DOI:10.19652/j.cnki.femt.2406186

基于深度强化学习的无线多址接入方法研究*

刘宇鹏[1,2] 雷少波[1] 樊浩研[1] 牛 虹[1]
(1.内蒙古电力(集团)有限责任公司电能计量分公司 呼和浩特 010010;
2.湖南大学电气与信息工程学院 长沙 410082)

摘 要:随机多址竞争接入技术的优化可以显著增强无线网关的处理能力,也是边缘计算应用的关键前提。针对无线物联网络中存在的异构协议多址接入系统吞吐量低的问题,提出了一种基于深度强化学习的智能自适应无线多址接入方法。首先通过信道感知、动作反馈和最小化损失机制进行接入状态的强化学习,然后采用改进的近端策略优化(PPO)算法评估最优信道接入策略,实现与传统的 TDMA、ALOHA 协议共存互补来减少接入时隙的碰撞,从而提高接入资源利用率和网络吞吐量。结果表明,改进算法能够使网络接入吞吐量相较于未使用强化学习时提升 26.6%,相比强化学习的深度 Q 网络(DQN)算法提升 2.6%,能有效降低异构多址接入问题的复杂性且显著提高无线网关的多址接入性能。

关键词:多址接入;深度强化学习;边缘计算;物联网
中图分类号: TP393;TN914.5 **文献标识码**:A **国家标准学科分类代码**: 520.60

Wireless multiple access method based on deep reinforcement learning

Liu Yupeng[1,2] Lei Shaobo[1] Fan Haoyan[1] Niu Hong[1]
(1. Electric Energy Measurement Branch of Inner Mongolia Power (Group) Co., Ltd., Hohhot 010010, China;
2. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China)

Abstract: The optimization of random multiple contention access can significantly enhance the power of wireless gateways and is also a key prerequisite for edge computing applications. Aiming at the problem of low throughput of heterogeneous protocol multiple access systems in wireless IoT networks, an intelligent adaptive wireless multiple access method based on deep reinforcement learning is proposed. First, the access state is reinforced through channel perception, action feedback and loss minimization mechanism. Then, the improved proximal strategy optimization PPO algorithm is used to evaluate the optimal channel access strategy, and complementation with traditional TDMA and ALOHA protocols are achieved to reduce the collision of access time slots, thereby improving access resource utilization and network throughput. The results show that the improved algorithm can increase the throughput by 26.6% compared with the case without reinforcement learning, and by 2.6% compared with the DQN algorithm. It can effectively reduce the complexity of multiple access and significantly improve the multiple access performance of wireless gateways.

Keywords: multiple access; deep reinforcement learning; edge computing; internet of things

0 引 言

随着互联网和无线通信技术的发展,无线物联网(internet of things, IoT)在分布式智能感知、故障诊断和决策控制系统等领域已取得显著成果[1-3]。同时,边缘计算(edge computing, EC)技术[4-6]将数据处理从云端推到检测边缘,进一步有效提高了分布式能源、时延敏感应用如电力系统监测工程应用等的数据分析服务能力,满足了工业系统更高的实时性和更强的安全性[7]需求。

在边缘计算中,不同协议物联网设备的上行接入要求带来了边缘接入服务器和物联节点之间的异构网络多址接入问题[8-9]。该问题是由多个不同协议的异构传感节点

收稿日期:2024-06-24
* **基金项目**:内蒙古电力(集团)有限责任公司科技项目(LX01234742)资助

上行竞争有限的无线信道资源所引起的，可能会导致信道冲突引起接入性能和系统吞吐量下降。另外，由于涉及到多协议接入的应用场景，即传感节点可能采用如 TDMA、ALOHA、CSMA 等不同的接入控制协议，这也进一步增加了网络管理的复杂性，需要有效感知信道接入状态、充分利用接入资源来满足不同节点的多址接入需求，以提升接入系统吞吐量。

针对多址接入问题，国内外学者进行了研究。Liu 等[10]提出利用一种利用移动边缘计算和机器学习的碰撞预测避免 MAC 协议，克服了车载自组网络中基于 TDMA 的时隙分配方法由于信道资源利用不足而引起的传输延迟和数据包碰撞问题。该方法是基于 TDMA 协议的，对于诸如 ALOHA 或 CSMA 等其他协议的适应性可能有限。李焕焕等[11]研究基于长短期记忆网络(long short-term memory，LSTM)的 MAC 协议识别方法，在充分考虑接收信号间时域相关性后实现了 4 种常见 MAC 协议的识别。该研究仅限于多个设备终端使用相同的 MAC 协议类别，对于多种协议共同接入的情况未进行深入探讨。Yu 等[12]基于深度 Q 网络(deep Q network，DQN)算法实现了基于深度强化学习的多址接入控制，为解决多址接入问题提供了一种新思路。

传统的多址接入控制机制，如 TDMA、FDMA、CDMA 和 ALOHA，通常基于固定规则，在面对动态变化的网络环境时缺乏灵活性和自适应性，导致资源利用率和系统吞吐量低。此外，ALOHA 机制容易发生资源冲突和碰撞，尤其在高流量情况下，数据包丢失和重传进一步降低了系统的整体性能[13]。TDMA 还存在公平性问题，难以在网络负载变化时优化资源分配。虽然基于 DQN 算法的多址接入控制具有一定的智能自适应性，但在实际应用中也存在诸多不足，如算法在训练过程中收敛速度较慢，计算复杂度较高，无法在实时通信系统中迅速优化策略[14]。此外，DQN 算法对于训练数据的依赖较大，如果训练数据不足或不够全面，可能导致算法难以有效适应复杂的网络环境，影响系统性能。

针对以上问题，本文提出了一种基于深度强化学习(deep reinforcement learning，DRL)的多址接入方法，以近端策略优化(proximal policy optimization，PPO)算法[15]为基础，能够与多个异构传感节点协同工作进行优化的信道接入策略选择，从而最大化整个网络的吞吐量。

1 系统模型

基于无线物联网的“端－边－云”拓扑结构如图 1 所示，主要由传感器节点、边缘计算服务器和云中心计算服务器组成。

边缘计算接入点提供了传感节点的数据收集和中继服务，可以进行数据本地处理或者转发至中心服务器，因此在边缘计算服务器的接入侧面临着随机多址竞争接入的决策问题。

边缘计算服务器基于无线接入信道的时隙控制模型[16]响应上行接入请求，其中两类节点分别采用传统的时隙 TDMA、随机 ALOHA 机制，另外一种采用了基于深度强化学习 PPO 算法进行竞争多址接入。本文将 PPO 节点称为智能节点，这些节点能记录历史信道状态信息，并不断学习得到最优的信道接入策略。智能节点可以通过充分利用 TDMA 和 ALOHA 的空闲时隙，避免多个传感节点接入时的时隙碰撞，从而优化提高网络接入性能和整体吞吐量。

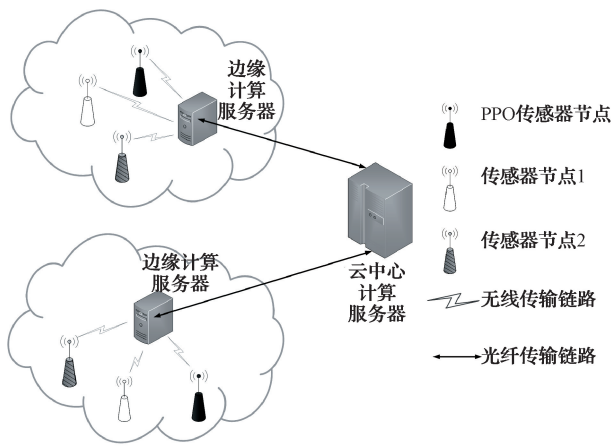


图 1 物联网“端－边－云”拓扑结构

Fig. 1 IoT "end-edge-cloud" topology

2 基于 PPO 的多址接入算法

本文所提算法以最大化网络吞吐量为优化目标，首先将多址接入问题建模为马尔科夫决策(Markov decision process，MDP)过程[17]，然后通过 PPO 算法进行接入状态感知和接入决策，从而提高多址接入概率。

2.1 MDP 建模

MDP 过程主要由智能节点的动作空间、信道观测值、状态空间和奖励值 4 个部分组成。

智能节点在每个时隙开始前选择一个动作 $a_t \in \{a_0, a_1\}$，其中 a_0 和 a_0 分别表示节点不发送数据和发送数据。不发送数据时会进行信道监听；而发送数据时，则会根据是否收到来自边缘服务器的 ACK 信号判断传输成功与否，从而进行下一步的传输决策判断。

$z_t \in \{z_y, z_f, z_n\}$ 表示采取行动后的信道观测值，其中 z_y 表示当前时隙有且只有一个节点成功传输数据；z_f 和 z_n 分别表示传输失败和信道空闲。

将 $t+1$ 时刻的动作－观测值对定义为 $c_{t+1} \triangleq (a_t, z_t)$，可能组合共有 4 种分别为 $\{a_1, z_y\}$，$\{a_1, z_f\}$，$\{a_0, z_y\}$ 和 $\{a_0, z_n\}$；将 $t+1$ 时刻的状态定义为 $s_{t+1} \triangleq [c_{t-M+2}, \cdots, c_t, c_{t+1}]$，其中 M 为需要监测的状态历史长度；将多个动作－状态值对 c_{t+1} 组合成一个状态序列，方便智能节点考虑更多的历史信息做出更好的决策。

智能节点在采取行动 a_t 后，从状态 s_t 转移到状态 s_{t+1} 并产生一个奖励值 r_t，定义如下：

$$r_t=\begin{cases}1, & z_t=z_y\\ 0, & z_t=z_f \text{ 或 } z_t=z_n\end{cases} \tag{1}$$

2.2 PPO 算法改进

PPO 算法采用 Actor-Critic[18] 架构，其最终优化目标如下：

$$\max_{\theta}\{J(\theta)\triangleq \mathbb{E}_S[V_\pi(S)]\} \tag{2}$$

式中：$V_\pi(S)$ 为状态价值函数；$\mathbb{E}_S[V_\pi(S)]$ 表示对所有可能的状态 s_t 求期望；θ 为神经网络参数；$J(\theta)$ 是参数 θ 下的性能指标。

PPO 算法是将 AC 架构中需要被优化的函数 $J(\theta)$ 近似为损失函数 $L^{CLIP}(\theta)$ 如下：

$$L^{CLIP}(\theta)=\hat{E}_t[min(r_t(\theta)\hat{A}_t,clip(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t)] \tag{3}$$

式中：$r_t(\theta)=\pi_\theta(a_t\mid s_t)/\pi_{\theta_{old}}(a_t\mid s_t)$ 定义为新策略和旧策略的概率之比，$\pi_\theta(a_t\mid s_t)$ 是新策略，$\pi_{\theta_{old}}(a_t\mid s_t)$ 是旧策略；$clip$ 函数为截断函数，目的是将新老策略的变化范围控制在 $[1-\varepsilon,1+\varepsilon]$ 之间；$\hat{E}_t[\cdot]$ 代表对多个样本求期望。优势函数 $\hat{A}_t$ 用于评估在状态 s 下采取动作 a 相对于平均行为的优势。

$$\hat{A}_t=\delta_t+(\gamma\lambda)\delta_{t+1}+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{4}$$

$$\delta_t=r_t+\gamma V(s_{t+1})-V(s_t) \tag{5}$$

式中：δ_t 表示在时间 t 的 TD 误差；r_t 表示采取动作后获得的奖励值；γ 表示奖励折扣因子；λ 表示 GAE 超参数，用于控制未来奖励的权重；T 表示总时间步数。

最后，使用梯度上升更新 PPO 算法中策略网络和价值网络的参数，更新方法如下：

$$w_{t+1}=w_t+\alpha^w\delta_t\nabla_w V(s_t,w_t) \tag{6}$$

$$\delta_t=r_t+\gamma V(s_{t+1},w_{t+1})-V(s_t,w_t) \tag{7}$$

$$\theta_{t+1}=\theta_t+\alpha^\theta\delta_t\nabla_\theta\lg\pi_{\theta_t}(a_t\mid s_t;\theta) \tag{8}$$

式中：w_t、θ_t 分别表示当前时刻策略网络和价值网络参数；w_{t+1}、θ_{t+1} 分别表示更新过后的网络参数；α^w、α^θ 是网络学习率；$\nabla_w V(s_t,w_t)$ 表示对 w 求梯度；$\nabla_\theta\lg\pi_{\theta_t}(a_t\mid s_t;\theta)$ 表示对 θ 求梯度。

采用的 PPO 算法具有在离散和连续动作空间中均有性能稳定的特点，因此被应用于本文 MAC 协议的决策。在竞争有限的通信资源时，智能节点可以通过 PPO 算法做出最佳决策来实现网络性能最大化，即根据当前环境状态灵活地选择最优的接入选择，从而优化整个网络的吞吐量。基于 PPO 的多址接入算法具体流程如下。

步骤 1）初始化环境状态 s_0，Actor 网络参数 θ，Critic 网络参数 w，目标 Actor 网络参数 θ^-、更新步数 F 和经验缓冲区；

步骤 2）$t=t_0\sim t_m$ 执行循环；

步骤 3）输入状态 s_t 到 Actor 网络，输出不同动作 a 的概率分布 $\pi(a\mid s_t,\theta)$；

步骤 4）根据 $\pi(a\mid s_t,\theta)$ 随机采样，选取动作 a_t；

步骤 5）智能节点采取动作 a_t 与环境交互，记录信道观察值 z_t 和奖励 r_t；

步骤 6）根据 s_t 和 a_t 计算出下一个状态 s_{t+1}；

步骤 7）将 (s_t,a_t,r_t,s_{t+1}) 依次存储；

步骤 8）当时间步 t 达到 F 的整数倍时，更新目标 Actor 网络参数，$\theta^-\leftarrow\theta$；

步骤 9）使用目标 Actor 网络选择动作 a_{old}，并计算概率 $\pi_{old}(a_{old}\mid s_t,\theta^-)$；

步骤 10）估计状态的价值 $V(s_t,w)$；

步骤 11）根据式(4)和(5)计算 $\hat{A}_t,\delta_t$；

步骤 12）对于每个样本数据，计算 $r_t(\theta)$；

步骤 13）根据式(3)计算 PPO 的损失函数，根据式(6)～(8)使用梯度下降方法分别更新 Actor 网络和 Critic 网络的参数 θ 和 w，并最小化损失函数 $L^{CLIP}(\theta)$。

3 实验结果与分析

3.1 仿真设置

本文使用 Keras 深度学习平台来训练神经网络，并且与基准 DQN 算法进行了对比分析。算法采用全连接层神经网络和 ReLU 函数作为激活函数。

为了更快找到适合模型的超参数，使用了贝叶斯超参数优化算法[19]，设置的采集函数是 EI（expected improved），概率代理模型是基于高斯随机过程（Gaussian process，GP）的概率代理模型。在一定的迭代次数内，采集函数以概率代理模型为先验信息，采集新的超参数用于更新模型，最后根据更新后的概率代理模型选出最佳的超参数组合，表 1 为使用贝叶斯优化得到的最佳超参数组合。

表 1 算法的超参数设置

Table 1 Algorithm hyperparameter Settings

超参数	值
折扣因子 γ	0.85
训练批次 B	10
状态历史长度 M	18
目标网络更新频率 F	20
目标网络更新权重 w	0.95
clip 函数中裁切值 ε	0.2
Adam 优化器学习率	0.000 08

3.2 性能指标

实验采用系统吞吐量作为性能指标，其定义为 每个时隙内成功传输的平均数据包，其公式如下：

$$T=\sum_{\tau=t-N+1}^{t}n_\tau/N \tag{9}$$

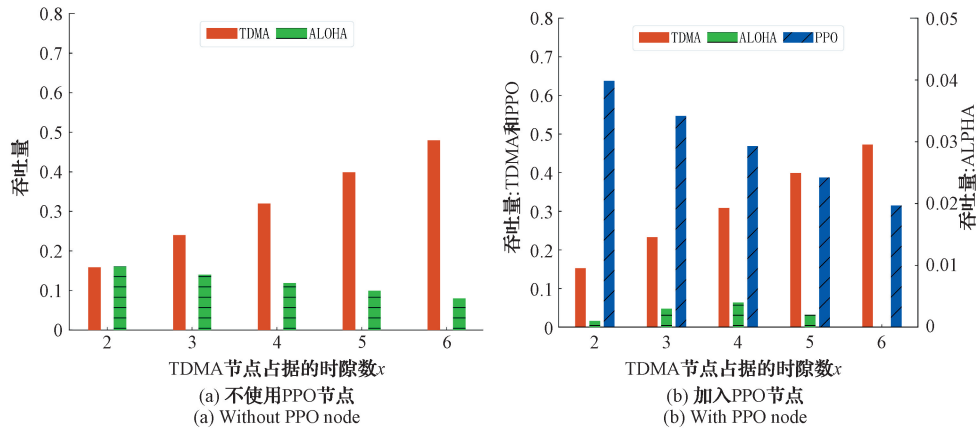
根据第 τ 个时隙中是否成功传输，n_τ 的值分别设为 1 和 0，N 设为 1 000，每个时隙持续 1 ms，这样能够反映节点在过去 1 s 内的吞吐量。

3.3 实验结果与讨论

实验中接入网关和各多址接入节点进行时序同步，将每帧设置为 10 个时隙。实验系统有 3 类节点 MAC 竞争接入机制：1）TDMA 节点，采用固定时隙接入，每次传输需要占用其中 x 个时隙；2）ALOHA 节点，采用固定概率接入，对每个时隙以概率 q 选择是否进行数据传输；3）所设计的 PPO 智能节点，无需先验知道其他传统节点的接入模式和当前数量，可以通过对信道接入状态的智能感知和强化学习来进行自主接入决策。实验对每种情况进行了 10 次实验，每次迭代 10 000 个轮次。

1）使用 PPO 的性能提升

首先对是否使用所提出的 PPO 智能节点两种情况进行了对比。当 ALOHA 节点 q 固定为 0.2 时网络吞吐量随 TDMA 节点的接入时隙 x 的变化如图 2 所示。从图 2(a)可以看出，TDMA 节点吞吐量会随着其 x 值的增大而增大，而 ALOHA 节点的吞吐量会有小幅度下降。从图 2(b)可以看出，PPO 智能节点加入后对于 TDMA 节点固定时隙占用影响较小，但是由于抢占了原先大部分低传输概率 ALOHA 节点的空闲时隙，因此可以充分利用接入系统资源，使得加入 PPO 节点后网络总吞吐量相比图 2(a)提升了 34.7%。



(a) 不使用PPO节点 (a) Without PPO node　(b) 加入PPO节点 (b) With PPO node

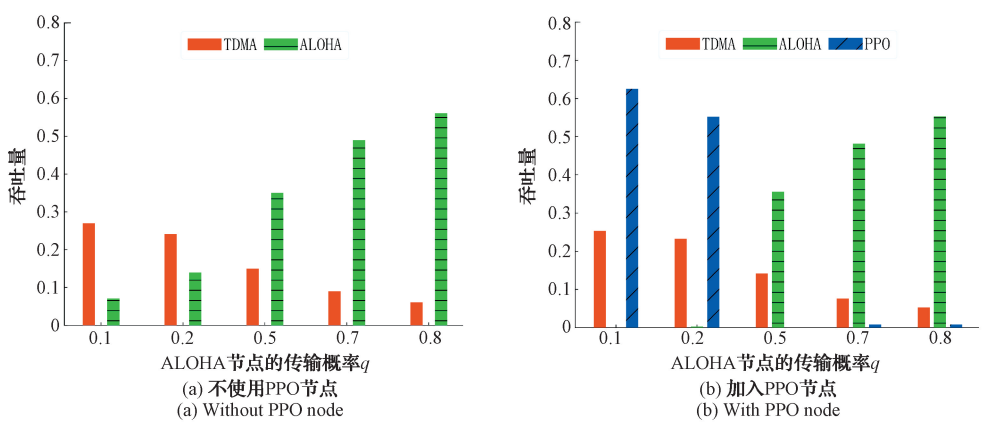
图 2　ALOHA 节点 $q=0.2$ 时网络吞吐量随 x 的变化

Fig. 2　Change of network throughput when ALOHA node $q=0.2$

当 TDMA 节点的 x 值固定为 3 时，系统是否使用 PPO 节点的网络吞吐量随 ALOHA 接入概率 q 值的变化情况如图 3 所示。从图 3(a)可以看出，随着接入概率 q 的不断增大，ALOHA 节点吞吐量不断增加，TDMA 节点吞吐量相应的减小。图 3(b)为加入 PPO 节点网络吞吐量的变化情况，当 q 值较小时，PPO 节点充分利用了接入空闲时隙有一定的吞吐量；而当 q 值较大时，PPO 节点吞吐量几乎为 0。对比图 3(a)和(b)可以发现，PPO 算法对于 TDMA 节点影响较小，但有效弥补了 ALOHA 节点 q 值较小时的系统时隙浪费，使得网络总吞吐量提升了 18.4%。

2）DQN 与 PPO 的性能对比

将本文算法与文献[12]的 DQN 算法方法进行性能



(a) 不使用PPO节点 (a) Without PPO node　(b) 加入PPO节点 (b) With PPO node

图 3　TDMA 节点 $x=3$ 时网络吞吐量随 q 的变化

Fig. 3　Change of network throughput when TDMA node $x=3$

对比。DQN 是一种常用的强化学习算法，具有支持连续空间和适合离线学习的优点。表 2 是在 ALOHA 的 q 值固定为 0.2 时得到实验结果，可以看出随着 TDMA 节点 x 值的变化。相较于 DQN 算法，采用 PPO 算法的系统总吞吐量分别提高了 3.4%、3.9%、1.4%、2.7%和 3.2%，表明在其他接入状态相同的情况下 PPO 算法比 DQN 能够更有效地改善网络的接入性能。

表 2　$q=0.2$ 时网络总吞吐量随 x 的变化

Table 2　Total network throughput when $q=0.2$

x	吞吐量/%	
	DQN[12]	本文方法
2	75.8	79.2
3	74.4	78.3
5	76.8	78.2
7	76.2	78.9
8	75.6	78.8

$q=0.2$ 时，网络中不同节点的吞吐量如图 4 所示，可以发现 PPO 节点的吞吐量要高于 DQN 节点，尤其在 $x=3$ 时两者之间的差值达到最大，相差 4%。此外，TDMA 节点和 ALOHA 节点的吞吐量基本保持一致，这说明表 2 中网络总吞吐量的差异主要是由于 PPO 与 DQN 节点之间的吞吐量区别引起的。

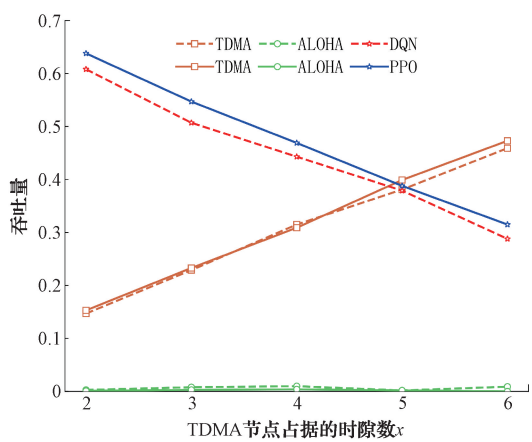


图 4　ALOHA 节点 $q=0.2$ 时网络中各节点的吞吐量对比

Fig. 4　Throughputs of nodes when ALOHA node $q=0.2$

还可以发现 TDMA 节点的吞吐量随着时隙 x 值增加而增加；而 ALOHA 节点的吞吐量几乎为 0，这是由于 ALOHA 节点的传输概率 q 较小时，即使选择发送也会因为同一时隙中其他两个节点占用而导致碰撞；而 PPO 节点需要考虑 ALOHA 节点和 TDMA 节点的发送数据情况，在总体吞吐量有限的情况下，随着 TDMA 节点吞吐量的增加，PPO 节点吞吐量整体呈下降趋势。

表 3 是在 TDMA 的 x 值固定为 3 时得到的实验结果，可以看出网络总吞吐量随着 ALOHA 节点 q 值变化的情况，表明使用 PPO 时的网络总吞吐量相较于 DQN 分别高出 2.1%、2.4%、1.9%、3.7%和 2.5%。并且，在 q 值为 0.5 时网络总吞吐量最小，这是由于随着 ALOHA 节点的发送概率增大，DQN 和 PPO 智能节点都选择减少数据发送以避免碰撞，出现了网络吞吐量下降。

表 3　$x=3$ 时网络总吞吐量随 q 变化

Table 3　Total network throughput when $x=3$

q	吞吐量/%	
	DQN[11]	本文方法
0.1	85.7	87.8
0.2	76.4	78.8
0.5	47.9	49.8
0.7	52.9	56.6
0.8	58.9	61.4

5 种 ALOHA 节点不同 q 值情况下的节点吞吐量对比如图 5 所示。随着 q 值增加，ALOHA 节点的吞吐量迅速上升，而相应 PPO 节点则呈现急剧下降的趋势。此外，q 值的增加也会影响网络中 TDMA 节点的吞吐量。具体来说，对比图 5 中 q 值较大(0.7、0.8)和 q 值较小(0.1、0.2)的情况，发现 q 值的增加导致 TDMA 节点吞吐量减小。因为当 q 值较大时，不论当前时隙是否被 TDMA 节点占用，ALOHA 节点都有很大的概率发送数据。如果当前时隙未被 TDMA 节点占用，较大的发送概率提高了网络总吞吐量；但若当前时隙已被 TDMA 节点占用，更大的发送概率导致两节点碰撞概率增大，因此降低了 TDMA 节点的吞吐量；另一方面，智能节点 DQN 和 PPO 节点则是感知信道中 ALOHA 接入增加而基于强化学习机制减少了发送自动避免接入冲突，因此 TDMA 节点和 PPO 节点的吞吐量会随着 q 值的增加而下降。

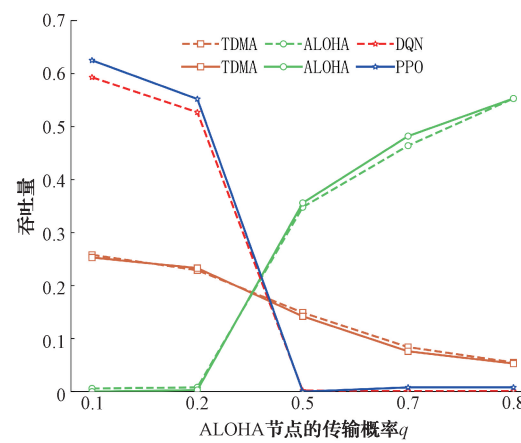


图 5　TDMA 节点 $x=3$ 时网络中各节点的吞吐量对比

Fig. 5　Throughputs of nodes when when TDMA node $x=3$

4　结　论

为了解决物联网中多址接入时系统吞吐量低的问题，

本文提出了一种基于DRL的多址接入方法，采用PPO算法进行异构多址节点的接入决策。实验结果表明，所提算法相比传统和基于DQN算法都能更有效地提高网络总吞吐量。未来将在提高吞吐量的基础上，进一步探讨多节点共存网络中的接入公平性问题，确保通信资源的公平分配。

参 考 文 献

[1] 徐钰龙，李君，李正权，等. 基于深度强化学习的无人机辅助物联网多目标优化[J]. 国外电子测量技术，2024，43(5)：26-35.
XU Y L，LI J，LI ZH Q，et al. Multi-objective optimization of unmanned aerial vehicle assisted internet of things based on deep reinforcement learning [J]. Foreign Electronic Measurement Technology，2024，43(5)：26-35.

[2] 马助兴，付炜平，李焱，等. 基于物联网技术的变电站智能安全管控系统的设计及实现[J]. 电子测量技术，2019，42(23)：6-14.
MA ZH X，FU W P，LI Y，et al. Design and implementation of substation intelligent safety management and control system based on internet of things technology [J]. Electronic Measurement Technology，2019，42(23)：6-14.

[3] TEMENE N，SERGIOU C，GEORGIOU C，et al. A survey on mobility in wireless sensor networks[J]. Ad Hoc Networks，2022，125：102726.

[4] 邓集检，张月霞. 基于用户意愿度D2D协助的工业物联网资源分配[J]. 国外电子测量技术，2024，43(2)：193-200.
DENG J J，ZHANG Y X. Resource allocation of industrial internet of things based on user willingness D2D assistance[J]. Foreign Electronic Measurement Technology，2024，43(2)：193-200.

[5] 王瑶，卢先领，沈义峰. 移动边缘计算中依赖型任务的调度模型研究[J]. 电子测量与仪器学报，2022，36(8)：60-68.
WANG Y，LU X L，SHEN Y F，et al. Research on scheduling model of dependent tasks in mobile edge computing[J]. Journal of Electronic Measurement and Instrumentation，2022，36(8)：60-68.

[6] LUO R，JIN H，HE Q，et al. Cost-effective edge server network design in mobile edge computing environment[J]. IEEE Transactions on Sustainable Computing，2022，7(4)：839-850.

[7] 吴钢，周金辉，李慧. 面向边缘增强分布式电力无线传感网的资源分配[J]. 中国电力，2023，56(8)：77-85，98.
WU G，ZHOU J H，LI H，et al. Resource allocation for edge-enhanced distributed power wireless sensor network[J]. Electric Power，2023，56(8)：77-85，98.

[8] GAMAL S，RIHAN M，HUSSIN S，et al. Multiple access in cognitive radio networks：From orthogonal and non-orthogonal to rate-splitting [J]. IEEE Access，2021，9：95569－95584.

[9] YANG X，WANG L，SU J，et al. Hybrid MAC protocol design for mobile wireless sensors networks[J]. IEEE Sensors Letters，2018，2(2)：1-4.

[10] LIU B，DENG D，RAO W，et al. CPA-MAC：A collision prediction and avoidance MAC for safety message dissemination in MEC-assisted VANETs[J]. IEEE Transactions on Network Science and Engineering，2022，9(2)：783-794.

[11] 李焕焕，彭盛亮，陈铮，等. 认知无线电中基于LSTM网络的MAC协议识别[J]. 信号处理，2019，35(5)：837-842.
LI H H，PENG SH L，CHEN ZH，et al. MAC protocol recognition based on LSTM network in cognitive radio [J]. Journal of Signal Processing，2019，35(5)：837-842.

[12] YU Y，WANG T，LIEW S C. Deep-reinforcement learning multiple access for heterogeneous wireless networks[J]. IEEE Journal on Selected Areas in Communications，2019，37(6)：1277-1290.

[13] 刘磊，李宇，张春华，等. 延迟容忍及冲突避免的水声网络S-Aloha协议[J]. 仪器仪表学报，2014，35(3)：513-519.
LIU L，LI Y，ZHANG CH H，et al. Delay tolerant and collision avoidance S-Aloha protocol for underwater acoustic networks[J]. Chinese Journal of Scientific Instrument，2014，35(3)：513-519.

[14] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529-533.

[15] CHEN Z，YIN B，ZHU H，et al. Mobile communications，computing，and caching resources allocation for diverse services via multi-objective proximal policy optimization[J]. IEEE Transactions on Communications，2022，70(7)：4498-4512.

[16] JAKOVETIĆ D, BAJOVIĆ D, VUKOBRATOVIĆ D, et al. Cooperative slotted aloha for multi-base station systems [J]. IEEE Transactions on Communications, 2015, 63(4): 1443-1456.

[17] LIN S, FAN R, FENG D, et al. Condition-based maintenance for traction power supply equipment based on partially observable Markov decision process[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(1): 175-189.

[18] BANERJEE C, CHEN Z, NOMAN N, et al. Optimal actor-critic policy with optimized training datasets[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(6): 1324-1334.

[19] CHO H, KIM Y, LEE E, et al. Basic enhancement strategies when using Bayesian optimization for hyperparameter tuning of deep neural networks[J]. IEEE Access, 2020, 8: 52588－52608.

作 者 简 介

刘宇鹏，博士研究生，高级工程师，主要研究方向为电力系统自动化和物联网技术。

雷少波，硕士，高级工程师，主要研究方向为电力网络运维技术。

E-mail: 353263598@qq.com

樊浩研，本科，工程师，主要研究方向为电能计量技术。

牛虹，本科，工程师，主要研究方向为电能计量技术。