Chen *et al. Chin. J. Mech. Eng.* (2021) 34:13
https://doi.org/10.1186/s10033-020-00520-9

Chinese Journal of Mechanical Engineering

ORIGINAL ARTICLE

Open Access

Model Parameter Transfer for Gear Fault Diagnosis under Varying Working Conditions

Chao Chen[1], Fei Shen[1], Jiawen Xu[1] and Ruqiang Yan[1,2*]

Abstract

Gear fault diagnosis technologies have received rapid development and been effectively implemented in many engineering applications. However, the various working conditions would degrade the diagnostic performance and make gear fault diagnosis (GFD) more and more challenging. In this paper, a novel model parameter transfer (NMPT) is proposed to boost the performance of GFD under varying working conditions. Based on the previous transfer strategy that controls empirical risk of source domain, this method further integrates the superiorities of multi-task learning with the idea of transfer learning (TL) to acquire transferable knowledge by minimizing the discrepancies of separating hyperplanes between one specific working condition (target domain) and another (source domain), and then transferring both commonality and specialty parameters over tasks to make use of source domain samples to assist target GFD task when sufficient labeled samples from target domain are unavailable. For NMPT implementation, insufficient target domain features and abundant source domain features with supervised information are fed into NMPT model to train a robust classifier for target GFD task. Related experiments prove that NMPT is expected to be a valuable technology to boost practical GFD performance under various working conditions. The proposed methods provides a transfer learning-based framework to handle the problem of insufficient training samples in target task caused by variable operation conditions.

Keywords: Gear fault diagnosis, Model parameter transfer, Varying working conditions, Least square support vector machine

1 Introduction

Gear has been used extensively in transmission system due to its large velocity ratio, strong bearing capacity, compactness and high efficiency [1–4]. Gear fault diagnosis (GFD) also becomes one of the most important research hotspots from both industrial and academic communities for ensuring the safe and efficient operation of gear transmission system. Till now, with the development of sensing methods (e.g., vibration, rotor speed, acoustic signal and others), data-driven methods [5, 6], which are based on analyzing measured data without need of a deep understanding of the mechanical drive systems, have become more and more attractive and been proved to be valid in the field of gear fault diagnosis. Generally, there are two steps in data-driven method: (1) constructing a classification model based on sampled data, and (2) using the well-trained model to predict the mechanical fault type. In many existing researches, the fault diagnostic task can be treated as a problem of pattern recognition, which usually is composed of two technical processes: (1) feature extraction, and (2) fault recognition. The purpose of feature extraction is to obtain low-dimensional fault descriptors from high-dimensional vibration data. There are many advanced signal processing methods that have been proposed to provide cognizable features, such as wavelet transform (WT) [7], principal components analysis (PCA) [8], singular value decomposition (SVD) [9], empirical mode decomposition

*Correspondence: ruqiang@seu.edu.cn
[1] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
Full list of author information is available at the end of the article

(EMD) [10], etc. Then, conventional machine learning methods (e.g., extreme learning machine, support vector machine and neural network) are employed to build a gear fault diagnostic model. However, these conventional methods usually work for GFD under constant speed and load conditions, thus having weak generality when facing with variable working conditions. Generally, gears are often working under time-varying operating conditions, for example, the running states of gas turbines or wind power generators change very often while working, and the operation parameters of planetary gearbox may vary correspondingly, thus inevitably resulting in a consequence that the extracted features in one time period might be different from those in the next time period. More importantly, the identical and independent distribution (IID) between training data and test data is required to ensure effective implementation of these conventional machine learning methods. Recently, these problems have aroused researchers' interest and received intensive attentions. For example, Song et al. [11] developed a new singular value decomposition interpolation (SVDI) based signal processing method, in which the time-domain and frequency-domain characteristic matrices extracted from vibration signals under discrete working conditions were firstly decomposed into singular vectors, rotation matrices and characteristic means with SVD, then these three parts were interpolated to reconstruct the target eigenmatrix for data augmentation. Han et al. [12] utilized empirical mode decomposition (EMD) to decompose vibration signals into several intrinsic mode functions (IMF), and extracted feature vectors that consist of time domain indexes, frequency domain indexes, energy domain characteristic parameter and fractal box dimension from the selected IMF to investigate the dynamic feature of vibration signal accurately and improve the robustness of feature vectors under different loads for GFD. Meanwhile, Zhao et al. [13] designed a synchrosqueezing transform (SST) and deep convolutional neural network (DCNN) based method for gearbox fault classification under varying operation conditions, where a new index, the envelope time-frequency representation (TFR), was calculated by using SST, then DCNN was adopted to dig underlying features of the TFRs and determine the fault type of planetary gearbox automatically. In general, most of these methods can achieve good results by exploring advanced feature extraction methods or building a complex network classifier, but they rely on sufficient labeled training dataset normally, which could degrade performance when facing with insufficient data. However, only a few number of labeled samples collected for training probably exist in many real-world applications, which hinder the promotion of these methods greatly.

Therefore, how to train a robust model with high accuracy under limited labeled data is important. Recently, transfer learning (TL), a fast-growing filed of machine learning, has been emerging due to its knowledge transfer ability [14]. To be delighted, the amount of labeled target data (termed as target domain, TD) maybe small, but there are still plenty of relevant data which can be obtained in machine industry from another time period (e.g., under another speed and load) or adjacent components (termed as source domain, SD). By utilizing the TL technology, useful information can be extracted from existing or previous task to boost the learning efficiency of target task. The model parameter transfer (MPT), one of the transfer learning architectures, is an effective tool to transfer the shared parameters or prior distributions of hyperparameters. Recently, most of these approaches are designed to work for multitask learning (MTL). For example, Lawrence et al. [15] succeeded in learning parameters from multiple tasks through the shared Gaussian process (GP) prior. Bonilla et al. [16] proposed a GP-based model to learn the shared model knowledge over tasks. Schwaighofer et al. [17] succeed in learning multi-tasks by utilizing the combination of hierarchical Bayesian framework (HB) and GP. Besides, Evgenious et al. [18] proposed a new algorithm by referencing HB idea to solve multitask learning in the frame of support vector machine (SVM). All these methods can be easily modified for TL. Strictly speaking, MTL tries to learn different tasks jointly and simultaneously, while TL prefer to improve the performance of TD task with the help of knowledge extracted and stored from SD data. Comparison between MTL and TL is shown in Figure 1. Intuitively, we may minimize the difference in parameters of classification hyperplane between TD and SD to transfer the knowledge obtained from SD, so that a robust GFD model with better performance in TD can obtained.

According to the above analysis, a novel model parameter transfer (NMPT) approach, which aims at excavating and further transferring the shared characteristic parameters of hyperplane for the problems of insufficient labeled training samples and non-IID between source and target domains, is developed to assist target gear fault identification using source domain data. Specifically, on this basis of controlling the empirical risk of source domain, the proposed method further integrates the advantage of the conventional MPT and TL together, which can be concluded that: (a) the least square support vector machine (LSSVM) based MPT can characterize the shared and domain-specific parameters over tasks; and (b) the idea of TL is introduced to dig and extract transferable knowledge and to minimize the

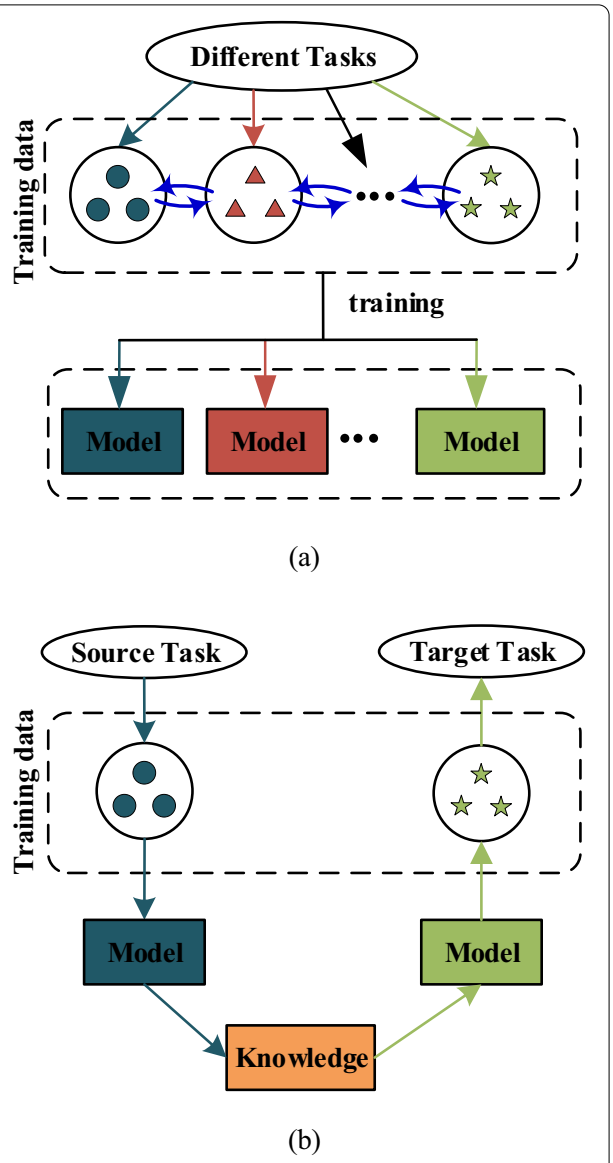

Figure 1 Comparison of multitask learning (MTL) and transfer learning (TL): **a** MTL; **b** TL

distributional discrepancies between source and target domains. To sum up, the novelties and main contributions of this paper can be summarized as:

- Based on controlling the empirical risk of source domain features in LSSVM framework, an improved TL model is proposed by further minimizing the discrepancies of separating hyperplanes between source and target domains, and then transferring both shared and domain-specific parameters over tasks to make use of source domain data to assist target diagnostic task;
- The model parameter transfer idea is innovatively introduced to the area of gear fault diagnosis, which provides a new idea for gear fault diagnosis under variable working conditions, especially when sufficient training data from target domain are not available.

The rest of this paper is organized as follows. In Section 2, the theoretical background is briefly presented. Section 3 concentrates on introducing details of the proposed NMPT method and then gives the whole framework of GFD. Section 4 illustrates the experimental study and proves that NMPT can achieve good results in GFD under variable working conditions. Finally, some conclusions drawn from this paper are listed in Section 5.

2 Theoretical Background

This study is going to leverage the NMPT model under LSSVM framework for GFD. Therefore, in this section, the fundamental theory of LSSVM as well as its improvement for MTL are briefly reviewed.

2.1 Least Squares Support Vector Machine (LSSVM)

First, the basic principle of training a SVM-based model for classification problem is to find the optimal separating hyperplane ($f = \boldsymbol{w}^*\phi(x) + b$) in a reproducing kernel Hilbert space (RKHS) [19]. According to structural risk minimization (SRM) principle, the optional $\boldsymbol{w}$ and b can be obtained by minimizing the following function:

$$\min\ R = \frac{1}{2}\|\boldsymbol{w}\|^2 + C \times R_{\text{emp}}, \tag{1}$$

where C is positive real regularized parameter, $\boldsymbol{w}$ is weight vector defining the orientation of separating hyperplane, R represents structural risk, R_{emp} denotes loss function which controls the error of separating hyperplane f on training data, and different kinds of R_{emp} can contribute to different forms of SVMs. By utilizing squared error function, the SRM problem in LSSVM is to compute the optimal decision-made separating hyperplane according to the vector $\boldsymbol{x}$ and its label $y\in\{-1,+1\}$ by minimizing the following function with a constraint, which can be formulated as:

$$\begin{aligned} &\min_{\omega,e,d}\ J(\boldsymbol{w},e) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \tfrac{C}{2}\sum_{i=1}^{N} e_i^2, \\ &\text{s.t.},\ y_i\{\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i) + b\} = 1 - e_i, \quad i = 1,2,\ldots,N, \end{aligned} \tag{2}$$

where e_i is error function, $\phi(\cdot)$ denotes a transform function that maps the input features x into RKHS, b is a bias term, N indicates the total number of training samples. Then a classification hyperplane $f = \boldsymbol{w}^*\phi(x) + b$ is constructed for this task.

2.2 Multi-Task LSSVM (MTLSSVM)

Given m learning tasks, the MTL aims to learn all tasks simultaneously rather than individually. Let each task $\forall i \in m$, we have n_i training samples $\{\boldsymbol{x}_{i,j}, y_{i,j}\}_{j=1}^{n_i}$, thus the total number of training samples is $N = \sum_{i=1}^{m} n_i$.

Based on the regularization framework and hierarchical Bayesian framework, some researchers assumed that all $\boldsymbol{w}_i$ can be rewritten as $\boldsymbol{w}_i = \boldsymbol{w}_0 + \boldsymbol{v}_i$, where $\boldsymbol{w}_0$ (playing the role of mean vector) and $\boldsymbol{v}_i$ carry the information of commonality and specialty over tasks [20, 21], respectively. That is to say, when m learning tasks are analogous to each other, the vectors $\boldsymbol{v}_i$ tend to be "small", otherwise, the vector $\boldsymbol{w}_0$ tends to be "small". To this end, the following optimization problem which is similar to LSSVM for single task is solved to estimate all $\boldsymbol{v}_i$ as well as $\boldsymbol{w}_0$ simultaneously:

$$\begin{aligned}
&\min_{\boldsymbol{w},e,d} \; J(\boldsymbol{w}_0, \{\boldsymbol{v}_i\}_{i=1}^{m}, \{\boldsymbol{e}_i\}_{i=1}^{m}) \\
&= \frac{1}{2}\|\boldsymbol{w}_0\|^2 + \frac{1}{2} \times \frac{\lambda}{m} \sum_{i=1}^{m} \|\boldsymbol{v}_i\|^2 + \frac{C}{2} \sum_{i=1}^{m} \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{e}_i, \\
&\text{s.t.,} \; (\boldsymbol{w}_0 + \boldsymbol{v}_i)^{\mathrm{T}} \boldsymbol{Z}_i + b_i \boldsymbol{y}_i = \boldsymbol{1}_{n_i} - \boldsymbol{e}_i, \quad i = 1, 2, \cdots, m,
\end{aligned} \tag{3}$$

where C and λ are positive real regularized parameters, $\boldsymbol{b} = \{b_1, b_2, \cdots, b_m\}^{\mathrm{T}}$, $\boldsymbol{e}_i = \{e_{i,1}, e_{i,2}, \cdots, e_{i,n_i}\}^{\mathrm{T}}$, $\boldsymbol{Z}_i = \{\varphi(\boldsymbol{x}_{i,1}) y_{i,1}, \varphi(\boldsymbol{x}_{i,2}) y_{i,2}, \cdots, \varphi(\boldsymbol{x}_{i,n_i}) y_{i,n_i}\}$, $\boldsymbol{y}_i = \{y_{i,1}, y_{i,2}, \cdots, y_{i,n_i}\}^{\mathrm{T}}$.

These previous works of LSSVM and MTLSSVM are not oriented to the target task where there exists the problem of insufficient training data or non-IID between training and testing data. Whereas, it is significant to derive useful information from these existed models to enhance the TD task. Therefore, different from the single task learning and multitask learning, the proposed NMPT utilizes SD data (related but different from TD) to solve target domain problems with a specific structure, which is introduced in the following section.

3 Proposed NMPT Framework for GFD

The proposed NMPT method via transferring the knowledge of classification hyperplane from SD to TD is presented in this section.

3.1 Basic Definition

Given SD and TD, the main purpose of NMPT can be described as: under LSSVM framework, NMPT aims to improve the performance of TD classification model $\boldsymbol{f}_t = \boldsymbol{w}_t^* \boldsymbol{x}_t + b_t$ by using the knowledge from source domain classifiers model $\boldsymbol{f}_s = \boldsymbol{w}_s^* \boldsymbol{x}_s + b_s$, where the SD and TD are different but similar in some aspects. In addition, the training data is set as follows:

$$\begin{aligned}
Ds &= \left\{ (\boldsymbol{x}_j^s, y_j^s) \right\}, j = 1, 2, \cdots, Ns, \\
Dt &= \left\{ (\boldsymbol{x}_i^t, y_i^t) \right\}, i = 1, 2, \cdots, Nt,
\end{aligned} \tag{4}$$

where Ds, Dt are SD and TD labeled data, respectively; $\boldsymbol{x}_j^s$, y_j^s denote the jth feature vector and corresponding label of SD data; $\boldsymbol{x}_i^t$, y_i^t denote the ith feature vector and corresponding label of TD data; Ns and Nt represent the number of SD and TD, in this paper, $Nt << Ns$.

3.2 NMPT Architecture

In this section, the proposed NMPT approach is discussed. As mentioned above, the method mainly utilizes the labeled data from SD and TD to solve the target GFD problem. First, inspired by the work of multitask LSSVM framework [21, 22], we assume that the parameters, $\boldsymbol{w}_t$ and $\boldsymbol{w}_s$ form both tasks can be separated into two parts, respectively:

$$\boldsymbol{w}_{\mathrm{t}} = \boldsymbol{w}_0 + \boldsymbol{v}_{\mathrm{t}}, \boldsymbol{w}_{\mathrm{s}} = \boldsymbol{w}_0 + \boldsymbol{v}_{\mathrm{s}} \tag{5}$$

where $\boldsymbol{w}_0$ is the shared parameter, $\boldsymbol{v}_s$ and $\boldsymbol{v}_t$ are the domain-specific parameters of SD and TD tasks, respectively. Then, based on previous transfer strategy that controls empirical risk of source domain, we want to find the knowledge from $\boldsymbol{w}_s$ and transfer it to $\boldsymbol{w}_t$ ulteriorly. As enough training data can prevent the model from overfitting, parameter $\boldsymbol{w}_0$ from $\boldsymbol{w}_s$ is set as one of transfer knowledge. In addition, by minimizing the term $\mu || \boldsymbol{v}_t - \boldsymbol{v}_s ||^2$ during the optimization process, we can also recognize and apply knowledge of $\boldsymbol{v}_s$ learned from SD. Hence, to achieve this goal, an extension of LSSVM to transfer learning case is built as follows:

$$\begin{aligned}
&\min_{\boldsymbol{w},e,d} J(w_0, v_t, v_s, e) \\
&= \tfrac{1}{2}\|\boldsymbol{w}_0\|^2 + \tfrac{1}{2} \times \tfrac{\lambda}{2}\left(\|\boldsymbol{v}_t\|^2 + \|\boldsymbol{v}_s\|^2\right) + \tfrac{Ct}{2} \sum_{i=1}^{Nt} e_i^2 \\
&+ \tfrac{Cs}{2} \sum_{i=Nt+1}^{Ns+Nt} e_i^2 + \mu \|\boldsymbol{v}_t - \boldsymbol{v}_s\|^2, \\
&\text{s.t.,} y_i^t \{(\boldsymbol{w}_0 + \boldsymbol{v}_t)^{\mathrm{T}} \varphi(\boldsymbol{x}_i^t) + b_t\} = 1 - e_i, \; i = 1, 2, \cdots, Nt, \\
&y_j^s \{(\boldsymbol{w}_0 + \boldsymbol{v}_s)^{\mathrm{T}} \varphi(\boldsymbol{x}_j^s) + b_s\} = 1 - e_j, \; j = 1, 2, \ldots, Ns,
\end{aligned} \tag{6}$$

where $\boldsymbol{w}_0$ and $\mu || \boldsymbol{v}_t - \boldsymbol{v}_s ||^2$ are transfer learning items, Cs, Ct, λ and μ are positive real regularized parameters. An illustration that describes the diagram of NMPT is presented in Figure 2.

As less tagged target training data will cause the corresponding classification model to show some tendency towards performance degradation, the decision boundary with parameter $\boldsymbol{w}_t$ from target task could suffer from this problem. However, by utilizing the knowledge of $\boldsymbol{w}_s$ from source domain, NMPT architecture can ensure a relatively small generalization error on the target domain by mainly focusing on achieving the following goals: (1) learning a more accurate $\boldsymbol{w}_0$ for target domain; (2) reducing the difference of

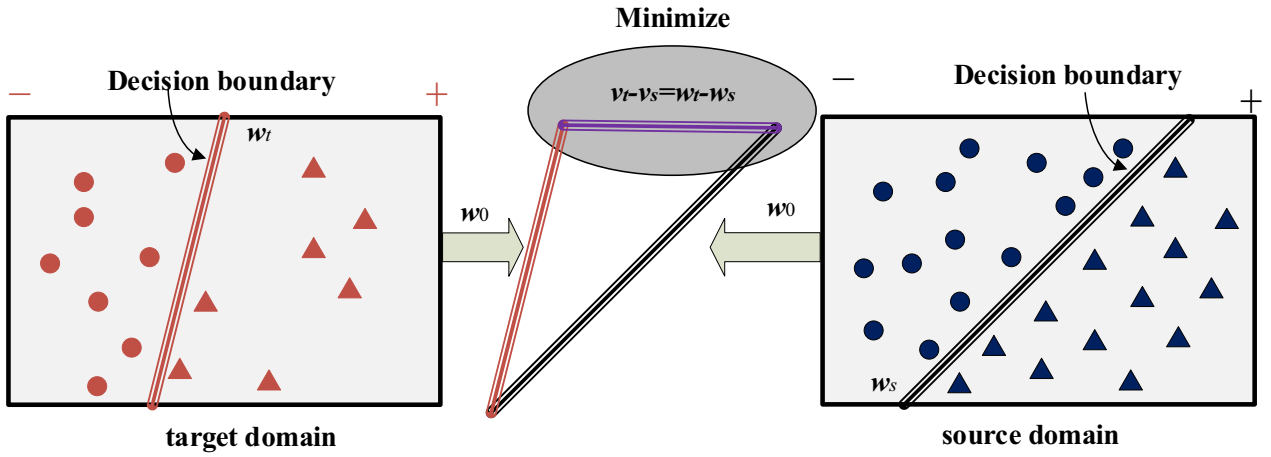

Figure 2 Schematic diagram of NMPT

model parameters by minimizing $\mu||\boldsymbol{v}_t-\boldsymbol{v}_s||^2$ (see the purple line in Figure 2). These two goals can make source domain model be applicable for target domain and ensure the leading role of *Dt* in building classification model for target task. In addition, by comparing eq. (2) with eq. (6), we find the NMPT model tries to make the separating hyperplane of SD be qualified for TD classification task from two aspects on the basis of SRM principle: one is to minimize the margin discrepancies of training data between SD and TD to adjust separating hyperplane, the other is to control loss function on SD data, simultaneously. All these two improvements can prove a good capability of generalization on TD.

Then, the solving process of NMPT optimization problem (c.f. Eq. (6)) is listed as follows:

First, the Lagrangian function for Eq. (6) is built as:

$$\begin{aligned}
&L(\boldsymbol{w}_0,\boldsymbol{v}_t,\boldsymbol{v}_s,b,e,a)\\
&=\frac{1}{2}\|\boldsymbol{w}_0\|^2+\frac{1}{2}\times\frac{\lambda}{2}\left(\|\boldsymbol{v}_t\|^2+\|\boldsymbol{v}_s\|^2\right)+\frac{Ct}{2}\sum_{i=1}^{Nt}e_i^2\\
&+\frac{Cs}{2}\sum_{i=Nt+1}^{Ns+Nt}e_i^2+\mu\|\boldsymbol{v}_t-\boldsymbol{v}_s\|^2\\
&-\sum_{i=1}^{Nt}a_i\left\{y_i^t\{(\boldsymbol{w}_0+\boldsymbol{v}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i^t)+b_t\}-1+e_i\right\}\\
&-\sum_{i=Nt+1}^{Nt+Ns}a_i\left\{y_i^s\{(\boldsymbol{w}_0+\boldsymbol{v}_s)^{\mathrm{T}}\varphi(\boldsymbol{x}_i^s)+b_s\}-1+e_i\right\},
\end{aligned}\tag{7}$$

where a_i is a Lagrange multiplier. Then, according to Karush–Kuhn–Tucker (KKT) conditions, the solutions for optimality are yielded as:

$$\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{w}_0}=0\rightarrow\boldsymbol{w}_0=\sum_{i=1}^{Nt}a_iy_i^t\varphi(\boldsymbol{x}_i^t)+\sum_{i=Nt+1}^{Nt+Ns}a_iy_i^s\varphi(\boldsymbol{x}_i^s),\\
&\frac{\partial L}{\partial \boldsymbol{v}_t}=0\rightarrow\frac{\lambda}{2}\boldsymbol{v}_t+2\mu(\boldsymbol{v}_t-\boldsymbol{v}_s)-\sum_{i=1}^{Nt}a_iy_i^t\varphi(\boldsymbol{x}_i^t)=0,\\
&\frac{\partial L}{\partial \boldsymbol{v}_s}=0\rightarrow\frac{\lambda}{2}\boldsymbol{v}_s+2\mu(\boldsymbol{v}_s-\boldsymbol{v}_t)-\sum_{i=Nt+1}^{Nt+Ns}a_iy_i^s\varphi(\boldsymbol{x}_i^s)=0,\\
&\frac{\partial L}{\partial b_t}=0\rightarrow\sum_{i=1}^{Nt}a_iy_i^t=0,\\
&\frac{\partial L}{\partial b_s}=0\rightarrow\sum_{i=1}^{Ns}a_iy_i^s=0,\\
&\frac{\partial L}{\partial e_i}=0\rightarrow a_i=Ce_i,\\
&\frac{\partial L}{\partial a_i}=0\rightarrow\begin{cases}y_i^t\{(\boldsymbol{w}_0+\boldsymbol{v}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i^t)+b_t\}-1+e_i=0\\(i=1,2,\ldots,Nt)\\y_i^s\{(\boldsymbol{w}_0+\boldsymbol{v}_s)^{\mathrm{T}}\varphi(\boldsymbol{x}_i^s)+b_s\}-1+e_i=0\\(i=Nt+1,Nt+2,\ldots,Nt+Ns),\end{cases}
\end{aligned}\tag{8}$$

where $\boldsymbol{v}_t$ and $\boldsymbol{v}_s$ can be derived as:

$$\begin{aligned}\boldsymbol{v}_t &= \frac{\left(1+\frac{4\mu}{\lambda}\right)\boldsymbol{w}_0 - \sum_{i=Nt+1}^{Nt+Ns} a_i y_i^s \varphi(\boldsymbol{x}_i^s)}{\frac{\lambda}{2}+4\mu} = \frac{\frac{4\mu}{\lambda}\left(\sum_{i=1}^{Nt} a_i y_i^t \varphi(\boldsymbol{x}_i^t) + \sum_{i=Nt+1}^{Nt+Ns} a_i y_i^s \varphi(\boldsymbol{x}_i^s)\right) + \sum_{i=1}^{Nt} a_i y_i^t \varphi(\boldsymbol{x}_i^t)}{\frac{\lambda}{2}+4\mu},\\ \boldsymbol{v}_s &= \frac{\left(1+\frac{4\mu}{\lambda}\right)\boldsymbol{w}_0 - \sum_{i=1}^{Nt} a_i y_i^t \varphi(\boldsymbol{x}_i^t)}{\frac{\lambda}{2}+4\mu} = \frac{\frac{4\mu}{\lambda}\left(\sum_{i=1}^{Nt} a_i y_i^t \varphi(\boldsymbol{x}_i^t) + \sum_{i=Nt+1}^{Nt+Ns} a_i y_i^s \varphi(\boldsymbol{x}_i^s)\right) + \sum_{i=Nt+1}^{Nt+Ns} a_i y_i^s \varphi(\boldsymbol{x}_i^s)}{\frac{\lambda}{2}+4\mu}.\end{aligned} \tag{9}$$

By eliminating $\boldsymbol{w}_0$, $\boldsymbol{v}_t$, $\boldsymbol{v}_s$ and e_i through substitution, one linear system can be obtained as follows:

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{Y}_1 \\ \boldsymbol{Y} & \boldsymbol{\Omega} \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \bar{\boldsymbol{I}} \end{bmatrix}, \tag{10}$$

where $\boldsymbol{a} = [a_1, a_2, \cdots, a_{Nt}, a_{Nt+1}, \cdots, a_{Nt+Ns}]^{\mathrm{T}}$, $\boldsymbol{b} = [b_t, b_s]^{\mathrm{T}}$, $\boldsymbol{Y}_1 = [y_1^t, y_2^t, \cdots, y_{Nt}^t, y_1^s, y_2^s, \cdots, y_{Ns}^s]$, $\boldsymbol{I} = [1, 1, \cdots, 1]_{(Nt+Ns)\times 1}$, $\boldsymbol{0} = [0, 0]$, $\boldsymbol{Y} = blockdiag(\boldsymbol{y}_s, \boldsymbol{y}_t)$, $\boldsymbol{y}_t = [y_1^t, y_2^t, \cdots, y_{Nt}^t]^{\mathrm{T}}$, $\boldsymbol{y}_s = [y_1^s, y_2^s, \cdots, y_{Ns}^s]^{\mathrm{T}}$, $\boldsymbol{\Omega}$ is $(Nt + Ns) \times (Nt + Ns)$ symmetric matrix $\boldsymbol{\Omega} = \Omega_0 + \Omega_1 + \frac{1}{C}\boldsymbol{I}_{Nt+Ns}$, $\Omega_1 = blockdiag(\Omega_t, \Omega_s)$, K represents the kernel function, the detail element in Ω is defined as:

$$\begin{aligned}&\Omega_{0ij} = \left(1 + \frac{4\mu}{\lambda} \Big/ \left(\frac{\lambda}{2}+4\mu\right)\right) y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j),\ y_i, y_j \in \boldsymbol{Y}_1,\\ &\left(\boldsymbol{x}_i,\ \boldsymbol{x}_j \in \left[\boldsymbol{x}_1^t,\ \boldsymbol{x}_2^t, \cdots,\ \boldsymbol{x}_{Nt}^t,\ \boldsymbol{x}_1^s,\ \boldsymbol{x}_2^s \cdots,\ \boldsymbol{x}_{Ns}^s\right],\right)\\ &\Omega_{tij} = \frac{1}{\frac{\lambda}{2}+4\mu} y_i^t y_j^t K(\boldsymbol{x}_i^t, \boldsymbol{x}_j^t),\ i, j \in [1, Nt],\\ &\Omega_{sij} = \frac{1}{\frac{\lambda}{2}+4\mu} y_i^s y_j^s K(\boldsymbol{x}_i^s, \boldsymbol{x}_j^s),\ i, j \in [1, Ns].\end{aligned} \tag{11}$$

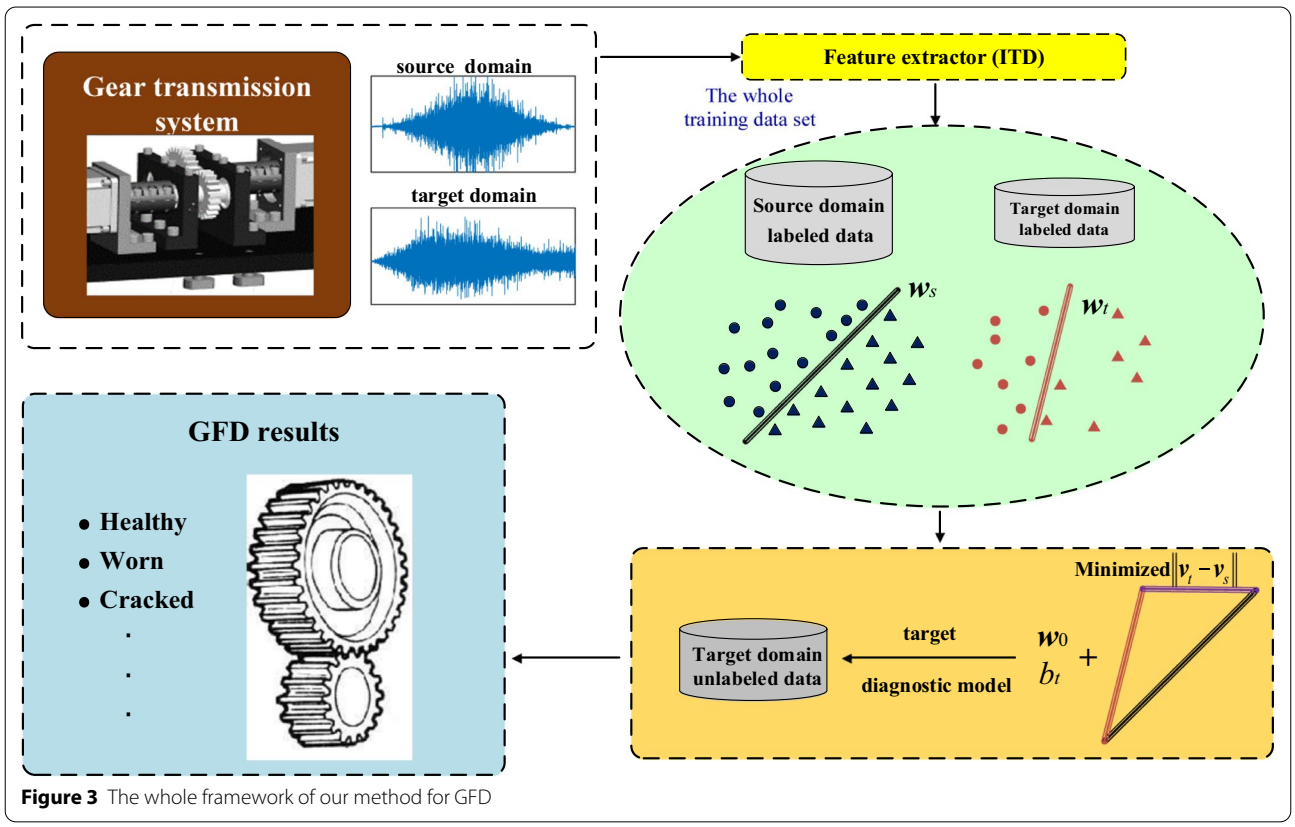

Figure 3 The whole framework of our method for GFD

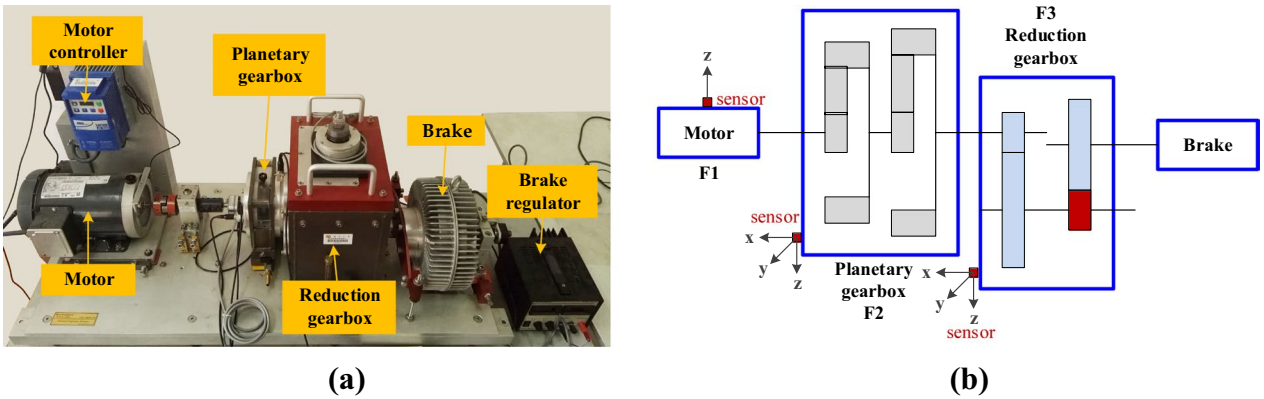

Figure 4 Spectra quest's drivetrain dynamics simulator: (a) The real chart of system; (b) The structure chart of system

Table 1 Gear fault type and working conditions

Fault types	C1 Healthy	C2 Chipped	C3 Missing	C4 Cracked	C5 Worn
Speeds (r/min)	S1 1200	S2 1800	S3 2400	–	
Load (N·m)	L1 0	L2 10.97	L3 14.63	–	
Fault location	F3 Reduction gearbox				

The best fit values of parameters $\boldsymbol{a}$, b_t and b_s can be finally worked out, then the corresponding decision function can be constructed as follows:

$$y = sgn\left[\left(1+\frac{4\mu}{\lambda}\Big/\left(\frac{\lambda}{2}+4\mu\right)\right)\times\left(\sum_{i=1}^{Nt}a_i y_i^t K(\boldsymbol{x}_i^t,\boldsymbol{x})+\sum_{i=Nt+1}^{Nt+Ns}a_i y_i^s K(\boldsymbol{x}_i^s,\boldsymbol{x})\right)+\frac{1}{\frac{\lambda}{2}+4\mu}\sum_{j=1}^{Nt}a_i y_j^t K(\boldsymbol{x}_j^t,\boldsymbol{x})+b_t\right]. \quad (12)$$

3.3 Complete Process of NMPT Model for Gear Fault Diagnosis

In the proposed framework, an intrinsic time-scale decomposition (ITD) architecture is first introduced to decompose a vibration signal into a set of proper rotation components (PRCs). Then, the energy parameter of each proper rotation component (PRC) is calculated to conduct dimensionality reduction and construct feature vectors. By structuring and solving the optimization problem of NMPT (c.f. Eq. (6)) using the learned fault representations, the parameters of NMPT model (including $\boldsymbol{w}_0$, $\boldsymbol{v}_s$ $\boldsymbol{v}_t$, b_s and b_t) can be learned simultaneously. Finally, the target data are fed into NMPT to output the predicted fault categories. Figure 3 gives the overall proposed framework for NMPT-based GFD.

4 Experiment and Discussion

4.1 Descriptions of Experimental Simulator and Datasets

To conduct experimental verification, the testing platform, drivetrain dynamics simulator (DDS), is shown in Figure 4. It includes driving motor, speed regulator, planetary gearbox, reduction gearbox, brake device, brake regulator. During data collection, the variety of speeds and loads can be implemented through speed regulator and brake regulator, respectively. Meanwhile, there are altogether 7 vibration sensors (model: 608A11, sample frequency: 5120 Hz) in the structure, one is mounted on the surface of motor to measure z-axial vibration signal of the motor (F1), the rest are as follows: three for planetary gearbox (F2) and three for reduction gearbox (F3). Except for the healthy gear (Healthy, C1), there are four different types of gear faults, denoted as a small piece of material breaking away from tooth (Chipped, C2), a tooth fracturing at the location of root (Missing, C3), the emergence of cracks on root cracked (Cracked, C4) and the loss of material from the contacting surface of tooth (Worn, C5). The descriptions of fault types and different experiment conditions are shown in Table 1.

4.2 Experimental Results and Analysis

4.2.1 Feature Extraction

Intrinsic time-scale decomposition (ITD) , proposed by Frei et al. [23], is a time frequency analysis method which can adaptively decompose a given vibration signal X into a series of proper rotation components (PRCs) and a

monotonous trend signal (remaining baseline signal) with low end effects and high efficiency, which can described as:

$$X=H^1+H^2+\cdots+H^p+L^p, \tag{13}$$

where p denotes the final decomposition level, H^i is the ith PRC, L^p is the remaining baseline signal.

Nevertheless, these obtained PRCs with ITD technology are too complex to be taken as fault vectors as inputs for conducting fault classification directly. Thus, the energies of first six level PRCs are calculated for dimensionality reduction of PRCs and fault feature design.

4.2.2 Experimental Study

In this part, the diagnostic performance of the proposed NMPT is first analyzed, then, in order to further demonstrate the superiority of NMPT, it is also compared with other methods:

LSSVM(non-transfer): Least squares support vector machine;
MTLSSVM (non-transfer): Multi-Task LSSVM;
TCA [24]: Transfer component analysis;
DSM [25]: Domain selection machine;
ELSSVM [26]: Enhanced LSSVM

For a fair comparison, all kernel-based methods use the Radial Basis Function (RBF) as the kernel function. In this study, 2000 sampled data points of original vibration signal under each specific working condition were fed into ITD model for feature extraction. Regardless in source or target domain, each gear fault category contains 200 samples under any chosen working condition. The datasets to perform experiments are set as follows: for LSSVM, 10 samples of each fault type are selected from target domain; for MTLSSVM and those transfer strategies, both the aforesaid 10 target domain samples and 100 source domain samples are arranged. Moreover, 100 testing samples from target domain are also arranged, and there is no overlap between training and testing samples in target domain. Therefore, the total size of training set is 50 and 550 for LSSVM and the rest methods, respectively; the total size of testing set is 500. In order to quantitatively describe the domain differences, the Kullback-Leibler (KL) divergence is calculated by:

$$\mathrm{KL}(Ds,Dt)=\frac{\mathrm{KL}(Ds||Dt)+\mathrm{KL}(Dt||Ds)}{2}, \tag{14}$$

where KL($\cdot||\cdot$) represents the KL divergence between Ds and Dt. Table 2 shows the descriptions of datasets (from DA1 to DA10) as well as their corresponding KL divergences. It shows that the KL indexes of all the data sets are larger than zero, which means there exists differences between SD and TD indeed. The signals that come from the same axis have relatively small KL divergence compared with those from different axes (e.g., transferring among different rotating speeds: DA1/DA3/DA4 vs DA2, different loads: DA5/DA7/DA8 vs DA8). Meanwhile, the KL divergence of nonadjacent mechanical components is larger than those adjacent to each other (DA10 vs DA9).

Table 2 Specific tests in experimental section

Test.	Source domain	Target domain	KL divergence
DA1	[S1, L1, F3]-x	[S2, L1, F3]-x	3.99
DA2	[S1, L1,F3]-x	[S2, L1, F3]-y	6.79
DA3	[S1, L1,F3]-z	[S2, L1, F3]-z	1.99
DA4	[S3, L1, F3]-z	[S2, L1, F3]-z	4.21
DA5	[S2, L2, F3]-x	[S2, L1, F3]-x	3.76
DA6	[S2, L2, F3]-x	[S2, L1, F3]-y	5.01
DA7	[S2, L2, F3]-z	[S2, L1, F3]-z	1.03
DA8	[S2, L3, F3]-z	[S2, L1, F3]-z	1.43
DA9	[S2, L1, F2]-z	[S2, L1, F3]-z	11.85
DA10	[S2, L1, F1]-z	[S2, L1, F3]-z	30.07

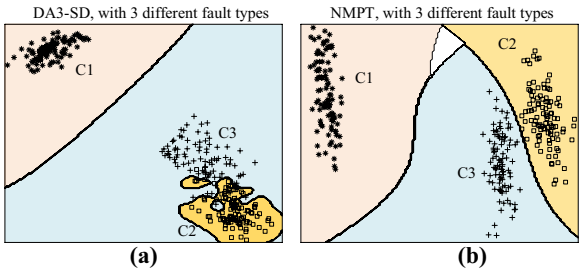

Figure 5 Classification hyperplane of DA3-SD: **a** Original DA3-SD; **b** After model parameter transfer

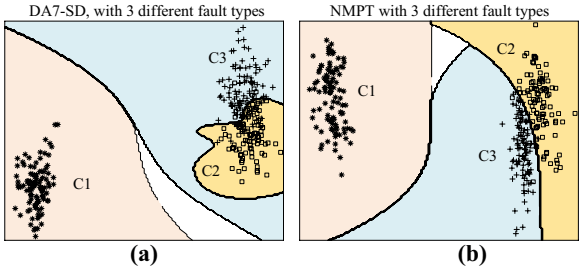

Figure 6 Classification hyperplane of DA7-SD: **a** Original DA7-SD; **b** After model parameter transfer

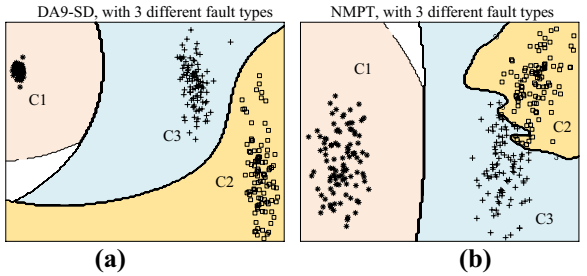

Figure 7 Classification hyperplane of DA9-SD: **a** Original DA9-SD; **b** After model parameter transfer

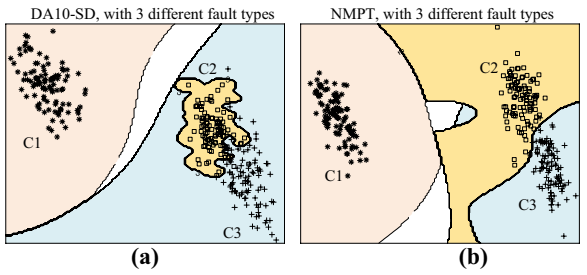

Figure 8 Classification hyperplane of DA10-SD: **a** Original DA10-SD; **b** After model parameter transfer

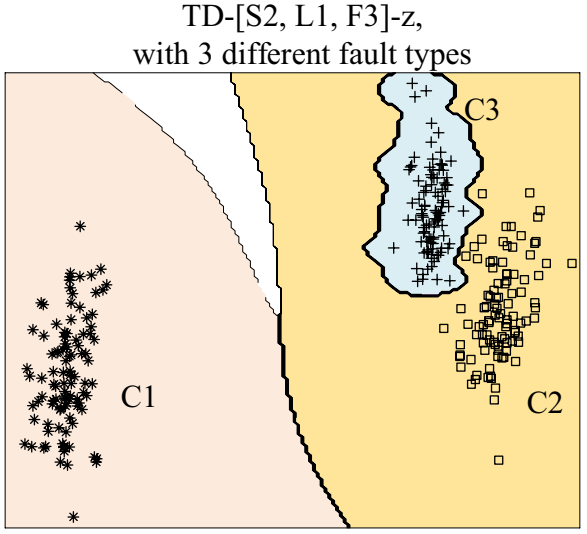

Figure 9 The original classification hyperplane of [S2,L1,F3]-z

First, Figures 5, 6, 7 and 8 give the visualized results of separating hyperplanes on four source domain datasets with three different fault types, including varying speeds (DA3), changing loads (DA7), adjacent mechanical parts (DA9 and DA10), to show the effectiveness of NMPT in minimizing the discrepancies of classification hyperplanes between SD and TD caused by operation conditions. Here, all datasets share the same target domain. By comparing these original classification hyperplanes, as is shown in Figure 5(a), Figure 6(a), Figure 7(a), Figure 8(a) and Figure 9, different working conditions can bring diversified results, which could easily cause erroneous diagnoses on target task when utilizing source domain samples as auxiliary training data directly. Whereas, NMPT tries to generalize the distinguishing ability from source domain to target domain, as shown in Figure 5(b), Figure 6(b), Figure 7(b) and Figure 8(b). Among them, Figure 5(b) and Figure 6(b) demonstrate similar results, which indicate that the proposed model are relatively more robust to transfer source domains from different speeds or loads compared with that from adjacent mechanical components.

Then, the performance of NMPT strategy for GFD from Test DA1 to DA10 are presented by confusion matrix, which are drawn in Figures 10, 11, and 12. In confusion matrix, the rows and columns show the actual and predicted fault types, respectively. The diagnostic accuracies of each fault type are shown in diagonal cells. Meanwhile, the misclassification rates are also listed outside the diagonal cells. Thus, from Figures 10, 11, 12 and Table 2, we can find that:

(1) Even though there exists relatively high domain differences between SD and TD in some data sets (e.g., DA9 and DA10), the NMPT model can still learn a precise classification for target task (e.g., Figure 12(a) and (b));

(2) The NMPT model investigated in this study shows very similar GFD accuracies among varying loads (from DA5 to DA8), similar conclusion can be found in changing speeds (from DA1 to DA4), which verify the robustness of NMPT to sensor axis factors. Meanwhile, the best performance of NMPT under different loads happens in diverse sensor axes (DA6). Whereas, transferring among the same axis can achieve performance improvement in the cases of varying rotating speeds (DA1 & DA3);

(3) The optimal classification performance occurs in the cases where source and target data come from the

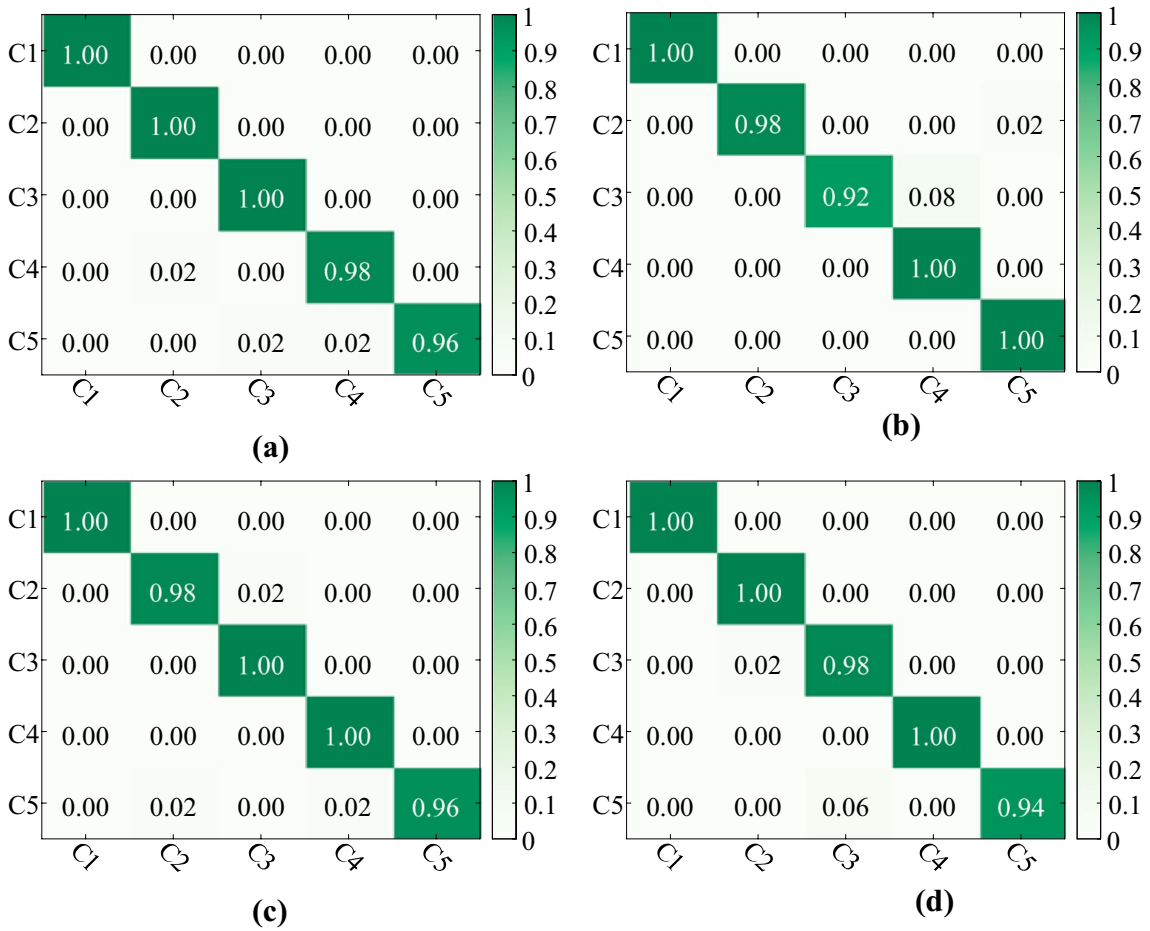

Figure 10 Confusion matrix of NMPT on transferring datasets with varying speeds

same gearbox (from DA1 to DA8), among them, the best classification accuracy of NMPT reaches 98.8% (DA1 & DA3). Besides, the performance of utilizing motor data to assist the fault recognition of reduction gearbox is lower than transferring between reduction gearbox and planetary gearbox;

(4) By comparing the accuracy and error rates in all data sets, there are many factors that can affect the model performance, among them, the mechanical components that contribute source data is the most crucial element.

In general, the classification accuracy of NMPT is always over 94%. Therefore, NMPT model can avoid overfitting of GFD under various working conditions by making reasonable use of abundant labeled data form another working condition or adjacent components.

After investigating the classification performances of NMPT method on all data sets, it is still meaningful to further compare NMPT with other methods. Table 3 lists the comparison results from DA1 to DA10, which are calculated over the whole categories. Among them, the classification performance of LSSVM model is the lowest mainly due to two things: (a) the LSSVM model is trained only by using the insufficient target domain samples, which will inevitably hinder the generalization performance according to the principles of structural risk minimization; and (b) the standard LSSVM model is lack of transferring knowledge among domains, while

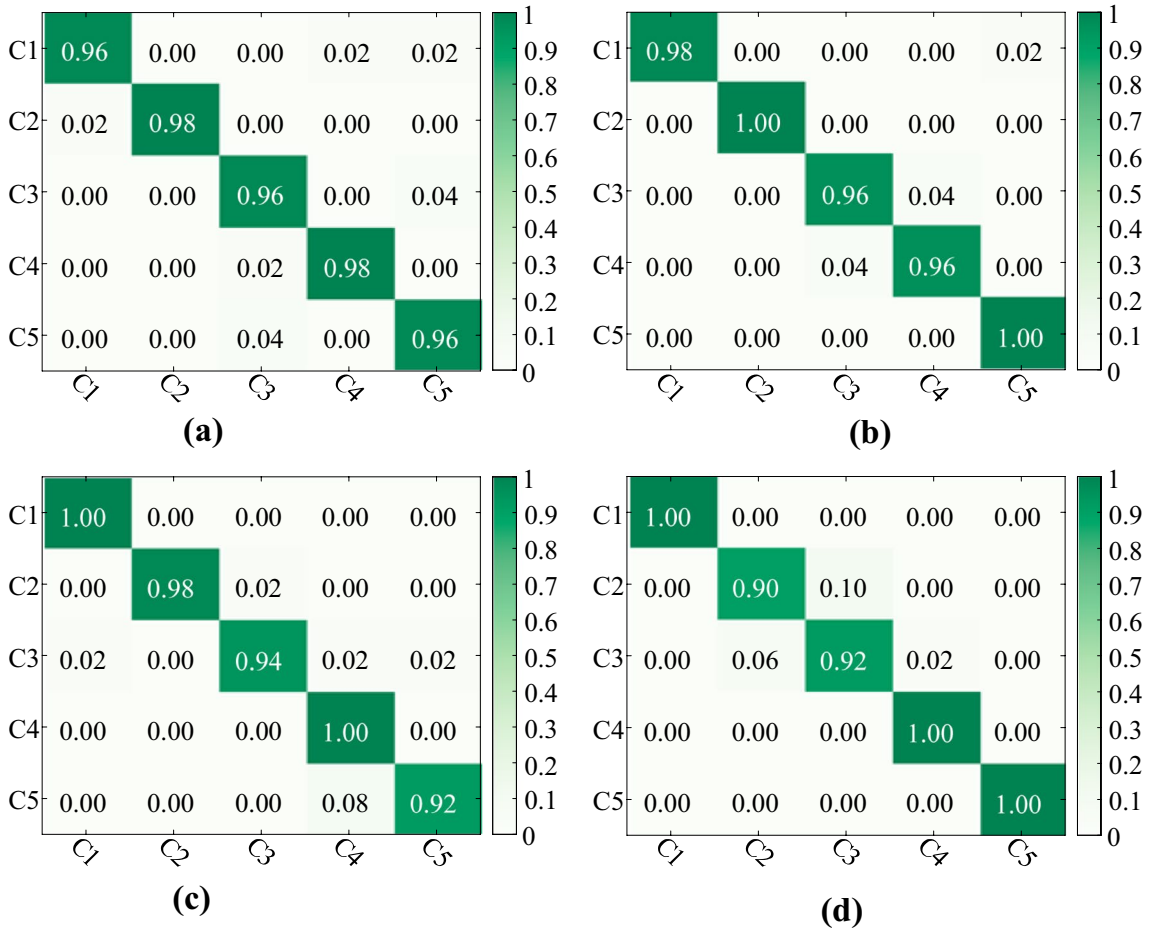

Figure 11 Confusion matrix of NMPT on transferring datasets with different loads

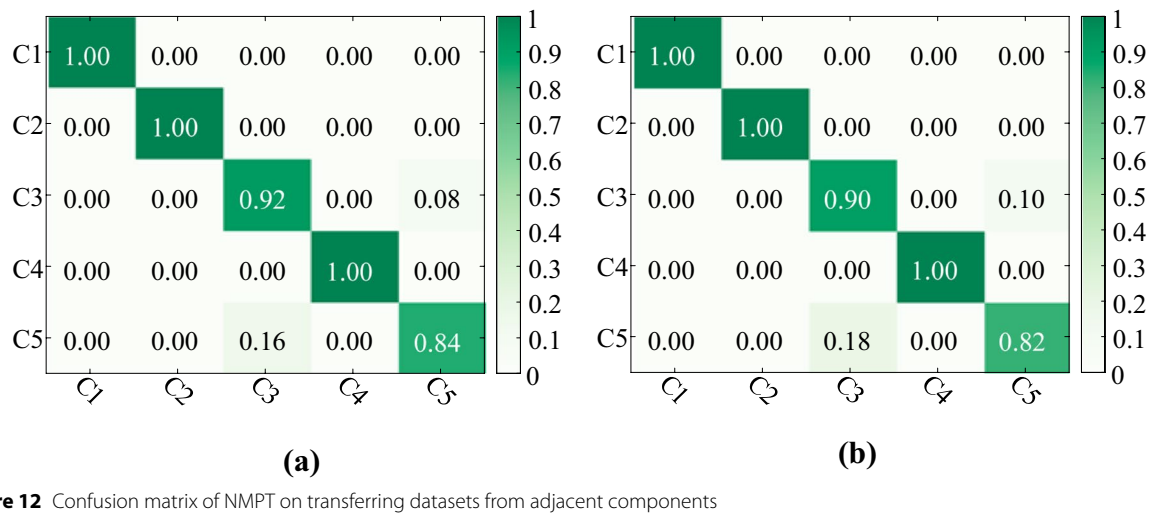

Figure 12 Confusion matrix of NMPT on transferring datasets from adjacent components

Table 3 Total GFD accuracies from test DA1 to DA10

Model	DA1	DA2	DA3	DA4	DA5	DA6	DA7	DA8	DA9	DA10
LSSVM	0.868	0.860	0.876	0.876	0.868	0.860	0.876	0.876	0.876	0.876
MTL	0.932	0.884	0.932	0.924	0.896	0.904	0.892	0.900	0.920	0.884
TCA	0.900	0.880	0.952	0.960	0.900	0.924	0.916	0.928	0.944	0.924
DSM	0.936	0.948	0.960	0.918	0.896	0.928	0.956	0.956	0.952	0.900
ELSSVM	0.936	0.884	0.956	0.936	0.928	0.896	0.948	0.940	0.928	0.900
NMPT	**0.988**	**0.980**	**0.988**	**0.984**	**0.968**	**0.980**	**0.968**	**0.964**	**0.952**	**0.944**

NMPT can make the best use of source domain samples to provide a performance improvement of diagnostic model for target task. Compared with other models, NMPT possesses the highest accuracy in the whole datasets (with the highest diagnostic accuracy: 98.8%), which proves the superiority of NMPT in utilizing source domain signals to assist GFD in target domain and provides a practical method for improving GFD performance.

5 Conclusions

(1) For the GFD problems under variable working conditions, the structure of a NMPT-theoretic strategy is presented, which utilizes ITD technology to structure fault characteristics for model parameter transferring. Experimental results indicate that the proposed method can achieve 97.16% diagnostic precision when the energies of first six level PRCs are set as feature vectors.

(2) The visualization results verify that NMPT can generalize the distinguishing ability from source domain to target domain, which is beneficial for GFD under various working conditions.

(3) With regard to the diagnostic performance, the NMPT model shows a strong robustness under different working conditions. Meanwhile, it can be found that the influence of working conditions on the GFD results is ordered by: rotating speed < load < location.

(4) The proposed model parameter transfer strategy show better performance than other popular methods, because NMPT can further minimize the discrepancy of two decision boundaries over tasks. Thus, the proposed strategy is expected to be an effective and feasible tool to solve GFD problem with less labeled target training data.

(5) In the future, we could explore the relationships between KL indicator, working condition factors and GFD results to improve the universality of the NMPT model.

Abbreviations

GFD: Gear fault diagnosis; MPT: Model parameter transfer; ITD: Intrinsic time-scale decomposition; LSSVM: Least squares support vector machine; MTLSSVM: Multi-task LSSVM; DDS: Drivetrain dynamics simulator.

Acknowledgements

Not applicable.

Authors' contributions

RY and JX designed the experiment, CC and FS analyzed the data, all the authors wrote and improved the paper. All authors read and approved the final manuscript.

Authors' information

Chao Chen received his B.Sc. and M.Sc. degree from Jiangsu University in 2011 and 2014 respectively. Now he is pursuing his PhD degree in School of Instrument Science and Engineering, Southeast University. His main research interest is machine fault diagnosis..

Fei Shen received his B.Sc. and M.Sc. degree from Southeast University in 2014 and 2016 respectively. Now he is pursuing his PhD degree in School of Instrument Science and Engineering, Southeast University. His main research interest is machine fault diagnosis..

Jiawen Xu is currently an associate researcher in School of Instrument Science and Engineering, Southeast University.

Ruqiang Yan received his B.Sc. and M.E. degree from University of Science and Technology of China in 1997 and 2002 respectively, and received his Ph.D. degree in 2007 from University of Massachusetts, Amherst. Now he is a professor and Ph.D. supervisor in Xi'an Jiaotong University. His main research interests include machine condition monitoring and fault diagnosis, signal processing, and wireless sensor networks.

Funding

Supported by National Natural Science Foundation of China (Grant No. 51835009).

Competing interests

The authors declare that they have no competing interests.

Author Details

[1] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China. [2] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China.

Received: 28 April 2020 Revised: 13 November 2020 Accepted: 19 November 2020
Published online: 18 January 2021

References

[1] F Shen, C Chen, R Q Yan, et al. A fast multi-tasking solution: NMF-theoretic co-clustering for gear fault diagnosis under variable working conditions. *Chinese Journal of Mechanical Engineering*, 2020, 33: 16.

[2] X H Jin, Y Sun, J H Shan, et al. Fault diagnosis and prognosis for wind turbines: An overview. *Chinese Journal of Scientific Instrument*, 2017, 38(5): 1041-1053. (in Chinese)
[3] L M Wang, Y M Shao. Crack fault classification for planetary gearbox based on feature selection technique and K-means clustering method. *Chinese Journal of Mechanical Engineering*, 2018, 31: 4.
[4] R N Liu, B Y Yang, E Zio, et al. Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, 2018, 108: 33-47.
[5] J Yu, Y He. Planetary gearbox fault diagnosis based on data-driven valued characteristic multigranulation model with incomplete diagnostic information. *Journal of Sound and Vibration*, 2018, 429: 63-77.
[6] Z Gao, C Cecati, S X Ding. A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, 2015, 62(6): 3757-3767.
[7] R Q Yan, R X Gao, X F Chen. Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal Processing*, 2014, 96(PART A): 1-15.
[8] S J Deng, L W Tang, X T Zhang. Gear fault diagnosis based on an adaptive neighborhood incremental PCA-LPP manifold learning algorithm. *Journal of Vibration and Shock*, 2017, 36(14): 111-132. (in Chinese)
[9] M Zeng, Y Yang, J S Cheng, et al. μ-SVD based denoising method and its application to gear fault diagnosis. *Journal of Mechanical Engineering*, 2015, 51(3): 95-103. (in Chinese)
[10] S Park, S Kim, J Choi. Gear fault diagnosis using transmission error and ensemble empirical mode decomposition. *Mechanical Systems and Signal Processing*, 2018, 108: 262-275.
[11] T Song, Y L Wang, M F Zhao, et al. Fault diagnosis for rotating machineries under variable operation conditions based on SVDI. *Journal of Vibration and Shock*, 2018, 37(19): 211-216. (in Chinese)
[12] D Y Han, N Zhao, P M Shi. Gear fault feature extraction and diagnosis method under different load excitation based on EMD, PSO-SVM and fractal box dimension. *Journal of Mechanical Science and Technology*, 2019, 33(2): 487-494.
[13] D Z Zhao, T Y Wang, F L Chu. Deep convolutional neural network based planet bearing fault classification. *Computers in Industry*, 2019, 107: 59-66.
[14] S J Pan, Q Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1345-1359.
[15] N D Lawrence, J C Platt. Learning to learn with the informative vector machine. *Proceedings of the 21th International Conference on Machine Learning*, Banff, Alberta, Canada, July 4-8, 2004: 65-72.
[16] E V Bonilla, K M A Chai, C K I Williams. Multi-task Gaussian process prediction. *Proceedings of the 22th Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 8-11, 2008: 153-160.
[17] A Schwaighofer, V Tresp, K Yu. Learning Gaussian process kernels via hierarchical Bayes. *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 13-18, 2004: 1209-1216.
[18] T Evgenious, M Pontil. Regularized multi-task learning. *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, USA, August 22-25, 2004: 109-117.
[19] L Chen, S Zhou. Sparse algorithm for robust LSSVM in primal space. *Neurocomputing*, 2018, 275: 2880-2891.
[20] R Q Yan, F Shen, C Sun, et al. Knowledge transfer for rotary machine fault diagnosis. *IEEE Sensors Journal*, 2020, 20(15): 8374-8393.
[21] S Xu, X An, X Qiao, et al. Multi-task least-squares support vector machines. *Multimedia Tools and Applications*, 2014, 71(2): 699-715.
[22] C A Micchelli, M Pontil. Kernels for multi-task learning. *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 13-18, 2004: 921-928.
[23] M G Frei, I Osorio. Intrinsic time-scale decomposition: time–frequency–energy analysis and real-time filtering of non-stationary signals. *Proceedings of the Royal Society A Mathematical Physical and Engineering Sciences*, 2007, 463(2078): 321-342.
[24] S J Pan, I W Tsang, J T Kwok, et al. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011, 22(2): 199-210.
[25] L X Duan, D Xu, S F Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 16-21, 2012: 1338–1345.
[26] C Chen, F Shen, R Q Yan. Enhanced least squares support vector machine-based transfer learning strategy for bearing fault diagnosis. *Chinese Journal of Scientific Instrument*, 2017, 38(1): 33-40. (in Chinese)