

基于图深度强化学习的电力网络通信自趋优控制方法

江颖洁, 魏永静, 孙超, 田安琪, 张彦, 朱尤祥

(国网山东省电力公司信息通信公司, 山东 济南 250002)

摘要: 在智能电网的发展框架下, 电力通信网络面临提升其运行实时性、确保传输可靠性和实现管理智能化的重大挑战。针对智能电网背景下电力通信网络面临的实时性、可靠性和智能化挑战, 提出一种融合风险量化与图深度强化学习(RQGRL)的自趋优控制策略。该策略通过整合图卷积神经网络(GCN)与深度Q学习算法(DQN), 旨在提升网络管理智能化水平, 尤其是在资源调度与路径规划中加强了对风险因素的考虑。具体而言, 模型将GCN嵌入深度强化学习(DRL)框架, 实时监测网络状态并作出智能决策, 强化网络韧性, 应对突发通信故障以提供最佳路由备选方案。通过量化风险, 并将其作为网络图中边的权重, 结合加权GCN在DRL框架中的应用, 模型能动态优化路径选择和资源分配, 以适应复杂多变的电网通信需求。仿真结果显示, 相比最先进的DRL解决方案, 所提出的RQGRL模型在带宽分配效率上提升了6.6%, 并在链路故障恢复方面展现出卓越的能力, 即使在高并发故障下仍能维持网络高效运行。

关键词: 深度强化学习; 图神经网络; 深度Q学习算法; 路径推荐

DOI: 10.11907/rjtk.251271

扫描二维码阅读全文:



中图分类号: G642.0

文献标识码: A

文章编号: 1672-7800(2025)010-0056-09

A Self-trending Optimization Control Method for Power Network Communication Based on Graph Deep Reinforcement Learning

JIANG Yingjie, WEI Yongjing, SUN Chao, TIAN Anqi, ZHANG Yan, ZHU Youxiang

(State Grid Shandong Electric Power Company Information and Communication Company, Ji'nan 250002, China)

Abstract: Under the framework of smart grid development, electric power communication networks face challenges in real-time operations, transmission reliability, and intelligent management. This paper proposes a self-convergent optimization control strategy, RiskQuant-GRL (RQGRL), which integrates risk quantization and graph deep reinforcement learning. By combining Graph Convolutional Networks (GCN) with Deep Q-Learning (DQN), RQGRL aims to enhance network intelligence in resource scheduling and path planning. Specifically, it embeds a GCN into a Deep Reinforcement Learning (DRL) framework to monitor the network state, make intelligent decisions, and provide optimal routing. By quantifying risk as edge weights and using weighted GCN, the model dynamically optimizes path selection and resource allocation. Simulation results show that RQGRL improves bandwidth allocation efficiency by 6.6% over state-of-the-art DRL solutions and excels in link failure recovery, maintaining efficient network operation even under high concurrency.

Key Words: deep reinforcement learning; graph neural network; deep Q-learning algorithms; path recommendation

收稿日期: 2025-05-21

基金项目: 国网山东省电力公司科技项目(520627240007)

作者简介: 江颖洁(1989-), 女, 国网山东省电力公司信息通信公司高级工程师, 研究方向为信号与信息处理; 魏永静(1981-), 女, 国网山东省电力公司信息通信公司高级工程师, 研究方向为电力系统及其自动化; 孙超(1993-), 男, 国网山东省电力公司信息通信公司工程师, 研究方向为电子与通信工程; 田安琪(1992-), 女, 国网山东省电力公司信息通信公司工程师, 研究方向为信息与通信工程; 张彦(1988-), 男, 国网山东省电力公司信息通信公司工程师, 研究方向为通信工程; 朱尤祥(1990-), 男, 国网山东省电力公司信息通信公司工程师, 研究方向为电子与通信工程。本文通讯作者: 孙超。

0 引言

随着能源数字化的加速推进,电力通信网作为能源互联网的神经中枢,承载着电网监控、数据传输、指令下达等关键功能,其稳定性和高效性直接影响着整个电网的运行安全和服务质量。光纤通信技术以其大容量、高速率、低损耗的优势,在电力通信体系中占据着核心地位。然而,随着智能电网的深入发展,其对通信网络的实时性、可靠性和智能化水平提出了更高要求,尤其是面对大规模分布式发电、多能源互补和电力市场交易等复杂场景,传统网络管理和优化策略逐渐显现出局限性^[1-2]。此外,启发式算法得到的解决方案远非最优。面对这一系列挑战,如何高效、动态地进行资源调度,确保网络在各种工况下的稳定运行,成为亟待解决的关键问题。

特别是在资源分配不均、网络拥塞频发及面对突发故障时的快速响应和恢复能力方面,现有技术难以满足日益增长的智能化管理需求。与此同时,深度强化学习(Deep Reinforcement Learning, DRL)已经在序贯决策和自动控制问题上取得了显著进步^[3-4]。因此,网络社区已将 DRL 作为关键技术展开研究,并将其应用于网络优化(如路由),以实现自动驾驶网络^[5-6]。在此背景下,将 DRL 的前沿成果引入电力通信网络管理与优化领域,显得尤为迫切^[7]。当前,电力通信网络的研究现状表明,尽管已有工作关注于利用先进的路径规划算法以提升网络效率与稳定性^[8]。但这些研究大多侧重于优化数据传输路径,而未能充分考虑网络运行中的风险因素,包括但不限于资源分配不均引发的潜在瓶颈、网络拥塞导致的服务质量下降,以及面对突发事件时的脆弱性和恢复速度。这种对风险评估的忽视限制了管理策略的有效性和网络整体的韧性。

针对上述研究局限,本文旨在突破传统方法的框架,提出一种新颖的电力通信网络自趋优控制策略。该策略创新性地融入风险量化方法,利用 DRL 的强大功能,实现对网络状态的实时监测与智能决策。本文主要贡献包括:

(1)提出了一种基于风险量化和深度强化学习的电力通信自趋优控制方法,针对性地解决了电力通信网络在路径规划和资源调度中的风险忽视问题。该方法采用图卷积网络(Graph Convolutional Networks, GCN),一种基于空间域采样的图神经网络变体,对节点特征进行层次化聚合,不仅关注节点间的流量调度,还引入了风险评估因子,其核心在于通过 GCN 模型对网络中各节点及其连接特性进行深层次特征挖掘,以获取更为精确的状态表示。在此基础上,设计了包含风险量化在内的多维评价体系,用以指导深度强化学习算法(DRL)进行智能决策。

(2)利用后置图神经网络嵌入强化学习框架,在融合光纤传输网架结构信息的同时,充分利用强化学习对不确定因素的适应性进行快速求解。为了克服传统强化学习

中可能因贪婪策略而导致的局部最优问题,本文方法引入了经验回放池(Experience Replay Buffer)。通过对一段时间内(时间节点的时序时间段内)的历史路径选择进行重放与评估,该机制有助于学习长期优化策略,避免了短期贪婪行为对整体路径选择优化的影响。经验回放池不仅增强了学习过程的鲁棒性,还提高了模型对复杂网络环境的适应能力,确保了路径选择既满足即时需求又兼顾长远利益。

(3)与传统方法相比,本文方法在路由选择与流量分配方面有显著提升,能够有效识别并避开受影响的链路,保持比理论基线更好的性能。

1 相关工作

1.1 基于图神经网络的风险态势感知

风险态势感知是指在电力通信网络管理中,持续监测、分析网络内外部环境,识别潜在威胁、评估安全风险,并预测未来可能出现的安全态势的能力^[9]。在智能电网的背景下,电力通信网络承载着关键的监控和控制信息,其稳定性和效率直接影响电网运行安全^[10]。风险态势感知使网络管理人员能够识别最脆弱的链接并有效分配资源,从而提高整个网络的复原力和可靠性。从电力通信网络中提取和学习潜在风险仍然是一项严峻挑战。Sheng 等^[11]提出一种利用双向门控递归单元(Bi-directional Gated Recurrent Units, BGRUs)的暂态稳定性评估方法,该方法增强了对具有挑战性的不稳定样本的学习能力。BGRU 是长短时记忆神经网络(Long Short-Term Memory, LSTM)的一种变体。图神经网络作为一种能有效处理包含复杂数据集的模型,自文献^[12-13]提出以来便发展出了许多变体,其特点非常适合电力通信网络中的特征处理^[14-16]。张沛等^[17]利用图神经网络对不确定因素的适应性设计前置图神经网络和后置图神经网络,并将其嵌入强化学习框架以提高模型对目标拓扑变化的适应性,进一步提高模型泛化和计算能力。郭创新等^[18]采用基于强化学习的图神经网络以适应新型电力系统中复杂的电网状态变化,从而实时调整和优化电网策略。

然而,上述方案均依赖历史数据训练,难以适应电力通信网络拓扑的动态重构,在处理电网的复杂拓扑结构、动态变化特性和高维度的风险指标时,计算精度和效率仍需进一步提升。同时,系统考虑的风险因子较少,在一定程度上限制了模型全面性和实用性。针对上述问题,本文提出了一种新方法,不仅考虑了基础风险因素,同时引入了针对这些风险因素的应急处理预案等,通过对各种风险因素进行量化处理,将其整合为综合风险得分,并在 DRL 算法中引入这一特征,从而增强了对复杂电网环境中各种风险因素的感知和应对能力。

1.2 同步数字体系中的路径选择与带宽分配

同步数字体系(Synchronous Digital Hierarchy, SDH)是一种国际标准的光纤通信系统,用于高效地传输、复用、路由和管理多种类型的数据。SDH作为电力通信网的核心传输技术之一,以其高度标准化、灵活可靠的特点,为电网提供了稳定、可靠的数据传输服务,是保障电力系统安全稳定运行的重要基础设施之一。随着智能电网的不断发展,其产生的数据流量快速增长,如何高效利用网络资源,特别是在动态变化的流量需求下实现最优带宽分配,是一项持续性挑战。强化学习(Reinforcement Learning, RL)被认为是实现网络资源动态分配的有效方法^[19]。例如,Suárez等^[20]探索了深度强化学习(Deep Reinforcement Learning, DRL)在光传输网络在线路由中的应用,提出一种细粒度的网络状态表示法,并设计了一种优化长期资源保护的智能策略。Almasan等^[8]将图神经网络集成到DRL代理中,并设计了一个针对特定问题的行动空间,增强了模型在任意网络拓扑中的泛化能力。然而,上述方法在应对风险因素方面的考虑仍然不足。在路由优化中仅以带宽利用率为目标函数,未能将诸如链路故障、节点失效和应急路径可用性等风险指标纳入奖励机制,导致策略在实际部署中易受突发故障影响。在光传输网络中,风险因素如链路故障、节点失效和突发流量等,对网络性能和稳定性具有重大影响。传统方法往往侧重于优化资源利用率和带宽分配,而对这些潜在的风险因素缺乏足够考虑,可能导致在实际应用中出现意外的网络性能下降或中断。本文提出了一种新方法,通过量化风险并将其作为特征纳入网络决策,从而解决了网络决策中的不确定性问题。具体而言,本文定义了一系列风险评估指标,如链路故障率、数据包丢失率和应急计划可用性^[18]。对这些指标进行加权 and 汇总,生成综合风险评分,反映当前网络状态的潜在

风险,并作为DRL代理在决策过程中的关键输入。

2 电力通信网络自趋优控制模型

电力通信网络自趋优控制模型主要包括特征提取模块、GRL代理模块。首先,将实际光传输网络抽象为图数据,图数据包含SDH的网络拓扑及其光缆特征数据 X 两部分。将抽象出来的数据输入至图卷积网络(Graph Convolutional Network, GCN)进行节点特征提取,输出节点特征向量;其次,将支路两端节点的特征向量和支路特征拼接,输入至由全连接层组成的支路特征提取网络,输出支路特征向量;最后,将支路特征向量输入至线性整流函数(Rectified Linear Unit, ReLU),输出该支路的风险分析结果。将该风险分析结果进行量化并作为边的权重。在下游的深度Q网络框架的eval net和target net中各嵌入一层加权GCN,用于提取网络拓扑结构信息与风险特征表征。为兼顾特征提取能力与训练稳定性,GCN仅使用一层卷积结构,输出节点特征经过高阶信息聚合后直接输入至下游Q网络模块。该GCN层随深度Q学习算法(Deep Q-Learning, DQN)主网络同步更新,最终由eval net输出 t 时刻动作值。

为提高训练过程的稳定性并抑制模型复杂度,本文采用轻量化GCN设计,仅设置单层卷积操作以完成网络特征提取。通过合理选择卷积邻域与激活函数,确保节点间局部邻接关系和风险特征得以有效建模,避免因过多叠加卷积层带来的特征过度拟合与训练震荡问题。此外,eval net与target net中GCN部分参数共享,以进一步提升模型鲁棒性与收敛效率。代理执行动作 a_t ,环境状态由当前状态 s_t ,切换为下一状态 s_{t+1} ,并反馈当前动作奖励值 r_t 供代理进行学习。本文模型架构如图1所示。

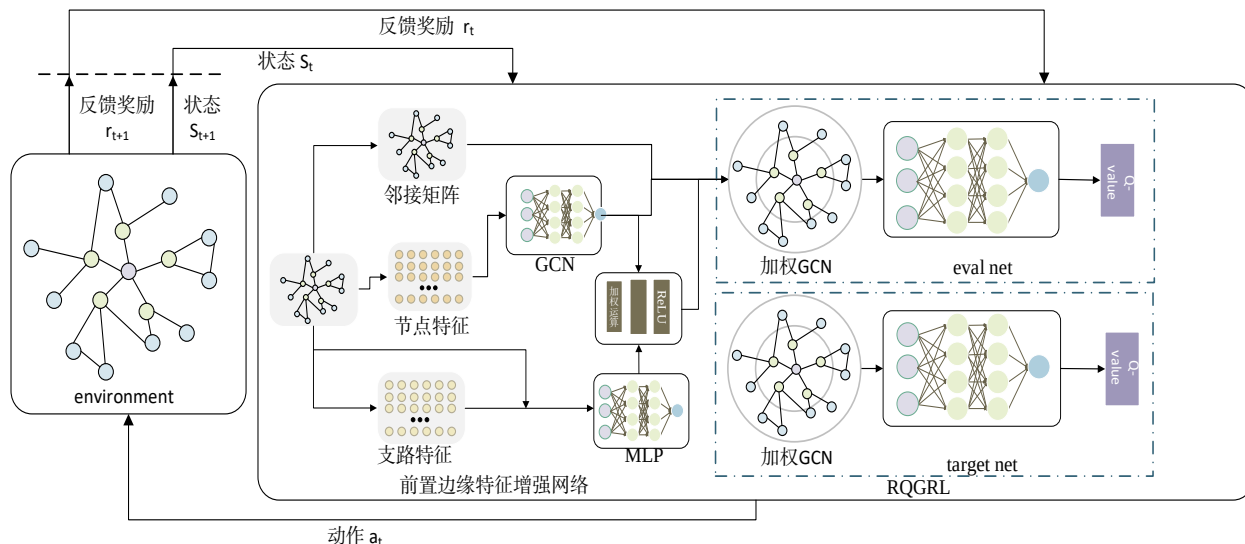


Fig. 1 Self-optimizing control model for power communication networks

图1 电力通信网络自趋优控制模型

2.1 问题定义与预备知识

2.1.1 问题定义

本文根据 ITU-T G.707 标准,将光纤通信网络抽象为 SDH 拓扑图 $G = (V, \varepsilon)^{[21]}$ 。其中, V 表示所有节点的集合, ε 表示所有支路的集合。针对本文提出的问题,该模型参考的特征向量包括:节点 u 的特征向量为 X_u , 包含节点类型 t_u 、设备状态 s_u 、容量 c_u 、丢包率 p_u 以及节点支持的应急预案等级 e_u ; 边 $u \rightarrow v$ 的特征向量 e_{uv} , 包括带宽 b 和信号质量 q , 以及是否配置有备用路径作为应急措施的二进制标记 a 。 B_{uv} 表示该支路的链路介数公式, 可表示为 $B_{uv} = \sum_{p \neq u \neq v \neq q} \frac{\sigma(uv)}{\sigma_{pq}}$ 。其中, B_{uv} 表示链路 uv 的介数, σ_{pq} 是从所有节点 p 到节点 q 的最短路径中经过链路 uv 的链路数量, 而 σ_{pq} 是从 p 到 q 的所有最短路径总数。本文所需特征类型如表 1 所示。

Table 1 Node and branch feature vectors

表 1 节点与边特征向量

类型	特征	含义
节点特征	t_u	节点类型
	s_u	设备状态
	c_u	容量
	p_u	丢包率
	e_u	节点支持的应急预案等级
支路特征	b	带宽
	q	信号质量
	a	是否配有备用路径
	B_{uv}	链路 uv 的链路介数

2.1.2 预备知识

本文采用 DQN 算法对模型进行训练学习。DQN 算法的理论基础依然建立在 Q-Learning 的贝尔曼方程之上, 但其核心在于使用深度神经网络作为 Q 值函数的近似器^[22]。如 Sutton 等^[19]所述, DQN 通过最小化时序差分误差实现 Q 值函数近似, 对于任意状态 s 和动作 a , DQN 的目标是学习一个函数 $Q(s, a; \theta)$ 。其中, θ 代表神经网络的参数, 使得其输出尽可能接近真实的 Q 值。对于任意状态 s , 动作 a 和下一个状态 s' , 存在最优 Q 函数 $Q^*(s, a)$ 满足:

$$Q^*(s, a) = E[R(s, a, s') + \gamma \max_{a'} Q(s', a') | s, a] \quad (1)$$

这里, $R(s, a; \theta)$ 表示通过行动 a 从状态 s 过渡到状态 s' 后立即获得的预期奖励, γ 是决定未来奖励重要性的权重因子。

DQN 通过最小化 Q 值网格输出与上述贝尔曼方程右侧目标值之间的均方误差, 进行网络参数更新:

$$L(\theta) = E_{(s, a, r, s') \sim U(D)}[(\gamma - Q(s, a; \theta))^2] \quad (2)$$

其中, $\gamma = r + \gamma \max_{a'} Q(s', a'; \theta)$ 是目标 Q 值, θ 代表目标网格的参数, D 是经验回放缓冲区, $U(D)$ 表示从 D 中均匀采样经验。

2.2 光纤网络的 GCN 驱动特征提取

本文特征提取部分可分为两部分: 节点特征提取和分

支特征提取。GCN 作为深度神经网络的经典类型, 成功地将卷积运算从常规数据(如图像)推广到图结构数据, 从而可以有效地处理复杂网络中的节点特征^[23]。其核心公式为:

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}} \hat{A} D^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (3)$$

其中: $H^{(l+1)}$ 为第 l 层 GCN 卷积处理后的输出信息; $H^{(l)}$ 为第 l 层 GCN 的节点输入信息, 其中, $H^{(0)} = X$, $X = [X_1, X_2, \dots, X_n]^T$; $\hat{A} = A + I_n$, 其中 A 为图的邻接矩阵, I_n 为单位矩阵; D 为度矩阵; $W^{(l)}$ 为第 l 层 GCN 的权重参数矩阵; $\sigma(\cdot)$ 为激活函数。 $L = D^{-\frac{1}{2}} \hat{A} D^{-\frac{1}{2}}$ 又称拉普拉斯矩阵, 其作用是防止在运算中出现数值不稳定的情况。通过上述两层 GCN 卷积运算, 在时间步长 t 处, 最终得到节点特征表示 h_t , 该特征向量可以反映网络中节点在时刻 t 的状态特征。

分支特征提取网络由全连接层组成, 其过程取决于节点特征提取的输出结果。在时间步 t 时, 对于支路 l , 节点特征网络输出节点 b 的特征向量为 $h_{b,t}$, 则当支路 l 两端的节点为节点 b_1 和节点 b_2 时, 支路 l 的支路特征为 $Y_{l,t} = (b_{l,t}, q_{l,t}, a_l)$, 全连接层的表达式为:

$$\begin{cases} d_{l,t}^0 = \text{CONCAT}(h_{b_1,t}, h_{b_2,t}, Y_{l,t}) \\ d_{l,t}^k = \sigma(W_s \cdot d_{l,t}^{k-1} + b_s) \end{cases} \quad (4)$$

其中: $d_{l,t}^k$ 为在时间 t 时网络中第 k 层全连接层输出的支路 l 特征向量, 且 $k=0$ 时为网络输入的节点特征与支路特征的拼接; W_s 和 b_s 为全连接层的参数。

经过上述全连接层处理后, 得到分支特征向量 $d_{l,t}$ 并输入输出网络。输出网络依次连接单个全连接层和激活函数, 激活函数的主要作用是根据分支特征向量计算分支 l 的风险值。在时间 t 时, 支路 l 的支路特征为:

$$R_{l,t} = \text{ReLU}(W_s d_{l,t} + b'_s) \quad (5)$$

其中, ω, b' 为单全连接层参数, $\text{ReLU}(\cdot)$ 为 ReLU 激活函数。

2.3 基于 GCN 和 DRL 的光纤网络代理设计

本文基于强化学习对图神经网络策略模型开展进阶训练。首先, 对强化学习的马尔可夫决策过程进行建模, 包括状态、动作、奖励函数等模块; 其次, 为了提高整体训练效率, 需要通过监督学习方法对初始模型进行预训练; 最后, 使用 DQN 强化学习算法训练策略模型, 使其具备处理复杂网络环境的能力。

为将上述 MDP 决策模型进行公式化描述, 给出状态空间、动作空间定义。

2.3.1 状态空间

依据强化学习标准框架, 状态空间需要包括光纤通信网络的拓扑图、链路占用情况、节点负载、路径状态等多维动态信息^[19]。特别引入链路介数作为拓扑特征的重要补充, 它表示可能穿过链路的路径数量。具体而言, 采用以下方式计算链路介数: 对于拓扑图中的每一对节点, 计算 k

条候选路径,并维持一个按链接计数器,该计数器指示经过链接的路径数量。因此,每个链接上的介数是跨越链接的端到端的路径数量除以路径总数。总而言之,状态空间由网络拓扑、链路占用率、节点负载和路径状态共同组成,其中链路调解器的数量反映了可能穿过链路的路径数量。

模型在时间步 t 下的状态为 S_t , 则:

$$S_t = \{G_t, O_{l,t}, L_{n,t}, P_{p,t}, H_t, R_t\} \quad (6)$$

其中:在时间步 t 下, G_t 表示网络拓扑图的连接状态; $O_{l,t}$ 为链路 l 在时间步 t 的占用带宽比例; $L_{n,t}$ 为节点 n 处理的数据流量; $P_{p,t}$ 为路径 p 的当前可用性和带宽预留情况; H_t 是节点特征向量; R_t 是每个分支的风险值。

2.3.2 动作空间

在大规模真实网络中, DRL 代理往往面临高维度的动作空间问题, 参考 Almasan 等^[8]的方法, 将源-目标节点对的候选路径限制为 K 条最短路径, 以降低动作空间维度。这样做既保持了一定程度的路由灵活性, 又显著减少了模型需要探索和评估的动作数量, 从而提高了学习效率和可行性。

为应对实际网络中多样化的带宽需求与路径选择耦合优化的挑战, 本文提出了一种分层离散化表征机制。该机制通过将连续带宽需求映射为有限离散空间中的符号化表示, 实现了资源分配决策与拓扑路径选择的联合建模。具体而言, 本文采用长度为 M 的独热编码向量对带宽需求进行离散化表征。通过将路径选择向量 (K 维独热编码) 与带宽分配向量 (M 维独热编码) 进行嵌套式拼接, 形成统一的联合动作向量。其中, 路径选择部分通过 $o_k \in R^K$ 编码从 K 条候选路径中选择最优路径, 带宽分配部分通过 $b_i \in R^M$ 指定所选带宽等级。最终动作向量可形式化为:

$$A_a = \text{CONCAT}(o_k, b_i) \quad (7)$$

其中, $o_k \in R^K$ 为路径选择的独热编码, $b_i \in R^M$ 为当前带宽等级的独热编码, 最终动作向量 $A_a \in R^{(K+M)}$ 表示复合动作向量。此方法保留了 K 条最短路径的灵活性, 同时在离散带宽选择的基础上, 保证了动作维度的一致性与可学习性。

2.3.3 奖励函数

奖励函数设计为奖励部分和惩罚部分。奖励函数设计如下:

$$\begin{aligned} r_t = & \omega_1 \times \text{BandWidthUtilization}(t) - \omega_2 \times \text{CongestionLevel}(t) \\ & + \omega_3 \times \text{PathSuccessRate}(t) + \omega_4 \times \text{RiskScore}_t(t) - \omega_5 \\ & \times \text{HopCountFactor}_t \end{aligned} \quad (8)$$

其中, $\text{BandwidthUtilization}(t)$ 为网络在时间步 t 时, 其带宽资源被利用的程度; $\text{CongestionLevel}(t)$ 为网络在时间步 t 时的拥塞程度; $\text{PathSuccessRate}(t)$ 为在时间步 t 时, 网络中数据传输或路由请求的成功比例; RiskScore 为风险评估得分, 通过风险因子的加权求和得到, 公式为: $\text{RiskScore} = \omega_q \times q_{l,t} + \omega_p \times p_u + \omega_a \times (1 - a)$, 其中 $\omega_q, \omega_p, \omega_a$ 是对应风

险指标的权重。HopCountFactor _{t} 为跳数因子, 是基于跳数的衰减函数, 公式为 $\text{HopCountFactor} = \omega_h \times e^{-\alpha \cdot \text{HopCount}}$ (ω_h 为跳数权重)。为了提升奖励函数的客观性和可复现性, 本文采用基于信息熵的自适应加权机制以确定各指标权重。具体而言, 对于奖励函数中的 5 项指标, 通过计算其在训练数据中的离散熵 H_i 以量化其信息量, 并据此定义权重。计算公式分别为:

$$\omega_i = \frac{1 - H_i}{\sum_{j=1}^5 (1 - H_j)} \quad (9)$$

$$H_i = -\sum_k p_{ik} \log(p_{ik}) \quad (10)$$

其中, p_{ik} 表示第 i 项指标的归一化概率分布。该机制依据信息量越高 (熵越小) 则权重越大的原则, 确保高区分度指标对整体决策起主导作用, 从而构建更合理的奖励函数结构。

2.3.4 融合 PER 和 n 步累计奖励的 DRL 代理

在构建 DRL 代理的过程中, 设计了两套核心网络: 评估网络 (Eval Net) 和目标网络 (Target Net)。其中, Eval Net 用于实时评估当前状态下各可选动作的 Q 值, 以指导代理决策; Target Net 则为训练提供相对稳定的目标 Q 值, 缓解学习过程中的振荡问题。

为提升学习效率与稳定性, 引入了经验回放池 (Replay Buffer), 每条记录存储当前状态、执行动作、未来 n 步累计奖励、新状态及初始化权重。通过引入 n 步累计奖励机制, 能够在决策评估中考虑更长时间尺度的收益, 缓解短期反馈噪声对训练的不利影响。累积奖励的计算公式为:

$$G_t = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n Q(s_{t+n}, a_{t+n}) \quad (11)$$

在此基础上, 进一步采用优先经验回放 (Prioritized Experience Replay, PER), 通过引入基于时间差分误差 (TD-Error) 的优先级排序机制, 使高价值经验被更频繁采样学习。具体而言, 经验条目的采样概率按如下比例设定:

$$P(i) = \frac{p_i^{\alpha'}}{\sum_j p_j^{\alpha'}} \quad (12)$$

其中, α' 是一个超参数, 控制着采样概率的偏斜程度。随着 α' 的增加, PER 机制会更加倾向于选择那些 TD 误差较大的经验。

由于 PER 机制导致的采样概率不均匀, 需要计算一个权重 ω'_i 以补偿采样偏差, 保证学习目标的无偏性。重要性采样权重 ω'_i 定义为:

$$\omega'_i = \left(\frac{1}{N \cdot P(i)} \right)^\beta \quad (13)$$

其中, N 是经验池中的经验总数, β 用于调整重要性采样强度。随着训练的进行, β 逐渐增加到 1, 以逐渐减少权重的偏差影响。每次更新网络之后, 都需要重新计算每条经验的 TD 误差, 并据此更新其在经验池中的优先级。这样可以确保算法始终关注那些最有价值的经验。

最后,在模型更新阶段,代理以加权小批量样本为基础,最小化加权时序差分损失函数,公式如下:

$$L(\theta) = E_i[\omega'_i(y_i - Q(s_i, a_i; \theta))^2] \quad (14)$$

其中, y_i 为基于 n 步累计奖励计算的目标 Q 值,并根据最新 TD 误差动态更新各经验条目的优先级,从而不断强化对关键经验的学习,提升代理在复杂动态环境下的决策鲁棒性。下文给出了训练过程的伪代码,详细说明了每轮训练步骤及其逻辑关系。

算法 1 DRL 训练主循环

输入: 无

输出: 训练完成的 DRL 模型

FOR episode = 1 TO E DO

Reset environment and obtain initial state S_0 .

FOR t = 0 TO T DO

a. 节点特征提取: Extract node feature h_i from G_t using GCN.

b. 分支特征提取: Compute branch feature vectors $d_{l,t}$ using ht and branch attributes $Y_{l,t}$.

c. 动作空间定义: Define action space A based on k shortest paths for each source-destination pair.

d. 动作选择: Select action at using ϵ -greedy policy from $Q(s, a; \theta)$.

e. 执行动作: Execute a_t , observe reward r_t and next state S_{t+1} .

f. 存储经验: Store transition (S_t, a_t, r_t, S_{t+1}) in replay buffer D.

IF $|D| > \text{batch size } B$ THEN

i. 采样小批量: Sample $\{(S_j, a_j, r_j, S_{j+1})\}$ from D using PER probabilities $P(i)$.

ii. 计算目标 Q 值: For each sample, compute the $n = 5$ -step return R_j^n (由式(2)定义).

iii. 计算 TD 误差: Compute $\delta_j = |R_j^n - Q(S_j, a_j; \theta)|$.

iv. 更新优先级: Update priority pi in D based on δ_j (由式(3)定义).

v. 计算重要性采样权重: Compute ω_i for each sample in mini-batch (由式(4)定义).

vi. 更新评估网络参数: Minimize the loss L (由式(5)定义).

vii. 更新目标网络参数: Synchronize $\theta^- \leftarrow \theta$ every $C = 5$

steps to stabilize training.

END IF

END FOR

END FOR

3 仿真算例与结果分析

为了验证本文模型的有效性,使用国网某省电力通信网络数据进行训练和测试。采用该数据集部分节点,包括电力通信领域涉及不同业务的物理节点 1 200 个,每个节点包括节点类型、设备状态、容量、丢包率,以及该通信节点支持的应急预案等级,还有光纤链路带宽、信号质量、是

否配置有备用路径作为应急措施的二进制标记。

3.1 算法参数设置

本文实验参数设置如表 2 所示。

Table 2 Hyperparameter settings

表 2 超参数设置

参数	设置
学习率 α	0.001
折扣因子 γ	0.92
探索率 ϵ	0.1
经验回放池大小 D	10 000
更新频率 C	5
目标网络更新率 τ	0.002
PER 衰减速度 α'	0.6
PER 权重影响因子 β	0.4
优先级最小值 ι	0.000 1
可用最短路径 k	4

3.2 算法调度结果

3.2.1 性能评估

在评估实验中,比较了 RQGRL(RiskQuant-GRL)、代理与最先进的基于 DRL 的解决方案,并考虑了 3 种流量需求类型(STM-1、STM-4 和 STM-16)^[21]。当 DRL 代理正确分配一个需求时,它会立即获得奖励,如果该需求被正确分配,则为当前通信需求的带宽,否则为零。由此定义:如果 DRL 代理所选择的所有路径上的链路都有足够可用容量来承载这种需求,则该需求被视为成功分配;反之,如果需求未能成功分配,则本次尝试被视为失败,并进入下一个时间步或开始新的仿真周期。

为验证所提 RQGRL 方法的有效性,本文选取了前沿深度强化学习模型作为对比基线,所有基线模型均在相同仿真环境下进行训练与测试。具体方法如下:

(1)GCN-DRL 模型^[17]。基于图卷积网络(GCN)与深度 Q 学习(DQN)的故障恢复决策方法,其采用两层 GCN 提取节点特征(设备状态、容量等),并通过 DQN 选择最短路径。

(2)RiskGNN 模型^[18]。一种动态风险聚合图神经网络,通过节点特征(丢包率、应急预案等级)的层次传播评估网络风险。

(3)Safe-DRL 模型^[24]。面向综合能源系统的安全约束强化学习方法,通过预定义的安全引导函数限制动作空间。

(4)TreeSearch-DRL 模型^[25]。结合蒙特卡洛树搜索(MCTS)与 DRL 的机组组合优化方法,通过树剪枝策略减少高维动作空间的探索复杂度。尽管其在长期规划中表现稳定,但决策延迟高达 162 ms(因搜索树回溯耗时)。

(5)PPO-OPF 模型^[26]。基于近端策略优化(PPO)的最优潮流求解器,采用凸松弛技术将非凸问题转化为二阶锥规划(SOCP)以提升计算效率。

在国网电力通信网络拓扑结构上对 6 种路由策略进行了评估,并比较了本文方法与基线方法的表现,结果如

图 2 所示。具体而言, RQGRL 的带宽分配效率达到 93.2%, 较次优的 RiskGNN 方法(89.5%)提升 3.7 个百分点, 较传统凸松弛方法 PPO-OPF(84.1%)提升近 10%。这一提升源于 RQGRL 独特的边权重量化机制——通过图卷积网络动态融合链路介数、信号质量与应急预案等级特征, 将

综合风险评分编码为边权重(见式(5)), 从而在路径规划中同步优化资源利用与风险规避。相比之下, GCN-DRL 仅依赖节点状态聚合, 忽视了链路层动态特性; 而 PPO-OPF 因对非凸问题的凸化近似, 导致解空间受限, 难以捕捉真实网络的多目标权衡关系。

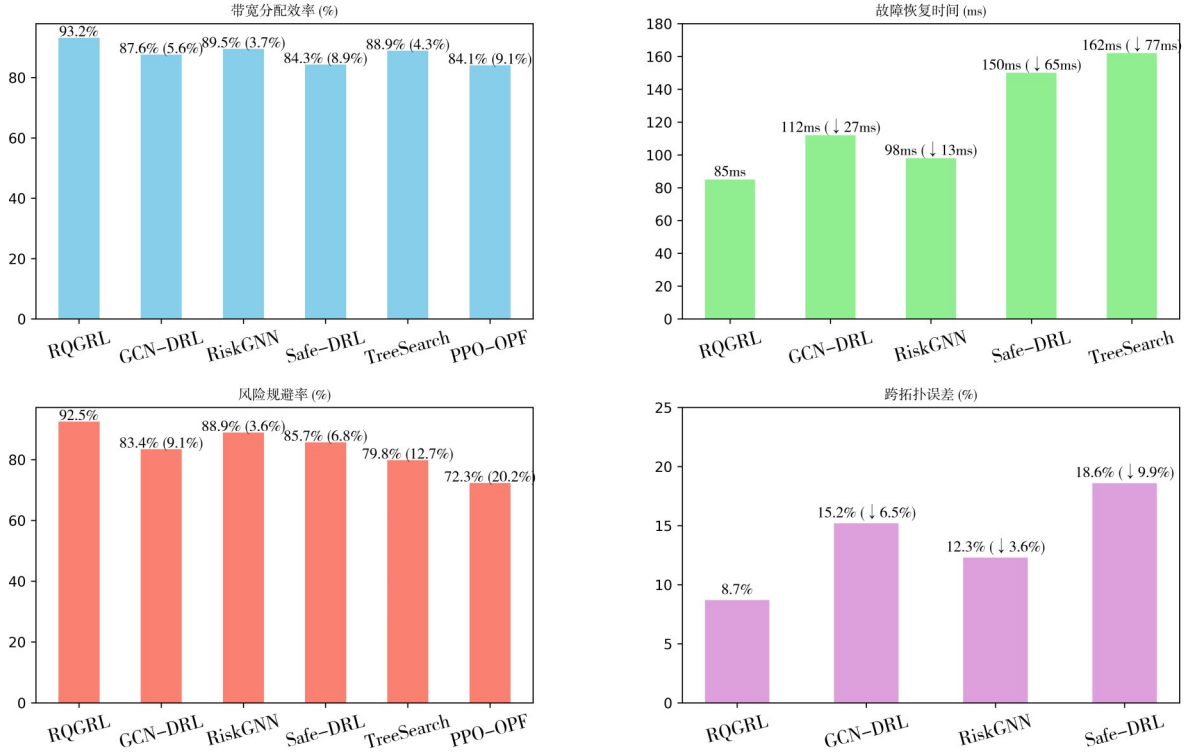


Fig. 2 Performance evaluation results

图 2 性能评估结果

在故障恢复能力方面, RQGRL 仅需 85 ms 即可完成路径切换, 较安全约束驱动的 Safe-DRL 方法(150 ms)提速 43.3%。这得益于其优先经验回放(PER)机制(见式(13)), 通过加权采样历史高风险链路故障案例, 强化代理对关键脆弱链路的替代路径学习效率。相比之下, Safe-DRL 依赖静态容量阈值约束动作空间, 无法动态感知突发故障的传播风险, 导致响应迟滞。此外, RQGRL 在风险规避率指标上达到 92.5%, 较 RiskGNN(88.9%)和树搜索剪枝方法 TreeSearch(79.8%)分别提升 3.6%、12.7%。这一差距凸显了边权重量化相对于节点中心化风险建模的优越性: RQGRL 通过实时更新链路风险权重(见式(5)), 精准识别并规避高丢包率、低备用路径的脆弱链路, 而 RiskGNN 的节点聚合策略难以捕捉跨链路风险关联性, TreeSearch 则完全缺乏显式风险建模能力。

在动态拓扑场景下, RQGRL 的跨拓扑泛化误差仅为 8.7%, 较 GCN-DRL(15.2%)和 RiskGNN(12.3%)降低 42.8%与 29.3%。这表明, 基于边权重的层次特征传递机制(见式(3))使模型能够适应节点与链路的随机增减, 维持拓扑结构的语义一致性。例如, 当 20% 的节点被随机移

除时, RQGRL 通过动态调整剩余链路的介数与风险权重, 仍能保持 90% 以上的带宽分配效率, 而 GCN-DRL 因过度依赖固定节点特征, 性能损失高达 15.2%。值得注意的是, PPO-OPF 和 TreeSearch 在此场景下完全失效(标记为“-”), 前者受限于凸松弛模型的静态假设, 后者则因树搜索的高延迟无法满足实时决策需求。这些结果共同验证了 RQGRL 在复杂动态环境中的技术突破——通过非凸动态规划与图网络驱动的自适应风险量化, 实现了效率、安全性与泛化性的协同优化。

3.2.2 链路故障恢复能力

该部分用于评估本文方法能否有效识别并避免网络中的风险因素。本文选择的危险影响因子为 $F = \{\text{链路拥塞}, \text{链路故障}\}$ 。链路拥塞和链路故障在同一场景下可以同时存在, 也可以单独存在, 具体情况取决于系统的运行状态和外部扰动。指标集 f_1, f_2, f_3 均为 F 的子集, f_1 表示代理只考虑链路拥塞指标, f_2 表示代理只考虑链路故障指标, $f_3 = \{f_{31}, f_{32}\}$, 表示代理同时考虑两种指标。 f_{31}, f_{32} 表示 f_3 情况下的链路拥塞指标、链路故障指标。

在评估中, 考虑了一系列可能经历多达 10 个链路故

障的情况。因此,RQGRL代理的任务是在仍然最大化分配总带宽的同时找到新的路由配置以避免受影响的链路。已有工作表明,现实世界的网络拓扑结构会在链路故障过程中发生变化,这些网络连接变化是不可预测的,而且会对协议收敛或实现网络优化目标产生重大影响^[27-28]。对国网电力通信网络拓扑结构进行以下操作:①在特定时间段内逐渐增加通过这些链路的流量,直至达到或超过链路容量的90%,以模拟高度拥塞状态;②随机选择网络中的链路进行故障模拟,模拟过程可以是瞬时故障(链路短暂中断后恢复)和永久故障(链路彻底失效),每次故障影响1至多条链路。比较了RQGRL代理相对于理论流模型所取得的分数(分配的带宽)。图3显示了平均值(y 轴)与链路故障数量(x 轴)之间的关系。可以看到,RQGRL的平均得分始终高于其他5种前沿算法,且随故障次数增加呈现平缓衰减趋势。当网络遭遇10次并发链路故障时,RQGRL仍能维持452分的带宽分配能力,较表现次优的Safe-DRL方法(406分)提升11.3%,验证了其在高风险环境中的强鲁棒性。具体而言,在低故障场景(0~4次)中,RQGRL通过边权重量化机制精准感知局部拥塞风险,其得分(如3次故障时783分)显著优于依赖节点风险评估的RiskGNN(770分);而在高故障场景(5~10次)中,其动态路径重规划能力有效规避了树搜索方法(TreeSearch)因动作空间爆炸导致的性能崩溃(10次故障时230分)。

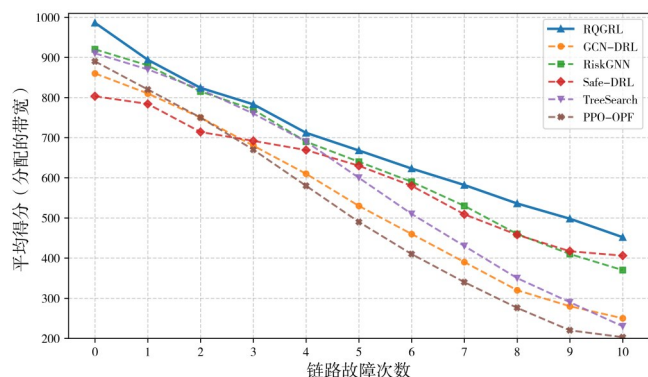


Fig. 3 Evaluation of use case with link failures

图3 存在链路故障的用例评估

实验结果表明,RQGRL的动态风险量化机制与图网络泛化能力是其性能优势的核心:边权重实时更新策略(见式5)使代理能同步响应链路拥塞与故障,而优先级经验回放(见式13)则通过强化高风险场景的学习,保障了长期收益优化。这些创新使得RQGRL在国网电力通信网络等动态复杂系统中展现出重要的应用潜力。

3.2.3 超参数敏感性实验

为进一步分析权重因子 β 对模型性能的影响,本文设计了超参数敏感性实验。 β 作为奖励函数中风险规避部分的权重参数,其取值直接影响到代理在优化资源利用与规避风险之间的权衡策略。若 β 取值较低,代理将倾向于追求带宽效率,可能忽视链路潜在风险;而当 β 取值较高

时,代理则更倾向于规避高风险路径,但可能导致资源利用率下降。因此,合理设定 β 对于提升整体性能具有重要意义。

实验中,保持其他超参数设置不变,将 β 的取值设置为 $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$,分别训练并测试RQGRL模型。对比评估带宽分配效率、故障恢复时间和风险规避率3个关键性能指标,结果如图4所示。可以观察到,随着 β 从0.2逐步增加至0.4,风险规避率不断提升,同时带宽效率保持在较高水平。当 $\beta = 0.4$ 时,模型达到最佳综合性能:带宽分配效率为93.2%,故障恢复时间为85 ms,风险规避率为92.5%。继续增加 β 至0.7、0.8时,虽然风险规避率进一步上升,但带宽效率下降明显,且故障恢复时间有所延长,表明代理策略在高 β 值下趋于过度保守,资源利用效率受损。因此,综合考虑资源利用、故障响应与风险控制三方面因素,本文最终选取 $\beta = 0.4$ 作为最优设置。

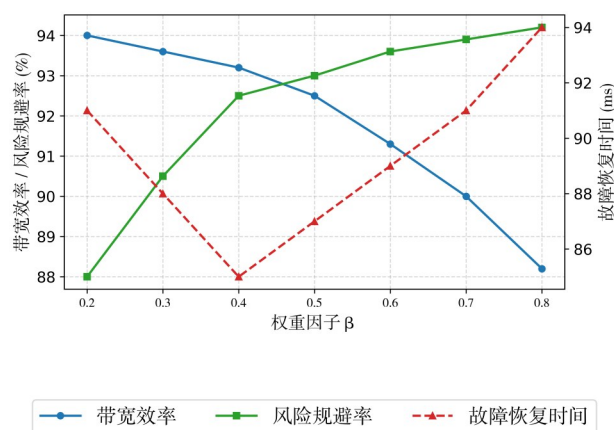


Fig. 4 Effect of weighting factor β on performance metrics

图4 权重因子 β 对性能指标的影响

4 结语

本文针对电力通信网络的优化控制难题,设计了一套新颖的自趋优控制策略,重点突破了传统方法忽视风险评估的局限。该策略的核心在于将深度强化学习与图神经网络相结合,通过风险量化方法动态调整网络策略,实现对通信资源的高效配置和路径规划的智能化决策。实验部分验证了模型的有效性,通过与当前最先进的DRL系统、理论流体模型和负载均衡路由策略对比,结果表明提出的RQGRL模型在带宽分配效率上取得了显著提升,模型在处理链路故障和高度拥塞时,仍能保持较高的带宽利用率和网络稳定性,证明了其在复杂网络环境下的优越性。

本文算法虽然在实验中取得了不错的效果,但仍有一些问题有待进一步研究。例如:如何提高算法泛化能力,使其能够应对更加复杂的网络环境;如何优化算法的动作空间设计,使其在避免高维动作空间的情况下更详细地描述动作。对这些问题的研究将会为DRL在实际应用中的

推广提供更多可能性。

参考文献:

- [1] HARTERT R, VISSICCHIO S, SCHAUS P, et al. A declarative and expressive approach to control forwarding paths in carrier-grade networks [J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 15–28.
- [2] KNIGHT S, NGUYEN H X, FALKNER N, et al. The Internet topology zoo [J]. IEEE Journal on Selected Areas in Communications, 2011, 29(9): 1765–1775.
- [3] MNH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529–533.
- [4] WANG M W, CUI Y, WANG X, et al. Machine learning for networking: workflow, advances and opportunities [J]. IEEE Network, 2017, 32(2): 92–99.
- [5] MESTRES A, RODRIGUEZ-NATAL A, CARNER J, et al. Knowledge-defined networking [J]. ACM SIGCOMM Computer Communication Review, 2017, 47(3): 2–10.
- [6] KALMBACH P, ZERWAS J, BABARCI P, et al. Empowering self-driving networks [C]//Proceedings of the Afternoon Workshop on Self-Driving Networks, 2018: 8–14.
- [7] FENG B, HU Y J, HUANG G, et al. Review of novel power system dispatching optimization methods based on deep reinforcement learning [J]. Automation of Electric Power Systems, 2023, 47(17): 187–199.
冯斌, 胡轶婕, 黄刚, 等. 基于深度强化学习的新型电力系统调度优化方法综述[J]. 电力系统自动化, 2023, 47(17): 187–199.
- [8] ALMASAN P, SUÁREZ-VARELA J, RUSEK K, et al. Deep reinforcement learning meets graph neural networks: exploring a routing optimization use case [J]. Computer Communications, 2022, 196: 184–194.
- [9] ENDSLEY M R. Toward a theory of situation awareness in dynamic systems [J]. Human Factors, 1995, 37(1): 32–64.
- [10] ZHANG Z G, KANG C Q. Challenges and prospects for constructing the new-type power system towards a carbon neutrality future [J]. Proceedings of the CSEE, 2022, 42(8): 2806–2818.
张智刚, 康重庆. 碳中和目标下构建新型电力系统的挑战与展望[J]. 中国电机工程学报, 2022, 42(8): 2806–2818.
- [11] SHENG Y, CAI D, WANG E, et al. Transient stability assessment of power systems based on semi-supervised active learning [C]//2023 5th Asia Energy and Electrical Engineering Symposium (AEEES), 2023: 758–763.
- [12] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model [J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61–80.
- [13] LI Y, TARLOW D, BROCKSCHMIDT M, et al. Gated graph sequence neural networks [C]//San Juan: The 4th International Conference on Learning Representations, 2016.
- [14] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks [C]//Vancouver: The 6th International Conference on Learning Representations, 2018.
- [15] RUSEK K, SUÁREZ-VARELA J, ALMASAN P, et al. RouteNet: leveraging graph neural networks for network modeling and optimization in SDN [J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2260–2270.
- [16] XU Z, TANG J, MENG J, et al. Experience-driven networking: a deep reinforcement learning based approach [C]//Honolulu: 2018 IEEE Conference on Computer Communications (INFOCOM), 2018: 1871–1879.
- [17] ZHANG P, CHEN Y X, WANG G H, et al. Fault recovery decision of distribution network based on graph reinforcement learning [J]. Automation of Electric Power Systems, 2024, 48(2): 151–158.
张沛, 陈玉鑫, 王光华, 等. 基于图强化学习的配电网故障恢复决策[J]. 电力系统自动化, 2024, 48(2): 151–158.
- [18] GUO C X, LIU Z P, LIU Y G, et al. GNN and rl based power system risk situation perception [J]. Power System and Clean Energy, 2023, 39(12): 41–49.
郭创新, 刘祝平, 刘永刚, 等. 基于图神经网络和强化学习的电网风险态势感知[J]. 电网与清洁能源, 2023, 39(12): 41–49.
- [19] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. Cambridge: MIT Press, 1998.
- [20] SUÁREZ V J, MESTRES A, YU J, et al. Routing in optical transport networks with deep reinforcement learning [J]. Journal of Optical Communications and Networking, 2019, 11(11): 547–558.
- [21] ITU-T. G. 707/Y. 1322 (03/07): Network node interface for the synchronous digital hierarchy (SDH) [S]. Geneva: ITU-T, 2007.
- [22] CHRISTOPHER J C H W, PETER D. Q-learning [J]. Machine Learning, 1992, 8(3–4): 279–292.
- [23] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [DB/OL]. <https://arxiv.org/abs/1609.02907>, 2016.
- [24] QIU D, DONG Z, ZHANG X, et al. Safe reinforcement learning for real-time automatic control in a smart energy-hub [J]. Applied Energy, 2022, 309: 118403.
- [25] DE MARS P, O'SULLIVAN A. Applying reinforcement learning and tree search to the unit commitment problem [J]. Applied Energy, 2021, 302: 117519.
- [26] ZHOU Y H, LEE W J, DIAO R, et al. Deep reinforcement learning based real-time AC optimal power flow considering uncertainties [J]. Journal of Modern Power Systems and Clean Energy, 2022, 10(5): 1098–1109.
- [27] FRANCOIS P, FILSFILS C, EVANS J, et al. Achieving sub-second IGP convergence in large IP networks [J]. ACM SIGCOMM Computer Communication Review, 2005, 35(3): 35–44.
- [28] JAIN S, KUMAR A, MANDAL S, et al. B4: experience with a globally-deployed software defined WAN [J]. ACM SIGCOMM Computer Communication Review, 2013, 43(4): 3–14.

(责任编辑: 孙 娟)