中国科学: 技术科学 2024 年 第 54 卷 第 10 期: 1991 ~ 2002
SCIENTIA SINICA Technologica techcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

论 文

基于深度强化学习的多智能体分布式事件触发优化控制

李方昱[1,2,3,4], 刘金溢[1,2,3], 黄琰婷[1,2,3,4], 韩红桂[1,2,3,4*]

1. 北京工业大学信息学部, 北京 100124;
2. 北京工业大学数字社区教育部工程研究中心, 北京 100124;
3. 北京工业大学计算智能与智能系统北京市重点实验室, 北京 100124;
4. 北京人工智能研究院, 北京 100124
* E-mail: rechardhan@bjut.edu.cn

收稿日期: 2023-11-01; 接受日期: 2024-01-05; 网络版发表日期: 2024-10-09
国家重点研发计划(编号: 2023YFB3307300)、国家自然科学基金项目(批准号: 62373014)、国家自然科学基金重大计划培育项目(编号: 92267107)、北京市自然科学基金项目(编号: KZ202110005009)和北京市青年学者基金项目(编号: 037)资助

摘要 为了解决多智能体系统在有障碍物的场景下协同运动控制中无线通信成本较高的问题, 本文设计了一种分布式事件触发优化控制方法. 该方法包括计算通信与控制的联合策略以及确定多智能体之间通信的触发条件, 使多智能体无需实时或按周期地进行通信, 从而有效降低数据传输总量, 实现优化通信机制, 达到降低通信成本的目的; 同时为了使智能体能够避开其他智能体与障碍物, 设计了一种基于指数函数的碰撞惩罚项, 使智能体在接近障碍物或其他智能体的过程中受到按照指数式增大的惩罚, 避免发生碰撞. 将该方法应用于多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)算法中, 在添加障碍物的多智能体粒子环境(multi-agent particle environment, MPE)上的仿真实验结果表明, 该方法可以在较好地完成多智能体协同控制任务的同时减少数据传输量, 达到了优化通信机制、降低通信成本的目的.

关键词 多智能体, 强化学习, 分布式事件触发控制, 协同运动控制

1 引言

多智能体协同运动控制是当今多智能体协同控制领域的一个热门话题, 在自主仓储物流、协同救援、协同探测等方面都有着丰富的应用价值[1]. 大多数多智能体协同运动控制都依赖于无线通信来共享观测结果. 然而, 传统的通信方式为连续通信或者周期性通信, 这对网络的带宽有很高的要求, 一定程度上也会造成能量的浪费, 提高了通信的成本[2]. 此外, 随着智能体和通信数据数量的增加, 由于数据包丢失和通信延迟, 会影响协同控制的性能[3]. 一些研究表明, 智能体的观测值虽然在不断变化, 但是从观测值中提取的有效信息非常相似, 导致一些智能体之间的通信和数据传输往往是冗余的[4]. 因此, 优化智能体之间的通信

引用格式: 李方昱, 刘金溢, 黄琰婷, 等. 基于深度强化学习的多智能体分布式事件触发优化控制. 中国科学: 技术科学, 2024, 54: 1991–2002
Li F Y, Liu J Y, Huang Y T, et al. Multi-agent distributed event-triggered optimization control based on deep reinforcement learning (in Chinese). Sci Sin Tech, 2024, 54: 1991–2002, doi: 10.1360/SST-2023-0346

机制，减少智能体间通信的同时使系统保持较高的协同控制性能是至关重要的[5].

为了优化多智能体系统的通信机制，达到降低通信成本的目的，一些学者将时间触发控制方法应用于多智能体系统中，Zhang和Tian[6]提出了时间触发算法来解决丢包和随机通信延迟问题. 此外，Zhang等人[7]提出了一种时间信息控制方法，通过比较智能体相邻时刻发送信息的相似度，从而判断智能体在当前时刻是否需要通信，进而大幅度减少智能体之间交换的信息量. 然而，时间触发策略需要在预先设定的时刻进行智能体之间的通信交流，这些时刻通常是由固定间隔分开的，无法确定当前时刻智能体之间是否需要通信, 可能会导致冗余的通信产生[8].

近年来的一些研究表明，与通常的时间触发方法相比，事件触发方法可以以显著更低的样本数量实现高性能控制[9]. Dimarogonas等人[10]在多智能体通信中引入了事件触发控制，每个智能体计算下一个更新时间，以确定与相邻智能体通信的定时. Trimpe和D'Andrea[11]提出了基于事件的分布式状态估计算法，以根据实际测量值和估计值之间的误差来确定定时和传输数据. Dohmann和Hirche[12]提出了一种基于事件触发通信的分布式控制策略，用于协同轨迹跟踪和抓取. 该方法在最大限度地减少从相邻智能体接收末端执行器的位置和速度的频率的同时能够很好地完成协同任务. 尽管上述方法能够实现减少智能体间通信的同时实现高性能的控制，然而，这些方法需要动力学模型，并且不能应用于较为复杂智能体系统中. 相比于有模型的方法，将强化学习应用到多智能体系统中的方法是无模型的，因此可以应用到更多的协同任务中[13].

一些研究将深度强化学习与事件触发控制方法相结合, 应用于智能体系统中. Baumann等人[14]将事件触发控制应用于深度强化学习中，将系统相邻通信时刻的状态偏移量作为事件触发条件，基于DDPG算法的策略网络计算通信决策和控制输入，通过判断通信决策是否满足触发条件，对系统的控制输入进行更新. Shibata等人[15]将该项研究扩展到了多智能体系统中, 设计了一种事件触发通信与控制的联合策略，有效地解决了多智能体协同运输中通信和控制策略的同步设计问题, 实现了将有效载荷运送到期望位置. Shibata等人[3]在后续的研究中对其提出的方法进行了优化，使多智能体系统在训练过程与实际应用场景中智能体数量不同的情况下，仍然能够降低通信成本并保持着良好的协同运输性能. Kesper等人[4]使用分层强化学习算法，使分布式事件触发控制方法能够适用于高维观测空间，避免因智能体数量过多而造成的维度灾难问题. Hu等人[13]设计了一种事件触发通信网络ETCNet, 包含事件触发发送网络ETSNet和事件触发接收网络ETRNet, 将有限带宽转换为事件触发策略的惩罚阈值决定智能体在每个时刻是否参与通信，从而保持了多智能体系统的协作性能并降低带宽. 尽管以上学者进行的研究能够实现对于通信成本的优化，然而，上述研究的实验场地均在空旷的场景下，而在实际的多智能体协同任务场景中，地形往往是复杂且有障碍物的, 因此，必须设计一种策略使智能体在实施协同任务的过程中能够拥有避障的能力.

为了解决多智能体系统避障的问题，一些学者在强化学习的奖励函数部分加入了碰撞障碍物的惩罚项. Wang等人[16]在奖励函数中设计了一种基于指数函数的智能体靠近障碍物的惩罚项，该方法使智能体越靠近障碍物惩罚越大，从而促使智能体在学习的过程中躲避障碍物. Li等人[17]设计了一种基于多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)的多智能体避障算法, 基于人工势场法对靠近智能体的障碍物进行惩罚. 然而, 上述研究采用集中式训练分布式执行的方式，并没有考虑智能体之间的通信，使得这些方法在一些多智能体协同任务中有一定的局限性. 综上，设计一种实现智能体在协同运动控制中降低通信成本，不影响协同控制性能的同时拥有避障能力的方法仍是一个挑战性难题.

综上所述, 本文的贡献可归纳为如下两点.

(1) 设计了一种基于深度强化学习的多智能体事件触发控制方法，该方法包含确定智能体之间通信的触发条件以及计算通信与控制的联合策略，能够优化多智能体系统的通信机制，实现了在不影响协同控制性能的同时极大地减少数据传输总量，达到了降低通信成本的目的.

(2) 设计了一种平衡控制性能和通信节省的奖励函数，对智能体靠近障碍物进行奖励，并对智能体间通信进行惩罚. 此外，设计了一种基于指数函数的碰撞惩罚项，对智能体靠近障碍物或其余智能体做出的

惩罚, 成功解决了多智能体系统避障问题.

2　背景知识

2.1　马尔可夫决策过程

马尔可夫决策过程(Markov decision process, MDP)可以用来对强化学习问题进行建模[18]. MDP指的是一个智能体根据策略采取相应动作与外部环境发生交互, 从而改变自己的状态和策略, 以最大化期望奖励的循环过程.

MDP可定义为一个四元组(S, A, R, P). 其中, S为智能体与环境交互过程中状态的集合, s表示智能体的某个特定的状态($s\in S$), 在文中表示智能体的位置、速度以及与其他智能体和障碍物之间的相对距离等信息的集合. A为有限动作的集合, a表示某个特定动作($a\in A$), 在文中表示智能体的控制输入和通信决策的集合. R为奖励函数, $R(s, a)$表示在当前状态s的情况下执行动作a获得的奖励值. P为状态转移函数, $P(s'|s, a)$表示基于当前状态s采取动作a转移到状态s'的概率. $\pi(a|s)$表示智能体状态空间S和动作空间A对应的映射策略, p为在当前状态s下采取动作a的概率, 可以公式化为

$$\pi(a\mid s)=p(a_t=a\mid s_t=s). \tag{1}$$

强化学习通过最大化长期总回报来学习和获得最优策略, 即要求MDP获得期望最大化的累积奖励[19]. 定义动作价值函数$Q^{\pi}(s_t, a_t)$为

$$\begin{aligned}Q^{\pi}(s_t,a_t)&=E_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\mid s_t=s,a_t=a\right]\\&=E_{\pi}\left[r_{t+1}+\gamma Q^{\pi}(s_{t+1},a_{t+1})\mid s_t=s,a_t=a\right],\end{aligned} \tag{2}$$

式中, r_{t+1}为智能体在t时刻做出动作a_t到达状态s_t获得的奖励, γ为折扣因子, E_{π}为智能体的策略为π时的总回报的期望. Q^{π}函数描述的是在某一个状态采取某一个动作, 它有可能得到的总回报的期望值.

由式(2)可得到最优动作价值函数$Q^{*}(s_t, a_t)$, 其计算公式为

$$\begin{aligned}Q^{*}(s_t,a_t)&=\max\nolimits_{\pi}Q^{\pi}(s_t,a_t)\\&=E^{*}\left[r_{t+1}+\gamma\max\nolimits_{\pi}Q^{*}(s_{t+1},a_{t+1})\mid s_t=s,a_t=a\right].\end{aligned} \tag{3}$$

多智能体深度强化学习的目标是对每个智能体给定一个MDP, 使其寻找最优的策略, 做出正确的动作到达期望的状态. 其中策略为状态到动作的映射, 正确的策略和动作会使智能体最终的累计回报最大化. 回报为智能体从每回合的初始状态到最终状态的奖励值衰减之和.

2.2　多智能体深度确定性策略梯度算法

传统的强化学习算法(如Q学习、利用Q值的存储和更新等方法)采用矩阵形式. 然而, 在复杂的多智能体状态空间下, 矩阵形式的计算代价过大, 效率较低, 而利用深度神经网络等非线性函数逼近器来近似地表示值函数或策略的深度强化学习框架能够解决此问题. MADDPG算法可以使多智能体在一个共享环境中学习, 集中式训练和分散式执行的方式使每个智能体可以根据其他智能体的动作和观测信息来改进自己的策略, 能够很好地实现多智能体系统的训练任务[20]. 其伪代码如算法1所示.

3　多智能体分布式事件触发控制

3.1　多智能体协同运动控制的深度强化学习设置

假设在多智能体运动控制任务中, 有N个智能体在二维的矩形地图区域内, 区域内有N_{ob}个障碍物, 每个智能体有M组观测信息. 多智能体系统在t时刻的深度强化学习变量如表1所示.

定义t时刻智能体i接收智能体j的观测信息为$o_i^j=\left[o_i^{j1},\dots,o_i^{jM}\right]^{\mathrm{T}}$; 智能体$i$自身持有的观测信息为$o_i^i=\left[o_i^1,\dots,o_i^M\right]^{\mathrm{T}}$; 最终输入到智能体$i$的观测信息为$O_i=\left[o_i^{1\mathrm{T}},\dots,o_i^{N\mathrm{T}}\right]^{\mathrm{T}}$. 对上述设置进行分析, 可知: 智能体$i$的观测信息包含绝对位置、速度、与其他智能体的相对位置、与障碍物的相对位置、期望位置、控制输入以及通信决策, 分别用x_i, v_i, x_i^a, x_i^{ob}, x_i^*, u_i以及c_i和d_i来表示, 与o_i^i中的每一项对应. 智能体i的策略和动作分别为π_i和a_i, 获得的奖励为r_i.

本文的目的是控制每个智能体的运动, 使其到达期望位置, 在此过程中需要躲避障碍物和其他智能体, 优化多智能体系统的通信机制, 减少智能体之间在每

算法1: 多智能体深度确定性策略梯度MADDPG

策略网络和评价网络的参数分别初始化为θ^{Q}和θ^{π}

初始化经验回放存储区D、状态s

设置经验回放区最大规模为N_r, 训练批次规模为N_b, 以及目标网络的更新频率为N_n

设置回合数episode最长为K, 每回合最大时间步长t为T, 智能体数量为N

for episode = 1 to K do

获取当前时刻N个智能体的状态s, $s = (O_1, \cdots, O_N)$

for t = 1 to T do

for i = 1 to N do

智能体i按照当前策略和观测生成动作$a_i = \pi_i(O_i)$

end for

多智能体执行$a = (a_1, \cdots, a_N)$, 得到$r = (r_1, \cdots, r_N)$和s'

将(s, a, r, s')保存至D中, 当$|D| \geq N_r$时, 替换旧的数据

$s \leftarrow s'$

for i = 1 to N do

从D中随机采样规模为N_b的数据(s^j, a^j, r^j, s'^j)

更新当前评价网络的参数, 定义损失函数L为

$$L(\theta_i) = \frac{1}{N_b}\sum_j\left(y^j - Q_i^{\pi}\left(s^j, a_1^j, a_2^j, \ldots, a_N^j\right)\right)^2$$

其中

$$y^j = r_i^j + \gamma Q_i^{\pi'}(s'^j, a_1', a_2', \ldots a_N') \Big|_{a_k' = \pi_k'(O_k^j)}$$

更新当前策略网络的参数, 定义评价函数$J(\theta_i)$:

$$\nabla_{\theta_i} J(\theta_i) \approx \frac{\sum_j \nabla_{\theta_i}\pi_i(o_i^j)\nabla_{a_i}Q_i^{\pi}\left(s^j, a_1^j, \ldots, a_i, \ldots, a_N^j\right)}{N_b}$$

其中

$$a_i = \pi_i(O_i^j)$$

end for

每N_n步对于每个智能体i, 软更新目标网络:

$$\begin{cases}\theta_i^{Q'} \leftarrow \tau\theta_i^{Q} + (1-\tau)\theta_i^{Q'}\\ \theta_i^{\pi'} \leftarrow \tau\theta_i^{\pi} + (1-\tau)\theta_i^{\pi'}\end{cases}$$

end for

end for

个回合传输的数据总量.

3.2 分布式事件触发控制方法架构

本文设计了一种分布式事件触发控制方法, 并将其用于多智能体协同运动控制场景中. 其在MADDPG算法中的应用架构如图1(a)所示.

分布式事件触发控制方法的触发条件部分由两个步骤组成: 首先判断智能体i是否满足与智能体j通信的触发条件, 其次判断是否满足接收来自智能体j的第m组观测数据的触发条件. 在基于分布式事件触发控制的多智能体强化学习算法中, 智能体i在当前时刻所要通信的智能体和接收的数据由上一时刻计算的通信决策所决定, 方法架构如图1(b)所示. 事件触发条件部分定义了2个变量, 分别为控制智能体i是否与智能体j通信的cr_i^j, 以及智能体i是否接收来自智能体j的第m组观测值o_j^m的dr_i^m. 其中, 智能体i始终接收自身的观测信息, o_i^{i1}至o_i^{iM}以及cr_i^i的门控开关始终闭合. 通过将这些变量引入分布式事件触发控制中, 实现对智能体系统通信机制的优化, 触发条件的计算公式为

表 1 多智能体协同运动控制的深度强化学习变量设置

Table 1 Deep reinforcement learning variables setting for multi-agent cooperative motion control

变量	设置
智能体的绝对位置	x
智能体的速度	v
智能体与其他智能体的相对位置	x^a
智能体与障碍物的相对位置	x^{ob}
智能体的期望位置	x^*
智能体的控制输入和通信决策	u和c, d
智能体获得的奖励	r
智能体执行的策略和动作	π和a

$$cr_i^j = \begin{cases}1, & \text{if } c_i^j > \lambda_1,\\ 0, & \text{else},\end{cases} \tag{4}$$

$$dr_i^j = \begin{cases}1, & \text{if } d_i^m > \lambda_2,\\ 0, & \text{else},\end{cases} \tag{5}$$

式中, λ_1和$\lambda_2 \in \mathbf{R}$, c_i^j和d_i^m均由联合策略网络计算得出.

基于上述事件触发条件, 最终输入到智能体i的观测信息O_i的元素可更新为

$$o_i^{jm} \leftarrow \begin{cases}o_j^m, & \text{if } cr_i^j \cdot dr_i^m = 1,\\ -1, & \text{else}.\end{cases} \tag{6}$$

将智能体i自身和所要接收的其他智能体的观测信息输入到联合策略π_i中, 智能体i的联合策略计算当前时刻的控制输入和通信决策, 随后将控制输入到智能体中, 并将通信决策作用于下一时刻. 通信与控制

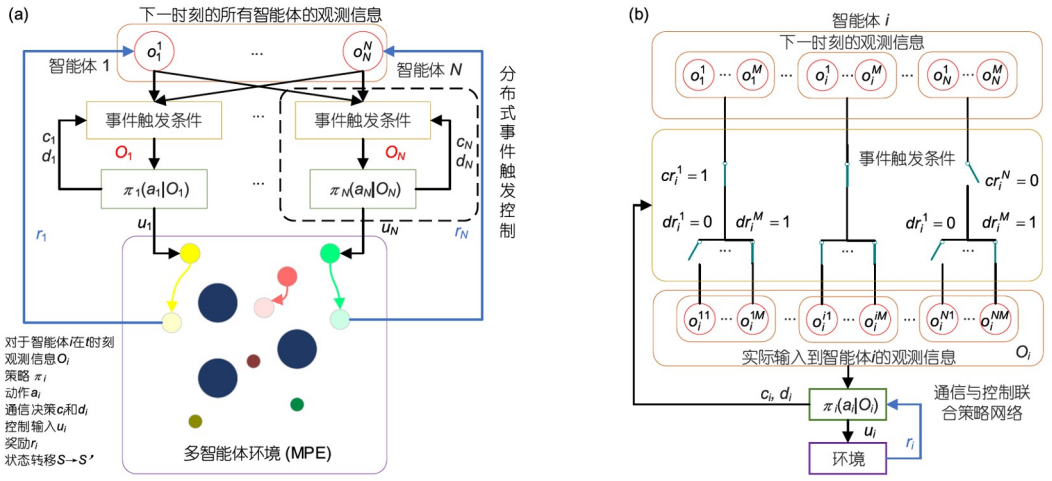

图 1　基于深度强化学习的多智能体分布式事件触发控制算法架构. (a) 分布式事件触发控制方法在MADDPG算法中的应用架构; (b) 分布式事件触发控制方法架构

Figure 1　Multi-agent distributed event-triggered control algorithm architecture based on deep reinforcement learning. (a) Architecture of distributed event-triggered control in MADDPG algorithm; (b) architecture of distributed event-triggered control.

联合策略计算公式为

$$\pi_i(u_i, c_i, d_i \mid O_i) = \pi_i(a_i \mid O_i). \tag{7}$$

分布式事件触发控制方法的通信与控制联合策略由深度神经网络DNN计算得出, 通过优化策略网络的结构和参数, 实现奖励值的快速收敛. 将该方法应用到MADDPG算法中, 其伪代码如算法2所示.

3.3　奖励函数设计

在深度强化学习中, 奖励函数的设计直接影响学习效果的好坏, 是影响多智能体协同运动控制性能的关键部分[21]. 奖励函数包含了对于智能体要学习的任务的量化描述, 指导训练的智能体向期望的方向学习[22].

在多智能体协同运动控制问题中, 希望智能体能够尽快到达各自的期望位置, 同时要求智能体在移动过程中能够避免与障碍物或其他智能体碰撞. 对于一般的多智能体协同运动控制任务, 一种简单的方法是智能体只有在到达期望位置时才会得到奖励[23]. 这种方法不适用于复杂的环境, 由于初始策略是随机生成的, 智能体将在充满障碍的复杂环境中以极小的概率到达目的地, 因此, 强化学习算法需要较长的时间才能收敛, 甚至无法收敛. 另一种方法是在起始位置与期望位置的过程中设置奖励[24], 这样可以更好地引导

算法2: 基于深度强化学习的多智能体分布式事件触发控制

for episode =1 to K do

　获取当前时刻N个智能体的状态s, $s = (O_1, \cdots, O_N)$

　for t = 1 to T do

　　for i = 1 to N do

　　智能体i按照当前策略和观测生成动作$a_i = \pi_i(O_i)$

　　通信与控制联合策略计算控制输入和通信决策:

$$\pi_i(u_i, c_i, d_i \mid O_i) = \pi_i(a_i \mid O_i)$$

　　智能体i根据当前的通信决策cr_i^j判断是否要与智能体j通信:

$$cr_i^j = \begin{cases} 1, & \text{if } c_i^j > \lambda_1 \\ 0, & \text{else} \end{cases}$$

　　智能体i根据当前的通信决策dr_i^j判断是否要接收智能体j的第m项观测信息:

$$dr_i^j = \begin{cases} 1, & \text{if } d_i^m > \lambda_2 \\ 0, & \text{else} \end{cases}$$

　　更新O_i中的元素:

$$o_i^{jm} \leftarrow \begin{cases} o_j^m, & \text{if } cr_i^j \cdot dr_i^m = 1 \\ -1, & \text{else} \end{cases}$$

　　end for

　　多智能体执行$a = (a_1, \cdots, a_N)$, 得到$r = (r_1, \cdots, r_N)$和s'

　　将(s, a, r, s')保存至D中, $|D| \geq$　N_r时, 替换旧的数据

　　$s \leftarrow s'$

　　调用MADDPG算法对多智能体系统进行训练

　end for

end for

智能体前往期望位置并避开其他智能体和障碍物，从而更好地引导智能体向期望回报最大的方向学习. 因此，本文基于第二种方法设计了一种奖励函数，使智能体在运输性能和通信节省之间取得平衡. 该奖励函数由3部分组成，分别为协同运动控制过程的奖励、碰撞障碍物或智能体的惩罚以及智能体之间相互通信的惩罚. 其中，智能体i的协同运动控制过程的奖励项设计为

$$r_i^{\mathrm{mot}} = -\sigma\left\|x_i^* - x_i\right\|_2, \tag{8}$$

式中, $\sigma > 0$, $\left\|x_i^* - x_i\right\|_2$是智能体与期望位置之间的直线距离. 该部分对智能体与期望位置的距离进行惩罚, 即相当于对智能体靠近期望位置进行正向奖励，鼓励智能体前往目的地，并在它们远离目标时对其进行惩罚. 此外, 在实际的协同运动控制任务中, 与障碍物相撞对智能体来说可能是灾难性的. 为了防止智能体离其他智能体或障碍物太近导致发生碰撞，本文将碰撞障碍物或智能体的惩罚项设计为

$$r_i^{\mathrm{col}} = -\sum_{k}^{N_{ob}}\sum_{j=1,j\neq i}^{N}\alpha\left(\mathrm{e}^{-\beta\left\|x_j - x_i\right\|_2} + \mathrm{e}^{-\beta\left\|x_k^{ob} - x_i\right\|_2}\right), \tag{9}$$

式中, $\alpha, \beta > 0$; $\left\|x_j - x_i\right\|_2$是2个不同智能体之间的直线距离; $\left\|x_k^{ob} - x_i\right\|_2$是智能体$i$与第$k$个障碍物之间的直线距离. 该部分对智能体与障碍物碰撞进行惩罚，每一时刻的观测值都记录了智能体位置和障碍物位置的相对距离, 智能体与障碍物越近, 惩罚力度越大.

为了使智能体之间减少通信，达到降低通信成本的目的，本文对智能体通信进行了惩罚，碰撞惩罚项公式设计为

$$r_i^{\mathrm{com}} = -\eta\left(\left\|cr_i\right\|_1 + \left\|dr_i\right\|_1\right), \tag{10}$$

式中, $\eta > 0$, 该部分对智能体通信进行惩罚旨在最小化每个时刻要通信的智能体的数量和接收的数据量.

智能体i每个时刻获得的总奖励为上述3部分之和, 其公式为

$$r_i = r_i^{\mathrm{mot}} + r_i^{\mathrm{col}} + r_i^{\mathrm{com}}. \tag{11}$$

在学习过程中，参数的合适与否可能导致不同的策略. 根据式(8)可知, 智能体尽量选择最短路线, 但是σ的值应较小, 否则智能体不会绕开障碍物或其他智能体; 式(9)中的α和β的值应该较大, 鼓励智能体避障; 式(10)中η的值应该小于σ，达到不影响协同控制性能的目的. 为了让智能体学习到期望的策略，平衡协同控制性能和通信成本之间的关系，需要对以上参数进行调试, 以最大化总回报.

4 多智能体协同运动控制仿真实验

4.1 实验场景

本文采用多智能体粒子环境(multi-agent particle environment, MPE)作为多智能体协同运动控制仿真环境, 为了模拟较为复杂的协同运动控制环境, 向环境中添加了一些障碍物, 来检测算法的避障能力. MPE是由OpenAI公司开发的一套时间离散、空间连续的二维多智能体粒子环境，通过控制二维空间中不同角色粒子的运动来完成一系列任务，目前被广泛用于各类多智能体强化学习算法的仿真验证.

在协同运动控制实验场景中，有3个智能体，如图1(a)所示，分别用绿色、红色和黄色的实心圆表示; 三者的期望位置分别用深绿色、深红色以及深黄色的较小的实心圆表示; 用较大的深蓝色实心圆来表示障碍物. 3个智能体的共同目标是通过交流智能体的观测信息，从而到达各自的期望位置，在此过程中还要避免与障碍物和其他的智能体发生碰撞. 每个智能体的控制输入u_i为5个可选动作，分别为原地不动、向上移动、向右移动、向下移动与向左移动.

在智能体进行协同运动控制任务中，当智能体与其他智能体或障碍物之间的距离小于安全距离$d_{\min}$时，会持续受到系统的惩罚，这种惩罚会随着距离的缩短而持续变大，如果智能体与障碍物或其余智能体发生碰撞，则会在受到一个极大的惩罚的同时结束该回合. 当智能体到达期望位置，则会持续受到一个较大的正向奖励，鼓励智能体停留在原地，从而完成协同运动控制任务.

4.2 实验设计

本文将基于深度强化学习的多智能体分布式事件触发控制算法应用到协同运动控制场景中，并设置了协同运动控制性能和通信成本2个指标来对算法进行评估. 对于智能体i，其协同运动控制性能和通信成本分别定义为

$$MC = \sum_{t}^{T}\sum_{i}^{N}\left\|x_i^* - x_i\right\|_2, \tag{12}$$

$$DC = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{\substack{j=1,\\ j\neq i}}^{N}\sum_{m=1}^{M} Or_i^{jm}, \tag{13}$$

式中, $t = 1, 2, \cdots, T$; MC为每回合所有智能体与其期望位置的直线距离的累加值, 本文用其表示多智能体系统的协同运动控制性能, MC值越小, 表明协同运动控制性能越好; DC为一个回合期间所有智能体接收其他智能体发送的数据传输总量, 本文用其表示多智能体系统的通信成本; Or_i^{jm}为智能体之间传输的数据量, 若智能体i在第t时刻接收到来自智能体j的第m组观测信息, 则Or_i^{jm}为该组观测信息的数据量, 反之, $Or_i^{jm} = 0$. DC值越小, 说明通信成本越低.

将式(4)和(5)中的参数设置为$\lambda_1 = \lambda_2 = 0$; 式(8)中$\sigma = 1.0$; 式(9)中$\alpha = 1.5$, $\beta = 5$; 式(10)中$\eta = 0.5$. 算法训练参数如下: 回合数最长$K = 400000$, 折扣因子$\gamma = 0.95$, 每回合的最大时间步长$T = 25$, 目标网络的更新频率$N_n = 100$, 训练批次规模$N_b = 1024$, 软更新参数$\tau = 0.01$, 经验回放区最大规模$N_r = N_b \times T = 2.56 \times 10^4$. 神经网络部分: 学习率设置为0.001, 设隐层为2层全连接神经元, 每层128个神经元, 激活函数为Relu函数. 当智能体与障碍物或其他智能体发生碰撞或者达到最大步长时, 视为任务失败或结束, 本轮训练回合终止, 根据设定的奖励函数返回奖励或惩罚值. 算法的训练流程如图2所示.

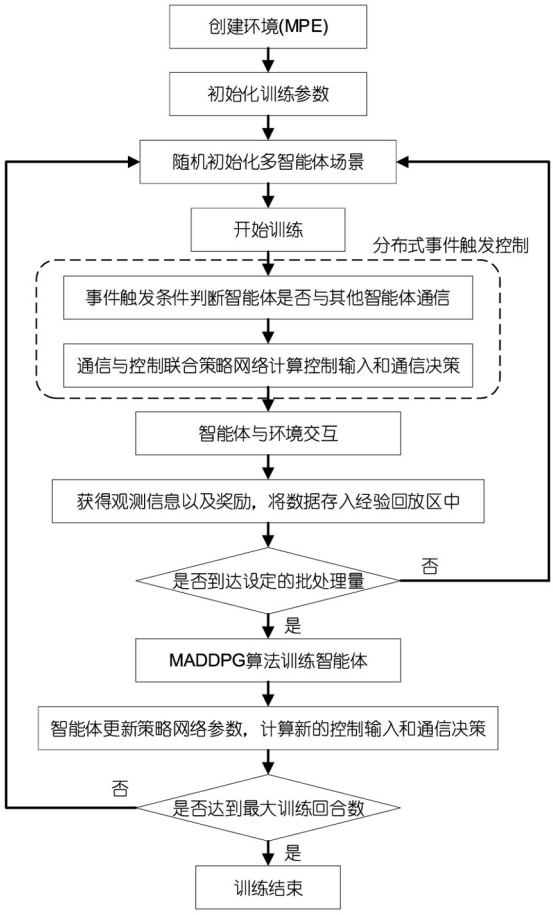

图 2　算法训练流程

Figure 2　Training process of the algorithm.

4.3　实验结果分析

为了验证本文提出的方法能够实现对智能体间通信机制进行优化, 达到通信节省的效果, 并且能够确保智能体完成协同运动控制任务; 本文将提出的方法与高频率通信、低频率通信、随机频率通信以及不通信几种方法进行比较. 上述方法定义如下.

(1) 高频率通信: 智能体每个时刻都接收其余智能体的所有观测信息.

(2) 低频率通信: 智能体每隔5个时刻进行一次通信, 并接收其余智能体的所有观测信息.

(3) 随机频率通信: 智能体每隔随机时刻进行一次通信, 并接收其余智能体的每一项观测信息.

(4) 不通信: 智能体不接收其他智能体的任何一项观测信息.

上述4种通信机制以及分布式事件触发控制方法的协同运动控制性能如图3所示.

结果表明, 在相同回合数的训练过程中, 分布式事件触发控制方法能够以极为接近高频率通信的速度收敛到和其几乎相同的值, 因此, 该方法的协同运动控制性能不亚于高频率通信. 此外, 分布式事件触发控制方法的收敛速度能够明显快于低频率通信、随机频率通信以及不通信, 且协同运动控制性能也明显优于这几种方法.

将分布式事件触发控制方法与其余几种通信机制的通信成本进行比较, 结果如图4所示. 实验结果表明, 分布式事件触发控制的通信成本远小于高频率通信成本, 因此可知该方法能够较好地优化智能体之间的通信机制, 降低通信成本. 尽管表2中的数据表明, 分布

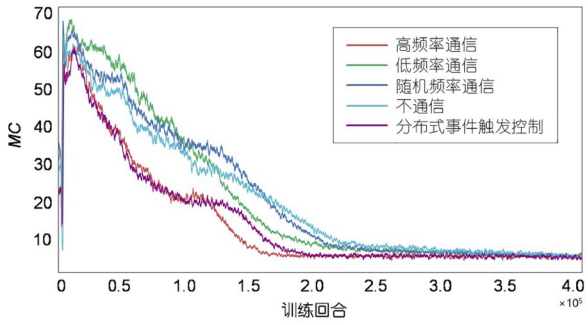

图 3 协同运动控制性能比较

Figure 3 Comparison of cooperative motion control performance.

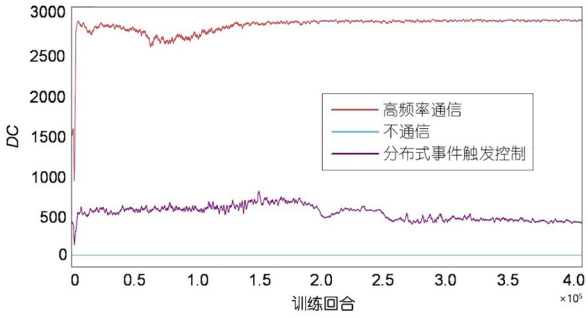

图 4 通信成本比较

Figure 4 Comparison of communication cost.

表 2 训练过程中分布式事件触发控制与低频率和随机频率通信的通信总成本

Table 2 Total communication cost of distributed event-triggered control with low-frequency and random-frequency communication during the training process

方法	通信总成本
分布式事件触发控制	2.14×10^9
低频率通信	2.32×10^9
随机频率通信	2.22×10^9

式事件触发控制方法相比于低频率通信和随机频率通信并没有明显地降低通信成本，但是结合图3可以看出，这几种方法的协同运动控制性能均不如前者. 因此可以确定: 本文提出的基于分布式事件触发控制的多智能体强化学习算法能够很好地平衡协同运动控制性能和通信节省.

将分布式事件触发控制方法与其余几种通信机制的奖励函数值变化曲线进行比较，如图5所示. 实验结

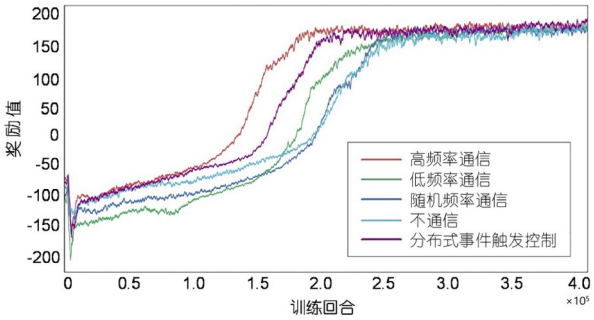

图 5 奖励函数曲线比较

Figure 5 Comparison of reward function curves.

果表明, 在分布式事件触发控制方法作用下, 智能体的奖励曲线上升速度仅次于高频率通信, 在第2×10^5训练回合左右到达收敛点, 快于剩余方法. 综上可知, 本文的方法需要较少的与环境交互的时间，以较低的通信成本实现极好的协同运动控制性能.

为了验证本文提出的方法有着良好的避障能力, 选取了4组较为复杂的场景进行了测试, 智能体协同运动控制轨迹如图6所示，图6(a)~(d)为多智能体在运动过程中的状态. 由智能体的轨迹可以看出，基于指数函数的碰撞惩罚项能够使智能体拥有良好的避障能力, 能够应用于较为复杂的协同运动控制场景.

为了对比几种方法的避障能力，使用图6的4种不同的测试场景，把训练好的模型运行1000次，每次都随机采取这4种场景中的一种，将每个回合结束时智能体与期望位置之间的直线距离小于0.01视为成功完成协同运动控制任务, 而智能体若发生碰撞, 则视为任务失败，比较几种方法的任务成功率以及碰撞率. 如表3所示，本文的方法实现了与高频率通信几乎一样高的任务完成率，且碰撞率也与高频率通信基本持平. 同时, 任务完成率和碰撞率均优于其余几种通信机制.

总之，与低频率通信、随机频率通信以及不通信的方法相比，本文设计的分布式事件触发控制方法展现出更优的协同运动控制性能，奖励值收敛速度更快; 同时拥有与高频率通信几乎一样好的协同传输性能、任务完成率和碰撞率，且通信成本远低于高频率通信, 能够应用于较为复杂的协同场景中. 这是因为分布式事件触发控制方法的触发条件能够减少智能体之间交流一些冗余的观测信息，其通信与控制的联合策略能够计算出智能体在当前状态下希望接收的某一观测信

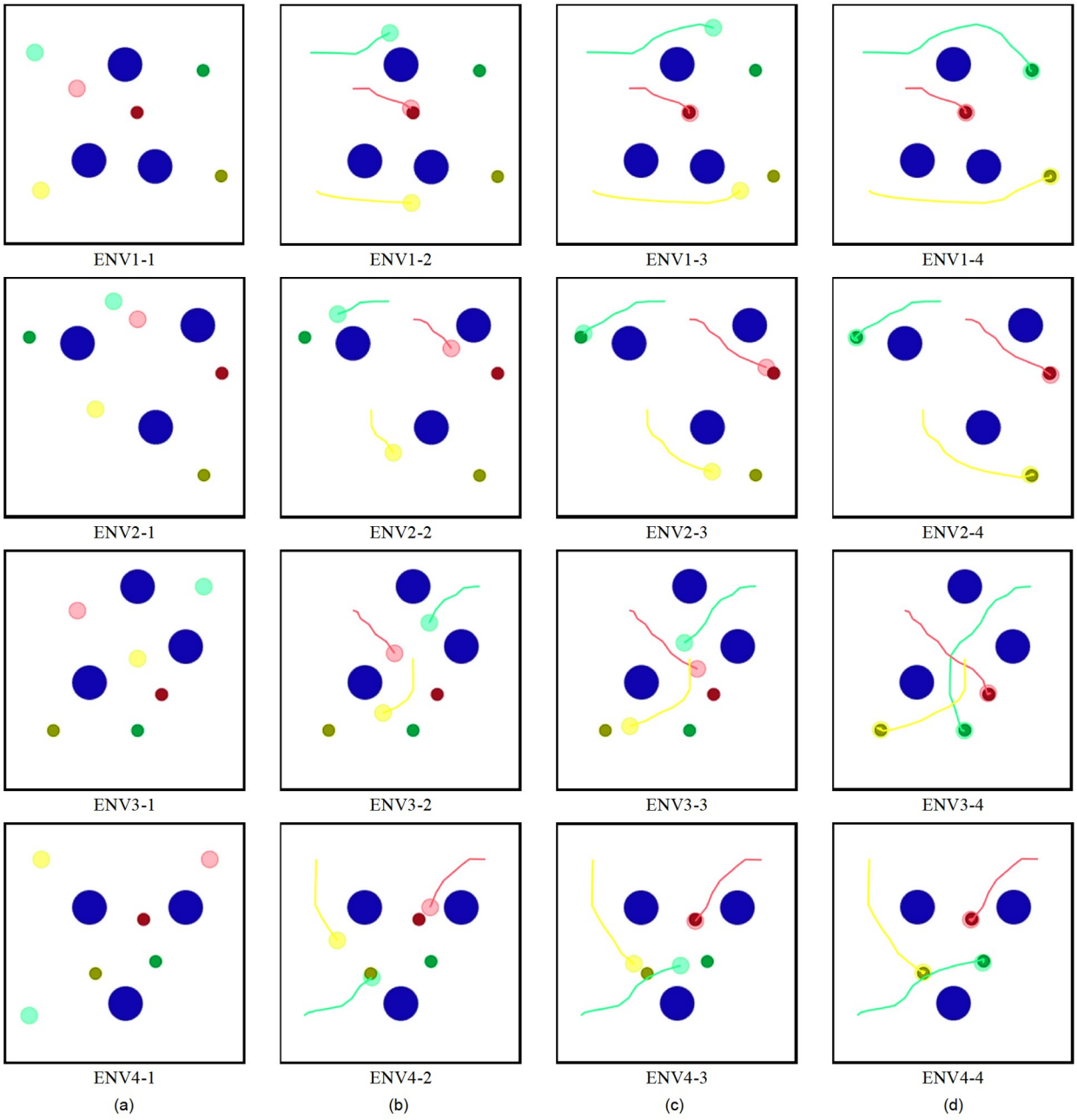

图 6　较为复杂的协同运动控制场景下的多智能体的路径. (a)~(d)为多智能体系统不同时刻的状态

Figure 6　Pathways of multi-agent in complex cooperative motion control scenarios. (a)–(d) are the states of the multi-agent system at different moments in time.

息, 减少智能体之间交流且不丢掉重要的观测信息. 而高频率通信方法虽然能够获取所有的重要信息，但是会收到许多冗余的信息，增加了通信负担. 低频率通信或者随机频率通信，由于其通信决策不是由智能体的策略网络计算得到的，因此很可能漏掉一些重要的观测信息并接收到大量的冗余信息，无法降低通信成本的同时还无法保证完成协同运动控制任务. 不通信虽然没有通信的成本，但是由于智能体之间不交流，导致一些重要信息无法接收，使得协同运动控制性能较低.

表 3 协同运动控制任务成功率和碰撞率比较[a)]

Table 3 Comparison of success and collision rates for cooperative motion control task

场景	任务成功率 (%)					碰撞率 (%)				
	分布式事件触发控制	高频率通信	随机频率通信	低频率通信	不通信	分布式事件触发控制	高频率通信	随机频率通信	低频率通信	不通信
ENV1	99.8	**99.9**	95.7	94.1	95.6	**0.0**	0.1	4.3	5.2	3.9
ENV2	**99.5**	99.0	98.5	98.0	97.4	0.2	**0.0**	1.2	0.9	2.0
ENV3	98.7	**99.5**	88.1	89.7	88.3	1.1	**0.5**	11.5	9.0	10.8
ENV4	99.1	**99.2**	92.0	94.5	94.3	**0.8**	0.8	6.6	4.7	5.7

a) 加粗字体为每种场景下性能指标的最优值

5 结论

针对多智能体协同运动控制任务中优化通信机制, 降低通信成本的问题, 本文设计了一种分布式事件触发控制方法, 并将其应用到多智能体系统中, 实现了在不降低协同运动控制性能的同时达到了通信节省的目的. 此外, 设计了一种基于指数函数的碰撞惩罚项, 使智能体拥有在复杂场景下的避障能力.

首先分析了马尔可夫决策过程的基本性质, 并进一步对MADDPG算法进行了详细的介绍; 随后考虑优化多智能体之间的通信机制, 使用分布式事件触发控制方法来减少智能体间的数据传输量, 从而降低通信成本, 同时为了解决通信决策的设计不能与控制输入分开, 采用了通信与控制的联合策略; 设计了一种基于指数函数的碰撞惩罚项, 以及平衡性能和通信成本的协同运动控制奖励项和通信惩罚项; 最后将上述智能体协同理论的模型和方法应用于MPE环境中, 通过实验验证该方法的协同运动控制性能、避障以及通信节省能力. 实验结果表明, 本文提出的基于深度强化学习的多智能体分布式事件触发控制算法能够优化多智能体系统的通信机制, 降低数据传输量, 实现了在不降低协同运动控制性能的同时达到了通信节省的目的, 所学习的奖励函数能够实现平衡协同运动控制性能和通信节省, 并且能够完成避障任务. 未来将基于事件触发控制方法, 以及考虑智能体之间通信可能存在信息丢失问题, 设计信息聚合策略, 利用更优的多智能体强化学习算法, 并对所学习到的奖励函数进行更完善的分析和解释, 在更高维的智能体观测空间中实现更多的多智能体协同任务.

参考文献

1 Gong X D, Ding B, Xu J, et al. Synchronous *n*-step method for independent *Q*-learning in multi-agent deep reinforcement learning. In: Proceedings of the 16th IEEE International Conference on Ubiquitous Intelligence and Computing. Leicester: IEEE, 2019. 460–467

2 Zhang W X, Ma L, Wang X D. Reinforcement learning for event-triggered multi-agent systems (in Chinese). CAAI Trans Intell Syst, 2017, 12: 82–87 [张文旭, 马磊, 王晓东. 基于事件驱动的多智能体强化学习研究. 智能系统学报, 2017, 12: 82–87]

3 Shibata K, Jimbo T, Matsubara T. Deep reinforcement learning of event-triggered communication and control for multi-agent cooperative transport. In: Proceedings of the IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021. 8671–8677

4 Kesper L, Trimpe S, Baumann D. Toward multi-agent reinforcement learning for distributed event-triggered control. In: Proceedings of the 5th Learning for Dynamics and Control Conference. Philadelphia, 2023. 1072–1085

5 Funk N, Baumann D, Berenz V, et al. Learning event-triggered control from data through joint optimization. IFAC J Syst Control, 2021, 16: 100144

6 Zhang Y, Tian Y P. Consensus of data-sampled multi-agent systems with random communication delay and packet loss. IEEE Trans Autom Control, 2010, 55: 939–943

7 Zhang S Q, Lin J, Zhang Q. Succinct and robust multi-agent communication with temporal message control. In: Proceedings of the 34th Neural Information Processing Systems. Vancouver, 2020. 17271–17282

8 Zhang Q, Yang Y, Xie X, et al. Dynamic event-triggered consensus control for multi-agent systems using adaptive dynamic programming. IEEE Access, 2022, 10: 110285–110293

9 Wang M Y, Yu B, Chen X. Hybrid-triggered consistency for uncertain multi-agent with network attack (in Chinese). Electron Opt Control, 2023, 30: 85–93 [王梦阳, 于冰, 陈侠. 网络攻击不确定多智能体混合触发一致性. 电光与控制, 2023, 30: 85–93]

10 Dimarogonas D V, Frazzoli E, Johansson K H. Distributed event-triggered control for multi-agent systems. IEEE Trans Autom Control, 2012, 57: 1291–1297

11 Trimpe S, D'Andrea R. An experimental demonstration of a distributed and event-based state estimation algorithm. In: Proceedings of the 18th International Federation of Automatic Control. Milan, 2011. 8811–8818

12 Dohmann P B, Hirche S. Distributed control for cooperative manipulation with event-triggered communication. IEEE Trans Robot, 2020, 36: 1038–1052

13 Hu G, Zhu Y, Zhao D, et al. Event-triggered communication network with limited-bandwidth constraint for multi-agent reinforcement learning. IEEE Trans Neural Netw Learn Syst, 2023, 34: 3966–3978

14 Baumann D, Zhu J J, Martius G, et al. Deep reinforcement learning for event-triggered control. In: Proceedings of the 57th IEEE International Conference on Decision and Control. Florida: IEEE, 2018. 943–950

15 Shibata K, Jimbo T, Matsubara T. Deep reinforcement learning of event-triggered communication and consensus-based control for distributed cooperative transport. Robot Auton Syst, 2023, 159: 104307

16 Wang C, Wang J, Shen Y, et al. Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach. IEEE Trans Veh Tech, 2019, 68: 2124–2136

17 Li S B, Song Q. Cooperative control of multiple AGVs based on multi-agent reinforcement learning. In: Proceedings of the 6th IEEE International Conference on Unmanned Systems. Hefei: IEEE, 2023. 512–517

18 Wang J R, Huang J H, Tang Y. Swarm intelligence capture-the-flag game with imperfect information based on deep reinforcement learning (in Chinese). Sci Sin Tech, 2023, 53: 405–416 [王健瑞, 黄家豪, 唐漾. 基于深度强化学习的不完美信息群智夺旗博弈. 中国科学: 技术科学, 2023, 53: 405–416]

19 Yu X, Wu W J, Luo J, et al. Identification method for collective consensus mechanism based on inverse reinforcement learning (in Chinese). Sci Sin Tech, 2023, 53: 258–267 [于鑫, 吴文峻, 罗杰, 等. 面向群体共识机制的逆强化学习辨识方法. 中国科学: 技术科学, 2023, 53: 258–267]

20 Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. California, 2017. 6382–6393

21 Xu Z W, Zhang B, Li D P, et al. Consensus learning for cooperative multi-agent reinforcement learning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, 2023. 11726–11734

22 Neary C, Xu Z, Wu B, et al. Reward machines for cooperative multi-agent reinforcement learning. In: Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems. Richland, 2021. 934–942

23 Xiao D, Tan A H. Cooperative reinforcement learning in topology-based multi-agent systems. Auton Agent Multi-Agent Syst, 2013, 26: 86–119

24 Harutyunyan A, Devlin S, Vrancx P, et al. Expressing arbitrary reward functions as potential-based advice. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. Texas, 2015. 2652–2658

Multi-agent distributed event-triggered optimization control based on deep reinforcement learning

LI FangYu[1,2,3,4], LIU JinYi[1,2,3], HUANG YanTing[1,2,3,4] & HAN HongGui[1,2,3,4]

[1] *Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;*
[2] *Engineering Research Center of Ministry of Digital Community, Ministry of Education, Beijing University of Technology, Beijing 100124, China;*
[3] *Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China;*
[4] *Beijing Institute of Artificial Intelligence, Beijing 100124, China*

In order to reduce the communication cost of the multi-agent cooperative control in obstacle environments, this paper designs a distributed event-triggered optimization control method. The method jointly designs the event-triggered communication mechanism and control policy, determines the communication triggering conditions, and reduces redundancy in the real-time or cyclic communication. Meanwhile, in order to avoid collision between the agent with other agents or obstacles, an exponential function-based collision penalty term is designed to give a penalty exponentially when the agent approaches obstacles or other agents. The method is applied to the multi-agent deep deterministic policy gradient (MADDPG) algorithm and simulated on the multi-agent particle environment (MPE) with added obstacles. Simulation results show that this method can complete the collision-free multi-agent cooperative control task with superior performance while reducing the amount of data transmission.

multi-agent, reinforcement learning, distributed event-triggered control, cooperative motion control

doi: 10.1360/SST-2023-0346