边缘智能下基于强化学习的车联网路由协议

刘冰艺[1,2]，刘煜昊[1]，韩玮祯[1,3]，夏振厂[1]，吴黎兵[4]，熊盛武[1]

（1. 武汉理工大学计算机科学与人工智能学院，湖北 武汉 430070；
2. 武汉理工大学三亚科教创新园，海南 三亚 572000；
3. 武汉理工大学重庆研究院，重庆 401135；
4. 武汉大学国家网络安全学院，湖北 武汉 430070）

摘 要：为实现复杂城市车联网环境下高可靠、自适应的数据包路由协议，提出一个端−边−云边缘智能架构，该架构包括终端用户层、边缘协作层和云计算层。在所提边缘智能架构的基础上，设计了一个基于多智能体强化学习的数据包路由协议。实验结果表明，相比于现有的紧急消息传输机制、基于交叉路口雾节点的分布式路由协议和基于双深度 Q 网络的路由协议，所提协议在消息传输时延和接收率方面分别取得 29.65%～44.06%和 17.08%～25.38%的优化。

关键词：边缘智能；车联网；多智能体强化学习；数据包路由

中图分类号：U495

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023187

Edge intelligence-assisted routing protocol for Internet of vehicles via reinforcement learning

LIU Bingyi[1,2], LIU Yuhao[1], HAN Weizhen[1,3], XIA Zhenchang[1], WU Libing[4], XIONG Shengwu[1]

1. School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China
2. Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China
3. Chongqing Research Institute, Wuhan University of Technology, Chongqing 401135, China
4. School of Cyber Science and Engineering, Wuhan University, Wuhan 430070, China

Abstract: To achieve a highly reliable and adaptive packet routing protocol in a complex urban Internet of vehicles, an end-edge-cloud edge intelligence architecture was proposed which consisted of an end user layer, an edge collaboration layer, and a cloud computing layer. Based on the proposed edge intelligence architecture, an packet routing protocol based on multi-intelligent reinforcement learning technologies was designed. The experimental results show that the proposed protocol could significantly improve the transmission delay and the packet reception rate in the interval of 29.65%～44.06% and 17.08%～25.38% compared to the state-of-the-art transmission mechanism for emergency data (TMED), intersection fog-based distributed routing protocol (IDR), and double deep Q-net based routing protocol (DRP).

Keywords: edge intelligence, Internet of vehicles, multi-agent reinforcement learning, packet routing

---

收稿日期：2023−05−29；**修回日期：**2023−07−27

通信作者：夏振厂，zcxia.cs@whut.edu.cn

基金项目：国家自然科学基金资助项目（No.62272357, No.62176194, No.62202348, No.U20A20177, No.62272348）；湖北省重点研发计划基金资助项目（No.2022BAA052）；海南省重点研发计划基金资助项目（No.ZDYF2021GXJS014）；重庆市科学基金资助项目（No.cstc2021jcyj-msxm4264）；武汉理工大学重庆研究院研究基金资助项目（No.ZD2021-04, No.ZL2021-05）

Foundation Items: The National Natural Science Foundation of China (No.62272357, No.62176194, No.62202348, No.U20A20177, No.62272348), The Key Research and Development Program of Hubei Province(No.2022BAA052), The Key Research and Development Program of Hainan Province (No.ZDYF2021GXJS014), The Science Foundation of Chongqing (No.cstc2021jcyj-msxm4264), The Research Project of Chongqing Research Institute of Wuhan University of Technology (No.ZD2021-04, No.ZL2021-05)

0　引言

车联网在促进城市交通系统的安全相关应用方面具有巨大的潜力[1]。在车联网中，车辆之间有效的信息传播使驾驶人员能够了解潜在的风险和交通异常，这对提高交通安全和效率有着重大意义[2]。高效、可靠的路由协议可以减少数据包的传输时延，降低数据包的丢失率，从而使车辆更加及时地获取交通信息，支持信号灯控制等智能交通系统应用[3]，避免潜在的交通拥堵和交通风险。在复杂的城市交通环境中，传统的基于 IEEE 802.11p 标准的路由协议暴露出很多问题[4]。首先，城市场景中的建筑物使无线信号严重衰减，导致数据包丢失，影响了路由协议的性能[5]。其次，传统的路由协议采用三次握手来寻找下一跳中继车辆，这使数据包的传输需要额外的控制信息交互作为支撑，带来了较高的传输时延和通信负载。另外，在大量数据包并存的环境中，由于数据包的起止点不同，它们难以重用已经构建好的传输链路。与此同时，多个数据包的同时传输也会引发拥塞等问题，影响路由协议的性能。

近年来，已有大量工作聚焦于车联网中的数据包路由协议研究。这些路由协议主要分为 2 个主流的研究方向：基于路径的路由协议和不基于路径的路由协议。在基于路径的路由协议中，数据包在传输的开始阶段计算一条完整的路径。文献[6]提出了一种增强地理源节点路由协议，使用蚁群算法计算网络上每个路段的权重，并根据带权重的道路拓扑图来选择具有最小权重的完整消息传输路径。文献[7]使用贪心转发策略向前转发探测数据包以侦测所有路径中最少耗时的路由。基于路径的路由协议可以减少路由计算的次数，从而减少数据包传输过程中的开销，但这类协议的目的在于找到一个全局的、临时性的最优路由，在高度动态的城市车联网环境中缺乏灵活性和自适应能力。

不基于路径的路由协议不需要在消息传输的开始阶段确定完整的路径，这类协议在每跳消息传输时确定下一跳节点。基于车辆聚类的消息传输协议可以显著提升车辆间的数据交互效率[8]。文献[9]在交叉路口进行车辆的聚类管理，并构建多跳链路。聚类的管理者进行数据包下一跳中继节点的决策。文献[10]根据车辆的运动状态来进行移动区域的划分，每个区域的队长节点从其他区域内选择合适的中继节点，中继节点的选取指标为中继节点与目的节点之间的距离。不基于路径的路由协议在数据包传输的过程中动态地选择下一跳节点。因此，这类协议可以更好地适应动态的车联网环境，在不断的环境感知过程中优化数据包传输路径。然而，不基于路径的路由协议无法在全局的视角下进行路径决策，其最终导出的完整路由可能是次优解。此外，文献[11]提出不使用 V2V（vehicle-to-vehicle）多跳传输数据包，而是直接在基站之间使用高速有线连接技术进行通信。基站间的有线连接可以提供更可靠和稳定的网络连接，确保数据的高传输速度和低时延，但是影响了 V2X（vehicle-to-everything）多跳通信的灵活性和便捷性，同时带来高昂的部署与维护成本。

随着强化学习[12]等人工智能算法在车联网领域的广泛应用[13]，近些年涌现出许多基于强化学习的路由协议。文献[14]提出一个分层路由框架，该框架使用基于 Q 学习的路由路径规划，同时设计了一个基于位置的下一跳节点选择算法。文献[15]设计了一个基于 Q 学习的交通感知路由协议，根据路段交通信息选择链路可靠性高的路段进行传输，同时设计一个 Q−贪婪算法，根据车辆的地理位置构建车辆间数据传输链路。文献[16]提出一个分层路由协议，该协议将地理区域划分成网格，使用 Q 价值表格进行下一跳网格选取。基于强化学习的路由协议可以根据动态的环境自适应调整数据包传输策略。然而，这些协议是针对单个数据包路由问题设计的，没有考虑到多个数据包并存的车联网环境中数据包之间的相互影响。

相比于传统的路由协议，基于强化学习的路由协议的应用也会带来新的挑战。首先，强化学习框架的训练和推理过程对计算资源有较高要求，而车辆等终端用户的计算能力有限，无法满足强化学习框架的高算力需求。其次，终端车辆用户的通信范围有限，且对环境的感知能力有限。环境感知的偏差会导致强化学习框架得出的数据包传输策略无法适应真实的环境。为解决上述挑战，本文设计了一个边缘智能[17]架构，并在此基础上提出了一个基于多智能体强化学习（MARL, multi-agent reinforcement learning）的数据包路由协议。本文的主要研究工作如下。

1) 提出一个三层的端−边−云的边缘智能架构用于车联网中的多跳消息路由，包括终端用户层、

边缘协作层和云端计算层。其中，终端用户层实现环境感知与数据上传，并完成与通信环境的交互；边缘协作层提供低时延的计算和存储服务；云端计算层完成深度强化学习的训练任务。

2) 基于边缘智能架构，提出了一种基于多智能体强化学习的数据包路由协议，旨在实现在车联网中高效智能的消息传输。该方法将数据包路由问题建模成一个马尔可夫博弈，使通信节点能够在与车联网环境的交互过程中自适应地推导出最佳的数据包路由策略。

3) 本文提出一种基于多智能体近端策略优化（MAPPO，multi-agent proximal policy optimization）算法的数据包路由协议，并进行了广泛的实验验证。实验结果表明，在边缘智能架构下，基于多智能体强化学习的数据包路由协议可以显著提高车联网中的消息传输性能。

1 系统模型

1.1 端–边–云边缘智能架构

本文提出了一种边缘智能架构来构建高效的边缘计算网络，在此基础上可以实现可靠的数据包路由协议。边缘智能架构下的数据包路由如图 1 所示。所提边缘智能架构包括终端用户层、边缘协作层和云端计算层。

1) 终端用户层。深度强化学习（DRL，deep reinforcement learning）等人工智能算法是实现高效数据包路由的有效方法。然而，在具有大量学习参数的强化学习框架的训练过程中，海量的数据是必不可少的。这些数据由广泛分布的车辆和路侧单元（RSU，road side unit）等车联网终端用户生成。因此，终端节点感知到的数据需要上传到边缘的数据中心和云服务器上[18]，以支持强化学习框架的训练过程。此外，单个终端用户无法满足基于强化学习的路由应用的高计算需求，为充分利用海量的数据资源，终端用户所感知的大量数据需要在邻近的边缘服务器中进行处理。所以，需要频繁地上传车辆运动状态和通信状况等感知数据，经过边缘服务器处理后再上传到云端，以支持基于强化学习的多跳路由协议。终端用户与边缘服务器之间的消息传输包括边缘服务器到车辆终端的模型分发和车辆终端到边缘服务器的数据上传等。值得一提的是，远距离的环境信息获取成本较高，需要耗费额外传输时间和信道资源，再加上通信时延和交付率无法保证，还会涉及新的路由问题。因此，所提协议限制了边缘服务器的环境感知范围。边缘服务器只采集其通信范围之内的环境信息，其通信范围之外的终端用户不与边缘服务器进行通信。

2) 边缘协作层。近年来，移动边缘计算与人工智能的融合为车联网中多跳数据传输协议的设计提供了一种新的技术方法，即边缘智能。通过在车联网中靠近终端用户的边缘端运行基于人工智能的消息传输应用程序，使边缘服务器甚至终端节点充分利用人工智能的便利，为用户提供实时可靠的计算和存储服务[19]。然而，基于强化学习等人工智能算法的应用程序在训练和推理过程中需要大量的计算资源，导致显著的能量损失和计算时延。另

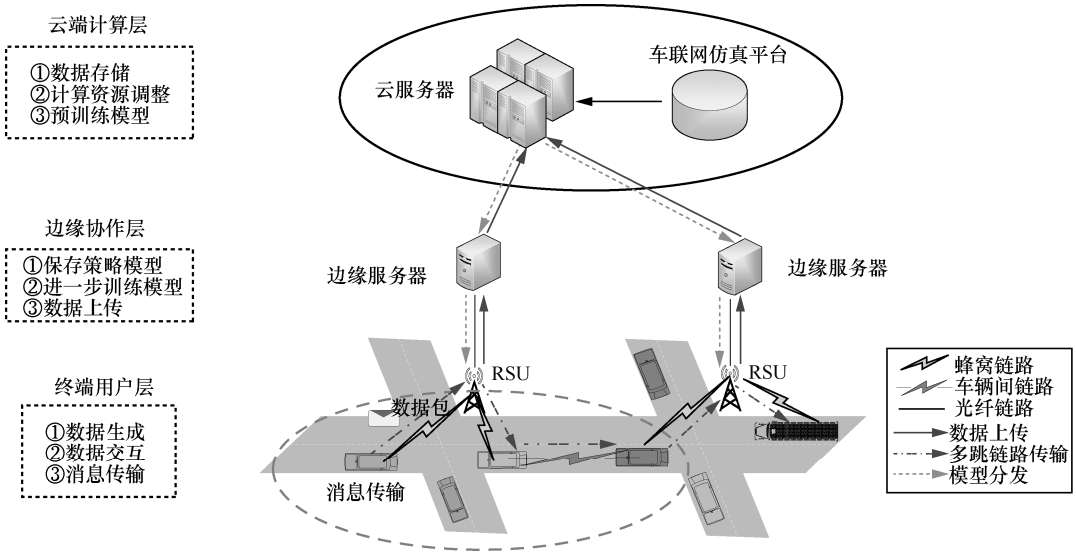

图 1 边缘智能架构下的数据包路由

一方面，由于移动边缘计算可以在终端用户附近提供低时延的计算和存储服务，因此强化学习算法的训练和推理过程可以嵌入边缘服务器，以提高基于强化学习的数据包路由协议的效率。在可预见的将来，自然语言处理、计算机视觉等人工智能技术将在车联网场景中得到广泛的应用。这对终端用户和边缘服务器的计算能力和存储能力提出了挑战。考虑到近年来云计算的繁荣，强大的云数据中心可以通过云−边缘的协调方式来处理计算密集型的任务，例如强化学习框架的训练等。边缘服务器可以充分利用云端的海量计算和数据资源，确保基于强化学习的路由协议的稳定性能。边缘服务器与云服务器之间的通信链路主要负责云服务器到边缘服务器的模型分发和边缘服务器到云服务器的数据上传等。简而言之，边缘协作层能够充分协调端−边−云各层的资源，在充分了解复杂的通信和交通条件的基础上，保障高效、智能的多跳数据包传输。

3) 云端计算层。近年来，世界各地的众多机构都致力于提供云计算服务，其中云数据中心的特点是具有丰富的计算和数据资源。因此，在边缘智能架构下基于强化学习的数据包路由协议不仅可以利用终端节点的数据和计算资源，还可以通过数据卸载和计算任务卸载[20]利用云计算服务。例如，针对数据包路由问题的强化学习框架的训练过程可以在云端完成，而其推理过程可以在边缘侧进行。此外，考虑到在真实环境中收集训练数据和训练强化学习框架的时间和计算消耗非常庞大，大量研究利用车联网仿真平台，通过模拟复杂多变的车联网环境，对强化学习框架进行预训练。车联网仿真平台需要强大的计算能力和大量的训练数据作为支撑，因此需要将其集成到云服务器中。

1.2　边缘智能架构下的数据包路由协议

在分层边缘架构下，本文提出一个基于多智能体强化学习的数据包路由协议。该协议提出在每个交叉路口部署 RSU 和边缘服务器，形成一个边缘节点。边缘节点之间通过路段连接彼此。边缘节点负责在其通信范围内构建多跳传输链路。同时，当数据包到达交叉路口后，边缘服务器通过训练好的强化学习模型指导数据包向最佳的路段上传输。

一些已有研究聚焦于分布式的多跳数据包传输协议，完全使用车辆节点来进行数据包的多跳转发。而本文选用 RSU 来指导数据包在交叉路口的传输方向，其原因可以总结为以下三点。第一，RSU 拥有更大的通信范围，具备更强大的信息收集能力，将它部署在交叉路口能够最大限度地感知路口交通信息，在云端构建完整的交通数据集。第二，面对计算密集型的计算任务，如强化学习模型的推理，单个车辆无法提供足够的算力和能源支撑。第三，在分布式的路由协议中，单个车辆需要通过广播周期性的探测数据包来感知并选取最优的下一跳中继节点。大量的周期性探测数据包将严重影响数据包的传输成功率。

针对数据包多跳传输问题，本文将数据包视为智能体。当数据包到达交叉路口时，数据包根据边缘节点的环境感知信息进行传输方向决策。值得一提的是，数据包作为智能体并没有计算和存储能力，因此传输距离较远的数据包必须利用存储在边缘节点的强化学习模型进行下一跳路口决策。其他传输距离较近的数据包也可能用到存储在边缘节点的多跳链路信息，但是不需要用到强化学习模型。当数据包在边缘节点之间的路段上传输时，由路段上的车辆节点进行分布式的中继转发。单个数据包的多跳传输问题可以建模成一个马尔可夫决策过程。具体而言，在每个时间步，数据包到达一个边缘节点，根据边缘节点感知的周围交通和通信状况进行传输方向决策，在该方向的路段上进行多跳传输后，环境转移到下一状态。同时，根据数据包是否到达终点，以及数据包的传输时延，环境将反馈给数据包一个即时奖励。考虑到多个数据包同时存在的真实车联网场景，本文将多个数据包的路由问题建模成一个马尔可夫博弈。

1.3　强化学习框架的部署与训练

深度强化学习结合了深度学习的优势，利用了深度神经网络（DNN，deep neural network）强大的数据表示能力和特征学习能力。然而，基于多智能体强化学习的数据包路由协议所需的算力远超单个智能车辆的限制。为此，本文基于端−边−云架构，在边缘服务器上部署多智能体强化学习中的深度神经网络（如状态价值网络、策略网络）。每个数据包都可以使用边缘节点的感知数据，并通过边缘服务器中的深度神经网络来推导数据包传输策略。由于边缘计算网络可以提供低时延的通信服务，这种端−边缘的协作方式可以满足数据包传输的低时延需求。

由于多智能体强化学习框架的训练需要大量的数据存储和时间消耗，直接在 RSU 上进行训练

十分困难。因此，本文选择在云服务器上的车联网仿真平台上完成多智能体强化学习框架的预训练过程，如图 1 所示。值得注意的是，仿真平台无法完全模拟真实车联网环境中可能存在的交通和通信条件，而现实环境中的复杂动态性也造成了仿真与现实的差距[21]，这将导致预训练的策略与真实车联网环境不兼容。为了使预训练的策略模型对真实车联网环境中的策略产生指导意义，本文从交通环境的设置层面和通信模型层面尽可能模拟真实车联网环境，此外，本文在仿真中引入了真实车联网环境中可能存在的不确定性因素，使仿真更贴近真实车联网环境。

2. 问题建模

2.1 马尔可夫博弈建模

在城市场景中，需要频繁地将大量的数据包传输到特定位置。然而，现有的路由协议通常是在数据包出现后才考虑特定的传输路由，因此很难复用其他数据包的路由。在城市场景中，由于数据包数量众多，这种方法效率低下且耗时。此外，大多数路由协议没有将路由过程视为多智能体问题。

因此，本文提出了一个基于多智能体强化学习的路由协议（MARP, MARL-based routing protocol），它采用了端−边−云系统模型。边缘服务器定期收集并上传本地信息到云端，将这些局部信息聚合成一个加权有向图形成全局环境，为实现 MARL 算法提供了前提条件。数据包的路由过程被抽象为在加权有向图中寻找最优路径的博弈。其中，消息的发送节点先将数据包发送给 RSU，RSU 根据强化学习算法训练得到的决策模型传输数据包到其他 RSU，当消息到达距离消息的目的节点最近的 RSU 后，再由该 RSU 将数据包传递给目的节点。由于数据包在 RSU 之间的传输将影响后续的决策，这可以被视为一个顺序决策问题，而多个数据包的顺序决策可以建模成一个马尔可夫博弈。本文采用 MAPPO 算法[22]来解决上述马尔可夫博弈，并推导出奖励最大化的数据包路径规划策略。马尔可夫博弈可以表示为一个元组 $(\mathcal{I},\mathcal{S},\mathcal{O},\mathcal{A},\gamma,\mathcal{R},\mathcal{P})$，其中每个元素定义如下。

1) $\mathcal{I}$ 表示智能体集合。在本文中，每个数据包是一个智能体，表示为 i，负责与环境进行交互。所有智能体组成一个智能体集合 $\mathcal{I}=\{1,\cdots,N\}$。

2) $\mathcal{S}$ 表示环境状态空间。时间步 t 的环境状态 s_t 包含三部分信息，分别是当前加权有向图 G、每个数据包的当前位置和目的地、交叉路口的车流密度。所有可能的环境状态组成一个环境状态空间 $\mathcal{S}$。

3) $\mathcal{O}$ 表示智能体的观测空间。在时间步 t，每个智能体 i 从环境获得一个观测，表示为 $o_{i,t}$。观测包含三部分信息，分别是当前位置、目的地位置、交叉路口车辆密度。所有可能的观测组成观测空间 $\mathcal{O}$。

4) $\mathcal{A}$ 表示智能体的动作空间。在时间步 t，每个智能体 i 根据当前策略 $\pi_{i,t}(a_{i,t}\mid o_{i,t})$ 和当前观测 $o_{i,t}$ 选择一个动作 $a_{i,t}$。智能体的动作 $a_{i,t}$ 表示下一时间步停留在当前位置或传输到下一个相邻的交叉路口。所有智能体的联合动作表示为 $\boldsymbol{a}_t=(a_{1,t},\cdots,a_{N,t})$，所有可能的动作组成动作空间 $\mathcal{A}$。

5) γ 表示折扣因子。$\gamma\in[0,1]$ 用来权衡未来的奖励对当前累计折扣奖励的影响。

6) $\mathcal{R}$ 表示奖励函数。在时间步 t，根据每个智能体 i 的观测 $o_{i,t}$ 和动作 $a_{i,t}$，环境给该智能体反馈一个奖励 $r_{i,t}$。该奖励与数据包的传输时延相关。本文旨在最大化所有智能体的累计折扣奖励，表示为 $R=\mathbb{E}\left[\sum_{i=0}^{N}\sum_{t=0}^{T}\gamma^t r_{i,t+1}\right]$。

7) $\mathcal{P}$ 表示状态转移概率矩阵。在时间步 t，根据环境状态 s_t 和联合动作 $\boldsymbol{a}_t$，当前环境状态转移到下一个环境状态 s_{t+1} 的概率可以表示为 $P(s_{t+1}\mid s_t,\boldsymbol{a}_t)$，所有状态转移概率的集合组成状态转移概率矩阵 $\mathcal{P}$。

2.2 多智能体强化学习框架设计

MAPPO 算法是 PPO（proximal policy optimization）算法在多智能体环境中的一种拓展方案，是经典的行动者−评价者（actor-critic）架构，采用集中式训练分布式执行的训练方法。因此，每个智能体拥有一个行动者（actor）网络和一个评价者（critic）网络。多智能体强化学习框架如图 2 所示。图 2 中变量省去了时间步 t 的下标。每个智能体 i 首先从环境中获取观测 o_i 并输入一个多层感知机，得到中间状态特征 x_i。然后，中间状态特征 x_i 被输入行动者网络，得到策略 π_i。智能体 i 根据策略 π_i 获取一个动作 a_i。同时，中间状态特征 x_i 作为评价者网络的输入，输出状态价值 v_i。状态价值 v_i 和环境奖励 r_i 被用来计算优势函数 A_i，以更新行动者网络[23]。

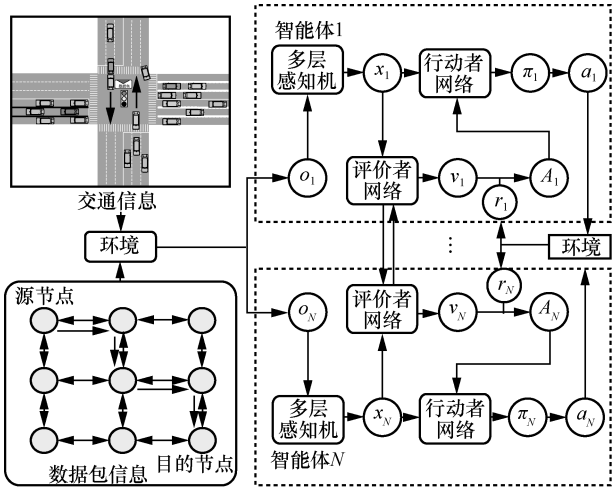

图 2　多智能体强化学习框架

2.3　多智能体强化学习训练算法

本文采用集中式训练分布式执行的多智能体强化学习架构。在集中式训练阶段，评价者网络可以使用所有智能体的信息。而在分布式执行阶段，每个智能体的行动者网络只能使用其本地信息。集中式训练算法如算法 1 所示。

算法 1　集中式训练算法

输入　每个智能体的初始化行动者网络参数 θ_i，评价者网络参数 φ_i

输出　每个智能体的近似最优策略网络 φ_i^*

1) while 训练未结束 do
2)　for 时间步 $1\sim T$ do
3)　　for 每个智能体 i do
4)　　每个智能体从环境中获取观测 $o_{i,t}$，并计算得到动作 $a_{i,t}$ 和动作价值 $v_{i,t}$
5)　　智能体与环境进行交互，并获得奖励 $r_{i,t}$ 和新的观测 $o_{i,t+1}$
6)　　使用式(1)计算优势函数值 $A_{i,t}$
7)　　将经验 $(\mathcal{X}_t,\mathcal{G}_t,\mathcal{C}_t^m,\mathcal{R}_t^{e,m},\mathcal{R}_t^{p,m})$ 存入经验池 $\mathcal{D}$
8)　　end for
9)　end for
10)　for 训练轮次 do
11)　　从经验池 $\mathcal{D}$ 中抽取数据 $\boldsymbol{d}$
12)　　for d_i in $\boldsymbol{d}$ do
13)　　if d_i 是成功传输的数据包经验 do
14)　　根据式(2)更新智能体 i 的行动者网络的参数 θ_i
15)　　根据式(3)更新智能体 i 的评价者网络的参数 φ_i
16)　　end if
17)　　end for
18)　end for
19) end while

该算法首先初始化每个智能体的行动者网络参数 θ_i 和评价者网络参数 φ_i。对于时间步 $1\sim T$，每个智能体 i 从环境中获取观测 $o_{i,t}$，并计算得到动作 $a_{i,t}$ 和动作价值 $v_{i,t}$。然后，智能体与环境交互并从环境中得到奖励 $r_{i,t}$（第 14)行），同时环境反馈一个新的观测 $o_{i,t+1}$。智能体 i 根据式(1)计算优势函数值 $\delta_{i,t}$，并将经验存入经验池 $\mathcal{D}$。

$$\delta_{i,t}=r_{i,t}+\gamma V_{\varphi_i}(s_{t+1})-V_{\varphi_i}(s_t) \tag{1}$$

其中，当前时间步为 t，s_t 表示时间步 t 的状态，状态价值函数 $V_{\phi_i}(s)=\mathbb{E}_{\phi_i}\left[\sum_{\sigma=t}^{T}\lambda^{\sigma-t}r_{i,t+1}\mid s_t=s\right]$ 表示从时间步 t 开始到时间步 T 结束的累计折扣奖励的期望。

在更新过程中，首先从经验池 $\mathcal{D}$ 中抽取经验，使用成功传输的数据包的经验，根据式(2)更新行动者网络参数。

$$J(\theta_i)=\mathbb{I}[\min(\mu_{i,t}(\theta_i)A_{i,t},\mathrm{clip}(\mu_{i,t}(\theta_i),1-\varepsilon,1+\varepsilon)A_{i,t})] \tag{2}$$

其中，$\mu_{i,t}(\theta_i)=\dfrac{\pi_{\theta_i}(a_{i,t}\mid o_{i,t})}{\pi_{\theta_i^{\mathrm{old}}}(a_{i,t}\mid o_{i,t})}$ 表示基于新策略采取动作的概率与基于旧策略采取动作的概率之比。MAPPO 算法使用重要性采样定理，应用历史策略采样得到的样本进行更新。为了增强训练的稳定性和可控性，MAPPO 算法限制了策略梯度的更新幅度。具体来说，$\mathrm{clip}(\mu_{i,t}(\theta_i),1-\varepsilon,1+\varepsilon)$ 函数对行动者网络参数 $\mu_{i,t}(\theta_i)$ 的更新幅度进行约束，$1-\varepsilon$ 和 $1+\varepsilon$ 分别为 $\mu_{i,t}(\theta_i)$ 的上下界约束。

接着，根据式(3)更新评价者网络参数。

$$L(\varphi_i)=\frac{1}{B}\sum_{k=1}^{B}\max[(V_{\varphi_i}(s_k)-R_k^i)^2,(\mathrm{clip}(V_{\varphi_i}(s_k),V_{\varphi_i^{\mathrm{old}}}(s_k)-\varepsilon,V_{\varphi_i^{\mathrm{old}}}^i(s_k)+\varepsilon)-R_k^i)^2] \tag{3}$$

其中，$R_k^i=\sum_{l=0}^{T-k}\gamma^l r_{i,k+l}$ 表示智能体 i 在时间步 k 的累计折扣奖励。

评价者网络的训练需要来自其他智能体的额外信息，并捕获智能体之间的交互，这导致了评价者网络训练过程的复杂性。因此，如第 1 节所述，本文在云服务器上对所有智能体的策略进行预训

练，然后通过云−边缘的通信链路将经过训练的策略模型分发到边缘服务器。为了确保策略能够更好地适应智能体的交通流量和通信条件，需要对预先训练过的策略在物理设备上进行进一步训练，以减少仿真平台与现实的差距。由于评价者网络不仅需要从其他智能体中获取额外的信息作为输入，而且需要捕获智能体之间的相互影响。因此，评价者网络中的深度神经网络规模庞大，且结构相对复杂。综上，评价者网络的训练需要来自云服务器的计算和存储资源，而为了满足边缘服务器的计算需求，行动者网络只使用本地信息，即在云中部署了评价者网络，在边缘部署了行动者网络。此外，经验回放池和计算奖励过程也被集成到云服务器中。

在进一步的训练过程中，智能体生成的观察结果和动作被上传到云服务器，云服务器计算奖励并等待下一个观察结果被上传。然后，云服务器将经验保存到经验回放池中。一旦在经验回放池中有足够的经验，云服务器就会更新评价者网络中的参数，用于更新边缘服务器中的行动者网络。

3 仿真实验与分析

3.1 仿真设置

为了评估所提基于多智能体强化学习路由协议的性能，本文在 OMNeT++（objective modular network testbed in C++）仿真器和 Veins（vehicles in network simulation）[24]框架的基础上实现车联网通信仿真平台，采用当前技术成熟的 IEEE 802.11p 标准作为车联网通信仿真协议。对于仿真环境，本文构建了一个虚拟地图和一个真实地图，分别如图 3 和图 4 所示。其中，真实地图取自苏州工业园区，使用 OpenStreetMap 进行路网生成。在虚拟地图中，交叉口之间的路段设置为长 1 km 的双向四车道。车辆的初始位置和轨迹由 SUMO（simulation of Urban mobility）随机生成。此外，车辆的速度受到路段限速的限制，虚拟地图中路段的限速设置为 15 m/s，真实地图中路段的限速与真实世界保持一致。在本文的仿真实验中，结合真实的通信场景，分别给 RSU 和车辆设置了不同的通信功率。其中，RSU 的通信功率设置为 15 mW，支持约 500 m 的可靠通信；车辆的通信功率设置为 6 mW，支持约 250 m 的可靠通信。本文的仿真场景考虑了真实环境下的驾驶行为多样性（掉头、靠边停车）、网络拥塞模拟（限制带宽、模拟数据包丢失）和数据包纠错机制（丢包检测、重传机制）。场景中随机生成 10 个数据包作为智能体，智能体的源节点和目的节点随机选取产生，每个智能体通过所提 MARP 进行远距离传输。场景中的其他数据包，包括用于环境感知的周期性数据包、用于多跳链路构建的探测数据包、用于下一跳路由选取的决策数据包等都被看作动态车联网环境的一部分。仿真参数如表 1 所示。实验使用的服务器具体参数配置为 Intel Xeon Gold 6226R CPU 和 NVIDIA RTX 3090 GPU。

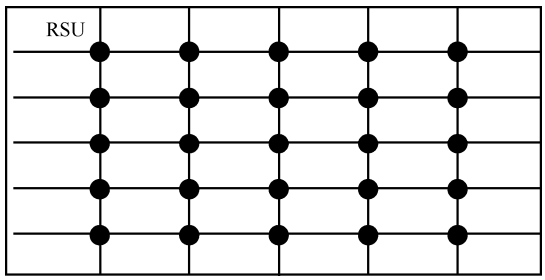

图 3 虚拟地图

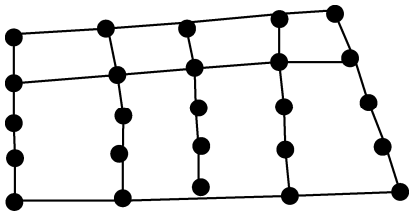

图 4 苏州工业园区

表 1 仿真参数

参数	值
虚拟地图大小	5 000 m×5 000 m
MAC 协议	IEEE 802.11p
车道数量	双向四车道
道路长度/m	1 000
最大车速/(m·s^{-1})	15
交叉路口数量	25
RSU 传输功率/ mW	15
车辆传输功率/ mW	6
智能体数量/个	10
每轮训练步长/步	200
最小信噪比阈值/ dBm	−110

为了构建端−边−云的边缘智能架构，将边缘服务器部署在RSU上。本文将MARL模型部署在RSU上的边缘服务器中。通过这种方式，数据包路由协议和消息传播策略可以通过云−边缘协作得到良好的训练。在每个时间步中，RSU 为终端用户分发消息路由策略。

为了验证所提 MARP 的性能，将其与紧急消息传输机制（TMED）[7]、基于交叉路口雾节点的分布

式路由协议（IDR）[9]和基于双深度 Q 网络（DDQN）的路由协议（DRP）进行对比。其中，DRP 结合本文提出的边缘架构，使用 DDQN 进行数据包传输方向的决策，同时利用边缘节点构建的多跳链路，实现多数据包的多跳路由。本文构建了不同车辆密度的场景，并引入数据包平均传输时延、平均接收率和传输跳数 3 个指标对所有路由协议进行分析。

3.2　结果分析

虚拟地图和真实地图中车辆数量对平均传输时延的影响分别如图 5 和图 6 所示。为了反映数据包的冲突导致的性能下降，在本文的实验设置中，冲突的数据包的传输时延被设置为 1 000 ms。从图 5 和图 6 可以看出，随着车辆数量的增加，几种路由协议的传输时延都呈上升趋势。因为车辆数量的增加带来了更多的数据包，这增加了数据包传输的冲突概率，发生冲突的数据包需要花费更多的时间重新传输。在不同交通流量的城市场景中，MARP 的平均传输时延最低。这是因为 TMED、IDR 和 DRP 需要耗费额外的时间进行多跳链路的构建和路由路径的选择，并且没有考虑数据包传输过程中的碰撞丢包问题，导致平均传输时延较高。而在本文所提协议中，边缘节点根据周围交通和通信状况，使用多智能体强化学习算法提前学习数据包传输策略，同时考虑了数据包之间的影响作用。这样大大降低了数据包的传输时延，降低时延达 29.65%～44.06%。

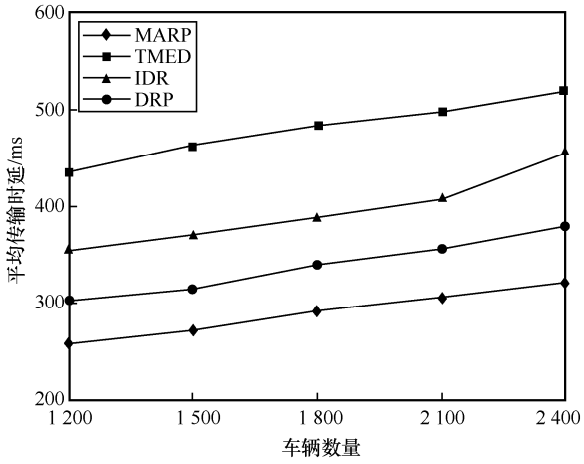

图 5　虚拟地图中车辆数量对平均传输时延的影响

在多跳数据包传输问题中，数据包转发次数，即传输跳数，是影响传输时延的重要因素。更多的传输跳数会带来更长的传输时延。表 2 展示了不同距离的传输任务中各方案的传输跳数。从表 2 可以看出，随着传输距离的增加，各方案的传输跳数随之增加。在不同距离的消息传输任务中，所提 MARP 的传输跳数少于其他协议。这是因为 MARP 利用强化学习技术，能够根据动态变化的交通环境来选择最优下一跳节点，从而减少了不必要的消息转发。

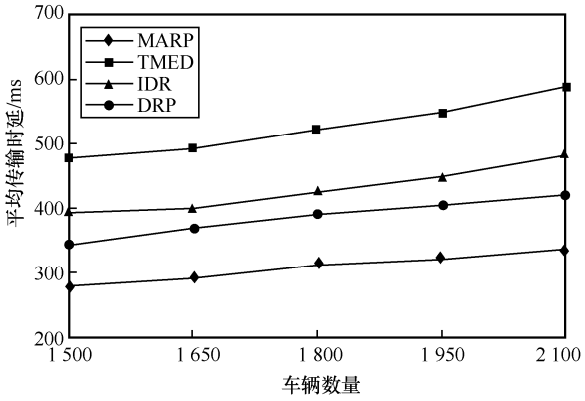

图 6　真实地图中车辆数量对平均传输时延的影响

表 2　不同距离的传输任务中各方案的传输跳数

传输距离/m	MARP	TMED	IDR	DRP
3 000	16	18	18	17
4 000	20	22	22	21
5 000	26	29	28	28
6 000	30	33	32	31

虚拟地图和真实地图中车辆数量对平均接收率的影响分别如图 7 和图 8 所示。随着车辆数量的增加，几种路由协议的平均接收率都呈下降趋势。这是因为车辆数量的增加带来了更多的数据包数量，增加了数据包传输的冲突概率。所提 MARP 的接收率明显优于其他路由协议。这是因为 TMED、IDR 和 DRP 都只考虑单个数据包的路由过程，没有考虑数据包之间的相互影响。这将导致更多的冲突和更高的丢包率。所提 MARP 相比于其他路由协议提升数据包接收率达 17.08%～25.38%。

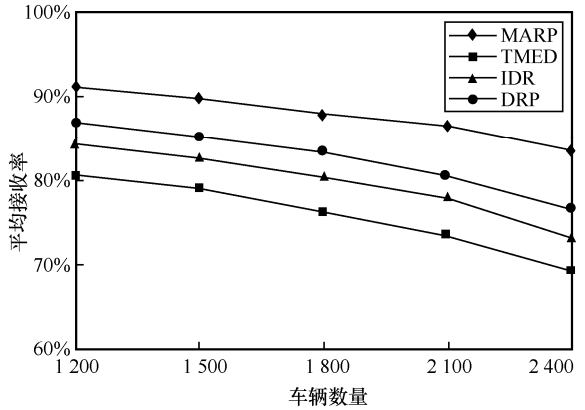

图 7　虚拟地图中车辆数量对平均接收率的影响

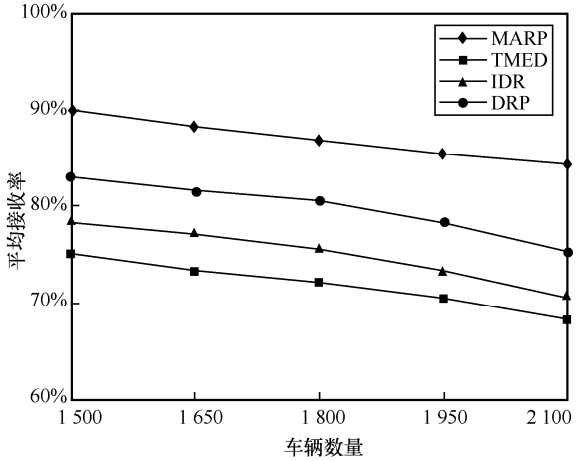

图 8 真实地图中车辆数量对平均接收率的影响

4 结束语

本文研究复杂城市交通场景下多数据包路由问题。首先，针对车辆节点计算资源有限、感知能力受限等挑战，本文提出了端−边−云的边缘智能架构。其中，终端用户层负责环境感知、数据生成、信息上传和车辆间多跳传输；边缘协作层负责强化学习模型的推理、数据处理和数据传输方向决策。云端计算层负责交通和通信环境的构建、强化学习框架的训练、数据包路由决策模型的分发。其次，针对多数据包同时传输的城市交通场景，本文将多个数据包的多跳传输问题建模成一个马尔可夫博弈，设计一个多智能体强化学习框架，并提出使用 MAPPO 算法进行训练。在所提出的边缘架构下，强化学习框架的训练过程在云服务器上完成。考虑到在大量的真实设备上进行强化学习框架的训练难以实现，本文提出在云端部署车联网仿真平台，完成强化学习框架的预训练过程。训练完成的强化学习模型部署在边缘服务器上，指导数据包的传输方向。本文在虚拟地图和真实地图中对所提算法进行了性能评估，实验结果表明，基于边缘智能的数据包路由协议可以显著提高车联网中的消息传输性能。

参考文献：

[1] YU Y F. Mobile edge computing towards 5G: vision, recent progress, and open challenges[J]. China Communications, 2016, 13: 89-99.

[2] LIU B Y, HAN W Z, JIANG W, et al. A novel V2V-based temporary warning network for safety message dissemination in urban environments[J]. IEEE Internet of Things Journal, 2022, 9(24): 25136-25149.

[3] LIU B Y, HAN W Z, WANG E S, 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, PP(99): 1-13.

[4] YAO L, WANG J, WANG X, et al. V2X routing in a VANET based on the hidden Markov model[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(3): 889-899.

[5] TAKAHASHI S, YOSHIDA M, RAMONET A G, et al. Shadowing-fading-based intersection geographic opportunistic routing protocol for urban VANETs[C]//Proceedings of 24th International Conference on Advanced Communication Technology (ICACT). Piscataway: IEEE Press, 2022: 179-184.

[6] GOUDARZI F, ASGARI H, AL-RAWESHIDY H S. Traffic-aware VANET routing for city environments—a protocol based on ant colony optimization[J]. IEEE Systems Journal, 2018, 13(1): 571-581.

[7] QIU T, WANG X, CHEN C, et al. TMED: a spider-web-like transmission mechanism for emergency data in vehicular ad hoc networks[J]. IEEE Transactions on Vehicular Technology, 2018, 67(9): 8682-8694.

[8] LIU B Y, FANG Z P, WANG W, et al. A region-based collaborative management scheme for dynamic clustering in green VANET[J]. IEEE Transactions on Green Communications and Networking, 2022, 6(3): 1276-1287.

[9] SUN G, ZHANG Y J, YU H F, et al. Intersection fog-based distributed routing for V2V communication in urban vehicular ad hoc networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(6): 2409-2426.

[10] LIN D, KANG J, SQUICCIARINI A, et al. MoZo: a moving zone based routing protocol using pure V2V communication in VANETs[J]. IEEE Transactions on Mobile Computing, 2017, 16(5): 1357-1370.

[11] BHOI S K, SAHU P K, SINGH M, et al. Local traffic aware unicast routing scheme for connected car system[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(6): 2360-2375.

[12] SUTTON R S, BARTO A G. Introduction to reinforcement learning[M]. Cambridge: MIT Press, 1998.

[13] LIU B Y, DENG D X, RAO W B, et al. CPA-MAC: a collision prediction and avoidance MAC for safety message dissemination in MEC-assisted VANETs[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(2): 783-794.

[14] LUO L, SHENG L, YU H F, et al. Intersection-based V2X routing via reinforcement learning in vehicular ad hoc networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 5446-5459.

[15] WU J Q, FANG M, LI H K, et al. RSU-assisted traffic-aware routing based on reinforcement learning for urban VANETs[J]. IEEE Access, 2020, 8: 5733-5748.

[16] LI F, SONG X Y, CHEN H J, et al. Hierarchical routing for vehicular ad hoc networks via reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2019, 68(2): 1852-1865.

[17] ZHAO N, WU H, YU F R, et al. Deep-reinforcement-learning-based latency minimization in edge intelligence over vehicular networks[J]. IEEE Internet of Things Journal, 2022, 9(2): 1300-1312.

[18] ZHOU Z, CHEN X, LI E, et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing[J]. Proceedings of the IEEE, 2019, 107(8): 1738-1762.

[19] SHAO X, HASEGAWA G, DONG M X, et al. An online orchestration mechanism for general-purpose edge computing[J]. IEEE Transactions on Services Computing, 2022, 16(2): 927-940.

[20] QIAO G H, LENG S P, ZHANG K, et al. Collaborative task offloading in vehicular edge multi-access networks[J]. IEEE Communications Magazine, 2018, 56(8): 48-54.
[21] HSU K C, REN A Z, NGUYEN D P, et al. Sim-to-Lab-to-Real: safe reinforcement learning with shielding and generalization guarantees[J]. Artificial Intelligence, 2023, 314: 103811.
[22] YU C, 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.
[23] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation[J]. arXiv Preprint, arXiv:1506.02438, 2015.
[24] SOMMER C, GERMAN R, DRESSLER F. Bidirectionally coupled network and road traffic simulation for improved IVC analysis[J]. IEEE Transactions on Mobile Computing, 2011, 10(1): 3-15.

[作者简介]

刘冰艺（1990− ），男，湖北武汉人，博士，武汉理工大学副教授、博士生导师，主要研究方向为无线网络、车载自组织网络、物联网等。

刘煜昊（1999− ），男，湖北潜江人，武汉理工大学硕士生，主要研究方向为车载自组织网络、强化学习。

韩玮祯（1996− ），男，江苏常州人，武汉理工大学博士生，主要研究方向为车载自组织网络、强化学习。

夏振厂（1987− ），男，河南周口人，博士，武汉理工大学讲师、硕士生导师，主要研究方向为车联网、网络拥塞控制、强化学习等。

吴黎兵（1972− ），男，湖北武汉人，博士，武汉大学教授、博士生导师，主要研究方向为分布式计算、网络安全、无线感知网络等。

熊盛武（1966− ），男，湖北武汉人，博士，武汉理工大学教授、博士生导师，主要研究方向为智能网联汽车、机器学习、数据挖掘等。