DOI:10.19651/j.cnki.emt.2210146

D2D 通信中基于深度强化学习的资源分配*

沈国丽[1] 李 君[2] 李正权[3,4]
(1.南京信息工程大学 南京 210044;2.无锡学院 无锡 214105;3.江南大学轻工过程先进控制教育部重点实验室 无锡 214122;4.北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘 要:设备到设备(D2D)通信能够以蜂窝设施为基础来提高资源利用率、用户吞吐量和节省电池能量。在 D2D 网络中,模式选择和资源分配是关键问题。为了提高 D2D 通信的和速率与频谱利用效率,提出一种联合模式选择、功率和资源块分配的方案。首先根据用户地理位置选定模式选择标准,帮助用户选择相应的通信模式;然后针对复用通信模式,使用基于深度强化学习的异步优势动作评价(A3C)算法为不同的 D2D 用户分配资源块和功率。仿真结果表明,本文提出的基于 A3C 算法的联合优化方案收敛速度快,并且性能相对于其他算法较好。

关键词:模式选择;功率分配;资源分配;D2D 通信;深度强化学习
中图分类号:TN929.5 **文献标识码**:A **国家标准学科分类代码**:510

Resource allocation based on deep reinforcement learning in D2D communication

Shen Guoli[1] Li Jun[2] Li Zhengquan[3,4]
(1. Nanjing University of Information Science & Technology, Nanjing 210044, China; 2. Wuxi University, Wuxi 214105, China; 3. Key Laboratory of Advanced Control of Light Industry Process, Ministry of Education, Jiangnan University, Wuxi 214122, China; 4. State Key Laboratory of Network and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Device to device (D2D) communication can be based on cellular facilities to improve resource utilization, user throughput and save battery energy. In D2D network, mode selection and resource allocation are the key issues. In order to improve the sum rate and spectrum efficiency of D2D communication, a scheme of joint mode selection, power and resource block allocation is proposed. Firstly, the mode selection criteria are selected according to the user's geographical location to help the user select the corresponding communication mode; Then, for the multiplexing communication mode, the asynchronous dominant action evaluation (A3C) algorithm based on deep reinforcement learning is used to allocate resource blocks and power to different D2D users. The simulation results show that the joint optimization scheme based on A3C algorithm proposed in this paper has fast convergence speed and better performance than other algorithms.

Keywords: mode selection; power distribution; resource allocation; D2D communication; deep reinforcement learning

0 引 言

如今,无线通信已经极大地改变了人们的生活。一方面用户对无线频谱资源的需求急剧膨胀,导致无线频谱资源的稀缺程度不断增大,另一方面用户要求更高的数据速率、更低的时延和能耗[1]。为了给用户提供更高标准和更多样的无线服务,人们倾向于在第五代(5G)移动通信中采用更先进的技术[2]。在这些技术中,设备到设备(device-to-device,D2D)通信越来越受到关注,通信中两个对等的移动用户之间可以直接进行信息传输,并能与蜂窝用户复用相同的频谱资源。

对蜂窝网络中的 D2D 通信的研究存在许多挑战。模式选择、资源块分配和功率控制是 D2D 通信中的几个关键问题,后两点是本文的重点关注问题。具体来说,D2D 通信可以使用三种模式:蜂窝模式、专用模式和复用模式[3],其中将后两种通信模式作为 D2D 通信模式[4]。复用模式

收稿日期:2022-05-28
* 基金项目:国家自然科学基金(61571108)、网络与交换技术国家重点实验室(北京邮电大学)开放课题资助项目(SKLNST-2020-1-13)资助

可以实现更高的频谱利用率,但是会产生同频干扰;专用模式下通过使用相交的频谱资源从而避免用户间干扰,但代价是降低了频谱利用率。根据信道条件、功率限制和用户服务质量要求,文献[5-8]中提出了不同的模式选择策略。本文将研究基于距离的模式选择策略,以减小 D2D 链路对蜂窝链路的干扰。

在资源分配研究中,每个用户需要选择通信资源块,同时优化其发射功率。近年来,有许多研究工作致力于解决 D2D 环境中的资源分配问题,通过采用基于组合优化[9]、随机几何[10]或博弈论[11]的集中式算法来联合优化资源分配,以提高和速率或总能效。然而在以上工作中,决策者拥有完整的信道状态信息(CSI),这在实际系统中通常是不可能的,因为这会导致巨大的开销,同时信道不停的快速变化会导致资源分配的不确定性。在文献[12-13]中,研究了将深度强化学习(deep reinforcement learning,DRL)应用到视频游戏和 Alpha Go 中,并取得了成功,之后又在文献[14-16]中取得了显著进展。文献[17]提出了一种联合小基站睡眠和功率控制的分布自适应 RL 算法以改善 EE。使用分布式深度强化学习进行资源分配,可以有效解决上述信令开销巨大等问题。如文献[18]中,将每个用户视为智能体,通过与未知车辆环境交互,实现了更好的资源共享策略。因此,本文研究使用多智能体深度强化学习方法来解决 D2D 用户的资源分配问题。

在本文中,为了解决 D2D 通信的模式选择、资源块分配和功率分配问题,提出了基于地理位置的 K 近邻查询算法和基于分布式框架的异步优势动作评价(asynchronous advantage actor-critic, A3C)算法。本文 A3C 算法中,将 D2D 用户作为智能体,采用集中式训练分布式执行的方法,合理分配资源块和功率。通过此方案能有效的减少 D2D 用户与蜂窝网络的系统间干扰,最大化和速率。

1 系统模型和问题描述

1.1 系统模型

本文考虑单小区上行蜂窝网络,包括蜂窝用户和 D2D 用户对在内的所有无线用户都位于该小区中,并且每个无线用户都配备单个天线。如图 1 所示,基站(BS)位于蜂窝小区中心,系统中随机分布两种用户,集合 $M=\{1,2,\cdots,M\}$,$K=\{1,2,\cdots,K\}$,$B=\{1,2,\cdots,B\}$ 分别表示蜂窝用户(Cue)集合,D2D 用户集合,频谱子带集合,其中 $B>M$。DueT 为 D2D 用户的发射端,DueR 为 D2D 用户的接收端。

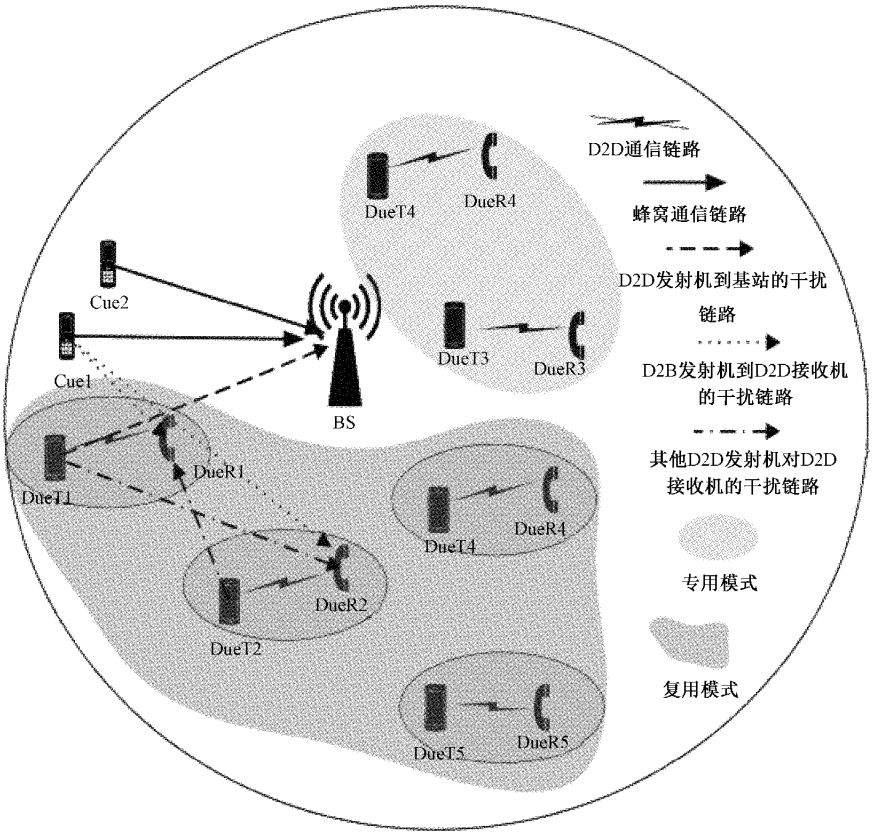

图 1 D2D 通信网络的系统模型

M 个蜂窝用户已经预先分配了 $B_R(M=B_R)$ 个正交频谱子带,K_D 个 D2D 用户复用剩下的 B_D 条频谱子带,K_R 个 D2D 用户与蜂窝用户共享频谱。每个蜂窝用户占用一个频谱子带,多个 D2D 用户可以共享相同资源块,一个 D2D 用户只能占用一个资源块。本文假设对等设备发现和会话设置在模式选择和资源分配之前完成。

当共享频谱时，用户间会相互干扰，这会大大影响系统和速率和通信可靠性。因此，本文的主要工作是为 D2D 用户设计有效的频谱共享方案，以便蜂窝(D2B)链路和 D2D 链路在考虑到移动用户的高动态环境下，以最小的信令开销实现各自的目标。

1.2　D2D 通信模式

D2D 的一个重要问题是用户之间的通信模式，因为合适的通信模式会增加系统和速率。在 D2D 用户的模式中，可以从以下两种通信模式中选择一种。公式中使用的所有符号在表 1 中给出。

表 1　符号及其定义

符号	定义
$p_m[b]$	占据第 b 个频谱子带的 D2B m 的发射功率
$p_k[b]$	占据第 b 个频谱子带的第 k 个 D2D 用户发射端的发射功率
$p_{k'}[b]$	占据第 b 个频谱子带的第 k' 个 D2D 用户发射端的发射功率
$g_{m,B}[b]$	链路(从第 m 个 CUE 到 BS)的信道增益
$g_{k,B}[b]$	链路(从第 k 个 D2D 发射端到 BS)的信道增益
$g_k[b]$	第 k 个 D2D 链路的信道增益
$g_{k',k}[b]$	链路(从第 k' 个 D2D 发射端到第 k 个 D2D 接收端)的信道增益
$g_{m,k}[b]$	链路(从第 m 个 CUE 到第 k 个 D2D 接收端)的信道增益
$\rho_k[b]$	二进制频谱分配指标，若 $\rho_k[b]=1$，意味着第 k 个 D2D 链路复用第 b 个频谱子带，否则 $\rho_k[b]=0$
σ^2	每个子信道上的加性高斯白噪声
W	每个子信道的带宽

1)复用模式

在复用模式下，多个 D2D 用户可以复用相同的蜂窝频谱子带。复用第 b 个频谱子带的第 k 个 D2D 链路接收到的信干噪比(signal to interference and noise，SINR)和数据传输速率可以分别表示为：

$$\gamma_k[b]=\frac{p_k[b]g_k[b]}{\sum\limits_{k'\in K,k'\neq k}\rho_{k'}[b]p_{k'}[b]g_{k',k}[b]+p_m[b]g_{m,k}[b]+\sigma^2} \tag{1}$$

$$R_k[b]=W\log(1+\gamma_k[b]) \tag{2}$$

占据第 b 个频谱子带的 D2B 链路接收到的 SINR 和数据传输速率可以分别表示为：

$$\gamma_m[b]=\frac{p_m[b]g_{m,B}[b]}{\sum_{k\in K}\rho_k[b]p_k[b]g_{k,B}[b]+\sigma^2} \tag{3}$$

$$R_m[b]=W\log(1+\gamma_m[b]) \tag{4}$$

2)专用模式

在专用模式下，D2D 用户单独占用子信道进行通信，此时没有蜂窝用户的参与，并同时允许多个 D2D 用户复用相同的子信道进行数据传输。占据第 b 个频谱子带的第 k 个 D2D 链路接收到的 SINR 和数据传输速率可以分别表示为：

$$\gamma_k^D[b]=\frac{p_k[b]g_k[b]}{\sum\limits_{k'\in K,k'\neq k}\rho_{k'}[b]p_{k'}[b]g_{k',k}[b]+\sigma^2} \tag{5}$$

$$R_k^D[b]=W\log(1+\gamma_k^D[b]) \tag{6}$$

1.3　问题描述

首先引入模式选择标准，确定 D2D 用户的工作模式。针对模式选择问题，本文以 D2D 发射端到基站的距离为模式选择标准，采用 K 近邻查询算法，选择距离基站最近的 K 个 D2D 用户作为专用模式。然后模式选择完成后，在复用模式的基础上，联合资源分配和功率分配。

蜂窝链路主要是支持具有移动性的高数据传输速率的娱乐服务，因此可以设计目标为最大化蜂窝用户的和速率，其定义为 $\sum_{m\in M}R_m[b]$。D2D 链路主要负责用户间进行可靠地安全信息交互，因此可以将此问题建模为在时间 T 内，大小为 D 的数据包的传输速率：

$$R_k=\sum_{t=1}^{T}\sum_{b=1}^{B}\rho_k[b]R_k[b,t],k\in K \tag{7}$$

在本文中，D2D 用户数据交互成功的要求为：$R_k\geqslant\frac{D}{T}$，由此可以将 D2D 链路的信息交互失败率表示为：

$$Pr\left\{R_k\leqslant\frac{D}{T}\right\},k\in K \tag{8}$$

因此，在满足蜂窝用户 SINR 的最低要求前提下，本文中研究的资源分配优化问题正式表述为：通过 D2D 用户的资源块分配 $\rho_k[b]$ 和发射功率控制 $p_k[b]$，同时最大化蜂窝链路和速率与 D2D 链路中数据传输速率来体现。因此，本文将时隙 t 下的优化问题用数学公式表示为：

$$\underset{\rho,p}{arg\max}\{\sum_{m\in M}R_m[b]+\sum_{k\in K}\sum_{b\in B}[\rho_k[b]((1-\lambda_k)R_k[b]+\lambda_kR_K^D[b])]\} \tag{9}$$

$$s.t.\ \gamma_m^t\geqslant\gamma_{th},\forall m\in M,\forall t\in T \tag{10}$$

$$\lambda_k\in\{0,1\},\forall k\in K \tag{11}$$

$$0\leqslant p_k[b]\leqslant p_k^{\max},\forall k\in K,\forall b\in B \tag{12}$$

$$\sum_{b\in B}\rho_k[b]\leqslant 1,\forall k\in K \tag{13}$$

$$\sum_{k\in K}\sum_{b\in B}\rho_k[b]=K \tag{14}$$

其中,$\lambda_k=1$ 表示 D2D 用户工作在专用通信模式,反之则工作在复用通信模式;γ_m^t 为 t 时刻蜂窝用户 m 的 SINR;γ_{th} 为蜂窝用户最小 SINR 要求。约束 (10)表示蜂窝用户的服务质量要求;约束 (11)保证了单个 D2D 用户只能工作在一种模式;条件 (12)表示 D2D 用户的发射功率约束;约束 (13)保证每个 D2D 用户可以复用的频谱子带最多为一个;约束条件 (14)代表每个 D2D 用户都能够有对应的资源块。

2 联合功率和资源块的分配算法

本文使用马尔可夫决策过程(markov decision process, MDP)对上节提出的优化问题进行建模。由于目标函数是一个非凸函数并且优化对象中存在二进制参数,因此其优化是一个混合整数非线性规划(MINLP)问题[19]。为了处理上述问题,可以通过使用深度强化学习算法来解决 MDP 问题。

2.1 深度强化学习

本文将选择复用模式进行通信的 D2D 用户作为智能体,同时与环境交互,并更新功率和资源块分配策略。为防止智能体间相互竞争,通过设置相同的奖励使其重心放在网络的整体性能上。本文采用集中式训练分布式执行的方法。训练时每个智能体的奖励为系统奖励,并通过 A3C 调整策略;执行时每个智能体只得到局部的环境信息,然后根据训练后的 A3C 网络来作决策。本文的状态空间、动作空间、策略、奖励函数定义如下:

1)状态空间

由 $S=\{s_k(t),k\in K_R,t\in T\}$ 表示状态空间。$s_k(t)$ 是第 k 个智能体在时隙 t 的状态,包括以下参数:

(1)系统中所有蜂窝用户的信干噪比:$\Gamma=\{\gamma_1[b],\gamma_2[b],\cdots,\gamma_m[b]\}_{b\in B_R,m\in M}$;

(2)单个智能体观测到的剩余数据包大小:D_k;

(3)单个智能体观测到的剩余传输时间:T_k;

(4)迭代数:E;

(5)ε-greedy 方法中的随机动作选择概率:ϵ;

因此,第 k 个智能体在时隙 t 的状态表示为:

$$s_k(t)=\{\Gamma,D_k,T_k,E,\epsilon\} \tag{15}$$

2)动作空间

由 $A=\{a_k(t),k\in K_R,t\in T\}$ 表示动作空间。在每个时隙 t,动作 $a_k(t)\in A$ 包括以下参数:

(1)智能体的频谱子带选择:$\rho_k[b]$;

(2)离散为 L 个的发射功率选择:$p_k^l=\dfrac{p_k^{max}}{L}l,l\in L$;

因此,第 k 个智能体在时隙 t 的动作表示为:

$$a_k(t)=\{\rho_k[b],p_k^l\} \tag{16}$$

3)策略

策略是从状态空间到动作空间的映射,可以表示为 $\pi(a_k(t)\mid s_k(t)):S\to A$。

4)奖励函数

为了利用 DRL 算法来解决本文制定的问题,需要将其转化为 DRL 框架的标准模型[20]。由问题描述中可知,本文的目标是双重的:最大化 D2B 链路和速率,同时在时间约束 T 内降低 D2D 用户数据包传输的失败率。对于第 1 个目标,在每个时隙 t 的奖励中,只包括所有 D2B 链路的瞬时和速率:$\sum_{m\in M}R_m[b,t]$;对于第 2 个目标,在数据包传输完成之前,本文设置奖励为 D2D 数据传输速率,传输完成后的奖励设置为一个正常数 φ。因此,在每个时隙 t 的部分奖励设置为:

$$U_k(t)=\begin{cases}\sum_{b\in B}\rho_k[b]R_k[b], & D_k>0\\ \varphi, & \text{其他}\end{cases} \tag{17}$$

其中,φ 是需要根据经验调整的超参数。在训练中,对 φ 进行调整,使其大于通过运行几个时隙而获得的最大 D2D 数据传输速率,并小于此最大值的两倍。在本文中,为了简化目标,当在数据包传输完成之前,设置奖励为 $R_k[b]/10$,传输完成后的奖励设置为 1。

除了以上两个目标的奖励设置外,本文为了满足蜂窝用户的服务质量,在文中设置了奖惩机制。当 D2B 链路的 SINR 小于阈值时,设置奖励为一个负常数 μ,反之则无奖励回馈,即奖励为 0。因此,在每个时隙 t 的部分奖励设置为:

$$L_k(t)=\begin{cases}0, & \gamma_m[b]\geqslant\gamma_{th}\\ \mu, & \text{其他}\end{cases} \tag{18}$$

为了使奖惩值与前面的奖励值为同一等级,此处 μ 值设置为 φ 的负数。因此,本文将每个时隙 t 的奖励设置为:

$$R(t)=\vartheta\sum_{m\in M}R_m[b,t]+(1-\vartheta)\sum_{k\in K}[U_k(t)+L_k(t)],\quad b\in B_R,t\in T \tag{19}$$

其中,ϑ 是平衡奖励的正权重。此处的奖励是所有智能体共享的奖励。

2.2 A3C 算法

如图 2 所示,A3C[21] 是在演员评论家(actor-critic, AC)算法[22]的基础上提出的,不同之处在于 A3C 算法采用多个 AC 并发工作,并异步训练多个 AC 的神经网络,从而能够显著加快收敛速度。在 A3C 框架中,包括一个全局网络和多个线程网络,它们的网络均为 AC 结构,唯一不同点在于全局网络不需要进行训练,仅用于存储 AC 结构的参数。当达到终端状态时,每个工人获取其梯度参数,并将其返回到全局网络,全局网络更新全局参数,并将其分发给工人以保证共享相同的策略。通过这种方式,一方面多线程运行提高了运行效率,另一方面参数的相关性被切断,因此不需要像传统的 DQN 算法那样采用经验回放技术,并且训练收敛速度大大提高。

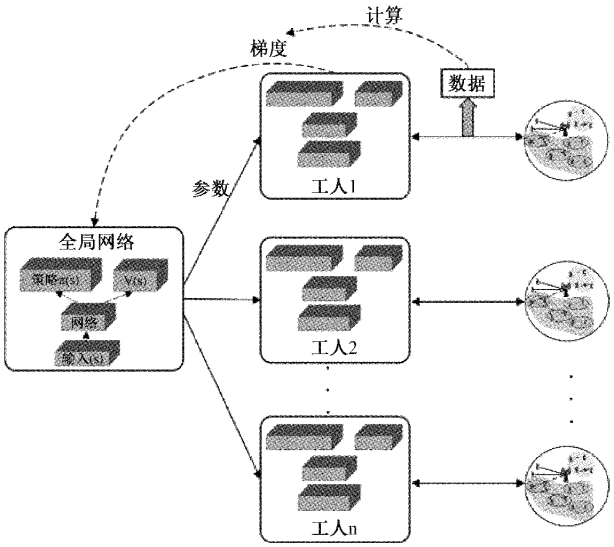

图 2　A3C 框图

2.3　基于 A3C 的联合资源分配框架

本文设置每个回合的长度为数据包传输的约束时间 T。每一回合都以随机初始化的环境状态和大小为 D 的完整 D2D 数据包开始传输，并持续到 T 结束。

在每个时隙 t，环境处于状态 s_t，估计的状态值为 $V(s_t;\theta_v)$。智能体在当前状态 s_t 下根据策略 $\pi(a_t \mid s_t;\theta)$ 执行动作 a_t，环境在一定的概率下将转移到下一状态 s_{t+1}，智能体将获得奖励 r_t。A3C 的状态值函数由下式给出：

$$V(s_t;\theta_v)=E[G_t \mid s=s_t,\pi]=E[\sum_{k=0}^{\infty}\gamma^k r_{(t+k)} \mid s=s_t,\pi] \tag{20}$$

其中，$G_t=\sum_{k=0}^{\infty}\gamma^k r_{(t+k)}$ 是状态累积折扣回报，$\gamma\in[0,1]$ 是折扣因子，表示未来奖励对当前状态值的影响情况。

A3C 使用 k 个时间步，累积奖励定义为：

$$R_t=\sum_{i=0}^{k-1}\gamma^i r_{(t+i)}+\gamma^k V(s_{t+k};\theta_v) \tag{21}$$

其中，k 上限为 t_{max}^{A3C}，并且策略和价值函数都在 t_{max}^{A3C} 或达到最终状态之后更新。

与 AC 算法相比，A3C 在策略梯度阶段采用优势函数 A_t，目的是减少估计的方差，由下式给出：

$$A(s_t,a_t;\theta,\theta_v)=R_t-V(s_t;\theta_v) \tag{22}$$

其中，θ 和 θ_v 分别是 Actor 和 Critic 网络的参数，R_t 是式(21)中定义的真实奖励，$V(s_t;\theta_v)$ 是估计的状态值。因此，优势函数 A_t 可以用来增强智能体的学习能力，以免高估或低估动作，从而提高决策能力。

在提出的基于 A3C 的算法中，Actor 和 Critic 网络都与深度神经网络(DNN)相关联。基于优势函数 A_t，Actor 的损失函数由下式给出：

$$f_\pi(\theta)=\log\pi(a_t \mid s_t;\theta)(R_t-V(s_t;\theta_v))+\beta H(\pi(s_t;\theta)) \tag{23}$$

其中，$H(\pi(s_t;\theta))$ 是熵，用于鼓励在训练过程中进行探索，从而避免陷入局部最优，超参数 β 用于控制熵正则化的强度，从而促进探索和利用之间的权衡。

Critic 网络中的损失函数定义为：

$$f_v(\theta)=(R_t-V(s_t;\theta_v))^2 \tag{24}$$

用于更新价值函数 $V(s_t;\theta_v)$，Critic 更新参数是基于以下累积梯度进行的：

$$\mathrm{d}\theta_v\leftarrow\mathrm{d}\theta_v+\frac{\partial(R_t-V(s_t;\theta_v))^2}{\partial\theta_v'} \tag{25}$$

其中，θ_v' 表示 Critic 网络中特定线程的参数向量。

Actor 更新参数为：

$$\mathrm{d}\theta\leftarrow\mathrm{d}\theta+\nabla_{\theta'}\log\pi(a_t \mid s_t;\theta')(R_t-V(s_t;\theta_v))+\delta\nabla_{\theta'}H(\pi(s_t;\theta')) \tag{26}$$

其中，θ' 表示 Actor 网络中特定线程的参数向量。

Actor 和 Critic 的参数根据它们的累积梯度进行更新，如式(26)和(25)所示。RMSProp 算法用于估计 Actor 和 Critic 网络的梯度，计算公式为：

$$g=\alpha g+(1-\alpha)(\Delta\theta)^2 \tag{27}$$

其中，α 是动量，$\Delta\theta$ 表示 Actor 或 Critic 网络中损失函数的累积梯度。可以将估计的梯度 g 设置为共享值或分离值，但共享值往往更稳健。基于估计的梯度 g 来优化 A3C 的参数：

$$\theta\leftarrow\theta-\eta\frac{\Delta\theta}{\sqrt{g+\varepsilon}} \tag{28}$$

其中，η 表示学习率，ε 是一个很小的正数，用于在分母为 0 时避免发生错误。

本文中，基于 DRL 的资源分配方案的伪代码如算法 1 所示。

算法 1 基于多智能体 DRL 的联合资源分配解决方案

初始化：初始化全局参数向量 θ 和 θ_v 以及特定线程的参数向量 θ' 和 θ_v'；初始化全局共享计数器 $T=0$ 和线程特定计数器 $t\leftarrow1$；分别初始化计数器的最大值 T_{max}^{A3C} 和 t_{max}^{A3C}；

重复：

　初始化全局智能体的梯度：$\mathrm{d}\theta\leftarrow0$ 和 $\mathrm{d}\theta_v\leftarrow0$；

　用全局参数同步每个 worker 的参数：$\theta'=\theta$ 和 $\theta_v'=\theta_v$；

　设置 $t_{start}=t$，并获取观察到的系统状态 s_t；

　重复：

　　根据策略 $\pi(a_t \mid s_t;\theta')$ 选择一个动作 a_t；

　　获得一个奖励 r_t，并且得到新的观测状态 s_{t+1}；

　　$t\leftarrow t+1$

　　$T\leftarrow T+1$

　直到终端 s_t 或 $t-t_{start}=t_{max}^{A3C}$

$$R=\begin{cases}0, & s_t\text{ 为终端状态}\\ V(s_t;\theta_v'), & s_t\text{ 非终端状态}\end{cases}$$

　for $i\in\{t-1,\cdots,t_{start}\}$

$R \leftarrow r_i + \gamma R$

通过式(21)和(20)更新 $\mathrm{d}\theta$ 和 $\mathrm{d}\theta_v$

结束

通过式(23)更新参数 θ 和 θ_v

直到 $T > T_{\max}^{A3C}$

3 仿真结果及分析

在本节中提供仿真结果来验证所提出的基于 A3C 的联合资源分配算法的性能。本文所用的系统模型由 3GPP TR 36.885 的城市案例给出,每条街有四个车道,模型网格的尺寸为 216 m×125 m,区域面积为 649 m×375 m。在本文的系统中,基站位于区域中心,而 M 个蜂窝用户和 K 对 D2D 用户均匀分布在各个车道上。本文设置的最大回合数为 1 500,以每回合的最大时隙数 $T = 100$ 来运行训练模型,在这些时隙上,D2D 用户传输生成的数据包任务。仿真中的参数设置如表 2 所示。

表 2 仿真参数设置

参数	值
基站数量	1
资源块数量(B)	13
蜂窝用户数量(M)	12
D2D 用户数量(K)	6～18
蜂窝用户的发射功率(p_m)	23 dBm
D2D 最大发射功率(p_k^{max})	23 dBm
发射功率级(L)	4
带宽(W)	10 MHz
蜂窝用户的最小信噪比(γ_{th})	6 dB
AWGN(σ^2)	−114 dBm
路径损耗模型	128.1+37.6log(R(km))
阴影标准方差	3 dB
载波频率(f_c)	2 GHz
D2D 数据包大小(D)	[1,2,…,8]×1 060 bytes
D2D 数据包传输时间范围(T)	100 ms

对于实验,本文使用具有六个内核的 AMD 5600X 处理器。在 Python 3.8 环境中使用 TensorFlow 2.5.0 进行实验。根据不同的仿真设置,每个 D2D 用户传输的数据包大小为 1 060～8 480 bytes,D2D 用户总数为 6～18。蜂窝用户数量为 12,上行带宽和资源块分别为 10 MHz 和 13 个,并将其中的 1 个资源块用作专用通信。用户分别以 10～15 m/s 的随机速度在各个车道上移动。A3C 网络使用深度神经网络,Actor 网络由一个输入层、一个隐藏层和一个输出层组成,其激活函数分别为 ReLU、ReLU 和 Softmax,神经元个数分别为 $M+4$、120 和 B_RL。Critic 网络由一个输入层、一个隐藏层和一个输出层组成,其激活函数为 ReLU,神经元个数分别为 $M+4$、120 和 1。由于 Critic 网络评估 Actor 网络的动作,需要 Actor 网络的收敛速度低于 Critic 网络,因此本文将 Actor、Critic 网络的学习率分别设置为 0.000 1 和 0.01,折扣因子为 $\gamma = 0.99$。在训练阶段,将 D2D 用户传输的数据包大小和 D2D 用户总数分别固定为 1 060 bytes 和 10 个,在测试阶段分别通过数据包大小和用户总数的改变来验证所提方法的稳定性。

为了突出所提方法的优势,本文提供了以下不同算法的性能比较:1)随机方法,用户将随机分配资源块和功率;2)基于深度学习的 Q-learning 算法;3)Actor-Critic 算法;4)A3C 算法,即本文所提方案。

3.1 模式选择测试

首先进行 D2D 通信的模式选择,前面已提到使用 K 近邻查询算法来选择距离基站最近的 k 个 D2D 用户工作在专用通信模式。根据设置不同的 k 值,其数据传输失败率如图 3 所示。本文中专用频谱子带设置为 1。选择专用模式进行通信的 D2D 用户均复用此专用频谱子带,从图 3 中可以看出,当 k 值为 1 和 2 时,所有专用模式 D2D 用户在约束时间内都可以成功的传输数据,当 k 值大于 2 时,部分专用模式 D2D 用户传输数据失败。为了尽可能的减轻 D2D 用户对基站的干扰,因此,在仿真中设置 k 值为 2,并给出 D2D 用户的模式选择结果,如图 4 所示。

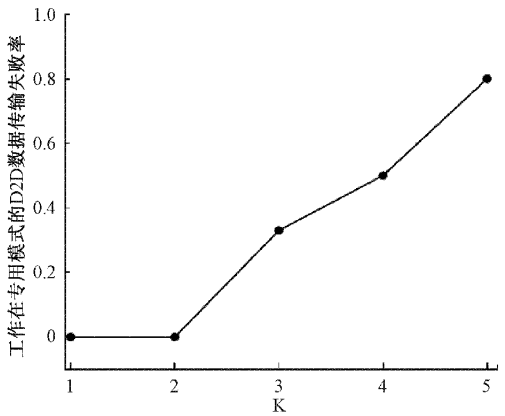

图 3 D2D 数据传输失败率与(专用模式)D2D 用户数量的关系图

3.2 资源分配测试

模式选择完成后,将选择复用模式的 D2D 用户作为智能体,使用基于 A3C 的联合资源分配算法来进行功率控制和资源块选择。

本文首先根据奖励,提供 3 种比较方法的收敛性能评估。从图 5 观察到,在 1 500 个训练回合内,除了 DQN 方法,其他算法均逐渐收敛到一定水平,但由于车辆的高移动性和策略探索,曲线有时会有所波动。本文所提方法比其他两种学习方法获得了更高的回报,并且在 390 回合左右就已经收敛。原因是所提方法采用多核的 CPU,不仅打

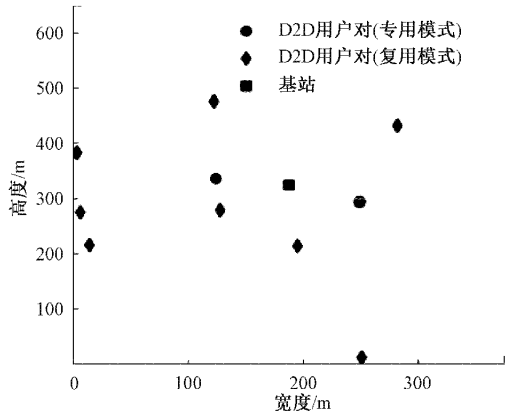

图 4　D2D 用户的模式选择分布图

破了数据的相关性，而且可以实现更快的收敛速度和更稳定的收敛效果。

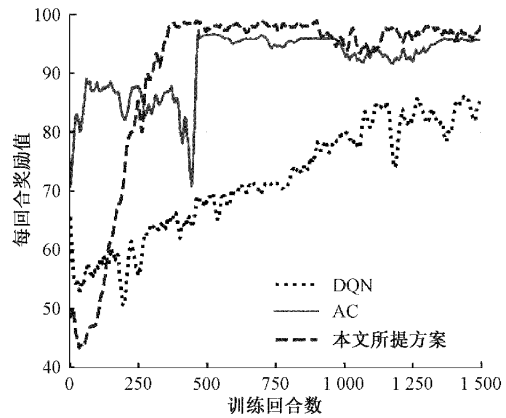

图 5　三种比较方法的收敛性能

图 6 和 7 分别显示了蜂窝用户和速率、D2D 用户平均速率与 D2D 所传输的数据包大小的关系。从图 6 中可以看出，随着 D2D 数据包的增加，蜂窝用户和速率不断减小。从图 7 中可以看出，随着 D2D 数据包的增加，D2D 的平均速率不断增加。这是由于数据包增大，D2D 用户为了实现用户间的数据包交付而需要更大的发射功率，这导致 D2D 的数据传输速率同样增大。相反的，D2B 链路因为受到的干扰增多，其性能则会呈现下降的趋势。从图 6 中可以发现，本文所提方案在不同的数据包大小上实现了比其他三种方法更好的性能。由于训练模型时采用的数据包大小为 1 060 bytes，通过测试时改变数据包的大小，从而证明了所提算法对任务数据量变化的稳定性。

在图 8 中，获得了不同 D2D 数据包下的数据包交付失败率。从图中可以看到，所有方法的数据包交付失败率都随着 D2D 数据包的增大而增大。但是，本文所提方案相比其它方法表现最好，而随机方法最差。这是因为随机算法随机选择资源块，可能会有较多的 D2D 用户占用同一频谱资源，从而导致各 D2D 用户间的干扰增强，那么数据包交付失败率也会随之增高。

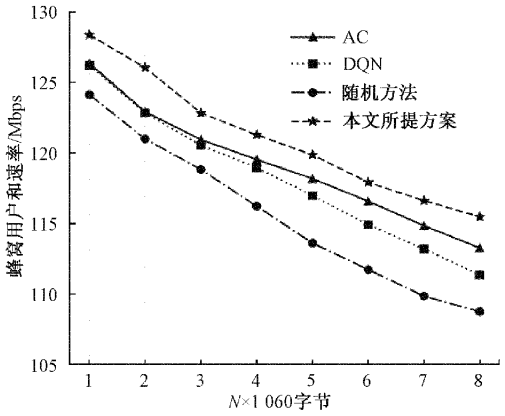

图 6　D2B 总速率与 D2D 数据包大小关系图

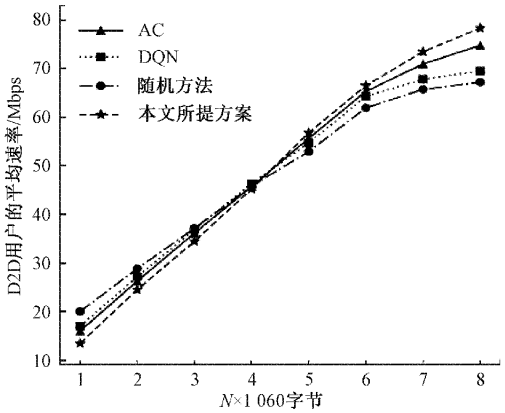

图 7　D2D 平均速率和 D2D 数据包大小关系图

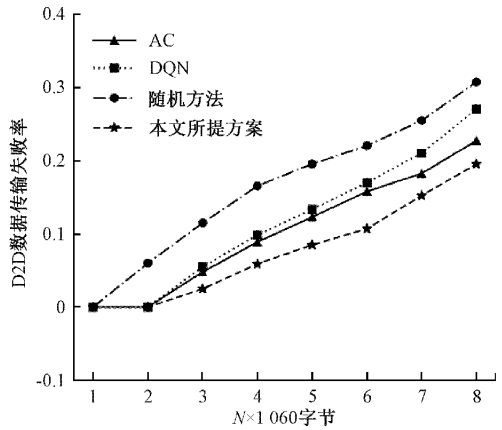

图 8　D2D 数据包交付失败率与 D2D 数据包大小关系图

将图 7 和 8 两张图进行对照，发现在数据包较小时，本文所提方案的 D2D 和速率相较于其他方法偏低，这是由于本文所提方法在时间约束的前几个时隙就以合适的速率将数据包成功交付，而其他方法需要更多的时隙去传输数据，这将导致累积的 D2D 速率增大。当数据包大于 5×1 060 bytes 时，本文所提方案的 D2D 和速率超过了其他方法。对应的，从图 8 可以观察到，当数据包大于 2×1 060 bytes 时，其他方法与所提方法相比，它们的交付失败率大大增加。由此不难得

出,由于交付失败率增加,更多的D2D用户在约束时间内一直在传输数据,但由于干扰增强导致每个时隙的D2D传输速率变的很低。在数据包较大时,本文所提方法还能以相对偏高的速率传输数据,表明所提方案的优越性。

图9描绘了当D2D用户传输的数据包为1 060 bytes时,不同数量的D2D用户的性能。可以观察到这4种方法的D2B性能随着D2D用户数量的增加而降低。随着更多的D2D用户导致更多的通信链路,所有链路都试图访问有限的频谱子带,因此在这种情况下D2B速率会下降。同时可以发现,当D2D用户数量适中(小于10)时,D2D用户交互成功的概率比较大,但一旦用户数量增加超过可接受的范围,交互失败率就会显著上升。出现这种现象的原因如下:严重的用户间干扰降低了通信质量,由于需要服务的通信链路数量较多,需要满足用户服务质量要求,所有的方法都无法在有限的频谱资源下完成大量的任务,从而导致链路传输成功的概率低。然而,在D2D用户数较多的场景下,本文所提方法仍在四种方法中取得了最好的性能。

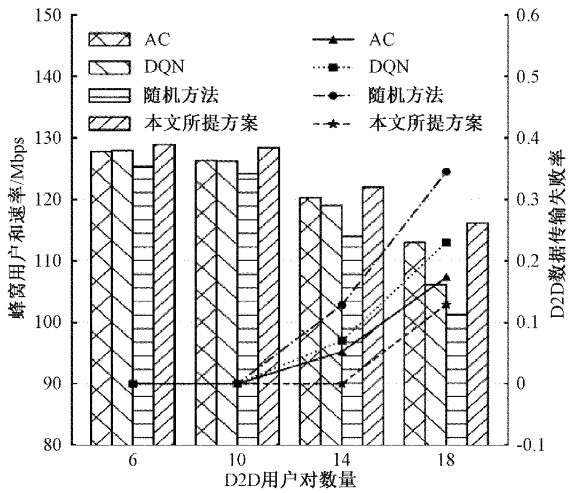

图9 蜂窝用户和速率、D2D数据包交付失败率与D2D用户数量关系图

4 结 论

在本文中,首先进行通信模式选择,再在复用模式的基础上联合资源分配以最大化蜂窝链路和速率,同时降低D2D链路传输失败率。具体来说,提出了一个A3C框架,以支持异步联合功率和资源块分配策略。仿真结果表明,所提出的A3C方法在学习效率和网络性能方面优于其它方法。未来,将考虑引入不同的算法分别获取连续的功率选择动作和离散的资源块选择动作,以更好的优化系统性能。

参考文献

[1] 李玉兵.未来移动通信系统中的D2D关键技术研究[D].成都:电子科技大学,2012.

[2] 尤肖虎,潘志文,高西奇,等.5G移动通信发展趋势与若干关键技术[J].中国科学:信息科学,2014(5):551-563.

[3] YU G, XU L, FENG D, et al. Joint mode selection and resource allocation for device-to-device communications [J]. IEEE Transactions on Communications, 2014, 62(11): 3814-3824.

[4] ZHAO K, SHENG M, LIU J, et al. Graph-based joint mode selection and resource allocation scheme for D2D and cellular hybrid network using SCMA[C]. 2016 8th International Conference on Wireless Communications & Signal Processing(WCSP), IEEE, 2016: 1-5.

[5] XU X D, ZHANG Y, SUN Z, et al. Analytical modeling of mode selection for moving D2D-enabled cellular networks[J]. IEEE Communications Letters, 2016, 20(6):1203-1206.

[6] LIU Z, PENG T, XIANG S, et al. Mode selection for device-to-device (D2D) communication under LTE-advanced networks [C]. 2012 IEEE International Conference on Communications(ICC), IEEE, 2012: 5563-5567.

[7] AKKARAJITSAKUL K, PHUNCHONGHARN P, HOSSAIN E, et al. Mode selection for energy-efficient D2D communications in LTE-advanced networks: A coalitional game approach[C]. 2012 IEEE International Conference on Communication Systems(ICCS), IEEE, 2012: 488-492.

[8] CHEN C Y, SUNG C A, CHEN H H. Capacity maximization based on optimal mode selection in multi-mode and multi-pair D2D communications[J]. IEEE Transactions on Vehicular Technology, 2019: 1-1.

[9] KAI C, LI H, XU L, et al. Joint subcarrier assignment with power allocation for sum rate maximization of D2D communications in wireless cellular networks[J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4748-4759.

[10] HUANG S, LIANG B, LI J. Distributed interference and delay aware design for D2D communication in large wireless networks with adaptive interference estimation [J]. IEEE Transactions on Wireless Communications, 2017, 16(6): 3924-3939.

[11] ZHOU Z, WANG B, GU B, et al. Time-dependent pricing for bandwidth slicing under information asymmetry and price discrimination [J]. IEEE Transactions on Communications, 2020, 68(11): 6975-6989.

[12] MNIH V, KAVUKCUOGLU K, SILVER D, et al.

Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[13] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search [J]. Nature, 2016, 529 (7587): 484-489.

[14] MAO H, ALIZADEH M, MENACHE I, et al. Resource management with deep reinforcement learning[C]. Proceedings of the 15th ACM Workshop on Hot Topics in Networks, 2016: 50-56.

[15] HE Y, ZHANG Z, YU F R, et al. Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks [J]. IEEE Transactions on Vehicular Technology, 2017, 66(11): 10433-10445.

[16] HE Y, YU F R, ZHAO N, et al. Secure social networks in 5G systems with mobile edge computing, caching, and device-to-device communications [J]. IEEE Wireless Communications, 2018, 25 (3): 103-109.

[17] 郑冰原，孙彦赞，吴雅婷，等. 基于深度强化学习的超密集网络资源分配[J]. 电子测量技术，2020(9):6.

[18] LIANG L, YE H, LI G Y. Spectrum sharing in vehicular networks based on multi-agent reinforcement learning [J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2282-2292.

[19] 赵家乐，赵凌霄，马丹，等. D2D 通信中联合模式选择和资源分配方案研究[J]. 现代电子技术，2019, 42(13): 33-37.

[20] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: A brief survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[21] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C]. International Conference on Machine Learning, PMLR, 2016: 1928-1937.

[22] GRONDMAN I, BUSONIU L, LOPES G A D, et al. A survey of actor-critic reinforcement learning: Standard and natural policy gradients [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C(Applications and Reviews), 2012, 42(6): 1291-1307.

作者简介

沈国丽，硕士研究生，主要研究方向为无线通信、深度强化学习方向。

E-mail:20201249231@nuist.edu.cn

李正权(通信作者)，教授，主要研究方向为无线通信、信号处理、信道编码译码方向。

E-mail:lzq722@jiangnan.edu.cn

李君，教授，主要研究方向为无线通信、资源分配、机器学习、编码译码等方向。

E-mail:07a0303105@cjlu.edu.cn