doi:10.3969/j.issn.1003-3114.2025.05.006
引用格式:郭歆莹,李明,朱春华.面向通信空白场景的 DRL 辅助 FANET 双跳信息增强路由协议[J].无线电通信技术,2025,51(5):929-939.[GUO Xinying, LI Ming, ZHU Chunhua. DRL-assisted FANET Double-Hop Information Enhanced Routing Protocol for Communication Blackout Scenarios[J]. Radio Communications Technology, 2025, 51(5): 929-939.]

面向通信空白场景的 DRL 辅助 FANET 双跳信息增强路由协议

郭歆莹[1,2,3],李　明[1,2,3],朱春华[1,2,3]
(1.河南工业大学 粮食信息处理与控制教育部重点实验室,河南 郑州 450001;
2.河南工业大学 河南省粮食仓储信息智能感知与决策重点实验室,河南 郑州 450001;
3.河南工业大学 河南省粮情智能监测与应用工程研究中心,河南 郑州 450001)

摘　要:针对飞行自组网(Flying Ad Hoc Network,FANET)在通信空白场景下存在的高时延问题,提出了一种深度强化学习(Deep Reinforcement Learning,DRL)辅助的双跳信息增强路由协议(Double-Hop Information Enhanced Routing Protocol,DHRP)。为了实现有效的路由决策,采用马尔可夫决策过程(Markov Decision Process,MDP)对路由行为进行建模,在状态空间设计中结合了节点位置信息与链路信道容量,并综合考虑了双跳范围内的网络信息,以深度值网络为核心,在融合实时网络状态动态调整机制的奖励函数引导下,做出最优下一跳路由决策。实验结果表明,在通信空白场景下,DHRP 相较于现有的路由方案,显著降低了 FANET 的平均端到端时延。此外,在不同节点规模和网络拥塞条件下,DHRP 均表现出优越的适应性和鲁棒性,通过对动态网络环境的实时感知与智能决策机制,有效保障了整体网络性能。

关键词:飞行自组网;通信空白;深度强化学习;双跳信息;路由协议

中图分类号:TN919.23　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:

文章编号:1003-3114(2025)05-0929-11

DRL-assisted FANET Double-Hop Information Enhanced Routing Protocol for Communication Blackout Scenarios

GUO Xinying[1,2,3], LI Ming[1,2,3], ZHU Chunhua[1,2,3]
(1. Key Laboratory of Grain Information Processing and Control of the Ministry of Education, Henan University of Technology, Zhengzhou 450001, China;
2. Key Experiment on Intelligent Perception and Decision-making of Grain Storage Information in Henan Province, Henan University of Technology, Zhengzhou 450001, China;
3. Henan Province Engineering Research Center for Intelligent Monitoring and Application of Grain Conditions, Henan University of Technology, Zhengzhou 450001, China)

Abstract: To address the challenge of high end-to-end delay in Flying Ad Hoc Network (FANET) under communication blackout scenarios, this paper proposes a Deep Reinforcement Learning (DRL)-assisted Double-Hop Information Enhanced Routing Protocol (DHRP). The proposed protocol models the routing process as a Markov Decision Process (MDP) to enable effective decision-making. In constructing the state space, it incorporates both node location information and link channel capacity, while considering network information within a two-hop neighborhood. Centered on a deep value network, the protocol employs a reward function that reflects real-time network dynamics to guide the agent in selecting the optimal next-hop node. Simulation results show that, compared to existing ap-

收稿日期:2025-05-20
基金项目:国家自然科学基金(61901159);河南工业大学青年骨干教师培养计划(21420104)
Foundation Item: National Natural Science Foundation of China (61901159); Cultivation Programme for Young Backbone Teachers in Henan University of Technology (21420104)

proaches, DHRP significantly reduces the average end-to-end delay in FANET under communication blackout conditions. Furthermore, DHRP demonstrates strong adaptability and robustness across various node densities and levels of network congestion by leveraging real-time environmental awareness and an intelligent decision-making mechanism to maintain overall network performance.

Keywords: FANET; communication blackout; DRL; double-hop information; routing protocol

0 引言

随着无人机(Unmanned Aerial Vehicle,UAV)技术的快速发展,FANET 因其灵活部署、快速响应及高机动性,在搜救勘探[1]、紧急通信[2]、智慧农业[3]及火灾监测[4]等领域得到广泛应用。FANET 由多个 UAV 节点构成,采用多跳通信模式进行数据传输[5],以扩展覆盖范围并增强网络连通性。然而,由于 UAV 节点的高速移动及网络拓扑动态变化,FANET 的稳定通信面临严峻挑战,尤其是在通信空白场景下,数据传输的稳定性和时延可能受到严重影响[6]。

通信空白区域通常是指因地形屏蔽、电磁干扰或远距离信号衰减导致无线信号无法有效覆盖的区域,如山区、海洋、战场或高楼密集的城市环境[7]。在此类场景下,FANET 的多跳通信机制难以维持稳定路由,表现为链路断裂频率增加、数据传输路径延长,导致平均端到端时延显著上升。此外,传统路由协议在此类环境下因路径重建频繁引发额外控制开销,进一步加剧网络拥塞并增加传输时延。在紧急通信和灾害救援等任务关键型应用中,高时延可能导致信息传输滞后,影响任务执行的效率与准确性。因此,如何设计面向通信空白场景的低时延 FANET 路由协议已经成为当前研究亟待解决的关键问题。

FANET 的路由协议设计已成为国际通信网络研究的重要议题之一[8],对于提升网络通信质量具有关键作用。虽然研究者尝试将部分移动自组网(Mobile Ad Hoc Network,MANET)和车载自组网(Vehicular Ad Hoc Network,VANET)中的经典路由协议[9-11]应用于 FANET 场景中,但由于 FANET 拓扑变化更为剧烈,这些协议的适应性仍然面临着诸多挑战。在高度动态的网络环境中,基于拓扑的传统路由协议因需持续维护下一跳信息而频繁更新路由表,导致信令负荷增加,通信效率下降,影响网络性能。

基于地理位置的路由协议不同于需要维护全局路由信息的传统方案,它仅需利用当前邻居和目标节点的位置信息,即可进行转发决策。以文献[12]提出的贪婪周边无状态路由(Greedy Perimeter Stateless Routing,GPSR)协议为例,该方法通过采用贪婪策略实现路由决策,在有效降低控制开销的同时,提升了数据传输性能。然而,在通信空白区域,GPS 信号可能受阻或失准,导致地理位置信息难以准确获取,进而影响路由决策的可靠性。由此可能导致数据包传输路径不稳定,平均端到端时延显著增加。

近年来,机器学习(Machine Learning,ML)与 DRL[13]技术被广泛应用于 FANET 路由优化,并取得了显著成果。例如,文献[14]提出的基于 Q 学习的地理路由协议(Q-learning-based Geographic Routing Protocol,QGeo)通过在路由决策过程中综合考虑 Q 值、链路质量和位置偏差,使协议在动态环境下更接近最优解。文献[15]提出的基于 GPSR 的 Q 网络增强地理路由协议(Q-Network Enhanced Geographic Routing Protocol Based on GPSR,QNGPSR)通过深度 Q 网络(Deep Q-Network,DQN)利用邻居节点拓扑信息辅助路由选择,提升协议在动态环境中的适应性。文献[16]提出的基于 DRL 的深度价值网络(Deep Value Network,DVN)路由协议利用地理信息辅助路由决策,降低了平均端到端时延。然而,该协议仅依赖位置信息,未充分考虑链路状态等关键网络信息,在复杂动态环境下的适应能力有限。文献[17]进一步提出了基于 DRL 的自适应可靠路由协议(Adaptive and Reliable Routing Protocol with Deep Reinforcement Learning,ARdeep),通过 DQN 引入链路状态信息,并结合邻居节点的 Q 值进行决策,以优化数据转发。然而,该方法在拓扑剧烈变化的环境下仍存在一定局限性。尽管上述研究在提升 FANET 网络服务质量(Quality of Service,QoS)方面取得了一定进展,但其主要关注高速移动环境下的拓扑动态变化,而对通信空白场景下 FANET 的平均端到端时延增加问题缺乏系统性研究。

针对上述问题,本文提出了一种 DRL 辅助 FANET 的 DHRP,旨在降低通信空白场景下的平均端到端时延。该协议将路由选择建模为 MDP,并结合节点地理位置与链路信道容量作为状态特征。DHRP 综合考虑双跳范围内的网络信息,采用深度

值网络通过历史飞行数据进行离线模型训练,并在奖励函数中引入实时网络状态动态调整机制,引导转发节点在动态环境中做出最优下一跳决策。在存在通信空白区域的 FANET 环境中,该方法能够显著地降低平均端到端时延,提升网络整体传输性能。

1 系统模型

1.1 网络模型

本文研究的 UAV 网络由若干架 UAV 和一个基站组成,如图 1 所示。基站作为通信的目标节点负责数据接收,一架指定的 UAV 作为源节点负责数据发送,其余的 UAV 则作为中继节点协助完成数据的多跳转发过程。本研究的核心问题是优化路由决策,使数据能够规避通信空白区域,以最低的平均端到端时延传输至目标节点,进而提升网络的稳定性和传输效率。

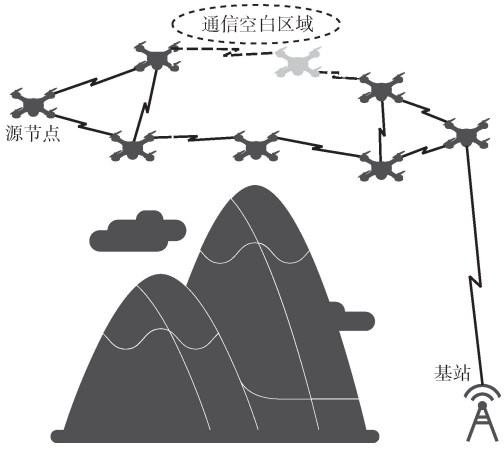


图 1 存在通信空白区域的 FANET 多跳路由

Fig. 1 Multi-hop routing in FANET with communication blackout areas

在 FANET 中,若 2 个 UAV 节点位于彼此的通信范围内,则可建立直接通信链路,实现高效数据传输。节点间的链路时延可表示为:

$$D_{\text{link}}(i,j) = \frac{d(i,j)}{c} + \frac{S}{R(i,j)}, \tag{1}$$

式中:$\frac{d(i,j)}{c}$ 表示传播时延,$\frac{S}{R(i,j)}$ 表示传输时延。$d(i,j)$ 为节点 i 与节点 j 之间的距离,c 为光速(取 3×10^8 m/s),S 为数据包大小,$R(i,j)$ 为 i 与 j 之间链路的数据传输速率,该速率受协议开销、干扰和重传机制等因素影响。然而,鉴于研究重点在于路由优化,为简化分析,本文假设信道容量可近似表示数据传输速率,并使用香农容量公式进行计算:

$$C(i,j) = B(i,j)\text{lb}[1 + SINR(i,j)], \tag{2}$$

式中:$B(i,j)$ 为 i 与 j 之间链路的信道带宽,$SINR(i,j)$ 为 i 与 j 之间链路的信干噪比(Signal to Interference plus Noise Ratio,SINR)。该假设使用香农容量作为数据速率的上界,不考虑媒体接入控制(Media Access Control,MAC)层竞争、协议开销等实际网络影响因素。因此,计算出的传输时延仅为理想估计,主要用于不同路由方案的相对性能评估,而非精确建模实际传输时延。

本文采用解码转发中继协议[18],假设 $D_{\text{queue}}(i)$ 表示 i 的排队时延。本研究的目标是确定最优路由 $\mathcal{P}^* = (i_1, i_2, \cdots, i_T)$,以最小化源节点 i_s 至目标节点 i_d 的平均端到端时延,其表达式如下:

$$\min_{\mathcal{P}^*} \sum_{t=1}^{T-1} [D_{\text{queue}}(i_t) + D_{\text{link}}(i_t, i_{t+1})]$$
$$\text{s.t. } d(i_t, i_{t+1}) \leqslant R_c, \forall t = 1,2,\cdots,T-1, \tag{3}$$

式中:节点 i_1 表示源节点,节点 i_T 表示目标节点,R_c 为通信半径,T 为决策时序总步数。若 $d(i_t, i_{t+1}) \leqslant R_c$,则表示节点 i_t 与节点 i_{t+1} 之间的距离在通信范围内,式(3)成立;反之,若 $d(i_t, i_{t+1}) > R_c$,则表示 2 个节点无法直接通信,式(3)不成立。

在 FANET 高度动态的拓扑环境下,传统的强化学习(Reinforcement Learning,RL)路由机制通常为每个节点独立训练智能体。然而,这种方式依赖静态路径选择,难以适应频繁变化的网络拓扑。为此,本文引入分布式智能路由机制,将决策过程分散至各节点,使其根据当前网络状态动态选择路径,而非依赖于预设路径。同时,智能体随数据包动态移动,并在网络中共享参数,以提高学习效率和网络适应性。

本文假设所有 UAV 节点仅能获取以下局部信息:自身位置、目标节点的位置、通信范围内候选节点的位置以及与候选节点之间链路的信道带宽和 SINR。在路由发现过程中,各节点通过周期性发送 HELLO 报文交换必要信息,并利用分布式 DRL 动态优化路由策略,以适应 FANET 高度动态的拓扑结构并降低平均端到端时延。

1.2 MDP

为适应动态变化的网络环境,本文将路由决策过程建模为 MDP,以四元组 $\langle \mathcal{S}, \mathcal{A}, P, R \rangle$ 表示。其中,$\mathcal{S}$ 为状态空间,$\mathcal{A}$ 为动作空间,P 为状态转移概率函数,描述智能体从当前状态转移至下一状态的概率分布,R 为即时奖励函数,用于评估智能体在特定状态下执行某一动作所获得的收益。在 RL 框

架下,智能体通过与环境的交互不断优化策略,以最大化累积奖励。其交互过程如图 2 所示,智能体首先观察当前状态 $\boldsymbol{s}_t$,选择并执行动作 a_t ,随后从环境中获得即时奖励 r_t ,并转移至新的状态 $\boldsymbol{s}_{t+1}$。

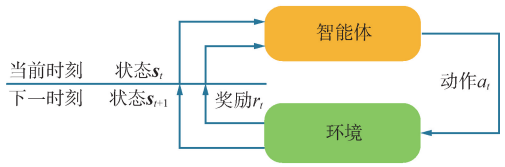


图 2 智能体与环境之间的交互过程

Fig. 2 Interaction process between the agent and the environment

在每个时间步长 Δt 中,路由决策的核心任务是从当前节点的邻居节点中选择最优下一跳节点并转发数据包。整个路由过程始于源节点生成数据包,并持续进行数据包的转发,直至数据包成功抵达目标节点或在超过最大跳数 $t_{\max}$ 后终止。令 i_t 表示时间步长 Δt 开始时数据包所在节点,$\boldsymbol{x}(i)=(u_i,v_i,w_i)$ 表示节点 i 的位置,数据包所在节点位置为 $\boldsymbol{x}(i_t)$,源节点和目标节点的位置分别为 $\boldsymbol{x}(i_{\mathrm{s}})$ 和 $\boldsymbol{x}(i_{\mathrm{d}})$ 。令 $\mathcal{N}_{i_t} \triangleq \{j \mid d(i_t,j) \leqslant R_{\mathrm{c}}\}$ 表示节点 i_t 的邻居节点集合。为降低动作空间的维度,仅考虑距离目标节点最近的前 K 个邻居节点作为候选集合,记为候选节点集合 $\mathcal{C}_{i_t} \triangleq \{i_t^1,i_t^2,\cdots,i_t^K\}$,其中 i_t^k 表示 i_t 的第 k 个候选节点。基于上述分析,可以建立以下 MDP。

状态空间 $\mathcal{S}$:最优路由决策依赖于当前节点及其候选节点的信息。本文旨在学习一种仅基于局部信息的路由策略,因此每个智能体的状态定义为其在时间步 t 观测到的联合网络状态,包括当前节点、目标节点和候选节点的位置,以及当前节点与候选节点之间链路的信道容量。当前 i_t 在 t 的状态可表示为:

$$\boldsymbol{s}_t=[\boldsymbol{x}(i_t),\boldsymbol{x}(i_t^1),\cdots,\boldsymbol{x}(i_t^K),\boldsymbol{x}(i_{\mathrm{d}}),C(i_t,i_t^1),\cdots,C(i_t,i_t^K)], \tag{4}$$

式中:$\boldsymbol{s}_t \triangleq \boldsymbol{s}(i_t)$, $C(i_t,i_t^k)$ 为 i_t 与第 k 个候选节点之间链路的信道容量。

动作空间 $\mathcal{A}$:在 Δt 中,i_t 接收到数据包后,需要从候选节点集合 $\mathcal{C}_{i_t} \triangleq \{i_t^1,i_t^2,\cdots,i_t^K\}$ 中选择一个节点进行转发,定义动作 $a_t \in \mathcal{C}_{i_t}$ 。

状态转移概率函数 P :状态转移受实际网络环境影响,因 UAV 网络拓扑高度动态,P 随时间变化且未知。

即时奖励函数 R :奖励函数旨在平衡网络时延与链路质量,以优化路由性能并提升网络资源利用率。其数学表达式如下:

$$R_t=\begin{cases} r_{\max},\text{下一跳 } i_{t+1} \text{ 为目标节点} \\ r_{\min},i_t \text{ 处于黑洞状态} \\ \omega[-D_{\mathrm{norm}}(i_t,i_{t+1})]+(1-\omega)C_{\mathrm{norm}}(i_t,i_{t+1}),\text{其他} \end{cases}, \tag{5}$$

$$D_{\mathrm{norm}}(i_t,i_{t+1})=\frac{D_{\mathrm{link}}(i_t,i_{t+1})+D_{\mathrm{queue}}(i_t)}{D_{\max}}, \tag{6}$$

$$C_{\mathrm{norm}}(i_t,i_{t+1})=\frac{C(i_t,i_{t+1})}{C_{\max}}, \tag{7}$$

式中:$\omega(0<\omega<1)$ 为权衡系数,$D_{\mathrm{norm}}(i_t,i_{t+1})$ 为归一化时延,$C_{\mathrm{norm}}(i_t,i_{t+1})$ 为归一化信道容量。本文基于实际场景对时延的敏感性,通过参数敏感性实验设 $\omega=0.75$,以在时延优化与链路质量之间实现较优平衡,提升策略的实用性与稳定性。当下一跳 i_{t+1} 为目标节点时,i_t 与 i_{t+1} 之间的链路将获得最大奖励 $r_{\max}$ 。当 i_t 处于黑洞状态时,即其所有候选节点到目标节点的距离均大于自身到目标节点的距离时,数据包的传播时延将显著增加。为避免因路由空洞导致的额外时延,给此类情况设定最小奖励 $r_{\min}$ 。在其他情况下,奖励值依据时延和信道容量动态调整,以优化整体网络性能。

2 路由协议设计

2.1 深度强化学习算法

相较于传统路由算法,基于 RL 的智能体能够更灵活地适应动态网络环境的变化。通过与环境的持续交互,智能体不断优化策略,以提升路由决策的有效性。

首先,定义策略函数:

$$\pi(a_t \mid \boldsymbol{s}_t)=\mathbb{P}(\mathcal{A}_t=a_t \mid \mathcal{S}_t=\boldsymbol{s}_t)。 \tag{8}$$

式(8)表示智能体在状态 $\boldsymbol{s}_t$ 下选择动作 a_t 的概率。智能体的目标是学习从状态到动作的映射,即策略函数 π ,以最大化折扣回报。

$$U_t=\sum_{l=t}^{T-1}\gamma^{l-t}R_l, \tag{9}$$

式中:$\gamma \in [0,1]$ 为折扣因子,用于平衡即时奖励与未来奖励对当前决策的影响,R_l 为第 l 个时间步长中的即时奖励。式(9)描述了智能体在 $T-t$ 个时间步长内可能获得的累积未来折扣奖励。

其次,引入动作价值函数:

$$Q_\pi(\boldsymbol{s}_t,a_t)=E\left[\sum_{l=t}^{T-1}\gamma^{l-t}R_l \mid \boldsymbol{s}_t,a_t,\pi\right], \tag{10}$$

式中:Q 表示在状态 $\boldsymbol{s}_t$ 下,智能体通过 π 选择动作

a_t 后,能够获得的期望折扣回报。在路由优化问题中,Q 值反映了选择 $i_t^{a_t}$ 作为下一跳节点后,转发节点与目标节点之间可能获得的期望累积奖励。

进一步地,最优动作价值函数定义为:

$$Q_*(\boldsymbol{s}_t,a_t)=\max_{\pi} Q_\pi(\boldsymbol{s}_t,a_t)。\tag{11}$$

则可得到最优策略函数:

$$\boldsymbol{\pi}^*=\underset{a\in\mathcal{A}_t}{\operatorname{argmax}} Q_*(\boldsymbol{s}_t,a)。\tag{12}$$

$Q_*(\boldsymbol{s}_t,a_t)$ 包含了确定最佳下一跳所需的全部信息。因此,智能体可以在训练中通过学习 $Q_*(\boldsymbol{s}_t,a_t)$ 获取最优策略函数 $\boldsymbol{\pi}^*$,以最大化累积奖励,实现高效路由。

$Q_*(\boldsymbol{s}_t,a_t)$ 可通过贝尔曼方程递归计算:

$$Q_*(\boldsymbol{s}_t,a_t)=E\left[R_t+\gamma\max_{A\in\mathcal{A}} Q_*(\boldsymbol{s}_{t+1},A)\mid \boldsymbol{s}_t,a_t,\boldsymbol{\pi}^*\right],\tag{13}$$

式中:A 表示动作随机变量。

在状态空间较大的情况下,直接存储和更新 Q 值的 Q-learning 方法难以适用。DQN 通过训练深度神经网络(Deep Neural Network,DNN)$Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)$ 学习逼近 $Q_*(\boldsymbol{s}_t,a_t)$,其中 $\boldsymbol{\theta}_Q$ 表示 $Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)$ 的参数。DNN 具有强大的拟合和泛化能力,能够从历史决策序列中学习经验,并推广至未见的新状态,提高路由算法的适应性。

在 DQN 训练阶段,最大化操作可能导致时间差分(Temporal Difference,TD)目标高估真实价值,而自举机制会使偏差在训练过程中不断传播,进而引发高估问题,影响算法的稳定性和最终策略的性能。为缓解这一问题,DQN 引入目标神经网络,该网络以较低频率更新当前神经网络的参数,显著降低目标值与估计值之间的相关性,减少训练过程中的估计偏差,提升算法的稳定性和收敛性能。

具体而言,DQN 在训练过程中采用双重神经网络架构[19],包括当前 Q 网络 $Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)$ 和目标 Q 网络 $Q'(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}'_Q)$,其中 $\boldsymbol{\theta}'_Q$ 表示 $Q'(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}'_Q)$ 的参数。算法通过最小化损失函数 $L(\boldsymbol{\theta}_Q)$ 训练 $Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)$:

$$L(\boldsymbol{\theta}_Q)=E[(\hat{y}_t-Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q))^2],\tag{14}$$

$$\hat{y}_t=r_t+\gamma Q'(\boldsymbol{s}_{t+1},a^*;\boldsymbol{\theta}'_Q),\tag{15}$$

$$a^*=\underset{a\in\mathcal{A}}{\operatorname{argmax}} Q(\boldsymbol{s}_{t+1},a;\boldsymbol{\theta}_Q),\tag{16}$$

式中:TD 目标 $\hat{y}_t$ 由目标 Q 网络计算,以近似真实的 Q 值;动作 a^* 由当前 Q 网络计算。

通过对 $\boldsymbol{\theta}_Q$ 微分,可以求得 $L(\boldsymbol{\theta}_Q)$ 的梯度:

$$\nabla_{\boldsymbol{\theta}_Q}L(\boldsymbol{\theta}_Q)=E[(r_t+\gamma Q'(\boldsymbol{s}_{t+1},a^*;\boldsymbol{\theta}'_Q)-Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q))\nabla_{\boldsymbol{\theta}_Q}Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)]。\tag{17}$$

利用式(17)结合梯度下降法对 $L(\boldsymbol{\theta}_Q)$ 进行调优。在训练过程中,$Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)$ 在每一步执行完后都会使用随机梯度下降最小化 $L(\boldsymbol{\theta}_Q)$,以学习率 α 更新训练自身参数 $\boldsymbol{\theta}_Q$:

$$\boldsymbol{\theta}_Q\leftarrow\boldsymbol{\theta}_Q-\alpha\nabla_{\boldsymbol{\theta}_Q}L(\boldsymbol{\theta}_Q)。\tag{18}$$

此外,$Q'(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}'_Q)$ 在每一步执行完后都会采用软更新机制对 $\boldsymbol{\theta}'_Q$ 进行更新:$\boldsymbol{\theta}'_Q=\tau\boldsymbol{\theta}_Q+(1-\tau)\boldsymbol{\theta}'_Q$。在训练结束后,$Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)$ 的参数 $\boldsymbol{\theta}_Q$ 将不再改变。

在测试阶段,智能体依据训练好的 $Q(\boldsymbol{s}_t,a_t;\boldsymbol{\theta}_Q)$ 执行最优动作:

$$a_t^*=\underset{a\in\mathcal{A}_t}{\operatorname{argmax}} Q(\boldsymbol{s}_t,a;\boldsymbol{\theta}_Q)。\tag{19}$$

而在训练阶段为平衡探索和利用,采用贪婪算法选择动作:

$$a_t=\begin{cases}\underset{a\in\mathcal{A}_t}{\operatorname{argmax}} Q(\boldsymbol{s}_t,a;\boldsymbol{\theta}_Q),p_1=1-\varepsilon\\ \text{均匀抽取}\{a\mid a\in\mathcal{A}_t\},p_2=\varepsilon\end{cases},\tag{20}$$

式中:p_1 表示选择当前最优动作(利用)的概率,p_2 表示随机选择动作(探索)的概率,贪婪算法的参数 ε 随训练进程逐步衰减,以增强智能体的学习能力。

此外,为打破训练数据的时序相关性,提高学习稳定性,本文引入经验回放[20]机制。智能体在每个 Δt 中生成经验元组 $e_t=[\boldsymbol{s}_t,a_t,r_t,\boldsymbol{s}_{t+1}]$,并存入经验回放池。在训练时,算法周期性地从经验回放池中随机采样小批量数据优化损失函数,以减少样本间的相关性,提高算法的收敛性和泛化能力。

2.2 DHRP

本文基于 DRL 设计了 DHRP,该协议由离线训练与在线决策 2 个阶段组成。DHRP 基于历史飞行数据以离线方式训练模型,综合考虑双跳范围内的链路特征信息,并以深度值网络为核心实现最优下一跳的选择。在路由过程中,转发节点能够在数据包转发前预先规划路径,规避通信空白区域,进而有效降低端到端路由时延,并提升策略的在线适应性与鲁棒性。

针对 FANET 中的路由问题,考虑给定的当前状态 $\boldsymbol{s}_t$ 和任意动作 a_t,由于在单个 Δt 内节点的移动变化较小且链路信道容量波动幅度有限,假设二者的影响可忽略不计,因此可以在转发节点发送数据包前预测下一状态 $\boldsymbol{s}_{t+1}$。具体而言,$\boldsymbol{s}_{t+1}$ 实际上是节点 $i_t^{a_t}$ 在当前时间步 t 观察到的状态:

$$\boldsymbol{s}_{t+1}=\boldsymbol{s}(i_t^{a_t})=[\boldsymbol{x}(i_t^{a_t}),\boldsymbol{x}(i_t^{a_t,1}),\cdots,\boldsymbol{x}(i_t^{a_t,K}),\boldsymbol{x}(i_{\mathrm{d}}),C(i_t^{a_t},i_t^{a_t,1}),C(i_t^{a_t},i_t^{a_t,2}),\cdots,C(i_t^{a_t},i_t^{a_t,K})],\tag{21}$$

式中:$i_t^{a_t,k}$ 表示节点 $i_t^{a_t}$ 的第 k 个候选节点。

在路径选择前,转发节点 i_t 与候选节点 $i_t^{a_t}$ 之间的信道容量 $C(i_t,i_t^{a_t})$ 可通过式(2)计算得到。其中,$SINR(i,j)$ 可由物理层测得,$B(i,j)$ 由 MAC 协议分配并可通过 MAC 层查询。在时延计算方面,i_t 与 $i_t^{a_t}$ 之间的链路时延 $D_{\text{link}}(i_t,i_t^{a_t})$ 可依据式(1)预先计算,而排队时延 $D_{\text{queue}}(i_t)$ 则可以由 i_t 依据自身队列状态测得。由于本文的深度值网络模型采用历史飞行数据进行离线训练,因此可以通过对历史飞行数据进行离线统计,计算得到时延最大值 $D_{\max}$ 和信道容量最大值 $C_{\max}$。

为了充分利用双跳范围内链路特征信息,本文引入中间状态价值函数,其定义如下:

$$V_\pi(\boldsymbol{s}_t)=E\left[\omega\left[-\frac{D_{\text{link}}(i_t,i_t^{a_t})}{D_{\max}}\right]+(1-\omega)C_{\text{norm}}(i_t,i_t^{a_t})+\sum_{l=t+1}^{T}R_l\mid \boldsymbol{s}_t,\pi\right]。\tag{22}$$

该函数衡量数据包在 i_t 处经历排队时延后,依据 π 进行转发并最终抵达目标节点的期望回报。相应地,最优状态价值函数定义为:

$$V_*(\boldsymbol{s}_t)=\max_{\pi}V_\pi(\boldsymbol{s}_t)。\tag{23}$$

鉴于 $V_*(\cdot)$ 和 $Q_*(\cdot)$ 的关系,可得:

$$V_*(\boldsymbol{s}_t)+E\left[\omega\left[-\frac{D_{\text{queue}}(i_t)}{D_{\max}}\right]\right]=\max_{a\in\mathcal{A}_t}Q_*(\boldsymbol{s}_t,a)。\tag{24}$$

进一步地,通过 $V_*(\cdot)$ 将 $Q_*(\cdot)$ 表示为:

$$Q_*(\boldsymbol{s}_t,a)=E\left[r_t+\omega\left[-\frac{D_{\text{queue}}(i_t^a)}{D_{\max}}\right]\right]+V_*(\boldsymbol{s}_{t+1})。\tag{25}$$

由此可见,$Q_*(\cdot)$ 的学习可通过 $V_*(\cdot)$ 来实现。将式(25)代入式(24),并考虑 $\boldsymbol{s}_{t+1}=\boldsymbol{s}(i_t^a)$,可推导出 $V_*(\cdot)$ 的贝尔曼方程:

$$V_*(\boldsymbol{s}_t)=\max_{a\in\mathcal{A}_t}\{E[r(i_t,i_t^a)]+V_*(\boldsymbol{s}(i_t^a))\},\tag{26}$$

$$r(i_t,i_t^a)\triangleq\omega\left[-\frac{D_{\text{link}}(i_t,i_t^a)}{D_{\max}}\right]+\omega\left[-\frac{D_{\text{queue}}(i_t^a)}{D_{\max}}\right]+(1-\omega)C_{\text{norm}}(i_t,i_t^a)。\tag{27}$$

2.3 离线训练阶段

依据以上分析可得,学习 $V_*(\boldsymbol{s}_t)$ 即可得到 $Q_*(\boldsymbol{s}_t,a_t)$。因此,本研究采用所有节点共享的 DNN 模型 $V(\boldsymbol{s}_t;\boldsymbol{\theta}_V)$ 学习 $V_*(\boldsymbol{s}_t)$。相较于 DQN,该模型不依赖于动作空间,因此网络规模更小,显著减少了训练参数的数量,提高了计算效率。

本文旨在训练一个 DRL 决策模型,以嵌入 UAV 网络的历史拓扑信息并优化路由策略。由于训练过程与数据包流量无关,假设在训练期间所有 UAV 节点的排队时延相同且恒定,即总排队时延仅由路径中的跳数决定。基于该假设,$V(\boldsymbol{s}_t;\boldsymbol{\theta}_V)$ 利用历史飞行数据进行离线训练,各节点采用 ε-Greedy 策略进行数据包转发。依据式(26),利用的动作由式(28)决定:

$$a_t=\underset{a\in\mathcal{A}_t}{\operatorname{argmax}}[r(i_t,i_t^a)+V(\boldsymbol{s}(i_t^a);\boldsymbol{\theta}_V)]。\tag{28}$$

本文构建的深度值网络模型如图 3 所示。经验元组按照式(29)的形式存储于经验回放池,算法在每一步执行完后都会从中随机采样一批经验样本 $\mathcal{B}$,并依据式(26)利用随机梯度下降最小化损失函数 $E[(\hat{y}_t-V(\boldsymbol{s}_t;\boldsymbol{\theta}_V))^2]$,最终通过式(30)更新模型参数 $\boldsymbol{\theta}_V$。

$$\tilde{\boldsymbol{e}}_t=[\boldsymbol{s}_t,\boldsymbol{s}(i_t^1),\boldsymbol{s}(i_t^2),\cdots,\boldsymbol{s}(i_t^K),r(i_t,i_t^1),r(i_t,i_t^2),\cdots,r(i_t,i_t^K)],\tag{29}$$

$$\boldsymbol{\theta}_V\leftarrow\boldsymbol{\theta}_V+\frac{\alpha}{|\mathcal{B}|}\nabla_{\boldsymbol{\theta}_V}\sum_{\tilde{\boldsymbol{e}}_l\in\mathcal{B}}[\hat{y}_l-V(\boldsymbol{s}_l;\boldsymbol{\theta}_V)]^2。\tag{30}$$

若满足条件 $i_l^{a_l}=i_{\text{d}}$,则式(32)成立;否则,式(33)成立。其中,a_l 计算如下:

$$a_l=\underset{a\in\mathcal{A}_l}{\operatorname{argmax}}[r(i_l,i_l^a)+V(\boldsymbol{s}(i_l^a);\boldsymbol{\theta}_V)],\tag{31}$$

$$\hat{y}_l=r(i_l,i_l^{a_l}),\tag{32}$$

$$\hat{y}_l=r(i_l,i_l^{a_l})+V'(\boldsymbol{s}(i_l^{a_l});\boldsymbol{\theta}'_V),\tag{33}$$

式中:$V'(\cdot;\boldsymbol{\theta}'_V)$ 为目标值网络。参数 $\boldsymbol{\theta}'_V$ 通过式(34)进行更新:

$$\boldsymbol{\theta}'_V\leftarrow\tau\boldsymbol{\theta}_V+(1-\tau)\boldsymbol{\theta}'_V。\tag{34}$$

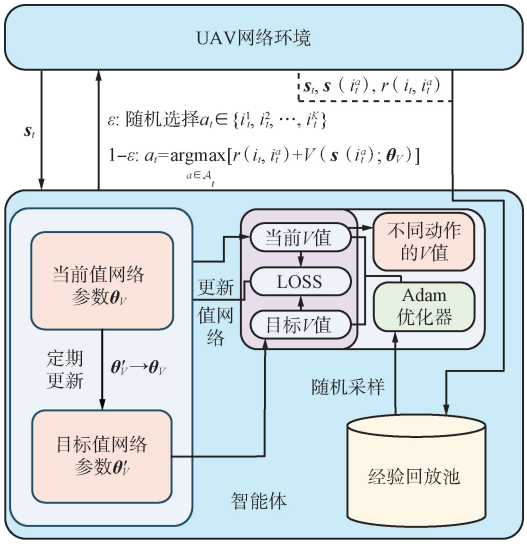


图 3 本文构建的深度值网络模型

Fig. 3 Deep value network model proposed in this paper

2.4 在线决策阶段

深度值网络在训练完成后将部署至每架 UAV,以支持在线路由决策。由于式(28)中的动作决策依赖于候选节点的状态等信息,即 $r(i_t,i_t^a)$ 和 $V(\boldsymbol{s}(i_t^a);\boldsymbol{\theta}_V)$。因此本文通过综合考虑双跳范围内链路特征信息,使转发节点能够在决策前获取关键数据,进而优化路径选择。具体而言,转发节点 i_t 在面临下一跳决策时,会先向其候选节点集合 $\mathcal{C}_{i_t} \triangleq \{i_t^1,i_t^2,\cdots,i_t^K\}$ 请求链路状态估计,即每个候选节点 i_t^a 估算式(27)中的相关变量,观察其状态 $\boldsymbol{s}(i_t^a)$ 并求得 $V(\boldsymbol{s}(i_t^a);\boldsymbol{\theta}_V)$。在完成 $r(i_t,i_t^a)+V(\boldsymbol{s}(i_t^a);\boldsymbol{\theta}_V)$ 的计算之后,每个 i_t^a 通过单跳广播将该值发送给 i_t。最终,i_t 基于接收到的信息选择最优动作:

$$a_t^* = \underset{a\in\mathcal{A}_t'}{\operatorname{argmax}}\left[r(i_t,i_t^a)+V(\boldsymbol{s}(i_t^a);\boldsymbol{\theta}_V)\right], \tag{35}$$

$$\mathcal{A}_t' \triangleq \{k \mid 1\leqslant k\leqslant K, i_t^k \neq i_1,i_2,\cdots,i_{t-1}\}, \tag{36}$$

式中:$\mathcal{A}_t'$ 用于避免路由循环。

为评估 DHRP 协议中反馈机制的通信负担,本文对单个候选节点反馈内容字段进行说明,如表 1 所示。每个候选节点需要反馈信道带宽、SINR、链路延迟、排队延迟以及状态价值 5 项信息,均为 32 位浮点型数据,共占用 20 byte。若每轮包含 10 个候选节点,额外的控制信息约为 200 byte。扩展型 HELLO 报文周期为 500 ms,即每秒发送 2 次,可得额外通信开销约为 400 B/s,远低于 FANET 中 1~5 MHz 信道带宽上限,带宽占用比例不到 1%,开销可接受,不会造成显著负担。

表 1 单个候选节点反馈内容字段说明

Tab. 1 Field description of feedback content from a single candidate node

字段	数据类型	字节数
信道带宽/MHz	float32	4
SINR/dB	float32	4
排队时延/ms	float32	4
链路时延/ms	float32	4
状态价值(本地计算)	float32	4

DHRP 通过结合双跳范围内的网络信息,突破了传统 DQN 路由算法仅依赖转发节点一跳信息进行路由决策的局限。该设计使转发节点在数据包转发前即可预规划路径,进而有效规避网络环境中的通信空白区域,提升算法的在线适应性。DHRP 的学习和决策过程如算法 1 所示。

算法 1 DHRP 选择算法

初始化参数 $\boldsymbol{\theta}_V$ 和参数 $\boldsymbol{\theta}_V'$

离线训练阶段

for *episode* = 1,2,…,N **do**

  在 UAV 历史飞行数据中随机抽取网络快照

  设置源节点位置 $\boldsymbol{x}(i_s)$ 和目标节点位置 $\boldsymbol{x}(i_d)$

  for t = 1,2,…,$t_{\max}$ **do**

    if 节点 $i_t^{a_t}$ 为目标节点或 t 超过最大跳数 $t_{\max}$ **then**

      break

    获取当前节点 i_t 的状态 $\boldsymbol{s}_t=\boldsymbol{s}(i_t)$

    以概率 ε 随机选择动作 $a_t\in\{i_t^1,i_t^2,\cdots,i_t^K\}$,或者以概率 $1-\varepsilon$ 根据式(28)选择动作 a_t。

    将由式(29)构成的经验元组 $\tilde{\boldsymbol{e}}_t$ 存储于经验回放池

    在经验回放池中随机采样一批经验样本 $\mathcal{B}$。

    依据式(30)和式(34)分别更新参数 $\boldsymbol{\theta}_V$ 和参数 $\boldsymbol{\theta}_V'$

在线决策阶段

Input:源节点位置 $\boldsymbol{x}(i_s)$,目标节点位置 $\boldsymbol{x}(i_d)$,参数 $\boldsymbol{\theta}_V$

for t = 1,2,…,$t_{\max}$ **do**

  if 当前节点 i_t 为目标节点 **then**

    break

  获取当前节点 i_t 的状态 $\boldsymbol{s}_t=\boldsymbol{s}(i_t)$

  for a = 1,2,…,K **do**

    获取每个候选节点 i_t^a 的状态 $\boldsymbol{s}(i_t^a)$,估算 $r(i_t,i_t^a)$ 进而得到 $r(i_t,i_t^a)+V(\boldsymbol{s}(i_t^a);\boldsymbol{\theta}_V)$,再将计算结果转发给当前节点 i_t

  依据式(35),当前节点 i_t 计算得到 a_t^*,并将数据包转发到节点 $i_t^{a_t^*}$

3 仿真分析

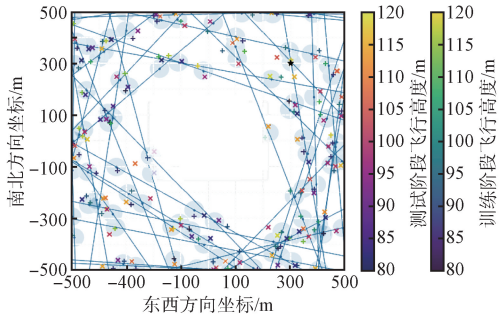
3.1 模拟环境

本文采用模拟生成的 UAV 网络快照作为训练和测试数据源,以精确控制网络环境的动态特性,包括节点位置、通信参数及拓扑结构等。通过构建多样化的仿真场景,为算法性能验证提供了可靠支撑,有效弥补了真实数据获取的局限性。

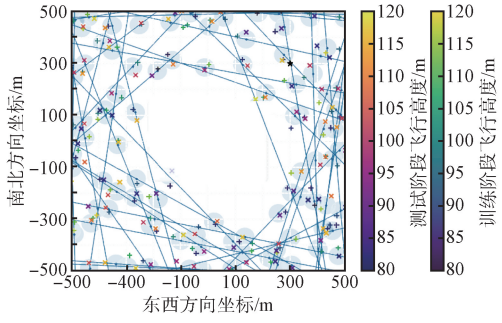
本文设定的三维空间范围为东西方向-500~500 m、南北方向-500~500 m、海拔 80~120 m,其二维投影如图 4 所示。其中,蓝色线条表示预先规划飞行路径,蓝色圆点表示预先规划节点位置,五角黑星表示基站位置,阴影区域表示某些 UAV

节点出现概率较高的区域。在该空间范围内,随机生成 40 条固定飞行路径,并在整个模拟过程中保持不变,以确保路径一致性和实验可重复性,为路由算法的客观评估提供稳定基准。此外,40 条飞行路径上均匀分布 100 架 UAV,同一路径上的 UAV 以同向匀速飞行,并保持安全间距,且每架 UAV 的飞行高度随机分布于海拔 80~120 m。如果 2 架 UAV 之间可建立直接通信链路,则其信道带宽和 SINR 值会根据节点间的距离动态赋值,且带宽和 SINR 值与距离成反比,以模拟无线信道的空间相关性和动态特性。

在实际应用中,UAV 可能因多种原因未能按时起飞或偏离预定飞行路径,导致历史飞行位置(训练数据)与当前飞行位置(测试数据)存在偏差。为模拟这一现象,本文在预定节点位置的基础上,引入服从瑞利分布的随机偏差,生成用于训练和测试路由算法的合成飞行位置,即图 4 中的十字符号和叉形符号,分别表示训练阶段和测试阶段的节点位置。此外 UAV 位置随着时间步的推移不断变化,且同一 UAV 在训练集和测试集中的位置不同。



(a) t_1 时刻飞行位置



(b) t_{50} 时刻飞行位置

图 4　不同时刻的 UAV 路径示例快照

Fig. 4　Example snapshots of UAV paths at different time steps

为评估路由算法在通信空白场景下的适应性,并反映典型的非均匀飞行密度分布,在构建 UAV 网络快照时引入了“禁飞区域”的建模方式。禁飞区域设定为 UAV 不可进入并且无法建立通信连接,用于抽象模拟由地形遮挡、政策限制或信号阻断等因素导致的链路不可用情形,进而功能性地还原通信空白区域对路由策略的影响。本文设定 2 个禁飞区域,分别位于(105,-105,0) m 和(-105,105,0) m,半径均为 150 m,海拔 60~120 m,如图 5 所示,其中,红色虚线表示通信空白区域。

本文共生成了 2 000 个 UAV 网络快照,其中训练集和测试集各占一半,用于评估算法在未知数据上的性能。在每个快照中,源节点随机选定,目标节点固定为位于(300,300,60) m 的地面基站。

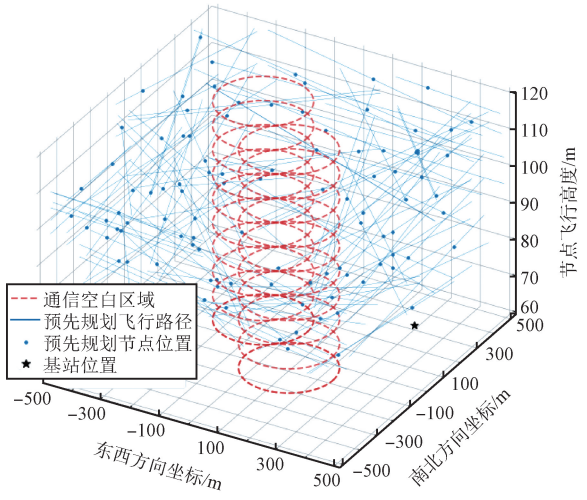


图 5　t_1 时刻的 3D UAV 路径示例快照

Fig. 5　3D Example Snapshot of UAV Paths at Time t_1

3.2　实验设置

本文使用 Python3.6 编写事件驱动模拟器,并以 Tensorflow 作为深度学习框架实现 DRL 路由算法,通过生成 UAV 网络快照来模拟 FANET 网络环境。每架 UAV 的通信半径设为 180 m,数据包大小设为 S = 128 kB ,链路间数据传输速率由信道容量决定,并用于计算数据包的传输时延。训练过程中,排队时延设置为 $D_{\text{queue}}(i)$ = 5 ms。

动作候选集大小设置为 K=10,深度值网络有 2 个隐藏层,神经元数量为 50×50。隐藏层采用 ReLU 激活函数,而输出层未使用激活函数。训练过程中,优化器选用 Adam,学习率设为 α = 0.000 1,目标网络参数更新的权重系数设为 τ = 0.001。更新过程中采用小批量梯度下降法,经验回放批次大小

为 32。在训练阶段,贪婪算法的参数在前 100 回合设置为$\varepsilon=1$,随后 400 回合内衰减至 $\varepsilon=0.1$,并在剩余回合中保持 $\varepsilon=0.1$。在测试阶段,贪婪算法的参数固定为 $\varepsilon=0$,并冻结深度值网络的参数。具体参数设置如表 2 所示。

表 2　仿真实验参数

Tab. 2　Simulation Parameters

仿真参数	设定值
仿真区域大小/m	1 000×1 000
节点飞行高度/m	80~120
UAV 节点数量	60~100
飞行路径数量	40
通信半径/m	180
数据包大小/kB	128
排队时延/ms	5
拥堵节点排队时延/ms	20
信道带宽/MHz	1~5
SINR/dB	10~40
经验回放批次大小	32
经验回放池大小	1 000 000
动作候选集大小 K	10
学习率 α	0.000 1
折扣因子 γ	1
权衡系数 ω	0.75
权重系数 τ	0.001
贪婪算法参数 ε	0~1

3.3　实验结果与分析

为评估 DHRP 路由协议在最小化平均端到端时延方面的性能,本文将其与 GPSR[12]、DQN[16] 和 DVN[16] 三种路由协议进行仿真对比。上述基于 RL 算法的路由协议均包含训练和测试 2 个阶段,在训练阶段选择总节点数量为 100 作为基准值。为验证模型的泛化能力,测试阶段采用与训练阶段不同的网络快照,以评估其在未见拓扑结构下的适应性和有效性。

图 6 展示了 4 种路由协议在训练过程中平均端到端时延的变化趋势。每种协议均独立训练 100 次,曲线表示平均值,阴影区域表示标准差。相比之下,DHRP 在学习效率方面表现优异,训练初期即可迅速降低时延,有效规避了 DQN 和 DVN 因早期决策不稳定带来的高时延问题。DHRP 在约 400 个回合后趋于收敛且波动较小,而 DVN 则需超过 500 个回合才能逐渐收敛。由于在决策中结合了链路质量信息,DHRP 在收敛后表现出最低的平均端到端时延,展现出良好的拓扑适应性,能够有效减少通信空白区域的影响,通过与环境的持续交互学习优化路由策略,进而显著降低平均端到端时延。

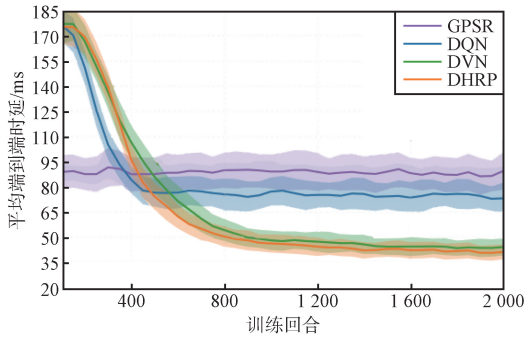


图 6　不同路由协议在训练阶段的学习曲线

Fig. 6　Learning curves of different routing protocols during the training phase

图 7 和图 8 分别展示了在不同节点数量条件下,4 种路由协议的平均端到端时延与端到端时延抖动对比结果。随着节点数量的增加,各协议的时延和抖动均有所波动,但 DHRP 始终保持最低的时延与最小的抖动,在各项指标上均显著优于 DQN 和 DVN,并远优于 GPSR。GPSR 依赖贪心转发策略,在通信空白场景下容易出现路径绕行或转发失败,导致端到端时延大幅上升。尽管 DQN 和 DVN 作为基于 DRL 的方法,在一定程度上缓解了此类问题,但在面对动态拓扑结构时,其时延波动仍比较明显。DHRP 通过引入双跳范围内的链路特征信息,使转发节点能够提前获取式(27)中的关键变量,并结合候选节点的局部观测,实现路径的预规划,进而有效减弱通信空白区域对路由决策的干扰,显著降低平均端到端时延。

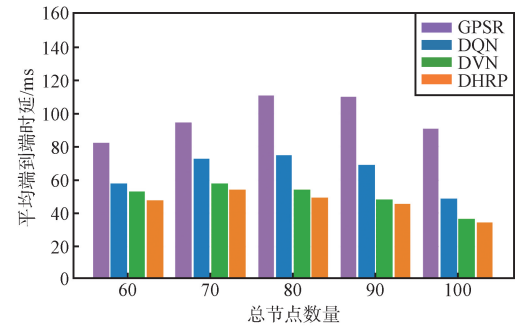


图 7　总节点数量不同时对应的平均端到端时延

Fig. 7　Average end-to-end delay under different total numbers of nodes

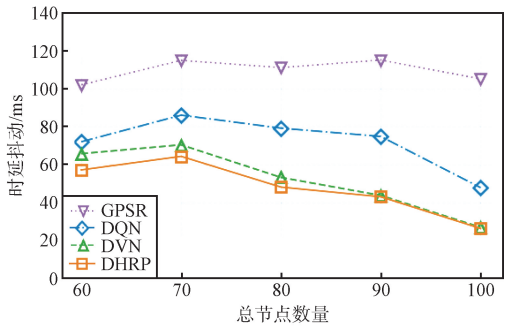


图 8　总节点数量不同时对应的端到端时延抖动

Fig. 8　End-to-end delay jitter under different total numbers of nodes

同时,DHRP 在奖励函数设计中综合考虑信道容量和时延 2 个关键指标,引导智能体优先选择高容量、低波动的链路,进一步降低由链路质量不稳定引起的端到端时延抖动。实验结果表明,DHRP 在节点规模变化的场景下,能够持续地做出高效、稳定的路由决策,进而提升网络整体的传输性能与鲁棒性。

图 9 和图 10 分别展示了在总节点数为 100 时,4 种路由协议在不同拥堵节点比例下的平均端到端时延和端到端时延抖动对比结果。实验结果表明,DHRP 在各类拥堵水平下均能保持较低的时延,其时延随拥堵程度增加而增长的幅度相对较小,且端到端时延抖动最低,展现出良好的稳健性。在高拥堵场景下,GPSR 的时延急剧上升,而 DQN 和 DVN 也表现出不同程度的性能退化。相比之下,DHRP 能实时感知网络环境变化,智能调整路由决策,使转发节点在数据转发前获取双跳范围内的链路状态信息,包括信道容量和时延信息,实现更合理的中继节点选择,规避通信空白区域的影响,进而寻找到时延更低且路径更稳定的路由,使数据包能够更快速、稳定地到达目标节点。

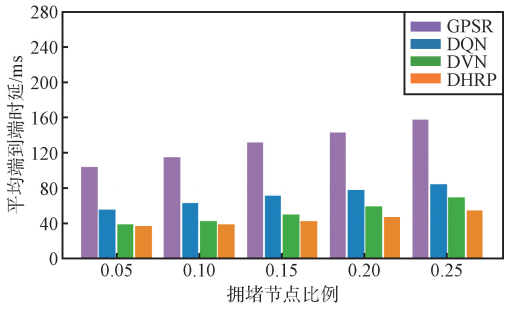


图 9　不同拥堵节点比例对应的平均端到端时延

Fig. 9　Average end-to-end delay under different congested node ratios

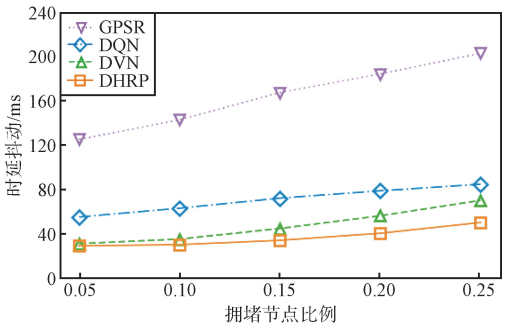


图 10　不同拥堵节点比例对应的端到端时延抖动

Fig. 10　End-to-end delay jitter under different congested node ratios

上述结果表明,DHRP 在面对不同拥堵节点比例时,依然保持良好的鲁棒性与适应性,能够在网络拥堵加剧的情况下有效降低平均端到端时延,进而提升整体网络性能。

4　结束语

在通信空白场景下,FANET 路由协议需要考虑高时延问题。本文提出了一种 DRL 辅助的 DHRP。该协议在奖励函数中结合了实时信道容量与时延信息,通过动态调整路由机制引导节点选择最优的下一跳。DHRP 基于历史飞行数据以离线方式训练深度值网络模型,综合考虑双跳范围内的网络信息,进而在动态环境中实现最优路由决策。实验结果表明,相比于其他协议,DHRP 在不同节点规模和网络拥堵条件下均能保持更低的平均端到端时延和端到端时延抖动,有效减少通信空白区域的影响,展现出优越的适应性和鲁棒性。本文所提方法目前仅在构建的高保真仿真环境中进行了验证,尚未部署于真实的 FANET 系统。未来研究将致力于将该方法推广至更贴近现实的半实物或硬件平台中,以进一步验证其性能并进行优化。

参 考 文 献

[1] KAKAMOUKAS G A, SARIGIANNIDIS P G, ECONOMIDES A A. FANETs in Agriculture-A Routing Protocol Survey[J]. Internet of Things, 2022, 18: 100183.

[2] FENG W M, TANG J, YU Y, et al. UAV-enabled SWIPT in IoT Networks for Emergency Communications[J]. IEEE Wireless Communications, 2020, 27(5): 140-147.

[3] BOURSIANIS A D, PAPADOPOULOU M S, DIAMANTOULAKIS P, et al. Internet of Things (IoT) and Agricultural Unmanned Aerial Vehicles (UAVs) in Smart Farming: A Comprehensive Review[J]. Internet of Things, 2022, 18: 100187.

[4] BUSHNAQ O M,CHAABAN A,AL-NAFFOURI T Y. The Role of UAV-IoT Networks in Future Wildfire Detection [J]. IEEE Internet of Things Journal, 2021, 8 (23): 16984-16999.

[5] KHAWAJA W, OZDEMIR O, GUVENC I. UAV Air-to-Ground Channel Characterization for mmWave Systems [C]//2017 IEEE 86th Vehicular Technology Conference (VTC-Fall). Toronto:IEEE,2017:1-5.

[6] ZHAO N, LU W D, SHENG M, et al. UAV-assisted Emergency Networks in Disasters [J]. IEEE Wireless Communications,2019,26(1):45-51.

[7] BEKMEZCI I, SAHINGOZ O K, TEMEL Ş. Flying Ad-Hoc Networks (FANETs): A Survey [J]. Ad Hoc Networks,2013,11(3):1254-1270.

[8] MALHOTRA A, KAUR S. A Comprehensive Review on Recent Advancements in Routing Protocols for Flying Ad Hoc Networks[J]. Transactions on Emerging Telecommunications Technologies,2022,33(3):e3688.

[9] 郭彦芳. 一种改进的基于能量优化的 AODV 路由协议[J]. 无线电通信技术,2016,42(4):25-28.

[10] PERKINS C E, BHAGWAT P. Highly Dynamic Destination-sequenced Distance-vector Routing (DSDV) for Mobile Computers[J]. ACM SIGCOMM Computer Communication Review,1994,24(4):234-244.

[11] JACQUET P, MUHLETHALER P, CLAUSEN T, et al. Optimized Link State Routing Protocol for Ad Hoc Networks[C]//Proceedings. IEEE International Multi Topic Conference,2001. IEEE INMIC 2001. Technology for the 21st Century. Lahore:IEEE,2001:62-68.

[12] KARP B,KUNG H T. GPSR:Greedy Perimeter Stateless Routing for Wireless Networks[C]//Proceedings of the 6th Annual International Conference on Mobile Computing and Networking. New York:ACM,2000:243-254.

[13] 郑莹,段庆洋,林利祥,等. 深度强化学习在典型网络系统中的应用综述[J]. 无线电通信技术,2020,46(6):603-623.

[14] JUNG W S,YIM J,KO Y B. QGEO:Q-learning-based Geographic Ad Hoc Routing Protocol for Unmanned Robotic Networks[J]. IEEE Communications Letters, 2017, 21(10):2258-2261.

[15] LYU N Q,SONG G H,YANG B W,et al. QNGPSR:A Q-Network Enhanced Geographic Ad-Hoc Routing Protocol Based on GPSR[C]//2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). Chicago:IEEE,2018:1-6.

[16] LIU D,CUI J J,ZHANG J K,et al. Deep Reinforcement Learning Aided Packet-routing for Aeronautical Ad-Hoc Networks Formed by Passenger Planes[J]. IEEE Transactions on Vehicular Technology,2021,70(5):5166-5171.

[17] LIU J M,WANG Q,HE C T,et al. ARdeep:Adaptive and Reliable Routing Protocol for Mobile Robotic Networks with Deep Reinforcement Learning[C]//2020 IEEE 45th Conference on Local Computer Networks (LCN). Sydney:IEEE,2020:465-468.

[18] FAREED M M,UYSAL M. On Relay Selection for Decode-and-Forward Relaying[J]. IEEE Transactions on Wireless Communications,2009,8(7):3341-3346.

[19] 米洪,郑莹. 基于双深度 Q 网络的车联网安全位置路由[J]. 无线电通信技术,2025,51(1):96-105.

[20] LIU Y Z,MATTAR M G,BEHRENS T E J,et al. Experience Replay is Associated with Efficient Nonlocal Learning[J]. Science,2021,372(6544):eabf1357.

作者简介:

郭歆莹 女,(1989—),博士,副教授,硕士生导师。主要研究方向:5G/6G 智能通信、无人机通信。

李 明 男,(2001—),硕士研究生。主要研究方向:无人机自组网、强化学习。

朱春华 女,(1976—),博士,教授,博士生导师。主要研究方向:宽带无线通信、通信信号处理、智能信息处理。