基于多智能体深度强化学习的多域协同抗干扰方法研究

张彪[1]，汪西明[2]，徐逸凡[1]，李文[1]，韩昊[1]，刘松仪[1]，陈学强[1]
（1. 陆军工程大学通信工程学院，江苏 南京 210007；2. 国防科技大学信息通信学院，湖北 武汉 430010）

摘 要：动态的传输需求和有限的缓存空间给恶意干扰环境下的无线数据传输带来巨大挑战。针对上述问题，从频域和时域的角度出发，研究了面向分布式物联网的协同抗干扰信道选择和数据调度联合决策方法，构建了基于多用户马尔可夫决策过程的数据传输模型，提出了基于多智能体深度强化学习的协同抗干扰信道和数据联合决策算法。仿真表明，所提算法可有效避开恶意干扰并避免同频互扰。相较于对比算法，网络吞吐量显著提高，丢包数量明显降低。

关键词：协同抗干扰；信道选择；数据调度；多智能体强化学习；深度学习
中图分类号：TN973.3；TP181
文献标志码：A
doi: 10.11959/j.issn.2096−3750.2022.00293

Multi-domain collaborative anti-jamming based on multi-agent deep reinforcement learning

ZHANG Biao[1], WANG Ximing[2], XU Yifan[1], LI Wen[1], HAN Hao[1], LIU Songyi[1], CHEN Xueqiang[1]
1. College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China
2. College of Information and Communication, National University of Defense Technology, Wuhan 430010, China

Abstract: Dynamic transmission requirements and the limited cache space bring great challenges to wireless data transmission in the malicious jamming environment. Aiming at the above problems, a collaborative anti-jamming channel selection and data scheduling joint decision method for distributed internet of things was studied from the perspective of frequency domain and time domain. A data transmission model based on multi-user Markov decision process was constructed and a collaborativeanti-jamming joint-channel-and-data decision algorithm based on multi-agent deep reinforcement learning was proposed. Simulation results show that the proposed algorithm can effectively avoid the malicious jamming and the co-channel interference. Compared with the comparison algorithm, the network throughput is significantly improved, and the number of packet dropout is significantly reduced.

Key words: collaborative anti-jamming, channel selection, data scheduling, multi-agent reinforcement learning, deep learning

0 引言

随着 5G 网络的普及和对未来 6G 技术的广泛研究，物联网等新型无线网络让人们的生活越来越便利且智能[1-3]。然而，随着“万物互联”的物联网规模不断扩大、传输需求越发复杂多样，频谱资源难以协调的问题日渐凸显。此外，恶意干扰的攻击将导致大量频谱不可使用，可以使物联网瘫痪，智能通信抗干扰逐渐成为亟须研究的热点问题[4-5]。

随着人工智能技术的迅速发展，将机器学习和抗干扰相结合的智能抗干扰方法得到了广泛研究[6-18]。然而，针对多用户场景的抗干扰方法研究仍然存在

收稿日期：2022−05−05；**修回日期：**2022−08−22
通信作者：汪西明，ximingw@nudt.edu.cn
基金项目：国家自然科学基金资助项目（No.62071488，No.61961010）
Foundation Items: The National Natural Science Foundation of China (No.62071488, No.61961010)

以下几个问题。

1) 现有研究假设用户一直有数据传输，并且按照固定长度的数据量在每个时隙进行传输，忽略了实际场景中用户有限的缓存空间和动态的传输需求。

2) 无线通信网络用户间存在复杂的耦合关系，每个用户的用频决策会对整个电磁频谱环境和其他用户的用频决策造成影响，此外干扰环境下难以维持可靠控制信道，协调用户之间的用频难度较大。

3) 用户有动态的传输需求但数据缓存空间有限，选择传输信道之后还需要根据缓冲区状态选择合适的数据量进行发送，传输数据太少会造成数据堆积，传输数据过多会增加被干扰的风险，决策空间和决策难度大。

综上考虑，本文研究恶意干扰环境下分布式物联网频域与时域联合的多用户协同抗干扰方法。频率域进行信道选择，时间域进行数据调度决策，将动态干扰环境下的多用户信道选择和数据调度联合决策问题构建为多用户马尔可夫决策过程。考虑恶意干扰环境下用户之间信息交互不可靠的实际情况，提出基于多智能体深度强化学习的协同抗干扰信道选择和数据调度联合决策算法，用户独立进行频谱感知、动作决策、网络训练，分布式地试探频谱环境得到有效的抗干扰数据传输策略。

本文的主要贡献如下。

1) 针对干扰环境下分布式物联网频谱资源难以协调的问题，从频域和时域联合的角度出发，引入多用户马尔可夫决策过程建模多用户信道选择和数据调度联合决策问题。

2) 提出基于多智能体深度强化学习的协同抗干扰信道选择和数据调度联合决策算法，改进深度神经网络的结构解决状态空间巨大问题。仿真结果表明，在恶意干扰环境下无须信息交互，该算法能有效避开恶意干扰和同频互扰，显著提高了网络吞吐量。

注意到，文献[19]研究了单发射机在扫频干扰环境下动态传输时间和数据调度问题，提出了基于 *Q*-learning 的抗干扰算法。然而，本文工作与该文献存在较大区别，主要概括如下。

1) 无线通信网络中往往同时有多个设备具有用频需求，频谱环境更为动态复杂，文献[19]所使用的 *Q*-learning 算法难以具体表征环境状态，使用深度神经网络能挖掘更多频谱环境中的特征信息。

2) 该文献研究单用户强化学习抗干扰算法，直接应用到多用户场景中无法解决同频互扰问题，收敛性难以保证。因此，本文所研究的基于多智能体深度强化学习的多用户数据传输方法更有实际意义。

1　相关工作

随着人工智能的迅速发展，机器学习赋能的通信抗干扰技术成为研究热点。基于强化学习的抗干扰方法可实现在线的环境感知、干扰预测和频谱接入，已被广泛应用于智能通信抗干扰领域[8-12]。例如，文献[8-9]将干扰视为频谱环境的一部分，通过 *Q*-learning 算法在线学习干扰规律，获得接近最优的抗干扰信道接入策略。但是，一旦干扰方动态调整干扰模式，频谱环境变得十分复杂，传统 *Q*-learning 算法在复杂频谱环境中难以收敛。针对该问题，文献[12]将深度强化学习方法引入到智能通信抗干扰领域，通过深度神经网络提取复杂频谱环境的状态信息，有效解决了 *Q*-learning 算法无法收敛问题。无线通信网络用户规模不断扩大，往往同时拥有多个用频设备，但上述文献研究的是单用户抗干扰方法，应用在多用户通信网络中很难解决用户间的同频互扰问题，研究多用户抗干扰方法更具实际意义。

多用户协同抗干扰方法被广泛研究[15-18,20]。文献[16]提出基于协同 *Q*-learning 的多用户通信抗干扰方法，用户之间通过交互 *Q* 表可获得有效的协同抗干扰策略。但是该方法仅适用于小规模用户，随着用户数量增加，该方法需要交互的学习信息和算法复杂度呈指数倍增长。文献[17]研究了超密集网络中多子网协同抗干扰频谱接入问题，为减小算法的复杂度，设计了一种平均场与强化学习结合的协同抗干扰算法，实现整个网络的协同抗干扰频谱接入。虽然文献[16-17]所提方法能够协同解决多用户信道选择问题，但是他们均需要可靠的控制信道来交互信息。由于电磁频谱的开放性，可靠控制信道难以获得，恶意干扰的存在更加剧这个问题，研究无须控制信道的分布式协同抗干扰方法十分有意义。

文献[18]针对多智能体协同学习算法依赖可靠控制信道进行用户间信息交互的问题，提出了一种分布式多智能体强化学习抗干扰算法，即交叉检测 *Q*-learning，仿真两个用户情况下取得了和拥有可靠控制信道的协同学习算法一样的通信效果。文献[20]研究了动态复杂干扰环境下自组织网络频谱资源协调问

题，提出了基于分布式学习的抗干扰算法，无须信息交互，每个智能体通过对环境的试探积累经验，能够快速找到多智能体协同抗干扰策略。但是，上述研究均假设用户一直有数据包需要发送，并且按照固定数量来发送数据，未考虑动态的传输需求和有限的缓存区大小。除上述研究外，文献[19]研究了单用户动态传输时间和数据调度问题，但无线通信网络用户数量庞大，单用户方法难以解决用户之间的协同问题。

综上所述，本文提出一种基于多智能体深度强化学习的协同抗干扰算法来解决恶意干扰环境下多用户信道选择和数据调度联合决策问题，所提方法无须信息交互，可快速收敛并显著提高了网络吞吐量。

2　系统模型与问题建模

2.1　系统模型

数据汇聚物联网系统模型如图 1 所示，该网络中包含 N 个发射机、一个数据接收中心和一个干扰机。发射机的集合为 $\mathcal{N}=\{0,1,\cdots,N\}$，每个发射机拥有感知设备对频谱环境进行实时感知；每个发射机有一个最大存储容量为 L 的缓冲区，数据包在到达后首先被存储到缓冲区中，然后再被发送。在每个时隙的开始，所有的发射机独立决策传输信道和传输数据包的数量。数据接收中心负责汇聚接收发射机传输的数据并反馈数据接收状态。干扰机的干扰模式为具有动态频谱特性的动态梳状干扰，通过发射高功率干扰信号意图破坏发射机到数据接收中心的通信链路。

2.1.1　时隙结构

通信时隙结构如图 2 所示，发射机以时隙为单位进行业务数据的传输。每个时隙可分为 5 个阶段：频谱瀑布图构建、动作决策、数据传输、确认字符（ACK, acknowledge character）传输和学习[21]。

- 频谱瀑布图构建阶段：发射机根据感知的频谱环境构建频谱瀑布图，用时记为 T_{S}。
- 动作决策阶段：发射机根据频谱环境状态和缓冲区数据包的数量来决策传输信道和本时隙传输数据包的数量，用时记为 T_{A}。
- 数据传输阶段：发射机向数据接收中心传输数据，用时记为 T_{D}。
- ACK 传输阶段：数据接收中心发送 ACK 信息反馈接收状态给发射机，用时记为 T_{ACK}。
- 学习阶段：发射机根据 ACK 信息反馈的接收状态学习干扰机和其他发射机的用频规律用时记为 T_{L}。

假设发射机时隙同步，发射机在每个时隙的开始根据频谱感知的结果构建频谱瀑布图，然后独立做出信道选择和数据调度决策。在数据传输阶段，发射机根据所做决策向数据接收中心传输数据，由于传输数据量的不同导致每个发射机的传输时长不一样，发射机数据传输完毕即可提前进入 ACK 接收阶段。数据传输阶段结束后，数据接收中心向发射机发送 ACK 信息来反馈数据接收情况。ACK 信号包含信息较少，采用具有较大扩频因子的扩频通信（SSC, spread spectrum communication）技术进

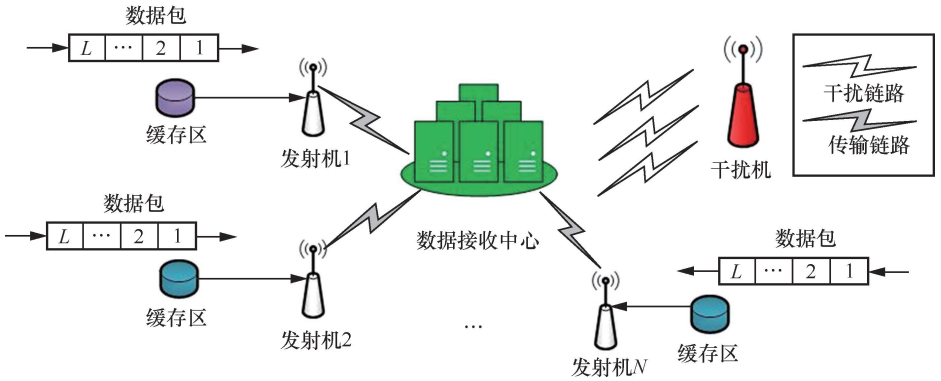


图 1　数据汇聚物联网系统模型

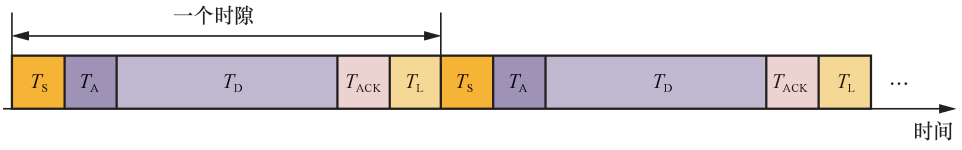


图 2　通信时隙结构

行传输，确保 ACK 信息完美传输，但是 SSC 传输速率较低，无法用于数据传输[22]。发射机接收到 ACK 信息后，根据所反馈的传输情况来进行算法的学习和更新。

2.1.2 业务模型

假设业务数据到达速率服从均值为 λ 的泊松分布，在第 t 时隙内发射机 n 收到 d_t^n 个业务数据包的概率为

$$P(d_t^n)=\frac{\lambda^{d_t^n}}{d_t^n!}\mathrm{e}^{-\lambda} \tag{1}$$

缓冲区最大存储数据包的数量为 L，假设 t 时隙传输开始时发射机 n 缓冲区内数据量为 l_t^n，发送的数据包数量为 l_s^n，接收到的数据包数量为 l_r^n。由于缓冲区存储空间有限，数据包被取出发送之后即删除，不考虑传输失败之后进行重传，所以，在第 t 时隙传输结束后，发射机 n 缓冲区内存储数据量为

$$l_{t+1}^n=\min(L,l_t^n+l_r^n-l_s^n) \tag{2}$$

当缓冲区剩余存储空间不足以存储新到达的数据包时，会造成缓冲区数据包溢出丢包。

2.1.3 传输模型

整个数据汇聚物联网共享一段频谱，该频谱被不重复地均分为 M 个带宽为 b 的传输信道，信道集合为 $\mathcal{M}=\{1,2,\cdots,M\}$，信道 $k\in\mathcal{M}$ 的频率范围为 $[f_k-b/2,f_k+b/2]$，其中 f_k 为信道 k 的中心频率。$U(f)$ 为发射机传输数据时的功率谱密度（PSD, power spectral density）方程，发射机的传输功率为

$$p=\int_{-b/2}^{b/2}U(f)\mathrm{d}f \tag{3}$$

发射机在恶意干扰环境下进行数据传输，既要协调信道选择避免和其他发射机造成冲突，还要决策传输数据包的数量防止恶意干扰。发射机 n 在信道 k 进行数据传输时的信干噪比（SINR, signal-to-interference-plus-noise ratio）为

$$\beta_{n,k}=\frac{p_n\partial_{n,k}}{I_{n,k}+J_{n,k}+\delta} \tag{4}$$

其中，p_n 为发射机 n 的传输功率，$\partial_{n,k}$ 为发射机 n 在信道 k 传输数据时的信道系数。

发射机 n 选择在信道 k 传输数据时受到其他发射机同频干扰总功率的值为

$$I_{n,k}=\int_{f_k-b/2}^{f_k+b/2}[\sum_{m\in N_n}\partial_{m,n}U_m(f-f_m)]\mathrm{d}f \tag{5}$$

其中，$U_m(f)$ 为发射机 m 的 PSD 方程，$\partial_{m,n}$ 为发射机 m 在发射机 n 所选传输信道上的信道系数，N_n 为除发射机 n 以外的其他发射机集合 $N_n=N\setminus\{n\}$，f_m 为发射机 m 选择传输信道的中心频率。

发射机 n 传输数据时受到恶意干扰的功率为

$$J_{n,k}=\int_{f_k-b/2}^{f_k+b/2}[\partial_{j,n}U_j(f-f_j)]\mathrm{d}f \tag{6}$$

其中，$\partial_{j,n}$ 为干扰信号在发射机 n 所选传输信道上的信道系数，$U_j(f)$ 为干扰机的 PSD 方程。

加性高斯白噪声的功率为

$$\delta=\int_{f_k-b/2}^{f_k+b/2}n(f)\mathrm{d}f \tag{7}$$

其中，$n(f)$ 为高斯白噪声的 PSD 方程。

此时发射机 n 在信道 k 上的数据传输速率 $V_{n,k}$ 表示为

$$V_{n,k}=b\cdot\mathrm{lb}(1+\beta_{n,k}) \tag{8}$$

当数据传输速率 $V_{n,k}\geqslant V_{\mathrm{th}}$ 时表示数据能够传输成功，其中，V_{th} 为数据传输时所需的最低速率门限，对应通信服务质量门限 β_{th}。根据数据传输阶段每个数据包传输时的 $V_{n,k}$ 值决定数据包能否发送成功，得到发射机 n 在时隙 t 成功传输的数据包数量 J_t^n。

$$J_t^n=h_t^n-h_{\mathrm{fail}}^n \tag{9}$$

其中，h_t^n 表示发射机 n 在时隙 t 发送数据包数量，h_{fail}^n 表示发送数据包因干扰和同频互扰而传输失败的数据包数量。

为了更好地认知频谱环境，考虑所有信号同时存在，发射机 n 的感知结果为

$$s_n(f)=\sum_{m\in N_n}\partial_{m,n}U_m(f-f_m)+\partial_{j,n}U_j(f-f_j)+n(f) \tag{10}$$

离散频谱采样值定义为

$$o_i=10\lg[\int_{i\Delta f}^{(i+1)\Delta f}s_n(f)\mathrm{d}f] \tag{11}$$

其中，Δf 为频谱分析的分辨率，一次频谱感知得到的结果为 $\boldsymbol{o}=[o_1,o_2,\cdots,o_X]^{\mathrm{T}}$，其中 X 为采样点个数。

2.2 问题建模

每个时隙所传输的数据量会对下一个时隙的缓冲区数据状态造成影响，所以该问题属于连续决策问题。因此，本文考虑连续决策优化问题，将长期累积回报值作为优化目标，对于发射机 $n(n=1,2,\cdots,N)$ 而言优化目标为 ψ^n。

$$\psi^n=\max\sum_{t=1}^{+\infty}J_t^n \tag{12}$$

其中，J_t^n 为发射机 n 在 t 时隙成功传输的数据包的数量。即发射机 n 的优化目标是最大化长期累积成功传输数据包的数量，优化目标值受数据包到达速率、数据调度决策、其他发射机的信道选择和干扰机的干扰模式等因素共同影响。

3 多域协同抗干扰方法

对于每个发射机而言，找到最优的数据传输策略十分困难，原因包括发射机之间存在强耦合关系，每一个发射机的决策会对整个频谱环境的状态造成影响；发射机之间无法进行可靠的信息交互，难以通过信息交互来协调用频；选择合适的信道之后还需要根据缓冲区内所存储数据包的数量决定发送数，传输数据包数量太少会造成频谱资源的浪费和缓冲区数据堆积溢出，传输数据包数量过多会增加被干扰的风险。

多用户马尔可夫决策过程擅长建模和分析动态环境中多用户之间的决策交互关系。因此，本文采用多用户马尔可夫决策过程建模动态干扰环境下，数据汇聚物联网的信道选择和数据调度联合决策过程，并提出基于多智能体深度强化学习的协同抗干扰信道选择和数据调度联合决策算法使多发射机自主学习用频和数据传输的有效协同策略。

3.1 多用户马尔可夫决策过程

定义： N 用户马尔可夫决策过程[23]可表示为 $<\mathcal{N},S,A_1,\cdots,A_N,R_1,\cdots,R_N,\Pr>$。其中 $\mathcal{N}=\{1,2,\cdots,N\}$ 为用户的集合，S 为状态空间，A_n 用户 n 的动作空间，$R_n:S\times A_1\times\cdots\times A_n\to\mathbb{R}$ 为用户 n 的奖励函数，$\Pr:S\times A_1\times\cdots\times A_n\to\mu(S)$ 为状态转移概率函数，其中 $\mu(S)$ 是状态空间 S 的概率分布集合。

将协同抗干扰问题建模成多用户马尔可夫决策过程，设计如下。

3.1.1 状态空间

状态设计为频谱瀑布图和缓冲区状态的组合。发射机需要挖掘频谱环境的规律来指导动作的选择，发射机通过感知得到频谱状态构建频谱瀑布图，频谱感知得到的信息包含了所有用户的频谱占用状况以及干扰机的干扰信道。假设每个发射机都能够感知到完整的频谱，为了充分发掘用频规律，采用时间拓展模型[12]，即频谱瀑布图为一段历史时间内感知到的频谱状态 $\boldsymbol{O}_t=[\boldsymbol{o}_t,\boldsymbol{o}_{t-1},\cdots,\boldsymbol{o}_{t-\boldsymbol{\Phi}+1}]^{\mathrm{T}}$，其中 $\boldsymbol{\Phi}$ 表示历史时长，频谱瀑布图示意图如图 3 所示。发射机的缓冲区状态为各自缓冲区的数据量，即发射机 n 在时隙 t 的缓冲区状态为 l_t^n。在时隙 t，发射机 n 的状态为 $S_t^n=(\boldsymbol{O}_t,l_t^n)$。

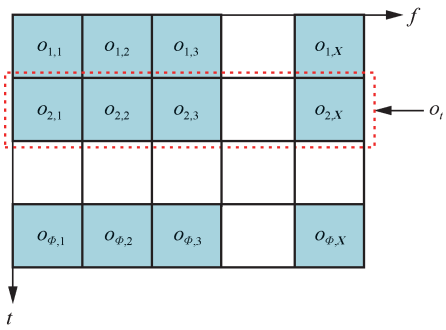


图 3 频谱瀑布图示意图

3.1.2 动作空间

将发射机的动作设计为选择的信道以及在该信道上传输数据包的数量。发射机 n 在 t 时隙选择的信道为 $c_t^n\in\{1,2,\cdots,M\}$，选择传输数据包的数量为 $h_t^n\in\{0,1,2,\cdots,H_{\max}\}$，$H_{\max}$ 为发射机一个时隙内最多能够传输的数据包数量。发射机 n 的动作表示为 $a_t^n=(c_t^n,h_t^n)$，发射机的动作空间大小为 $(M\times H_{\max}+1)$。

3.1.3 奖励函数

奖励函数与成功传输数据量、传输失败数据量和缓冲区剩余数据量有关，当数据传输速率大于数据传输所需的速率门限值 V_{th} 时表示传输成功。发射机 n 在 t 时隙执行动作 a_t^n，传输 h_t^n 个数据包，成功传输的数据包数表示为 h_{succ}^n，缓冲区存储空间有限，缓冲区剩余数据包越多，下一个时隙数据到达时因数据溢出而造成丢包的概率越大。定义成功传输的奖励因子 ε，失败传输的惩罚因子 ϑ，缓冲区折扣因子 η。

发射机执行动作获得的奖励定义为

$$r_t^n=\begin{cases}-0.5 & ，h_t^n>l_t^n\\ \varepsilon\times h_{\text{succ}}^n+\eta\times l_{t+1}^n，& h_t^n=l_t^n\\ \vartheta\times h_t^n & ，其他\end{cases}\tag{13}$$

决策的动作错误时，即决策传输数据包数量比缓冲区内存储数据包的数量多时，发射机不传输数据，并反馈一个 –0.5 的回报值；本时隙传输的数据包全部成功传输给与正奖励 $\varepsilon\times h_{\text{succ}}^n$，并加上缓冲区折扣回报 $\eta\times l_{t+1}^n$；发射机传输的数据包存在传输失败时，给予失败传输的惩罚 $\vartheta\times h_t^n$。

3.2　信道选择和数据调度联合决策算法

数据汇聚物联网中干扰机的干扰模式动态未知，发射机之间无法进行可靠的信息交互，所以无法获取准确的状态转移概率函数。因此，本文采用无模型强化学习方法[24]解决信道和数据调度联合决策问题。由于状态和动作空间维度都很大，建立 Q 表来存储状态价值函数很难维护，采用深度 Q-learning 来处理频谱空间巨大问题[25]。集中式深度强化学习算法需要可靠控制信道进行信息交互的条件无法满足，因此本文设计的信道和数据联合决策的学习算法基于分布式多智能体深度强化学习。

3.2.1　分布式多智能体深度强化学习算法

每个发射机拥有一个结构相同的卷积神经网络，通过训练卷积神经网络对高维连续状态下的状态动作值函数进行拟合。发射机 $n(n=1,2,\cdots,N)$ 在时隙 t 的动作值函数表示为 $Q_t^n(S_t^n,a^n)$，表示发射机在状态 S_t^n 下执行动作 a 能获得的最大长期累积奖励值。采用 Bellman 方程进行对 Q 值进行更新。

$$\begin{aligned}&Q_t^n(S_t^n,a^n)\leftarrow Q_t^n(S_t^n,a^n)+\\&\alpha[r_t+\gamma\max_a Q_{t+1}^n(S_{t+1}^n,a^n)-Q_t^n(S_t^n,a^n)]\end{aligned}\tag{14}$$

其中，α 是学习率，γ 是折扣因子。

定义发射机 n 的深度 Q 网络的损失函数为

$$L(\theta_t^n)=[\text{Target}\,Q-Q(S_t^n,a^n,\theta_t^n)]^2\tag{15}$$

其中，θ_t^n 表示发射机 n 的预测卷积网络的权重参数，目标 Q 值（Target Q ）为

$$\text{Target}\,Q=r+\gamma\max_{a^n}Q_t^n(S_{t+1}^n,a^n;\bar{\theta}_t^n)\tag{16}$$

其中，$\bar{\theta}_t^n$ 是目标卷积神经网络模型的权重参数。

采用梯度下降法训练卷积神经网络，损失函数的梯度为

$$\nabla_{\theta_t^n}L(\theta_t^n)=[\text{Target}\,Q-Q(S_t^n,a_t^n,\theta_t^n)]\nabla_{\theta_t^n}Q(S_t^n,a_t^n,\theta_t^n)\tag{17}$$

传统深度强化学习算法常采用 $\varepsilon-$贪婪策略，使用户能够充分地探索外界环境，分布式深度强化学习算法每个用户作为一个独立的智能体，采用 $\varepsilon-$贪婪策略会使环境动态不稳定。因此，本文所提信道和数据联合决策的学习算法,前 K 时隙采用随机策略选择为经验池填充数据，之后决策选择纯贪婪策略。K 值的选取与算法目标网络更新速率等有关，本文 K 值选取为 100。

3.2.2　神经网络结构

本文采用的卷积神经网络结构如图 4 所示。频谱瀑布图和缓冲区存储数据量作为卷积神经网络的输入，每个发射机拥有相同的频谱瀑布图和差异的缓冲区状态。缓冲区状态被向量化处理输入到全连接层 1，全连接层 1 的神经元个数为

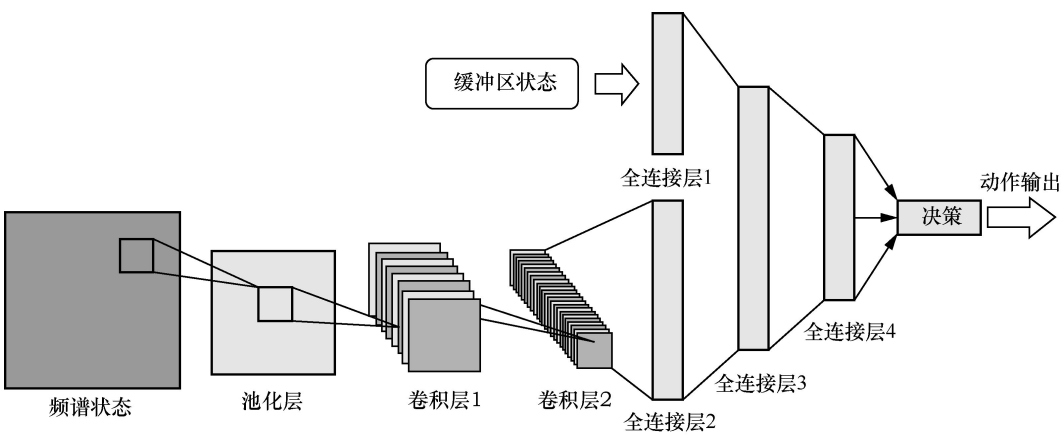


图 4　卷积神经网络结构

L，缓冲区数据量化为全连接层 1 从第一个神经元起数值为 1 的神经元的个数。使用最大池化层处理频谱瀑布图中的冗余信息，过滤器的维度为 2×2，步长为 2。接着使用两层卷积层进行频谱状态特征的提取，第一层卷积层使用 32 个大小为 4×4 的卷积核，步长为 4；第二层卷积层使用 32 个大小为 3×3 的卷积核，步长为 2。卷积层处理后，提取的特征信息通过全连接层 2，然后与全连接层 1 进行连接并输入到神经元个数为 256 的全连接层 3，最后连接全连接层 4。全连接层 4 神经元的个数与动作空间大小相同，神经元输出的值即对应估计动作的 Q 值。

3.2.3 算法复杂度分析

根据文献[26]，本文所提算法所采用的神经网络一次前向传播的计算复杂度 Time 为

$$\text{Time}=O\left(\sum_{l} m_l\cdot f_l^2\cdot c_{l-1}\cdot c_l+\sum_{L}2I_L\cdot E_L-1\right) \tag{18}$$

其中，前一项表示卷积层的计算复杂度，l 为卷积层的数量，m 为每个卷积核输入特征图的面积，f 为卷积核的边长，c_{l-1} 和 c_l 分别为卷积层 l 的输入特征图通道数和输入特征图通道数；式(18)中后一项表示全连接层的计算复杂度，L 为全连接层的数量，I 为全连接层的输入神经元数量，E 为全连接层的输出神经元数量。

基于多智能体深度强化学习的协同抗干扰信道选择和数据调度联合决策算法见算法 1，根据算法流程和网络结构，下面将分析本文所提算法的复杂度。在一次迭代中包含一次策略计算和一次网络训练。在策略计算时，本文采用的是纯贪婪策略，即线性搜索当前状态下的最大 Q 值，该过程的复杂度相较于神经网络的计算复杂度可以忽略不计。在网络训练过程时，用户从记忆重放单元中采样 B 个样本用于更新网络。计算损失值 $L(\theta)$ 时包含 $2B$ 次前向传播，复杂度为 $O(2B\cdot\text{Time})$。最后更新网络时进行一次后向传播，后向传播的复杂度可近似看作和前向传播的复杂度相同，因此网络训练过程每一次迭代的计算复杂度为 $O((2B+2)\cdot\text{Time})$。当网络训练好后，测试阶段迭代一次的计算复杂度为 $O(\text{Time})$。根据仿真环境的参数设置，计算得到训练阶段一次迭代过程的计算量为 2.34×10^{10} FLOPS（floating point operations per second），测试阶段一次迭代的计算量为 1.8×10^{8} FLOPS。现在 GPU 设备的算力能够达到 10^{12} FLOPS 以上，很容易满足网络训练所需的算力。

算法 1 基于多智能体深度强化学习的协同抗干扰信道选择和数据调度联合决策算法

初始化：生成卷积神经网络 Q^n，网络权重 θ^n 随机赋值，初始状态为 S_0^n，$n=0,1,\cdots,N$。

for $t=1,2,\cdots,T$

  if $t\leqslant K$

    发射机 $n(n=1,2,\cdots,N)$ 根据当前状态 S_t^n 随机选择一个动作 a_t^n；

  else

    基于贪婪策略选择动作 a_t^n。

  执行动作 a_t^n 获得奖励值 r_t^n，状态转移到下一个状态 s_{t+1}^n。

  将经验 $(S_t^n,a_t^n,r_t^n,s_{t+1}^n)$ 存入发射机 n 的经验池 E_n 中。

  for $n=1,2,\cdots,N$

    从经验池 E_n 中随机批量采样 $(S_t^n,a_t^n,r_t^n,\ s_{t+1}^n)$；

    根据式（15）更新 Target $Q=r+\gamma\max\limits_{a^n}Q_t^n(S_{t+1}^n,a_{t+1}^n;\overline{\theta}_t^n)$；

    计算损失函数 $\nabla_{\theta_t^n}L(\theta_t^n)$ 并更新 θ_t^n；

  end

end

4 仿真结果与分析

4.1 仿真参数设置

4.1.1 通信参数

通信参数见表 1。考虑数据汇聚网络中发射机数量 $N=5$，共享 20 MHz 带宽的频段，该频段被均分成 10 个不重复的传输信道。发射机以 100 kHz 的分辨率每毫秒对频谱进行一次全波段感知，产生 200 个频谱采样点。历史时长 $\Phi=200$ ms，即发射机保存 200 ms 内的频谱信息构建频谱瀑布图，频谱瀑布图的尺寸为 200 px × 200 px。数据传输采用升降系数为 0.5 的升余弦波，发射机的发送功率为 0 dBm，数据传输时所需的最低速率门限 $V_{\text{th}}=7$ Mbit/s，即通信服务质量门限 $\beta_{\text{th}}=10$ dB，每个传输时隙的长度为 15 ms。

表 1　　通信参数

名称	参数值
用户数量	$N=5$
信道数量	$M=10$
缓冲区长度	$L=10$ 个
历史时长	$\Phi=200$ ms
数据包到达速率	$\lambda=[1,2,3,4,5,6,7]$ 个/时隙
单时隙最大传输数据量	$H_{\max}=5$ 个
发送功率	0 dBm
最低速率门限	$V_{\text{th}}=7$ Mbit/s
通信服务质量门限	$\beta_{\text{th}}=10$ dB
成功传输奖励因子	$\varepsilon=0.5$
失败传输惩罚因子	$\vartheta=-0.1$ 、$\eta=-0.2$

4.1.2　干扰参数

干扰参数见表 2。干扰信号采用升降系数为 0.5 的升余弦波，干扰机的发射功率为 30 dBm，干扰模式为动态切换的梳状干扰，每个时刻拥有两个固定频率的干扰信号，干扰信号的带宽为 2 MHz，即同时干扰 4 个信道，干扰信号切换速度为 15 ms。背景噪声水平为 –90 dBm / Hz 。

表 2　　干扰参数

名称	参数值
干扰信道数量	4
干扰机功率	30 dBm
背景噪声功率	–90 dBm / Hz
干扰切换速度	15 ms

4.1.3　算法参数

算法参数见表 3。算法迭代次数为 2 000 次，强化学习的折扣因子 $\gamma=0.9$ 。采用学习率为 0.02 的 ADAM 优化器训练卷积神经网络，每个发射机在每次迭代时从经验池中批次采样 64 个经验进行训练。

表 3　　算法参数

名称	参数值
迭代次数	2 000
折扣因子	$\gamma=0.9$
学习率	$\alpha=0.02$
采样批次大小	$B=64$ 个

4.2　仿真分析

不同数据包到达速率下的网络吞吐量如图 5 所示。每 50 个时隙计算一次网络吞吐量，网络吞吐量为 50 个时隙内所有发射机平均成功传输数据包所需发送时间与 50 个时隙总时长之比。可以看出，随着迭代次数的增加，多种数据包到达速率下的网络吞吐量都在快速提高并收敛。

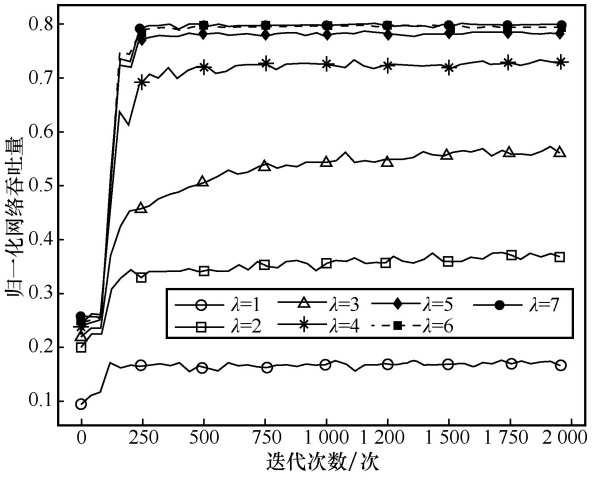


图 5　不同数据包到达速率下的网络吞吐量

为验证所提算法的有效性，本文与以下两种方法进行了对比。

1) 基于干扰感知的 Q-learning 算法[19]

为了让该方法适用于仿真所采用的场景，每个发射机根据上一时隙感知结果使用独立 Q-learning 算法独立进行信道选择和数据调度联合决策，仿真中采用了与本文一样的纯贪婪策略。

2) 智能信道决策抗干扰算法[20]

每个发射机根据频谱环境使用分布式 DQN 算法选择传输信道并以固定数量的数据包 $H(H=2,3,4,5)$ 在每个时隙进行传输，当缓冲区内数据包数量小于 H 则发送缓冲区内所有数据包。

不同数据包达到速率下，网络吞吐量对比如图 6 所示，可以看出，数据包到达速率 λ 为 1 和 2 时，需要传输的数据包较少，各种算法均倾向于把所存储的数据包都发送出去，所以所提算法与对比算法的网络吞吐量十分接近。随着数据包到达速率的增大，业务数据到达越来越多，网络吞吐量整体呈现不断提高的趋势。智能信道决策算法 $H=2$ 时，每个时隙传输数据量较少，数据包到达速率 $\lambda\geqslant 3$ 以后网络吞吐量不再提升。智能信道决策算法 $H=3$ 时，因为数据包到达速率 $\lambda\geqslant 4$ 以后，缓冲区数据包不断堆积，传输信道的调整造成吞吐量产生一定的降低。智能信道决策算法 $H=4$ 时，能较好平衡缓冲区数据堆积和传输信道选择，取得较高的吞吐量。智能信道决策算法

$H=5$ 时，在数据包到达速率 $\lambda \geqslant 4$ 之后，因为每个时隙发送数据包过多，恶意干扰造成丢包增多，传输信道的调整又带来同频互扰，所以网络吞吐量不断下降。*Q*-learning 算法的网络吞吐量随着数据包到达速率的增大逐渐上升到 0.62 左右。所提算法根据频谱环境和缓冲区数据量，智能进行信道和数据调度联合决策，在躲避干扰和避免互扰的同时选择合适数量的数据包进行传输，吞吐量优势明显，在恶意干扰环境下取得更高的网络吞吐量。

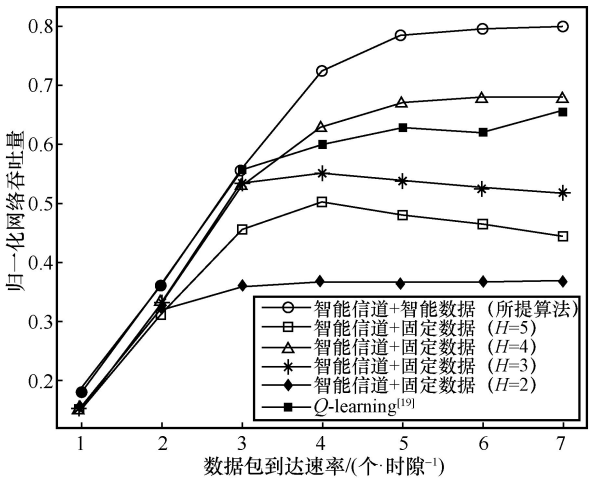


图 6　网络吞吐量对比

由于仿真了 7 种数据到达速率下的网络吞吐量变化趋势，数据包到达速率 $\lambda=4$ 时的网络吞吐量如图 7 所示。随着迭代次数的增加，所提算法与对比算法的网络吞吐量不断提高并收敛，所提算法的收敛水平显著高于对比算法。

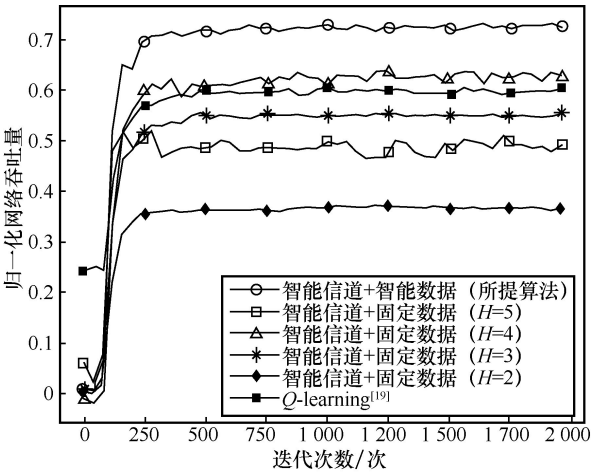


图 7　数据包到达速率 $\lambda=4$ 时的网络吞吐量

不同数据包达到速率下，平均丢包数量对比如图 8 所示。随着到达速率的增大，需要传输的数据包数量不断提高，但缓冲区存储空间和每个时隙的最大传输容量有限，数据包到达速率提高到一定值以后，平均丢包数量几乎以直线上升。所提算法对频谱环境的利用和数据调度决策更加合理，平均丢包数量最少。

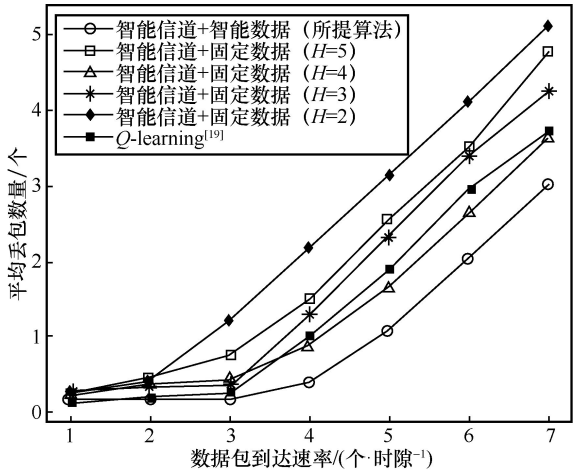


图 8　平均丢包数量对比

数据传输过程中避开恶意干扰和避免同频互扰的效果是衡量抗干扰算法性能的重要指标，数据传输需要消耗发射机的能量资源，干扰造成丢包的同时也浪费了传输所消耗的能量。不同数据包到达速率下，干扰丢包数量对比如图 9 所示，可以看出，智能信道决策算法在各种固定传输数据量的情况下都有一定数量的数据包因干扰而丢失，且每时隙固定传输数据量越多，因干扰造成丢包数量越多，智能信道决策算法 $H=5$ 时，因为干扰和同频互扰造成的丢包情况最为严重。*Q*-learning 算法的抗干扰性能比较优异，因干扰造成的丢包数量很少。所提算法对信道和数据调度联合决策，合理规划每个时隙传输数据包的数量，在各种数据包到达速率下收敛阶段几乎能完全避开恶意干扰和同频互扰，避免了传输失败造成的能量浪费。

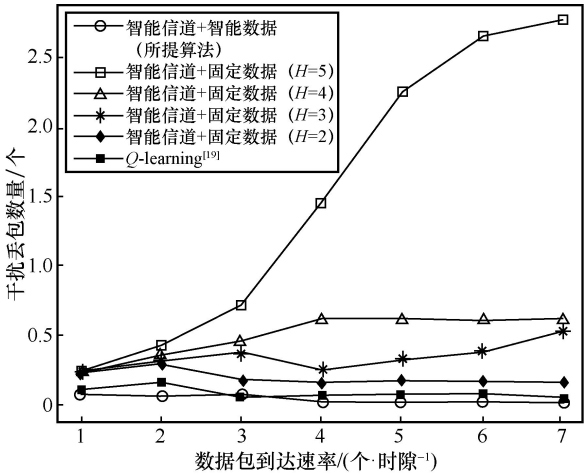


图 9　干扰丢包数量对比

不同数据包到达速率下，平均缓冲区长度对比如图 10 所示。数据包到达速率较小时，业务数据能够很快被传输，缓冲区数据量堆积较少。随着数据包到达速率的提高，需要传输的业务数据不断增多，缓冲区内所堆积的数据包数量不断上升。

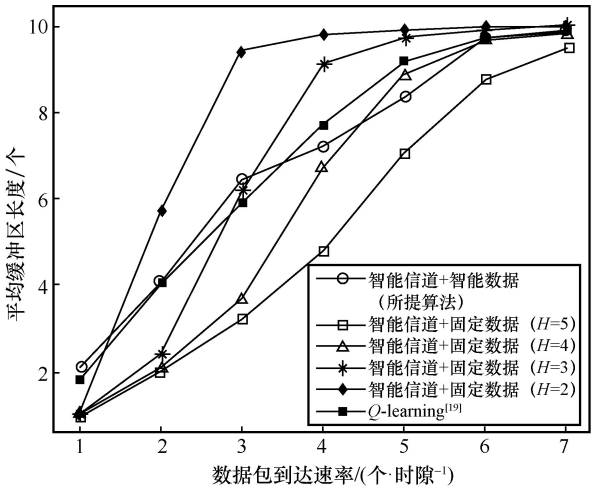


图 10　平均缓冲区长度对比

缓冲区内堆积的数据包数量越多，下个时隙业务数据到来时，存储空间不足造成溢出丢包的风险增大，不同数据到达速率下，缓冲区溢出丢包数量对比如图 11 所示，智能信道决策算法溢出数据包数量与每时隙固定发送数据包数量和数据包到达速率密切相关。智能信道决策算法每时隙固定发送数据包数量 H 越大，缓冲区溢出造成丢包数量越少，智能信道决策算法 $H=5$ 缓冲区溢出数据最少，因为每个时隙都尽可能发送最多的数据包，但干扰造成丢包数量过多网络吞吐量很低。Q-learning 算法与智能信道决策算法 $H=3$ 时的缓冲区溢出状况相似。所提算法平均缓冲区溢出造成丢包数量与智能信道决策算法 $H=4$ 相似，并拥有最高的网络吞吐量。不同数据包到达速率下，网络吞吐量随用户数量的变化如图 12 所示。随着用户规模的增大，各种数据包到达速率下的网络吞吐量均呈现下降趋势。用户规模在 6 个及以下时，用户数量不多于可用信道的数量，网络吞吐量下降较少。当用户规模超过 6 个之后，网络吞吐量开始大幅度下降。所提算法在用户数量不大于可用信道数量的情况具有可靠的分布式性能。

不同数据包到达速率下，平均奖励值对比如图 13 所示，可以看出，奖励值呈现先上升后下降的趋势，奖励值与成功传输数据包数量、缓冲区内数据量有关。数据包到达速率较小时，成功传输数据量较少所以平均奖励值较小；随着数据到达速率提高，奖励值不断上升；平均奖励值上升到一个最大值之后，缓冲区数据出现堆积，因为奖励值存在缓冲区折扣，所以平均奖励值下降。智能信道决策算法平均奖励值的最大值与每时隙发送数据数量 H 密切相关，除 $H=5$ 受到干扰较为严重，其余智能信道决策算法的最大奖励值出现在 $H=\lambda$ 处。Q-learning 算法在数据包达到速率 $\lambda=4$ 时到达最大值，随后开始缓慢下降。所提算法拥有较高的奖励值，但是平均奖励稍低于对比算法 $H=4$，因为对比算法以固定数据包数量进行数据传输不需要考虑数据调度决策，而所提算法需要选择合适的数据包数量发送，当决策错误时会收到一个较大的负奖励值，故奖励值相对比对比算法 $H=4$ 时略低。

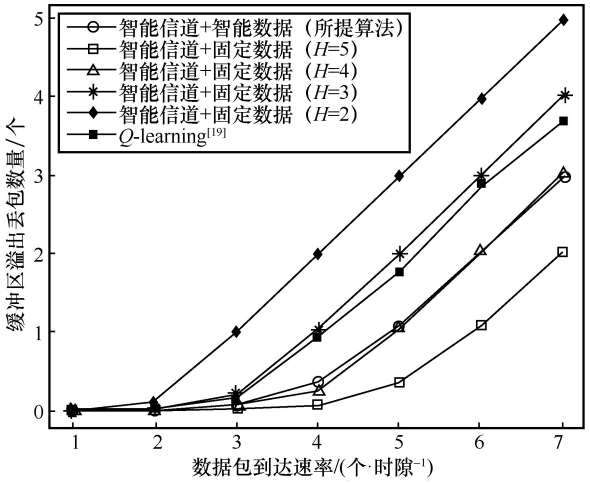


图 11　缓冲区溢出丢包数量对比

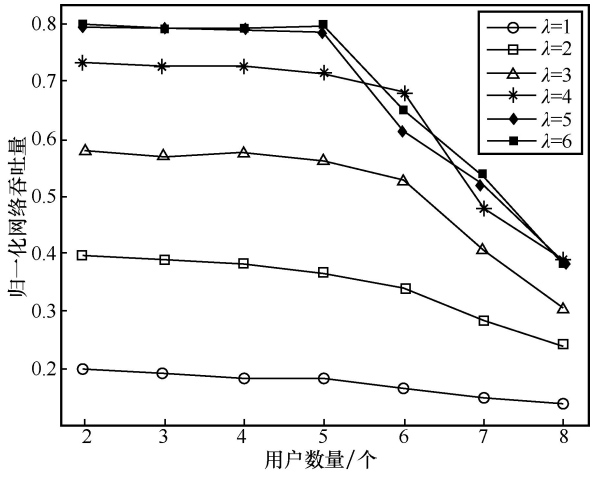


图 12　网络吞吐量随用户数量的变化

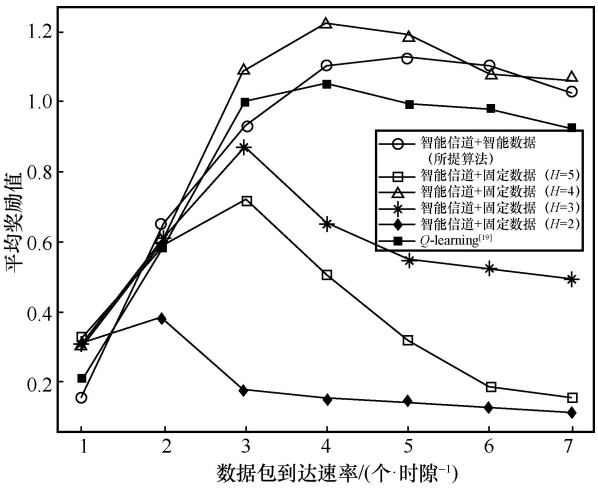


图 13　平均奖励值对比

不同数据包到达速率下，网络吞吐量随用户数量的变化如图 13 所示。随着用户规模的增大，各种数据包到达速率下的网络吞吐量均呈现下降趋势。用户规模在 6 个及以下时，用户数量不多于可用信道的数量，网络吞吐量下降较少。当用户规模超过 6 个之后，网络吞吐量开始大幅度下降。所提算法在用户数量不大于可用信道数量的情况具有可靠的分布式性能。

5　存在的问题及挑战

本文对分布式物联网协同抗干扰数据传输问题做了一定的研究，但是还存在诸多不足，以下几个方面的问题有待进一步研究。

1) 异构数据业务传输问题：万物互联的时代各种异构的业务数据存在传输需求，固定带宽信道的传输不是最合理的传输方案，研究可变带宽信道的数据传输方案将会是一个有前景的方向[27]。

2) 异步传输问题：分布式无线网络进行全网时隙同步难度较大，研究异步数据传输方法是一项有意义并富有挑战性的工作。

3) 并行无线信道数据传输问题[28]：单信道认知无线网络已有许多研究，但并行无线信道的研究成果较少。并行信道能成倍提高传输效率，但并行信道存在系统复杂和抗干扰能力弱的问题，因此在并行无线信道的研究取得突破十分有意义。

4) 传输时延问题[29]：时延敏感型业务和时延不敏感型业务的传输需求存在差异[30]，当两种业务同时存在时，如何进行传输调度分配来降低时延又减少频繁占用通信时隙也是一个值得探讨的问题。

6　结束语

本文研究了恶意干扰环境下分布式物联网的协同抗干扰问题。考虑发射机之间信息交互不可靠的实际情况和频谱环境的动态不确定性，将问题建模成多用户马尔可夫决策过程，每个发射机的优化目标是最大化长期累积传输成功数据包的数量；提出基于多智能体深度强化学习的协同抗干扰信道和数据联合决策算法；根据问题定义了系统状态、用户动作、奖励函数和深度 Q 网络的结构。每个发射机独立进行频谱感知、动作决策以及神经网络的训练更新，在无须信息交互情况下能够快速收敛，显著提高了网络吞吐量。最后，通过仿真验证了算法的性能和收敛性。

参考文献：

[1] CHOWDHURY M Z, SHAHJALAL M, AHMED S, et al. 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions[J]. IEEE Open Journal of the Communications Society, 2020(1): 957-975.

[2] ZHANG L, LIANG Y C, NIYATO D. 6G Visions: mobile ultra-broadband, super internet of things, and artificial intelligence[J]. China Communications, 2019, 16(8): 1-14.

[3] AL-FUQAHA A, GUIZANI M, MOHAMMADI M, et al. Internet of things: a survey on enabling technologies, protocols, and applications[J]. IEEE Communications Surveys &Tutorials, 2015, 17(4): 2347-2376.

[4] PIRAYESH H, ZENG H C. Jamming attacks and anti-jamming strategies in wireless networks: a comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2022, 24(2): 767-809.

[5] KARAGIANNIS D, ARGYRIOU A. Jamming attack detection in a pair of RF communicating vehicles using unsupervised machine learning[J]. Vehicular Communications, 2018, 13: 56-63.

[6] 王海超, 王金龙, 丁国如, 等. 空天地一体化网络中智能协同抗干扰技术[J]. 指挥与控制学报, 2020, 6(3): 185-191.
WANG H C, WANG J L, DING G R, et al. Intelligent cooperative anti-jamming technology in space-air-ground integrated networks[J]. Journal of Command and Control, 2020, 6(3): 185-191.

[7] 冉雨, 程郁凡, 陈大勇, 等. 采用BP神经网络的智能抗干扰决策引擎研究[J]. 信号处理, 2019, 35(8): 1350-1357.
RAN Y, CHENG Y F, CHEN D Y, et al. Intelligent anti-jamming decision engine based on BP neural network[J]. Journal of Signal Processing, 2019, 35(8): 1350-1357.

[8] KONG L J, XU Y H, ZHANG Y L, et al. A reinforcement learning approach for dynamic spectrum anti-jamming in fading environment[C]//Proceedings of 2018 IEEE 18th International Conference on Communication Technology. Piscataway: IEEE Press, 2018: 51-58.

[9] PEI X F, WANG X M, YAO J N, et al. Joint time-frequency anti-jamming

communications: a reinforcement learning approach[C]//Proceedings of 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP). Piscataway: IEEE Press, 2019: 1-6.

[10] HAN H, WANG X M, GU F L, et al. Better late than never: GAN-enhanced dynamic anti-jamming spectrum access with incomplete sensing information[J]. IEEE Wireless Communications Letters, 2021, 10(8): 1800-1804.

[11] XIAO L, WAN X Y, LU X Z, et al. IoT security techniques based on machine learning: how do IoT devices use AI to enhance security? [J]. IEEE Signal Processing Magazine, 2018, 35(5): 41-49.

[12] LIU X, XU Y H, JIA L L, et al. Anti-jamming communications using spectrum waterfall: a deep reinforcement learning approach[J]. IEEE Communications Letters, 2018, 22(5): 998-1001.

[13] XU Y F, XU Y H, REN G C, et al. Play it by ear: context-aware distributed coordinated anti-jamming channel access[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 5279-5293.

[14] XU Y F, XU Y H, DONG X, et al. Convert harm into benefit: a coordination-learning based dynamic spectrum anti-jamming approach[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 13018-13032.

[15] XU Y F, REN G C, CHEN J, et al. A one-leader multi-follower Bayesian-stackelberg game for anti-jamming transmission in UAV communication networks[J]. IEEE Access, 2018(6): 21697-21709.

[16] YAO F Q, JIAL L. A collaborative multi-agent reinforcement learning anti-jamming algorithm in wireless networks[J]. IEEE Wireless Communications Letters, 2019, 8(4): 1024-1027.

[17] WANG X M, XU Y H, CHEN J, et al. Mean field reinforcement learning based anti-jamming communications for ultra-dense Internet of Things in 6G[C]//Proceedings of 2020 International Conference on Wireless Communications and Signal Processing (WCSP). Piscataway: IEEE Press, 2020: 195-200.

[18] ELLEUCH I, POURRANJBAR A, KADDOUM G. A novel distributed multi-agent reinforcement learning algorithm against jamming attacks[J]. IEEE Communications Letters, 2021, 25(10): 3204-3208.

[19] LI W, XU Y H, GUO Q J, et al. A Q-learning-based channel selection and data scheduling approach for high-frequency communications in jamming environment[C]//Machine Learning and Intelligent Communications, 2019: 145-160.

[20] WANG X M, CHEN X Q, WANG M, et al. Decentralized reinforcement learning based anti-jamming communication for self-organizing networks[C]//Proceedings of 2021 IEEE Wireless Communications and Networking Conference. Piscataway: IEEE Press, 2021: 1-6.

[21] PEI X F, WANG X M, RUAN L, et al. Joint power and channel selection for anti-jamming communications: a reinforcement learning approach[C]//Machine Learning and Intelligent Communications, 2019: 551-562.

[22] XUE C J. Anti-interference performance of multi-path direct sequence spread spectrum wireless communication system[C]//Proceedings of 2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT). Piscataway: IEEE Press, 2010(1): 461-464.

[23] ORORBIA M E, WARN G P. Design synthesis through a Markov decision process and reinforcement learning framework[J]. Journal of Computing and Information Science in Engineering, 2022, 22(2): 021002.

[24] FEINBERG V, WAN A, STOICA I, et al. Model-based value estimation for efficient model-free reinforcement learning[EB]. 2018.

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

[26] HE K M, SUN J. Convolutional neural networks at constrained time cost[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 5353-5360.

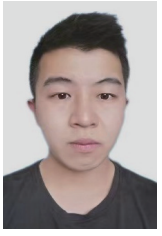
[27] ZHANG X B, WANG H, RUAN L, et al. Joint channel, power and bandwidth optimization for anti-jamming communications: amulti-agent Q-learning approach[C]//Proceedings of 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP). Piscataway: IEEE Press, 2021: 1-6.

[28] 陈昕, 徐彤, 向旭东, 等. 具有并行信道的认知无线网络性能评价研究[J]. 计算机研究与发展, 2013, 50(10): 2126-2132.
CHEN X, XU T, XIANG X D, et al. Performance evaluation of cognitive radio networks with parallel channels[J]. Journal of Computer Research and Development, 2013, 50(10): 2126-2132.

[29] LI J, HAN Y. Optimal resource allocation for packet delay minimization in multi-layer UAV networks[J]. IEEE Communications Letters, 2017, 21(3): 580-583.

[30] KAWABATA A, CHATTERJEE B C, BA S, et al. A real-time delay-sensitive communication approach based on distributed processing[J]. IEEE Access, 2017(5): 20235-20248.

[作者简介]

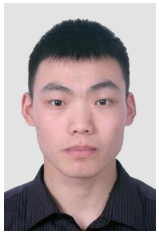


张彪（1999– ），男，陆军工程大学通信工程学院硕士生，主要研究方向为智能通信抗干扰和强化学习。

汪西明（1993– ），男，博士，国防科技大学信息通信学院讲师，主要研究方向为智能通信抗干扰、无线资源优化、多智能体决策理论等。

徐逸凡（1995– ），男，博士，陆军工程大学通信工程学院讲师，主要研究方向为无线通信和智能通信抗干扰等。

李文（1996− ），男，陆军工程大学通信工程学院博士生，主要研究方向为智能抗干扰通信、强化学习、博弈论和动态频谱接入等。



韩昊（1996− ），男，陆军工程大学通信工程学院博士生，主要研究方向为智能频谱对抗、智能通信抗干扰、博弈论、机器学习等。

刘松仪（1995− ），男，陆军工程大学通信工程学院博士生，主要研究方向为机器学习、智能抗干扰通信、无线通信资源优化等。

陈学强（1985− ），男，博士，陆军工程大学通信工程学院副教授，主要研究方向为认知无线电、无线频谱资源优化等。