

面向异构工业任务高并发计算卸载的深度强化学习算法

刘晓宇^{1), 2), 3), 4)} 许驰^{1), 2), 3)} 曾鹏^{1), 2), 3)} 于海斌^{1), 2), 3)}

¹⁾(中国科学院沈阳自动化研究所, 机器人学国家重点实验室, 沈阳 110016)

²⁾(中国科学院网络化控制系统重点实验室, 沈阳 110016)

³⁾(中国科学院机器人与智能制造创新研究院, 沈阳 110169)

⁴⁾(中国科学院大学, 北京 100049)

摘 要 进入工业 4.0 时代, 大规模互联分布式智能工业设备产生了海量的具有时延敏感和计算负载差异的异构工业任务, 终端侧有限的计算能力难以支持任务的实时高效处理。通过工业无线网络将任务卸载到网络边缘侧服务器进行多接入边缘计算成为解决终端侧算力受限问题的一种有效手段。然而, 工业无线网络有限的时频资源难以支持大规模分布式工业设备的高并发任务卸载。本文充分考虑异构工业任务高并发计算卸载中有限时频资源约束和建模难的问题, 提出一种基于深度强化学习的动态优先级并发接入算法(Deep Reinforcement Learning-based Concurrent Access with Dynamic Priority, CADP-DRL)。该算法首先分析异构工业任务的时延敏感性和计算负载时变性, 为工业设备分配不同的优先级, 动态地改变工业设备接入信道进行计算卸载的概率。然后, 利用 Markov 决策过程形式化动态优先级高并发计算卸载问题, 并采用深度强化学习方法建立高维状态空间下状态到动作的映射关系。针对动态优先级和并发卸载的多目标决策问题, 设计了包含优先级奖励和卸载奖励的复合奖励函数。为保证训练数据的独立同分布, 同时提高算法收敛速度, 提出了带经验权重的经验回放方法。对比实验结果表明, CADP-DRL 能够快速收敛, 实时响应, 在实现最小卸载冲突的情况下为高优先级工业设备提供最高的成功卸载概率保证, 性能优于 slotted-Aloha、DQN、DDQN 和 D3QN 算法。

关键词 多接入边缘计算; 工业无线网络; 动态优先级; 任务卸载; 深度强化学习

中图法分类号 TP18

Deep Reinforcement Learning-based High Concurrent Computing Offloading for Heterogeneous Industrial Tasks

Liu Xiao-yu^{1), 2), 3), 4)} Xu Chi^{1), 2), 3)} Zeng Peng^{1), 2), 3)} Yu Hai-bin^{1), 2), 3)}

¹⁾(State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016)

²⁾(Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang 110016)

³⁾(Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169)

⁴⁾(University of Chinese Academy of Sciences, Beijing 100049)

本课题得到国家重点研发计划(2020YFB1710900); 国家自然科学基金(61803368, 61972389, 61903356, U1908212); 中国博士后科学基金(2019M661156); 辽宁省“兴辽英才计划”项目(XLYC1801001); 中国科学院青年创新促进会资助。刘晓宇, 博士研究生, 主要研究领域为工业实时通信、深度强化学习. E-mail: liuxiaoyu1@sia.cn。许驰(通信作者), 博士, 副研究员, 硕士生导师, 中国计算机学会(CCF)会员(52851M), 主要研究领域为工业无线网络、5G URLLC. E-mail: xuchi@sia.cn。曾鹏, 博士, 研究员, 博士生导师, 主要研究领域为工业通信、边缘计算. E-mail: zp@sia.cn。于海斌(通信作者), 博士, 研究员, 博士生导师, 主要研究领域为工业通信与实时系统理论、分布控制系统技术、网络协同与智能制造. E-mail: yhb@sia.cn。

Abstract With the rapid development of Industry 4.0, massive distributed intelligent industrial devices are interconnected by industrial wireless networks, and generate a large number of heterogeneous industrial tasks with different delay sensitivity and computing load during smart manufacturing. Real-time and efficient processing of industrial tasks is the key factor affecting the safety and efficiency of industrial manufacturing production. However, the limited local computing capacity of industrial devices cannot support the real-time and efficient processing of industrial tasks, and the common industrial cloud computing paradigm results in uncertain communication delays and additional network security issues. It is an effective method to offload industrial tasks to Multi-access Edge Computing servers deployed in base stations, access points and other network edge infrastructures through industrial wireless networks. Nevertheless, the limited time-frequency resources of industrial wireless networks cannot support the high concurrent computing offloading of industrial tasks. With consideration of the difficulty in modeling high concurrent computing offloading of industrial tasks, a Deep Reinforcement Learning-based Concurrent Access with Dynamic Priority (CADP-DRL) algorithm is proposed in this paper. Firstly, the industrial devices are assigned with dynamic priorities according to the delay sensitivity and computing load of their industrial tasks, and the access offloading probabilities of these industrial devices are changed dynamically depending on their priorities. Then, the Markov decision process is utilized to formulate the dynamic priority concurrent computing offloading problem. As both dynamic priority and concurrent computing offloading of massive industrial devices result in the explosion of state space, deep reinforcement learning is used to establish a mapping relationship from states to actions in the high-dimensional state space. Next, the long-term cumulative reward is maximized to obtain an effective dynamic priority concurrent computing offloading policy. Especially, with the aim at the multi-objective decision of dynamic priority and concurrent offloading, a novel compound reward function with joint priority reward and offloading reward is designed. The priority reward is used to ensure reliable offloading of high-priority industrial devices, and the offloading reward is employed to minimize offloading conflicts. In order to guarantee the independent and identical distribution of training data while accelerating the convergence of CADP-DRL, an experience replay with experience-weight is designed. Experiences are classified as high-weighted and low-weighted experiences depending on their weights, and stored in different experience memories respectively. Experiences are randomly sampled as the training data for CADP-DRL, and the sampling probabilities of experiences in different experience memories vary dynamically to break the time correlation among experiences while speeding up the convergence. The expensive training overhead of CADP-DRL is consumed in the offline training phase, and the trained CADP-DRL can make an effective computing offloading decision in real-time in the online execution phase. Slotted-Aloha algorithm is chosen as the benchmark algorithm in the field of communications, and DQN, DDQN and D3QN algorithms are chosen as the benchmark algorithms in the field of deep reinforcement learning. Extensive experiments show that, compared with these benchmark algorithms, CADP-DRL converges quickly, and performs well in generalization. Meanwhile, CADP-DRL always guarantees the highest successful offloading probabilities of high-priority industrial devices with minimum offloading conflicts.

Key words multi-access edge computing; industrial wireless networks; dynamic priority; task offloading; deep reinforcement learning

1 引言

随着工业 4.0 进程的逐渐加速, 工业现场大规模的分布式智能工业设备通过工业无线网络实现了互联互通, 传输着工业生产过程中的海量异构工

业任务, 包括控制消息, 音视频流, 感知信息等^[1,2], 其具有明显的时延敏感和计算负载差异。然而, 工业任务能否在本地有限算力下实时高效处理是影响工业安全能效生产的关键因素。

为了缓解本地算力不足的问题, 自 2005 年起, 云计算^[3]技术开始出现, 并得到广泛应用。云计算

通过集中式部署计算单元实现强大的算力支持。分布式工业设备通过网络将本地数据卸载到云端进行计算处理,云端处理完成后再将结果回传给分布式工业设备,极大地缓解了本地算力不足问题。但是,由于云计算通常部署在距离工业现场较远的地方,长距离的网络传输会造成不确定的传输时延和附加的网络安全问题。为了提升实时性,将算力迁移到网络边缘的多接入边缘计算(Multi-access Edge Computing, MEC)应运而生^{[4][5]}。在 MEC 中,通过将 MEC 服务器部署在网络边缘处(例如基站,网关等),实现算力下沉。分布式工业设备将本地数据卸载到 MEC 服务器进行计算处理,既解决了本地算力不足问题,又有效地降低了传输时延,提高了网络安全性。

然而,大量的工业任务并发卸载到 MEC 服务器要求网络支持高并发的信道接入,成为工业无线网络面临的核心挑战。当前,国际主流的工业无线网络技术及标准,例如 WirelessHART、ISA100.11a、WIA 等^[6-10],均采用多信道联合时分多址接入的方式,并不足以支持未来海量工业设备的高并发接入和计算卸载。为此,将动态多信道接入^[11-13]应用于 MEC,可望缓解工业无线网络有限时频资源约束问题,实现海量异构工业任务的高并发计算卸载。

目前,动态多信道接入方法主要分为模型驱动法和免模型法。对于模型驱动的动态多信道接入,需要提前获知工业无线网络中关于设备数量、卸载任务、信道质量等的完整系统信息,建立准确的系统模型。在此基础上,采用不同方法求解动态多信道接入的最优策略。因此,最优策略具有特定的模型依赖,仅适用于特定情况下的网络,缺乏处理复杂多变网络的灵活性和可扩展性。相反,免模型的动态多信道接入不需要提前获知完整系统信息,而是通过工业设备与工业无线网络的交互获知部分系统信息,逐步估计系统模型,逼近最优策略。

海量异构工业任务的高并发计算卸载会导致系统模型的状态空间维灾问题,而且状态间的耦合也会加剧。考虑到人工智能(Artificial Intelligence, AI)技术在处理大状态空间和特征耦合问题(例如,图像识别、自然语言处理等)上的卓越表现,设计基于 AI 的免模型动态多信道接入算法成为学术界新的研究热点。强化学习(Reinforcement Learning, RL),作为 AI 中的一种自学习方法,不需要已知的系统模型,利用工业设备与工业无线网络的持续交互近似系统模型,求解动态多信道接入问题。

分布式工业设备向 MEC 服务器的并发卸载过程中,设备与信道接入关系是严格一一对应的,即动作空间是离散的。RL 中通常使用 Q-learning 处理离散动作问题。Q-learning 维护着一个存储历史时隙状态、动作和奖励的表格(称为 Q-table),通过最大化长期累积奖励获得最优策略^[14-17]。然而随着设备数量的增加, Q-table 中的元素数量剧增,状态空间维灾问题加剧,导致基于表格搜索的 Q-learning 求解低效。使用深度学习(Deep Learning, DL)估计系统模型来取代 Q-table,是处理状态空间维灾问题的有效手段。因此,将 RL 和 DL 相结合,使用深度强化学习(Deep Reinforcement Learning, DRL)^[14]实现海量异构工业任务的高并发计算卸载,既发挥了 RL 自学习的优势,又能利用 DL 处理状态空间维灾和特征耦合问题。

另外, MEC 中任务卸载是动态的。异构工业任务的时变性导致工业设备在不同时隙具有不同的卸载需求。然而,现有工作^[16,18-21]很少考虑工业设备的动态卸载需求,即任意时隙全部设备接入信道的概率是相等的。但是,在工业生产中,不同工业设备产生的工业任务具有明显的时延敏感和计算负载差异。而且,任务的时延敏感和计算负载是时变的。因此,本文根据工业任务时延敏感和计算负载的变化,给工业设备分配动态优先级。不同优先级的工业设备接入信道进行计算卸载的概率也是动态的、不相等的。

在动态多信道接入的基础上引入动态优先级进一步加剧了状态空间维灾问题。另外,工业设备仅能获知自己接入信道进行计算卸载是否成功,无法获知其他设备的接入卸载是否成功。综上,状态空间的维灾问题和信息的局部可观测是面向异构工业任务高并发计算卸载的主要挑战。

因此,本文提出基于深度强化学习的动态优先级并发接入算法(Deep Reinforcement Learning-based Concurrent Access with Dynamic Priority, CADP-DRL),其主要目标包括:为时延敏感和计算负载各异的工业设备划分动态优先级;为工业设备分配信道以最小化卸载冲突,保证高优先级设备的实时可靠卸载。

本文首先建立面向异构工业任务高并发计算卸载的系统模型,应用 Markov 决策过程(Markov decision process, MDP)将问题形式化。然后,为 CADP-DRL 构造了兼顾动态优先级和并发卸载的复合奖励函数求解动态优先级并发接入;设计了带

经验权重的经验回放用于加速算法收敛。接下来,详细阐述了 CADP-DRL 的算法结构、训练流程、计算复杂度。最后,与 slotted-Aloha^[18-21]、DQN^[14]、DDQN^[14]、D3QN^[20,21]等算法进行实验比较,对比不同工业场景下的卸载冲突、长期平均累积奖励、优先级权重卸载效用、实时性等性能,验证了 CADP-DRL 的有效性。

本文首次利用深度强化学习方法求解异构工业任务高并发计算卸载问题,主要贡献总结如下:

1、考虑工业生产中时延敏感和计算负载各异的异构工业任务,为工业设备设计不同的优先级。工业设备的优先级是动态的,取决于工业任务时变的时延敏感和计算负载约束。工业设备接入信道进行计算卸载的概率也是动态的,取决于设备优先级。

2、设计兼顾动态优先级和并发卸载的复合奖励函数,在最小化卸载冲突的情况下为高优先级工业设备提供更高的成功卸载概率保证。

3、为保证训练数据的独立同分布,同时加快 CADP-DRL 收敛,设计带经验权重的经验回放方法。根据经验权重将经验分别存储在不同经验池中,动态调整不同经验池的采样概率。

4、与 slotted-Aloha、DQN、DDQN 和 D3QN 比较,对比不同工业场景下的卸载冲突、长期平均累积奖励、优先级权重卸载效用、实时性等指标,验证了 CADP-DRL 的有效性。

本文第 2 节对 MEC 中模型驱动和免模型的动态多信道接入等相关研究工作进行总结。第 3 节和第 4 节分别对工业无线网络中异构工业任务的高并发计算卸载进行建模和 MDP 形式化。第 5 节对 CADP-DRL 算法进行介绍,详细说明了算法结构、训练流程以及计算复杂度。第 6 节对实验结果进行分析,并与 slotted-Aloha、DQN、DDQN 和 D3QN 对比,从卸载冲突、长期平均累积奖励、优先级权重卸载效用、实时性等方面说明 CADP-DRL 处理异构工业任务高并发计算卸载的有效性。第 7 节对本文工作进行了总结。

2 相关工作

工业无线网络内的时频资源是有限的,其所能承载的任务能力也是有限的。动态多信道接入可以提高工业无线网络的时频资源利用率,为更多的工业设备提供服务。

2.1 模型驱动的动态多信道接入方法

对于模型驱动的动态多信道接入方法,需要提前获知完整的系统信息来建立准确的系统模型。在此基础上,使用不同的方法获得最优策略。

短视法(Myopic policy)和 Whittle-Index 法将多信道接入问题转化为多臂匪徒问题。文献[22-25]通过已知的系统信息获得 Markov 转换矩阵最大化多信道接入的吞吐量和算法收敛速度。文献[26]基于任务队列、能量约束和信道状态,采用 Markov 决策过程求解 MEC 中多用户环境下的任务卸载最佳策略以降低能耗和时延。

博弈论关注冲突中的个体决策问题,需要提前获知个体的决策收益。有限时频资源约束下的异构工业任务的高并发计算卸载是典型的冲突决策问题,完整的系统信息能够获得个体决策收益。文献[27,28]使用博弈论解决多用户多接入下的速率最优问题。文献[29,30]在 MEC 资源受限的情况下,提出了基于博弈论多接入卸载算法,能够实现快速收敛、时延最小化和能源效率最大化。

总之,模型驱动的动态多信道接入方法需要完整准确的系统模型,解得的最优策略具有特定的模型依赖,缺乏处理复杂多变网络的灵活性和可扩展性。

2.2 免模型的动态多信道接入方法

对于免模型的动态多信道接入方法,不需要提前获得完整系统信息和准确系统模型。工业设备通过与工业无线网络的交互获知系统的部分信息,可以逐渐估计系统模型。随着交互次数的增加,策略逐渐逼近最优。

文献[31]使用随机采样的蒙特卡洛方法将认知无线电中的信道接入建模为部分可观的 Markov 决策过程,最大化信道接入成功率。与随机采样不同,Q-learning 通过最大化长期累积奖励获取最优策略。文献[15-17]使用 Q-learning 最大化认知无线电中的频谱感知和接入成功率。文献[32]将 MEC 卸载决策建模为多标签分类过程,使用两层全连接神经网络进行训练。文献[33]将 MEC 服务器和物联网设备间的时频资源分配建模为双向拍卖,结合 RL 加速纳什均衡解的求解速度。

但是,状态空间维灾会导致 Q-learning 求解的低效。使用 DL 估计系统模型是处理状态空间维灾的有效手段。文献[18,19]利用深度 Q-learning 网络(Deep Q-learning Network, DQN)最大化无线网络中

单用户的信道接入成功率和传输速率。对于多用户的无线网络,文献[20,21]应用 D3QN 最大化用户个体吞吐量和系统全局吞吐量。文献[34]将 DRL 应用于异构子网内移动用户的多任务卸载,实现移动用户卸载任务的时延最小化。

3 系统模型

工业无线网络系统由工业设备、工业基站和 MEC 服务器组成,如图 1 所示。工业设备分布式部署,产生时延敏感和计算负载各异的异构工业任务,并观测自身局部信息,上报工业基站;工业基站收集系统内全部工业设备上报的信息;MEC 服务器与工业基站连接,利用工业基站收集的信息,离线训练 CADP-DRL,实时处理卸载数据,为分布式工业设备提供算力支持。MEC 服务器中训练好的 CADP-DRL 下载到工业基站中,在线执行异构工业任务的高并发计算卸载。当网络拓扑出现较大变化时,工业基站重新收集各工业设备信息进行 CADP-DRL 的训练更新。

工业生产过程中,工业任务是异构的,具体表现为不同任务的时延敏感和计算负载的差异性、时变性。一般而言,工业任务可以划分为控制消息、音视频流、感知信息等。其中,控制消息(如机器人定位、电机制动等)具有最高的时延敏感性,音视频流(如物料识别、实时监控等)需要高带宽和计算负载保证,感知信息(如设备组态等)需要在约束时间内完成传输^[35,36]。为了动态地调整工业设备的信道接入,保证强时延敏感和高计算负载任务的实时卸载。本文建立工业设备的动态优先级机制,考虑工业生产中时延敏感和计算负载各异的异构工业任务,为工业设备设计了不同的优先级。工业设备的优先级是动态的,取决于工业任务时变的时延敏感和计算负载约束。工业任务的时延越敏感、计算负载越高,工业设备的优先级越高,其成功接入信道进行计算卸载的概率越高。

工业无线网络覆盖范围的 N 台工业设备组成设备集合 $N=\{1,2,\dots,N\}$ 。工业设备根据工业任务的时延敏感和计算负载约束,确定不同的优先级 $p(p \in P, P=\{0,1,\dots,P\})$ 。 p 越大,优先级越高。

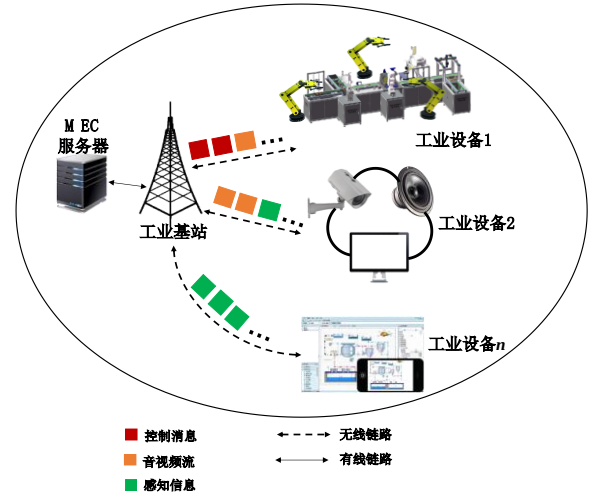


图 1 系统模型

分布式工业设备之间、工业设备与工业基站之间使用工业无线网络连接。卸载所需的时频资源由时域和频域组成,其中时隙 t 是时域的基本单位,信道 c 是频域基本单位,全部信道组成信道集合 $C=\{1,2,\dots,C\}$ 。

由于工业生产是持续进行的,本文假设在任意时隙,全部工业设备都有任务数据产生,且均有接入卸载请求。在时隙 t ,工业设备 n 接入信道 c 成功进行计算卸载需要满足以下约束: 1)只有工业设备 n 接入信道 c ; 2)工业任务卸载的数据量满足信道容量约束; 3)卸载数据被 MEC 服务器成功接收并处理。如果任一约束未满足,卸载将会失败。在本文模型中,重点关注多设备动态优先级的并发接入,即算法对约束 1)的有效性。因此,本文假设工业设备卸载任务的数据量均满足信道容量要求,且只要成功接入信道就能成功卸载任务数据至 MEC 服务器。

当系统内时频资源充足时,即工业设备数小于等于信道数($N \leq C$),所有工业设备都能实时卸载,不会发生卸载冲突。但是当系统内时频资源有限时,即工业设备数大于信道数($N > C$),多个工业设备竞争卸载,卸载冲突不可避免。在每次卸载结束时,工业设备 n 会收到一个局部可观测信息 o_n , $o_n \in \{0,1\}$,指示卸载是否成功。其中, $o_n=1$ 表示卸载成功, $o_n=0$ 表示卸载失败。

在时隙 t 开始时,工业设备 n 根据卸载任务的时延敏感和计算负载需求确定优先级 p ,选择信道 c 接入。在时隙 t 结束时,工业设备 n 会收到局部可观测信息 $o_n(t)$ 。例如,在时隙 t ,工业设备 n 是控制设备,传输具有严格时延敏感的控制消息,拥有高优先级。另外,工业设备 $n+1$ 是音视频设备,传输音视频流。相比于控制消息,音视频流时延敏感

要求低,因此工业设备 $n+1$ 拥有低优先级。时隙 t 结束时,工业设备 n 以其较高的接入概率成功接入信道,获得卸载成功的局部可观测信息,即 $o_n(t)=1$ 。相反,工业设备 $n+1$ 由于较低的接入概率无法接入信道,获得卸载失败的局部可观测信息,即 $o_n(t)=0$ 。

4 MDP 形式化

为了实现异构工业任务高并发计算卸载,工业设备与工业无线网络交互,执行不同的动作实现状态转移。在不同状态下,工业设备执行不同的动作能够获得不同的奖励。最大化长期累积奖励可以获得最优策略。本文使用 MDP 形式化上述状态转移过程,由状态 x 、动作 a 、奖励 r 、状态转移函数 f 描述。

4.1 状态

在时隙 t , $x(t)$ 表示工业设备所处的状态,所有可能存在的状态组成状态空间 X , $x(t) \in X$ 。在本文问题中,工业设备在时隙 t 开始时选择接入的信道 $c(t)$ 、优先级 $p(t)$ 以及全部信道的占用率 $cf(t)$ 组成了工业设备在时隙 t 的状态 $x(t)$ 。状态 $x(t)$ 是一维向量,元素个数为 $2N+C$,如式(1)所示,

$$x(t)=[c(t), p(t), cf(t)] \quad (1)$$

其中, $c(t)$ 是前 N 个元素的集合 V_c 。在时隙 t 开始时,当工业设备 n 选择信道 c 接入卸载, V_c 的第 n 个元素值为 c 。 $p(t)$ 是第 N 个元素到第 $2N$ 个元素的集合 V_p 。当工业设备 n 优先级为 p 时, V_p 的第 n 个元素值为 p 。 $cf(t)$ 是最后 C 个元素的集合 V_{cf} 。每当有工业设备选择信道 c 接入卸载, V_{cf} 的第 c 个元素值加 1。 V_{cf} 的第 c 个元素数值越高,表明选择信道 c 的工业设备越多。

4.2 动作

在时隙 t , $a(t)$ 表示工业设备所执行的动作,所有可能存在的动作组成动作空间 A , $a(t) \in A$ 。在本文问题中,工业设备在时隙 t 结束时实际接入的信道为动作 $a(t)$,即

$$a(t)=[a_1(t), a_2(t), \dots, a_n(t), \dots, a_N(t)] \quad (2)$$

其中 $a_n(t) \in \{0, 1, \dots, C\}$ 。 $a_n(t)=0$ 表示工业设备 n 未能卸载, $a_n(t) \in \{1, 2, \dots, C\}$ 表示工业设备 n 成功卸载。

4.3 复合奖励

在时隙 t , $r(t)$ 表示工业设备在状态 $x(t)$ 执行动作 $a(t)$ 转移到下一状态 $x(t+1)$ 所获得的奖励。奖励为策略更新提供方向,策略要向着奖励增加的方向更新。针对并发卸载和动态优先级的多目标决策,本文设计了包括卸载奖励和优先级奖励的复合奖励函数,以实现最小化卸载冲突和高优先级设备的实时可靠卸载。

4.3.1 卸载奖励

针对并发卸载目标,本文从网络全局和设备个体角度出发,设计了全局卸载奖励和个体卸载奖励。

在全局卸载奖励中,需要考虑全部信道的接入卸载情况。当全部信道都有工业设备接入时给予奖励,当有信道空闲浪费时给予惩罚。本文定义卸载冲突 col ,即工业设备总数 N 与网络中实时卸载的工业设备数 m ($m \in \mathbb{N}_+$ and $m \leq N$) 之间的差,用于衡量全部信道的接入卸载情况。当全部信道均有设备接入卸载时,给予奖励;反之,当任一信道处于空闲,未有设备接入卸载时,给予惩罚。

根据以上定义,当网络中实时卸载的工业设备数等于信道数时(即 $m=C$),卸载冲突最小,即 $col=N-C$ 。当 $col=N-C$ 时,全局卸载奖励是一个很大的正奖励 R_g ;当 $col>N-C$ 时,则表示有信道处于空闲,全局卸载奖励为 0,如式(3),

$$r_g(t) = \begin{cases} 0, & \text{if } col > N - C \\ R_g, & \text{if } col = N - C \end{cases} \quad (3)$$

相比于全局卸载奖励,个体卸载奖励用于奖励工业设备个体的成功卸载。在个体卸载奖励中,当工业设备 n 在时隙 t 成功卸载时,收到正奖励;反之会收到负奖励。个体卸载奖励 $r_i(t)$ 如式(4),

$$r_i(t) = \sum_{n=0}^{N_s} r_n(t) \quad (4)$$

其中, $r_n(t) = \begin{cases} 0, & \text{if } o_n(t) \neq 1 \\ 1, & \text{if } o_n(t) = 1 \end{cases}$; N_s 表示成功接入信道进行计算卸载的工业设备。

4.3.2 优先级奖励

工业设备优先级是动态的。工业任务的时延越敏感、计算负载越高,对应地工业设备优先级越高,成功卸载的概率越高。因此优先级奖励 $r_p(t)$ 与工业设备优先级为正相关关系,即接入卸载的设备优先级越高,优先级奖励越大,如式(5),

$$r_p(t) = \sum_{n=0}^{N_s} p_n(t) \quad (5)$$

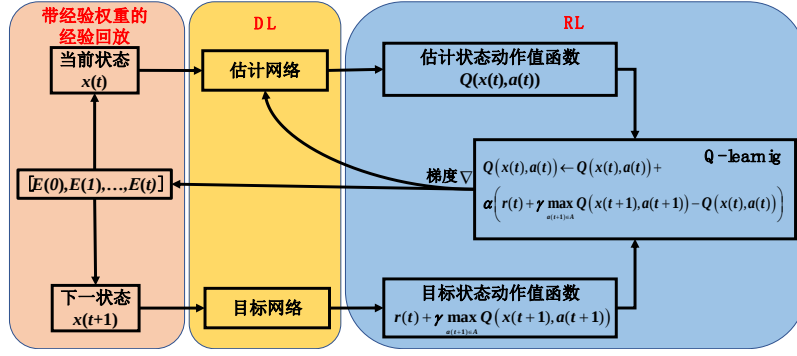


图2 CADP-DRL 算法结构

其中， $p_n(t)$ 表示工业设备 n 在时隙 t 的优先级 p 。

综上，定义工业设备在时隙 t 的奖励如式(6)，

$$r(t) = r_g(t) + r_i(t) + r_p(t) \quad (6)$$

由于工业生产过程的连续性和异构任务的时变性，工业设备在不同时隙的信道选择与优先级是随机的、时变的。不同状态下，设备采取不同的动作会收到差异显著的奖励。我们通过最大化长期累积奖励即可获得异构工业任务高并发计算卸载的最优策略，如式(7)所示，

$$R(t) = \sum_{\tau=t_0}^t \gamma^{\tau-t_0} r(\tau) \quad (7)$$

其中， $R(t)$ 表示长期累积奖励； γ 表示折扣因子，表明过往时隙奖励对当前时隙奖励的影响程度； τ 表示时隙； $r(\tau)$ 为从状态 $x(\tau)$ 到状态 $x(\tau+1)$ 获得的奖励。

4.4 状态转移函数

在时隙 t ，状态转移函数 $f(t)$ 表示工业设备执行动作 $a(t)$ 由状态 $x(t)$ 转移到状态 $x(t+1)$ 的概率，即 $f(x(t+1)|x(t), a(t))$ 。随着交互次数增加，通过最大化长期累积奖励， $f(t)$ 逐渐收敛到 $f^*(t)$ ，

$$f^*(t) \leftarrow \max \sum_{\tau=0}^t R(\tau) \quad (8)$$

5 CADP-DRL 算法

基于上述 MDP 形式化，我们提出 CADP-DRL 算法求解异构工业任务高并发计算卸载的最优策略。CADP-DRL 算法由带经验权重的经验回放、DL、RL 组成，结构如图 2 所示。

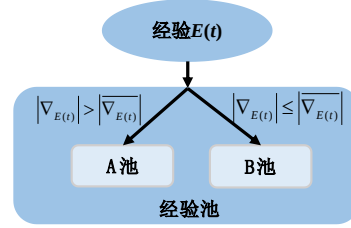


图3 带经验权重的经验池

5.1 带经验权重的经验回放

将工业设备与工业无线网络交互产生的状态、动作、奖励等作为经验 E 存储在经验池 H 中。经验将作为深度神经网络训练数据。在时隙 t ，经验 $E(t)$ 和经验池 $H(t)$ 表示为

$$E(t) = (x(t), a(t), r(t), x(t+1)) \quad (9)$$

$$H(t) = (E(0), E(1), \dots, E(t)) \quad (10)$$

为了保证深度神经网络的训练数据是独立同分布的，本文采用经验回放从经验池中随机采样获得经验进行训练，以此来打破经验池中的经验的时间相关性。

但是，标准经验回放中每条经验被采样的概率是相同的，导致算法收敛较慢。本文考虑到不同的经验对收敛的贡献不同，将每条经验的下降梯度 ∇ 作为经验权重^[37]。然后构造了两个经验池 A 池、B 池，分别存放高权重经验和低权重经验，如图 3 所示。相比于 Sum Tree^[37]方法，经验采样和更新的计算复杂度由 $O(\log_2 N)$ 下降到 $O(1)$ 。

在训练开始阶段，从 A 池、B 池中采样的概率是相同的。为了加快收敛，随着训练次数的增加，逐渐增加 A 池中高权重经验的采样概率，同时逐渐减少 B 池中低权重经验的采样概率。经验采样概率如式(11)所示，

$$g_{A,B}^r = g_{A,B}^{ini} - g_{A,B}^{dec} \quad (11)$$

其中 $0 < g_{A,B}^r \leq 1$, 表示 A 池、B 池的经验采样概率; $g_{A,B}^{ini}$ 表示 A 池、B 池的初始采样概率; $g_{A,B}^{dec}$ 表示 A 池、B 池的采样衰减率。

5.2 深度神经网络

状态遵循策略 π 实现到动作的映射, 即 $a(t) \sim \pi(x(t))$ 。本文使用状态动作值函数 $Q_\pi(x(t), a(t))$ 来评价策略。 $Q_\pi(x(t), a(t))$ 表示遵循策略 π 的情况下, 在状态 $x(t)$ 执行动作 $a(t)$ 获得的长期累积奖励。

当并发卸载任务较少时, 采用基于 Q-table 的 Q-learning 方法可有效实现计算卸载。但是, 当网络内并发卸载任务数量剧增时, Q-learning 求解变得低效。因此, CADP-DRL 采用深度神经网络估计状态动作值函数 $Q_\pi(x, a)$, 即 $Q_\pi(x, a; \theta) \approx Q_\pi(x, a; \theta^*)$, 其中, θ 为神经网络的参数集合。

估计网络和目标网络的网络结构都如图 4 所示。它们结构相同, 但是参数不同。估计网络和目标网络分别用于获得当前时隙 t 的估计状态动作值函数 $eval_Q = Q_\pi(x(t), a(t); \theta)$ 和下一时隙 $t+1$ 的目标状态动作值函数 $target_Q = r(t) + \gamma \max_{a(t+1) \in A} Q_\pi(x(t+1), a(t+1); \theta^-)$ 。估计网络参数 θ 是实时更新的, 目标网络参数 θ^- 是估计网络 S 步之前的旧网络参数, 参数的异步更新能够避免模型震荡和发散。

CADP-DRL 的损失函数定义为:

$$L(\theta) = E[(target_Q - eval_Q)^2] \quad (12)$$

接下来, 基于公式(12)计算参数 θ 的梯度, 如式(13)。最后使用随机梯度下降更新参数 θ , 最终获得最优策略 π^* 。

$$\nabla_\theta L(\theta) = E[(target_Q - eval_Q) \nabla_\theta eval_Q] \quad (13)$$

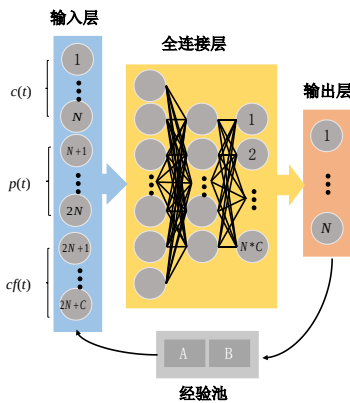


图 4 深度神经网络结构

5.3 Q-learning

状态 $x(t)$ 在经历两层全连接层特征提取后转变为状态 $x(t)$ 中全部设备 N 在全部动作下的状态动作值函数 $Q_\pi(x(t), \forall a(t) \in A)$ 。然后, 对于任意设备, 输出层使用 ε -greedy 方法从其全部动作中选择一个具体的动作, 如式(14),

$$a(t) = \begin{cases} \text{random select} & \varepsilon \\ \max_{a(t)} Q_\pi(x(t), \forall a(t) \in A) & 1 - \varepsilon \end{cases} \quad (14)$$

其中, ε 表示探索概率, 用于平衡探索新动作和利用已知最大状态动作值的动作之间的比例关系。

通过 ε -greedy 方法选择动作 $a(t)$ 后, 在状态 $x(t)$ 执行 $a(t)$ 转移至下一状态 $x(t+1)$ 。随着状态从 $x(t)$ 转移到 $x(t+1)$, $Q_\pi(x(t), a(t))$ 按照贝尔曼方程^[38]更新,

$$Q_\pi(x(t), a(t)) \leftarrow Q_\pi(x(t), a(t)) + \alpha \left(r(t) + \gamma \max_{a(t+1) \in A} Q_\pi(x(t+1), a(t+1)) - Q_\pi(x(t), a(t)) \right) \quad (15)$$

其中 α 表示学习速率, 取值范围为 $(0, 1]$ 。

随着训练次数的不断增加, 状态动作值函数会逐渐收敛, 策略 π 也趋向最优策略 π^* , 即

$$\pi^* \leftarrow \max_{\pi} \sum_{\tau=t_0}^t Q_\pi(x(\tau), a(\tau)) \quad (16)$$

5.4 算法训练

在 CADP-DRL 中, 工业设备与工业无线网络交互产生的经验分类存储在经验池中, 带经验权重的经验回放从经验池中随机采样经验作为训练数据。在训练数据中, $x(t)$ 输入估计网络产生估计状态动作值函数 $eval_Q$, $x(t+1)$ 输入目标网络产生目标状态动作值函数 $target_Q$ 。Q-learning 利用 $eval_Q$ 和 $target_Q$ 产生梯度 ∇ 实时训练估计网络。同时, 经验池中的经验得到维护和更新。另外, 目标网络的训练更新与估计网络是异步的, 它使用估计网络 S 步之前的旧网络参数。完整的训练过程如算法 1 所示。

算法 1. 训练过程.

初始化: 训练次数 U , batch_size M , 经验池 H , 估计网络参数 θ , 目标网络参数 θ^- , 更新步数 S

输入: 工业设备状态 x

输出: 工业设备动作 a

for $u = 0, \dots, U$:

 从经验池 A 池、B 池采样 M 条经验作为训练数据;

 for $s = 0, \dots, S$:

经验中当前状态 $x(t)$ 输入估计网络获得 $Q_\pi(x(t), \forall a(t) \in A)$;

根据公式(14)选择动作 $a(t)$ ，执行动作 $a(t)$ 获得奖励 $r(t)$ ，到达下一状态 $x(t+1)$;

存储或更新经验 $(x(t), a(t), r(t), x(t+1))$ 到经验池 H 中;

将下一状态 $x(t+1)$ 输入目标网络获得 $Q_\pi(x(t+1), \forall a(t+1) \in A)$ ，根据公式(14)选择动作 $a(t+1)$;

根据公式(13)更新估计网络的参数 θ ;

状态转移， $x(t) = x(t+1)$;

end for

目标网络的参数 θ^- 每 S 次迭代后复制估计网络参数 θ ;

end for

5.5 计算复杂度

本文使用计算复杂度衡量 CADP-DRL 的执行效率。在离线训练和在线执行阶段，状态到动作的映射都是通过深度神经网络实现的。首先，我们定义深度神经网络的计算复杂度 $O(G)$ 。 $O(G)$ 与神经网络层数和每层的神经元数量相关^[21],

$$O(G) = O\left(\sum_{f=1}^F d_f d_{f+1}\right) \quad (17)$$

其中，表示 F 表示神经网络层数； d_f 表示第 f 层神经网络的神经元数。

在离线训练阶段，使用 M 条采样经验训练 U 次，且每隔 S 步更新目标网络参数，CADP-DRL 的计算复杂度为 $O(GMU^S)$ 。但是，算法训练是在 MEC 服务器上离线进行的，不会对算法的实时在线执行产生干扰。因此，CADP-DRL 在线执行时的计算复杂度仅为 $O(G)$ 。

6 实验结果分析

本文实验考虑工业无线网络中信道资源紧张情况，即工业设备数大于信道数的情况。实验场景设置如下：

- 场景 1：生产监控系统，系统内包括音视频监控设备、感知监测设备等。各类设备通过工业无线网络相连，网络内包含 6 台工业设备，3 个信道。网络内主要传输监控产生的音视频流和感知信息等，优先级在 0~3 范围内随机分布；
- 场景 2：离散制造系统，系统内包括控制设备、

物流设备、音视频监控设备、感知监测设备等。各类设备通过工业无线网络相连，网络内包含 10 台工业设备，5 个信道。网络内传输的任务数据涵盖控制信息、音视频流和感知信息等，优先级在 0~7 范围内随机分布。

为验证 CADP-DRL 的有效性，本文选择通信领域经典 slotted-Aloha 算法^[18-21]，DRL 领域的标准 DQN 算法^[14]、Double DQN(DDQN)算法^[14]、Double Dueling DQN(D3QN)^[20,21]算法作为基线算法。基于 Intel Xeon W2245、NVIDIA Titan RTX 和 Tensorflow-GPU-1.14 的实验平台，本文通过实验对比分析 CADP-DRL 算法与基线算法的性能差异，CADP-DRL 的参数设置如表 1。

表 1 实验参数

参数	值
学习速率 α	0.001
折扣因子 γ	0.9
探索概率 $\mathcal{E}$	0.1
A 池容量	10000
B 池容量	10000
A 池初始采样概率 g_A^{ini}	0.5
B 池初始采样概率 g_B^{ini}	0.5
A 池采样衰减率 g_A^{dec}	-0.001/每 1000 次训练
B 池采样衰减率 g_B^{dec}	0.001/每 1000 次训练
Batch size M	128

6.1 卸载冲突

为了便于观测卸载冲突的变化趋势，本文使用平均卸载冲突描述卸载冲突随着训练次数增加的变化趋势，定义如下：

$$\overline{col} = \frac{\sum_{i=0}^I col_i}{I} \quad (18)$$

其中， $\overline{col}$ 表示在 I 次训练内的平均卸载冲突； col_i 表示第 i 次迭代时的卸载冲突。

特别地，本文定义最优卸载概率：

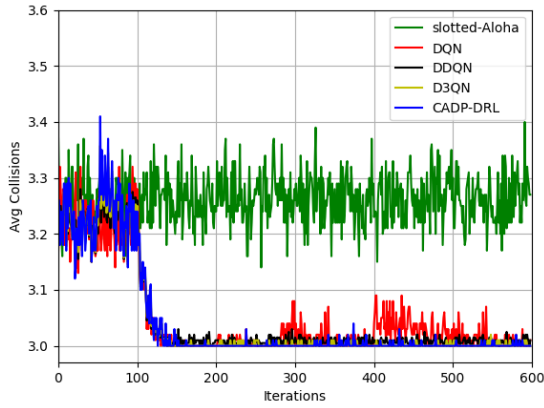
$$K = \frac{\text{count}(N - C | col_I)}{I} \quad (19)$$

其中， K 表示最优卸载概率； col_I 表示 I 次训练产生的卸载冲突集合； $\text{count}(N - C | col_I)$ 表示 I 次训练中最小卸载冲突出现的次数。最优卸载概率越高，算法收敛性越好。

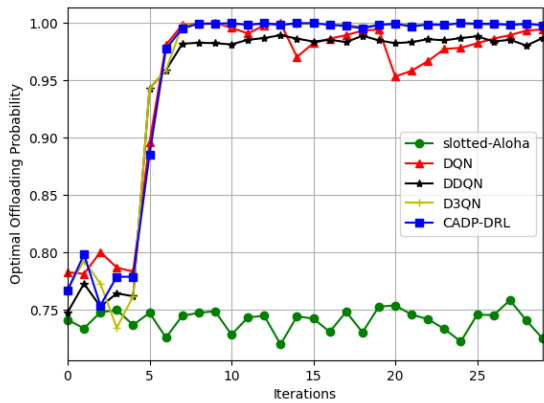
图 5、6 所示的是在场景 1、2 中应用不同算法求解异构工业任务高并发计算卸载的平均卸载冲

突和最优卸载概率的变化情况。

如图 5 所示,在场景 1 的训练初始阶段,slotted-Aloha、DQN、DDQN、D3QN 和 CADP-DRL 的平均卸载冲突均较大,且波动剧烈。slotted-Aloha 的最优卸载概率在 70%~75%左右,而 DQN、DDQN、D3QN 和 CADP-DRL 的最优卸载概率迅速提升到 90% 以上。随着训练次数的增加,slotted-Aloha 的随机接入特性使其平均卸载冲突仍然波动较大,无法收敛到最小卸载冲突,最优卸载概率仍然在 70%~75%左右。相反,DQN、DDQN、D3QN 和 CADP-DRL 迅速收敛到最小卸载冲突附近,仅在小范围内波动。相比于 DQN、DDQN, CADP-DRL 和 D3QN 的平均卸载冲突收敛更紧致,波动范围更小。同时, CADP-DRL 和 D3QN 的最优卸载概率接近 100%。

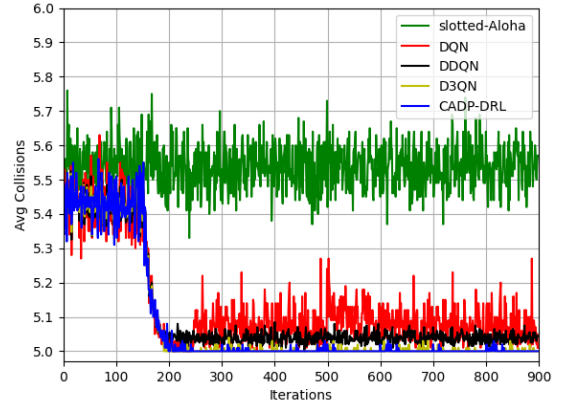


(a) 平均卸载冲突 $\overline{col}$

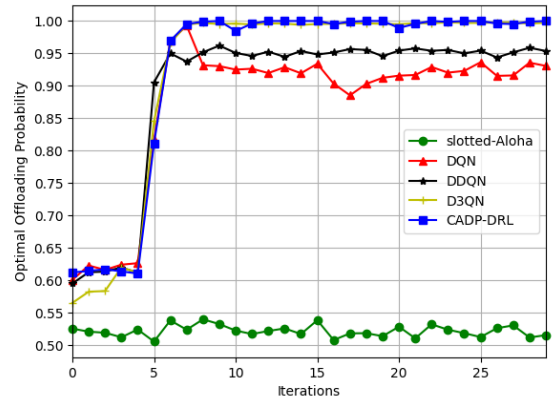


(b) 最优卸载概率 K

图 5 场景 1 下的平均卸载冲突和最优卸载概率



(a) 平均卸载冲突 $\overline{col}$



(b) 最优卸载概率 K

图 6 场景 2 下的平均卸载冲突和最优卸载概率

场景 2 中,由于工业场景中设备、信道、优先级规模的增加,状态空间维度增加。如图 6 所示,在场景 2 的训练初始阶段,slotted-Aloha、DQN、DDQN、D3QN 和 CADP-DRL 的平均卸载冲突均较大,且波动剧烈。由于 slotted-Aloha 可随机选择的状态增加,导致最优卸载概率低于 55%,收敛较差。但是,DQN、DDQN、D3QN 和 CADP-DRL 的最优卸载概率仍然能迅速提升到 90% 以上。随着训练次数的增加,slotted-Aloha 的随机接入特性使其平均卸载冲突仍然波动较大,无法收敛到最小卸载冲突,最优卸载概率仍然低于 55%。相反,DQN、DDQN、D3QN 和 CADP-DRL 迅速收敛到最小卸载冲突,仅在小范围内波动。相比于 DQN、DDQN, CADP-DRL 和 D3QN 的平均卸载冲突收敛更紧致,波动范围更小。同时, CADP-DRL 和 D3QN 的最优卸载概率接近 100%。

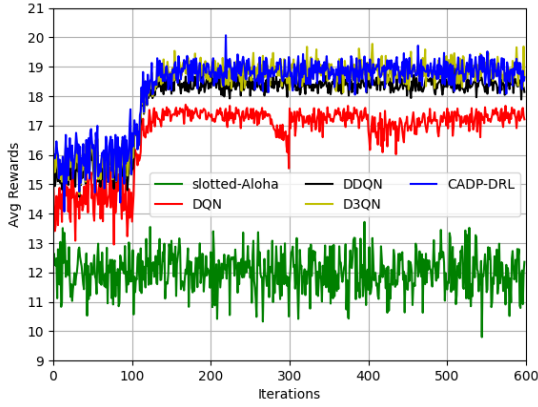
6.2 长期平均累积奖励

长期累积奖励用于衡量策略的有效性。本文使用长期平均累积奖励验证策略的有效性，定义如下：

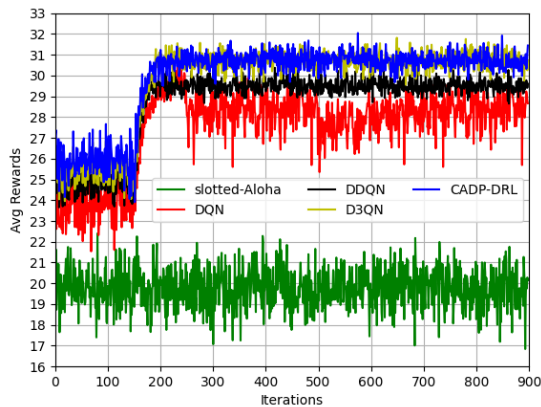
$$\overline{R(I)} = \frac{R(I)}{I} \quad (20)$$

其中， $\overline{R(I)}$ 表示长期平均累积奖励； $R(I)$ 表示 I 次迭代的长期累积奖励。

如图 7 所示，无论是在场景 1 还是场景 2 中，slotted-Aloha 都具有较低的长期平均累积奖励且发散。相比于 slotted-Aloha，DQN、DDQN、D3QN 和 CADP-DRL 的长期平均累积奖励能够收敛。CADP-DRL 的收敛速度更快，更紧致，与 D3QN 性能相近。在相同场景下，CADP-DRL 的长期平均累积奖励普遍高于 DQN、DDQN。



(a) 长期平均累积奖励 $\overline{R(I)}$ (场景 1)



(b) 长期平均累积奖励 $\overline{R(I)}$ (场景 2)

图 7 长期平均累积奖励

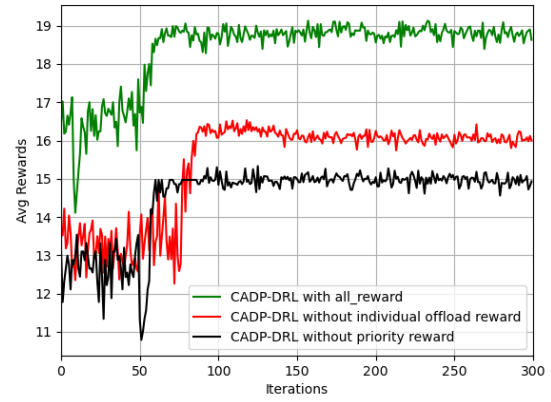
复合奖励函数中不同奖励对并发卸载和动态优先级具有不同的影响。本文在场景 1 环境下，通

过消融实验验证了个体卸载奖励、优先级奖励对于卸载冲突收敛和优先级权重卸载的贡献。

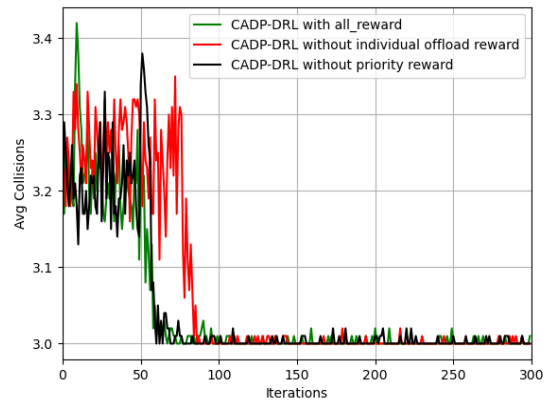
如图 8(a)(b)所示，取消个体卸载奖励后，CADP-DRL 的长期平均累积奖励值下降且收敛速度明显变慢。但是由于全局卸载奖励的存在，CADP-DRL 仍然能够收敛到最小卸载冲突。

另外，如图 8(a)(c)所示，取消优先级奖励后，CADP-DRL 的长期平均累积奖励显著下降，但收敛速度与全部奖励均存在时的收敛速度相同。CADP-DRL 仍然能够收敛到最小卸载冲突。但是，CADP-DRL 的优先级权重卸载效用值增加，且与 slotted-Aloha 的效用值很接近。

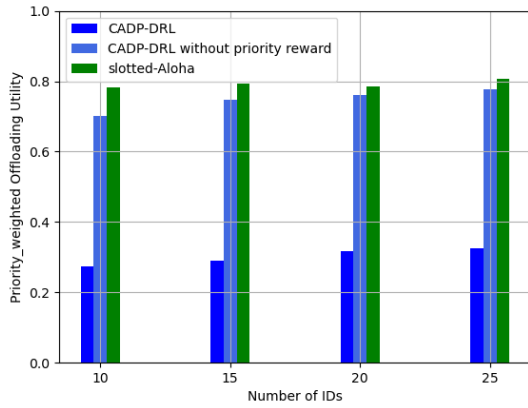
综上，全局卸载奖励和个体卸载奖励作用于卸载冲突收敛，实现最小卸载冲突；优先级奖励保证高优先级设备的优先接入卸载，为高优先级工业设备提供更高的卸载概率保证。



(a) 长期平均累积奖励



(b) 平均卸载冲突



(c) 优先级权重卸载效用

图8 奖励贡献分析

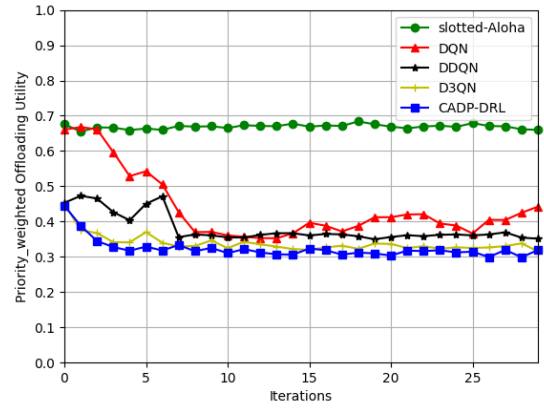
6.3 优先级权重卸载效用

在实现全部信道的成功接入后, 还要保证高优先级工业设备具有更高的成功卸载概率。本文使用优先级权重卸载效用衡量 slotted-Aloha、DQN、DDQN、D3QN 和 CADP-DRL 的优先级卸载保证。

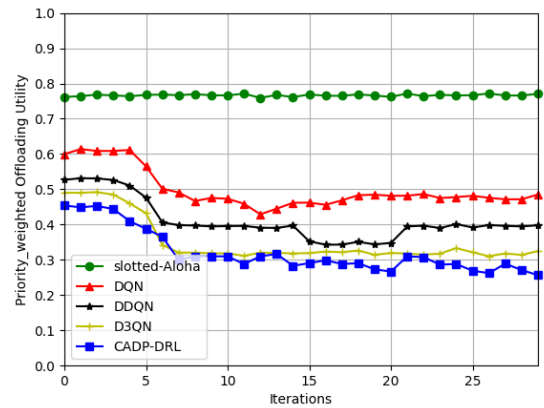
优先级权重卸载效用充分考虑了成功卸载设备的优先级在全部工业设备优先级中的权重等级, 其计算方法如下: 首先, 将时隙 t 系统内全部工业设备的优先级降序排列, 生成序列 p^1 ; 然后, 将时隙 t 系统内成功卸载的工业设备的优先级降序排列, 生成序列 p^2 , 序列长度为 b ; 依次从序列 p^2 中取出元素, 找到该元素在序列 p^1 中第一次出现时的索引, 将其存储在索引列表 $index$ 中, 并将序列 p^1 中该索引位置的元素置为 -1; 依次计算 $index$ 和序列 p^1 中前 b 个索引的平方差, 累积求和得到 sum ; 最后, 对 sum 进行归一化处理。

根据以上优先级权重卸载效用计算过程, 效用值越接近 0, 卸载设备的优先级越高。反之, 效用值越接近 1, 卸载设备的优先级越低。

图 9(a)所示为场景 1 下优先级权重卸载效用。slotted-Aloha 的效用值较大, 仅能为高优先级设备提供约 33%的卸载概率保证。随着训练次数的增加, DQN、DDQN、D3QN 和 CADP-DRL 的效用值逐渐收敛。DQN、DDQN 和 D3QN 能够为高优先级设备提供约 60%、65%和 68%的卸载概率保证。特别地, CADP-DRL 效用值最低, 能够为高优先级设备提供约 70%的卸载概率保证。



(a) 优先级权重卸载效用(场景 1)



(b) 优先级权重卸载效用(场景 2)

图9 优先级权重卸载效用

类似地, 图 9(b)所示为场景 2 下优先级权重卸载效用。由于设备数量和优先级范围的增加, slotted-Aloha 效用值相较于场景 1 更大, 仅能为高优先级设备提供约 23%的卸载概率保证。随着训练次数的增加, DQN、DDQN、D3QN 和 CADP-DRL 的效用值逐渐收敛。DQN、DDQN 和 D3QN 能够为高优先级设备提供约 53%、60%和 68%的卸载概率保证。特别地, CADP-DRL 的效用值仍然是最低的, 能够为高优先级用户提供约 70%的卸载概率保证。

综上, CADP-DRL 能够为高优先级工业设备提供更高的卸载概率保证。

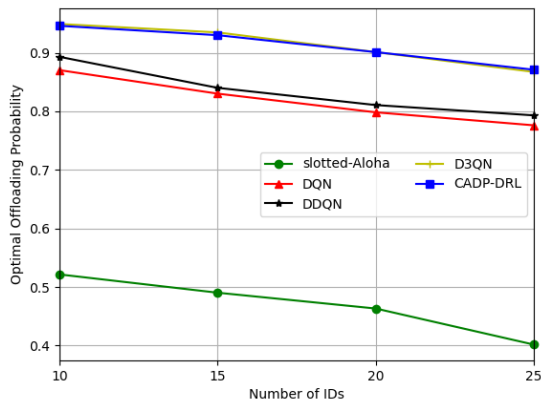
6.4 算法泛化性与实时性

为了验证 CADP-DRL 具有较强的可扩展性, 能够适应网络环境的动态变化。本文使用场景 2 中训练好的 CADP-DRL, 固定信道数量和优先级范围, 改变设备数量, 验证其在设备数分别为 $N=15$, $N=20$, $N=25$ 下的最优卸载概率和优先级权重卸载效用等性能。

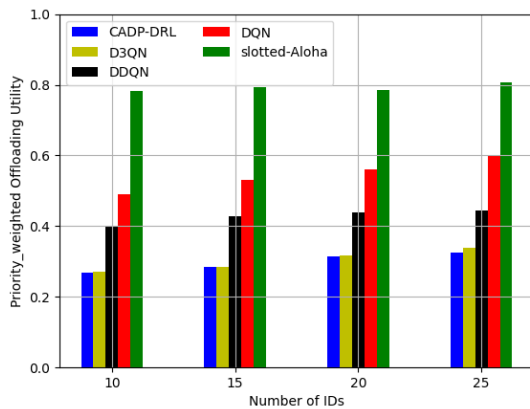
图 10(a)所示为不同设备数下的最优卸载概率。随着设备数量的增加, 卸载冲突发生的概率逐渐增加, 导致最优卸载概率逐渐下降。但是, 相比于 slotted-Aloha、DQN 和 DDQN, CADP-DRL 和 D3QN 能够提供最高的最优卸载概率。即使当设备数量远大于信道数量时, CADP-DRL 的最优卸载概率仍为 87% 左右。

类似的, 图 10(b)所示为不同设备数下的优先级权重卸载效用。随着设备数量的增加, 卸载冲突发生的概率逐渐增加, 高优先级设备卸载面临更严重的信道竞争, 效用值逐渐增加。但是, 相比于 slotted-Aloha、DQN、DDQN 和 D3QN, CADP-DRL 的效用值仍然是最低的(即最高的卸载概率保证), 能够为高优先级设备提供 70% 左右的卸载概率保证。

综上, 在不同规模的工业无线网络环境中, CADP-DRL 能够最大程度实现最小卸载冲突, 同时为高优先级工业设备提供更高的卸载概率保证。



(a) 不同设备数下的最优卸载概率



(b) 不同设备数下的优先级权重卸载效用

图 10 算法泛化性

表 2 各算法模型加载时间和物理执行时间

实验场景	算法	模型加载时间 (ms)	物理执行时间 (ms)
$N=6, C=3, P=3$	DQN	668	1.687
	CADP-DRL	671	1.701
	DDQN	761	1.833
	D3QN	863	1.963
$N=10, C=5, P=7$	DQN	668	1.701
	CADP-DRL	671	1.719
	DDQN	761	1.861
	D3QN	863	2.037

由于异构工业任务的高并发计算卸载需要实时在线进行, 本文分别测试了场景 1 和场景 2 环境下各算法的模型加载时间和物理执行时间, 如表 2 所示。

由表 2 可知, DQN 和 CADP-DRL 的模型加载时间和物理执行时间几乎相同, 是各算法中最短的。D3QN 的模型加载时间和物理执行时间都是最长的。DDQN 的模型加载时间和物理执行时间介于 CADP-DRL 和 D3QN 之间。其原因如下: DQN 和 CADP-DRL 仅使用全连接层作为隐藏层, 算法结构简单, 其模型加载时间和物理执行时间最短。DDQN 在 DQN 基础上增加了一个 DQN, 形成复杂的 Double 结构, 造成 DDQN 模型加载时间和物理执行时间显著增加。类似地, D3QN 在 DDQN 基础上又增加了 Dueling 结构, 导致模型加载时间和物理执行时间进一步增加。

特别地, 从 6.1~6.3 节可知, CADP-DRL 和 D3QN 在卸载冲突、长期平均累积奖励、优先级权重卸载效用等方面性能相近。但是, CADP-DRL 算法结构更加轻量化。相比于 D3QN, CADP-DRL 的模型加载时间缩短 192ms, 物理执行时间提效 15% 以上。随着网络规模的增加, CADP-DRL 的物理执行时间提升更为明显, 实时性能明显优于 D3QN, 更适合异构工业任务的高并发计算卸载。

7 结束语

海量异构工业任务向 MEC 服务器的高并发计算卸载造成了状态空间维灾和信息局部可观测问题, 成为工业无线网络有限时频资源约束下影响工业生产安全的关键。针对上述问题, 本文提出基于深度强化学习的动态优先级并发接入算法 (CADP-DRL)。首先建立了面向异构工业任务高并

发计算卸载模型。通过分析工业任务的时延敏感和计算负载,为工业设备分配不同的优先级,动态地改变工业设备接入卸载的概率。然后,利用 Markov 决策过程形式化动态优先级高并发计算卸载问题,并采用深度强化学习方法建立高维状态空间下状态到动作的映射。针对动态优先级和并发卸载的多目标决策问题,设计了包含优先级奖励和卸载奖励的复合奖励函数。为保证训练数据的独立同分布,同时提高算法收敛速度,设计了带经验权重的的经验回放方法。对比实验结果表明, CADP-DRL 能够快速收敛,实时响应,在实现最小卸载冲突的情况下为高优先级工业设备提供最高的成功卸载概率保证,性能优于 slotted-Aloha、DQN、DDQN 和 D3QN 算法。

致 谢 感谢国家重点研发计划(2020YFB1710900)、国家自然科学基金(61803368, 61972389, 61903356, U1908212)、中国博士后科学基金(2019M661156)、辽宁省‘兴辽英才计划’项目(XLYC1801001)和中国科学院青年创新促进会对本研究的支持。

参 考 文 献

- [1] Yaqoob I, Hashem I A T, Ahmed A, et al. Internet of Things forensics: Recent advances, taxonomy, requirements, and open challenges. *Future Generation Computer Systems*, 2019, 92(1): 265-275.
- [2] Yao X, Zhou J, Lin Y, et al. Smart manufacturing based on cyber-physical systems and beyond. *Journal of Intelligent Manufacturing*, 2019, 30(8): 2805-2817.
- [3] Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. *Communications of the ACM*, 2013, 53(4): 50-58.
- [4] Porambage P, Okwuibe J, Liyanage M, et al. Survey on multi-access edge computing for Internet of Things realization. *IEEE Communications Surveys and Tutorials*, 2018, 20(4): 2961-2991.
- [5] Shi W, Cao J, Zhang Q, et al. Edge computing: vision and challenges. *IEEE Internet of Things Journal*, 2016, 3(5): 637-646.
- [6] International Electrotechnical Commission, Industrial Communication Networks—Wireless Communication Network and Communication Profiles—WirelessHart, IEC 62591 Edition 2.0, 2016.
- [7] International Electrotechnical Commission, Industrial Networks—Wireless Communication Network and Communication Profiles—ISA100.11a, IEC 62734 Edition 1.1, 2019.
- [8] Liang W, Zhang X, Xiao Y, Wang F, Zeng P, and Yu H. Survey and experiments of WIA-PA specification of industrial wireless network. *Wireless Communications and Mobile Computing*, 2010, 11(8): 1197-1212.
- [9] Liang W, Zheng M, Zhang J, Shi H, Yu H, et al. WIA-FA and its applications to digital factory: A wireless network solution for factory automation. *Proceedings of the IEEE*, 2019, 107(6): 1053-1073.
- [10] Raza M, Aslam N, Le-Minh H, et al. A critical analysis of research potential, challenges, and future directives in industrial wireless sensor networks. *IEEE Communications Surveys and Tutorials*, 2017 20(1): 39-95.
- [11] Incel O D. A survey on multi-channel communication in wireless sensor networks. *Computer Networks*, 2011, 55(13): 3081-3099.
- [12] Kaur A and Kumar K. A comprehensive survey on machine learning approaches for dynamic spectrum access in cognitive radio networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 2020, 1(10): 1-40.
- [13] Ahmad W S H M W, Radzi N A M, Samidi F S, et al. 5G technology: towards dynamic spectrum sharing using cognitive radio networks. *IEEE Access*, 2020, 8(1): 14460-14488.
- [14] Liu Jian-wei, Gao Feng and Luo Xiong-lin. Survey of deep reinforcement learning based on value function and policy gradient. *Chinese Journal of Computers*, 2019, 42(6):1406-1438. (in Chinese).
- (刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述. *计算机学报*, 2019, 42(6):1406-1438.)
- [15] Li F, Lam K Y, Sheng Z, et al. Q-learning-based dynamic spectrum access in cognitive industrial internet of things. *Mobile Networks and Applications*, 2018, 23(6): 1636-1644.
- [16] Slimeni F, Chtourou Z, Scheers B, et al. Cooperative Q-learning based channel selection for cognitive radio networks. *Wireless Networks*, 2019, 25(7): 4161-4171.
- [17] Khurana S and Upadhyaya S. Spectrum management in cognitive radio ad-hoc network using Q-learning. *International Journal of Information Technology*, 2020, 12(2): 599-604.
- [18] Wang S and Lv T. Deep reinforcement learning based dynamic multichannel access in HetNets. //Proceedings of the 2019 IEEE Wireless Communications and Networking Conference (WCNC). Marrakech, Morocco, 2019: 1-6.
- [19] Wang S, Liu H, Gomes P H, et al. Deep reinforcement learning for dynamic multichannel access in wireless networks. *IEEE Transactions on Cognitive Communications and Networking*, 2018, 4(2): 257-265.
- [20] Naparstek O and Cohen K. Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks. //Proceedings of the 2017 IEEE Global Communications Conference. Singapore, 2017: 1-7.
- [21] Naparstek O and Cohen K. Deep multi-user reinforcement learning for distributed dynamic spectrum access. *IEEE Transactions on Wireless Communications*, 2018, 18(1): 310-323.
- [22] Zhao Q, Krishnamachari B and Liu K. On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance. *IEEE Transactions on Wireless Communications*, 2008, 7(12): 5431-5440.
- [23] Wang K, Lin C and Yu J. On optimality of myopic policy in

- multi-channel opportunistic access. IEEE Transactions on Communications, 2016, 65(2): 677-690.
- [24] Liu K and Zhao Q. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. IEEE Transactions on Information Theory, 2010, 56(11): 5547-5567.
- [25] Meshram R, Manjunath D, and Gopalan A. On the whittle index for restless multiarmed hidden Markov bandits. IEEE Transactions on Automatic Control, 2018, 63(9): 3046-3053.
- [26] Liu J, Mao Y, Zhang J, et al. Delay-optimal computation task scheduling for mobile-edge computing systems. //Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT). Barcelona, Spain, 2016: 1451-1455.
- [27] Cohen K and Leshem A. Distributed game-theoretic optimization and management of multichannel Aloha networks. IEEE/ACM Transactions on Networking, 2015, 24(3): 1718-1731.
- [28] Bistritz I and Leshem A. Approximate best-response dynamics in random interference games. IEEE Transactions on Automatic Control, 2017, 63(6): 1549-1562.
- [29] Wang Y, Lang P, Tian D, et al. A game-based computation offloading method in vehicular multi-access edge computing networks. IEEE Internet of Things Journal, 2020, 7(6): 4987-4996.
- [30] Chen X, Jiao L, Li W, et al. Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Transactions on Networking, 2015, 24(5): 2795-2808.
- [31] Wang X, Wong A and Ho P. H. Dynamic Markov-chain Monte Carlo channel negotiation for cognitive radio. //Proceedings of the 2010 INFOCOM IEEE Conference on Computer Communications Workshops. San Diego, USA, 2010: 1-5.
- [32] Yu S, Wang X and Langar R. Computation offloading for mobile edge computing: A deep learning approach. //Proceedings of the 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications. Montreal, Canada, 2017: 1-6.
- [33] Li Q, Yao H, et al. Reinforcement and belief learning-based double auction mechanism for edge computing resource allocation. IEEE Internet of Things Journal, 2020, 7(7): 5976-5985.
- [34] Li Y, Qi F, Wang Z, et al. Distributed edge computing offloading algorithm based on deep reinforcement learning. IEEE Access, 2020, 8(1): 85204-85215.
- [35] Danielis P, Skodzik J, Altmann V, et al. Survey on real-time communication via ethernet in industrial automation environments. //Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA). Barcelona, Spain, 2014: 1-8.
- [36] Wollschlaeger M, Sauter T, Jasperneite J. The future of industrial communication: automation networks in the era of the internet of things and industry 4.0. IEEE Industrial Electronics Magazine, 2017, 11(1): 17-27.
- [37] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay. arXiv preprint arXiv:1511.05952, 2015.
- [38] Liu Quan, Zhai Jian-wei, Zhang Zogn-chang, et al. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1): 1-27 (in Chinese).
- (刘全, 翟建伟, 章宗长等. 深度学习综述. 计算机学报, 2018, 41(1): 1-27)



Liu Xiao-yu, Ph.D. His main research interests include industrial real-time communication and deep reinforcement learning

Xu Chi, Ph.D., associate professor. His main research interests include industrial wireless network and 5G URLLC.

Zeng Peng, Ph.D., professor. His main research interests include industrial communication and edge computing.

Yu Hai-bin, Ph.D., professor. His main research interests include industrial communication and real-time systems, distributed control system and network collaboration and intelligent manufacturing.

Background

Massive interconnected distributed industrial devices generate a large number of heterogeneous industrial tasks with different delay sensitivity and computing load. The limited local computing capacity of industrial devices cannot support the real-time and efficient processing of industrial tasks. It is an effective method to offload industrial tasks to multi-access edge computing servers through industrial wireless networks. However, the limited time-frequency resources of industrial wireless networks cannot support the high concurrent computing offloading.

Combining dynamic multi-channel access and dynamic priority in industrial wireless networks can solve limited time-frequency resources in some extent. Considering the difficulty in modeling high concurrent computing offloading for heterogeneous industrial tasks, a Deep Reinforcement Learning-based Concurrent Access with Dynamic Priority (CADP-DRL) algorithm is proposed in this paper.

Firstly, the industrial devices are assigned with dynamic priorities according to the delay sensitivity and computing load of their industrial tasks, and the access-offloading probabilities of industrial devices are changed dynamically. Then, the Markov decision process is utilized to formulate the dynamic priority concurrent computing offloading problem. As both dynamic priority and concurrent computing offloading of massive industrial devices result in the explosion of state space, deep reinforcement learning is used to establish a mapping

relationship from states to actions in the high-dimensional state space. With the aim at the multi-objective decision of dynamic priority and concurrent offloading, a compound reward function with joint priority reward and offloading reward is designed. In order to accelerate the convergence of the algorithm, the classification storage and dynamic sampling strategies of the experience pool are designed. Experiences are stored in different experience pools according to the experience-weight, and the sampling probabilities of different experience pools are dynamically adjusted. In experiments, compared to slotted-Aloha, DQN, DDQN and D3QN, CADP-DRL converges quickly, and can make a computing offloading decision in real-time. Meanwhile, CADP-DRL guarantees the highest successful offloading probabilities of high-priority industrial devices within minimum offloading conflicts.

This paper is supported by National Key Research and Development Program of China (2020YFB1710900), National Natural Science Foundation of China (61803368, 61972389, 61903356, U1908212), China Postdoctoral Science Foundation (2019M661156), LiaoNing Revitalization Talents Program (XLYC1801001) and Youth Innovation Promotion Association, Chinese Academy of Sciences.