基于多智能体近端策略优化的多信道动态频谱接入

陈平平[1],张 旭[1],谢肇鹏[1*],丘毓萍[2],方 毅[3]
(1. 福州大学先进制造学院,福建晋江 362251;2. 福州大学物理与信息工程学院,福建福州 350108;
3. 广东工业大学信息工程学院,广东广州 510006)

摘 要: 为了在多用户多信道通信场景中应用动态频谱接入(Dynamic Spectrum Access,DSA)技术提高通信效率,保证用户公平,本文基于多智能体近端策略优化(Multi-Agent Proximal Policy Optimization,MAPPO)提出了MAPPO-DSA算法. 该算法首先针对单信道接入在多个信道同时空闲时存在的频谱浪费问题,使用多信道接入作为解决方案. 同时,多信道接入导致状态空间与动作空间指数增长,计算成本高,学习难度大. 为此本文引入MAPPO深度强化学习(Deep Reinforcement Learning,DRL)算法,在复杂环境中高效学习和优化接入策略. 通过设计优化MAPPO中观测及奖励等强化学习要素和共享网络参数来保证用户公平. 最后,在不同场景下的实验结果表明,所提出的MAPPO-DSA能够学习到近似最优的接入策略,部分场景中的网络吞吐量逼近理论上限,显著优于现有算法,且有效保证用户公平.

关键词: 动态频谱接入;深度强化学习;多智能体近端优化;多信道接入
基金项目: 国家自然科学基金(No.62171135,No.62322106,No.62071131);福建省自然科学基金(No.2022J06010)
中图分类号: TP317.4 **文献标识码:** A **文章编号:** 0372-2112(2024)06-1824-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230663

Multi-Channel Dynamic Spectrum Access Based on Multi-Agent Proximal Policy Optimization

CHEN Ping-ping[1], ZHANG Xu[1], XIE Zhao-peng[1*], QIU Yu-ping[2], FANG Yi[3]
(1. *School of Advanced Manufacturing, Fuzhou University, Jinjiang, Fujian* 362251, *China*;
2. *College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian* 350108, *China*;
3. *School of Information Engineering, Guangdong University of Technology, Guangzhou, Guangdong* 510006, *China*)

Abstract: To enhance communication efficiency and ensure user fairness in multi-user multi-channel communication scenarios, based on multi-agent proximal policy optimization (MAPPO) for the application of dynamic spectrum access (DSA) technology, this paper proposes the MAPPO-DSA algorithm. The algorithm addresses the issue of spectrum waste in single-channel access when multiple channels are simultaneously idle by using multi-channel access as a solution. However, multi-channel access leads to an exponential increase in the state and action spaces, resulting in high computational costs and learning difficulties. To tackle this, the paper introduces the MAPPO deep reinforcement learning (DRL) algorithm to efficiently learn and optimize access strategies in complex environments. The design of MAPPO incorporates reinforcement learning elements such as observation and reward, as well as shared network parameters to ensure user fairness. Experimental results in different scenarios demonstrate that the proposed MAPPO-DSA algorithm can learn near-optimal access strategies, and approach the theoretical throughput limit in some scenarios, outperforming the existing algorithms significantly and effectively ensuring user fairness.

Key words: dynamic spectrum access; deep reinforcement learning; multi-agent policy optimization; multi-channel access

Foundation Item(s): National Natural Science Foundation of China (No.62171135, No.62322106, No.62071131); Natural Science Foundation of Fujian Province (No.2022J06010)

收稿日期:2023-07-14;修回日期:2024-01-16;责任编辑:李勇锋
*通讯作者:谢肇鹏

1 引言

当今频谱资源已成为宝贵的稀缺资源[1],传统的频谱管理方法在提高频谱利用效率方面面临诸多挑战[2]. 动态频谱接入(Dynamic Spectrum Access,DSA)应运而生. DSA技术可以在时变的频谱环境中找到频谱空洞,机会式地进行通信[3,4]. DSA分为三类,分别是动态专用模型、开放共享模型和分层接入模型[5]. 其中分层接入模型将用户划分为主用户(Primary User,PU)和次用户(Second User,SU),其中次用户也被称为认知用户(Cognitive User,CU). 分层接入模型允许不同用户以不同的优先级进行频谱接入,在保证PU通信质量的前提下,能满足多样化的通信需求[6,7].

然而在实际的通信环境具有高度的复杂性和不确定性,DSA需要有自主学习的能力,来找到合适的接入策略. 为此引入深度强化学习(Deep Reinforcement Learning,DRL). 文献[8]提出了一种基于DRL的多小区非正交多址接入能效优化功率分配算法. 文献[9]将多智能体DRL算法应用于移动边缘计算系统. 文献[10]基于Deep Q Network(DQN)[11]提出了一种自适应DSA算法,获得与已知信道模型相同的最优性能. 近年来多智能体强化学习技术的发展,使得分布式智能动态频谱分配成为可能[12],文献[13]提出了一种基于DQN的多认知用户分布式DSA算法. 文献[14]提出了一种基于DQN的多认知用户多信道异构网络DSA算法,认知用户通过历史观测学习接入策略. 文献[15]提出了一种基于QMIX[16]的多认知用户单信道DSA算法,采用集中式训练分布式执行(Centralized Training Decentralized Execution,CTDE)框架.

上述研究基于单信道接入,当多个信道同时可用时,会导致频谱资源的浪费. 而多信道接入能够有效解决该问题. 但多信道接入的动作空间大小随着信道数量指数级增加,学习难度较大. 此外,在上述多用户场景DSA的研究中,大多是将单认知用户DSA算法直接应用于多认知用户场景[10,13,14]. 每个认知用户独立采用单用户DSA算法,将其他用户视作环境的一部分,破坏了环境的平稳性,且无法有效保障用户间公平.

据此,本文基于多用户多信道通信场景,利用多信道接入减少频谱资源浪费,在单信道接入中,单个用户网络吞吐率上限仅有1. 而多信道接入的网络吞吐率上限为信道数量K,大幅提升了性能上限. 针对多信道场景学习难度大的问题,引入多智能体近端策略优化(Multi-Agent Proximal Policy Optimization,MAPPO)[17],提出了MAPPO-DSA算法. 基于CTDE框架,分布式执行,集中式训练. 认知用户完全根据自身的独立观测进行接入决策. 实验结果表明本文提出的MAPPO-DSA算法在不同场景下均能有效提升网络吞吐量,在部分场景逼近理论上限值,相比现有算法提升明显,且用户公平性良好.

2 系统模型

考虑一种基于时隙Aloha协议[18]的多信道MAC异构无线网络(Heterogeneous Wireless Networks,HetNet). 系统存在多个正交信道,时间被划分为离散时隙并分配到不同的帧中,每一帧包含10个时隙. 用户分为主用户与认知用户,主用户采用预先设定的MAC协议向接入站(Access Point,AP)发送数据包. 认知用户根据接入策略机会地选择信道向AP发送数据包. AP通过广播返回ACK信号. 多个用户同时访问一个信道则会发生碰撞. 每个用户都配备一个有限缓冲区用于存储将要发送的数据包,数据包在每一帧的开始时按照泊松分布随机到达用户的缓冲区.

系统中主用户共有三种类型:

(1)TDMA用户:在每一帧的固定时隙发送数据包;

(2)Q-Aloha用户:在每个时隙以固定概率q发送数据包;

(3)Fixed-Window Aloha(FW-Aloha)用户:在发送数据包后产生一个随机整数$w \in [0, W]$,等待w个时隙后进行下一次发送.

认知用户通过频谱感知收集信息,执行频谱接入决策. 其重点在于认知用户需要提前预判频谱空洞产生,并进行接入,如图1所示. 图中白色方块代表该时隙没有用户接入该信道,处于空闲状态. 灰色方块代表该时隙下该信道被占用.

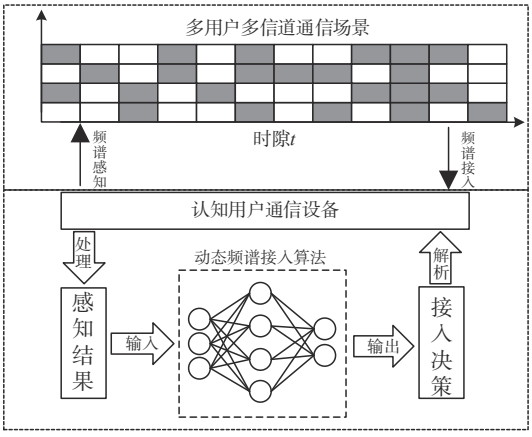

图1　认知用户接入模型

本文的目标是设计一个多信道多认知用户场景下的动态频谱接入算法,最大化网络吞吐量,同时保证认知用户间的公平性.

3 多智能体近端策略优化

3.1 近端策略优化

DRL是一种在线学习方法，基于部分可观测马尔科夫决策过程(Partially observable Markov decision processes，POMDP)[19]建模，智能体在与环境的交互过程中，观测环境状态，根据当前策略执行动作，根据反馈信息优化自身策略. 最终智能体在不断地尝试和反馈中找到最优的行为策略，其中策略表示为在状态s下执行动作a的条件概率，即

$$\pi(a,s)=p(a|s) \tag{1}$$

智能体在当前环境状态s_t下获得观测o_t，根据其策略π每个动作对应的概率，随机进行采样，得到一个动作a_t并执行，环境根据智能体执行的动作反馈奖励r_t，环境进入下一个状态，智能体获得下一个观测，不断重复该过程，找到最优策略π^*，最大化长期折扣奖励G.

$$G=\sum_{k=0}^{\infty}\gamma^k r_k \tag{2}$$

$$\pi^*=\arg\max_{\pi}\left(\sum_{k=0}^{\infty}\gamma^k r_k\right) \tag{3}$$

近端策略优化(Proximal Policy Optimization，PPO)[20]是行动者-评论家(Actor-Critic，AC)[21]框架的一种DRL算法. Actor输入状态s_t，输出每个动作a对应的概率$p(a|s_t)$. Critic输入当前状态s_t，输出值函数$V(s_t)$，代表状态s_t下的长期折扣奖励[22]. 在PPO中，Actor的损失函数L_t^{CLIP}如下：

$$L_t^{\mathrm{CLIP}}=-\hat{\mathbb{E}}_t\left[\min\left(k_t(\theta',\theta)\hat{A}_t,\mathrm{clip}\left(k_t(\theta',\theta),1-\varepsilon,1+\varepsilon\right)\hat{A}_t\right)\right] \tag{4}$$

$$k_t(\theta',\theta)=\frac{\pi_{\theta'}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} \tag{5}$$

$$\hat{A}_t=\delta_t+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{6}$$

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

其中，θ为旧策略参数，θ'为新策略参数，$k_t(\theta',\theta)$为重要性采样权重，$\hat{A}_t$为动作优势函数估计值，ε为设定的裁剪参数，δ_t为值函数的时序差分误差.

Critic的损失函数L_t^{VF}如下：

$$L_t^{\mathrm{VF}}=\left(V_{\theta}(s_t)-\left(\hat{A}_t+V_{\theta}(s_t)\right)\right)^2 \tag{8}$$

为了避免陷入局部最优，PPO引入了熵正则化项，即策略π_θ在状态s_t下的熵值$S_\theta(s_t)$. 最终，PPO的损失函数如式(9)所示：

$$L_t=L_t^{\mathrm{CLIP}}+c_1L_t^{\mathrm{VF}}-c_2S_\theta(s_t) \tag{9}$$

其中，c_1与c_2取0.5.

3.2 多智能体近端策略优化

MAPPO在PPO的基础上采用CTDE框架，将智能体的学习过程分为训练和执行阶段[17]. 在训练阶段集中训练策略网络Actor和价值网络Critic，执行时智能体仅利用自身观测和策略网络Actor进行动作的选择.

4 MAPPO-DSA

本文提出MAPPO-DSA算法，在每个认知用户端建立策略网络Actor$_\theta$，共享网络参数θ. 在服务器端建立一个价值网络Critic$_\varphi$. 执行阶段，认知用户执行动作，收集数据. 训练阶段，认知用户集中训练，更新策略网络和价值网络，如图2所示. 多次训练后，结束学习，进入测试阶段，认知用户仅根据策略网络执行动作.

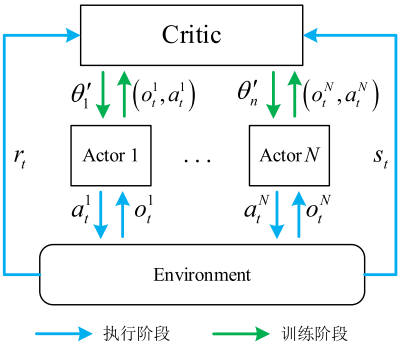

图2 MAPPO算法示意图

MAPPO-DSA算法的核心由CTDE框架、损失函数、强化学习要素(动作、观测和奖励)、神经网络组成. 其中CTDE框架与损失函数上文已经介绍.

4.1 强化学习要素

动作：定义a_t^n认知用户n在时隙t所有信道上的联合动作，具体如下：

$$\boldsymbol{a}_t^n=\left[a_{t,1}^n,a_{t,2}^n,\cdots,a_{t,K}^n\right] \tag{10}$$

其中，$a_{t,k}^n\in\mathcal{A}\equiv\{0,1\}$，0代表等待，1代表发送.

观测：定义o_t^n为认知用户n在时隙t获得的环境观测. 频谱感知结果$\boldsymbol{b}_t=\left[b_t^1,b_t^2,\cdots,b_t^K\right]$是认知用户利用频谱感知法对所有信道进行检测得到的结果，其中，b_t^k为信道k的频谱感知结果，当信道空闲时为0，否则为1.

认知用户可以根据收到的ACK信号判断出所有认知用户的发送情况，令$v_{t,k}^n$表示认知用户n在信道k上的等待时间. 当认知用户n在信道k上发送成功时，$v_{t,k}^n=0$，否则$v_{t,k}^n=v_{t-1,k}^n+1$. 观测$\boldsymbol{o}_t^n=\left[a_{t-1}^n,b_t,v_t^n,v_t^{-n}\right]$. 其中，$\boldsymbol{v}_t^n=[v_{t,1}^n,v_{t,2}^n,\cdots,v_{t,K}^n]$表示认知用户$n$的等待时间，$\boldsymbol{v}_t^{-n}=[v_t^1,\cdots,v_t^{n-1},v_t^{n+1},\cdots,v_t^N]$表示除了认知用户$n$外，其他认知用户的等待时间.

全局状态：定义s_t为时隙t的全局状态信息，具体

如下:

$$\boldsymbol{s}_t=\left[a_{t-1},b_t,v_t\right] \tag{11}$$

奖励:定义r_t为时隙t的全局奖励,如下:

$$r_t=\sum_{k=1}^{K}r_t^k \tag{12}$$

其中,时隙t信道k对应的奖励r_t^k如下:

$$r_t^k=\begin{cases}1, & \text{认知用户}n\text{发送成功且}v_{t,k}^n=\max\limits_i v_{t,k}^i\\ \dfrac{v_{t,k}^n}{\sum\limits_{i=1}^{N}v_{t,k}^i}, & \text{认知用户}n\text{发送成功且}v_{t,k}^n\neq\max\limits_i v_{t,k}^i\\ -2, & \text{碰撞}\\ -0.5, & \text{浪费}\\ 0, & \text{其他}\end{cases} \tag{13}$$

4.2　神经网络

本节所采用的神经网络结构如图3所示. 策略网络的输入为观测o_t^n和当前时隙策略网络隐状态h_t^n,输出为熵$S_\theta(o_t^n)$、动作a_t^n、动作概率对数值$\ln\pi_\theta(a_t^n|o_t^n)$和下个时隙策略网络隐状态$h_{t+1}^n$. 价值网络的输入为全局状态$s_t$和当前时隙价值网络隐状态$c_t^n$,输出为状态值$V_\varphi(s_t)$和下一时隙价值网络隐状态$c_{t+1}^n$. 需要注意的是,在策略网络中,为了得到最终的动作概率分布,需要在神经网络的输出端增加一个softmax层.

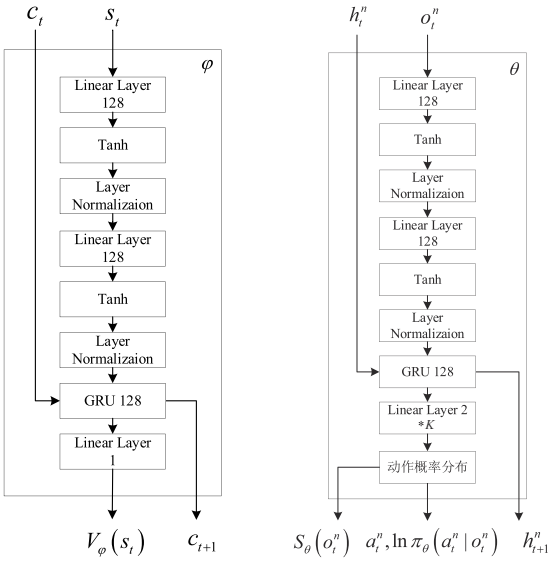

图3　神经网络

4.3　算法流程

MAPPO-DSA分为学习与测试两个阶段. 学习阶段细分为执行与训练两个阶段,交替进行. N个认知用户,在执行阶段的时隙t,认知用户n获取观测o_t^n,根据观测o_t^n和策略网络Actor$_\theta$选择动作a_t^n并执行. 随后接入站结合各用户的接入情况反馈ACK信号,进入下一时隙,重复上述过程.

在训练阶段,各认知用户将数据上传,接入站根据认知用户记录的数据计算对应时隙的奖励r_t、全局状态s_t,根据价值网络Critic$_\varphi$计算状态值$V_\varphi(s_t)$和隐状态c_t. 随后将认知用户和接入站记录的数据进行组合、计算,存储在经验池D中. 学习阶段的伪代码如算法1所示. 测试阶段与学习阶段的区别是不再进行训练.

算法1　MAPPO-DSA(学习阶段)

初始化:观测o_t^n、策略网络参数θ、价值网络参数φ、策略网络隐状态h_t^n、价值网络隐状态c_t、经验池D、学习率α、奖励折扣率γ、优势折扣率λ、裁剪参数ε、数据分组N_{batch}、更新次数N_{update}、数据块长度T、执行阶段长度M、流量到达率λ、训练次数E、用户数量N、信道数量K、价值误差参数c_1、熵正则项参数c_2.

for i=1 to E do
  执行阶段开始
  for t=1 to M do
    if t mod F then
      用户流量遵循泊松分布 到达
    end if
    for n=1 to N do
      认知用户n将观测o_t^n、隐状态h_t^n输入策略网络θ,策略网络θ输出动作a_t^n、$\ln\pi_\theta(a_t^n|o_t^n)$、$S_\theta(o_t^n)$、$h_t^n$.
      执行动作a_t^n
    end for
    for n=1 to N do
      认知用户n获得观测o_{t+1}^n
      记录$(o_t,a_t,o_{t+1},\ln\pi_\theta(a_t^n|o_t^n),h_t^n)$
    end for
  end for
  训练阶段,认知用户上传数据,中心接入基站获得数据,存储于经验池D
  for x=1 to N_{update} do
    经验池内每T个时隙对应的数据分为一组后随机打乱
    for y=1 to N_{batch}
      随机抽取$MT//N_{\text{batch}}$组数据
      损失函数$L=L^{\text{CLIP}}+0.5L^{\text{VF}}-0.5S$
    更新策略网络参数为θ',更新价值网络参数为φ'
    end for
  end for
更新认知用户端策略网络$\theta\leftarrow\theta'$
end for

5　仿真结果及分析

本文在不同场景下应用MAPPO-DSA,给出学习阶段和测试阶段的吞吐量曲线图,对比算法采用随机接入与文献[14]使用的独立深度Q网络(Independent

Deep Q Network Dynamic Spectrum Access，IDQN-DSA)[14,23]，超参数设置如表1所示.

表1 超参数设置表

超参数名称	超参数设置
最大训练次数E	10 000
执行阶段长度M	100
帧内时隙数量F	10
用户流量缓冲区大小	50
优势折扣率λ	0.95
奖励折扣率γ	0.9
裁剪参数ε	0.2
数据分组N_{batch}	2
更新次数N_{update}	5
数据块长度T	20
价值误差参数c_1	0.5
熵正则项参数c_2	0.001
学习率α	0.000 5
优化器	RMSprop

本章将展示不同场景下MAPPO的学习、测试阶段吞吐量曲线图和不同算法的对比图. 其中，Sum代表系统总吞吐量，PU_i代表主用户i的吞吐量，CU_n代表认知用户n的吞吐量.

接下来，本文比较下不同方案的计算复杂度[24]. 令Z为神经网络的参数数量. 假设神经网络计算一次的复杂度为$O(Z)$，训练次数为E，用户数量N，每一次训练的执行阶段长度为M，即训练阶段计算复杂度为$O(\text{ZENM})$. 在本文中，由于，MAPPO-DSA与IDQN-DSA在执行策略上不同，整体的训练架构一样. 因此，他们的复杂度均为$O(\text{ZENM})$.

5.1 仿真实验

场景1：2个信道，4个认知用户，2个TDMA主用户. 主用户1、2一帧发送5个数据包. 认知用户数据流量饱和，主用户流量到达遵循泊松分布$P(5)$. 图4是MAPPO-DSA测试阶段的吞吐量曲线图. 场景1主用户总吞吐量理论上限为1.0，系统总吞吐量理论上限为2.0. 在保证认知用户间公平的情况下，所有认知用户吞吐量理想值均为0.25. 图4中测试阶段所有认知用户吞吐量均接近理想值. 图5是不同算法的总吞吐量对比，可以看到MAPPO-DSA收敛速度快，吞吐量逼近理论上限，稳定性显著优于IDQN-DSA.

场景2：10个信道，2个认知用户，10个TDMA主用户，主用户1、2固定5个时隙发送数据，主用户3、4固定4个时隙发送数据，主用户5、6固定3个时隙发送数据，主用户7、8固定2个时隙发送数据，主用户9、10固定1个时隙发送数据. 认知用户数据流量饱和，主用户流量到达遵循泊松分布$P(\lambda)$，到达率等于自身在一帧内计

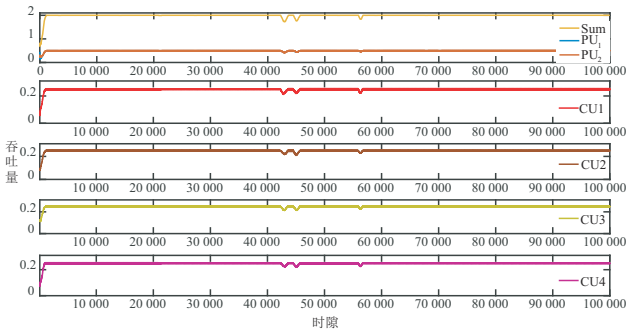

图4 MAPPO-DSA场景1测试阶段吞吐量

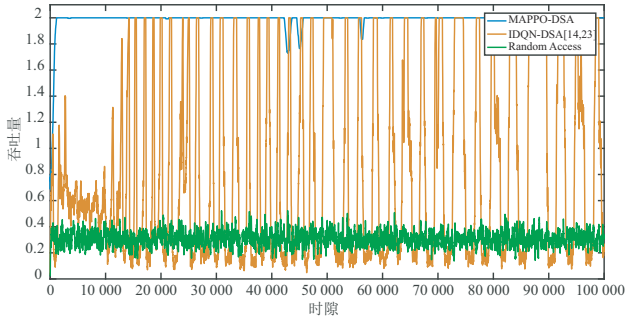

图5 场景1不同算法总吞吐量对比

划发送的数据包数量. 图6是MAPPO-DSA测试阶段的吞吐量曲线. 主用户总吞吐量理论上限为3.0，系统总吞吐量理论上限为10.0，认知用户吞吐量理想值为3.5，均接近理想值. 图7是不同算法对比，MAPPO-DSA快速逼近吞吐量上限10，而RA的吞吐量不足5，IDQN-DSA的吞吐量则约在5至9区间内反复波动. 相比于IDQN-DSA和RA，MAPPO-DSA网络吞吐量显著提升.

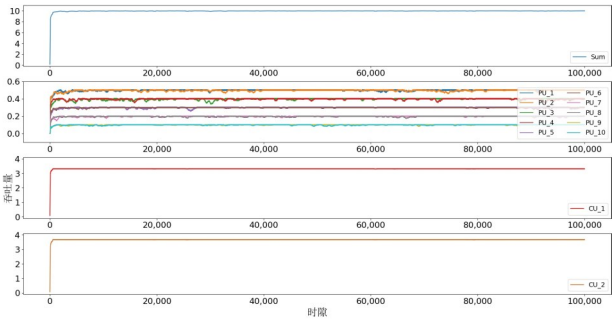

图6 MAPPO-DSA场景2测试阶段吞吐量

场景3：4个信道，4个认知用户，4个TDMA主用户. 主用户1、2一帧发送5个数据包，主用户3、4一帧发送4个数据包. 认知用户数据流量饱和，主用户流量到达遵循泊松分布$P(\lambda)$，到达率一帧内计划发送的数据包数量. 图8是MAPPO-DSA场景3测试阶段吞吐量曲线图. 场景3主用户总吞吐量理论上限为1.8，系统总吞吐量理论上限为4.0. 在保证认知用户间公平的情况下，认知用户吞吐量理想值为0.55. 图8中认知用户吞

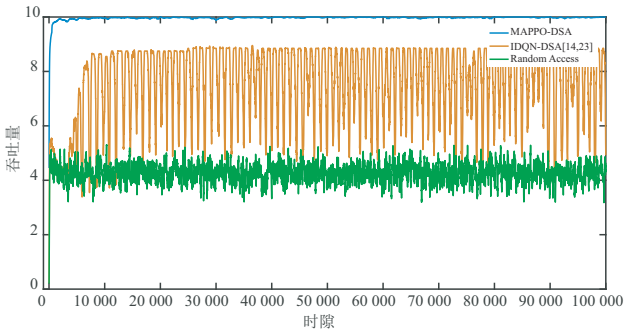

图7　场景2不同算法总吞吐量对比

吐量均为0.54左右,接近理想值. 图9是场景三不同算法总吞吐量对比,场景3中认知用户数量为4,RA算法在认知用户数量增加时会出现严重的碰撞,吞吐量仅有0.75左右,不足性能上限的20%. IDQN-DSA仍然存在不稳定现象,其网络吞吐量仅在少数时隙接近4.0.

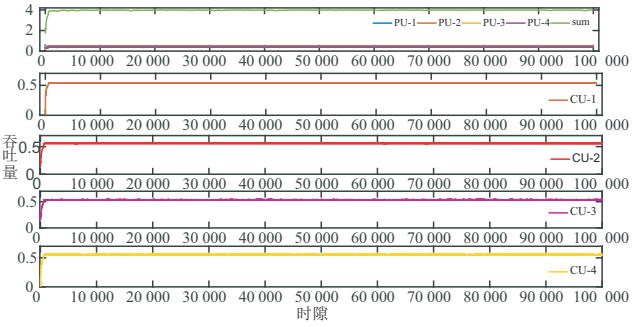

图8　MAPPO-DSA场景3测试阶段吞吐量

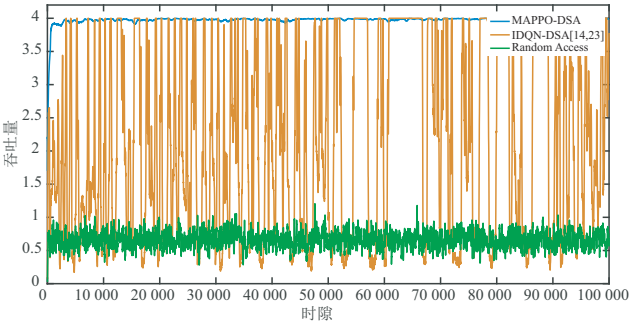

图9　场景3不同算法总吞吐量对比

场景4:3个信道,2个认知用户,3个主用户. 主用户1采用TDMA协议,一帧发送5个数据包. 主用户2采用Q-Aloha协议,以$q=0.5$的概率在每个时隙发送数据. 主用户3采用FW-Aloha协议,窗口大小$W=4$. 各用户到达率为饱和流量下一帧内平均尝试发送的数据包数量. 图10、11是MAPPO-DSA在场景四测试阶段吞吐量曲线图. 测试阶段,MAPPO-DSA系统总吞吐量约为2.21. 主用户1、2、3吞吐量分别约为0.498、0.421、0.110. 两个认知用户吞吐量均在0.599左右.

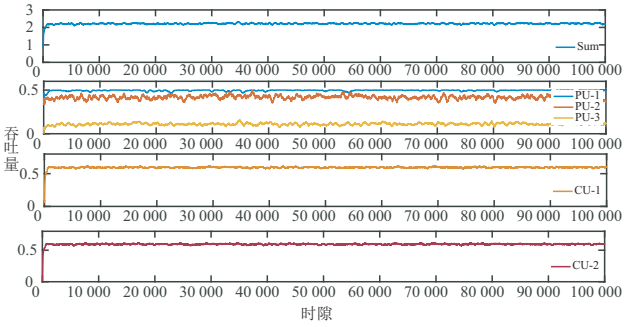

图10　MAPPO-DSA场景4测试阶段吞吐量

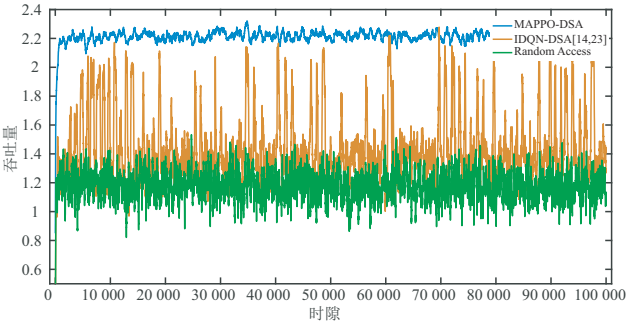

图11　场景4不同算法总吞吐量对比

5.2　结果分析

场景1~3中MAPPO-DSA总吞吐量均接近信道数量K,频谱利用率接近100%. 相比IDQN-DSA和RA,MAPPO-DSA有显著优势,且用户间公平性良好. 保证公平性的主要原因有以下两点.(1)所有策略网络使用了相同的结构与参数,即所有认知用户的接入策略都相同.(2)奖励的设置. 根据式(13),在信道k出现频谱空洞时,需要让信道k上等待时间最长的认知用户发送数据,才能最大化奖励,由此保证了用户公平.

MAPPO-DSA通过CTDE框架能够学习到性能优异且稳定性收敛性强的接入策略. 所有场景中MAPPO-DSA无论是系统总吞吐量还是稳定性都要优于IDQN-DSA和RA,也能有效保证认知用户间的公平性.

6　结论

本文提出了MAPPO-DSA算法,首先利用多信道接入机制来解决单信道接入存在的频谱浪费问题,随后引入MAPPO算法帮助认知用户在多信道接入的复杂场景下学习接入策略. 在学习过程中采用CTDE框架,将学习阶段细分为执行和训练阶段. 在执行阶段认知用户分布式执行策略并收集经验数据. 在训练阶段,认知用户上传经验数据至服务器集中训练. 经过一定次数的训练后,进入测试阶段,不再进行训练. 测试阶段中认知用户完全分布式执行自身策略,无需额外交换信息. 实验结果表明,MAPPO-DSA在不同的多认知用户多信道场景下收敛性良好,网络参数共享与强化学

习要素的设计能够有效保证认知用户间的公平性,系统吞吐量逼近理论上限.

参考文献

[1] HU F, CHEN B, ZHU K. Full spectrum sharing in cognitive radio networks toward 5G: A survey[J]. IEEE Access, 2018, 6: 15754-15776.

[2] AKYILDIZ I F, LEE W Y, VURAN M C, et al. A survey on spectrum management in cognitive radio networks[J]. IEEE Communications Magazine, 2008, 46(4): 40-48.

[3] TANDRA R, MISHRA S M, SAHAI A. What is a spectrum hole and what does it take to recognize one?[J]. Proceedings of the IEEE, 2009, 97(5): 824-848.

[4] 蒋师, 屈代明, 吴露露, 等. 动态频谱接入技术的分类和研究现状[J]. 通信技术, 2008, 41(11): 20-22.
JIANG S, QU D M, WU L L, et al. A taxonomy of dynamic spectrum access technologies and current research progress[J]. Communications Technology, 2008, 41(11): 20-22. (in Chinese)

[5] ZHAO Q, SWAMI A. A survey of dynamic spectrum access: Signal processing and networking perspectives[C]// 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2007: IV-1349-IV-1352.

[6] ZHAO Q, SADLER B M. A survey of dynamic spectrum access[J]. IEEE Signal Processing Magazine, 2007, 24(3): 79-89.

[7] 葛雨明, 孙毅, 蒋海, 等. 基于认知无线电技术的动态频谱分配方案研究[J]. 计算机学报, 2012, 35(3): 446-453.
GE Y M, SUN Y, JIANG H, et al. Research on dynamic spectrum allocation using cognitive radio technologies[J]. Chinese Journal of Computers, 2012, 35(3): 446-453.

[8] 胡浪涛, 毕松姣, 刘全金, 等. 基于深度强化学习的多小区NOMA能效优化功率分配算法[J]. 电子科技大学学报, 2022, 51(3): 384-391.
HU L T, BI S J, LIU Q J, et al. Multi-cell NOMA energy efficiency optimization power allocation algorithm based on deep reinforcement learning[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(3): 384-391. (in Chinese)

[9] 李保罡, 石泰, 陈静, 等. 基于强化学习的非正交多址接入和移动边缘计算联合系统信息年龄更新[J]. 电子与信息学报, 2022, 44(12): 4238-4245.
LI B G, SHI T, CHEN J, et al. Age of information updates in non-orthogonal multiple access-mobile edge computing system based on reinforcement learning[J]. Journal of Electronics & Information Technology, 2022, 44(12): 4238-4245. (in Chinese)

[10] WANG S, LIU H, GOMES P H, et al. Deep reinforcement learning for dynamic multichannel access in wireless networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(2): 257-265.

[11] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[EB/OL]. [2023]. http://arxiv.org/abs/1312.5602.pdf.

[12] 宋波, 叶伟, 孟祥辉. 基于多智能体强化学习的动态频谱分配方法综述[J]. 系统工程与电子技术, 2021, 43(11): 3338-3351.
SONG B, YE W, MENG X H. Review of multi-agent reinforcement learning based dynamic spectrum allocation method[J]. Systems Engineering and Electronics, 2021, 43(11): 3338-3351. (in Chinese)

[13] NAPARSTEK O, COHEN K. Deep multi-user reinforcement learning for distributed dynamic spectrum access[J]. IEEE Transactions on Wireless Communications, 2019, 18(1): 310-323.

[14] 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.

[15] GUO Z, CHEN Z, LIU P, et al. Multi-agent reinforcement learning-based distributed channel access for next generation wireless networks[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(5): 1587-1599.

[16] RASHID T, SAMVELYAN M, DE WITT C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. The Journal of Machine Learning Research, 2020, 21(1): 7234-7284.

[17] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.

[18] NAMISLO C. Analysis of mobile radio slotted ALOHA networks[J]. IEEE Journal on Selected Areas in Communications, 2006, 2(4): 583-588.

[19] CASSANDRA A R. A survey of POMDP applications[C]//AAAI 1998 Symposium on Planning with Partially Observable Markov Decision Processes. California: AAAI, 1998: 1724.

[20] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. [2023]. http://arxiv.org/abs/1707.06347.pdf.

[21] PETERS J, SCHAAL S. Natural actor-critic[J]. Neurocomputing, 2008, 71(7-9): 1180-1190.
[22] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction[M]. Massachusetts: MIT Press, 2018.
[23] DIALLO E A O, SUGIYAMA A, SUGAWARA T. Learning to coordinate with deep reinforcement learning in doubles pong game[C]//16th IEEE International Conference on Machine Learning and Applications (ICMLA). Piscataway: IEEE, 2017: 14-19.
[24] NAPARSTEK O, COHEN K. Deep multi-user reinforcement learning for distributed dynamic spectrum access[J]. IEEE Transactions on Wireless Communications, 2019, 18(1): 310-323.

作者简介

陈平平　男,1986年12月生,教授,福州大学先进制造学院,主要研究方向为强化学习、压缩感知、信道编码与无线通信等.中国电子学会会员编号:E190021215M.

张　旭　男,1997年9月生,2023年毕业于福州大学先进制造学院,获得硕士学位,主要研究方向为强化学习与无线通信.

谢肇鹏　男,1995年7月生,讲师,福州大学先进制造学院,主要研究方向为强化学习、信道编码与无线通信等.中国电子学会会员编号:E190156454M.
E-mail: xzp_fzu@163.com

丘毓萍　女,1999年10月生,福州大学物理与信息工程学院信息与通信工程博士研究生,主要研究方向为强化学习、多址接入、信道编码等.

方　毅　男,1986年8月生,教授,广东工业大学信息工程学院,研究方向为信息论与信道编码、无线通信、数据存储编码.中国电子学会会员编号:E190028682M.