DOI:10.19651/j.cnki.emt.2106047

基于联邦深度强化学习的车联网资源分配*

王晓昌[1,2,3] 吴 璠[4] 孙彦赞[1,2,3] 吴雅婷[1,2,3]
(1.上海大学 上海先进通信与数据科学研究院 上海 200444;2.上海大学 特种光纤与光接入网重点实验室 上海 200444;
3.上海大学 特种光纤与先进通信国际合作联合实验室 上海 200444;4.上海大学经济学院 上海 200444)

摘 要:车辆通信(V2X)能够有效地提高交通安全性和移动性,是车辆部署场景中的关键技术之一。V2X 通信链路需要满足不同应用的服务质量(QoS)要求,如车对车(V2V)链路的延迟和可靠性要求。面向车辆高速移动性导致的无线信道快速变化,为保证不同车辆链路的 QoS 约束和车辆动态网络的鲁棒性,提出一种基于联邦深度强化学习(FDRL)的频谱分配和功率控制联合优化框架。框架首先根据不同车辆链路需求提出了对应的优化问题,并定义了强化学习的状态空间、动作空间和奖励函数;然后介绍了联邦深度强化学习的训练框架;最后,通过分布式的车辆端强化学习和基站聚合平均训练,找到最佳的频谱分配和功率控制策略。仿真结果表明,与其他对比算法相比,所提出算法能够提高车对基站(V2I)的总用户信道容量,并保证了新加入车辆时动态网络的鲁棒性。

关键词:车辆通信;深度强化学习;资源分配;联邦学习
中图分类号:TN929.5 **文献标识码**:A **国家标准学科分类代码**:510.5015

Internet of vehicles resource management based on federal deep reinforcement learning

Wang Xiaochang[1,2,3] Wu Fan[4] Sun Yanzan[1,2,3] Wu Yating[1,2,3]
(1. Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China;
2. Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, Shanghai 200444, China;
3. Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai University, Shanghai 200444, China; 4. School of Economics of Shanghai University, Shanghai 200444, China)

Abstract: Vehicle to everything (V2X) communication, which can effectively improve traffic safety and mobility, is one of the key technologies in vehicle deployment scenarios. V2X communication links need to meet different quality of service (QoS) requirements for different applications, such as the latency and reliability requirements of vehicle to vehicle (V2V) links. Considering rapid changes in wireless channels due to vehicles high mobility, while ensuring QoS constraints for different vehicle links and improving the robustness of dynamic networks, a joint optimization framework based on federal deep reinforcement learning (FDRL) for spectrum allocation and power control is proposed. The framework first proposes the corresponding optimization problems according to different vehicle link requirements, and defines the state space, action space and reward function for reinforcement learning. Then the joint deep learning reinforcement learning training framework is given. Finally, the optimal spectrum allocation and power control strategies are found by distributed vehicle-side reinforcement learning and base station aggregation averaging training. Simulation results show that the proposed framework can improve the total transmission rates of all the vehicle-to-infrastructure (V2I) users and guarantee the robustness of the network when new vehicles are connected to the network compared with other comparative algorithms.

Keywords: vehicle communication; deep reinforcement learning; resource allocation; federated learning

0 引 言

近年来,车辆通信(vehicle to everything,V2X)以其提高交通安全性和移动性的解决方案而受到业界和学术界的广泛关注[1]。在高移动性车辆网络中,存在着不同类型的连接,可以将其分为车对基站(vehicle to infrastructure,

收稿日期:2021-03-18
* 基金项目:国家重点研发计划(2019YFE0196600)、国家自然科学基金(61501289,61671011,61420106011,61701293)项目资助

V2I)链路和车对车(vehicle to vehicle,V2V)链路[2]。V2I链路为车辆与基站的通信链路,主要用以支持交通效率改善和提供信息娱乐服务,因此要求较大的信道容量[3]。另一方面,V2V链路为车车之间的直接通信链路,主要用于车辆行驶安全关键信息的传输,因此要求信息传输的超低时延和超高可靠性。在LTE V2V标准中,V2V链路共享使用V2I的上行链路进行直连通信,以提高网络的频谱使用效率,但同样会造成V2I链路和V2V链路之间的相互干扰问题[4]。

为了协调V2I链路和V2V链路之间的相互干扰,需要有效的无线电资源管理策略。文献[5]中提出一种干扰限制区域控制方案,以保护设备到设备(device to device,D2D)接收机免受蜂窝用户干扰。文献[6-7]考虑了多个蜂窝和D2D用户的频谱和功率分配设计,文献[8]提出了基于基站(base station,BS)控制的D2D功率控制算法,以实现D2D链路的信干噪比(signal interference plus noise ratio,SINR)最大化,同时控制蜂窝链路所受干扰在一定水平之下。文献[9]提出了一种三阶段的功率控制和频谱分配算法,以满足蜂窝用户和D2D用户的最小SINR约束要求,并最大限度地提高系统吞吐量。当前D2D网络中的干扰协调和资源管理方法多是基于已知的信道状态信息,然而V2X网络中车辆的高移动性导致无线信道的快速变化,此种方法已不再适用于网络拓扑高速动态变化的V2V网络。为了应对这种实时的动态资源分配,基于深度强化学习的无线资源分配算法受到了广泛的关注[10-11]。

强化学习(reinforcement learning,RL)在解决马尔可夫决策过程的问题时具有优势,目前广泛应用于各种通信决策过程。文献[12]提出了一种多智能体Q学习算法,用于解决蜂窝D2D网络的联合模式选择和功率自适应问题。文献[13]研究了多跳V2I通信,提出了一种基于Q学习的路由选择算法,以实现提高网络吞吐量和降低通信延迟。文献[14]提出了一种利用Q学习算法以支持D2D云无线接入网络的模式选择和子信道分配。上述研究多是面向网络节点不多的简单网络场景,但当网络中的节点数量众多时,基于传统RL的算法存在算法不稳定和计算效率较低的问题。文献[15]研究表明,联合利用神经网络的深度强化学习(deep reinforcement learning,DRL)算法有望降低计算复杂度和学习成本,并提高算法收敛速度。然而上述工作中所提出的DRL模型都是基于集中式的,车联网中严格的时延要求和车辆有限的本地训练数据给DRL模型的训练带来了巨大的挑战。此外,若DRL模型没能有效训练,将恶化新激活V2V对根据此模型参数做出的局部决策,并降低动态车辆网络的鲁棒性。为解决上述问题,本文提出了基于联邦深度强化学习(federal deep reinforcement learning,FDRL)的联合资源分配优化框架,同时考虑了不同链路的服务质量(quality of service,QoS)要求,以实现V2I链路的总容量最大化,并保证V2V对的可靠性要求。

本文贡献主要包括:首先综合考虑不同车辆链路的QoS需求,对基于频谱分配和功率控制的联合优化问题进行数学建模,以最大化V2I链路容量,并保证V2V链路可靠性。然后,针对所提优化问题中存在的二进制变量特性,提出了基于联邦深度强化学习框架的频谱分配和功率控制联合优化算法,并定义了强化学习中的状态、动作和奖励函数。最后,为提高训练稳定性,进一步引入了目标Q网络和经验重放策略,以提升车辆网络V2I链路总信道容量和提高动态车辆网络的鲁棒性。

1 系统模型及优化问题

1.1 系统模型

考虑由BS和多个VUE组成的车辆网络,如图1所示。其中蜂窝网络中BS沿高速公路部署,支持V2X通信的路边单元连接到每个基站,车辆沿直线公路行驶。本文将时间维划分为以 $t \in \{1,2,\cdots\}$ 为索引的时隙,其长度为LTE子帧长度,即1 ms。频域上一个子带和时域上的一个时隙长度构成一个可调度的最小资源块(resource block,RB)。定义V2I链路中对应的车辆为蜂窝用户设备(cellular user equipment,CUE),V2V链路对应的车辆为车辆用户设备(vehicle user equipment,VUE)。车辆网络中所有 M 个CUE组成的集合为 $\mathcal{M}=\{1,2,\cdots,M\}$,所有 K 对VUE组成的集合为 $\mathcal{K}=\{1,2,\cdots,K\}$。为简化分析,第 m 个CUE用户提前占用第 m 个RB进行通信。V2V链路包括1个单天线的V2V发射机(VUE Tx)和1个单天线的V2V接收机(VUE Rx)。为了提高频谱利用效率,V2V链路共享V2I的上行链路频谱。

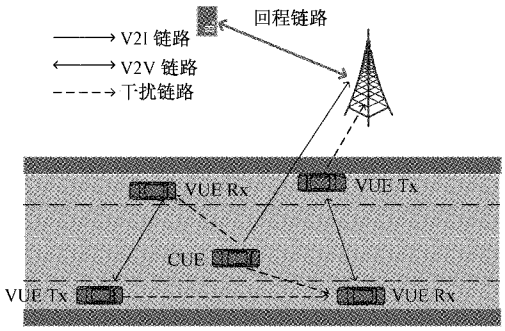

图1 车辆网络场景示意图

1.2 问题建模

假设在时隙 t,第 k 对V2V链路的信道增益 $h_k^d(t)$ 为:

$$h_k^d(t)=\alpha_k^d g_k^d(t),t=1,2,\cdots \quad (1)$$

其中,$g_k^d(t)$ 是信道小尺度衰落分量,并假设为单位均值指数分布。α_k^d 是信道大尺度衰落分量,且 $\alpha_k^d=G_k\beta_k^d(R_k^d)^{-\varphi_k}$,其中 G_k 是路损常数,$R_k^d=$

$\sqrt{(x_{Tx}-x_{Rx})^2+(y_{Tx}-y_{Rx})^2}$ 是第 k 对 VUE Tx 和 VUE Rx 之间的欧氏距离，φ_k 是功率衰减常数。同理，CUE m 到 BS 的信道增益，VUE k 到 BS 的干扰信道增益，CUE m 到 VUE k 的干扰信道增益，以及从 VUE k' 到 VUE k 的干扰信道增益可分别表示为 $h_m^c(t)=\alpha_m^c g_m^c(t)$，$h_{k,B}(t)=\alpha_{k,B}g_{k,B}(t)$，$h_{m,k}(t)=\alpha_{m,k}g_{m,k}(t)$，$h_{k',k}(t)=\alpha_{k',k}g_{k',k}(t)$，则第 m 个 V2I 链路的 SINR 可表示为：

$$\gamma_m^c(t)=\frac{p_m^c h_m^c(t)}{\sigma^2+\sum_{K}^{} s_{m,k}^t p_k^d h_{k,B}(t)} \tag{2}$$

其中，σ^2 为白噪声功率，p_m^c 和 p_k^d 分别表示 CUE m 和 VUE k 发射功率，$s_{m,k}^t$ 为频谱复用标志的二进制变量，其中 $s_{m,k}^t=1$ 表示 VUE k 复用了 CUE m 的频谱，反之，$s_{m,k}^t=0$。由香农公式计算，第 m 个 V2I 链路信道容量大小为：

$$R_m^c(t)=W\log_2(1+\gamma_m^c(t)) \tag{3}$$

其中，W 是传输信道带宽。

第 k 个 VUE 对的 SINR 可表示为：

$$\gamma_k^d(t)=\frac{p_k^d h_k^d(t)}{\sigma^2+I_k^d(t)} \tag{4}$$

其中，$I_k^d(t)=\sum_{m=1}^{M} s_{m,k,l}p_m^c h_{m,k}(t)+\sum_{k'\neq k} s_{m,k',l}p_{k'}^d h_{k',k}(t)$ 表示使用相同频谱的 V2I 链路和其余 V2V 对的干扰之和。

车辆通信中不同车载应用对通信链路 QoS 的要求也各不相同。CUE 承担带宽要求高的流量应用，因此定义 V2I 链路的 QoS 要求为保证通信链路的最小容量要求，可表示为：

$$R_m^c(t)\geqslant R_0^c \tag{5}$$

其中，R_0^c 表示 V2I 链路的最小容量需求。

同时，V2V 链路主要用于实时分发车辆行驶安全相关的关键消息，因此通信可靠性为其 QoS 的关键要求。本文通过控制中断事件的概率，以保证 V2V 链路的可靠性。定义中断事件为接收到的 SINR γ_k^d 低于预定阈值 γ_0^d，则通信可靠性要求可表示为：

$$\Pr\{\gamma_k^d\leqslant\gamma_0^d\}\leqslant p_0 \tag{6}$$

其中，γ_0^d 为 VUE 建立可靠链路所需的最小 SINR，$\Pr\{\cdot\}$ 为输入的概率，p_0 为 V2V 链路的最小可中断概率。参考文献[16]，瑞利衰落条件下，可进一步把可靠性约束转化为：

$$\gamma_k^d\geqslant\gamma_{th}=\frac{\gamma_0^d}{\ln\dfrac{1}{1-p_0}} \tag{7}$$

其中，γ_{th} 为有效中断阈值。

为了最大化 V2I 链路的总信道容量并确保 V2V 对的可靠性要求，本文的优化问题可建模为：

$$\begin{aligned}&\max_{\boldsymbol{P},\boldsymbol{S}}\sum_{m=1}^{M}R_m^c(t)\\&\text{s.t. } C1:R_m^c(t)\geqslant R_0^c,\forall m\in\mathcal{M},\forall t\\&\quad C2:\gamma_k^d\geqslant\gamma_{th},\forall k\in\mathcal{K},\forall t\\&\quad C3:\sum_{m=1}^{M}s_{m,k}^t\leqslant 1,\forall k\in\mathcal{K},\forall t\\&\quad C4:0\leqslant p_k^d\leqslant p_{\max}^d,\forall k\in\mathcal{K}\\&\quad C5:s_{m,k}^t\in\{0,1\},\forall m\in\mathcal{M},k\in\mathcal{K},\forall t\end{aligned} \tag{8}$$

其中，$\boldsymbol{P}$ 和 $\boldsymbol{S}$ 分别表示为 VUE 的发射功率和对应的频谱分配方案。$p_{\max}^d$ 为 VUE 的最大发射功率。约束 $C1$ 和 $C2$ 分别表示每个 CUE 和 VUE 的最小容量和可靠性要求。约束 $C3$ 表示 VUE 只能访问单个 CUE 的频谱。$C4$ 表示 VUE 的发射功率不能超过其最大限度。由于存在二进制变量，式(8)是一个混合整数非线性规划问题，传统的解决方式具有很高的计算复杂度。

2　基于 FDRL 的频谱和功率联合控制框架

本节中，首先将优化问题式(8)转化为马尔可夫决策过程(markov decision process，MDP)，同时使用深度强化学习解决 MDP 问题。然后描述了强化学习使用智能体不断与环境交互找到最优策略所对应的基本元素，包括状态空间、动作空间和奖励函数；最后，针对车联网中时延要求和训练数据不足的问题，提出了基于 FDRL 的频谱分配和功率控制联合优化框架。

2.1　强化学习的基本元素

强化学习(RL)主要用于解决可描述为马尔可夫决策过程的问题。在 RL 问题中，智能体可以周期性的学习采取行动 a_t，观察最大回报 r_t，并自动调整策略 π 以获得最优策略 π^*。将每个 V2V 链路视为一个智能体，多个智能体在与车联网环境的交互中进行学习。多智能体在竞争博弈的情况下，可达到局部最优，但不能满足整体网络性能最大化。为了达到优化问题的目标，将多智能体问题转化为合作博弈，对所有智能体使用相同的奖励函数。下面本文将提出的优化问题转化为 RL 问题，定义了车辆通信中多智能体资源分配问题的状态空间、动作空间和奖励函数，表示如下。

1) 状态空间：在每个时隙 t，每个 V2V 链路作为一个智能体，与其他的智能体进行交互选择不同的状态空间来进行表示当前环境。状态空间包含 V2V 在上一个时隙接收的干扰功率 $I_k^d(t-1)$，V2V 接收机到相应发射机和基站之间的信道增益 $h_k^d(t)$，V2I 链路对于 V2V 链路的干扰信道增益 $h_{m,k}(t)$，自身的发射功率 p_k^d 以及上一个时隙中的 RB 块所选用的次数 $N(t-1)$。因此在时隙 t 中，第 i 个 V2V 链路的状态空间 s_i^t 表示为：

$$\begin{aligned}s_i^t=\{&\{I_k^d(t-1),h_k^d(t),h_{m,k}(t),P_k^d\mid\forall m\in\mathcal{M},k\in\mathcal{K}\},\\&\{N(t-1)=\{N_1(t-1),\cdots,N_m(t-1)\mid\forall m\in\mathcal{M}\}\}\end{aligned} \tag{9}$$

2) 动作空间：每个 V2V 对智能体动作空间为联合优化对应的频谱分配 $s_{m,k}^t$ 和功率控制 p_k^d。其中本文采用离散功率控制方案，并假设 VUE 的发射功率具有 N_p 个级别。因此 V2V 对的动作空间大小为 $M\times N_p$，在时隙 t 中，

动作空间 a_i^t 可表示为:

$$a_i^t = \{\{s_{m,k}^t \in \{0,1\} \mid \forall m \in \mathcal{M}, k \in \mathcal{K}\},$$
$$\{p_k^d = \frac{p_{\max}^d f}{N_p - 1} \mid f \in \{0,1,\cdots,N_p-1\}, \forall k \in \mathcal{K}\}\} \quad (10)$$

3) 奖励函数:强化学习利用奖励函数来对应约束问题的优化目标,因此,在时隙 t 时,奖励函数 r_i^t 定义如下:

$$r_i^t = \begin{cases} \lambda_c \sum_m R_m^c(t) + \lambda_d \sum_k (\gamma_k^d - \gamma_{th}), & R_m^c(t) \geqslant R_0^c \\ \sum_m (R_m^c(t) - R_0^c), & \text{其他} \end{cases} \quad (11)$$

其中,当 V2I 链路满足最小信道容量 R_0^c 时,奖励函数对应为 $\lambda_c \sum_m r_m^c(t) + \lambda_d \sum_k (\gamma_k^d - \gamma_{th})$,其中第 1 项表示为 V2I 用户总信道容量,第 2 项为 V2V 链路的可靠性保证,λ_c 和 λ_d 用于平衡两部分对奖励函数的贡献。当未达到最小信道容量,奖励函数添加惩罚项 $\sum_m (R_m^c(t) - R_0^c)$,使其能够满足速率约束条件。

2.2 深度强化学习策略

在马尔可夫决策过程中,智能体以离散的时间步长不断地从与环境的交互中学习和做出决策。在每一个时隙 t 中,智能体观察环境的当前状态 $s_t \in \mathcal{S}$,并根据策略 π 选择并执行动作 $a_t \in \mathcal{A}$。在此基础上,智能体将从环境得到 $r_t = r(s_t, a_t) \in \mathcal{R}$,并根据环境 $p(s_{t+1} \mid s_t, a_t)$ 的转移概率到下一个状态 $s_{t+1} \in \mathcal{S}$。智能体的目标是找到使其获得的期望折扣奖励最大化的最优策略。在策略 π 下,长期奖励函数定义为从状态 s_t 开始的期望折扣奖励:

$$V^\pi(s_t, a_t) = \mathbb{E}_\pi[\sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i) \mid s_t, a_t] \quad (12)$$

其中,$T \to \infty$ 是所采取的总时间步数,$\gamma \in [0,1]$ 是折扣因子。$\sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)$ 为获得的长期折扣奖励,$r(s_t, a_t)$ 是即时奖励。

最优策略 $\pi^*(s)$ 可以通过最大化长期奖励 $V^{\pi*}(s_t, a_t)$ 获得。Q 学习中通过维护值函数表 $Q(s_t, a_t)$ 来表示长期奖励值。那么,最优策略 $\pi^*(s)$ 可以表示为:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s_t, a_t) \quad (13)$$

智能体将从与环境的实际交互中学习,并在接收其行为的结果时调整其行为,从而使预期的折扣奖励最大化。根据贝尔曼方程,最优的 Q 值更新状态动作函数为:

$$Q(s_t, a_t) = (1-\alpha) Q(s_t, a_t) + \alpha[r_t(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t-1})] \quad (14)$$

其中,α 是学习率,$\delta_t = r_t(s_t, a_t) + \gamma \max_{a_{t|1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$ 为时间差分(temporal difference, TD)误差。

DQN 利用由 θ 参数化的 DNN 来近似动作值函数 $Q(s, a; \theta) \approx Q(s, a)$。为了解决在 RL 中函数逼近的不稳定性问题,采用了一个有效大小的经验重放缓冲区 $\boldsymbol{D}$,它存储了智能体在每个时隙 t 的经验样本 (s_t, a_t, r_t, s_{t+1})。最久没有使用过的样本将被删除从而得到新样本的空间,以保证缓冲区 $\boldsymbol{D}$ 始终为最新。每次迭代中,Q 网络随机抽取一小批经验样本 $(s_t, a_t, r_t, s_{t+1}) \sim U(\boldsymbol{D})$ 来更新网络参数 θ 以最小化损失函数 $L(\theta)$:

$$L(\theta) = E[(y_t - Q(s_t, a_t \mid \theta))^2] \quad (15)$$

$$y_t = r_t(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1} \mid \theta') \quad (16)$$

DRL 算法为了进一步提高 RL 的稳定性,使用了目标网络。在训练过程中,Q 值会发生偏移。因此,如果使用一组不断移动的值来更新主网络参数,则值估计可能会失控。这导致了算法的不稳定。为了解决这个问题,目标网络被用来频繁但缓慢地更新主网络的参数值。这样,目标值和估计值之间的相关性显著降低,从而稳定了算法。在原神经网络中的参数 θ_i 使用软目标进行更新,其中 $\tau \ll 1$:

$$\theta_{i'} \leftarrow \tau \theta_i + (1-\tau) \theta_{i'} \quad (17)$$

利用目标网络,约束目标值缓慢变化,使 Q 值学习更接近于有监督学习。同时为了能够使目标值在训练过程中保证稳定,DRL 网络采用待训练网络进行参数训练循环一定次数才进行网络更新,使目标网络得到稳定。最后,智能体采用 ε 贪婪策略,意味着智能体以 ε 的概率任意选择动作 $a_t \in \mathcal{A}$,而以 $(1-\varepsilon)$ 的概率选择最佳的动作 $a_t = \arg\max_a Q(s_t, a; \theta)$,其中 ε 为探索因子。

2.3 聚合平均算法

FDRL 由两部分组成:1)V2V 对的分布式训练;2)BS 的聚合平均计算。对于参与深度强化学习的 V2V 对,其模型参数 θ_l 为局部模型,V2V 对在深度强化学习训练后,根据对应的位置坐标使用联邦平均学习更新到 V2V 对的平均局部模型参数中,然后将得到的全局网络反馈下载给整个成员 V2V 对。其中上传参数为每个 V2V 对的模型参数,表示为:

$$\theta_{g,t-1} = \frac{D^k}{D} \theta_{l,t+1}^m \quad (18)$$

其中,$\theta_{g,t-1}$ 和 $\theta_{l,t|1}^m$ 分别是 V2V 对的全局 Q 网络和局部 Q 网络的权重,D 代表样本的大小。

在训练过程中,基站将该区域内所有的 V2V 对网络模型参数进行聚合平均,完成后将所得的全局网络反馈下载给对应的 V2V 对。

本文提出的基于联邦 DRL 的频谱分配和功率控制的算法流程如算法 1 所示。

算法 1　FDRL 的频谱分配和功率控制算法

1. 初始化车辆网络状态，初始化系统状态，动作；
2. 初始化经验缓冲区 $\boldsymbol{D}$，初始化权重 θ_t 的在线网络 $Q(s_t,a_t\mid\theta_t)$；
3. **for** $episode = 1:M$：
4. 初始化网络，接收从 BS 端下载全局模型 θ_α；
5. **for** $t = 1:T$：
6. 智能体根据系统状态 s_t 利用 ε-贪婪策略选择动作 a_t；
7. 执行动作 a_t 调整 V2V 车辆的发射功率和 RB 选择，由式(11)得到奖励 r_t；
8. 智能体接收下一个状态 s_{t+1}；
9. 存储经验 (s_t,a_t,r_t,s_{t-1}) 到缓冲区；
10. **if** $\boldsymbol{D}$ 的容量大于 N：
11. 从经验缓冲区 $\boldsymbol{D}$ 中随机抽取批量经验样本 $\mathcal{B}$，其中经验 (s_i,a_i,r_i,s_{i+1}) 表示批量样本 $\mathcal{B}$ 的第 i 个经验样本；
12. y_i 根据式(16)计算而得；
13 通过式(15)最小化损失函数 $L(\theta)$，更新在线网络权重 θ_t；
14 根据位置坐标上传模型参数 θ_t 到 BS 端，根据式(18)进行聚合平均；
15. **end**
16. **end**

3　系统仿真与分析

本节对所提出的频谱分配和功率控制算法进行了仿真验证，仿真场景考虑的是一个高速公路车辆场景，仿真参数参考 3GPP TR. 36. 885 设置[17]。仿真场景为单个小区的双车道高速公路场景，其中基站位于中心，车辆的位置分布根据空间泊松过程产生。车辆发射功率采用离散化的 4 个等级分布，取值为{10，15，20，23}。DQN 网络由 3 个完全连接的隐藏层组成，每层神经元个数为{800，450，120}。RMSPro 优化器被用于更新网络参数，学习率为 0. 001。其余实验参数如表 1 所示。

表 1　仿真参数

参数	取值
载波频率/GHz	2
信道带宽/M	10
基站天线高度/m	25
基站信道增益/dBi	8
车辆最大发射功率/dBm	23
车辆天线增益/dBi	3
网络大小/m×m	750×500
道路宽度/m	4

为了更好地说明所提算法的性能，本文将联邦 DRL 算法与 Q 学习算法、车辆分组算法[18]和随机算法进行了对比分析。

图 2 所示为 V2I 总用户速率与不同 V2V 车辆对数的关系。由图 2 可知，随着 V2V 车辆对数的增加，4 种算法对应的 V2I 总用户速率都逐渐降低。这是由于随着 V2V 车辆对数增加，越来越多的 V2V 链路共享 V2I 的频谱，由于 V2V 链路产生的干扰量的增加，V2I 总和速率会降低，使得整体网络性能变差。同时，随着车辆对数的增加，所提出的算法降低趋势变为平缓，这是因为随着智能体与环境的交互，使得增加 V2V 车辆对能够选择较好的方案。由仿真结果可以看出，随着网络环境的变化，与其他算法相比，所提出算法车辆网络根据不同链路的 QoS 具有较好的自适应能力，同时具有更好的网络性能。

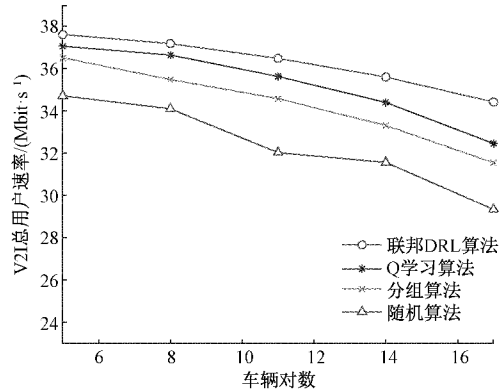

图 2　V2I 总用户速率与 V2V 车辆对数的关系

图 3 所示为车辆网络场景在新加入 V2V 对两种算法的训练过程对比。车辆场景新加入 V2V 对后，需要对以前的网络模型进行重新训练。由图 3 可知，新加入 V2V 车辆对后，不同模型在多次循环迭代过程中网络场景性能逐渐达到稳定。联邦 DRL 算法将预先训练好的网络模型下载到新加入的 V2V 车辆对中，使车辆对选择最佳动作。在循环迭代次数大约为 60 时，V2I 总用户速率最大且达到稳定。图 3 同时对比了无迁移学习的 DRL 算法。在联邦 DRL 模型的基础上对新加入的 V2V 对进行局部训练，使

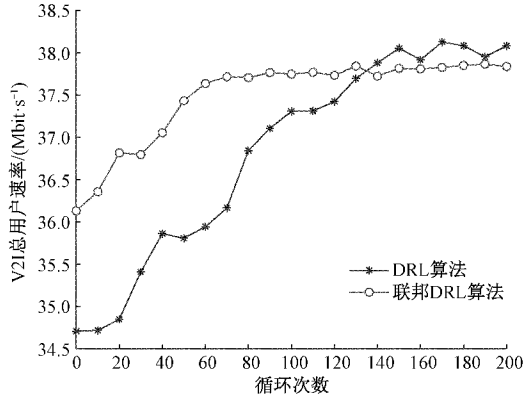

图 3　V2I 总用户速率与新加入 V2V 车辆对的关系

得不同 V2V 对的训练数据更多地被考虑,对应的全局模型对车辆环境更加鲁棒。从收敛结果后可以看出,联邦 DRL 算法性能稳定,而 DRL 算法性能波动较大。

图 4 为本文所提联邦 DRL 算法与典型的 Q 学习算法的收敛速度对比结果。从图 4 中可以看出,随着迭代次数的增加,本文所提算法在训练前期由于需要聚合不同网络模型的参数,性能低于 Q 学习算法。后期由于模型参数的平均和采用,本文所提算法不仅具有更快的收敛速度,同时能取得更高的网络性能,同时在收敛后的网络性能更加稳定。

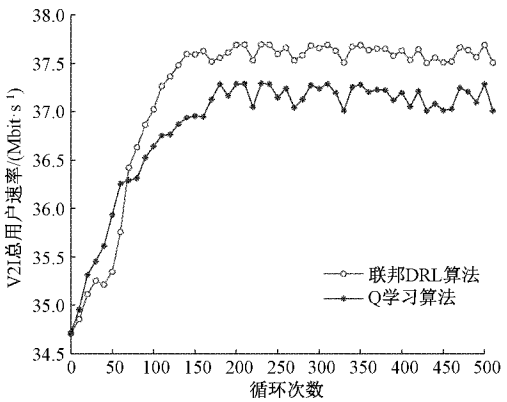

图 4 算法的收敛仿真图

4 结 论

在车辆网络环境中,为了满足各种链路的不同 QoS 要求,同时提高对于动态车辆环境的鲁棒性,本文提出了一种基于联邦深度强化学习的频谱分配和功率控制方案。在满足 V2V 链路可靠性要求的前提下,联合频谱分配和功率控制,以提高 V2I 链路的总用户速率。每辆车辆都可以根据观测状态进行分布式资源管理,都可根据位置坐标将各自网络参数上传到 BS 进行聚合平均计算,然后下载到各 V2V 对进行参数训练。仿真结果表明,本文提出的算法能够有效地最大化 V2I 链路的总信道容量。同时,对于新加入车辆,网络性能能够在较小的训练次数中达到收敛,且鲁棒性更好。

参考文献

[1] ARANITI G, CAMPOLO C, CONDOLUCI M, et al. LTE for vehicular networking: A survey[J]. IEEE Communications Magazine, 2013, 51(5): 148-157.

[2] KARAGIANNIS G, ALTINTAS O, EKICI E, et al. Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions [J]. IEEE Communications Surveys & Tutorials, 2011, 13(4): 584-616.

[3] CHENG H T, SHAN H, ZHUANG W. Infotainment and road safety service support in vehicular networking: From a communication perspective[J]. Mechanical Systems and Signal Processing, 2011, 25(6): 2020-2038.

[4] LU N, CHENG N, ZHANG N, et al. Connected vehicles: Solutions and challenges[J]. IEEE Internet of Things Journal, 2014, 1(4): 289-299.

[5] LIU P, WANG C, FU T, et al. Exploiting opportunistic coding in throwbox-based multicast in vehicular delay tolerant networks[J]. IEEE Access, 2019, 7: 48459-48469.

[6] WEN Q, HU B J, ZHENG L. Outage-constrained device-to-device links reuse maximization and its application in platooning [J]. IEEE Wireless Communications Letters, 2019, 8(6): 1635-1638.

[7] MIN H, LEE J, PARK S, et al. Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks [J]. IEEE Transactions on Wireless Communications, 2011, 10(12): 3995-4000.

[8] JANIS P, KOIVUNEN V, RIBEIRO C, et al. Interference-aware resource allocation for device-to-device radio underlaying cellular networks[C]. VTC Spring 2009-IEEE 69th Vehicular Technology Conference. IEEE, 2009: 1-5.

[9] FENG D, LU L, YUAN-WU Y, et al. Device-to-device communications underlaying cellular networks[J]. IEEE Transactions on Communications, 2013, 61(8): 3541-3551.

[10] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[11] 廖晓闽,严少虎,石嘉,等. 基于深度强化学习的蜂窝网资源分配算法[J]. 通信学报, 2019, 40(2): 11-18.

[12] LIANG L, YE H, LI G Y. Multi-agent reinforcement learning for spectrum sharing in vehicular networks[C]. 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), IEEE, 2019: 1-5.

[13] WU C, YOSHINAGA T, JI Y, et al. Computational intelligence inspired data delivery for vehicle-to-roadside communications[J]. IEEE Transactions on Vehicular Technology, 2018, 67(12): 12038-12048.

[14] 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.

[15] BUONIU L, BABUŠKA R, DE SCHUTTER B.

Multi-agent reinforcement learning: An overview[J]. Innovations in Multi-agent Systems and Applications-1, 2010: 183-221.

[16] LIANG L, XIE S, LI G Y, et al. Graph-based resource sharing in vehicular communication[J]. IEEE Transactions on Wireless Communications, 2018, 17(7): 4579-4592.

[17] 3GPP. TR. 36. 885. Technical specification group radio access network: Study on LTE-based V2X services (Release 14)[S]. Jun, 2016.

[18] ASHRAF M I, BENNIS M, PERFECTO C, et al. Dynamic proximity-aware resource allocation in vehicle-to-vehicle(V2V) communications[C]. 2016 IEEE Globecom Workshops(GC Wkshps), IEEE, 2016: 1-6.

作者简介

王晓昌，硕士研究生，主要研究方向为车辆网络、资源管理。
E-mail: cmailforwang@shu. edu. cn

吴璠(通信作者)，在站博士后，主要研究方向为资源与能源经济学、优化理论建模、机制设计。
E-mail: fanwu6@shu. edu. cn

孙彦赞，副教授，主要研究方向为无线通信资源管理、干扰协调、绿色通信。
E-mail: yanzansun@shu. edu. cn

吴雅婷，副教授，主要研究方向为无线通信 OFDM、MIMO 系统资源管理。
E-mail: yt-wu@shu. edu. cn