计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202440404
61(11): 2806−2820, 2024

C-V2X 环境下基于队友模型的多智能体通信切换优化

刘冰艺[1,2] 王东东[1] 施海勇[1] 王恩澍[3] 吴黎兵[3] 汪建平[4]
[1](武汉理工大学计算机与人工智能学院 武汉 430070)
[2](武汉理工大学三亚科教创新园 海南三亚 572000)
[3](武汉大学国家网络安全学院 武汉 430070)
[4](香港城市大学计算机科学系 香港九龙)
(byliu@whut.edu.cn)

Optimization of Multi-Agent Handover Based on Team Model in C-V2X Environments

Liu Bingyi[1,2], Wang Dongdong[1], Shi Haiyong[1], Wang Enshu[3], Wu Libing[3], and Wang Jianping[4]
[1](*School of Computer Science and Artificial Intelligence*, *Wuhan University of Technology*, *Wuhan* 430070)
[2](*Sanya Science and Education Innovation Park*, *Wuhan University of Technology*, *Sanya*, *Hainan* 572000)
[3](*School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430070)
[4](*Department of Computer Science*, *City University of Hong Kong*, *Kowloon*, *Hong Kong*)

Abstract Cellular vehicle-to-everything (C-V2X) communication technology is a crucial component of future intelligent transportation systems (ITS). Millimeter wave (mmWave), as one of the primary carriers for C-V2X technology, offers high bandwidth to users. However, due to limited propagation distance and sensitivity to obstructions, mmWave base stations must be densely deployed to maintain reliable communication. This requirement causes intelligent connected vehicle (ICV) to frequently switch communications during travel, easily leading to local resource shortages, thus degrading service quality and user experience. To address these challenges, we treat each ICV as an agent and model the ICV communication switching issue as a cooperative multi-agent game problem. To solve this problem, we propose a cooperative reinforcement learning framework based on a teammate model. Specifically, we design a teammate model to quantify the interdependencies among agents in complex dynamic environments. Furthermore, we propose a dynamic weight allocation scheme that generates weighted mutual information among teammates for the input of the mixing network, aiming to assist teammates in switching to base stations that provide satisfactory QoS and QoE, thereby achieving high throughput and low communication switching frequency. During the algorithm training process, we design an incentive-compatible training algorithm aimed at aligning the individual goals of the agents with collective goals, enhancing communication throughput. Experimental results demonstrate that this algorithm achieves a 13.8% to 38.2% increase in throughput compared with existing communication benchmark algorithms.

Key words C-V2X; resource allocation; handover; multi-agent reinforcement learning; cooperative multi-agent reinforcement learning

收稿日期: 2024-05-31; **修回日期**: 2024-07-31

基金项目: 国家自然科学基金项目(62272357, 62302326, 62202348, U20A20177); 湖北省重点研发计划基金项目(2022BAA052); 香港研究资助局 NSFC/RGC 项目(N_CityU140/20)

The work was supported by the National Natural Science Foundation of China (62272357, 62302326, 62202348, U20A20177), the Key Research and Development Program of Hubei province (2022BAA052), and the Project of Hong Kong Research Grant Council under NSFC/RGC (N_CityU140/20).

通信作者: 王恩澍(wanges17@whu.edu.cn)

摘　要　蜂窝车联网（cellular vehicle-to-everything, C-V2X）通信技术是未来智能交通系统（intelligent transportation systems, ITS）的重要组成部分. 毫米波（millimeter wave, mmWave）作为 C-V2X 通信技术的主要载体之一，可以为用户提供高带宽. 然而，由于传播距离有限和对遮挡的敏感性，毫米波基站必须密集部署以维持可靠的通信，这使得智能联网车辆（intelligent connected vehicle, ICV）在行驶过程中不得不频繁地进行通信切换，极易造成局部资源短缺，进而导致服务质量低下和用户体验不佳. 为了应对这些挑战，每辆 ICV 被视为一个智能体，并将 ICV 的通信切换问题建模为一个合作型多智能体博弈问题. 为了解决这一问题，提出了一个基于队友模型的合作型强化学习框架. 具体来说，首先设计了一个队友模型，用于量化复杂动态环境下智能体之间的相互依赖关系；接着提出了一种动态权重分配方案，生成了队友间的加权互信息，用于混合网络的输入，旨在帮助队友切换到可以提供良好 QoS 和 QoE 的基站，以获得高吞吐量和低通信切换频率. 在算法训练过程中，设计了一种激励相容训练算法，旨在协调智能体的个体目标与集体目标的一致性，提升通信吞吐量. 实验结果显示，提出的方法在不同规模车辆的场景中均展示出了卓越的性能，相较于现有的基于通信基准方法有 13.8%~38.2% 的吞吐量提升.

关键词　蜂窝车联网；资源分配；通信切换；多智能体强化学习；合作多智能体强化学习

中图法分类号　TP391

随着物联网（IoT）和人工智能（AI）的迅猛发展，智能物联网（artificial intelligence of things, AIoT）逐渐成为现代科技的重要领域，将 AI 的强大分析和决策能力与 IoT 设备的广泛连接和数据收集能力相结合，推动着各行各业的智能化升级. 尤其在交通领域，AIoT 与蜂窝车联网（cellular vehicle-to-everything, C-V2X）技术的结合促进了智能网联汽车新兴应用和服务的发展[1]. 毫米波（millimeter wave, mmWave）作为其主要载体，提供了大规模网络覆盖、高数据传输速率和容量、极低的通信延迟、更高的能源效率，并支持多种设备和服务[2]. 为了向用户提供更广泛的网络覆盖、更高的数据传输速率以及更优秀的连接体验，基于 mmWave 的通信技术采纳了诸如网络密集化、异构网络（HetNets）和大规模多输入多输出等关键技术[3-4]. 在 C-V2X 环境中，这些技术的应用不仅使车辆间的直接通信更加高效，而且促进了车辆与基础设施之间的互联互通，为智能交通系统（intelligent transportation systems, ITS）和自动驾驶技术的实现提供了强大的网络支持[5-6].

由于毫米波的波长较短，其传播存在较大的路径损耗且难以有效穿透墙壁或其他障碍，因此需要密集部署毫米波基站以确保服务的有效性[7]. 然而，这种部署策略也导致智能联网车辆（intelligent connected vehicle, ICV）在高速移动时，尤其是在城市快速路上需频繁进行通信切换[8]. 在 HetNets 中，用户设备数据在跨区域或网络间的传输涉及复杂的计算和高信号开销，这对网络性能产生了显著影响[9]. 此外，在城市中心或大型活动现场等高流量区域，每辆车都可能需要网络连接以进行数据传输、位置更新、接收导航信息等操作，原有的网络资源如基站的带宽和处理能力可能会迅速达到饱和点[10].

传统的通信切换方案[11-13]依赖于车辆与当前基站的信号质量或其他预设的单一条件来决定是否进行切换，在车辆密集且动态变化的环境中可能出现资源过载现象，进而导致网络响应速度下降、数据传输延迟增加，甚至可能出现通信中断的情况. 现有的基于规则的通信切换方案[14-20]通常是静态的，缺乏对环境变化的适应性. 由于基于规则的方案通常只在满足条件后才触发切换，这可能导致响应延迟. 在高移动性的通信环境中，这种延迟可能导致连接中断或服务质量下降. 在信号边界地带，设备可能不断在 2 个基站之间来回切换，消耗额外的网络资源并影响用户体验.

人工智能技术的发展为实现智能通信切换打开了新思路，许多研究[21-27]应用强化学习方案来解决通信切换问题，取得了一定的成果. 目前，基于单智能体的方法依赖于局部观测做出决策，缺乏对全局网络状态的了解，在动态通信场景中的性能不佳. 而基于多智能体的方法通常采用去中心化的顺序执行策略，多个智能体独立进行训练和执行，没有考虑到智能体之间的协调性. 在多智能体系统中，智能体之间的有效合作直接影响整个系统的效能和效果. 当每个智能体都试图最大化自己的即时奖励时，可能会忽视长期合作的收益，从而导致合作关系不稳定，甚至出现对集体目标不利的竞争行为.

针对这些挑战，本文提出一个基于队友模型的

合作型多智能体强化学习框架. 具体来说, 首先将 C-V2X 异构网络场景中智能联网车辆的通信切换和资源利用不均匀问题转化为一个多智能体马尔可夫博弈问题, 其中每一辆车被建模为一个智能体. 为了解决多个车辆在异构网络场景中的通信切换问题, 本文在算法 QMIX 的基础上改进了一个多智能体架构自主学习智能通信切换策略. 针对通信资源利用不均匀问题, 本文首先设计了一个队友模型, 队友模型利用智能体的本地观测和队友编号生成针对特定队友互信息, 互信息的意义是给出自己的观测中对队友行为的建议, 在不扩大队友的观测空间的前提下, 为智能体推荐高 QoS 和 QoE 的基站. 接着, 本文提出了一个动态权重分配方案, 通过智能体的速度信噪比以及和队友间的距离对互信息进行权重分析, 筛选有用的信息, 协助智能体选择最优的基站, 避免资源利用不均. 此外, 本文设计了激励相容模块, 将个体的优化目标和集体优化目标纳入决策过程, 当智能体的选择偏离了集体利益的时候, 本文通过拉格朗日乘数放大这种影响, 作为对智能体行为的约束, 确保每个智能体在追求个人最大化利益的同时, 也会考虑其行为对整体系统的影响, 从而促进智能体之间的合作.

综上所述, 本文的贡献包括 3 个方面:

1)将 C-V2X 异构网络场景中智能联网车辆的通信切换和资源利用不均匀问题形式化为一个合作多智能体博弈问题. 为了解决该问题, 本文提出了一个基于队友模型的合作型多智能体强化学习框架.

2)设计了一个在 QMIX 基础上改进的多智能体强化学习架构和队友模型, 解决基站密集场景下多车智能通信切换和通信资源利用不均问题.

3) 为了评估本文的方法, 搭建了异构车联网协同仿真场景, 并进行了实验. 实验结果表明, 本文方法性能优于最好的基线方法.

1 相关工作

1.1 传统通信切换方法

车辆通信切换问题已经受到广泛的关注, 有许多研究致力于解决移动车辆在通信过程中的通信切换挑战[14-20]. 为了降低通信切换开销, 相关研究不仅对通信切换参数进行了调整[15], 还优化了通信切换过程[16]. Isa 等人[15] 提出了一种基于自组织网络概念的 LTE 网络自组织通信切换过程, 通过优化通信切换参数 Hysteresis(Hys)和 Time-To-Trigger(TTT), 网络性能在吞吐量、延迟和抖动方面有显著改善, 优化这些参数可以显著提高网络性能. Alhabo 等人[16] 提出了一种基于灰色关联分析的通信切换方法——GRA-通信切换, 用于密集小区异构网络中. 该方法结合了分析层次过程(analytic hierarchy process, AHP)来确定通信切换指标的权重, 并使用灰色关联分析方法对可用小区进行排序以选择最佳通信切换目标. 受此启发, Hasan 等人[17] 提出了一种针对超密集异构网络中频繁通信切换问题的缓解算法. 该算法基于用户的移动行为, 将经历频繁通信切换的用户分为快速移动用户和乒乓用户. 对于快速移动用户, 算法将其切换到宏层网络, 而乒乓用户则通过调整通信切换参数进行管理. 如果调整参数无法避免乒乓现象, 这些用户也会被切换到宏层网络. 该方法显著提高了通信切换性能并减少了成本. 此外, 考虑到车辆移动性的复杂性和服务需求的多样化, 有研究提出利用移动边缘计算(MEC)的强大计算能力[18-20]. Qi 等人[20] 通过在基站附近部署 MEC 并采用 LSTM 方案, 能预测车辆在候选网络的 QoS 覆盖区内的停留时间. 同时, 在 MEC 服务器上部署的 5G gNB 选择算法使得 MEC 服务器可以作为最终通信切换决策的协调者, 从而优化整个网络的运行效率.

基于规则的方法在环境发生变化时表现不佳, 尤其是面对网络条件、用户行为或设备特性的变化时, 这些方法依赖于预先定义的逻辑和参数, 在复杂或未知环境中缺乏灵活性.

1.2 基于强化学习的通信切换方法

随着强化学习(reinforcement learning, RL)技术在决策控制领域的发展, 该技术逐渐被引入到解决车辆通信切换问题中[21-27]. Wang 等人[22] 通过引入一个双层框架, 结合了异步多用户强化学习和监督学习技术, 来优化移动物联网和传统蜂窝用户的通信切换过程. 这个框架首先将用户设备根据移动模式聚类, 然后在每个簇内使用深度神经网络控制通信切换, 目的是降低通信切换频率并保证系统吞吐量. Nasim 等人[23] 提出了一种基于多臂老虎机学习模型的方法, 用于跟踪和优化毫米波波束形成向量, 并使用汤普森采样算法来选择最佳波束, 以提高移动用户的通信容量. 该方法在毫米波 COSMOS 实验室中进行了实验评估, 验证了其有效性. Wei 等人[24] 应用深度 Q 网络(deep Q-network, DQN)扩展了预连接切换(predictive HO, PHO)机制, 以支持 5G 网络中严格的服务质量要求. 通过利用 DQN 辅助的 P 通信切换管理方案, 针对目标小区的选择问题进行了顺序决

策，基于候选小区的参考信号接收质量值及其变化率进行排序，以优化 P 通信切换的成功率. Lin 等人[25]设计了一种异构车辆集群迁移框架，该框架利用路侧单元（road side unit, RSU）协作和车辆中继来改善连接性. Lee 等人[26]的工作针对毫米波信号的特性提出了一种基于双 DQN 的深度强化学习框架，根据过去接受的信号数据进行阻塞预测. Wang 等人[27]提出了一种称为 O-MAPPO 的毫米波场景下的动态切换方案. 具体来说，MAPPO 做出触发决策后，基于信道模型计算用户设备（user equipment, UE）和基站（base station, BS）之间的信噪比（signal to interference plus noise ratio, SINR），还对波束和 BS 的选择以及带宽分配进行优化和评估，并计算出所有 UE 的吞吐量和时延；然后将计算结果作为观测和状态传递给 MAPPO，根据奖励函数评估切换触发决策. Guo 等人[28]提出了一种新的多智能体强化学习方法，用于处理移动用户设备的通信切换和功率分配问题. 该方法采用完全协作的多代理框架，所有代理都有相同的目标. 文献 [28] 的作者引入了一种集中式训练与分散式执行的框架，并使用近端策略优化算法来训练每个 UE 的策略. 在训练阶段使用全局信息，此外，引入的反事实基线帮助解决了信用分配问题.

然而，上述解决方案仍然存在一些缺陷. 首先，单个代理通常只能处理其直接感知到的信息，这限制了它可以考虑的环境因素的范围. 单个代理可能只知道当前基站的信号强度和自身的位置，而不了解其他基站的负载情况或周围其他车辆的行为. 这种局限性会导致决策过程中缺乏全局视角，从而无法最优化整体网络性能. 此外，单个代理可能只优化其自身的通信切换性能，而忽视对其他车辆的影响. 这种自私的行为可能导致其他车辆的服务质量下降，甚至引发更多的通信切换，增加网络负担.

同时，在非合作多代理系统中，代理独立行动而不进行协调，也可能导致策略冲突. 多个代理可能同时竞争相同的资源，如同一个基站的带宽或信道，这不仅会导致资源分配不均，还可能引发竞争问题，导致通信质量的下降. 尤其是在高密度车辆环境中，这种冲突会变得更加严重，甚至导致频繁的通信切换失败、连接中断，严重影响整体网络的稳定性和吞吐量. 此外，由于缺乏协调，非合作多代理系统可能无法有效利用环境中的可用信息，从而无法充分优化整体网络性能.

为了克服这些挑战，需要更协调的多代理系统，能够在各代理之间实现有效的合作和信息共享. 通过合作，代理可以获得更全面的环境信息，做出更加全局的优化决策，从而提高网络的整体性能和用户体验质量. 合作多代理系统不仅能够优化资源分配、减少冲突，而且能够通过共享信息提高通信切换决策的准确性，减少不必要的通信切换次数和通信切换失败的风险.

基于以上考虑，本文提出的合作算法解决方案不同于传统的集中式训练分散式执行框架. 本文设计了一种队友模型，该模型利用队友之间的通信来解决传统方法的缺陷. 车辆会利用队友模型根据本地的观察和队友的编号生成对特定队友的互信息，给出自身观察下的队友的动作选择，互信息生成后，为了避免无用的信息对智能体选择的干扰和信道的浪费，本文对生成的互信息进行权重分析，由于安全和隐私问题，车辆不能直接广播观察和行动信息. 本文直接将生成的通信信息应用于队友的价值函数，同时为了更好地促进智能体之间的合作，与传统的合作算法将优化目标等同于最大化总体利益不同，本文设计了激励相容约束，这些约束在代理最大化自己利益的同时考虑了总体利益，当智能体的行为偏离了总体的目标时，本文会放大这种影响，作为一种惩罚，使智能体在考虑个体目标的时候兼顾总体的目标，从而更好地平衡了个体利益和集体利益.

2 C-V2X 通信模型

在本节中，本文介绍所使用的 C-V2X 通信场景建模以及 C-V2X 通信信道建模.

2.1 场景模型概览

在车联网环境中，通信切换是确保车辆在移动过程中能够进行连续通信的关键技术. 特别是在高速行驶的情况下，需要在不同的基站之间进行通信切换，以维持稳定的通信连接. 基站类型包括提供广覆盖和较长服务半径的宏基站（macro base stations, MBS）以及提供高带宽和低延迟通信服务的毫米波小基站（small base stations, SBS）. 如图 1 所示，当智能网联汽车从一个基站的覆盖区域移动至同一系统内另一个基站的覆盖区时会发生通信切换. 具体来说，当车辆即将驶出基站 A 的服务范围时，需要在基站 B 和基站 C 中做出选择. 其中，基站 B 距离更远但是服务的车辆少，基站 C 更近但是服务的车辆更多. 车辆对基站的选择直接影响到通信服务质量，在 C-V2X 的密集通信场景下，这一问题尤为显著.

在本文的 C-V2X 通信场景中，本文考虑了由 SBS

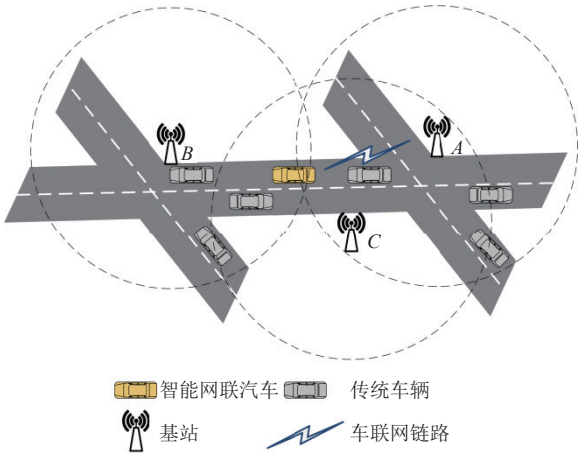


Fig. 1 Handover during vehicle movement

图 1 车辆移动过程中的通信切换

和 MBS 共同组成的异构网络. SBS 使用毫米波频率进行通信, 通常在 24 GHz 至 100 GHz 频率范围内运行, 其中 28 GHz、39 GHz 和 60 GHz 是最常用的频率. 毫米波技术提供的带宽比传统微波频率宽, 通常达到几百 MHz 甚至 GHz, 且允许更高的数据传输速率, 能够支持高达 20 Gbps 的峰值速率. MBS 是蜂窝网络中的关键基础设施, 提供广泛的覆盖和连接服务, 通常在 300 MHz 到 6 GHz 的微波频段内运行, 且 MBS 的覆盖范围可从几公里到数十公里不等. 虽然与 SBS 相比, MBS 的数据传输速率较低, 但仍然能提供稳定的网络连接. MBS 特别适用于需要广泛覆盖并服务大量用户的环境, 如市中心、郊区住宅区和主要交通路线.

在本文环境的模型中, MBS 负责提供广泛覆盖和基本的通信服务, 而 SBS 则负责在高流量区域提供高带宽和低延迟的通信服务. 车辆在移动过程中可以根据实际需求选择连接 SBS 或 MBS. 如果车辆处于高速移动状态, 可能更适合连接覆盖范围更大的 MBS, 以减少通信切换次数和提高连接稳定性. 而在低速移动或静止状态下, 连接 SBS 可以获得更高的数据传输速率和更好的用户体验. 由于 MBS 拥有更大的通信范围, 可以覆盖到全局, 本文假设 MBS 在训练过程中可以获取到全局信息.

2.2 系统信道建模

C-V2X 技术使用的通信频段(24~100 GHz)具有极高的数据传输速率和容量. 然而, 其使用的毫米波信号传播距离较短, 容易受到物理障碍物的影响, 导致信号衰减严重. 为了能准确地建模 C-V2X 信道, 本文在信道建模的时候引入了 Close-In(CI)模型.

CI 模型是一种在无线通信中使用的传播损失预测模型, 适用于毫米波和微波频段. 该模型基于在某一特定参考距离(如 1 m)内的自由空间路径损失(free space path loss, FSPL)以及超出该参考距离的额外衰减构建, 能够考虑到信号强度随距离的变化, 并能够区分视距和非视距场景. 此外, CI 模型还提供了一种有效的方法来估计复杂环境中的信号传播损失, 在无线通信领域得到了广泛应用. 路径损失计算为

$$L(f,d)=F(f,d_0)+10n\lg\left(\frac{d}{d_0}\right)+\chi_\sigma^c,\ \text{where}\ d\geqslant d_0, \tag{1}$$

其中, $F(f,d_0)$表示自由空间路径损失, 即在频率 f 和距离 d_0 下的路径损失, 用分贝(dB)表示. d 表示接收器和发射器之间的实际距离; d_0 是参考距离, 是建立路径损失模型时的起始点距离; $10n\lg(d/d_0)$用于描述距离 d 对路径损失的影响, n 是路径损失指数, 反映环境对信号衰减的影响程度, 该项通过取距离比的对数并乘以 n, 将距离的影响转换为对路径损失的贡献. χ_σ^c是一个额外项, 考虑随机变化, 表示特定环境中阴影衰落的标准偏差.

自由空间路径损失是衡量电磁波通过自由空间传播时能量衰减的一种度量, 在无线通信中是一个非常重要的概念, 用于计算从一个点到另一个点传输的信号在没有任何阻碍物的情况下的衰减. 自由空间路径损失计算为.

$$F(f,d_0)=20\lg\left(\frac{4\pi f d_0\times 10^9}{c}\right), \tag{2}$$

其中, c 是光速. 本文选择 1 m 作为临近自由空间的参考距离, 可以为户外城市微观和城市宏观环境提供很好的参数稳定性和模型精确性, 从而建立了一种标准化的建模方法.

3 问题建模

在本文中, 本文将智能网联汽车的通信切换决策问题转化为一个马尔可夫博弈问题, 其中每一辆智能网联汽车被建模为一个智能体. 马尔可夫博弈中各要素的具体定义为:

1) 智能体. 分布在基站附近的 N 辆智能网联汽车被建模为智能体 $I=\{agt_1, agt_2, \cdots, agt_N\}$.

2) 观测. 在每个时间步 t, 每个用户设备 u 都会到一个局部观测 o_t^u, 其中包括它的移动速度 v^u、移动的方向 d^u、使用的基站编号 b^u、信号测量 $X^u(t)$以及基站服务的用户设备数量 n_b^u. 信号测量表示为 $X^u(t)=(X_0^u(t), X_1^u(t), \cdots, X_M^u(t))$, 表示用户设备在时间步 t 从各个基站接收到的信号强度, 其中 M 代表基站的总数.

3）动作. 在每个时间步 t, 智能体需要选择一个基站进行通信. 动作可以表示为 $A=\{a_0,a_1,\cdots,a_M\}$, 其中 $b_t\in A=\{a_0,a_1,\cdots,a_M\}$ 是选中的基站.

4）奖励. 在每个时间步 t 结束时, 智能体 i 收到一个奖励 $r^u(o_t^u,a_t)=\alpha R_b^u(o_t^u,a_t)-\beta P_h^u(o_t^u,a_t)$, 其中 $R_b^u(o_t^u,a_t)$ 代表智能体的吞吐量大小, $P_h^u(o_t^u,a_t)$ 代表智能体因发生通信切换受到的惩罚. 系数 α 和 β 是权重因子.

为了计算 MBS 和 SBS 的吞吐量, 本文使用了 2 个公式. MBS 的吞吐量计算公式为:

$$R_{\mathrm{mb}}(o_t^u,a_t)=w_0^u\mathrm{lb}(1+SINR_{u,\mathrm{mb}}), \tag{3}$$

其中, $R_{\mathrm{mb}}(o_t^u,a_t)$ 表示 MBS 向 UE 提供的吞吐量, 而 $SINR_{u,\mathrm{mb}}$ 是 UE 在 MBS 覆盖下的信噪比. 对于 SBS 的吞吐量计算公式为:

$$R_{\mathrm{mm}}(o_t^u,a_t)=w_0^u\mathrm{lb}(1+SINR_{u,\mathrm{mm}}), \tag{4}$$

在这种情况下, $R_{\mathrm{mm}}(o_t^u,a_t)$ 表示 SBS 向 UE 提供的吞吐量, 而 $SINR_{u,\mathrm{mm}}$ 表示 UE 在 SBS 覆盖下的信噪比.

量化 UE 在基站间通信切换的影响至关重要. 本文设计了一个被称为通信切换惩罚 $P_h^u(o_t^u,a_t)$ 的度量, 用于评估由于 UE 的通信切换而引起的服务中断. 对于时间步 t 的 UE u, 通信切换惩罚定义为

$$P_h^u(o_t^u,a_t^u)=1\{b_t^u\neq b_{t-1}^u\}, \tag{5}$$

其中, $P_h^u(o_t^u,a_t)$ 表示 UE u 的通信切换惩罚, 而 b_t^u 和 b_{t-1}^u 分别代表 UE 在时间步 t 和 t–1 的服务基站, 如果时间步 t 的服务基站与时间步 t–1 的服务基站不同, 表明发生了一次通信切换事件, 因此会有相应的惩罚. 如果 UE 在这些时间点上保持与同一基站的连接, 函数值为 0, 表示没有发生通信切换, 因此没有惩罚.

4 基于队友模型的多智能体强化学习

本节介绍本文提出的基于队友模型的多智能体强化学习框架, 该框架在 QMIX 的基础上增加了队友模型和互信息权重分析, 并在训练阶段引入了激励相容性原则, 本节主要内容包括 3 个部分, 分别为 VHBTM 框架的系统概览、队友模型以及算法训练介绍.

4.1 系统概览

本文将基站密集场景下车辆通信切换问题建模为一个多智能体马尔可夫博弈问题, 其中每一辆车被建模为一个智能体. 如图 2 所示, 每一个智能体内部结构包括 2 个分支, 分别为一个策略模型、一个队友模型. 在训练过程中本文加入激励相容模块用于指导训练, 在策略中, 在每个时间步 t 智能体 i 的局部观测状态 o_i^t 和上一个时间步的动作 a_i^{t-1} 拼接, 然后送入一个多层感知(multilayer perception, MLP)层得到局部观测状态表示 $x_{i,t}$. 接着, $x_{i,t}$ 被输入到一个 GRU 层, 得到融合历史状态信息特征表示 o_i^t. 最后, 一个 Q_{net} 层将历史状态信息特征表示 o_i^t 作为输入并输出智能体的动作价值. 在队友模型中, 通过智能体的局部观测 o_i^t 和队友编号 d_j 进行编码, 处理后的数据作为多元高斯分布的参数进行处理, 然后采样结果, 得到队友信息, 在拿到队友信息之后本文对队友信息进行 AHP 分析, 最后得到针对智能体 j 的互信息 x_{ij}, 互信息直接作用于队友的 Q 值, 最后将所有的 Q 值输入到混合网络. 混合网络接受每个智能体的局部 Q

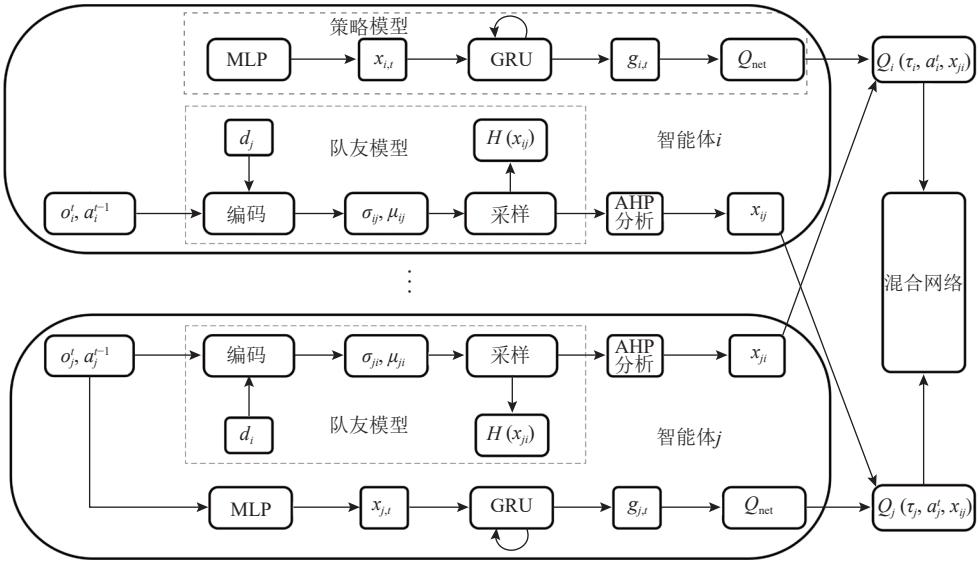


Fig. 2 VHBTM algorithm framework

图 2 VHBTM 算法框架

值作为输入，并输出一个代表整个团队的全局 Q 值. 这个全局 Q 值用于指导所有智能体的联合行动，确保所选行动最大化整个团队的长期收益.

为了解决完全分布式的执行方案不能协调智能体行为的问题，本文设计了队友模型，该模型使智能体能够根据自身观察和队友的标识来预测队友的行为，并据此在自己的观察中提供关于队友行为的建议，通过队友模型产生的互信息直接影响智能体的价值函数. 这种方法避免了隐私问题，并且没有扩大智能体的观察空间，从而避免了训练过程的复杂化. 同时，队友模型又能避免智能体只能根据自身的局部观察来做出基站选择决策的局限性.

同时，为了确保互信息的有效性，即队友车辆提供的行为建议是有意义的，本文还需要利用 AHP 模型来分析互信息的权重. 权重分析包括考虑智能体之间的距离、智能体的速度和信噪比，这些因素影响通信切换的可能性及智能体间局部观察的重叠程度. 当车辆发生通信切换的可能性比较大，并且队友之间可选的基站范围存在重叠部分，此时车辆之间的通信切换的相互影响更大，彼此之间的互信息应该有更高的权重，通过这种权重分析，本文可以剪除无用的互信息，防止冗余信息的干扰和信道资源的浪费.

在传统的多智能体强化学习中，合作程度通常被定义为所有智能体所获得奖励的总和. 尽管这种定义方式提供了灵活的训练方法，也并不能充分反映智能体之间的合作程度，因为智能体往往更关注最大化自身的利益. 为了更准确地体现智能体之间的合作，本文在系统中设计了激励相容模块. 本文希望智能体在追求个人目标的同时，也能考虑到整体策略的协调性，从而实现对所有智能体有益的状态. 这种方法成功地克服了传统方法的局限，提升了系统的整体吞吐量和效率.

4.2 队友模型

为了指导队友的决策，VHBTM 框架学习了针对特定智能体的队友模型，利用这些模型基于单个智能体的局部观察来指导队友的决策. 为了描述智能体 i 与智能体 j 之间的相互依赖性，即智能体 i 的行为如何影响其队友，本文使用多元高斯分布来建模队友之间的关系. 智能体 i，通过其局部观察 o_i^t 和队友的标识符 d_j 来获取队友模型. 编码器接收原始数据并将其转换为特征向量 $\boldsymbol{v}_{ij}$，该向量直接用于参数化多元高斯分布. 本文对处理过的数据进行采样，采样结果是智能体 i 关于智能体 j 的互信息表示 x_{ij}. 为了提高队友模型的准确性和增强智能体间协作的效果，本文引入了正则化机制，这种机制有助于明确指导队友模型的生成过程，确保模型能够更精确地反映智能体之间的相互作用. 本文使用信息理论中的熵(entropy)和条件熵(conditional entropy)的概念来量化智能体之间的互信息. 熵用于衡量智能体对环境状态的不确定性，而条件熵则描述了在已知某些条件的情况下系统不确定性的减少.

通过计算智能体间互信息的增益，本文可以定量地评估智能体之间信息交换的有效性，从而在学习过程中引导智能体学习如何更好地协调行动，优化队友模型效果. 具体地，本文使用熵 $H(x_{ij}|o_i,d_j)$ 来表示在智能体 i 的局部消息 o_i^t 和队友 j 的身份 d_j 已知的情况下，智能体 j 的队友表示 x_{ij} 的不确定性. 此外，条件熵描述了在智能体 i 知道队友 j 的行动 a_j 的情况下 x_{ij} 的不确定性. 互信息 I 被定义为这 2 个熵之间的差异，计算公式为

$$I(x_{ij},a_j|o_i,d_j)=H(x_{ij}|o_i,d_j)-H(x_{ij}|o_i,a_j,d_j). \tag{6}$$

如果智能体 j 的行动 a_j 没有提供关于互信息 x_{ij} 的额外信息，那么条件熵将等价于无条件熵，这种情况下 $H(x_{ij}|o_i,d_j)=H(x_{ij}|o_i,a_j,d_j)$. 这时候，互信息变为 0. 本文的目标是最大化互信息，这相当于在智能体 i 的局部信息条件下最小化学习到的队友模型的不确定性. 这使得智能体 i 在建模队友 j 时能够获得更准确和更有力的队友表示 x_{ij}. 在生成互信息之后，提出的框架需要对互信息进行权重分析来评估其对队友智能体的有效性. 由于分析权重需要综合考虑车辆的速度、信噪比等因素，本文将权重分析问题建模为一个多个互信息的权重分配问题，为了使分析准确高效，本文设计了 AHP 来进行权重分析.

AHP 的核心在于将复杂的决策问题分解为层次结构中的多个更小、更易评估的部分. 首先，本文定义决策的目标，即评估互信息的有效性. 由于互信息传递的目的是协调智能体之间的基站选择，本文考虑的子标准应直接和通信切换相关，同时智能体可选择的基站范围应存在重合部分，即智能体的选择相互影响，受文献 [16] 的启发，本文将子标准分为：速度、信噪比和距离. 其中速度和信噪比直接影响通信切换的频率，距离越近，多智能体的通信切换决策对彼此的影响越大. 另外 AHP 的成对比较法，要求决策者评估各因素对达成总目标的相对重要性. 在这一过程中，决策者需要对每对因素进行比较，判断它们对实现决策目标的贡献大小，并给出一个定量评

分，通常是 1~9 的标度，其中“1”表示 2 个因素同等重要，“9”表示一个因素对目标贡献极大地超过另一个因素. 本文将速度和信噪比的评分定为 1，距离的评分定为 2. 这些成对比较的结果被整合到一个成对比较矩阵中，随后通过数学方法计算出每个因素的权重，这些权重表示了每个因素在整个决策过程中的相对重要性. 通过一致性检验，本文确保这些评估的一致性和逻辑性，避免了决策偏差. 我们构建 AHP 的具体步骤如下：

1)构建层次结构. 本文的目标是计算车辆之间互信息的权重，考虑 3 个标准：速度、信噪比和距离. 备选层由不同的队友车辆组成.

2) 构建 AHP 判断矩阵. 本文使用 1~9 的量表来量化这些关系，其中“1”表示 2 个元素同等重要，更高的值表示更大的重要性. 基于上述描述，本文可以构建一个成对比较矩阵. 比较结果呈现在一个称为 AHP 矩阵的方阵 $\boldsymbol{X}=(w_{ij})_{n\times n}$ 中，其中 $w_{ij}=1$, $(i,j=1,2,\cdots,n)$, 这里的 n 表示标准的数量，而 w_{ij} 是根据表 1 得到的.

Table 1 Simulation Parameters

表 1 仿真参数

参数	值
仿真地图大小	1 000 m×1 000 m
车道数量	双向四车道
宏基站总带宽/MHz	20
宏基站覆盖范围/km	2
毫米波基站总带宽/MHz	100
毫米波基站覆盖范围/m	200
宏基站噪声功率密度/(dBm·Hz^{-1})	–174
毫米波基站噪声功率/(dBm·Hz^{-1})	–164
宏基站发射功率/W	120
毫米波基站发射功率/W	3

3)计算每个元素的权重. 本文使用以下公式计算每个标准的几何平均数 GM_i：

$$GM_i=\left(\prod_{j=1}^{n}a_{ij}\right)^{\frac{1}{n}}, \tag{7}$$

其中 a_{ij} 代表成对比较矩阵的元素，n 是标准的数量，然后通过式(8)对标准的权重 W_i 进行归一化：

$$W_i=\frac{GM_i}{\sum_{k=1}^{n}GM_k}, \tag{8}$$

其中 GM_i 是第 i 个标准的几何平均数，分母是所有几何平均数的总和.

4)一致性检验. 评估 AHP 中成对比较矩阵的一致性，需要进行一致性检验. 首先计算矩阵的最大特征值 $\lambda_{\max}$，然后计算一致性指数(CS)：

$$CS=\frac{\lambda_{\max}-n}{n-1}, \tag{9}$$

其中 n 是矩阵的阶数，对应于标准的数量. 随机一致性指数(RI)是从随机生成的矩阵得到的预定平均一致性指数，其值根据标准的数量而变化. 一致性比率(CR)的计算公式为：

$$CR=\frac{CS}{RI}. \tag{10}$$

如果 $CR\leqslant 0.1$，则被认为是可以接受的，表明矩阵具有满意的一致性；如果 $CR>0.1$，则应审查并修正成对比较矩阵中的判断.

4.3 算法训练

激励相容约束在多智能体系统的通信切换中扮演重要角色. 这些约束确保所有智能体根据其实际的通信需求和能力作出决策，从而提升整个系统的效率和效果. 为了将这些约束整合进优化目标，本文采用了拉格朗日乘数法，形成了以下的双重目标：

$$J(C_t|i\in C_t)=\max_{\pi_i^M}\left[\sum_{j\in C_t}V_j(o_{j,t},C_t)+\lambda_t^i(o_{i,t})\cdot V_i(o_{i,t},C_t)\right], \tag{11}$$

其中，C_t 表示时间步 t 时车辆的集合，π_t^M 表示所有智能体在时间步 t 的混合策略，$V_j(o_{j,t},C_t)$ 指的是给定其观察 $o_{j,t}$ 和智能体集合 C_t 时智能体 j 的价值函数. 而 $\lambda_t^i(o_{i,t})$ 是智能体 i 的拉格朗日乘数，强调了智能体 i 在团队中的影响，通过调整这个权重，可以模拟智能体 i 行动的变化对团队其他成员的潜在影响. 确保在优化目标中考虑智能体 i 的激励相容约束. 本文使用策略梯度来最大化该目标：

$$\begin{aligned}L\left(\theta_i^M\right)=-E\Big[&\sum_{j\in C_t}(r_{j,t}+\gamma V_{\sigma,j}(o_{t+1},C_{t+1})-V_{\sigma,j}(o_t,C_t))+\\&\lambda_t^i(o_{i,t})(r_{i,t}+\gamma V_{\sigma,i}(o_{t+1},C_{t+1})-\\&V_{\sigma,i}(o_t,C_t))\Big]\log\pi_{\theta_i^M}(o_{i,t},C_t),\end{aligned} \tag{12}$$

式(12)表示的是给定策略下长期回报的期望值，包括了所有智能体 j 的累计损失，$r_{j,t}$ 代表智能体 j 在时间步 t 获得的即时奖励，$V_{\sigma,j}(o_{t+1},C_{t+1})$ 代表智能体 j 预测在下一个状态 o_{t+1} 下，根据当前的策略，团队预期的未来价值. 智能体 i 在当前和下一时间步的预期价值差，本文使用 $\lambda_t^i(o_{i,t})$ 进行调整，通过策略梯度优化策略，旨在找到一个平衡点，让智能体在追求个人

最大利益的同时,也能考虑整体团队的效益,从而优化整个团队的表现.为了更新拉格朗日乘数,本文采用了对偶梯度下降法:

$$\log\lambda_t^i(o_{i,t})=\log\lambda_t^i(o_{i,t})-\mu\left(V_\sigma^M(o_t,C_t)-V_\sigma^M(o_t,C_{t\backslash\{i\}})\right). \tag{13}$$

这个更新过程是为了平衡智能体 i 对整个系统的贡献,以确保激励兼容性.μ是学习率控制更新的步长,决定乘数更新的速度.本文利用考虑智能体 i 和不考虑智能体 i 时的预测函数差值来反映智能体 i 对整体系统的影响,表示为 $V_\sigma^M(o_t,C_t)-V_\sigma^M(o_t,C_{t\backslash\{i\}})$,当智能体 i 存在时系统性能提升或降低,通过这个差值来量化.拉格朗日乘数作为一种权重,对智能体 i 的行为进行奖惩,以使其行为与集体目标一致.当智能体 i 的存在导致系统整体价值增加时,通过减小拉格朗日乘数,减少对其行为的奖励,促使其行为更加注重集体利益.相反,如果智能体 i 的行为导致系统价值降低,通过增大拉格朗日乘数,增加对其行为的惩罚,迫使其改变行为以更好地符合集体目标.通过动态调整每个智能体的拉格朗日乘数,可以灵活地调节每个智能体对系统性能的贡献,使得每个智能体在追求个体最优策略的同时,不会损害到系统的整体效益.

本文在算法 1 中详细介绍了 VHBTM 的整体训练算法.算法 1 旨在优化多智能体环境中的团队合作和决策过程,以高效地实现共同目标.算法 1 从初始化模型参数θ和ϑ开始.此外,还初始化了回放缓冲区 B,用于存储智能体的观察、动作和奖励数据(行①).环境 S 也为训练过程做好准备(行②).在训练期间(行③~⑰),对于总时间步 T 内的每一个时间步 t(行④~⑫),算法会从环境中收集一系列的观察 $o_{1,t},\cdots,o_{n,t}$,(行⑤).然后基于这些观察和队友索引 j,使用队友模型生成互信息 x_{ij}(行⑥).计算 x_{ij} 的权重 w_{ij}(行⑦),并应用于 Q_j,影响动作选择过程(行⑧).当前价值 Y_{tot} 定义为即时奖励 r 和折现未来奖励的总和,后者由 $\gamma\max Q_{\text{tot}}$ 给出(行⑨).通过根据总预期奖励调整算法优化策略.执行动作 $a_{1,t},\cdots,a_{n,t}$ 并收集奖励(行⑩),所有相关数据包括观察、动作和奖励都存储在回放缓冲区 B 中(行⑪),以供未来参考.在后续的训练周期中(行⑬~⑰),更新每个智能体的目标网络参数 θ 以反映最新学习(行⑭).还对拉格朗日乘数 λ 进行调整(行⑮),优化团队的整体表现.通过持续更新策略并适应环境,算法 1 旨在提升多智能体系统中的团队合作和决策能力.

算法 1. VHBTM 训练算法.

输入:环境状态空间 S,智能体动作空间 A,奖励函数 r,每个回合的最大时间步 T,折扣因子 γ,学习率 η;

输出:集合的价值函数 Q_{tot} 代表所有智能体的性能.

① 初始化网络参数 θ, ϑ 和经验回放缓冲区 B;
② 初始化训练环境;
③ while 训练未结束 do
④ for 1~T 时间步 do
⑤ 从环境中收集一系列观察值 $o_{1,t},\cdots,o_{n,t}$;
⑥ 使用本地观测和队友的编号 j 以及队友模型生成互信息 x_{ij},由式(6)计算互信息 I;
⑦ 由式(7)~(10)计算 x_{ij} 的权重 w_{ij};
⑧ 将加权过后的 x_{ij} 应用到队友的价值函数 Q_j;
⑨ 定义当前价值为 $Y_{\text{tot}}\leftarrow r+\gamma\max Q_{\text{tot}}$;
⑩ 在环境中执行动作 $a_{1,t},\cdots,a_{n,t}$ 并收集奖励 $r_{1,t},\cdots,r_{n,t}$;
⑪ 将 $(o_{1,t},\cdots,o_{n,t},a,r_{1,t},\cdots,r_{n,t})$ 存储到经验缓冲区 B;
⑫ end for
⑬ for 训练轮次 do
⑭ 由式(11)(12)更新每个智能体的目标网络参数 θ;
⑮ 由式(13)更新拉格朗日乘数 λ;
⑯ end for
⑰ end while

5 仿真实验与分析

本节首先介绍了本文使用的场景和基准方法,以及所采用的参数设置.然后,在包括智能联网车辆和传统车辆的混合交通场景中评估了提出的 VHBTM 框架的性能.

5.1 仿真设置

本文假设智能体在每个学习阶段开始时随机分布在区域内,位置随机但保证均匀分布,采用随机方向模型来模拟智能网联汽车的移动模式.智能体的速度在 3~20 m/s 的范围内均匀分布,模拟慢速至快速移动车辆的各种速度.此外,模拟区域内还随机行驶着 30 辆无关的车辆,它们的路径和速度也是随机生成,增加环境的复杂性和不可预测性.

本文在一个城市区域内考虑了一个包含 1 个宏基站和 4 个毫米波基站的两层异构网络.宏基站位于模拟区域的中心,具体坐标为(0, 0),覆盖距离可达

1 000 m, 总带宽设为 20 MHz. 4 个毫米波基站均匀分布在宏基站的周围, 每个基站的坐标分别设定为(400, 0), (–400, 0), (0, 400)和(0, –400), 毫米波基站提供更高的数据传输速率, 总带宽设为 100 MHz. 宏基站的噪声功率密度设定为–174 dBm/Hz, 毫米波基站的噪声功率密度则设定为–164 dBm/Hz, 以反映真实城市场景信号的传播特性. 毫米波基站和宏基站的默认总发射功率分别为 $PS = 3$ W 和 $PM = 120$ W.

本文工作与以下基准算法进行比较:

1) QMIX.由 Rashid 等人[29] 提出, QMIX 是一种用于强化学习中多智能体协作问题的算法, 特别适用于需要智能体共同合作以实现共同目标的场景. 它是值分解网络(VDNs)的扩展, 旨在解决单一全局价值函数在复杂多智能体环境中不够有效的问题. QMIX 的核心思想是将全局价值函数(Q 值)分解为每个智能体的个体贡献, 同时确保整体策略优于或等于任何智能体策略子集的组合.

2) QPLEX. 由 Wang 等人[30] 提出, QPLEX 是专门为多智能体系统中的合作问题设计的. 它在 QMIX 的基础上进行了发展, 旨在解决 QMIX 在处理高度动态和复杂交互环境时面临的一些限制. QPLEX 引入了双重对决网络结构, 增强了对智能体间关系的建模能力, 从而更有效地学习多个智能体之间的合作策略.

3) NDQ. 由 Wang 等人[31] 提出, NDQ 是对值分解多智能体强化学习算法的改进. 现有方法侧重于学习完全去中心化的价值函数, 这对于需要通信的任务并不高效. 为了解决这个限制, NDQ 引入了一种通信机制, 其中智能体偶尔向其他智能体发送消息以实现有效的协调. 此外, NDQ 还引入了正则化器, 以最大化本地 Q 函数与用于通信的消息之间的互信息, 并最小化消息的信息熵.

4) MAPPO. 由 Yu 等人[32] 提出, MAPPO 是将众所周知的单智能体 PPO 算法扩展到多智能体环境中, 以适应合作和竞争状态. 通过引入平滑的策略更新机制, MAPPO 旨在增强学习过程的稳定性, 并促进智能体之间的有效合作, 共同应对复杂任务. 这种方法侧重于保持单智能体训练的优势, 同时探索如何通过集中训练和分散执行的框架实现智能体之间的协调和通信.

5) MAIC. 由 Yuan 等人[33] 提出, MAIC 是一种基于通信的多智能体强化学习算法, 引入了一种新颖的队友建模网络和消息生成器, 能够提供更有效的通信方式. 智能体学习预测队友的行动选择, 并基于此向它们传递信息. 此外, 加入一个正则项来限制信息传输, 提高传输效率.

5.2 数值分析

图 3 展示了本文的算法与其他算法在相同异构网络环境下的吞吐量比较. 数据表明, 本文的 VHBTM 算法明显优于其他算法, 而 QMIX 在此测试中表现最差. QMIX 尽管考虑了智能体之间的合作, 但它的单调性约束未能充分考虑智能体之间的复杂交互, 这限制了其在联合价值函数中表达复杂策略的能力. 当智能体数量增加时, 智能体间交互的复杂性增加, QMIX 的吞吐量逐渐下降, 从开始的快速增长至最终在 150 Mbps 左右稳定.

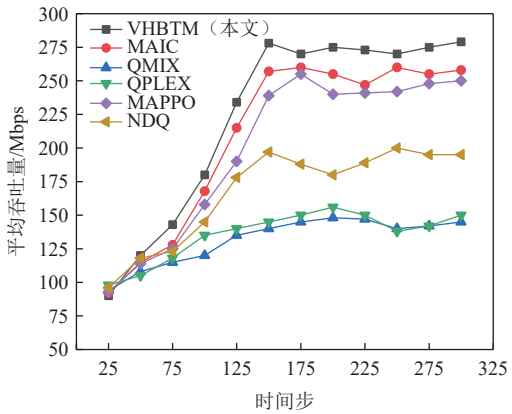


Fig. 3 Throughput variation over time steps

图 3 吞吐量随时间步的变化

相比之下, 尽管 QPLEX 在 QMIX 的基础上提出了新的价值分解方法, 并通过双重对决结构增强了智能体之间策略协同的一致性, 但从图 3 看出, QPLEX 和 QMIX 在实际表现上差距不大, 均在 150 Mbps 上下波动, 这是因为 QPLEX 通过双重对决网络来尝试分解总体价值函数为个体智能体的贡献, 然而这种结构在实际操作中不能完全捕捉智能体间复杂的交互动态. 尤其在环境中智能体的行为依赖关系非常强烈时, 简单的双层结构可能无法充分表达这种复杂性. 但当智能体数量变多, 算法需要处理的状态和行动空间急剧膨胀, 这可能导致学习效率下降和策略质量变差. 大规模环境中的节点间依赖和相互作用更加复杂, QPLEX 的表现可能受到影响.

NDQ 旨在通过引入通信机制来解决智能体之间的协调问题. 这种算法通过允许智能体间偶尔的信息交流, 试图提高整个系统的决策质量和效率. 相对于 QMIX, NDQ 在复杂环境中的决策能力得到了提升, 算法最终稳定在 200 Mbps 左右, 然而, 尽管 NDQ 的设计初衷是通过增强智能体间的协同作用来改进

性能, 其效果却受到了限制, NDQ 使用了 2 种正则化器来改变智能体的决策过程, 这些正则化器旨在最小化发送的消息量. 这种方法虽然减少了通信量, 但同时也限制了智能体决策时可以利用的策略范围. 这种限制可能会阻碍智能体在学习过程中的效率. 此外, 由于消息内容本身的不确定性, 智能体在复杂环境中使用这些消息做出决策的能力也受到了影响.

MAPPO 是基于单智能体的 PPO 算法扩展而来的. 单智能体 PPO 因其在连续动作空间中的稳定性和有效性而广受欢迎, 特别是通过截断的重要性采样技术解决策略更新过程中的高方差问题. 然而, 当这些技术应用于多智能体合作环境时, MAPPO 虽然试图维持 PPO 的优点, 但其完全去中心化的执行策略在多智能体协调和合作方面存在局限, 尤其是在需要精细调整和同步的复杂场景中, 这种方法仍存在一定的不足. 因为多智能体环境中智能体间的相互依赖和动态交互增加了策略协调的复杂性, 而 MAPPO 的去中心化执行可能无法有效处理这些复杂的交互动态. MAIC 专注于解决通信问题, 它有一个定制化的消息机制, 相对于其他的方案, MAIC 可以更好地适应多智能体合作环境, 最终吞吐量在 260 Mbps 左右, 但它没有考虑个体目标与整体目标之间的平衡, 导致智能体可能选择更高的短期利益, 忽视了整体利益.

VHBTM 引入了队友模型, 该模型使得智能体能够基于自身的局部观察以及对队友行为的预测来做出决策. 通过这种方式, 智能体不仅可以根据当前的通信需求选择最合适的基站, 而且可以预见并适应复杂的动态环境. 这种预测能力是传统单智能体和简单多智能体算法所缺乏的, 因为它们通常只响应当前的环境状态而不考虑长期的策略和队友行为的影响. 通过 AHP 来确定信息的权重, 确保了信息交换的效率和目标性, 有效避免了信息冗余和资源浪费. 此外, VHBTM 引入激励相容机制确保每个智能体在追求个人最优策略的同时也能兼顾整体的目标, 这对于维持系统中所有智能体的协调一致性至关重要. 这种机制能有效地调和个体利益与集体利益的矛盾, 从而促进更协调和有效的群体决策.

图 4 展示了随着权重系数 α 变化时不同算法的吞吐量表现, VHBTM 在较高的 α 值时表现出强大的适应性和鲁棒性. MAIC 表现出与 VHBTM 类似的趋势, 增长幅度约为 8.3%. 其他算法对 α 的增加不敏感, 其性能基本稳定(增长幅度从 4.61% 升到 7.13%), 这表明每个算法在处理权重系数增加时采用了不同的策略, 在高度动态的环境中, 智能体之间的关系和相互作用可能迅速变化, QMIX 和 QPLEX 都不能很好地适应环境, 虽然 QPLEX 试图通过价值分解让每个智能体的行动反映其对全局目标的贡献, 但实际中可能难以确保所有关键信息都被有效整合. 这导致算法在实际应用中未能充分利用所有可用信息, 从而影响决策质量.

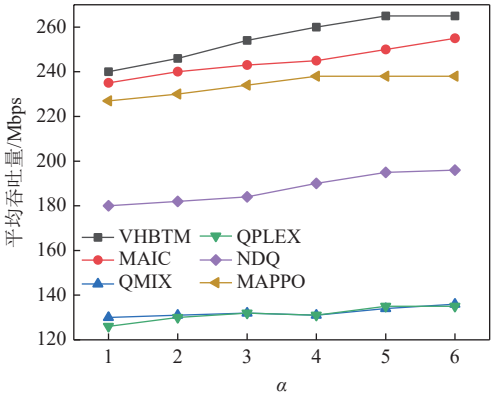


Fig. 4 Throughput varies with weight coefficient

图 4 吞吐量随权重系数的变化

图 5 展示了随着通信切换权重系数 β 的变化, 各种算法的通信切换发生的次数. 随着 β 的增加, 通信切换次数有所下降. 当权重系数开始上升时, 通信切换次数显著下降(范围从 8.3% 到 20%), 并在降至一定程度后趋于稳定. 不同算法对参数的敏感性各不相同. VHBTM 和 MAIC 在减少不必要的通信切换方面更为有效. 图 4 和 图 5 的结果为本文调整通信切换和吞吐量的平衡提供了有意义的参考.

表 2 展示了智能体的数量对算法吞吐量的影响, 可以看出随着智能体数量的增加, 智能体之间的交互变得频繁, 吞吐量都有不同程度的减少, 其中 VHBTM 表现得最好. 在智能体的数量偏少时各个算

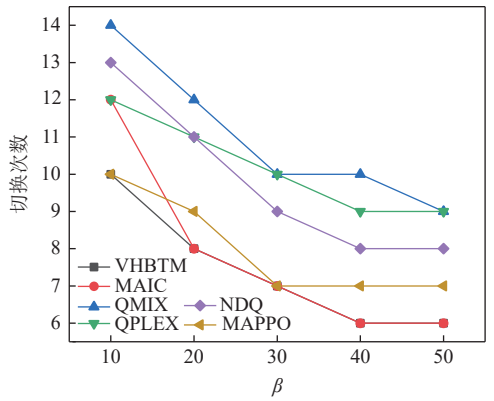


Fig. 5 Variation of handover number with weight coefficient

图 5 切换次数随权重系数的变化

法的性能几乎没有明显的区别，但是当智能体数量增加一倍时 VHBTM 的吞吐量只下降了约 13%，说明当环境变得复杂时，本文提出的算法更有优势.

Table 2 Effect of Different Numbers of Agents on Algorithms' Throughput

表 2 不同数量的智能体对算法吞吐量的影响 Mbps

智能体数量	VHBTM	MAIC	QMIX	QPLEX	NDQ	MAPPO
5	314	312	294	310	300	276
7	281	262	171	191	230	249
10	271	255	136	136	196	238

图 6 展示了不同算法随时间变化的通信切换趋势，在开始阶段算法处于探索状态，通信切换次数较多，随着时间步的增加，算法通信切换次数减少，不同的下降速率反映了算法对环境的适应性和算法具有很快的收敛速度. 其中，VHBTM(57.3%)的表现最为突出，而 QMIX 和 QPLEX 虽然也能随时间改善性能，但它们达到性能平衡点的速度和稳定性较差.

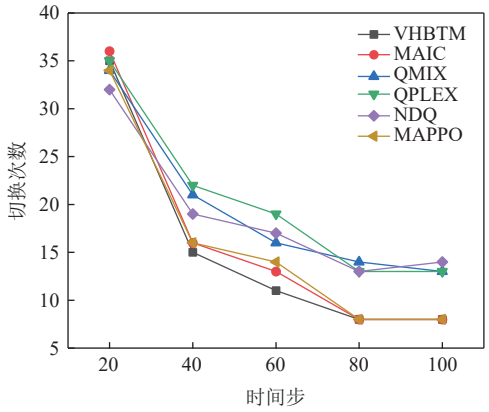


Fig. 6 Handover number varies with time steps

图 6 切换次数随时间步的变化

图 7 展示了随着传统车辆数量的增加，不同算法的吞吐量变化情况. 当车辆增加时，需要更多的车辆间交互，基站可用资源将变得更少，算法的吞吐量将显示不同程度的降低. 随着环境变得更加复杂，智能体间的交互变得更加频繁，尽管不同算法在较少智能体场景中有类似的表现，QMIX 和 QPLEX 不能很好地适应复杂的交互场景，吞吐量下降最多(分别为 53.74% 和 52.12%)，而 NDQ 尽管具有通信机制，但由于通信机制的限制性，其对环境的适应性有限，导致在其他车辆数量增加到一定水平时，吞吐量迅速下降至 150 Mbps 以下. 因此，VHBTM 更适合于具有较多智能体的通信切换的场景.

图 8 和图 9 展示了当基站的数量增多时，智能体

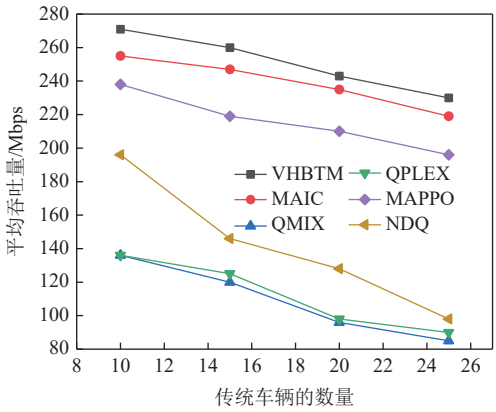


Fig. 7 Impact of traditional vehicle numbers on throughput

图 7 传统车辆数量对吞吐量的影响

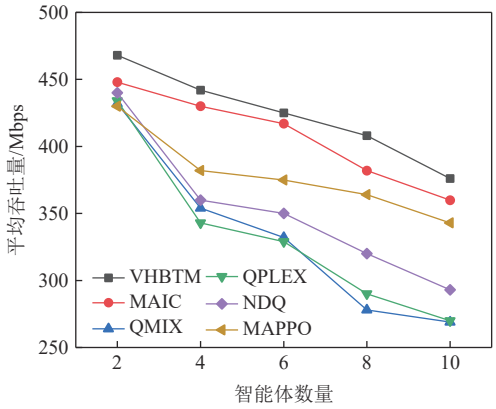


Fig. 8 Impact of the number of agents on throughput when increasing the number of base stations

图 8 基站数量增加时智能体数量对吞吐量的影响

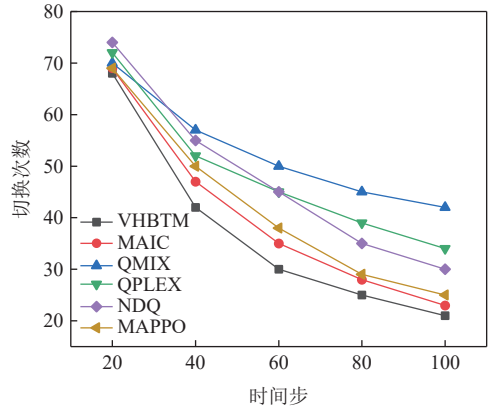


Fig. 9 Variation of handover number with time step when increasing the number of base stations

图 9 基站数量增加时切换次数随时间步的变化

数量对吞吐量的影响和通信切换次数随时间步的变化情况. 本文在模拟中增设了坐标分别为(200, 200)，(−200, −200)，(−200, 200)，(200, −200)的 4 个基站，这样的配置旨在模拟一个更加密集的网络环境.

在图 8 中，基站数量的增加使得环境的复杂度上

升, 相应地, 智能体需要在更大的策略空间中进行学习和决策, 这增加了学习的难度. 从数据中可以观察到, 当智能体数量增加到 6 个时, QMIX 和 QPLEX 算法的性能显著下降, 吞吐量从 430 Mbps 左右降至 340 Mbps 左右. 这种现象表明, 尽管 QMIX 和 QPLEX 通过分解全局价值函数来处理多智能体间的协作问题, 在智能体数量较少时这种方法效果显著. 但是, 随着智能体数量的进一步增加, 这种基于全局价值函数的分解方法难以准确捕捉和处理智能体间日益复杂的相互作用和影响, 导致性能下降.

在图 9 中, 观察到随着基站数量的增加, 智能体在进行通信切换时的次数也显著增加. 这种增加主要是由于更多的基站增加了网络的复杂性, 智能体需要在更多的基站之间做出选择. 在这种情况下, 本文提出的 VHBTM 算法显示出了优越的适应能力, 它通过有效地管理智能体之间的交互和协调, 能够在更加复杂的网络环境中保持稳定的通信性能. 这表明 VHBTM 算法不仅能够处理基本的通信场景, 还能够在基站数量增多导致的复杂交互环境中实现更高效的协作和优化通信切换策略, 显著提高整体网络的操作效率和智能体的通信质量.

6 结束语

本文提出一种基于队友模型的合作型强化学习框架. 首先将 C-V2X 异构网络场景中通信切换和资源利用不均匀问题转化为一个多智能体马尔可夫博弈问题, 并将车辆建模为智能体. 具体来说, 为了解决车辆在异构网络场景中的通信切换问题, 本文设计了一个在 QMIX 基础上改进的多智能体架构自主学习通信切换策略. 针对通信资源利用不均匀问题, 本文首先设计了一个队友模型, 该模型利用智能体的本地观察和队友编号对特定的队友生成互信息, 为队友推荐高 QoS 和稳定 QoE 的基站. 接着, 本文提出了一个动态权重分配方案, 通过速度、信噪比和智能体之间的距离筛选更有用的互信息, 互信息加权后直接作用在队友的价值函数上, 协助智能体选择最优的基站, 避免资源利用不均. 此外, 本文设计了激励相容训练算法, 以协调智能体的个体目标与集体目标, 当个体目标偏离集体目标的时候, 我们使用拉格朗日乘数放大这种影响, 以此促进智能体之间的合作. 仿真结果显示, 本文提出的算法在 C-V2X 异构网络场景中优于最好的基线方法, 展示了良好的吞吐量和通信切换性能. 未来我们会结合空天地一体化网络, 在面向卫星通信、低空通信和地面通信等多种场景, 针对我们的算法优化频谱分配和管理策略.

作者贡献声明: 刘冰艺提出了总体方案和技术路线; 王东东和施海勇设计了实验并撰写论文; 王恩澍完善了技术方案和实验方案; 吴黎兵和汪建平提出论文指导意见并修改论文.

参 考 文 献

[1] Olwal T O, Djouani K, Kurien A M. A survey of resource management toward 5G radio access networks[J]. IEEE Communications Surveys & Tutorials, 2016, 18(3): 1656−1686

[2] Qi Yanli, Zhou Yiqing, Liu Ling, et al. MEC coordinated future 5G mobile wireless networks[J]. Journal of Computer Research and Development, 2018, 55(3): 478−486 (in Chinese)
(齐彦丽, 周一青, 刘玲, 等. 融合移动边缘计算的未来 5G 移动通信网络[J]. 计算机研究与发展, 2018, 55(3): 478−486)

[3] Zhou Yuxuan, Yang Xu, Qin Chuanyi, et al. HYBRID-D2D-MIMO (HDM) network architecture and hybrid data distributing strategy (HDDS)[J]. Journal of Computer Research and Development, 2020, 57(9): 1911−1927 (in Chinese)
(周玉轩, 杨絮, 秦传义, 等. HDM 网络架构与混合式数据分发策略[J]. 计算机研究与发展, 2020, 57(9): 1911−1927)

[4] Liu Ling, Zhou Yiqing, Zhuang Weihua, et al. Tractable coverage analysis for hexagonal macrocell-based heterogeneous UDNs with adaptive interference-aware CoMP[J]. IEEE Transactions on Wireless Communications, 2018, 18(1): 503−517

[5] Liu Bingyi, Han Weizhen, Wang Enshu, et al. Multi-agent attention double actor-critic framework for intelligent traffic light control in urban scenarios with hybrid traffic[J]. IEEE Transactions on Mobile Computing, 2023, 23(1): 660−672

[6] Wang Enshu, Liu Bingyi, Lin Songrong, et al. Double graph attention actor-critic framework for urban bus-pooling system[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(5): 5313−5325

[7] Abuajwa O, Roslee M B, Yusoff Z B. Simulated annealing for resource allocation in downlink NOMA systems in 5G networks[J]. Applied Sciences, 2021, 11(10): 4592

[8] Ibrahim L F, Salman H A, Taha Z F, et al. A survey on heterogeneous mobile networks planning in indoor dense areas[J]. Personal and Ubiquitous Computing, 2020, 24: 487−498

[9] Lü Huazhang, Chen Dan, Fan Bin, et al. Standardization progress and case analysis of edge computing[J]. Journal of Computer Research and Development, 2018, 55(3): 487−511 (in Chinese)
(吕华章, 陈丹, 范斌, 等. 边缘计算标准化进展与案例分析[J]. 计算机研究与发展, 2018, 55(3): 487−511)

[10] Rappaport T S, Xing Y, MacCartney G R, et al. Overview of millimeter wave communications for fifth-generation (5G) wireless networks—With a focus on propagation models[J]. IEEE

Transactions on Antennas and Propagation, 2017, 65(12): 6213−6230

[11] Jameel F, Haider M A A, Butt A A. Massive MIMO: A survey of recent advances, research issues and future directions[C]//Proc of 2017 Int Symp on Recent Advances in Electrical Engineering (RAEE). Piscataway, NJ: IEEE, 2017: 1−6

[12] Stojmenovic I. Handbook of Wireless Networks and Mobile Computing[M]. Hoboken, NJ: John Wiley & Sons, 2002

[13] Kim Y, Kim Y, Oh J, et al. New radio (NR) and its evolution toward 5G-advanced[J]. IEEE Wireless Communications, 2019, 26(3): 2−7

[14] Marquez-Barja J M, Ahmadi H, Tornell S M, et al. Breaking the vehicular wireless communications barriers: Vertical handover techniques for heterogeneous networks[J]. IEEE Transactions on vehicular Technology, 2014, 64(12): 5878−5890

[15] Isa I N M, Baba M D, Yusof A L, et al. Handover parameter optimization for self-organizing LTE networks[C]//Proc of 2015 IEEE Symp on Computer Applications & Industrial Electronics (ISCAIE). Piscataway, NJ: IEEE, 2015: 1−6

[16] Alhabo M, Zhang Li, Nawaz N. GRA-based handover for dense small cells heterogeneous networks[J]. IET Communications, 2019, 13((13):): 1928−1935

[17] Hasan M M, Kwon S, Oh S. Frequent-handover mitigation in ultra-dense heterogeneous networks[J]. IEEE Transactions on Vehicular Technology, 2018, 68(1): 1035−1040

[18] Yang Bing, Yang Xuan, Ge Xiaohu, et al. Coverage and handover analysis of ultra-dense millimeter-wave networks with control and user plane separation architecture[J]. IEEE Access, 2018, 6: 54739−54750

[19] Simsek M, Bennis M, Güvenc I. Context-aware mobility management in HetNets: A reinforcement learning approach[C]//Proc of 2015 IEEE Wireless Communications and Networking Conf (WCNC). Piscataway, NJ: IEEE, 2015: 1536−1541

[20] Qi Weijing, Song Qingyang, Wang Shupeng, et al. Social prediction-based handover in collaborative-edge-computing-enabled vehicular networks[J]. IEEE Transactions on Computational Social Systems, 2021, 9(1): 207−217

[21] Luong N C, Hoang D T, Gong S, et al. Applications of deep reinforcement learning in communications and networking: A survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3133−3174

[22] Wang Zhi, Li Lihua, Xu Yue, et al. Handover control in wireless systems via asynchronous multiuser deep reinforcement learning[J]. IEEE Internet of Things Journal, 2018, 5(6): 4296−4307

[23] Nasim I, Skrimponis P, Ibrahim A S, et al. Reinforcement learning of millimeter wave beamforming tracking over COSMOS platform[C]// Proc of the 16th ACM Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization. New York: ACM, 2022: 40−44

[24] Wei Yao, Lung C H, Ajila S, et al. Deep Q-networks assisted pre-connect handover management for 5G networks[C]//Proc of 2023 IEEE 97th Vehicular Technology Conf (VTC2023-Spring). Piscataway, NJ: IEEE, 2023: 1−6

[25] Lin Yan, Zhang Zhengming, Huang Yongming, et al. Heterogeneous user-centric cluster migration improves the connectivity-handover trade-off in vehicular networks[J]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 16027−16043

[26] Lee C, Jung J, Chung J M. Intelligent dual active protocol stack handover based on double DQN deep reinforcement learning for 5G mmWave networks[J]. IEEE Transactions on Vehicular Technology, 2022, 71(7): 7572−7584

[27] Wang Ruiyu, Sun Yao, Zhang Chao, et al. A novel handover scheme for millimeter wave network: An approach of integrating reinforcement learning and optimization[J/OL]. Digital Communications and Networks, 2023[2024-03-05]. https://www.sciencedirect.com/science/article/pii/S2352864823001360

[28] Guo Delin, Tang Lan, Zhang Xinggan, et al. Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 13124−13138

[29] Rashid T, Samvelyan M, de Witt C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(178): 1−51

[30] Wang Jianhao, Ren Zhizhou, Liu Terry, et al. QPLEX: Duplex dueling multi-agent Q-learning[J]. arXiv preprint, arXiv: 2008. 01062, 2020

[31] Wang Tonghan, Wang Jianhao, Zheng Chongyi, et al. Learning nearly decomposable value functions via communication minimization[J]. arXiv preprint, arXiv: 1910. 05366, 2019

[32] Yu Chao, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611−24624

[33] Yuan Lei, Wang Jianhao, Zhang Fuxiang, et al. Multi-agent incentive communication via decentralized teammate modeling[C]//Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2022, 36(9): 9466−9474



Liu Bingyi, born in 1990. PhD, professor and PhD supervisor. His main research interests include wireless networking, vehicular networking, autonomous driving, and Internet of things.

刘冰艺, 1990 年生. 博士, 教授, 博士生导师. 主要研究方向为无线网络、车联网、自动驾驶、物联网.



Wang Dongdong, born in 2000. Master candidate. His main research interests include vehicular networking, autonomous driving, and reinforcement learning.

王东东, 2000 年生. 硕士研究生. 主要研究方向为车联网、自动驾驶、强化学习.



Shi Haiyong, born in 1998. PhD candidate. His main research interests include vehicular networking, autonomous driving, and reinforcement learning.

施海勇, 1998 年生. 博士研究生. 主要研究方向为车联网、自动驾驶、强化学习.



Wang Enshu, born in 1990. PhD, associate professor. His main research interests include vehicular networking, autonomous driving, and reinforcement learning.

王恩澍, 1990 年生. 博士, 副教授. 主要研究方向为车联网、自动驾驶、强化学习.



Wu Libing, born in 1970. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include autonomous driving, distributed computing, network security, and wireless sensor networks.

吴黎兵, 1970 年生. 博士, 教授, 博士生导师. CCF 杰出会员. 主要研究方向为自动驾驶、分布式计算、网络安全、无线感知网络.



Wang Jianping, born in 1975. PhD, professor. IEEE Fellow and AAIA Fellow. Her main research interests include autonomous driving, reliable Internet, cloud computing, optical networks, and service-oriented networks and data center networking.

汪建平, 1975 年生. 博士, 教授. IEEE Fellow, AAIA Fellow. 主要方向有自动驾驶、可靠互联网、云计算、光网络、面向服务的网络和数据中心网络.