文章编号：1671-7449(2024)01-0071-08

无人机辅助物联网中基于Safe Actor-Critic的信息年龄最小化研究

魏宪鹏，付　芳*，张志才

（山西大学 物理电子工程学院，山西 太原 030006）

摘　要：无人机作为一种新的通信设备，有望在物联网数据采集、监控等业务中发挥关键作用。为保证所采集数据的时效性，利用信息年龄来衡量无人机从物联网设备接收到的数据新鲜度。通过联合优化无人机轨迹和无人机与物联网设备的关联策略以最小化信息年龄加权和，并保证无人机累积飞行能量消耗满足预算要求。由于上述问题同时受短期和长期约束条件的限制，将问题建模为受约束的马尔可夫决策过程（CMDP），并利用Safe Actor-Critic来求解。仿真结果表明，所提算法在最小化信息年龄的同时，能有效保证能量预算。

关键词：无人机；信息年龄；物联网；Safe Actor-Critic

中图分类号：TP181　　**文献标识码**：A　　**doi:** 10.3969/j.issn.1671-7449.2024.01.011

引用格式：魏宪鹏，付芳，张志才．无人机辅助物联网中基于Safe Actor-Critic的信息年龄最小化研究[J]．测试技术学报，2024，38(1)：71-78.

WEI Xianpeng，FU Fang，ZHANG Zhicai. Age of information minimization for UAV-Enabled internet of things networks：A safe actor-critic approach[J]. Journal of Test and Measurement Technology，2024，38(1)：71-78.

Age of Information Minimization for UAV-Enabled Internet of Things Networks： A Safe Actor-Critic Approach

WEI Xianpeng，FU Fang*，ZHANG Zhicai

（College of Physics and Electronic Engineering，Shanxi University，Taiyuan 030006，China）

Abstract：As a novel communication device，unmanned aerial vehicle（UAV）is expected to play a key role in data collection，monitoring and applications in the Internet of Things（IoT）. To ensure the timeliness of the collected data，this paper employs the Age of Information（AoI）to measure the freshness of the data received by the UAV from IoT devices. This paper jointly optimizes the UAV's trajectory and UAV-devices association to minimize the weighted sum of AoI and ensure that the cumulative flight energy consumption of the UAV meets the energy budget requirement. Because the problem is confined by both short-term and long-term constraints，the problem is modeled as a constrained Markov decision process（CMDP）and solved by the Safe Actor-Critic approach. The simulation results show that our proposed scheme can effectively decrease AoI while guaranteeing the energy requirement.

Key words：unmanned aerial vehicles；age of information；internet of things；Safe Actor-Critic

收稿日期：2022-12-06

基金项目：山西省基础研究计划自然科学研究面上资助项目(202103021224024)；山西省基础研究计划青年科学研究资助项目(20210302123021)；山西省重点研发计划资助项目(202202020101004)

作者简介：魏宪鹏(1997－)，男，硕士生，主要从事物联网资源分配研究。E-mail：536340431@qq.com。

***通信作者**：付　芳(1985－)，女，讲师，博士，主要从事网络资源分配研究。E-mail：fufang0621@sxu.edu.cn。

0 引 言

由于无人机的灵活性、机动性和低成本，其在物联网(Internet of Things，IoT)网络中实时应用发挥着关键作用，如智能交通[1]、灾难救援[2]、野火预防[3]等。在这些应用程序中，要求将IoT设备生成的实时数据尽可能新鲜地传递给接收器。例如，智能交通中复杂的数据和过时数据可能会导致错误的操作，甚至造成灾难性的后果[4]。因此，保证接受数据的及时性对无人机辅助物联网网络至关重要。信息年龄(Age of Information，AoI)是一种有效的性能指标，其定义为自生成接收器的最新更新以来经过的时间量[5]，其中最新收到的数据包的年龄值较小，因此，可以通过最小化AoI来保证接收数据的时效性。

基于深度强化学习(Deep Reinforcement Learning，DRL)的无人机轨迹设计被认为是处理无人机路径规划问题的有效方法[6-9]，其中无人机被视为“智能体”，通过与环境直接交互获得最优轨迹。例如，Fu F等[8]提出了一种基于好奇心驱动的DQN路径规划方法；Wang L等[9]提出了一种基于深度确定性策略梯度算法的无人机路径设计方法，以降低分布式边缘计算系统中用户的能量开销。然而，这些优化问题大都受短期限制条件约束。众所周知，无人机的飞行能量预算对无人机的路径规划有很大影响，然而，他们忽略了飞行的能耗成本。考虑到无人机承载能量的局限性，Hu X等[10]提出了一种最小化无人机能耗的无人机轨迹规划方案；Liao Y等[11]提出了一种多目标优化方案，以最小化AoI和无人机的能耗成本；Sun M等[12]通过优化无人机的飞行路径和频谱分配，在AoI和飞行能量成本之间找到平衡。上述工作可以有效降低能耗，但不能保证无人机累积飞行能耗不超过总能耗预算。此外，在这些方案中，无人机的可用能量通常没有得到充分利用，难以获得最优的无人机路径规划方案，从而导致高AoI。因此，如何充分利用无人机的能量做出更合理的决策是一个值得研究的问题。

本文研究无人机的路径规划和用户关联问题，以在满足长期飞行能量约束的同时最小化AoI加权和。

1 系统模型与假设

1.1 系统模型

无人机辅助物联网场景如图1所示。IoT设备随机部署在室外区域，在该区域中，无人机从起点到目的地巡航，旨在收集IoT设备的状态信息尽可能新鲜。本文考虑的模型中，无人机在采集IoT设备信息时会处于悬停状态，因此不会产生多普勒频移现象[13]。令$K=\{1,2,\cdots,K\}$表示所有物联网设备的集合，设备k的位置由$q_k=(x_k,y_k,0),\forall k\in K$表示。UAV的巡航时间分为$T$个时隙，每个时隙的长度为$\tau$s。假设UAV在固定的高度$H$上运动，相应地，设$q[t]=(x[t],y[t],H),\forall t\in T$表示UAV在第$t$个时隙的位置，$q[0]=(x_{\text{ori}},y_{\text{ori}},H)$表示UAV的初始位置，$q[T]=(x_{\text{dest}},y_{\text{dest}},H)$表示UAV的目的地。

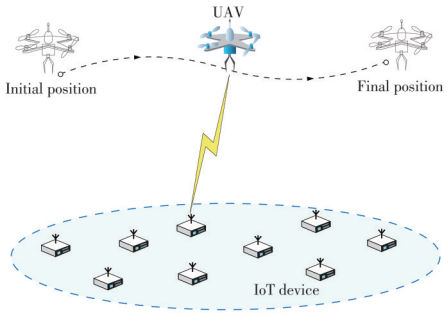

图1 系统场景

Fig. 1 System scenario

每个IoT设备在每个时隙的开始处更新其状态数据，该状态数据在调度设备时直接传输到UAV。假设利用时分多址(Time Division Multiple Access，TDMA)方案[13]，即在每个时隙中，UAV最多为一个设备提供服务。设z_k^t表示调度t时隙的设备k，即$z_k^t\in\{0,1\},\forall k\in K,t\in T$，其中$z_k^t=1$时，表示设备$k$被UAV调度，否则$z_k^t=0$。因此，有

$$\sum_{k=1}^{K}z_k^t\leqslant 1,\forall k\in K,t\in T。\tag{1}$$

1.2 飞行能量消耗模型

无人机的能源推进成本通过式(2)计算

$$P_{\text{fly}}[t]=p_0\left(1+\frac{3v^2[t]}{U_{\text{tip}}^2}\right)+P_1\left(\sqrt{1+\frac{v^4[t]}{4v_0^4}}-\frac{v^2[t]}{2v_0^2}\right)^{1/2}+\frac{1}{2}z_0\rho\mu\xi v^3[t],\tag{2}$$

式中：P_0为悬停状态恒功率；P_1为诱导功率；U_{tip}为叶片的叶尖速度；v_0为悬停状态的转子平均诱导速度；z_0和ρ分别为机身阻力比和空气密度；μ和ξ分别为转子坚固度和转子盘面积。为了UAV保留足够的能量以执行其他功能，UAV的机动性必须满足以下能量约束

$$E_{fly}[T]=\sum_{t=1}^{T}P_{fly}[t]\tau\leqslant E_{max},\tag{3}$$

式中：$E_{fly}[T]$为整个巡航期间累计推进能耗；E_{max}为UAV最大允许推进能耗[14]。

1.3　无线传输与AoI模型

令G_{k2U}表示从装置k到位置为$q[t]$的UAV的平均信道增益，其在LoS和非LoS（NLoS）链路[15]下求平均，计算公式为

$$G_{k2U}\left(q[t]\right)=20\log\left(4\pi f_c d_{k2U}\left(q[t]\right)\rho^{-1}\right)+\eta_{LoS}\Lambda_{LoS(q[t])}+\eta_{NLoS}\left(1-\Lambda_{LoS(q[t])}\right),\tag{4}$$

式中：f_c为载波频率；ρ为光速；Λ为选择概率；$d_{k2U}\left(q[t]\right)$为从设备k到UAV的距离

$$d_{k2U}\left(q[t]\right)=\sqrt{\left(x[t]-x_k\right)^2+\left(y[t]-y_k\right)^2+H^2}。$$

设备k与UAV之间可实现的数据速率

$$R_{k2U}^t=B\log_2(1+SINR_{k2U}^t),\tag{5}$$

式中：B为信道带宽；$SINR_{k2U}^t$为设备k与无人机之间的信噪比

$$SINR_{k2U}^t=P_A[t]G_{k2U}\left(q[t]\right)/\sigma^2,$$

式中：$P_A[t]$为$k2U$的发射功率；σ^2为设备k处的高斯白噪声功率。

另外，本文采用AoI作为量化信息新鲜度的指标，用D_k^t表示UAV从设备k第t时隙采集信息的数据量。$D_k^t=z_k^t R_{k2U}^k\tau$，若$D_k^t\geqslant D_{min}$，更新信息年龄为$A_k^t=1$，这意味着设备$k$的当前状态数据成功传输到无人机，其中$D_{min}$为成功恢复或解码接收到的数据所需的最小数据量[1]；如果$D_k^t\leqslant D_{min}$，则为传输失败，在t时刻物联网设备k处的AoI更新为$A_k^t=A_k^{t-1}+1$。

1.4　问题建模

通过联合优化UAV的轨迹$q[t]$以及调度策略$z[t]=\{z_1^t,\ z_2^t,\ \cdots,\ z_k^t\}$，在满足能量约束的前提下，使整个飞行周期内的长期AoI加权和最小。问题表述为

$$\min_{q[t],z(t)}\sum_{t=1}^{T}\sum_{k=1}^{K}\omega_k A_k^t(q[t],z[t]),\tag{6}$$

$$q[0]=(x_{ori},y_{ori},H),q[T](x_{dest},y_{dest},H),\tag{7}$$

$$\left\|q[t]-q[t-1]\right\|\leqslant V_{max}\tau,\tag{8}$$

$$z_k^t\in\{0,1\},\forall k\in K,t\in T,\tag{9}$$

$$\sum_{k=1}^{K}z_k^t\leqslant 1,\forall k\in K,t\in T,\tag{10}$$

$$E_{fly}[T]=\sum_{t=1}^{T}P_{fly}[t]\tau\leqslant E_{max},\tag{11}$$

式中：ω_k为A_k^t在式(6)中的权重，表示设备信息的相对重要性。UAV的初始和最终位置在式(7)中给出。式(8)为无人机的速度约束，其中V_{max}为UAV的最大速度。式(9)和式(10)保证UAV在每个时间段内最多调度一个IoT设备。式(11)表示$E_{fly}[T]$整个巡航期间的累计推进能量消耗不能大于E_{max}。接下来，我们将式(6)建模为一个CMDP，然后采用一种新的DRL算法，即Safe-Actor-Critic[16]来解决此CMDP问题。

2　约束性马尔可夫决策过程

本节将上述优化问题(6)建模为CMDP。将CMDP一个元组表示为$\langle S,A,P,s_0,r,c,c_0\rangle$，每个元素具体描述如下：

$S=S'\cup S_{dest}$为环境状态特征空间，其中S'为瞬态空间，S_{dest}为最终状态空间。S'包括3个部分：无人机在t时隙的坐标$q[t]=\left(x[t],y[t],H\right)$；物联网设备的位置$q_k=\left(x_k,\ y_k,\ 0\right)$；物联网设备的AoI值$A_k^t,\ \forall k\in K,\ t\in T$。$S_{dest}$最终状态空间为$q[T]=\left(x_{dest},y_{dest},H\right)$。

A为动作空间，包括无人机的速度v_t和方向，以及无人机的调度策略$z[t]$。

P为状态转移概率函数。无人机的坐标根据$p[t]=v_t*\tau+p[t-1]$进行转移，$v_t\leqslant V_{max}$，v_t为无人机在t时刻的飞行速度。

$s_0\in S$为初始状态，其中包括$q[0]=\left(x_0,\ y_0,\ H\right)$及$A_k^0,\ \forall k\in K$。

r为奖励函数，定义为

$$\begin{cases}-\sum_{k=1}^{K}\omega_k A_k^t, & \text{if } q[t]\neq(x_{dest},y_{dest},H),\\ -\sum_{k=1}^{K}\omega_k A_k^t+\Omega, & \text{otherwise},\end{cases}\tag{12}$$

式中：Ω为一个正常数，用于将无人机诱导到最终位置[17-18]。

c为立即约束代价，定义为$c(s,\ a)=P_{\text{fly}}[t]\tau$，$c_0$为约束代价上限，根据式(11)有$c_0=E_{\max}$。

设$\Pi(s)=\left\{\pi(\cdot|s)\middle|\sum_{a\in A}\pi(a|s)=1\right\}$表示策略集，给定$s_0$和$\pi(\forall\pi\in\Pi(s))$，agent的长期奖励为

$$R^{\pi}(s_0)=E\left\{\sum_{t=0}^{T^*-1}r(s_t,a_t)|s_0,\pi\right\},\tag{13}$$

式中：T^*为从起始状态s_0到目的地首次成功的时间。安全约束为

$$C^{\pi}(s_0)=E\left\{\sum_{t=0}^{T^*-1}c(s_t,a_t)|s_0,\pi\right\}。\tag{14}$$

解决CMDP问题的方法是找到最优策略π^*，使长期收益最大化，且满足安全约束。CMDP的优化问题被公式化为

$$\pi^*(\cdot|s)=\arg\max_{\pi\in\Delta}\left\{R^{\pi}(s_0)\middle|C^{\pi}(s_0)\leqslant c_0\right\}。\tag{15}$$

如何将长期约束$C^{\pi}(s_0)$转化为可行的单步策略集是求解CMDP的关键。

3 Safe Actor-Critic

3.1 安全策略集

本节利用Lyapunov函数理论来构建安全策略集。首先，假设可以获得式(15)的可行策略，用$\pi_b(\cdot|s)\in\Pi$表示。给定初始状态s_0和约束阈值c_0，Lyapunov函数定义集为

$$\begin{aligned}\Gamma_{\pi b}(s_0,c_0)=\{&\ell(s):B_{\pi b,c}[\ell](s)\leqslant\ell(s),\forall s\in S';\\&\ell(s)=0,\forall s\in S\backslash S';\ \ell(s_0)\leqslant s_0\},\end{aligned}\tag{16}$$

式中：$B_{\pi b,\ c}[\ell](s)$为贝尔曼函数计算，即

$$B_{\pi b,\ c}[\ell](s)=\sum_{a\in A}\pi_b(a|s)\left[c(s,\ a)+\gamma\sum_{s'\in S'}P(s'|s,\ a)\ell(s')\right],\ \forall s\in S,\ \pi_b\in\Pi。$$

对于$\forall\ell(s)\in\Gamma_{\pi b}(s_0,\ c_0)$，Lyapunov函数诱导的安全策略集为

$$F_{\ell}(s)=\left\{\pi(\cdot|s)\in\Pi(s):B_{\pi,c}[\ell](s)\leqslant\ell(s)\right\},\tag{17}$$

式中：$\ell(s_0)\leqslant c_0$，$\forall\pi(\cdot|s)\in F_{\ell}(s)$为式(15)的可行性策略。从式(17)中可以看出，较大的ℓ意味着可以获得较大的$F_{\ell}(s)$，因此，下面的关键工作是构造一个合适的Lyapunov函数ℓ。

根据文献[16]中的引理1，关于π^*的长期约束$C^{\pi^*}(s)$可以转化为π_b诱导的Lyapunov函数，写为

$$\begin{aligned}\ell_{\Delta}(s)=C^{\pi^*}(s)=E\left\{\sum_{n=0}^{T^*-1}\left[c(s_t)+\Delta(s_t)\right]|\pi_b,s\right\},\\\forall s\in S',\ell_{\Delta}(s)=0,\forall s\in S\backslash S',\end{aligned}\tag{18}$$

式中：$\Delta(s_t)$为每一步中可用的附加约束成本，用于扩展可行的操作空间并改进策略。然而，在没有π^*的先验知识的情况下构建$\Delta(s_t)$是具有挑战性的。为了降低计算复杂度[19]，$\Delta(s_t)$近似为

$$\Delta=\Delta(s_n)=\left(c_0-C^{\pi b}(s_0)\right)/E\left[T^*|s_0,\pi_b\right],\tag{19}$$

式中：$c_0-C^{\pi b}(s_0)$为从s_0到最终状态可用的总辅助约束成本；$E\left[T^*|s_0,\ \pi_b\right]$为UAV从开始位置到目的地的预期首次成功时间。通过这种方式，可以在规划轨迹的同时充分利用UAV的推进能量预算。根据式(18)可以得到$\ell_{\Delta}(s)$是可以由$\ell_{\Delta}(s)=\sum_{a\in A}\{\pi(a|s)Q_{\ell\Delta}(s,\ a)\}$计算，其中$Q_{\ell\Delta}(s,a)=Q_C(s,a)+\Delta(s)Q_T(s,a)$为$\ell_{\Delta}$的状态-动作值，$Q_C(s,a)$为约束值，$Q_T(s,a)$为从$s$到最终状态的残差步长，$\Delta(s)Q_T(s,a)$表示约束成本的其余部分。为保证策略$\pi(a|s)$安全，必须满足$\left[\pi(a|s)-\pi_b(a|s)\right]^{\mathrm{T}}Q_{\ell\Delta}(a|s)\leqslant\Delta(s)$，这意味着由$\pi(a|s)$引起的额外成本$\left[\pi(a|s)-\pi_b(a|s)\right]^{\mathrm{T}}Q_{\ell\Delta}(a|s)$不能超过$\Delta(s)$。然后，由$\ell_{\Delta}(s)$诱导的安全策略集(17)可以写为

$$\begin{aligned}F_{\ell\Delta}(s)=\{&\pi(\cdot|s)\in\Pi(s):\\&\left[\pi(a|s)-\pi_b(a|s)\right]^{\mathrm{T}}Q_{\ell\Delta}(a|s)\leqslant\Delta(s)\}。\end{aligned}\tag{20}$$

3.2 critic部分

以下采用actor-critic框架来解决问题(15)。在critic部分，使用DNN分别评估$Q(s,a)$，$Q_C(s,a)$和$Q_T(s,a)$。

在每步中，新生成的数据被保存在经验池中，即$D\leftarrow(s,a,r,c,s')\cup D$，通过从经验池中随机采样一批样本$(s,a,r,c,s')$来训练DNN，并通过式(21)更新参数

$$Loss(\vartheta)=E\left[\left(y-Q(s,a;\vartheta)\right)\right]^2,\tag{21}$$

式中：$y=r(s,a)+\max_{a\in A}\hat{Q}\left(s',a;\hat{\vartheta}\right)$为目标状态值，$Q(s,a;\vartheta)$为当前状态值，参数$\vartheta$更新为

$$\vartheta=\vartheta-\alpha_{c,t}\left(y-Q(s,a;\vartheta)\right)\cdot\nabla_{\vartheta}Q(s,a;\vartheta),\tag{22}$$

式中：$\alpha_{c,\ t}$为critic的学习率。目标状态值

$\hat{Q}\left(s',a;\hat{\vartheta}\right)$的参数$\hat{\vartheta}$经过几步后被更新为$\hat{\vartheta}=\vartheta$。

同样$Q_C(s,a)$和$Q_T(s,a)$也分别通过DNN近似器$Q(s,a;\vartheta_C)$和$Q(s,a;\vartheta_T)$进行评估。参数ϑ_C和ϑ_T通过以下方式更新

$$\vartheta_C=\vartheta_C-\alpha_{c,t}\left(y_C-Q_C(s,a;\vartheta_C)\right)\cdot\nabla_\vartheta Q_C(s,a;\vartheta_C),\tag{23}$$

$$\vartheta_T=\vartheta_T-\alpha_{c,t}\left(y_T-Q_T(s,a;\vartheta_T)\right)\cdot\nabla_\vartheta Q_T(s,a;\vartheta_T),\tag{24}$$

式中：$y_C=c(s)+\pi\left(a|s'\right)^T\hat{Q}_C\left(s',a;\hat{\vartheta}_C\right)$，$y_T=1+\pi\left(a|s'\right)^T\hat{Q}_T\left(s',a;\hat{\vartheta}_T\right)$，因此，式(18)转换为

$$\Delta(s)=\frac{\left(c_0-\pi_b(\cdot|s_0)^TQ_C(s_0,\cdot;\vartheta_C)\right)}{\pi_b(\cdot|s_0)^TQ_T(s_0,\cdot;\vartheta_T)}。\tag{25}$$

3.3　actor 部分

基于上节获得的$Q_C(s,a)$和$Q_T(s,a)$以及在式(20)中构建的安全策略集，可以计算出式(15)的最优行动概率为

$$\pi'(a|s)=\arg\max_{\pi\in\Delta}\{R^\pi(s):\left[\pi(a|s)-\pi_b(a|s)\right]^TQ_{\ell\Delta}(s,a)\leqslant\Delta(s)\},\tag{26}$$

式中：$R^\pi(s)=\pi(a|s)^TQ(s,a)$。

3.4　Safe Actor-Critic 算法

Safe Actor-Critic 算法的框架如图 2 所示。

算法收敛性可以在文献[16]中找到。该算法包括了三部分：actor 部分，critic 部分以及经验池，其学习率$\alpha_{c,t}$和$\alpha_{a,t}$满足

$$\sum_{t=0}^{\infty}\alpha_{c,t}=\infty,\ \sum_{t=0}^{\infty}\alpha_{c,t}^2<\infty,\ \sum_{t=0}^{\infty}\alpha_{a,t}=\infty,\ \sum_{t=0}^{\infty}\alpha_{a,t}^2<\infty,\ \lim_{t\to m}\frac{\alpha_{a,t}}{\alpha_{c,t}}=0。\tag{27}$$

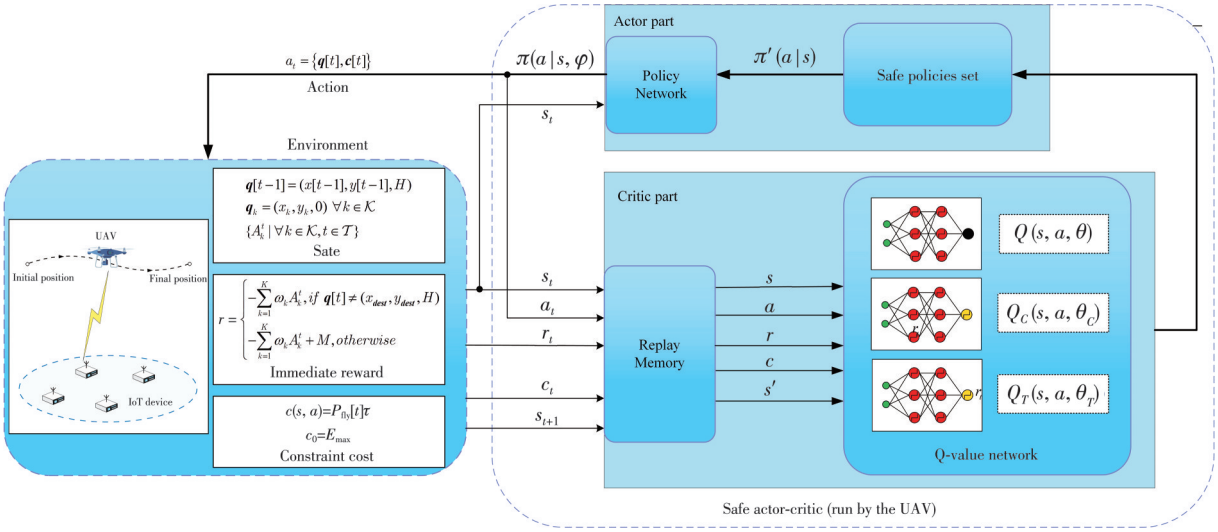

图 2　Safe Actor-Critic 框架

Fig. 2　The framework of Safe Actor-Critic

4　仿真结果与讨论

模拟基于Python的模拟器上实现，其中环境的参数设置如下：在600 m × 600 m的面积上随即部署K个物联网设备，无人机在该区域上空巡航，接收设备产生的数据，其悬停高度固定为H=100 m。传输速率的参数设置为f_c=5.9 GHz，B=1 MHz[20]，p_{k2U}=0.1 W($\forall k\in K$)[20]，σ^2=−110 dBm，信道参数的值为δ=9.61，β=0.16，η_{LoS}=1 dB，η_{NLoS}=20 dB[15]。无人机的能源推进成本参数设置为：P_0=3.4 W，P_1=118 W，U_{tip}=60 m/s，$V_{\max}$=30 m/s，v_0=5.4 m/s，ρ=1.225 km/m^2，μ=0.03，z_0=0.3，ξ=0.28 m^2[21]。

图 3 为所提算法不同 Actor 学习率之间的收敛性能，其满足等式(26)且通过反复试验来设置。在这一部分中，Critic 的学习率被设定为$\alpha_{c,t}=5\times10^{-4}$。算法总共运行 500 回合，每个回合中包括 100 步。

由图 3 可知，当学习率为$\alpha_{a,t}=5\times10^{-4}$，曲线大约 150 回合处达到收敛，这是因为学习率过高，总会导致高方差和低奖励。然而，当学习率下降为$\alpha_{a,t}=1\times10^{-5}$时，学习速率变慢。相

比 $\alpha_{a,\,t}=1\times10^{-5}$ 和 $\alpha_{a,\,t}=5\times10^{-4}$，学习率为 $\alpha_{a,\,t}=5\times10^{-5}$ 是最佳的学习率，该学习率在平均收益和方差方面具有良好的性能。

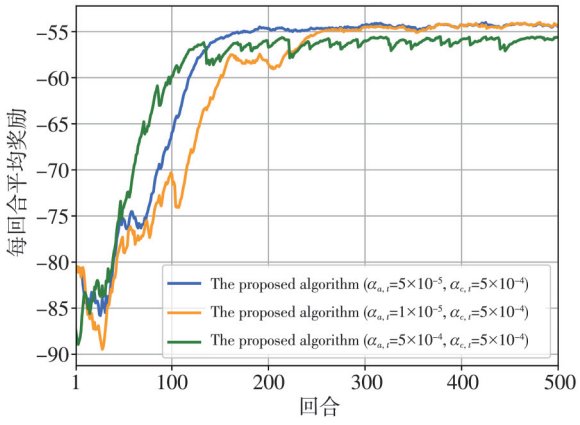

图 3 不同 Actor 学习率奖励表现

Fig. 3 The reward performance comparison with different actor's learning rates

图 4 为不同 Critic 学习率之间的收敛性能，这里 Actor 的学习率被固定为 $\alpha_{a,\,t}=5\times10^{-5}$。同样发现算法的收敛性能对学习率非常敏感，学习率为 $\alpha_{c,\,t}=5\times10^{-3}$ 导致显著方差，而 $\alpha_{c,\,t}=3\times10^{-4}$ 导致较长的学习时间，Critic 的最佳学习率为 $\alpha_{c,\,t}=5\times10^{-4}$。因此，在下面的部分中，$\alpha_{a,\,t}$ 和 $\alpha_{c,\,t}$ 分别被设为 $\alpha_{a,\,t}=5\times10^{-5}$ 和 $\alpha_{c,\,t}=5\times10^{-4}$。

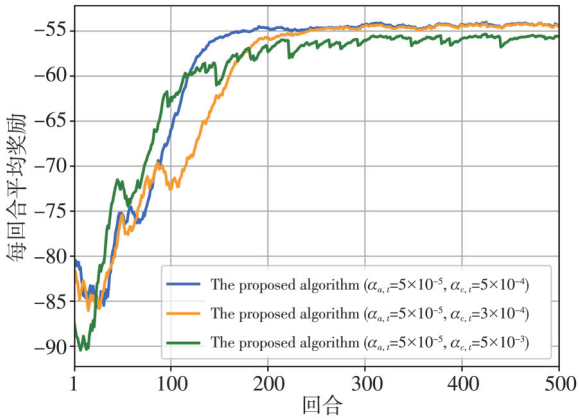

图 4 不同 Critic 学习率奖励表现

Fig. 4 The reward performance comparison with different critic's learning rates

为了显示所提出的基于 Safe Actor-Critic（SAC）算法的高效率，还模拟了基于 Safe DQN 的算法（SDA）[7] 和基于拉格朗日 Actor-Critic 的算法（LAC）[21]。图 5 为无人机在不同的总能量预算下每次 SAC、SDA 和 LAC 的累积推进能量消耗。从图 5 可以看出，当 $E_{\max}=1.1\times10^{4}$ J 时，SAC 的总推进能量成本在收敛后小于 1.1×10^{4} J，SDA 的能耗成本同样小于 1.1×10^{4} J。当 $E_{\max}=2.6\times10^{4}$ J 时，SAC 的能耗约 2.5×10^{4} J。这是因为 SAC 基于能量预算 $E_{\max}$ 为无人机构建了一个安全策略集，因此，总推进能量成本不会超过预算 $E_{\max}$。当 $E_{\max}=1.1\times10^{4}$ J 时，LAC 的能量消耗约为 1.5×10^{4} J。这是因为 LAC 的策略不可能受到长期能源约束的严重限制，即 UAV 的每回合的总推进能量成本可能超过总能量预算。

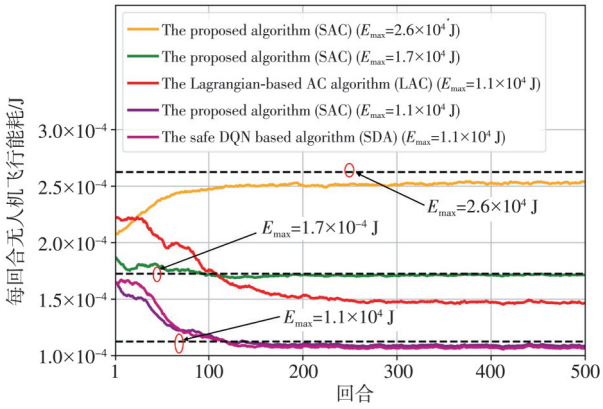

图 5 不同总能量预算下无人机飞行的累积推进能耗

Fig. 5 The UAV's cumulative propulsion energy consumption per episode with different total energy budgets

图 6 为每回合 SAC、LAC 和 SDA 在不同总能量预算下的奖励表现。

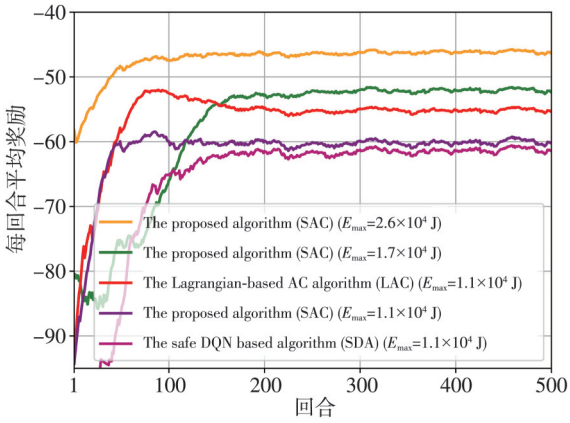

图 6 不同能量预算下每回合奖励表现

Fig. 6 The reward performance per episode with different total energy budgets

从图 6 中可以看到，当 $E_{\max}$ 从 1.1×10^{4} J 增加到 $E_{\max}=2.6\times10^{4}$ J 时，SAC 的奖励明显增加，这是因为 $E_{\max}$ 越大，则无人机的可行动空间越大，获得最优策略的机会越多[22]，获得的奖励也越高。当 $E_{\max}=1.1\times10^{4}$ J 时，LAC 的奖励比 SAC 高，这是因为 LAC 的策略并不严重受限于图 5 所示的能量预算。尽管在图 5 中，SDA 同样受到能量约束，但是从图 6 中可以看出当 $E_{\max}=1.1\times10^{4}$ J 时，SDA 的奖励低于 SAC，因此，根

据图 5 和图 6 可知，与 SDA 和 LAC 相比，提出的 SAC 可严格满足推进能量消耗预算要求，并且收敛性能最佳。

图 7 所示为每个回合中不同的总能量预算下不同物联网设备数目的 AoI 值，可见随着物联网设备数目的增加，AoI 加权和显著增加。这是因为无人机在每个时隙最多连接一台设备，部署的设备越多，平均每台设备享受的服务越少，AoI 之和也随之增加。此外，当能量预算增加时，固定数量物联网设备的 AoI 会减少，这是因为有了更多的推进能量预算，无人机可以进行更灵活的轨迹规划，以接收更高 AoI 值的设备。

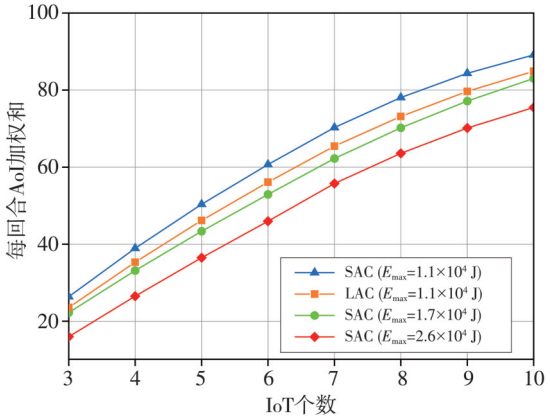

图 7　不同 IoT 个数的 AoI 加权和

Fig. 7　The weighted sum AoI of different devices

图 8 显示了每一阶段的平均加权和 AoI 与 UAV 飞行高度的关系，可见当无人机的高度增加时，AoI 值增加。由于物联网设备到无人机的信道增益主要取决于两者之间的距离，因此在带宽和发射功率一定的情况下，飞行高度越高，信道条件越弱，传输速率越低。

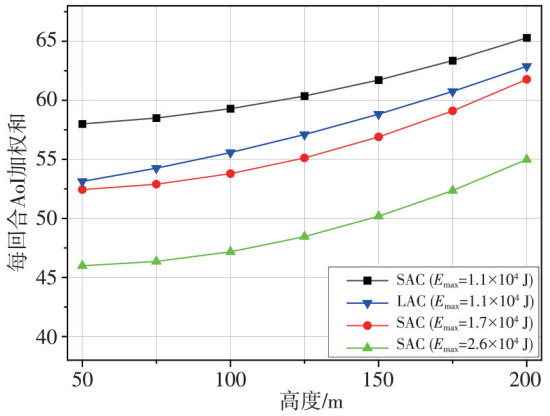

图 8　不同高度下的建立表现

Fig. 8　The reward performance versus height

5　结论和展望

本研究的贡献总结如下：

1）联合优化无人机的轨迹和物联网设备调度策略以最小化网络的加权和 AoI，其中无人机累积飞行能量成本受能量预算限制。

2）由于优化目标受一组短期约束和长期能量约束的限制，该问题被建模为约束马尔可夫决策过程（CMDP）。

3）采用 Safe Actor-Critic 来求解该 CMDP，为保证策略安全，利用 Lyapunov 函数构建安全策略集，并基于此策略集训练策略网络。

在未来的工作中，我们将利用多智能体 DRL 方法讨论多无人机场景下的 AoI 最小化问题。

参考文献：

[1] SAMIR M，ASSI C，SHARAFEDDINE S，et al. Age of information aware trajectory planning of UAVs in intelligent transportation systems：a deep learning approach[J]. IEEE Transactions on Vehicular Technology，2020，69(11)：12382-12395.

[2] LIN N，LIU Y，ZHAO L，et al. An adaptive UAV deployment scheme for emergency networking[J]. IEEE Transactions on Wireless Communications，2021，21(4)：2383-2398.

[3] SUN L，WAN L，WANG X. Learning-based resource allocation strategy for industrial IoT in UAV-enabled MEC systems[J]. IEEE Transactions on Industrial Informatics，2020，17(7)：5031-5040.

[4] SU C，YE F，WANG L C，et al. UAV-assisted wireless charging for energy-constrained IoT devices using dynamic matching[J]. IEEE Internet of Things Journal，2020，7(6)：4789-4800.

[5] ZHANG L，ANSARI N. Latency-aware IoT service provisioning in UAV-aided mobile-edge computing networks[J]. IEEE Internet of Things Journal，2020，7(10)：10573-10580.

[6] 钱志鸿，田春生，郭银景，等. 智能网联交通系统的关键技术与发展[J]. 电子与信息学报，2020，42(1)：2-19.
QIAN Zhihong，TIAN Chunsheng，GUO Yinjing，et al. The key technology and development of intelligent and connected transportation system[J]. Journal of Electronics and Information，2020，42(1)：2-19.(in Chinese)

[7] WANG L，WANG K，PAN C，et al. Deep Q-

network based dynamic trajectory design for UAV-aided emergency communications[J]. Journal of Communications and Information Networks, 2020, 5(4): 393-402.

[8] FU F, JIAO Q, YU F R, et al. Securing UAV-to-vehicle communications: a curiosity-driven deep Q-learning network (C-DQN) approach [C]//International Conference on Communications Workshops (ICC Workshops), 2021: 1-6.

[9] WANG L, WANG K, PAN C, et al. Deep reinforcement learning based dynamic trajectory control for UAV-assisted mobile edge computing [J]. IEEE Transactions on Mobile Computing, 2022, 21(10): 3536-3550.

[10] HU X, WONG K K, YANG K, et al. UAV-assisted relaying and edge computing: scheduling and trajectory optimization[J]. IEEE Transactions on Wireless Communications, 2019, 18(10): 4738-4752.

[11] LIAO Y, FRIDERIKOS V. Energy and age pareto optimal trajectories in UAV-assisted wireless data collection[J]. IEEE Transactions on Vehicular Technology, 2022, 71(8): 9101-9106.

[12] SUN M, XU X, QIN X, et al. AoI-energy-aware UAV-assisted data collection for IoT networks: a deep reinforcement learning method[J]. IEEE Internet of Things Journal, 2021, 8(24): 17275-17289.

[13] FANG Z, WANG J, REN Y, et al. Age of information in energy harvesting aided massive multiple access networks[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(5): 1441-1456.

[14] JEONG S, SIMEONE O, KANG J. Mobile edge computing via a UAV-mounted cloudlet: optimization of bit allocation and path planning[J]. IEEE Transactions on Vehicular Technology, 2017, 67(3): 2049-2063.

[15] ZHANG L, ANSARI N. Latency-aware IoT service provisioning in UAV-aided mobile-edge computing networks[J]. IEEE Internet of Things Journal, 2020, 7(10): 10573-10580.

[16] CHOW Y, NACHUM O, DUENEZ-GUZMAN E, et al. A lyapunov-based approach to safe reinforcement learning [C]//32nd Conference on Neural Information Processing Systems(NeurIPS 2018), 2018: 31-41.

[17] DONG D, CHEN C, CHU J, et al. Robust quantum-inspired reinforcement learning for robot navigation[J]. IEEE/ASME Transactions on Mechatronics, 2010, 17(1): 86-97.

[18] 陈兴国, 高阳, 范顺国, 等. 基于核方法的连续动作 Actor-Critic 学习[J]. 模式识别与人工智能, 2014(2): 103-110.
CHEN Xingguo, GAO Yang, FAN Shunguo, et al. Kernel-based continuous-action Actor-Critic learning [J]. Journal of Pattern Recognition and Artificial Intelligenceon, 2014(2): 103-110. (in Chinese)

[19] HU H, XIONG K, QU G, et al. AoI-minimal trajectory planning and data collection in UAV-assisted wireless powered IoT networks[J]. IEEE Internet of Things Journal, 2020, 8(2): 1211-1223.

[20] ZENG Y, XU J, ZHANG R. Energy minimization for wireless communication with rotary-wing UAV[J]. IEEE Transactions on Wireless Communications, 2019, 18(4): 2329-2345.

[21] ALTMAN E. Constrained Markov decision processes with total cost criteria: lagrangian approach and dual linear program [J]. Mathematical Methods of Operations Research, 1998, 48(3): 387-391.

[22] ZHAO Y, LI Z, CHENG N, et al. Joint UAV position and power optimization for accurate regional localization in space-air integrated localization network[J]. IEEE Internet of Things Journal, 2020, 8(6): 4841-4854.

声 明