doi:10.3772/j.issn.1002-0470.2021.01.001

基于 Transformer 神经网络的滚动轴承故障类型识别①

邱大伟②*** 刘子辰* 周一青③*** 龙 隆*** 谭雯雯*** 曹 欢***

(*中国科学院计算技术研究所移动计算与新型终端北京市重点实验室 北京 100190)

(**中国科学院大学 北京 100049)

(***北京化工大学信息科学与技术学院 北京 100029)

摘 要 工程应用中的滚动轴承故障类型识别要求同时具有较高的识别准确度和时间效率,基于上述需求提出基于 Transformer 神经网络的滚动轴承故障类型识别方法。所提方法结合小波包变换时频域能量特征和快速傅里叶变换频域特征生成满足 Transformer 神经网络的输入样本矩阵,解决 Transformer 神经网络的输入问题。同时,提出应用于滚动轴承故障类型识别的归一化位置编码方法,解决 Transformer 神经网络在滚动轴承故障分析领域的位置编码问题。在此基础上,提出 Transformer 神经网络双向输入样本矩阵处理机制和算法训练过程中错误样本权重增强机制,提升所提方法的鲁棒性。使用 KAt 数据中心的滚动轴承数据集验证所提方法的识别性能,与现有常用深度学习方法相比,所提方法在时间效率和准确度性能上均有一定的优势,其中,准确度能够提升 11% 以上,单个样本的平均处理时间小于 1 ms。

关键词 滚动轴承;故障类型识别;Transformer 神经网络;前向特征矩阵;后向特征矩阵;归一化位置编码;权重增强

0 引 言

滚动轴承作为机械设备的重要组成部分,其健康的运行状态是机械设备正常运行的基础。然而,在实际工程应用中,如在无人拖拉机的应用中[1],滚动轴承经常由于长时间运行或者恶劣的环境因素而出现各种故障,如塑形变形,故障严重时会导致整个轴承烧毁,甚至导致机械设备的损坏。但是在对滚动轴承的日常维护中,对于出现损伤的轴承往往进行整体替换,无法针对具体的故障部位进行有效更换,在频繁维修时会造成轴承的浪费和经济损失。因此,本文针对上述问题对工业应用中常见的滚动轴承故障进行具体分析,为滚动轴承的维护和保养提供有力保障。

基于以上需求,研究人员展开对滚动轴承故障类型识别的研究。现有研究方法中,有很多文献针对滚动轴承故障类型检测进行研究。由于支持向量机(support vector machine,SVM)在分类方面取得一定成果,很多研究者将 SVM 应用于滚动轴承故障诊断,如使用簇间距离优化的 SVM 进行滚动轴承故障诊断[2]、使用最大类分离性的二元粒子群优化算法的 SVM 进行滚动轴承的故障诊断[3]等。SVM 属于早期的机器学习算法,对样本量的需求相对较少,传统数据模式下均能取得较好的效果,但是在物联网和大数据时代,样本量剧增,SVM 算法由于其轻量

① 北京市自然科学基金(L172049),中国科学院科技促进经济社会发展 STS 项目(KFJ-STS-ZDTP-057)和中国科学院智能农业机械装备工程实验室(GC201907-02)资助项目。

② 男,1991 年生,博士生;研究方向:农机设备故障诊断分析;E-mail: qiudawei@ict.ac.cn

③ 通信作者,E-mail: zhouyiqing@ict.ac.cn

(收稿日期:2019-12-23)

化的设计难以再提高算法的准确度,因而需要更复杂的方法结合大量数据来提高算法的准确性。在这样的背景驱动下,有研究者提出了具备深度堆叠结构可学习的最小二乘支持向量机用于滚动轴承故障诊断,能够做到自动提取故障诊断的敏感特征[4]。因此,在大数据时代,具备强大的学习能力及不同要素映射能力的算法将更有发展空间。

物联网、大数据和人工智能技术的快速发展,为滚动轴承的故障类型识别提供了大数据量和智能化处理的基本保证;同时,随着信息通信技术的快速发展[5-7],也为数据的实时传输提供了保证。但是如何在这样的背景下进行准确、高效的滚动轴承故障类型识别依然是一个挑战。现有大量研究方法被提出并用于滚动轴承的故障类型识别。如文献[8]提出使用经验模式分解进行特征提取,然后使用人工神经网络(artificial neural network,ANN)进行故障诊断。随着深度学习技术越来越成熟,研究者将深度学习算法应用于滚动轴承的故障诊断,如使用深度神经网络进行滚动轴承故障诊断[9],但是深度神经网络存在参数膨胀的问题,导致其工程应用性不强。为了解决这个问题,目前提出的比较典型的方法是使用卷积神经网络(convolutional neural network,CNN)进行滚动轴承的故障诊断[10-13],其能够通过局部连接和权值共享减少神经网络需要训练的参数个数,具备简单化的结构设计,使网络更容易训练,能有效应对大数据量的问题,同时能够保证较高的模型准确度性能。但是 CNN 对于输入时间序列的远距离特征捕获能力较差,远距离捕获能力是提升模型性能的有效手段,而增加其远距离特征捕获能力的代价是增加神经网络的深度,无疑进一步增加了算法的复杂度。为了有效解决神经网络的远距离捕获能力,长短时记忆(long-short term memory,LSTM)网络被提出,并用于滚动轴承的故障诊断[14]。也有研究者将 CNN 和 LSTM 结合,提出了卷积双向 LSTM 网络[15]进行滚动轴承故障诊断,使用 CNN 进行特征提取,然后使用双向 LSTM 对特征进行分类。

LSTM 具有强大的时间序列远距离特征捕获能力得益于它的长短时记忆结构和循环处理结构,但是这种结构使它很难做到并行计算,导致算法的时间效率低,因此在工程中应用的并不多。那么,如何做到既能捕获时间序列的远距离特征又能进行并行计算保证较高的时间效率一直是研究者面临的难点问题。在此需求下,Transformer 神经网络应运而生。Transformer 是一种基于注意力机制的神经网络,包含的注意力机制可以很轻松地解决时间序列的远距离特征捕获问题,而位置编码能够使神经网络在处理输入序列时考虑时间序列的顺序性问题。截至目前,Transformer 已经在自然语言处理的多个领域取得较好的效果,比如机器翻译、语义抽取及阅读理解等。但是,其独特的结构限制了直接应用于滚动轴承故障分析领域,如滚动轴承故障分析领域的数据与自然语言的词向量间的有效映射和自然语言处理领域的位置编码并不适合于滚动轴承故障分析领域。本文针对上述限制展开研究,将 Transformer 神经网络由自然语言处理领域引入到滚动轴承故障分析领域,进行滚动轴承的故障类型识别,属于 Transformer 神经网络的全新应用,为其在工业领域增加了新的应用场景。

为了解决上述限制和问题,本文提出了基于 Transformer 神经网络的滚动轴承故障类型识别方法,论文的贡献总结如下。

(1) 提出结合小波包变换时频域能量特征和快速傅里叶变换频域特征生成满足 Transformer 神经网络的输入样本矩阵,解决 Transformer 神经网络的输入问题。

(2) 提出应用于滚动轴承故障类型识别的归一化位置编码方法,解决 Transformer 神经网络在滚动轴承故障分析领域的位置编码问题。

(3) 提出 Transformer 神经网络双向输入样本矩阵处理机制并提出算法训练过程中错误样本权重增强机制,提升了算法的鲁棒性。

论文的其余部分组织如下。第 1 节给出滚动轴承故障类型识别方法,同时,介绍本文使用的特征提取方法及构成样本矩阵的流程。第 2 节介绍本文在进行滚动轴承故障类型识别时使用的 Transformer 神经网络。第 3 节给出所提方法的性能评估结果。第 4 节给出论文总结和未来工作方向。

1 滚动轴承故障类型识别方法

1.1 故障类型识别框架

本文使用 Transformer 神经网络识别滚动轴承故障类型。设 ftd 是滚动轴承真实故障类型,s 是滚动轴承振动数据。借助 Transformer 神经网络可以使用 s 计算 ftd,计算公式为

$$ftd = \mathrm{T}\left(\left(\begin{bmatrix}\vec{f}_1(s)\\ \vdots \\ \vec{f}_n(s)\end{bmatrix}_{i\times j} + \boldsymbol{E}_{i\times j}\right), \left(\begin{bmatrix}\overleftarrow{f}_n(s)\\ \vdots \\ \overleftarrow{f}_1(s)\end{bmatrix}_{i\times j} + \boldsymbol{E}_{i\times j}\right)\right) \pm \sigma \tag{1}$$

$$ftd \approx \mathrm{T}(\cdot) \tag{2}$$

式中,$f_1,\cdots,f_n$ 是滚动轴承振动数据 s 的特征,所有特征组成特征矩阵,矩阵尺寸为 $i\times j$,$\vec{f}$和$\overleftarrow{f}$为特征的前向和后向排列。$\boldsymbol{E}$ 是本文提出的归一化位置编码矩阵,其尺寸与特征矩阵的尺寸一致,为 $i\times j$。T 是 Transformer 神经网络故障类型识别模型。σ 表示模型预测值与真实值的偏差,在工程应用中,σ 为随机变量,很难进行预测,因此,故障类型识别模型得到的识别结果近似为滚动轴承故障类型,如式(2)所示。总之,本文的目标是设计特征提取方法构建前向特征矩阵、后向特征矩阵和设计位置编码方法构建位置信息,使用 Transformer 神经网络从前向和后向处理输入数据得到滚动轴承故障类型识别结果。

本文提出的使用 Transformer 神经网络的滚动轴承故障类型识别的通用架构如图 1 所示,图中所示架构包含 3 个部分,分别是数据预处理、识别模型构建和滚动轴承故障类型识别。原始振动数据经过数据预处理后得到前向和后向排列的特征矩阵。数据预处理包含 4 个数据处理部分,分别是时频域特征提取、频域特征提取、前向特征矩阵构建和后向特征矩阵构建。识别模型构建是滚动轴承故障类型识别的核心部分,主要包括创建模型和训练模型两部分。本文设计归一化位置编码使 Transformer 神经网络可以分别捕获前向特征矩阵和后向特征矩阵数据的位置信息,在构建完成 Transformer 神经网络故障类型识别模型后,前向特征矩阵和后向特征矩阵均被送入 Transformer(构成双向数据处理机制)进行模型的训练,模型训练时使用提出的错误样本权重增强机制。训练过程中,当使用验证数据集得到最大的验证准确度时,保存故障类型识别模型为最优模型。最后,可以使用测试数据集测试最优滚动轴承故障类型识别模型的性能。

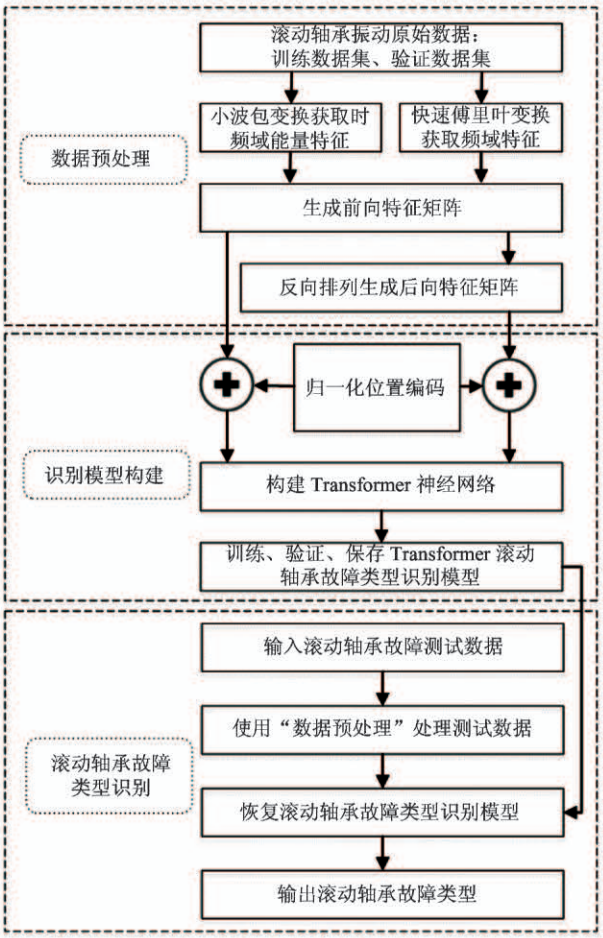

图 1 滚动轴承故障类型识别通用架构

1.2 数据预处理

1.2.1 时频域特征提取

轴承滚动运转过程中由于发生故障(如剥落、松动、裂纹及断裂)将导致动态振动信号的非平稳性的现象(信号的统计特征,如均值、方差等,随时间的变化而变化)。而这些非平稳性信号是表征某些故障最直接的工具。因此,进行故障类型识别的关键是对轴承动态信号的非平稳性进行有效的分析。已有研究表明[16],小波包变换能够在时频域中对信号进行分析,滚动轴承的非平稳振动信号包含

的频谱分布与滚动轴承的特征结构、故障类型密切相关，非平稳振动信号经小波包分解后在最底层上不同正交小波包空间的能量分布可以表示故障特征频率谱，是滚动轴承运行状况的本质特征。因此，本文引用小波包变换技术获取轴承序列数据的小波包能量特征，作为滚动轴承运行状态提取的时频域特征。

小波包分解及其系数重构可参看文献[16]。

本文所使用的小波包空间能量算子 $E(j, n)$ 定义如下：

$$E(j, n) = \sqrt{\sum_{k \in Z} [| p_s(n, j, k) |]^2} \quad (3)$$

其中，$p_s(n, j, k)$ 是小波包变换系数，s 为原始信号，j 为小波包分解层数，n 为第 j 层节点索引。

对原始信号 $s(t)$ 进行 J 层小波包分解，在分解层数 J 上信号 $s(t)$ 的小波包能量特征表示如下：

$$C(J, s) = [E(J, 0), E(J, 1), \cdots, E(J, 2^J - 1)] \quad (4)$$

本文采用的小波包变换提取时频域能量特征过程如算法 1 所示。

算法 1　小波包变换时频域能量特征提取

输入：滚动轴承振动原始数据 *data*；
输出：滚动轴承能量特征 *f_tf*；
1：加载 *data*；
2：for $i \in \{1, \cdots, \text{size}(data, 1)\}$ do
3：　使用 db10 小波对 $data(i, :)$ 进行 6 层小波包分解，得到小波树 *tree*；
4：　for $j \in \{1, \cdots, 2^6\}$ do
5：　　重构 *tree* 第 6 层 j 处的小波包系数，得到 *coef*；
6：　　使用 *coef* 利用式(3)计算小波包空间能量 $energy(j)$；
7：　end for
8：　*energy* 标准化为均值为 0、标准差为 1 的标准化数据 $f\_tf(i, :)$；
9：end for
10：return *f_tf*.

1.2.2　频域特征提取

虽然上述小波包变换得到的能量特征能够在时频域有效地区分滚动轴承非平稳振动信号，但是为了更直观地表示振动信号的频率特征，本文使用快速傅里叶变换对滚动轴承振动信号进行频域信号特征提取，快速傅里叶变换的具体变换公式可参看文献[17]。本文采用的快速傅里叶变换提取频域特征过程如算法 2 所示。

算法 2　快速傅里叶变换频域特征提取

输入：滚动轴承振动原始数据 *data*，采样点数 *N*；
输出：滚动轴承频域特征 *f_f*；
1：加载 *data*；
2：for $i \in \{1, \cdots, \text{size}(data, 1)\}$ do
3：　使用快速傅里叶变换处理 $data(i, :)$，得到变换值 *Y*；
4：　取 *Y* 的幅值，并换算成实际的幅度 *A*；
5：　取 *A* 的前半部分作为傅里叶变换特征值 $f\_f(i, :)$；
6：end for
7：return *f_f*.

1.2.3　特征矩阵构建

为了满足 Transformer 神经网络的输入需求，将上述时频域特征和频域特征进行组合，构成前向特征矩阵。前向特征矩阵的生成过程如图 2 所示。

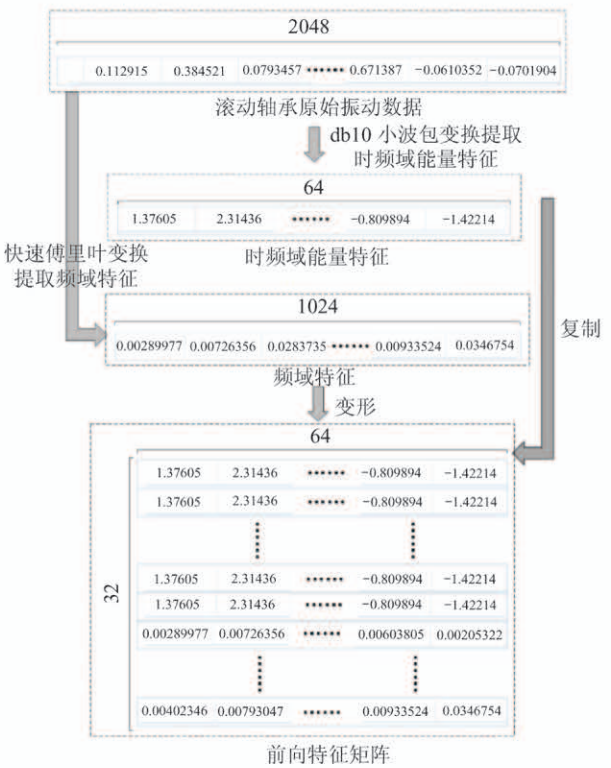

图 2　前向特征矩阵构造

图 2 所示使用 db10 类型的小波基函数对原始数据进行小波包变换，获取时频域能量特征，使用的

小波包分解层数为 6 层,得到 64 个时频域特征。使用快速傅里叶变换提取原始信号的频域特征,特征点数为 1024。对时频域能量特征使用复制机制和对频域特征使用变形机制,构成尺寸为 32 ×64 的前向特征矩阵。

对于上述过程需要说明的是,由于滚动轴承振动信号是离散信号,所以小波基函数选择的原则是必须具有离散小波变换能力,而选用不同的小波基函数,小波包变换的结果也有差异。本文所使用的 db10 小波基函数仅提供示例性应用。对于小波包分解层数的选择也可由读者自行尝试使用不同的层数。对于特征矩阵的尺寸也不仅仅局限于 32 ×64。

得到前向特征矩阵后,通过式(5)得到后向特征矩阵,然后即可将前向特征矩阵和后向特征矩阵输入 Transformer 神经网络进行 Transformer 故障类型识别模型的训练。

$$\begin{bmatrix} f_{1,1} & \cdots & f_{1,16} \\ \vdots & \ddots & \vdots \\ f_{32,1} & \cdots & f_{32,26} \end{bmatrix} \Rightarrow \begin{bmatrix} f_{32,26} & \cdots & f_{32,1} \\ \vdots & \ddots & \vdots \\ f_{1,26} & \cdots & f_{1,1} \end{bmatrix} \tag{5}$$

2 Transformer 神经网络故障类型识别模型

2.1 模型构建

完整结构的 Transformer 神经网络具有编码器-解码器结构[18]。编码器用于将输入序列的符号表示映射为序列的连续表示,解码器即将连续表示恢复为符号表示。完整的编码器-解码器结构更多地用于语言模型和翻译模型的建模。结合滚动轴承故障类型识别任务的特点,只需要计算出输入序列的连续表示,即可得到故障类型。所以本文中只需使用 Transformer 的编码器部分即可完成滚动轴承的故障类型识别。本文使用的基于 Transformer 的滚动轴承故障类型识别模型如图 3 所示,模型由 4 部分组成,即归一化位置编码、前向数据 Transformer 处理层、后向数据 Transformer 处理层和全连接 & Argmax 层。归一化位置编码使 Transformer 在处理滚动轴承特征数据时考虑矩阵的位置信息,关于归一化位置编码的详细介绍将在下节给出。

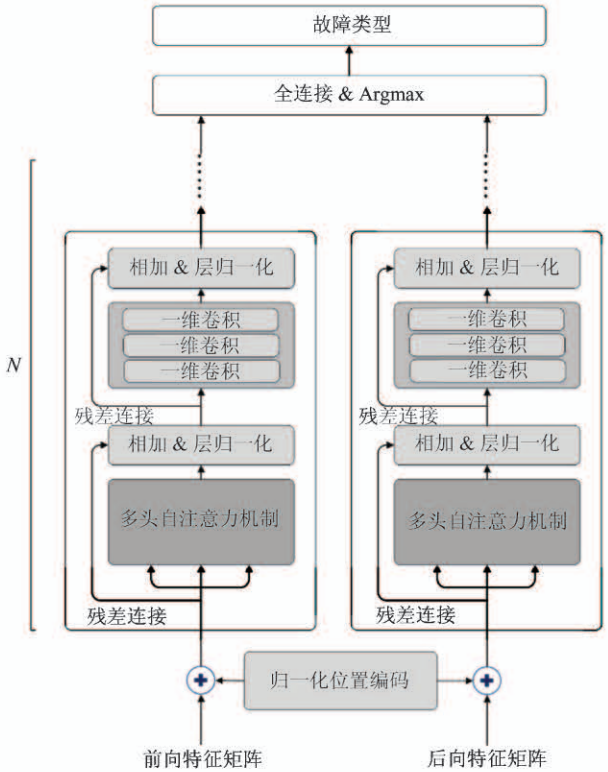

图 3 基于 Transformer 的滚动轴承故障类型识别模型

前向数据 Transformer 处理层和后向数据 Transformer 处理层在结构上是相同的,区别仅在于输入数据的不同。每个 Transformer 处理层由 $N = 12$ 的相同层堆叠排列而成。每一个相同层包含两个子层,分别是多头自注意力机制层和一维卷积层。为了使网络结构具备通过增加深度来提高准确率的能力,即解决增加深度带来的副作用(退化问题,指层数逐渐增加后,出现准确率反而下降的情况),每个子层内部均使用残差连接[19],同时,每个子层末端使用层归一化[20]以提高神经网络的训练速度和泛化性能。因此,每个子层的输出可表示为

$$o = \mathrm{LayerNorm}(x + \mathrm{Sublayer}(x)) \tag{6}$$

其中,Sublayer(x)是每个子层内部的函数,本文中为自注意力机制层处理函数和一维卷积层处理函数。LayerNorm(·)为层归一化处理函数。

多头自注意力机制可以描述为将查询和一组键-值对映射到输出,输出是一组值的加权和,其中分配给每个值的权重由使用键值进行查询计算得到。多头注意力机制允许模型在不同位置共同关注来自不同表示子空间的信息。多头自注意力机制的

计算过程如下：

$$\begin{aligned}&\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Concat}(head_1,\cdots,head_h)\cdot\boldsymbol{W}^O\\&head_i = \mathrm{Attention}(\boldsymbol{Q}\cdot\boldsymbol{W}_i^Q\cdot\boldsymbol{K}\cdot\boldsymbol{W}_i^K,\ \boldsymbol{V}\cdot\boldsymbol{W}_i^V)\\&\mathrm{Attention}(\boldsymbol{Q},\ \boldsymbol{K},\ \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V}\end{aligned}\tag{7}$$

其中，参数矩阵 $\boldsymbol{W}_i^Q\in\mathbb{R}^{d_{\mathrm{model}}\times d_k}$，$\boldsymbol{W}_i^K\in\mathbb{R}^{d_{\mathrm{model}}\times d_k}$，$\boldsymbol{W}_i^V\in\mathbb{R}^{d_{\mathrm{model}}\times d_k}$，$\boldsymbol{W}_i^O\in\mathbb{R}^{d_{\mathrm{model}}\times d_k}$。$h$ 是头的数量，d_k 是查询或键的尺寸。将同时对一组查询计算注意力函数，并将它们打包到矩阵 $\boldsymbol{Q}$ 中。键和值也打包到矩阵 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 中。Concat 函数用于拼接多头注意力计算得到的输出。softmax 函数用于获得值的权重。

一维卷积由 3 个一维卷积串联而成，构成全连接前馈网络。由于需要从多头自注意力机制的输出片段中获取感兴趣的特征，同时该特征在片段中的位置不具有高度相关性，所以本文使用 3 个一维核尺寸为 1 的卷积网络对多头自注意力机制的输出进行处理。

滚动轴承特征矩阵数据经过 Transformer 层处理后，最后再经过全连接 &Argmax 层处理。该层对 Transformer 层输出的前向特征矩阵的处理结果和后向特征矩阵的处理结果进行非线性映射，得到最终的滚动轴承故障类型识别结果。

2.2 归一化位置编码

对于 Transformer 来说，为了能够保留输入数据之间的相对位置信息，必须要添加额外的处理。目前主流的神经网络均具有位置信息编码能力，如 LSTM 的循环结构本身已经包含数据的位置信息，CNN 使用的搜索窗也包含了数据的位置信息。原始 Transformer 神经网络在设计时考虑了引入位置信息编码，为每个输入数据增加一个位置编码信息。在自然语言处理领域使用 Transformer 神经网络时，有很多的位置编码信息可供使用，如固定位置编码[21]、正余弦位置编码和 2D 位置编码[22]。不同的位置编码方法在不同的应用中效果均有差别，目前尚没有定论说明某种位置编码方法适合于某种具体的应用。由于滚动轴承特征数据的长度固定，同时为了直观且简单地体现滚动轴承特征矩阵数据间的位置关系，本文提出一种归一化位置编码方法，其计算公式为

$$\begin{aligned}&p = \{1,\ 2,\ \cdots,\ n\}\\&PE_i = \frac{p_i - \min}{\max - \min}\\&E = \mathrm{reshape}(PE)\end{aligned}\tag{8}$$

其中，n 是滚动轴承特征总数，min 是 p 的最小值，max 是 p 的最大值。reshape(·)函数将 PE 变换为与滚动轴承特征矩阵的尺寸相同的矩阵。将位置编码矩阵和滚动轴承特征矩阵进行相加即可使 Transformer 神经网络的输入具备位置编码信息。

2.3 模型训练

在模型训练过程中，对于训练数据集中识别错误的样本采用权重增强机制处理，增加模型对错误样本的处理次数。在模型训练时，由模型识别值和标准值计算 softmax 交叉熵并求取平均值作为损失值，然后使用 ADAM[23] 进行梯度更新并最小化损失值，完成模型的训练。模型训练过程如算法 3 所示。

算法 3 模型训练

输入：全部训练样本 S；上一次迭代后训练样本的故障类型识别结果 R_{-1}；上一次迭代后训练样本的真实故障类型 L_{-1}；本次迭代后训练样本的故障类型识别结果 R；本次迭代后训练样本的真实故障类型 L；损失值 $loss$；学习率 lr；每次迭代选择的训练样本数量 B；本次训练时选择的训练样本 S_w。

输出：训练后的故障类型识别模型；

比较 R_{-1} 和 L_{-1}，得到识别错误样本 E_s 错误数量 n；

if $n=0$ then

　S_w←从训练样本集 S 随机选择 B 个样本；

else

　S_w←从训练样本集 S 随机选择 $B-n$ 个样本 + E_s；

end if　% 训练错误样本的权重增强机制

向故障类型识别模型中输入 S_w；

得到本次迭代故障类型识别结果 R；

$loss$ = reduce_mean(softmax_cross_entropy_with_logits(R, L))；　% 由模型识别值和标准值计算 softmax 交叉熵并求取平均值作为 $loss$

train = AdamOptimizer(lr).minimize($loss$)；　% 使用 ADAM 进行梯度更新并最小化损失值

return 故障类型识别模型；

3 模型仿真和分析

3.1 数据描述及参数配置

为了验证本文所提方法在实际应用中的有效性,数据集使用 KAt 数据中心的滚动轴承数据[24]。本文所用数据集均来自真实故障,所用轴承数量为15,其中 5 个为健康状态的轴承,5 个为外圈损坏故障的轴承,5 个为内圈损坏故障的轴承。所有轴承数据在获取时的转速为 900 rpm,负载扭矩为 0.7 Nm,径向力为 1000 N。表 1 列出训练数据集和测试数据集的划分。其中,使用 9 个轴承数据组成训练数据集,使用 6 个轴承数据组成测试数据集。

表 1 滚动轴承数据集

轴承状态	轴承	训练/测试
健康	K001	训练
	K004	训练
	K005	训练
	K002	测试
	K003	测试
外圈损伤故障	KA04	训练
	KA22	训练
	KA30	训练
	KA15	测试
	KA16	测试
内圈损伤故障	KI04	训练
	KI18	训练
	KI21	训练
	KI14	测试
	KI16	测试

上述所用数据集经过分割后生成数据样本,每个数据样本的数据点数为 2048。经过特征提取后,生成的前向特征矩阵和后向特征矩阵的尺寸均是 32×64。归一化位置编码的尺寸为 32×64。需要两个 Transformer 神经网络分别处理前向特征数据和后向特征数据,每个 Transformer 神经网络的层数为 12。Transformer 神经网络内部多头注意力机制的头数为 8。Transformer 神经网络内部的 3 个一维卷积的维度分别为 128、192 和 64,核尺寸均为 1。仿真实验所用硬件环境为 NVIDIA GeForce GTX 1050 Ti GPU 和 Intel(R) Core(TM) i7-8750H CPU @2.20 GHz。

3.2 模型检测仿真

本文算法在经过 80 000 次训练迭代后,模型训练约耗时 47.7 min。本文模型在训练过程中使用的训练迭代次数为预先设置的固定值80 000,此值根据实际经验获取,能够保证模型识别结果在经过这个迭代次数后完全收敛。为了能够有效降低模型的训练时间,可进一步开展对模型收敛性的研究,可促使模型达到收敛条件后及时停止模型训练,达到降低模型训练时间的目的,保存训练过程中验证准确度最大的模型作为最优模型。再使用轴承测试数据集测试所提算法的滚动轴承故障类型识别性能。

滚动轴承测试数据集的故障类型识别结果如表 2所示,表内的数字表示故障类型真实值与模型预测值间的数量统计。表中所示模型对于健康状态的识别准确度最高,为 0.9851,对于外圈损伤故障状态和内圈损伤故障状态的识别准确度分别为 0.8126和 0.8088。其中,内圈损伤故障误判为外圈损伤故障的样本数较多,为 999 个样本,外圈损伤故障误判为内圈损伤故障的样本也达到 257 个,分析原因是内圈损伤故障和外圈损伤故障的样本存在一定的相似性。综合考虑,模型对测试样本的综合故障类型识别结果的准确度为 0.8600。

表 2 模型识别结果

识别值 \ 真实值	健康	外圈故障	内圈故障
健康/个	4161	628	92
外圈损伤故障/个	36	3838	999
内圈损伤故障/个	27	257	4614
识别准确度	0.9851	0.8126	0.8088

3.3 对比和分析

本文所提滚动轴承故障类型识别方法主要包含两个部分,分别是故障类型识别模型的构建和特征矩阵的构建,下面将分别针对这两部分进行一系列的对比仿真来验证所提方法的有效性。

3.3.1 深度学习模型对比分析

为了验证所提基于 Transformer 神经网络的滚动轴承故障类型识别方法的有效性,本节使用一系

列的深度学习模型进行对比仿真，分别从模型准确度和模型复杂度进行分析。对比结果如表 3 所示。对比模型分别为 CNN 模型、LSTM 模型和 CNN + LSTM 模型。对比模型的参数说明如下。

CNN 模型使用 3 个卷积层、3 个池化层和 1 个全连接层，3 个卷积层的卷积核尺寸均为 5 × 5，卷积核的数量分别为 50、100 和 150，池化层的窗口大小均为[1, 2, 2, 1]，全连接层具有 128 个神经元。LSTM 模型由 2 个 LSTM 层组成，每层的 LSTM 神经元数量是 128。CNN + LSTM 模型由 3 个卷积层、3 个池化层和 4 个 LSTM 层组成，每层的参数配置分别与 CNN 模型和 LSTM 模型对应层的配置参数一致。以上对比模型在设置参数时考虑的因素是每个模型的总变量数与本文所提模型的变量数一致。所有对比模型使用的数据集均为本文所使用的滚动轴承前向特征矩阵。

表 3　不同对比模型的滚动轴承故障类型识别结果

故障类型识别模型	CNN 模型	LSTM 模型	CNN + LSTM 模型	本文模型
识别准确度	0.7698	0.7685	0.7706	0.8600
模型总变量数	1 116 465	1 049 859	1 011 889	935 427
单测试样本的平均处理时间/ms	0.264604	16.983734	0.932315	0.964127

由表 3 可以看出，本文所提模型与表中所列其他深度学习模型相比在识别准确度上均有较大程度的提高。其中，与 CNN 模型相比，识别准确度可提高 11.7%。由于 CNN 对于输入时间序列的远距离特征捕获能力较差，作为对比算法在参数设置时并没有设置较深的网络层数，因此其识别效果不及本文所提方法。本文模型与 LSTM 模型相比，识别准确度可提高 11.9%，原因是 LSTM 模型不存在注意力机制，不能区别不同样本对结果的贡献程度。同时，与 CNN + LSTM 模型相比，本文模型识别准确度可提高 11.6%。可见本文模型在滚动轴承故障类型识别上的有效性良好。

本文模型在保证较高准确度的同时，也保持着较高的时间效率，表 3 中所列模型中，CNN 模型、CNN + LSTM 模型和本文模型的单个样本的平均处理时间均在 1 ms 以下，其中 CNN 模型由于其卷积化结构使其时间效率最佳。而 LSTM 模型由于具有循环结构，内部神经元的处理为顺序处理，导致其时间效率很低。

3.3.2　特征提取方法对比分析

本文所用小波包变换 + 快速傅里叶变换构成的特征矩阵供滚动轴承故障类型识别模型处理，不同的特征提取方法对结果有不同的影响。本节给出了使用本文提出的神经网络模型再结合不同的特征矩阵时的轴承故障类型识别结果与其他特征提取方法的对比。对比结果如图 4 所示。其中不同的特征提取方法分别为小波包变换、小波变换、快速傅里叶变换、原始数据和本文使用的小波包变换 + 快速傅里叶变换。

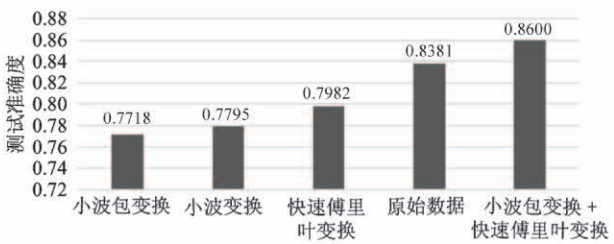

图 4　不同特征提取方法时的故障类型识别结果

从图 4 可以看出，本文所用小波包变换 + 快速傅里叶变换构建的特征矩阵得到较好的故障类型识别性能，证明本文使用的特征提取方法适用于本文所提模型的滚动轴承故障类型识别。使用原始数据得到的故障类型识别准确度为 0.8381，性能仅次于本文所用特征组合。而当分别使用单独的小波包变换得到的特征、小波变换得到的特征和快速傅里叶变换得到的特征时，测试准确度均在 0.8 以下，均不能取得较理想的效果。

3.3.3　特征提取尺寸对比分析

本文所提基于 Transformer 神经网络的滚动轴

承故障类型识别模型所用输入前向特征矩阵和后向特征矩阵的尺寸均受滚动轴承特征总数量的限制,不同的特征矩阵(包括前向特征矩阵和后向特征矩阵)的尺寸对滚动轴承故障类型识别的性能具有不同的影响,本节对比了采用不同的特征矩阵尺寸时的滚动轴承故障识别结果,对比结果如图 5 所示。其中所用特征矩阵尺寸分别为 8 × 256、16 × 128 和 32 × 64。

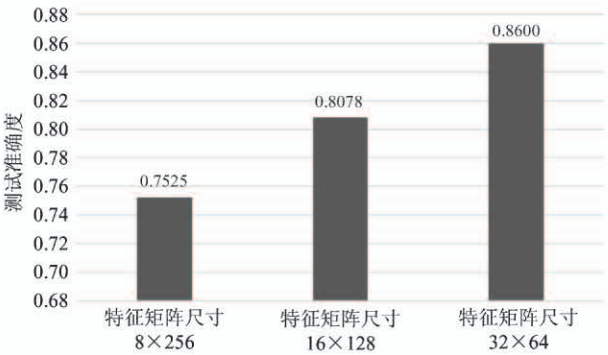

图 5　使用不同特征提取尺寸时得到的滚动轴承故障类型识别结果

图 5 表明,采用不同的特征矩阵,滚动轴承故障类型识别结果存在明显差别,特征矩阵尺寸为 8 × 256 时,得到的故障类型识别准确度为0.7525;当特征矩阵尺寸为 16 × 128 时,得到的故障类型识别准确度是 0.8078,识别性能明显提升。当采用本文所用的特征矩阵尺寸 32 × 64 时,故障类型识别准确度是 0.8600。在实际应用中需要综合考虑不同特征矩阵尺寸下的故障类型识别性能,选择最合适的特征矩阵尺寸。

3.3.4　位置编码方法对比分析

为了使 Transformer 神经网络在处理滚动轴承特征数据时包含数据的位置信息,本文提出了一种应用于轴承特征数据的归一化位置编码方法。

为了验证所提位置编码方法的有效性,进行如下对比实验,分别使用正余弦位置编码、固定位置编码和 2D 位置编码及本文所提故障类型识别模型完成故障类型识别,测试结果见图 6。

如图 6 所示,使用正余弦位置编码时,测试准确度为 0.8350。固定位置编码使用 One-hot 表示,测试准确度为0.8393。2D位置编码方法性能较差,

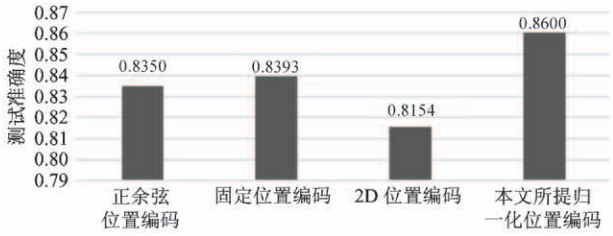

图 6　使用不同位置编码时得到的滚动轴承故障类型识别结果

测试准确度只能达到 0.8154。本文提出的归一化位置编码方法效果最好,测试准确度达到 0.8600。需要说明的是,当应用于滚动轴承的故障类型检测时,不同的位置编码方法获得不同的结果,很难解释哪种位置编码方法最好。因此,只能使用最合适的位置编码方法。

4　结　论

本文针对实际应用中的滚动轴承故障类型识别要求同时具有较高的识别准确度和时间效率的问题,提出了基于 Transformer 神经网络的滚动轴承故障类型识别方法,所提方法使用前向特征矩阵和后向特征矩阵完成数据的双向处理,同时所设计的归一化位置编码方法使 Transformer 神经网络在处理时考虑到特征矩阵的位置信息,模型训练时的错误样本权重增强能够进一步增加模型的鲁棒性。为了验证所提方法的性能,使用 KAt 数据中心的滚动轴承数据集对模型进行验证,整体识别准确度能够达到 0.86。与现有常用深度学习方法相比,本文所提方法在时间效率上和准确度性能上均有一定的优势,其中,准确度能够提升 11% 以上,单个样本的平均处理时间小于 1 ms。

本文的研究结果可以直接应用于工程实践,能够同时满足准确度和时间效率的最低要求。但是,本文所提滚动轴承故障类型识别方法在效果上依然存在一些问题,比如所提方法不能满足某些准确度要求更高的应用中,因此,接下来的工作是进一步提高模型的准确度,满足更高准确度的需求,同时保证算法的时间效率。此外,也将开展对模型训练过程中收敛性的研究,促使模型达到收敛条件后,及时停止模型的迭代,降低模型训练时间。

参考文献

[1] Hou K, Zhang Y C, Shi J L, et al. Motion planning based on artificial potential field for unmanned tractor in farm land[C]//International Conference on Applied Human Factors and Ergonomics, Orlando, USA, 2018:153-162

[2] Zhang X Y, Liang Y T, Zhou J Z, et al. A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM [J]. *Measurement*, 2015, 69:164-179

[3] Ziani R, Felkaoui A, Zegadi R. Bearing fault diagnosis using multiclass support vector machines with binary particle swarm optimization and regularized Fisher's criterion[J]. *Journal of Intelligent Manufacturing*, 2017, 28(2):405-417

[4] Li X, Yang Y, Pan H Y, et al. A novel deep stacking least squares support vector machine for rolling bearing fault diagnosis[J]. *Computers in Industry*, 2019, 110: 36-47

[5] Zhou Y Q, Tian L, Liu L, et al. Fog computing enabled future mobile communication networks: a convergence of communication and computing[J]. *IEEE Communication Magazine*, 2019, 57(5): 20-27

[6] Liu L, Zhou Y Q, Yuan J H, et al. Economically optimal MS association for multimedia content delivery in cache-enabled heterogeneous cloud radio access networks[J]. *IEEE Journal on Selected Areas in Communications*, 2019, 37(7): 1584-1593

[7] Liu L, Zhou Y Q, V. Garcia, et al. Load aware joint CoMP clustering and inter-cell resource scheduling in heterogeneous ultra dense cellular networks[J]. *IEEE Transactions on Vehicular Technology*, 2018, 67(3): 2741-2755

[8] Ali J B, Fnaiech N, Saidi L, et al. Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals[J]. *Applied Acoustics*, 2015, 89(3):16-27

[9] Lu W N, Wang X Q, Yang C C, et al. A novel feature extraction method using deep neural network for rolling bearing fault diagnosis[C]//Proceedings of the Chinese Control and Decision Conference, Qingdao, China, 2015:2427-2431

[10] Guo X J, Chen L, Shen C Q. Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis[J]. *Measurement*, 2016, 93:490-502

[11] Li S B, Liu G K, Tang X H, et al. An ensemble deep convolutional neural network model with improved D-S evidence fusion for bearing fault diagnosis[J]. *Sensors*, 2017, 17(8):1729

[12] Ding X X, He Q B. Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis[J]. *IEEE Transactions on Instrumentation and Measurement*, 2017, 66(8):1926-1935

[13] Li X, Zhang W, Ding Q, et al. Multi-layer domain adaptation method for rolling bearing fault diagnosis[J]. *Signal Processing*, 2019, 157:180-197

[14] Zhao H T, Sun S Y, Jin B. Sequential fault diagnosis based on LSTM neural network[J]. *IEEE Access*, 2018: 12929-12939

[15] Zhao R, Yan R Q, Wang J J, et al. Learning to monitor machine health with convolutional bi-directional LSTM networks[J]. *Sensors*, 2017,17:273

[16] Gokhale M Y, Khanduja D K. Time domain signal analysis using wavelet packet decomposition approach[J]. *International Journal of Communications Network and System Sciences*, 2010, 3(3):321-329

[17] 马学娟. 基于快速傅里叶变换(FFT)和小波变换的大型风机机械振动故障的分析[J]. 科技与创新,2016(11):121,125

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *arXiv*:1706.03762, 2017

[19] He K M, Zhang X Y, Ren A Q, et al. Deep residual learning for image recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016:770-778

[20] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. *arXiv*:1607.06450, 2016

[21] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[J]. *arXiv*:1705.03122v2, 2017

[22] Wang Z L, Liu J C. Translating mathematical formula images to LaTeX sequences using deep neural networks with sequence-level training[J]. *arXiv*:1908.11415, 2019

[23] Kingma D P, Ba J. Adam: a method for stochastic optimization[J]. *arXiv*:1412.6980, 2014

[24] Lessmeier C, Kimotho J K, Zimmer D, et al. Condition

monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: a benchmark data set for data-driven classification[C]// Proceedings of the European Conference of the Prognostics and Health Management Society, Bilbao, Spain, 2016: 1-17

A novel fault type detection method of rolling bearing using Transformer neural networks

Qiu Dawei[*,**], Liu Zichen[*], Zhou Yiqing[*,**], Long Long[*,**], Tan Wenwen[***], Cao Huan[*,**]

([*] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**] University of Chinese Academy of Sciences, Beijing 100049)

([***] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029)

Abstract

Rolling bearing fault type detection under practical applications requires both high detection accuracy and high time efficiency. Based on the above requirements, a rolling bearing fault type detection method using Transformer neural network is proposed. The proposed method combines wavelet packet transform time-frequency domain energy features and fast Fourier transform frequency domain features to generate input feature matrices, and solve the input problem of the Transformer neural network. At the same time, a normalization positional encoding method applied to rolling bearing fault type detection is proposed to solve the position coding problems of Transformer neural network in the field of rolling bearing fault analysis. A method for processing the input feature matrix from the bidirectional by the Transformer neural network and an error sample weight enhancement mechanism during the model training process are proposed to improve the robustness of the proposed method. The detection performance of the proposed method is verified using the rolling bearing dataset of the KAt data center. Compared with existing deep learning methods, the proposed method has certain advantages in both time efficiency and accuracy performance. The accuracy can be improved by more than 11%, and the average processing time of each sample is less than 1 ms.

Key words: rolling bearing, fault type detection, Transformer neural network, forward feature matrix, backwardfeature matrix, normalization positional encoding, weight enhancement