文章编号:1671-4598(2025)08-0301-08 DOI:10.16526/j.cnki.11-4762/tp.2025.08.037 中图分类号:TP393.04 文献标识码:A

基于 PPO 的移动自组网自适应信道接入算法

王飞越，宋政育，秦鑫彤，孙 昕
(北京交通大学 电子信息工程学院，北京 100044)

摘要：针对移动自组网中 p 坚持 CSMA 协议的信道接入难题，提出一种基于 PPO 的自适应信道接入算法；首先构建以最大化节点信道利用率为目标的 p 坚持 CSMA 竞争接入优化问题；将优化问题建模为马尔可夫决策过程并设计包含网络状态信息、多维动作空间以及效用奖励函数的强化学习框架；采用 PPO 动态优化节点竞争概率、竞争概率增长因子以及准入节点数量关键变量，实现自适应信道接入；经过仿真验证，所提算法的收敛速度优于基于 DDPG 的接入方案；所提算法的信道利用率相较于固定准入节点数、固定竞争概率以及预设竞争概率方案分别提高 33.3%、48.1%和 18.9%，并且在 35 节点以内的网络规模下始终优于其他方案；所提算法还具有业务优先级区分机制，收敛后高优先级业务数据节点接入成功率可达 90%以上。

关键词：MANET；p 坚持 CSMA 协议；深度强化学习；信道利用率；优先级区分

Adaptive Channel Access Algorithm for MANETs Based on PPO

WANG Feiyue, SONG Zhengyu, QIN Xintong, SUN Xin
(School of Electronic Information Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: To address channel access problem in mobile ad hoc networks (MANETs), p-persistent CSMA protocol is used to propose an adaptive channel access algorithm based on proximal policy optimization (PPO). Firstly, a p-persistent CSMA competitive access optimization problem with the goal of maximizing node channel utilization is constructed. Then, this problem is modeled as a Markov decision process (MDP) with specifically designed network state information, multi-dimensional action space, and reward functions. Finally, the PPO algorithm is employed to dynamically optimize node competition probability, competition probability growth factor, and number of permitted access nodes, thereby achieving adaptive channel access. Simulation results demonstrate that the proposed algorithm has a notable advantage over the scheme based on deep deterministic policy gradient (DDPG) in the convergence speed. In terms of channel utilization, the proposed algorithm improves 33.3%, 48.1%, and 18.9% compared to fixed permitted-node-number schemes, fixed competition probability schemes, and preset probability schemes, respectively, and outperforms other solutions in networks with up to 35 nodes. Additionally, the algorithm incorporates a service priority differentiation mechanism, and the access success rate for high-priority service nodes reaches over 90% after convergence.

Keywords: MANET; p-persistent CSMA protocol; deep reinforcement learning; channel utilization; priority differentiation

0 引言

移动自组网（MANET，mobile Ad hoc network）具有分布式架构、自组织特性以及灵活动态的组网方式，在近年来吸引了学术界以及工业界的深入研究与广泛应用。MANET 的媒体接入控制（MAC，media access control）协议决定了网络资源的接入、分配和调度方式，负责解决多用户之间高效、合理地共享有限信道资源的问题，对于保证 MANET 的性能至关重要[1-2]。由于自组网节点频繁移动，网络拓扑持续动态变化，且业务需求呈现多样化特征，传统 MAC 协议在这种高动态网络环境难以保持较优性能，因此需要采用灵活高

收稿日期:2025-02-26； **修回日期**:2025-03-27。
基金项目:国家自然科学基金项目(61901027)。
作者简介:王飞越(1999-),男,硕士研究生。
孙 昕(1967-),女,博士,教授。
通讯作者:宋政育(1984-),男,博士,副教授。
引用格式:王飞越,宋政育,秦鑫彤,等.基于 PPO 的移动自组网自适应信道接入算法[J].计算机测量与控制,2025,33(8):301-308,318.

效、具有自适应性以及能够满足差异化业务需求的信道接入算法。

MANET 信道接入协议可以分为基于竞争机制的 MAC 协议（例如 IEEE 802.11 MAC 协议）、基于分配机制的 MAC 协议（例如时分多址协议，TDMA，time division multiple access）以及混合类 MAC 协议（例如 Z-MAC 协议，Zebra Media Access Control）[3]。现有研究通常针对 MAC 协议中的单一参数进行优化[4]。文献[5]根据碰撞次数动态调整退避范围，在不同网络负载下优化退避策略，提升了 CSMA/CA 协议在 ZigBee 网络中的性能；文献[6]通过监控网络丢包率，动态调整退避窗口大小，可适应不同网络流量条件，改进了 Z-MAC 协议的吞吐量和信道利用率。随着 IEEE 802.11 CSMA/CA 协议的出现，p 坚持 CSMA 协议因其简单性和碰撞规避能力得到了许多学者的关注。文献[7]通过理论分析和数学建模推导出了 p 坚持 CSMA 最优竞争概率的解析公式，并且提出了一种动态调整机制，该机制可通过监测网络中的碰撞和空闲时间，自动调整 p 值以接近最优容量。文献[8]通过建立多目标优化问题，表征了 p 坚持 CSMA 系统稳定性区域的边界，并且通过监测队列长度和信道状态，动态调整用户的重传概率，以保持系统的稳定性。虽然上述研究提供了多种性能优化方法，但是大多数方法面临环境先验信息依赖性强、计算开销大、动态适应性不足等瓶颈。

近年来，深度强化学习（DRL，deep reinforcement learning）在无线网络中的研究与应用为解决上述问题提供了新思路。利用 DRL 强大的函数逼近能力和自适应学习优势可以有效提高网络性能。文献[9]将深度 Q 网络（DQN，deep Q-network）和双深度 Q 网络（DDQN，double DQN）技术应用于无线自组网，通过深度强化学习为每个 IoT 设备选择时隙和传输波束，以最大化网络吞吐量；文献[10]以最小化所有节点的加权端到端时延为目标建立优化问题，基于 AC 强化学习算法优化 TDMA 时隙调度策略，减少低优先级流量的加权端到端延迟，同时保证高优先级流量的服务质量（QoS，quality of service）需求；文献[11]提出一种指针一评论家架构的 DRL 模型用以解决未授权频段中的无线网络共存问题，以最大化小基站节点的占用时间和小基站节点间的公平性为目标，对是否共享当前窗口给小基站节点以及具体的共享对象进行优化，提高了小基站节点间的公平性。文献[12]在异构无线网络场景下，基于历史动作和信道观测结果，采用 DQN 方法优化了节点在每个时隙中是否传输数据或者等待的动作，有效提升了系统的总吞吐量并且实现了一公平性。然而现有的 DRL 方案聚焦的优化目标大多为吞吐量和时延等，缺乏对节点优先级（如接入失败的节点应在下一时隙优先接入）和信道利用率的考虑。此外，MANET 的高动态性要求算法具备快速收敛能力以及策略稳定性，但传统的 DRL 方法（如 DQN）因策略更新震荡以及样本效率低下等缺点而难以满足需求[13]。近年来，近端策略优化（PPO，proximal policy optimization）算法作为一种创新的策略梯度算法，在应对复杂环境和任务时表现出卓越的性能，在强化学习领域受到了研究者的广泛关注。PPO 算法基于 Actor-Critic 框架，采用神经网络拟合价值函数和策略函数，能够同时学习价值网络和策略网络，可以有效处理非线性、高维以及连续的状态和动作空间；PPO 算法还引入了裁剪目标函数和重要性采样技术，有效提升了模型训练的稳定性以及样本效率，能够在动态的 MANET 环境下快速适应环境变化并做出稳健的决策，适用于处理 MANET 的信道接入问题。与深度确定性策略梯度（DDPG，deep deterministic policy gradient）等其他强化学习算法相比，PPO 在策略稳定性方面具有明显优势，实现简单且易于调参，计算效率较高，适用于资源受限的 MANET 节点部署。

此外，在 MANET 的应用中，为了满足差异化的 QoS 需求，MAC 协议需要具备业务优先级区分能力[14]。文献[15]通过引用访问类别来区分不同业务的优先级，其中每个访问类别都对应独立的退避实体，并分别以不同的退避参数来实现差异化的信道访问优先级；文献[16]通过二进制签名数字区分数据包的优先级，规定高优先级数据包发送突发反馈信号，并且限制低优先级数据包节点进入后续竞争阶段，以保证高优先级流量的 QoS。在上述研究中，优先级参数往往是预先配置的固定值，难以适应复杂多变的网络环境，并且均未从节点接入成功率的角度分析优先级区分机制的效果。此外，传统的优先级机制往往独立工作，节点需要为此专门维护相关的信息，造成了额外的控制开销，也增加了 MAC 协议整体的复杂性。

针对上述挑战，本文研究基于 p 坚持 CSMA 协议的移动自组网信道接入问题，以信道利用率最大化为目标，提出基于近端策略优化的自适应信道接入（PPO-ACA，proximal policy optimization based adaptive channel access）算法，对节点竞争概率、竞争概率增长因子以及准入的节点数量进行联合优化，同时对业务进行优先级区分。仿真结果表明，本文提出的算法可以有效利用信道资源，与基于 DDPG 的方案相比收敛速度更快，同时能够保证高优先级业务的节点具有较高的接入成功率。

1 系统模型及问题建模

1.1 系统模型

假设一个规模为节点的无中心分布式移动自组网，

拓扑结构如图 1 所示。网络中运行 p 坚持 CSMA 协议，每个网络节点均工作在频分双工模式下，可同时作为源节点竞争接入以及作为目的节点接收数据。

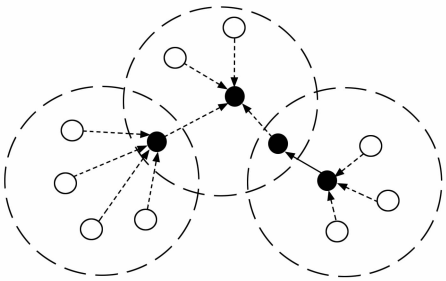

图 1　移动自组网拓扑结构

移动自组网中 p 坚持 CSMA 协议的时帧结构如图 2 所示。记一个时隙的长度为，每个时隙被划分为参数重置、竞争接入、信息广播以及数据传输 4 个阶段，其中数据传输阶段又被进一步划分为若干微时隙。

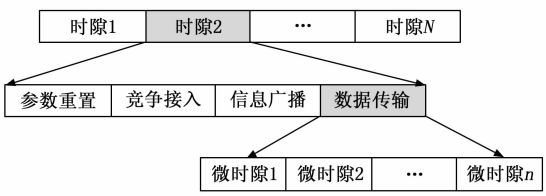

图 2　p 坚持 CSMA 协议的时帧结构

区别于传统的 p 坚持 CSMA 机制，本文在每个时隙的开始引入了参数重置阶段。网络节点在该阶段内根据网络环境的变化自适应地重置协议参数，从而在后续阶段中以更优的信道接入策略参与竞争，进而改善信道资源的利用情况。对于网络节点 m，在时隙 n 的参数重置阶段中将会对竞争接入时间 $t_{m,n}^{C}$、竞争接入概率 $p_{m,n}$ 和竞争概率增长因子 $v_{m,n}$ 等参数进行重新配置。

在时隙 n 的竞争接入阶段，节点 m 的任意一跳邻居节点都可在竞争接入时间 $t_{m,n}^{C}$ 内参与竞争以获得数据传输阶段的接入机会。当时间超出 $t_{m,n}^{C}$ 范围或者已接入节点达到准入节点数上限 $K_{m,n}$ 时，竞争接入阶段结束。

在信息广播阶段，节点 m 将上一阶段中竞争成功的节点信息广播至所有邻居节点，并为这些节点分配数据传输阶段中的微时隙。

在数据传输阶段，准入节点在分配的微时隙中与节点 m 进行数据传输。

本文假设在时隙 m 中业务数据包以相同的到达率 m 到达各节点，并且数据包到达服从泊松分布。尽管实际的 MANET 环境中不同节点的业务数据包到达情况可能存在差异，但相同到达率的泊松分布已经被用于描述 MANET 业务到达[17]，该假设简化了理论分析，并且有助于聚焦协议机制本身。记欲与节点 m 接入而参与竞争的节点数量为 $R_{m,n}=F_{m,n-1}+D_{m,n-1}$，其中：$R_{m,n-1}$ 为上一时隙中竞争失败的节点数量；$R_{m,n-1}$ 为上一时隙中有新到达数据包的节点数量。

由于竞争失败的节点数是所有参与竞争的节点 $R_{m,n-1}$ 中除去准入节点数上限 $K_{m,n-1}$ 的部分，因此上一时隙中竞争失败的节点数 $F_{m,n-1}$ 可具体表示为：

$$F_{m,n-1}=R_{m,n-1}-K_{m,n-1} \tag{1}$$

由于不同网络节点待传输的数据包具有不同的业务优先级，因此 $F_{m,n-1}$ 还可以表示为：$F_{m,n-1}=F_{m,n-1}^{H}+F_{m,n-1}^{L}$，其中：$F_{m,n-1}^{H}$ 为竞争失败的节点中待传输高优先级业务数据包的节点数；$F_{m,n-1}^{L}$ 为竞争失败的节点中待传输低优先级业务数据包的节点数。以待传输高优先级业务数据的节点为例，竞争失败的节点将在参数重置阶段修改其竞争概率 $p_{m,n}^{H}$ 为：

$$p_{m,n}^{H}=(1+a^{H}v_{m,n})p_{m,n} \tag{2}$$

其中：a^{H} 为高优先级业务的重要性系数，描述了高优先级业务相对低优先级业务的重要程度比值，可以根据业务的实际情况进行调整；$p_{m,n}$ 为时隙 n 中网络节点 m 的基础竞争概率。同时，低优先级业务的节点也依据其业务的重要性系数对其竞争概率进行类似的修改操作。对竞争失败的高、低优先级业务节点进行不同的竞争概率调整，可与节点的其他参数重置操作同时进行，因此可以在不额外增加控制开销的情况下实现对高、低优先级业务的区分。

定义信道中连续两次有效数据传输之间的间隔时间为虚拟传输间隔[18]。以时隙 n 为例，记 $t_{m,j,n}$ 为节点 m 第 $i-1$ 次和第 i 次成功传输之间的时间间隔，信道在该时间间隔内可能包含若干次传输碰撞事件或者空闲侦听时段，直至出现新的成功传输事件，虚拟传输间隔如图 3 所示。

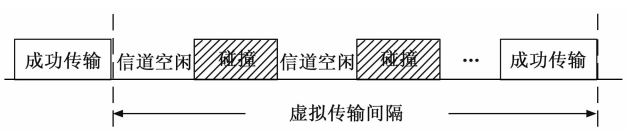

图 3　虚拟传输间隔

为了对虚拟传输间隔进行建模，需要分别表示出 $t_{m,j,n}$ 范围内信道经历的空闲和碰撞次数以及单次空闲状态和碰撞状态的平均时长。

在 $t_{m,j,n}$ 中发生碰撞的平均次数 $E[N_C]$ 为：

$$E[N_C]=\frac{1-(1-p_{m,n})^{D_{m,n-1}}(1-p_{m,n}^{H})^{F_{m,n-1}^{H}}(1-p_{m,n}^{L})^{F_{m,n-1}^{L}}}{D_{m,n-1}(1-p_{m,n})^{D_{m,n-1}-1}(1-p_{m,n}^{H})^{F_{m,n-1}^{H}}(1-p_{m,n}^{L})^{F_{m,n-1}^{L}}+\varphi}-1 \tag{3}$$

$$\varphi=(F_{m,n-1}^{H}+F_{m,n-1}^{L})$$

$$(1-p_{m,n}^{H})^{F_{m,n-1}^{H}-1}(1-p_{m,n}^{L})^{F_{m,n-1}^{L}}(1-p_{m,n})^{D_{m,n-1}} \quad (4)$$

如图 3 所示，易知 $t_{m,j,n}$ 中信道空闲的平均次数为 $E[N_C]+1$。

信道中单次空闲状态的平均时长 $E[Idle]$ 为：

$$E[Idle]=T_{Idle}\frac{(1-p_{m,n})^{D_{m,n-1}}(1-p_{m,n}^{H})^{F_{m,n-1}^{H}}(1-p_{m,n}^{L})^{F_{m,n-1}^{L}}}{1-(1-p_{m,n})^{D_{m,n-1}}(1-p_{m,n}^{H})^{F_{m,n-1}^{H}}(1-p_{m,n}^{L})^{F_{m,n-1}^{L}}} \quad (5)$$

其中：T_{Idle} 表示数据传输阶段微时隙的长度。

信道中单次碰撞状态的平均时长 $E[coll]$ 为：

$$E[coll]=T_{\text{req}}+T_{\text{BIFS}} \quad (6)$$

其中：T_{req} 为发送接入请求数据包所需的时间长度；T_{BIFS} 为退避帧间间隔，用于避免冲突和实现公平的信道访问，此处为发生碰撞后的退避时间开销。

成功传输 $E[Suc]$ 的平均时间长度[19]为：

$$E[Suc]=T_{\text{req}}+T_{\text{SIFS}}+T_{\text{ACK}}+T_{\text{BIFS}} \quad (7)$$

其中：T_{ACK} 为发送确认数据包所需的时间长度；T_{SIFS} 为短帧间间隔，用于需要立即响应的情况（如 ACK 帧的发送），此处为发送 req 后的等待时间开销。

由上述公式，虚拟传输间隔 $t_{m,j,n}$ 的平均时长 $E[t_{m,i,n}]$ 表示为：

$$E[t_{m,i,n}]=E\Big[\sum_{m=1}^{N_C}(Idle_m+Coll_m)\Big]+E[Idle_{N_C+1}]+E[Suc]=(E[N_C]+1)E[Idle]+E[N_C]E[Coll]+E[Suc] \quad (8)$$

综上所述，节点 m 在时隙 n 中的竞争接入时间 $t_{m,n}^{c}$ 的平均时长 $E[t_{m,n}^{c}]$ 为：

$$E[t_{m,n}^{C}]=E\Big[\sum_{i=1}^{K_{m,n}}t_{m,j,n}\Big]=K_{m,n}E[t_{m,j,n}] \quad (9)$$

1.2 优化问题建模

本文在移动自组网场景下，以最大化 p 坚持 CSMA 协议下网络节点的信道利用率为目标建立优化问题，优化每个时隙中的节点竞争概率、竞争概率增长因子以及准入节点数量，并将时隙长度限制、参与竞争的节点数量关系以及竞争概率等参数的取值范围作为约束条件。该优化问题可表示为：

$$\begin{aligned}&\max_{K_{m,n},p_{m,n},\nu_{m,n}}\frac{1}{N}\sum_{n=1}^{N}K_{m,n}\tau/T\\ &\text{s.t. } C_1: E[t_{m,n}{}^{C}]+K_{m,n}\tau\leqslant T,\forall m,n\\ &\quad C_2: R_{m,n}=R_{m,n-1}-K_{m,n-1}+D_{m,n-1},\forall m,n\\ &\quad C_3: 0<p_{m,n}\leqslant 1,\forall m,n\\ &\quad C_4: 0<p_{m,n}^{H}\leqslant 1,\forall m,n\\ &\quad C_5: 0<p_{m,n}^{L}\leqslant 1,\forall m,n\\ &\quad C_6: 0<\nu_{m,n}\leqslant 1,\forall m,n\end{aligned} \quad (10)$$

其中：$\frac{1}{N}\sum_{n=1}^{N}K_{m,n}\tau/T$ 表示在 N 个时隙内的平均信道利用率；τ 表示一个时隙内单次数据传输的平均时长。条件 C_1 为时隙长度约束，表示竞争接入时段和数据传输时段必须严格限定在时隙框架内；条件 C_2 为竞争节点数量约束，表示当前时隙中参与竞争的节点数由历史竞争失败节点与新到达数据包节点数量共同决定；条件 C_3 至 C_5 表示竞争概率取值约束；条件 C_6 表示竞争概率增长因子取值约束。

为了确保满足上述约束条件，采取如下措施：在系统执行本文算法时，当上述条件不满足时会显著影响到协议性能，算法会在当前步的奖励中加入惩罚项，以惩罚当前的错误信道接入策略（即动作）。在实际仿真中，为了提高程序运行效率，对于约束条 C_1，程序将判断每个时间步中条件 C_1 是否满足，若不满足条件则重置当前步；对于其他约束条件，程序将严格控制各条件中的数量关系及取值范围以满足条件。

2 算法设计

本文以移动自组网 p 坚持 CSMA 协议的信道接入数学模型为理论基础，构建了如式（10）所示的优化问题。由于部分约束条件的非线性和优化变量之间的耦合性导致了问题（10）具有非凸性，使用传统优化算法可能陷入局部最优并且难以直接求解[20]。相比之下，深度强化学习方法通过构建智能化决策框架，通过数据驱动实现动态策略优化，无需显示建模复杂约束关系，具备处理高维状态空间与连续动作决策的能力，并且可以通过从状态到动作的直接映射机制避免解析求解复杂约束。为此，本文采用深度强化学习中的近端策略优化（PPO，proximal policy optimization）构建问题（10）的求解模型，提出基于 PPO 的自适应信道接入算法。该算法通过动态优化节点竞争概率、竞争概率增长因子以及准入节点数量，实现无线信道资源的高效利用。

为了利用深度强化学习方法有效求解问题（10），需要将其建模为马尔可夫决策过程（MDP，Markov decision process），利用 MDP 框架形式化问题的时序决策特性，使智能体能够在动态环境中捕捉状态转移的马尔可夫性。在 MDP 框架下，首先需要定义状态、动作并设计奖励函数。

状态：状态是环境动态特征的即时反映，环境状态构成了节点行为决策的基础。在时隙 n 中，节点 m 的初始状态 $s_{m,n}$ 可表示为：$s_{m,n}=\{R_{m,n-1},K_{m,n-1},D_{m,n-1}\}$，$R_{m,n-1}$ 为前时隙的竞争节点数，$K_{m,n-1}$ 前时隙竞争成功的节点数，$D_{m,n-1}$ 为前时隙中有新到达数据包的节点数，这些参数共同反映了当前的信道争用状况。

动作：在时隙 n，节点 m 通过环境感知获取状态信息 $s_{m,n}$，根据当前策略输出动作分布，采样得到动作 $a_{m,n}=\{K_{m,n},P_{m,n},\upsilon_{m,n}\}$，$a_{m,n}$ 包括准入节点数 $K_{m,n}$、

节点竞争概率 $P_{m,n}$以及竞争概率增长因子 $v_{m,n}$，随后节点执行动作 $a_{m,n}$与环境进行交互。

奖励：在时隙 n，节点 m 通采取动作 $a_{m,n}$与环境进行交互，环境根据动作 $a_{m,n}$的结果生成奖励作为策略效果的评估信号。基于优化问题（10）和状态空间定义，设计即时奖励函数 $r_{m,n}$为：

$$r_{m,n} = K_{m,n}\tau/T + \psi \tag{11}$$

其中：$K_{m,n}\tau/T$ 表示在时隙 n 中与节点 m 进行数据传输的总时长占比，用于衡量当前时隙内信道资源的使用效率；τ 为单次数据传输的平均时长；ψ 为不满足问题（10）约束条件时施加的惩罚因子。

在时隙 n 中，节点 m 首先对网络环境进行感知得到当前状态信息 $s_{m,n}$，在此状态下依据当前策略生成动作的概率分布并采样得到动作 $a_{m,n}$，执行动作并根据环境反馈获得奖励信号 $r_{m,n}$，随后再次感知环境获得下一状态信息 $s_{m,n+1}$，上述过程即为马尔可夫链，完整地表征了 MDP 的链式状态转移特性。每当进行一次上述过程，保存 $\{s_{m,n}, a_{m,n}, r_{m,n}, s_{m,n+1}\}$ 作为一条样本数据；而当上述过程进行到一定次数时，就将这些样本数据作为一条轨迹（Trajectory）进行保存至轨迹集合，后续将收集到的轨迹数据存储至经验缓存中用于训练网络。

PPO 算法采用典型的 Actor－Critic 框架。Actor 网络即策略网络，用于拟合策略函数 $\pi(s)$，其输入为智能体观测到的状态 $s_{m,n}$，输出为动作的概率分布 $\pi_\theta(a_{m,n} \mid s_{m,n})$。PPO 算法结合了重要性采样的思想，构造了新、旧两个 Actor 网络（参数分别为 θ 和 θ_{old}），分别代表新、旧策略以提高样本的利用率[21]。Actor 网络的更新方式是通过梯度上升法更新新 Actor 网络参数 θ 并定期将参数赋给旧 Actor 网络以更新参数 θ_{old}。以时间步为例，Actor 网络的损失函数[22]为：

$$L_{\text{CLIP}}(\theta) = \hat{E}_t\left\{min\left\{\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t, clip\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}, 1-\varepsilon, 1+\varepsilon\right]\hat{A}_t\right\}\right\} \tag{12}$$

其中：$\hat{A}_t$ 表示优势函数，基于广义优势估计（GAE，generalized advantage estimation）计算得到；$clip$ 为裁剪函数，当新旧策略比值小于 $1-\varepsilon$ 或大于 $1+\varepsilon$ 时分别取值为 $1-\varepsilon$ 和 $1+\varepsilon$；ε 为截断超参数。

优势函数 $\hat{A}_t$ 可以表示为：

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{k-t-1}\delta_{k-1} = \sum_{l=t}^{k-1}(\gamma\lambda)^{l-t}\delta_l \tag{13}$$

其中：γ 为折扣系数；λ 为 GAE 参数；$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_1)$。

Critic 网络即价值网络（参数为 φ），用于拟合价值函数 $V(s)$，其输入是状态 $s_{m,n}$，输出为估计的状态值。通过优势函数 $\hat{A}_t$ 评价动作价值。Critic 网络的更新方式是采用均方误差作为损失函数，并通过梯度下降法更新参数。Critic 网络的损失函数为：

$$L^{VF}(\varphi) = [V_\varphi(s_t) - V_{\text{tar}}(s_t)]^2 \tag{14}$$

其中：$V_\varphi(s_t)$ 为 Critic 网络估计的状态价值，$V_{\text{tar}}(s_t)$ 为状态目标值。

本文实现的 PPO 算法中，Actor 网络包括一个输入层、3 个隐藏层和一个输出层。隐藏层的神经元数量分别为 512、256、10，使用 ReLU 激活函数，输出层生成动作的分布并根据该分布对动作采样。Critic 网络与 Actor 网络的结构基本相同，只是其输出层以单个神经元输出价值估计。PPO 算法中的具体超参数设置见下一节的仿真参数表。

基于上述分析，本文提出的基于 PPO 的自适应信道接入算法可总结为算法 1，其具体工作流程如图 4 所示。

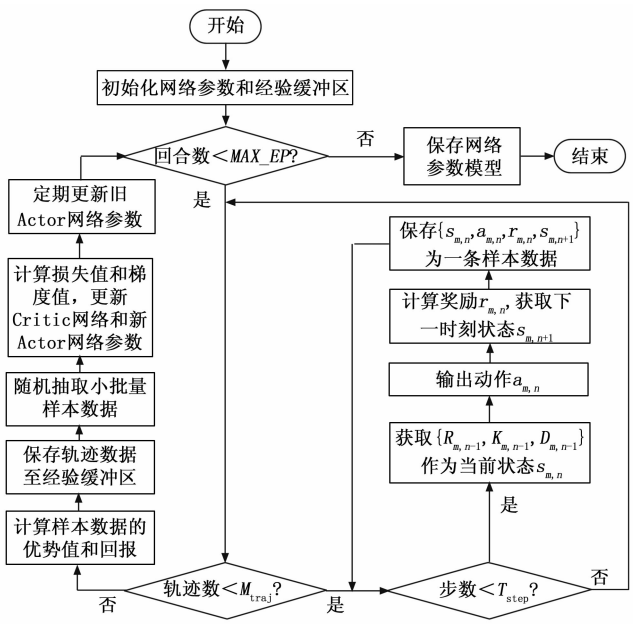

图 4　PPO-ACA 算法的流程

算法 1：

初始化新 Actor 网络 θ、旧 Actor 网络 θ_{old}、Critic 网络 φ 和经验回放缓冲区。

for episode ∈ {1, …, MAX_Ep}

重置状态为 $s_{m,n=0}$

for trajectory ∈ {1, …, M_{traj}}

for timestep ∈ {1, …, T_{step}}

观测环境，获取状态 $s_{m,n}$

Actor 网络选择动作 $a_{m,n}$

执行动作 $a_{m,n}$，根据式（12）计算奖励值 $r_{m,n}$，观察下一状态 $s_{m,n+1}$

将 $\{s_{m,n}, a_{m,n}, r_{m,n}, s_{m,n+1}\}$ 保存为样本

end for

end for

保存所有轨迹数据至经验缓存

划分经验缓冲区数据，并从中抽取小批量的数据进行训练

根据式（14）计算优势函数 $\hat{A}_t$

根据式（13）训练新 Actor 网络，并更新网络参数 θ

根据式（15）训练 Critic 网络，并更新网络参数 φ

定期更新旧 Actor 网络参数 θ_{old}，$\theta_{\text{old}} \leftarrow \theta$

end for

算法 1 的核心为 PPO，在处理形如式（10）的优化问题时相比于传统的优化算法计算复杂度更低。在 PPO 的框架下，假设 Actor 网络和 Critic 网络结构中分别有 L_π 和 L_V 个全连接层，则算法 1 的计算复杂度可以表示为：

$$O\left[(E+1)M_{\text{traj}}T_{\text{step}}\left(\sum_{i=0}^{L_\pi-1}n_i^{\text{actor}}n_{i+1}^{\text{actor}}+\sum_{i=0}^{L_V-1}n_i^{\text{critic}}n_{i+1}^{\text{critic}}\right)\right] \tag{15}$$

其中：E 为 PPO 参数更新阶段时的优化轮数；M_{traj} 为数据收集阶段策略网络生成的轨迹数量；T_{step} 为每个轨迹中包含的步数；n_i^{actor} 和 n_i^{critic} 分别为 Actor 网络和 Critic 网络中第 i 层的神经元数量。当 E 为常数时（通常取值为 3～10），主导项为：

$$O\left[M_{\text{traj}}T_{\text{step}}\left(\sum_{i=0}^{L_\pi-1}n_i^{\text{actor}}n_{i+1}^{\text{actor}}+\sum_{i=0}^{L_V-1}n_i^{\text{critic}}n_{i+1}^{\text{critic}}\right)\right] \tag{16}$$

可以看出，算法 1 的计算复杂度主要取决于 Actor 和 Critic 的网络结构以及样本数量，与自组网的网络规模和时隙划分情况无关。由于这种特性，当网络规模扩大时，在不改变神经网络结构和样本数量情况下，算法 1 仍然可以保持较低的计算复杂度。

由于 PPO 算法是一种改进的策略梯度算法，使用随机梯度上升优化 Actor 网络目标函数，根据 PPO 算法理论[22]以及 Sutton 等人对策略梯度方法的局部收敛性理论[23]可知，PPO 算法生成的策略网络参数序列 $\{\theta_k\}$ 能够收敛至局部最优策略，满足：$\lim_{k\to\infty}\nabla_\theta L_{\text{CLIP}}(\theta_k)=0$。

3 仿真结果与分析

本节对所提出的 PPO-ACA 算法的性能进行仿真验证。首先将相同环境下本文算法在不同学习率设置下的信道利用率曲线与基于 DDPG 的信道接入算法[17]进行了对比；其次，针对时隙长度、数据包到达率以及网络节点数量等关键参数的影响进行研究，分析了上述参数与信道利用率之间的关系，并对比了本文 PPO-ACA 算法与其他 3 种信道接入方案的信道利用率；最后，验证了业务优先级区分机制对于节点的接入成功率的影响。仿真参数见表 1。

表 1 仿真参数

参数名称	参数取值
批量大小	$e=64$
折扣系数	$\gamma=0.95$
裁剪参数	0.2
GAE 参数	$\lambda=0.95$
通信时隙数	$N=10$
通信时延	$\tau=0.03$ s
退避帧间间隔	$T_{\text{BIFS}}=2.5$ μs
短帧间间隔	$T_{\text{SIFS}}=7.5$ μs
请求包发送时长	$T_{\text{req}}=22.2$ μs
确认包发送时长	$T_{\text{ACK}}=7.5$ μs
Actor 网络层数	$L_\pi=3$
Critic 网络层数	$L_V=3$
Actor 网络单层神经元数量	$n_i^{\text{actor}}=10\sim512$
Critic 网络单层神经元数量	$n_i^{\text{critic}}=10\sim512$

图 5 展示了信道利用率随迭代次数的变化曲线对比。可以看出，随着迭代次数的增加，各方案下的信道利用率逐渐增大并最终趋于稳定。本文提出的基于 PPO 的算法相比基于 DDPG 的信道接入算法具有更快的收敛速度和相对更高的信道利用率。此外，当学习率为量级时本文算法能够最快收敛，学习过程中表现出最好的策略稳定性，并且达到了几种方案中最高的信道利用率。

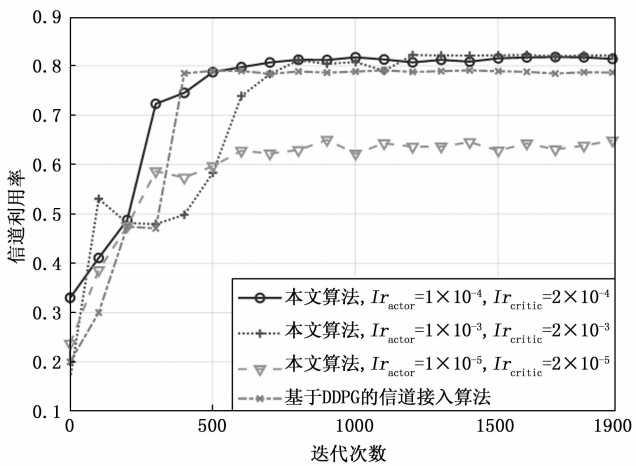

图 5 不同自适应信道接入算法的信道利用率对比

为分析不同时隙长度、数据包到达率以及网络节点数与信道利用率之间的关系，本文选取了固定准入节点数量方案、固定节点竞争概率方案和预设竞争概率方案[24]与本文 PPO-ACA 算法进行对比。在固定准入节点数量方案下，限定单时隙最大接入节点阈值，当实际参与竞争的节点数低于该门限时，系统可实现全节点无冲突接入，相当于为所有节点固定分配通信时隙的 TDMA 机制；在固定节点竞争概率方案下，每个时隙中，各节点的竞争概率为固定值，即传统 p 坚持 CSMA 的

竞争机制；在预设竞争概率方案下，每个时隙中竞争节点从预设概率集合中选取竞争概率。

信道利用率与时隙总长度之间的关系如图 6 所示。从图中容易看出，本文算法和其他 3 种信道接入方案下的信道利用率均随时隙增长而呈现下降趋势。虽然时隙长度增加，但是数据包的到达率并未改变，这些新增的时隙长度中并没有伴随更多的节点通信需求，反而使得信道处于空闲状态的时间增加，进而导致信道利用率下降。本文算法能够基于上一时隙观察到的信道争用情况以及新增的传输需求，自适应调整本时隙内准入节点数量以及节点的竞争概率，从而以更符合当前环境的策略进行信道接入，因此相较于其他 3 种相对固定的信道接入方案获得了更高的信道利用率。在时隙长度的情况下，本文算法的信道利用率分别比固定准入节点数量、固定节点竞争概率以及预设竞争概率方案高 33.3%、48.1%和 18.9%。

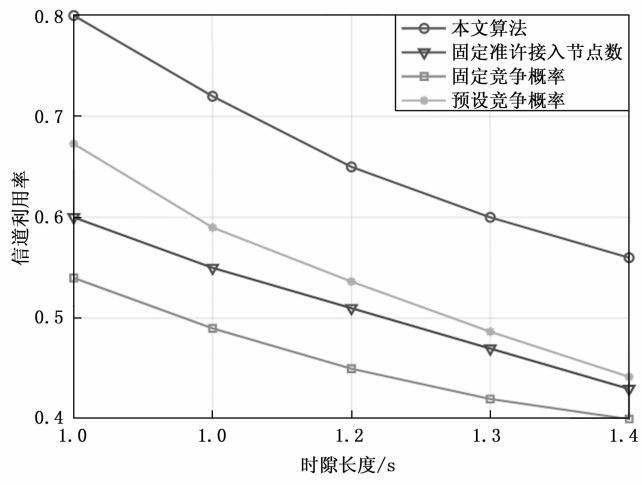

图 6　不同时隙长度的信道利用率对比

图 7 为信道利用率与数据包到达率之间的关系。可以看出，随着数据包到达率增加，4 种方案下的信道利用率均呈现上升趋势，这是因为节点的数据传输需求变得更加频繁，信道中有更多的时间用于通信，因此信道利用率上升；而当数据包到达率高于 3 时，除了固定准入节点数量的信道利用率停止增长之外，其他几种方案的信道利用率仍然保持增加但增速相对下降。这是由于网络中的节点更加频繁地参与竞争，导致信道中发生更多的碰撞从而增加了竞争时间，此时因竞争成功而获得接入机会的节点也逐渐饱和，导致信道利用率增速变缓。本文算法对环境的观测中包含了上一时隙中的数据包到达情况，并且由此对本时隙的节点接入策略进行调整，因此其信道利用率高于其他 3 种信道接入方案。此外，由于固定准入节点数量方案限制了每个时隙中能够接入的节点数量上限，随着数据包到达率的增加，竞争成功的节点数量达到该上限后便不再增加，导致信道利用率也保持不变。

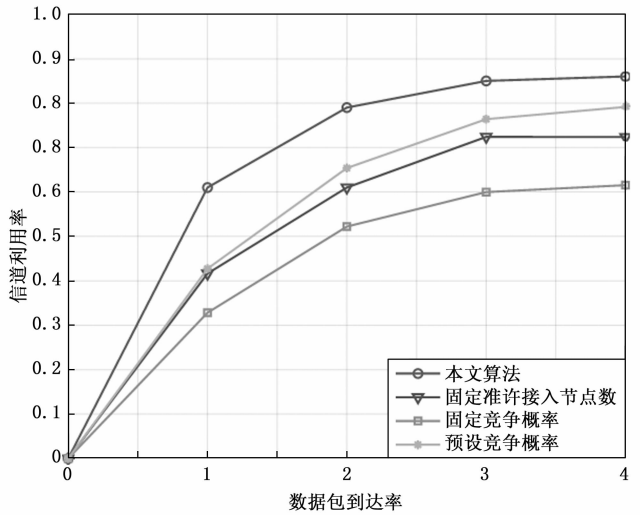

图 7　不同数据包到达率的信道利用率对比

图 8 为信道利用率与网络节点数之间的关系。随着网络节点数增加，所有方案下的信道利用率均有所提高，这是因为此时参与竞争的节点数量在合理范围内逐渐增多，信道中用于传输数据的时间占比上升。但是当网络节点数达到一定规模时，信道中产生的碰撞显著增加，竞争时间占比增大，此时信道利用率反而开始下降。由于本文算法根据网络中实时的竞争情况自适应调整准入节点数量、节点竞争概率以及竞争概率增长因子，能够有效提升 p 坚持 CSMA 的性能，所以在 35 节点以内的网络中始终可以实现最高的信道利用率。在固定准入节点数量方案下，由于当节点数增大到一定规模后便会达到准入节点数上限，并且准入节点均可获得传输数据的机会，因此其信道利用率不会随着节点数增加而下降。综上所述，当网络规模在 35 节点以内时，采用本文算法可有效提升信道利用率；而在更大的网络规模下，应使用固定准入节点数即 TDMA 接入方案以保

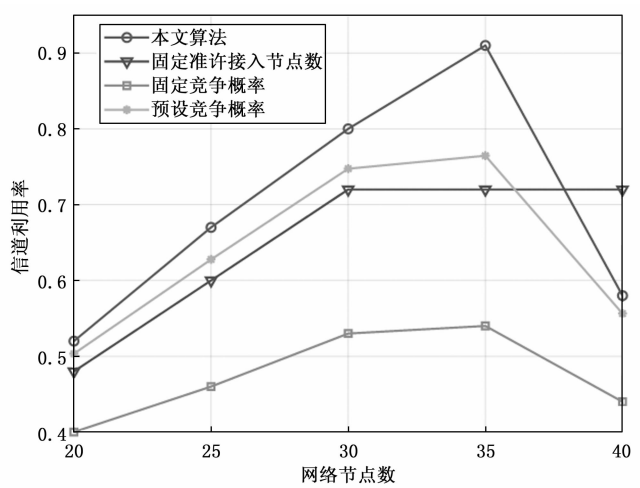

图 8　不同网络节点数的信道利用率对比

持较高的信道利用率。

假设网络中存在高、低两种优先等级的业务，其数据包随机到达网络中的节点。本节分别将启用和不启用优先级区分机制时具有高、低优先级业务的节点接入成功率进行了对比。图 9 展示了不同业务优先级下节点的接入成功率，可以看出随着迭代次数增加，本文算法对节点竞争接入的调控逐渐趋于稳定，此时高、低优先级节点都获得了较高的接入成功率，其中高优先级节点的接入成功率可达到 90%以上。在启用优先级区分机制的情况下，高优先级业务的竞争失败节点往往能够获得更高的竞争概率，而低优先级业务节点的竞争概率相较于基础竞争概率仅些微提升，因此前者的接入成功率最高，而后者的接入成功率在算法收敛时与不启用优先级区分机制的情况下较为接近，这保障了高优先级业务的稳定传输。此外，由于本文算法中的业务优先级区分机制工作在参数重置阶段，与节点的参数重置同时进行，因此无需独立维护额外的控制信息，不会增加协议的复杂性和控制开销。

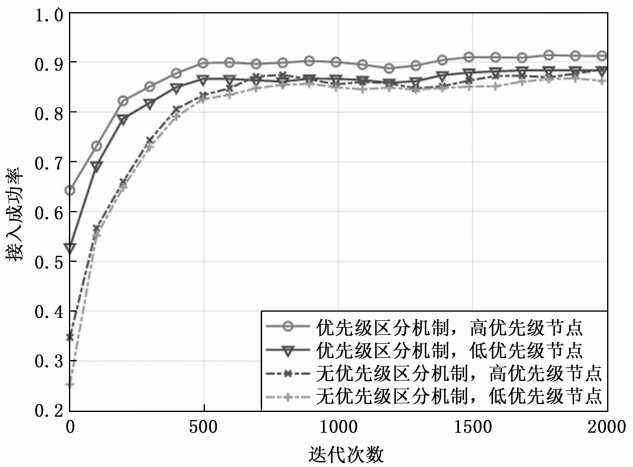

图 9 不同优先级业务的节点接入成功率对比

4 结束语

针对基于 p 坚持 CSMA 协议在移动自组网（MANET）中的信道接入问题，本文提出一种基于 PPO 的自适应信道接入（PPO-ACA）算法。本文以最大化节点的信道利用率为目标建立优化问题，采用深度强化学习中的 PPO 方法来求解该问题，对节点竞争概率、竞争概率增长因子和准入节点数量进行联合优化。本文所提出的 PPO-ACA 算法根据对本地环境的观测做出决策，自适应地调控网络节点的信道接入行为，可以有效应对动态变化的移动自组织网络环境；由于 PPO 实现相对简单，计算效率高，适合分布式部署，网络节点可从路由表中收集邻居节点的相应信息，独立运行 PPO-ACA 算法，从而满足了移动自组网中动态、分布式、自适应的组网需求；此外，本文算法还对不同业务优先等级进行了区分，提升了高优先级业务节点的接入成功率。仿真结果表明，本文算法相比基于 DDPG 的信道接入算法具有更快的收敛速度，相比固定准入节点数量、固定节点竞争概率和预设竞争概率的信道接入方案可以更加有效地利用信道资源，并且可以保证具有高优先级业务的节点接入成功率更高。由于忽略了网络节点的具体物理特性，因此在下一步工作中可以考虑节点的天线类型、工作频段以及能耗等因素对系统整体性能的影响以满足未来移动自组网中节点多样化的服务需求。

参考文献：

[1] 董 超，陶 婷，冯斯梦，等. 面向无人机自组网和车联网的媒体接入控制协议研究综述 [J]. 电子与信息学报，2022，44 (3)：790－802.

[2] 米 阳. 无人机自组网低时延 MAC 协议研究与设计 [D]. 西安：西安电子科技大学，2020.

[3] 卢玫欣，苏胜君，施伟斌，等. Z-MAC 协议改进算法 [J]. 软件导刊，2017，16 (12)：78－80.

[4] 闫 涛，赵一帆，高明虎，等. 移动自组网中的自适应 MAC 协议研究综述 [J]. 计算机工程与应用，2023，59 (11)：46－56.

[5] 赵梦华. 大规模移动自组织网络媒体接入控制协议关键技术研究 [D]. 北京：北京交通大学，2020.

[6] XIE Z，XU Y. Research on OTA optimization of wireless sensor networks based on CSMA/CA improved algorithm [C] // IEEE，Chengdu，2018：331－335.

[7] BRUNO R，CONTI M，GREGORI E. Optimal capacity of p-persistent CSMA protocols [J]. Communications Letters IEEE，2003，7 (3)：139－141.

[8] SEO J B，JIN H. Stability region of p-persistent CSMA systems [J]. IEEE Communications Letters，2017，21 (3)：652－655.

[9] KIM N，NA W，LAKEW D S，et al. DQN-based directional mac protocol in wireless Ad Hoc network in internet of things [J]. IEEE Internet of Things Journal，2024，11 (7)：12918－12928.

[10] WI G，SON S，PARK K J. Delay-aware TDMA scheduling with deep reinforcement learning in tactical MANET [C] //IEEE，Jeju，2020：370－372.

[11] AL-TAM F，MAZAYEV A，CORREIA N，et al. Deep PC-MAC：a deep reinforcement learning pointer-critic media access protocol [C] //IEEE，Pisa，2020：1－6.

[12] 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.

（下转第 318 页）