DOI:10.19652/j.cnki.femt.2305683

异构网络中基于深度强化学习的用户关联与资源分配策略*

符平博[1] 陶 旭[1,2] 张 见[1,2] 李 晖[1,2]
(1.南京信息工程大学电子与信息工程学院 南京 210044;
2.无锡学院江苏省集成电路可靠性技术及检测系统工程研究中心 无锡 214105)

摘 要:由于异构网络非凸性和组合性的特点,联合用户关联和资源分配来实现能量效率(energy efficiency,EE)和频谱效率(spectral efficiency,SE)同时最大化的最优全局策略仍然是非常具有挑战性的。基于深度强化学习(deep reinforcement learning,DRL)的方法成为在保证异构网络下行链路用户设备(user equipments, UEs)服务质量(quality of service,QoS)的同时实现联合 EE-SE 性能最大化的必要解决方案。此外,为解决状态-动作空间下计算量大的问题,引入了多智能体架构的深度强化学习算法(MAD3QN)来获得近乎最优控制策略。仿真结果表明,MAD3QN 算法在系统容量方面比 DDQN 算法和 DQN 算法分别提高了 9.2%和 18.2%,在联合 EE-SE 性能方面分别提高了 8.5%和 16.6%,提升了系统的有效性。

关键词:深度强化学习;用户关联;资源分配;能量效率;频谱效率
中图分类号: TN915.81 **文献标识码**:A **国家标准学科分类代码**: 510.5030

Strategy of user association and resource allocation based on deep reinforcement learning in heterogeneous networks

Fu Pingbo[1] Tao Xu[1,2] Zhang Jian[1,2] Li Hui[1,2]
(1. College of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. Jiangsu Integrated Circuit Reliability Technology and Testing System Engineering Research Center, Wuxi University, Wuxi 214105, China)

Abstract: Due to the non convexity and combinatorial characteristics of heterogeneous networks, it is still very challenging to combine user association and resource allocation to achieve the optimal global strategy that maximizes both energy efficiency (EE) and spectral efficiency (SE) simultaneously. The method based on deep reinforcement learning (DRL) has become a necessary solution for maximizing joint EE-SE performance while ensuring the quality of service (QoS) of user equipments (UEs) in heterogeneous networks. In addition, to solve the problem of high computational complexity in the state action space, the double DQN algorithm with multi-agent dueling architecture (MAD3QN) was introduced to obtain almost optimal control strategies. The simulation results show that the MAD3QN algorithm has increased system capacity by 9.2% and 18.2% respectively compared to the DDQN algorithm and DQN algorithm, and improved joint EE-SE performance by 8.5% and 16.6% respectively, enhancing the effectiveness of the system.

Keywords: deep reinforcement learning; user association; resource allocation; energy efficiency; spectral efficiency

0 引 言

随着移动设备的快速增长和物联网(internet of things,IoT)的出现,下一代无线网络面临着应对无线应用激增的巨大挑战[1]。最有希望的解决方案是使用具有不同传输功率和覆盖范围的微蜂窝(picocells)和家庭蜂窝

收稿日期:2023-10-23
* **基金项目**:国家自然科学基金(61661018)、江苏省基础研究计划青年基金(BK20210064)、无锡市科技创新创业资金(WX03-02B0137-022200-34)项目资助

(femtocells)来强化现有的蜂窝网络。这些异构网络(heterogeneous networks, HetNets)可以将用户设备(user equipments, UEs)从宏基站(macro base stations, MBSs)卸载到传输功率和覆盖范围不同的微基站(picocell base stations, PBSs)和家庭基站(femtocell base stations, FBSs)上[2]。在HetNet中,终端用户可能与宏基站、微基站或家庭基站相关联,从而导致不同的用户体验。因此,用户关联是HetNets中的一个重要问题。此外,为了联合实现HetNets的高频谱能量效率,PBSs和FBSs可以与MBSs重用和共享相同的信道。

这种HetNets存在一些优化问题,如用户关联和资源分配[3-4]。现有的HetNets用户关联研究大多侧重于速率最大化[5],通过联合优化用户关联和各层频谱分配来实现覆盖率和速率效用最大化。文献[6]提出了联合小区关联和带宽分配方案,以最大化网络和日志率,以实现比例公平。文献[7]提出群体智能算法设计了一个带有BS开/关操作的用户关联机制,以最大化加权长期速率的总和。资源分配(resource allocation, RA)中可以通过优化频谱和功率分配来减少功耗、减轻干扰,从而提高网络的能量利用效率和频谱资源利用效率。文献[8]建立了整个异构蜂窝网络中的能量效率表达式,为提高网络的能量效率提供了理论指导。文献[9]研究节能功率分配和无线回程带宽分配的设计是为了在一定服务质量(quality of service, QoS)约束下使下行HetNets的系统EE最大化,从而提出了一种近似最优的迭代资源分配算法,以实现最大的能量效率(energy efficiency, EE)。文献[10]对密集异构网络中的资源进行分布式分配以减轻干扰从而提高整个网络的频谱效率。文献[11]在设备到设备辅助蜂窝网络场景下提出了通过资源块的联合和功率控制的分配算法来实现最大化频谱效率。文献[12]研究了在异构CRAN中最大化EE性能的无线电资源管理方法。文献[13-15]在EE和频谱效率(spectral efficiency, SE)之间进行了权衡,并满足了用户对服务质量QoS的需求。以上的文献考虑的是固定的环境状态来求解其效用函数。但是,在大而复杂的状态和动作下,特别是当网络性能随着环境的变化而变化时,这种方法很难得到最优解。文献[16]在考虑到网络高度复杂情况下,提出了基于深度Q网络(deep Q-network, DQN)结合分布式协调学习的算法来有效学习优化后的智能资源管理策略。文献[17]将能效优化问题设计为一个多级决策问题,并采用多智能体Actor-Critic(MAAC)算法对问题进行求解,从而最大限度提高异构网络的能量利用率。因此,本文主要研究强化学习(Reinforcement Learning, RL)方法来解决这一具有挑战性的问题。

RL的智能体不断与未知环境交互,学习最优控制策略。Q-learning是一种广泛用于估计状态-动作值函数的强化学习算法[18]。在Q-learning中,状态-动作值是以表格形式计算的,对于大的状态-动作空间函数,很难找到最优值。为了避免不可行性问题,引入了DQN,将RL与深度神经网络相结合成深度强化学习。DQN通过智能体来感知环境状态,并在与环境交互过程中得到最大收益目标[19]。最近,深度强化学习(deep reinforcement learning, DRL)在优化无线网络中的RA问题上表现出了巨大的进步,如动态信道接入[20]、移动卸载[21]、RA管理[22]等。然而,开发一种基于DRL的方法来解决联合优化问题的研究还很少。

本文提出的用户关联和资源分配算法为解决HetNets下行链路中的EE-SE性能最大化问题提供了一种有效的方法。提出了一种基于多智能体的DRL算法联合UARA(user association & resource allocation)优化方法,在保证终端用户的QoS要求的同时,获得最优的长期网络效用,同时通过联合终端用户和基站,为终端用户分配信道以最终得到最优解。提出了EE-SE在下行HetNets中的联合优化性能,通过无单位加权系数将复杂的EE-SE联合优化问题求解为单个目标函数。考虑到联合优化问题的非凸性和组合性特点,提出了多智能体DRL方法,为UEs定义了状态、动作和奖励函数,并最终提出了MAD3QN(multi-agent dueling double DQN)算法,通过不断与环境交互进行学习与优化来实现近乎最优的策略。通过仿真表明,本文提出的算法在保证终端用户QoS要求的前提下比其他算法具有更快的收敛速度,更大的系统容量和更好的EE-SE联合优化性能。

1 系统模型及问题表述

1.1 系统模型

本文模拟了异构网络场景,如图1所示,由D_m个MBSs、D_p个PBSs、D_f个FBSs和N个随机的终端用户UEs组成。通常部署PBS是为了从具有中等传输功率的宏小区分流业务。FBS通常用于具有更好的QoS和更高的数据速率的小覆盖区域。用$BS=\{mbs_1,\cdots,mbs_{D_m},pbs_1,\cdots,pbs_{D_p},fbs_1,\cdots,fbs_{D_f}\}$来表示所有BSs的集合,BSs的索引集合为$\boldsymbol{B}=\{0,1,2,\cdots,F-1\}$,HetNets中基站总数量$F=D_m+D_p+D_f$。同时假设BSs在$W$

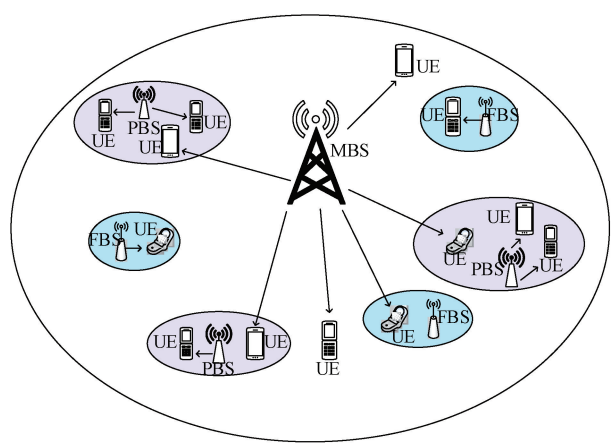

图1 系统模型

Fig. 1 System model

个共享的正交信道上工作。

考虑到不同的 BSs 可能服务于不同的 UEs，第 m 个用户关联向量被定义为 $b_m^l(t)=(b_m^0(t),b_m^1(t),\cdots,b_m^{F-1}(t))$，$l\in\boldsymbol{B}$，$m\in\boldsymbol{N}$，其中 $\boldsymbol{N}=\{1,2,\cdots,N\}$，$F$ 为基站总数。当第 m 个用户选择与第 i 个 BS 关联时，$b_m^l(t)=1$，$l=i$ 否则 $b_m^l(\mathrm{t})=0$，$l\neq i$，其中 $i\in\boldsymbol{B}$。假设每个 UE 每次最多只能选择一个 BS，所以有：

$$\sum_{l=0}^{F-1}b_m^l(t)\leqslant 1,\quad \forall m\in\boldsymbol{N} \tag{1}$$

此外，当 UE 与 BS 关联后，每个 BS 的信道资源分配给 UE。对于第 m 个 UE，信道分配向量定义为 $c_m^w(t)=(c_m^0(t),c_m^1(t),\cdots,c_m^W(t))$，$w\in\omega$，$m\in\boldsymbol{N}$，其中 $\omega\in\{1,2,\cdots,W\}$，$W$ 为共享的正交信道总数。如果第 m 个用户在时间 t 上选择信道 k，$c_m^w(t)=1$，$w=k$ 否则 $c_m^p(t)=0$，$w\neq k$。虽然在同一信道上可以同时运行的 UEs 是无限个的，但为了讨论的简单性，本文假设每个终端用户最多只能选择一个信道，所以有：

$$\sum_{w=1}^{W}c_m^w(t)\leqslant 1,\quad \forall m\in\boldsymbol{N} \tag{2}$$

参考文献[23]的系统模型，则 MBSs 和 PBSs 的路径损耗为：

$$Loss(d_{m,l})=35+42\lg d_{m,l}\ \mathrm{dB} \tag{3}$$

FBSs 的路径损耗为：

$$Loss(d_{m,l})=38+32\lg d_{m,l}\ \mathrm{dB} \tag{4}$$

式中：$d_{m,l}$ 代表着用户 UE 到基站 BS 的距离。遵循文献[23]系统信道增益模型定义：

$$h_l^{m,w}(t)=10^{-Loss(d_{m,l})/20}\sqrt{\varphi_l^{m,w}s_l^{m,w}} \tag{5}$$

式中：$\varphi_l^{m,w}$ 为用户与基站之间的天线增益；$s_l^{m,w}$ 为用户与基站之间的阴影系数。

由于 PBSs 和 FBSs 位于 MBSs 的覆盖区域内，因此需要考虑共享信道干扰。定义 $p_{l,m}^w(t)=(p_{l,m}^1(t),p_{l,m}^2(t),\cdots,p_{l,m}^w(t))$，$w\in\omega$，$m\in\boldsymbol{N}$，$l\in\boldsymbol{B}$ 为 t 时刻第 m 个 UE 与其相关联的 BS_l 之间的信道 C_w 上使用的发射功率矢量。然后，第 m 个 UE 与其关联的 BS_l 在信道 C_w 上的信干噪比(signal to interference plus noise ratio，SINR)为：

$$\gamma_{l,m}^w(t)=\frac{b_m^l(t)c_m^w(t)h_l^{m,w}(t)p_{l,m}^w(t)}{\sum\limits_{g\in B,g\neq l}b_m^g(t)c_m^k(t)h_g^{m,w}(t)p_{g,m}^w(t)+W_0N_0} \tag{6}$$

式中：$h_l^{m,w}(t)$ 为 t 时刻在信道 C_w 上工作的 BS_l 与第 m 个终端用户之间的信道增益；W_0 为信道带宽；N_0 为噪声功率谱密度。则信道 C_w 上工作的 BS_l 与第 m 个终端用户之间的下行容量为：

$$r_{l,m}^w(t)=W_0\log_2(1+\gamma_{l,m}^w(t)) \tag{7}$$

则第 m 个用户终端的总传输容量为：

$$r_m(t)=\sum_{l=0}^{F-1}\sum_{w=1}^{W}r_{l,m}^w(t)=\sum_{l=0}^{F-1}\sum_{w=1}^{W}W_0\log_2(1+\gamma_{l,m}^w(t)) \tag{8}$$

1.2 问题描述

本文在保证下行链路 HetNets 终端用户最小服务质量 QoS 需求的同时，从其所选的基站获得最大传输容量和实现 EE-SE 性能最大化的目标。因此，第 m 个 UE 的 SINR $\gamma_m(t)$ 应不低于最低 QoS 要求 Ω_m，有：

$$\gamma_m(t)=\sum_{l=0}^{F-1}\sum_{w=1}^{W}\gamma_{l,m}^w(t)\geqslant\Omega_m \tag{9}$$

而且，考虑到 BS_l 的发射功率为 $p_{l,m}^w(t)$，那么第 m 个 UE 相关的总传输成本为：

$$\varphi_m(t)=\sum_{l=0}^{F-1}\lambda_l b_m^l(t)\sum_{w=1}^{W}c_m^w(t)p_{l,m}^w(t) \tag{10}$$

其中 λ_l 为 BS_l 发射功率的单价成本，则总功耗为：

$$P_{total}=\varphi_m(t)+\Psi_m \tag{11}$$

其中 Ψ_m 表示动作选择成本，$\Psi_m>0$，因为考虑到动作的选择可能会消耗一定的成本。

首先定义 EE 和 SE，然后再解决 EE-SE 的联合优化问题，其中 EE 表示为：

$$EE(t)=r_m(t)/P_{total} \tag{12}$$

SE 表示为：

$$SE(t)=r_m(t)/W_0 \tag{13}$$

从式(12)和(13)可以看出，最大化 EE 和最大化 SE 通常是相互矛盾的。最大化 SE 意味着增加总功耗 P_{total}，这可能导致 EE 的降低。同样，最大化 EE 也可能导致 SE 的减少。因此，有必要研究 EE 和 SE 之间的权衡作为联合优化。由于两者的量纲不同，将 $W_0/p_{l,m}^w$ 与 SE 相乘以确保加权求和时 EE 和 SE 单位一致。最后，为了调整客观度量，使用无量纲参数 $\beta\in(0,1]$，则 EE-SE 联合优化问题表示为：

$$\max\sum_{t=1}^{T}K(t)=\sum_{t=1}^{T}\left[(1-\beta)EE(t)+\beta\frac{W_0}{P_{l,m}^w}SE(t)\right] \tag{14}$$

约束条件为：

$$\sum_{l=0}^{F-1}\sum_{w=1}^{W}\gamma_{l,m}^w(t)\geqslant\Omega_m \tag{15}$$

$$W_0\log_2(1+\gamma_{l,m}^w(t))\geqslant r_m^{\min}(t),\forall m\in\boldsymbol{N},\forall t\in\boldsymbol{T} \tag{16}$$

式中：$r_m^{\min}(t)$ 为每个 UE 的最低用户数据速率。式(15)为保证满足终端用户 UE 的最小服务质量 QoS 需求，式(16)为 UE 的数据速率必须大于每个用户的最小数据速率。

2 算法设计

各用户通过选择基站和信道来最大化 EE-SE 的性能并保证用户的 QoS 需求。RA 问题由于其非凸性和组合性而成为 NP-Hard 问题。因此，用传统的方法很难得到最优解。所以，本文将动态 EE-SE 的 RA 问题表示为马尔可夫决策过程(Markov decision process，MDP)。MDP 的目标是找到一个最优策略 π^* 以最大化未来的回报。本文采用 DRL 来解决联合优化问题。

2.1 强化学习

在强化学习中，智能体与未知环境持续交互，观察当前状态并采取行动。在与环境互动后，反馈以奖励的形式产生，则得到一个 MDP 的四元组($\boldsymbol{S}$,$\boldsymbol{A}$,$\boldsymbol{R}$,$\boldsymbol{P}$)，其中 $\boldsymbol{S}$ 是一组可能的状态，$\boldsymbol{A}$ 是 UEs 行为的集合，$\boldsymbol{R}$ 是 UEs 的奖励函数，$\boldsymbol{P}$ 代表着采取动作 a 时从状态 s 到 s' 的状态转移概率。

动态规划(dynamic planning，DP)和 Q-learning 是解决 MDP 问题的有效方法。然而，在 DP 中，需要一个完善的模型来解决 MDP 问题，这反映了较高的计算成本。因此，考虑用 Q-learning 方法来解决 MDP 问题。MDP 问题通过定义状态空间、行动空间和奖励函数来表述如下。

1) 状态空间

考虑状态空间中的两个组成部分信道增益 $g_l^{m,w}(t)$ 和 $z_n(t)$，即：

$$s_t=[g_1^{1,1}(t),\cdots,g_F^{N,W}(t),z_1(t),\cdots,z_N(t)] \tag{17}$$

式中：$z_n(t)$ 表示每个用户在时间 t 上是否满足用户 QoS 的需求，$z_n(t)\in\{0,1\}$。$z_n(t)=1$ 代表第 n 个终端用户满足最低 QoS 要求，即 $\gamma_{l,m}^w(t)\geqslant\Omega_n$；否则 $z_n(t)=0$，$\gamma_{l,m}^w(t)<\Omega_n$。

2) 动作空间

在 t 时刻上，所有 UEs 需要选择一个基站和信道，所以每个 UE 的动作空间 $\boldsymbol{A}_m$ 可定义为：

$$a_{l,m}^w(t)=\{b_m^l(t),c_m^w(t)\} \tag{18}$$

式中：$b_m^l(t)\in\{0,1\}$，$b_m^l(t)\in\{b_m^0(t),\cdots,b_m^{F-1}(t)\}$；$c_m^w(t)\in\{0,1\}$，$c_m^w(t)\in\{c_m^1(t),\cdots,c_m^W(t)\}$。

因为有 F 个 BSs 和 W 个信道，所以每个终端用户的可能动作数为 $F\times W$ 个，且随着 F 和 W 的增加，动作数可能会越来越大，同时影响状态的变化。在任意 t 时刻，另一些 $N-1$ 个 UEs 的动作向量会被定义为 $\boldsymbol{A}_{-m}=\{a_1(t),\cdots,a_{m-1}(t),a_{m+1}(t),\cdots,a_N(t)\}=\{b_m^l(t),c_m^w(t)\}$。

3) 奖励函数

奖励是智能体根据状态进行下一步动作的函数，用于评估动作的好坏[24]。本文在考虑到其他用户行为 $a_{-m}(t)$ 的同时，第 m 个用户采取动作 $a_{l,m}^w(t)$ 后可以获得即时奖励 $\boldsymbol{R}_m$ 为：

$$\boldsymbol{R}_m(t)=\boldsymbol{R}_m(s,a_m^*,a_{-m}^*)=$$
$$\sum_{t=1}^{T}\left[(1-\beta)EE(t)+\beta\frac{W_0}{P_{l,m}^w}SE(t)\right],\quad\forall a_m\in\boldsymbol{A}_m \tag{19}$$

每个 UE 用户都已经采用了他们的最佳响应策略，没有任何一个用户可以通过单方面的改变来获得更好的结果。也就是说在稳定状态时，所有用户的策略都相互适应，没有人可以通过单独的行动来获得奖励。如果每个 UE 都获得了关于奖励函数和状态转移的信息，则可以用整数规划方法找到最佳策略从环境中获得累计奖励。而本文这些信息对于 UE 来说都是未知的，所以为了克服这一挑战采用强化学习获得最佳策略 π^*。

2.2 联合优化问题的多智能体 DRL

在 Q-learning 中，agent 根据当前状态获得最优策略动作 $a_m=\pi_m(s)\in\boldsymbol{A}_m$。每个 UE 迭代地将自己的最优策略和状态信息发送到只有 1 位(0 或 1)的关联 BS，通过回程通信链路在不同基站之间传递消息，以获取所有终端用户的全局状态信息和协同决策策略。Q 值函数的期望收益可以表示为：

$$Q_m(s_t,a_m)=Q_m(s_t,a_m)+\gamma[A(t)+\mu\max_{a'_m\in\boldsymbol{A}_m}Q_m(s'_t,a'_m)-Q_m(s_t,a_m)] \tag{20}$$

式中：$A(t)=\left[(1-\beta)EE(t)+\beta\frac{W_0}{P_{l,m}^w}SE(t)\right]$；$\mu$ 为折扣因子；γ 为 $Q_m(s_t,a_m)$ 更新的学习率，通过适当的设置 γ。Q-learning 在更新 $Q_m(s_t,a_m)$ 时可以趋于收敛。

对于 Q-learning 将状态－动作值以表格的形式存储，非常适合低维状态－动作空间值。然而，在本文提出的多层异构网络中，由于状态空间和动作空间较大，使用 Q-learning 算法收敛速度较慢，为高维度状态－动作值找到最优策略总是很有挑战性的[25]。为了解决低维问题，采用神经网络(neural network，NN)逼近 Q 值函数 DQN。在 DQN 中，使用神经网络来表示状态－动作空间值。此外，神经网络将 Q 值函数近似为：

$$Q_m(s_t,a_m;\theta)=Q_m(s_t,a_m)$$

式中：θ 表示在线网络权值。DQN 还利用目标网络来稳定整个网络的性能。然后对神经网络进行更新，使目标网络与当前 Q 值函数之间的损失函数最小为：

$$L_m(\theta)=E[y^t-Q_m(s_t,a_m;\theta)]^2 \tag{21}$$

其中，y^{DQN} 为目标函数：

$$y^{DQN}=[A(t)+\mu\max_{a'_m\in\boldsymbol{A}_m}Q_m(s'_t,a'_m;\theta^-)] \tag{22}$$

式中：θ^- 表示目标函数的权值，该权值是从前面迭代的在线网络权值 θ 复制而来的，这降低了训练样本的相关性。但是，从同一个正在更新的网络中计算目标值，可能会导致学习不稳定。因此，在训练中每步过后，网络转换经验 $\{s_t,a_m,A(t),s'_t\}$ 被存储在经验回放池 $\boldsymbol{D}$ 中，RL 代理将从可用数据集中随机抽取批次样本来训练神经网络。DQN 和 Q-learning 都使用相同的动作值来选择评估 Q 值函数，这有时会导致动作值 Q 函数高估了其真实值，这会使算法做出次优决策。为了解决此问题，本文提出 DDQN(Double DQN)策略，策略将目标函数的动作选择和评估分离为：

$$y^{DDQN}=A(t)+\mu Q_m(s'_t,\text{argmax}Q_m(s'_t,a'_m;\theta);\theta^-) \tag{23}$$

DDQN 策略的强化学习过程如图 2 所示。具体来说，目标网络通过状态 s'_t 和在线网络选择出的动作来计算 $Q_m(s'_t,a'_m;\theta)$，然后利用折扣因子 μ 和奖励 $A(t)$，得到目标值 y^{DDQN}，最后通过目标值减去在线网络预测的当前值 $Q_m(s_t,a_m;\theta)$ 来计算误差，然后反向传播来更新权重。

由式(23)可知，在线网络是用来选择动作的，而目标

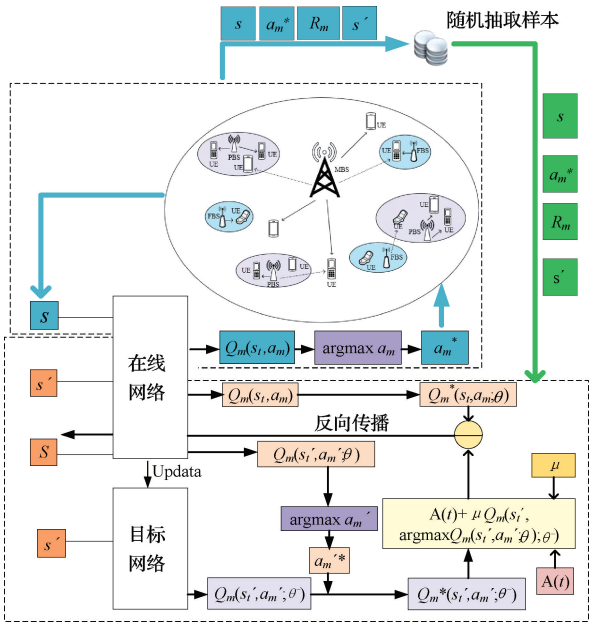

图 2　基于 DDQN 策略的强化学习

Fig. 2　Reinforcement learning based on DDQN strategy

网络是用来逼近动作值的。仅考虑动作选择的近似动作值并不能解释当前状态值。所以，有必要确定在特定状态下行动的优势。为了保证在某一特定状态下行动选择的优势，进一步提出 Dueling DQN 策略，将优势函数 $A(s_t,a_m)$ 分解为状态和状态行为价值函数[26]，以衡量更有价值的行动，可以表示为：

$$A(s_t,a_m)=Q_m(s_t,a_m)-V(s_t) \tag{24}$$

式中：$V(s_t)$ 为动作的优势值。

利用 Dueling 框架，在 DDQN 上构建了 D3QN(Dueling Double DQN)。然后将 $Q_m(s_t,a_m)$ 分解为两个部分，第 1 个部分计算动作的优势，第 2 部分计算状态值。所以估计动作值函数表示为：

$$Q(s_t,a_m)=V(s_t)+\left(A(s_t,a_m)-\frac{1}{|A|}\sum_{a_m}A(s_t,a_m)\right) \tag{25}$$

式中：$|A|$ 决定了整个动作空间的优势。

将 D3QN 扩展为多智能体 DRL。本文提出了多智能体 D3QN 的方法，算法 1 为 MAD3QN 的算法流程。

算法 1：MAD3QN 算法

输入：所有 Ues 允许采取的操作

输出：实现所有终端用户 QoS 要求的资源分配。

1. 初始化经验回放内存。

2. 用网络 θ 和网络 θ^- 初始化在线网络 $Q_m(s_t,a_m;\theta)$ 和 $Q_m(s'_t,a'_m;\theta^-)$。

3. For $episode=1,2,3,\cdots,E$ do

4. 观察状态空间(17)

5. For $step=1,2,3,\cdots,T$ do

6. For $Agent=1,2,3,\cdots,M$ do

7. 根据 $\varepsilon-greedy$ 策略从 $Q_m(s_t,a_m;\theta)$ 中选择动作 a_m

8. 终端用户向基站发送请求以访问所选信道，如果基站收到请求并向终端发送反馈信号，则表示有可用的信道，则用户终端获得即时奖励 $\boldsymbol{R}_m(s_t,a_m)$。

否则，基站将不予回复，用户终端将获得负奖励。

9. Agent m 执行 a_m 后观察下一个状态 s'_t

10. Agent m 将 $\{s_t,a_m,A(t),s'_t\}$ 存储在 $\boldsymbol{D}$ 中

11. Agent m 按最小批次从 $\boldsymbol{D}$ 中随机取样

12. Agent m 根据(23)计算目标网络值 y^{DDQN}

13. 执行梯度下降以最小化损失函数

14. 经过每一步都更新 $\theta^-=\theta$

15. End For

16. End For

17. End For

3　仿真实验与数据分析

3.1　仿真试验设置

基于用户关联和联合 UARA 动态性能分析，本文提出 MAD3QN 算法通过使用 Python 平台的 Pytorch 框架实现。本文在仿真中，网络由 2 个 MBSs、10 个 PBSs、16 个 FBSs 和 50 个终端 UE 组成。MBSs、PBSs 和 FBSs 的半径分别是 500、100 和 30 m，此网络分布如图 3 所示，然后仿真实验中所提到的仿真参数定义如表 1 所示。MAD3QN 算法的超参数如表 2 所示，采用 ReLU(rectified linear unit)作为激活函数，然后采用 RMSProp (root mean square propagation)优化方法来获得最优随机梯度下降。除了超参数，深度学习由输入层、3 个隐藏层(64、32 和 32 个神经元)和输出层($F\times W$ 个神经元)组成，其中 F 和 W 分别为基站数和信道数。

表 1　仿真实验参数设置

Table 1　Simulation experiment parameter settings

参数名称	参数值
信道带宽 W_0/kHz	180
终端用户数量 N	50
信道数量 K	30
基站传输功率/dBm	{40,30,20}
基站半径/m	{500,100,30}
噪声功率密度 N_0/(dBm/Hz)	−174
动作选择成本 ψ_m	0.001
QoS 要求/dB	3
发射功率的单价成本 λ_l	0.000 5
天线增益 $\varphi_l^{m,w}$/dBi	9
阴影系数 $\varsigma_l^{m,w}$/dB	8

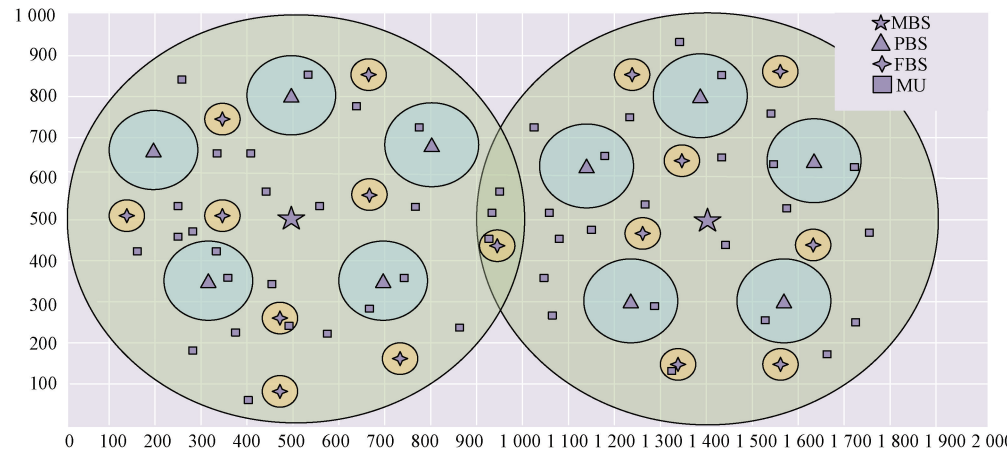

图 3 MBSs、PBSs、FBSs 和 UEs 的网络分布

Fig. 3 Network distribution diagram of MBSs, PBSs, FBSs, and UEs

表 2 MAD3QN 算法超参数

Table 2 MAD3QN algorithm hyperparameters

参数名称	参数值
训练次数	500
每轮训练最大步数	500
折扣率 μ	0.9
回放经验池大小	500
学习率 γ	0.05
优化器	RMSProp
激活函数	ReLU

3.2 仿真结果及分析

用各种优化策略对 MAD3QN 方法的性能进行了评价。RMSProp[27]、Adam[28]、AdaGrad[29] 3 种优化策略的训练平稳步数如图 4 所示。在学习过程的开始阶段，这 3 种情况下的训练步数都非常大。随着事件数的增加，收敛速度有加快的趋势。RMSProp 优化策略比其他两种策略收敛速度更快。因此，在本文 MAD3QN 方法中选择 RMSProp 优化策略。

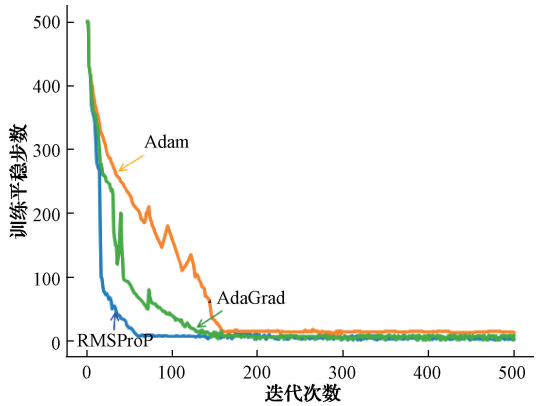

图 4 不同优化策略下的训练平稳步数

Fig. 4 Training stationary steps under different optimization strategies

不同优化算法下的训练平稳步数如图 5 所示。随着迭代次数的增加，除了 Greedy 算法无法收敛外，其余几种情况的收敛速度有加快的趋势。总体而言，几种 DRL 方案满足所有 UE 的 QoS 要求所需的训练步数较小。本文提出的 MAD3QN 优化算法收敛速度最快，在迭代次数还没达到 100 次时就收敛稳定下来，DDQN 算法次之、DQN 算法最差。

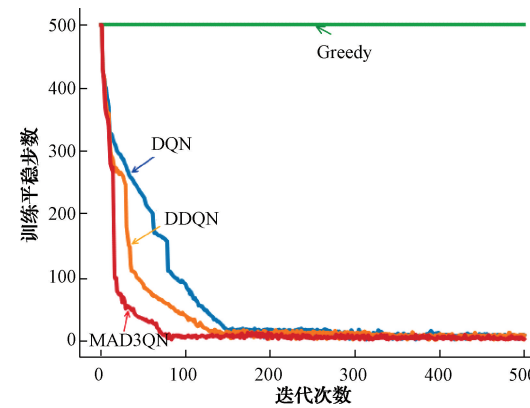

图 5 训练平稳步数随迭代次数变化

Fig. 5 The number of stable training steps varies with the number of iterations

使用不同优化算法对系统容量(即和速率)的影响如图 6 所示。在保证满足用户 QoS 需求时，本文所使用的 MAD3QN 优化方法的系统容量比 DQN 和 DDQN 高。

不同用户个数在 3 种方法的学习曲线如图 7 所示，随着终端用户数量的不断增加，几种 DRL 方法可以实现更高的系统容量，其中本文所使用的 MAD3QN 策略在用户数量不同时获得了最高的系统容量。

3 种情况不同 β 下 EE-SE 的性能如图 8 所示。从图 8 可以看出，经过 500 次的迭代，随着 β 数值的增大，EE-SE 的性能下降。当 $\beta=0$ 时 EE-SE 平均值为 11，但随着 β 的增大，$\beta=0.5$ 时 EE-SE 平均值为 8，当 $\beta=1$ 时，EE-SE 平

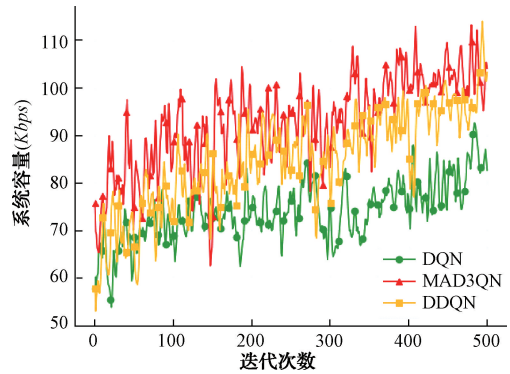

图 6　系统容量随迭代次数变化

Fig. 6　The system capacity varies with the number of iterations

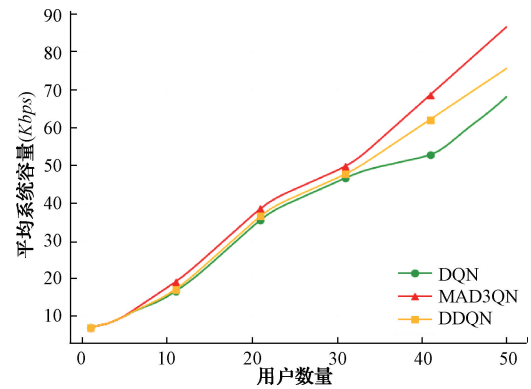

图 7　不同用户下的平均系统容量

Fig. 7　Average system capacity under different users

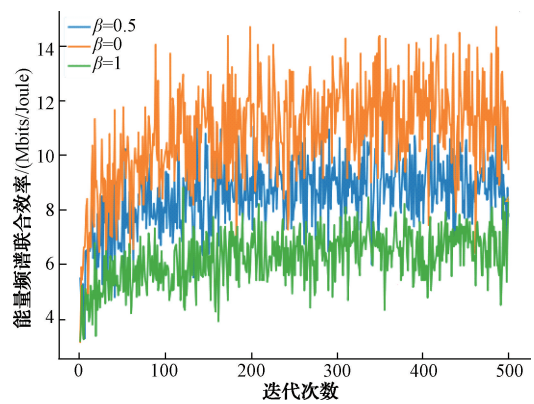

图 8　能量－频谱联合效率随迭代次数变化

Fig. 8　The joint efficiency of energy-spectrum varies with the number of iterations

均值下降到 6，EE-SE 的性能下降了大约 83.3%。

随着终端数量的不断增加，由于所有终端的 QoS 要求都得到了保证，则可以用几种 DRL 方法来实现更高的 EE-SE 性能。如图 9 所示，EE-SE 的最优性能随着终端用户 UE 数量的增加而增加。与其他 DRL 方案相比，MAD3QN 在用户数量增加的情况下可以获得更好的性能。在 MAD3QN 中，Agent 可以利用 Dueling 框架来观察动作的当前状态效应，这使得其 Q 值比 DDQN 和 DQN 的更有效。

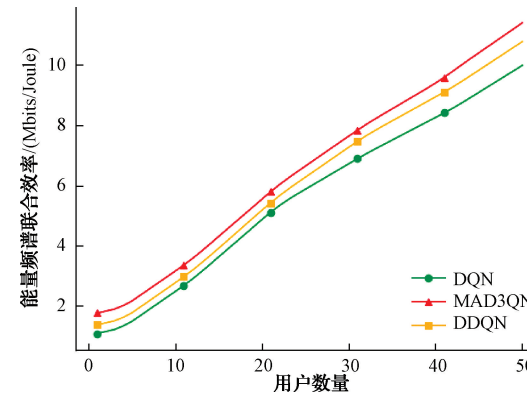

图 9　不同用户下的能量频谱联合效率

Fig. 9　Joint efficiency of energy spectrum under different users

4　结　论

本文提出了多智能体 DRL 方法，以获得 HetNets 的联合最优 UARA 策略。优化问题的制定是为了在保证终端的 QoS 要求的同时最大限度的提高 EE-SE 的性能。考虑到该联合优化问题的非凸性和组合性特点，本文提出一种联合关联 UEs 和 BSs 并为 UEs 分配信道的多智能体深度强化学习方法。此外，通过使用 Dueling 框架在 DDQN 上构建 D3QN。最后，仿真结果表明，MAD3QN 方法比其他 DRL 方法具有更快的收敛速度以及在满足用户 QoS 需要的前提下获得最优的 EE-SE 性能。

本文没有考虑到所需的计算资源，后面可以对各种 DRL 算法所需算力进行研究。先了解各种算法所需硬件资源（如 CPU、GPU、TPU）、内存、存储、训练数据集的大小等计算资源，然后再了解模型的规模对计算资源需求产生的影响，其模型中包括神经网络的层数和神经元数量等参数。

参 考 文 献

[1] HUANG Y, TAN J, LIANG Y C. Wireless big data: Transforming heterogeneous networks to smart networks[J]. Journal of Communications and Information Networks, 2017, 2(1): 19-32.

[2] LIEN S Y, HUNG S C, CHEN K C, et al. Ultra-low-latency ubiquitous connections in heterogeneous cloud radio access networks[J]. IEEE Wireless Communications, 2015, 22(3): 22-31.

[3] YE Q, RONG B, CHEN Y, et al. User association for load balancing in heterogeneous cellular net-

works[J]. IEEE Transactions on Wireless Communications, 2013, 12(6): 2706-2716.

[4] SHEN K, YU W. Distributed pricing-based user association for downlink heterogeneous cellular networks[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(6): 1100-1113.

[5] HATTAB G, CABRIC D. Coverage and rate maximization via user association in multi-antenna HetNets[J]. IEEE Transactions on Wireless Communications, 2018, 17(11): 7441-7455.

[6] WANG N, HOSSAIN E, BHARGAVA V K. Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays[J]. IEEE Transactions on Wireless Communications, 2016, 15(5): 3251-3268.

[7] ZHOU T, FU Y, QIN D, et al. Joint user association and BS operation for green communications in ultra-dense heterogeneous networks[J]. IEEE Transactions on Vehicular Technology, 2023:1-14.

[8] 金明录, 郭楠. 基于 Thomas 簇过程的异构蜂窝网能量效率分析[J]. 通信学报, 2019, 40(10):149-156.
JIN M L, GUO N. Energy efficiency analysis of heterogeneous cellular networks based on Thomas clustering process[J]. Journal of Communications, 2019, 40(10): 149-156.

[9] ZHANG H, LIU H, CHENG J, et al. Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks[J]. IEEE Transactions on Communications, 2017, 66(4): 1705-1716.

[10] LI L, ZHOU Z, SUN S, et al. Distributed optimization of enhanced intercell interference coordination and resource allocation in heterogeneous networks[J]. International Journal of Communication Systems, 2019, 32(6): e3915.

[11] PHUNCHONGHARN P, HOSSAIN E, KIM D I. Resource allocation for device-to-device communications underlaying LTE-advanced networks[J]. IEEE Wireless Communications, 2013, 20(4): 91-100.

[12] LIU Q, HAN T, ANSARI N, et al. On designing energy-efficient heterogeneous cloud radio access networks[J]. IEEE Transactions on Green Communications and Networking, 2018, 2(3): 721-734.

[13] XU S, LI R, YANG Q. Improved genetic algorithm based intelligent resource allocation in 5G Ultra Dense networks[C]. 2018 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2018: 1-6.

[14] COSKUN C C, AYANOGLU E. Energy-and spectral-efficient resource allocation algorithm for heterogeneous networks[J]. IEEE Transactions on Vehicular Technology, 2017, 67(1): 590-603.

[15] COSKUN C C, AYANOGLU E. Energy-spectral efficiency tradeoff for heterogeneous networks with QoS constraints[C]. 2017 IEEE International Conference on Communications (ICC). IEEE, 2017: 1-7.

[16] YANG H, ZHAO J, LAM K Y, et al. Distributed deep reinforcement learning-based spectrum and power allocation for heterogeneous networks [J]. IEEE Transactions on Wireless Communications, 2022, 21(9): 6935-6948.

[17] 张茜茜,李君,李正权,等.基于多智能体 Actor-Critic 算法的异构网络能效优化[J].电子测量技术,2022, 45(22):12-18.
ZHANG Q Q, LI J, LI ZH Q, et al. Energy efficiency optimization of heterogeneous networks based on multi-agent Actor Critic algorithm [J]. Electronic Measurement Technology, 2022, 45 (22): 12-18.

[18] THAM M L, IQBAL A, CHANG Y C. Deep reinforcement learning for resource allocation in 5G communications[C]. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019: 1852-1855.

[19] 康守强,刘哲,王玉静,等.基于改进 DQN 网络的滚动轴承故障诊断方法[J].仪器仪表学报,2021,42(3): 201-212.
KANG SH Q, LIU ZH, WANG Y J, et al. Fault diagnosis method for rolling bearings based on improved DQN network [J]. Chinese Journal of Scientific Instrument, 2021, 42(3): 201-212.

[20] IQBAL A, THAM M L, CHANG Y C. Double deep Q-network-based energy-efficient resource allocation in cloud radio access network [J]. IEEE Access, 2021, 9: 20440-20449.

[21] XIAO L, LI Y, HUANG X, et al. Cloud-based malware detection game for mobile devices with offloading[J]. IEEE Transactions on Mobile Computing, 2017, 16(10): 2742-2750.

[22] CHALLITA U, DONG L, SAAD W. Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective[J]. IEEE Transactions on Wireless Communications, 2018, 17 (7): 4674-4689.

[23] SHI Y, ZHANG J, LETAIEF K B. Group sparse beamforming for green cloud-RAN[J]. IEEE Transactions on Wireless Communications, 2014, 13(5): 2809-2823.

[24] 刘子怡,李君,李正权.多用户蜂窝网络中基于深度强化学习的功率分配[J].国外电子测量技术,2023,42(3):30-35.
LIU Z Y, LI J, LI ZH Q. Power allocation based on deep reinforcement learning in multi-user cellular networks[J]. Foreign Electronic Measurement Technology, 2023,42(3): 30-35.

[25] YANG H, ZHAO J, LAM K Y, et al. Deep reinforcement learning based resource allocation for heterogeneous networks[C]. 2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2021: 253-258.

[26] 李国燕,史东雨,张宗辉.基于改进Dueling DQN的多园区网络动态路由算法[J].电子测量与仪器学报,2022,36(11):211-220.
LI G Y, SHI D Y, ZHANG Z H. Multi campus network dynamic routing algorithm based on improved Dueling DQN [J]. Journal of Electronic Measurement and Instrumentation, 2022, 36 (11): 211-220.

[27] MUKKAMALA M C, HEIN M. Variants of rmsprop and adagrad with logarithmic regret bounds[C]. International Conference on Machine Learning. PMLR, 2017: 2545-2553.

[28] 陈波杰,蔡乐才,刘星,等.一种优化LSTM神经网络模型的预测方法[J].四川轻化工大学学报(自然科学版),2022,35(5):78-86.
CHEN B J, CAI L C, LIU X, et al. A prediction method for optimizing LSTM neural network models[J]. Journal of Sichuan University of Light Industry and Chemical Technology (Natural Science Edition), 2022, 35(5): 78-86.

[29] 张旭,韦洪旭.基于AdaGrad自适应策略的对偶平均方法[J].舰船电子工程,2022,42(9):41-44,53.
ZHANG X, WEI H X. Dual averaging method based on AdaGrad adaptive strategy[J]. Ship Electronic Engineering, 2022, 42(9): 41-44,53.

作者简介

符平博,硕士研究生,主要研究方向为深度强化学习、网络资源优化。
E-mail:2680794348@qq.com

陶旭(通信作者),博士,讲师,主要研究方向为功率器件电子学、光子检测技术及其应用等。
E-mail:tx_tju_nju@163.com

张见,博士,讲师,主要研究方向为光通信材料与器件、通信信号处理等。
E-mail:408677197@qq.com

李晖,教授,主要研究方向为星地互联网络、6G移动通信、异构网络优化等。
E-mail: hitlihui1112@163.com