基于 D3QN 的认知物联网动态频谱接入

于 越[1] 陈玲玲[1] 刘文刚[1] 冯 琦[1]
YU Yue CHEN Lingling LIU Wengang FENG Qi

摘 要 认知无线电（cognitive radio，CR）融入物联网有利于减少大规模物联网部署的频谱稀缺性，而大规模物联网部署的核心技术是设计有效分配频谱的频谱接入算法。然而，随着认知物联网（cognitive-internet of things，C-IoT）网络的部分可观测信道和用户数量的增加，次用户以避免干扰和快速获取频谱状态信息。文章提出了一种基于深度强化学习（deep reinforcement learning，DRL）的动态频谱接入（dynamic spectrum access，DSA）算法，该算法改进网络结构为双决斗深度 Q 网络（dueling double deep Q network，D3QN），适用于具有多个用户和信道的 C-IoT 网络。通过与 Q-learning 和 DQN 算法对比，仿真结果表明，该算法能够快速准确地进行 DSA 决策，显著提高网络信道容量。

关键词 认知物联网；动态频谱接入；深度强化学习；C-IoT 网络；D3QN

doi：10.3969/j.issn.1672-9528.2024.12.015

0 引言

如今物联网在智能交通、智能制造、智慧城市等领域得到了广泛应用 [1]。但物联网发展的一个主要技术限制是因大量物联网用户和设备连接到网络而导致的频谱短缺，这可能对其发展构成一大挑战。频谱资源在实际网络中是动态的、稀缺的、极其有限的 [2]。由于大量无线电频谱未得到充分利用，而另一部分频谱资源又过于拥挤，导致可用频谱受到资源和业务分配不均的情况。因此，有效地共享现有频谱已成为满足频谱需求和提高频谱利用率的主要研究方向。通过认知无线电（cognitive radio，CR）接入未充分利用的授权频谱，可以有效缓解频谱稀缺问题，增加频谱接入机会，最大限度地利用未充分利用的授权频谱，提高网络容量，从而支持物联网网络的海量设备的连接 [3]。

认知物联网（cognitive internet of things，C-IoT）网络中的授权频谱利用率不足，会导致频谱资源严重浪费。而对于日益增长的频谱需求，一个可行的解决方案是设计灵活高效的频谱接入。由于授权频谱占用的动态变化，必须要访问这些通信信道 [4]，在这种情况下，需要 C-IoT 设备通过频谱感知来检测信道占用情况，并利用其结果来实现最优频谱接入策略。在捕获空闲频谱后，需要设计既能防止对主用户（primary user，PU）的干扰，又能管理对次用户（secondary user，SU) 干扰的频谱接入方法，有效地将可用频谱分配给 SU[5]。频谱接入主要包括静态频谱分配和动态频谱接入（dynamic spectrum access，DSA）。在静态频谱分配中，通信系统仅在集中资源管理器为频率预先分配的频谱上运行 [6]。但由于无线频谱资源的限制，会导致频谱利用效率低下。所在 DSA 技术中，一旦 PU 需要频谱， SU 应该迅速腾出信道 [7]。因此，DSA 技术的关键问题是在不干扰 PU 通信的情况下，确保 SU 能够动态访问频谱空洞，这对频谱访问决策的速度提出了更严格的要求。

1 相关工作

认知物联网通过感知环境、学习和适应来实现智能化的物联网系统 [8]。在认知物联网中，设备可以根据当前环境的情况自主做出决策，达到最大化网络性能和资源利用效率。动态频谱接入作为认知物联网中的一个重要问题，涉及到如何智能地选择和管理频谱资源，以满足不同设备和应用的需求 [9]。由于频谱资源的稀缺性和动态性，传统静态频谱分配方法无法满足物联网系统对频谱资源高效利用的需求，因此需要设计新的动态频谱接入算法来解决这一问题。

深度强化学习（deep reinforcement learning，DRL）是一种结合了深度学习和强化学习的方法，在许多领域取得了显著成果。强化学习的特点是学习网络可以与变化的、不确定的环境进行交互以获得知识，这在处理动态系统方面提供了卓越的性能。利用 Q 学习作为一种新型频谱管理框架，使 DSA 用户能够独立、智能地管理频谱资源。为了提高效率，本文使用神经网络来实现 Q 学习过程，创建了深度 Q 网络（deep Q network，DQN）解决频谱管理问题 [10]。Oshri 等人 [11] 在文献中，探讨了多信道无线网络中的动态频谱接入问题，在这种网络中，用户选择信道并以一定的概率传输数据包，目的是在用户之间无需协调或消息交换的情况

1. 吉林化工学院信息与控制工程学院 吉林吉林 132022

下最大限度地提高网络利用率。由于状态空间大以及状态的部分可观测性，因此该问题在计算上非常昂贵。为了解决这个问题， Li 等人[12] 在文献中提出了一种基于 DRL 的 DSA 方案，结合多种访问方法，以最大限度地提高系统吞吐量。SU 在 DSA 网络中采用的访问策略直接影响系统的性能。因此，引入了 DRL 来帮助 SU 在动态环境中学习最佳访问策略。

针对物联网设备增长带来的部分可观测信道和密集的计算负担，本文提出了一种高效的 DSA 算法，在抑制干扰的同时分配空闲频谱，从而改善 C-IoT 网络的频谱管理。D3QN 算法（dueling double deep q-network）是一种改进的 Q 学习算法，通过使用深度神经网络来近似 Q 值函数，并结合 Dueling DQN 和 DDQN 来提高算法的稳定性并加快收敛速度。通过 D3QN 算法来学习设备在不同状态下选择合适的频谱资源的策略，并根据奖励函数来调整策略以最大化系统的性能。

2 系统模型

本文考虑一个多用户多信道的且信道部分可观测的认知物联网场景，其中包含 M 个主用户和 N 个次用户，为每个主用户分配一个单独的信道。如图 1 所示，当 PU 和 SU 或多个 SU 同时使用同一无线信道时，可能会产生正常的通信链路和干扰链路。用 (X_m, Y_m) 和 (X_n, Y_n) 分别表示主次用户通信时两端发射机和接收机的位置信息，其中 $m \in \{1, 2, \cdots, M\}$，$n \in \{1, 2, \cdots, N\}$，并用 $d = \sqrt{(X_m - X_n)^2 + (Y_m - Y_n)^2}$ 计算通信距离。

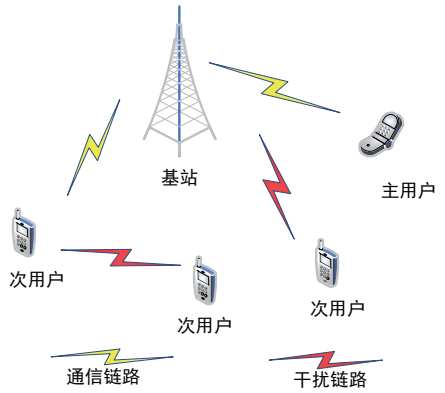

图 1 系统模型

在传播距离为 d 的情况下，采用 WINNER II 信道模型来计算通信链路的路径损耗：

$$\mathrm{PL}(d, f_c) = \mathrm{PL}_\alpha + \mathrm{PL}_\beta \log_{10}(d[\mathrm{m}]) + \mathrm{PL}_\gamma \log_{10}\left(\frac{f_c[\mathrm{GHz}]}{5}\right) \quad (1)$$

式中：f_c 是无线信道的载波频率；PL_α、PL_β 和 PL_γ 分别表示参考距离的路径损耗、路径损耗指数和路径损耗的频率相关性；$d[m]$ 表示在通信模型中通信双方的发射机和接收机的通信距离。

假设发射机与接收机之间存在视距路径，采用 Rician 信道模型推导，固信道增益 h 可表示为：

$$h = \sqrt{\frac{k}{k+1}}\sigma e^{j\theta} + \sqrt{\frac{k}{k+1}}\mathrm{CN}(0, \sigma^2) \quad (2)$$

式中：k 代表视距环境下接收信号功率和散射路径功率方差的比值；σ 由路径损耗决定；θ 表示接受信号的相位；CN(·) 为圆对称复高斯随机变量。因此，SU 接收信号的信干噪比（SINR）可表示为：

$$\mathrm{SINR} = \frac{|h_{nm}|^2 P_{nm}}{|h_{mm}|^2 P_{mm} + \sum_{n=1, n\neq i}^{N} |h_{nm}|^2 P_{nm} + \delta_o^2} \quad (3)$$

式中：$|h_{nm}|^2$、$|h_{mm}|^2$ 分别表示次用户 N 和主用户 M 在信道 m 上的增益；P_{nm}、P_{mm} 分别表示次用户 N 和主用户 M 在信道 m 上发射功率；δ_o^2 代表噪声功率。

3 基于 D3QN 的 DSA 算法

SU 必须获取授权信道状态并搜索空闲信道进行数据传输，这样可以提高频谱效率，保证 PU 的通信质量。由于用户在每个时间间隔内的环境观测是不够的，本文将 DSA 问题表述为一个部分可观察马尔可夫决策过程（partially observable markov decision process，POMDP）模型，其目标是根据之前观测信道的状态，智能体在每个时隙中选择一个动作，然后观察即时奖励和当前观察结果，以做出下一个决策。

具体而言，每个 SU 通过能量检测对所有 M 个通道独立进行频谱感知。这样，就确定了每个时隙开始时信道的占用情况，并采用硬决策规则进行决策。PU 的每个通道有空置（1）和占用（0）两种状态，SU 可以访问处于空闲状态的通道，但禁止访问处于占用状态的通道。

在每个时隙开始时，每个 SU 对所有 M 个信道进行频谱感知，检测信道状态。设时隙 t 处的传感结果为：$S(t)=[S_1(t), S_2(t), \cdots, S_n(t), \cdots, S_N(t)]$，其中 $S_n(t)=[s_1(t), s_2(t), \cdots, s_m(t)]$ 表示第 N 个次用户对 m 个信道的感知结果，且 $s(t)=\{0,1\}$，代表检测到不同的信道状态，其中 0 代表空闲，1 代表忙碌即 PU 正在占用信道。但在复杂的认知物联网环境下，感知到的信道状态信息不一定是准确的，$s(t)$ 可能包含错误，设感知错误的概率为 P_f。第一个 SU 唯一已知的信息 $s_1(t)$ 表示环境中的观察状态和 DQN 的输入。

在进行频谱感知后，每个 SU 根据感知结果决定最多访问一个信道或空闲。第 n 个 SU 的动作记为：$a_n(t) \in \{0, 1, \cdots, M\}$。其中，$a_n(t)=0$ 表示第 n 个 SU 在时隙 t 内不选择信道进行接入，$a_n(t)=m$ 且 $m\neq 0$ 表示第 n 个 SU 在时隙 t 内选择第 m 个信道进行接入。

在 SU 发射机进入信道后，信道状态根据其马尔可夫链发生变化。然后对应的 SU 接收机将接收到的信噪比反馈给发射机。作为结果，第 m 个通道上的第 n 个 SU 的奖励函数可以表示为：

$$r_n(t)=\begin{cases}-C & ，产生碰撞\\ \log_2(1+\mathrm{SINR}) & ，其它\end{cases} \tag{4}$$

式中：$C>0$，代表与 PU 或其他 SUs 碰撞时给到的惩罚。

动态频谱访问策略是分布式的，使得感知结果和访问决策信息不会在 SU 之间共享。且 SU 不知道信道状态的转移概率和感知误差的概率。只能通过访问后获得的接收到的 SINR 来学习如何访问频道，并以此为基础制定访问策略，以最大化自己的累积折扣奖励 R。

D3QN 是结合决斗深度 Q 网络和双深度 Q 网络，其网络结构如图 2 所示。

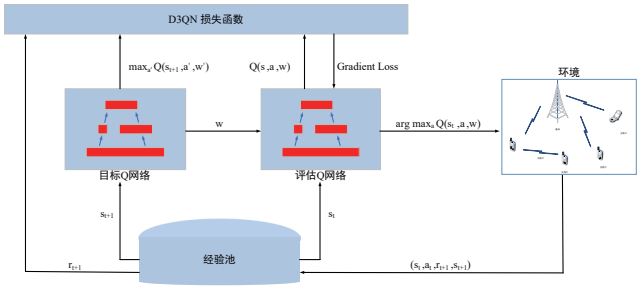

图 2 基于 D3QN 的 DSA 算法结构图

以下是基于 D3QN 的动态频谱接入算法的流程：

（1）初始化两个深度 Q 网络：目标 Q 网络和评估 Q 网络，它们具有相同的结构和初始权重。初始化经验回放缓冲区，用于存储状态转换。设置训练参数，如学习率、折扣因子、ε-greedy 策略参数等。

（2）训练过程：在每个时间步，每个用户（SU）基于当前状态选择动作：用户使用 ε-greedy 策略从网络中选择动作，并根据 ε-greedy 策略得到的奖励值大小决定。用户执行动作并观察到奖励以及下一个状态。将状态转换（当前状态、动作、奖励、下一个状态）存储到经验回放缓冲区中。从经验回放缓冲区中随机抽取一批数据作为训练样本并周期性地更新网络参数。

（3）动态频谱接入决策： 根据当前状态输入到 D3QN 网络中，获取它们的 Q 值。根据 ε-greedy 策略，选择 Q 值较大的动作作为频谱接入的决策。

（4）更新：当系统状态发生变化或者需要更新网络参数时，根据获取到的新数据对 D3QN 网络进行增量式更新，以适应动态环境的变化。

4 仿真分析

在本节中，给出了基于分布式 D3QN 的动态频谱接入的实验结果。首先随机选择一个 150 m×150 m 正方形中的 SU 和 PU 的位置。采用 WINNER II 模型和 Rician 模型分别计算路径损耗和推导的通道模型。对于 WINNER II 模型，设置 $f_c=5$ GHz、$\mathrm{PL}_\alpha=41$、$\mathrm{PL}_\beta=22.7$、$\mathrm{PL}_\gamma=20$。对于 Rician 模型，设置 $k=8$，σ^2 由 WINNER II 模型得到的路径损失决定。则单个 SU 接收机接收到的信噪比如式（3）所示，其中带宽 $B=1$ MHz，噪声谱密度 $\sigma_o^2=10^{-14.7}$ mW/Hz，单个 SU 的发射功率为 20 mW，单个 PU 的发射功率为 40 mW，设置与 PU 产生碰撞时的惩罚 $C=2$。

为证明所提算法的优越性，采用 Python 进行仿真设计，将 PU 的两种状态建立为动态的马尔科夫链，并采用 Q-learning 方法作为参考比较，本文考虑的认知物联网环境中有 6 个 SU，20 个 PU 和信道。Q-learning 算法已被广泛用于解决多用户动态频谱接入问题，能够适应环境的变化和不确定性，通过不断地探索和利用来更新策略，从而应对不同的环境和任务，但仍存在一些局限性。在多用户动态频谱接入问题中，状态空间往往非常庞大，而且可能是连续的、高维的，这导致了 Q-learning 算法在处理大规模状态空间时会面临维度灾难问题。通过与 Q-learning 和 DQN 算法对比

如图 3 所示的平均接入成功率比较，基于 Q-learning 算法在整个训练过程中变化不大，这是由于所设计的环境用户数过大使其在训练过程中无法有效学习。本文所考虑的 D3QN 算法在收敛速度和平均接入成功率方面显著优于其他两种算法。

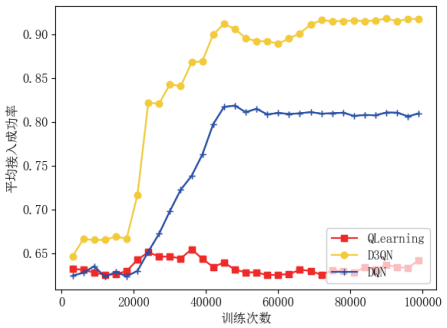

图 3 平均接入成功率

如图 4 所示，D3QN 算法在训练初期迅速获得较高的平均奖励值，并在训练后期趋于稳定，最终达到约 5.2。DQN 算法则在约 40 000 次训练后显著提升平均奖励值，最终稳定在约 4.4。而 Q-learning 算法的平均奖励值变化较小，始终保持在 3.0 左右，表现相对较差。总体而言，D3QN 算法得到的平均奖励值显著优于其他两种算法。

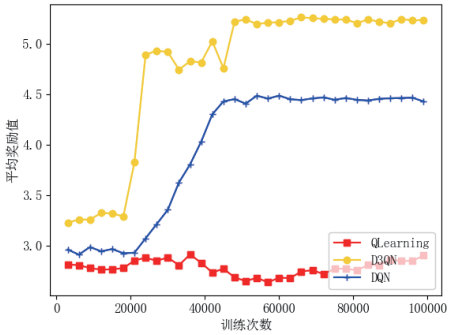

图 4 平均奖励值

如图 5 所示，Q-learning 算法随着训练次数的增加在逐渐降低与 PU 的碰撞概率，而 DQN 和 D3QN 算法的收敛速度更快，且 D3QN 最终的训练结果也明显由于其他两种算法，最后的平均碰撞概率稳定在 0.13 左右。

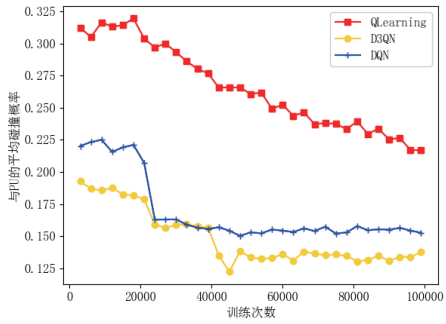

图 5 与 PU 的平均碰撞概率

如图 6 所示在与 SU 的平均碰撞概率方面，D3QN 和 DQN 算法表现出显著的优势，尤其是 D3QN 几乎完全避免了 SU 之间的碰撞，而 Q-learning 在这一指标上表现最差。

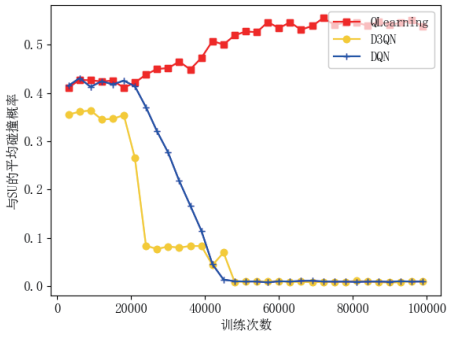

图 6 与 SU 的平均碰撞概率

5 总结

本文研究了不完全频谱感知和无集中控制器下的分布式 DSA 网络的频谱接入策略。提出了基于 D3QN 的 DSA 策略，结合 DDQN 和 Dueling DQN 实现 D3QN 来适配 DSA 策略。每个 SU 能够仅依靠它们自己的频谱感知结果以及学习结果来分布式地做出正确的频谱访问决策。通过仿真实验验证所提出的策略的性能，实验结果表明，与 Q-learning 和 DQN 相比，本文基于 D3QN 的 DSA 方法在多用户多信道下的收敛速度更快，性能更好。

参考文献：

[1] 席广亮，甄峰，钱欣彤，等 .2021 年智慧城市建设与研究热点回眸 [J]. 科技导报，2022,40(1):196-203.

[2] 王诗，朱笑莹，孙浩，等 . 认知物联网基于边缘计算的频谱感知 - 分配方法 [J]. 电子测量与仪器学报 ,2023,37(7):81-92.

[3] 原帅前，贾向东，尚通健，等 . 认知无线电系统多接入用户信息新鲜度研究 [J]. 计算机应用研究 ,2024,41(3):894-899.

[4] 李姣军，喻涛，周继华，等 . 动态不确定场景下认知工业物联网的资源分配策略 [J]. 浙江大学学报 (工学版), 2024, 58(5): 960-966.

[5] 殷晓虎，谢豪，李加美 . 基于 Markov 模型的认知车联网频谱感知方法 [J]. 无线电工程 ,2023,53(3):563-569.

[6]MOSLEH S, MA Y, REZAC J D, et al. Dynamic spectrum access with reinforcement learning for unlicensed access in 5G and beyond[C/OL]//2020 IEEE 91st Vehicular Technology Conference(VTC2020-Spring).Piscataway:IEEE, 2020[2024-06-19].https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=928979.

[7]BENEDETTO F, MASTROENI L, QUARESIMA G. Auction-based theory for dynamic spectrum access: a review[C]//2021 44th International Conference on Telecommunications and Signal Processing (TSP).Piscataway: IEEE, 2021: 146-151.

[8] 贺兴，艾芊，邱才明，等 . 泛在电力物联网数据挖掘体系建设综述及数据驱动认知框架探究 [J]. 低压电器，2019(19):1-14.

[9] 杨秋静 . 认知物联网中基于吞吐量最大化的中继节点的选择算法 [J]. 传感技术学报 ,2022,35(1):132-137.

[10] SONG H, LIU L J, JONATHAN A, et al. A deep reinforcement learning framework for spectrum management in dynamic spectrum access[J]. IEEE internet of things journal, 2021, 8(14): 11208-11218.

[11] OSHRI N, KOBI C. Deep multi-user reinforcement learning for distributed dynamic spectrum access[J]. IEEE transactions on wireless communications, 2019, 18(1): 310-323.

[12] LI Z Q, LIU X, NING Z L. Dynamic spectrum access based on deep reinforcement learning for multiple access in cognitive radio[J]. Physical communication, 2022, 54: 101845.

【作者简介】

于越（1998—），男，山西晋城人，硕士研究生，研究方向：动态频谱接入。

陈玲玲（1980—），女，吉林长春人，博士，教授，研究方向：认知无线电。

刘文刚（1997—），男，安徽阜阳人，硕士研究生，研究方向：动态频谱共享。

（收稿日期：2024-08-22）