min	2 kN	0.3 mm

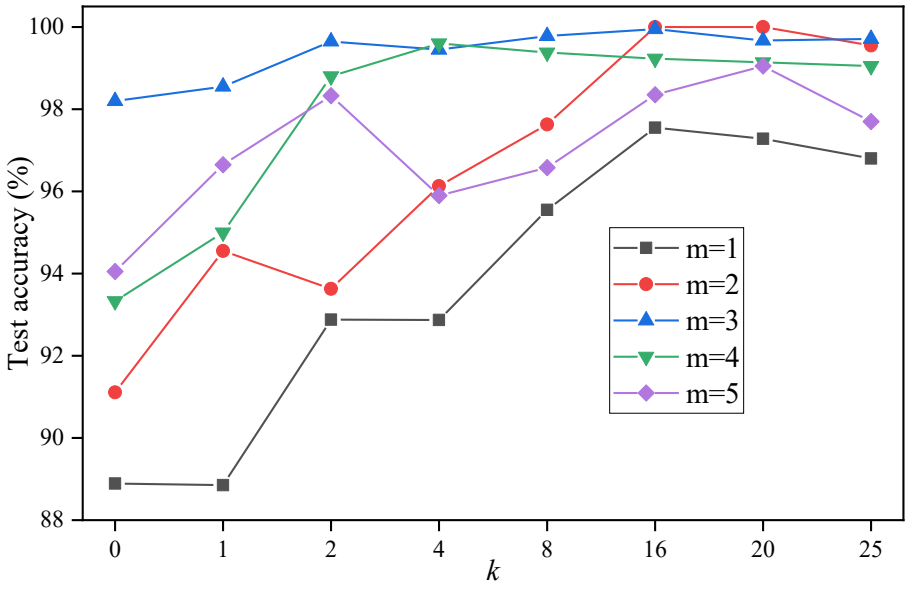

Fig. 9. The testing accuracy of I-Softmax in different margins.

DDTLN is used to precisely diagnose the fault.

Considering that the pseudo label is not equal to the true-ground label, the parameter γ is set to $\gamma = \lambda$ 0.1in Eqs. (25) and (26). This setting is able to reduce the effect of IJDA loss and target-domain I-Softmax loss during DDTLN training. In Eq. (27), the learning rate is set to 0.001. Training epochs are set to 300 and the batch size is set to 256. In addition, the DDTLN is trained on the Tensorflow platform with the GPU of NVIDIA 1050Ti.

4.3. Effectiveness analysis on I-Softmax loss

According to Eq. (18), the parameters (k, m) control the decision margin. Therefore, it is extremely important to improve the accuracy and performance of DDTLN. However, the I-Softmax loss will lose its ability to separate and compact the learned features if it is set to a relatively small value. On the contrary, if it is set to a relatively large value, the DDTLN will not converge. Thus, the value of the I-Softmax loss has to first be experimentally confirmed. To improve the test performance of I-Softmax, all datasets in Table 2 are combined into a single dataset. Then, the dataset is divided into the training dataset and the test dataset in the ratio of 7:3. The test results of different margins are illustrated in Fig. 9. When $m = 3$, the test accuracy changes with k slightly compared to other values of m, and it arrives at the maximum value when $k = 16$. Therefore, $m = 3$ and $k = 16$ are chosen in the subsequent diagnostic experiments. It should be noted that typical A-Softmax loss [20] and L-Softmax loss [19] are used to contrastively verify the superiority of I-Softmax. However, they are not able to converge. Thus, their testing results are not listed. Similarly, compared with the Soft-margin Softmax [28], the proposed I-Softmax loss has the more flexible margin to control the decision boundary and has a higher diagnostic accuracy, as shown in Fig. 9.

To intuitively demonstrate the discriminative power of features learned from different margins, these features are projected from the last FC layer into the unit sphere in Fig. 10. It can be observed that the I-Softmax leads to a more rigorous decision boundary and more discriminative distribution. Compared with the original Softmax and the Soft-margin Softmax, the I-Softmax explicitly makes the

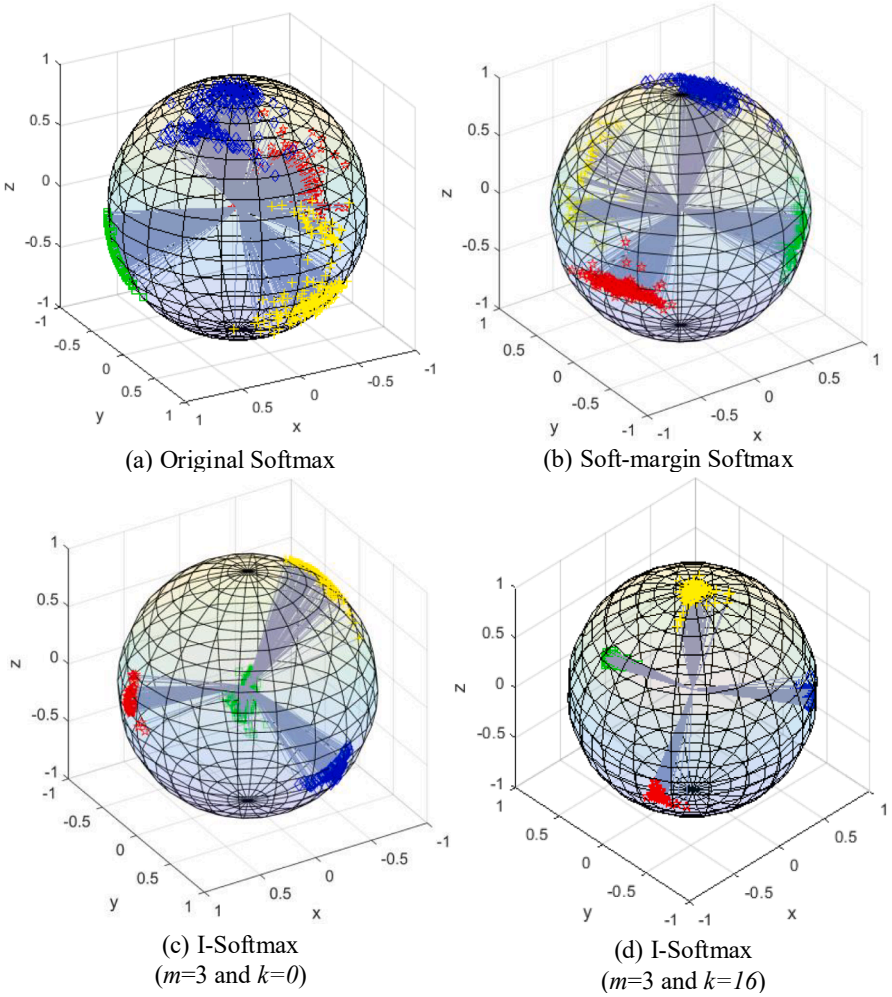

Fig. 10. Visualization of learned features projected onto the unit sphere.

intra-class distance smaller and the inter-class distance larger.

4.4. Experimental results and discussion

To further test the effectiveness and superiority of the proposed DDTLN, several well-known DA methods are used for comparison, such as DDC [6], DCORAL [8], DANN [4], MCD [29], FTNN [30] and JDA [31]. DDC, DCORAL, FTNN and JDA are famous distance metric-based DA models, and MCD and DANN are the typical adversarial mechanism-based DA models. Like the proposed DDTLN, MCD and JDA can also achieve the fine-grained class-wise distribution alignment. To verify the advantage of the proposed IJDA mechanism, the IJDA that includes the original Softmax loss and the IJDA loss is also tested by the six transfer tasks. The backbone network and the training rule of these compared methods are the same as the DDTLN.

Six cross-machine transfer tasks are implemented to demonstrate the diagnostic accuracy and robustness of DDTLN. To ensure the reliability of DDTLN, each method is performed ten times within each transfer task. Average diagnostic accuracies and corresponding standard deviations of ten methods are illustrated in Table 3. The average accuracy of the proposed IJDA mechanism is 6.37 % higher than the original IDA mechanism, which reflects the effectiveness of the IJDA mechanism. In addition, for showing the effectiveness of the proposed DDM and I-Softmax, the ablation experiments are performed. Without using the I-Softmax, IJDA (MMD), IJDA(CORAL) and IJDA (DDM), which are respectively based on MMD, CORAL and DDM, are applied to the fault transfer diagnosis. From the Table 3, we can clearly know that the proposed metric DDM has a better performance in IJDA mechanism. In particular, the average accuracy of the proposed DDTLN is over 90 %, and it is 30.83 % higher compared with other methods. It should be noted that the DDTLN is the highest in each transfer task. In conclusion, the proposed DDTLN method has a better diagnostic ability than typical DA methods.

To intuitively show the advantage of DDTLN, *t*-distributed stochastic neighbor embedding (*t*-SNE) [32] is used to map the learned high-dimensional features into two-dimension space. For task A → B, the *t*-SNE maps obtained by five models are illustrated in Fig. 11. The proposed DDTLN model can obtain the minimum intra-class distance and the maximum inter-class distance. This is mainly because DDTLN can better align the marginal and conditional distributions from the target and source domains than the existing DA models. In other words, DDTLN can learn more category-discriminative and domain-invariant features. The comparative results further prove that the DDTLN model has higher accuracy than typical DA methods.

4.5. Further experimental study

Although DDTLN model shows its excellent diagnosis performance on three bearing datasets, the faults in these datasets are produced by artificial machining, whose shapes are usually regular. It follows that the fault impacts in the three datasets may be similar. IMS public dataset [33] is a well-known open dataset, which was collected by the University of Cincinnati. In the tests, the radial load (6000 lbs) was directly applied to the shaft and the bearing via a spring mechanism, the sampling rate was set as 20000 Hz, and the input speed was 2000 r/min. Compared with A, B and C, the faults in IMS are naturally produced during the bearing life-cycle tests, and their shapes are irregular. Thus, IMS has a big difference with A, B and C. To further evaluate the effectiveness and advantage of DDTLN, IMS which has the actual faults is used for building other six cross-machine transfer tasks, and they consist of IMS → A, A → IMS, IMS → B, B → IMS, IMS → C, and C → IMS. Similarly, the samples obtained under four healthy conditions (NC, IF, BF, and OF) in the IMS dataset are used.

The experiment results are shown in Table 4. It can be seen from the Table 4 that the average accuracy of DDTLN obviously outperform other diagnosis models, and its diagnostic accuracy is over 84 %. However, it is 5.97 % lower than that in Table 4. This may be because that the faults in the IMS dataset are irregular and have big differences with those in A, B and C. The comparative results again verify that the MWSAN model possesses a stronger generalization ability than other diagnosis models in the cross-machine transfer diagnosis.

5. Conclusions

In this paper, a new transfer learning network named DDTLN was proposed to implement the cross-machine fault diagnosis. The DDTLN mainly consists of the IJDA mechanism and an I-Softmax loss. In IJDA, a new distribution discrepancy metric composed of MMD and CORAL was constructed to enhance domain confusion. Furthermore, an improved CDA mechanism was proposed to boost the distribution matching between the source domain and the target domain to a larger degree. Compared with the original Softmax, the I-Softmax loss has a stronger ability in learning more separable features. In addition, it can flexibly control the decision boundary and can be conveniently optimized. With the IJDA mechanism and I-Softmax loss, the DDTLN obtains more separable but domain-invariant features. The DDTLN can achieve an average accuracy of over 90 % in six cross-machine transfer tasks. Lastly, experimental results also verify that DDTLN has a stronger diagnostic ability than the well-known DA methods. The related code can be downloaded from https://qinyi-team.github.io/#blog.

This research has the following limitations, including the interpretability of DDTLN and the evaluation of transferability between the source and target domains. In future works, we will combine some signal processing algorithms into the transfer learning neural network for enhancing its interpretability, and explore how to evaluate the transferability between two domains.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

Table 3
Experimental results.

Methods	Cross-machine transfer tasks (%)						Avg
	A ± B	B ± A	A ± C	C ± A	B ± C	C ± B	
DDC	78.54 ± 4.10	82.05 ± 4.97	51.57 ± 4.57	52.00 ± 4.56	55.13 ± 5.37	38.87 ± 4.36	59.69
DCORAL	80.29 ± 6.06	91.19 ± 2.70	39.19 ± 3.73	43.50 ± 6.00	45.39 ± 7.14	48.52 ± 4.96	58.01
DANN	65.23 ± 8.34	52.61 ± 5.14	35.69 ± 8.16	40.68 ± 6.34	39.54 ± 6.59	51.31 ± 7.68	47.51
MCD	66.65 ± 4.21	50.61 ± 5.05	40.51 ± 4.22	44.15 ± 4.12	30.69 ± 6.87	31.52 ± 6.32	44.02
FTNN	81.34 ± 3.67	82.66 ± 2.91	43.87 ± 3.94	50.05 ± 2.98	56.87 ± 4.39	40.52 ± 4.95	59.22
JDA	78.93 ± 4.57	84.42 ± 3.68	54.40 ± 3.87	43.29 ± 5.00	54.63 ± 4.44	41.49 ± 4.41	59.53
IJDA(MMD)	82.73 ± 5.85	83.22 ± 2.99	56.35 ± 4.35	55.57 ± 5.97	55.64 ± 6.08	39.57 ± 5.66	62.18
IJDA(CORAL)	88.83 ± 3.54	85.01 ± 2.98	56.74 ± 2.99	49.57 ± 4.31	51.89 ± 5.62	47.03 ± 5.81	63.18
IJDA(DDM)	84.17 ± 7.15	88.39 ± 2.01	56.04 ± 3.98	54.43 ± 5.21	57.42 ± 6.51	54.95 ± 5.26	65.90
DDTLN	**98.08 ± 2.33**	**95.12 ± 1.92**	**82.36 ± 4.35**	**87.09 ± 6.69**	**91.55 ± 6.32**	**88.92 ± 8.01**	**90.52**

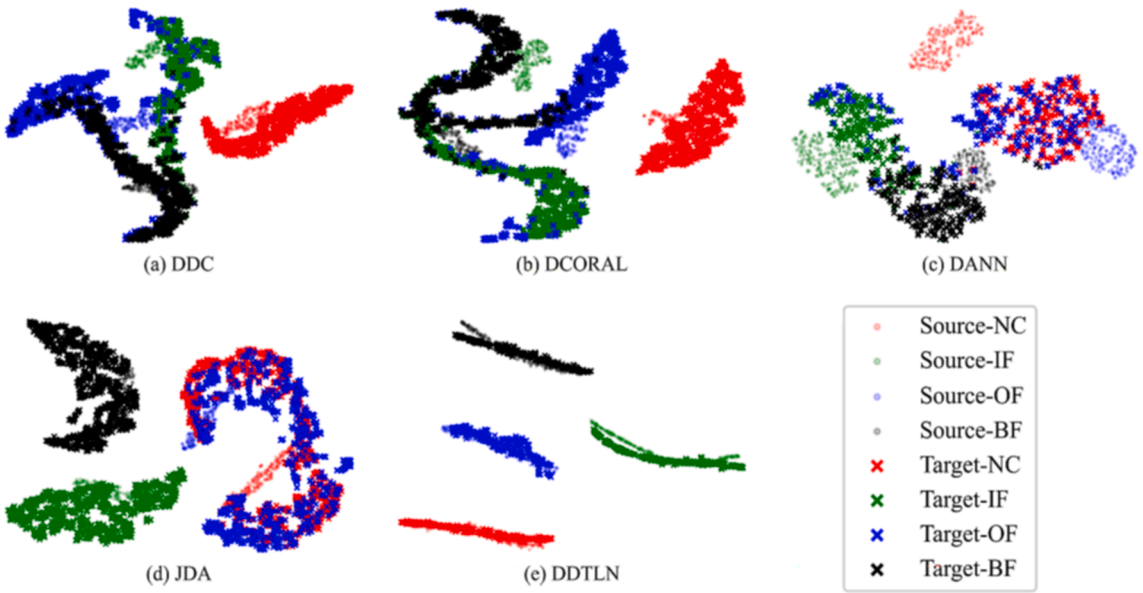

Fig. 11. The *t*-SNE mappings of learned features obtained by five DA models.

Table 4
Experimental results based on IMS.

Methods	Cross-machine transfer tasks (%)						Avg
	IMS ± A	A ± IMS	IMS ± B	B ± IMS	IMS ± C	C ± IMS	
DDC	65.72 ± 1.23	60.81 ± 1.21	81.56 ± 1.14	76.14 ± 2.11	39.64 ± 1.65	33.10 ± 1.79	59.50
DCORAL	63.74 ± 1.55	68.31 ± 1.66	**84.31 ± 1.24**	73.61 ± 1.05	35.04 ± 1.10	31.35 ± 1.33	59.39
DANN	64.99 ± 1.87	58.37 ± 1.98	63.10 ± 1.99	70.33 ± 1.67	32.65 ± 2.21	28.41 ± 1.97	52.98
MCD	66.12 ± 2.34	60.63 ± 2.22	68.41 ± 1.65	67.55 ± 2.13	33.14 ± 1.79	30.27 ± 1.86	54.35
FTNN	63.53 ± 1.20	58.05 ± 1.36	79.36 ± 1.33	74.66 ± 1.54	38.87 ± 1.06	40.02 ± 1.63	59.08
JDA	68.36 ± 1.43	61.53 ± 1.28	77.14 ± 1.24	72.51 ± 1.39	33.14 ± 1.82	35.41 ± 1.75	57.68
DDTLN	**81.15 ± 2.57**	**85.38 ± 3.84**	83.49 ± 2.88	**88.60 ± 3.44**	**83.01 ± 2.08**	**85.67 ± 2.25**	**84.55**

influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

The work described in this paper was supported by the National Natural Science Foundation of China (nos. 52175075 and

62033001), Chongqing Research Program of Basic Research and Frontier Exploration (no. cstc2021ycjh-bgzxm0157), and Open Foundation of State Key Laboratory of Mechanical Transmission, Chongqing University, China (SKLMT-MSKFKT-202020).

References

[1] Y. Qin, Q. Qian, Y. Wang, J. Zhou, Intermediate distribution alignment and its application into mechanical fault transfer diagnosis, IEEE Trans. Ind. Inf. (2022).
[2] K. Zhong, M. Han, T. Qiu, B. Han, Fault diagnosis of complex processes using sparse kernel local Fisher discriminant analysis, IEEE Trans. Neural Networks Learn. Syst. 31 (2019) 1581–1591.
[3] Y. Qin, X. Wu, J. Luo, Data-model combined driven digital twin of life-cycle rolling bearing, IEEE Trans. Ind. Inf. 3 (18) (2022) 1530–1540.
[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (2017) 2030–2096.
[5] Y. Feng, J. Chen, S. He, T. Pan, Z. Zhou, Globally localized multisource domain adaptation for cross-domain fault diagnosis with category shift, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1/15.
[6] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance, arXiv preprint arXiv:1412.3474, (2014).
[7] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, Int. Conf. Mach. Learn. (2015) 97–105.
[8] B. Sun, K. Saenko, Deep CORAL: Correlation Alignment for Deep Domain Adaptation, European conference on computer vision, 2016, pp. 443-450.
[9] L. Wen, L. Gao, X. Li, A new deep transfer learning based on sparse auto-encoder for fault diagnosis, IEEE Trans. Syst. Man Cybernet. Syst. 49 (2019) 136–144.
[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inform. Process. Syst. 27 (2014) 7354–7363.
[11] Q. Qian, Y. Qin, Y. Wang, F. Liu, A new deep transfer learning network based on convolutional auto-encoder for mechanical fault diagnosis, Measurement 178 (2021), 109352.
[12] W. Li, Z. Chen, G. He, A novel weighted adversarial transfer network for partial domain fault diagnosis of machinery, IEEE Trans. Ind. Inf. 17 (2020) 1753–1762.
[13] C. Shen, X. Wang, D. Wang, Y. Li, J. Zhu, M.J.I.T.o.I. Gong,, Dynamic joint distribution alignment network for bearing fault diagnosis under variable working conditions, IEEE Trans. Instrum. Meas. 70 (2021) 1–13.
[14] P. Panareda Busto, J. Gall, Open set domain adaptation, Proceedings of the IEEE international conference on computer vision, 2017, pp. 754-763.
[15] W. Zhang, X. Li, H. Ma, Z. Luo, X. Li, Universal domain adaptation in fault diagnostics with hybrid weighted deep adversarial learning, IEEE Trans. Ind. Inf. 17 (2021) 7957–7967.
[16] Z. Chai, C. Zhao, B. Huang, Multisource-refined transfer network for industrial fault diagnosis under domain and category inconsistencies, IEEE Trans. Cybern. (2021) 1–13.
[17] M. Ragab, Z. Chen, M. Wu, H. Li, C.-K. Kwoh, R. Yan, X. Li, Adversarial multiple-target domain adaptation for fault classification, IEEE Trans. Instrum. Meas. 70 (2020) 1–11.
[18] M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, Transfer Feature Learning with Joint Distribution Adaptation, Proceedings of the 2013 IEEE International Conference on Computer Vision, 2013, pp. 2200-2207.
[19] W. Liu, Y. Wen, Z. Yu *et al.*, Sphereface: Deep hypersphere embedding for face recognition, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 212-220.
[20] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, International Conference on Machine Learning, 2016, pp. 7.
[21] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2005, pp. 539-546.
[22] Z. Wu, H. Jiang, T. Lu, K. Zhao, A deep transfer maximum classifier discrepancy method for rolling bearing fault diagnosis under few labeled data, Knowl.-Based Syst. 196 (2020), 105814.
[23] M. Lucic, K. Kurach, M. Michalski et al., Are gans created equal? a large-scale study, arXiv preprint arXiv:1711.10337, (2017).
[24] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5265-5274.
[25] R. Ranjan, C.D. Castillo, R. Chellappa, L2-constrained softmax loss for discriminative face verification, arXiv preprint arXiv:.09507, (2017).
[26] W.A. Smith, R.B. Randall, Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study, Mech. Syst. Sig. Process. 64 (2015) 100–131.
[27] X. Hong, H. Gao, Y. Sun, H. Song, Q. Liu, Ball Screw Stability Degradation Stages Evaluation Based on Deep Residual Neural Network and Multi-sensor Fusion, 2018 Prognostics and System Health Management Conference (PHM-Chongqing), IEEE, 2018, pp. 785-790.
[28] X. Liang, X. Wang, Z. Lei, S. Liao, S.Z. Li, Soft-margin softmax for deep classification, in: International Conference on Neural Information Processing, Springer, 2017, pp. 413–421.
[29] Z. Wu, H. Jiang, T. Lu, K. Zhao, A deep transfer maximum classifier discrepancy method for rolling bearing fault diagnosis under few labeled data, Knowl.-Based Syst. 196 (2022), 105814.
[30] B. Yang, Y. Lei, F. Jia, S. Xing, An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings, Mech. Syst. Sig. Process. 122 (2020) 692–706.
[31] T. Han, C. Liu, W. Yang, D. Jiang, Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application, ISA Trans. 97 (2020) 269–281.
[32] V.D.M. Laurens, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[33] H. Qiu, J. Lee, J. Lin, G. Yu, Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics, J. Sound Vibrat. 289 (2006) 1066–1090.