马志昕，骆淑云. 基于 Advanced-TD3 算法的卫星探索控制策略[J]. 智能计算机与应用,2024,14(4):83-88. DOI:10.20169/j.issn.2095-2163.240411

基于 Advanced-TD3 算法的卫星探索控制策略

马志昕，骆淑云
(浙江理工大学 计算机科学与技术(人工智能)学院，杭州 310018)

摘 要：卫星控制算法在卫星控制领域拥有十分重要的地位,而深度强化学习则是当前前沿的卫星控制算法之一。针对目前太空环境日渐复杂的问题,提出了基于 TD3 算法的改进 TD3(Advanced-TD3)算法,实现控制卫星到达预定目标区域。在开源环境中进行仿真实验,实验结果验证了该算法的空间探索能力,拥有较高的鲁棒性,可以较为精确地帮助卫星完成控制问题,增强卫星对复杂空间中的控制能力,提高卫星的运行效率。

关键词：深度强化学习；Advanced-TD3 算法；卫星控制；空间探索
中图分类号：TP181/P185.18 **文献标志码**：A **文章编号**：2095-2163(2024)04-0083-06

Satellite exploration control strategy based on Advanced-TD3 algorithm

MA Zhixin, LUO Shuyun
(School of Computer Science and Technology(Artificial Intelligence), Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Satellite control algorithms have a critical position in the field of satellite control, Deep reinforcement learning is one of the current cutting-edge satellite control algorithms, In response to the current problem of an increasingly complex space environment, Improved TD3 (Advanced-TD3) algorithm a satellite control method based on the improved TD3 algorithm is proposed, which is used to control the satellite to reach the target point automatically. The algorithm is simulated in an open-source environment under Python, and the experimental results verify the space exploration capability of the algorithm, which possesses high robustness and can help the satellite to complete the control problem more accurately in order to enhance the control capability of the satellite in the complex space and improve the operation efficiency of the satellite.

Key words: deep reinforcement learning; Advanced-TD3 algorithm; satellite control; space exploration

0 引 言

航天技术是人类历史上最为复杂的技术之一,而探索太空资源、研究航天战略部署一直都是中国的重要方针。人造卫星大量部署在近地轨道、中地球轨道和地球同步转移轨道中,而这些轨道本身为极其复杂的环境,制定好的卫星控制系统就成为了太空安全的重要课题。传统的卫星控制方法依赖地面中心各种观测设备的预判以及计算,并针对计算结果对自动控制方法的参数进行修改控制,通常耗时、耗能且延迟很高,都没达到即时控制行为,给卫星安全带来了极大的隐患。控制的指令因为某些特殊因素导致无法准确传达,有可能造成无法挽回的损失。

近年来,强化学习在卫星控制领域得到了广泛研究和应用,如卫星姿态控制、卫星导航、卫星遥感图像处理、卫星任务调度等,但在卫星自主控制领域中应用较少。使用强化学习来实现卫星自主控制有适应性强、自主性强、灵活性高、效率高等优势,可以提高卫星控制任务的适应性、性能和效率。

本文基于以上问题和结论提出 Advanced-TD3 算法,运用该算法成功实现控制卫星到达预定目标区域的变轨控制。

1 研究现状与知识背景

卫星变轨是指卫星改变轨道的过程,包括卫星升轨、卫星降轨、卫星偏转轨道等操作。卫星变轨技术可以用于卫星的姿态控制、轨道调整、避让碰撞、

作者简介：马志昕(1996-),男,硕士研究生,主要研究方向:强化学习,任务调度。
通讯作者：骆淑云(1986-),女,博士,讲师, 主要研究方向:工业边缘计算,区块链,强化学习。Email:Shuyunluo@ zstu.edu.cn

收稿日期：2023-03-16

重新进入大气层等,是卫星运行中必不可少的技术手段之一。保证卫星在目标区域运动的技术被称为“轨道控制技术”,主要包括卫星的姿态控制、轨道控制和动量管理等,以保持卫星在预定轨道上稳定运行。

在机械结构的控制算法中,包括主/被动磁姿态控制系统的控制算法、拥有反作用力轮的姿态控制、带有滑膜控制的姿态控制等,其拥有独特的控制结构,针对这些结构的控制来控制卫星姿态的变化[1-3]。在宏观的卫星姿态控制算法中,有 PID(Proportional、Integral、Differential)算法、线性二次调节(LQR)算法、反演控制(Backstepping Control)算法、非线性自适应控制算法等,这些算法都在现实中拥有着大量的应用[4-7]。

太空任务日渐复杂,常规的姿态控制不能满足复杂的太空任务,需要在软硬件中对细节参数进行复杂的调整,难以适应动态复杂的任务需求[8]。

当前,基于强化学习的应用已经成功的完成了大量富有挑战的任务,而在卫星姿态控制以及通信任务中也有诸多成功的应用[9]。TD3 算法(Twin Delayed Deep Deterministic Policy Gradient Algorithm)是一种用于连续动作空间的强化学习算法,由 Scott Fujimoto 等在 2018 年提出[10]。TD3 算法基于 DDPG(Deep Deterministic Policy Gradient)算法,通过引入双网络和延迟更新等技术,解决了 DDPG 算法中出现的 Q 值过估计和目标网络更新不稳定的问题,提高了算法的稳定性和性能。在强化学习解决卫星问题中,有大量研究者成功将其应用,并证明了强化学习解决相关问题的可行性,苗峻等[11]为解决无模型深度强化学习的探索和扩展平衡问题,设计了 ε-imitation 动作选择策略方法,最终提出了基于 ADDPG 的卫星编队控制策略;Shi Z 等[12]提出了一种将深度强化学习与预定时间稳定性相结合的姿态跟踪控制方法,不仅提高了卫星的自主决策能力,而且保证了卫星姿态控制的可靠性;刘冰雁等[13]为解决航天器与非合作目标的空间交会问题,缓解深度强化学习在连续空间的应用限制,提出了一种基于分支深度强化学习的追逃博弈算法,以获得与非合作目标的空间交会策略;Liu H 等[14]提出了一种两步自适应控制策略,在不了解每颗卫星动力学的情况下迭代求解最优编队控制问题。除此之外,Zhang Z B 等[15]在卫星追踪领域、Qu X 等[16]在卫星追捕、Gao D 等[17]在卫星通信领域和 Chu Z 等[18]在卫星姿态控制等领域均有深度研究应用。

本文将基于 Advanced-TD3 算法来实现卫星的轨道控制,以到达目标区域。

2 基本理论

2.1 轨道六根数定位模型

在模拟环境中采用轨道六根数模型来定位卫星的轨道信息。轨道六根数空间模型是物体在空间中的位置和速度的数学表示,使用 6 个参数或元素定义物体围绕天体(例如地球)的轨道,包括半长轴、离心率、倾角、升交点赤经、近拱点角和真近地点角。该模型可用于预测物体在空间中的未来位置和速度。

半长轴(a):物体围绕天体轨道半径的最大长度;

离心率(e):物体轨道的扁平程度,定义为轨道长轴与圆轨道长轴的比值;

倾角(i):物体轨道与地球的赤道面的夹角;

升交点赤经($Raan$):物体轨道上升交点(即物体从地球背面穿过地球赤道面)与地球赤道面的投影在地球赤道面上的位置;

近拱点角(Ω):物体从近拱点(即物体离地球最近的点)移动到升交点的过程中所绕天体的轨道旋转的角度;

真近地点角(M):物体围绕天体的轨道上,从物体上一次经过近拱点到下一次经过近拱点所经过的平均角度。

轨道六根数示意图如图 1 所示。

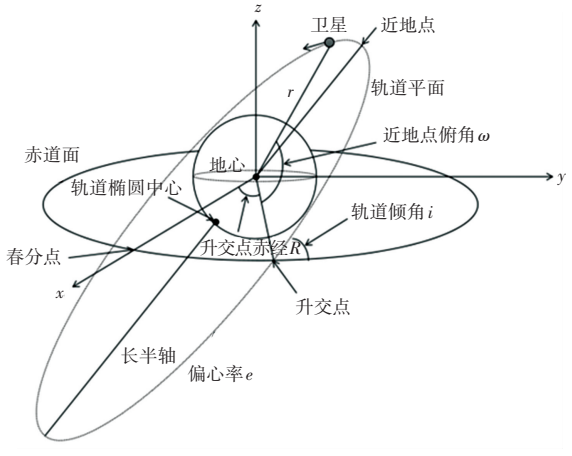

图 1 轨道六根数示意图

Fig. 1 Schematic diagram of the number of six tracks

通常只需要半长轴(a)、离心率(e)、倾角(i)、升交点赤经($Raan$)、近拱点角(Ω) 便可计算出轨道,再通过真近地点角(M) 可以得出卫星在轨道上的具体位置。

2.2　卫星运动模型

为了说明卫星运动的方程，引入了惯性坐标系（ECI）。这个参考系是人造卫星以地球为椭圆焦点的旋转系，具体如图 2 所示。本文以地心为原点，将卫星速度定义由 $\boldsymbol{x}$、$\boldsymbol{y}$、$\boldsymbol{z}$ 3 个正交基向量组成。

详细计算如公式（1）所示：

$$\begin{cases} \boldsymbol{x} = r \times \sin(i) \times \cos(\varphi) \\ \boldsymbol{y} = r \times \sin(i) \times \sin(\varphi) \\ \boldsymbol{z} = r \times \cos(i) \end{cases} \tag{1}$$

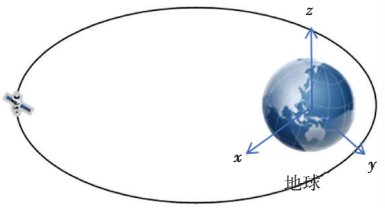

图 2　地心惯性坐标系结构示意图

Fig. 2　Structure of geocentric inertial coordinate system

2.3　强化学习

强化学习（Reinforcement Learning，RL）是机器学习的一种重要分支，相比于其他机器学习方法，强化学习更加适用于环境动态、复杂、不确定或者难以建模的场景。强化学习的学习过程中，主要研究如何通过智能体（agent）与环境（environment）的交互学习最优的行为策略，在不断的试错和学习中，找到最优的策略，以最大化累积奖励目标。智能体这种学习方法的模式称作马尔科夫决策过程（Markov Decision Process，MDP），提供了描述强化学习交互过程的一种理论框架。马尔科夫决策过程如图 3 所示。

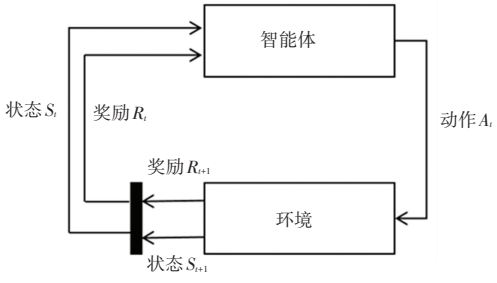

图 3　马尔科夫决策过程示意图

Fig. 3　Markov decision process diagram

3　基于 Advanced-TD3 算法的目标探索策略

TD3 是一种基于深度学习的策略梯度算法，基本思想是使用神经网络来学习控制策略。在基于 TD3 的卫星控制解决方法中，通过练神经网络来使卫星学习环境中的控制策略，通过网络输出提前界定好的动作空间数组作为在当前环境下的动作。然而智能体在状态空间趋于无穷大，且许多数据的维度不在同一量级，导致探索能力极差，训练效果差等问题。Shi 等[19]提出了一种结合深度强化学习 TD3 和传统卫星姿态控制器的运动目标跟踪观测姿态跟踪控制方法，并将 LSTM 集成到 TD3 中，从目标的图像位置学习目标的运动状态，并实时获得所需的姿态；Zhang 等[15]基于 PID-Guide TD3 算法提出了一种无模型的深度强化学习控制器，可以根据环境的反馈不断学习，实现航天器的高精度姿态控制，而无需反复调整控制器参数。

本文基于 TD3 算法，针对目标问题提出 Advanced-TD3 算法（Advanced Delayed Deep Deterministic Policy Gradients）。首先，引入 RuningMeanStd 归一化方法来解决数据维度差异过大的问题；其次，加入长短期记忆网络（LSTM，Long Short-Term Memory）来提高学习能力。算法伪代码如图 4 所示。

Algorithm 1: Advanced-TD3算法

使用随机参数θ_1, θ_2, ϕ初始化actor网络 $A_{\theta_1}, A_{\theta_2}$，critic网络$C_{\theta_1}, C_{\theta_2}$;
初始化target网络$\theta_1' \leftarrow \theta_1, \theta_2' \leftarrow \theta_2, \phi' \leftarrow \phi$;
在初始化actor网络和critic网络时，初始化 $LSTM$ 网络 L;
初始化Reply Buffer $\mathcal{B}$;
确定最大动作 maxaction;
for $t = 1$ **to** T **do**
　从Actor网络中输出一个动作，并带上随机探索噪声，再收敛动作在最大最小动作上下限内 $(a \sim \pi_\phi(s) + \epsilon)$.clip(noise,-maxaction, maxaction),$\epsilon \sim \mathcal{N}(0, \sigma)$;
　将动作a输入到环境中获得环境奖励 $r(s,a)$ 和新的 state s';
　将从环境中得到的元组 (s,a,r,s') 用 $RuningMeanStd$ 进行归一化;
　将归一化后的元组 (s,a,r,s') 存入Reply Buffer $\mathcal{B}$ 中;
　从 Reply Buffer $\mathcal{B}$ 中随机采样若干 mini-batch 经验 (s,a,r,s');
　给target action 加入噪声扰动 $\tilde{a} \leftarrow \pi_{\phi'}(s') + nosie$，　$noise \sim$ clip $(\mathcal{N}(0,\tilde{\sigma}), -\text{maxaction}, \text{maxaction})$;
　$y \leftarrow r + \gamma \min_{i=1,2} Q_{\theta_i'}(s', \tilde{a})$;
　更新 critics 网络 $\theta_i \leftarrow \text{argmin}_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(s,a))^2$;
　if t mod d **then**
　　通过确定性策略梯度更新 ϕ:
　　$\nabla_\phi J(\phi) = N^{-1} \sum \nabla_a Q_{\theta_1}(s,a)\big|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$;
　　更新 target 网络:
　　$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$;
　　$\phi' \leftarrow \tau\phi + (1-\tau)\phi'$;
　end if
end for

图 4　Advanced-TD3 伪代码

Fig. 4　Advanced-TD3 pseudo-code

在 Advanced-TD3 中，噪声扰动是一个重要的元素。噪声扰动是在策略评估步骤中加入的一个随机噪声，使得 Advanced-TD3 能够更好地评估策略的质量，并减少策略评估的偏差。扰动的加入使得 Advanced-TD3 能够更加高效地提升学习效率，并且提高了算法的鲁棒性。在智能体更新网络时对目标动作的取值加以适当的扰动可以提升探索效率，所以在多次试验后，本文给予了随着网络更新次数变化而变化的衰减噪声扰动 γ_n，详细公式如公式（2）所示：

$$\gamma_n = \gamma_{n-1} \times 0.95 \tag{2}$$

3.1 状态空间与动作空间

状态集合 s 由卫星在轨道六根数表达下位置、卫星速度构成。状态空间具体公式如公式(3)所示：

$$s_i = \{(a_i, e_i, I_i, \omega_i, \Omega_i, \varphi_i), (v_x, v_y, v_z)_i\} \tag{3}$$

其中，(v_x, v_y, v_z) 代表卫星在 x, y, z 三轴上的速度。

动作空间集合 A 由强化学习智能体网络中输出的卫星在 x, y, z 三轴加速度 a 组成，动作空间集合具体公式如公式(4)所示：

$$A_i = (a_x, a_y, a_z)_i \tag{4}$$

三轴加速度大小有所约束，其上下限为三轴上正反最大加速度，如公式(5)所示：

$$\begin{cases} a_x \in (-a^r_{x\max}, a^r_{x\max}) \\ a_y \in (-a^r_{y\max}, a^r_{y\max}) \\ a_z \in (-a^r_{z\max}, a^r_{z\max}) \end{cases} \tag{5}$$

3.2 奖励函数设计

奖励函数是强化学习算法的重要组成部分，对强化学习具有至关重要的意义。本文一方面考虑智能体在安全可靠的范围内运行，不能发生碰撞，不能超速导致偏心率大于 1 而失控；另一方面，要使得其在探索中能在稀疏环境中探索到目标区域，并且在到达目标之前要学会避障行为。在所有的奖励中，本文设置一个权重系数使不同维度的奖励尽可能在正确引导的情况下趋于平滑。

在安全性方面，碰撞惩罚 r_c 和偏心惩罚 r_{ep} 总和 R_p 设置为触发条件后就会终止训练，并且计算当前所有奖励，并赋予惩罚，σ 为碰撞惩罚系数，τ 为偏心系数，如公式(6)所示：

$$R_{\text{punish}_i} = \sigma r_{c_i} + \tau r_{\text{ep}_i} (r_{c_i} \in \{-10, 0\}, r_{\text{ep}_i} \in \{-10, 0\}) \tag{6}$$

在正常的训练任务中，根据实际情况卫星的信息将提前赋予智能体，存于奖励中。智能体会根据卫星本身的六根数坐标和目标区域的六根数坐标作为对照，做差值 R_{punish_i} 作为负奖励，详细公式如公式(7)和公式(8)所示：

$$R_{\text{award}_i} = \alpha r_{x_i} + \beta r_{y_i} + \gamma r_{z_i} \tag{7}$$

$$\begin{cases} r_{xi} = x_{\text{target}} - x_i \\ r_{yi} = y_{\text{target}} - y_i \\ r_{zi} = z_{\text{target}} - z_i \end{cases} \tag{8}$$

其中，α 为 x 轴差值奖励的权重因子；β 为 y 轴差值奖励的权重因子；γ 为 z 轴差值奖励的权重因子。

最终总的奖励如公式(9)所示：

$$R_i = R_{\text{award}_i} + R_{\text{punish}_i} \tag{9}$$

4 实验仿真与分析

4.1 RuningMeanStd 归一化

本文实验中存在大量维度不一的数据，例如在宇宙空间中卫星可以飞离地心数千万米的距离，但是卫星的速度一般不会超过 10 000，二者明显数据维度不在一个量级，此时如将该状态作为网络的输入传入强化学习网络中，可能出现梯度爆炸或者梯度消失的现象，导致无法正常完成训练。因此，本文对输入的数据进行归一化，在归一化方法中加入了经验步长 n，并将所有归一化数据作为网络经验也进行单独存放，从而在经验回放中正确归一化，如公式(10)和公式(11)所示：

$$X_{\text{normaliztion}} = \frac{x - \mu}{\sigma} \tag{10}$$

$$\begin{cases} \mu = \mu_{\text{ord}} + \dfrac{x - \mu_{\text{ord}}}{n} \\ \sigma = \left(\dfrac{S}{n}\right)^2, (S = S_{\text{ord}} + (x - \mu_{\text{ord}}) \times (x - \mu)) \end{cases} \tag{11}$$

其中，μ 为输入数据的均值；σ 为标准差；S 为均值差比。

4.2 实验环境

本文使用 Python3 语言制作并使用基于 VPython 库的卫星运动轨道控制仿真系统，所有所需的环境底层函数均由 VPython 构建，并搭建强化学习所需的环境。

卫星运动轨道控制仿真系统结构如图 5 所示，分为强化学习主函数和环境模块。训练开始时，首先会由强化学习主函数调用初始化模块，再由初始化模块去对基础环境进行初始化；完成初始化后，强化学习主函数会输出一个一元数组作为卫星的三轴加速度动作给环境模块，由环境模块中的动作模块捕获动作进行处理，再交由状态更新模块去调用基础环境模块，处理当前环境状态更新；最后，由奖励计算模块获取所有状态和动作信息，传给强化学习主函数，完成马尔科夫过程的闭环。

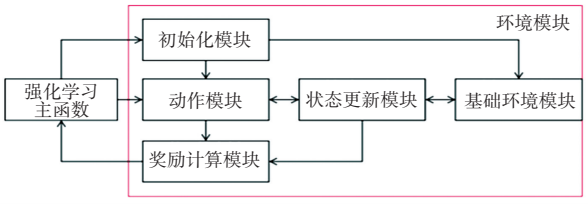

图 5 卫星运动轨道控制仿真系统结构

Fig. 5 Structure of satellite motion orbit control simulation system

4.3 实验参数

在经过多次调优实验后，最终实验中的详细参数见表 1～表 3。

表 1　奖励函数系数

Table 1　Reward function coefficient

参数名	参数数值	参数名	参数数值
a	0.001	γ	5
β	0.001	τ	5
δ	0.001		

表 2　强化学习网络系数

Table 2　Reinforcement learning network factor

参数名	参数数值	参数名	参数数值
动作集合大小	3	x 轴动作空间	[-100 , 100]
状态集合大小	8	y 轴动作空间	[-100, 100]
噪声因子	0.250 0	z 轴动作空间	[-100, 100]
衰减噪声因子	0.950 0	最大步长	500
Actor 网络学习率	0.000 1	网络宽度	256
Critic 网络学习率	0.000 1	Batch 大小	256
Max_episode	13 000	Max_steps	3 000
dt	1	tau	0.005

表 3　卫星初始参数

Table 3　Satellite initial parameters

参数名	参数数值	参数名	参数数值
长	20/m	初始 x	300 000
宽	10/m	初始 y	480 000
高	10/m	初始 z	4 551 000
重量	4 474/kg	目标距离	2 808 346/m

4.4 实验以及结果分析

卫星的初始点和重点的位置必须通过绕行才可达到，否则会碰撞地球。在本算法的网络中，Actor 网络采用了三层全连接层，每层之间选用 tanh 作为激活函数，三层全连接层的神经网络个数分别为：8、256，256、256，256、3，Actor_target 网络采用和 Actor 网络一样的结构。Critic 网络采用了 6 层全连接层，前 3 层输出 $Q1$，后 3 层输出 $Q2$，每层之间采用 Relu 作为激活函数，每 3 层全连接层的神经网络个数分别为：11、256，256、256，256、1，Critic_target 网络采用和 Critic 网络一样的结构。经过训练，如图 6 所示，在本算法设定 1e-4 学习率下与随机数方法的比较中，证明了本算法的有效性。

同时，Actor 网络在 1e-4 学习率下的训练误差如图 7 所示，在约 4×10^7 次训练后，损失率基本稳定在较低区间浮动，再次证明该算法的可行性。

学习率对于强化学习的影响很大，尤其在 TD3 算法中尤其明显。采集每 1 000 个 episode 为一组的千轮命中率如图 8 所示，可见在 3e-4、1e-4、1e-6 的学习率下，都可以在 6 000 轮 episode 逐渐学习探索到目标区域，在 12 000 轮左右收敛。其中，3e-4 的效果和 1e-4 的效果类似，均好于 1e-6 下的学习率；在 1e-4 学习率下，稳定性会略优于 3e-4。

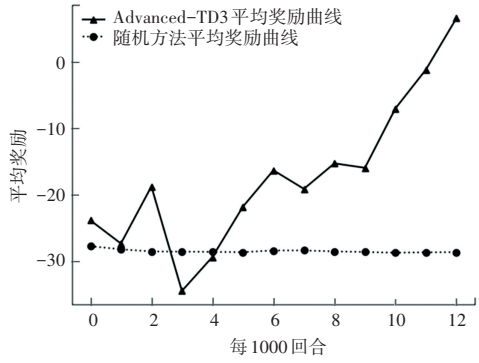

图 6　1e-4 学习率下与随机数方法奖励对比

Fig. 6　Comparison of rewards with random number methods under 1e-4 learning rate

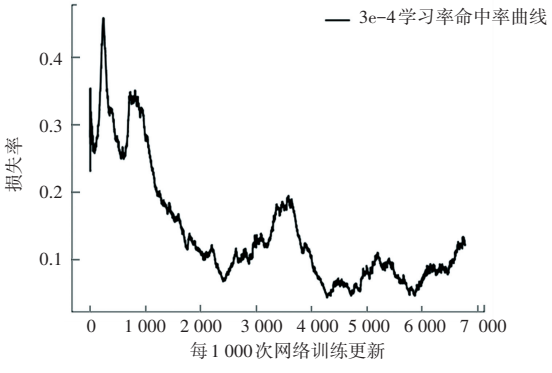

图 7　1e-4 学习率下的训练误差

Fig. 7　Training error at 1e-4 learning rate

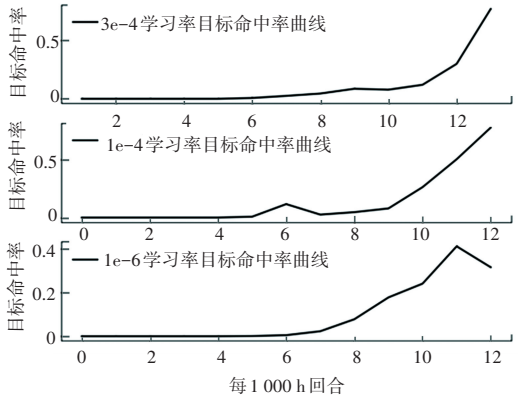

图 8　各学习率下目标命中率

Fig. 8　Target hit rate at each learning rate

为了进一步研究 Advanced-TD3 算法在各个学习率下的对比，在训练中采集了各学习率下每 1 000 轮 episode 中的平均奖励结果如图 9 所示，可见 Advanced-TD3 在该实验背景下，使用 1e-4 学习率的效果更好。

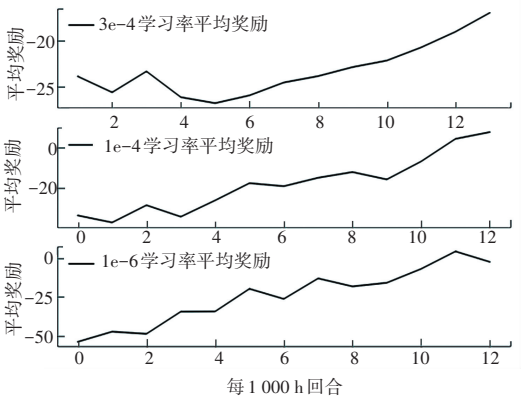

图 9 各学习率下平均奖励

Fig. 9 The average award under each study rate

5 结束语

本文针对卫星的空间探索问题,提出了基于 TD3 算法的改进 TD3(Advanced-TD3)算法,可以成功实现对卫星的控制,完成变轨到达目标区域的问题。Advanced-TD3 算法可以更好地适应未知环境中的变轨控制问题,对复杂空间拥有一定的鲁棒性。因为状态空间十分巨大,卫星的探索能力有限,因此需要对已有的学习样本进行更好地处理,因此引入了 LSTM 网络以提高训练能力。入网络的状态数据存在多个维度,差异巨大,如果直接传入网络会引发梯度消失或梯度爆炸问题,因此使用了 RuningMeanStd 归一化,同时将归一化内的参数如网络模型进行存储,方便后期的迁移学习。

参考文献

[1] OVCHINNIKOV M Y, ROLDUGIND S. A survey on active magnetic attitude control algorithms for small satellites [J]. Progress in Aerospace Sciences, 2019, 109: 100546.

[2] GOLZARI A, PISHKENARI H N, SALARIEH H, et al. Quaternion based linear time-varying model predictive attitude control for satellites with two reaction wheels [J]. Aerospace Science and Technology, 2020, 98: 105677.

[3] LIU F, YE L, LIU C, et al. Micro-thrust, low-fuel consumption, and high-precision east/west station keeping control for geo satellites based on synovial variable structure control[J]. Mathematics, 2023, 11(3): 705.

[4] ALEKSANDROV A Y, TIKHONOV A A. Application of a PID-like control to the problem of triaxial electrodynamic attitude stabilization of a satellite in the orbital frame [J]. Aerospace Science and Technology, 2022, 127: 107720.

[5] FAKOOR M, NIKPAY S, KALHORA. On the ability of sliding mode and LQR controllers optimized with PSO in attitude control of a flexible 4-DOF satellite with time-varying payload [J]. Advances in Space Research, 2021, 67(1): 334-349.

[6] VAFAMANDN. Adaptive robust neural-network-based backstepping control of tethered satellites with additive stochastic noise [J]. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(5): 3922-3930.

[7] XIAN B, WANG S, YANG S. Nonlinear adaptive control for an unmanned aerial payload transportation system: theory and experimental validation [J]. Nonlinear Dynamics, 2019, 98: 1745-1760.

[8] 许轲,吴凤鸽,赵军锁.基于深度强化学习的软件定义卫星姿态控制算法[J].北京航空航天大学学报,2018,44(12):2651-2659.

[9] ZHOU J, XUE S, XUE Y, et al. A novel energy management strategy of hybrid electric vehicle via an improved TD3 deep reinforcement learning[J]. Energy, 2021, 224: 120118.

[10] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]//Proceedings of International Conference on Machine Learning. PMLR, 2018: 1587-1596.

[11] 苗峻,涂歆滢,殷建丰,等. 基于 ADDPG 策略的超立方体卫星编队控制[J].中国空间科学技术,2023,43(4):24-34.

[12] SHI Z, ZHAO F, WANG X, et al. Satellite attitude tracking control of moving targets combining deep reinforcement learning and predefined-time stability considering energy optimization[J]. Advances in Space Research, 2022, 69(5): 2182-2196.

[13] 刘冰雁,叶雄兵,王新波,等. 基于分支深度强化学习的非合作目标追逃博弈策略求解[J].航空学报,2020,41(10):348-358.

[14] LIU H, CHEN Z, WANG X, et al. Optimal formation control for multiple rotation-translation coupled satellites using reinforcement learning[J]. Acta Astronautica, 2023, 204: 583-590.

[15] ZHANG Z B, LI X H, AN J P, et al. Model-free attitude control of spacecraft based on PID-guide TD3 algorithm[J]. International Journal of Aerospace Engineering, 2020, 2020: 1-13.

[16] QU X, GAN W, SONG D, et al. Pursuit-evasion game strategy of USV based on deep reinforcement learning in complex multi-obstacle environment [J]. Ocean Engineering, 2023, 273: 114016.

[17] GAO D, ZHANG H, LI C, et al. Satellite attitude control with deep reinforcement learning [C]//Proceedings of 2020 Chinese Automation Congress (CAC). IEEE, 2020: 4095-4101.

[18] CHU Z, SUN B, ZHU D, et al. Motion control of unmanned underwater vehicles via deep imitation reinforcement learning algorithm[J]. IET Intelligent Transport Systems, 2020, 14(7): 764-774.

[19] SHI Z, ZHAO F, WANG X, et al. Satellite attitude tracking control of moving targets combining deep reinforcement learning and predefined-time stability considering energy optimization[J]. Advances in Space Research, 2022, 69(5): 2182-2196.