基于强化学习的异构超密度网络资源分配算法

吴 锡,任正国,孙 君
(南京邮电大学 江苏省无线通信重点实验室,江苏 南京 210003)

摘 要:为了保证下行链路用户服务质量(Quality of Service,QoS),提升异构超密度网络的频谱利用率(Spectrum Efficient,SE)和能源效率(Energy-Efficient,EE),提出了一种基于多智能体强化学习(Deep Reinforcement Learning,DRL)的频谱和功率联合分配算法。首先,以频谱利用率和能源效率为优化目标,用户服务质量为约束,得到资源分配优化函数。然后定义多智能体用户状态空间,奖励以及动作空间,通过较小的通信开销获得状态空间信息,得到一维状态空间数据,减少网络的输入数据量,用户利用自身的信道状态信息(Channel State Information,CSI)而不依赖全局信道状态信息,再根据状态空间信息得到频谱和功率分配策略。最后,通过训练深度神经网络找到最佳的资源分配策略。仿真结果表明,该算法可以实现较快的收敛速度,对比贪婪算法以及其他强化学习方法,能源效率均提升 20% 以上,频谱利用率分别提升 27% 和 11%。

关键词:异构超密度网络;强化学习;资源分配;功率分配;用户服务质量
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2023)01-0114-07
doi:10.3969/j.issn.1673-629X.2023.01.018

Resource Allocation Algorithm for Heterogeneous Ultra-dense Networks Based on Reinforcement Learning

WU Xi,REN Zheng-guo,SUN Jun
(Jiangsu Key Laboratory of Wireless Communications,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:In order to ensure the quality of service (QoS) for downlink users and improve the spectrum efficiency (SE) and energy efficiency (EE) of heterogeneous ultra-dense networks,a multi-agent based joint spectrum and power allocation algorithm of deep reinforcement learning (DRL) is proposed. Firstly,we obtain the resource allocation optimization function with spectrum utilization and energy efficiency as the optimization goal. Secondly,the user state space,reward,and action space are defined,and state space information is obtained through relatively small communication overhead,which is one-dimensional data,and the amount of input data to the network is reduced. Users use their own channel state information (CSI) instead of relying on the global channel state information and then obtain spectrum and power allocation strategies based on the state information. Finally,the best resource allocation strategy is found by training a deep neural network. The simulation results show that the proposed algorithm can achieve a faster convergence speed. Compared with the greedy algorithm and other reinforcement learning methods,the energy efficiency is increased by more than 20%,and the spectrum utilization rate is increased by 27% and 11%,respectively.

Key words:heterogeneous ultra-dense network;reinforcement learning;resource allocation;power allocation;QoS

0 引 言

随着无线设备的急剧增加,现有的蜂窝网络已经无法满足爆炸式增长的无线业务需求。异构超密度网络使用具有不同传输功率和覆盖范围的微小区和毫微小区来增强现有的蜂窝网络,这些异构网络(HetNet)可以将用户设备(User Equipment,UE)从宏基站(Macro Base Station,MBS)转移到微基站(Pico BS,PBS)和毫微基站(Femto BS,FBS)。此外,PBS 和 FBS 可以重复使用 MBS 并与 MBS 共享相同的信道,实现异构网络的高频谱效率。因此,异构超密度网络

收稿日期:2022-01-17 **修回日期**:2022-05-18
基金项目:国家自然科学基金项目(61771255);中科院重点实验室开放课题(20190904)
作者简介:吴 锡(1996-),男,硕士生,研究方向为物联网关键技术研究;通讯作者:孙 君(1980-),女,博士后,副教授,研究方向为无线资源管理与无线频谱理论研究、无线通信环境与信道建模、终端直接蜂窝网络、未来移动通信网络与技术、物联网资源管理与关键技术研究。

被认为是增加未来无线通信系统容量的方案之一。此类异构网络中存在一些问题亟待优化,例如小区间干扰,资源浪费以及能源消耗大的问题。设计节能高效的无线通信系统已成为一种新趋势[1-3],合理的频谱分配和功率分配策略能显著地提升能源效率和系统容量。文献[4]提出了异构网络中用户关联和资源分配的联合优化方案,但是,考虑到非凸性和组合性,获得联合优化问题的全局最优策略是具有挑战性的。为了提升异构小区的能源效率,满足用户的服务质量,文献[5]研究了基于凸优化的方法分配传输功率以及带宽。文献[6-7]分别研究了博弈论方法、线性规划方法解决联合用户关联和资源分配问题。但是,这些方法需要几乎完整的信道状态信息,而完整的信道状态信息通常很难得到。近年来深度强化学习成为人工智能应用中一种新的研究趋势,并且正在成为解决无线通信系统动态资源分配问题的可行工具。该文专注于利用强化学习方法来解决这一难题。

在强化学习算法中,强化学习代理考虑最大化长期奖励,而不是简单地获得当前的最佳奖励,这对于解决动态的资源分配问题十分有效。在提升系统吞吐量方面,文献[8]研究了一种基于强化学习的媒体访问控制协议,用以学习最佳的信道访问策略。文献[9]考虑了一种多信道无线网络中网络效用最大化的动态频谱访问问题,用户能够从他们的确认字符(Acknowledge character,ACK)信号中学习频谱接入策略;文献[10]提出了基于生成对抗网络(Generative Adversarial Networks,GAN)的深度强化学习方法,用以找到一个最优的带宽共享解决方案。但是上述文献仅考虑频谱资源的分配,并没有涉及功率控制,不够完善。文献[11]研究了多智能体的强化学习方法用以解决无线网络中的功率分配问题,用户根据相邻用户的信道状态信息和 QoS 来调整自己的发射功率,文章虽然以功率控制为出发点,但是并没有考虑系统的能源效率问题,有些欠妥。文献[12]考虑超密集异构网络中的同层干扰和跨层干扰问题,提出了基于强化学习的资源分配方案。文献[13]结合深度学习和强化学习构建神经网络,根据环境状态动态调整信道和功率分配。但是文献[12-13]均未对信道信息有较高的要求。文献[14]研究了 REINFORCE、DQL(Deep Q Learning)和 DDPG(Deep Deterministic Policy Gradient)等方法在多小区功率分配上的性能表现,但是也只是从系统总容量出发,对于用户 QoS 要求、能源效率等方面欠考虑,同时对于信道状态信息要求较高。文献[15-16]研究了超密度蜂窝小区的资源分配问题,提出了基于模型的深度强化学习方法,但是在集中式结构下分析的资源分配问题,同样需要完整的信道状态信息;文献[17]以能源效率为目标,通过与环境的交互来学习混合能源驱动的异构网络中用户调度和资源分配的最优策略,但是算法收敛速度较慢,且不稳定。文献[18]提出了基于强化学习的方法来解决联合用户关联基站和频谱分配问题。虽然考虑到用户的 QoS 要求,但仅从最大化系统容量的角度出发,并没有分析系统的频谱效率。

上述基于文献资源分配主要存在以下问题:(1)全局信道状态信息难以得到;(2)维度问题:由于上述文献中通常需要全局或整个小区内的信道状态信息,导致其神经网络输入输出维度与小区数量、信道数量、用户数成正比,且状态空间随着输入输出维度呈指数增长。此外,在高维空间中的探索效率低下,因此学习可能不切实际。

综上所述,为了在有限的信道状态信息下解决异构超密度网络的下行链路中的频谱和功率联合分配问题,由于联合优化问题的非凸性和组合性,且是一个 NP-hard 问题,提出了一种新的基于多智能体强化学习的分布式优化算法。该算法以满足用户 QoS 为基本要求,提升系统的频谱利用率和能源效率为主要目标。通过强化学习算法训练神经网络,得到接近最优的联合频谱和功率分配策略。现有的工作大多将基站作为资源分配的决策者,在一定情况下增加基站的负担,且全局的信道状态信息的要求较高,该文将资源分配的决策放在用户侧,在算法收敛较快的情况下所造成的计算负担和能耗是可以接受的。

1　系统模型

考虑一个具有 M_m 个宏基站(MBS),M_p 个微基站(PBS)以及 M_f 个毫微基站 $M_m + M_p + M_f = M$ 和 N 个移动用户的异构超密度网络的下行链路,每个小区 BS 位于每个小区的中心,其授权移动用户随机分布在小区。每两个相邻的小区之间有重叠的区域。为了最大程度地利用无线电资源,将频率复用因子设置为 1,为了避免小区内干扰,假设每个小区中的每个用户仅分配一个子信道,因此所有用户信号在同一小区中子信道是正交的。小区中使用的 K 个正交子信道可以在每个相邻小区中重复使用。然而,重叠区域中的用户由最近的小区 BS 服务,由于可能使用相同的频谱资源,他们可能遭受严重的小区间干扰(Inter Carrier Interference,ICI)。网络模型结构如图 1 所示。

令 $d_{m,n}$ 表示基站 $m \in \boldsymbol{M} = \{1,2,\cdots,M\}$ 与用户 $n \in \boldsymbol{N} = \{1,2,\cdots,N\}$ 之间的关联关系,$d_{m,n} = 1$ 表示基站 m 与用户 n 关联,假设用户与具有最高边际效用的基站相关联,则有以下资源分配:

频谱分配:当用户 n 与基站 m 上的子信道 $k(k \in$

$K=\{1,2,\cdots,K\}$）相关联时，频谱状态可以定义为 $c^k_{m,n}(t)=1$，反之 $c^k_{m,n}(t)=0$，表示用户 n 不使用信道 k。

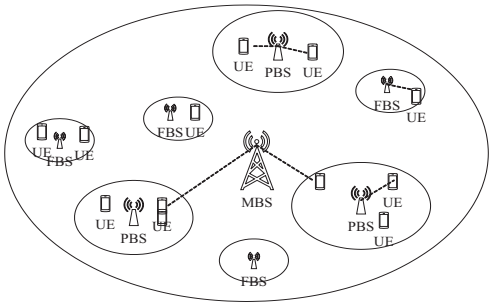

图 1 网络模型结构

功率分配：用 $p^k_{m,n}(t)$ 表示用户 n 与基站 m 在子信道 k 上的传输功率。应确保每个小区 BS 的发射功率在给出的功率限制（式 1）$P^{\max}_m$ 之下。

$$0 \leqslant p^k_{m,n} \leqslant P^{\max}_m, \quad \forall m \in \boldsymbol{M}, \forall k \in \boldsymbol{K} \tag{1}$$

考虑具有平坦衰落的单个频带，采用块衰落模型来表示在时隙 t 中用户 n 到基站 m 的下行链路信道增益 $g_{m,n,k}(t)$ 为：

$$g_{m\to n}(t)=|h^t_{m,n,k}|^2\beta_{m,n,k} \tag{2}$$

其中，$\beta_{m,n,k}$ 表示同时考虑几何衰减和阴影衰落的大尺度衰落分量，它们在不同时隙发生变化但在同一时隙内保持不变。$h^{(t)}_{m,n,k}$ 为小尺度衰落分量，小尺度平坦衰落可以建模为一阶复高斯-马尔可夫过程：

$$h^t_{m,n,k}=\rho h^{(t-1)}_{m,n,k}+n^t_{m,n,k} \tag{3}$$

其中，$n^t_{m,n,k}\sim \mathrm{CN}(0,1-\rho^2)$，$h^1_{m,n,k}\sim \mathrm{CN}(0,1)$，相关系数 ρ 由下式定义：

$$\rho=J_0(2\pi f_d T_s) \tag{4}$$

其中，$J_0(\bullet)$ 是第一类零阶贝塞尔函数；f_d 是最大多普勒频率；T_s 是相邻时刻之间的时间间隔。

不同小区中的用户分配相同的子信道，例如，基站 m 在子信道 k 上服务的用户 n 的 ICI 可以表示为：

$$I^k_{m,n}(t)=\sum_{m'=1}^{M}\sum_{n'=1,n'\neq n,}^{N} d_{m',n}(t)c^k_{m',n}(t)p^k_{m',n}(t)g^k_{m',n}(t) \tag{5}$$

其中，$p^k_{m',n}(t)$ 表示 t 时刻在子信道 k 上基站 m' 到用户 n' 的发射功率，$g^k_{m',n}(t)$ 是在子信道 k 上从基站 m' 到用户 n 的信道增益的平方。因此，基站 m 在子信道 n 上服务的用户 k 的信号干扰加噪声比（Signal to Interference plus Noise Ratio，SINR）由下式给出：

$$\mathrm{SINR}^k_{m,n}(t)=\frac{c^k_{m,n}(t)p^k_{m,n}(t)g^k_{m,n}(t)}{WN_0+I^k_{m,n}(t)} \tag{6}$$

其中，N_0 是从基站 m 到用户 n 的加性高斯白噪声的功率密度；当基站 m 的用户 n 和基站 m' 的用户 n' 同时被分配了子信道 k 时，$p^k_{m',n}(t)$ 将干扰基站 m 的用户 n，且 $m'\neq m$。用户 n 从基站 m 在子信道 k 上的下行链路容量表示为：

$$\Gamma_n(t)=W\log_2(1+\mathrm{SINR}^k_{m,n}(t)) \tag{7}$$

其中，W 为子信道的带宽。

定义能源效率函数 EE 为：

$$\mathrm{EE}=\frac{W\log_2(1+\mathrm{SINR}^k_{m,n}(t))}{p^k_{m,n}(t)} \tag{8}$$

定义频谱效率函数 SE 为：

$$\mathrm{SE}=\frac{\Gamma_n(t)}{W}=\log_2(1+\mathrm{SINR}^k_{m,n}(t)) \tag{9}$$

考虑所有用户想要满足其各自最小的 QoS 要求 Ω，因此，假设用户 n 下行速率 $\Gamma_n(t)$ 不小于最小 QoS 要求 Ω_n，即：

$$\Gamma_n(t)\geqslant\Omega_n \tag{10}$$

为了联合单个用户优化频谱效率和网络能源效率，效用函数可以定义为：

$$\eta_n(t)=\psi\cdot\log_2(1+\mathrm{SINR}^k_{m,n}(t))+(1-\psi)\cdot\frac{B_{m,n}\log_2(1+\mathrm{SINR}^k_{m,n}(t))}{p^k_{m,n}(t)} \tag{11}$$

参数 β 是为了考虑频谱效率和能源效率的折中，该文的目标是在保证用户 QoS 前提下，提升频谱效率和能源效率，则联合优化问题表示为：

$$\begin{aligned}&\max_{\{c^k_{m,n},p^k_{m,n}\}}\sum_{n=1}^{N}\eta_n(t)\\&\text{s.t.}\\&C1:\Gamma_n(t)\geqslant\Omega_n\\&C2:0\leqslant p^k_{m,n}\leqslant P^{\max}_m,\ \forall m\in M,\forall k\in K\end{aligned} \tag{12}$$

2 基于强化学习的联合资源分配

上述异构场景下的联合资源分配问题可以表示为马尔可夫决策过程（Markov Decision Processes，MDP）。Q 学习算法是解决 MDP 问题的最有效的算法之一。然而，异构超密度网络规模庞大，拓扑结构复杂，使得算法的计算复杂度难以控制，DRL 能够很好地解决此类复杂的问题，网络实体经过不断与环境交互，通过学习可以进行自主决策，同时 DNN（Deep Neural Networks）的引入能够在具有大的状态空间和动作空间的问题求解上具有显著优势。为了解决在全局信道状态信息不可知的问题，引入了多智能体的方法，每个智能体只根据自己的信道状态信息以及极小的信息传递便可做出决策。因此，提出了基于多智能体强化学习的联合资源分配框架。本节分别定义了联合资源分配的状态空间、动作空间和奖励函数，然后提出了基于多智能体的强化学习算法解决联合资源分配问题。

2.1 强化学习三要素定义

在强化学习中，智能体（代理）基于策略做出决

策,选择动作对环境造成影响,得到反馈。状态空间、动作空间和奖励函数是强化学习的三要素。对于该文所考虑的异构超密度网络,将用户作为智能体,定义状态空间、动作空间和奖励函数如下:

状态空间:状态空间 S 包括 $t+1$ 个状态 $\{s(0),s(1),\cdots,s(t)\}$。代理每次观察到的用于表征网络环境的状态均由两部分组成 $\{s^q(t),s^g(t)\}$,第一部分定义为用户 QoS 状态,用以指示是否每个用户的当前策略的接收信干噪比在 t 时刻达到事先设定的 QoS 要求,$s^q(t)=\{s_1(t),s_2(t),\cdots,s_N(t)\}$,其中 $s_n(t)\in\{0,1\}$。$s_n(t)=0$ 表示用户 n 的接收信号干噪比不能满足预设的最小 QoS 要求,即 $\mathrm{SINR}_n(t)<\Omega_n$;$s_n(t)=1$ 表示 $\mathrm{SINR}_n(t)\geqslant\Omega_n$。$s^g(t)$ 定义为当前用户当前时刻以及当前用户前一时刻的信道状态信息,即 $s^g(t)=\{g^k_{m,n}(t),g^k_{m,n}(t-1)\}$。由于每个用户仅需要得到自己的信道状态信息,减少了消息传递的开销。

动作空间:根据当前状态,代理可以基于决策策略 π 在 $a\in A$ 处采取动作,符号 A 表示动作集。文中该动作包括选择子信道和相应的传输功率。一旦做出决定,每个授权的移动用户将此决策发送给基站请求下发资源。因此,可以将动作表示为 $a(t)=\{c(t),p(t)\}$。动作 $c(t)$ 是信道选择策略,$c(t)\in\{c^1_{m,n}(t),c^2_{m,n}(t),\cdots,c^K_{m,n}(t)\}$,$c^k_{m,n}(t)\in\{0,1\}$。$p_t$ 定义为下行信道的功率策略。在式(1),下行链路功率是连续变量,并且受到最大功率约束。但由于 DQN(Deep Q Network)的动作空间必须为有限个,因此将可能的发射功率以 $|B|$ 级别量化,该文设置的可选择的功率等级为:

$$B=\{P_m^{\min},P_m^{\min}(\frac{P_m^{\max}}{P_m^{\min}})^{\frac{1}{|B|-1}},P_m^{\max}\} \tag{13}$$

其中,$P_m^{\min}$ 是 BS m 非零的最小发射功率。

每一个代理可能选择的动作数为 $K\times|B|$,随着 K 和 $|B|$ 增大的可选择的动作数会变得很大。此外,不同行为的选择通常会影响状态的演变。在未知的环境中,每个代理使用强化学习的方法获得最优的策略 $\pi_n^*:S\to A$。根据分布式特性,可以将协作式多智能体强化学习与本地状态一起考虑。具体的,所有的用户都尝试通过消息传递基于状态空间来学习最优策略,每个用户仅用一个比特将其自己的状态信息(0 或 1,这里仅发送 $s^q(t)$ 部分的信息。即,当前用户的接收信干噪比是否满足自身最小 QoS 要求)发送到相关联的基站,通过回程通信链路在基站之间传递的信息,其数据为一维,大小仅 N 比特,相较于全局的信道状态信息的三维数据,通信开销可忽略[19]。

奖励:在采取行动后,代理可以计算环境的回报 $r_n(t)$。智能体的唯一目标是最大化总回报。因为 $a_n(t)$ 的行为对奖励 $r_n(t)$ 有直接影响,所以发送给代理的奖励定义了对代理而言是好是坏的行为。在这种情况下,利用效用函数 $\eta_n(t)$ 得到奖励函数 $r_n(t)$,达到系统最大化频谱效率和能源效率的近似最优解。

$$r_n(t)=\psi\cdot\log_2(1+\mathrm{SINR}^k_{m,n}(t))+(1-\psi)\cdot\frac{B_{m,n}\log_2(1+\mathrm{SINR}^k_{m,n}(t))}{p^k_{m,n}(t)}-\zeta \tag{14}$$

其中,ψ 为频谱效率和能源效率的折中因子,ζ 为未到达 QoS 要求的用户数目。

2.2　多智能体联合资源分配策略

在 t 时刻,每个智能体通过观测状态 $s(t)\in S$,按照既定的策略 π 选择相应的动作 $a(t)\in A$,并和环境产生交互,然后得到即时奖励 $r(t)$,进入下一个状态 $s(t+1)$。智能体的目标是学习策略 $\pi:s(t)\in S\to a(t)\in A$,根据其当前状态 $s(t)$ 来选择下一个动作 $a(t+1)$,该策略会产生最大可能的预期累积奖励。

智能体和环境交互,以寻求最大化奖励,使用值函数来评估当前环境的状态和策略,方程式为:

$$G_t=\sum_{l=0}^{\infty}\gamma^l r(t+l+1) \tag{15}$$

其中,$\gamma\in[0,1)$ 是确定未来奖励权重的折现率。若折现因子为 0,则只考虑当前奖励,意味着采取短视的策略,若 $\gamma\in(0,1)$,表示将长远的未来收益考虑到了当前行为产生的价值中。

状态值函数,用以描述遵循策略 π 时一个状态的值。

$$V_n^{\pi}(s)=E_{\pi}[G_t\mid s(t)=s_n]=E_{\pi}[r(t+1)+\gamma V^{\pi}(s(t+1))\mid s(t)=s_n] \tag{16}$$

类似的得到状态行为对的价值函数:

$$q_n^{\pi}(s,a)=E_{\pi}[r(t+1)+\gamma q^{\pi}(S_{t+1},A_{t+1})\mid s(t)=s_n,a(t)=a_n] \tag{17}$$

对于任何 MDP 问题,总存在一个确定性的最优策略;同时如果知道最优行为价值函数,则表明找到了最优策略。

最优状态价值函数是所有状态价值函数中的最大值,为:

$$V_n^*(s)=\max_{\pi}E_{\pi}\{[r(t+1)+\gamma V_n^{\pi}(s(t+1))\mid s(t)=s_n]\} \tag{18}$$

针对 $V^*(s)$,一个状态的最优价值等于从该状态出发采取的所有行为产生的行为价值中最大的那个行为价值:

$$V_n^{\pi *}(s)=\max_{a_n\in A}Q^*(s_n,a_n) \tag{19}$$

于是可以通过找到最优行为价值函数来寻找最佳

策略 $\pi*$。根据 Q 学习算法,通过以下公式更新 Q 值 $Q_n(s_n,a_n)$:

$$Q_n(s_n,a_n)\leftarrow Q_n(s_n,a_n)+\alpha_k[r(t+1)+\gamma\max_{a\in A}(Q_n(s_n',a_n)-Q_n(s_n,a_n))] \tag{20}$$

其中,α_k 是学习率,s_n' 表示用户 n 下一状态。由于传统的 Q 学习算法使用 Q 表来储存给定状态与动作时的价值,在异构超密度网络中,由于基站密集部署,网络环境变得复杂,使得动作空间大小随着基站的数量指数增加,很难通过 Q 表值的方法找到最优策略。将深度学习与 Q 学习算法结合的深度强化学习网络算法很好地解决了 Q 学习维度受限的问题。

在 DQN 中,DNN 用来表示动作和状态空间,DNN 输入是当前的状态,输出是当前可执行状态的动作的 Q 值表的近似,具有权重 θ 的 NN(Neural Networks)函数逼近器 $Q_n(s_n,a_n;\theta)\approx Q_n(s_n,a_n)$。DQN 使用目标网络和在线网络来稳定整体网络性能,目标网络是在线网络 $Q_n(s_n,a_n;\theta)$ 的副本,但其权重在数次迭代中固定不变。目标网络的权重每经过一定次数的迭代,更新为在线网络中的权重。损失函数定义为:

$$L_n(\theta)=E[(y_n^{\mathrm{DQN}}-Q_n(s_n,a_n;\theta))^2] \tag{21}$$

其中,$y_n^{\mathrm{DQN}}=r(t+1)+\gamma\max_{a\in A}Q_n(s_n',a_n;\theta^-)$,$\theta^-$ 表示目标网络的权重。其中,由于每一个智能体所处的环境不同,其权重也不同。使用 ε 贪心策略从在线网络 $Q_n(s_n,a_n;\theta)$ 中选择动作 a_n,用以权衡开发与探索的策略。开发是对于当前时刻,本着最大化动作价值的原则选择最优的动作;探索则是从长远角度考虑可能带来的最大化总收益。

在 DQN 中,为了克服学习的不稳定性,使用了经验重播策略。将元组 (s_n,a_n,r,s_n') 存储在体验重播内存 D 中。在学习过程中,不仅使用当前的体验 (s_n,a_n,r,s_n'),可以通过从回放内存 D 随机抽样均匀的小批经验来训练 NN。通过减少训练样本之间的相关性,体验重播策略可确保最佳策略不能被收敛到局部最小值。

此外,由于在 Q 学习和 DQN 方法中使用相同的值来选择和评估动作,因此可能会过于乐观地估计 Q 值函数。因此,DDQN(Double DQN)用于通过将目标 y_n^{DQN} 替换为以下目标 y_n^{DDQN} 来缓解上述问题。

$$y_n^{\mathrm{DDQN}}=r(t+1)+\gamma Q_n(s_n',\arg\max_{a\in A}Q_n(s_n',a_n;\theta^-);\theta^-) \tag{22}$$

为了实现更好的策略估计,引入对决神经网络[20]获得优势函数 $A_n(s_n,a_n)=Q_n(s_n,a_n)-V_n(s_n)$,动作 a_n 对比其他动作的优势可用 $A_n(s_n,a_n)$ 表示。因此,在决斗架构中,DDQN 的最后一层由 $V_n(s_n)$ 和 $A_n(s_n,a_n)$ 两个子网络组成。通过组合 $V_n(s_n)$ 和 $A_n(s_n,a_n)$,可以估计动作价值函数 $Q_n(s_n,a_n)$。

该文所用的强化学习模型如图 2 所示。每一个用户拥有一个代理,每一个代理拥有两个 DQN 网络,一个是在线网络,一个是目标网络。在线网络的主要作用是根据用户得到的状态信息,包括用户的当前功率分配和信道选择策略是否满足预先设置的下行速率,以及自身的信道状态信息,输出功率分配和信道分配的决策,然后用户将此策略发送给基站,请求基站根据此策略分配资源,基站根据自身可用的资源拒绝或同意用户的请求。目标网络的作用是辅助在线网络的权重参数更新。

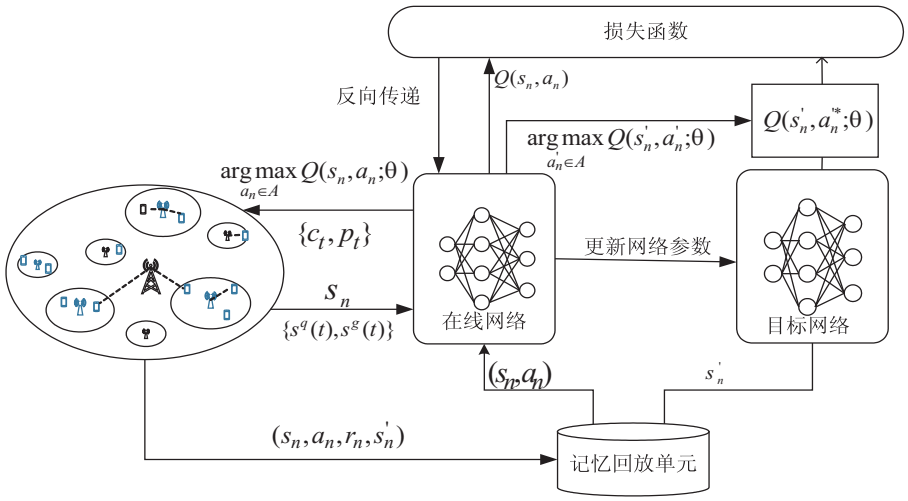

图 2 强化学习模型

算法 1 描述了联合优化问题的多智能体强化学习方法。在每个训练情节开始时,初始化状态信息,每个用户向其关联的基站报告其自身的当前状态(仅为 $s^q(t)$),通过消息回程通信链路在基站间传递,获得所有用户的 QoS 信息。然后,基站将该信息发送给所有用户。每个情节持续 T 步。例如,在每个情节的步骤中,用户 n 使用 ε 贪婪策略从估计的 Q 值 $Q_n(s_n,a_n)$ 中选择执行动作。然后,每个用户将其子信道分配和

功率分配请求发送到该用户已经选择与其关联的基站。该请求包含所需子信道的索引和下行发射功率。然后,基站根据其可用资源来接受或拒绝来自用户 n 的请求。如果基站接受用户 n 的请求,基站将以用户请求的功率向用户 n 发送反馈信号。此时用户可以根据得到的反馈计算下行链路速率,进而计算出奖励函数。反之基站拒绝用户请求,不向用户的请求做出反馈,此时用户得到奖励为负。然后,在获得立即奖励 $r_n(s_n,a_n)$ 和更新下一状态之后,每个用户更新 Q 值。当满足所有用户的 QoS 或达到最大步长 T 时,当前情节结束。

算法 1

1. 初始化重现存储 D ,以及目标网络替换步长 N^- 。

2. 初始化在线网络 $Q_n(s_n,a_n;\theta)$ 和权重 θ ,初始化在线网络 $Q_n(s_n^{'},a_n^{'};\theta^-)$ 并使权重 $\theta^- = \theta$ 。

3. 重复 500 个情节,每一个情节重复 500 步,对于每一步进行以下操作:

(a)所有的 UE 根据当前状态信息并使用 ε 贪婪策略选择出决策 $\{C_{m,n}^k,p_{m,n}^k\}$ 。

(b)所有的 UE 更新环境 $s(t+1)$,得到奖励 $r(t+1)$ 。

(c)所有的 UE 将各自的 $(s_n,a_n,r_n,s_n^{'})$ 存储到各自的记忆回放单元 D 中。

(d)所有的 UE 从各自的记忆回放单元 D 中随机抽取样本,计算损失函数 $L_n(\theta)$,并更新权重。

4. 每隔 N^- 步,所有的 UE 将各自的目标网络参数 θ^- 替换为在线网络权重 θ 。

5. 当所有的 UE 满足 QoS 条件,或者达到最大迭代步骤,结束当前情节。

3　仿真分析

在这一节中,给出了所提多智能体强化学习资源分配算法在异构超密度网络中的下行链路中的性能表现,并给出了该算法与其他 RL 算法[13]以及贪婪算法的对比。采用 tensorflow 平台实验仿真,仿真设置宏基站的数量为 2,微基站的数量为 8,毫微基站的数量为 12,以及用户数 $N \in \{20,25,30,35,40\}$,且用户随机分布在各个小区的范围内。宏基站和微基站的覆盖半径分别 500 m、100 m,最大传输功率分别为 38 dbm、30 dbm,二者的路径损耗模型均为 $34 + 40 * \log_{10}(d)$,毫微基站的覆盖半径为 30 m,最大传输功率为 20 dbm,其路径损耗模型为 $37 + 30 * \log_{10}(d)$ 。信道带宽为 180 kHz,噪声功率密度 N_0 为−174 dBm/Hz。重现存储 D 的大小为 500,抽样批次的长度为 32,学习率参数为 0.000 05。网络布置如图 3 所示。

图 4 给出了该算法在不同学习率收敛的回合数训练效率的表现,在学习过程开始时,训练步骤都非常大,这是因为经过初始化,代理没有之前学习的经历,很难找到令所有用户满足 QoS 要求的策略,需要经过很长的迭代步骤才能收敛,甚至达到预设的最大回合数都不能收敛。但随着情节数目的增加,代理经过学习,收敛速度加快,对比不同的学习率,当学习率为 0.000 05时,在 40 回合后,不到 10 步就能够收敛,而学习率为 0.001 时,收敛较慢。这是因为学习率对于网络来说太大了,只有合适的学习率才能使收敛更快。

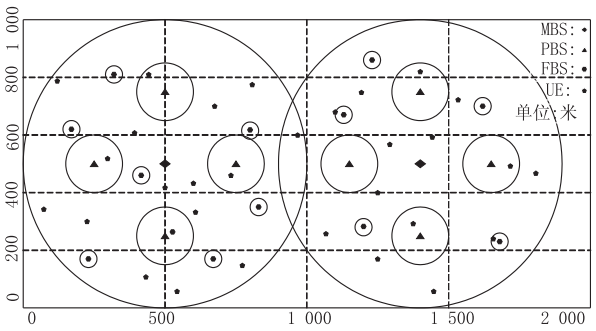

图 3　网络布置

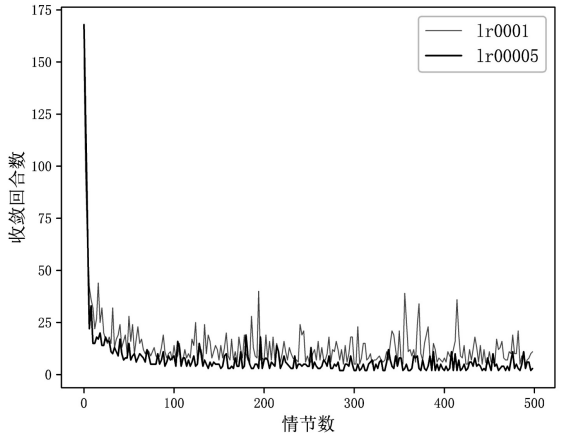

图 4　训练步骤与回合数曲线

图 5 给出了不同算法在不同功率分配等级条件下系统能源效率的表现。相较于贪婪算法和文献[13]中算法,该文得到的能源效率分别提升了 26.43% ~ 43.47% 和 22.68% ~ 33.25%。随着功率分级数量的提升,在一定区间内会提升能源效率,但过高的分级数量会增加计算复杂度,且提升的能量效率有限。

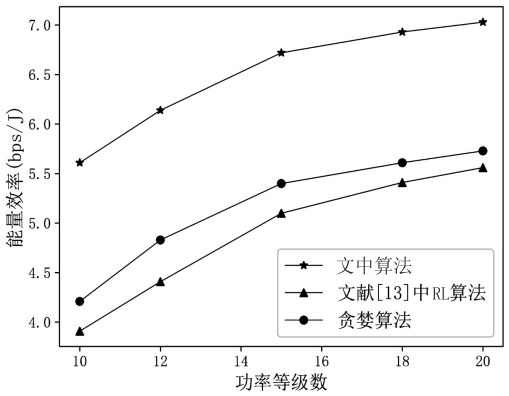

图 5　能源效率与功率分配等级关系曲线

图 6 给出了三种不同算法在不同子信道数的条件下系统频谱效率的表现,随着子信道数量的增加,三种算法的频谱利用率均有下降。这是因为随着子信道数量的增加,在相同的用户数量下,信道的复用效率降低。对比文献[13]中的算法和贪婪算法,该算法的频谱利用率均有所提升。

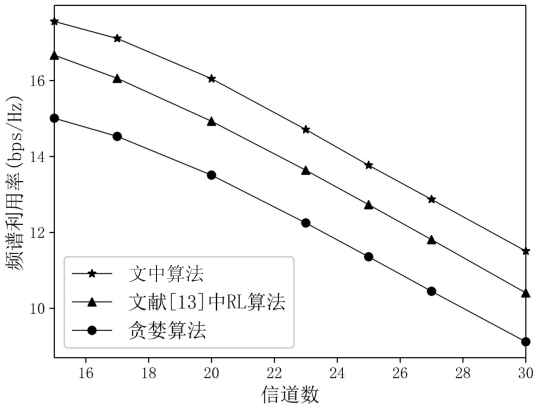

图 6 频谱效率与信道数量关系曲线

在图 7 中,比较了三种算法在不同用户数量条件下系统吞吐量的变化。可以看出随着用户数量的增加,系统总的吞吐量增大,但是增长的趋势有所减缓,在用户数与信道数相等时,多个用户共用同一条信道的情况较少,干扰较小,这时所提方案平均每个用户的下行速率最大值位于初始点,达到 2 Mpbs/s,此时系统总容量为 40.21 Mbps/s,随着用户数量增加,干扰增大,平均每个用户的下行速率减少。相比文献[13]中的算法和贪婪算法,文中算法系统速率提升最大达到 14.11% 和 25.65% 。

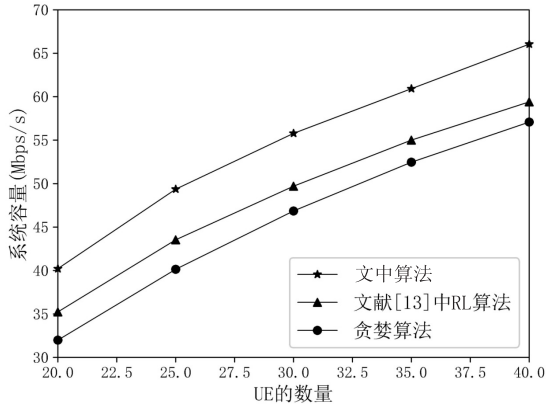

图 7 系统容量与用户数量关系曲线

4 结束语

在异构超密度网络,为了满足用户最小 QoS 要求,提升系统频谱利用率以及能源效率,提出了基于多智能体强化学习框架的分布式资源管理算法。将能源效率和频谱效率作为奖惩值,并通过有限的消息传递得到状态信息,再根据状态信息分配频谱和功率策略,然后反复训练更新神经网络,使得到的策略趋向于最优策略。仿真结果表明,该算法可以满足用户需求,提升网络能效,有效解决复杂动态网络下的资源分配问题。

参考文献:

[1] JIANG C X, ZHANG H J, REN Y, et al. Energy-efficient non-cooperative cognitive radio networks: micro, meso, and macro views[J]. IEEE Communication Magazine, 2014, 52(7): 14-20.

[2] YU F R, ZHANG X, LEUNG V C M. Green com-munications and networking[M]. Boca Raton: CRC Press, 2012.

[3] XU C, SHENG M, YANG C G, et al. Pricing-based multi-resource allocation in OFDMA cognitive radio networks: an energy efficiency perspective[J]. IEEE Transactions on Vehicular Technology, 2014, 63(5): 2336-2348.

[4] HAN Q N, YANG B, MIAO G W, et al. Backhaul-aware user association and resource allocation for energy-constrained HetNets[J]. IEEE Transactions on Vehicular Technology, 2017, 66(1): 580-593.

[5] 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, 2018, 66(4): 1705-1716.

[6] BAYAT S, LOUIE R H Y, HAN Z, et al. Distributed user association and femtocell allocation in heterogeneous wireless networks[J]. IEEE Transactions on Communications, 2014, 62(8): 3027-3043.

[7] ELSHERIF A R, CHEN W P, ITOA, et al. Resource allocation and inter-cell interference management for dual-access small cells[J]. IEEE Journal on Selected Areas in Communications, 2015, 33(6): 1082-1096.

[8] 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.

[9] 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.

[10] HUA Y, LI R, ZHAO Z, et al. GAN-based deep distributional reinforcement learning for resource management in network slicing[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(2): 334-349.

[11] NASIR Y S, GUO D. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2239-2250.

[12] 郑冰原,孙彦赞,吴雅婷,等. 基于 DQN 的超密集网络能效

(下转第 129 页)