面向变工况机械设备智能故障诊断的可解释三特征提取器迁移网络

陈 凯[1,2], 丁传仓[1,2], 王报祥[1,2], 黄伟国[1,2], 朱忠奎[1,2]

（1. 苏州大学轨道交通学院, 江苏 苏州 215131; 2. 江苏省智慧城轨工程研究中心, 江苏 苏州 215131）

摘要: 针对深度神经网络可解释性低及目前可解释网络无法实现跨域诊断的问题, 提出了一种可解释三特征提取器迁移网络（interpretable triple feature extractor transfer network, ITFETN）。针对可解释性问题, 建立了多层稀疏编码模型, 推导了多层稀疏编码模型的迭代求解算法, 通过展开快速迭代软阈值算法, 得到稀疏编码模型求解算法的等效网络形式, 并将其作为特征提取器, 形成具有可解释性的算法结构等效网络; 针对跨域迁移诊断问题, 构建了三特征提取器策略用于提取源域、目标域的共享特征以及各自的私有特征, 并基于特征对抗思想设计了迁移诊断任务的损失函数用于 ITFETN 的有效训练, 有效提取出源域和目标域中距离最小化的共享特征进行跨域诊断, 实现可解释迁移诊断任务。试验结果表明, ITFETN 在两个实例分析中的平均准确率和鲁棒性相较于对比方法均有所提升, 能够有效实现具有可解释性的跨域诊断。

关键词: 智能故障诊断; 可解释网络; 迁移学习; 稀疏编码; 三特征提取器

中图分类号: TP18; TH133 **文献标志码:** A **文章编号:** 1004-4523（2025）06-1232-10

DOI: 10.16385/j.cnki.issn.1004-4523.2025.06.011

Interpretable triple feature extractor transfer network for intelligent fault diagnosis of mechanical equipment under variable working conditions

CHEN Kai[1,2], DING Chuancang[1,2], WANG Baoxiang[1,2], HUANG Weiguo[1,2], ZHU Zhongkui[1,2]

（1.School of Rail Transportation，Soochow University，Suzhou 215131，China；
2.Intelligent Urban Rail Engineering Research Center of Jiangsu Province，Suzhou 215131，China）

Abstract: To address the limitations of deep neural networks in terms of interpretability and the inability of current interpretable networks to perform cross-domain diagnosis tasks, this paper proposes an interpretable triple feature extractor transfer network(ITFETN). For the interpretability challenge, a multi-layer sparse coding model is established, and its iterative solving algorithm is derived. By unrolling the fast iterative soft thresholding algorithm, an equivalent network form of the sparse coding model soving algorithm is obtained. This equivalent network then serves as a feature extractor, forming an interpretable algorithm-structure-equivalent network. To tackle the problem of cross-domain tranfer diagnosis, a triple feature extractor strategy is constructed. This strategy is designed to extract the shared features from the source and target domains, as well as their respective private features. Based on the concept of feature adversarial learning, a loss function for the transfer diagnosis task is designed for the effective training of ITFETN. This effectively extracts shared features with minimized distance between the source and target domains for cross-domain diagnosis, thereby achieving interpretable transfer diagnosis tasks. Experimental results demonstrate that ITFETN exhibits improved average accuracy and robustness in two case studies compared to benchmark methods. This confirms its effectiveness in achieving interpretable cross-domain diagnosis.

Keywords: intelligent fault diagnosis; interpretable network; transfer learning; sparse coding; triple feature extractor

故障诊断是对系统及设备的数据进行采集、分析和处理, 从而实现潜在故障准确诊断的技术。开展机械设备故障诊断研究具有重大意义[1]。

在故障诊断领域中, 传统信号处理方法表现出优异的状态特征提取及故障诊断能力, 例如傅里叶变换、经验模态分解和谱峭度分析等。信号处理方法具有丰富的理论支撑, 利用信号处理方法获得的故障诊断结果具有明确的可解释性。然而, 当数据

收稿日期: 2024-12-11; **修订日期:** 2025-04-15

基金项目: 国家自然科学基金资助项目（52205119, 52275157, 52405124）; 江苏省自然科学基金资助项目（BK20220497）; 江苏省智慧城轨工程研究中心开放课题（SDGC2414）

量较大且分布情况较为混乱时，信号处理方法由于缺乏优异的大批量数据处理及非线性学习能力，导致故障诊断的效率和准确率大幅度下降。随着人工智能的快速发展，基于深度学习的故障诊断技术由于其优异的特征非线性表征学习和故障诊断能力得到了广泛应用。目前，在机械故障诊断中广泛采用的代表性深度神经网络，包括人工神经网络[2]、卷积神经网络[3]和残差神经网络等。同时，许多学者在这些方法的基础上，针对特定的故障诊断任务做了改进和创新，以提升故障诊断准确性。例如，SHI 等[4]提出了基于图嵌入的深度泛化学习系统，通过逐步编码和解码机制学习振动信号中的高级抽象特征，可解决旋转机械设备的不平衡数据故障诊断问题。CHEN 等[5]提出了一种基于持续学习的双分支自适应聚合残差网络，利用知识蒸馏函数克服了灾难性遗忘问题，实现故障增量诊断任务。尽管深度神经网络模型能够通过强大的学习能力实现智能诊断，然而其自身的算法结构缺乏可靠的理论支撑，内部层与层之间缺乏明确的理论映射关系。这些传统的深度神经网络模型往往缺乏可解释性，被称为黑箱模型[6]。因此，如何提高深度神经网络模型的可解释性是一个重要的课题。

深度神经网络的可解释性研究主要分为事前可解释网络模型构建和事后可解释性分析两种。事前可解释网络模型构建是通过特定的结构设计使模型具有可解释性，而事后可解释性分析是指在网络模型训练的过程中或训练结束之后，根据网络架构及学习步骤事后分析可解释性。本文将重点讨论事前可解释深度神经网络模型。

事前可解释网络模型主要分为物理知识嵌入模型、功能框架嵌入模型和算法结构等效模型。物理知识嵌入模型是最直接的事前可解释模型，它在损失函数中直接引入与问题求解相关的约束，通过遵循物理定律或者经验公式，使得深度神经网络学习到额外的物理先验知识。WANG 等[7]提出了一种物理信息神经网络，通过引入与经验退化公式、空间状态方程等相关的变量，可准确估计电池的健康状态。功能框架嵌入模型通过在网络中加入可解释的功能性框架，使得诊断模型具有可解释性。LI 等[8]提出了一种新的小波驱动网络，将卷积神经网络的第一个卷积层替换为连续小波卷积层，可实现具有局部可解释性的机械故障诊断。然而，物理知识嵌入模型仅适用于具有物理背景相关的领域，功能框架嵌入模型也仅仅具有局部的可解释性。

相较于以上两类模型，算法结构等效模型提供了一种更为优异的可解释网络模型构建方案，利用神经网络来等效传统信号处理方法中的解析结构，是一种完全可解释的网络模型。算法结构等效模型以传统的信号处理算法来驱动网络训练，在网络结构的设计和特征提取器算法的推导中体现出网络的可解释性，本文研究重点也在于此。其中，算法展开是构建算法结构等效模型最常采用的一种方法[9]。算法展开理论将具有先验知识的传统迭代算法近似等效到深度神经网络中，可构建高效而可解释的深度网络架构。作为一种完全可解释的深度模型，算法结构等效模型已经被应用于多个研究领域。例如，SHAO 等[10]基于稀疏解耦模型，提出隐式变量展开网络，实现相较于与其他解耦方法更好的可解释性和稳定性。AN 等[11]利用算法展开框架提出了事前可解释的对抗算法展开网络，同时开展了网络的事后可解释性分析，并将其成功应用于机械异常检测。LI 等[12]通过推广传统全变分梯度域正则化算法并展开，得到深度展开去模糊网络，在满足效率和可解释性的同时，显著提升了图像去模糊的实际性能。PENG 等[13]提出了一种可解释的个性化联邦学习框架，并为该框架提出了一种可解释的人工智能机制，该机制应用决策树匹配本地联邦学习模型的输入和输出，利用 t-SNE 可视化全局聚合前后的本地模型。ZHANG 等[14]提出了一种基于可解释卷积编码的损伤定位方法，该方法采用多层迭代软阈值算法求解多层卷积稀疏编码模型并展开，能够充分提取损伤特征。ZHOU 等[15]通过算法展开的方法求解非负稀疏编码模型，将先验知识嵌入深度神经网络中，能够准确有效地定位和重构冲击力。

尽管基于算法展开的结构等效模型在机械故障诊断中取得了良好的效果，然而其缺乏跨域迁移学习能力，无法适用于变工况下的机械故障诊断。具体来说，现有的基于算法展开的可解释网络结构大多只能在单一工况下完成故障诊断任务，当机械系统的工况发生变化时，模型的故障诊断能力会下降，无法完成迁移学习的任务。

针对当前深度神经网络无法实现可解释跨域智能诊断的问题，本文提出了可解释三特征提取器迁移网络（interpretable triple feature extractor transfer network，ITFETN）。具体地，针对可解释性问题，建立了多层稀疏编码模型，推导了多层稀疏编码模型的迭代求解算法，通过算法展开得到求解算法的等效网络形式，将其作为 ITFETN 的特征提取器；此外，针对跨域迁移诊断问题，构建了三特征提取器策略用于提取源域、目标域的共享特征以及各自的私有特征，并设计了迁移诊断任务的损失函数用于 ITFETN 的有效训练，实现可解释迁移诊断任务。

1 算法展开理论

稀疏表示[16]是一种简洁而高效的数据表征与特征提取方法，广泛应用于图像处理、目标识别和机器视觉等领域。近年来，稀疏表示方法也被应用于机械旋转部件的故障诊断中，能够有效提取机械设备振动信号中的状态特征。

通常来说，从机械旋转部件上采集到的信号可以看作由状态特征和其他干扰两部分组成[17]：

$$\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{n}=\boldsymbol{D\gamma}+\boldsymbol{n} \tag{1}$$

式中，$\boldsymbol{y}$为采集到的信号；$\boldsymbol{x}$为状态特征；$\boldsymbol{n}$为其他干扰；$\boldsymbol{D}$为学习字典；$\boldsymbol{\gamma}$为字典编码。

通过寻找最优的学习字典和稀疏编码，能够达到信号最优的保真性及编码稀疏性，进而准确提取信号中的状态特征。为了实现这个目标，构建如下优化问题：

$$\begin{aligned}&\min_{\boldsymbol{\gamma}}\ \|\boldsymbol{\gamma}\|_0\\&\text{s.t.}\ \|\boldsymbol{y}-\boldsymbol{D\gamma}\|_2^2\leqslant\varepsilon\end{aligned} \tag{2}$$

式中，ε为重构误差阈值。

式(2)中的优化问题是NP-hard问题，难以利用普通方法求解。因此，通常采用L1范数取代L0范数，则优化问题相应地转变为：

$$\min_{\boldsymbol{\gamma}}\ \lambda\|\boldsymbol{\gamma}\|_1+\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D\gamma}\|_2^2 \tag{3}$$

式中，λ为正则化参数。

在交替方向乘子法(alternating direction method of multipliers, ADMM)框架下，通过引入辅助变量并设置增广拉格朗日项约束变量，能够求解式(3)中的优化问题。算法1列出了式(3)中优化问题的迭代求解算法。其中，K为迭代次数，$soft(\cdot)$为软阈值算子，上标“T”表示矩阵转置，上标“−1”表示矩阵求逆，u与γ为具有约束的两个对偶变量，μ为对偶变量约束正则化参数，d为对偶变量约束误差，k为迭代展开次数，$\mathbf{I}$为单位矩阵。

算法1：迭代求解算法

初始化： $\mu>0$, d

for k=0: K−1

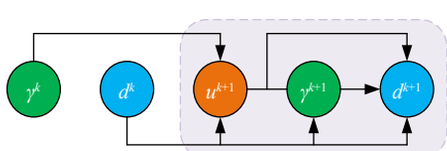
$u^{k+1}\leftarrow soft(\gamma^k+d^k,\lambda/\mu)$

$\gamma^{k+1}\leftarrow(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}+\mu\mathbf{I})^{-1}[\boldsymbol{D}^{\mathrm{T}}\boldsymbol{y}+\mu(u^{k+1}-d^k)]$

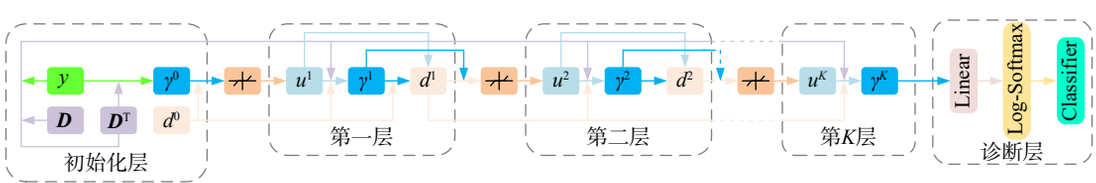
$d^{k+1}\leftarrow d^k-u^{k+1}+\gamma^{k+1}$

end for

在算法1中，将依次求解$\boldsymbol{u}$、$\boldsymbol{\gamma}$、$\boldsymbol{d}$的过程视为一个迭代单元，将该迭代单元展开成K层网络结构，即算法展开，如图1所示。每个迭代单元的计算过程依赖上个迭代单元的计算结果，直至计算出最终层结果。



(a) 迭代求解单元
(a) The iterative solving unit



(b) 迭代求解算法网络展开形式
(b) The network unrolling form of iterative solving algorithm

图1 迭代求解算法展开网络

Fig. 1 The unrolling network of iterative solving algorithm

2 基于ITFETN的故障诊断

本节主要介绍提出ITFETN的构建过程及基于ITFETN的可解释智能故障诊断方法。首先，建立了用于挖掘信号中状态特征的多层稀疏编码模型，利用快速迭代软阈值算法推导了多层稀疏编码模型的迭代求解算法，并通过展开求解算法得到其网络等效形式，作为ITFETN的特征提取器。同时，为了实现跨域故障诊断中的特征提取，利用三特征提取器分别提取源域、目标域的共享特征以及各自的私有特征，并将共享特征作为故障分类器的输入。最后，设计了包含预测损失和特征对抗损失两项的损失函数，用于实现ITFETN的网络训练，从而实现可解释跨域迁移诊断。ITFETN的框架图如图2所示，ITFETN利用可解释特征提取器提取源域、目标域的共享特征以及各自的私有特征；设计特征对抗模块计算特

征对抗损失，缩小双域共享特征距离，扩大域内私有及共享特征距离；设计分类器计算预测损失，提高诊断准确率；最后为两类损失函数分配合理的权重参数计算总损失，以确保网络的有效训练。

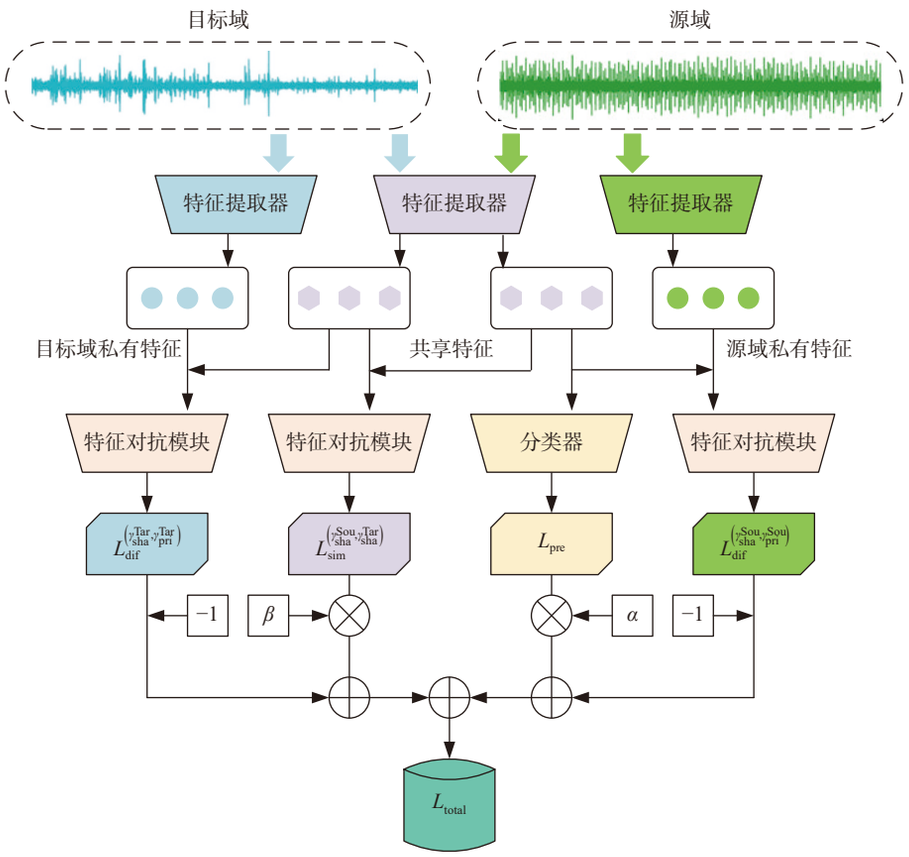


图 2　ITFETN 的框架图

Fig. 2　The framework diagram of ITFETN

首先，介绍多层稀疏编码模型及其迭代求解算法，用于状态特征可解释提取。

一般地，式 (1) 所示的单层字典的稀疏编码模型及迭代求解算法只能提取信号中的浅层特征。当信号量大且复杂时，信号中的深层特征无法被单层稀疏编码模型准确提取。所以，需要扩展稀疏编码模型中的字典层数，从而提取出隐含在信号中的深层特征。

具体地，对于式 (1) 来说，状态特征$\boldsymbol{x}$可以转化为多层字典的级联结构：

$$\boldsymbol{x}=\boldsymbol{D}_1\boldsymbol{\gamma}_1=\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{\gamma}_2=\boldsymbol{D}_1\boldsymbol{D}_2\cdots\boldsymbol{D}_L\boldsymbol{\gamma}_L \tag{4}$$

式中，状态特征$\boldsymbol{x}$被展开成L层结构，$\boldsymbol{D}_i$和$\boldsymbol{\gamma}_i$分别为第i层中的学习字典和与之匹配的稀疏编码。

相应地，式 (3) 中的优化问题转化为如下多层稀疏编码模型：

$$\begin{aligned}\min_{\boldsymbol{\gamma}_L}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D}_1\boldsymbol{D}_2\cdots\boldsymbol{D}_L\boldsymbol{\gamma}_L\|_2^2+\lambda_1\|\boldsymbol{D}_2\cdots\boldsymbol{D}_L\boldsymbol{\gamma}_L\|_1+\\ \lambda_2\|\boldsymbol{D}_3\cdots\boldsymbol{D}_L\boldsymbol{\gamma}_L\|_1+\cdots+\lambda_L\|\boldsymbol{\gamma}_L\|_1\end{aligned} \tag{5}$$

式中，λ_i为第i层中的稀疏正则化参数。式 (5) 旨在保证信号保真度的同时，通过在每一层加入稀疏性约束，以确保多层稀疏编码模型的全局稀疏性。

接下来，将详细介绍多层稀疏编码模型的求解算法推导过程。为了更方便地描述求解算法推导过程，本文以层数$L=2$为例。当层数$L=2$时，优化问题转变为：

$$\min_{\boldsymbol{\gamma}}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{\gamma}\|_2^2+\lambda_1\|\boldsymbol{D}_2\boldsymbol{\gamma}\|_1+\lambda_2\|\boldsymbol{\gamma}\|_1 \tag{6}$$

利用交替方向乘子法能够有效求解式 (6) 中的优化问题。ADMM 的核心思想是将复杂优化问题转换为多个简单优化问题，并迭代交替求解简单优化问题。

具体地，对于两层稀疏展开模型，首先引入辅助变量$\boldsymbol{\gamma}_1$，则式 (6) 中的优化问题转换为：

$$\begin{aligned}&\min_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{\gamma}_2\|_2^2+\lambda_1\|\boldsymbol{\gamma}_1\|_1+\lambda_2\|\boldsymbol{\gamma}_2\|_1\\ &\text{s.t. }\boldsymbol{\gamma}_1-\boldsymbol{D}_2\boldsymbol{\gamma}_2=0\end{aligned} \tag{7}$$

随后，通过增加增广拉格朗日项来保证约束，式 (7) 进一步转化为如下无约束优化问题：

$$\begin{aligned}\min_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\boldsymbol{v}}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{\gamma}_2\|_2^2+\lambda_1\|\boldsymbol{\gamma}_1\|_1+\\ \lambda_2\|\boldsymbol{\gamma}_2\|_1+\frac{\theta}{2}\|\boldsymbol{\gamma}_1-\boldsymbol{D}_2\boldsymbol{\gamma}_2+\boldsymbol{v}\|_2^2\end{aligned} \tag{8}$$

式中，θ 为正则化平衡参数；$\boldsymbol{v}$ 为与 $\boldsymbol{\gamma}$ 具有约束的对偶变量。

根据 ADMM 算法框架可知，式 (8) 中优化问题的解可以通过迭代求解以下问题得到：

$$\boldsymbol{\gamma}_2 \leftarrow \arg\min_{\boldsymbol{\gamma}_2} \|\boldsymbol{y}-\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{\gamma}_2\|_2^2 + \frac{\theta}{2}\|\boldsymbol{\gamma}_1-\boldsymbol{D}_2\boldsymbol{\gamma}_2+\boldsymbol{v}\|_2^2+\lambda_2\|\boldsymbol{\gamma}_2\|_1 \tag{9}$$

$$\boldsymbol{\gamma}_1 \leftarrow \arg\min_{\boldsymbol{\gamma}_1} \frac{\theta}{2}\|\boldsymbol{\gamma}_1-\boldsymbol{D}_2\boldsymbol{\gamma}_2+\boldsymbol{v}\|_2^2+\lambda_1\|\boldsymbol{\gamma}_1\|_1 \tag{10}$$

$$\boldsymbol{v} \leftarrow \boldsymbol{v}+\theta(\boldsymbol{\gamma}_1-\boldsymbol{D}_2\boldsymbol{\gamma}_2) \tag{11}$$

对于式 (9) 与 (10)，可以利用快速迭代软阈值算法（fast iterative shrinkage-thresholding algorithm, FISTA）优化内层稀疏编码 $\boldsymbol{\gamma}_i$。算法 2 列出了展开次数为 K 时，利用两层 FISTA 求解方法优化内层稀疏编码 $\boldsymbol{\gamma}_i$ 的过程。其中，$\boldsymbol{z}$ 为中间变量，$\hat{\boldsymbol{\gamma}}_i$ 为第 i 层中的理想编码，θ_i 为第 i 层中的重构误差正则化参数，t_k 为动量因子。

算法2: 两层FISTA求解算法

输入: 信号$\boldsymbol{y}$, 字典$\boldsymbol{D}_i$, 稀疏正则化参数λ_i

$\forall k$: **初始化** $\boldsymbol{\gamma}_0^k=\boldsymbol{y}$, $\boldsymbol{z}=0$

for $k=1:K$

$\hat{\boldsymbol{\gamma}}_i \leftarrow \boldsymbol{D}_{(i,L)}\boldsymbol{z},\ \forall i\in[0,L-1]$

for $i=1:2$

$\boldsymbol{\gamma}_i^{k+1} \leftarrow Soft_{\lambda_i/\theta_i}\left[\hat{\boldsymbol{\gamma}}_i-\theta_i\boldsymbol{D}_i^{\mathrm{T}}(\boldsymbol{D}_i\hat{\boldsymbol{\gamma}}_i-\boldsymbol{\gamma}_{i-1}^{k+1})\right]$

end for

$t_{k+1} \leftarrow \dfrac{1+\sqrt{1+4t_k^2}}{2}$

$z \leftarrow \boldsymbol{\gamma}_L^{k+1}+\dfrac{t_k-1}{t_{k+1}}(\boldsymbol{\gamma}_L^{k+1}-\boldsymbol{\gamma}_L^k)$

end for

为了方便网络训练，将算法 2 中参数 $\boldsymbol{\gamma}_i^{k+1}$ 的求解转换为如下形式：

$$\boldsymbol{\gamma}_i^{k+1}=\mathrm{ReLU}\left[\boldsymbol{\gamma}_i^k-\theta_i\boldsymbol{D}_i^{\mathrm{T}}\left(\boldsymbol{D}_i\boldsymbol{\gamma}_i^k-\boldsymbol{\gamma}_{i-1}^{k+1}\right)+\boldsymbol{b}_i\right] \tag{12}$$

其中，软阈值函数 *Soft* 转换为网络中常用的 ReLU 函数，偏置参数 $\boldsymbol{b}_i$ 是阈值参数 λ_i 的等价表达形式，学习字典 $\boldsymbol{D}_i$ 可以看作深度神经网络中的卷积核。因此，式 (12) 的表达形式可以被直接应用到神经网络中，作为特征提取器中前向计算的重要算法公式。

式 (12) 是算法展开网络中最核心的方程，可以作为网络的一个基本迭代单元，其中位于第 i 层编码的第 $k+1$ 迭代结果 $\boldsymbol{\gamma}_i^{k+1}$ 由上一层编码的第 $k+1$ 次迭代结果 $\boldsymbol{\gamma}_{i-1}^{k+1}$ 以及本层编码的第 k 次迭代结果 $\boldsymbol{\gamma}_i^k$ 共同决定，具体如图 3 所示。

根据前面的关于层数 L=2、展开次数为 K 时优化问题 (6) 的求解过程，将其进行推广泛化，从而能够得到多层稀疏编码模型 (5) 的迭代求解算法，即 ITFETN 的特征提取算法，如算法 3 所示。

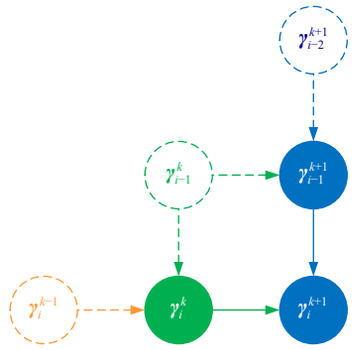


图 3 展开网络基本迭代单元

Fig. 3 The basic iteration unit of unrolling network

算法3: ITFETN特征提取算法

输入: $\boldsymbol{y}\in\mathbf{R}^{\mathrm{N}}, K\in\mathbb{N}^+, L\in\mathbb{N}^+, \boldsymbol{D}_i, \boldsymbol{\theta}_i, \boldsymbol{b}_i$

初始化 $\boldsymbol{\gamma}_0^k=\boldsymbol{y}$, $\forall k\in[0,K]$

$\boldsymbol{\gamma}_i^0=0$, $\forall i\in[0,L]$

for $k=0,1,\cdots,K$ **do**

for $i=1,2,\cdots,L$ **do**

$\boldsymbol{\gamma}_i^{k+1}=\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\gamma}_{i-1}^k$

end for

for $i=1,2,\cdots,L$ **do**

$\boldsymbol{\gamma}_i^{k+1}=\mathrm{ReLU}\left[\boldsymbol{\gamma}_i^k-\boldsymbol{\theta}_i\boldsymbol{D}_i^{\mathrm{T}}(\boldsymbol{D}_i\boldsymbol{\gamma}_i^k-\boldsymbol{\gamma}_{i-1}^{k+1})+\boldsymbol{b}_i\right]$

end for

end for

最终赋值 $\hat{\boldsymbol{\gamma}}_L=\boldsymbol{\gamma}_L^{K+1}$

ITFETN 特征提取器算法的结构如图 4 所示。

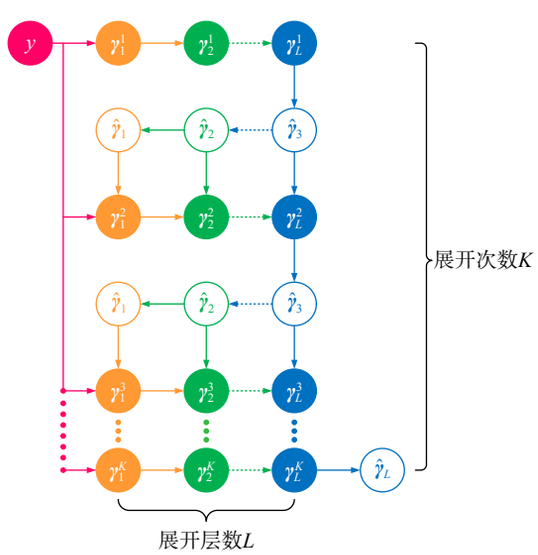


图 4 ITFETN 特征提取器算法结构

Fig. 4 Structure of the feature extractor algorithm in ITFETN

针对跨域故障诊断中的特征提取，ITFETN 引入了三特征提取器策略，即利用三个特征提取器分别提取源域和目标域的共享特征及各自的私有特征，并将共享特征作为故障分类器的输入，如图 2 所示。

接下来，详细介绍三特征提取器策略下 ITFETN 的损失函数设计。ITFETN 的损失函数 L_{total} 包含预测损失函数 L_{pre} 和特征对抗损失函数 L_{adv} 两部分：

$$L_{\text{total}} = \alpha L_{\text{pre}} + L_{\text{adv}} \tag{13}$$

式中，α为预测损失函数与特征对抗损失函数的权重比例。

预测损失函数L_{pre}旨在根据交叉熵损失，缩小预测标签与真实标签的距离，从而提高网络提取特定故障特征的能力；特征对抗损失函数L_{adv}旨在通过二分类交叉熵损失及相应的鉴别器，缩小共享特征之间的分布差异，加大提取的共享及私有特征之间的分布差异。具体地，特征对抗损失函数L_{adv}包含三部分，分别为：从源域和目标域提取的共享特征之间对抗产生的损失函数、从源域提取的共享特征和私有特征之间对抗产生的损失函数、从目标域提取的共享特征和私有特征之间对抗产生的损失函数，表示为：

$$L_{\text{adv}} = \beta L_{\text{sim}}^{\left(\gamma_{\text{sha}}^{\text{Sou}},\gamma_{\text{sha}}^{\text{Tar}}\right)} - L_{\text{dif}}^{\left(\gamma_{\text{sha}}^{\text{Sou}},\gamma_{\text{pri}}^{\text{Sou}}\right)} - L_{\text{dif}}^{\left(\gamma_{\text{sha}}^{\text{Tar}},\gamma_{\text{pri}}^{\text{Tar}}\right)} \tag{14}$$

式中，β为双域共享特征对抗损失函数与域内共享及私有特征之间对抗损失函数的权重比例。

权重参数对(α, β)的确定过程如下：对于权重参数α，ITFETN 的总损失函数包含预测损失和特征对抗损失两部分，而特征对抗损失对于共享特征的精确提取及网络的有效训练起到了更为重要的作用，所以α应小于 1，在预试验中，将α的权值确定在[0.2,0.8]的区间内，并以步长为 0.1 进行遍历；对于权重参数β，由于输入分类器的是双域的共享特征，所以共享特征之间的对抗损失权重应更大，在预试验中，将β的权值确定在[2,15]的区间内，并以步长为 1 进行遍历。经过一系列的预试验，发现当$\alpha=0.5$、$\beta=10$时，所提出网络的诊断准确率处于极值点，故将权重参数设置为$(\alpha, \beta)=(0.5,10)$。

3　试验验证及结果分析

本节利用 2 个数据集验证所提方法 ITFETN 在变工况可解释智能诊断中的有效性。2 个数据集分别为凯斯西储大学轴承数据集和苏州大学轮对轴承数据集，基本信息如下：

（1）凯斯西储大学轴承数据集：该公开数据集从轴承试验台上采集得到，采样频率为 12 kHz。如图 5 所示，该试验台由电机、测功机和负载传感器组成。该数据集包含内圈故障（IRF）、滚动体故障（BF）和外圈故障（ORF）三种故障类型，每种故障类型的直径分别设置为 0.007、0.014 和 0.021 in。因此，加上健康无故障状态，该数据集一共设有 10 种健康状态。试验台轴承分别被施加 0、1、2 和 3 HP 四种恒定负载，因此共有 10×4=40 个数据文件用于变工况故障诊断。对每个文件中的数据进行标准化预处理，并利用长度为 1024 的滑动窗口对文件进行分割。

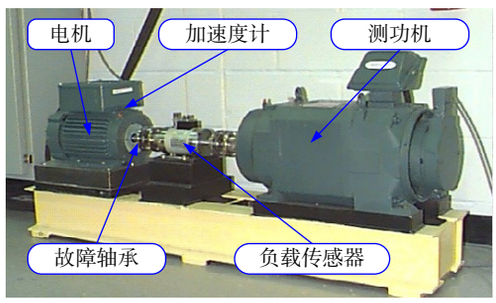


图 5　凯斯西储大学轴承试验台

Fig. 5　Bearing test rig of CWRU

（2）苏州大学轮对轴承数据集：该私有数据集从苏州大学综合轮对传动试验台采集得到。如图 6 所示，该试验台由驱动电机、轴承、减速器和加速度计组成，试验台轴承类型为 NJ204ET NSK，通过调节螺母控制施加负载大小。该数据集在恒转速为 400 r/min 下采集，包含内圈故障（IRF）和外圈故障（ORF）两种故障类型，且每种故障类型的直径分别设为 0.2、0.4 和 0.6 mm。因此，加上无故障状态，该数据集一共设有 7 种健康状态。对于每种健康状态，测试中分别施加 0、0.8、1.6 和 2.4 kN四种恒定负载，因此共有 7×4=28 个数据文件。对于该数据集，同样对每个文件中的数据进行标准化预处理，并利用长度为 1024 的滑动窗口对文件进行分割。

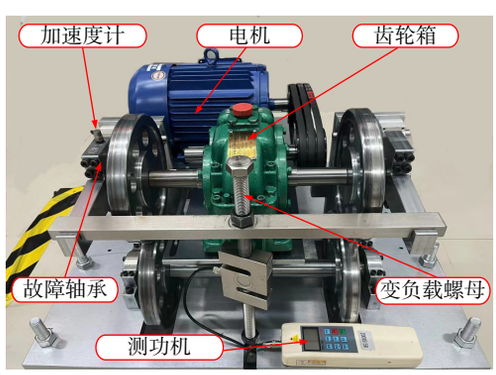


图 6　苏州大学轮对轴承试验台

Fig. 6　Wheelset bearing test rig of Soochow University

不同负载下采集的数据具有不同的分布形式，因此在本节中每一个负载下采集的数据被作为 1 个域，每个数据集包含 4 个域。对于 2 个数据集，均设置 12 个迁移诊断任务，用于验证提出方法的有效性。表 1 和 2 分别描述了 2 个数据集的迁移诊断任务。

表 1　凯斯西储大学轴承数据集迁移诊断任务

Tab. 1　Transfer diagnosis tasks for bearing dataset of CWRU

源域	迁移诊断任务		
A(0 HP)	T_1: A→B	T_2: A→C	T_3: A→D
B(1 HP)	T_4: B→A	T_5: B→C	T_6: B→D
C(2 HP)	T_7: C→A	T_8: C→B	T_9: C→D
D(3 HP)	T_{10}: D→A	T_{11}: D→B	T_{12}: D→C

表 2 苏州大学轮对轴承数据集迁移诊断任务

Tab. 2 Transfer diagnosis tasks for wheelset bearing dataset of Soochow University

源域	迁移诊断任务		
E(0 kN)	Q_1: E→F	Q_2: E→G	Q_3: E→H
F(0.8 kN)	Q_4: F→E	Q_5: F→G	Q_6: F→H
G(1.6 kN)	Q_7: G→E	Q_8: G→F	Q_9: G→H
H(2.4 kN)	Q_{10}: H→E	Q_{11}: H→F	Q_{12}: H→G

为了展现本文提出的 ITFETN 在可解释迁移诊断的优越性，LBP-Net[18]、ML-LISTA[19]、All-Free-Model[20]、DANN[21]、Coral[22] 和 MMD[23] 六种方法被作为对比方法，用于完成迁移诊断任务。在进行对比方法的试验中，设定的超参数的值和 ITFETN 几乎相同：网络训练过程中使用的是 SGD（随机梯度下降）优化器，其中学习率为 0.002，动量值为 0.9，权重衰减率为 10^{-5}；网络训练的批次大小为 32，训练轮次为 20；对于源域数据集，60% 的样本用于训练，剩余 40% 用以保持数据平衡；对于目标域数据集，60% 用于训练，20% 用于验证，剩余 20% 用于测试。LBP-Net 是普通的基追踪网络，其每一层展开单层基追踪迭代，相比之下，所提出方法展开整个多层基追踪问题的迭代；ML-LISTA 是具有双卷积字典滤波器通道的算法展开模型，比所提出方法具有更大规模的卷积权重；All-Free-Model 是未进行算法展开的、与 ML-LISTA 具有相同深度和类似循环架构的网络模型，其中不同层间的所有卷积滤波器都可以自由学习；DANN 是对抗迁移学习的代表方法，其将卷积神经网络与域适应策略结合，使用梯度翻转层，能够完成迁移分类任务；Coral 是领域自适应的经典方法，其通过对齐源域和目标域特征分布的二阶统计量协方差矩阵，从而减少域间差异；MMD 也是领域自适应的经典方法，其通过计算两个分布样本在再生希尔伯特空间中的均值差异，来衡量分布之间的差异。本文提出方法及对比方法的试验结果如表 3 和 4 所示，并以直方图的形式展现在图 7 和 8 中。

对于凯斯西储大学轴承数据集上的迁移诊断任

表 3 不同方法在凯斯西储大学轴承数据集迁移诊断任务中的诊断准确率（单位：%）

Tab. 3 Diagnosis accuracies of different methods for transfer diagnosis tasks in bearing dataset of CWRU (Unit: %)

迁移诊断任务	LBP-Net	ML-LISTA	All-Free-Model	DANN	Coral	MMD	ITFETN
T_1	94.66	91.55	94.76	92.70	96.18	91.46	97.76
T_2	95.93	94.67	92.12	90.81	94.99	96.33	97.47
T_3	97.24	92.32	92.32	94.99	94.81	91.48	98.41
T_4	93.18	91.38	92.32	89.83	92.75	92.68	97.21
T_5	96.42	96.51	93.07	96.04	90.64	95.54	99.37
T_6	96.51	95.09	93.65	94.71	93.57	95.16	98.88
T_7	95.47	92.12	91.63	93.32	95.17	90.84	97.36
T_8	96.67	93.54	96.38	95.59	95.87	92.22	98.41
T_9	98.35	95.97	95.93	89.33	93.54	92.77	99.40
T_{10}	95.65	90.92	91.87	91.65	92.40	94.66	96.91
T_{11}	94.62	94.38	95.27	91.28	95.87	96.26	97.78
T_{12}	97.76	95.15	94.69	91.72	94.85	96.51	99.50
平均值	96.04	93.63	93.67	92.66	94.22	93.83	98.21

表 4 不同方法在苏州大学轮对轴承数据集迁移诊断任务中的诊断准确率（单位：%）

Tab. 4 Diagnosis accuracies of different methods for transfer diagnosis tasks in wheelset bearing dataset of Soochow University (Unit: %)

迁移诊断任务	LBP-Net	ML-LISTA	All-Free-Model	DANN	Coral	MMD	ITFETN
Q_1	89.90	91.16	82.89	85.47	91.13	84.97	94.87
Q_2	95.07	90.81	87.52	94.35	93.95	86.89	97.69
Q_3	96.62	90.85	89.72	91.87	93.22	86.74	98.23
Q_4	90.93	85.16	84.35	91.09	94.89	94.45	95.53
Q_5	97.10	93.17	92.10	93.17	94.71	90.71	98.82
Q_6	94.31	89.19	90.67	93.59	91.32	86.44	98.00
Q_7	92.74	89.73	86.18	91.02	94.11	92.93	95.97
Q_8	93.58	84.56	88.37	91.03	87.41	94.18	97.43
Q_9	94.36	93.35	91.94	93.62	90.72	89.68	98.71
Q_{10}	93.46	92.41	82.89	88.25	88.33	93.07	96.82
Q_{11}	90.66	85.83	82.38	85.33	93.13	92.22	96.17
Q_{12}	96.58	93.63	94.30	94.59	89.05	93.22	99.07
平均值	93.75	89.99	87.78	91.12	91.83	90.45	97.28

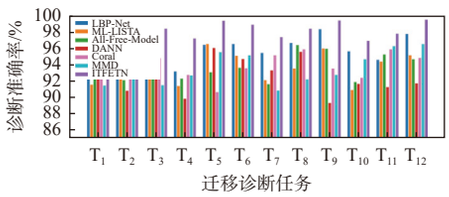


图 7　不同方法在凯斯西储大学轴承数据集迁移诊断任务中的诊断准确率

Fig. 7　Diagnosis accuracies of different methods for transfer diagnosis tasks in bearing dataset of CWRU

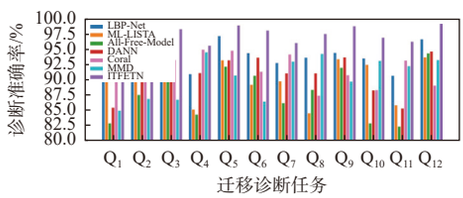


图 8　不同方法在苏州大学轮对轴承数据集迁移诊断任务中的诊断准确率

Fig. 8　Diagnosis accuracies of different methods for transfer diagnosis tasks in wheelset bearing dataset of Soochow University

务，本文所提方法达到了极高的准确率。对于每一个迁移诊断任务，ITFETN 的诊断准确率均在 96.5% 以上，其平均准确率比表现最好的对比方法 LBP-Net 提高了 2.17%。ML-LISTA、All-Free-Model、DANN、Coral、MMD 五种方法的诊断准确率均集中在 90%~97% 之间。而对于 LBP-Net，虽然准确率大多数都已超过 95%，但是每一个迁移诊断任务的准确率依然没有所提方法 ITFETN 高，证明了本文所提方法能够更优异地完成迁移诊断任务。

对于苏州大学轮对轴承数据集上的迁移诊断任务，本文所提方法依然表现优异。ITFETN 只有一个任务的准确率低于 95%，且有多个任务的准确率超过了 98%，其平均准确率比表现最好的对比方法 LBP-Net 提升了 3.53%。6 种对比方法中，均有个别迁移诊断任务的准确率低于 90%，甚至低于 85%，其表现远远不如本文所提方法。

综上，在 2 个轴承数据集设置的迁移诊断任务中，ITFETN 相较于对比方法表现均更加优异。LBP-Net 虽然也是为了解决多层基追踪问题而设计的网络结构，但是其每层内部的基追踪过程是封闭的，层与层之间没有直接联系，而 ITFETN 层与层之间保持着紧密联系，是一个整体的基追踪过程，这保证了所提网络的结构完整性与诊断精确性。ML-LISTA 虽然也是一个算法展开模型，但其具有的双卷积通道反而增加了模型的复杂度，在有限的迭代次数和训练数量下，同时准确学习双卷积通道的权重的难度比学习单通道权重的难度要大得多，这大大限制了 ML-LISTA 的诊断能力。All-Free-Model 虽然有着和 ITFETN 类似的深度和循环结构，但是其未进行算法展开，所有滤波器通道的权重都是自由学习的，而 ITFETN 是一个算法展开模型，其各通道权重的学习过程有着内在的逻辑关系和数值约束，这会使得整个学习过程更加严谨和快速。DANN 虽然是对抗迁移学习的代表方法，但是其模型缺乏可解释性，且迁移诊断表现不如 ITFETN 优异，体现出所提方法的有效性。Coral 和 MMD 虽然是领域自适应的经典方法，但是遇到不同的数据集，尤其是环境干扰较大的数据集，其迁移诊断准确率并未达到理想效果，鲁棒性也不如所提方法。

此外，ITFETN 以及对比方法在凯斯西储大学轴承数据集迁移诊断任务 T4 和苏州大学轮对轴承数据集迁移诊断任务 Q9 上的可解释跨域诊断任务结果的混淆矩阵分别如图 9 和 10 所示。可以看出，ITFETN 在不同的可解释跨域诊断任务中，出现错误诊断的概率小于其他方法。

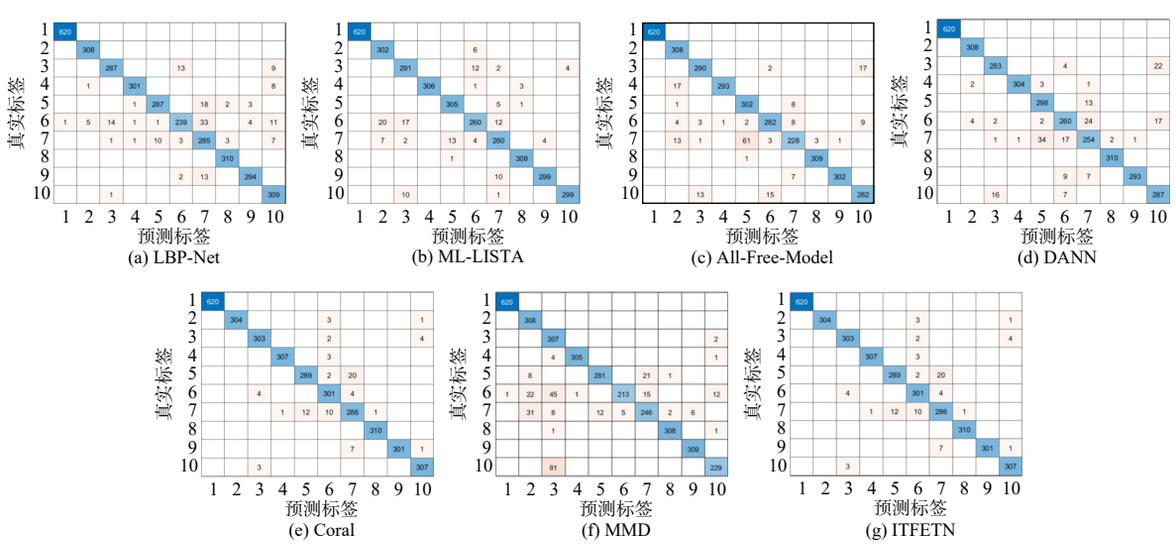


图 9　不同方法在凯斯西储大学轴承数据集迁移诊断任务 T_4 上的混淆矩阵

Fig. 9　Confusion matrixes of different methods for transfer diagnosis task T_4 in bearing dataset of CWRU

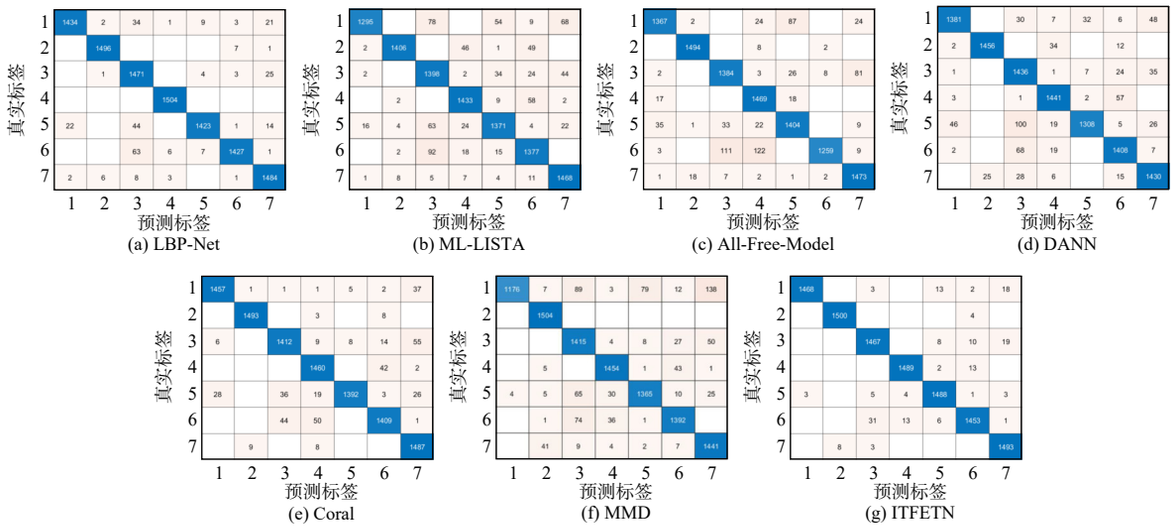


图 10 不同方法在苏州大学轮对轴承数据集迁移诊断任务 Q_9 上的混淆矩阵

Fig. 10 Confusion matrixes of different methods for transfer diagnosis task Q_9 in wheelset bearing dataset of Soochow University

通过以上分析可以看出，ITFETN 不仅诊断准确率高，而且对于不同数据集上的不同迁移诊断任务完成出色，具有高准确率和高鲁棒性的特点。

4 结 论

针对变工况可解释智能诊断，本文提出了可解释三特征提取器迁移网络 ITFETN。一方面，本文建立了多层稀疏编码模型，推导了多层稀疏编码模型的迭代求解算法，通过算法展开得到求解算法的等效网络形式，将其作为特征提取器，用于解决可解释性问题；另一方面，构建了三特征提取器策略用于分别提取源域、目标域的共享特征、私有特征，用于解决跨域迁移诊断问题。通过分析 2 个数据集上跨域迁移诊断任务的试验结果，验证了提出方法 ITFETN 相较于 6 种对比方法，在可解释迁移诊断中具有更高的准确率与鲁棒性，更为优异地完成了可解释迁移诊断任务。

本文的研究也存在部分局限性，如：并未从特征可视化、因果分析、逻辑推理等角度对所提出网络进行事后的可解释性分析。在未来的研究中，一方面，可在训练过程中或训练结束后对所提取的特征进行可视化操作，将其与故障脉冲信号进行相关性分析；另一方面，可利用类激活映射等方法进行显著性分析，计算出提取特征在故障模式识别中起到的贡献值。

参考文献：

[1] ZHAO Z B，LI T F，AN B T，et al. Model-driven deep unrolling：towards interpretable deep learning against noise attacks for intelligent fault diagnosis[J]. ISA Transactions，2022，129：644-662.

[2] ZHANG Y，MALIK O P，CHEN G P. Artificial neural network power system stabilizers in multi-machine power system environment[J]. IEEE Transactions on Energy Conversion，1995，10(1)：147-155.

[3] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[4] SHI M K，DING C C，WANG R，et al. Graph embedding deep broad learning system for data imbalance fault diagnosis of rotating machinery[J]. Reliability Engineering & System Safety，2023，240：109601.

[5] CHEN B J，SHEN C Q，SHI J J，et al. Continual learning fault diagnosis：a dual-branch adaptive aggregation residual network for fault diagnosis with machine increments[J]. Chinese Journal of Aeronautics，2023，36(6)：361-377.

[6] ZHAO Z B，LI T F，WU J Y，et al. Deep learning algorithms for rotating machinery intelligent diagnosis：an open source benchmark study[J]. ISA Transactions，2020，107：224-255.

[7] WANG F J，ZHAI Z，ZHAO Z B，et al. Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis[J]. Nature Communications，2024，15(1)：4332.

[8] LI T F，ZHAO Z B，SUN C，et al. WaveletKernelNet：An interpretable deep neural network for industrial intelligent diagnosis[J]. IEEE Transactions on Systems，Man，and Cybernetics：Systems，2022，52(4)：2302-2312.

[9] ZHAO Y N，LI Y L，ZHANG H C，et al. Deep，convergent，unrolled half-quadratic splitting for image deconvolution[J]. IEEE Transactions on Computational Imaging，2024，10：574-588.

[10] SHAO Y T，LIU Q C，XIAO L. IVIU-Net：implicit variable iterative unrolling network for hyperspectral sparse

unmixing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2023，16：1756-1770.

[11] AN B T，WANG S B，QIN F H，et al. Adversarial algorithm unrolling network for interpretable mechanical anomaly detection[J]. IEEE Transactions on Neural Networks and Learning Systems，2023，35(5)：6007-6020.

[12] LI Y L，TOFIGHI M，GENG J Y，et al. Efficient and interpretable deep blind image deblurring via algorithm unrolling[J]. IEEE Transactions on Computational Imaging，2020，6：666-681.

[13] PENG Y B，JIANG F B，DONG L，et al. GAI-enabled explainable personalized federated semi-supervised learning[J]. arXiv:2410.08634v1，2024.

[14] ZHANG H，HUA J D，LIN J，et al. Damage localization with Lamb waves using dense convolutional sparse coding network[J]. Structural Health Monitoring，2023，22(2)：1180-1192.

[15] ZHOU R，QIAO B J，JIANG L L，et al. A model-based deep learning approach to interpretable impact force localization and reconstruction[J]. Mechanical Systems and Signal Processing，2025，224：111977.

[16] ZHANG Z，XU Y，YANG J，et al. A survey of sparse representation：algorithms and applications[J]. IEEE Access，2015，3：490-530.

[17] AN B T，WANG S B，ZHAO Z B，et al. Interpretable neural network via algorithm unrolling for mechanical fault diagnosis[J]. IEEE Transactions on Instrumentation and Measurement，2022，71：3517011.

[18] PAPYAN V，ROMANO Y，ELAD M. Convolutional neural networks analyzed via convolutional sparse coding[J]. arXiv:1607.08194，2016.

[19] GREGOR K，LECUN Y，GREGOR K，et al. Learning fast approximations of sparse coding[C]// Proceedings of the 27th International Conference on International Conference on Machine Learning. 2010：399-406.

[20] SULAM J，ABERDAM A，BECK A，et al. On multi-layer basis pursuit，efficient algorithms and convolutional neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42(8)：1968-1980.

[21] GANIN Y，LEMPITSKY V，GANIN Y，et al. Unsupervised domain adaptation by backpropagation[C]// Proceedings of the 32nd International Conference on Machine Learning. ACM，2015：1180-1189.

[22] SUN B C，FENG J S，SAENKO K，et al. Return of frustratingly easy domain adaptation[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. ACM，2016，2058-2065.

[23] SMOLA A J，GRETTON A，BORGWARDT K. Maximum mean discrepancy[C]// 13th International Conference. ICONIP，2006.

第一作者：陈　凯(1999—)，男，硕士研究生。
E-mail：2966104038@qq.com

通信作者：朱忠奎(1974—)，男，博士，教授。
E-mail：zhuzhongkui@suda.edu.cn