

doi: 10.12052/gdutxb.220159

面向多用户动态频谱接入的改进双深度Q网络方法研究

何一汕, 王永华, 万频, 王磊, 伍文韬

(广东工业大学 自动化学院, 广东 广州 510006)

摘要: 随着移动通信技术的飞速发展, 有限的频谱利用资源与大量频谱通信需求之间的矛盾也日益加剧, 需要新的智能方法来提高频谱利用率。本文提出了一种基于分布式优先经验池结合双深度Q网络的多用户动态频谱接入方法。通过该方法, 次用户可以在动态变化的认知无线网络环境下根据自己感知信息来不断地学习, 选择空闲信道完成频谱接入任务来提高频谱利用率。该方法采用分布式强化学习框架, 将每个次用户视为一个智能体, 各个智能体采用标准单智能体强化学习方法进行学习以降低底层计算开销。另外, 该方法在神经网络训练的基础上加入优先级采样, 优化了神经网络的训练效率以帮助次用户选择出最优策略。仿真实验结果表明该方法能提高接入信道时的成功率、降低碰撞率和提升通信速率。

关键词: 动态频谱接入; 分布式强化学习; 优先经验池; 深度强化学习

中图分类号: TN929.5

文献标志码: A

文章编号: 1007-7162(2023)04-0085-09

An Improved Double Deep Q Network for Multi-user Dynamic Spectrum Access

He Yi-shan, Wang Yong-hua, Wan Pin, Wang Lei, Wu Wen-tao

(School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: With the rapid development of mobile communication technology, the contradiction between the limited spectrum utilization resources and the demand of a lot of spectrum communication is increasingly aggravated. New intelligent methods are needed to improve the utilization rate of spectrum. A multi-user dynamic spectrum access method based on distributed priority experience pool and double deep Q network is proposed. This method can help the secondary users to continuously learn according to their perceived environment information in the dynamic environment, and choose the idle channel to complete the spectrum access task for improving the spectrum utilization rate. In this method, a distributed reinforcement learning framework is adopted, and each secondary user is regarded as an agent. Each agent learns by using standard single-agent reinforcement learning method to reduce the underlying computing overhead. In addition, the method adds priority sampling on the basis of neural network training, and then optimizes the training efficiency of neural network to help sub-users choose the optimal strategy. The simulation results show that this method can improve the success rate, reduce the collision rate and improve the communication rate.

Key words: dynamic spectrum access; distributed reinforcement learning; prioritized experience pool; deep reinforcement learning

近年来,随着无线设备和移动应用的不断普及,数据流量呈爆炸式增长,有效利用频谱资源是未来

无线通信网络应对这种数据流量高速增长的关键。为了应对频谱通信资源短缺的问题,动态频谱接入

收稿日期: 2022-10-19

基金项目: 国家自然科学基金资助项目(61971147)

作者简介: 何一汕(1998-),男,硕士研究生,主要研究方向为认知无线网络和深度强化学习

通信作者: 王永华(1979-),男,副教授,博士,主要研究方向为认知无线网络、机器学习, E-mail: wangyonghua@gdut.edu.cn

(Dynamic Spectrum Access, DSA) 技术可以发挥有效的作用。动态频谱接入模型主要分为3种: 动态开放模型、开放共享模型和分层接入模型^[1]。3种模型之中分层接入模型与现有频谱接入方式兼容性较好, 其侧重点在于将频谱资源划分给主用户(Primary User, PU) 与次用户(Secondary User, SU), 主用户在进行频谱通信时候拥有优先权。在分层接入模型之中的覆盖式模式即机会式频谱接入, 可以帮助次用户在感知频谱空穴之后接入空闲频谱, 这样可以有效地降低频谱资源短缺带来的压力。

DSA技术作为频谱共享、提高频谱利用率的关键技术之一, 近年来备受关注。研究人员已经为提升频谱效率提出了不少解决办法。文献[2-4]介绍了基于图论和博弈论方法来解决频谱分配问题, 但是这类方法需要所有用户提前知道所处环境的频谱信道状态信息, 而大多时候次用户是无法获取系统的完整信息的, 只能基于部分感知信息选择信道。不过强化学习中基于无模型策略的学习方法, 可以在不需要系统环境的先验知识基础上, 通过智能体与环境不断交互学习使得智能体找到最优策略, 比如文献[5-6]将Q学习方法运用到物联网通信中, 但Q学习方法收敛缓慢且不能处理高维度的状态和动作空间, 为此研究人员通过结合了深度学习与Q学习的深度Q学习方法来满足对于认知无线环境的自适应管理^[7-8]。目前深度Q学习在频谱接入技术研究中已经取得了一些进展。例如: 文献[9]提出了一种基于深度Q学习的频谱分配方法, 以支持D2D用户无线接入网络的子信道分配和模式选择。文献[10]也提出了基于深度Q学习的频谱分配方法, 帮助次用户在多个微信蜂窝基站中进行关联和资源分配。虽然文献[9-10]中深度Q学习方法一定程度上克服了Q学习方法所带来的缺点, 但是该类方法主要使用了集中式的深度强化学习算法来帮助次用户进行频谱分配, 并没有考虑到随着次用户数量增加、动作空间变大的情况带来的“维数灾难”问题^[11-12]。并且这类方法在训练神经网络过程中对于数据的采集方式为传统的均匀采样, 没有充分利用数据池中比较“有效”的数据以帮助智能体进行更高效的学习。

因此本文提出了一种基于分布式优先经验池结合双深度Q网络(Distributed Priority Experience Pool and Double Deep Q Network, DPEP-DDQN) 的多用户动态频谱接入方法, 用来帮助次用户根据自己的通信需求和所处环境信息来进行信道选择。其贡献

如下。

(1) 为了解决动作空间增大带来的维数灾难问题, 采用了分布式强化学习框架, 将每个次用户视为一个智能体来对动态频谱接入进行建模分析。

(2) 为了使神经网络模型可以更充分利用有效数据来进行训练, 采用优先经验回放算法对神经网络中的训练样本进行优化, 加快收敛速度。

1 系统模型

如图1所示, 假设一个微小区的环境中随机分布若干主用户和次用户。主用户集合为 $M = \{m_1, m_2, \dots, m_J\}$ 一个主用户由一个主用户发射机和主用户接收机组成, J 为主用户总数。次用户集合为 $N = \{n_1, n_2, \dots, n_L\}$, 其中一个次用户由一个次用户发射机和次用户接收机组成, L 为次用户总数。假设存在 J 条信道正交授权可以均匀分配给各个主用户正常数据通信, 主用户通信时不需要考虑其余次用户是否存在, 只需根据自身传输需求进行传输即可。次用户在进行频谱接入的时候不可以打扰到主用户的正常通信, 即主用户在该条信道进行数据通信时次用户不可以接入, 只有当主用户不使用该频谱信道的情况下次用户才可以通过机会式接入主用户的授权信道进行通信。此外, 若主用户所在信道发生碰撞, 主用户会将信道碰撞的相关警告信息广播给次用户以对自身保护。因此次用户需要根据自身感知情况所获得的信息, 通过自身学习来选择信道。

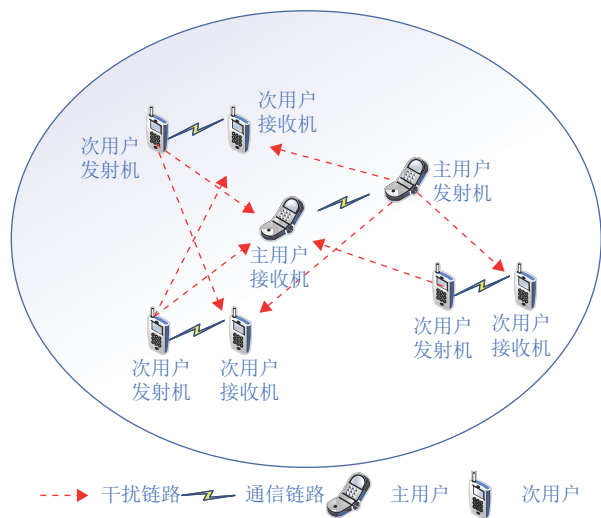


图1 认知网络环境

Fig.1 Cognitive network environment

1.1 信道状态模型

次用户要想成功接入信道, 需要感知想选择信

道的状态。每条信道可能处于两种状态:空闲状态(1)或者活跃状态(0)。若信道处于活跃状态,表明有主用户在使用这条信道,次用户在此期间不能接入该条信道,否则该时间段主用户没有使用这条信道,这时候次用户可以采用机会式接入该条信道进行通信。第 j 条信道的活动状态可被描述为马尔科夫链,如图2所示,其状态转移概率可表示为

$$\xi_j = \begin{bmatrix} \xi_{00}^j & \xi_{01}^j \\ \xi_{10}^j & \xi_{11}^j \end{bmatrix} \tag{1}$$

式中: $\xi_{ab}^j = \Pr\{\text{下一状态}b|\text{当前状态}a\}, (a, b \in 0, 1)$ 。

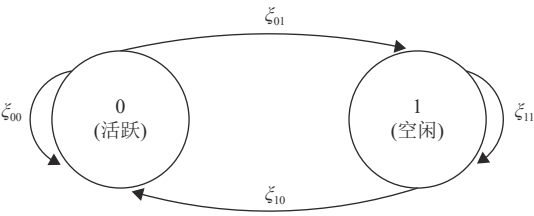


图2 信道状态的马尔科夫链
Fig.2 Markov chain of channel states

1.2 传输速率模型

涉及到的系统模型参数如表1所示。

表 1 系统模型参数
Table 1 System model parameters

符号	含义
J	主用户总数
L	次用户总数
p_{jj}/mW	第 j 条信道上主用户 j 的传输功率
p_{lj}/mW	第 j 条信道上次用户 l 的传输功率
p_{kj}/mW	第 j 条信道上次用户 k 的传输功率
$ h_{ll} ^2$	次用户信号发射机 l 与信号接收机 l 之间的信道增益
$ h_{jl} ^2$	主用户信号发射机 j 与次用户信号接收机 l 之间的信道增益
$ h_{kl} ^2$	次用户信号发射机 k 与次用户信号接收机 l 之间的信道增益
B/MHz	信道带宽
$\theta/(\text{mW} \cdot \text{Hz}^{-1})$	噪声频谱密度
f_c	无线信道的载波频率
$\bar{Z}$	参考距离处的路径损耗
Z	路径损耗
A_w	路径损耗指数
$\mathcal{B}_w$	路径损耗频率相关性
κ	视线路径和散射路径的接收机信号功率之间的比率
δ_j^l/dB	第 l 个次用户选择信道 j 后信干噪比
T_l/Mbps	第 l 个次用户的理论传输速率

从图1可知 J 个主用户和 L 个次用户在一个微小区的二维的空间中随机分布,且在知道发射机和接收机位置距离的情况下,可以通过统计信息的方法

来建立信道模型^[13],即在已知传播距离的情况下,根据不同场景的概率分布,通过WINNER II模型^[14]来定义其中的路径损耗:

$$Z(d, f_c) = \bar{Z} + A_w \lg d + \mathcal{B}_w \lg(f_c/5) \tag{2}$$

因此可以获得期望信号的路径损耗 $Z(d_{ll}, f_c)$,以及干扰信号的路径损耗 $Z(d_{jl}, f_c)$ 和 $Z(d_{kl}, f_c)$,其中 d_{ll} 、 d_{jl} 、 d_{kl} 分别为第 l 个次用户发射机与接收机之间的通信链路距离,第 j 个主用户发射机对第 l 个次用户接收机的干扰链路距离和第 k 个次用户发射机对第 l 个次用户接收机的干扰链路距离。假设发射机和接收机之间存在强视线路径(Line of Sight, LoS),则可采用Rician信道模型推导信道增益,可表示为

$$h = \sqrt{\frac{\kappa}{\kappa+1}} \sigma e^{j\theta} + \sqrt{\frac{1}{\kappa+1}} \text{CN}(0, \sigma^2) \tag{3}$$

$$\sigma^2 = 10^{-[(\bar{Z} + A_w \lg d + \mathcal{B}_w \lg(f_c/5))/10]} \tag{4}$$

式中: σ^2 取决于路径损耗, κ 为视线路径和散射路径的接收机信号功率之间的比率。 $\theta \sim U(0, 1)$ 为强视线路径上到达信号的相位,取值于0和1之间的均匀分布。 $\text{CN}(\cdot)$ 表示一个圆对称的复高斯随机变量。

因此,第 l 个次用户选择信道 j 后接收机收到信号的信干噪比(Signal to Interference plus Noise Ratio, SINR)可表示为

$$\delta_j^l = \frac{p_{lj}|h_{ll}|^2}{p_{jj}|h_{jl}|^2 + \sum_{k=1, k \neq l}^L p_{kj}|h_{kl}|^2 + B\theta} \tag{5}$$

根据香农定理,第 l 个次用户的理论传输速率为

$$T_l = B \lg(1 + \delta_j^l) \tag{6}$$

式中: B 为信道的带宽。本文研究目的是最大化所有次用户的总传输速率,故研究问题的优化目标可简化为

$$\max \sum_{l=1}^L T_l \tag{7}$$

2 基于强化学习的频谱接入建模

强化学习是一种通过让智能体与环境不断交互进行学习的方法^[15],深度强化学习是在强化学习的基础上引入深度学习,使得强化学习拥有强大的算力,从而帮助智能体适应动态环境进行自行学习。如图3所示,强化学习主要有5大要素:环境、智能体、状态 s 、动作 a 和奖励 r 。智能体想要得到的策略是由环境给出的状态 s 和智能体所做出的动作 a 以及得到的奖励 r 所决定的,而强化学习的目的就在于通

过对环境进行有限的交互次数找到最优策略 $\pi_*(s)$, 最大化奖励值, 其中 $\pi_*(s)$ 可表示为

$$\pi_*(s) = \operatorname{argmax}_a Q(s, a) \quad (8)$$

式中: Q 为值函数。

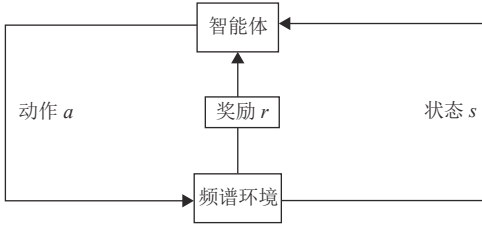


图3 智能体与环境交互模型

Fig.3 Agent and environment interaction model

本文研究的动态频谱接入过程的本质是一个马尔可夫决策过程, 可以利用强化学习来帮助解决。次用户机会式访问决策过程如图4所示, 次用户接入策略是由双深度Q网络和当前频谱感知结果共同决定的。次用户根据自身频谱接入策略接入无线信道进行数据传输, 然后次用户接收机根据实际无线传输质量得到反馈奖励, 反馈奖励由次用户发射机存储, 并作为双深度Q网络的训练数据用来更新频谱接入策略。

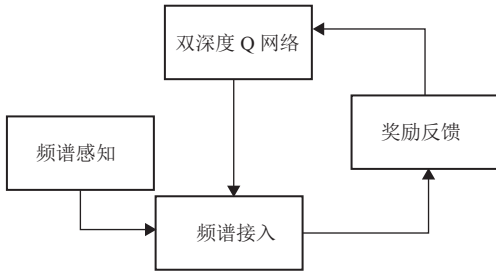


图4 动态频谱接入过程

Fig.4 Dynamic spectrum access process

2.1 状态集合

在每个时刻开始时, 信道状态感知结果可作为状态值, 则在 t 时刻状态集合可以表示为

$$S(t) = [S_1(t), \dots, S_l(t), \dots, S_L(t)] \quad (9)$$

式中: $S_l(t)$ 为 l 个次用户所感知到的信道状态, 是一个 J 维向量 $[s_1^l(t), \dots, s_j^l(t), \dots, s_J^l(t)]^T$, 并且 $s_j^l(t) \in \{0, 1\}$, $s_j^l(t) = 0$ 表示第 l 个次用户感知到第 j 条信道处于活跃状态; 反之 $s_j^l(t) = 1$ 表示第 l 个次用户感知到第 j 条信道处于空闲状态, 次用户可以选择机会式接入该信道。并且考虑到次用户在进行频谱感知的时候可能会存在一定的误差, 因此给 $s_j^l(t)$ 的感知结果加上一

定错误的概率 E_l^j , 故假设第 l 个次用户在第 j 个信道的真实情况为

$$P\{T_l^j(t) = s_l^j(t)\} = 1 - E_l^j \quad (10)$$

2.2 动作集合

在次用户感知完信道状态后, 次用户可以基于感知结果进行信道访问, 但是每个次用户最多访问一个信道, 其中次用户 t 时刻访问的动作集合表示为

$$A(t) = [A_1(t), \dots, A_l(t), \dots, A_L(t)] \quad (11)$$

式中: $A_l(t) \in \{0, \dots, j, \dots, J\}$, $A_l(t) = j$ 为 l 个次用户选择了第 j 个信道, 若 $A_l(t) = 0$, 则第 l 个次用户没有选择任何信道。

2.3 奖励函数

次用户 l 在 t 时刻根据信道状态做出信道 j 选择后, 为了实现传输速率最大化, 可以将奖励函数按照如下方法建立, 如式(12)所示。

$$R_l(t) = \begin{cases} -2, & \text{与主用户发生碰撞} \\ 0, & \text{不做任何选择} \\ Blb(1 + \delta_j^l), & \text{否则} \end{cases} \quad (12)$$

(1) 次用户访问信道时没有主用户正在使用, 并且只有一个次用户接入信道, 可以用理论传输速率 $Blb(1 + \delta_j^l)$ 当作奖励, 并且记成功接入一次。

(2) 次用户访问信道时主用户正在使用, 与主用户发生碰撞, 给予-2作为惩罚, 并记碰撞一次。

(3) 次用户访问信道时没有主用户正在使用, 但同时有多个次用户访问相同信道, 给予 $Blb(1 + \delta_j^l)$ 作为奖励, 记碰撞一次。

(4) 次用户决定不访问任何信道, 不给予奖励, 成功次数和碰撞均不记。

3 DPEP-DDQN算法描述

本文提出的DPEP-DDQN算法, 是在进行强化学习建模的基础上结合了分布式独立Q学习、优先经验回放算法和双深度Q网络的方法形成的。

3.1 分布式独立Q学习

为了减少集中式训练所带来的底层开销, 本文在进行强化学习建模的基础上加入分布式独立Q学习方法。在此方法中, 每个智能体都是独立学习的主体, 它们将其他智能体视为环境的一部分, 并认为联合状态可观测, 采用标准单智能体强化学习方法进行学习^[16]。如图5所示, 在与环境互动的过程中, 每个智能体可以只需要考虑自己的状态对环境做出反馈, 而不需考虑其余智能体的动作选择, 故这种情况

更贴近于认知无线网络环境中次用户不会在选择信道之前进行相互协商的情况。其中分布式独立Q学习方法的主要目标可简化为在满足次用户不对主用

户造成干扰的情况下,提高所有次用户的传输速率。故需要系统找到最优策略 $\pi_* = \pi_*^1 + \pi_*^2 \cdots + \pi_*^l + \cdots + \pi_*^L$, 其中 π_*^l 为第 l 个次用户信道选择的最优策略。

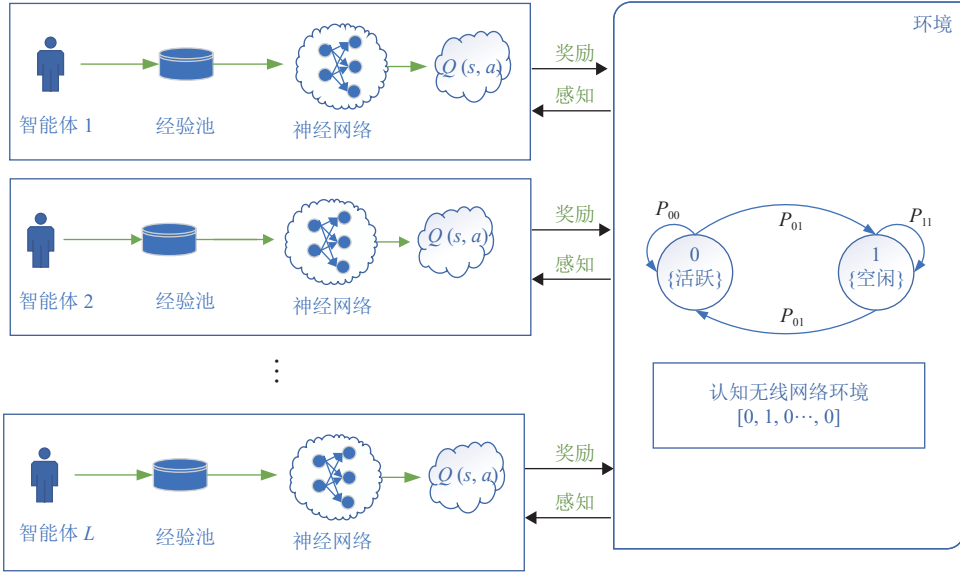


图5 分布式独立Q学习

Fig.5 Distributed independent Q learning

3.2 双深度Q网络

在准备分布式训练框架之后,神经网络将采用双深度Q网络(Double Deep Q Network, Double DQN), 该神经网络属于深度Q网络(Deep Q Network, DQN)的一种变体,相对于DQN其特点就是使用2次Q学习来寻找最优动作,通过解耦目标Q值动作的选择和目标Q值的计算,达到消除过度估计的目的^[17]。神经网络结构图如图6所示。

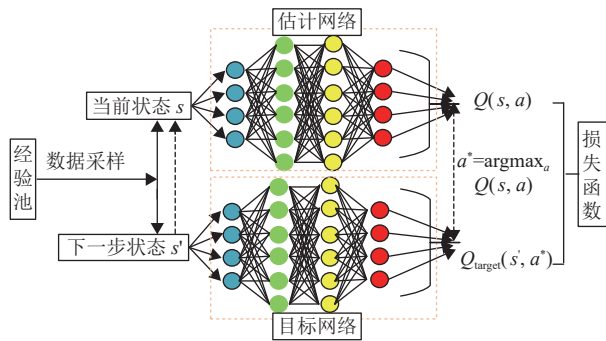


图6 神经网络结构图

Fig.6 Neural network structure diagram

3.3 优先经验回放

为了提升神经网络训练效率,本文采用了优先经验回放算法。在传统DQN算法中,为了打破样本之间的关联,通常的做法是在经验池中随机抽取样本训练进而更新参数来构建神经网络,但是这常常会

出现奖励稀疏的情况,例如智能体需要进行很多步探索才得到奖励,使得经验池中能让智能体学习的有效经验十分缺少,这个时候如果仍然采取随机抽取经验池样本的方式会使得效率降低。因此,本文将采取一种二叉树的结构来对数据进行存储,该数据存储结构在一定程度上可以在数据采样的时候节省时间,而且可以定义出便于神经网络训练的采样规则,比如可以对样本设置不同的优先级。

如图7所示,本文将采用一种基于SumTree即累计求和的二叉树来作为经验池储存结构。SumTree中每个底层的叶子节点存储样本数据以及对应的优先级,中继节点和根节点不储存样本数据而储存下级节点的优先级之和。叶子节点中存储的数据值区间的数值越大意味着优先级越高,能够存储的数字区间越长。在保证优先级采样的情况下也需要保障最低优先级非零概率采样,故设定次用户 l 选择采样概率计算方法为

$$H_l(x) = \frac{[h_l(x)]^\alpha}{\sum_y [h_l(y)]^\alpha} \quad (13)$$

式中: $[h_l(x)]^\alpha$ 和 $[h_l(y)]^\alpha$ 为样本 x 和样本 y 的优先级, α 为优先级指数, $\alpha = 0$ 时为均匀采样。 $[h_l(x)]^\alpha$ 的计算方法为

$$[h_l(x)]^\alpha = |\Phi_l(x)| + \varepsilon \quad (14)$$

$$\Phi_l(x) = R_l(x) - Q(S_l(x-1), A_l(x-1)) + \gamma Q_{\text{target}}(S_l(x), \arg\max_a Q(S_l(x), A_l(x))) \quad (15)$$

式中: $\Phi_l(x)$ 为样本 x 的时序差分(Temporal Difference, TD) 误差, Q 和 Q_{target} 分别为估计网络和目标网络输出的 Q 值, $R_l(x)$ 为样本 x 的奖励. ε 为一个非常小的正数, 其目的是为了保证采样概率 $H_l(x)$ 大于 0.

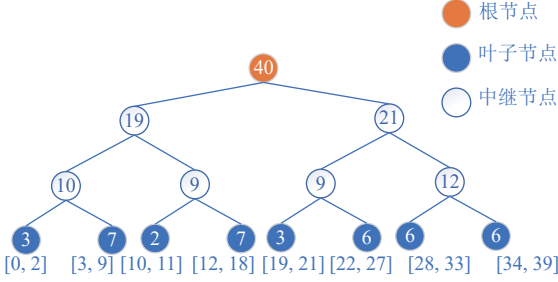


图7 SumTree结构图

Fig.7 SumTree structure

但是该采样机制可能会带来偏差, 为了使系统更加稳定, 故设定样本重要权重来纠正偏差:

$$w_l(x) = \left(\frac{1}{O_l} \times \frac{1}{H_l(x)} \right)^\beta \quad (16)$$

式中: $w_l(x)$ 为次用户 l 的标识权重系数, O_l 为次用户 l 的经验池大小, β 为标识非均匀补偿系数, 当 $\beta = 1$ 时就完全补偿了 $H_l(x)$.

3.4 DPEP-DDQN算法流程

算法1 DPEP-DDQN算法伪代码

1. 初始化: 主次用户的发射机与接收机位置, 各次用户神经网络和经验池的超参数
2. For 次用户 $l = 1$ to L do
3. 次用户 l 感知频谱环境, 初始化状态 $S_l(0)$ 和选择动作 $A_l(0)$
4. For 训练步数 $t = 1$ to T do
5. 次用户 l 以贪婪算法选择动作, $1 - \varepsilon$ 概率选择动作 $A_l(t) = \arg\max_a Q(S_l(t), A_l(t))$, 否则随机选择动作 $A_l(t)$
6. 获取环境反馈即时奖励 $R_l(t)$ 和下一时刻状态 $S_l(t+1)$
7. 将样本数据 $\{(S_l(t), A_l(t), R_l(t), S_l(t+1))\}$ 存入到经验池 D_l 并赋予优先级
8. 状态迭代 $S_l(t) = S_l(t+1)$
9. IF t 到达训练点 then
10. For $x = 1$ to y do
11. 从经验池从 D_l 中抽取样本并遵循 $x \sim H_l$ 的规律
12. 计算采样权重:

$$w_l(x) = \left(\frac{1}{O_l} \times \frac{1}{H_l(x)} \right)^\beta \max_y w_l(y)$$

13. 计算TD误差:

$$\Phi_l(x) = R_l(x) + \gamma Q_{\text{target}}(S_l(x), \arg\max_a Q(S_l(x), A_l(x))) - Q(S_l(x-1), A_l(x-1))$$

14. 更新样本优先级: $h_l(x) = |\Phi_l(x)|$

15. 累计权重变化: $\Delta_l = \Delta_l + w_l(x) \cdot \Phi_l(x) \cdot \nabla_{\theta} Q(S_l(x-1), A_l(x-1))$

16. End For

17. 更新神经网络权重 $\theta_l = \theta_l + \eta_l \cdot \Delta_l$, 重置权重差值 $\Delta_l = 0$

18. 更新目标网络参数 $\hat{\theta}_l = \theta_l$

19. End IF

20. 每隔一定步数随机更新环境的参数

21. End For

22. End For

4 实验环境介绍

本节中将设置系统模型的仿真参数。参考文献[14]中关于城市微小区场景中的WINNER II模型, 假设超参数 $f_c = 2.4$ GHz, $\bar{Z} = 41$, $A_w = 22.7$, $B_w = 20$, $\kappa = 8$, σ^2 由路径损耗所决定。其中次用户接收机所接收到的SINR由式(5)给出, 其中信道带宽 $B = 1$ MHz, 次用户的传输功率设置为50 mW, 主用户的传输功率设置为65 mW。

如动态频谱接入模型即信道状态模型(见图2), 将主用户的状态分割成2个独立的马尔科夫链: 主用户不在使用(1) 和主用户已经在使用(0)。为了初始化每个马尔科夫链, 分别从 $[0.7, 1]$, $[0, 0.3]$ 区间上的均匀分布中随机选取 ξ_{11} 和 ξ_{00} , 那么 $\xi_{10} = 1 - \xi_{11}$, $\xi_{01} = 1 - \xi_{00}$ 就可以计算得出。之所以选择这2个区间范围的原因是 ξ_{11} 和 ξ_{00} 大多数授权频段的利用率偏低, 所以 ξ_{00} 的值是偏低的, 并且 ξ_{11} 的值偏高。

为了更好地实验演示, 本次仿真还加入了myopic方法作为参考方法进行对比。在myopic方法中, 智能体将选择最大的即时奖励。为了计算预期的即时回报, myopic方法需要信道的转移概率和感知错误概率的信息。在myopic中, 第 l 个次用户首先计算信道状态概率, 其中 $G_j(t) = \Pr\{\text{在} t \text{时刻第} j \text{个信道状态} = 1\}$, 其表达式为

$$G_j(t) = s_l^j(t)(1 - E_l^j) + (1 - s_l^j(t))E_l^j \quad (17)$$

式中: $s_l^j(t)$ 为第 l 个次用户于第 j 个信道的感应结果, E_l^j 为第 l 个次用户于第 j 个信道感知错误的概率, 假设其为0.2。因此myopic方法可以计算第 j 个信道的奖励函数为

$$R_j(t) = G_j(t) [\xi_{10}^j(-2) + \xi_{11}^j B \text{lb}(1 + \delta_j^l)] + (1 - G_j(t)) [\xi_{00}^j(-2) + \xi_{01}^j B \text{lb}(1 + \delta_j^l)] \quad (18)$$

5 实验仿真与分析

本节将对所提方案进行仿真实验和性能评估,实验将以myopic方法为参考基准,此外,需要注意的是,每个主用户都拥有自己的信道,并且假设所有主用户不会同时使用信道。实验使用python语言编写,以Tensorflow进行建模,构建如图6所示的神经网络结构。构建的神经网络有2个隐藏层,每个隐藏层中的神经元数目都是64,激活函数都是tanh函数。训练总次数是378 000,记忆库D的容量O为5 400。

5.1 位置分布

图8描述了在假设的一个半径为100 m的圆形二维空间中,9个主用户和8个次用户随机位置的分布,其中主用户发射机、接收机随机分布,次用户接收机与发射机之间的距离在20~50 m之间随机选择。

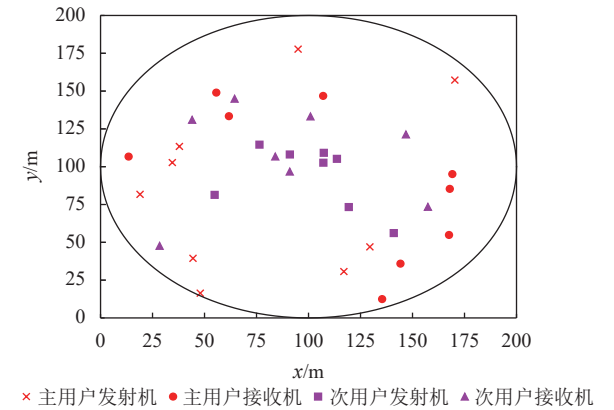


图8 位置分布图(9个主用户与8个次用户)
Fig.8 Location distribution map (9 PU and 8 SU)

5.2 平均成功率

图9描述了在378 000次迭代中,次用户在进行频谱接入时的平均成功率,每隔27 000次进行一次平均统计。从图中比较算法可以得出:当存在感知错误可能性时,作为参考的Myopic表现得不太好。虽然Myopic算法将信道的转移概率和感知误差概率视为已知信息,但是该算法完全不考虑未来奖励,故成功率一直维持在较低水平。Q-learning需要密集的Q表更新,故其在面对传输数据较大时,环境反馈的复杂性提高,智能体难以理解环境,从而达不到好的效果,其成功率也仅在0.41左右波动。相比较于加入神经网络进行训练的深度Q学习算法,在神经网络的优化下,深度Q学习算法可以取得一个很好的学习结果,其中按照性能升序排序依次是:DQN, Double DQN, DPEP-DDQN。DQN 算法由于本身的神经网络

学习效果差,使得接入成功率只能提升到0.58左右。当训练网络更换成Double DQN后,每个智能体变成使用2次Q学习来寻找最优动作,这使得智能体可以更好地学习,使得成功率可以达到0.85左右。虽然DPEP-DDQN与Double DQN在收敛的时候达到了相似的成功率,但是加上了优先经验回放的DPEP-DDQN算法的收敛速度会比普通Double DQN快。

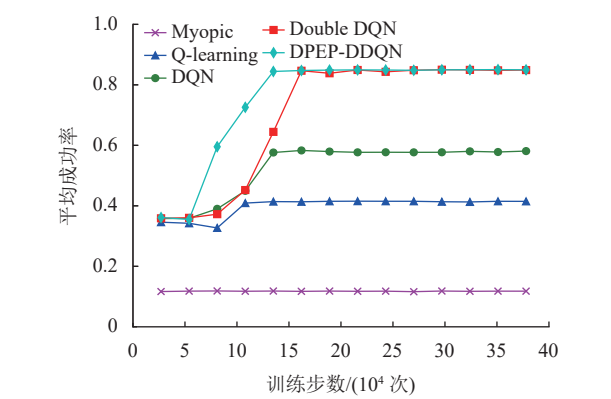


图9 平均成功率(9个主用户与8个次用户)
Fig.9 Average success rate (9 PU and 8 SU)

5.3 平均碰撞率

图10描述了次用户在进行信道争夺时产生的平均碰撞率。对比图9可以发现虽然用户可以选择接入信道,但是在以传输速率为期望回报的情况下,次用户都会努力尝试接入信道,故成功率和碰撞率总和接近1。从图中可知Myopic算法和Q-learning算法都表现得不太好,信道的碰撞率较高,这种情况下会造成整个认知无线网络环境通信的干扰加深,影响整个网络的传输速率。相比之下DPEP-DDQN和Double DQN算法虽然存在差异但是最终都可达到一个较低的碰撞率,尤其是DPEP-DDQN算法,不仅最终收敛的碰撞率维持最小,而且最终曲线收敛最快。

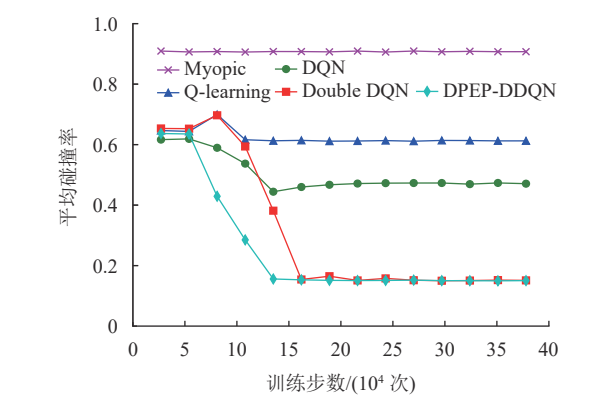


图10 平均碰撞率(9个主用户与8个次用户)
Fig.10 Average collision rate (9 PU and 8 SU)

5.4 平均奖励

图11描述了次用户在进行信道争夺时产生的平均奖励值,即信道通信的理论传输速率。在参数相同的模拟环境下DPEP-DDQN可以以最快的速度达到最大奖励。这是因为加入了SumTree累计求和的二叉树作为经验池储存结构之后,智能体就不会像传统模型那样进行随机采样,而是对数据加入相应的优先级,这使得神经网络训练时能够更加有效地采集样本,不仅可以优化数据库的容量,而且可以更好地保证整个网络的传输速率。

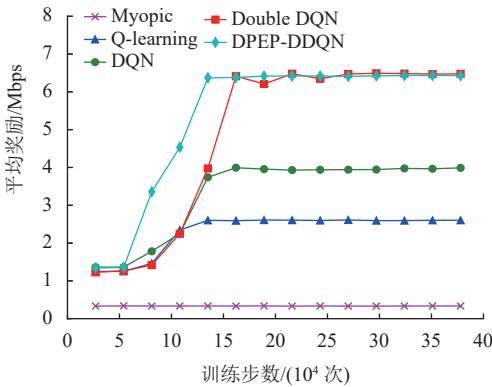


图 11 平均奖励(9个主用户与8个次用户)
Fig.11 Average reward (9 PU and 8 SU)

5.5 不同次用户数目下的最终平均奖励

图12描述了在存在9个主用户的情况下,次用户的数目分别从1增加到8时次用户最终收敛时的平均奖励。可以看出,在这5种算法中最终平均奖励都随着次用户数目的增加而呈现减少趋势。这是因为随着次用户数目的增加,通信资源逐渐短缺,可以使用的信道数目也会逐渐减少,因此理论传输速率就会随之下降。不难看出,在通信资源充足的情况下,各类算法最终都可以达到较好的奖励值,可是随着频谱资源的逐渐减少,各类算法之间的差距也开始渐

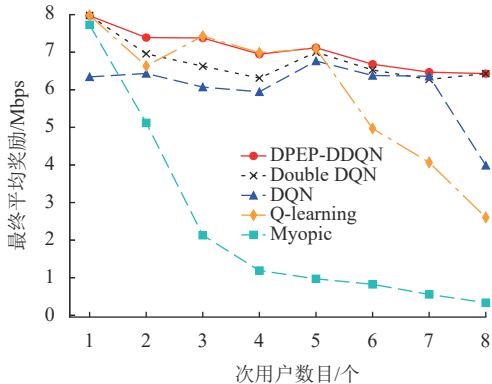


图 12 不同次用户数目下的最终平均奖励

Fig.12 The final average reward for different number of SU

渐增加。由此可以看出本文提出的DPEP-DDQN算法在面临通信资源减少的时候仍然可以保持一个较好的传输速率。

6 结论

为了解决认知无线网络中多个次用户需要快速且准确接入频谱信道的问题,本文提出了一种基于分布式优先经验池结合双深度Q网络的多用户动态频谱接入方法,用来帮助次用户在不完美感知频谱环境的情况下,机会式地做出适当的频谱访问决策。实验结果表明,该方法不仅可以保证最高的接入成功率和传输速率,并且在次用户数目不增加的情况下,该算法的最大化理论传输速率和收敛速度仍然可以保持良好水平。不仅可以帮助认知无线网络降低干扰,还有效地提高了整个网络的传输速率。

参考文献:

[1] SU H, ZHANG X. Cross-layer based opportunistic MAC protocols for QoS provisionings over Cognitive radio wireless networks [J]. *IEEE Journal on Selected Areas in Communications*, 2008, 26(1): 118-129.

[2] WANG J, HUANG Y, JIANG H. Improved algorithm of spectrum allocation based on graph coloring model in cognitive radio[C]//2009 WRI International Conference on Communications and Mobile Computing. Kunming: IEEE, 2009: 353-357.

[3] GAO L, DUAN L, HUANG J. Two-sided matching based cooperative spectrum sharing [J]. *IEEE Transactions on Mobile Computing*, 2017, 16(2): 538-551.

[4] 刘新浩, 马昕睿, 王大为. 基于图论模型的认知无线电频谱分配仿真建模研究[J]. *电脑与电信*, 2021(3): 16-20.

LIU X H, MA X R, WANG D W. Simulation modeling of cognitive radio spectrum allocation based on graph theory model [J]. *Computer & Telecommunication*, 2021(3): 16-20.

[5] LIU X, SUN C, ZHOU M, *et al.* Reinforcement learning based dynamic spectrum access in cognitive internet of vehicles [J]. *China Communications*, 2021, 18(7): 58-68.

[6] 郑思远, 崔苗, 张广驰. 基于强化学习的无人机安全通信轨迹在线优化策略[J]. *广东工业大学报*, 2021, 38(4): 59-64.

ZHENG S Y, CUI M, ZHANG G C. Reinforcement learning-based online trajectory optimization for secure UAV communications [J]. *Journal of Guangdong University of Technology*, 2021, 38(4): 59-64.

[7] XU F, YANG F, BAO S, *et al.* DQN inspired joint computing and caching resource allocation approach for software defined information-centric Internet of Things network [J]. *IEEE Access*, 2019, 7: 61987-61996.

[8] CHEN Y, LI Y, XU D, *et al.* DQN-based power control for

IoT transmission against jamming[C]//2018 IEEE 87th Vehicular Technology Conference (VTC Spring) . Portugal: IEEE, 2018: 1-5.

[9] SUN Y, PENG M, POOR H V. A distributed approach to improving spectral efficiency in uplink device-to-device-enabled cloud radio access networks [J]. *IEEE Transactions on Communications*, 2018, 66(12): 6511-6526.

[10] KAI C H, MENG X W, MEI L S, *et al.* Deep reinforcement learning based user association and resource allocation for d2d-enabled wireless networks[C]//2021 IEEE/CIC International Conference on Communications in China (ICCC) . Xiamen: IEEE, 2021: 1172-1177.

[11] ASMUTH J, LI L, LITTMAN M L, *et al.* A bayesian sampling approach to exploration in reinforcement learning [J]. *Eprint Arxiv*, 2009, 58(7): 1805-1810.

[12] ZHANG R B, ZHONG Y, GU G C. A new accelerating algorithm for multi-agent reinforcement learning [J]. *Journal of Harbin Institute of Technology*, 2005, 12(1): 48-51.

[13] VIRIYASITAVAT W, BOBAN M, TSAI H M, *et al.* Vehicular communications: survey and challenges of channel and propagation models [J]. *IEEE Vehicular Technology Magazine*, 2015, 10(2): 55-66.

[14] MEINIL J, KYSTI P, JMS T, *et al.* WINNER II channel models[M]. New Jersey: John Wiley & Sons, Ltd, 2009.

[15] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 2018.

[16] 傅波. 基于交替跟踪的分布式多智能体合作学习算法研究[D]. 长沙: 中南大学, 2014.

[17] 郭理清, 陈锋. 干线动态协调控制的深度Q网络方法[J]. *信息技术与网络安全*, 2020, 39(6): 1-6.

GUO H Q, CHEN F. A deep Q network method for dynamic arterial coordinated control [J]. *Cyber Security and Data Governance*, 2020, 39(6): 1-6.

(责任编辑: 杨耀辉)

(上接第66页)

[4] 吴凯, 杨新志. 浅述人工智能的深度学习[J]. *石河子科技*, 2020 (6) : 44-45

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. *Communications of the ACM*, 2017, 60(6): .84-90.

[6] 周佳禾, 宫元九. 利用卷积神经网络检测锂电池电极缺陷[J]. *电子测量技术*, 2019, 42(19) : 182-186.

ZHOU J H, GONG Y J, Detection of electrode defects in lithium batteries based on convolutional neural networks[J]. *Electronic Measurement Technology*, 2019, 42(19) : 182-186.

[7] 周佳禾, 宫临凡, 王盛瑶. 方形卷绕式锂电池电极褶皱检测方法研究[J]. *仪表技术与传感器*, 2021(5): 107-111.

ZHOU J H, GONG L F, WANG S Y. Research on detection of electrode fold of spirally wound flattened lithium batteries [J]. *Instrument Technique and Sensor*, 2021(5): 107-111.

[8] 黄梦涛, 连一鑫. 基于改进 Canny 算子的锂电池极片表面缺陷检测[J]. *仪器仪表学报*, 2021, 42(10): 199-209.

HUANG M T, LIAN Y X. Lithium battery electrode plate surface defect detection based on improved Canny operator [J]. *Chinese Journal of Scientific Instrument*, 2021, 42(10): 199-209.

[9] 张旭. 基于深度卷积神经网络的铝材表面缺陷检测[D]. 上海: 华东理工大学, 2020.

[10] 章云, 王晓东. 基于受限样本的深度学习综述与思考[J]. *广东工业大学学报*, 2022, 39(5): 1-8.

ZHANG Y, WANG X D. A review and thinking of deep learning with a restricted number of samples [J]. *Journal of Guangdong University of Technology*, 2022, 39(5): 1-8.

[11] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Fleet D. European Conference on Computer Vision. Cham: Springer, 2014: 818-833.

[12] SANDLER M, HOWARD A, ZHU M, *et al.* Mobilenetv2: inverted residuals and linear bottlenecks[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510-4520.

[13] HE K, ZHANG X, REN S, *et al.* Deep residual learning for image recognition[C]// The IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[14] 黄剑航, 王振友. 基于特征融合的深度学习目标检测算法研究[J]. *广东工业大学学报*, 2021, 38(4): 52-58.

HUANG J H, WANG Z Y. A research on deep learning object detection algorithm based on feature fusion. [J]. *Journal of Guangdong University of Technology*, 2021, 38(4): 52-58.

[15] DENG J, DONG W, SOCHER R, *et al.* Imagenet: a large-scale hierarchical image database[C]//2009 IEEE Conference On Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009: 248-255.

(责任编辑: 杨耀辉)