极端不均衡分布下强化学习驱动的 TBM 主轴承故障辨识方法

张志瑶[1]，于川博[2]，苗栩凯[3]，乐明宇[3]，梅元元[1,4]，张蒙祺[1†]，莫继良[1]

(1. 西南交通大学 机械工程学院，成都 610031；2. 西南交通大学 利兹学院，成都 611756；
3. University of Leeds，School of Mechanical Engineering，West Yorkshire；
4. 中铁工程服务有限公司，成都 610036)

摘　要: 全断面隧道掘进机 (TBM) 主轴承故障辨识直接关系到整机掘进安全与效率, 但工程中高可靠性要求和主动维护策略的实施使监测数据呈现极端类别不均衡 (故障样本⩽1%), 致使稀疏故障特征难以学习, 漏检风险极高. 为此, 本文提出一种深度强化学习驱动的 TBM 主轴承故障辨识模型 (DRLimb), 将传统静态分类问题重构为强化学习的序贯决策优化问题. 该方法首先将故障辨识过程建模为马尔可夫决策过程, 通过双网络架构 (在线 Q 网络与目标 Q 网络) 及软更新机制确保策略学习过程的稳定收敛. 继而, 设计了与决策历史相关的非对称奖励机制, 对故障样本的正确辨识给予更高回报、对漏检施加更强惩罚, 迫使智能体聚焦于稀疏但关键的故障模式, 提升对少数类故障模式的辨识灵敏度. 理论分析证明, 将多数类奖励系数设置为类别不均衡比率, 可实现类间梯度贡献的均衡化. 在多个极端不均衡比率的 TBM 主轴承数据集上的实验表明, DRLimb的 G-mean 值与 F1-Score 均稳定超过 93.2%, 显著优于主流不均衡学习诊断模型与基线模型.

关键词: TBM 主轴承；故障辨识；极端不均衡分布；深度强化学习；非对称奖励函数

中图分类号: TH17；TP277　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1142

引用格式: 张志瑶, 于川博, 苗栩凯, 等. 极端不均衡分布下强化学习驱动的 TBM 主轴承故障辨识方法 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

Main bearing fault identification of TBM driven by deep reinforcement learning under extreme imbalance conditions

ZHANG Zhi-yao[1]，*YU Chuan-bo*[2]，*MIAO Xu-kai*[3]，*YUE Ming-yu*[3]，*MEI Yuan-yuan*[1,4]，*ZHANG Meng-qi*[1†]，*MO Ji-liang*[1]

(1. School of Mechanical Engineering， Southwest Jiaotong University，Chengdu 610031，China；2. SWJTU-Leeds Joint School， Southwest Jiaotong University， Chengdu 611756， China； 3. School of Mechanical Engineering， University of Leeds， West Yorkshire， UK； 4. China Railway Engineering Service Co.， Ltd.，Chengdu 610036，China)

Abstract: Main-bearing fault identification in full-face Tunnel Boring Machines (TBMs) impacts excavation safety and efficiency. In practice, stringent reliability requirements and proactive maintenance yield extremely imbalanced monitoring data (fault samples ⩽1%), resulting in difficult-to-learn sparse fault features and a very high missed-detection risk. To address this issue, we propose DRLimb, a deep reinforcement learning driven TBM main bearing fault identification model that reformulates the traditional static classification problem as a sequential decision optimization problem under a reinforcement learning framework. Specifically, the process is modeled as a Markov decision process; a dual-network architecture (online and target Q-networks) with soft updates is used to ensure stable convergence of the policy learning process. Moreover, we design a decision-history-aware asymmetric reward mechanism that assigns higher rewards to correct fault identifications and stronger penalties to missed detections, forcing the agent to focus on sparse yet critical fault patterns and enhancing the identification sensitivity to minority fault modes. Theory shows that setting the majority-class reward coefficient proportional to the imbalance ratio balances inter-class gradient contributions. Experiments on multiple TBM main-bearing datasets with extreme

收稿日期: 2025-11-03；**录用日期:** 2026-01-29.

基金项目: 国家自然科学基金项目 (52405220, 52475218); 中国博士后科学基金面上项目 (2025M771349); 高端装备机械传动全国重点实验室开放基金项目 (SKLMT-MSKFKT-202530).

†通信作者. E-mail: mzhang@swjtu.edu.cn.

imbalance show that DRLimb achieves G-mean and F1-score above 93.2%, outperforming state-of-the-art imbalance-aware diagnostic models and baselines.

Keywords: TBM main bearing; fault identification; extreme imbalanced distribution; deep reinforcement learning; asymmetric reward function

0 引 言

全断面隧道掘进机 (TBM) 主轴承是 TBM 关键承力部件, 一旦发生故障, 则可能引发整机停工、工程延误、经济损失乃至安全事故, 因此实现精准的 TBM 主轴承故障辨识至关重要[1]. 然而, 实际工程中, 主轴承故障的发生频率极低, 状态监测数据中故障样本稀疏, 占比远低于正常样本. 这种极端不均衡的数据分布特性严重制约了数据驱动故障诊断模型的性能, 致使模型泛化性能严重不足, 对故障的漏检率居高不下[2]. 传统基于监督学习的故障辨识方法在训练阶段依赖于一个共同的假设, 即模型通过最小化一个在全体训练数据上的经验损失函数来寻求最优解. 在极端类别不均衡的数据集上, 这一标准流程将导致损失函数及其梯度被多数类样本所主导, 使得模型参数更新几乎不响应少数类的特征, 其本质是优化目标与期望的均衡分类性能存在偏差[3-4].

为缓解类别不均衡对模型训练的影响, 现有研究主要遵循数据增强[5]、元学习[6-7]、表征学习[8-9]、迁移学习[10] 和代价敏感学习[11-12] 等技术路线. 在 TBM 主轴承故障辨识这类极端不均衡场景下, 这些方法的局限性尤为突出. 数据增强方法 (如生成对抗网络) 在故障样本极其稀少的条件下, 难以学习到真实的故障特征分布, 易产生模式崩溃或生成低质量样本, 反而引入干扰. 重采样技术 (如过采样与欠采样) 在类别比率低于0.01时面临两难: 多数类欠采样会严重损失工况多样性, 而少数类过采样则极易导致模型过拟合. 代价敏感学习通过调整损失权重来平衡类别影响, 但在极端不均衡下, 最优代价矩阵的设定缺乏理论指导, 且对超参数过于敏感, 模型鲁棒性难以保证. 值得注意的是, 现有不均衡故障诊断研究多集中于不均衡比率高于0.02的场景, 对于比率低于0.01的极端情况, 尚缺乏普适而鲁棒的解决方案.

面对 TBM 主轴承数据的极端不均衡问题, 基于监督学习的传统静态分类将样本视为独立同分布, 通过一次前向传播完成决策, 其优化目标难以在多数类的淹没效应下聚焦稀疏故障样本. 相比之下, 深度强化学习 (DRL) 提供了一种基于交互反馈的序贯决策新范式. 该范式将故障辨识建模为马尔可夫决策过程, 通过奖励塑形机制自适应地调整对故障类的关注度[13]. 例如, Wang 等[14] 基于卷积注意力机制构建深度 Q 网络, 提出了一种强噪声与复合故障条件下的滚动轴承故障诊断方法; Li 等[15] 使用改进 DenseNet121 作为智能体的策略和价值网络, 并结合少数类过采样技术与优势演员-评论家框架, 重新设计了强化学习框架中的故障诊断交互过程, 提升了模型在类别不均衡数据中的决策能力; Kang 等[16] 提出了一种双经验池 DRL 模型, 采用平衡交叉采样技术按比例从双经验池中选择样本来训练双残差网络模型, 引导智能体在多数类与少数类间实现更均衡的特征学习. 王辉等[17] 提出了一种基于多尺度深度注意 Q 网络 (MSDAQN) 的 DRL 模型, 通过多尺度卷积神经网络提取多尺度故障特征, 并利用自适应通道注意力进行加权融合突出关键信息. 柏林等[18] 提出了一种域泛化 D3QN 故障诊断方法, 通过在竞争网络和双 Q 网络基础上引入了域识别网络, 从强背景工况噪声中, 分离并甄别出微弱的故障状态信息, 并配合创新的奖励机制设计, 实现了鲁棒的跨工况故障辨识方法. 然而, 这些基于 DRL 的故障辨识方法大多仍需借助外部数据平衡技术来构造相对均衡的训练环境, 未能从 DRL 的序贯决策本质出发, 设计一种完全内生的、由奖励机制直接驱动的样本聚焦策略, 以从根本上应对极端不均衡问题.

本文聚焦解决极端不均衡下 TBM 主轴承故障辨识的两个核心问题: (1) 如何在不依赖外部重采样或复杂代价调整的前提下, 使模型在训练过程中自动且持续地关注稀疏故障样本; (2) 如何设计一种与 DRL 序贯决策特性深度契合的激励机制, 从梯度优化层面提升对少数类样本的灵敏度. 为此, 本文提出一种 DRL 驱动的 TBM 主轴承故障辨识模型 (DRLimb), 将 TBM 主轴承故障辨识问题重构为一个序贯决策过程, 智能体在每个采样时刻根据主轴承状态监测数据输出动作, 以最大化智能体在整个决策轨迹中获得的期望折扣累积回报为目标驱动模型的训练. DRLimb 中引入了一种与决策轨迹深度绑定的非对称奖励机制, 依据动作的历史序列与当前状态对识别故障类样本的行为施加显著更高的正向奖励 (或对漏检施加更严厉的惩罚), 从而在无需额外平衡数据分布的情况下, 引导智能体在策略优化过程中自发地、迭代地聚焦于稀疏故障特征的学习, 在策略优化的梯度层面, 促使智能体提升对稀疏

故障样本的辨识灵敏度.

1　DRLimb 模型

针对 TBM 主轴承故障辨识中的极端样本不均衡问题, 本文将故障辨识问题重构为一个序贯决策过程, 提出一种基于 DRL 的故障辨识框架, 称为 **DRLimb**. 本方法的核心动机在于利用 DRL 的序贯决策特性, 设计一种与决策历史深度绑定的非对称奖励机制, 使智能体在训练过程中能够依据长期回报, 自发、持续地聚焦于稀疏的故障样本, 从而在无需显式重采样或复杂代价矩阵调整的情况下, 从策略优化的源头增强对少数类样本的辨识能力. DRLimb 的核心执行流程如伪代码 Algorithm 1 所示, 其主要符号定义见表 1.

Algorithm 1: DRLimb的交互流程.

输入: 训练数据集$\mathcal{D}$, 经验回放池$\mathcal{M}$, 环境Env

初始化在线Q网络参数θ, 目标Q网络参数$\phi \leftarrow \theta$, 总训练步数$count \leftarrow 0$

当 $count < T_{\max}$执行

　随机打乱$\mathcal{D}$得到序列$\{(x_1, l_1), \ldots, (x_T, l_T)\}$

　初始化状态$s_1 \leftarrow x_1$

　对于$t = 1 \rightarrow T$ **且** $count < T_{max}$**执行**

　　选择动作:

$$a_t = \begin{cases} \text{随机动作} & \text{以概率}\epsilon \\ \arg\max_a Q(s_t, a; \theta) & \text{否则} \end{cases}$$

　　与环境交互:

$$(r_t, Terminal_t) \leftarrow Env.step(x_t, l_t, a_t)$$

　　转移状态$s_{t+1} \leftarrow x_{t+1}$

　　存储经验$(s_t, a_t, r_t, s_{t+1}, Terminal_t)$到$\mathcal{M}$

　　如果 $\mathcal{M}$样本数>= $|\mathcal{B}|$ **则**

　　采样批次$\mathcal{B} \sim \mathcal{M}$

　　对于每个经验$j \in \mathcal{B}$**执行**

　　计算目标值:

$$y_t = r_t + (1 - Terminal_t)\gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a'; \phi)$$

　　结束

　　梯度更新:

$$\theta \leftarrow \theta - \alpha \nabla_\theta \frac{1}{|\mathcal{B}|} \sum_j (y_j - Q(s_j, a_j; \theta))^2$$

$$count \leftarrow count + 1$$

　　结束

　　如果$Terminal_t = 1$**则**

　　提前终止episode

　　结束

　　软更新目标网络: $\phi \leftarrow (1 - \tau)\phi + \tau\theta$

　结束

结束

表1　主要符号及其含义

符号	含义
S	状态空间集合
x_t	t时刻正常/故障样本
s_t	t时刻的状态
A	动作空间集合
a_t	状态s_t时对应的的动作
l_t	状态s_t时对应的标签
R	奖励函数
r_t	状态s_t时对应的奖励值
E	从初始状态到终止状态的完整决策轨迹
T	决策轨迹E的长度
t	对决策轨迹E的数据索引
G_t	状态s_t时的累计折扣回报
y_t	状态s_t时的目标动作价值
$\mathcal{P}$	状态转移分布
λ	针对正常样本的奖励值
π	策略函数
γ	折扣因子
Q	动作价值函数
$\mathbb{E}_\pi$	在策略π下的期望算子
$\mathcal{M}$	经验回放池
ϵ	探索率

DRLimb 的核心由智能体、经验回放池与目标 Q 网络三大模块构成, 其核心创新在于所设计的非对称奖励函数, 该函数根据智能体动作的历史序列与当前状态, 对成功辨识故障样本的行为赋予显著更高的正向奖励, 并对漏检或误检施加差异化惩罚. 此机制从梯度更新层面, 引导智能体的策略网络优先优化对稀疏故障特征的响应, 从而在极端不均衡条件下实现更高的故障召回率与鲁棒性.

1.1　DRLimb 的交互环境定义

DRLimb 智能体与环境的交互基于马尔可夫决策过程建模, 其核心要素定义如下:

(1) 状态空间$\mathcal{S}$: 将每个训练样本x_t视为一个状态s_t. 每轮训练通过随机打乱样本序列初始化状态轨迹, 起始状态s_1对应打乱后的首个样本.

(2) 动作空间$\mathcal{A}$: 动作a_t为样本x_t的预测标签/类别判定$\hat{l}_t$. 在主轴承故障辨识任务中, $\mathcal{A} = \{0, 1\}$, 其中0代表正常样本 (多数类), 1代表故障样本 (少数类).

(3) 奖励函数$\mathcal{R}$: 奖励r_t ($r_t \in [-1, 1]$) 用于评估动作a_t (预测标签) 与真实标签l_t的一致性. 其核心设计原则是对故障类样本的识别结果施加更高的奖惩值, 以引导智能体聚焦于样本稀疏的故障类别, 同时避免智能体策略收敛于局部最优.

(4) 策略π: 策略函数$\pi(a|s_t)$定义了在状态s_t下选择每个可能动作$a_t \in \mathcal{A}$的概率分布, 是智能体的决策核心.

(5) 状态转移$\mathcal{P}$与终止条件: 状态转移是确定性的, 依照预设的样本序列顺序进行. 一个回合 (episode) 对应于一条从s_1到s_T的决策轨迹$E=\{s_1,a_1,r_1,\ldots,s_T,a_T,r_T\}$. 当以下任一条件满足时, 回合终止: 1) 训练集中所有样本处理完毕; 2) 智能体错误识别了任一故障样本.

根据上述定义, 基于 DRLimb 的主轴承故障辨识问题可形式化表述为: 求解最优策略$\pi^*:\mathcal{S}\to\mathcal{A}$, 旨在最大化智能体在整个决策轨迹中获得的期望折扣累积回报.

1.2 DRLimb 主轴承故障辨识模型框架

如图 1 所示, DRLimb 主要由智能体、经验回放池以及目标 Q 网络三大核心模块构成. 其中, (1) 智能体是决策中心, 其核心是一个用于逼近最优动作价值函数$Q^*(s,a)$的在线 Q 网络Q_θ. 该网络接受状态s_t并输出各动作的价值, 智能体基于ϵ-贪婪策略选择动作a_t与环境交互; (2) 经验回放池模块以固定容量存储智能体的交互经验$(s_t,a_t,r_t,s_{t+1},Terminal_t)$. 训练时, 通过从池中均匀采样批次数据 (3) 目标 Q 网络模块Q_ϕ是与在线 Q 网络结构相同但参数更新的滞后副本, 用于计算时序差分学习中的目标 Q 值y_t, 并定期软更新其参数ϕ.

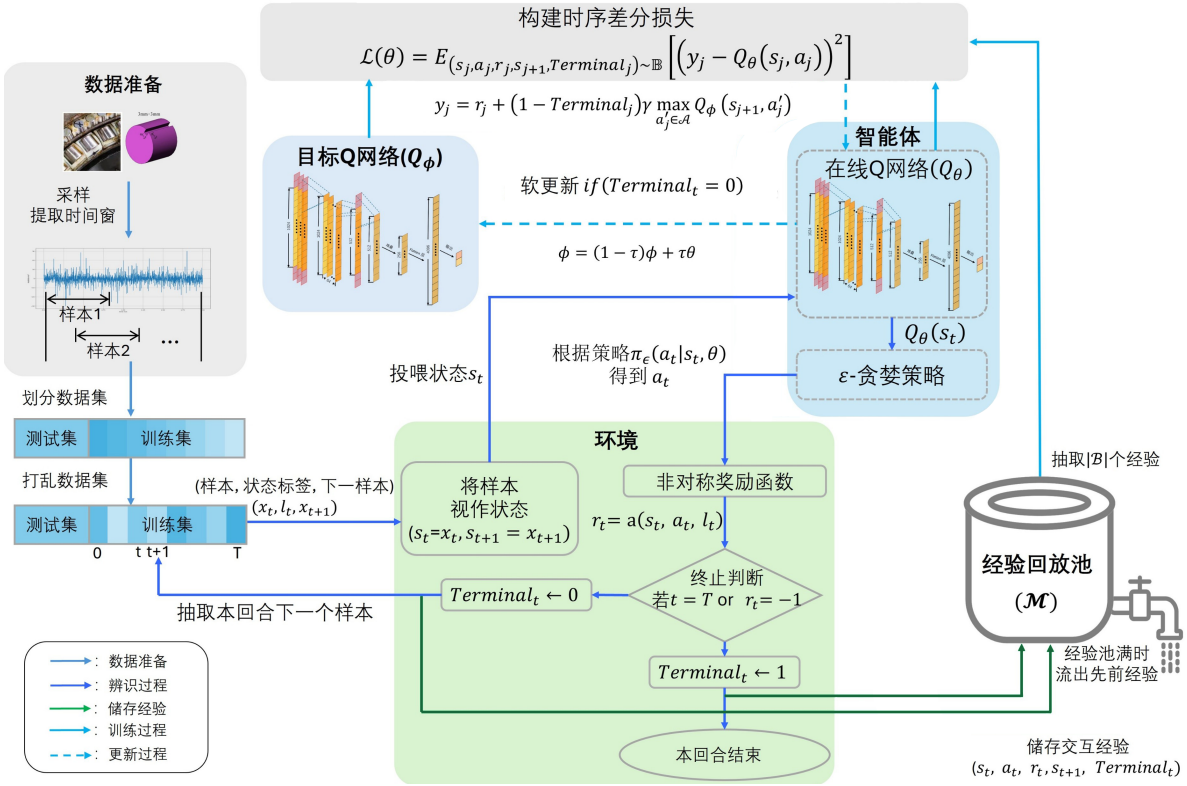

图1 框架与环境交互示意图

1.2.1 智能体模块

智能体是 DRLimb 的决策核心, 通过与环境的交互学习优化故障辨识策略. 如式 (1) 所示, 智能体通过策略π进行决策. 该策略定义了在给定状态$s_t\in\mathcal{S}$时, 智能体选择动作$a_t\in\mathcal{A}$的条件概率分布:

$$\pi(a_t|s_t)=\mathbb{P}(a_t=a\mid s_t=s). \tag{1}$$

为引导智能体在极端不均衡数据中聚焦于主轴承稀疏的故障样本, 引入非对称的奖励函数$R(s_t,a_t,l_t)$:

$$R(s_t,a_t,l_t)=\begin{cases}+1, & a_t=l_t \text{ 且} s_t\in D_P\\ -1, & a_t\neq l_t \text{ 且} s_t\in D_P\\ +\lambda, & a_t=l_t \text{ 且} s_t\in D_N\\ -\lambda, & a_t\neq l_t \text{ 且} s_t\in D_N\end{cases}; \tag{2}$$

其中, $l_t\in 0,1$为样本x_t的真实状态标签 (1表示故障, 0表示正常), D_P与D_N分别为故障与正常样本集, $\lambda\in[0,1]$为调节多数类奖励权重的系数.

为最大化训练样本的识别准确率并增强对稀有故障样本的检测能力, 智能体的优化目标被设定为最大化累积折扣回报G_t, 其定义为:

$$G_t=\sum_{k=t}^{T}\gamma^{k-t}r_k, \tag{3}$$

其中, T为回合终止时刻, γ为折扣因子. 由于正确识别样本可获得正向奖励, 最大化G_t即等价于优化故障识别的整体性能, 促使智能体在决策过程中更加关注故障样本的正确分类.

智能体的优化目标可以等价于优化动作价值函

数Q^π. Q^π评估在状态s_t下执行动作a_t并后续遵循策略π所能获得的期望累积回报:

$$Q^\pi(s_t,a_t)=\mathbb{E}_\pi[G_t\mid s_t,a_t]. \tag{4}$$

Q^π服从贝尔曼方程, 该方程揭示了其跨时间的递归关系:

$$\begin{aligned}Q^\pi(s_t,a_t)=\mathbb{E}_{s_{t+1}\sim\mathcal{P}(s_{t+1}|s_t,a_t)}[r(s_t,a_t)+\\ \gamma\,\mathbb{E}_{a_{t+1}\sim\pi}[Q^\pi(s_{t+1},a_{t+1})]],\end{aligned} \tag{5}$$

其中$r(s_t,a_t)$为状态s_t下执行动作a_t的即时奖励, $\mathcal{P}(s_{t+1}|s_t,a_t)$为状态转移分布, 即从状态$s_t$采取动作$a_t$转移到状态$s_{t+1}$的概率. 该方程揭示了动作价值函数$Q^\pi$的递归分解特性: 当前状态-动作对的评估值等于即时奖励与后继状态折扣期望值的叠加.

通过贝尔曼最优方程求解最优动作价值函数Q^*. Q^*表示在所有策略中能获得的最大期望回报:

$$Q^*(s_t,a_t)=\max_\pi Q^\pi(s_t,a_t). \tag{6}$$

贝尔曼方程最优方程将策略期望$\mathbb{E}_{a_{t+1}\sim\pi}$替换为$\max_{a'}[\cdot]$算子, 直接求解状态-动作对的内在最优价值$Q^*(s_t,a_t)$:

$$\begin{aligned}Q^*(s_t,a_t)=\\ \mathbb{E}_{s_{t+1}\sim\mathcal{P}}[r(s_t,a_t)+\gamma\max_{a'_t\in\mathcal{A}}Q^*(s_{t+1},a'_t)].\end{aligned} \tag{7}$$

基于Q^*的贪婪策略即为最优策略. 但为在训练过程中平衡探索与利用, 采用ϵ-贪婪策略作为行为策略:

$$\pi_\epsilon(a_t|s_t)=\begin{cases}1-\epsilon+\dfrac{\epsilon}{|\mathcal{A}|} & a_t=\arg\max\limits_{a'_t\in\mathcal{A}}Q(s,a'_t)\\ \dfrac{\epsilon}{|\mathcal{A}|} & \text{otherwise}\end{cases}; \tag{8}$$

该策略π_ϵ以概率$1-\epsilon$选择最优动作 (利用), 以概率ϵ均匀随机选择所有可能动作a (探索). ϵ值通常随训练过程衰减, 初期注重探索, 后期偏向利用, 最终收敛于贪婪策略π_{greedy} (即$\epsilon=0$的情形), 如式 (9) 所示:

$$\pi_{greedy}(a_t|s_t)=\begin{cases}1 & a_t=\arg\max\limits_{a'_t\in\mathcal{A}}Q(s_t,a'_t)\\ 0 & \text{otherwise}\end{cases}. \tag{9}$$

为在连续状态空间中逼近最优动作价值函数Q^*, 采用一维卷积神经网络Q_θ作为函数逼近器, 称为在线 Q 网络 (Online Q-Network)Q_θ:

$$Q_\theta(s_t,a_t)\approx Q^*(s_t,a_t), \tag{10}$$

该网络以主轴承状态监测信号x_t, 即s_t, 为输入, 通过多层一维卷积, ReLU 激活函数与池化操作逐步捕获不同时间尺度上的抽象特征表示, 最终输出各动作$a\in\mathcal{A}$的价值估计.

1.2.2　经验回放池模块

经验回放池 (Experience Replay Buffer) $\mathcal{M}$是 DRLimb 框架中的核心存储组件, 旨在通过存储并重复利用历史交互经验, 解决训练过程中因经验样本间时序相关性而导致的训练不稳定性问题, 这对于处理 TBM 主轴承的序贯监测数据尤为重要. 该回放池$\mathcal{M}$的容量上限固定为$|\mathcal{M}|_{max}$, 并遵循先进先出 (First In First Out, FIFO) 原则, 持续存储智能体交互产生的五元组经验数据:

$$\mathcal{M}=\mathcal{M}\cup\{(s_t,a_t,r_t,s_{t+1},Terminal_t)\}, \tag{11}$$

其中, $Terminal_t$为终止标志 (Terminal indicator), 当s_{t+1}为终止状态s_T或智能体收到惩罚时取值为 1, 否则为 0. 该设计确保了在回合终止时不计算后续状态的 Q 值. 该终止机制旨在优先保障对稀疏故障样本的识别灵敏度, 以应对极端类别不均衡的工程安全需求. 在故障样本极少时, 单个回合可能提前终止, 但通过跨回合的经验累积与探索策略 (如 ε-贪婪) 协同, 经验回放池仍能维持多样性, 从而支持稳定学习. 当队列容量$|\mathcal{M}|=|\mathcal{M}|_{max}$时, 最早存入的经验将从经验回放池$\mathcal{M}$中流出/释放.

在线 Q 网络训练时, 通过从$\mathcal{M}$中均匀随机采样批次经验样本数据$\mathcal{B}$更新 Q 网络参数:

$$\mathcal{B}=\{(s_j,a_j,r_j,s_{j+1},Terminal_j)\}_{j=1}^{|\mathcal{B}|}\sim\mathcal{U}(\mathcal{M}), \tag{12}$$

其中$\mathcal{U}$表示均匀随机采样. 该均匀随机采样机制可以打破连续经验样本间的时序相关性, 使得用于梯度更新的训练数据近似满足独立同分布假设, 从而提升极端不均衡数据分布下深度 Q 网络训练的稳定性和样本利用效率.

1.2.3　目标 Q 网络模块

目标 Q 网络模块通过引入一个独立的目标网络 (Target Q-Network)Q_ϕ作为独立函数逼近器, 解决时序差分学习 (Temporal Difference Learning) 中固有的目标值非平稳性问题. 该网络与在线 Q 网络Q_θ结构相同, 但其参数ϕ通过软更新机制 (Soft Update Mchanism) 进行滞后更新, 与Q_θ的参数更新解耦, 以期为训练提供一个稳定的目标值估计. 在达到最终状态s_T之前, Q_ϕ的参数ϕ通过式 (13) 以较小更新系数τ缓慢跟踪在线网络参数θ:

$$\phi=(1-\tau)\phi+\tau\theta,\ \tau\in(0,1). \tag{13}$$

策略网络Q_θ通过最小化如下时序差分损失$\mathcal{L}(\theta)$进行优化, 该损失衡量了$Q_\theta(s_j,a_j)$与由目标网络Q_ϕ计算的目标值y_j之间的差异:

$$\mathcal{L}(\theta)=\mathbb{E}_{(s_j,a_j,r_j,s_{j+1},Terminal_j)\sim\mathcal{B}}[(y_j-Q_\theta(s_j,a_j))^2], \tag{14}$$

其中目标动作价值y_j由贝尔曼方程算出:

$$y_j=r_j+(1-Terminal_j)\gamma\max_{a'_j\in\mathcal{A}}Q_\phi(s_{j+1},a'_j). \tag{15}$$

在线网络参数θ通过梯度下降法进行更新. 通过经验回放采样、目标值计算与参数更新的交替迭代, 目标 Q 网络模块与在线 Q 网络协同工作, 共同确保Q_θ能够稳定且渐进地逼近最优动作价值函数Q^*.

1.3 奖励调节系数最优性分析

在 DRLimb 中, 奖励调节系数λ的取值对平衡两类样本的梯度贡献至关重要. 本节从梯度下降的视角出发, 分析λ的最优取值. 为简化分析, 考虑在线 Q 网络Q_θ的损失函数$\mathcal{L}(\theta)$ (见式 (14)) 的梯度. 在第k次更新时, 梯度可分解为故障类样本和正常类样本的贡献之和:

$$\begin{aligned}\nabla_\theta\mathcal{L}(\theta_k)=&\\ &-2\sum_{m=1}^{P+N=|B|}((1-t_m)\gamma\max_{a'_m}Q(s'_m,a'_m;\theta_{k-1})\\ &-Q(s_m,a_m;\theta_k))\nabla_{\theta_k}Q(s_m,a_m;\theta_k)\\ &\underbrace{-2\sum_{i=1}^{P}(-1)^{1-I(a_i=l_i)}\nabla_{\theta_k}Q(s_i,a_i;\theta_k)}_{\text{故障类贡献,}P\text{项}}\\ &\underbrace{-2\lambda\sum_{j=1}^{N}(-1)^{1-I(a_j=l_j)}\nabla_{\theta_k}Q(s_j,a_j;\theta_k),}_{\text{正常类贡献,}N\text{项}}\end{aligned} \tag{16}$$

其中, P和N分别表示当前批次$\mathcal{B}$中故障类和正常类样本的数量, 且$|\mathcal{B}|=P+N$. 第二项 (P项) 对应故障样本的梯度贡献, 第三项 (N项) 对应正常样本的梯度贡献. 当$\lambda=1$时, 正常样本因数量优势 ($N\gg P$) 导致第三项的绝对值显著大于第二项. 时序差分误差$(y_m-Q(s_m,a_m;\theta_k))$对于同类样本具有相似的统计特性, 并可近似为常数. 同时, 关注梯度方向的主要差异来源于奖励函数的设计. 故障类样本和正常类样本的奖励绝对值分别为 1 和λ. 因此, 梯度贡献的幅度大致与奖励绝对值成比例. 于是, 为使两类样本的梯度贡献均衡, 需满足P项与N项梯度值相等, 即:

$$|\sum_{i=1}^{P}\nabla_{\theta_k}Q(s_i,a_i;\theta_k)|=\lambda|\sum_{j=1}^{N}\nabla_{\theta_k}Q(s_j,a_j;\theta_k)|, \tag{17}$$

进一步, 假设对于同类样本, 其梯度期望值相近, 即$\mathbb{E}[\nabla_{\theta_k}Q(s_i,a_i;\theta_k)]\approx\mathbb{E}[\nabla_{\theta_k}Q(s_j,a_j;\theta_k)]$, 且时序差分误差的期望也相近, 则上式可简化为:

$$P\cdot c\approx\lambda\cdot N\cdot c, \tag{18}$$

其中c为常数. 则上式可以转换为:

$$\lambda\approx\frac{P}{N}. \tag{19}$$

基于上述梯度均衡分析, 为确保在整个训练过程中 (而不仅是单个批次$\mathcal{B}$内) 多数类与少数类样本的梯度贡献保持平衡, 本文将奖励调节系数λ设置为整个训练集的不均衡比率ρ, 即:

$$\lambda=\frac{|D_P|}{|D_N|}=\rho. \tag{20}$$

由此, 为 DRLimb 模型得到一个关键的理论设计准则: 当奖励调节系数λ等于训练数据的类别不均衡比率ρ时, 正常类与故障类样本对模型参数更新的期望梯度贡献在理论上达到均衡. 该准则确保了模型不会在训练过程中被多数类样本主导, 为提升主轴承稀疏故障样本的辨识能力提供了理论保障.

2 试验分析

2.1 试验设置与数据描述

为验证所提模型的有效性, 本研究基于 TBM 主轴承损伤模拟试验台 (如图 2 所示) 采集的振动数据集进行分析. 为模拟真实工业场景中噪声与非平稳性共存的挑战, 在训练中主动向数据注入背景噪声, 以检验模型在此复合条件下的鲁棒性. 为模拟实际工况中的极端不均衡场景, 构建了六种不同不均衡比率ρ的实验 (Case A ~ F), 其中故障样本数量随ρ变化 (40~80 个), 正常样本固定为 8 000 个, 故障样本规模如表 2 所示. 采用均衡测试集评估模型效果, 各类样本均为 4 000 个.

表3 六个 Case 对应的故障样本数目情况

Case	不均衡率	故障样本	Case	不均衡率	故障样本
Case A	0.010	80	Case B	0.009	72
Case C	0.008	64	Case D	0.007	56

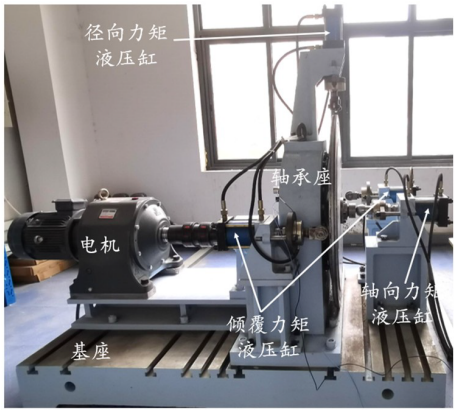

图2 TBM 主轴承损伤模拟试验台示意图

Case E	0.006	48	Case F	0.005	40

所有试验在搭载 NVIDIA V100-SXM2-32GB 显卡的计算服务器上开展. DRLimb 的双 Q 网络采用四层一维卷积架构, 每层均由卷积 (核 5/填充 2)、ReLU 及 2 倍下采样池化构成, 通道数由 32 增至 64; 特征图经展平后输入至含 256 维隐藏层的全连接层输出 Q 值. 探索率采用线性衰减策略 $\epsilon_t = \epsilon_0 - t \cdot (\epsilon_0 - \epsilon_{\min})/T_{\max}$, 其中 t 为当前训练步数, $T_{\max}$ 为最大训练步数, 该策略确保探索率从初始值 $\epsilon_0 = 1.0$ 逐步降至 $\epsilon_{\min} = 0.01$, 在训练前期促进充分探索, 后期侧重策略利用. DRLimb 的其他关键参数配置如表 3 所示. 此外, 为进一步提升透明性与可复现性, 已将 DRLimb 的实现代码公开发布: https://github.com/Leeds-Anomalymous/DQNimb.git

为客观评估模型在极端不均衡数据上的性能, 选取 G-mean (式 21) 与 F1 Score (式 22) 作为评价指标, 两者值越大, 表明算法性能越优. 所有试验均重复 10 次, 结果以"指标均值±标准差 (Std)"报告.

$$\mathrm{G-mean} = \sqrt{\mathrm{Recall} \times \mathrm{Specificity}}, \tag{21}$$

$$\mathrm{F1\ Score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{22}$$

2.2　结果分析

2.2.1　基于 DRLimb 的 TBM 主轴承故障辨识结果

图 3 和表 4 分别展示了六个 Case 的混淆矩阵和性能指标. 分析可知: (1) 在极端不均衡场景, DRLimb 模型各项指标保持相对稳定, G-mean 均高于 93.20%, 表明模型具有优异的泛化能力. 并且模型对稀疏的少数类 (故障样本) 的识别率稳定在 88.7%~91.8%. 这表明模型在确保多数类精度不受损的前提下, 显著提升了对 TBM 主轴承故障样本的辨识能力. (2) 随着 ρ 的降低 (即可供学习的故障样本数量减少), 模型对故障类的识别准确率并未出现显著衰减. 这验证了 DRLimb 框架在应对极端不均衡问题上的固有优势: 其通过非对称奖励函数对少数类施加持续、强劲的学习信号, 迫使智能体在样本极度稀缺的条件下依然能自主学习到鲁棒的故障特征, 而非简单地依赖数据量. 此外, 对比少数类 (故障), 多数类 (正常) 中存在的部分误判, 可归因于奖励函数为极力提升故障召回率而固有的"宁错判, 不漏判"的设计倾向, 这也符合工程实际的需要.

表3　DRLimb 关键参数配置

参数类型	参数名称	符号	取值
训练参数	批量大小	$\|B\|$	64
	最大训练步数	$T_{\max}$	120, 000
	学习率	α	0.000 25
	目标Q网络更新率	τ	0.05
探索策略	初始探索率	ϵ_{init}	1.0
	最小探索率	$\epsilon_{\min}$	0.01
	衰减方式	-	线性衰减
经验回放	缓冲区最大容量	$\|M\|_{max}$	50, 000
奖励函数	折扣因子	γ	0.1
	多数类奖励系数	λ	$\lambda = \rho$

表4　DRLimb 在六个极端不均衡场景的故障辨识结果

	G-mean	F1 Score

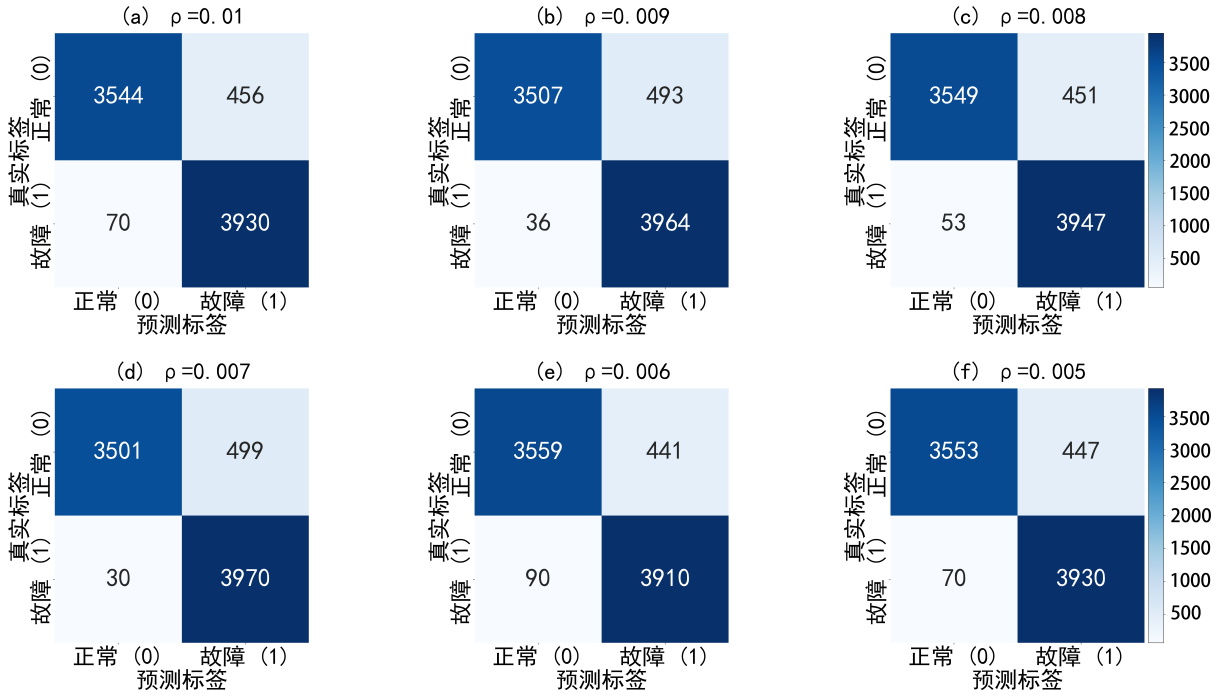

图3　**DRLimb** 在六个极端不均衡场景的故障辨识混淆矩阵

试验		
Case A ($\rho=0.01$)	0.9330 ± 0.0022	0.9343 ± 0.0021
Case B ($\rho=0.009$)	0.9329 ± 0.0011	0.9341 ± 0.0009
Case C ($\rho=0.008$)	0.9354 ± 0.0051	0.9366 ± 0.0046
Case D ($\rho=0.007$)	0.9320 ± 0.0116	0.9336 ± 0.0121
Case E ($\rho=0.006$)	0.9327 ± 0.0021	0.9338 ± 0.0019
Case F ($\rho=0.005$)	0.9342 ± 0.0014	0.9352 ± 0.0013

2.2.2 对比试验结果

选取了针对不均衡问题的八类故障辨识方法: 基于元学习自适应的方法 MW-Net[19], 基于特征约束机制的方法 Novel-CNN[4], 基于显式加权损失函数的方法 Focal-loss-CNN[20] 与改进的动态标签分布感知边界正则化方法 DLMR-CNN[10], 基于固定预设规则的方法 Normalized-CNN[3], 基于过采样的方法 OREM[21], 基于表征学习的 BBN[8] 和基于元学习平衡策略的 BALMS[7]; 以及三类基线模型 (ResNet32* (经典 ResNet32 的一维卷积形式)、BiLSTM 和 Transformer) 做对比分析, 结果如表 5、表 6 以及图 4 所示. 分析图 4 可知, 随着ρ的降低 (即故障样本数量减少), 三个基线模型的 G-mean 出现显著波动, 尤其在$\rho=0.006$时, BiLSTM 和 Transformer 模型波动剧烈 (±17%). 该现象表明, 传统架构在极端不均衡数据分布条件 TBM 主轴承故障辨识任务中难以保持稳定性能. 相比之下, DRLimb 性能最优且稳定.

分析表 5 和 表 6 可知, (1) 在所有不均衡比率的 Case 里, DRLimb 在 G-mean 与 F1 Score 两项指标上均稳定保持最优 (高于93.2%). (2) OREM 方法通过识别干净子区域生成样本, 但在极端不均衡条件 (Case E ($\rho=0.006$)) 下候选区域估计准确性下降, 性能受限 (G-mean 为 86.55%). MW-Net 借助元学习自适应调整样本权重, 但其权重网络对元训练数据质量敏感, 在 Case F ($\rho=0.005$) 时性能下降 (G-mean 为 78.49%). 尽管 Novel-CNN 与 Normalized-CNN 均针对滚动轴承故障辨识设计了特征约束机制, 但在极端不均衡数据分布下的 TBM 主轴承故障辨识任务中仍表现不佳. Focal-loss-CNN 与 DLMR-CNN 分别通过损失函数中的样本加权及分类边界重校准策略进行了改进. 然而, Focal-loss-CNN 在极

表5 Case A ~ Case C 中 DRLimb 与其他方法的对比

方法	Case A (ρ=0.01)		Case B (ρ=0.009)		Case C (ρ=0.008)	
	G-mean	F1-score	G-mean	F1-score	G-mean	F1-score
ResNet32*	0.6920±0.1482	0.7311±0.1386	0.6965±0.2037	0.7297±0.1588	0.7100±0.1708	0.7391±0.1373
BiLSTM	0.7137±0.1484	0.7386±0.1305	0.8118±0.0402	0.7912±0.1373	0.7730±0.0680	0.7900±0.0739
Transformer	0.8743±0.0330	0.8832±0.0592	0.8677±0.0559	0.8721±0.0548	0.8750±0.0166	0.8829±0.0245
MW-Net	0.8868±0.0464	0.8919±0.0442	0.8804±0.0319	0.8861±0.0308	0.8686±0.0486	0.8755±0.0461
Novel-CNN	0.9235±0.0085	0.9255±0.0081	0.9203±0.0145	0.9225±0.0142	0.9103±0.0094	0.9134±0.0092
Normalized-CNN	0.7972±0.2804	0.8353±0.1768	0.7918±0.2797	0.8304±0.1766	0.7133±0.3767	0.7839±0.2385
Focal-loss-CNN	0.8252±0.1787	0.8425±0.1457	0.7939±0.1851	0.8150±0.1530	0.6563±0.2038	0.6986±0.1659
OREM	0.8935±0.0401	0.8971±0.0384	0.8858±0.0224	0.8895±0.0214	0.8872±0.0314	0.8909±0.0306
DLMR-CNN	0.9254±0.0045	0.9275±0.0042	0.9289±0.0025	0.9308±0.0024	0.9261±0.0050	0.9281±0.0045
BBN	0.9266±0.0107	0.9249±0.0114	0.9277±0.0099	0.9257±0.0105	0.9144±0.0217	0.9117±0.0234
BALMS	0.9287±0.0044	0.9265±0.0047	0.9353±0.0008	0.9336±0.0009	0.9327±0.0055	0.9310±0.0061
DRLimb	0.9330±0.0022	0.9343±0.0021	0.9329±0.0011	0.9341±0.0009	0.9354±0.0051	0.9366±0.0046

表6 Case D ~ Case F 中 DRLimb 与其他方法的对比

方法	Case D (ρ=0.007)		Case E (ρ=0.006)		Case F (ρ=0.005)	
	G-mean	F1-score	G-mean	F1-score	G-mean	F1-score
ResNet32*	0.6755±0.2119	0.6999±0.1994	0.5454±0.1784	0.6091±0.1632	0.3796±0.1050	0.5033±0.1325
BiLSTM	0.6812±0.0926	0.6976±0.1269	0.5101±0.3153	0.6074±0.2080	0.5666±0.1804	0.6279±0.1658
Transformer	0.8143±0.0854	0.8020±0.1210	0.7735±0.1766	0.7573±0.2104	0.8280±0.0411	0.8097±0.1319
MW-Net	0.8463±0.0510	0.8555±0.0483	0.8589±0.0596	0.8670±0.0562	0.7849±0.0647	0.8012±0.0600
Novel-CNN	0.8847±0.0219	0.8899±0.0210	0.8339±0.0427	0.8442±0.0402	0.8711±0.0610	0.8778±0.0572
Normalized-CNN	0.6764±0.3620	0.7507±0.2269	0.5272±0.4540	0.6641±0.2850	0.5254±0.3939	0.6439±0.2470
Focal-loss-CNN	0.6310±0.2137	0.6789±0.1724	0.4557±0.4028	0.6022±0.2586	0.5104±0.2272	0.5846±0.1835
OREM	0.9095±0.0253	0.9115±0.0244	0.8655±0.0531	0.8715±0.0508	0.8712±0.0407	0.8759±0.0394
DLMR-CNN	0.9194±0.0113	0.9217±0.0103	0.8801±0.0287	0.8859±0.0262	0.9080±0.0054	0.9114±0.0051
BBN	0.9224±0.0139	0.9202±0.0148	0.9243±0.0202	0.9229±0.0215	0.9064±0.0145	0.9028±0.0155
BALMS	0.9309±0.0009	0.9295±0.0009	0.9245±0.0186	0.9222±0.0192	0.9182±0.0122	0.9152±0.0132
DRLimb	0.9320±0.0116	0.9336±0.0121	0.9327±0.0021	0.9338±0.0019	0.9342±0.0014	0.9352±0.0013

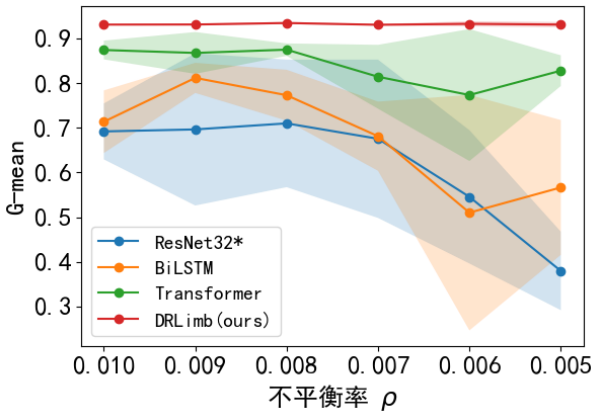

图4　DRLimb 和基线模型的故障辨识结果

度不均衡条件 (Case E, $\rho = 0.006$) 下表现出鲁棒性不足, G-mean 仅为 0.4557.DLMR-CNN 虽然在各测试情景下整体表现稳健, 但在该极端不均衡条件下性能亦出现轻微回落, 显示出其对数据分布仍具有一定的敏感性. (3) DRLimb 表现出最优的鲁棒性, 相比之下, 现有不均衡处理方法 (如 OREM、MW-Net) 的标准差普遍在 2%~6%, 且对比优秀的基线模型 (如 BBN、BALMS), DRLimb 具有更加卓越的性能. 而传统模型 (如 BiLSTM、Normalized-CNN) 在极端不均衡场景下指标值波动剧烈, 标准差超过 15%, 揭示了 DRLimb 通过交互式学习与动态策略优化, 从根本上增强了对数据分布剧变的适应能力, 而非依赖于预设的、在极端条件下易失效的启发式规则.

2.2.3　参数敏感性分析

为探究 DRLimb 模型中关键参数对性能的影响, 对 Q 网络卷积层数Num, 正常样本奖励值λ以及折扣因子γ进行敏感性分析.

Q 网络卷积层数 (Num): 如图 5 所示, 卷积层数显著影响模型性能. 层数过少 (1~3 层) 导致感受野有限, 特征提取不充分; 4~5 层时可均衡局部与全局特征, G-mean 达到最优; 进一步增加层数引发梯度消失与过拟合, 性能下降. 基于此, 设定卷积层数为 4, 在模型性能与训练效率间取得均衡.

正常样本奖励值 (λ): λ是均衡健康类别 (多数类) 和故障类别 (少数类) 梯度贡献的关键参数. 如图 6 所示, λ远小于ρ时 G-mean 较低; λ接近ρ时 G-mean 迅速达到峰值, 验证了当$\lambda = \rho$时梯度均衡的最优条件 (式 20); 而λ过大 (如超过10ρ) 时 G-mean 下降, 表明过高奖励会使优化过度偏向正常样本.

折扣因子 (γ): γ衡量智能体对远期奖励的重视程度. 如图 7 所示, γ增大时 G-mean 下降且标准差扩大. 结合图 8, $\gamma = 0.9$时损失值剧烈波动, 收敛困难. 这表明近期奖励包含更直接有效的决策信息, 过

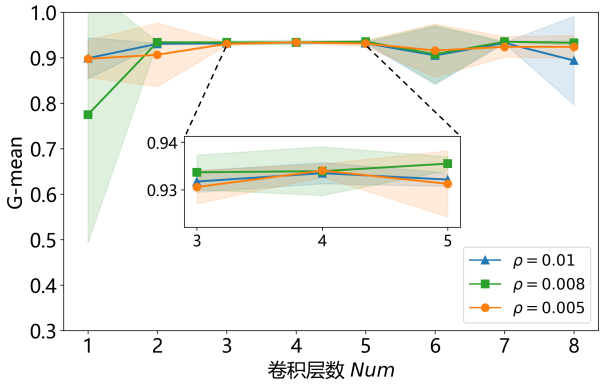

图5　Q 网络卷积层数 (Num) 参数敏感性分析

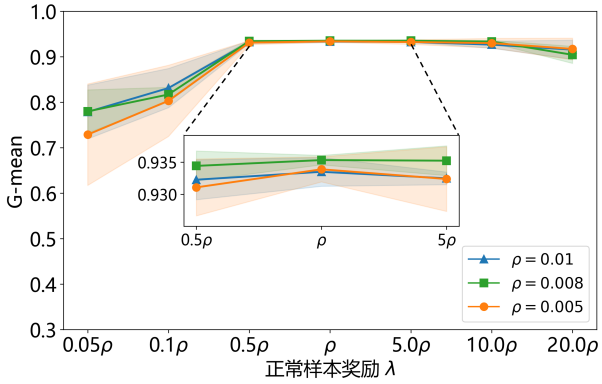

图6　正常样本奖励 (λ) 参数敏感性分析

高γ会引入累积误差和训练不稳定. 但γ过小会导致智能体过于短视, 可能无法学习考虑长期效益的最优策略. 因此选取$\gamma = 0.1$, 避免训练振荡与估计偏差.

2.2.4　跨工况泛化性能验证

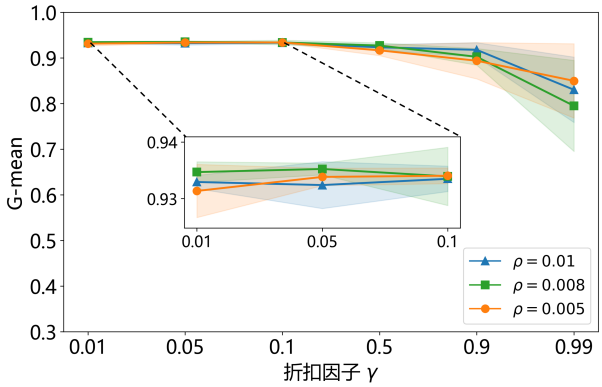

图7　折扣因子 (γ) 参数敏感性分析

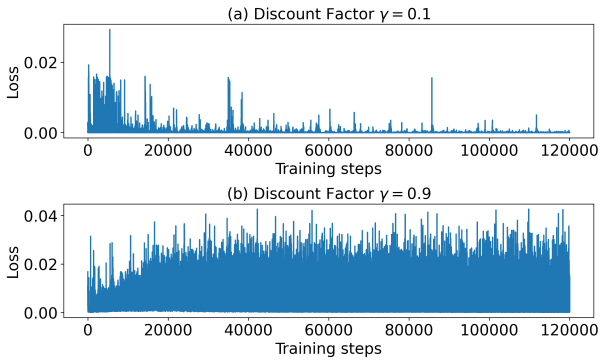

图8　折扣因子 (γ) 为 0.1 和 0.9 下的 Q 网络损失图

为验证所提方法在未见工况下的泛化能力, 基于 Case A(ρ=0.01) 设计了面向轴向载荷和转速的跨工况实验, 如表 7 所示.

表7 跨工况实验设置

实验	训练工况	测试工况
CrossKN	10 kN, 20 kN	30 kN
CrossKN-Reversed	30 kN	10 kN, 20 kN
CrossR	1 rpm, 2 rpm	3 rpm
CrossR-Reversed	3 rpm	1 rpm, 2 rpm

图 9 展示了 DRLimb 的跨工况实验结果. 在轴向载荷维度, 当训练工况从低负荷迁移到高负荷 (CrossKN 到 CrossKN-Reversed), 宏平均 F1-score 的中位数从 0.925 提升至 0.930, 表明模型在高负荷训练后能有效泛化到低负荷工况. 在转速维度, 从低转速迁移到高转速 (CrossR 到 CrossR-Reversed), 宏平均 F1-score 的中位数从 0.925 提升至 0.935, 显示对转速变化的良好泛化能力. 同时, G-mean 在各项实验中保持稳定分布, 进一步验证了模型在未见工况下的泛化性能.

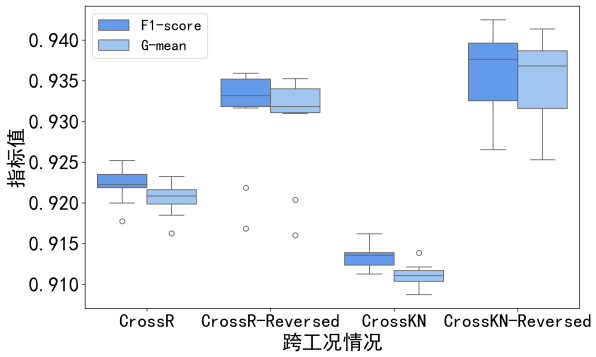

图9 DRLimb 模型的跨工况实验结果

3 结 论

本文针对 TBM 主轴承故障辨识中类别极端不均衡这一核心挑战, 提出了一种基于深度强化学习的 DRLimb 模型. 通过将故障辨识重构为序贯决策过程, 构建在线 Q 网络与目标 Q 网络协同的双网络架构, 并基于理论证明的最优条件, 即将奖励调节系数设为类别不均衡比率, 有效均衡了类间梯度贡献. 实验表明, 该方法在各类极端不均衡场景下 G-mean 均高于 93.2%, 显著优于传统监督学习方法, 验证了深度强化学习通过交互决策与非对称奖励机制自适应调整分类边界, 在解决极端不均衡故障辨识任务中的巨大潜力, 为 TBM 主轴承故障辨识提供了新的技术途径.

参考文献 (References)

[1] Zheng J Y, Hu S, Ji J C, et al. A review of fatigue failure and structural design of main bearings in tunnel boring machines based on engineering practical examples[J]. Engineering Failure Analysis, 2024, 163: 108611.

[2] Fu X C, Tao J F, Qin C J, et al. A roller state-based fault diagnosis method for tunnel boring machine main bearing using two-stream CNN with multichannel detrending inputs[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 3527812.

[3] Zhao B, Zhang X M, Li H, et al. Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions[J]. Knowledge-Based Systems, 2020, 199: 105971.

[4] Xing Z Y, Zhao R Z, Wu Y C, et al. Intelligent fault diagnosis of rolling bearing based on novel CNN model considering data imbalance[J]. Applied Intelligence, 2022, 52(14): 16281-16293.

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

[6] Ye X H, Zhang X M, He B C, et al. Rolling bearing fault diagnosis with variable load and few samples based on multifeature fusion meta-learning[C]. 2023 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes. Yibin, 2023: 1-6.

[7] Ren J W, Yu C J, Sheng S N, et al. Balanced meta-softmax for long-tailed visual recognition[J/OL]. 2020, arXiv: 2007.10740.

[8] Zhou B Y, Cui Q, Wei X S, et al. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 9716-9725.

[9] 李康, 李爽, 高小永, 等. 多变量时序标记 Transformer 及其在电潜泵故障诊断中的应用[J]. 控制与决策, 2025, 40(4): 1145-1153.
(Li K, Li S, Gao X Y, et al. Multivariate time series tokenized Transformer and its application in fault diagnosis of electric submersible pump[J]. Control and Decision, 2025, 40(4): 1145-1153.)

[10] 陈洋洋, 周谧, 张永斌. 基于多粒度对齐和证据推理的多源域自适应故障诊断方法[J]. 控制与决策, 2025, 40(11): 3403-3414.
(Chen Y Y, Zhou M, Zhang Y B. Multi-source domain adaptation fault diagnosis method based on multigranularity alignment and evidential reasoning[J]. Control and Decision, 2025, 40(11): 3403-3414.)

[11] He C B, Fu Z Y, Chen P, et al. A confident cross-domain mixup–based network with dynamic label-distribution-aware margin regularization for bearing fault diagnosis under variable working conditions[J]. Engineering

Applications of Artificial Intelligence, 2026, 163: 112964.

[12] 石佳, 郭鹏, 张志瑶, 等. 融合多传感器时空特征演化与头尾部梯度竞争均衡的电机长尾数据故障诊断[J]. 控制与决策, 2026, 41(2): 393-404.
(Shi J, Guo P, Zhang Z Y, et al. A fault diagnosis method for long-tailed motor data integrating multi-sensor spatiotemporal feature evolution and head-tail gradient competition equilibrium[J]. Control and Decision, 2026, 41(2): 393-404.)

[13] Ding Y, Ma L, Ma J, et al. Intelligent fault diagnosis for rotating machinery using deep *Q*-network based health state classification: A deep reinforcement learning approach[J]. Advanced Engineering Informatics, 2019, 42: 100977.

[14] Wang R X, Jiang H K, Zhu K, et al. A deep feature enhanced reinforcement learning method for rolling bearing fault diagnosis[J]. Advanced Engineering Informatics, 2022, 54: 101750.

[15] Li Y H, Wang Y P, Zhao X, et al. A deep reinforcement learning-based intelligent fault diagnosis framework for rolling bearings under imbalanced datasets[J]. Control Engineering Practice, 2024, 145: 105845.

[16] Kang Y X, Chen G, Pan W P, et al. A dual-experience pool deep reinforcement learning method and its application in fault diagnosis of rolling bearing with unbalanced data[J]. Journal of Mechanical Science and Technology, 2023, 37(6): 2715-2726.

[17] 王辉, 徐佳文, 严如强. 基于多尺度注意力深度强化学习网络的行星齿轮箱智能诊断方法[J]. 机械工程学报, 2022, 58(11): 133-142.
(Wang H, Xu J W, Yan R Q. Multi-scale attention based deep reinforcement learning for intelligent fault diagnosis of planetary gearbox[J]. Journal of Mechanical Engineering, 2022, 58(11): 133-142.)

[18] 柏林, 何牧耕, 陈兵奎, 等. 基于域泛化 D3QN 的跨工况故障诊断方法[J]. 机械工程学报, 2024, 60(22): 165-178.
(Bo Lin, He M G, Chen B K, et al. Domain generalization D3QN for machinery fault diagnosis across different working conditions[J]. Journal of Mechanical Engineering, 2024, 60(22): 165-178.)

[19] Shu J, Xie Q, Yi L X, et al. Meta-weight-net: Learning an explicit mapping for sample weighting[C]. Neural Information Processing Systems. Vancouver, 2019: 1-12.

[20] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 2999-3007.

[21] Zhu T F, Liu X W, Zhu E. Oversampling with reliably expanding minority class regions for imbalanced data learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(6): 6167-6181.

作者简介

张志瑶 (1993−), 女, 助理教授, 博士, 主要研究方向为智能装备运维、设备状态监测、故障诊断, E-mail: zhiyaozhang@swjtu.edu.cn;

于川博 (2004−), 男, 本科生, 主要研究方向为强化学习与深度学习在工业制造系统优化中的理论与应用, E-mail: mn23c2y@leeds.ac.uk;

苗栩凯 (2004−), 男, 本科生, 主要研究方向为深度学习与信号处理, E-mail: mn23xm@leeds.ac.uk;

乐明宇 (2004−), 男, 本科生, 主要研究方向为强化学习与机械运动控制方法, E-mail: hlld8753@leeds.ac.uk;

梅元元 (1988−), 男, 高级工程师, 主要研究方向为盾构机再制造技术、智能化与数字化技术、隧道施工技术, E-mail: 418473556@qq.com;

张蒙祺 (1989−), 男, 副研究员, 博士, 博士生导师, 主要研究方向为摩擦学、接触力学、数值计算方法、机械破岩理论与应用, E-mail: mzhang@swjtu.edu.cn;

莫继良 (1982−), 男, 研究员, 博士, 博士生导师, 主要研究方向为摩擦学与动力学行为分析、界面/结构损伤失效评估与优化设计、振动与噪声控制、故障诊断与智能化等, E-mail: jlmo@swjtu.cn.