中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于强化学习的挖掘机时间最优轨迹规划

张韵悦, 孙志毅, 孙前来, 王银

引用本文:
张韵悦,孙志毅,孙前来,王银. 基于强化学习的挖掘机时间最优轨迹规划[J]. 控制与决策, 2024, 39(5): 1433−1440.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0811

您可能感兴趣的其他文章

Articles you may be interested in

基于MCPDDPG的智能车辆路径规划方法及应用

The method and application of intelligent vehicle path planning based on MCPDDPG

控制与决策. 2021, 36(4): 835−846 https://doi.org/10.13195/j.kzyjc.2019.0460

移动机器人运动规划中的深度强化学习方法

Deep reinforcement learning for motion planning of mobile robots

控制与决策. 2021, 36(6): 1281−1292 https://doi.org/10.13195/j.kzyjc.2020.0470

基于Frenet坐标系的自动驾驶轨迹规划与优化算法

Trajectory planning and optimization algorithm for automated driving based on Frenet coordinate system

控制与决策. 2021, 36(4): 815−824 https://doi.org/10.13195/j.kzyjc.2019.0748

基于正态云模型的状态转移算法求解多目标柔性作业车间调度问题

State transition algorithm based on normal cloud model for solving multi−objective flexible job shop scheduling problem

控制与决策. 2021, 36(5): 1181−1190 https://doi.org/10.13195/j.kzyjc.2019.1233

阴影条件下基于迁移强化学习的光伏系统最大功率跟踪

Transfer reinforcement learning based maximum power point tracker of PV systems under partial shading condition

控制与决策. 2020, 35(12): 2939−2949 https://doi.org/10.13195/j.kzyjc.2019.0412

基于强化学习的挖掘机时间最优轨迹规划

张韵悦, 孙志毅†, 孙前来, 王 银

(太原科技大学 电子信息工程学院，太原 030024)

摘 要: 针对挖掘机的自主作业场景,提出基于强化学习的时间最优轨迹规划方法. 首先,搭建仿真环境用于产生数据,以动臂、斗杆和铲斗关节的角度、角速度为状态观测变量,以各关节的角加速度值为动作信息,通过状态观测信息实现仿真环境与自主学习算法的交互;然后,设计以动臂、斗杆和铲斗关节运动是否超出允许范围、完成任务总时间和目标相对距离为奖励函数对策略网络参数进行训练;最后,利用改进的近端策略优化算法 (proximal policy optimization, PPO) 实现挖掘机的时间最优轨迹规划. 与此同时,与不同连续动作空间的强化学习算法进行对比,实验结果表明:所提出优化算法效率更高,收敛速度更快,作业轨迹更平滑,可有效避免各关节受到较大冲击,有助于挖掘机高效、平稳地作业.

关键词: 挖掘机；自主作业；轨迹规划；多智能体；PPO算法；智能决策

中图分类号: TP241 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0811

引用格式: 张韵悦, 孙志毅, 孙前来, 等. 基于强化学习的挖掘机时间最优轨迹规划 [J]. 控制与决策, 2024, 39(5): 1433-1440.

Time optimal trajectory planning of excavator based on deep reinforcement learning

ZHANG Yun-yue, SUN Zhi-yi†, SUN Qian-lai, WANG Yin

(Department of Electronics and Information Engineering，Taiyuan University of Science and Technology，Taiyuan 030024，China)

Abstract: Aiming at the autonomous operation scenarios of excavators, a time optimal trajectory planning method based on reinforcement learning is proposed. This method builds a simulation environment to generate data. The angle and velocity of the boom, arm and bucket joints are used as state observation variables, and the angle acceleration of each joint is used as action information, and the simulation environment and autonomous learning are realized through the state observation information. The interaction of the algorithm is designed to train the policy network parameters using whether the joint motion of the boom, arm and bucket exceeds the allowable range, the total time to complete the task and the relative distance of the target as the reward function to train the policy network parameters. Finally, using the improved proximal policy optimization (PPO) realizes the time optimal trajectory planning of the excavator. At the same time, compared with the results of the different reinforcement learning algorithms with continuous action spaces, the experimental results show that the proposed optimization algorithm has higher efficiency, faster convergence speed, and smoother operation trajectory, which can effectively avoid the large impact on each joint and contribute to the efficient and stable operation of the excavator.

Keywords: excavator；autonomous operation；trajectory planning；multi-agent；PPO algorithm；intelligent decision-making

0 引 言

液压挖掘机作为用途广泛的机械装备,应用于采矿、农业、林业和建筑等行业. 由于挖掘机处于危险、恶劣的工作环境中,在长期高强度作业模式下,一方面使得工作人员的操作具有挑战性,另一方面驾驶员疲劳驾驶易引起低效率作业、不必要的能量损耗以及设备故障等问题[1]. 针对上述情况,为了保障人身安全和提高设备的生产力,对液压挖掘机进行智能化

收稿日期: 2022-05-10；**录用日期:** 2023-02-18.

基金项目: 山西省重点研发计划项目 (201903D121130)；山西省自然科学基金项目 (201901D111265)；山西省研究生创新项目 (2021Y670)；太原科技大学科研启动基金项目 (20192014).

责任编委: 刘德荣.

†通讯作者. E-mail: zys8128@163.com.

改造成为国内外学者重点关注的问题[2-3]. 自主化作业的挖掘机不仅可以带来可观的经济和社会效益,且可消除危险环境对工作人员的安全隐患,有助于实现挖掘机在核辐射、太空以及水下等区域的应用[4-5]. 因此,挖掘机向自动化、智能化发展已成为必然趋势[6].

目前,在国内外的相关研究工作中,澳大利亚机器人中心[7]和韩国首尔国立大学[8]建立了自主挖掘系统,国内山河集团和浙江大学[9-10]创建了挖掘机器人实验平台,挖掘机逐步实现了由人工操作到自主化作业的改造. 其中,在挖掘机的诸多智能化技术中,作业任务轨迹规划作为低层规划技术,不仅直接决定挖掘机的工作效率和能耗大小,且规划轨迹的可重复精度是保证设备在作业过程中平稳性和可靠性的重要基础. 因此,作为实现挖掘机自主化作业的核心技术,对轨迹进行合理地规划具有重要意义. 张文佳等[11]提出了S型-梯形速度轨迹规划方法,实现了高效率的点到点任务规划. Kim等[12]利用递归几何算法优化了B样条插值的时间-力矩最优轨迹. 白云飞等[13]采用了4次多项式插值方法,结合自适应粒子群求取能耗最优轨迹. 潘双夏等[14]利用了遗传算法求取修整平面作业的最优平滑轨迹. 综合上述研究,求取满足约束条件和优化目标的最优轨迹主要利用数值优化[15-18]和智能算法[19-20]来实现. 但是,该方法在优化过程中存在计算量大、易收敛于局部最优解和无法实现实时作业任务轨迹规划等问题.

随着人工智能技术的不断革新和成熟,近几年,以强化学习和深度学习为主的自学习算法被应用于挖掘机的任务轨迹规划研究. Egli等[21]利用TRPO (trust region policy optimization)算法实现了数据驱动挖掘机铲斗齿尖末端高精度跟踪目标轨迹. Kurinov等[22]利用具有协方差矩阵自适应近端策略优化算法(proximal policy optimization with covariance matrix adaptation, PPO-CMA)求解了挖掘机的端到端最优时间卸料轨迹. Hodel[23]采用了TRPO算法优化求解平滑的修整平面作业轨迹.

本文以PC 1012型液压挖掘机为研究对象,利用改进的PPO求取挖掘机工作装置的时间最优轨迹. 在虚拟环境中构建机械臂模拟挖掘机的工作装置,结合多智能体PPO的优化理论知识,提出自适应权重的采样机制和集中训练-分布执行训练方式,建立挖掘机时间最优运动轨迹的强化学习系统. 其中,多智能体系统由挖掘机动臂、斗杆和铲斗3个关节杆件建立的神经网络组成,在对多智能体系统的神经网络进行自主训练时,以机械臂和目标点构建的环境作为训练神经网络的数据支撑,即动臂、斗杆和铲斗关节角度、角速度以及目标角度状态为神经网络的输入,根据状态变量计算输出相应的各关节加速度值,结合初始角度和速度信息,由加速度得出角度和速度值,进一步以关节角度是否超限、总运动时间和目标相对距离为奖励函数对网络输出值进行评价,依据评价值对神经网络的参数进行修正,通过不断地训练学习,最终产生高奖励值的最优策略.

1 运动学描述

1.1 基于PoE的运动学正解

为了避免D-H坐标法繁琐的建模过程,本文采用旋量理论的指数积公式法[24](the product of exponentials formula, PoE)建立挖掘机的运动学模型,如图1所示. 图1中:1为回转平台,2为动臂关节,3为斗杆关节,4为铲斗关节. 由图1可见:动臂、斗杆和铲斗关节的运动主要依据基础坐标系U和末端工具坐标系T进行描述;$\xi_i\ (i=1,2,3,4)$为第i个关节的运动旋量坐标;θ_i为第i个关节的角度值;$a_j\ (j=0,1,2,3,4)$为机械臂的杆长,其中$a_0=0.69$,具体参数值如表1所示.

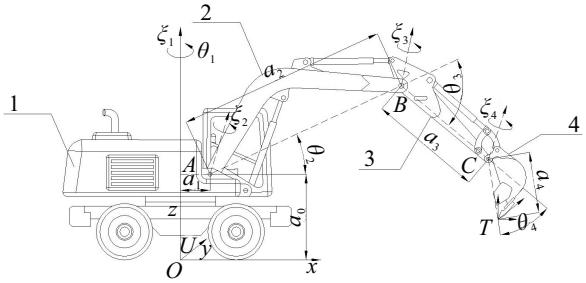

图1 基于PoE方法的挖掘机结构

表1 PC 1012挖掘机结构参数

关节i	杆长a_i/m	关节角度范围θ_i/(°)
回转	0.53	$-180\sim180$
动臂	1.475	$-53.83\sim54.62$
斗杆	0.797	$-156.61\sim-32.2$
铲斗	0.425	$-165.4\sim14.6$

首先,当各关节处于0位置时,即$\theta_i=0\ (i=1,2,3,4)$,建立T坐标系相对于U坐标系的齐次变换矩阵,如下式所示:

$$g_{ST}(0)=\begin{bmatrix}1&0&0&a_1+a_2+a_3+a_4\\0&1&0&0\\0&0&1&a_0\\0&0&0&1\end{bmatrix}. \tag{1}$$

各关节的旋量坐标ξ_i表达式为

$$\xi_i=\begin{bmatrix}\omega_i\\ \nu_i\end{bmatrix},\ i=1,2,3,4. \tag{2}$$

其中: ω_i、ν_i为运动旋量, $\omega_i \in R^3$为第i个关节旋转轴上的单位矢量, $\omega=(\omega_x,\omega_y,\omega_z)^{\mathrm{T}}\in R^3$, $\nu_i=\omega_i\times q_i$为第i个关节移动方向的单位矢量, $q_i\in R^3$为第i个关节旋转轴上的单位矢量.

然后, 通过将回转、动臂、斗杆和铲斗关节运动进行组合, 得到T坐标系相对于U坐标系的位姿描述, 即

$$g_{\mathrm{ST}}(\theta)=\mathrm{e}^{\hat{\xi}_1\theta_1}\mathrm{e}^{\hat{\xi}_2\theta_2}\mathrm{e}^{\hat{\xi}_3\theta_3}\mathrm{e}^{\hat{\xi}_4\theta_4}g_{\mathrm{ST}}(0). \tag{3}$$

其中: θ_i为第i个关节的转动量; $\hat{\xi}_i$为第i个关节的运动旋量, $\hat{\xi}_i=\begin{bmatrix}\hat{\omega} & \nu\\ 0 & 0\end{bmatrix}\in \mathrm{se}(3)$为特殊欧几里德群SE(3)对应的李代数se(3), $\hat{\omega}=\begin{bmatrix}0 & -\omega_z & \omega_y\\ \omega_z & 0 & -\omega_x\\ -\omega_y & \omega_x & 0\end{bmatrix}\in \mathrm{so}(3)$为三维特殊正交群SO(3)对应的李代数so(3).

最后, 由式(1)～(3)和Chasles定理[25], 计算得到挖掘机的正运动学方程为

$$g_{\mathrm{ST}}(\theta)=\begin{bmatrix}R(\theta) & P(\theta)\\ 0 & 1\end{bmatrix}. \tag{4}$$

其中: $R(\theta)=\begin{bmatrix}c_1c_{234} & -c_1s_{234} & s_1\\ s_1s_{234} & -s_1c_{234} & -c_1\\ s_{234} & c_{234} & 0\end{bmatrix}$为工作装置的位姿矩阵, $P(\theta)=\begin{bmatrix}c_1(a_4c_{234}+a_3c_{23}+a_2c_2+a_1)\\ s_1(a_4c_{234}+a_3c_{23}+a_2c_2+a_1)\\ a_4s_{234}+a_3s_{23}+a_2s_2+a_0\end{bmatrix}$依次为铲斗齿尖末端在$U$坐标系的$x$、$y$、$z$坐标值. 式中: $c_1=\cos\theta_1, c_2=\cos\theta_2, s_{23}=\sin(\theta_2+\theta_3), s_1=\sin\theta_1, s_2=\sin\theta_2, c_{234}=\cos(\theta_2+\theta_3+\theta_4), s_{234}=\sin(\theta_2+\theta_3+\theta_4), c_{23}=\cos(\theta_2+\theta_3)$.

1.2　运动学逆解

针对挖掘机自主化作业的轨迹规划任务, 在给定目标点的前提下, 通过在关节空间中设计满足优化目标和约束条件的动臂、斗杆以及铲斗关节运动曲线, 从而使得铲斗齿尖末端形成相应的作业轨迹. 为了实现任务规划前期铲斗齿尖末端目标点到工作装置各关节角度值的转化, 采用数值分析法进行运动学逆解, 如下式所示:

$$\tan\theta_1=\frac{y}{x}\Rightarrow\theta_1=\arctan 2\Big(\frac{y}{x}\Big), \tag{5}$$

$$\theta_2=\arctan\frac{A}{B}, \tag{6}$$

$$\cos\theta_3=\frac{C}{D}, \tag{7}$$

$$\sin\theta_3=\pm\sqrt{1-\cos^2\theta_3}, \tag{8}$$

$$\theta_3=\arctan 2(\sin\theta_3,\cos\theta_3), \tag{9}$$

$$\theta_4=\zeta-\theta_2-\theta_3. \tag{10}$$

其中

$$A=(a_2+a_3c_3)(z-a_0-a_4s_{234})-a_3s_3\Big(\frac{x}{c_1}-a_1-a_4c_{234}\Big),$$

$$B=(a_2+a_3c_3)\Big(\frac{x}{c_1}-a_1-a_4c_{234}\Big)+a_3s_3(z-a_0-a_4s_{234}),$$

$$C=\Big(\frac{x}{c_1}-a_1-a_4c_{234}\Big)^2+(z-a_0-a_4s_{234})^2-a_2^2-a_3^2,$$

$$D=2a_2a_3,$$

(x,y,z)为铲斗齿尖末端的坐标值, ζ为铲斗齿尖末端的姿态角度, θ_1、θ_2、θ_3、θ_4依次为回转平台、动臂、斗杆和铲斗关节角度值.

2　多智能体自主学习轨迹规划

对于自主作业的挖掘机, 在驱动空间与关节空间可实现相互转换的前提下, 如何规划工作装置各关节的运动以实现挖掘机高效率和平稳作业成为重点研究问题. 本文针对挖掘机端到端轨迹规划任务, 以作业时间为优化目标, 将动臂、斗杆和铲斗关节作为独立决策的智能体, 搭建深度神经网络对目标任务轨迹规划策略进行逼近, 利用改进的PPO自主学习算法求取动臂、斗杆和铲斗关节一系列最大奖励值的动作决策行为, 最终, 铲斗齿尖末端形成的时间最优轨迹为3个关节的组合决策序列.

2.1　PPO算法

PPO作为新型的policy gradient算法[26], 其实质为将策略π以参数为μ的神经网络进行表示, 通过神经网络与环境的交互, 形成包含L个步骤的序列τ, 即

$$\tau=\{s_1,a_1,s_2,a_2,s_3,a_3,\ldots,s_L,a_L\}. \tag{11}$$

其中: $s_t\in R^n\ (t=1,2,\ldots,L)$为当前环境的状态向量, $a_t\in R^m\ (t=1,2,\ldots,L)$为状态$s_t$对应神经网络的动作输出向量. 在同一个状态下, 神经网络输出的动作满足参数为μ的概率分布, 因此, 序列τ是不确定的, 序列发生的概率为

$$p_\mu(\tau)=p(s_1)p_\mu(a_1|s_1)p(s_2|s_1,a_1)p_\mu(a_2|s_2)p(s_3|s_2,a_2)\ldots=p(s_1)\prod_{l=1}^{L}p_\mu(a_l|s_l)p(s_{l+1}|s_l,a_l). \tag{12}$$

其中:$p(s_1)$表示当前环境初始状态为s_1的概率;$p_\mu(a_l|s_l)$表示在环境状态为s_l、参数为μ的神经网络,输出动作为a_l的概率;$p(s_{l+1}|s_l,a_l)$表示在状态为s_l下执行动作a_l时,新的环境状态为s_{l+1}的概率.

由于序列τ在每个阶段均获得相应的奖励,由此可计算得出在策略π的情况下,神经网络获得的期望奖励,有

$$\bar{R}_\mu=\sum_\tau R(\tau)p_\mu(\tau)=E_{\tau\sim p_\mu(\tau)}[R(\tau)]. \tag{13}$$

其中:$R(\tau)$为序列τ的总奖励值;$p_\mu(\tau)$表示神经网络参数为μ时,序列τ的概率分布. 为了使得策略网络获得最大化的期望奖励,利用梯度提升方法更新网络参数μ,通过更新策略π获取最大奖励,期望奖励的梯度为

$$\nabla\bar{R}_\mu=\frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L}A_\mu(s_l^n,a_l^n)\nabla\log p_\mu(a_l^n|s_l^n). \tag{14}$$

其中:N为采样得到的序列个数,$A_\mu(s_l^n,a_l^n)$为动作a_l在状态s_t的优势函数,L为第n个序列包含决策的个数,s_l^n、a_l^n依次为第n个序列中第l个状态和动作决策.

为了提高采样数据的重复利用率,加快训练速度,选择off-policy的学习方式,该方法对应的梯度变为

$$\begin{aligned}&\nabla\bar{R}_\mu=\\&E_{s_l,a_l\sim\pi_\mu}[A_\mu(s_l,a_l)\nabla\log p_\mu(a_l|s_l)]=\\&E_{s_l,a_l\sim\pi_{\mu'}}\Big[\frac{p_\mu(a_t|s_t)}{p_{\mu'}(a_t|s_t)}A_{\mu'}(s_t,a_t)\nabla\log p_\mu(a_l|s_l)\Big].\end{aligned} \tag{15}$$

其中:μ'为旧策略的神经网路参数;μ为新策略的神经网络参数;$E_{s_l,a_l\sim\pi_\mu}$表示神经网络参数为μ,由(s_l,a_l)计算得到的期望值. 为了在相同状态下,神经网络输出动作的概率分布相差小,将目标函数进行转换,其表达式如下式所示:

$$\begin{aligned}J_{\mu'}^{\text{clip}}(\mu)=&E_{s_l,a_l\sim\pi_{\mu'}}[\min(q_l(\mu)A_{\mu'}(s_t,a_t),\\&\text{clip}\,(q_l(\mu),1-\varepsilon,1+\varepsilon)A_{\mu'}(s_t,a_t))],\\\text{clip}\,(x,x_{\min},x_{\max})=&\begin{cases}x,\ x_{\min}\leqslant x\leqslant x_{\max};\\x_{\min},\ x<x_{\min};\\x_{\max},\ x_{\max}<x.\end{cases}\end{aligned} \tag{16}$$

其中:$q_l(\mu)=\dfrac{p_\mu(a_t|s_t)}{p_{\mu'}(a_t|s_t)}$,$p_\mu$为新策略在状态$s_l$执行动作$a_l$的概率,$p_{\mu'}$为旧策略在状态$s_l$执行动作$a_l$的概率. 该目标函数通过信赖域修正的方法实现梯度下降,可减小适应性修正的需求.

2.2 基于PPO算法的时间最优轨迹

基于第2.1节的理论知识,为了提高算法的优化性能,对PPO算法中采样机制和训练方式进行改进,并将其应用于求取高效率的自主作业轨迹.

2.2.1 优先采样机制

将动臂、斗杆和铲斗关节能够到达目标点的序列数据按照4元组$(s_l,a_l,s_{l+1},r_{l+1})$的形式存储于经验池. 其中:$s_l$为状态信息,$a_l$为智能体的动作信息,$s_{l+1}$为智能体执行新动作后的状态信息,$r_{l+1}$为执行该动作获得的奖励值. 在经验池中进行采样时,针对随机采样机制无法有效利用高质量的样本以及易使得模型陷入局部最优解的问题,本文采用自适应权重的优先采样机制,该机制在样本数量不断变化的情况下,可保证训练结果收敛.

策略网络的损失函数(16)会受到优势函数的影响,因此,在设计自适应权重时,将提高优势函数值对采样权重的影响. 依次计算每个智能体采样样本的优势值,并对其取绝对值,将绝对值进行由大至小的排序. 根据所有样本的采样概率和等于1,将样本自适应权重计算公式设计如下:

$$P_e=\frac{\dfrac{1}{e}}{\sum\limits_{e=1}^{M}\dfrac{1}{e}}. \tag{17}$$

其中:M为一个智能体的总样本数量,e为样本的序号,P_e为第e号样本的采样概率. 式(17)既提高了绝对值大的优势值样本采样概率,使得处于边界的奖励值样本也能影响策略网络的训练,又兼顾了探索与利用的关系,实现对不同的采样样本概率的平衡.

特别地,在多智能体系统中,为了协调智能体间的合作关系,将每个智能体作为单独个体产生的样本进行权重值计算,并按照各自的权重值采集设定的样本,用于更新策略网络参数.

2.2.2 集中训练-分布执行机制

在动臂、斗杆和铲斗关节组成的多智能体系统中,多智能体环境的状态转移由所有智能体的动作共同决定,每个智能体获得的回报与其他智能体相关,即改变单个智能体的策略会直接影响其他智能体最优决策的选择和值函数估计的精确性. 因此,为了保证多智能体系统算法的收敛性,本文采用集中训练-分布执行的架构进行仿真,如图2所示. 联合动作与整个系统的状态转移和奖励值函数有关,且在动作决策时智能体间存在耦合关系,因此,集中训练将所有智能体的状态信息s和动作信息$a_1\sim a_n$作为联合动作值函数的输入,有助于防止单个智能体的策略变化

影响其他智能体的决策.

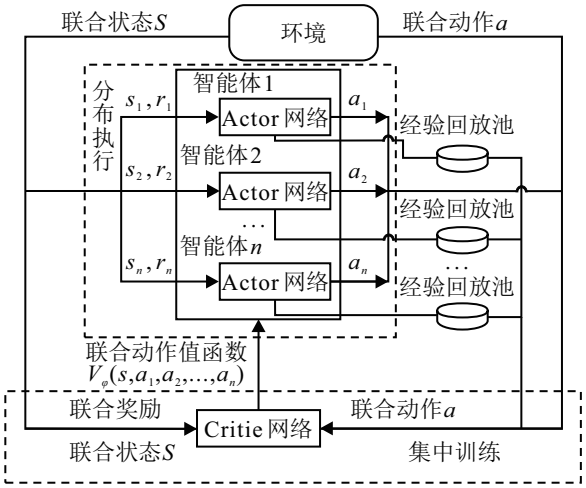

图 2　集中训练-分布执行结构

然而,在实际执行时,对于单个智能体而言只能获取到部分状态和动作信息,无法观测到其他智能体的信息,甚至有可能无法获得联合状态. 因此,针对该情况,对每个智能体采用分布式执行操作,将独立的智能体观测的信息作为决策网络的输入,输出则对应该智能体的决策动作,该方式有效弥补了模型探索性弱的缺点.

2.2.3　时间最优轨迹规划

所提出PPO算法采用Actor-Critic框架,因此,对动臂、斗杆和铲斗关节每个智能体建立Actor策略网络用于策略迭代更新;Critic评估网络为共享网络,用于计算全局动作值函数. 本文的核心工作为建立自学习系统,通过训练使得挖掘机工作装置具备自主规划时间最优轨迹的能力. 为了实现挖掘机工作装置在其允许的工作范围内高效率作业,设计智能体的奖励函数为

$$\begin{cases} r_{11}=-10\times|(\theta_2-(-53.83))\times(\theta_2<(-53.83))|,\\ r_{12}=-10\times|(\theta_2-54.62)\times(\theta_2>54.62)|,\\ r_{21}=\\ -10\times|(\theta_3-(-156.61))\times(\theta_3<(-156.61))|,\\ r_{22}=-10\times|(\theta_3-(-32.2))\times(\theta_3>(-32.2))|,\\ r_{31}=-10\times|(\theta_4-(-165.4))\times(\theta_4<(-165.4))|,\\ r_{32}=-10\times|(\theta_4-14.6)\times(\theta_4>14.6)|,\\ r_t=-t-\min(D_t). \end{cases} \tag{18}$$

$$r=r_{11}+r_{12}+r_{21}+r_{22}+r_{31}+r_{32}+r_t. \tag{19}$$

其中:θ_2、θ_3、θ_4依次为动臂、斗杆和铲斗关节角度值;r_{11}和r_{12}表示动臂关节运动是否超出允许运动范围获得的奖励;$\theta_2<-53.83$和$\theta_2>54.62$为布尔表达式,即当动臂关节角度值在允许运动范围内时,布尔表达式结果为0,反之,当动臂关节角度值超过允许范围时,布尔表达式的结果为1;同理,r_{21}和r_{22}、r_{31}和r_{32}依次为斗杆和铲斗关节运动是否超出允许运动范围获得的奖励;同样地,$\theta_3<(-156.61)$和$\theta_3>(-32.2)$、$\theta_4<(-165.4)$和$\theta_4>14.6$为布尔表达式;D_t为铲斗齿尖末端当前位置与目标终点的距离;t为作业的总时间. 由式(18)和(19)可知,当各关节超过允许的运动范围时,奖励会减少;当运动的总时间越长以及铲斗齿尖末端距离目标点的距离越大时,奖励也会越少. 考虑到各关节为独立的智能体,定义每个智能体与环境交互获得的奖励值是相同的,共享的评价网络会受到所有智能体动作的影响.

综上所述,基于PPO算法的时间最优轨迹规划过程如算法1所示.

算法1　近端策略优化算法.

step 1: 初始化策略网络Actor参数μ和评估网络Critic参数φ;

step 2: for　$h=0,1,\ldots$ do;

step 3: 运行随机策略$\pi_h=\pi(\mu_h)$获得完成目标任务的完备轨迹$J_h=\{\tau_l\}$;

step 4: 计算对应轨迹的奖励值r_l;

step 5: 由评价网络计算状态值函数V_{φ_h},并进一步计算优势函数$A_{\pi_{\mu_h}}$;

step 6: 采用Adam更新策略网络参数

$$\mu_{h+1}=\arg\max_{\mu}\frac{1}{|J_h|L}\sum_{\tau\in J_h}\sum_{l=0}^{L}\Big(\frac{p_\mu(a_t|s_t)}{p_{\mu_h}(a_t|s_t)}A_{\pi_{\mu_h}}(s_t,a_t),\ \mathrm{clip}\Big(\frac{p_\mu(a_t|s_t)}{p_{\mu_h}(a_t|s_t)},1-\varepsilon,1+\varepsilon\Big)A_{\pi_{\mu_h}}(s_t,a_t)\Big);$$

step 7: 更新评估网络参数

$$\varphi_{h+1}=\arg\min_{\varphi}\frac{1}{|J_hL|}\sum_{\tau\in J_h}\sum_{l=0}^{L}(V_{\varphi_h}(s_l)-r_l);$$

step 8: end　for.

3　仿真实验及分析

本文以PC 1012型挖掘机自主平整地面任务为优化目标,选取作业的目标点以及对应的运动学逆解,结果如表2所示. 对目标任务进行网络训练的硬件配置CPU为i7-7700 HQ、显卡为GTX 1060、内存为16 GB.

表 2　目标点由位姿空间到关节空间的转换

目标点/m	动臂关节角度/(°)	斗杆关节角度/(°)	铲斗关节角度/(°)
(2.85, 0, 0)	4.886 8	−37.578 2	−32.308 6
(1.35, 0, 0)	14.394 8	−131.644 6	−7.750 2

3.1 模型参数配置

利用PPO算法训练强化学习智能体前,对模型的基本要素进行定义.

1) 状态s. 定义可到达目标点的动臂、斗杆和铲斗关节角度、由策略网络输出的动作信息角加速度计算得到的角速度和目标角度值作为观测信息. 同时,以上信息作为策略网络的输入,对其进行了归一化处理.

2) 动作a. 定义策略网络的输出为各关节的加速度值,且采取的动作满足$a_l \sim N(0,1)$的正态分布. 本文为了降低决策动作的维度,对输出信息进行离散化.

3) 奖励函数r. 奖励由两部分组成: 第1部分为针对各关节是否超出允许的活动范围获取的奖励; 第2部分由完成任务的总时间和当前铲斗齿尖末端点与给定目标点的距离值作为奖励.

4) 网络设计. Actor与Critic网络结构基本相同,采用双隐层结构的全连接网络,隐藏层包含512个神经元,ReLu函数为激活函数. 其中: Actor网络接收归一化的状态观测信息,经过全连接层后,设置Softmax函数作为神经网络最后一层,将输出结果转化为概率分布向量,形成离散化的输出信息; Critic网络输出一维状态值函数.

5) 超参数设置. 采用Adam网络优化器,学习率为0.000 25,折扣率为0.99,裁剪率为0.2,批大小数量为128,经验库容量设定为4 000,开始训练样本数量为1 800.

3.2 实验结果分析

由于当前对强化学习算法优劣的评价指标没有统一标准,本文将从两方面进行评价: 1) 奖励值的曲线走势,奖励值越大表明算法越好,曲线收敛速度越快表明算法的收敛性越好; 2) 依据算法的最终优化结果,即各关节完成任务的时间长短,完成任务效率越高,算法性能越好. 因此,本文将强化学习中适用于求解连续动作空间的DDPG (deep deterministic policy gradient)算法、TRPO算法、传统PPO算法和改进的PPO算法用于求解挖掘机的时间最优轨迹.

首先,为了表明所提出改进PPO算法的优越性,在相同条件下依次利用DDPG算法、TRPO算法和传统PPO算法求取时间最优轨迹. 各算法的学习时间如表3所示. 由表3可见,所提出改进的PPO算法训练时间最短,优化效率更高.

表 3 不同算法的训练时间对比

算法	训练时间/h
DDPG	36.2
TRPO	33
传统的PPO	34.5
本文改进的PPO	21

然后,获得各算法的平均奖励值曲线,如图3所示. 由图3可见,随着训练周期的进行,4种算法获得的奖励值逐渐增加,且最终均趋于稳定,表明通过策略网络与环境的交互,4种方法利用奖励函数不断地修正网络参数,最终使得网络参数收敛,均训练得出最优策略. 但是,相比于DDPG、TRPO与传统PPO算法,所提出方法奖励值波动较小且获得的奖励值较大,收敛速度快,算法训练过程中效率高.

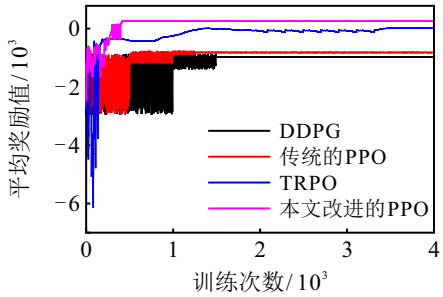

图 3 奖励值曲线对比

最后,综合考虑上述各算法的学习训练时间和奖励值等因素,选择了训练时间较短和奖励值较大的TRPO算法和所提出改进的PPO算法得出的最优策略用于求解得到各关节自主作业的角度变化曲线,如图4和图5所示.

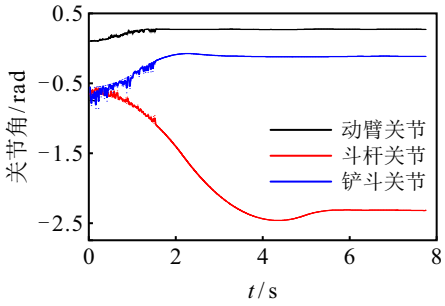

图 4 TRPO算法的关节角度曲线

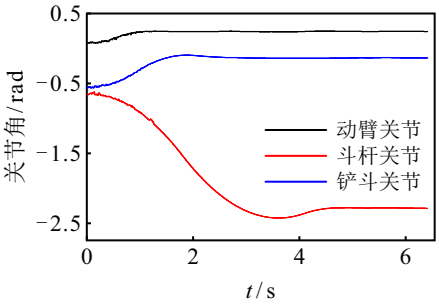

图 5 改进PPO算法的关节角度曲线

由图4和图5可见,两种算法的规划结果前期均存在抖动,原因是开始阶段存在较小概率错误动作的输出,导致产生了抖动现象,但是,随着训练周期的增

加,曲线均趋于平滑. 相比而言,所提出算法训练得出的关节角度曲线更平滑,各关节实现以小幅度变化到达目标点,有助于保护液压驱动装置;此外,所提出改进的PPO算法完成平整地面任务所用时间为6.4 s,而TRPO算法所用时间为7.695 s,因此,所提出方法规划的轨迹效率更高.

与此同时,在作业时间已知的条件下,将最优策略产生的各关节角度值代入挖掘机模型中,铲斗齿尖末端产生的轨迹如图6所示. 由图6可见,TRPO算法训练得到的时间最优轨迹不平滑,易对各关节带来较大冲击;所提出改进的PPO算法求得的作业轨迹不仅效率高,且连续平滑,验证了所提出方法的有效性.

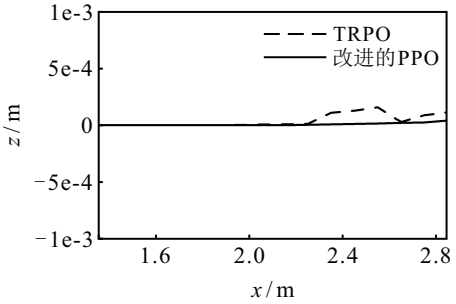

图 6　平整作业对比

4　结　论

1) 在利用旋量理论对挖掘机工作装置建立运动学方程的基础上,提出了多智能体系统自主学习的轨迹规划方法.

2) 采用基于自适应权重采样机制和集中训练-分布执行框架的PPO算法对神经网络进行训练,最终得到了高效率、平稳的自主作业轨迹.

3) 通过与连续动作空间的强化学习算法的优化结果进行对比,表明了所提出算法收敛性更快、效率更高.

参考文献(References)

[1] Dadhich S, Bodin U, Andersson U. Key challenges in automation of earth-moving machines[J]. Automation in Construction, 2016, 68: 212-222.

[2] Jung T, Raduenz H, Krus P, et al. Boom energy recuperation system and control strategy for hydraulic hybrid excavators[J]. Automation in Construction, 2022, 135: 104046.

[3] Vahdatikhaki F, Langroodi A K, Scholtenhuis L O, et al. Feedback support system for training of excavator operators[J]. Automation in Construction, 2022, 136: 104188.

[4] Koivo A J, Thoma M, Kocaoglan E, et al. Modeling and control of excavator dynamics during digging operation[J]. Journal of Aerospace Engineering, 1996, 9(1): 10-18.

[5] 李运华, 范茹军, 杨丽曼, 等. 智能化挖掘机的研究现状与发展趋势[J]. 机械工程学报, 2020, 56(13): 165-178.
(Li Y H, Fan R J, Yang L M, et al. Research status and development trend of intelligent excavators[J]. Journal of Mechanical Engineering, 2020, 56(13): 165-178.)

[6] 葛磊, 董致新, 李运华, 等. 系列化液压挖掘机数字样机研究[J]. 机械工程学报, 2019, 55(14): 186-196.
(Ge L, Dong Z X, Li Y H, et al. Research on digital prototypes of serial hydraulic excavators[J]. Journal of Mechanical Engineering, 2019, 55(14): 186-196.)

[7] Assadzadeh A, Arashpour M, Brilakis I, et al. Vision-based excavator pose estimation using synthetically generated datasets with domain randomization[J]. Automation in Construction, 2022, 134: 104089.

[8] Yoo S, Park C G, Lim B, et al. Bandwidth maximizing design for hydraulically actuated excavators[J]. Journal of Vibration and Control, 2010, 16(14): 2109-2130.

[9] 任钊民. 基于SLAM和IMU融合定位的履带式液压挖掘机的行走轨迹控制研究[D]. 杭州: 浙江大学, 2020: 74-95.
(Ren Z M. Rsearch on walking trajectory control of tracked hydraulic excavator based on SLAM and IMU fusion positioning[D]. Hangzhou: Zhejiang University, 2020: 74-95.)

[10] 张杰. 挖掘机器人轨迹跟踪电液伺服控制策略研究[D]. 杭州: 浙江大学, 2020: 46-60.
(Zhang J. Research on electro-hydraulic servo control strategy for trajectory tracking of robotic excavator[D]. Hangzhou: Zhejiang University, 2020: 46-60.)

[11] 张文佳, 尚伟伟. 2自由度绳索牵引并联机器人的高速点到点轨迹规划方法[J]. 机械工程学报, 2016, 52(3): 1-8.
(Zhang W J, Shang W W. High-speed point-to-point trajectory planning of a 2-DOF cable driven parallel manipulator[J]. Journal of Mechanical Engineering, 2016, 52(3): 1-8.)

[12] Kim Y B, Ha J, Kang H, et al. Dynamically optimal trajectories for earthmoving excavators[J]. Automation in Construction, 2013, 35: 568-578.

[13] 白云飞, 张奇峰, 范云龙, 等. 基于能耗优化的深海电动机械臂轨迹规划[J]. 机器人, 2020, 42(3): 301-308.
(Bai Y F, Zhang Q F, Fan Y L, et al. Trajectory planning of deep-sea electric manipulator based on energy optimization[J]. Robot, 2020, 42(3): 301-308.)

[14] 潘双夏, 季炳伟, 童永峰. 基于操纵平稳性的液压挖掘机轨迹规划方法[J]. 浙江大学学报: 工学版, 2006,

40(8): 1311-1314.
(Pan S X, Ji B W, Tong Y F. Trajectory planning for hydraulic excavating machine based on manipulating stability[J]. Journal of Zhejiang University: Engineering Science, 2006, 40(8): 1311-1314.)

[15] Boryga M, Graboś A. Planning of manipulator motion trajectory with higher-degree polynomials use[J]. Mechanism and Machine Theory, 2009, 44(7): 1400-1419.

[16] 庄宇飞, 马广富, 黄海滨. 欠驱动刚性航天器时间最优轨迹规划设计[J]. 控制与决策, 2010, 25(10): 1469-1473.
(Zhuang Y F, Ma G F, Huang H B. Time-optimal motion planning of an underactuated rigid spacecraft[J]. Control and Decision, 2010, 25(10): 1469-1473.)

[17] Seo J, Lee S, Kim J, et al. Task planner design for an automated excavation system[J]. Automation in Construction, 2011, 20(7): 954-966.

[18] 文郁, 黄江帅, 江涛, 等. 安全平滑的改进时间弹性带轨迹规划算法[J]. 控制与决策, 2022, 37(8): 2008-2016.
(Wen Y, Huang J S, Jiang T, et al. Safe and smooth improved time elastic band trajectory planning algorithm[J]. Control and Decision, 2022, 37(8): 2008-2016.)

[19] 龙腾, 李恩, 杨国栋, 等. 基于PSO非均匀样条插值的混合结构柔性臂抑振轨迹规划[J]. 控制与决策, 2018, 33(6): 978-988.
(Long T, Li E, Yang G D, et al. Trajectory planning of vibration suppression for hybrid structure flexible manipulator based on PSO non-uniform spline interpolation[J]. Control and Decision, 2018, 33(6): 978-988.)

[20] 訾斌, 徐锋, 唐锴, 等. 基于机器视觉的喷涂机器人轨迹规划与涂装质量检测研究综述[J]. 控制与决策, 2023, 38(1): 1-21.
(Zi B, Xu F, Tang K, et al. Trajectory planning for spray-painting robot and quality detection of paint film based on machine vision: A review[J]. Control and Decision, 2023, 38(1): 1-21.)

[21] Egli P, Hutter M. Towards RL-based hydraulic excavator automation[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, 2020: 25-29.

[22] Kurinov I, Orzechowski G, Hämäläinen P, et al. Automated excavator based on reinforcement learning and multibody system dynamics[J]. IEEE Access, 2020, 8: 213998-214006.

[23] Hodel B J. Learning to operate an excavator via policy optimization[J]. Procedia Computer Science, 2018, 140: 376-382.

[24] 魏武, 李艳杰, 廖志鹏, 等. 基于旋量理论的蛇形机器人运动学建模[J]. 华南理工大学学报: 自然科学版, 2019, 47(2): 1-8.
(Wei W, Li Y J, Liao Z P, et al. Kinematics modeling of snake-like robot based on screw theory[J]. Journal of South China University of Technology: Natural Science Edition, 2019, 47(2): 1-8.)

[25] 阳涵疆, 李立君, 高自成. 基于旋量理论的混联采摘机器人正运动学分析与试验[J]. 农业工程学报, 2016, 32(9): 53-59.
(Yang H J, Li L J, Gao Z C. Forward kinematics analysis and experiment of hybrid harvesting robot based on screw theory[J]. Transactions of the Chinese Society of Agricultural Engineering, 2016, 32(9): 53-59.)

[26] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J/OL]. 2017, arXiv: 1707.06347.

作者简介

张韵悦(1991−), 女, 博士生, 从事设备智能化技术及应用等研究, E-mail: 2284061147@qq.com;

孙志毅(1959−), 男, 教授, 博士生导师, 从事深度学习缺陷检测、图像处理、设备智能化技术及应用等研究, E-mail: zys8128@163.com;

孙前来(1976−), 男, 副教授, 博士, 从事图像处理和基于机器视觉的缺陷检测等研究, E-mail: sqlsun@tyust.edu.cn;

王银(1982−), 男, 副教授, 博士, 从事计算机视觉和智能控制等研究, E-mail: xpw417@163.com.