Email:ycyk704@163.com TEL:010-68382327 010-68382557

基于 Q-Learning 的深度神经网络自适应退避策略*

毛中杰[1]，俞 晖[1]，麻智超[2]，王 政[1]
（1 上海交通大学 上海 200240
2 北京遥测技术研究所 北京 100076）

摘要：针对无人机自组织网络，结合 Q-Learning 和深度神经网络，提出一种自适应退避策略，以提高基于竞争的 MAC 协议通信性能。以 Matlab 为仿真平台，仿真比较了自适应退避策略与二进制指数退避策略的性能。

关键词：无人机；Q-Learning；深度神经网络；退避策略
中图分类号：TN830.6 **文献标识码：**A **文章编号：**CN11-1780(2021)01-0019-07

An adaptive back-off strategy based on deep Q-Learning neural network

MAO Zhongjie[1], YU Hui[1], MA Zhichao[2], WANG Zheng[1]
（1. Shanghai Jiao Tong University, Shanghai 200240, China;
2. Beijing Research Institute of Telemetry, Beijing 100076, China）

Abstract: An adaptive back-off strategy based on Q-Learning and deep neural network is proposed to improve the communication performance of MAC protocol based on competition for unmanned aerial vehicle self-organizing network. In the experiment, Matlab is used as the simulation platform to compare the performance of adaptive back-off strategy and binary exponential back-off strategy.

Key words: Unmanned aerial vehicle; Q-Learning; Deep neural network; Back-off strategy

引 言

无人机自组织网络由于拓扑和业务变化的多变性，常常使用基于竞争的 MAC 层协议进行通信，在基于竞争的 MAC 层协议中，退避策略是影响协议性能的重要因素。随着近年来无线传感器网络的逐渐发展，网络节点和业务类型不断增加，传统的退避算法如二进制指数型退避算法 BEB（Binary Exponential Back-off Algorithm）等在变化的网络环境下表现不佳，针对这一问题，出现许多优化 MAC 协议的退避算法。

Q-Learning 算法近年来被广泛应用于通信领域。文献[1-3]利用 Q-learning 算法优化退避算法中的部分参数，如最大重传次数和最大退避次数等，取得了一定的性能提升，但是也存在着一些缺点，比如训练效率较低，Q 表内存占用过大，查表时间过长，从而影响网络性能；同时，各个节点仅考虑自身情况，各节点接入信道的公平性不能得到保障。

人为设置 MAC 协议参数门限值以调整协议性能是另一种优化方法。文献[4]通过调整 MAC 层中的重传上限次数，来降低由于不当触发 TCP 拥塞控制机制导致的性能下降。缺点是使用手动设置门限对重传次数进行调整，不具有普适性。文献[5]充分利用了树状网络的分层以及跳数等信息，利用数学原理对不同层的不同节点的退避窗口 CW（Contention Window）进行划分，通过仿真对比，证明了其有效性。缺点是该方法仅针对静态网络有效，无法根据环境变化对 CW 进行动态调整。

部分文献提出使用增强学习对网络节点接入信道的概率进行优化。文献[6]针对 p-CSMA 在认知无线电场景中的应用提出了一种强化学习方案，网络节点在正常退避结束后，以概率 p 接入信道。文献[7]提出了一种针对 p-CSMA 的 Q-Learning 学习策略，不同于文献[6]的场景，网络中的节点均是平等的，

*基金项目：国防基础科研计划“十三五”项目（NO. JCKY2017203B082）
收稿日期：2020-07-20

节点在没有先验网络信息的情况下，通过历史网络指标，学习出各个节点独立的策略——p 的大小，以优化网络性能。

目前，大部分使用增强学习的文献受限于算法复杂度和空间平稳变化的因素，仅考虑了使用 Q-Learning 算法对通信系统进行简单的参数优化，其数学建模不完备，缺乏对环境信息的利用。而在无人机通信网络中，因其具备高速移动、拓扑变化快等特点，在通用的 Q-Learning 学习算法中，使用了有限大小 Q 表来存储状态-动作所对应的值 $Q(s, a)$，因此，智能体可以存储的 $Q(s, a)$值是有限个数的，这导致 Q-Learning 算法无法满足复杂网络的变化，而使用较大 Q 表存储将导致 Q 表内存占用过大，查表时间过长，从而影响网络性能。为了解决这一缺陷，本文提出一种适用于连续状态变化的基于 Q-Learning 的深度神经网络自适应退避算法。

1　系统模型

在通信系统中，网络节点能够实时统计一些通信指标，如传输成功率和传输时延等。不同的协议参数设置将引起通信指标的变化，与增强学习中的智能体产生动作，并从环境中获得反馈的特征相契合。基于此，本文提出一种适用于基于 Q-Learning 的深度神经网络自适应退避策略，该策略采用与环境交互的方式，基于增强学习的原理，改变退避算法的参数从通信环境中获得通信指标的反馈，并以此更新策略，最终获取奖励最大化的退避策略。该策略使用的系统模型如图 1 所示。

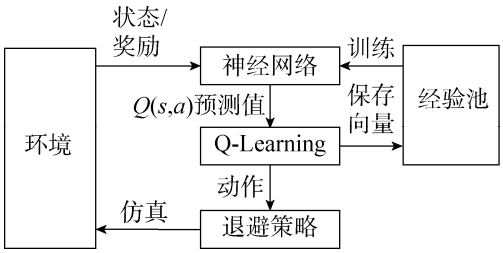

图 1　系统模型示意图

Fig. 1　Diagram of system model

在系统模型中，智能体使用退避策略在环境中进行通信，获得单位时间内的通信指标及性能，智能体将其转化为状态、奖励值，输入神经网络，神经网络输出 Q 预测值，根据 Q-Learning 算法，将期望值、预测值、输入输出转化为向量保存在经验池中；同时，输出新的动作调整退避策略，退避策略更新后，继续在环境中进行通信，循环往复。神经网络定时从经验池中抽取向量进行训练，提高神经网络预测的准确度。

1.1　退避策略

在无人机自组织网络中，节点使用基于竞争的 MAC 协议接入信道，以 CSMA/CA 协议为例，节点以随机的方式接入信道，为了避免冲突，节点在监听到信道空闲时，将退避一段时间，在这段时间内，节点将持续监听信道，当整段时间内信道均为空闲时，节点接入信道并发送数据；当节点发送数据包失败时，将根据退避策略调整退避窗口的大小，进行退避重传，重传次数到达上限时，数据包将被丢弃。相比二进制指数退避策略，本文提出一种动态调整参数的退避策略：

当节点通信冲突时，退避窗口 CW 按 m 倍增长；当节点通信完成时，退避窗口 CW 按 n 倍减少，其中常数 m、n 由基于 Q-Learning 的神经网络确定，以控制退避窗口变化的速度。

$$CW_{t+1}=\begin{cases}\max\{CW_{\min},CW_t/n\} & \text{成功传输后}\\ \min\{CW_{\max},CW_t*m\} & \text{传输失败后}\end{cases} \tag{1}$$

1.2　数学模型

基于马尔科夫决策过程，将通信系统建模如下：

状态 S：退避算法中的参数 m、n 以及网络中的业务负载 $Load$。状态 S 定义为（m，n，$Load$），其物理意义是在业务负载 $Load$ 下，退避参数设置为（m，n）。本文中 $m\in[1, 3]$，$n\in[1, 3]$。业务负载 $Load$ 具体为一跳范围内邻居节点的所有业务队列中的数据包的个数（假设每个数据包大小相同，不同大小的业务由不同个数的数据包组成），由于网络环境的多样性，业务负载 $Load$ 没有取值范围。

动作 A：本模型中的动作为参数 m、n 的大小，即 $A=\{[m_1,n_1],[m_2,n_2],\cdots,[m_k,n_k]\}$。

奖励 R：在采取动作 A 后，通信系统进行单位时间的通信，得出这一段时间的传输成功率变化量 PDR、平均时延变化量 DE 以及网络负载 $Load$ 信息，然后将这些指标与之前进行对比并归一化，为了量化节

点发送业务的公平性，我们使用网络负载标准差σ_{Ld}并引入变异系数C_v来反映各节点网络负载的波动性，将按比例计入 R 的计算。

$$PDR=\frac{PDR_{\mathrm{now}}-PDR_{\mathrm{past}}}{PDR_{\mathrm{past}}},\ DE=\frac{delay_{\mathrm{now}}-delay_{\mathrm{past}}}{delay_{\mathrm{past}}} \tag{2}$$

$$C_v=\frac{\sigma_{\mathrm{Ld}}}{Load_{\mathrm{now}}}*100\%,\ C_{\mathrm{Fairness}}=\frac{C_v^{\mathrm{past}}-C_v^{\mathrm{now}}}{C_v^{\mathrm{past}}} \tag{3}$$

$$R=\mu*PDR+\theta*DE+(1-\mu-\theta)*C_{\mathrm{Fairness}}.\quad \mu,\theta\in[0,1] \tag{4}$$

上式中，μ, θ为奖励权重因子，表示传输成功率变化量和平均时延在奖励中的占比；PDR_{now}，PDR_{past}分别表示当前单位时间内的传输成功率和前一个单位时间内的传输成功率；$delay_{\mathrm{now}}$，$delay_{\mathrm{past}}$分别表示当前单位时间内的平均时延和前一个单位时间内的平均时延；σ_{Ld}，$Load_{\mathrm{now}}$分别表示当前单位时间内的网络负载标准差和网络负载；C_v^{now}，C_v^{past}分别表示当前单位时间内的网络负载变异系数和前一个单位时间内的网络负载变异系数。

转移概率 P：在本模型中，假设通过广播发送的控制信息不会丢失，且将被各节点严格执行，理想情况下，状态转移概率全部为 1。

2　基于 Q-Learning 的神经网络设计

本文采用深度神经网络 DNN（Deep Neural Networks）模型进行训练，DNN 可以理解为有很多隐藏层的神经网络，DNN 内部的神经网络层可以分为三类：输入层，隐藏层，输出层。小的局部模型与神经网络相同，即一个线性关系$z=\sum\omega_i x_i+b$和一个激活函数$\sigma(z)$。

2.1　前向传播算法

神经网络中由输入到输出使用的是前向传播算法，对于本系统而言，输入的参数是退避算法的（m，n）参数以及业务负载共 3 个参数，最终输出的是每个动作的 $Q_{\mathrm{predict}}=Q(s,a)$，根据Q-Learning的更新公式 $Q(S,A)\leftarrow Q(S,A)+\alpha[r+\gamma\max_a Q(S',a)-Q(S,A)]$，其含义是对参数为 S,A 下的 Q 值进行更新；作用是找出参数为 S,A 的状态下，下一步采取何种动作 a 才能使得 Q 值最大化。当采取的动作策略逐渐最优化时，Q 值的更新逐渐减少，即 $r+\gamma\max_a Q(S',a)-Q(S,A)\rightarrow 0$，因此，期望的训练样本输出是 $Q_{\mathrm{real}}=r+\gamma*Q(S',a)$。

前向传播流程如图 2 所示。

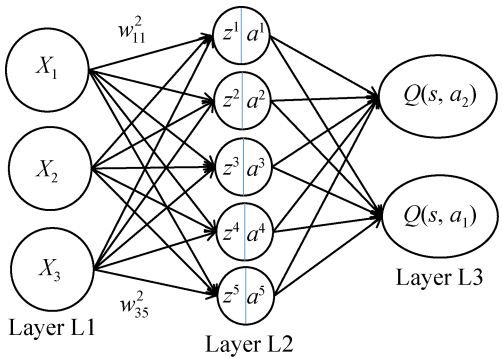

图 2　神经网络前向传播示意图

Fig. 2　Diagram of neural network forward propagation

图 2 演示一种 3 层的神经网络，输入有 3 个参数，对应业务负载和退避算法参数，输出为不同动作的 Q 值，对应了动作 A 集合中的不同动作。前向传播算法的具体流程为：

输入：总层数 L，所有隐藏层和输出层对应的矩阵 $\boldsymbol{W}$，偏移变量 b，输入值向量 x

输出：输出层的输出 a^L

1) 初始化 $a^1=x$

2) for l=2 to L，计算：

$$a^l=\sigma(z^l)=\sigma(\boldsymbol{W}^l a^{l-1}+b^l) \tag{5}$$

在前向传播算法中，总层数体现了深度神经网络有多少个隐藏层数，一般情况下，隐藏层层数越多，对结果的预测效果越好。

2.2　损失函数

本文使用均方差函数度量损失，对于每个样本，损失函数定义为

$$J(\boldsymbol{W},b,x,y)=\frac{1}{2}\|a^L-y\|_2^2 \tag{6}$$

上式中，a^L 为第 L 层的输出，即 Q 的预测值，y 为 Q 值的期望值，根据 Q-Learning 的更新公式 $Q(S, A)\leftarrow Q(S, A)+\alpha[r+\gamma\max_a Q(S', a)-Q(S, A)]$，与之对应，在本文中 $a^L=Q_{\text{predict}}$, $y=r+\gamma*Q(S', a)$。其中，r 表示奖励 R，γ表示折扣因子，意义为之前 state 的 Q 值对当前 Q 值的影响大小。

2.3　反向传播算法

为了最小化损失函数，使用批量梯度下降法作为反向传播算法，对神经网络每一层的参数 $\boldsymbol{W}$、b 进行更新。反向传播算法流程如下：

输入：总层数 L，以及各隐藏层与输出层的神经元个数，激活函数，损失函数，迭代步长β，最大迭代次数 MAX 与停止迭代阈值，输入的 q 个训练样本

$$\{(x_1,y_1),(x_2,y_2),\cdots,(x_q,y_q)\}$$

输出：各隐藏层与输出层的线性关系系数矩阵 $\boldsymbol{W}$ 和偏倚变量 b

1）初始化各隐藏层与输出层的线性关系系数矩阵 $\boldsymbol{W}$ 和偏倚向量 b 的值为一个随机值。

2） for iterator to 1 to MAX：

2-1) for i =1 to q：

a) 将 DNN 输入 a^1 设置为 x_i

b) for l=2 to L，进行前向传播算法计算

$$a^{i,l}=\sigma(z^{i,l})=\sigma(\boldsymbol{W}^l a^{i,l-1}+b^l) \tag{7}$$

c) 通过损失函数计算输出层的$\delta^{i,L}$

d) for l=L-1 to 2，进行反向传播算法计算

$$\delta^{i,l}=(\boldsymbol{W}^{l+1})^{\mathrm{T}}\delta^{i,l+1}\cdot\sigma'(z^{i,l}) \tag{8}$$

2-2) for l=2 to L，更新第 l 层的 $\boldsymbol{W}^l$，b^l：

$$\boldsymbol{W}^l=\boldsymbol{W}^l-\beta\sum_{i=1}^{q}\delta^{i,l}(a^{i,l-1})^{\mathrm{T}},b^l=b^l-\beta\sum_{i=1}^{q}\delta^{i,l} \tag{9}$$

2-3) 如果所有 $\boldsymbol{W}$、b 的变化值都小于停止迭代阈值，则跳出迭代循环到步骤 3。

3）输出各隐藏层与输出层的线性关系系数矩阵 $\boldsymbol{W}$ 和偏倚变量 b。

2.4　ε-greedy 策略

神经网络输出 Q 预测值后，智能体选出其中 Q 值最大的动作，使用贪婪策略ε-greedy 来决策下一步采取什么样的动作。贪婪策略的作用是维持探索和利用两个过程的平衡，以ε的概率进行探索，即对每个动作 a 以均等概率进行选取；以 1–ε的概率进行利用，即采用 Q 预测值中最大值对应的动作。选择动作的数学表达式如下：

$$\pi(a\mid s)=\begin{cases}\dfrac{\varepsilon}{z}+1-\varepsilon, & \text{if } a^*=\underset{a\in A}{\operatorname{argmax}}\,Q(s,a)\\ \dfrac{\varepsilon}{z}, & \text{otherwise}\end{cases} \tag{10}$$

z 表示全部可选的行为，1/z 代表从 z 个行为中随机选择行为 a 的概率，$\dfrac{\varepsilon}{z}$代表以ε的概率在 z 个可选行为中选中行为 a 的概率。最好的行为 a^*有两种情况下可能被选到，一是以 Q 值表选取，二是随机选取。

ε由以下公式确定：

$$\varepsilon=\begin{cases}0.9 & \text{if}\ \dfrac{N_{\text{iteration}}}{N_{\text{decay}}}<0.25\\ 0.7 & \text{if}\ 0.25\leqslant\dfrac{N_{\text{iteration}}}{N_{\text{decay}}}<0.5\\ 0.5 & \text{if}\ 0.5\leqslant\dfrac{N_{\text{iteration}}}{N_{\text{decay}}}<0.75\\ 0.3 & \text{if}\ 0.75\leqslant\dfrac{N_{\text{iteration}}}{N_{\text{decay}}}\leqslant1.0\end{cases}\tag{11}$$

$N_{\text{iteration}}$ 是迭代次数，N_{decay} 是预设值，即程序迭代终止条件。随着传输数据包的数目增加，探索的概率将逐步降低，利用的概率将逐渐增加，直至迭代终止。在仿真中，为了训练神经网络，探索的概率应始终大于 0。

3　仿真及结果分析

3.1　仿真参数设置

本文使用 Matlab 软件作为平台进行仿真性能分析，采用的 MAC 协议为载波侦听多路访问/冲突避免（CSMA/CA）协议，模拟 N=100 个网络节点的通信仿真，允许旧节点的退网以及新节点的入网，节点业务到达服从参数为λ的泊松分布，仿真中不考虑路由协议带来的路径选择问题，以及最大化退避策略带来的性能影响。神经网络设置为 5 层，每层神经元个数为[3, 32, 64, 36, 9]，经验池大小设置为 40，具体仿真参数如表 1 所示。

表 1　仿真参数表

Table 1　Simulation parameter table

Data_rate(bps)	$6*10^{6}$
Packet_size(bit)	200
Slot_size(s)	$9*10^{-6}$
N	100
λ	500
Max_delay(s)	0.1
Simulation_time(s)	100
Sampling interval(s)	1
N_{decay}	400

Data_rate 表示数据传输速率，以 bps 为单位；Packet_size 表示数据包的大小，以 bit 为单位；Slot_size 表示时隙的大小，以 s 为单位；N 表示节点数量，以个数为单位；λ表示服从泊松分布的业务流的参数；Max_delay 表示数据包的最大容许时延，超过该值数据包直接丢弃，以 s 为单位；Simulation_time 表示单次通信系统仿真的时间，以 s 为单位；Sampling interval 表示在单次通信系统仿真中的采样间隔，采样获得的参数将存入经验池，用于神经网络的训练，以 s 为单位。N_{decay} 表示迭代次数的预设值。

3.2　仿真结果分析

取μ=0.5，θ=0.5，在相同的仿真参数下进行，对比了二进制指数(BEB)退避策略和基于 Q-Learning 的深度神经网络 DQN (Deep Q-Learning Neural Network)自适应退避策略的仿真性能，选取了传输成功率和平均时延为评判指标，仿真结果如图 3、图 4 所示。

通过图 3 和图 4 可以看出，在迭代次数较少的情况下，两种退避策略的传输成功率和时延几乎没有差别，这表明在样本数目较少的情况下，基于 DQN 的自适应退避策略对于网络性能的提升微弱；随着迭代次数的增加，基于 Q-Learning 的深度神经网络自适应退避策略的传输成功率高于二进制指数退避策略，平均时延低于二进制指数退避策略，这证明了基于 Q-Learning 的深度神经网络自适应退避策略可以改善网络的性能。而在仿真次数到达一定数量后，神经网络的训练趋于饱和，网络性能稳定下来，与预期结果相符。

本文同时仿真了不同奖励比例（μ，θ）下的 DQN 性能，选择训练后的传输成功率和平均时延作为评判标准，仿真结果如图 5、图 6 所示。

从图 5、图 6 中可以看出，所有参数下的基于 Q-Learning 的深度神经网络自适应退避策略在训练后的传输成功率均高于训练前的传输成功率，训练后的平均时延均低于训练前的平均时延，这证明了基于 Q-Learning 的深度神经网络自适应退避策略的普适性。同时，在不同的奖励比例下，算法的性能有所差

别，证明了此算法对网络具有调控能力。

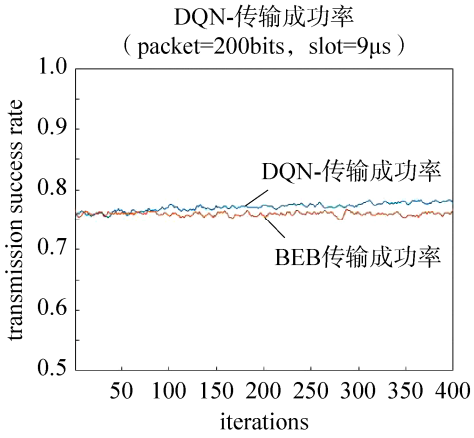

图 3　DQN-BEB 传输成功率

Fig. 3　DQN-BEB transmission success rate graph

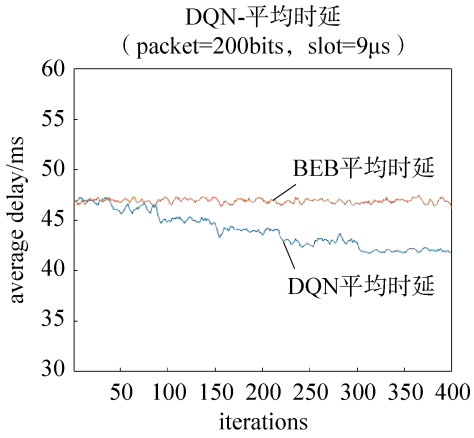

图 4　DQN-BEB 平均时延

Fig. 4　DQN-BEB average delay graph

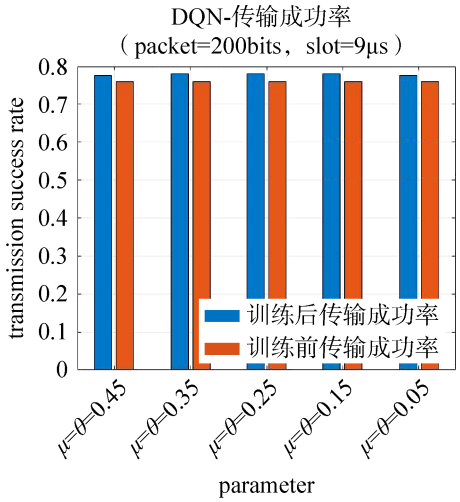

图 5　DQN 传输成功率柱状图

Fig. 5　DQN transmission success rate histogram

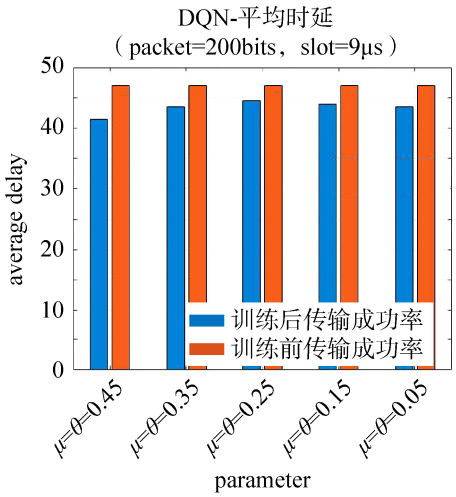

图 6　DQN 平均时延柱状图

Fig. 6　DQN average delay histogram

4　结束语

综合以上仿真结果可知，本文使用的基于 Q-Learning 的深度神经网络自适应退避策略在一定的迭代次数下，可以有效改善网络的平均时延和传输成功率等性能；同时，调整奖励函数中的比例参数，可以有效调控网络的性能，以适应不同的通信要求。本文依托于无人机配备的强大的计算资源，实施建模完备的神经网络算法，对复杂的网络环境信息进行分析，从而进行退避策略的调整，最终获得了优秀的网络通信性能。

参考文献

[1]　刘元安，李尚南，张洪光，等. 一种无线传感器网络参数自适应调节方法[P]. CN109462858A, 2019-03-12.

[2]　杜艾芊. 基于多智能体 Q 学习的车载通信 MAC 层信道接入技术研究[D]. 南京：南京邮电大学, 2017.

[3]　ANDREAS P, SHENG Z, FALAH A, et al. (2018) Contention-based learning MAC protocol for broadcast Vehicle-to-Vehicle communication[C]. IEEE Vehicular Networking Conference (VNC), 2017.

[4]　LOHIER S, GHAMRIDOUDANE Y, PUJOLLE G. MAC-layer adaptation to improve TCP flow performance in 802.11 wireless networks[C]//WiMob'2006. Montreal: IEEE Press, 2006: 427–433.

[5] CHENG B, CI L, TIAN C, et al. Contention Window-Based MAC protocol for wireless sensor networks[C]//IEEE International Conference on Dependable, 2014.
[6] BAO S, FUJII T. Q-learning based p-persistent CSMA MAC protocol for secondary user of cognitive radio networks[C]// 2011 Third International Conference on Intelligent Networking and Collaborative Systems (INCoS). IEEE Computer Society, 2011.
[7] BAYAT Y H, SHAH M V, KEBRIAEI H. A multi-state Q-learning based CSMA MAC protocol for wireless networks[J]. Wireless Networks, 2018, 24(4): 1251–1264.

[作者简介]
毛中杰　1996 年生，在读硕士研究生，研究方向为无人机网络接入协议设计优化。
俞　晖　1969 年生，硕士，高级工程师，研究方向为无线通信。
麻智超　1985 年生，硕士，高级工程师，研究方向为无线通信与组网技术。
王　政　1997 年生，在读硕士研究生，研究方向为无人机网络接入协议设计优化。

《遥测遥控》编委会换届公告

《遥测遥控》编委会换届会议暨第六届编委会第一次工作会议于 2020 年 12 月 4 日在山东威海举办。第六届编委会组成名单如下：

顾　　问　张履谦　包为民　叶培健　吴伟仁　童庆禧　苏东林　李　洪
　　李艳华　李小平　张晓林　黄大庆
主　任委员　李凉海
副主任委员　刘彦明　卢满宏
编　　委　（按姓氏笔画排序）
　　于　勇　王　宇　王　鑫　王正鹏　王立辉　王星来　司伟建
　　邢孟道　刘　昊　刘庆会　刘佳琪　刘荣科　刘衍军　安建平
　　孙厚军　孙道恒　李　波　李　聪　李军伟　汪　勃　宋振飞
　　张　睿　陈　力　陈伟平　武　楠　金　铭　孟俊敏　孟维晓
　　赵小翔　赵晓群　胡　苏　胡小工　姜秋喜　贺凯飞　贺峥光
　　卿　立　高　勇　高成臣　高晓明　黄蜀玲　曹桂兴　常　青
　　商　建　谌　明　彭红梅　韩　帅　蓝　鲲　廉保旺　翟宇梅
　　樊　昀　潘时龙　戴旭初　戴保平