基于多智能体深度强化学习的车联网频谱共享

王为念[1],苏 健[1*],陈 勇[2],张建照[2],唐 震[1]
(1. 南京信息工程大学计算机与软件学院,江苏南京 210044;2. 国防科技大学第六十三研究所,江苏南京 210007)

摘 要: 针对高动态车联网环境中基站难以收集和管理瞬时信道状态信息的问题,提出了基于多智能体深度强化学习的车联网频谱分配算法. 该算法以车辆通信延迟和可靠性约束条件下最大化网络吞吐量为目标,利用学习算法改进频谱和功率分配策略. 首先通过改进DQN模型和Exp3策略训练隐式协作智能体. 其次,利用迟滞性Q学习和并发体验重放轨迹解决多智能体并发学习引起的非平稳性问题. 仿真结果表明,该算法有效载荷平均成功交付率可达95.89%,比随机基线算法提高了16.48%,可快速获取近似最优解,在降低车联网通信系统信令开销方面具有显著优势.

关键词: 车联网;分布式频谱共享;多智能体;深度强化学习
基金项目: 国家自然科学基金(No.61802196,No.62131005)
中图分类号: TP393.1 **文献标识码:** A **文章编号:** 0372-2112(2024)05-1690-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220320

Multi-Agent Reinforcement Learning Enabled Spectrum Sharing for Vehicular Networks

WANG Wei-nian[1], SU Jian[1*], CHEN Yong[2], ZHANG Jian-zhao[2], TANG Zhen[1]
(1. *School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, Jiangsu* 210044, *China*;
2. *The* 63*rd Research Institute, National University of Defense Technology, Nanjing, Jiangsu* 210007, *China*)

Abstract: Aiming at the problem that it is difficult for base stations to collect and manage instantaneous channel state information in high dynamic vehicle networking environment, a spectrum allocation algorithm for vehicle networking based on multi-agent deep reinforcement learning is proposed. The algorithm aims to maximize the network throughput under the constraints of vehicle communication delay and reliability, and uses the learning algorithm to improve the spectrum and power allocation strategy. Firstly, the implicit cooperative agent is trained by improving DQN model and EXP3 strategy. Secondly, the nonstationary problem caused by multi-agent concurrent learning is solved by using hysteretic Q-learning and concurrent experience replay trajectory. The simulation results show that the average successful delivery rate of the payload of the proposed algorithm can reach 95.89%, which is 16.48% higher than the random baseline algorithm. It can quickly obtain the approximate optimal solution, and has significant advantages in reducing the signaling overhead of the Internet of vehicles communication system.

Key words: Vehicular network; Distributed spectrum sharing; Multi agent; Deep reinforcement learning
Foundation Item(s): National Natural Science Foundation of China (No.61802196, No.62131005)

1 引言

车联网(Internet of Vehicles,IoV)是智能交通系统在物联网的典型应用[1]. 自2000年开始,学术界提出车到万物(Vehicle to Everything,V2X)的概念,将车联网拓展为支持车辆与一切道路实体间进行通信的网络,能够支持多样的安全和娱乐服务[2]. V2X两种典型的操作模式是车间通信(Vehicle to Vehicle,V2V)和车到基础设施通信(Vehicle to Infrastructure,V2I). V2V承载可靠性约束的安全信息传输,V2I面向高速率大容量需求的娱乐相关应用服务. 随着车联网业务类型多样化发展,服务质量(Qualify of Service,QoS)需求日益增加,车联网中通信道干扰也愈发严重. 在具有高动态网络拓扑、

收稿日期:2022-03-28;修回日期:2022-07-13;责任编辑:王天慧
*通讯作者:苏健

高移动性通信车辆节点等特性的车联网环境下,如何合理高效地对有限频谱资源进行分配,降低同信道干扰问题的影响,提升系统吞吐量和服务质量是车载通信架构下的重要挑战. 车联网资源共享研究中传统的资源优化方案居多,但面临城市交通环境信道状态信息(Channel Status Information,CSI)不准确等问题,往往难以实现优化目标. 文献[3~5]已经证明了深度强化学习(Deep Reinforcement Learning,DRL)在解决资源分配问题的潜力. 近年来,利用DRL开发车联网频谱资源共享的学习算法受到学界和业界广泛关注.

车联网无线信道具有快时变、非平稳的衰落特性,对资源分配有更高的实时性和准确性的要求,因此基于DRL的车联网资源分配多是采用基于DQN的频谱分配方法. 例如,针对单播和广播场景下V2V用户延时保障问题,文献[6]提出一种基于DRL的分布式车辆网络资源分配机制. 针对C-V2X通信的传输模式选择和资源分配的联合问题,文献[7]提出了一种基于DRL的分布式算法,以同时满足V2I和V2V用户的QoS要求. 但上述算法是针对静态环境,并未充分考虑无线网络环境动态变化引起的高方差和奖励值估计,存在算法鲁棒性不足的问题. 文献[8]利用SAC强化学习理论建立神经网络,以熵最大化和累计奖励和最大化为目标训练智能体,使得V2V用户能够获得较优的频谱分配决策. 文献[9]为减少网络信令开销,车辆用户使用一个深度神经网络压缩其观察信息,这些信息随后被反馈给集中决策单元并采用深度Q网络分配资源. 上述两种方法是基于全局信息分配信道资源,而在复杂时变的车联网环境中全局信息获取与维护开销巨大. 此外,文献[6~9]都属于传统的单智能体强化学习算法(Single-Agent Reinforcement Learning,SARL),很难适用于多用户的车联网场景. 如果将SARL扩展到多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)设置,多智能体并发探索引起的非平稳性会显著阻碍训练并降低性能[10]. 为了解决非平稳性问题,直接从多智能体的角度来研究车辆网络中的信道接入问题. 文献[11]利用DQN理论提出了一种分布式算法来优化车辆网络中的频谱和功率分配,并提出了基于指纹的重放缓冲区来解决非平稳性问题,但该算法没有考虑环境动态变化的影响,导致性能较差. 文献[12]研究了多智能体传动系统中车辆的最优访问控制,并提出了一种将统计学习方法和动态规划技术相结合的分布式访问算法,由于该算法使用了基于表格的动态规划方法,因此状态必须被量化为离散水平,限制了其在高维问题中的适用性. 此外,文献[13]利用多智能体深度确定性策略梯度方法建模和处理非正交多址技术条件下V2I用户和V2V用户频谱资源分配问题,使得V2I用户和速率最大化且同时满足V2V通信严格的延迟和可靠性约束,但采用的是确定性策略,其算法稳定性较差. 文献[14]提出利用双对抗深度循环Q网络和公共奖励训练隐式协作智能体的模型,缓解环境动态变化引起的不稳定奖励估计问题,但对车辆速度的鲁棒性较差. 文献[12~14]虽然在不同方面取得了较好的优化效果,但依旧存在信令开销较大、算法健壮性不足等问题.

针对上述问题,为进一步提升复杂电磁环境下的频谱共享效率,本文提出了一种基于多智能体深度强化学习的V2X频谱分配算法. 具体贡献如下.

(1)针对高速移动环境下频谱高效利用需求的挑战,提出了一种改进DQN模型,利用中长短时记忆网络LSTM和决斗网络架构实现高效的特征表示和价值近似.

(2)针对多智能体深度强化学习的非平稳性问题,利用迟滞Q学习方法训练其他智能体的负并发行为,结合并发体验重放轨迹CERT,同步多智能体训练过程,从而有效协调多智能体学习. 此外,采用Exp3策略进行动作选择,更好估计每个动作的奖励.

(3)通过仿真实验验证该算法收敛性和收敛效果,且所提算法在信道总容量和有效载荷交付率等方面优于现有代表性算法,能够高效利用车联网频谱资源完成更多的通信任务.

2　系统模型与问题描述

车联网通信场景如图1所示,包括单个基站(BS)和多个车辆用户. 车辆用户根据其通信需求的不同分为V2I用户和V2V用户,V2I用户需要大链路容量以支持面向信息娱乐的应用,V2V用户需要高链路可靠性传输安全相关信息,V2I用户和V2V用户的集合分别表示为$M=\{1,2,\cdots,m\}$和$K=\{1,2,\cdots,k\}$.

为保证V2I用户高质量的传输,主要关注车联网上行信道. 车辆通信网络中上行信道资源利用率不高,且BS端干扰更易于管理,V2V用户可复用V2I用户的上行信道资源,以提高子信道的利用效率. 且假定V2I用户预先分配好具有固定传输速率的正交信道$S=\{1,2,\cdots,s\}$,其中s表示子信道的数量.

为V2V用户设计有效的频谱共享方案,使V2I和V2V用户在高动态网络环境下,以最小信令开销实现各自目标. 假定信道衰落在一个子信道内大致相同,且在不同子信道之间是独立的,则在时隙t内,V2V用户对k在信道s的信道功率增益为

$$g_{k,k}^{s,t}=a_k^{s,t}h_{k,k}^{s,t} \tag{1}$$

其中,$a_k^{s,t}$表示与频率无关的大尺度衰落效应,即阴影效应和路径损耗;$h_{k,k}^{s,t}$表示与频率相关的小尺度衰落信道

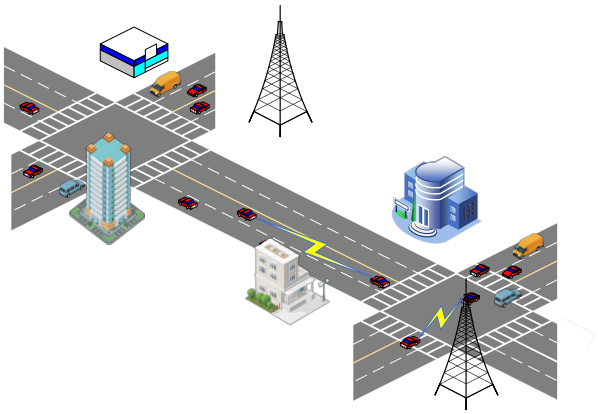

图1 车联网通信场景

增益. 在时隙t内,V2V用户k对V2V用户k'在信道s的干扰信道增益$g_{k,k'}^{s,t}$,V2V用户k对BS在信道s的干扰信道增益$g_{k,B}^{s,t}$,V2I用户m对BS在信道s的通道增益$g_{m,B}^{s,t}$,以及V2I用户m到V2V用户k在信道s的干扰信道增益$g_{m,k}^{s,t}$等均可类似定义.

考虑V2X用户不同的传输需求,将正交信道进行动态分配,采用二进制指示变量$\alpha_{m,s}$和$\beta_{k,s}$分别描述V2I用户和V2V用户是否占用子信道s,即

$$\alpha_{m,s},\beta_{k,s}=\begin{cases}1,\text{信道分配给用户}\\0,\text{其他}\end{cases} \tag{2}$$

假定每个用户只能同时占用一个信道,且一个信道最多分配给一个V2I用户,多个V2V用户可以共享一个信道. 于是,在时隙t内,V2I用户m在信道s的信噪比(Signal to Interference plus Noise Ratios,SINRs)为

$$\gamma_{m,s,t}^{M}=\frac{\alpha_{m,s}P_m^M g_{m,B}^{s,t}}{\sigma^2+\sum\limits_{k\in K}\beta_{k,s}P_k^K g_{k,B}^{s,t}} \tag{3}$$

与此同时,在时隙t内,V2V用户k在信道s的SINRs为

$$\gamma_{k,s,t}^{K}=\frac{\beta_{k,s}P_k^K g_{k,k}^{s,t}}{\sigma^2+I_{m,k}} \tag{4}$$

其中,P_m^M和P_k^K分别表示V2I用户m和V2V用户k的传输功率,σ^2表示噪声功率,以及

$$I_{m,k}=\sum_{m\in M}\alpha_{m,s}P_m^M g_{m,k}^{s,t}+\sum_{k'\in K,k'\neq k}\beta_{k,s}P_k^K g_{k,k'}^{s,t} \tag{5}$$

表示V2V用户k对所有信道的干扰功率.

根据香农定理,获得V2I用户m和V2V用户k在信道s时隙t的数据速率分别为

$$R_{m,s,t}^{M}=W\log\left(1+\gamma_{m,s,t}^{M}\right) \tag{6}$$

和

$$R_{k,s,t}^{K}=a_{k,s,t}W\log\left(1+\gamma_{k,s,t}^{K}\right) \tag{7}$$

其中,W表示每个子频带的带宽. 此外,引入矩阵$\boldsymbol{A}=\left\{a_{k,s,t}|k\in K\right\}$指示V2V用户链路是否可靠. $a_{k,s,t}$为V2V用户k在数据传输的可靠性能指标,即

$$a_{k,s,t}=\begin{cases}1, & \gamma_{k,s,t}^{K}\geqslant\gamma_0^K\\0, & \gamma_{k,s,t}^{K}<\gamma_0^K\end{cases} \tag{8}$$

其中,γ_0^K表示V2V用户的SINRs门限值.

如前所述,V2I用户被设计为支持移动高数据速率的娱乐服务,为满足Qos需求将V2I用户的吞吐量需求定义为

$$C_M=\sum_{m\in M}\sum_{s\in S}\sum_{t\in T}\alpha_{m,s,t}R_{m,s,t}^{M}\geqslant C_{\min}^{M} \tag{9}$$

其中,$C_{\min}^{M}$表示可容忍的最小吞吐量.

与此同时,V2V用户负责可靠地传播安全关键信息,即延时和可靠性需求. 这些信息根据车辆机动性以不同频率定期生成,建模为在时间预算T内成功交付大小为B的数据包,即

$$\Pr\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\beta_{k,s,t}R_{k,s,t}^{K}\geqslant B_k/\Delta T\right\},k\in K \tag{10}$$

其中,B_k表示周期性生成V2V用户有效载荷的大小,$\Delta T=T_{\max}-\left(t\bmod T_{\max}\right)$表示交付时间,假定生成周期等于可容忍延时$T_{\max}$.$\left(t\bmod T_{\max}\right)$表示生成最新消息所需时间.

因此,系统整体目标是V2V用户如何合理选择自身传输参数,即占用的子信道和采用的传输功率,从而提高V2I用户和V2V用户的通信性能. 基于此,车联网频谱资源分配问题定义如下:

$$\begin{aligned}&\max_{P,L}\ C_M\\&\text{s.t.}\quad \text{C1, C2: (9),(10)}\\&\qquad \text{C3: }\sum_{s\in S}\beta_{k,s}\leqslant 1,\beta_{k,s}\in\{0,1\},\forall k\in K\\&\qquad \text{C4: }\sum_{s\in S}\alpha_{m,s}\leqslant 1,\alpha_{m,s}\in\{0,1\},\forall m\in M\\&\qquad \text{C5: }P_k^K\leqslant P_{\max}^K,\forall k\in K\end{aligned} \tag{11}$$

其中,$P=\left\{P_k^K|k\in K\right\}$表示V2V用户传输功率集. $L=\left\{\beta_{k,s}|k\in K,s\in S\right\}$表示信道选择指标集. 约束C1和C2表示多样性Qos需求;约束C3和C4表示每个V2X用户同时只能占用一个信道. C5表示V2V用户可以使用的最大传输功率.

3 频谱共享方案设计

针对车联网频谱资源分配全局优化中CSI收集维护开销巨大的问题,采用分布式V2X资源分配机制. 分布式架构下面临的主要挑战是如何协调多个V2V用户行动,使其不会为提升各自性能而损害整个系统的性能. 此外,在式(10)中定义的V2V用户数据包传输率涉及在时间约束T内跨多个相干时隙进行顺序决策,指数级复杂性给传统的优化方法带来了困难.

为应对上述挑战,利用多智能体深度强化学习机制,建模为部分可观察马尔科夫决策. 多智能体深度强化学习模型能有效建模高动态网络环境下序贯决策问题,多个智能体通过与复杂未知环境不断交互以试错寻求累积奖励最大的频谱选择策略,改进频谱分配和功率控制.

3.1　多智能体环境建模

在基于多智能体深度强化学习的频谱资源共享方案中,V2V用户作为一个智能体探索未知环境. 建立部分可观察马尔科夫决策模型,如图2所示,每个时隙t内,给定当前环境状态S_t,V2V用户k接收环境的观察$z_k(t)$确定为$z_k(t)=O(S_t,k)$,采取动作$a_k(t)$,形成联合动作A_t. 此后,智能体接收奖励R_{t+1},环境演化到概率为p的下一个状态S_{t+1}. 与此同时,每个智能体都会接收到新的观察值$z_k(t+1)$. 其中,智能体k的策略π依赖于局部观察.

3.1.1　状态和观测空间

环境状态S_t包括全局信道条件和所有智能体的行为. V2V用户只能通过观察获得底层环境的状态,观测空间包含本地信道信息$G_{m,k,s}$和V2V用户k对所有信道的干扰功率$I_{m,k}$. 综上所述,相关信道的观测空间为

$$O(S_t,k)=\left\{\left\{I_{m,k}\right\},\left\{G_{m,k,s}\right\}\right\},\quad m\in M,k\in K,s\in S \tag{12}$$

其中,

$$G_{m,k,s}=\left\{g_{k,k}^{s,t},g_{k,B}^{s,t},g_{k,k'}^{s,t},g_{m,k}^{s,t}\right\} \tag{13}$$

然而,针对车辆网络中获取与维护CSI开销巨大的问题,改用V2V用户k对所有信道的干扰功率$I_{m,k}$,剩余的V2V有效载荷B_k和剩余的时间预算T_k描述观测空间. 此外,当V2V用户的数量增加时,在观测空间中添加不同用户标识有助于智能体学习不同的策略. 因此,观测空间的维度为$4\times K$. 其中,V2V用户k的观察函数为

$$z_k(t)=\left\{I_{m,k},B_k,T_k,k\right\} \tag{14}$$

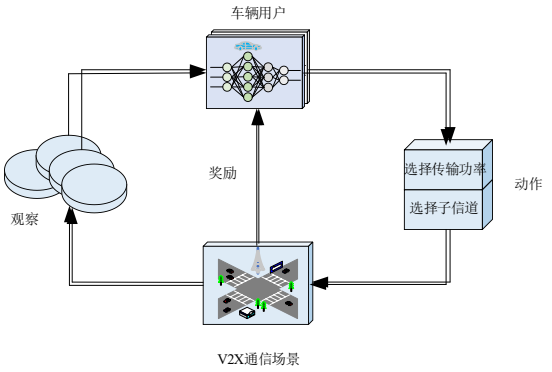

图2　多智能体车联网部分可观察马尔科夫决策过程模型

3.1.2　动作空间

在全局优化问题中,V2V用户优化子信道和传输功率选择的联合动作,所有V2V用户都具有相同的动作空间A. 具体来说,V2V用户重用了V2I用户所占据的子信道,V2V用户的可用信道集对应于S. 将可用的传输功率空间离散为多个层次(仿真中采用$\{23,10,5,-100\}$). 因此,动作空间的维度为$4\times S$,表示为

$$a_k(t)=\left\{(s,p)|s\in S,p\in A_p\right\} \tag{15}$$

3.1.3　奖励值设计

如图3所示,智能体根据观察到的环境状态采取行动,环境将立即返回智能体一个奖励. 然后在学习阶段,智能体根据收到的奖励更新资源分配策略,直到算法收敛. 奖励函数根据式(11)表示的优化问题来设计. V2X频谱共享优化目标有两个:最大化V2I容量,同时在V2I时间约束内增加V2V有效负载交付的成功概率. 因此,为了最大限度地增加V2X用户在t时隙满足QoS要求完成的任务数量,定义以下两个奖励元素. 由式(9)可知V2I用户的总容量为C_M. 此外,将每个V2V用户k在时隙t有效传输速率大小设置成奖励值,并且当所有载荷传输完成后该奖励值又被设置为常数ξ,表示为

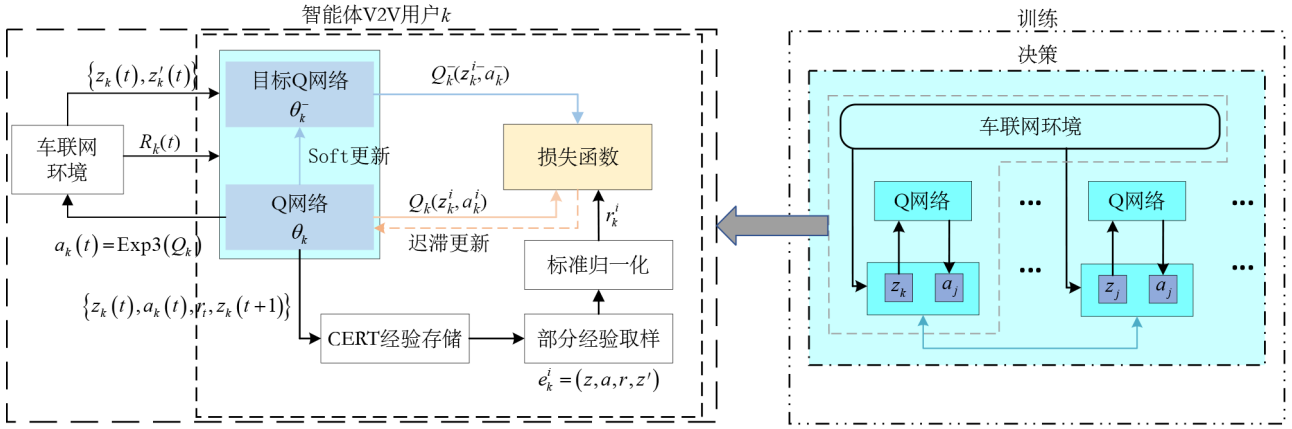

图3　算法整体框架

$$F_K=\begin{cases}\sum_{k\in K}\sum_{s\in S}\sum_{t\in T}\beta_{k,s,t}R^K_{k,s,t}, & \text{如果}B_k\geqslant 0\\ \xi, & \text{其他}\end{cases} \tag{16}$$

其中，ξ是一个调优的超参数，大于迄今为止获得的最大V2V传输速率．因此，V2V用户越早完成传输，即B_k越早减少到0，V2V用户能获得的奖励就越多．

综上所述，在时隙t对应的系统奖励设置定义为

$$R_{t+1}=\lambda_m C_M+\lambda_k F_K \tag{17}$$

由于V2X用户的吞吐量分布随着车辆的移动性而变化，在式(17)中定义的奖励分配也可以发生变化．其中，λ_m和λ_k表示平衡V2I用户和V2V用户传输速率目标的权重．

3.2 V2X-DQSA算法设计

算法设计面临的3个挑战：(1)信道状态快速变化和部分可观测性；(2)多智能体并发训练引起非平稳性；(3)环境动态变化导致训练有效性的不准确评估．

为应对以上挑战，设计了基于多智能体深度强化学习的V2X频谱接入算法V2X-DQSA，该算法实现分布式频谱资源分配策略的优化，整体框架如图3所示．V2X-DQSA算法的主要思想是建立产生近似行为策略和策略价值判断的改进DQN网络，将车联网中V2V用户在时隙t产生的状态、行为、奖励存储在CERT记忆库中，通过优化网络损失函数反向训练神经网络以获得性能较佳的资源分配策略．

3.2.1 算法架构

传统RL方法都是针对静态环境而设计的，在V2X频谱共享问题中，环境的动态分布可能会发生变化．具体来说，式(17)中设计的奖励分配随着车辆的流动性而波动，引起的高方差和偏差奖励估计导致系统性能降低．为解决该问题，V2X-DQSA算法融合改进DQN和Double-DQN技术，并在训练过程中结合独立学习者(Independent Learners，ILs)训练范式、并发经验回放轨迹CERTs以及Exp3策略关键技术解决多智能体学习算法稳定性问题．

为了能够实现更加高效的特征表示和价值近似，采用深度神经网络(Deep Neural Network，DNN)与递归神经网络(Recurrent Neural Network，RNN)及决斗网络相结合的改进DQN架构，如图4所示．

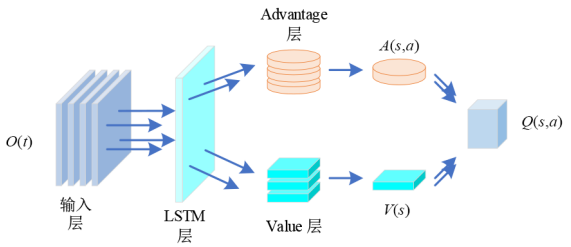

图4 改进DQN架构

该架构中，DNN通常被用作一个函数近似器计算Q值，即$Q^\theta(z,a)$，其中θ表示DNN的参数．具体地，DNN利用循环隐藏层保持内部隐藏状态，自动聚合过去的观察结果．RNN基于所获得部分观察结果估计全局状态的能力，导致有效的学习POMDP策略[15]．此外，RNN的预测能力使其适合于快速变化问题．利用LSTM[16]作为隐藏层，负责学习如何随着时间的推移积累经验，通过保持一个内部状态，并随着时间的推移聚合观测结果，使得网络能够使用进程的历史记录估计真实状态．

决斗网络(Dueling Network)架构[17]，该架构下两个附带网络共存：一个由θ参数化的网络用以估计状态值函数$V(z|\theta)$，另一个由参数化的网络用以估计优势动作函数．通过式(18)对两个网络进行聚合，以近似Q值函数．

$$\begin{aligned}&Q(z,a|\theta,\theta')\\&=V(z|\theta)+\left(A(z,a|\theta')-\frac{1}{|\Delta_\pi|}\sum_{a'}A(z,a'|\theta')\right)\end{aligned} \tag{18}$$

其中，评估采用从Q函数中减去状态相对于所采取动作平均值的方法．

在随机环境中，Q-learning使用最大动作值作为最大期望动作值的近似，引入额外的积极偏差对动作值造成高估．为此，结合Double-DQN[18,19]算法将动作选择与评估分离，采用双估计方法解决价值高估问题．Double-DQN更新的损失函数为

$$\underset{(s,a,r,s')\sim U(D)}{E}\left[\left(Y_t-Q^\theta(z,a)\right)^2\right] \tag{19}$$

$$Y_t=R_{t+1}+\gamma Q\left(z_{t+1},\arg\max Q(z_{t+1},a;\theta_t),\theta_t^-\right) \tag{20}$$

其中，Y_t表示目标值，θ_t^-表示静态目标网络的参数，使用目标网络直接更新．此外，将$Y_t-Q^\theta(z,a)$用表示时间误差δ.

此外，算法遵循独立学习者(Independent learners，ILs)训练范式，智能体以一种分散的方式学习自己行为反馈．但由于在探索阶段其他智能体的行为不可预测，导致ILs存在非平稳性．为解决这一问题，结合迟滞Q学习[20](Hysteretic Q-learning)，根据一个联合行动结果和估计两种学习率的状态值，分别用于高估和低估的时间误差δ，可表示为

$$Q(z,a)\leftarrow\begin{cases}Q(z,a)+\phi\delta, & \text{如果}\delta\leqslant 0\\ Q(z,a)+\alpha\delta, & \text{否则}\end{cases} \tag{21}$$

其中，$0<\phi<\alpha<1$．在实践训练中，随着训练过程的推进而逐渐增长，适应学习速度，以实现智能体在训练初期有效的对抗负面更新．

针对本地经验非并发性质引起的多智能体算法非平稳性问题，引入并发经验回放轨迹CERTs[21]．如图5

所示,将每个体验元组可视化为一个立方体,在执行每个学习集e时,智能体k在时隙t收集经验元组,每一集经验都被存储在沿时间轴t排列的序列中.

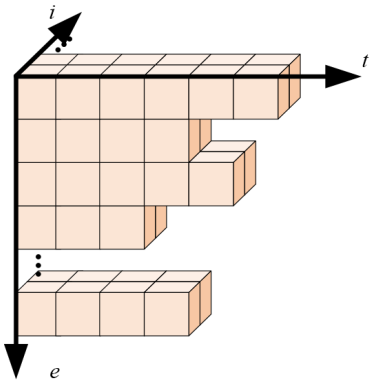

图5　CERT结构

在训练过程中,所有智能体分别沿着事件轴e和智能体i轴同时存储经验. 当需要时,对所有智能体进行同步的小批量采样,以执行Q值更新.

此外,行动选择采取Exp3策略[22],将$Q_k(a)$的分布向量p_k与均匀分布进行混合,以确保算法尝试所有动作,并对每个动作的奖励进行良好估计. 具体计算方法为

$$\Pr\left(a_k(t)=a\right)=(1-\upsilon)\,p_k+\upsilon/K,\ \forall a\in S$$

$$p_k=\frac{e^{\omega Q_k(a)}}{\sum_{\tilde{a}=1}^{K}e^{\omega Q_k(\tilde{a})}} \tag{22}$$

其中,$\upsilon>0$,ω表示温度参数. 此外,υ在实际应用中很小,随着时间的推移变为零,因此算法在选择具有高估计Q值的动作时更加贪婪.

3.2.2　V2X-DQSA算法步骤

每个时隙内,各智能体执行基于多智能体深度强化学习的V2X频谱接入算法,多次迭代后选择概率最大策略作为V2V用户当前时隙的频谱分配策略. 基于多智能体深度强化学习的V2X频谱分配算法详细步骤如算法1所示. 该算法主要包含初始化、环境交互、模型更新3个步骤.

(1)初始化:重启V2X环境模拟,将每个V2V用户的Q网络参数θ_k初始化. 同时,当前迭代次数初始化为0. 本步骤对应算法1中的1~4步.

(2)环境交互:在每个时间片内,每个V2V用户根据Exp3策略选择动作,获得近似的奖励值. 将观察结果$Z_t\{z_1(t),z_2(t),\cdots,z_k(t)\}$存入CERT中. 本步骤对应算法1中的5~9步.

(3)模型更新:从CERT中抽取一批经验进行训练. 采用双DQN更新和滞后学习,通过神经网络梯度反向传播更新Q网络所有参数. 目标网络以较低频率进行更新,以稳定评估网络更新. 本步骤对应算法1中的10~14步.

算法1　V2X-DQSA

输入: 学习率α、迟滞率ϕ、折扣率γ、探索率υ,温度ω,批处理大小B、采样轨迹长度L_T、目标网络更新频率U

输出: Q网络参数θ_k

1. 启动V2X环境模拟器,创建V2I用户和V2V用户
2. 随机初始化每个智能体的Q网络参数θ_k
3. 从第n个训练集开始迭代
4. 从时隙t开始,初始化环境S_t
5. 从智能体k开始,$k\in K$,获取当前观察结果
$Z_t\{z_1(t),z_2(t),\cdots,z_k(t)\}$
6. 在Q网络中使用Z_t作为输入,得到Q网络所有动作对应的Q值输出,采用式(22)得到动作分布
7. 通过Exp3策略选择动作
$A_t\{a_1(t),a_2(t),\cdots,a_k(t)\}$
8. 所有的智能体采取动作A_t获取奖励$R(t)$,并且获得下一个观察结果
$Z_{t+1}\{z_1(t+1),z_2(t+1),\cdots,z_k(t+1)\}$
9. 将$Z_t\{z_1(t),z_2(t),\cdots,z_k(t)\}$添加到CERTs缓冲区中
10. 从CERTs缓冲区中取出大小为B,迹线长度为L_T的经验数据
$E_t\{e_1,e_2,\cdots,e_k\}$
11. 根据式(20)计算每次经验数据$e_k^i=(z,a,r,z')$,$i\in L_T$的目标Q值Y_t
12. 采用迟滞更新
$\delta_e=Y_t-Q^\theta(z,a)$
$\tilde{\delta}_e=\max\{\delta_e,\phi\delta_e\}$
13. 通过神经网络梯度反向传播更新Q网络的所有参数,即
$\theta_t\leftarrow\theta_t+\frac{\alpha}{B}\sum_e\tilde{\delta}_e\nabla Q^\theta(s,a)$
14. 如果$e\%U=0$,则更新目标Q网络参数,$\theta_k^-\leftarrow\theta_k$
15. 结束迭代

4　性能仿真分析

实验中使用3GPP中定义的两车道城市场景进行模拟仿真城市案例[23],其中详细描述了车辆下降模型、密度、速度、移动方向、车辆通道、V2V数据流量等. 车辆在一定范围内以随机速度初始化后保持均匀运动,根据道路拓扑结构移动,当到达一个十字路口时,会选择直转或以相等的概率转弯. 此外,与文献[8]类似,将模拟面积缩小了2倍,以实现模拟的可处理性. M条V2I用户由X型车辆启动,K条V2V用户在每个车辆及其周围的邻居之间形成,假定V2V用户、V2I用户数量和子信道数量S相等,即S=K=M. 表1中列出了仿真中的参数配置,表2列出V2I用户和V2V用户的信道模型参数,训练程序中采用的调优参数见表3.

表1 环境仿真参数

参数名称	值
V2I用户的数量	4, 8
V2V用户的数量	4, 8
载波频率	2 GHz
子载波带宽	4 MHz
BS天线高度	25 m
BS天线增益	8 dBi
车辆天线高度	1.5 m
车辆天线增益	3 dBi
车辆接收机噪声因数	9 dB
车辆移动速度	[10,15] km/h
车辆衰落和移动模型	文献[23]中A.1.2的城市情况
V2I用户传输功率	25 dBm
V2V用户传输功率	[23,10,5,−100] dBm
V2V用户SINRs门限γ_0^K	10 dB
噪声功率	−114 dBm
V2V载荷传输时间限制T	20 ms
V2V起始载荷大小	[2,3,4]×1 060 byte

表2 V2I和V2V用户的信道模型参数

参数名称	V2I用户	V2V用户
路径损失模型	$128.1+37.6\log_{10}d$	LOS in WINNER + B1 Manhattan[24]
阴影分布	对数正太分布	对数正太分布
阴影标准差	8 dB	3 dB
解相关距离	50 m	10 m
路损和阴影更新	100 ms	100 ms
快速衰落	瑞丽衰落	瑞丽衰落
快衰更新	1 ms	1 ms

在训练过程中，为提高评估的准确性，逐渐减少探索率υ，并增加迟滞率ϕ平衡积极和消极样本之间的更新. 执行阶段，每个智能体感知局部观察，根据单个训练模型选择一个概率最大的动作. 此外，与文献[9]类似，在训练阶段，选择最大有效负载，即L=6×1 060 byte，该情况是V2V用户完成数据包交付的最具挑战性设置，以验证该算法的鲁棒性. 此外，为了获得多样化的训练样本，在训练过程中定期重新初始化车辆的位置.

通过提供多智能体应用场景下评估V2X-DQSA算法收敛的有效性，并且定义了信道总容量和有效载荷交付率作为所提算法与对比算法的性能指标. 对比算法具体包括D3RQN[14]、DQN[11]、SAC[8]、随机基线法(random)以及集中式暴力搜索(centralized maxV2V). 随机基线法在每个时间片以随机方式选择每个V2V用户的频谱子波段和传输功率以最大化V2V用户总和率. 集中式暴力搜索(centralized maxV2V)搜索所有V2V用户的动作空间，以获得最大的V2V用户总吞吐量. 该算法虽然实用性不高，但可以为V2I和V2V用户提供性能上界，以分析所提算法的最优性. 此外，为了验证该算法的可伸缩性，进一步研究了该算法在不同车辆速度下的性能.

表3 调优超参数

参数名称	值
学习率α	0.000 1
折扣率γ	0.95
探索率υ	0.05→0
迟滞率ϕ	0.2→0.8
温度ω	1→20
总勘探集	15 000
训练集	20 000
轨迹长度L_T	20
样本大小B	32
目标网络更新频率U	4
CERTs大小	1 000
奖励比重$\{\lambda_m,\lambda_k\}$	{0.1, 1}

图6和图7展示了当车辆用户分别为4和8时的累积奖励曲线与训练损失曲线，其中累积奖励曲线由训练过程中获得的相应最大值归一化. 可以观察到，损失函数随着迭代次数的进一步增加近似收敛，并且累积奖励随着训练的继续而提高，表明了所提算法的收敛有效性和鲁棒性.

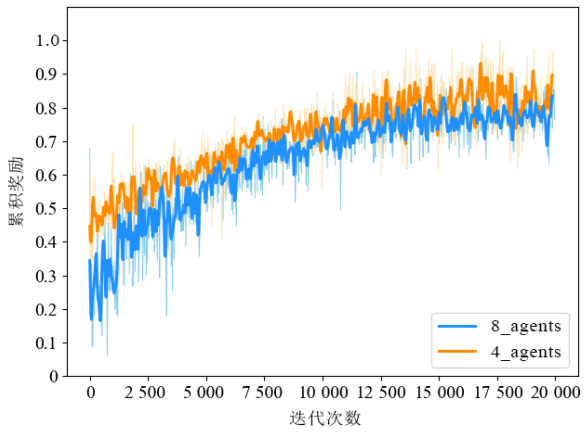

图6 V2V用户训练期间累计奖励

为了对不同算法进行准确和稳健的验证，基于20种不同的随机数生成器种子进行仿真，并以95%的置信区间进行了说明.

图8所示为各种算法V2I用户相对于V2V用户不同有效负载的总吞吐量. 由于有效载荷大小的增加必然导致V2V用户传输时间更长，不可避免地对V2I用户带来更强的干扰，因此所有方法的性能都会下降. 虽然V2X-DQSA算法在观测空间中没有使用完全CSI，但在有效载荷大小增加的情况下，其信道吞吐量均大于其

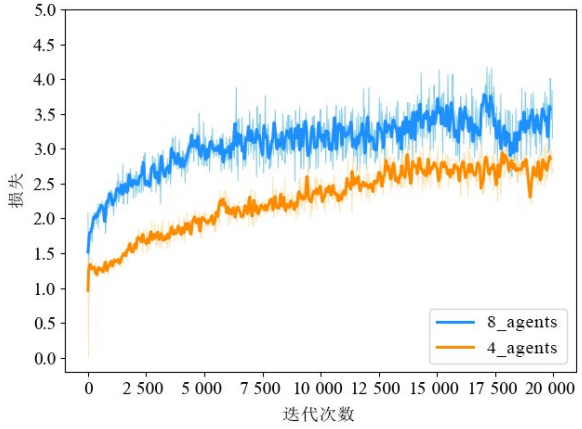

图7　V2V用户训练损失

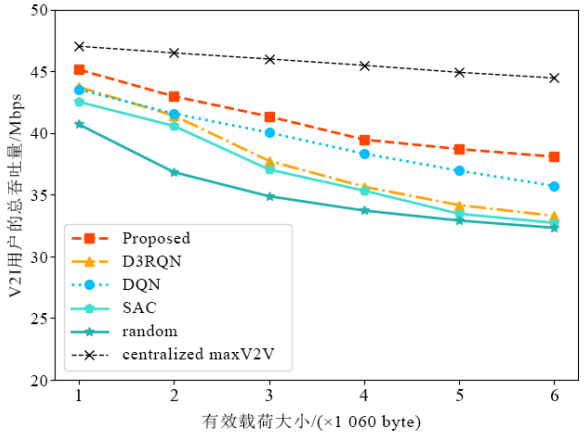

图8　4条V2I用户信道总容量随载荷数量的变化

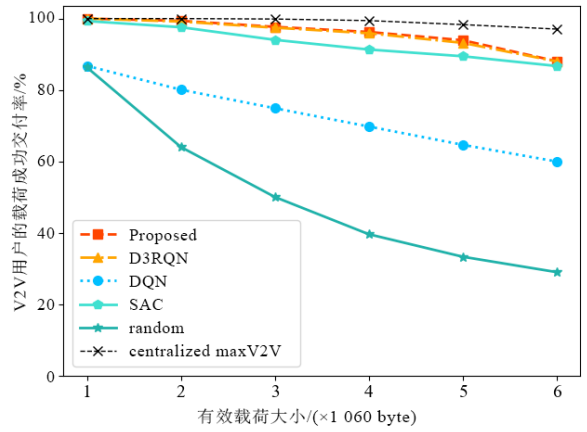

图9　4条V2V用户载荷成功交付率随载荷数量的变化

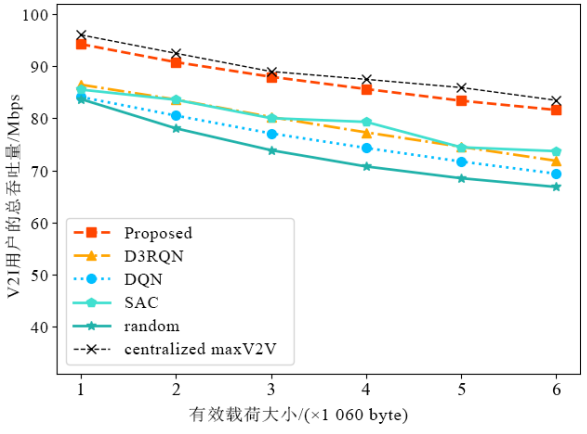

图10　8条V2I用户信道总容量随载荷数量的变化

他算法. 这是因为该算法采用随机行为策略,结合DNN和RNN两种思想,能够在短时间内为智能体选择最优行为,能更好地适用于车联网环境. 同时随着有效载荷的增加,所有算法的V2I用户总吞吐量均降低,这是因为所有V2V用户共享频谱资源,V2I用户总吞吐量和V2V用户包传输比率之间存在权衡. 图9显示所有算法V2V用户相对于增加有效载荷大小的成功交付率. 随着有效载荷大小的增加,不同算法的交付率均在下降并且趋于稳定. 初始交付率高是由于系统开始运行时信道资源充足且需要传输的有效载荷小,可以通过合理分配频谱资源完成传输任务. 随着系统中传输任务的增加,对比算法的交付率明显低于所提算法,表明了该算法在跟踪变化环境动态的有效性.

图10和图11展示了当车辆用户为8时,各种算法随有效载荷变化的吞吐量和交付率性能对比. 从图中可以看出,各种算法的总吞吐量和交付率随着有效载荷的增加呈下降趋势. 此时V2X-DQSA算法下降趋势和对比算法相比较为平稳,并且总吞吐量和交付率性能依旧优于其他算法,这是因为该算法Exp3策略的采用使得智能体在完成当前任务的基础上尽可能地行为

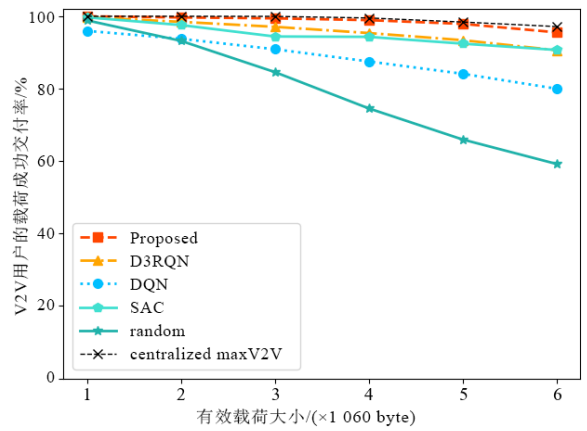

图11　8条V2V用户载荷成功交付率随载荷数量的变化

随机化以获得近似最优的多种选择,提高了智能体在环境中的探索能力,也同时提高了算法在动态环境中的稳定性.

进一步考虑车速对该算法性能的影响,在车辆低速的情况下,对V2V用户的SINRs门限值约束会降低某些车辆成功接入网络的可能性. 而在车辆高速的情

况下，网络用户间的耦合干扰低，V2V用户的SINRs门限值约束要求可以得到满足，更容易成功接入网络. 图12和图13展示了当车辆用户是4和8，有效载荷为6×1 060 byte时，该算法的V2I用户总吞吐量和V2V用户包传输率之和. 为了评价该算法的泛化能力，用速度[10，15]m/s训练的相同模型. 车辆速度从五个不同范围生成，即[10，15]m/s、[15，20]m/s、[20，25]m/s和[25，30]m/s、[30，35]m/s. 从图中可以看出，随着车辆速度的增加，所提算法的性能并没有明显波动，表明了该算法在高动态网络下的有效性和稳定性.

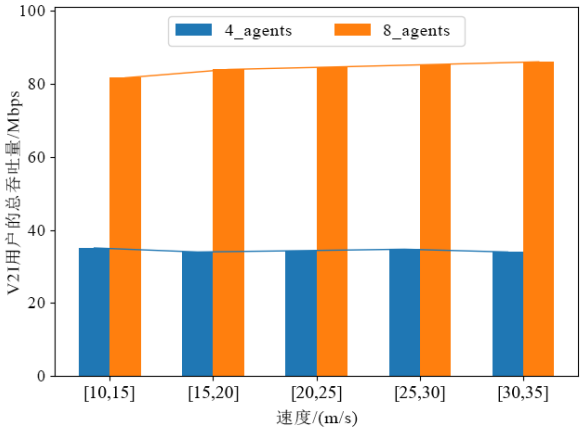

图12 车速对V2I用户总吞吐量的影响

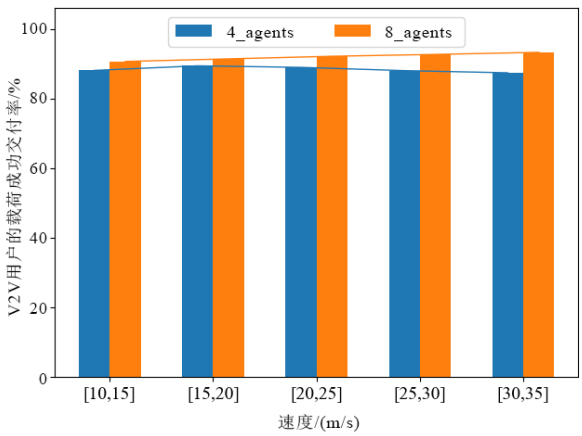

图13 车速对V2V用户载荷成功交付率的影响

V2X-DQSA算法计算复杂度实际上取决于已实现的DNN结构，并随着隐藏层数和相应神经元数量的增加而增加. 具体来说，每个智能体DNN由一个包含64个神经元完全连接的输入层、一个包含128个单元隐藏LSTM层和一个包含64个神经元完全连接的输出层组成. 该算法用Python和PyTorch实现. 如果能够在执行中引入更先进的加速技术，如模型压缩、量化、GPU，甚至专用FPGA、硬件加速等，还能够进一步提升效率.

5 结论

本文提出了一种基于多智能体深度强化学习的V2X频谱分配算法，该算法通过改进型DQN模型设计，支持V2V用户仅基于局部CSI观测结果优化联合子信道和传输功率的选择，并在没有智能体间通信的情况下进行隐式协调. 在满足V2V用户延迟和可靠性要求的同时，最大限度地提高V2I用户的总吞吐量. 仿真实验结果表明，所提算法能够获得稳定和具有更好收敛性能的模型，同时算法也降低了信令开销，并具备较好的可扩展性. 未来改进方向包括设计适应实际环境的复杂交通、信道和车辆移动模型，进一步提高所提算法可扩展性泛化、训练效率和鲁棒性等.

参考文献

[1] GYAWALI S, XU S, QIAN Y, et al. Challenges and solutions for cellular based V2X communications[J]. IEEE Communications Surveys & Tutorials, 2020, 23(1): 222-255.

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

[3] WANG S, LIU H, GOMES P H, et al. Deep reinforcement learning for dynamic multichannel access in wireless networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(2): 257-265.

[4] NAPARSTEK O, COHEN K. Deep multi-user reinforcement learning for distributed dynamic spectrum access[J]. IEEE Transactions on Wireless Communications, 2018, 18(1): 310-323.

[5] 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.

[6] YE H, LI G Y, JUANG B H F. Deep reinforcement learning based resource allocation for V2V communications[J]. IEEE Transactions on Vehicular Technology, 2019, 68(4): 3163-3173.

[7] ZHANG X, PENG M, YAN S, et al. Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications[J]. IEEE Internet of Things Journal, 2019, 7(7): 6380-6391.

[8] 黄煜梵, 彭诺蘅, 林艳, 等. 基于SAC强化学习的车联网频谱资源动态分配[J]. 计算机工程, 2021, 47(9): 34-43.
HUANG Y F, PENG N H, LIN Y, et al. Dynamic al location of spectrum resources of Internet of vehicles based on sac reinforcement learning[J]. Computer Engineering, 2021, 47(9): 34-43. (in Chinese)

[9] WANG L, YE H, LIANG L, et al. Learn to compress CSI and allocate resources in vehicular networks[J]. IEEE Transactions on Communications, 2020, 68(6): 3640-3653.
[10] 许新操, 刘凯, 刘春晖, 等. 基于势博弈的车载边缘计算信道分配方法[J]. 电子学报, 2021, 49(5): 851-860.
XU X C, LIU K, LIU C H, et al. Channel allocation method for vehicle edge computing based on potential game [J]. Acta Electronica Sinica, 2021, 49(5): 851-860. (in Chinese)
[11] 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.
[12] LE T D, KADDOUM G. A distributed channel access scheme for vehicles in multi-agent V2I systems[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(4): 1297-1307.
[13] XU Y H, YANG C C, HUA M, et al. Deep deterministic policy gradient (DDPG) -based resource allocation scheme for NOMA vehicular communications[J]. IEEE Access, 2020, 8: 18797-18807.
[14] XIANG P, SHAN H, WANG M, et al. Multi-agent rl enables decentralized spectrum access in vehicular networks [J]. IEEE Transactions on Vehicular Technology, 2021, 70(10): 10750-10762.
[15] XU Y, YU J, BUEHRER R M. The application of deep reinforcement learning to distributed spectrum access in dynamic heterogeneous environments with partial observations[J]. IEEE Transactions on Wireless Communications, 2020, 19(7): 4494-4506.
[16] HAUSKNECHT M, STONE P. Deep recurrent Q-learning for partially observable MDPs[J]. AAAI Fall Symposium - Technical Report, 2015, 3: 29-37.
[17] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]//International Conference on Machine Learning. New York: ACM, 2016: 1995-2003.
[18] HASSELT H. Double Q-learning[J]. Advances in Neural Information Processing Systems, 2010, 23: 2613-2621.
[19] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2016: 2094-2100.
[20] MATIGNON L, LAURENT G J, LE FORT-PIAT N. Hysteretic Q-learning: An algorithm for decentralized reinforcement learning in cooperative multi-agent teams [C]//2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2007: 64-69.
[21] OMIDSHAFIEI S, PAZIS J, AMATO C, et al. Deep decentralized multi-task multi-agent reinforcement learning under partial observability[C]//International Conference on Machine Learning. New York: ACM, 2017: 2681-2690.
[22] AUER P, CESA-BIANCHI N, FREUND Y, et al. Gambling in a rigged casino: The adversarial multi-armed bandit problem[C]//Proceedings of IEEE 36th Annual Foundations of Computer Science. Piscataway: IEEE, 1995: 322-331.
[23] LG Electronics, Deutsche Telekom. WF on SLS evaluation assumptions for eV2X[EB/OL]. (2016)[2022]. https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_85/Docs/R1-165704.zip.
[24] MARTIN D, WERNER M, AFIF O. WINNER II channel models[M]//Radio Technologies and Concepts for IMT-Advanced. New York: Wiley, 2010: 39-92.

作者简介

王为念　男,1996年出生,江苏徐州人. 2020年获得南京信息工程大学学士学位,目前在南京信息工程大学计算机软件学院攻读硕士学位. 主要研究方向为车联网、强化学习和频谱共享等.
E-mail: 1334079243@qq.com

苏　健　男,1986年出生,湖北荆州人. 2012年和2016年分别在华中师范大学和电子科技大学获工学硕士和工学博士学位. 现为南京信息工程大学副教授,发表sci期刊论文50余篇. 主要从事物联网技术、无线传感器网络及反向散射通信等方面的研究工作.
E-mail: sj890718@gmail.com

陈　勇　男,1975年出生,湖南衡阳人. 2000年获解放军理工大学硕士学位. 现为国防科技大学第六十三研究所研究员. 主要研究方向为认知无线网络、频谱管理.
E-mail: chy63s@126.com

张建照　男,1985年出生,河南南阳人. 2012获解放军理工大学工学博士学位. 现为国防科技大学第六十三研究所高级工程师. 主要研究方向为认知无线电、智能频谱管理.
E-mail: lgzjz2007@126.com

唐　震　男,1998年出生,江苏南通人. 2020年获得南京信息工程大学学士学位. 目前在南京信息工程大学计算机软件学院攻读硕士学位. 主要研究方向为物联网、信号处理和深度学习等.
E-mail: 903414103@qq.com