Journal of System Simulation

Volume 34 | Issue 10 Article 16

10-18-2022

Reinforcement-learning-based Adaptive Tracking Control for a Space Continuum Robot Based on Reinforcement Learning

Da Jiang
1.Dalian University of Technology, Dalian 116024, China;, ziangdar@sina.com

Zhiqin Cai
1.Dalian University of Technology, Dalian 116024, China;, zhqcai@dlut.edu.cn

Zhongzhen Liu
1.Dalian University of Technology, Dalian 116024, China;

Haijun Peng
1.Dalian University of Technology, Dalian 116024, China;2.State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian 116024, China;

See next page for additional authors

Reinforcement-learning-based Adaptive Tracking Control for a Space Continuum Robot Based on Reinforcement Learning

Abstract

Abstract: Aiming at the tracking control for three-arm space continuum robot in space active debris removal manipulation, *an adaptive sliding mode control algorithm based on deep reinforcement learning is proposed. Through BP network, a data-driven dynamic model is developed as the predictive model to guide the reinforcement learning to adjust the sliding mode controller's parameters online, and finally realize a real-time tracking control.* Simulation results show that the proposed data-driven predictive model can accurately predict the robot's dynamic characteristics with the relative error within ±1% to random trajectories. Compared with the fixed-parameter sliding mode controller, the proposed adaptive controller has a lower overshoot and shorter settling time and can achieve a better tracking performance.

Keywords

space continuum robot, reinforcement learning, predictive control, sliding mode control, trajectory tracking

Authors

Da Jiang, Zhiqin Cai, Zhongzhen Liu, Haijun Peng, and Zhigang Wu

Recommended Citation

Da Jiang, Zhiqin Cai, Zhongzhen Liu, Haijun Peng, Zhigang Wu. Reinforcement-learning-based Adaptive Tracking Control for a Space Continuum Robot Based on Reinforcement Learning[J]. Journal of System Simulation, 2022, 34(10): 2264-2271.

第 34 卷第 10 期 系统仿真学报© Vol. 34 No. 10
2022 年 10 月 Journal of System Simulation Oct. 2022

基于强化学习的连续型机械臂自适应跟踪控制

江达[1]，蔡志勤[1*]，刘忠振[1]，彭海军[1,2]，吴志刚[2]
(1. 大连理工大学，辽宁 大连 116024；2. 工业装备结构分析国家重点实验室，辽宁 大连 116024)

摘要：针对空间主动碎片清除操作中连续型三臂节机器人系统跟踪问题，*提出一种基于强化学习的自适应滑模控制算法。基于数据驱动的建模方法，采用BP神经网络对三臂节连续型机械臂进行建模，并作为预测模型指导强化学习实时调节所提出滑模控制器的控制参数，从而实现连续型机器人运动的实时跟踪控制。*仿真结果表明：提出的数据驱动的预测模型对随机轨迹预测的相对误差保持在±1%以内，能够高精度地反映系统动态特性。对比固定参数的滑模控制器，提出的自适应控制器在保证系统达到控制目标的同时具有更低的超调量和更短的调节时间，表现出更好的控制效果。

关键词：空间连续型机器人；强化学习；预测控制；滑模控制；轨迹跟踪
中图分类号：TP273.2 文献标志码：A 文章编号：1004-731X(2022)10-2264-08
DOI: 10.16182/j.issn1004731x.joss.21-0632

Reinforcement-learning-based Adaptive Tracking Control for a Space Continuum Robot Based on Reinforcement Learning

Jiang Da[1], *Cai Zhiqin*[1*], *Liu Zhongzhen*[1], *Peng Haijun*[1,2], *Wu Zhigang*[2]
(1. Dalian University of Technology, Dalian 116024, China;
2. State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian 116024, China)

Abstract: Aiming at the tracking control for three-arm space continuum robot in space active debris removal manipulation, *an adaptive sliding mode control algorithm based on deep reinforcement learning is proposed*. *Through BP network, a data-driven dynamic model is developed as the predictive model to guide the reinforcement learning to adjust the sliding mode controller's parameters online, and finally realize a real-time tracking control.* Simulation results show that the proposed data-driven predictive model can accurately predict the robot's dynamic characteristics with the relative error within ±1% to random trajectories. Compared with the fixed-parameter sliding mode controller, the proposed adaptive controller has a lower overshoot and shorter settling time and can achieve a better tracking performance.

Keywords: space continuum robot; reinforcement learning; predictive control; sliding mode control; trajectory tracking

引言

持续增长的空间碎片对在轨航天器构成了重大的威胁，空间环境的不断恶化已经成为当今国际社会面临的全球性挑战。随着空间探索的不断深入，空间主动碎片清除技术的重要性愈发显著。相比刚性机械臂，连续型机械臂具有占用空间小，柔软灵活等特点，可以通过主动变形在有限的工作空间内完成复杂的动作，将末端执行器在复杂

收稿日期：2021-07-07 修回日期：2021-09-12
基金项目：国家自然科学基金重大研究计划重点项目(91748203)；国家自然科学基金优秀青年项目(11922203)
第一作者：江达(1992-)，男，博士生，研究方向为空间机器人动力学与控制。E-mail：ziangdar@sina.com
通讯作者：蔡志勤(1961-)，女，博士，教授，研究方向为空间机器人动力学与控制。E-mail：zhqcai@dlut.edu.cn

的空间环境中安全稳定地送至抓捕点附近，在空间主动碎片清除任务中表现出广阔的应用前景。

然而，与刚性机器人相比，自身柔软灵活的特点使其呈现出高度非线性的动力学特征，这使得动力学模型建立及控制器设计非常困难，在参数摄动、外部干扰等不确定因素下，最终的控制精度会进一步降低。因此，合理且有效地建模和控制十分必要。为进一步提高控制效果，近年来大量基于神经网络的建模及控制方法被应用到连续型机器人研究上。文献[1-2]采用了前馈神经网络分别拟合连续型机械臂的正逆运动学模型，达到了较高的精度。Thuruthel 等分别采用前馈神经网络[3]、递归神经网络[4]学习连续型机械臂的动力学模型，用以拟合机械臂的动态响应并进行评估，并据此构建开环控制策略。此类需要大量的监督数据，并限制了机器人的运动轨迹。虽然神经网络的引入使得连续型机器人此类具有高度非线性特征模型的建模过程更加简单，但由于拟合模型的精度依赖于监督数据的完备性，导致模型仍不可避免的会受到过拟合问题的影响。同时开环控制策略进一步限制了控制精度及应用场景，需要合理地嵌入到执行器控制中，来保证全局控制策略的稳定性。

对于具有强非线性动力学特征的机器人跟踪控制问题，模型预测控制方法受到了广泛的认可。Li 等[5]提出了一种机器人运动规划网络 MPC-MPNet，经监督数据训练后，网络生成满足动力学约束的可行路径，再由模型预测控制实现局部转向的避障功能，该算法只在规划过程中执行正向路径扩展，不适合在动态障碍物环境中进行实时规划；Ouyang 等[6]提出了一种具有指数加权预测范围的模型预测控制器，通过建立接触过程中机器人驱动空间和变形空间的线性近似模型，来实现在接触力作用下的连续型机器人主动变形控制，该方法控制效果依赖接触变形近似模型的精度；Tang 等[7]针对连续型机器人提出了一种迭代学习模型预测控制方法，通过伪刚体模型对执行器的变形进行初步预测，利用迭代学习不断降低模型误差，最后由模型预测控制实现机器人变形，该方法适用于具有一定刚度的软管式连续型机器人，对弯曲特性较明显的连续型机器人拓展性不强。上述方法可以看出良好的动态模型预测精度直接影响模型预测控制的效果，然而依赖监督数据的神经网络控制器需要完备的数据样本，且容易陷入局部最优解，此类方法不具有探索外部环境的能力，无法实时对外部反馈信息进行进一步处理，不能直接大规模拓展到连续型机器人控制中。

为进一步提高预测模型对控制策略的引导效果，提升环境探索能力。学者们提出将深度强化学习引入模型预测控制中滚动优化的奖励策略内，以提升控制精度，此类混合控制方法正成为机器人控制的新趋势。Frazelle 等[8]采用 Actor-Critic 框架的策略搜索方法对连续型机器人进行运动学控制，该方法采用离散空间下的状态和动作框架，限制了控制精度，导致该方法不适用于被控模型更为复杂的动力学控制问题；Shin 等[9]针对外科手术过程中手术机械臂与软组织接触时的控制问题，采用神经网络学习视觉空间下软组织受力时的动力学模型，并预测其动态响应，然后由基于模型预测控制的强化学习来对机械臂进行操纵，该方法的控制精度同样受离散动作训练模型和演示数据数量的影响；Thuruthel 等进一步拓展了文献[3-4]工作，提出了一种基于模型的连续型机器人机械手闭环预测控制的策略学习算法[10]，采用递归神经网络拟合前向动力学模型，采用引导策略搜索的强化学习算法进行轨迹优化，然后由滚动优化和监督学习推导出闭环策略。然而随机打靶法进行轨迹采样的方法需要大量的数据，且不具有完备性，无法在大范围跟踪运动控制中获得合理解。

对于具有强非线性动力学模型的空间连续型机器人，神经网络运算速度快，训练方便，可以结合实际数据训练形成具有一定反映系统动态的预测模型，并具有较好的迁移能力。但对于空间碎片清除任务中对精度具有较高要求的执行机构

DOI: 10.16182/j.issn1004731x.joss.21-0632

跟踪控制器，仍需要借助数学模型建造基于模型的动力学控制器。基于此，本文结合经验模型和数值模型优势，根据模型预测控制原理，提出一种数据驱动的多层前馈神经网络模型，将该模型用于非线性系统建模，预测系统在当前控制策略下的预期动态响应。然后考虑外部干扰和建模误差的情况下，设计变结构控制器，并在双延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient，TD3)算法的基础上，引入模型预测控制原理，根据系统长期性能表现，在连续动作空间内实时优化滑模控制器参数，从而进一步提升控制器控制能力，最终实现空间主动碎片清除操作中连续型机械臂的自适应鲁棒控制。

1 空间连续型机器人动力学模型

针对本文研究的空间主动碎片清除场景，假设机器人系统已被送至碎片附近，并进行了初步位姿调整使得连续型机械臂已进入碎片的可抓捕范围内，机械臂末端装有用于识测的传感器及带有抓捕功能的末端执行器。提出的平面三臂节空间连续型机器人以柔性支撑梁为骨架，每个臂节各个模块处置有固定节盘，初始节盘处置有电机，通过穿插的驱动线来驱动整个机械臂的运动。针对此模型需要作出如下假设：①节盘与驱动线之间光滑无摩擦；②柔性支撑处无外部碰撞情况；③各臂节变形服从等曲率假设。相邻两臂节构型关系如图1所示。

基于如上假设，采用广义坐标$\boldsymbol{q}=(\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$描述机器人运动，其中连续型机器人几何位形为平面运动。其中β_i和(x_i, y_i)分别为当前臂节$i(i=1, 2, 3)$对应的局部坐标系$O_iX_iY_i$相对于全局坐标系OXY的转角和坐标，α_i为臂节i的弯曲变形角度，则对于臂节i内任意点p_i对应的弦O_ip_i，其对应弯曲角度为$\varphi_i(\varphi_i\in[0, \alpha_i])$。同理$\alpha_{i+1}$，$\beta_{i+1}$和$(x_{i+1}, y_{i+1})$为局部坐标系$O_{i+1}X_{i+1}Y_{i+1}$中的对应项。

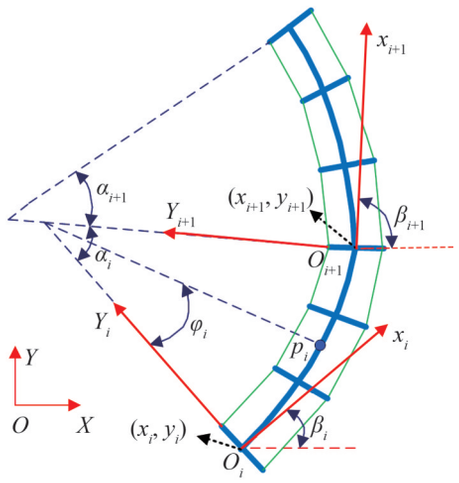

图1 连续型空间机器人构型

Fig. 1 Continuum space robot configuration

本文采用文献[11]描述的空间连续型机器人动力学模型。基于此，机器人系统动能T为

$$T = T^{\mathrm{d}} + T^{\mathrm{s}} = (1/2)\dot{\boldsymbol{q}}^{\mathrm{T}}\boldsymbol{M}\dot{\boldsymbol{q}} \tag{1}$$

式中：T^{d}和T^{s}分别为节盘动能和柔性支撑动能；$\boldsymbol{M}$为机器人系统的质量矩阵；$\dot{\boldsymbol{q}}$为$\boldsymbol{q}$的一阶导数。

连续型机器人系统弹性力$\boldsymbol{Q}_{\mathrm{e}}$对应的虚功为

$$\delta W_{\mathrm{e}} = -\int_0^l\iint_A E\varepsilon\delta\varepsilon\mathrm{d}A\mathrm{d}s = -\boldsymbol{Q}_{\mathrm{e}}^{\mathrm{T}}\delta\boldsymbol{q} \tag{2}$$

式中：E为柔性支撑的弹性模量；A为其截面面积；l为其长度；ε为中性层的弯曲应变；s为局部坐标系下弹性力作用点到原点的弧长。

连续型机器人系统驱动力$\boldsymbol{Q}_{\mathrm{a}}$对应的虚功为

$$\delta W_{\mathrm{a}} = \boldsymbol{Q}_{\mathrm{a}}^{\mathrm{T}}\delta\boldsymbol{q} \tag{3}$$

由此可得空间连续型机器人系统动力学方程：

$$\boldsymbol{M}\ddot{\boldsymbol{q}} = -\boldsymbol{Q}_{\mathrm{e}} + \boldsymbol{Q}_{\mathrm{a}} + \boldsymbol{Q}_{\mathrm{v}} \tag{4}$$

式中：$\boldsymbol{Q}_{\mathrm{v}} = -\dot{\boldsymbol{M}}\dot{\boldsymbol{q}} + (\partial T/\partial\boldsymbol{q})^{\mathrm{T}}$，$\dot{\boldsymbol{M}}$为质量阵的导数；$\ddot{\boldsymbol{q}}$为$\boldsymbol{q}$的二阶导数。为了更好地表示驱动力向量$\boldsymbol{\tau}=[\tau_1, \tau_2, \tau_3]^{\mathrm{T}}$，并考虑模型参数摄动和空间环境下的外部扰动，将式(4)整理为

$$\boldsymbol{M}_0(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}_0(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \boldsymbol{\tau} + \boldsymbol{f}(t) \tag{5}$$

式中：$\boldsymbol{M}_0(\boldsymbol{q})\in\mathbb{R}^{3\times3}$，$\boldsymbol{C}_0(\boldsymbol{q}, \dot{\boldsymbol{q}})\in\mathbb{R}^{3\times1}$；$\boldsymbol{f}(t)$为表征外部干扰$d(t)$和建模误差$\Delta\boldsymbol{M}_0\ddot{\boldsymbol{q}} + \Delta\boldsymbol{C}_0\boldsymbol{q}$的列向量

$$\boldsymbol{f}(t) = d(t) + \Delta\boldsymbol{M}_0\ddot{\boldsymbol{q}} + \Delta\boldsymbol{C}_0\boldsymbol{q} \tag{6}$$

由此，则可以通过驱动线改变各臂节变形，

进而控制整个机械臂的运动。

2 空间连续型机器人滑模控制器

定义系统的角度跟踪误差向量为$\boldsymbol{e}=\boldsymbol{q}-\boldsymbol{q}_{\mathrm{d}}$，其中，$\boldsymbol{q}_{\mathrm{d}}$为期望轨迹向量；系统角速度跟踪误差向量为$\dot{\boldsymbol{e}}=\dot{\boldsymbol{q}}-\dot{\boldsymbol{q}}_{\mathrm{d}}$。假设期望轨迹$\boldsymbol{q}_{\mathrm{d}}$及其导数$\dot{\boldsymbol{q}}_{\mathrm{d}}$、$\ddot{\boldsymbol{q}}_{\mathrm{d}}$连续有界。定义滑模面如下：

$$\boldsymbol{h}=\dot{\boldsymbol{e}}+\delta\boldsymbol{e} \tag{7}$$

式中：δ是正常数。由局部线性化原理可得：

$$\dot{\boldsymbol{h}}=\ddot{\boldsymbol{e}}+\delta\dot{\boldsymbol{e}}=\boldsymbol{M}_0^{-1}(\boldsymbol{\tau}+\boldsymbol{f}-\boldsymbol{C}_0\boldsymbol{q})-\ddot{\boldsymbol{q}}_{\mathrm{d}}+\delta\dot{\boldsymbol{e}} \tag{8}$$

为了保证闭环系统的运动在有限的时间内到达滑模面，设计滑模控制趋近律如下：

$$\dot{\boldsymbol{h}}=-k\mathrm{sgn}(\boldsymbol{h})-\delta\boldsymbol{h} \tag{9}$$

式中：k是正常数。将式(9)代入式(8)，可得：

$$\boldsymbol{\tau}=\boldsymbol{M}_0[\ddot{\boldsymbol{q}}_{\mathrm{d}}-2\delta\dot{\boldsymbol{e}}-\delta^2\boldsymbol{e}-k\mathrm{sgn}(\dot{\boldsymbol{e}}+\delta\boldsymbol{e})]+\boldsymbol{C}_0\boldsymbol{q}-\boldsymbol{f}$$

定义$\boldsymbol{f}_{\mathrm{c}}$为$\boldsymbol{f}$的估计向量，则本文提出的基于滑模控制跟踪控制律可写为

$$\boldsymbol{\tau}=\boldsymbol{M}_0[\ddot{\boldsymbol{q}}_{\mathrm{d}}-2\delta\dot{\boldsymbol{e}}-\delta^2\boldsymbol{e}-k\mathrm{sgn}(\dot{\boldsymbol{e}}+\delta\boldsymbol{e})]+\boldsymbol{C}_0\boldsymbol{q}-\boldsymbol{f}_{\mathrm{c}} \tag{10}$$

将式(10)代入式(9)，可得：

$$\dot{\boldsymbol{h}}=-k\mathrm{sgn}(\boldsymbol{h})-\delta\boldsymbol{h}-\boldsymbol{M}_0^{-1}(\boldsymbol{f}-\boldsymbol{f}_{\mathrm{c}})$$

选择Lyapunov函数如下：

$$V=(1/2)\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h} \tag{11}$$

对式(11)求取一阶导数，则有：

$$\dot{V}=\boldsymbol{h}^{\mathrm{T}}\dot{\boldsymbol{h}}=-\boldsymbol{h}^{\mathrm{T}}k\mathrm{sgn}(\boldsymbol{h})-\delta\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}-\boldsymbol{h}^{\mathrm{T}}\boldsymbol{M}_0^{-1}(\boldsymbol{f}-\boldsymbol{f}_{\mathrm{c}})$$

定义$\bar{\boldsymbol{f}}_{\mathrm{c}}=\boldsymbol{M}_0^{-1}\boldsymbol{f}_{\mathrm{c}}$，$\bar{\boldsymbol{f}}=\boldsymbol{M}_0^{-1}\boldsymbol{f}$，假设$\bar{\boldsymbol{f}}$有界：$\bar{\boldsymbol{f}}_{\mathrm{l}}\leqslant\bar{\boldsymbol{f}}\leqslant\bar{\boldsymbol{f}}_{\mathrm{u}}$

通过选取$\bar{\boldsymbol{f}}_{\mathrm{c}}=(\bar{\boldsymbol{f}}_{\mathrm{u}}+\bar{\boldsymbol{f}}_{\mathrm{l}})/2-(\bar{\boldsymbol{f}}_{\mathrm{u}}-\bar{\boldsymbol{f}}_{\mathrm{l}})\mathrm{sgn}(\boldsymbol{h})/2$，有：

$$\bar{\boldsymbol{f}}-\bar{\boldsymbol{f}}_{\mathrm{c}}=\bar{\boldsymbol{f}}-[(\bar{\boldsymbol{f}}_{\mathrm{u}}+\bar{\boldsymbol{f}}_{\mathrm{l}})/2-(\bar{\boldsymbol{f}}_{\mathrm{u}}-\bar{\boldsymbol{f}}_{\mathrm{l}})\mathrm{sgn}(\boldsymbol{h})/2)]=\begin{cases}\bar{\boldsymbol{f}}-\bar{\boldsymbol{f}}_{\mathrm{l}}>0, & \boldsymbol{h}>0\\ \bar{\boldsymbol{f}}-\bar{\boldsymbol{f}}_{\mathrm{u}}<0, & \boldsymbol{h}<0\end{cases}$$

则$\boldsymbol{h}^{\mathrm{T}}\boldsymbol{M}_0^{-1}(\boldsymbol{f}-\boldsymbol{f}_{\mathrm{c}})>0$。由$-\boldsymbol{h}^{\mathrm{T}}k\mathrm{sgn}(\boldsymbol{h})-\delta\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}\leqslant0$最终可得$\dot{V}<0$。

3 基于强化学习的滑模控制器

考虑由状态向量$\boldsymbol{s}$，动作向量$\boldsymbol{a}$，状态转移概率分布$\boldsymbol{p}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)$和奖励函数$r$组成的标准马尔可夫过程。强化学习旨在通过不断学习来最大化长期期望奖励$\boldsymbol{R}(\boldsymbol{s},\boldsymbol{a})=\sum_{i=t}^{T}\gamma^{i-t}r_t(\boldsymbol{s}_t,\boldsymbol{a}_t)$，以此获得最优策略$\pi^{\theta}$，其中$r_t(\boldsymbol{s}_t,\boldsymbol{a}_t)$表示在状态向量$\boldsymbol{s}_t$和动作向量$\boldsymbol{a}_t$下的奖励值，$\gamma$为折扣系数。动作值函数为

$$Q(\boldsymbol{s}_t,\boldsymbol{a}_t)=\mathrm{E}[R_t|\boldsymbol{s}=\boldsymbol{s}_t,\boldsymbol{a}=\boldsymbol{a}_t]=\mathrm{E}[\sum_{i=t}^{T}\gamma^{i-t}r_t(\boldsymbol{s}_t,\boldsymbol{a}_t)]$$

TD3算法是典型的基于Actor-Critic框架算法[12]，主网络包含2个由θ^{Qk}(k=1, 2)参数化的Critic网络$Q(\boldsymbol{s},\boldsymbol{a}|\theta^{Qk})$，以及一个由$\theta^{\mu}$参数化的Actor网络$\mu(\boldsymbol{s}|\theta^{\mu})$。并以惩罚系数$\rho$通过滑动平均法[13]更新目标网络参数：

$$\theta'=\rho\theta+(1-\rho)\theta'$$

为避免Actor-Critic框架算法普遍存在的过拟合问题导致的次优动作策略，更新过程中TD3算法始终选取2个Critic网络中的最小值，并以t_{d}的间隔进行延迟策略更新。对于经验池存储的随机N_{b}组数据，每个Critic网络各自的代价函数如下：

$$L=(1/N_{\mathrm{b}})\sum_{j=1}^{N_{\mathrm{b}}}[y_{Qj}^{k}-Q^{k}(\boldsymbol{s}_j,\boldsymbol{a}_j|\theta^{Qk})]^2 \tag{12}$$

$$y_{Qj}^{k}=r_j+\gamma\min[Q^{k'}(\boldsymbol{s}_{j+1},\varepsilon+\boldsymbol{\mu}'(\boldsymbol{s}_{j+1}|\theta^{\mu'})|\theta^{Qk'})]$$

$$\varepsilon\sim\mathrm{clip}(N(0,\sigma),-c,c)$$

其中，ε服从高斯噪声$N(0,\sigma)$，TD3算法通过在迭代过程中引入随机噪声来进一步增加智能体探索环境的能力。Actor网络由策略梯度算法更新：

$$\nabla_{\theta^{\mu}}J\approx(1/N_{\mathrm{b}})\sum_{i=1}^{N_{\mathrm{b}}}\nabla_{\boldsymbol{a}}Q(\boldsymbol{s},\boldsymbol{a}|\theta^{Q})|_{\boldsymbol{s}_i,\mu(\boldsymbol{s}_i)}\nabla_{\theta^{\mu}}\boldsymbol{\mu}(\boldsymbol{s}|\theta^{\mu})|_{\boldsymbol{s}_i}$$

算法1 MPC-TD3算法

初始化Actor/Critic主网络、目标网络及经验池。

for episode=1 to M_E do

初始化动作探索噪声、获得初始状态s_1。

for t=1 to T do

依策略网络选择动作a_t并据此获得r_t及s_{t+1}。

将数据(s_t, a_t, r_t, s_{t+1})存入经验池。

从经验池中进行采样并据此更新 Critic 网络。

if t mod t_d then

更新 Actor 网络及全部目标网络。

end

本文研究的自适应滑模控制器，不能直接采用强化学习进行自适应参数调整。因为对于任意轨迹的跟踪问题，传统强化学习的动作策略无法在短期内表现出明显的奖励差异，导致网络难以据此获得有效的策略输出。其次，在每个时间步，不适合频繁调用此类非线性系统动力学方程，容易造成计算负担，不适用于实际场景。因此本文引入了数据驱动的学习方法，该方法计算量小，不需要精确的动力学模型信息，可针对不同的环境采用对应的数据进行训练，具有良好的环境实时交互和迁移能力，适用于仿真计算及地面实验。基于此，本文引入预测模型来进一步考虑长期策略收益，通过搭建 BP 神经网络[14]预测连续型机器人在当前控制策略下的动态响应，以计算预测域内的预测奖励，再由滚动优化实时调节滑模控制器的控制参数。从基于趋近律的滑模控制角度来看，适当调整趋近速度参数δ和符号函数增益参数k能够优化系统向滑模面的趋近速度和克服摄动/外部扰动的能力，从而改善控制系统动态性能，因此本文选用δ和k作为强化学习的自适应优化参数。本文所提出的基于模型预测控制的双延迟确定性策略梯度算法(model predictive control-based twin delayed deep deterministic policy gradient, MPC-TD3)如算法 1，算法更新如图 2 所示。

4 仿真校验

如图 1 所示的三臂节空间连续型机械臂系统控制仿真中，机械臂的节盘质量m^d和半径r^d分别为 0.01 kg 和 0.02 m；柔性支撑长度l和质量m^s分别为 0.2 m 和 1 kg；转动惯量$I_z = 7.85 \times 10^{-9}$ kg · m^2；弹性模量$E = 6.2 \times 10^9$ Pa。所提出的基于 MPC-TD3 的滑模控制器控制流程如图 3 所示。其中，Actor、Critic 和预测网络均采用 3 层架构的 BP 神经网络，中间层包含 100 个节点，随机初始化网络参数。

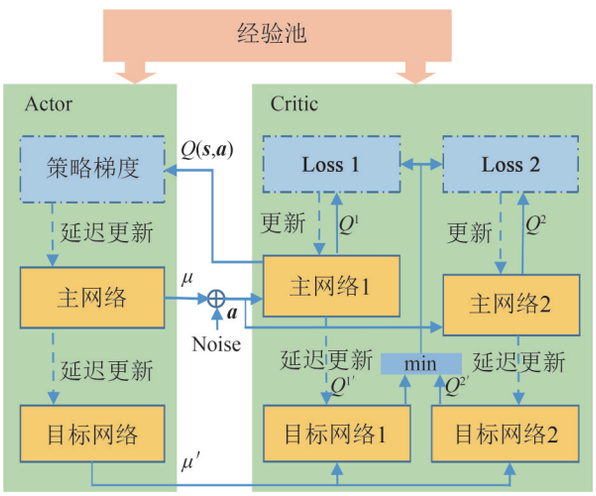

图 2 算法更新示意图

Fig. 2 Update sketch of MPC-TD3 algorithm

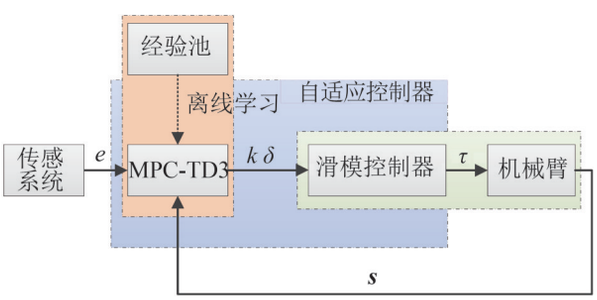

图 3 基于 MPC-TD3 的自适应滑模控制器流程

Fig. 3 Control flow of the MPC-based adaptive sliding mode controller

合理地选取网络状态信息是良好控制的保证，过度冗余的网络输入会导致网络输出对系统状态变化不敏感，降低学习网络的性能；而关键输入信息的缺失则导致网络不能有效地反映系统动态变化。本文中动作向量$\boldsymbol{a}$选取为滑模控制器的控制参数k和δ。状态向量$\boldsymbol{s}$信息包含各节角度、角速度、角度跟踪误差、角速度跟踪误差及下一时刻预期角度、角速度、角加速度，来合理地表征机械臂系统与目标轨迹的动态信息。奖励函数r选取为

$$r = -d_e + h_{goal}$$

式中：d_e为跟踪误差向量$\boldsymbol{e}$的二范数；h_{goal}为持续达到预期跟踪效果时的额外奖励值。文中设置当

控制器稳定跟踪期望轨迹，保持跟踪误差小于0.02 rad超过25个时间步长，获得额外附加奖励值，文中设置为1。

仿真过程中，预测网络输入包括当前时刻控制力向量$\boldsymbol{\tau}$及各臂节角度、角速度，输出为下一时刻各臂节角度、角速度。选取由式(5)随机生成的3 000组数据作为训练样本，外部干扰和建模误差$\boldsymbol{f}$由高斯噪声$N(0, 0.001)$表征。训练持续300代，网络训练过程中的代价值如图4所示。可以看出，设计的预测模型训练过程中代价值快速下降，经70代训练后已趋于稳定。在测试集中选取随机轨迹进行测试，三臂节角度的相对误差变化如图5所示。仿真结果表明：所设计的基于BP网络的预测模型可以将拟合的各臂节弯曲角度的相对误差保持在±1%以内，验证了该预测模型的准确性。

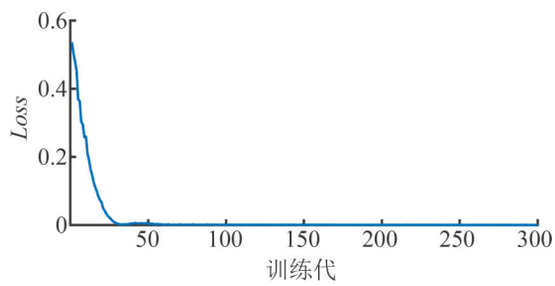

图4 预测模型训练过程中的代价值

Fig. 4 Loss value in predict model's training process

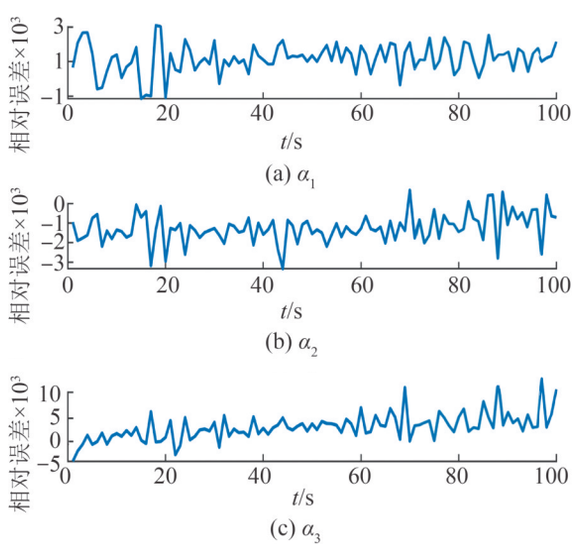

图5 预测模型测试过程中关于角度的相对误差

Fig. 5 Relative error of manipulators' bending angle in predict model's testing process

将该预测模型嵌入到基于TD3算法的滑模控制器中，滚动优化滑模控制器控制参数(k, δ)。强化学习训练持续300代，每代随机生成期望轨迹和机械臂初始状态，步长为100步，前10代数据进入经验池。预测域为5，外部干扰和建模误差$\boldsymbol{f}$同样选取为高斯噪声$N(0, 0.001)$。为验证所提出控制器的跟踪效果，随机选取三臂节弯曲角度的初始值为(0.2π rad, 0.15π rad, 0.17π rad)。在机械臂抓捕空间碎片的作业过程中，假设控制系统已通过逆运动学求解器生成了抓捕空间碎片对应的各臂节所需角度，则空间碎片抓捕问题则转化为机械臂跟踪该轨迹的跟踪控制问题。本文随机选取期望轨迹$\boldsymbol{q}_{\mathrm{d}} = (q_{\mathrm{d}_1}, q_{\mathrm{d}_2}, q_{\mathrm{d}_3})^{\mathrm{T}}$如式(13)：

$$\begin{cases} q_{\mathrm{d}_1} = (\pi/9)\sin[(\pi/9)t] + \pi/4 \\ q_{\mathrm{d}_2} = (\pi/9)\sin[(\pi/9)t] + \pi/5 \\ q_{\mathrm{d}_3} = (\pi/9)\sin[(\pi/9)t] + \pi/4 \end{cases} \tag{13}$$

由于每代训练过程中均对动作进行随机探索并随机初始化环境，为合理地展现训练效果，本文在每代训练结束后对同一轨迹进行策略评估，图6展示了每代训练后跟踪同一轨迹时的平均奖励值变化，可以看出，在受建模误差和外部干扰$\boldsymbol{f}$影响下，每代评估时的平均奖励值在逐步提升，策略在逐步优化。需要说明的是，由于$\boldsymbol{f}$的影响，在每代中的每个step对应的奖励值，尤其是取得额外奖励的h_{goal}的时间会存在差异，导致平均奖励值会存在小幅震荡，但由图可知，约经过180代训练后，稳定到达目标跟踪效果，持续获得额外奖励h_{goal}，训练效果趋于稳定。

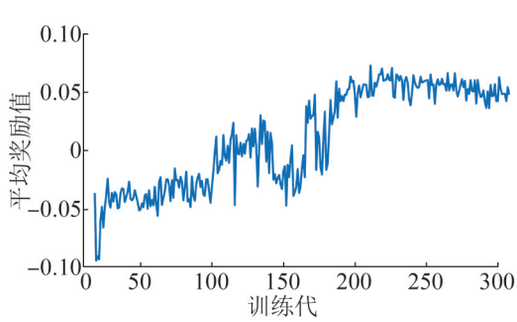

图6 每代训练后策略评估的平均奖励值

Fig. 6 Average reward in policy evaluation after each episode

DOI: 10.16182/j.issn1004731x.joss.21-0632

图 7 对比了本文提出的基于 MPC-TD3 的滑模控制器和固定参数滑模控制器[15]$(k,\delta)=(0.08,4.5)$的跟踪效果。各臂节控制力矩限制在 3 N 内，采样步长为 0.1 s。可以看出，2 种控制器都能顺利完成跟踪任务,但所提控制器通过根据系统状态自适应调节控制参数，约于 1.7 s 时便达到稳定跟踪状态。图 8 直观地描述了基于 MPC-TD3 滑模控制器下的连续型机械臂的跟踪效果，由该图可知随着时间变化，受控机械臂逐渐跟踪上期望轨迹。综上可以看出，相比固定参数滑模控制器，本文提出的控制器明显具有更低的超调量和更短的调节时间。同时也说明本文算法对外部扰动和建模误差具有更强的抑制能力，体现了本文控制系统的鲁棒性。

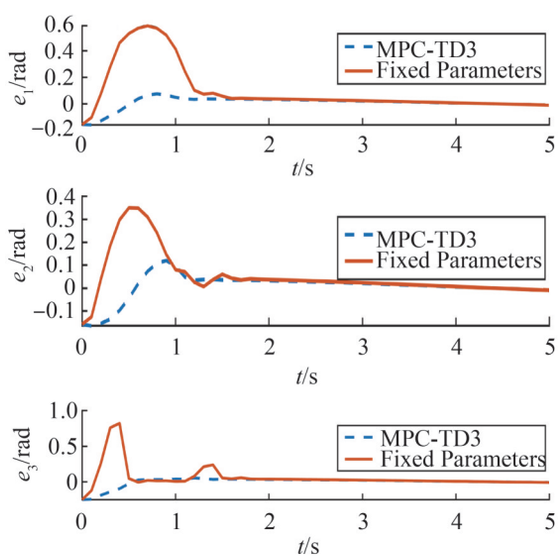

图 7　三臂节角度的跟踪误差对比

Fig. 7　Comparison for tracking error of the three arms

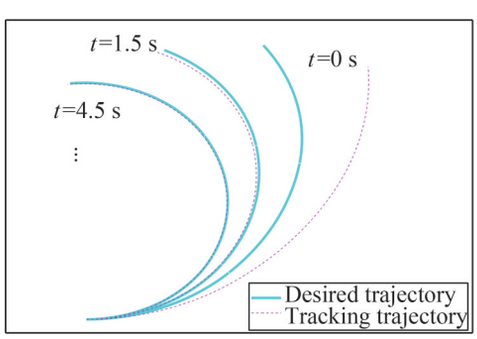

图 8　轨迹曲线

Fig. 8　Trajectory curve

5　结论

本文针对空间碎片清除操作中三臂节连续型机械臂跟踪控制问题，提出一种基于强化学习的自适应滑模控制器。在 TD3 算法的基础上，引入数据驱动的预测模型在线预测系统动态响应，并由滚动优化方法实时调整滑模控制器的控制参数。

仿真结果表明：针对受空间环境外部干扰和建模误差影响的轨迹跟踪控制问题，本文提出的预测模型对随机轨迹预测的相对误差保持在±1%以内，显示了较高精度的预测能力。同时对比固定参数滑模控制器，所提出的控制器具有更低的超调量和更快的跟踪速度，能够更有效地抑制外界干扰和建模误差带来的影响，表现出更好的控制效果，同时验证了所提出控制器的鲁棒性。未来的工作将集中在考虑摩擦和碰撞问题的地面实验验证。

参考文献：

[1] Grassmann R, Modes V, Burgner-Kahrs J. Learning the Forward and Inverse Kinematics of a 6-DOF Concentric Tube Continuum Robot in SE(3) [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York: IEEE, 2018: 5125-5132.

[2] Lai J, Huang K, Chu H K. A Learning-based Inverse Kinematics Solver for a Multi-Segment Continuum Robot in Robot-Independent Mapping[C]//2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). New York: IEEE, 2019: 576-582.

[3] Thuruthel T G, Falotico E, Manti M, et al. Stable Open-Loop Control of Soft Robotic Manipulators[J]. IEEE Robotics and Automation Letters(S2377-3766), 2018, 3 (2): 1292-1298.

[4] Thuruthel T G, Falotico E, Renda F, et al. Learning Dynamic Models for Open Loop Predictive Control of Soft Robotic Manipulators[J]. Bioinspiration & Biomimetics (S1748-3182), 2017, 12(6): 066003.

[5] Li L, Miao Y, Qureshi A H, et al. MPC-MPNet: Model-Predictive Motion Planning Networks for Fast, Near-Optimal Planning Under Kinodynamic Constraints [J]. IEEE Robotics and Automation Letters (S2377-3766), 2021, 6(3): 4496-4503.

[6] Ouyang B, Mo H, Chen H, et al. Robust Model

Predictive Deformation Control of a Soft Object by Using a Flexible Continuum Robot[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). New York: IEEE, 2018: 613-618.

[7] Tang Z Q, Heung H L, Tong K Y, et al. A Novel Iterative Learning Model Predictive Control Method for Soft Bending Actuators [C]//2019 International Conference on Robotics and Automation (ICRA). New York: IEEE, 2019: 4004-4010.

[8] Frazelle C, Rogers J, Karamouzas I, et al. Optimizing a Continuum Manipulator's Search Policy Through Model-Free Reinforcement Learning [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York: IEEE, 2020: 5564-5571.

[9] Shin C, Ferguson P W, Pedram S A, et al. Autonomous Tissue Manipulation via Surgical Robot Using Learning Based Model Predictive Control [C]//2019 International Conference on Robotics and Automation (ICRA). New York: IEEE, 2019: 3875-3881.

[10] Thuruthel T G, Falotico E, Renda F, et al. Model-Based Reinforcement Learning for Closed-Loop Dynamic Control of Soft Robotic Manipulators [J]. IEEE Transactions on Robotics (S1552-3098), 2019, 35(1): 124-134.

[11] 邱小璐, 蔡志勤, 刘忠振, 等. 空间连续型机器人自适应鲁棒容错控制[J]. 计算力学学报, 2021, 38(1): 46-50.
Qiu Xiaolu, Cai Zhiqin, Liu Zhongzhen, et al. Adaptive Robust Fault Tolerant Control of A Space Continuum Robot[J]. Chinese Journal of Computational Mechanics, 2021, 38(1): 46-50.

[12] Fujimoto S, Van H, Meger D. Addressing Function Approximation Error in Actor-Critic Methods [C]//2018 International Conference on Machine Learning(ICML). New York: PMLR, 2018: 1587-1596.

[13] Gu S, Holly E, Lillicrap T, et al. Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-policy Updates [C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). New York: IEEE, 2017: 3389-3396.

[14] Braganza D, Dawson D M, Walker I D, et al. A Neural Network Controller for Continuum Robots [J]. IEEE Transactions on Robotics (S1552-3098), 2007, 23(6): 1270-1277.

[15] Wang J, Zhou Y, Bao Y, et al. Trajectory Tracking Control Using Fractional-Order Terminal Sliding Mode Control With Sliding Perturbation Observer for a 7-DOF Robot Manipulator [J]. IEEE/ASME Transactions on Mechatronics (S1083-4435), 2020, 25(4): 1886-1893.

DOI: 10.16182/j.issn1004731x.joss.21-0632