采用马尔科夫转移场和图注意力网络的滚动轴承故障诊断方法

雷春丽[1,2]，薛林林[1,2]，夏奔锋[1,2]，焦孟萱[1,2]，史佳硕[1,2]

（1.兰州理工大学机电工程学院，甘肃 兰州 730050；
2.兰州理工大学数字制造技术与应用省部共建教育部重点实验室，甘肃 兰州 730050）

摘要：针对实际工程环境复杂多变而导致模型识别准确率不高的问题，提出了一种融合马尔科夫转移场和图注意力网络（Markov transition field and graph attention networks，MTF-GAT）的滚动轴承故障诊断模型。利用MTF保留信号时间相关性的优点，将一维信号转换为二维特征图并定义图的节点和边；利用图注意力层可自适应地对邻近节点分配不同权重的特点，提高模型捕获有用故障特征的能力，并采用深层卷积模块进一步提取图的抽象信息；通过模拟实际工程环境，将各类故障信号输入到训练好的MTF-GAT模型进行故障诊断，并在两个数据集上进行试验验证。结果表明，本文所提出的模型在多种环境下均能准确地完成故障分类任务，相较于其他常用的深度学习模型，MTF-GAT模型具有更好的识别精度和泛化性能。

关键词：故障诊断；滚动轴承；图注意力网络；多头注意力机制；马尔科夫转移场
中图分类号：TH165^{+}.3；TH133.33 **文献标志码：**A **文章编号：**1004-4523(2024)12-2158-10
DOI：10.16385/j.cnki.issn.1004-4523.2024.12.018

引 言

滚动轴承作为核心零部件被广泛应用于现代工业机械设备，滚动轴承出现损伤可能会导致整个设备的严重损坏，造成大量的经济损失[1]。因此，开展滚动轴承故障诊断研究和故障状态的精准识别具有重要的工程意义。

近年来，深度学习因其强大的自动特征提取能力得到了学者们的青睐，在计算机视觉、语音识别、自然语言处理等领域已取得了丰硕的成果[2-4]。随着深度学习和故障诊断技术的发展，越来越多的学者将其引用到故障诊断领域中。文献[5-6]综述了深度学习方法及其在机械装备健康监测方面的相关工作，囊括了目前主流的智能诊断模型。宫文峰等[7]对卷积神经网络（convolutional neural networks，CNN）超参数的选择和训练技巧进行了深度分析，提高了模型结构的通用性和可操作性。CHEN等[8]提出了一种基于循环谱相干和卷积神经网络（cyclic spectral coherence and convolutional neural networks，CSC-CNN）的故障诊断方法，并将其应用于滚动轴承故障识别中，提高了模型的识别准确率并具有良好的泛化性能。孟宗等[9]通过对不平衡数据进行二次数据增强，再利用改进的CNN进行信号的特征提取，提高了模型在轴承故障诊断的通用性。董绍江等[10]利用注意力机制可对通道特征进行权重分配的优点，将其引入CNN中，实现了变工况滚动轴承损伤程度的识别。贾峰等[11]利用迁移学习（transfer learning，TL）和自适应加权方法克服额外故障状态样本的影响，有效实现滚动轴承故障诊断。LIANG等[12]利用一种并行卷积神经网络（parallel convolutional neural networks，P-CNN），融合了时域和频域特征，在少量数据集规模情况下取得了较高的识别准确率。ZHANG等[13]提出了一种深度半监督网络，通过有标签样本和无标签样本相互配合的方式，提高了小样本下故障分类模型的识别准确率。然而，以上方法虽然在轴承故障诊断中取得较好的效果，但并未涉及多种试验条件（变工况、小样本、早期故障识别）下的轴承故障诊断研究。因此，故障诊断模型会在实际应用中出现识别能力降低和泛化能力较弱的现象，其分类稳定性不能得到充分保障。

基于上述分析，本文提出了一种融合马尔科夫转移场和图注意力网络的滚动轴承故障诊断模型。首先，使用MTF将一维信号转换为二位特征矩阵，以保留时间相关性，并定义图的节点和边；其次，将图输入到可以提高模型的特征学习能力的图注意力

收稿日期：2022-12-27；**修订日期：**2023-03-03
基金项目：国家自然科学基金资助项目（51465035）；甘肃省自然科学基金资助项目（20JR5RA466）。

层；然后，通过深层卷积模块进一步提取图的抽象信息；最后，通过模拟实际工程环境进行滚动轴承故障诊断试验，证明了所提方法的有效性与优越性。

1　理论基础

1.1　马尔科夫转移场

马尔科夫转移场（MTF）是一种通过马尔科夫转移概率来表达一维时域数据中信息的方法，可将原始一维信号转化为二维图像[14-15]。该方法通过考虑每个分位数与时间步长之间的依赖关系，保留了原始信号在不同时间间隔内的时间相关性。

假设存在一维时序信号 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，将 $\boldsymbol{X}$ 划分到 Q 个分位数单元中，每个数据点相应的分位数为 $q_j\left(j\in[1,Q]\right)$。然后，沿着时间轴以一阶马尔科夫链的方式计算分位数之间的跃迁来构造 $Q\times Q$ 的马尔科夫转移矩阵 $\boldsymbol{W}$，其表达式为：

$$\boldsymbol{W}=\begin{pmatrix} w_{11|P(x_t\in q_1|x_{t-1}\in q_1)} & \cdots & w_{1Q|P(x_t\in q_1|x_{t-1}\in q_Q)} \\ w_{21|P(x_t\in q_2|x_{t-1}\in q_1)} & \cdots & w_{2Q|P(x_t\in q_2|x_{t-1}\in q_Q)} \\ \vdots & & \vdots \\ w_{Q1|P(x_t\in q_Q|x_{t-1}\in q_1)} & \cdots & w_{QQ|P(x_t\in q_Q|x_{t-1}\in q_Q)} \end{pmatrix} \tag{1}$$

式中　w_{ij} 表示分位数 q_i 位于分位数 q_j 后的概率，即 $w_{ij}=P(x_t\in q_i|x_{t-1}\in q_j)$。

通过考虑时间因素，构建矩阵 $\boldsymbol{M}$ 以获取位置与时间步长之间的依赖关系。矩阵 $\boldsymbol{M}$ 根据分位数与时间步长之间的关系，通过沿时间顺序排列每个概率来扩展矩阵 $\boldsymbol{W}$，保留了额外的时间信息。矩阵 $\boldsymbol{M}$ 的表达式为：

$$\boldsymbol{M}=\begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nn} \end{pmatrix}= \begin{pmatrix} m_{ij|x_1\in q_i,x_1\in q_j} & \cdots & m_{ij|x_1\in q_i,x_n\in q_j} \\ m_{ij|x_2\in q_i,x_1\in q_j} & \cdots & m_{ij|x_2\in q_i,x_n\in q_j} \\ \vdots & & \vdots \\ m_{ij|x_n\in q_i,x_1\in q_j} & \cdots & m_{ij|x_n\in q_i,x_n\in q_j} \end{pmatrix} \tag{2}$$

式中　m_{ij} 表示分位数 q_i 转移到分位数 q_j 的转移概率，即 $m_{ij}=P(q_i\rightarrow q_j)$，其中 $i,j\in[1,Q]$。

1.2　图注意力网络

图注意力网络（GAT）是从空间上考虑图结构的模型，即考虑目标节点和其他节点的几何关系，可以自适应地对邻近节点分配不同的权重，从而将MTF特征矩阵节点的相关性更好地融入到故障诊断模型中。GAT的核心在于注意力机制，对作用较大的节点给予更好的权重，在处理局部信息的时候同时关注整体的信息，其表达式为：

$$Attention(\boldsymbol{Q},\boldsymbol{S})=\sum<\boldsymbol{Q},\boldsymbol{S}>\cdot\boldsymbol{S} \tag{3}$$

式中　$\boldsymbol{S}$ 为所有邻近节点的特征向量；$\boldsymbol{Q}$ 为当前中心节点的特征向量；$\sum$ 为加权求和；故注意力机制就是对所有的节点特征进行加权求和，权重是中心节点与邻近节点特征之间的相关度。

GAT中的图注意力层（graph attention layer，GAL）的输入和输出是一系列节点的特征向量，可表示为：

$$\boldsymbol{h}_0=\{\boldsymbol{h}_{01},\boldsymbol{h}_{02},\cdots,\boldsymbol{h}_{0N}\},\boldsymbol{h}_{0i}\in\mathrm{R}^{F_0} \tag{4}$$

$$\boldsymbol{h}_1=\{\boldsymbol{h}_{11},\boldsymbol{h}_{12},\cdots,\boldsymbol{h}_{1N}\},\boldsymbol{h}_{1i}\in\mathrm{R}^{F_1} \tag{5}$$

式中　$\boldsymbol{h}_0$ 为每个输入节点的特征向量；$\boldsymbol{h}_1$ 为每个输出节点的特征向量；N 为节点数；F_0 和 F_1 分别为输入和输出节点的特征维度。

GAL的结构如图1所示。假设中心节点为 u_i，其邻近节点 u_j，图中中心节点 u_i 有3个邻近节点，本文只考虑一阶邻近节点。通过自注意力机制 att 来计算输入向量的注意力权重系数，再进行Softmax归一化处理，最终得到注意力权重系数 a_{ij}，其操作如下式所示：

$$e_{ij}=att(\boldsymbol{E}\boldsymbol{h}_i,\boldsymbol{E}\boldsymbol{h}_j) \tag{6}$$

$$a_{ij}=\mathrm{Softmax}(e_{ij})=\frac{\exp(e_{ij})}{\sum\limits_{u_k\in N(u_i)}\exp(e_{ij})} \tag{7}$$

式中　e_{ij} 为节点 i 与节点 j 之间的注意力权重；a_{ij} 表示经过归一化后的注意力权重，表示节点 j 对节点 i 的重要程度；$\boldsymbol{E}$ 表示节点从输入特征维度转换为输出特征维度的权重参数矩阵，且 $\boldsymbol{E}\in\mathrm{R}^{F_0\times F_1}$。

自注意力机制 att 可由一个权重向量 $\boldsymbol{a}$ 参数化，并利用激活函数进行非线性化，故注意力权重的表达式可进一步推导为：

$$a_{ij}=\frac{\exp(L(\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{E}\boldsymbol{h}_i,\boldsymbol{E}\boldsymbol{h}_j)))}{\sum\limits_{u_k\in N(u_i)}\exp(L(\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{E}\boldsymbol{h}_i,\boldsymbol{E}\boldsymbol{h}_k)))} \tag{8}$$

式中　$\boldsymbol{a}^{\mathrm{T}}$ 为注意力权重向量的转置，且 $\boldsymbol{a}\in\mathrm{R}^{2F_1}$；$L$ 表示LeakyReLU激活函数。

在获得注意力权重系数后，通过加权求和便可得到中心节点 u_i 的输出特征向量为：

$$\boldsymbol{h}_{1i}=\sigma\Big(\sum_{u_j\in N(u_i)}a_{ij}(\boldsymbol{E}\boldsymbol{h}_j)\Big) \tag{9}$$

式中　$\boldsymbol{h}_{1i}$ 为节点 u_i 新的特征向量；σ 为激活函数，通常使用ReLU函数。

GAL使用了一种多头自注意力机制用以捕获不同的信息，其操作如图2所示。多头自注意力机制通过独立计算 K 组注意力从而获得更全面的信息。注意力头的融合方式一般分为两种，即拼接操作和平均操作，其表达式为：

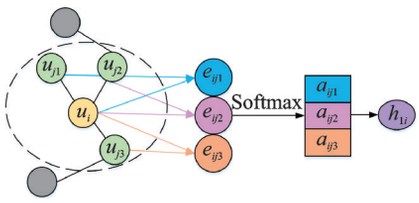


图1 图注意力层

Fig. 1 Graph attention layer

$$\boldsymbol{h}_{1i}=\begin{cases}\Big\Vert_{k=1}^{K}\sigma\Big(\sum\limits_{u_j\in N(u_i)}a_{ij}^{k}(\boldsymbol{E}^{k}\boldsymbol{h}_j)\Big), \text{拼接}\\ \sigma\Big(\dfrac{1}{K}\sum\limits_{k=1}^{K}\sum\limits_{u_j\in N(u_i)}a_{ij}^{k}(\boldsymbol{E}^{k}\boldsymbol{h}_j)\Big), \text{平均}\end{cases}\tag{10}$$

式中 K为注意力头的数量；$\Vert$ 表示拼接操作；a_{ij}^{k}和$\boldsymbol{E}^{k}$分别为第k组自注意力机制的注意力系数和权重参数矩阵。

GAT的中间层通常采用拼接操作，用以提升注意力层的表达能力；而为了避免扩大特征维度，最后一层通常采用平均操作。图2展示了注意力头数量为3时的操作过程，3条实线代表3个相互独立的注意力系数，该操作提高了模型的特征学习能力，同时降低了过拟合的风险。

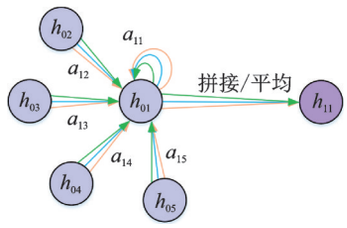


图2 多头自注意力机制

Fig. 2 Multi-head self-attention mechanism

2 滚动轴承故障诊断方法

2.1 MTF-GAT故障诊断模型的构建

本文提出了一种基于马尔科夫转移场与图注意力网络的滚动轴承故障诊断方法，结构如图3所示。首先将原始一维振动信号转化为MTF特征图，保留了原始信号在不同时间间隔内的时间依赖性，并将MTF特征矩阵定义为图；然后将图输入到图注意力层，通过自注意力机制自适应调整节点之间的

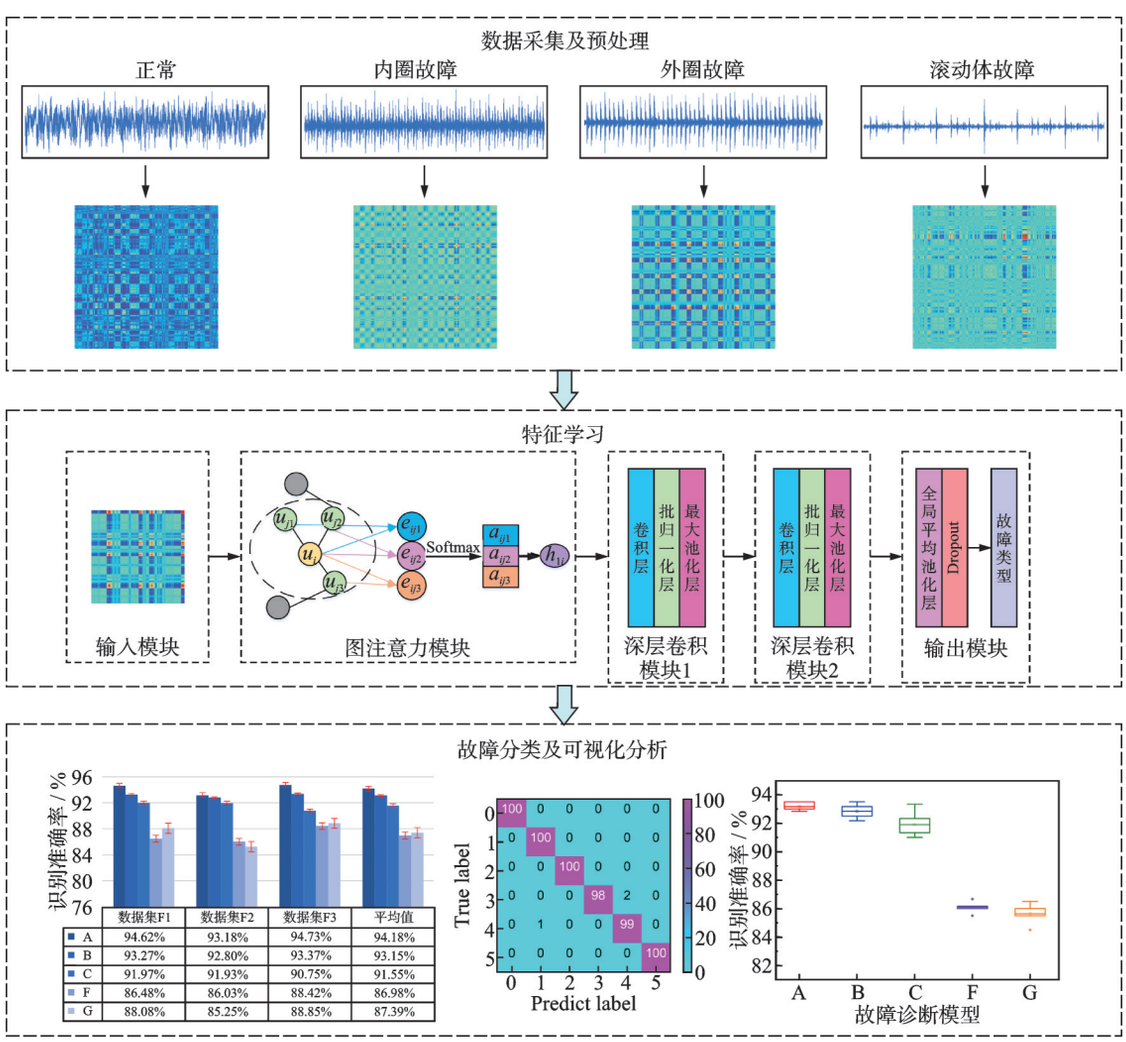


图3 MTF-GAT故障诊断模型

Fig. 3 MTF-GAT fault diagnosis model

注意力系数，增强模型的泛化能力；再经过两个深层卷积模块继续提取信号的深层抽象特征；最后经过全局平均池化层和Dropout层实现对故障的识别分类，其参数如表1所示。

表1　MTF-GAT模型的最优参数

Tab. 1　Optimal parameters of MTF-GAT model

层名	超参数
MTF图像转换层	输入大小为2048，输出大小为[128，128]
模型输入层	输入大小为[128，128]
GAT	注意力头数为4，输入节点特征维度为16
卷积层1	卷积核数目为64，大小为[5，5]，步长为[1，1]
最大池化层1	池化核大小为[2，2]
卷积层2	卷积核数目为128，大小为[5，5]，步长为[1，1]
最大池化层2	池化核大小为[2，2]
全局平均池化层	—
Dropout层	Dropout率大小为0.5
Softmax层	—

2.2　故障诊断流程

本文所提出基于MTF-GAT的滚动轴承故障诊断方法的具体流程如下：

步骤1：采集已知故障滚动轴承原始振动信号，用于模型的训练；

步骤2：按设置的样本长度以重叠采样的数据增强方式随机分割振动信号，如图4所示，并转化为MTF；

步骤3：构建MTF-GAT网络模型，并初始化模型参数；

步骤4：把训练样本输入MTF-GAT模型进行预训练，逐层前向传播获得误差；

步骤5：将步骤4中获得的误差利用Softmax分类函数进行反向传播，选用Adam优化器更新网络参数，使交叉熵损失函数的值达到最小，若达到最优值则进行步骤6，否则跳转到步骤4，直到模型获得最优参数，并保存最佳MTF-GAT模型；

$$J_i=\frac{\mathrm{e}^{\varepsilon_i}}{\sum_{r=1}^{R}\mathrm{e}^{\varepsilon_r}} \tag{11}$$

式中　J_i为每一个Softmax分类函数输出的概率，所有J_i之和为1，$i=1,2,\cdots,R$；ε_i为前层第i个节点输出；R为输出节点的个数，即分类的类别个数。

$$L_{\mathrm{CE}}=-\frac{1}{n}\sum_{\delta}\sum_{c=1}^{N}y_{\delta c}\cdot\ln p_{\delta c} \tag{12}$$

式中　L_{CE}为交叉熵损失函数；N为类别的数量；$y_{\delta c}$为符号函数，若样本δ的真实类别等于c则取1，否则取0；$p_{\delta c}$为观测样本δ属于类别c的预测概率。

步骤6：先将测试样本实行与步骤2相同的操作，然后输入到训练好的MTF-GAT模型中进行滚动轴承故障诊断。

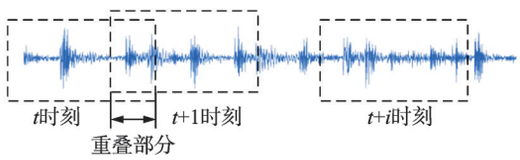


图4　重叠采样操作示意图

Fig. 4　Overlapped sampling operation diagram

3　试验数据集构建

3.1　MFS数据集

数据集1为本实验室MFS试验台上测得的故障数据，试验台实物如图5所示。驱动端故障轴承型号为ER-16K的深沟球轴承，轴承故障采用激光蚀刻技术加工而成，如图6所示，共分为内圈故障、外圈故障和滚动体故障三种故障类型。数据集1采集了轴承转速分别为1200，1300和1400 r/min三种不同工况下的振动信号，信号采样频率为15.36 kHz，采样时间为8 s。试验所用轴承内、外圈故障宽度和滚动体故障孔径分别为1.2和1.8 mm，故障深度均为0.25 mm，共计6种故障类型，根据转速不同可制作成F1，F2和F3三种数据集。

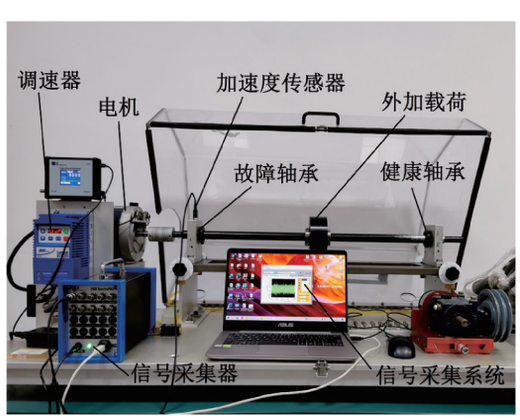
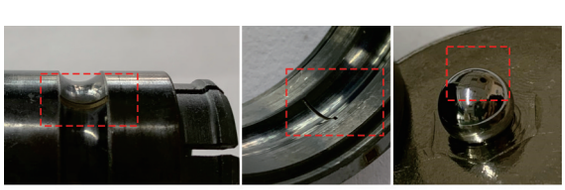


图5　机械故障模拟试验台

Fig. 5　Mechanical fault simulation test bench



(a) 内圈故障 (a) Inner race fault　(b) 外圈故障 (b) Outer race fault　(c) 滚动体故障 (c) Rolling element fault

图6　ER-16K滚动轴承故障部位

Fig. 6　ER-16K rolling bearing fault location

3.2 XJTU-SY 数据集

数据集 2 所用数据为西安交通大学和昇阳科技有限公司联合实验室测得的 XJTU-SY 故障数据集[16-17]，试验台如图 7 所示，主要由交流电动机、电动机转速控制器、加速度传感器、转轴、液压加载系统和测试轴承等组成，试验轴承为 LDK-UER204 滚动轴承，其参数如表 2 所示。由于载荷施加在水平方向，故数据集 2 选用水平加速度传感器采集的振动信号，信号采样频率为 25.6 kHz，采样间隔为 1 min，每次采样时长为 1.28 s。本文选用了外圈故障、混合故障（内圈、外圈）、内圈故障和保持架故障共 4 种故障类型，该数据集主要用于早期故障的分类识别。

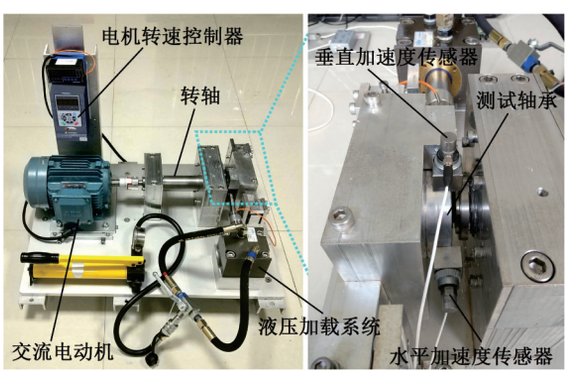


图 7 XJTU-SY 滚动轴承加速寿命试验台

Fig. 7 XJTU-SY rolling bearing accelerated life test bench

表 2 LDK-UER204 滚动轴承参数

Tab. 2 LDK-UER204 rolling bearing parameters

参数名称	数值	参数名称	数值
额定动载荷/N	12820	轴承中径/mm	34.55
接触角/(°)	0	滚动体个数	8
额定静载荷/kN	6.65	滚动体直径/mm	7.92

4 试验验证与分析

本文试验的软件环境为 PyCharm 2020.1.2 中的 Keras 框架，硬件环境为 Intel(R) Xeon-(R) Silver 4110 CPU @2.10 GHz 2.10 GHz 双处理器和 NVIDIA Quadro P4000 显卡。

设置本文所提模型迭代轮次为 60，初始学习率为 5×10^{-4}，注意力头数为 4，卷积核尺寸为 5，每个卷积层均采用批归一化和 ReLU 激活函数，选用 Adam 自适应优化器作为模型优化参数的算法。在本节中，MFS 试验台轴承最低转速为 1200 r/min，XJTU-SY 数据轴承最低转速为 2100 r/min，其采样频率分别为 15.36 和 25.6 kHz。依据公式 $N=f_s\times60/n_s$（N 表示一个周期的采样点数；f_s 表示采样频率；n_s 表示转速）可知，其一个样本周期内的采样点数分别最多为 768 和 731 个。因此，为保证每个样本中故障信息的完整性，设置每个样本数据点数为 2048，MTF 图像大小为 128 × 128。所有试验结果均取 10 次试验的平均值。

4.1 模型验证

4.1.1 输入节点特征维度对模型的影响分析

输入节点特征维度是图注意力层的关键参数之一，因此，需要选择合适的特征维度来保证模型的故障诊断性能。由于论文篇幅原因，本文对特征维度分别为 2，4，8，16 和 24 时模型的诊断性能进行分析，选用数据集 F1 进行试验，设置训练集样本量为 30，测试集样本量为 100，并以识别准确率、四分位数和标准差为评价指标，其结果如表 3 所示。

表 3 不同输入节点特征维度的诊断效果

Tab. 3 Diagnostic effect of different input node feature dimensions

输入节点特征维度	评价指标		
	识别准确率/%	四分位数差/%	标准差/%
2	99.30	0.42	0.25
4	99.33	0.29	0.21
8	99.43	0.30	0.18
16	99.72	0.16	0.15
24	99.62	0.29	0.25

从表 3 可以看出，在输入节点特征维度为 2 时，识别准确率已达到 99.30%，四分位数差和标准差分别为 0.42% 和 0.25%，随着输入节点特征维度的增加，模型的诊断效果也逐步提高，当输入节点特征维度为 16 时模型的分类效果达到最佳，其识别准确率达到 99.72%，四分位数差和标准差分别仅为 0.16% 和 0.15%；随着输入节点特征维度继续增加，模型的诊断效果开始下降，四分位数差和标准差也随之升高。以上现象说明在输入节点特征维度为 16 时可以更精确稳定地完成识别分类任务，故在后续试验中，选取输入节点特征维度为 16。

4.1.2 模型的有效性验证

为验证本文所提 MTF-GAT 模型在故障诊断中的优势，以本文所提模型作为主干网络，建立 4 种故障诊断对比模型，其中 MTF-CNN1 网络模型中将 GAT 中的图注意力层替换为卷积层；MTF-CNN2 网络模型中去除了 GAT 的图注意力层；WDCNN[18] 以原始信号作为输入的宽核一维卷积神经网络。以上所有 CNN 模型的参数设置均相同，对比试验在数据集 F3 上进行，设置训练集样本量为 30，测试集样本量为 100，其结果如表 4 所示。

表 4　不同模型的诊断效果

Tab. 4　Diagnostic effects of different models

模型	评价指标		
	识别准确率/%	四分位数差/%	标准差/%
MTF-GAT	99.50	0.26	0.13
MTF-CNN1	88.67	1.67	1.47
MTF-CNN2	77.47	3.04	2.90
WDCNN	98.33	0.63	0.38

从表 4 可以看出，在数据集 F3 中，MTF-GAT 模型识别准确率最高，达到 99.50%；同时，其四分位数差和标准差最低，分别仅为 0.26% 和 0.13%，说明了所提模型可以准确识别出不同故障类型并具有良好的稳定性。MTF-GAT 模型的识别准确率分别比 MTF-CNN1 和 MTF-CNN2 高 10.83% 和 22.03%，说明了图注意力层可显著提高模型的诊断效果。对比 MTF-GAT 和 WDCNN，前者的识别准确率比后者高 1.17%，四分位数差和标准差分别比后者低 0.37% 和 0.25%，说明了在滚动轴承故障识别中，以 MTF 作为输入的 2D-CNN 相较于以原始信号作为输入的 1D-CNN 具有一定的优势。

为进一步考察 MTF-GAT 模型故障分类的优越性，引入混淆矩阵进行量化分析，如图 8 所示。由图 8 可知，本文所提模型的故障类型区分能力最佳，其他模型均有不同程度误判，尤其是 MTF-CNN1 和 MTF-CNN2，对多种故障产生了较大程度的错分，已不能够准确地完成故障识别任务。综上分析验证了图注意力层在提高模型滚动轴承故障诊断性能上的有效性。

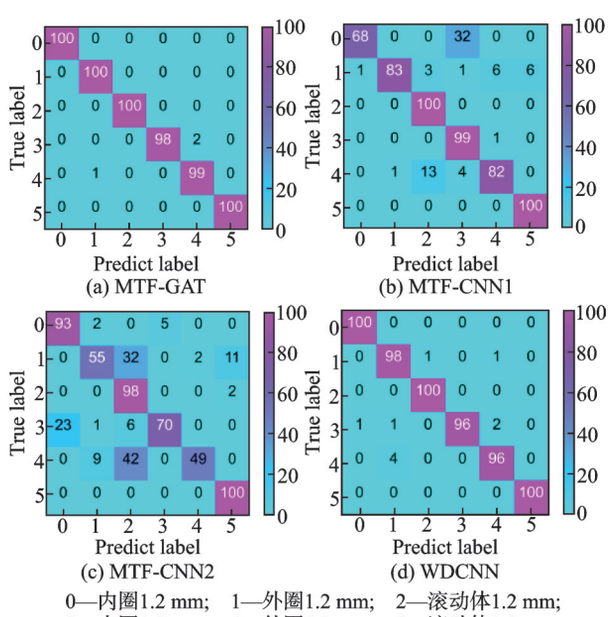


图 8　不同模型故障分类结果的混淆矩阵

Fig. 8　Confusion matrix of fault classification results of different models

4.2　泛化性能分析

机械设备的运行状态复杂多变，测得信号的特征存在明显差异，为验证 MTF-GAT 模型在变工况环境下识别轴承损伤程度的能力，设置了泛化性能试验，通过与 4 种常用的深度学习故障诊断模型进行对比分析。其中 A 表示本文所提模型，采用 MTF-GAT 模型的滚动轴承故障诊断方法；B 表示 2D-CNN 模型[8]，是一种改进的 LeNet-5 模型，在变负载情况下具有较好的表现；C 表示 MSACNN 模型[19]，是一种多尺度 CNN，具有良好的泛化性能；D 表示 WDCNN 模型[18]，具有较好的鲁棒性；E 表示 MCCNN[20]模型，是一种多通道 CNN。在本对比试验中，模型 B 和 C 均以 MTF 作为输入，模型 D 和 E 均以原始信号作为输入，设置训练集样本量为 30，测试集样本量为 100。试验结果如表 5 所示，表中如 F1→F2 表示数据集 F1 用于模型训练，数据集 F2 用于测试。

表 5　不同模型在变工况下的故障识别效果

Tab. 5　Fault identification effect of different models under variable working conditions

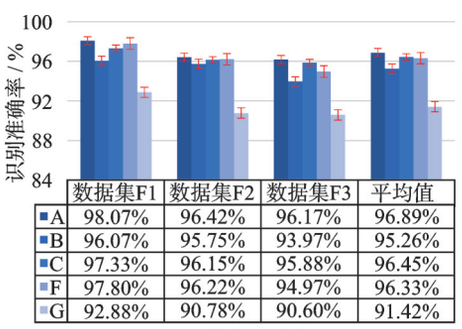
试验工况	识别准确率/%				
	A	B	C	D	E
F1→F2	99.60	98.57	97.93	97.17	97.93
F1→F3	99.55	97.77	98.28	87.22	96.65
F2→F1	99.38	98.55	98.35	92.80	97.63
F2→F3	98.73	98.30	98.55	98.65	98.63
F3→F1	99.42	98.35	98.48	91.82	91.63
F3→F2	98.94	98.42	98.58	98.55	98.55
平均值	99.27	98.33	98.36	94.37	96.84

从表 5 可以看出，本文所提出的 MTF-GAT 模型在 6 种变转速工况下的平均识别准确率为 99.27%，在所有 5 种故障诊断模型中效果最好，分别比模型 B，C，D 和 E 高 0.94%，0.91%，4.90% 和 2.43%。在所有工况中，本文所提模型在 F1→F2 时分类效果最佳，识别准确率达到 99.60%，比其余诊断效果最好的模型 B 高 1.03%；在 F2→F3 时识别准确率最低，为 98.73%，但仍比其余诊断效果最好的模型 D 高 0.08%。此外，模型 A 在 6 种变工况下识别准确率均为 5 种模型中最高，反观其余模型的适应性表现欠佳，如 4 种对比模型中平均识别准确率最高的模型 C，在 F1→F3，F3→F1 和 F3→F2 时的识别性能均优于模型 B，D 和 E，但在 F1→F2 和 F2→F1 时不如模型 B，在 F2→F3 时不如模型 D 和 E。经过上述分析，本文所提出的 MTF-GAT 模型具有优异的变转速自适应能力，

说明引入图注意力层可有效提高模型的泛化性能。

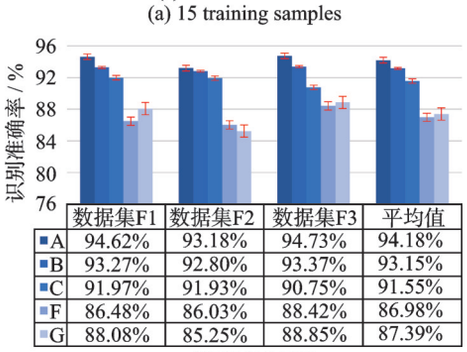
4.3 小样本下模型性能分析

在实际工况中，机械装备长期处于正常运行状态，收集带有故障的数据样本代价昂贵。因此，在小样本条件下模型能否较好完成分类任务是评价模型诊断性能的重要因素。本节将对比分析本文所提出模型和4种深度学习模型在小样本下的故障分类能力。F表示DFCNN模型[21]，G表示ICNN[22]模型，所有模型均以MTF作为输入。分别设置训练样本量为15和8，测试集样本量均为100，试验结果如图9所示。



	数据集F1	数据集F2	数据集F3	平均值
A	98.07%	96.42%	96.17%	96.89%
B	96.07%	95.75%	93.97%	95.26%
C	97.33%	96.15%	95.88%	96.45%
F	97.80%	96.22%	94.97%	96.33%
G	92.88%	90.78%	90.60%	91.42%

(a) 训练样本量为15
(a) 15 training samples



	数据集F1	数据集F2	数据集F3	平均值
A	94.62%	93.18%	94.73%	94.18%
B	93.27%	92.80%	93.37%	93.15%
C	91.97%	91.93%	90.75%	91.55%
F	86.48%	86.03%	88.42%	86.98%
G	88.08%	85.25%	88.85%	87.39%

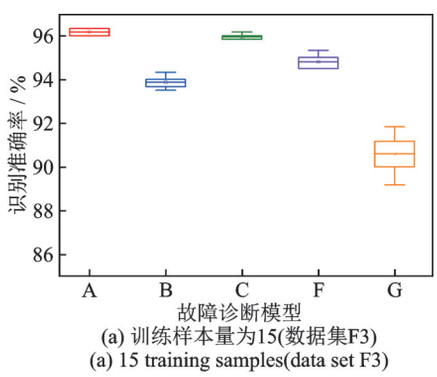
(b) 训练样本量为8
(b) 8 training samples

图9 小样本下不同模型的识别效果柱状图

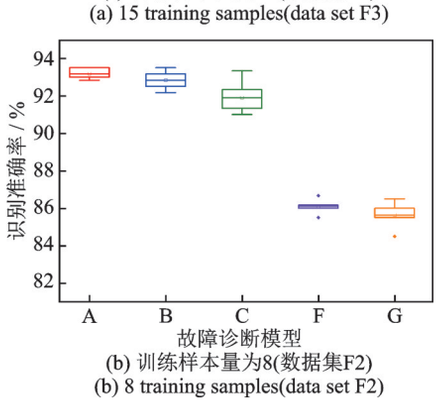
Fig. 9 Diagnostic effect histogram of different models under small samples

从图9(a)可以看出，当训练样本量为15时，本文所提出模型的平均识别准确率最高，达到96.89%，比其余模型分别高1.63%，0.44%，0.56%和5.47%。模型A在数据集F3的分类效果是3个数据集中相对较差的，但其识别准确率仍比4种对比模型中诊断性能最好的模型C高0.29%。从图9(b)可以看出，当训练样本量为8时，本文所提出模型的识别性能最佳，其平均识别准确率为94.18%，比4种对比模型中分类效果最好的模型B高1.03%。而模型F和模型G的平均识别准确率降低至90%以下，已不能较好地完成对滚动轴承故障的精准分类。

在样本量较小时能否对故障类型实现稳定分类是评价模型性能的必要因素，故采用箱型图来说明模型识别性能的稳定性，结果如图10所示。本文所提出的GAT网络模型在小样本下具有良好的识别效果，并且方差较小，而其他4种模型无法在两种训练样本量情况下同时保持足够的稳定性；此外，模型F和模型G在训练样本量为8时还伴有异常值的出现，已无法稳定识别不同的故障类型。综上分析，本文所提出的GAT网络模型在小样本条件下取得了较好的分类效果，同时显著提高了故障识别能力的鲁棒性，验证了图注意力层在提升模型学习能力和特征提取能力的优越性。



(a) 训练样本量为15(数据集F3)
(a) 15 training samples(data set F3)



(b) 训练样本量为8(数据集F2)
(b) 8 training samples(data set F2)

图10 小样本下不同模型的识别效果箱型图

Fig. 10 Box diagram of diagnostic effect of different models under small sample

4.4 早期故障识别性能分析

由于机械设备具有多样性，导致滚动轴承故障发生的时间往往难以确定，若能精确识别轴承早期故障，对设备故障预警和提高轴承甚至整个机械设备的使用寿命具有重要意义。为此，本节通过对比分析检验模型对早期故障的识别性能。文献[23]通过计算JS散度相关系数矩阵清晰地看出不同故障状态的跃迁，并结合相邻时刻的平均相似性，得到了早期故障发生时间点。本节采用该文献计算所得的初始故障时间对应的数据作为本试验的故障数据集，具体故障

样本分布如表6所示，该样本既包含了缓慢失效故障信号又包含了突发失效故障信号。试验设置训练集样本量为8，测试集样本量为100，试验结果如图11所示。

表6　XJTU-SY数据集早期故障样本分布[23]

Tab. 6　Early fault sample distribution of XJTU-SY data set[23]

故障类型	轴承编号	初始故障点/min	试验总时长/min	转速/(r·min^{-1})
外圈	Bearing1_1	77	123	2100
混合故障	Bearing 1_5	33	52	2100
内圈	Bearing 2_1	454	491	2250
保持架	Bearing 2_3	325	533	2250

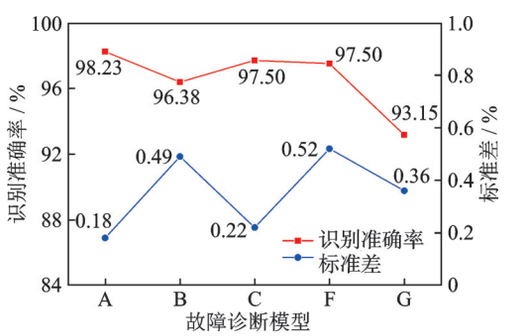


图11　不同模型的早期故障识别效果

Fig. 11　Early fault identification effect of different models

从图11可以看出，在训练样本量仅为8个时，本文模型可以较好地实现对滚动轴承早期故障的分类任务，其识别准确率达到98.23%，比其余模型分别高1.85%、0.53%、0.73%和5.08%；同时，模型A具有更好的稳定性，其标准差仅为0.18%。反观其他模型，除了模型C同时具有较高的识别准确率和较低的标准差，其余模型无法在诊断效果和稳定性上保持一致性，如对比模型F和模型G，虽然前者的识别准确率比后者高4.35%，但模型F的标准差高于模型G。

为了更好地观察不同模型早期故障诊断效果存在差异的原因，通过训练好的模型对测试样本进行故障分类，与实际标签对比，计算出各类健康状态识别准确率，结果如表7所示。从表7中可以看出，各模型均能对混合故障实现准确识别，造成模型诊断性能不同的原因主要是对内圈的分类效果存在较大差异，如模型B和模型G对内圈故障的识别准确率均在90%以下，而模型A的识别准确率虽然不足95%，但依旧比其他效果最好的模型C高1.70%。综上验证了本文所提出的GAT网络模型可以更稳定地完成滚动轴承早期故障的精准识别，且对各类健康状态的分类识别具有更高的准确率。

表7　不同模型对各类故障的识别准确率

Tab. 7　Diagnosis accuracy of different models for various faults

模型	各类故障识别准确率/%				
	外圈	混合故障	内圈	保持架	总识别准确率
A	100	100	94.38	98.54	98.23
B	98.70	100	89.67	97.15	96.38
C	100	100	92.68	98.12	97.70
F	100	100	92.00	98.00	97.50
G	97.53	100	80.73	94.34	93.15

5　结　论

(1)本文提出了一种基于马尔科夫转移场与图注意力网络的滚动轴承故障诊断模型，利用图注意力网络提高模型的特征提取能力并降低了过拟合风险，提高了模型在滚动轴承故障诊断性能上的有效性，为实际工业中轴承的故障诊断提供了方法。

(2)所提MTF-GAT模型在变工况条件下对滚动轴承进行识别分类，在MFS数据集上的平均识别准确率达到99.27%，提高了滚动轴承故障识别的变工况自适应能力；在小样本条件下，GAT网络模型显著提高了模型的稳定性和故障识别分类效果。与其他常用的深度学习模型相比，MTF-GAT模型在变工况和小样本下具有更好的自适应性和鲁棒性，验证了图注意力层在提高模型泛化性能和特征提取能力的优越性。

(3)所提GAT网络可以较好地完成对滚动轴承早期故障的分类任务，在XJTU-SY数据集上的识别准确率达到98.23%，相较于其他深度学习网络模型，GAT网络可以更稳定地完成滚动轴承早期故障的精准识别。

参考文献：

[1] SHEN C Q, QI Y M, WANG J, et al. An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder[J]. Engineering Applications of Artificial Intelligence, 2018, 76: 170-184.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60

(6): 84-90.

[3] NODA K, YAMAGUCHI Y, NAKADAI K, et al. Audio-visual speech recognition using deep learning[J]. Applied Intelligence, 2015, 42(4): 722-737.

[4] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing[J]. IEEE Computational Intelligence Magazine, 2017, 13(3): 55-75.

[5] ZHAO R, YAN R Q, CHEN Z H, et al. Deep learning and its applications to machine health monitoring [J]. Mechanical Systems and Signal Processing, 2019, 115: 213-237.

[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] 宫文峰, 陈辉, 张泽辉, 等. 基于改进卷积神经网络的滚动轴承智能故障诊断研究[J]. 振动工程学报, 2020, 33(2): 400-413.
GONG Wenfeng, CHEN Hui, ZHANG Zehui, et al. Intelligent fault diagnosis for rolling bearing based on improved convolutional neural network[J]. Journal of Vibration Engineering, 2020, 33(2): 400-413.

[8] CHEN Z Y, MAURUCIO A, LI W H, et al. A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks[J]. Mechanical Systems and Signal Processing, 2020, 140: 106683.

[9] 孟宗, 关阳, 潘作舟, 等. 基于二次数据增强和深度卷积的滚动轴承故障诊断研究[J]. 机械工程学报, 2021, 57(23): 106-115.
MENG Zong, GUAN Yang, PAN Zuozhou, et al. Fault diagnosis of rolling bearing based on secondary data enhancement and deep convolution network[J]. Journal of Mechanical Engineering, 2021, 57 (23): 106-115.

[10] 董绍江, 裴雪武, 吴文亮, 等. 改进抗干扰CNN的变负载滚动轴承损伤程度识别[J]. 振动、测试与诊断, 2021, 41(4): 715-722.
DONG Shaojiang, PEI Xuewu, WU Wenliang, et al. Damage degree identification of rolling bearings under variable load with improved anti-interference CNN[J]. Journal of Vibration, Measurement & Diagnosis, 2021, 41(4): 715-722.

[11] 贾峰, 李世豪, 沈建军, 等. 采用深度迁移学习与自适应加权的滚动轴承故障诊断[J]. 西安交通大学学报, 2022, 56(8): 1-10.
JIA Feng, LI Shihao, SHEN Jianjun, et al. Fault diagnosis of rolling bearings using deep transfer learning and adaptive weighting[J]. Journal of Xi'an Jiaotong University, 2022, 56(8): 1-10.

[12] LIANG M X, CAO P, TANG J. Rolling bearing fault diagnosis based on feature fusion with parallel convolutional neural network[J]. The International Journal of Advanced Manufacturing Technology, 2021, 112: 819-831.

[13] ZHANG K, TANG B P, QIN Y, et al. Fault diagnosis of planetary gearbox using a novel semi-supervised method of multiple association layers networks[J]. Mechanical Systems and Signal Processing, 2019, 131: 243-260.

[14] WANG Z G, OATES T. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks[C]//Workshops at the Twenty-ninth AAAI Conference on Artificial Intelligence, Austin, 2015: 1-7.

[15] 雷春丽, 夏奔锋, 薛林林, 等. 基于MTF-CNN的滚动轴承故障诊断方法[J]. 振动与冲击, 2022, 41(9): 151-158.
LEI Chunli, XIA Benfeng, XUE Linlin, et al. Rolling bearing fault diagnosis method based on MTF-CNN [J]. Journal of Vibration and Shock, 2022, 41(9): 151-158.

[16] WANG B, LEI Y G, LI N P, et al. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings[J]. IEEE Transactions on Reliability, 2018, 69(1): 401-412.

[17] 雷亚国, 韩天宇, 王彪, 等. XJTU-SY滚动轴承加速寿命试验数据集解读[J]. 机械工程学报, 2019, 55(16): 1-6.
LEI Yaguo, HAN Tianyu, WANG Biao, et al. XJTU-SY rolling element bearing accelerated life test datasets: a tutorial[J]. Journal of Mechanical Engineering, 2019, 55(16): 1-6.

[18] ZHANG W, PENG G L, LI C H, et al. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals [J]. Sensors, 2017, 17(2): 425.

[19] 丁雪, 邓艾东, 李晶, 等. 基于多尺度和注意力机制的滚动轴承故障诊断[J]. 东南大学学报(自然科学版), 2022, 52(1): 172-178.
DING Xue, DENG Aidong, LI Jing, et al. Fault diagnosis of rolling bearing based on multi-scale and attention mechanism[J]. Journal of Southeast University (Natural Science Edition), 2022, 52(1): 172-178.

[20] 刘好博, 郝洪涛, 丁文捷. 基于IMCKD和MCCNN的滚动轴承故障诊断方法[J]. 振动与冲击, 2022, 41(7): 241-249.
LIU Haobo, HOU Hongtao, DING Wenjie. Fault diag-

nosis method of rolling bearing based on IMCKD and MCCNN[J]. Journal of Vibration and Shock, 2022, 41(7): 241-249.

[21] ZHANG J Q, SUN Y, GUO L, et al. A new bearing fault diagnosis method based on modified convolutional neural networks[J]. Chinese Journal of Aeronautics, 2020, 33(2): 54-62.

[22] 许同乐，孟良，孔晓佳，等. 基于EEMD的ICNN故障诊断方法[J]. 北京邮电大学学报，2022，45(2)：110-116.
XU Tongle, MENG Liang, KONG Xiaojia, et al. ICNN fault diagnosis method based on EEMD[J]. Journal of Beijing University of Posts and Telecommunications, 2022, 45(2): 110-116.

[23] 黄如意，李霁蒲，王震，等. 基于多任务学习的装备智能诊断与寿命预测方法[J]. 中国科学：技术科学，2022，52：123-137.
HUANG Ruyi, LI Jipu, WANG Zhen, et al. Intelligent diagnostic and prognostic method based on multi-task learning for industrial equipment[J]. Scientia Sinica Technologica , 2022, 52: 123-137.

Rolling bearing fault diagnosis method based on Markov transition field and graph attention network

LEI Chun-li[1,2], *XUE Lin-lin*[1,2], *XIA Ben-feng*[1,2], *JIAO Meng-xuan*[1,2], *SHI Jia-shuo*[1,2]

(1.School of Mechanical and Electronical Engineering, Lanzhou University of Technology, Lanzhou 730050, China;
2.Key Laboratory of Digital Manufacturing Technology and Application, Ministry of Education, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: Aiming at the problem that the recognition accuracy of the model is not high due to the complex and variable engineering environment, a rolling bearing fault diagnosis model integrating Markov transition field and graph attention networks (MTF-GAT) is proposed in this paper. Using the advantage of MTF to retain the time correlation of the signal is applied to transform one-dimensional signals into two-dimensional feature maps, and the nodes and edges of the graph are defined. The graph attention layer can adaptively assign different weights to adjacent nodes to improve the ability of the model to capture useful fault features, and the abstract information of the graph is further extracted through the deep convolution module. By simulating the actual engineering environment, the various fault signals are input into the trained MTF-GAT model for fault diagnosis, and the model is verified by experiments on two data sets. The results show that the proposed model in this paper can accurately complete the task of fault classification in a variety of environments. Compared with other deep learning models, the MTF-GAT model has better recognition accuracy and generalization performance.

Key words: fault diagnosis; rolling bearings; graph attention networks; multi-head attention mechanism; Markov transition field

作者简介：雷春丽(1977—)，女，博士，教授。 E-mail：lclyq2004@163.com。

通讯作者：夏奔锋(1996—)，男，硕士研究生。 E-mail：xbf3511826@163.com。