时序数据多尺度缺失值填充方法

王智杰[1]，陈超[2]，陈东月[1]，骆天逸[1]，贺淳禹[1]，胡清华[1]，李东[1]
1. 天津大学人工智能学院，天津 300350；
2. 中国汽车技术研究中心有限公司，天津 300300

摘要
在工业物联网与智能监测技术快速发展的背景下，设备健康管理中的故障诊断、剩余寿命预测及微弱故障早期检测等任务，高度依赖传感器采集的时序数据。然而，实际应用中数据缺失问题普遍存在，严重影响着后续分析任务的有效性。现有缺失值填充方法往往侧重恢复数据整体趋势，却忽视了对细节特征的精准重构，导致填充后的数据难以满足复杂故障诊断的需求。为此，本文提出一种基于小波变换的缺失值填充（wavelet transform-based missing value imputation，WTMI）框架。该框架通过小波分解获取数据的多尺度表征，结合深度自动编码器实现分层重构，最终经小波逆变换完成缺失值预测。进一步地，为解决尺度间相关性建模与最优尺度选择问题，提出多尺度缺失值填充（multi-scale missing value imputation，MSMI）方法，其利用深度堆叠网络架构整合跨尺度信息，通过瓶颈层特征与低尺度数据的融合策略，显著提升填充精度。在化工过程仿真、动力设备运行及空气质量监控等多源实测数据集上的实验表明，相较于现有主流方法，本文提出的方法在保持填充精度相当的同时，将故障诊断准确率提升了2%，尤其对微弱故障的诊断精度提升了达5%，验证了多尺度分析在时序数据缺失值处理中的有效性。

关键词
缺失值；多尺度分析；小波变换；传感器数据；故障诊断

中图分类号：TP183　　文献标志码：A　　doi：10.11959/j.issn.2096-0271.2026041

Time series data multi-scale missing value imputation

Wang Zhijie[1], Chen Chao[2], Chen Dongyue[1], Luo Tianyi[1], He Chunyu[1], Hu Qinghua[1], Li Dong[1]
1. School of Intelligence and Computing, Tianjin University, Tianjin 300350, China
2. China Automotive Technology and Research Center, Tianjin 300300, China

Abstract
Due to the widely available missing data in data, missing value imputation has become an extremely important issue in data processing. Equipment health management, such as fault diagnosis, life prediction, and early detection of weak faults, relies on sensor data, which can seriously affect the analysis results of such methods when missing values are present in the sensor data. Most existing missing value imputation methods focus on overall trends while ignoring data details, but the details often reflect the true situation of the device, especially in weak fault detection. To address these issues, a wavelet transform-based missing value imputation method (WTMI) is proposed. The method first decomposes the data by wavelet to obtain multi-scale data, then reconstructs the multi-scale data separately using an auto-encoder, and finally obtains the final prediction value by using the inverse wavelet transformation. Considering the correlation between different scales and the problem of scale selection, we further proposed a multi-scale missing value imputation

(MSMI). MSMI uses a stacked neural network to link data at different scales, so that the information at different scales can be fully utilized and more accurate predictions can be obtained. The experimental results of imputation and fault diagnosis on artificial and practical datasets show that the imputation accuracy of this method is similar to existing methods, and it improves the fault diagnosis accuracy by 2% compared with existing filling methods, especially by 5% for weak faults.

Key words
missing value, multi-scale analysis, wavelet transform, sensor data, fault diagnosis

0 引言

数据作为数字时代的基石，在设备健康管理领域扮演着关键角色。故障诊断、剩余寿命预测及微弱故障早期检测等核心任务均高度依赖传感器时序数据的完整性与准确性。然而，受传感器失效、通信链路中断等因素影响，数据缺失在工业监测场景中普遍存在[1-3]。现有缺失值填充方法可归纳为两类：基于统计理论的传统方法（如多重插补、K 最近邻算法）与基于机器学习的智能方法（如深度自动编码器、循环神经网络）。尽管这些方法在一定程度上实现了缺失值恢复，但普遍存在“重整体趋势、轻细节特征”的局限性，导致填充后的数据丢失了对故障诊断至关重要的细微特征，难以满足复杂工业场景的分析需求。

目前常见的缺失值填充方法主要包括基于统计、基于机器学习和多重插补方法[4-5]。基于统计的填充方法利用数据的统计参数来填补缺失值[6]，虽然这些方法通常不会改变数据的分布，但可能引入较大的误差[7]。分层均值填充是该类方法的改进，它先对数据进行聚类，生成数据层，然后利用每层的均值来填补属于该层的样本的缺失值[8]。基于机器学习的填充方法通常将包含缺失值的特征视为决策特征，而将其他完整特征视为变量特征，学习器学习两者之间的关系，以便对缺失值进行预测。常见的方法包括KNN、EM、NN、DAE、RNN 等[9]。多重插补方法通过构建多个不同的填充数据集，对每个缺失值生成多组合理估计值，再综合多组填充结果得到最终补全数据，能有效量化缺失值带来的不确定性，提升统计推断与模型结果的可靠性[10]。总体而言，传统统计方法通过数据分布假设或邻域相似性实现填充，却难以捕捉时序数据的非线性动态特性；基于深度学习的方法虽能学习数据隐含特征，但缺乏对多尺度信息的显式建模，忽略了不同时间尺度下数据特征的关联性。例如，主成分分析（PCA）及其衍生方法（如MSPCA）虽尝试通过多尺度分解提升填充效果，但因未充分整合跨尺度信息，仍存在细节重构精度不足的问题。此外，现有方法在处理长序列缺失或多传感器协同填充时，常因尺度选择的盲目性导致预测偏差。尽管现有的填充方法在一定程度上取得了良好的效果，但缺失值填充的主要目的是支持后续的数据分析工作，如设备故障诊断、早期微弱故障检测以及寿命预测等[11]，数据中的细节特征对于准确诊断和可靠预测至关重要。然而现有的传感器缺失值填充方法大多注重重构数据的整体趋势，导致填充后的数据丢失了对于后续数据分析任务而言最关键的特征[12]。

针对上述挑战，本文提出一种融合小

波变换与深度神经网络的多尺度缺失值填充框架。本文首先提出了基于小波变换的缺失值填充（wavelet transform-based missing value imputation，WTMI）框架，小波神经网络虽然能够在多个尺度上对数据进行分析[13]，但如果以重构整体数据为目标，模型应当兼顾重构数据的概貌特征和细节特征，因此在模型中需独立进行不同尺度的数据重构。WTMI框架首先利用小波分解分别对传感器数据进行预处理以获取数据在多个尺度上的特征，接下来使用深度自动编码器分别对不同尺度的特征进行重构，然后利用小波逆变换重构数据。目前多尺度分解和重构方法常存在以下3个问题：①不同尺度的数据相互关联，数据的细节特征往往与趋势特征密不可分，如果将两者单独处理，信息的丢失会造成泛化性能下降；②小波变换的层数难以确认，导致模型可扩展性较差；③根据下游任务场景，不同尺度数据的重要性不同，例如故障诊断中对于不同的故障，数据趋势与细节特征的重要性不同，模型需要高度关注对故障诊断更加有效的尺度数据。为解决以上问题并进一步提高模型在各类任务场景上的性能，基于WTMI框架，本文进一步提出一种多尺度缺失值填充（multi-scale missing value imputation，MSMI）方法，尝试在多个尺度上对数据进行重构，以获得对数据分析更有效的填充值。MSMI方法在小波变换时选择了更多尺度的数据，一方面能够增加模型输入的数据量，另一方面不需要再提前设定分解层数。另外，MSMI方法参考了深度堆叠网络（deep stacking network，DSN）[13]，将更大尺度分量重构网络的瓶颈层特征与低尺度数据合并后作为低尺度重构网络的输入，从而在重构低尺度数据时，模型可以利用更大尺度的数据作为辅助，得到更准确的填充值。同时，通过将每个尺度重构网络的瓶颈层特征拼接后再输入故障诊断模型，MSMI方法很容易扩展为多尺度故障诊断方法。

该方法的技术创新点如下：

- 通过小波变换将时序数据分解为不同频率的分量，结合深度自动编码器实现分层重构，解决传统方法对细节特征保留不足的问题；
- 设计跨尺度信息融合机制，利用堆叠网络整合多尺度特征，提升长序列缺失场景下的填充鲁棒性；
- 在重构过程中引入故障诊断任务导向的正则化约束，实现填充精度与诊断性能的协同优化。

1 先验知识

1.1 小波变换

处理信号时，需要对信号的成分进行分析。傅里叶变换把信号利用基函数的线性组合表示出来，进而分析信号中的各个成分。短时傅里叶变换通过加窗来获取信号的时频变换。小波变换使用小波基函数，具备局部分析、非平稳信号处理的能力[14]，通过伸缩和平移等运算可对函数或信号进行多尺度的细化分析。离散小波变换的定义如式（1）所示。

$$\mathrm{DWT}(a,\tau)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty}f(t)*\varphi(\frac{t-\tau}{a})\mathrm{d}t \quad (1)$$

其中，尺度a控制小波函数的伸缩，平移量τ控制小波函数的平移。

1.2 基于深度自动编码机的缺失值填充

基于机器学习的缺失值填充方法将完

整值和缺失值分别视为特征变量和决策变量，并构建模型学习两者间的关联，其目标函数的一般形式为：

$$\underset{w,b}{\operatorname{argmin}}\ L\left(h\left(X_{\mathrm{c}},w,b\right)-X_{\mathrm{m}}\right) \tag{2}$$

其中，X_{c}和X_{m}分别表示数据的完整部分和缺失部分的变量，w和b为模型的可学习参数，$L(\cdot)$表示损失函数。

在实际应用中，任何传感器在任何时候都可能出现缺失值，对缺失值进行填充时应考虑所有可能性。如果系统中有n个传感器，那么缺失值存在2^n种情况，如果针对每一种情况设计一个填充模型，即需要训练2^n个模型才能满足需求，这种方法显然不合理。针对该问题人们提出了多种解决方案，如Huseyin等[15]将局部化添加到模型中，因此可以在单个统计框架中处理各种类型的缺失值。自编码器（autoencoder，AE）通过恢复所有信号来对缺失值进行填充[16-17]。AE是由编码层和解码层组成的多层神经网络，编码层和解码层通过瓶颈层相连，瓶颈层关键特征需通过解码层将原始数据完全恢复出来，因此将其称为原始数据的嵌入特征[18]。当传感器时序数据中存在缺失值时，自编码器可以利用其重构特性对缺失值进行填充[19]。另外，自编码器最小化重构误差的同时重建所有传感器数据，从而通过一次模型训练处理不同的数据缺失类型。Jia等[20]提出结合数据中的时序信息可以进一步提升深度自编码器（deep autoencoder，DAE）的预测精度。基于自编码器的缺失值填充方法的目标函数为：

$$\underset{w,b}{\operatorname{argmin}}\ \left\| h(X,w,b)-X \right\|^2 \tag{3}$$

其中，h是深度自编码器。$z^l=\{z_1^l,z_2^l,\cdots,z_n^l\}$为第$l$层神经元状态，$a^l=\{a_1^l,a_2^l,\cdots,a_n^l\}$为第$l$层神经元激活值，$\delta_i^l=\dfrac{\partial E}{\partial z_i^l}$为每层的传导误差，$f$为Sigmoid激活函数，$L$为DAE的最大层数，则式（3）优化过程中所需的计算包括以下几项：

$$\begin{aligned}
&\delta_i^L=-(X-a^L)\odot f'(z_i^L)\\
&\delta^l=(\textstyle\sum_{j=1}W^{l+1}\delta^{l+1})\odot f'(z^l)\\
&\frac{\partial E}{\partial W^l}=\delta^l(a^{l-1})^{\mathrm{T}}\\
&\frac{\partial E}{\partial b^l}=\delta^l
\end{aligned} \tag{4}$$

2 多尺度缺失值填充方法

2.1 WTMI框架

WTMI框架示意图如**图1**所示。WTMI框架利用小波分析将数据分解为多尺度数据[21]，并针对每个尺度的数据分别使用DAE进行重构，最后使用小波逆变换实现对缺失数据的填充。小波分解过程中的高频信息表示的是细节特征，而低频信息表示的是全局形态特征[22]，WTMI框架通过对两类分量的分别重构，实现同时兼顾数据细节和全局特征两个方面的能力。

在实际任务中难以确认哪类尺度信息或哪种细节对于目标来说更重要，因此模型存在选择分解层数的问题。本文在进行小波变换时会选择尽可能多的层数以避免该参数的选择。此外，在每一次进行小波变换的过程中，概貌信号只有一个，本文进行了多次小波分解，并将每一次的概貌信号都输入模型。为了使模型更加充分地利用数据中的信息，从而得到更准确的填充值，更细节尺度的数据也会被保留。例如，分解层数为3时，原始数据、概貌信号、2层分解与3层分解时的数据将作为不同尺度下的信息统一输入模型中进行训练。

设原始时间序列数据为$X\in R^{T\times D}$，其中

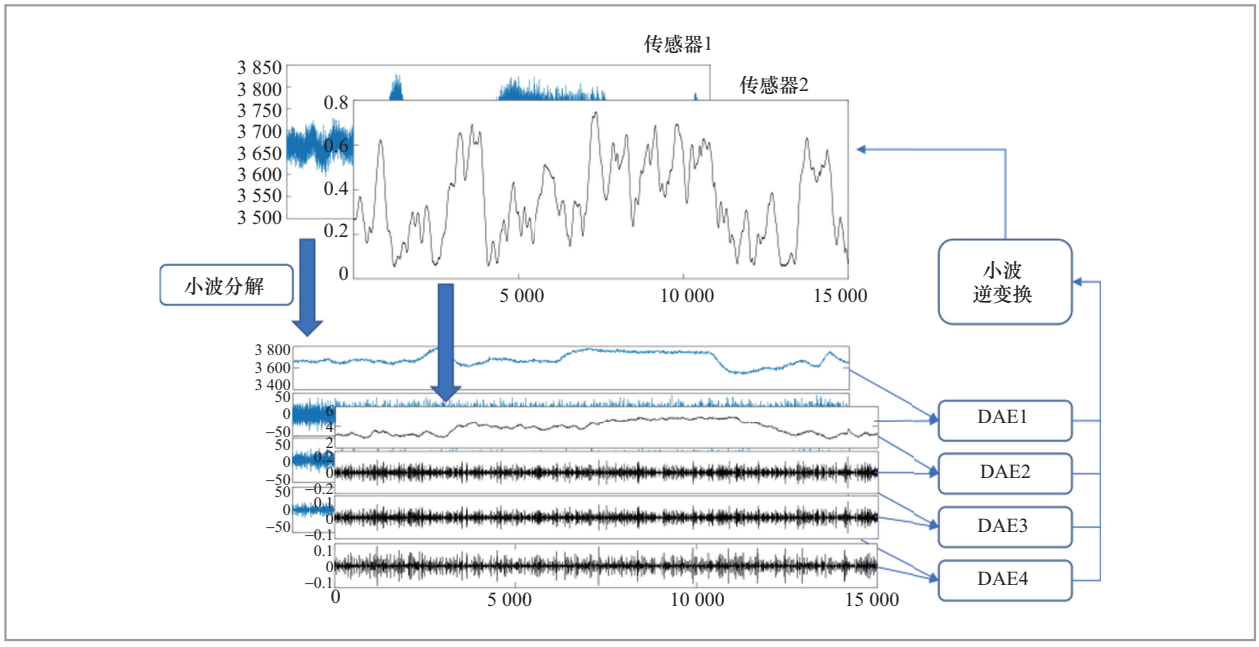


图1 WTMI框架示意图

T为时间步长，D为特征维度。定义掩码矩阵 $\boldsymbol{M}\in\{0,1\}^{T\times D}$，其中 $m_{t,d}=0$ 表示位置 (t,d) 缺失，$m_{t,d}=1$ 表示观测值存在。学习填充函数 $f(X\odot\boldsymbol{M},\boldsymbol{M})\to\hat{X}$，最小化重构误差 $L(\hat{X},X)$。

对每个特征维度 $d\in\{1,\cdots,D\}$ 独立执行 J 层离散小波变换（DWT）；分解形式如下：

$$\left\{A_d^{(J)},D_d^{(1)},D_d^{(2)},\cdots,D_d^{(J)}\right\}=\mathrm{DWT}(x_d,J) \quad (5)$$

其中，$x_d\in R^T$ 是第 d 维原始信号，$A_d^{(J)}\in R^{T/2^J}$ 为第 J 层概貌系数，$D_d^{(j)}\in R^{T/2^j}$ 是第 j 层细节系数。在小波分解过程中，为了减少参数的引入，尽可能选择多的分解层数作为输入，但过多的层数会导致计算量上升，本文将小波分解层数定为 $J=\mathrm{lb}T$。同时在小波基函数的选择上，使用 Haar/Daubechies 等紧支撑小波避免边界效应。在对数据做完小波分解后，重组所有尺度的系数为三维张量，如式（6）所示。

$$\mathcal{C}=[A^{(J)}\|D^{(J)}\|D^{(J-1)}\|\cdots\|D^{(1)}]\in R^{(J+1)\times K\times D} \quad (6)$$

其中，$K=\max\left(\frac{T}{2^J},\frac{T}{2^{J-1}},\cdots,T\right)$ 为各尺度系数的最大序列长度。对每个尺度 $s\in\{0,1,\cdots,J\}$，$s=0$ 时对应概貌层 $A^{(J)}$，并行执行GRU-AE填充，计算式为：

$$\hat{\mathcal{C}}_s=\mathrm{GRU-AE}(\mathcal{C}_s,\boldsymbol{M}_s) \quad (7)$$

其中，$\mathcal{C}_s\in R^{K\times D}$ 是第 s 尺度系数矩阵，$\boldsymbol{M}_s$ 是下采样后的掩码矩阵，GRU－AE编码、解码结构为：

$$\begin{cases} h_s^{\mathrm{enc}}=\mathrm{GRU}_{\mathrm{enc}}\left(\mathcal{C}_s\odot\boldsymbol{M}_s\right),\text{编码器} \\ \hat{\mathcal{C}}_s=\mathrm{GRU}_{\mathrm{dec}}\left(h_s^{\mathrm{enc}}\right),\text{解码器} \end{cases} \quad (8)$$

对填充后的系数执行逆变换（IDWT）；重构原始信号：

$$\hat{x}_d=\mathrm{IDWT}(\hat{A}_d^{(J)},\hat{D}_d^{(J)},\cdots,\hat{D}_d^{(1)}) \quad (9)$$

优化目标包含两个部分，首先是尺度系数重建损失定义为：

$$\mathcal{L}_{\mathrm{scale}}=\sum_{s=0}^{J}\lambda_s\left\|\hat{\mathcal{C}}_s-\mathcal{C}_s^{\mathrm{true}}\right\|_F^2 \quad (10)$$

其中，平衡权值 λ_s 按尺度重要性加权，例如 $\lambda_0 > \lambda_1 > \cdots > \lambda_J$ 强调概貌层。其次是原始信号重建损失定义为：

$$\mathcal{L}_{\text{orig}} = \left\| (\hat{\boldsymbol{X}} - \boldsymbol{X}) \odot \boldsymbol{M} \right\|_F^2 \tag{11}$$

总损失函数定义为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{orig}} + \gamma \mathcal{L}_{\text{scale}} \tag{12}$$

2.2 MSMI 方法

多尺度重构过程中整体和分量间数据及各个层次间数据包含隐式信息和关联关系，而目前层次间数据缺乏交互，并且在整体和分量数据的重构相互独立的前提下，引入的整体重构误差可能导致信息丢失。此外，当模型与后续故障诊断、寿命预测等任务结合时，面对不同应用场景填充的同样的缺失值无法很好地适应数据分析的需求。基于这些原因，本文提出了 MSMI 方法。在 DAE 中，由于不同隐藏层的数据是对数据不同尺度的描述，深度堆叠网络将之前层的输出拼接后输入隐含层，从而实现了不同尺度信息的相互关联[23]。WTMI 框架通过小波分解可以得到不同尺度的数据，因此，MSMI 方法可以利用不同尺度数据的信息进行集成学习，从而实现数据间的交互，以提高预测的准确性。

图 2 为 MSMI 的整体框架。MSMI 主要借鉴了 DSN 的思想，将不同尺度的数据重构过程堆叠起来。MSMI 首先将数据进行多次小波分解并获得不同尺度的分量数据。一般而言，更大尺度的数据对更小尺度的数据重构具有指导作用，而细节数据往往对大尺度的数据重构没有太多的辅助作用。因此，MSMI 将更大尺度的数据重构模型的中间层输出集成到小尺度重构模型的输入中，通过这种方式实现了利用大尺度的数据指导细节尺度数据重构的目的。与 DSN 不同的是，MSMI 通过堆叠 DAE 实现了不同尺度的数据重构。

在训练或测试的数据中可能存在缺失值，特别是在时序数据中会出现持续一段时间数据缺失的情况，这会导致小波分解本身不准确，进而影响模型最终的结果。如**图 3** 所示，MSMI 在数据输入模型之前进行了预处理，先通过线性插值的方法给出一个初步的预测，以减少分解时带来的误差。

同前述方法获得小波分解后的多尺度张量：

$$\mathcal{C} = [A^{(J)} \| D^{(J)} \| \cdots \| D^{(1)}] \in \mathbb{R}^{(J+1)\times K \times D} \tag{13}$$

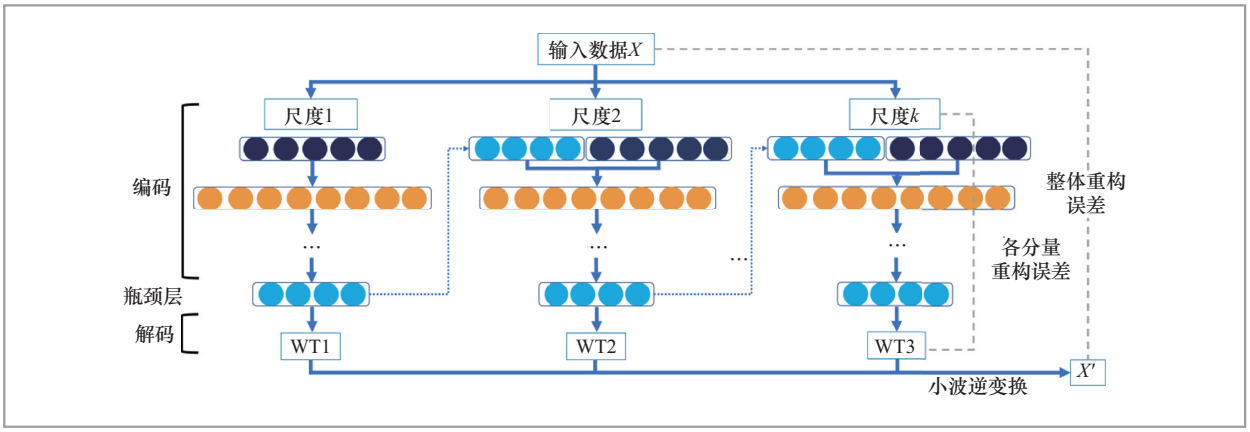


图 2 MSMI 整体框架

接下来进行堆叠式跨尺度编码器的构建，MSMI的目标为建立尺度间依赖关系，通过自顶向下由粗尺度到细尺度的堆叠GRU编码器来实现；过程如式（14）所示。

$$\begin{aligned} &h_0^{\text{enc}} = \text{GRU}_{\text{enc}}^0(\mathcal{C}_0), \text{顶层概貌系数处理} \\ &h_s^{\text{enc}} = \text{GRU}_{\text{enc}}^s(\mathcal{C}_s, \phi(h_{s-1}^{\text{enc}})), \forall s = 1, \cdots, J \\ &\phi(h) = \text{Upsample}(\boldsymbol{W}_\phi h), \text{跨尺度特征投影} \end{aligned} \tag{14}$$

其中，$\mathcal{C}_s$是第s尺度系数矩阵，$\phi: R^{d_{s-1}} \to R^{d_s}$是跨尺度特征映射函数，Upsample是双线性插值上采样匹配维度。粗尺度特征h_{s-1}^{enc}经上采样投影$\phi(\cdot)$作为细尺度GRU的初始状态，通过将低频信息的重构特征引入高频信号的重构过程中，使低频概貌信息指导高频细节恢复。接下来将粗尺度与细尺度重构网络的瓶颈层特征进行融合；过程如下：

$$\boldsymbol{H} = [h_0^{\text{enc}} \| h_1^{\text{enc}} \| \cdots \| h_J^{\text{enc}}] \in \mathbb{R}^{J \times d_h} \tag{15}$$

$$z = \frac{1}{J} \sum_{s=0}^{J} h_s^{\text{enc}} \tag{16}$$

$$g = \sigma(W_g z) \tag{17}$$

$$f = W_f(g \odot \boldsymbol{H}) \tag{18}$$

其中，σ为Sigmoid激活函数，$\odot$为特征级Hadamard乘积，W_g和W_f为可学习模型参数，$f \in \mathbb{R}^{d_f}$为压缩后的融合特征$(d_f \ll d_h)$。通过压缩多尺度特征为通道描述符，并学习尺度间重要性权重，可实现参数高效的特征重组。接下来利用融合特征指导各尺度重建，将融合特征注入并行GRU解码器，各尺度重构过程如下：

$$\begin{aligned} &\hat{\mathcal{C}}_s = \text{GRU}_{\text{dec}}^s(\psi_s(f)), \forall s = 1, \cdots, J \\ &\psi_s(f) = W_s^{\text{proj}} f + b_s \end{aligned} \tag{19}$$

此外，该方法通过构建跨尺度一致性约束，使$\hat{\mathcal{C}}_{s-1}$的逆变换经小波分解应重构$\hat{\mathcal{C}}_s$，强制相邻尺度重建结果满足小波变换

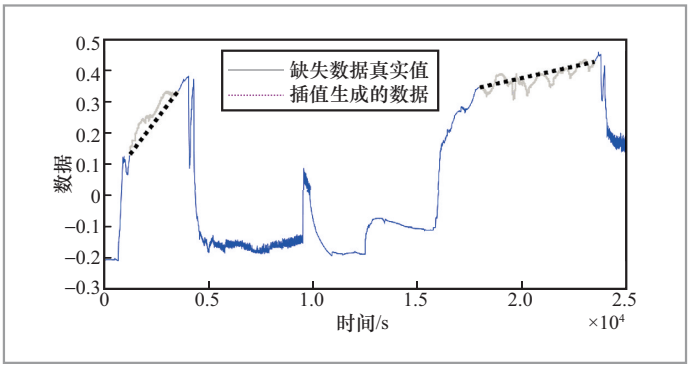


图3 输入数据插值示意图

的数学约束，解决传统方法中尺度间重建结果不兼容的问题，跨尺度一致性约束定义为：

$$\mathcal{L}_{\text{consist}} = \sum_{s=1}^{J} \left\| \mathcal{C}_s - \text{DWT}_s(\text{IDWT}(\hat{\mathcal{C}}_{s-1})) \right\|_F^2 \tag{20}$$

最后总的目标函数为：

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \underbrace{\| (\hat{X} - X) \odot \boldsymbol{M} \|_F^2}_{\mathcal{L}_{\text{orig}}} + \\ & \gamma \underbrace{\sum_{s=0}^{J} \lambda_s \| \hat{\mathcal{C}}_s - \mathcal{C}_s \|_F^2}_{\mathcal{L}_{\text{scale}}} + \delta \mathcal{L}_{\text{consist}} \end{aligned} \tag{21}$$

该方法通过显式建模尺度间依赖（堆叠编码器）、自适应特征融合（瓶颈层）、数学驱动约束（$\mathcal{L}_{\text{consist}}$），有效解决了多尺度方法中的信息孤立问题。

2.2.1 缺失值填充

基于MSMI框架对传感器缺失值进行填充的流程如下：将原始数据输入MSMI框架获取数据重构值，并将原始数据和重构数据在n个传感器上的数据依次进行k次采样，表示为$\{X_{i=1:n, j=1:k}\}$。由于实时系统中往往不会给出失效传感器或缺失值的位置，本文采用阈值判定的方式来确定哪些传感器信号是需要被替换的。阈值η的选取方式为将模型训练过程中的重构误差的3倍标准差作为判

定阈值，3倍标准差的位置可以保证挑选出99%以上的失效传感器，且不会用重构信号替换正常信号，最终获得填充后的信号X^r。缺失值填充过程如算法1所示。

算法1 MSMI缺失值填充算法

输入：将原始数据X输入已训练好的MSMI模型获取重构数据X'

输出：填充后的数据X^r

For $i=[1,2,\cdots,n]$

For $j=[1,2,\cdots,k]$

$$\bar{u}=\sum_{j=1}^{k}\left|x_{ij}-x'_{ij}\right|/k$$

$$\eta=3\sqrt{\frac{1}{k-1}\sum_{j=1}^{k}\left(\left|x_{ij}-x'_{ij}\right|-\bar{u}\right)^2}$$

IF $\left|X_{ij}-X'_{ij}\right|>\eta$ THEN

$$X^r_{ij}=X'_{ij}$$

ELSE

$$X^r_{ij}=X_{ij}$$

2.2.2 MSMI故障诊断任务

MSMI能够直接与任务进行融合，本文以故障诊断方法为例。如果在重构模型的每一个模型上加入分类任务的输出，可以得到一个多尺度的故障诊断分类器，且该模型可以同时实现缺失值填充。对应的联合优化目标函数为：

$$\underset{H}{\operatorname{argmin}}\quad L_2+\theta^*\|Y-T\|_2 \tag{22}$$

其中，Y是模型输出，T是目标类别值，模型通过θ在缺失值重构与故障诊断任务间进行平衡。在故障诊断任务中，Y由瓶颈层的输出通过Softmax网络层得到，即在同种解码层额外加上一个分类器以得到分类误差，并集成到模型反向传播过程中。

3 实验验证及分析

本节通过在人工和实际数据集上的实验结果来验证所提方法的有效性。使用的对比方法包括经典的缺失值填充方法，如KNN、NN（3层隐藏层网络）、MisGAN[24]、BRITS[25]、DAE、WTMI、本文所提的多尺度关联重构方法MSMI-1以及使用插值初始化后的MSMI-2。MSMI的隐含层网络神经元数量一般为输入参数数量的3倍左右；在处理故障诊断任务时，瓶颈层神经元数量一般使用PCA算法进行测试，本文实验中选取PCA信息保持率在99%以上时的维数作为瓶颈层神经元数量；堆叠网络结构不宜设置过深的网络层数，因此实验过程中网络隐藏层数量设置为5层。

本节所用的实验数据包括仿真数据集及实际传感器数据集。仿真数据集和实际传感器数据集主要用于验证缺失值填充的有效性，此外还有一个大型动力装置仿真故障数据集，用来验证在填充后数据上进行故障诊断的精度。数据集的详细情况见**表1**。在多个部件上安装传感器得到的数据，在实验中直接使用即可。实验中将不同地理位置同一时刻的数据拼接成一个样本。仿真数据中一般不存在缺失值，在实验过程中仿真数据集缺失率均设置为10%，数据为单点缺失。实际运行数据中一般已存在缺失值，例如，燃气轮机传感器数据缺失率约为3%；KDD CUP 2018数据集中PM2.5/NO_2/O_3/SO_2缺失2万条数据，CO缺失5万条数据，PM10缺失9万条数据；贵阳市空气日报数据中缺失率约为6%。模型训练过程中并不需要刻意

表1 数据集描述

数据集名称	数据集描述	样本数	特征数
TE过程仿真数据(TE)	利用TE化工过程的MATLAB仿真程序产生的数据,在稳定运行的情况下模拟了多种参数变化下的运行监测数据,其中控制参数有11个	18 000	41
燃气轮机运行传感器数据(GT)	某型燃气轮机实际运行数据	626 351	141
大型动力装置仿真数据(SPP-1)	通过仿真程序仿真了大型动力装置多个工况之间的切换,以及稳态运行的传感器数据	304 484	121
核电厂仿真传感器数据(SPP-2)	新型核电厂动力装置传感器仿真数据	18 452	58
大型动力装置实验台架仿真数据(SPP-3)	模拟了大型动力装置多种升工况的传感器数据	10 478	13
贵阳市空气日报数据(GZ)	提供贵阳市各地区每日空气质量监测数据,包含监测站点、主要污染物等信息	11 984	11
上海北京空气质量监测数据(SH-BJ)	数据集由北京和上海空气质量监测站的一年(2013年2月8日至2014年2月8日)空气质量数据组成[26]	236 863	8
KDD CUP 2018数据集(KDD Cup)	北京、伦敦各空气监测站点2017—2018年每天每小时的空气质量监测数据,以及当时附近的环境质量数据	310 000	8
大型动力装置仿真故障数据(SPP-F)	船舶动力装置的模拟传感器数据,包括67种故障数据	402 673	121

将缺失值挑选出来，经验证可知带有少量缺失值的训练数据能够使模型更鲁棒。

由于传感器数据的量纲并不统一，为了准确地对数据进行评价，实验过程中所有的数据都进行了归一化，并且在归一化过程中记录了相关参数，方便数据恢复到原始的量纲。在实验过程中笔者发现，MAE、RMSE及NRMSE等评价方法表现相同，因此，本文实验过程中使用的评价指标仅选择了RMSE。RMSE定义为

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum(h(x_i)-y_i)^2} \tag{23}$$

3.1 缺失值填充精度对比

表2展示了不同方法在缺失率固定为10%时的填充误差。从**表2**可知，WTMI、MSMI虽然在GT、SPP-3等数据上表现弱于NN和BRITS方法，但相比其他方法，整体上取得了更好的填充效果，且相比基础方法来说，显著降低了填充值的误差，这充分说明了MSMI的有效性。此外，在SPP-F数据集上，本文所提方法相比其他方法在效果上提升非常明显，这是因为SPP-F数据本身存在故障，常规方法无法对数据进行更好的重构，而在实际应用过程中系统故障时数据填充的重要性更加凸显，这反映出MSMI在实际应用过程中的优势。

从MSMI与WTMI的对比实验可以发现，不同尺度间的关联对于降低缺失值填充的误差起到了一定的作用。MSMI-1与MSMI-2的对比实验表明，使用插值初始化的输入数据对降低填充误差起到了一定作用，在整体上有一定的提升效果。随着连续缺失值出现的时长和比例增加，提升的效果会更为明显，这也反映了插值初始化的有效性。然而实际应用过程中此类情况占比较低，因此表中对比实验结果只采用了随机缺失情况下的填充误差。

WTMI、MSMI的主要作用是在恢复

表2 缺失率为10%时缺失值填充误差对比

数据算法	GT	TE	SPP-1	SPP-2	SPP-3	GZ	SH-BJ	KDD Cup	SPP-F
KNN	0.1378	0.0933	0.1629	0.1307	0.1954	0.1866	0.2204	0.3033	0.3487
NN	*0.0907*	0.0806	0.1598	0.1347	0.1592	0.1305	0.2895	0.2864	0.3801
DAE	0.1407	0.0706	0.1536	0.1234	0.1377	0.1433	0.2984	0.2607	0.3909
BRITS	0.1192	0.0663	0.1275	0.1294	*0.1005*	**0.1082**	0.1978	0.2427	0.3652
MisGAN	0.1035	0.0712	0.1309	0.1186	0.1246	0.1219	0.2059	0.2203	0.3729
WTMI	0.1147	0.0725	0.1264	0.1139	0.1168	0.1163	0.2174	0.236	0.2904
MSMI-1	0.1011	**0.0645**	**0.1235**	**0.1077**	0.1185	0.1154	**0.1947**	**0.2401**	**0.2806**
MSMI-2	**0.0925**	*0.0623*	*0.1104*	*0.1023*	**0.1029**	*0.1037*	*0.1822*	*0.2159*	*0.2655*

注：粗斜体为误差最低，粗体为次低误差。

数据概貌的同时关注恢复数据的细节。**图4**所示为不同填充方法下的缺失值填充误差。从**图4**可以看出，本文所提方法相比其他方法具有更稳定的预测值。这是因为本文所提方法关注了数据中的细节部分，从而使误差更集中。**图5**和**图6**所示分别为DAE和MSMI在大型动力装置仿真数据集上的缺失值填充效果对比，可以看出，本文提出的多尺度缺失值填充方法的预测值能够更好地拟合数据的细节特征，在数据的波动捕捉与局部特征刻画上表现更精准。相比之下，DAE仅能还原数据的整体变化趋势，对高频波动与细微结构的还原能力较弱，而本文MSMI在保持全局趋势的同时，实现了对数据局部细节与动态变化的有效建模，填充结果更贴近真实数据分布。

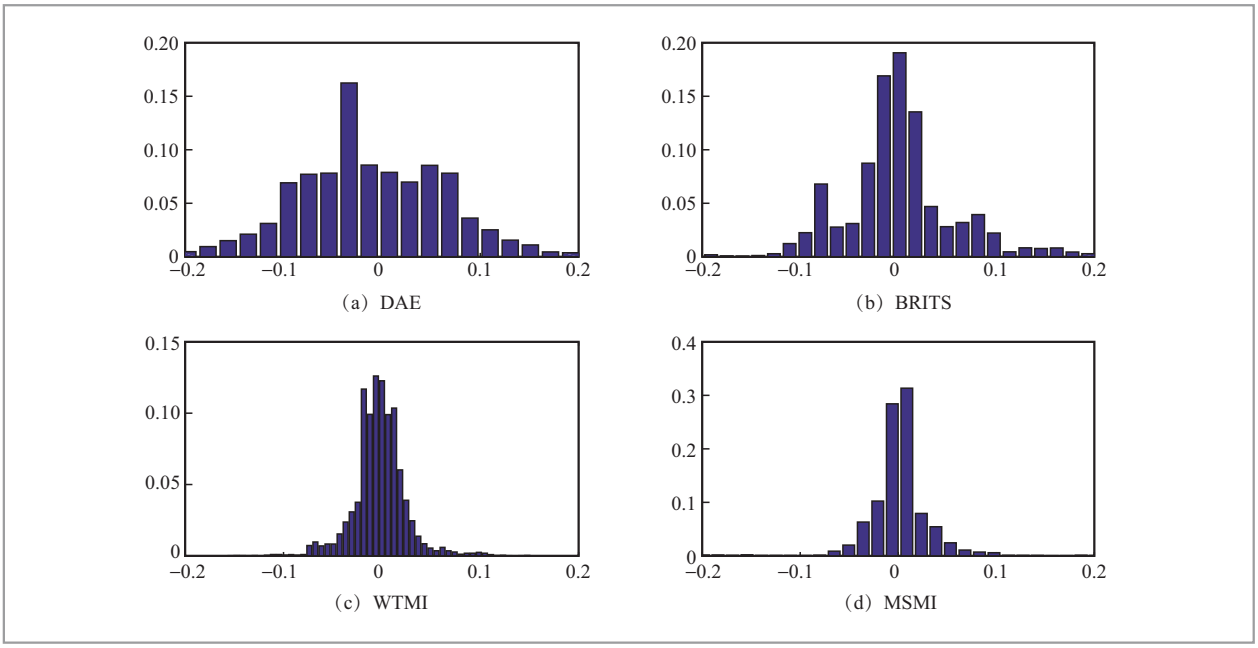


图4 仿真数据填充误差统计直方图对比

3.2 故障诊断精度对比

本节对比了不同缺失率下填充、不填充及多种填充方法的故障诊断精度。为了保证对比的有效性，实验过程中的故障诊断方法统一使用SVM，数据使用五折交叉进行验证，并通过参数寻优选取诊断准确率最优的模型进行测试。此外，传感器数据是时序数据，因此在实验过程中要切割成时间序列，然后在拼接的数据上进行实验。实验数据本身并不存在缺失值，可以直接得到无缺失值数据情况下的故障诊断精度作为基准评价，实验过程中模拟了10%的随机缺失，即数据缺失的时刻与传感器都是随机分布的。为了验证填充后的数据对故障诊断有提升作用，笔者也在不进行填充的设定下开展了故障诊断实验。最后实验对比了使用多种缺失值填充方法在数据上的故障诊断精度。

表3给出了故障诊断的结果。从结果可知，使用填充后的数据整体上对故障诊断有明显的提升作用，但一般难以达到原始数据的精度。填充值不准确会极大地影响故障诊断的精度，如SPP-F数据在使用KNN填充后的数据上进行故障诊断，精度下降明显。使用MSMI填充后的数据获得了最好的故障诊断精度，且接近原始数据的诊断精度。相比一般方法，MSMI利用了多尺度信息对数据进行重构和故障诊断，因而可以获得更高的诊断精度，**表3**的结果也印证了这一结论。模型在TE过程数据上提升较为明显，这是因为在TE数据中的故障主要表现为性能下降（如反应器液面降低10%等典型故障），这从侧面反映出MSMI对于需要利用细节特征进行分析的任务来说具有较大意义。

为了更直观地体现填充值对故障诊断的影响，故障诊断实验将所有故障总结为4种典型情况，包括部件故障、性能下降、

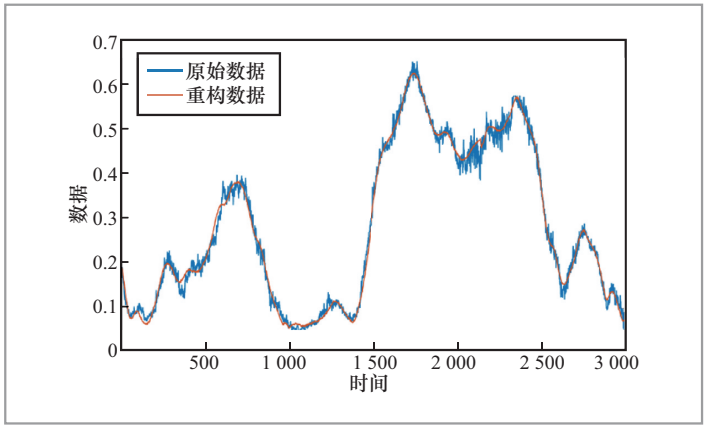


图5 DAE重构效果

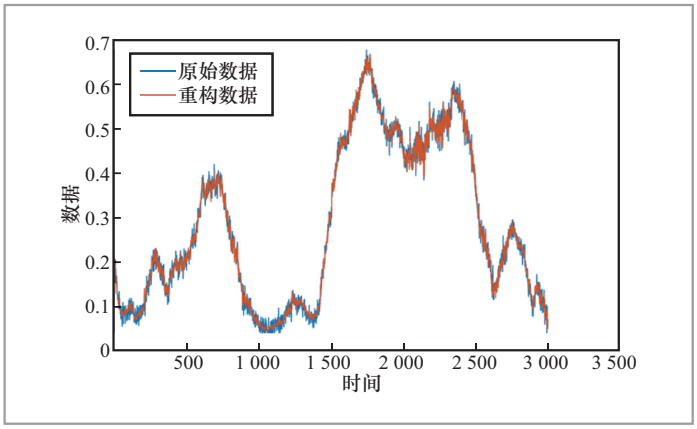


图6 MSMI重构效果图

表3 仿真故障数据故障诊断准确率

算法	SPP-F	TE
无缺失值	89.32%±2.96%	92.1%±3.35%
不填充	83.45%±1.98%	85.45%±2.97%
KNN	73.13%±3.12%	86.91%±2.42%
NN	85.11%±2.09%	86.34%±2.19%
DAE	84.7%±2.39%	89.15%±2.78%
BRITS	86.91%±1.72%	90.06%±3.19%
MisGAN	85.23%±2.84%	89.84%±2.83%
WTMI	84.96%±2.68%	90.54%±3.07%
MSMI-1	87.39%±1.66%	91.25%±2.64%
MSMI-2	**87.91%±1.48%**	**91.44%±3.26%**

微弱故障、混合故障，并分别进行分析。**图7**为在原始数据、不填充数据、DAE填充数据及MSMI填充数据上这4类故障的诊断精度对比。从**图7**可以看出，缺失值的存在会较大地影响故障诊断的精度，而填充缺失值则会抑制缺失值对故障诊断任务带来的影响。本文所提方法整体上比常规方法对故障诊断更有效。此外，对于性能下降、早期微弱故障的识别，MSMI能够得到与原始数据更接近的诊断结果，这说明了本文方法的有效性。

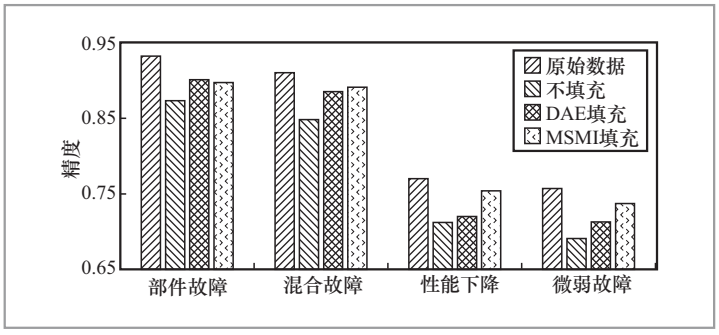


图7 仿真数据不同故障类型的诊断精度

3.3 参数影响

WTMI与MSMI主要涉及各分量重构效果与整体重构效果之间的平衡参数β，而在故障诊断任务中则增加了故障诊断效果与重构效果间的平衡参数θ，实验中分别针对这两种情况进行讨论。

图8分别展示了SPP-F数据在不同参数设置下的填充误差与故障诊断精度。**图8（a）**展示了在缺失值填充任务中β对缺失值填充误差的影响，当β较小时，其对整体的重构误差影响较小。随着β增大，模型更关注整体重构效果，忽视了细节特征和尺度间的关联，造成填充误差增大。**图8（b）**展示了故障诊断任务中θ参数对诊断精度及缺失值填充的影响，随着θ增大，模型会逐步向故障诊断任务倾斜，故障诊断准确率随之提升，但缺失值填充误差同步增大、填充精度相应降低。究其原因，缺失值填充精度下降带来的负面影响会与故障诊断性能提升的效果相互抵消，

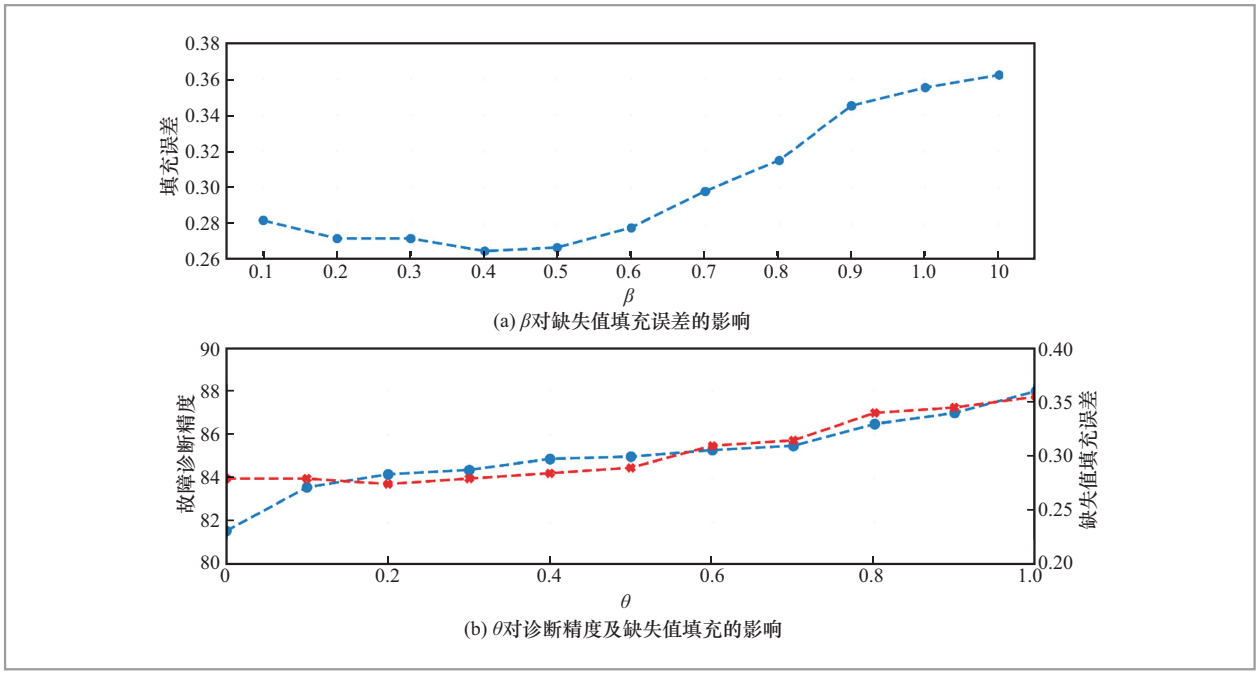


图8 参数对填充误差及故障诊断精度的影响

且前者的影响更大。当θ取值较小时，模型需同时兼顾高缺失值填充精度与高故障诊断精度，二者未产生明显冲突；随着θ持续增大，模型过度聚焦故障诊断任务，反而弱化了对缺失值填充的优化，最终导致填充精度持续下滑。

4 结束语

本文针对现有缺失值填充方法过度关注传感器数据概貌信号、却忽视细节特征的问题，提出了一种多尺度缺失值填充方法MSMI。在仿真数据集与实际数据集上的实验结果表明，MSMI 的缺失值填充效果较现有方法均有一定提升；得益于对细节特征的重点关注，该方法的缺失值预测结果也更加稳定。此外，故障诊断任务的实验结果显示，相较于其他方法，MSMI 在处理含缺失值数据时，对早期微弱故障识别、设备性能退化检测等任务具有明显优势。目前，本文仅围绕故障诊断相关任务对MSMI开展了初步探究。该方法可适配多类任务场景，还能灵活扩展为任务驱动型缺失值填充方法。针对多尺度信息对不同任务重要程度不一致的问题，可在缺失值填充过程中引入加权策略，重点强化某一尺度信息的作用。在不同任务场景下，方法性能与各类权重参数的影响关系还需进一步研究。

参考文献：

[1] Psychogyios K, Ilias L, Ntanos C, et al. Missing value imputation methods for electronic health records[J]. IEEE Access, 2023, 11: 21562-21574.

[2] Marlin B M, Zemel R S, Roweis S T, et al. Recommender systems: missing data and statistical model estimation[C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. New York: ACM, 2011: 2686-2691.

[3] Miller B N, Albert I, Lam S K, et al. MovieLens unplugged: experiences with an occasionally connected recommender system[C]//Proceedings of the 8th International Conference on Intelligent User Interfaces. New York: ACM, 2003: 263-266.

[4] de Brevern A G, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering[J]. BMC Bioinformatics, 2004, 5: 114.

[5] Andridge R R, Little R J A. A review of hot deck imputation for survey non-response[J]. International Statistical Review, 2010, 78(1): 40-64.

[6] Sullivan D, Andridge R. A hot deck imputation procedure for multiply imputing nonignorable missing data: the proxy pattern-mixture hot deck[J]. Computational Statistics & Data Analysis, 2015, 82: 173-185.

[7] Buck S F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1960, 22(2): 302-306.

[8] Little R J A. Survey nonresponse adjustments for estimates of means[J]. Revue Internationale de Statistique, 1986, 54(2): 139.

[9] 吴天舒, 尹宏鹏, 赵丹丹, 等. 基于迁移学习的零样本故障诊断[J]. 电子学报, 2023, 51(9): 2572-2577.
Wu T S, Yin H P, Zhao D D, et al. Zero sample fault diagnosis based on transfer learning[J]. Acta Electronica Sinica, 2023, 51(9): 2572-2577.

[10] Choudhury S J, Pal N R. Imputation of missing data with neural networks for classification[J]. Knowledge-Based Systems, 2019, 182: 104838.

[11] 张旭龙, 瞿晓阳, 谢骏飞, 等. 人工智能生成式内容技术概述[J]. 大数据, 2026, 12(1): 146-173.
Zhang X L, Qu X Y, Xie J F, et al. AIGC: a survey of AI-generated content technologies[J]. Big Data Research, 2026, 12(1): 146-173.

[12] 冯杨洋, 汪庆, 舒继武. 大模型时代下的存储系统挑战与技术发展[J]. 大数据, 2025, 11(1): 79-91.
Feng Y Y, Wang Q, Shu J W. Challenges and technical development of storage systems in the era of large language models[J]. Big Data Research, 2025, 11(1): 79-91.

[13] 杨彪, 熊赟, 傅玲, 等. 工业数字化转型: 故障诊断方法研究进展[J]. 大数据, 2024, 10(1): 110-126.
Yang B, Xiong Y, Fu L, et al. Industrial digital transformation: research onfault diagnosis methods[J]. Big Data Research, 2024, 10(1): 110-126.

[14] Mallat S G. A theory for multiresolution signal decomposition: the wavelet representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(7): 674-693.

[15] Ren L J, Wang T, Sekhari Seklouli A, et al. A review on missing values for main challenges and methods[J]. Information Systems, 2023, 119: 102268.

[16] Luo R F, Misra M, Himmelblau D M. Sensor fault detection *via* multiscale analysis and dynamic PCA[J]. Industrial & Engineering Chemistry Research, 1999, 38(4): 1489-1495.

[17] Luo R F, Misra M, Qin S J, et al. Sensor fault detection via multiscale analysis and nonparametric statistical inference[J]. Industrial & Engineering Chemistry Research, 1998, 37(3): 1024-1032.

[18] Nader S, Kjetil K P. Multi-scale deep network for fault detection: Ep3494284a1[P]. 2019.

[19] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3): 1-58.

[20] Misra M, Yue H H, Qin S J, et al. Multivariate process monitoring and fault diagnosis by multi-scale PCA[J]. Computers & Chemical Engineering, 2002, 26(9): 1281-1293.

[21] Wickerhauser M V. Adapted wavelet analysis from theory to software[M]. Piscataway: IEEE Press, 1994.

[22] Palanivinayagam A, Damaševičius R. Effective handling of missing values in datasets for classification using machine learning methods[J]. Information, 2023, 14(2): 92.

[23] Deng L, Yu D, Platt J. Scalable stacking and learning for building deep architectures[C]//Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2012: 2133-2136.

[24] Ozkan H, Pelvan O S, Kozat S S. Data imputation through the identification of local anomalies[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2381-2395.

[25] Samad M D, Abrar S, Diawara N. Missing value estimation using clustering and deep learning within multiple imputation framework[J]. Knowledge-Based Systems, 2022, 249: 108968.

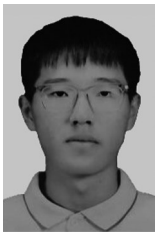
[26] Zheng Y, Liu F R, Hsieh H P. U-Air: when urban air quality inference meets big data[C]//Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013.

作者简介

王智杰（1996– ），男，天津大学人工智能学院博士生，主要研究方向为机器学习和图神经网络。

陈超（1989– ），男，中汽数据有限公司智能业务部智能技术室高级主管，主要研究方向为智能网联数据库建设与应用、功能安全和SOTIF研究。

陈东月（1993– ），女，博士，天津大学人工智能学院助理研究员，主要研究方向为故障诊断和图神经网络。



骆天逸（2006– ），男，天津大学人工智能学院本科生，主要研究方向为机器学习。

贺淳禹（2005– ），男，天津大学人工智能学院本科生，主要研究方向为机器学习。

胡清华（1976– ），男，博士，天津大学人工智能学院博士生导师，主要研究方向为机器学习和数据挖掘。

李东（1991– ），男，博士，天津大学人工智能学院助理研究员，主要研究方向为异常检测、小样本故障诊断。

收稿日期: 2025-11-04

基金项目: 中国博士后科学基金会与天津市联合资助项目(No.2024T016TJ)；国家自然科学基金资助项目(No.62406219，No.U23B2049)；中国博士后科学基金会与天津市联合资助项目(No.2023T014TJ)；中国博士后科学基金资助项目(No.2025M771494)

Foundation Items: China Postdoctoral Science Foundation-Tianjin Joint Support Program(2024T016TJ), The National Natural Science Foundation of China(No. 62406219, No. U23B2049), China Postdoctoral Science Foundation-Tianjin Joint Support Program(No. 2023T014TJ), China Postdoctoral Science Foundation (No.2025M771494)