控制与决策
Control and Decision
ISSN 1001-0920,CN 21-1124/TP

《控制与决策》网络首发论文

题目：事件触发式多智能体分层安全强化学习运动规划
作者：孙辉辉，胡春鹤，张军国
DOI：10.13195/j.kzyjc.2023.1288
收稿日期：2023-09-11
网络首发日期：2024-06-06
引用格式：孙辉辉，胡春鹤，张军国．事件触发式多智能体分层安全强化学习运动规划[J/OL]．控制与决策. https://doi.org/10.13195/j.kzyjc.2023.1288

网络首发时间：2024-06-06 14:33:05
网络首发地址：https://link.cnki.net/urlid/21.1124.TP.20240605.1313.011

事件触发式多智能体分层安全强化学习运动规划

孙辉辉[1,2]，胡春鹤[1,3†]，张军国[1,3]

(1. 北京林业大学 工学院，北京 100083；2. 淮南师范学院 机械与电气工程学院，安徽 淮南 232038；
3. 林木资源高效生产全国重点实验室，北京 100083)

摘　要: 针对深度强化学习序贯决策过程中面临的动作安全性问题,研究一种事件触发式多智能体分层安全强化学习运动规划方法. 首先,基于受限马尔可夫决策模型,构建一种具备安全约束的多智能体深度确定性策略梯度框架,该框架针对不同状态空间,以事件触发的方式实现运动策略的分层学习;然后,通过引入李雅普诺夫评价网络,建立带有条件约束的目标动作选择机制,并利用拉格朗日乘子法,解决多目标约束求解困难的问题,保证机器人内部决策的安全性;最后,在多机器人强化学习场景中对所提出方法进行实验. 实验结果表明:触发式多智能体分层安全强化学习方法使得机器人的状态轨迹从危险状态中快速恢复至安全空间,增强了策略的安全性和多机协同运动规划能力.

关键词: 强化学习；安全约束；运动规划；多智能体；事件触发

中图分类号: TP24　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.1288

引用格式: 孙辉辉,胡春鹤,张军国. 事件触发式多智能体分层安全强化学习运动规划[J]. 控制与决策, xxxx, xx(x): 1-xxxx.

Multi-agent event triggered hierarchical security reinforcement learning

SUN Hui-hui[1,2]，*HU Chun-he*[1,3†]，*ZHANG Jun-guo*[1,3]

(1. School of Technology，Beijing Forestry University，Beijing 100083，China；2. School of Mechanical and Electrical Engineering，Huainan Normal University，Huainan 232038，China；3. State Key Laboratory of Efficient Production of Forest Resources，Beijing 100083，China)

Abstract: In order to address the security issues that may arise in the sequential decision-making process of deep reinforcement learning, this paper studies a motion planning method based on multi-agent event triggered hierarchical security reinforcement learning (MEHSRL) method. Firstly, this method constructs a multi-agent twin delayed deep deterministic policy gradient algorithm based on the constrained Markov decision model. The model uses state security events as trigger conditions to implement hierarchical reinforcement learning in different state spaces. Then, by introducing a Lyapunov evaluation network, additional safety constraint rules are constructed for the reinforcement learning network, and the safety of robot decision is ensured by multi constraint objective optimization learning. Finally, the proposed method is tested in the security reinforcement learning scenario. The results show that proposed method achieves the goal of restoring the state trajectory from the dangerous state to the safe space in a limited time, improving the security of the strategy, and the effect of motion planning is better than the comparison method.

Keywords: reinforcement learning；security constraint；motion planning；multi-agent；event triggered

0　引　言

深度强化学习以数据驱动的方式在诸多序贯决策问题中获得优异表现,为移动机器人智能化进程提供了一套标准化学习范式[1-2]. 但是,在深度强化学习的应用和改进案例中,智能体往往将追求最大化奖励作为策略优化的唯一目标,并未充分考虑运动控制安全性因素的影响[3]. 有些深度强化学习方法甚至在学习过程中加入了随机探索以期获得更好的近似最优解,从而忽略了决策的潜在风险[4-5]. 特别是当遭遇连续时变风险未知的非结构场景时,缺乏安全约束的运动策略不仅会使得机器人陷入未知的风险状态,且易导致被控对象的物理损伤,进而带来不可预测的灾难性后果[6]. 研究人员发现,基于奖励最优准则的强化学习方法并不完全适用于充满未知信息的非安

收稿日期: 2023-09-11；**录用日期:** 2024-01-16.
基金项目: 国家自然科学基金项目(61703047)；河北省高等学校科学技术研究项目(QN2021312).
†通讯作者. E-mail: huchunhe@bjfu.edu.cn.

全场景,机器人的决策安全性也因而受到了严峻的挑战. 因此,如何将安全性考虑到移动机器人的运动规划的过程中,去限制和避免这些不安全的情况发生,成为了深度强化学习在机器人领域亟待研究的重要议题之一[7]. 为了解决这一问题,当前的强化学习策略常常应用边界约束条件对智能体的行为进行限制,以确保其在任务执行中遵守特定的条件. 这种限制可按照不同的方式实现: 1) 可将限制条件作为一组约束添加到智能体的决策过程中,以强制约束动作的输出范围; 2) 可通过设计不同奖励函数来引导智能体的行为逐渐趋近安全状态[8]. 这类方法虽然能够避免一些不安全的情况发生,但却不能完全解决安全性与奖励协调的问题[9]. 若约束条件设置过小,则易被智能体直接忽视; 若约束设置过大,则智能体易表现过于保守,从而错过或无法找到最优解[10-11].

不同于当前的动作约束型运动规划策略,研究者们重新提出了一类安全强化学习运动规划方法[12]. 安全强化学习方法可在满足安全约束条件的前提下,使用受限马尔可夫决策过程来描述运动模型,并通过最大化期望回报值来寻找最优策略[13]. 该方法并非在动作空间上直接约束[14],而是在动作评价网络的基础上增加了安全代价条件作为附加限制,共同指导策略网络的更新[15],进而保证机器人的行为安全性和合规性[16-17]. 在移动机器人领域中,特别是面向实时控制的无人驾驶、智能交通[18]、智能仓储等行业,安全强化学习系统已被广泛应用[19-22]. 虽然安全强化学习被提出,但是在实际应用中仍然面临一些问题[23-24]: 1) 具备安全约束的智能体往往过度注重当前动作的安全约束,却忽视了对于未来长期代价的考虑,使得策略出现了过于短视的问题,丧失了对最优策略的探索,导致机器人任务成功率的偏低; 2) 安全强化学习网络结构较为复杂、灵活性差,无法针对不同的运动状态采取合适的安全策略,在网络的快速学习和动作的安全约束问题上无法同时兼顾; 3) 带有安全约束的马尔可夫决策过程面临着运算繁琐、求解困难的问题,无法实现目标约束条件随着策略梯度同步更新[25].

针对以上问题,本文提出一种基于事件触发式多智能体分层安全强化学习运动规划方法(MEHSRL). 首先,所提出方法以双延迟深度确定性策略梯度为基础,为机器人构建一种具有安全保证的多智能体强化学习框架. 然后,该框架以安全阈值划分系统所处的状态空间,并为此构建事件触发式分层安全强化学习网络: 上层进化网络可实现策略的快速学习; 下层恢复网络基于李雅普诺夫安全约束条件,保证状态轨迹可在有限时间内从危险状态中恢复至安全空间. 最后,在安全强化学习仿真环境中,对所提出方法进行实验. 实验结果表明: 所提出MEHSRL方法在满足动作约束的情况下输出安全的动作策略,且能够提升多机器人的协同作业能力和任务成功率.

1 多智能体安全强化学习

传统的强化学习以最大化长期期望奖励为目标,多智能体安全强化学习方法在受限马尔可夫决策过程中寻找最优策略. 受限马尔可夫决策过程可使用六元组 (S, A, P, R, C, N) 来描述. 其中: N 为机器人总数, S 为状态空间集合, A 为一系列的动作组成的动作空间集合, P 为状态转移概率, R 为交互即时奖励, C 为即时代价. 安全强化学习问题为在满足安全约束的情况下,求解最大化期望目标的最优策略 π^*,即

$$\pi^* = \arg\max_{\pi\in\Omega_C} J(\pi). \tag{1}$$

其中: Ω_C 为满足安全约束的安全策略集, $J(\pi)$ 为目标函数.

为了约束智能体的动作,并考虑非安全因素带来的影响,在学习动作价值函数的同时,需要另外一个评价函数来评估当前策略的安全性,即非负安全约束函数 $c(s,a)$. 安全约束函数用于评估状态动作对的安全程度,即基于当前策略 π 和状态 s 选择动作 a 后获得的代价值.

安全约束下的马尔科夫决策过程主要关注长期累计期望下的安全代价值. 对于给定的策略 π,轨迹 τ 的累计折扣安全代价为

$$c(\tau) = \sum_{t=0}^{\infty} \gamma^t c_t. \tag{2}$$

其中: c_t 为单步即时代价, τ 为根据策略得到的状态动作轨迹.

在安全强化学习过程中,智能体的安全代价值由当前状态和动作决定. 受限马尔可夫决策的目标是最大化累计期望收益的同时,保证长期累计代价小于安全代价阈值. 最优策略 π^* 的求解过程可转化为一个带有条件约束的目标优化问题,即在约束条件下 $J_c(\pi)\leqslant m$,求解可最大化累计期望的策略 π^*,有

$$\begin{aligned} &\max \underset{s\sim\rho^\pi}{E}\Big[\sum_t \gamma^t r_t\Big];\\ &\text{s.t. } J_c(\pi)\leqslant m. \end{aligned} \tag{3}$$

其中: m 为安全代价的阈值, $J_c(\pi)$ 为代价函数的约束函数.

在与环境互动过程中,智能体每次会收到一个即时奖励R和一个代价C,安全策略目标是在不超过安全代价阈值的条件下最大化长期奖励,通过最大化满足安全约束条件的奖励函数,实现奖励值和代价值的最优匹配.

2　触发式分层安全强化学习规划方法

事件触发的分层安全策略包括两个部分,分别为进化策略和安全恢复策略. 前者负责安全策略的不断学习和更新,力求快速达到最优解;后者的重点在于确保机器人在接近或违反安全约束时能够提供安全约束,并将机器人带回安全状态. 状态的分层标准遵循“安全触发器”. 触发事件为李雅普诺夫安全状态函数$L_\pi(s)$. 当触发事件的输出值高于设定的特定阈值时,会触发安全触发器,系统从进化策略切换至安全恢复策略,从而确保机器人动作输出的安全性,具体如图1所示. 在当前状态s下,机器人在执行动作a后,若触发事件的输出值大于等于阈值,即$L_\pi(s)\geqslant m$,则机器人将被认定为进入了非安全状态. 运动规划系统将被切换至安全恢复策略,机器人的输出动作将会受到安全条件的约束. 触发事件函数可通过当前状态下的累计代价得到,具体表达式为

$$L_\pi(s)=E_\pi(c_t+\gamma c_{t+1}+\gamma^2 c_{t+2}+\ldots\mid s=s_t)=E_{a\sim\pi}[L_c(s,a)]=\sum_a \pi(a\mid s)L_c(s,a). \quad (4)$$

其中:$L_c(s,a)$为李雅普诺夫评价函数,c_t为当前动作和状态下的代价值.

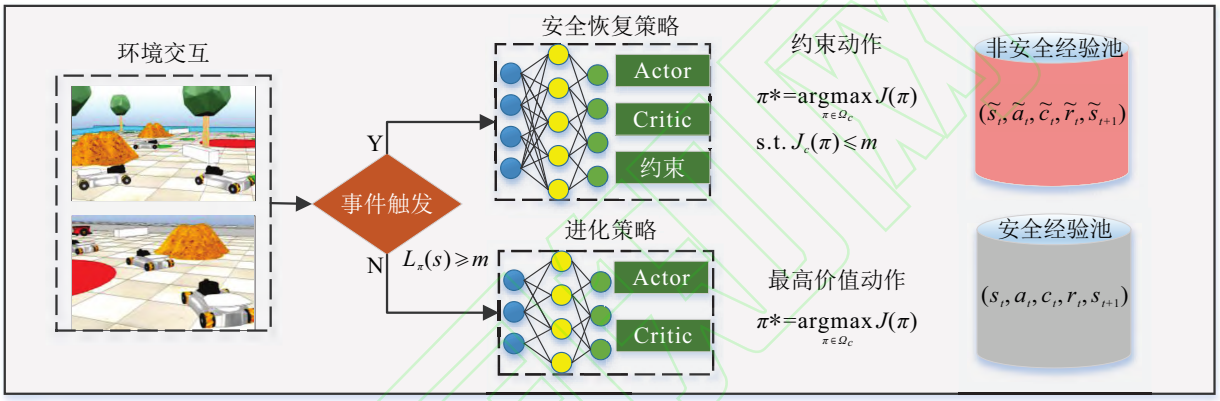


图 1　事件触发式分层安全策略

触发阈值m由经验样本中即时代价的均值得到,具体如下所示:

$$m=\frac{1}{N}\sum_{j=1}^{N}c_t^j, \quad (5)$$

其中N为经验样本数量.

根据不同的策略分别构建不同的经验样本池:非安全经验空间U_D和安全样本空间U_S. 智能体执行安全恢复策略时,非安全状态经验样本$(\tilde{s}_t,\tilde{a}_t,\tilde{c}_t,\tilde{r}_t,\tilde{s}_{t+1})$将被存入非安全经验空间$U_D$;智能体执行进化策略时,安全样本$(s_t,a_t,c_t,r_t,s_{t+1})$将被存入安全经验样本空间$U_S$.

1) 进化策略.

在事件触发的分层安全策略中,进化策略采用多智能体并行的结构. 每个智能体具有独立的Actor-Critic网络,同时使用双重Critic网络降低动作价值的高估误差. Critic网络统一使用全局信息进行学习,Actor在进行决策时使用观察到的局部信息进行决策. 所有智能体的联合策略可表示为$\pi=[\mu_1,\mu_2,\ldots,\mu_n]$,其策略参数可表示为$\theta=[\theta_1,\theta_2,\ldots,\theta_n]$.

在进化策略中:Actor网络根据观测状态输出机器人的动作,Critic网络负责评估当前动作的价值. 评价策略优劣的标准是执行当前动作后获得的奖励大小,累计期望奖励值越高,当前策略的性能越优. 因此,选择动作价值函数的累计期望作为优化目标,有

$$J(\mu)=E[Q_\mu(S,A)]. \quad (6)$$

其中:S为全局观测向量,A为所有智能体的动作集合,μ为确定性策略函数. 然后,使用求解策略梯度的方式对智能体的目标函数进行优化,即

$$\nabla J_\theta(u)=E_{s,a}[\nabla_\theta\mu(a|s)\nabla_a Q_\mu(S,A)|_{a=u(s)}]. \quad (7)$$

评价网络Critic采用集中式更新的方式,并利用TD误差进行网络更新优化,Critic网络更新的损失函数可表示为

$$L(w)=E[(y-Q_\mu(S,A|w))^2]. \quad (8)$$

其中:$y=r_t+\gamma Q_{\mu'}(S',A'|w')$为智能体的目标动作价值函数,$w$为当前Critic价值网络参数,$w'$为目标Critic价值网络参数.

在Actor-Critic架构的强化学习方法中,求解价

值函数时易产生高于真实值的偏差. 为了解决这个问题,额外增加了 1 组附加 Critic 网络来降低这种高估的偏差. 两个网络同时对动作价值进行评估,并始终选取较小值进行更新,从而有效抑制动作价值高估的影响. 基于双 Critic 网络生成的目标动作价值 y_c[26] 可表示为

$$\begin{cases} y_c = r_t + \min \gamma Q_{\mu'}(S', u'(S') + \varepsilon), \\ \varepsilon \sim \text{clip}(N(0,\sigma), -b, b). \end{cases} \tag{9}$$

其中: ε 为微噪声, b 为噪声的边界值.

这种微噪声的平滑操作,可增加算法的泛化能力,缓解过拟合问题,减少过高估计的不良状态对策略的干扰.

2) 安全恢复策略.

进化策略网络中的 Critic 评价网络仅包含了动作的奖励信息,却缺少了对安全因素的评价,因此易产生过分激进的策略使得机器人陷入危险的环境. 为了解决此问题,本节构建了带有安全约束的安全恢复网络,如图 2 所示.

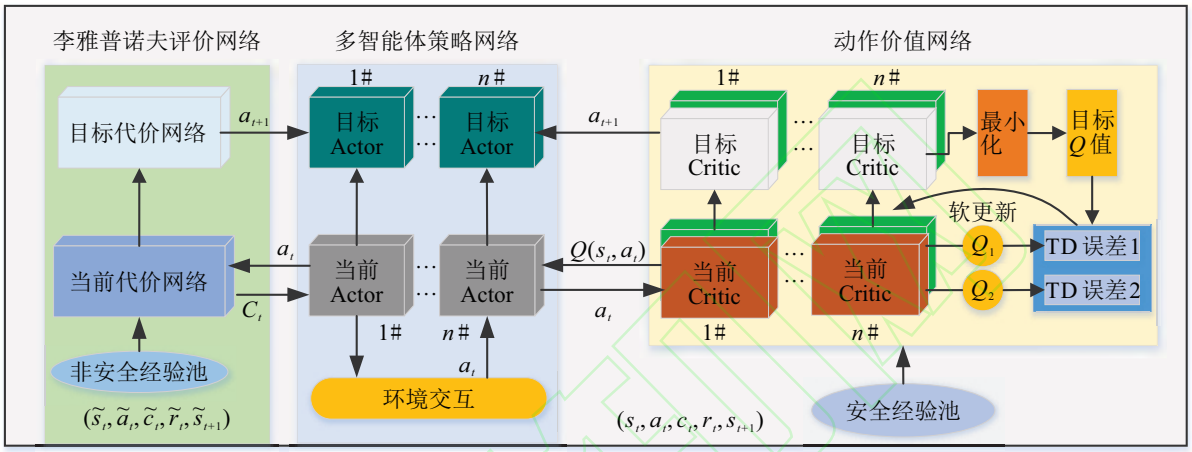


图 2 带有安全约束的多智能体网络架构

在多智能体分布式 Actor-Critic 架构的基础上,增加了用于安全约束的李雅普诺夫评价网络,并输出李雅普诺夫评价函数 $L_c(s,a)$ 来评估机器人每个决策的期望代价值,以此约束动作的决策过程. 3 类网络均包括当前网络和目标网络: 当前网络用来评估和决策,目标网络用来抑制高估问题. 李雅普诺夫评价函数 $L_c(s,a)$ 是一个概念性的函数,其值取决于所对应的状态和动作,与状态评价函数 $L_\pi(s)$ 的关系为

$$L_\pi(s) = E_{a\sim\pi}[L_c(s,a)]. \tag{10}$$

李雅普诺夫评价函数的计算方式与动作价值函数类似,当前代价值与未来代价值间满足贝尔曼方程,即

$$L_c(s,a) = c(s,a) + \gamma_c E_{a'\sim u(|s')}[L'_c(s',a')]. \tag{11}$$

其中: $c(s,a)$ 为即时非负安全约束函数, $L'_c(s',a')$ 为目标李雅普诺夫评价函数.

为了更新策略函数 μ, MEHSRL 需要最大化目标函数,考虑到边界安全约束条件 $J(\mu)$,最大化的目标价值为

$$\begin{aligned} &\max_{\mu} J(\mu) = E\Big[\sum_{t=0} \gamma^t r(s,a)\Big]; \\ &\text{s.t. } J_c(\mu) = E\Big[\sum_{t=0} \gamma^t c(s,a)\Big] \leqslant m. \end{aligned} \tag{12}$$

带有约束的目标函数优化过程是复杂的,为了简化求解过程,本文采用了拉格朗日乘子法来解决多目标优化问题. 首先,通过引入拉格朗日乘子,将约束条件融入目标函数中,将有约束问题转化为无约束问题来求解. 然后,基于动作价值函数 $Q(s,a)$ 和安全代价函数 $L_c(s,a)$,为策略网络构建了带有拉格朗日乘子的无约束优化目标,如下所示:

$$\begin{aligned} J_{\text{Actor}} = E_{U_D}[-Q_\mu(s,a)] + \xi E_{U_D}[L_c(s',a') - \\ (L_c(s,a) - k\cdot c(s,a))]. \end{aligned} \tag{13}$$

其中: ξ 为拉格朗日乘子, k 为安全系数, U_D 为非安全经验池. 接着,通过求解目标函数的梯度来最小化目标函数,进而优化安全策略网络的参数 θ. 目标函数的梯度可表示为

$$\begin{aligned} \nabla_\theta J_{\text{Actor}} = E_{U_D}[-\nabla_a Q_w(s,a)\nabla_\theta \mu_\theta(s)] + \\ \xi E_{U_D}[-\nabla_a L_c(s,a)\nabla_\theta \mu_\theta(s)], \end{aligned} \tag{14}$$

其中 μ_θ 表示在当前策略. 另外,李雅普诺夫评价函数通过最小化损失函数来更新,其损失函数 $\text{Loss}(\phi)_{\text{safe}}$ 可表示为

$$\begin{aligned} \text{Loss}(\phi)_{\text{safe}} = E_{U_D}[(y_s - L_c(s,\mu_\theta(s)))^2] = \\ \frac{1}{m_l}\sum_{j=1}^{N}(y_s - L_c(s,\mu_\theta(s)))^2. \end{aligned} \tag{15}$$

这里: m_l 为每个批次从危险记忆池中抽取的样本数量, y_s 为目标代价函数. 最后, 对拉格朗日乘子的值进行调整, 利用策略梯度方法使得以下目标最大化:

$$J_c(\xi) = \xi E_{U_D}[L_c(s', a') - (L_c(s, a) - k \cdot c(s, a))]. \tag{16}$$

随着网络的更新, 拉格朗日乘子 ξ 被不断优化, 保证了动作输出的安全性.

3　实验仿真和验证

3.1　实验初始化设置

本节将在安全强化学习环境中对所提出算法的性能表现进行训练和测试. 场景基于专业的机器人物理仿真器 CoppeliaSim 建立, 主要面向多机器人编队导航和目标搜索任务. 实验过程中, 所提出多智能体事件触发式分层安全强化学习方法 (MEHSRL) 各网络超参数定义如表 1 所示. 双层网络的切换的触发事件为李雅普诺夫状态函数 $L(s)$.

表 1　网络参数

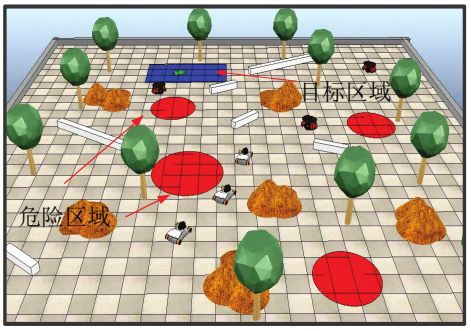
参数	介绍	数值
l_a	策略网络学习率	0.000 1
l_c	评价网络学习率	0.000 1
l_s	李雅普诺夫网络	0.000 1
Batch size	单批抽样本数量	512
γ	折扣因子	0.99
τ_1, τ_2, τ_3	网络软更新因子	0.01
U_D, U_s	经验回放池容量	10^6

表 1 中: l_a、l_c 和 l_s 分别为策略网络、评价网络和李雅普诺夫网络的学习率; U_D 和 U_s 为经验池容量, 用来存储机器人与环境的交互样本; Batch size 表示每个回合从经验池中采样并用于梯度计算的样本数量; γ 为折扣因子, 表示对未来奖励和未来代价的关注; τ_1、τ_2 和 τ_3 为目标网络的软更新因子.

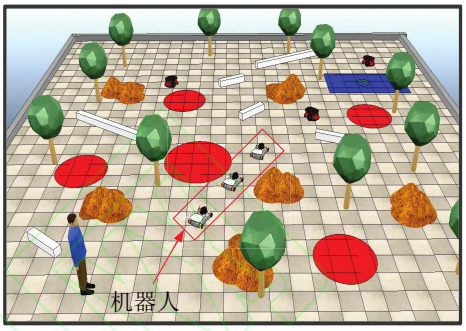
3.2　安全强化学习训练场景

为了测试多机器人在安全强化学习场景中的运动规划效果, 建立了机器人编队导航运动规划场景, 如图 3 所示. 环境分为训练场景和测试场景, 其中包括 3 组机器人和多个动静态障碍物. 首先在训练场景中进行策略学习, 待策略收敛后, 迁移至测试场景中对机器人的运动规划效果进行测试. 训练过程中, 多机器人需要导航至目标区域, 并保持一定的队形, 通过布满障碍和危险的区域. 机器人的目的地位于目标区域内, 每个训练回合结束后机器人的目标位置均会实时发生变化.

在多机器人场景中, 分别基于所提出方法 (MEHSRL)、信赖域方法 CPO 和柔性动作-评价软更新方法 (SAC) 进行多个轮次的训练, 对比分析它们的



(a)　训练场景



(b)　测试场景

图 3　多机器人安全强化学习环境

协同作业能力以及安全性的提升效果. 其中: CPO 为安全强化学习方法; SAC 为普通强化学习方法, 没有安全约束条件. 训练过程中, 机器人将自身状态和环境状态作为输入, 并以此获得动作输出. 策略训练共进行了 10 000 个回合, 每个回合最大训练步数为 500. 若机器人陷入局部最优, 则当达到 500 步时, 该回合将被强制性结束. 训练完成后, 得出了不同场景的平均回合奖励变化情况、平均回合代价函数变化情况和任务成功率. 具体结果如图 4 和图 5 所示.

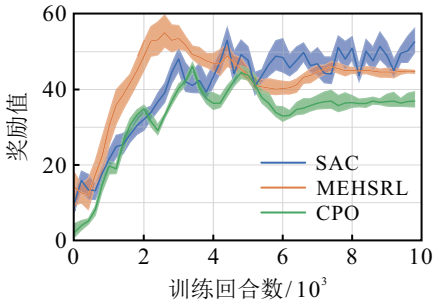


图 4　训练环境中不同方法的奖励值情况

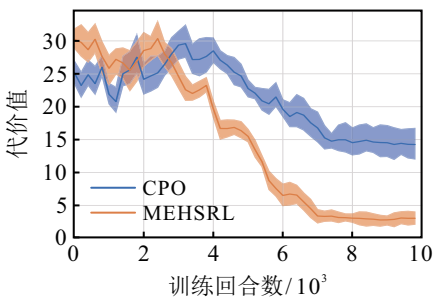


图 5　训练环境中不同方法的代价值情况

由图4可见,所提出方法在场景中奖励值表现稳定. 在3 000轮次以前,奖励值较快上升,且高于其他两种方法. 在3 000轮次后,由于安全李雅普诺夫函数的作用,机器人开始离开最高奖励区域,选择更加安全的策略,奖励值出现了稍微降低. 同样的情况,也出现在CPO方法中,这是安全强化学习方法中代价函数的约束作用. 3 000轮次后奖励值基本达到了收敛状态,所提出方法的奖励值大于CPO安全强化学习方法,但是略低于SAC方法. SAC算法的最终奖励值虽然最高,但是策略以追求最大奖励为唯一目标,没有安全李雅普诺夫评价函数的约束,从而忽略了对危险区域的躲避,造成了任务成功率的降低. 因此,SAC强化学习方法的安全性低于另外两种安全强化学习方法.

由图5可见,所提出MEHSRL方法的代价值可顺利收敛至安全阈值以下,且明显低于CPO方法. 代价值的差异表明所提出方法的安全约束性能优于CPO方法.

在实验训练中,3个机器人需要保持一定的队形进行导航,并通过不断调整自身姿态,保证自身速度稳定在设定速度. 为了分析所提出方法对机器人的自主编队的规划能力,实验记录了机器人的自身运动状态的变化情况,具体如图6和图7所示.

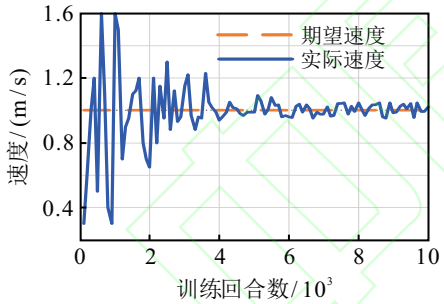


图 6　多机器人的编队速度变化

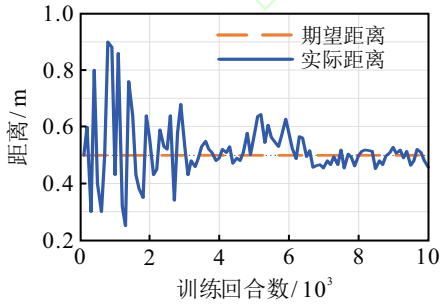


图 7　多机器人的编队距离变化

由图6和图7可见,机器人的速度在前期有较大的波动. 但是随着训练的进行,波动振幅开始逐渐减小,在约4 000回合后,机器人的速度逐渐趋近设定速度,且基本上可稳定在设定速度附近. 同样地,机器人间的平均距离也得到了较好的控制,在策略收敛后,机器人间的距离可稳定在期望距离附近,从而可以保持设定的队形完成导航任务. 从运动状态的变化趋势可以看出所提出方法具有较好的多机器人编队能力.

在完后训练后,策略模型将被迁移至测试环境中进行测试. 相比于训练环境,在测试环境中增加了动态障碍物,并更改了静态障碍物的布局. 通过2 000轮次的测试实验后,实验统计了3种方法的奖励值、代价值和任务成功率的变化情况. 实验结果如图8和图9所示.

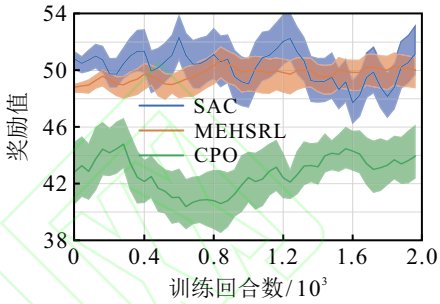


图 8　测试环境中不同方法的奖励值情况

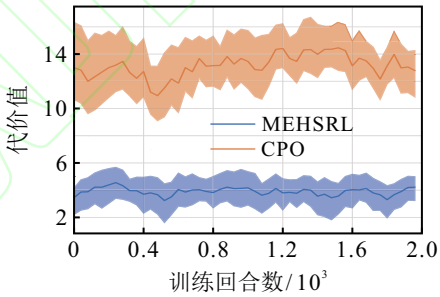


图 9　测试环境中不同方法的代价值情况

图8为3种方法在测试场景中的奖励值变化情况. 由图8可见,3种强化学习方法的奖励值表现均较为稳定,运动策略基本达到了收敛状态. 所提出MEHSRL方法的奖励值稳定在50上下变化;而普通安全强化学习CPO方法的奖励值仅为42左右;反观没有安全约束的SAC方法,其奖励值平均值略大于MEHSRL方法,但是却呈现出较大的波动性,预示着运动规划过程中动作决策进入危险状态的次数较多,任务成功率将受到较大影响.

图9为两种安全强化学习方法的代价值收敛情况. 代价值越低,表明对于动作安全性约束效果越好. 由图9可见:在整个测试过程中,所提出算法代价值始终几乎稳定在5以下;而CPO方法的代价值较高,处于13左右. 这表明在测试过程中所提出方法的涉险次数远低于CPO方法. 同时,实验记录了运动规划过程中的任务成功率,如图10所示.

由图10可见,所提出方法的任务成功率为96.80%,远大于其他两种方法. 综合多种实验结果,

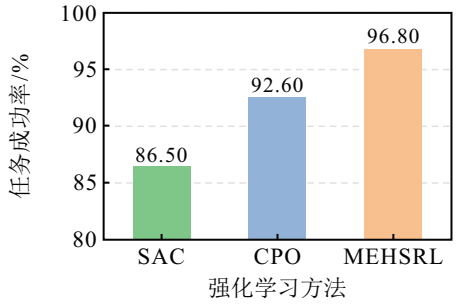


图 10　测试环境中不同方法的任务成功率

通过分析对比不同运动规划方法的任务成功率、奖励值和代价值的差异可以明显得出:所提出多智能体事件触发式分层安全强化学习方法MEHSRL在复杂的非安全环境中展现出较好的安全性和稳定性,机器人受到非安全因素影响较小,代价值和奖励值可保持在一个合理的区域内,机器人的导航、编队和搜索的任务成功率得到了全面提高,增强了在动态复杂的非安全环境中良好的运动规划能力和环境的适应性.

4　结　论

本文提出了一种基于事件触发的多智能体分层安全强化学习运动规划方法(MEHSRL)来改善多机器人运动规划中安全约束不足的问题. 所提出方法基于受限马尔可夫决策过程,建立了分层多智能体安全强化学习框架,实现了策略网络的快速学习与安全约束的平衡;通过引入李雅普诺夫代价函数网络,构建了带有条件约束的动作策略优化目标,同时利用拉格朗日乘子法优化了策略的求解过程,保证了状态轨迹可在有限时间内从危险状态中恢复至安全空间. 通过与其他方法的实验对比结果可以发现,所提出运动规划方法任务执行成功率高、安全性强,为保证复杂环境下多机器人安全运动规划提供了理论参考.

参考文献(References)

[1] 温广辉, 杨涛, 周佳玲, 等. 强化学习与自适应动态规划: 从基础理论到多智能体系统中的应用进展综述[J]. 控制与决策, 2023, 38(5): 1200-1230.
(Wen G H, Yang T, Zhou J L, et al. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems[J]. Control and Decision, 2023, 38(5): 1200-1230.)

[2] 夏家伟, 朱旭芳, 张建强, 等. 基于多智能体强化学习的无人艇协同围捕方法[J]. 控制与决策, 2023, 38(5): 1438-1447.
(Xia J W, Zhu X F, Zhang J Q, et al. Research on cooperative hunting method of unmanned surface vehicle based on multi-agent reinforcement learning[J]. Control and Decision, 2023, 38(5): 1438-1447.)

[3] 隋丽蓉, 高曙, 何伟. 基于多智能体深度强化学习的船舶协同避碰策略[J]. 控制与决策, 2023, 38(5): 1395-1402.
(Sui L R, Gao S, He W. Ship cooperative collision avoidance strategy based on multi-agent deep reinforcement learning[J]. Control and Decision, 2023, 38(5): 1395-1402.)

[4] 赵莉, 李炜, 李亚洁. 自适应事件触发通信机制下机理解析与数据驱动融合的ICPS双重安全控制[J]. 控制与决策, 2024, 39(1): 206-218.
(Zhao L, Li W, Li Y J. Dual security control based on fusion of mechanism analysis and data-driven under adaptive event-triggered communication scheme for ICPS[J]. Control and Decision, 2024, 39(1): 206-218.)

[5] Konar A, Baghi B H, Dudek G. Learning goal conditioned socially compliant navigation from demonstration using risk-based features[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 651-658.

[6] 王宇霄, 刘敬玉, 李忠飞, 等. 基于强化学习与安全约束的自动驾驶决策方法[J]. 交通运输研究, 2023, 9(1): 31-39.
(Wang Y X, Liu J Y, Li Z F, et al. An autonomous driving decision making method based on reinforcement learning and safety constraints[J]. Transport Research, 2023, 9(1): 31-39.)

[7] Riley J, Calinescu R, Paterson C, et al. Utilising assured multi-agent reinforcement learning within safety-critical scenarios[J]. Procedia Computer Science, 2021, 192: 1061-1070.

[8] Shi L, Wang X S, Cheng Y H. Safe reinforcement learning-based robust approximate optimal control for hypersonic flight vehicles[J]. IEEE Transactions on Vehicular Technology, 2023, 72(9): 11401-11414.

[9] Brunke L, Greeff M, Hall A W, et al. Safe learning in robotics: From learning-based control to safe reinforcement learning[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5: 411-444.

[10] Zhang J X, Liu W, Li Y M. Optimal formation control for second-order multi-agent systems with obstacle avoidance[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(2): 563-565.

[11] Sreenivas N K, Rao S. Safe deployment of a reinforcement learning robot using self stabilization[J]. Intelligent Systems with Applications, 2022, 16: 200105.

[12] Hsu K C, Ren A Z, Nguyen D P, et al. Sim-to-Lab-to-Real: Safe reinforcement learning with shielding and generalization guarantees[J]. Artificial Intelligence, 2023, 314: 103811.

[13] Hu Y F, Fu J J, Wen G H. Safe reinforcement learning for model-reference trajectory tracking of uncertain autonomous vehicles with model-based acceleration[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(3): 2332-2344.

[14] 王康, 李琼琼, 王子洋, 等. 考虑侧倾的无人车NMPC轨迹跟踪控制[J]. 控制与决策, 2022, 37(10): 2535-2542.

(Wang K, Li Q Q, Wang Z Y, et al. Trajectory tracking control for automated vehicle based on NMPC considering vehicle rolling motion[J]. Control and Decision, 2022, 37(10): 2535-2542.)

[15] Brunke L, Greeff M, Hall A W, et al. Safe learning in robotics: From learning-based control to safe reinforcement learning[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5: 411-444.

[16] Liu S H, Liu L J, Yu Z. Safe reinforcement learning for discrete-time fully cooperative games with partial state and control constraints using control barrier functions[J]. Neurocomputing, 2023, 517: 118-132.

[17] Li T X, Zhu K, Luong N C, et al. Applications of multi-agent reinforcement learning in future Internet: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2022, 24(2): 1240-1279.

[18] Mavrogiannis C I, Knepper R A. Multi-agent path topology in support of socially competent navigation planning[J]. The International Journal of Robotics Research, 2019, 38(2/3): 338-356.

[19] 王雪松, 王荣荣, 程玉虎. 安全强化学习综述[J]. 自动化学报, 2023, 49(9): 1813-1835.
(Wang X S, Wang R R, Cheng Y H. Safe reinforcement learning: A survey[J]. Acta Automatica Sinica, 2023, 49(9): 1813-1835.)

[20] Yang Y J, Jiang Y X, Liu Y C, et al. Model-free safe reinforcement learning through neural barrier certificate[J]. IEEE Robotics and Automation Letters, 2023, 8(3): 1295-1302.

[21] Liu Y Q, Gao Y F, Zhang Q C, et al. Multi-task safe reinforcement learning for navigating intersections in dense traffic[J]. Journal of the Franklin Institute, 2023, 360(17): 13737-13760.

[22] 刘俊辉, 单家元, 荣吉利, 等. 自适应学习率的增量强化学习飞行控制[J]. 宇航学报, 2022, 43(1): 111-121.
(Liu J H, Shan J Y, Rong J L, et al. Incremental reinforcement learning flight control with adaptive learning rate[J]. Journal of Astronautics, 2022, 43(1): 111-121.)

[23] 陈晋音, 章燕, 王雪柯, 等. 深度强化学习的攻防与安全性分析综述[J]. 自动化学报, 2022, 48(1): 21-39.
(Chen J Y, Zhang Y, Wang X K, et al. A survey of attack, defense and related security analysis for deep reinforcement learning[J]. Acta Automatica Sinica, 2022, 48(1): 21-39.)

[24] 李保罗, 蔡明钰, 阚震. 线性时序逻辑引导的安全强化学习[J]. 控制与决策, 2023, 38(7): 1835-1844.
(Li B L, Cai M Y, Kan Z. Linear temporal logic guided safe reinforcement learning[J]. Control and Decision, 2023, 38(7): 1835-1844.)

[25] 代珊珊, 刘全. 基于动作约束深度强化学习的安全自动驾驶方法[J]. 计算机科学, 2021, 48(9): 235-243.
(Dai S S, Liu Q. Action constrained deep reinforcement learning based safe automatic driving method[J]. Computer Science, 2021, 48(9): 235-243.)

[26] Fujimoto S, Hoof H V, Meger D. Addressing function approximation error in actor-critic methods[C]. International Conference on Machine Learning. Stockholm, 2018: 1587-1596.

作者简介

孙辉辉 (1989–), 男, 副教授, 博士, 从事智能机器人及其控制方法、多机器人协同运动控制等研究, E-mail: cumtsunhui@126.com;

胡春鹤 (1986–), 男, 副教授, 博士, 从事无人机自主控制、多无人机协同控制及其应用等研究, E-mail: huchunhe@bjfu.edu.cn;

张军国 (1978–), 男, 教授, 博士生导师, 从事智慧林业监测与信息处理、无人飞行器及林业特种机器人等研究, E-mail: zhangjunguo@bjfu.edu.cn.