

# 基于 Transformer 的知识继承式模糊神经网络的 故障诊断方法

李孟威, 卢伟<sup>†</sup>

(大连理工大学控制科学与工程学院, 辽宁 大连 116024)

**摘要:** 为了构建兼具高性能与内在可解释性的故障诊断模型, 本文提出了一种基于 Transformer 的知识继承式模糊神经网络 (TKI-FNN) 模型. 首先由训练好的先验 Transformer 分类器模块生成 logit 向量 (未归一化的概率分布). 随后, 知识蒸馏模块通过调节蒸馏温度参数, 将 logit 向量转换为携带类间相似性知识的软标签. 这些软标签不仅包含目标类别信息, 还隐含刻画不同故障类别之间的潜在关联关系. 最后, 知识继承的 Takagi-Sugeno-Kang (TSK) 模块通过其后件接收软标签, 并执行可解释的故障诊断推理. 该模型采用梯度下降方法优化由交叉熵、软标签正则化构成的复合目标函数. 这种全新的知识继承范式使得所提 TKI-FNN 在通过软标签正则化有效避免过拟合、提升故障诊断性能的同时, 仍能够保持下游 TSK 模块的语义可解释性不受破坏. 在涵盖真实工业过程的一系列实验中, 所提出模型在故障诊断精度与可解释性方面均表现出显著优势.

**关键词:** 故障诊断; 深度神经网络; Takagi-Sugeno-Kang; 知识蒸馏; 可解释性; 软标签

中图分类号: TP273 文献标志码: A

DOI: 10.13195/j.kzyjc.2026.0067

引用格式: 李孟威, 卢伟. 基于 Transformer 的知识继承式模糊神经网络的故障诊断方法 [J]. 控制与决策

## Transformer-based knowledge-inheriting fuzzy neural network for fault diagnosis

LI Meng-wei, LU Wei<sup>†</sup>

(College of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China)

**Abstract:** To develop a fault diagnosis model that combines high performance with intrinsic interpretability, this paper proposes a Transformer-based knowledge-inheritance fuzzy neural network (TKI-FNN). Specifically, a pre-trained Transformer classifier is first employed to generate logit vectors (i.e., unnormalized probability distributions). Subsequently, the knowledge distillation module transforms these logit vectors into soft labels by adjusting the temperature parameter, thereby embedding inter-class similarity information. These soft labels not only contain target class information but also implicitly characterize latent relationships among different fault categories. Finally, the knowledge-inherited Takagi-Sugeno-Kang (TSK) module receives soft labels through its consequent part and performs interpretable fault diagnosis inference. The model is optimized using gradient descent based on a composite objective function consisting of cross-entropy loss and soft-label regularization. This novel knowledge-inheritance paradigm enables the proposed TKI-FNN to effectively mitigate overfitting and improve fault diagnosis performance through soft-label regularization, while preserving the semantic interpretability of the downstream TSK module. In a series of experiments involving real industrial processes, the proposed model demonstrates significant advantages in both fault diagnosis accuracy and interpretability.

**Keywords:** fault diagnosis; deep neural network; soft label; Takagi-Sugeno-Kang; knowledge distillation; interpretability

## 0 引言

设备与系统的故障诊断对于保障生产安全和提

升系统可靠性具有关键作用<sup>[1-4]</sup>. 其核心任务在于通

过对系统监测数据的分析, 快速识别并定位故障, 并

收稿日期: 2026-01-20; 录用日期: 2026-04-21.

基金项目: 国家自然科学基金项目 (62473074, 62073056, 61876029); 辽宁省重点研发计划项目 (2024JH2/102400006).

责任编委: 方华京.

<sup>†</sup>通信作者. E-mail: luwei@dlut.edu.cn.

制定高效的维护策略. 美国化学安全与危害调查委员会数据显示, 在 2020 年 5 月至 2024 年 8 月期间, 共发生 25 起重大化学事故, 造成 7 人死亡、23 人重伤, 直接财产损失约 10 亿美元<sup>[6]</sup>. 此外, 2025 年 1 月, 美国宾夕法尼亚州斯普林代尔的 PPG 工厂发生爆炸<sup>[7]</sup>, 据报道该事故由设备故障引起, 造成多名员工受伤. 再如, 2023 年年底, 一艘韩国籍油船在上海港宝山航道深水航道延伸段突发舵机故障, 导致船舶失控<sup>[8]</sup>. 因此, 上述真实事故充分说明了研发高效且可靠的故障诊断方法, 对于保障生产安全、降低事故风险具有重要的现实意义.

基于深度学习 (DL) 的故障诊断方法已经被广泛应用于真实工业场景, 其通过分层堆叠多个特征转换模块从海量监测数据中提取丰富且具有区分度的特征. 随后通过其他专用模块对这些特征进行分析, 以判定故障发生并确定故障类型<sup>[9-11]</sup>. Li 等<sup>[12]</sup> 向 Transformer 引入稀疏约束, 构建了针对旋转机械振动脉冲段感知的故障诊断模型. Liu 等<sup>[14]</sup> 提出了一种融合多尺度卷积神经网络 (MRCNN) 与长短期记忆 (LSTM) 网络的并行融合故障诊断模型, 以应对模型退化问题. Chen 等<sup>[15]</sup> 设计了一种顺序集成多尺度卷积模块、通道注意力模块和 Transformer 编码器模块的混合故障诊断方法. 尽管基于 DL 的故障诊断方法已经广泛地验证是有效的, 但其固有的黑箱本质严重制约了在实际工业环境中的应用潜力.

近年来, 面向故障诊断任务, 基于深度模糊神经网络 (DFNN) 的高精度、可解释建模逐渐受到广泛关注, 已发展为该领域的重要研究方向之一<sup>[16,17]</sup>. 整体上, 这一技术路线通常通过串行式的集成结构来构建. 具体而言, 监测数据首先由深度神经网络 (DNN) 处理以提取故障判别性特征, 随后由模糊神经网络 (FNN), 尤其是具备语义可解释的 Takagi-Sugeno-Kang (TSK)<sup>[18]</sup>, 利用这些精炼特征完成故障类型的判别. Hu 等<sup>[19]</sup> 提出了一种名为多尺度模糊变分自编码器 (MFVAE) 的齿轮箱故障诊断模型, 以应对振动信号中的不确定性及模型可解释性不足的问题. 在该架构中, 首先采用两个不同规模的变分自编码器从振动信号中提取多尺度特征. 随后, 将融合后的特征分别输入到 FNN 和全连接网络中进行故障诊断. Li 等<sup>[20]</sup> 提出了一种基于粒计算的 DFNN 方法, 以应对长尾故障诊断的挑战. 该方法先使用模糊深度学习网络从原始数据中提取特征, 然后通过信息颗粒化模块将特征重组为颗粒球, 最后将这些颗粒球嵌入到 TSK 模型的前件中以执行故障诊断.

Jhang 等<sup>[21]</sup> 开发了一种专门用于旋转机械故障诊断的模型, 通过顺序集成卷积神经网络 (CNN) 和 TSK 实现. 总体而言, 现有的 DFNN 方法不仅巧妙地融合了 DNN 在复杂特征提取方面的优势, 使其性能可与主流 DNN 网络相媲美, 更重要的是, 还一定程度地保留了 FNN 所固有的语义可解释性.

尽管上述方法在故障诊断任务中表现出较好的有效性, 但其仍不可避免地存在若干局限之处: FNN 通过隶属函数的语言描述提取信息, 但难以准确刻画 DNN 生成的复杂故障特征, 从而掩盖了故障类间的微妙差异. 此外, 当 FNN 对前置 DNN 提取的故障特征进行处理时, 由于这些特征本身具有不可理解性, 其固有的可解释性会受到削弱. 总的来说, 串行集成方式并非 DFNN 的最优实现策略. 为解决基于 DFNN 的故障诊断方法固有的瓶颈, 本文提出了一种基于 Transformer 的知识继承式模糊神经网络 (TKI-FNN) 模型. 该框架突破了传统的串行范式, 创新性地引入知识蒸馏技术作为 DNN 与 FNN 的中介, 将二者有机融合, 旨在同时提升故障诊断性能与模型的可解释性. 所提方法作为面向各类故障诊断场景的通用模型, 专门针对多维监测特征和多类别故障诊断任务而设计. 具体来说, 在架构层面, TKI-FNN 模型包括三个主干模块: 先验 Transformer 分类器 (PTC) 模块, 知识蒸馏 (KD) 模块, 以及知识继承 TSK (KITSK) 模块. 首先, 输入数据在先验 PTC 模块中生成非归一化的概率分布知识. 随后, KD 模块对该知识进行进一步精炼, 以放大预测负类所携带的隐式信息, 从而生成软标签. 最后, 基于原始特征的 KITSK 模块通过后件接收这些软标签以执行故障识别. 该模型通过复合交叉熵损失与软标签正则化 (SLR) 损失实现参数优化. 值得注意的是, PTC 模块经过预训练, 为下游的 KITSK 模块提供知识, 因此 TKI-FNN 模型本质上是一种知识增强型 TSK 模型.

本文主要贡献如下: TKI-FNN 模型代表了一种全新的故障诊断实现范式, 通过独特的 KD 机制, 将先验 PTC 模块的分类相关知识传递 (继承) 至 KITSK 模块. 在该框架中, TSK 学习的是更易理解的软标签, 而非抽象的深度故障特征 (如 DFNN 的串行架构), 不仅增强了对负类别的感知能力, 还通过软标签正则化有效防止过拟合, 从而提升整体的诊断性能. 此外, 这种集成策略还完全维持了 TSK 的高度语义可解释性不被侵蚀. 更重要的是, 借助训练完善的 PTC 模块, 该方法无需从头训练, 进一步提升了模型效率. 因此, 无论相比于现有基于

DL 的方法还是基于 DFNN 的方法, 所提模型都在多个维度展现出显著优势, 代表了故障诊断方法的突破性进展。

## 1 理论基础

### 1.1 故障诊断原理

故障诊断的目标在于: 基于训练数据集, 构建一个诊断模型  $\mathcal{G}$ , 以学习过程观测数据与运行模式之间的映射关系, 从而实现未知样本运行状态的准确识别. 在实际工业环境中, 所采集的数据一般具有多变量形式, 并呈现典型的时序特性. 假设一个符合上述描述的数据集  $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, 2, \dots, N\}$ , 其包含一个正常类和  $K - 1$  个故障类.  $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,n}]^T \in \mathbb{R}^{n \times 1}$  表示第  $i$  个样本的  $n$  维过程观测向量,  $y_i \in \{1, 2, \dots, K\}$  为其对应的正常类或故障类别标签,  $K$  表示系统中考虑的类总数 (包括正常工况与多种故障类型). 根据这些类标签,  $\mathcal{D}$  被划分为一系列类特定的子集, 即  $\mathcal{D}^k = \{(\mathbf{x}_i, y_i) | y_i = k, i = 1, 2, \dots, N_k\}$ , 其中  $\mathcal{D} = \bigcup_{k=1}^K \mathcal{D}^k$  且  $N = \sum_{k=1}^K N_k$ .  $N_k$  表示第  $k$  个类特定子集中的样本基数.

从本质上讲, 故障诊断问题可抽象为一个多类别判别问题. 具体而言, 对于一个未见过的测试样本  $\mathbf{x}^{\text{test}}$ , 诊断模型应输出其对应的类预测结果  $y^{\text{test}}$ :

$$y^{\text{test}} = \mathcal{G}(\mathbf{x}^{\text{test}} | \mathbf{x}^{\text{train}}, y^{\text{train}}), \quad (1)$$

式 (1) 中,  $\mathbf{x}^{\text{train}}$  and  $y^{\text{train}}$  表示训练样本以及其对应的类标签.

### 1.2 基于 Transformer 编码器的特征提取

Transformer 编码器由多层自注意模块和前馈神经网络组成, 每层还包含残差连接和层归一化, 用于捕获序列的全局依赖并增强特征表示能力<sup>[12,13,15]</sup>. 在工业过程故障诊断任务中, 引入 Transformer 编码器可有效提取多维观测数据的全局与局部特征, 为判别模型提供信息丰富的表示, 从而提升诊断性能和泛化能力.

#### 1.2.1 位置编码

在 Transformer 中由于注意力机制的并行特性会导致序列的相对位置信息丢失, 因此在执行一般的特征嵌入后加入位置编码操作, 如下表达式:

$$\tilde{\mathbf{X}} = \mathbf{X} \mathbf{W}_p + \mathbf{b}_p + \mathbf{E}_{\text{pos}}, \quad (2)$$

其中  $\mathbf{X} \in \mathbb{R}^{t \times n}$  为通用的原始输入特征,  $t$  为序列长度.  $\mathbf{W}_p \in \mathbb{R}^{n \times d_{\text{model}}}$  表示可学习的线性投影权重矩阵, 用于将输入序列数据从原始特征空间映射到嵌入空间,  $d_{\text{model}}$  为嵌入维度.  $\mathbf{b}_p \in \mathbb{R}^{d_{\text{model}}}$  表示对应的偏置向量.  $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{t \times d_{\text{model}}}$  为位置编码矩阵, 用于引入序列的位置信息.  $\tilde{\mathbf{X}} \in \mathbb{R}^{t \times d_{\text{model}}}$  表示的是经过映射与位置编

码后的特征表示.

#### 1.2.2 多头注意力机制

在 Transformer 的第  $l$  个 ( $l = 1, 2, \dots, L$ ) 编码器层中, 多头注意力模块通过  $H$  个独立的单头注意力并行计算得到. 首先对第  $l - 1$  层的输出特征  $\tilde{\mathbf{X}}^{l-1}$  执行线性变换, 生成第  $h$  个头的查询  $\mathbf{Q}_h^l$ 、键  $\mathbf{K}_h^l$  和值  $\mathbf{V}_h^l$ , 其中  $h = 1, 2, \dots, H$ ,

$$\begin{aligned} \mathbf{Q}_h^l &= \tilde{\mathbf{X}}^{l-1} \mathbf{W}_{Q,h}^l, \\ \mathbf{K}_h^l &= \tilde{\mathbf{X}}^{l-1} \mathbf{W}_{K,h}^l, \\ \mathbf{V}_h^l &= \tilde{\mathbf{X}}^{l-1} \mathbf{W}_{V,h}^l. \end{aligned} \quad (3)$$

上式中,  $\mathbf{W}_{Q,h}^l, \mathbf{W}_{K,h}^l, \mathbf{W}_{V,h}^l \in \mathbb{R}^{d_{\text{model}} \times d_v}$  为第  $l$  层中第  $h$  个注意力头对应的查询、键和值的线性投影参数矩阵, 均为可学习参数.  $d_v$  为值向量维度. 对第  $h$  个头, 计算查询与键的相似度矩阵, 并通过 Softmax 得到注意力权重, 再与值矩阵加权求和,

$$\text{head}_h^l = \text{Softmax}\left(\frac{\mathbf{Q}_h^l \mathbf{K}_h^{lT}}{\sqrt{d_k}}\right) \mathbf{V}_h^l, \quad (4)$$

其中  $d_k$  是缩放因子, 用于防止点积值过大导致梯度消失或梯度爆炸. 然后将  $H$  个头拼接并执行线性映射:

$$\text{MultiHead}(\tilde{\mathbf{X}}^{l-1}) = \text{Concat}(\text{head}_1^l, \dots, \text{head}_H^l) \mathbf{W}_O^l, \quad (5)$$

其中  $\mathbf{W}_O^l \in \mathbb{R}^{H d_v \times d_{\text{model}}}$  用于将拼接后的多头输出映射回原始维度.

#### 1.2.3 前馈神经网络

前馈网络由两层线性映射和中间的 ReLU 激活函数组成, 其计算公式为:

$$\begin{aligned} \text{FFN}(\tilde{\mathbf{X}}^{\text{MHA},l}) &= \\ &\text{ReLU}(\tilde{\mathbf{X}}^{\text{MHA},l} \mathbf{W}_{\text{FFN1}} + \mathbf{b}_{\text{FFN1}}) \mathbf{W}_{\text{FFN2}} + \mathbf{b}_{\text{FFN2}}, \end{aligned} \quad (6)$$

其中,  $\mathbf{W}_{\text{FFN1}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ ,  $\mathbf{W}_{\text{FFN2}} \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$  为可学习权重矩阵.  $\mathbf{b}_{\text{FFN1}} \in \mathbb{R}^{d_{\text{ff}}}$ ,  $\mathbf{b}_{\text{FFN2}} \in \mathbb{R}^{d_{\text{model}}}$  为偏置向量.  $\text{ReLU}(\cdot)$  为逐元素非线性激活函数.  $\tilde{\mathbf{X}}^{\text{MHA},l}$  为多头注意力组件输出与输入残差融合后的中间结果, 其数学表示为,

$$\tilde{\mathbf{X}}^{\text{MHA},l} = \text{LayerNorm}(\text{MultiHead}(\tilde{\mathbf{X}}^{l-1}) + \tilde{\mathbf{X}}^{l-1}), \quad (7)$$

其中,  $\text{LayerNorm}(\cdot)$  为层归一化操作.

最终, Transformer 编码器利用残差连接和层归一化将多头注意力输出部分与前馈网络输出部分进行融合, 得到最终第  $l$  层编码器输出特征表示. 计算过程为:

$$\tilde{\mathbf{X}}^l = \text{LayerNorm}(\text{FFN}(\tilde{\mathbf{X}}^{\text{MHA},l}) + \tilde{\mathbf{X}}^{\text{MHA},l}), \quad (8)$$

### 1.3 目标知识蒸馏

目标知识蒸馏通过引入教师-学生学习范式<sup>[9,22]</sup>, 将教师模型中的隐性判别知识迁移至轻量化学生模型, 在提升性能与泛化能力的同时保持模型简洁性. 在目标知识蒸馏中, 教师模型对样本的判别信息主要体现在其输出层之前的 logit 向量上. logit 向量是指分类器在 Softmax 归一化之前的原始输出, 反映了模型对各类别的未归一化置信程度, 其中不仅包含最终预测类别的信息, 还隐含刻画了不同类别之间的相对判别关系, 因而被视为承载隐性判别知识的重要载体.

为了充分挖掘 logit 向量中所蕴含的隐性判别知识, 并缓解硬标签监督下类别间信息利用不足的问题, 引入蒸馏温度对原始 logit 向量进行平滑处理, 从而生成平滑化的 logit 向量. 设第  $i$  个样本对应的原始 logit 向量为  $\boldsymbol{\kappa}'_i = [\kappa'_{i,1}, \kappa'_{i,2}, \dots, \kappa'_{i,K}]^T \in \mathbb{R}^{K \times 1}$ . 则通过引入蒸馏温度系数  $\lambda > 0$ , 可得到平滑后的 logit 向量  $\boldsymbol{\kappa}'_i$ , 其计算方式如下:

$$\boldsymbol{\kappa}'_i = \frac{\boldsymbol{\kappa}_i}{\lambda} = \frac{[\kappa_{i,1}, \kappa_{i,2}, \dots, \kappa_{i,K}]^T}{\lambda}, \quad (9)$$

其中, 蒸馏温度  $\lambda (\lambda > 0)$  用于调节 logit 向量分布的平滑程度. 当  $\lambda > 1$  时, logit 差异被压缩, 弱化主导类别并增强非主导类别响应, 使模型能够刻画跨类相似关系. 在此基础上, 将平滑后的 logit 向量  $\boldsymbol{\kappa}'_i$  进一步通过 Softmax 函数映射为概率形式的软标签, 以显式揭示跨类相似性知识,

$$\bar{\kappa}'_{i,k}(\boldsymbol{\kappa}'_i) = \frac{\exp(\kappa'_{i,k})}{\sum_{s=1}^K \exp(\kappa'_{i,s})}, \quad (10)$$

其中,  $\bar{\kappa}'_{i,k}$  对应于软标签向量  $\bar{\boldsymbol{\kappa}}'_i = [\bar{\kappa}'_{i,1}, \bar{\kappa}'_{i,2}, \dots, \bar{\kappa}'_{i,K}]^T \in [0, 1]^{K \times 1}$  的第  $k$  个元素.

## 2 所提模型

本节逐步介绍了所提出的 TKI-FNN 模型的架构设计、训练流程及其可解释性, 全面呈现了该框架的核心逻辑与运行机制. 整体框架如图 1 所示.

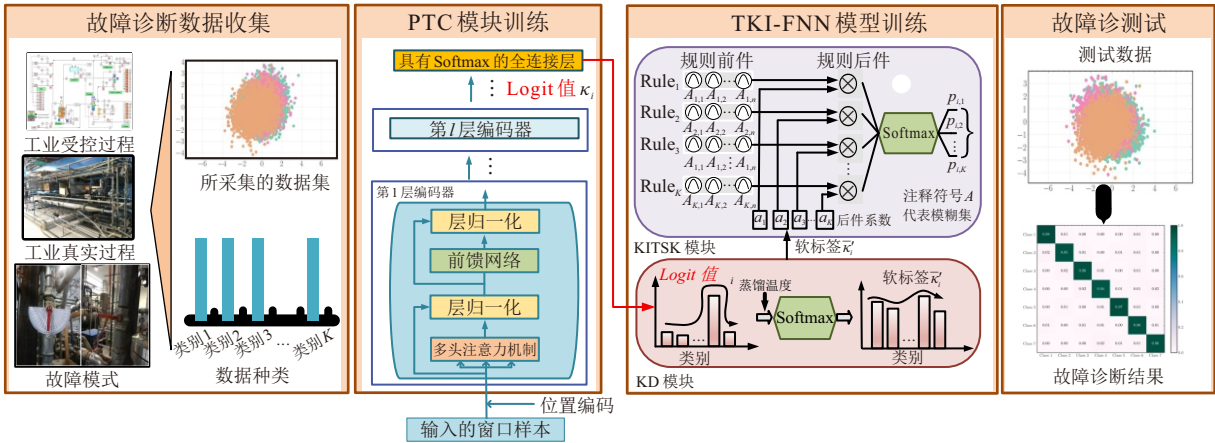


图1 TKI-FNN 模型架构

### 2.1 模型的提出

首先, 对数据集  $\mathcal{D}$  进行预处理, 以使其符合 PTC 模块的输入格式要求. 具体来说, 对每个数据子集系统地应用滑动窗口策略 (窗口宽度为  $w$ , 步长为 1), 以生成  $K$  个窗口集. 由第  $k$  个类特定子集形成如下窗口集,

$$\begin{cases} \mathcal{W}^k = \{\mathbf{W}_j^k | j = 1, 2, \dots, N_k - w + 1\}, \\ \text{其中 } \mathbf{W}_j^k = \{(\mathbf{x}_{i_j+z-1}^k, \mathbf{y}_{i_j+z-1}^k) \in \mathcal{D}^k | z = 1, 2, \dots, w\} \end{cases} \quad (11)$$

这里,  $\mathcal{W}^k$  表示复合窗口集,  $\mathbf{W}_j^k$  表示第  $k$  窗口集中的第  $j$  个窗口, 符号  $i_j$  表示第  $j$  个窗口的起始样本索引. 所有窗口集被整合为一个由  $K$  个类组成的复合窗口数据集, 记作  $\tilde{\mathcal{D}} = \bigcup_{k=1}^K \mathcal{W}^k = \{(\tilde{\mathbf{x}}_j, \tilde{\mathbf{y}}_j) | j = 1, 2, \dots, \tilde{N}\}$ , 其中  $\tilde{N} = \sum_{k=1}^K N_k - w + 1$  表示  $\tilde{\mathcal{D}}$  的基

数, 而  $\tilde{\mathbf{x}}_j \in \mathbb{R}^{w \times n}$  表示该数据集中第  $j$  个样本.

所提出的 TKI-FNN 模型建立在一个训练充分的 PTC 模块基础之上. 该模块采用第 1.2 节所述的 Transformer 编码器结构, 对窗口化数据集  $\tilde{\mathcal{D}}$  中的潜在特征进行提取, 并通过全连接网络 (FCN) 将其映射为非归一化的概率分布向量 (logit 向量)  $\boldsymbol{\kappa}_j = [\kappa_{j,1}, \kappa_{j,2}, \dots, \kappa_{j,K}]^T \in \mathbb{R}^{K \times 1}$ ,

$$\boldsymbol{\kappa}_j = \mathbf{W}_c(\mathbf{f}_j). \quad (12)$$

其中,  $\mathbf{W}_c \in \mathbb{R}^{K \times H}$  表示 FCN 的权重. 随后, Softmax 函数计算各类的概率值:

$$\bar{p}_{i_j}^k(\boldsymbol{\kappa}_j) = \frac{\exp(\kappa_{j,k})}{\sum_{s=1}^K \exp(\kappa_{j,s})}. \quad (13)$$

其中,  $\bar{p}_{i_j}^k$  表示样本  $\mathbf{x}_i \in \mathcal{D}$  归属于第  $k$  类的概率,

$k = 1, 2, \dots, K$ . PTC 模块输出的样本最终预测概率为:

$$\bar{y} = \arg \max_{k=1,2,\dots,K} (\bar{p}_{i_j}^k). \quad (14)$$

在训练阶段, PTC 模块的参数通过最小化交叉熵损失进行优化,

$$\min : \mathcal{L}_{\text{PTC}}(p^k, \bar{p}^k) = -\frac{1}{B} \sum_{i=1}^B \sum_{k=1}^K p_i^k \log(\bar{p}_{i_j}^k). \quad (15)$$

其中,  $B$  表示批处理量. 二值指示符  $p_i^k = 1$  表示样本  $\mathbf{x}_i$  属于第  $k$  类, 而  $p_i^k = 0$  则相反.

KD 模块将上游 PTC 模块生成的决策知识传递至下游 KITSK 模块以执行故障诊断, 其具体流程如第 1.3 节所述. 该模块通过引入蒸馏温度参数  $\lambda$ , 对潜在类区分的 logit 向量 (公式 (12) 所示) 进行平滑处理, 从而生成能够揭示更丰富类特定潜在信息的软标签, 尤其包含跨类相似性信息. 必须要强调的是, 随着蒸馏温度  $\lambda$  的升高, 软标签分布趋于更加平坦和柔和, 从而增强了对负标签所承载知识的关注. 然而, 过高的蒸馏温度可能掩盖不同故障类之间的差异, 使下游 KITSK 模块在识别某些难以区分的故障时性能下降. 因此, 合理调整蒸馏温度  $\lambda$  对模型性能至关重要.

KITSK 模块以原始数据集  $\mathcal{D}$  的特征作为前件输入, KD 模块提供的软标签作为后件输入变量, 从而生成所提 TKI-FNN 模型的最终故障诊断结果. 作为传统 TSK 的知识增强版本, KITSK 模块不仅融合了 PTC 模块的类特定知识以实现高精度故障诊断, 还能够凭借其内在的 If-Then 规则提供语义层面的解释. 值得注意的是, 与传统 TSK 模糊模型不同, 该模块中的模糊规则数量不再被简单视为网络超参数. 相反, 我们对其施加约束, 使规则数与数据集的标签类别保持一致. 在此配置下, 每条规则均得到充分利用, 并被明确用于判别样本是否属于其所对应的特定标签类别. 该模块中第  $k'$  个规则形式化如下:

$$\begin{aligned} & \text{Rule}_{k'} : \\ & \text{If} : x_1 \text{ is } A_{k',1} \text{ and } x_2 \text{ is } A_{k',2} \dots x_n \text{ is } A_{k',n} \\ & \text{Then} : g_{k'} = a_{k',1} \bar{\kappa}'_1 + a_{k',2} \bar{\kappa}'_2 + \dots + a_{k',K} \bar{\kappa}'_K. \end{aligned} \quad (16)$$

其中,  $A_{k',s}$  表示基于高斯模糊集, 其对应于第  $k'$  条规则中的第  $j$  个变量,  $k' = 1, 2, \dots, K$ ,  $s = 1, 2, \dots, n$ .  $\mathbf{a}_{k'} = [a_{k',1}, a_{k',2}, \dots, a_{k',K}]^T \in \mathbb{R}^{K \times 1}$  表示规则的后件参数.  $g_{k'}$  表示第  $k'$  个规则的后件输出, 其由软标签的线性组合构成. Rule $_{k'}$  的激活度如式 (17) 所示:

$$\varepsilon_{k'}(\mathbf{x}; \boldsymbol{\varphi}_{k',s}, \boldsymbol{\gamma}_{k',s}) = \prod_{s=1}^n \exp\left(-\frac{(x_s - \varphi_{k',s})^2}{2(\gamma_{k',s}^2)}\right). \quad (17)$$

其中,  $\varphi_{k',s}$  和  $\gamma_{k',s}$  分别表示模糊集  $A_{k',s}$  的中心与宽度. 向量  $\boldsymbol{\varphi}_{k'} \in \mathbb{R}^{n \times 1}$  与  $\boldsymbol{\gamma}_{k'} \in \mathbb{R}^{n \times 1}$  分别表示第  $k'$  条规则的均值与方差参数. Rule $_{k'}$  的归一化激活度由式 (18) 给出,

$$\bar{\varepsilon}_{k'} = \varepsilon_{k'} / \sum_{v=1}^K \varepsilon_v. \quad (18)$$

其中,  $\bar{\varepsilon}_{k'}$  表示第  $k'$  个规则的归一化规则激活度.

具体来说, 对于输入数据  $\mathbf{x}_i \in \mathcal{D}$ , KITSK 模块基于式 (19) 执行规则推理, 以计算类概率值:

$$\hat{p}^{k'}(\mathbf{x}_i, \bar{\boldsymbol{\kappa}}'_i) = \frac{\exp(S_{k'}(\mathbf{x}_i, \bar{\boldsymbol{\kappa}}'_i))}{\exp\left(\sum_{v=1}^K S_v(\mathbf{x}_i, \bar{\boldsymbol{\kappa}}'_i)\right)}. \quad (19)$$

其中,

$$S_{k'}(\mathbf{x}_i, \bar{\boldsymbol{\kappa}}'_i) = \bar{\varepsilon}_{k'}(\mathbf{x}_i) g_{k'}(\bar{\boldsymbol{\kappa}}'_i). \quad (20)$$

最终, 所提的 TKI-FNN 模型对输入样本  $\mathbf{x}_i$  的预测标签为:

$$\hat{y} = \arg \max_{k'=1,2,\dots,K} (\hat{p}^{k'}). \quad (21)$$

## 2.2 构建模型损失函数

TKI-FNN 模型融合了一个经过充分训练的 PTC 模块. 因此, 所提出模型的训练本质上主要集中在 KITSK 模块的参数优化上. 训练过程通过最小化一个由两部分组成的复合损失函数来完成: (1) 交叉熵损失确保了基本的故障诊断性能. (2) SLR 损失: 模型以软标签为引导, 促进 PTC 模块向 KITSK 模块传递更多知识, 同时有效抑制过拟合. 指导模型优化的复合损失如下所示:

模型损失:

$$\begin{aligned} \mathcal{L}(p^k, \hat{p}^k, \bar{\boldsymbol{\kappa}}'_k, \boldsymbol{\varphi}, \boldsymbol{\gamma}) &= (1 - \alpha) \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{slr}} = \\ &= -\frac{(1 - \alpha)}{P} \sum_{i=1}^P \sum_{k=1}^K p_i^k \log(\hat{p}_i^k) + \\ &= \frac{\alpha}{P \times K} \sum_{i=1}^P \sum_{k=1}^K (\hat{p}_i^k - \bar{\kappa}'_{i,k})^2. \end{aligned} \quad (22)$$

其中,  $\mathcal{L}_{\text{cls}}$ 、 $\mathcal{L}_{\text{slr}}$  分别表示交叉熵 (分类) 损失与 SLR 损失.  $P$  表示运算批大小. 概率变量  $p_i^k$  和  $\hat{p}_i^k$  分别对应真实与预测的类别概率.  $\alpha$  表示软标签正则化的权重. 具体而言, 当  $\alpha = 0$  时, KITSK 模块弃用软标签正则化. 当  $\alpha = 1$  时, 模块完全依赖于从 PTC 模块传递的知识进行训练参数. 值得注意的是, 在完全由正则化主导的训练模式下, PTC 模块中所存在的预测偏差或类识别不足等固有缺陷都无法通过真实标签加以修正. 因此, 合理地  $\alpha$  参数是保证高精度的故障

诊断性能的必要条件.

### 2.3 故障诊断流程

TKI-FNN 故障诊断流程从知识蒸馏的视角出发,对 DNN 与 FNN 进行有机集成,旨在有效缓解故障诊断中性能与可解释性之间的权衡问题.其整体算法流程如图 2 所示,具体步骤如下.

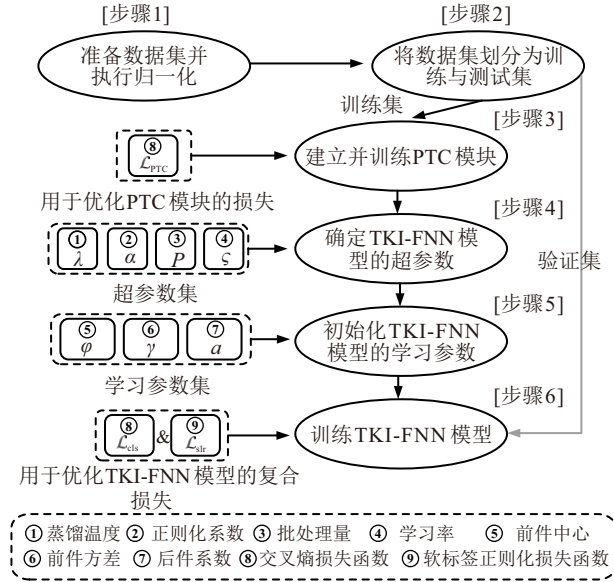


图2 训练流程图

**Step 1: 数据集预处理:** 对给定数据集  $\mathcal{D}$  执行最大最小归一化,消除不同维度间的数值差异,完成数据预处理.

**Step 2: 数据集划分:** 将归一化后的数据集拆分为训练集和验证集,验证集用于后续模型超参数的调优.

**Step 3: PTC 模块训练:** 依据式 (15) 的优化目标训练 PTC 模块,训练完成的 PTC 作为先验模型,为下游 KITSK 模块提供知识支撑.

**Step 4: 确定模型超参数并初始化学习参数:** 所提出的 TKI-FNN 模型需要指定两类超参数.结构参数与训练参数.结构参数通过设定蒸馏温度  $\lambda$  来决定模型的基本结构配置;相应地,训练超参数包括 SLR 损失项的权重系数  $\alpha$  (如式 (22) 所示)、学习率  $\varsigma$  以及运算批大小  $P$ .随后,训练过程首先在各自预定义的搜索空间内初始化 TKI-FNN 模型中的所有可学习参数:KITSK 模块中的参数  $\varphi_{k',s}$  与  $\gamma_{k',s}$  在区间  $[0, 1]$  内随机初始化,  $s = 1, 2, \dots, n$ ,  $k' = 1, 2, \dots, K$ .后件参数  $a_{k',k}$  服从均匀概率分布  $U[0, 1]$ .

**Step 5: TKI-FNN 模型训练:** 结合小批量梯度下降与 Adam 优化器<sup>[27]</sup>,最小化由交叉熵损失和 SLR 损失构成的复合损失函数(式 (22)).在每个训练迭代中,基于训练集随机采样的批次数据,通过链式法则

计算损失梯度,并沿梯度反方向更新模型参数.每次迭代后评估验证集性能,只有当复合损失相较于上一次迭代下降幅度超过预设阈值  $\epsilon\%$  时才进行参数更新,直至验证集性能达到最优:

$$\frac{L_{\text{prev}} - L_{\text{current}}}{L_{\text{prev}}} \times 100\% > \epsilon\%. \quad (23)$$

在上式中,  $L_{\text{prev}}$  代表上一次迭代或上一次参数更新时的复合损失值,  $L_{\text{current}}$  表示当前迭代计算得到的复合损失值,  $\epsilon\%$  作为预设的复合损失下降百分比阈值,用于判断是否显著改善并决定是否更新参数,在本文  $\epsilon$  设置为 0.5.上述训练过程被归纳在算法 1 中.此外,TKI-FNN 模型的算法复杂度为  $O(E(K^2 + Kn))$ ,其中  $E$  表示迭代次数.

训练完成后,TKI-FNN 模型通过顺序推理执行故障诊断.对于当前时间步的输入样本,该模型会结合前若干观测值形成观测窗口,输入到 PTC 模块生成 logit 向量,再由 KD 模块得到软标签.最终,KITSK 模块以原始样本为前件变量、软标签为后件变量,通过规则推理机制生成故障诊断决策.

#### 算法1 TKI-FNN模型训练流程

- 1: **输入:** 原始数据集  $\mathcal{D}$ , 蒸馏温度  $\lambda$ , 软标签正则化系数  $\alpha$ , 批次大小  $P$ ;
- 2: **输出:** 训练完成的 TKI-FNN 模型学习参数集  $\Theta$ ;
- 3: 对数据集  $\mathcal{D}$  进行最大最小归一化,并划分训练集  $\mathcal{D}^{\text{val}}$  与验证集  $\mathcal{D}^{\text{val}}$ ;
- 4: 利用滑动窗口策略生成复合窗口数据集  $\tilde{\mathcal{D}}$  (式(11));
- 5: 训练 PTC 模块: 最小化交叉熵损失  $\mathcal{L}_{\text{PTC}}$  (式(15));
- 6: 使用 PTC 模块生成 logit 向量  $\kappa_j$  (式(12));
- 7: 根据蒸馏温度  $\lambda$  生成软标签  $\bar{\kappa}_j$ ;
- 8: 初始化 KITSK 模块参数:
- 9: 前件参数  $\varphi_{k',s}, \gamma_{k',s} \in [0, 1]$  随机初始化;
- 10: 后件参数  $a_{k',k} \sim U[0, 1]$ ;
- 11: **for**  $e = 1$  **to** 最大迭代次数  $E$  **do**
- 12: **for** 每个批次  $b = 1$  **to** 批次大小  $B$  **do**
- 13: 计算 KITSK 模块规则前件激活度  $\varepsilon_{k'}$  (式(17)) 并归一化规则激活度  $\bar{\varepsilon}_{k'}$  (式(18));
- 14: 生成类别概率分布  $\hat{p}^{k'}$  (式(19) 和(20));
- 15: **end for**
- 16: 计算联合训练损失  $\mathcal{L}$  (联合交叉熵与 SLR 损失, 式(22));
- 17: 计算损失梯度并更新 KITSK 模块学习参数;
- 18: 在验证集  $\mathcal{D}^{\text{val}}$  上评估模型性能;
- 19: **if** 复合损失下降显著(式(23)) **then**
- 20: 更新最优学习参数集  $\Theta_{\text{optimal}}$ ;
- 21: **end if**
- 22: **end for**

23: 返回: 训练完成的TKI-FNN模型学习参数集 $\Theta_{optimal}$ .

### 2.4 模型可解释性

LI-FNN-KD 模型的可解释性主要来源于其规则化结构以及推理过程的透明化, 同时能够对规则进行语义层面的解读. 具体而言, 对于模型给出的预测类别, 可以在 KITSK 模块构建的规则库 (见式 (16)) 中定位与之最相关的类特定 If-Then 模糊规则. 所检索到的规则可从定性与定量两个层面对模型决策进行解释: 定性层面: 在所检索规则的前件部分, 各输入特征对应的模糊集均具备明确的语义术语, 用以刻画输入特征如何影响预测结果. 具体而言, 通过将模糊集中心映射至相应输入特征空间, 并依据其数值范围赋予相应的语义标签, 从而实现语义解释. 由此, 模型不仅能够指明输入特征所属的语义区域, 还阐明了其被识别为特定故障模式的机制. 定量层面: 在规则的后件部分, 通过对软标签施加线性组合权重, 刻画各类别对最终诊断结果的贡献方向与强度, 使模型的决策过程具备可量化的解释能力. 综上, 通过上述的分析方式, LI-FNN-KD 模型能够为故障诊断决策过程提供定性与定量相结合的深入解释.

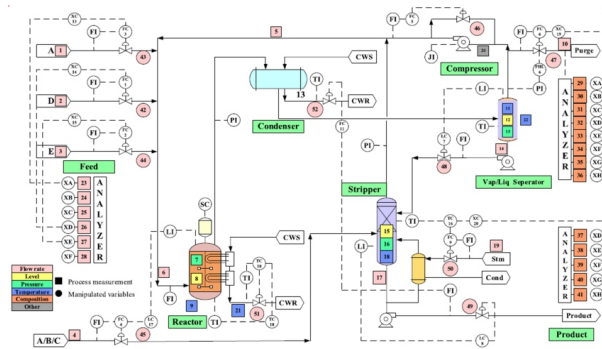
## 3 实验与结果分析

本节选取田纳西伊斯曼 (TE)、三相流 (TPF)、连续搅拌罐反应器 (CSTR) 和船舶柴油机 (MDE) 四个工业过程作为基准, 对所提出的 TKI-FNN 模型进行全面评估. 这四个工业过程的示意图如图 3 所示. 实验内容包括与基线模型的性能对比、模型可解释性分析, 以及关键组件对模型性能影响的评估.

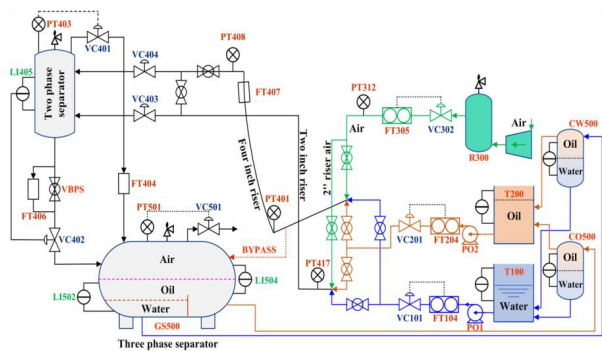
### 3.1 数据集

1) TE 数据集<sup>[14]</sup>: 美国伊斯曼化学公司开发的 TE 过程模拟平台是一套具有实际应用价值的化工过程仿真系统. 该流程由五个核心操作模块组成: 反应器、产品冷凝器、气液分离器、循环压缩机和产品汽提塔. 整个过程包含 41 个测量变量、12 个操作变量以及 18 种预设故障模式. 构建的 TE 数据集在训练集中包含 480 个正常类型样本, 以及每种故障类型各 500 个样本. 验证集中包含 1450 个正常类型样本, 以及每种故障类型各 360 个样本. 测试集的样本数量与验证集相同.

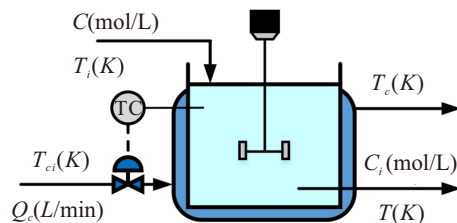
2) TPF 数据集<sup>[24]</sup>: TPF 过程基准来源于克兰菲尔德大学实验室的一个工业规模、全自动化的高压 TPF 系统. 该数据集包含 1 种正常状态和 5 种故障状态, 例如气路堵塞、水路堵塞以及分离器入口堵塞等. 在故障诊断实验中, 前 60% 的数据用于训练, 其余 40% 的数据则被随机均分为验证集和测试集.



(a) TE 过程



(b) TPF 过程



(c) CSTR 过程



(d) MDE 过程

图3 实验所需的四个工业过程示意图

3) CSTR 数据集<sup>[25]</sup>: CSTR 流程主要用于实现连续化学反应过程. 具体而言, CSTR 的故障模拟环境包括 10 个测量变量 (如入口浓度、入口温度、反应器温度、冷却流量等) 以及 10 种故障模式 (例如催化剂失活、传热系数失效和反应物浓度传感器偏差等). 在 CSTR 数据集中, 训练集中的每个类包含 1200 个样本, 验证集中的每个类包含 600 个样本. 测试集

的样本量与验证集相同。

4) MDE 数据集<sup>[28]</sup>: MDE 数据集来源于真实船用柴油机系统, 该系统采用 MAN 公司研制的 6S35ME-B9 型柴油机. 作为新一代智能柴油机, 该机型在输出功率、转速稳定性及燃油经济性方面具有显著优势. 实验数据采集于 2016 年 5 月, 共包含 15 个多传感器过程变量. 本研究考虑了排气管堵塞故障和空气冷却器冷却性能不足故障. 在 MDE 数据集中, 正常运行状态下共包含 12324 个样本, 而排气管堵塞和空气冷却器冷却不足两类故障样本数量分别为 4343 个和 3725 个. 实验过程中, 前 60% 用于模型训练, 其余 40% 等比例分配至验证集与测试集.

### 3.2 对比实验说明

#### 3.2.1 比较方法介绍

在对比方法的选取方面, 本文将所提出模型的故障诊断性能与十二种具有代表性的先进方法进行了比较, 覆盖了当前故障诊断领域中多种主流与前沿技术范式. 具体而言, 对比方法首先包括三类具有代表性的 FNN 模型, 即 Type-1 TSK、TSK-MUB<sup>[27]</sup> 和 RFNN<sup>[29]</sup>. 同时, 引入了四种典型的 DL 方法, 分别为 DBN<sup>[30]</sup>、SSAE-Softmax<sup>[31]</sup>、MRCNN-LSTM<sup>[14]</sup> 以及 ACEL<sup>[25]</sup>. 此外, 比较分析还纳入了五种具有代表性的 DFNN 方法, 包括 MFVAE<sup>[19]</sup>、DFM-FNCN<sup>[32]</sup>、FFT-FFR-RBFC<sup>[33]</sup>、GC-FDNN<sup>[20]</sup> 和 FFDNN<sup>[33]</sup>.

#### 3.2.2 超参数搜索区间

首先, 对训练集中所有特征进行归一化处理, 以消除不同量纲带来的影响. 随后, 设定超参数搜索空间并开展超参数优化. 具体而言, 蒸馏温度  $\lambda$  在 1, 10, 20, ..., 90 范围内选取; 正则化参数从区间  $\alpha \in \{0, 0.1, 0.2, \dots, 1.0\}$  中搜索; 批量大小  $P$  从 32, 64, 128, 256, 512, 1024 中选择; 学习率  $\zeta$  在  $1 \times 10^{-4}$  至  $1 \times 10^{-3}$  范围内调节. 在上述搜索空间基础上, 采用随机搜索策略对所提出模型的超参数进行精细调优, 最优参数配置详见表 1.

表1 所提模型关键超参数的最佳配置

工业数据集	知识蒸馏温度 $\lambda$	软标签正则化系数 $\alpha$
TE数据集	20	0.6
TPF数据集	30	0.5
CSTR数据集	50	0.7
MDE数据集	10	0.2

在十二个对比基线中, 对于 FNN 类方法, Type-1 TSK、TSK-MB 和 RFNN 的规则数目在区间  $[10, 100, 10]$  内进行搜索, 其中 10, 100 和 10 分别代表搜索区间的下界上界和步长. 对于 DL 类方法,

DBN 的网络层数和每层神经元数目分别在区间  $[1, 5, 1]$  和  $[10, 300, 10]$  内进行搜索. SSAE-Softmax 的编码器参数设置与 DBN 相同, 其解码器与编码器呈镜像对称结构. 在 MRCNN-LSTM 中, 三个并行深度卷积网络的卷积核数搜索区间分别为  $[1, 3, 1]$ 、 $[4, 6, 1]$  和  $[7, 9, 1]$ , 以提取多尺度空间特征. LSTM 部分的层数与每层神经元数目则在  $[1, 5, 1]$  和  $[10, 150, 10]$  区间内搜索确定. ACEL 与 MRCNN-LSTM 具有相似的整体结构, 其主要超参数的搜索区间与 MRCNN-LSTM 保持一致. 在 DFNN 类方法中, MFVAE 的变分自编码器层数和每层特征维度分别在区间  $[1, 5, 1]$  和  $[4, 256, 4]$  内进行搜索, 其解码器与编码器呈对称结构. FNN 部分的规则数目在区间  $[10, 100, 10]$  内搜索. 在 DFM-FNCN 中, 深度卷积网络的层数、每层卷积核宽度以及通道数目分别在区间  $[1, 5, 1]$ 、 $[1, 7, 1]$  和集合  $\{8, 16, 32, 64\}$  中进行设置, 其中 FNN 部分采用自动生成机制, 无需手动设定规则数目. FFT-FFR-RBFC 由  $L$  个后件为非线性的 TSK 模块级联构成, 其中  $L$  表示级联结构的总层数, 其取值在区间  $[1, 5, 1]$  内搜索, 且  $l \in 1, 2, \dots, L$  表示第  $l$  个 TSK 模块. 对于与原始输入直接相连的首层 TSK 模块, 其规则数目限定在  $[10, 100]$  范围内. 当  $l > 1$  时, 该范围进一步约束为  $[10 - (l - 1), M_{l-1}]$ , 以确保学习到低维且类可分的故障特征, 其中  $M_{l-1}$  为第  $l - 1$  个非线性 TSK 模块的规则数目. 在 GC-FDNN 中, 蒸馏层数目  $E$  设置在区间  $[1, 5, 1]$  内, 其中  $E$  表示蒸馏层的总层数, 且  $e \in 1, 2, \dots, E$  表示第  $e$  层蒸馏层. 首层神经元数量限定在  $[10, 300]$  范围内. 当  $e > 1$  时, 该范围同样约束为  $[10 - (e - 1), \widetilde{M}_{e-1}]$ , 以确保学习到低维且鲁棒的故障特征, 其中  $\widetilde{M}_{e-1}$  为第  $e - 1$  层蒸馏层的神经元数目. 此外, GC-FDNN 中 FNN 部分的规则数目与类别数保持一致. 对于 FFDNN, 其 DNN 部分的网络层数和每层神经元数目分别在区间  $[1, 5, 1]$  和  $[10, 300, 10]$  内进行搜索, 而 FNN 部分的规则数目在区间  $[10, 100, 10]$  内搜索. 上述所有对比方法中, 批处理大小与学习率的搜索区间均与所提 TKI-FNN 模型保持一致. 此外, 对比基线与所提方法均进行了 10 次重复实验以减轻潜在的随机性.

#### 3.2.3 性能评估指标

在比较实验中, 故障诊断性能通过三项指标进行评估: 宏平均准确率 (Macc)、宏平均 F1 分数 (MF1) 和宏平均召回率 (MRecall), 其数学表示如式 (24), (25) 和 (26) 所示.

表2 TE 与 TPF 数据集的故障诊断性能比较 (均值±标准差)

模型	TE数据集			TPF数据集		
	Macc	MF1	MRecall	Macc	MF1	MRecall
Type-1 TSK	0.593 ± 0.039	0.607 ± 0.038	0.571 ± 0.037	0.403 ± 0.063	0.394 ± 0.077	0.411 ± 0.079
TSK-MUB <sup>[27]</sup>	0.681 ± 0.013	0.683 ± 0.015	0.674 ± 0.015	0.582 ± 0.034	0.602 ± 0.039	0.628 ± 0.039
RFNN <sup>[29]</sup>	0.668 ± 0.014	0.674 ± 0.015	0.682 ± 0.015	0.614 ± 0.084	0.635 ± 0.072	0.653 ± 0.063
DBN <sup>[30]</sup>	0.704 ± 0.019	0.706 ± 0.010	0.723 ± 0.009	0.648 ± 0.030	0.650 ± 0.039	0.662 ± 0.035
SSAE-Softmax <sup>[31]</sup>	0.708 ± 0.016	0.701 ± 0.019	0.703 ± 0.021	0.702 ± 0.039	0.703 ± 0.040	0.723 ± 0.030
MRCNN-LSTM <sup>[14]</sup>	0.736 ± 0.044	0.774 ± 0.027	0.783 ± 0.026	0.703 ± 0.049	0.697 ± 0.053	0.713 ± 0.045
ACEL <sup>[25]</sup>	0.791 ± 0.016	0.777 ± 0.025	0.791 ± 0.022	0.769 ± 0.043	0.757 ± 0.038	0.757 ± 0.039
MFVAE <sup>[19]</sup>	0.750 ± 0.022	0.753 ± 0.025	0.752 ± 0.022	0.720 ± 0.034	0.711 ± 0.039	0.741 ± 0.035
DFM-FNCN <sup>[32]</sup>	<b>0.793 ± 0.033</b>	0.798 ± 0.033	0.803 ± 0.031	0.788 ± 0.033	0.782 ± 0.034	0.787 ± 0.044
FFT-FFR-RBFC <sup>[33]</sup>	0.690 ± 0.024	0.677 ± 0.027	0.672 ± 0.026	0.672 ± 0.041	0.663 ± 0.042	0.687 ± 0.034
GC-FDNN <sup>[20]</sup>	0.791 ± 0.019	0.807 ± 0.015	<b>0.819 ± 0.013</b>	0.730 ± 0.039	0.731 ± 0.039	0.727 ± 0.040
FFDNN <sup>[33]</sup>	0.724 ± 0.008	0.707 ± 0.006	0.712 ± 0.007	0.789 ± 0.040	0.785 ± 0.035	<b>0.813 ± 0.032</b>
TKI-FNN	0.791 ± 0.010	<b>0.809 ± 0.006</b>	0.817 ± 0.003	<b>0.824 ± 0.001</b>	<b>0.804 ± 0.001</b>	0.798 ± 0.001

表3 CSTR 与 MDE 数据集的故障诊断性能比较 (均值±标准差)

模型	CSTR数据集			MDE数据集		
	Macc	MF1	MRecall	Macc	MF1	MRecall
Type-1 TSK	0.689 ± 0.021	0.687 ± 0.021	0.666 ± 0.018	0.836 ± 0.031	0.757 ± 0.081	0.758 ± 0.052
TSK-MUB <sup>[27]</sup>	0.746 ± 0.011	0.757 ± 0.010	0.730 ± 0.009	0.828 ± 0.003	0.753 ± 0.007	0.749 ± 0.005
RFNN <sup>[29]</sup>	0.788 ± 0.007	0.791 ± 0.009	0.761 ± 0.009	0.850 ± 0.002	0.806 ± 0.005	0.793 ± 0.006
DBN <sup>[30]</sup>	0.805 ± 0.021	0.813 ± 0.021	0.799 ± 0.020	0.847 ± 0.009	0.795 ± 0.022	0.784 ± 0.021
SSAE-Softmax <sup>[31]</sup>	0.936 ± 0.004	0.941 ± 0.004	0.931 ± 0.005	0.861 ± 0.005	0.824 ± 0.007	0.809 ± 0.007
MRCNN-LSTM <sup>[14]</sup>	0.893 ± 0.016	0.899 ± 0.015	0.891 ± 0.014	0.913 ± 0.008	0.896 ± 0.011	0.873 ± 0.012
ACEL <sup>[25]</sup>	0.827 ± 0.021	0.831 ± 0.024	0.814 ± 0.024	0.924 ± 0.072	0.910 ± 0.084	0.890 ± 0.090
MFVAE <sup>[19]</sup>	0.918 ± 0.003	0.926 ± 0.002	0.911 ± 0.003	0.921 ± 0.003	0.907 ± 0.005	0.887 ± 0.007
DFM-FNCN <sup>[32]</sup>	0.893 ± 0.090	0.902 ± 0.090	0.879 ± 0.101	0.875 ± 0.076	0.839 ± 0.095	0.806 ± 0.106
FFT-FFR-RBFC <sup>[16]</sup>	0.899 ± 0.006	0.906 ± 0.006	0.888 ± 0.006	0.860 ± 0.005	0.820 ± 0.009	0.805 ± 0.009
GC-FDNN <sup>[20]</sup>	0.878 ± 0.005	0.890 ± 0.005	0.870 ± 0.006	0.927 ± 0.005	0.916 ± 0.006	0.904 ± 0.009
FFDNN <sup>[33]</sup>	0.924 ± 0.003	0.931 ± 0.003	0.914 ± 0.004	0.927 ± 0.003	0.915 ± 0.004	0.895 ± 0.006
TKI-FNN	<b>0.955 ± 1.2 × 10<sup>-4</sup></b>	<b>0.958 ± 1.1 × 10<sup>-4</sup></b>	<b>0.952 ± 1.3 × 10<sup>-4</sup></b>	<b>0.941 ± 9.6 × 10<sup>-5</sup></b>	<b>0.934 ± 8.4 × 10<sup>-5</sup></b>	<b>0.925 ± 7.9 × 10<sup>-5</sup></b>

$$\text{Macc} = \frac{1}{K} \sum_{k=1}^K \frac{\text{TP}_k + \text{TN}_k}{\text{TP}_k + \text{TN}_k + \text{FP}_k + \text{FN}_k}, \quad (24)$$

$$\text{MF1} = \frac{1}{K} \sum_{k=1}^K \frac{2\text{TP}_k}{2\text{TP}_k + \text{FP}_k + \text{FN}_k}, \quad (25)$$

$$\text{MRecall} = \frac{1}{K} \sum_{k=1}^K \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}. \quad (26)$$

上式中  $K$  表示类别总数,  $\text{TP}_k$ 、 $\text{TN}_k$ 、 $\text{FP}_k$  和  $\text{FN}_k$  分别表示第  $k$  类的真阳性、真阴性、假阳性和假阴性。

### 3.3 实验结果分析

表 2 和 3 汇报了所有方法在四个数据集上的性能. 与 FNN 方法 (包括 Type-1 TSK、TSK-MUB 和 RFNN) 在四类工业数据集上的对比结果表明, TKI-FNN 模型在故障诊断性能上实现了全面超越. 例如, 在 TE 数据集上, TKI-FNN 的 Macc 达到 79.1%, 分别较 Type-1 TSK、TSK-MUB 和 RFNN 提升 33.39%、

16.15% 和 18.41%; MF1 达到最优值 80.9%, 较三者分别提升 33.28%、18.45% 和 20.03%, 同时 MRecall 也显著提升. 相比之下, 传统 FNN 方法由于网络结构较为单一, 难以有效捕捉工业故障信号中复杂的非线性耦合关系.

针对 DBN、SSAE-Softmax、MRCNN-LSTM、ACEL 等主流 DL 模型的对比实验结果表明, TKI-FNN 在保持深度学习强特征表征能力的同时, 进一步提升了故障诊断的准确性与稳定性. 以 CSTR 数据集为例, TKI-FNN 的 Macc 达到 95.5%, 相比性能最优的 DL 模型 SSAFE-Softmax 提升了 2.03%; 其在 MF1 和 MRecall 指标下同样取得最优结果, 体现出更强的故障分类能力. 上述性能增益主要归功于引入的 KD 和 SLR 技术, 这些设计有效促进了判别性知识从上游 PTC 模块向 KITSK 模块的迁移, 同时抑

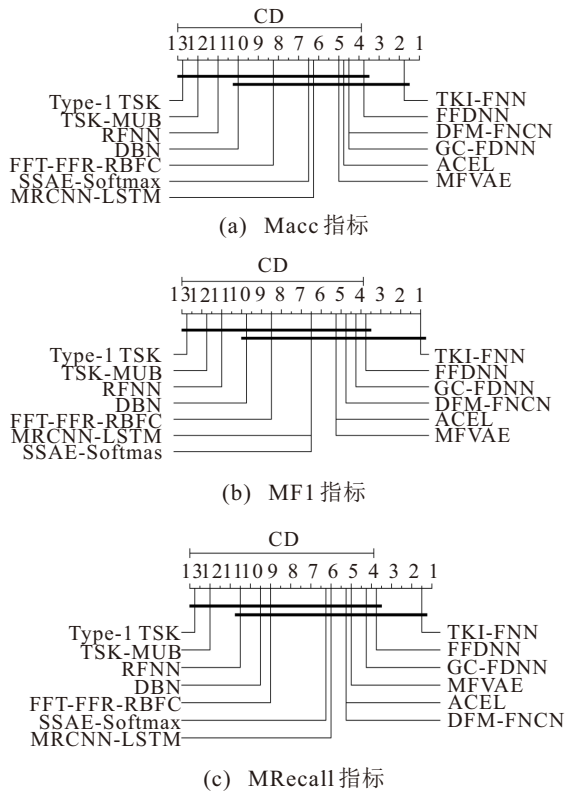


图4 Friedman-Nemenyi 事后检验结果 (基于四个工业数据集)

制了过拟合现象,从而提升了模型的稳定性与泛化能力.更为重要的是,TKI-FNN在保持高性能的同时,还在模型透明度和可解释性方面展现出显著优势.

在与MFVAE、DFM-FNCN、FFT-FFR-RBFC、GC-FDNN和FFDNN等DFNN方法的比较中,TKI-FNN在四类数据集的全部评价指标上均保持明显领先.所提出的TKI-FNN模型通过引入KD机制,实现了DNN与FNN的有效融合,相较于当前先进的DFNN方法,在故障诊断性能方面取得了可量化且

稳定的提升.

为进一步验证所提出TKI-FNN模型的优越性能,本研究同时开展了Friedman检验和Nemenyi事后检验,在全部四个数据集上将该模型与十二种基线方法进行了系统比较.Friedman检验用于判断各算法之间是否存在统计显著差异.针对三种评价指标,其对应的Friedman统计量的 $p$ 值分别为 $1.03e^{-04}$ , $1.12e^{-04}$ 和 $2.16e^{-07}$ ,均远小于显著性水平( $\beta = 0.05$ ),表明在不同评价指标下,各方法之间均存在统计显著差异.基于排序结果的Nemenyi事后检验通过图4中的临界差异(CD)图进行了可视化.具体而言,CD的数学表达式为,

$$CD = q^\beta \sqrt{\frac{k^*(k^* + 1)}{6N^*}} \quad (27)$$

其中, $k^*$ 表示算法的数量, $N^*$ 表示数据集的数量, $q^\beta$ 为在显著性水平 $\beta$ ( $\beta = 0.05$ )下由学生化极差分布确定的临界值.Nemenyi事后检验结果表明,传统FNN方法(Type-1 TSK、TSK-MUB和RFNN)整体表现最差,在三项评价指标上的平均排名均大于10.相比之下,DL方法(如DBN、SSAE-Softmax等)的性能显著优于FNN方法.所提出的TKI-FNN模型在四个数据集的三项指标上均取得最优表现,其平均排名明显优于其他模型,且性能优势具有统计学显著性.

此外,模型效率指标,包括训练时间、图形处理器(GPU)消耗、参数量、浮点运算量(FLOPs)以及推理时间,可从多个维度全面衡量模型在工业场景中的实际落地能力.FLOPs和推理时间均在单个测试样本层面进行评估.表4汇总了各基线模型在不同数据集上的训练时长与GPU消耗情况.结果表明,

表4 所有数据集在训练时间(s)和GPU(MiB)消耗上的比较结果

模型	TE数据集		TPF数据集		CSTR数据集		MDE数据集	
	训练时间	GPU消耗	训练时间	GPU消耗	训练时间	GPU消耗	训练时间	GPU消耗
Type-1 TSK	12.05	215	27.50	189	14.61	191	15.23	205
TSK-MUB <sup>[27]</sup>	11.66	189	27.36	168	74.71	238	12.85	198
RFNN <sup>[29]</sup>	33.59	170	47.49	155	191.67	163	38.72	168
DBN <sup>[30]</sup>	38.78	219	77.74	176	138.20	174	42.35	188
SSAE-Softmax <sup>[31]</sup>	19.23	168	46.15	177	146.15	176	22.58	182
MRCNN-LSTM <sup>[14]</sup>	713.45	3099	1723.99	3213	4538.73	930	5216.38	1050
ACEL <sup>[25]</sup>	34.98	330	61.38	312	490.20	272	38.65	295
MFVAE <sup>[19]</sup>	24.41	261	60.79	263	57.20	201	62.85	225
DFM-FNCN <sup>[32]</sup>	80.45	1415	85.82	558	201.28	292	225.63	320
FFT-FFD-RBFC <sup>[16]</sup>	322.23	224	184.72	243	297.97	201	315.80	235
GC-FDNN <sup>[20]</sup>	142.79	464	91.10	336	176.29	300	189.55	340
FFDNN <sup>[33]</sup>	50.16	568	100.25	505	72.25	425	78.90	455
TKI-FNN	18.13	429	74.48	334	23.32	254	19.93	227

表5 所有数据集在参数量 (M)、FLOPs (M) 和推理时间 (ms) 上的比较结果

模型	TE数据集			TPF数据集			CSTR数据集			MDE数据集		
	参数量	FLOPs	推理时间	参数量	FLOPs	推理时间	参数量	FLOPs	推理时间	参数量	FLOPs	推理时间
Type-1 TSK	0.0085	0.0066	$8.2 \times 10^{-5}$	0.0158	0.0115	$2.8 \times 10^{-4}$	0.0047	0.0027	$5.9 \times 10^{-6}$	0.0005	0.0013	$3.5 \times 10^{-6}$
TSK-MUB <sup>[27]</sup>	0.0392	0.0345	0.0002	0.0119	0.0087	0.0001	0.0011	0.0016	0.0002	0.0043	0.0025	0.0007
RFNN <sup>[29]</sup>	0.0162	0.0187	0.0016	0.0004	0.0002	0.0028	0.0193	0.0167	0.0008	0.0143	0.0183	0.0001
DBN <sup>[30]</sup>	1.1864	1.1792	0.0016	0.3715	0.3698	0.0002	0.3592	0.3576	0.0003	0.3618	0.3601	0.0003
SSAE-Softmax <sup>[31]</sup>	1.2129	1.2100	0.0018	0.3944	0.3926	0.0001	0.3807	0.3789	0.0004	0.3852	0.3834	0.0001
MRCNN-LSTM <sup>[14]</sup>	0.2378	3.1653	0.0017	0.2249	2.2924	0.0026	0.2189	0.9682	0.0005	0.2223	0.0986	0.0001
ACEL <sup>[25]</sup>	1.9423	2.1443	0.0041	0.5487	0.4072	0.0003	1.9076	2.1152	0.0005	0.1851	0.4758	0.0028
MFVAE <sup>[19]</sup>	10.1571	0.3878	0.0040	3.0458	0.1617	0.0033	9.6811	0.1510	0.0004	9.4978	0.0591	0.0003
DFM-FNCN <sup>[32]</sup>	10.7798	3.7481	0.0015	7.6355	2.4421	0.0030	4.2286	0.9228	0.0005	2.2926	0.0313	0.0003
FFT-FFD-RBFC <sup>[33]</sup>	12.3642	4.1257	0.0032	8.9516	2.8763	0.0035	5.1369	1.1042	0.0007	3.1428	0.0426	0.0015
GC-FDNN <sup>[20]</sup>	8.7526	1.8634	0.0029	6.2184	1.3592	0.0031	7.8653	0.5871	0.0005	8.1247	0.1263	0.0002
FFDNN <sup>[33]</sup>	10.6937	0.6145	0.0027	10.3544	0.4557	0.0038	9.8293	0.2122	0.0006	9.4867	0.0512	0.0001
TKI-FNN	0.0017	0.0004	0.0002	0.0003	$4.8 \times 10^{-5}$	0.0001	0.0004	0.0001	0.0001	0.0001	$1.5 \times 10^{-5}$	0.0002

表6 消融实验结果 (均值±标准差)

模型	TE数据集			TPF数据集		
	Macc	MF1	MRecall	Macc	MF1	MRecall
TKI-FNN-w/oPTC & KD	$0.654 \pm 0.025$	$0.662 \pm 0.032$	$0.682 \pm 0.021$	$0.587 \pm 0.038$	$0.604 \pm 0.042$	$0.631 \pm 0.043$
TKI-FNN-w/o KD & KITSK	$0.777 \pm 0.018$	$0.793 \pm 0.018$	$0.802 \pm 0.016$	$0.764 \pm 0.030$	$0.747 \pm 0.023$	$0.759 \pm 0.023$
TKI-FNN-w/o KD	$0.782 \pm 0.010$	$0.785 \pm 0.015$	$0.799 \pm 0.011$	$0.768 \pm 0.012$	$0.751 \pm 0.016$	$0.763 \pm 0.038$
TKI-FNN	<b><math>0.791 \pm 0.010</math></b>	<b><math>0.809 \pm 0.006</math></b>	<b><math>0.817 \pm 0.003</math></b>	<b><math>0.824 \pm 0.001</math></b>	<b><math>0.804 \pm 0.001</math></b>	<b><math>0.798 \pm 0.001</math></b>
模型	CSTR数据集			MDE数据集		
	Macc	MF1	MRecall	Macc	MF1	MRecall
TKI-FNN-w/o PTC & KD	$0.768 \pm 0.018$	$0.747 \pm 0.028$	$0.781 \pm 0.032$	$0.831 \pm 0.015$	$0.749 \pm 0.019$	$0.751 \pm 0.009$
TKI-FNN-w/o KD & KITSK	$0.948 \pm 0.005$	$0.953 \pm 0.004$	$0.947 \pm 0.004$	$0.886 \pm 0.015$	$0.852 \pm 0.023$	$0.830 \pm 0.023$
TKI-FNN-w/o KD	$0.935 \pm 0.008$	$0.949 \pm 0.007$	$0.948 \pm 0.005$	$0.891 \pm 0.010$	$0.854 \pm 0.015$	$0.845 \pm 0.016$
TKI-FNN	<b><math>0.955 \pm 1.2 \times 10^{-4}</math></b>	<b><math>0.958 \pm 1.1 \times 10^{-4}</math></b>	<b><math>0.952 \pm 1.3 \times 10^{-4}</math></b>	<b><math>0.941 \pm 9.6 \times 10^{-5}</math></b>	<b><math>0.934 \pm 8.4 \times 10^{-5}</math></b>	<b><math>0.925 \pm 7.9 \times 10^{-5}</math></b>

与 Type-1 TSK 和 TSK-MUB 相比, 所提出模型的训练过程耗时更长, 这主要源于其在优化过程中需要同时平衡多项损失函数. 在 GPU 消耗方面, 该模型在部分数据集上略高于 DBN、SSAE-Softmax 等方法. 尽管计算资源需求有所增加, 但这带来了更优的故障诊断性能和更强的语义可解释性. 除此之外, 从

表 5 中参数量、FLOPs 以及推理时间的结果可以看出, 所提模型进一步体现出较好的轻量性优势. 特别是在与 DL 类故障诊断方法和 DFNN 类方法的对比中, 所提模型表现出显著优势. 总体而言, 通过多方面模型效率指标的综合评估, 所提模型在实际应用中展现出更高的综合价值.

表7 模糊规则中前件均值 $\varphi$ , 方差 $\gamma$ 和后件系数 $\alpha$ 

规则	模糊规则中前件均值 $\varphi$ , 方差 $\gamma$ 和后件系数 $\alpha$
Rule Normal	If: $\varphi_1 = (1.4116, 0.3091, 0.5824, 0.4625, \dots, 1.2215)$ $\gamma_1 = (0.6952, 0.8411, 0.6843, 0.1268, \dots, 0.3448)$ Then: $\alpha_1 = (0.06615, 0.0529, 0.0034, \dots, -0.0414)$
Rule Fault1	If: $\varphi_2 = (1.5854, 0.4475, 0.5478, 0.3031, \dots, 1.8255)$ $\gamma_2 = (0.6743, 0.1683, 0.7351, 0.4318, \dots, 0.9005)$ Then: $\alpha_2 = (-0.0005, 0.1042, 0.011, \dots, -0.0656)$
	⋮
Rule Fault18	If: $\varphi_{19} = (0.7413, 0.6515, 0.5822, 0.4461, \dots, 1.6463)$ $\gamma_{19} = (0.9039, 1.3985, 0.7856, 0.5412, \dots, 0.0256)$ Then: $\alpha_{19} = (-0.0293, -0.0116, -0.0812, \dots, -0.0112)$

### 3.4 消融实验

在本节中,设计了三种消融实验,TKI-FNN-w/o PTC & KD、TKI-FNN-w/o KD & KITSK 和 TKI-FNN-w/o KD,以系统地验证关键组件对模型性能 的贡献.其中,TKI-FNN-w/oPTC & KD (其中 w/o 表示 “without”)表示前件和后件输入均使用原始特征 的一型 TSK;TKI-FNN-w/o KD & KITSK 仅保留先 验 PTC 模块;TKI-FNN-w/o KD 则将 PTC 模块 的 logit 向量直接输入到 KITSK 的后件,而不经 过 KD 技术处理(即  $\lambda = 1$ ),同时从复合损失中剔除 SLR 项.表 6 给出了消融实验结果,可见 TKI-FNN 在各项指标上均优于全部消融变体.相比 TKI-FNN-w/o PTC & KD,TKI-FNN 的显著性能提升主要来源于 DL 网络对复杂数据知识的学习,这些知识为下游 KITSK 模块精确识别各种复杂故障提供了有力 支持.关于消融变体 TKI-FNN-w/o KD & KITSK 和 TKI-FNN-w/o KD 的定量结果进一步表明了 KD 与 软标签正则化协同作用的重要性:KD 深入挖掘了 DL 网络中的类别相似性知识,而软标签正则化通过 生成更平滑的目标分布,有效抑制了过拟合现象.

### 3.5 模型可解释性分析

为了展示所提出的 TKI-FNN 模型的可解释性, 我们在单个测试样本上的故障诊断决策解释,考虑 样本  $\boldsymbol{x} = [0.5956, 0.5073, \dots, 0.1117]^T \in \mathbb{R}^{33 \times 1}$ . 对应地,由 PTC 模块生成的 logit 向量为  $\boldsymbol{\kappa} = [-2.4537, 5.637, \dots, -3.5623]^T \in \mathbb{R}^{19 \times 1}$ . 而经过 KD 模块处理得到的软标签为  $\boldsymbol{\kappa}' = [0.0528, 0.0539, \dots, 0.0526]^T \in \mathbb{R}^{19 \times 1}$ . 训练完成的 TKI-FNN 模型 的规则库,更具体地说是 KITSK 模块的规则库,包 含 19 条模糊规则.表 7 给出了这些规则的参数设定, 其中  $\varphi_k$  与  $\gamma_k$  分别表示第  $k$  条模糊规则 If 部分的参 数( $\varphi_k$  为高斯隶属函数的均值,  $\gamma_k$  为方差),而  $\boldsymbol{a}_k$  则 对应第  $k$  条规则 Then 部分的参数.

进一步地,为了赋予规则语义,我们将模糊集 的中心映射到对应的输入特征维度,并根据数值大小 进行升序排序.具体而言,排序后的区间 [1, 7]、 [8, 13]、[14, 19]、[20, 26] 和 [27, 33] 分别对应语义标 签 “非常低”、“低”、“中”、“高”和 “非常高”.规则 库的可视化结果如图 5 所示.在图 5 中,图中  $x_1 \sim x_{33}$  表示原始的 33 个输入变量,其中  $x_1$  表示 A 进料的流量,  $x_2$  表示 D 进料的流量,  $x_3$  表示 E 进 料的流量,而  $x_4$  表示 A 与 C 混合进料的流量,  $x_{33}$  表 示搅拌器转速.

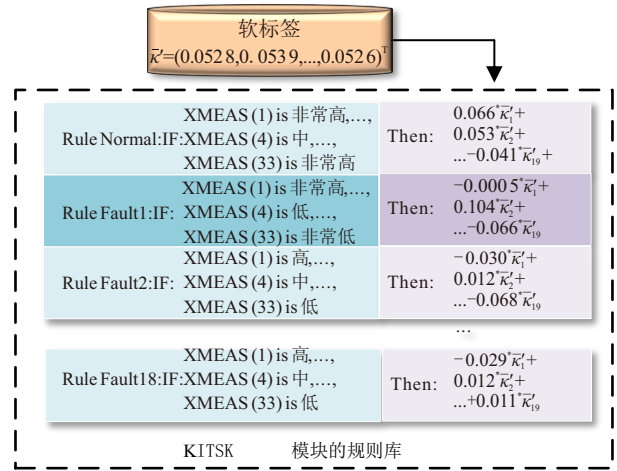


图5 TKI-FNN 模型的语言学可解释性

经过训练的模型预测输入样本  $\boldsymbol{x}$  属于 Fault 1 类, 该故障表示为 A 与 C 原料流量发生变化, 而 B 组分 含量保持不变. 根据预测结果, KITSK 模块的规则库 中对应的 Rule Fault 1 被激活, 表示正常类与 Fault 1 类的规则分别如公式 (28) 和 (29) 所示,

$$\left\{ \begin{array}{l} \text{Rule Normal:} \\ \text{If } x_1 \text{ is 非常高, } x_2 \text{ is 低, } \dots, x_4 \text{ is 中, } \dots, \\ \quad x_{33} \text{ is 非常高,} \\ \text{Then } g_1 = 0.0661\kappa'_1 + 0.0529\kappa'_2 + \dots - \\ \quad - 0.0414\kappa'_{19}, \end{array} \right. \quad (28)$$

$$\left\{ \begin{array}{l} \text{Rule Fault1:} \\ \text{If } x_1 \text{ is 非常高, } x_2 \text{ is 低, } \dots, x_4 \text{ is 低, } \dots, \\ \quad x_{33} \text{ is 非常低,} \\ \text{Then } g_2 = -0.0005\kappa'_1 + 0.1042\kappa'_2 + \dots - \\ \quad 0.0656\kappa'_{19}. \end{array} \right. \quad (29)$$

从定性角度来看,当输入样本的原始特征落在 Rule Fault1 所定义的语义区域内——该区域以  $\varphi_2$  为中 心,并呈现 “非常高  $\times$  低  $\times$  ...  $\times$  非常低” 的特征模式, 模型会将该样本判定为 Fault1 类.更详细地说,根据 公式 (29) 并结合相关输入特征的实际物理含义可以 发现,当 A 进料流量处于较高水平(特征分量  $x_1$ ), 当 D 进料流量处于较低水平(特征分量  $x_2$ ), A 与 C 混合进料流量处于较低水平(特征分量  $x_4$ ), 且搅 拌器转速处于非常低水平(特征分量  $x_{33}$ ) 时,样本  $\boldsymbol{x}$  被判定为 “A 与 C 原料流量发生变化, 而 B 组分含 量保持不变” 故障类别.相较于表示正常类型的公 式 (28), 该故障中 A 与 C 混合进料流量(特征分量  $x_4$ ) 由正常的中等水平下降至较低水平, 呈现出明显 偏离正常工况的变化.因此可以较为直观地判断,该 故障的诱因之一在于 A 与 C 混合进料流量的异常 降低.

从定量角度来看,后件中的软标签系数共同决

定了各个软标签分量对最终故障诊断结果的影响方向和强度, 其数值大小及正负符号体现了贡献的程度. 根据图 5 关于 Rule Fault1 所展示的各后件系数, 软标签分量  $\bar{\kappa}'_2$  (表示对样本属于 Fault1 类的预测概率) 对分类结果具有明显的正向推动作用, 其较大的正系数表明贡献显著; 而  $\bar{\kappa}'_1$  (对应样本属于正常类的概率) 表现出抑制效应, 即其增大会降低样本被判定为 Fault1 类的可能性.

## 4 结论

在本文中, 我们提出了一种全新的故障诊断框架—TKI-FNN 模型. 针对 DL 方法可解释性不足以及 DFNN 方法在串行架构中的局限性, 本研究创新性地从知识蒸馏的独特视角出发, 将 DNN 与 FNN 有机集成. 具体而言, PTC 模块提取的类特定知识在 KD 模块中进一步加工, 生成揭示类间相似性的软标签, 并输入 KITSK 模块以实现语义可解释的故障诊断. 该全新范式通过引入知识蒸馏与软标签正则化机制, 不仅显著提升了故障诊断性能, 同时保持了 FNN 固有的语义可解释性. 此外, 由于知识蒸馏的作用, 算法的计算效率相比传统 DFNN 方法得以大幅提升. 在三个工业受控过程及一个真实柴油机过程上的综合实验结果表明, 所提出的模型在整体性能上显著优于现有最先进的 FNN、DNN 与 DFNN 方法, 实现了当前最先进水平的故障诊断性能和良好的模型可解释性.

尽管该模型在故障诊断任务中展现出显著优势, 但仍需指出, 下游的 KITSK 模块主要学习并继承了上游 PTC 模块的类(目标)相关判别知识. PTC 模块复杂的中间特征表示中仍蕴含着更为细粒度的潜在信息, 这些信息可能对区分部分难以辨识的故障类型起到关键作用, 亟待通过针对性的算法设计进一步挖掘与利用. 因此, 后续研究可主要围绕以下两个方向展开: (1) 从上游 DL 模型中蒸馏更加丰富以及多粒度的知识, 以进一步提升模型性能; (2) 持续拓展模型的应用场景与适用范围.

## 参考文献 (References)

[1] 楼嗣威, 张徐杰, 杨越麟, 等. 高炉炼铁过程故障检测与诊断综述: 回顾, 现状与展望[J]. *自动化学报*, 2025, 51(8): 1739-1759.  
(Lou S W, Zhang X J, Yang Y L, et al. Review of fault detection and diagnosis research in blast furnace ironmaking process: Retrospective, status, and prospects[J]. *Acta Automatica Sinica*, 2025, 51(8): 1739-1759.)

[2] Zhang X, Zhao D X, Ma Z H, et al. Vibration fault detection in high speed trains based on continuous

Twavelet transform and lightweight vision transformer[J]. *Control and Decision*, DOI: 10.13195/j.kzyjc.2025.1071.

[3] 褚菲, 王建文, 马小平. 复杂工业过程安全运行控制方法研究综述: 现状、挑战与展望[J]. *控制与决策*, 2026, 41(6): 1489-1508.  
(Chu F, Wang J W, Ma X P. A review of safety operation control methods for complex industrial processes: Current status, challenges and future prospects[J]. *Control and Decision*, 2026, 41(6): 1489-1508.)

[4] Zhong K, Han B, Han M, et al. Hierarchical graph convolutional networks with latent structure learning for mechanical fault diagnosis[J]. *IEEE/ASME Transactions on Mechatronics*, 2023, 28(6): 3076-3086.

[5] Li Y F, Xu X H, Hu L, et al. A centroid contrastive multi-source domain adaptation method for fault diagnosis with category shift[J]. *Measurement*, 2025, 248: 116801.

[6] U. S. Chemical Safety and Hazard Investigation Board. Incident reports: Events reported to the csb under the accidental release reporting rule[R]. Washington DC, 2025. [Online].

[7] Spagnoletti Law Firm. Explosion at PPG industries plant in spring dale[R]. Pennsylvania, 2025. [Online].

[8] 沈悦, 周豪杰, 沈毅康, 等. 一起船舶舵机故障事故调查引发的思考[J]. *航海*, 2025(3): 45-48.  
(Shen Y, Zhou H J, Shen Y K, et al. Reflections on the investigation of a ship steering gear fault accident[J]. *Navigation*, 2025(3): 45-48.)

[9] 王嘉铭, 蔡浩原, 柳雅倩, 等. 变负载条件下电机故障的 Transformer-DANN 诊断方法研究[J]. *控制与决策*, 2025, 40(10): 3096-3105.  
(Wang J M, Cai H Y, Liu Y Q, et al. Research on motor faults under variable load conditions based on Transformer-DANN model[J]. *Control and Decision*, 2025, 40(10): 3096-3105.)

[10] 周旷, 覃文博, 孙天宇. 基于可信多源领域自适应的跨域滚动轴承故障诊断[J]. *控制与决策*, 2025, 40(7): 2251-2260.  
(Zhou K, Qin W B, Sun T Y. Cross-domain fault diagnosis of rolling bearings based on trusted multi-source domain adaptation[J]. *Control and Decision*, 2025, 40(7): 2251-2260.)

[11] Liu L, Zheng Y, Liang S J. Variable-wise stacked temporal autoencoder for intelligent fault diagnosis of industrial systems[J]. *IEEE Transactions on Industrial Informatics*, 2024, 20(5): 7545-7555.

[12] Li Y S, Zhou Z, Sun C, et al. Variational attention-based interpretable transformer network for rotary machine fault diagnosis[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(5): 6180-6193.

[13] 潘瑞东, 孔维健, 齐洁. 基于预训练模型与知识蒸馏的法律判决预测算法[J]. *控制与决策*, 2022, 37(1): 67-76.  
(Pan R D, Kong W J, Qi J. Legal judgment prediction

- based on pre-training model and knowledge distillation[J]. *Control and Decision*, 2022, 37(1): 67-76.)
- [14] Liu K, Lu N Y, Wu F, et al. Model fusion and multiscale feature learning for fault diagnosis of industrial processes[J]. *IEEE Transactions on Cybernetics*, 2023, 53(10): 6465-6478.
- [15] Chen Y Q, Zhang R D. Deep multiscale convolutional model with multihead self-attention for industrial process fault diagnosis[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025, 55(4): 2503-2512.
- [16] Zang Z S, Yin R, Lu W, et al. A linguistically interpretable deep fuzzy classification system with feature transformation and reconstruction[J]. *IEEE Transactions on Fuzzy Systems*, 2024, 32(8): 4297-4311.
- [17] Li S, Ji J C, Feng K, et al. Composite neuro-fuzzy system-guided cross-modal zero-sample diagnostic framework using multisource heterogeneous noncontact sensing data[J]. *IEEE Transactions on Fuzzy Systems*, 2025, 33(1): 302-313.
- [18] 胡磊, 韩敏. 基于核共轭梯度演化模糊系统的混沌时间序列在线预测[J]. *控制与决策*, 2024, 39(9): 3099-3107.  
(Hu L, Han M. Online prediction of chaotic time series based on kernel conjugate gradient evolving fuzzy system[J]. *Control and Decision*, 2024, 39(9): 3099-3107.)
- [19] Hu H X, Cai Y C, Meng Q, et al. MFVAE: A multiscale fuzzy variational autoencoder for big data-based fault diagnosis in gearbox[J]. *IEEE Transactions on Fuzzy Systems*, 2025, 33(1): 180-191.
- [20] Li M W, Zang Z S, Lu W, et al. Granular computing-based fuzzy deep neural network for long-tailed fault diagnosis: Design and analysis[J]. *Expert Systems with Applications*, 2026, 295: 128806.
- [21] Jhang J Y, Lin C J, Kuo S W. Convolutional Takagi-Sugeno-Kang-type fuzzy neural network for bearing fault diagnosis[J]. *Sensors and Materials*, 2023, 35(7): 2355.
- [22] Zhong H Y, Yu S, Trinh H, et al. A novel small-sample dense teacher assistant knowledge distillation method for bearing fault diagnosis[J]. *IEEE Sensors Journal*, 2023, 23(20): 24279-24291.
- [23] Wang M Y, Yang Y X, Wei L X, et al. A lightweight gear fault diagnosis method based on attention mechanism and multilayer fusion network[J]. *IEEE Transactions on Instrumentation and Measurement*, 2024, 73: 3503011.
- [24] Zhang N, Xu Y, Zhu Q X, et al. Novel regularization double preserving integrated with neighborhood locality projections for fault diagnosis[J]. *IEEE Transactions on Industrial Informatics*, 2023, 19(10): 10478-10488.
- [25] Zhao S Y, Duan Y L, Roy N, et al. A deep learning methodology based on adaptive multiscale CNN and enhanced highway LSTM for industrial process fault diagnosis[J]. *Reliability Engineering & System Safety*, 2024, 249: 110208.
- [26] Zhang L, Xiong G L, Liu H S, et al. Bearing fault diagnosis using multi-scale entropy and adaptive neuro-fuzzy inference[J]. *Expert Systems with Applications*, 2010, 37(8): 6077-6085.
- [27] Cui Y Q, Wu D R, Huang J. Optimize TSK fuzzy systems for classification problems: Minibatch gradient descent with uniform regularization and batch normalization[J]. *IEEE Transactions on Fuzzy Systems*, 2020, 28(12): 3065-3075.
- [28] Zhong K, Han M, Qiu T, et al. Fault diagnosis of complex processes using sparse kernel local fisher discriminant analysis[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(5): 1581-1591.
- [29] Zhang L J, Shi Y, Chang Y C, et al. Robust fuzzy neural network with an adaptive inference engine[J]. *IEEE Transactions on Cybernetics*, 2024, 54(5): 3275-3285.
- [30] Wang Y L, Pan Z F, Yuan X F, et al. A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network[J]. *ISA Transactions*, 2020, 96: 457-467.
- [31] Liu J P, Xu L C, Xie Y F, et al. Toward robust fault identification of complex industrial processes using stacked sparse-denoising autoencoder with softmax classifier[J]. *IEEE Transactions on Cybernetics*, 2023, 53(1): 428-442.
- [32] Juang C F, Cheng Y W, Lin Y M. Visually interpretable fuzzy neural classification network with deep convolutional feature maps[J]. *IEEE Transactions on Fuzzy Systems*, 2024, 32(3): 1063-1077.
- [33] Deng Y, Ren Z Q, Kong Y Y, et al. A hierarchical fused fuzzy deep neural network for data classification[J]. *IEEE Transactions on Fuzzy Systems*, 2017, 25(4): 1006-1012.

## 作者简介

李孟威 (1997-), 男, 博士生, 主要研究方向为深度学习、故障诊断、知识发现和表示等研究, E-mail: [lmw1997614@mail.dlut.edu.cn](mailto:lmw1997614@mail.dlut.edu.cn);

卢伟 (1976-), 男, 教授, 博士, 博士生导师, 从事粒计算、计算智能、知识发现和表示等研究, E-mail: [luwei@dlut.edu.cn](mailto:luwei@dlut.edu.cn).