https://doi.org/10.1051/jnwpu/20234120389

多智能体编队控制中的迁移强化学习算法研究

胡鹏林, 潘泉, 郭亚宁, 赵春晖

(西北工业大学 自动化学院, 陕西 西安 710129)

摘 要:针对多障碍环境下的多智能体系统协同编队避障与防撞问题,提出一种迁移学习与强化学习相结合的编队控制算法。在源任务学习阶段,利用值函数近似方法避免 Q-表格求解法所需的大规模存储空间问题,有效降低对存储空间的需求,提升算法求解速度;在目标任务学习阶段,采用高斯聚类算法对源任务进行分类,根据聚类中心和目标任务之间的距离,选择最优的源任务类进行目标任务学习,有效避免了负迁移现象,进而提升了强化学习算法的泛化能力及收敛速度。仿真实验结果表明,所提方法能使多智能体系统在复杂的障碍环境下有效地形成并保持编队构型,同时实现避障与防撞。

关 键 词:多智能体系统;迁移强化学习;值函数近似;编队控制;高斯聚类

中图分类号:TP13 **文献标志码**:A **文章编号**:1000-2758(2023)02-0389-11

多智能体系统由多个同构或异构的智能体构成,通过与环境之间的交互,共同完成单个智能体由于能力不够而无法完成的复杂任务。由于多智能体系统的鲁棒性、可靠性、灵活性等独特优势,在航空航天、工业生产、交通运输等领域获得了深入研究和广泛应用[1-3]。编队控制是指智能体在运动过程中,通过建立、保持、变换智能体的空间构型,在克服干扰约束的同时完成特殊的任务规划,保证编队的安全性和高效性。编队控制作为多智能体系统的重要研究内容之一,在军事和民用领域有着广泛应用。

在多智能体编队控制算法的发展过程中,出现了多种控制算法,主要有基于行为的方法、虚拟结构的方法、领航者-跟随者方法、一致性理论等多种控制理论与方法。基于行为法是模仿自然界的动物行为进行编队控制,一般采用鲁棒的分布式控制,具有很强的扩展能力。例如,基于大雁[4]、鸽子[5]、鱼群[6]等生物,提出了大量的编队控制算法。虚拟结构法采用虚拟的刚体结构描述多智能体编队构型,通过相对误差进行控制,具有很高的控制精度。例如,基于虚拟结构法的无人机编队控制[7]、无人船编队控制[8]、多机器人编队控制[9]。领航者-跟随者算法将智能体分为领航者和跟随者,双方之间保持一定距离,保证编队的稳定和安全。基于领航者-跟随者算法研究了无人机编队控制[10]、无人船编队控制[11]、三维空间中编队控制问题[12]。一致性理论是将图论和代数理论结合,通过处理系统误差实现编队控制,例如,基于一致性理论实现无人机的编队和避障控制[13],以及使得非对称移动机器人能够以期望的编队构型移动[14]。

上述多智能体编队控制算法通常需要精确的数学模型来设计控制律,在实际应用中,由于传感器误差、环境干扰等随机因素,通常难以获得精确模型,给多智能体编队控制造成巨大困难。

强化学习(reinforcement learning,RL)不需要太多先验知识和精确的数学模型,成为解决智能体系统控制问题的重要途径[15]。例如,将 Q-learning 和策略梯度算法结合,提出了多层结构的编队控制算法,使得机器人能够实现设定的编队构型[16]。通过自适应 Q-learning 算法,实现了障碍环境中编队控制[17]。通过基于 RL 的多智能体编队控制框架,消除了建模和控制器设计的繁琐工作,解决了复杂环境中编队控制问题[18]。传统的强化学习算法不仅需要大量的样本进行训练,还要求训练和测试数据属于同一个域,随着智能体数量的增加,状态空间呈

收稿日期:2022-06-15 **基金项目**:国家自然科学基金(61790552,62073264)资助

作者简介:胡鹏林(1996—),西北工业大学博士研究生,主要从事博弈论、强化学习及多智能体最优控制研究。

通信作者:潘泉(1961—),西北工业大学教授,主要从事无人机信息安全与多源信息融合研究。e-mail:quanpan@nwpu.edu.cn

指数增长,给存储空间和计算能力带来巨大挑战。

迁移强化学习(transfer reinforcement learning, TRL)从简单的源任务中获得知识,求解更复杂的目标任务,并且任务之间相似程度越高,知识的传递就更加容易,从而提高系统的学习效率[19]。按照学习方法可以分为基于样本的迁移、基于模型的迁移、基于特征的迁移,以及基于关系的迁移等多种模式[20]。除了计算机视觉、文本分类、自然语言处理等传统领域外,迁移强化学习逐渐应用在许多新兴的领域。在医学图像领域,由于医学图像的标记通常依赖于有经验的医生,因此,收集足够的训练数据是非常昂贵和困难的,迁移学习技术能很好地帮助医学影像分析[21-22]。在生物信息学领域,生物体之间的组成发生了变化,但其功能可能保持不变,可以借助迁移学习算法来进行生物序列的分析[23-24]。在交通运输领域,迁移学习可以帮助电子监控系统进行交通场景图像的理解,以及驾驶员行为的建模[25-26]。在个性化推荐系统领域,往往训练数据是稀疏的,迁移学习算法可以利用来自其他推荐系统的数据来帮助构建目标推荐系统[27-28]。

本文基于迁移强化学习算法,研究了复杂障碍环境下多智能体编队控制问题。在训练过程中,利用值函数近似算法,解决了任务规模不断扩大带来的存储和计算问题。采用高斯混合模型(Gaussian mixture model, GMM)对源任务进行聚类分析,避免出现负迁移现象,提高了迁移强化学习效率。文章内容安排如下:第 1 节介绍智能体模型和求解问题的数学描述,第 2 节介绍基于迁移强化学习的多智能体编队控制算法,第 3 节通过仿真实验验证了算法的有效性,第 4 节给出全文总结与未来研究方向。

1 问题描述

多智能体编队控制系统由 N 个智能体组成,智能体通过复杂的障碍物环境,保持一定的队形到达目标点,同时保证智能体不发生碰撞。智能体 i 的模型为

$$\begin{aligned} x_i(t+1) &= x_i(t) + v\cos(\Delta\varphi_i(t)) \\ y_i(t+1) &= y_i(t) + v\sin(\Delta\varphi_i(t)) \end{aligned} \tag{1}$$

式中:$i \in \boldsymbol{N} = \{1,2,\cdots,N\}$,$v$ 代表智能体的移动速度;$\varphi_i(t) \in [0,2\pi]$ 表示 t 时刻智能体 i 的航向角,即智能体 i 的移动方向与 x 坐标轴正方向的夹角,为了避免出现较大的转弯动作,航向角增量 $\Delta\varphi_i(t)$ 满足

$$\Delta\varphi_i(t) = |\varphi_i(t) - \varphi_i(t-1)| \leqslant \frac{\pi}{4} \tag{2}$$

用 $s_i(t) = [x_i(t), y_i(t)]$ 表示 t 时刻智能体 i 的坐标位置,障碍物集合为 $\boldsymbol{N}_o = \{1,2,\cdots,l\}$。

在多智能体编队中,最优性能指标不是单个智能体的策略达到最优,而是整个编队的策略达到最优,即编队中智能体之间构成合作博弈关系,因此,将多智能体控制任务描述为马尔科夫博弈过程

$$\boldsymbol{\Gamma} = \{S, A, \pi, R, V\} \tag{3}$$

式中:$S = \prod_{i \in N} s_i$ 表示 N 个智能体的联合状态空间;s_i 表示智能体 i 的状态空间。$A = \prod_{i \in N} a_i$ 表示 N 个智能体的联合动作空间,a_i 表示智能体 i 的动作空间。π 表示策略,即智能体在联合状态空间 S,根据概率分布 $P = S \times A \rightarrow S'$ 执行联合动作 A 之后,转移到状态 S' 获得奖励 R。通过奖励值的期望(4),对控制策略的性能进行评价

$$V_i^{\pi}(S) = \mathrm{E}_{\pi}\left\{ \sum_{t=0}^{\infty} \gamma^t R_i(t) \mid S_t = S \right\},\ i \in \boldsymbol{N} \tag{4}$$

式中,γ 表示折扣因子。

由于在学习过程中,奖励函数的作用非常重要,定义智能体 $i \in \boldsymbol{N}$ 的奖励函数 R_i 为

$$R_i(t) = (r_g)_i^t + (r_o)_i^t + (r_c)_i^t + (r_w)_i^t \tag{5}$$

式中,$(r_g)_i^t$ 表示智能体 i 关于目标点的奖励函数

$$(r_g)_i^t = \begin{cases} r_1 > 0, & \text{if } \| s_i(t+1) - s_g \|_2 \leqslant d_g \\ w_g(\| s_i(t) - s_g \|_2 - \| s_i(t+1) - s_g \|_2), & \text{otherwise} \end{cases} \tag{6}$$

式中,s_g 表示目标点的状态,智能体与目标点之间的距离不大于 d_g 时表示到达了目标点,w_g 为常系数。$(r_o)_i^t$ 表示智能体 i 关于障碍物的奖励函数

$$(r_o)_i^t = \begin{cases} r_2 < 0, & \text{if } \| s_i(t+1) - s_l \|_2 \leqslant d_o \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

式中,$l \in \boldsymbol{N}_o$,用 s_l 表示障碍物的状态,d_o 表示障碍

物的安全距离。$(r_c)_i^t$ 表示智能体 i 关于其他智能体 j 的奖励函数

$$(r_c)_i^t=\begin{cases}r_3<0, & \text{if } \|s_i(t+1)-s_j(t+1)\|_2\leqslant d_s\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

式中,d_s 表示智能体的安全距离。为了满足(2)式的要求,用 $(r_w)_i^t$ 表示对智能体 i 转向速度过快的惩罚

$$(r_w)_i^t=\begin{cases}r_4<0, & \text{if 不满足(2)式}\\ w_w\left(\dfrac{\pi}{4}-|\Delta\varphi_i(t)|\right), & \text{其他}\end{cases} \tag{9}$$

式中,w_w 为常系数。在多智能体编队中,智能体 i 和其他智能体之间进行交互,因此,智能体 i 的性能指标由自身和其他智能体共同决定,用 $\boldsymbol{\pi}_{-i}$ 表述其他智能体的策略,在考虑其他智能体的策略时,智能体 i 的性能指标为

$$V_i^{\pi}(s_i,\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i})=R_i(s_i,\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i})+\gamma\sum_{s_i'\in S}^{\infty}P(S,A,S')\sum_{a_i'\in A}\boldsymbol{\pi}(S')V_i^{\pi}(s_i',\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i}) \tag{10}$$

则最优价值函数为

$$V_i^{*}(s_i,\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i})=R_i(s_i,\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i})+\gamma\sum_{s_i'\in S}^{\infty}P(S,A,S')\max_{a_i'\in A}\sum_{A\setminus\{a_i\}}\boldsymbol{\pi}(S')V_i^{\pi}(s_i',\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i}) \tag{11}$$

多智能体博弈的策略取决于环境中智能体的联合行为,在其他智能体策略保持不变的情况下,智能体 i 在状态 s_i 的策略

$$\boldsymbol{\pi}_i^{\text{Nash}}=\sum_{a_i\in A}\boldsymbol{\pi}(S)V_i^{\pi}(s_i,\boldsymbol{\pi}_i\cdot\boldsymbol{\pi}_{-i}) \tag{12}$$

如果 $\boldsymbol{\pi}_i^{\text{Nash}}$ 等于最优响应 $\boldsymbol{\pi}_i^{\max}$

$$\boldsymbol{\pi}_i^{\max}=\max_{a_i\in A}\sum_{A\setminus\{a_i\}}\boldsymbol{\pi}(S)V_i^{*}(s_i,\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i}) \tag{13}$$

则(12)式为智能体 i 的 Nash 均衡策略,基于 Nash 均衡策略的价值函数为

$$V_i^{\text{Nash}}(s_i)=\sum_{a_i\in A}\{\boldsymbol{\pi}_i^{\text{Nash}}\cdot\boldsymbol{\pi}_{-i}^{\text{Nash}}V_i(s_i)\} \tag{14}$$

在多智能体编队过程中,采用价值函数

$$V_i^{\text{Nash}}(s_i)=R_i(t)+\gamma V_i^{\text{Nash}}(s_i') \tag{15}$$

智能体能够在优化自身策略的同时考虑其他智能体的影响,使得编队的整体性能最优。智能体 i 的最优性能为 $V_i(s_i,\boldsymbol{\pi}_i^{\text{Nash}},\boldsymbol{\pi}_{-i})$,其他智能体的最优性能为 $V_{-i}(s_i,\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i}^{\text{Nash}})$,二者与编队整体最优性能 $V(s_i,\boldsymbol{\pi}_i^{\text{Nash}},\boldsymbol{\pi}_{-i}^{\text{Nash}})$ 之间满足

$$\begin{aligned}V_{-i}(s_i,\boldsymbol{\pi}_i,\boldsymbol{\pi}_{-i}^{\text{Nash}})&\leqslant V(s_i,\boldsymbol{\pi}_i^{\text{Nash}},\boldsymbol{\pi}_{-i}^{\text{Nash}})\\ V_i(s_i,\boldsymbol{\pi}_i^{\text{Nash}},\boldsymbol{\pi}_{-i})&\leqslant V(s_i,\boldsymbol{\pi}_i^{\text{Nash}},\boldsymbol{\pi}_{-i}^{\text{Nash}})\end{aligned} \tag{16}$$

综上,多智能体编队控制问题可以描述为

$$\begin{aligned}&\mathscr{J}(S,\boldsymbol{\pi}^{*})=\max\prod_{i=1}^{N}V_i^{\text{Nash}}(s_i,\boldsymbol{\pi}_i^{*})\\ &\text{s.t. }\lim_{t\to\infty}\|s_i^{g}-s_i(t)\|_2=0\\ &\qquad\|s_i(t)-s_j(t)\|_2>d_s\\ &\qquad\|s_i(t)-s_l\|_2>d_o\\ &\qquad\Delta\varphi_i(t)\leqslant\frac{\pi}{4},\ l\in\boldsymbol{N}_o,\forall i,j\in\boldsymbol{N}\end{aligned} \tag{17}$$

式中,$\boldsymbol{\pi}_i^{*}$ 是使得编队整体的联合策略达到最优时的智能体 i 的最优策略,s_i^g 表示智能体 i 的目标位置。多智能体编队控制问题的目标为求解联合最优策略 $\boldsymbol{\pi}^{*}=(\boldsymbol{\pi}_1^{*},\boldsymbol{\pi}_2^{*},\cdots,\boldsymbol{\pi}_N^{*})$,使得最优性能指标 $\mathscr{J}(S,\boldsymbol{\pi}^{*})$ 最大化。

2　基于迁移学习的多智能体编队控制

通过前文获得了多智能体编队控制问题的数学描述以及优化目标。为了应对复杂的多智能体编队环境,提升强化学习速度以及泛化能力,本文引入迁移强化学习算法。如图 1 所示,迁移强化学习涉及到源任务和目标任务。

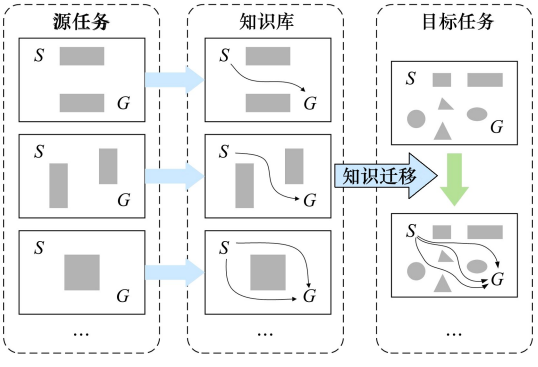


图 1　迁移强化学习示意图

基于马尔科夫博弈的描述,多智能体编队控制可以分解为源任务 $\boldsymbol{M}_s=\{N_s,S_s,A_s,R_s\}$ 和目标任务 $\boldsymbol{M}_t=\{N_t,S_t,A_t,R_t\}$ 2 个子任务。迁移强化学习过程包括 2 个阶段,首先是知识迁移阶段 A_t,根据目标任务 U_g 和源任务 U_s 之间的相关性生成合适的知识迁移模型 U_t

$$A_t: U_s \times U_g \rightarrow U_t \tag{18}$$

其次是学习阶段 A_1，结合迁移模型 U_t 和目标任务 U_g，实现目标任务的学习

$$A_1: U_t \times U_g \rightarrow U_g \tag{19}$$

下面从源任务和目标任务学习两方面进行论述。

2.1 基于值函数近似的源任务学习

在源任务学习中，为了突出迁移强化学习对训练速度的提升作用，采用 Q-Learning 算法进行源任务训练，基于值函数近似方法，避免 Q 表格带来的大规模存储问题。根据(15)式智能体 i 的值函数为

$$V_i^{\text{Nash}}(s_i(t)) = \mathrm{E}\Big\{\sum_{k=t}^{\infty}\gamma^k R_i(k)\Big\} = R_i(t) + \gamma V_i^{\text{Nash}}(s_i(t+1)) \tag{20}$$

值函数可以用线性函数表述为

$$\hat{V}_i^{\text{Nash}}(s_i(t), \boldsymbol{w}_i(t)) = \boldsymbol{\phi}_i(s_i(t))\boldsymbol{w}_i^{\mathrm{T}} \tag{21}$$

式中：$\boldsymbol{\phi}_i(s_i(t)) = [\phi_{i1}(s), \cdots, \phi_{im}(s)] \in \mathbf{R}^{m\times 1}$ 是由智能体 i 的状态构成的 m 维特征向量；T 表示矩阵的转置，用 $\boldsymbol{w}_i = [w_{i1}, \cdots, w_{im}] \in \mathbf{R}^{m\times 1}$ 表示联合权重向量，使用含有参数 $\boldsymbol{w}_i$ 的值函数表示其真值为

$$V_i^{\text{Nash}}(s_i(t)) = \hat{V}_i^{\text{Nash}}(s_i(t), \boldsymbol{w}_i(t)) \tag{22}$$

则 $\hat{V}_i^{\text{Nash}}(s_i(t), \boldsymbol{w}_i(t))$ 和 $V_i^{\text{Nash}}(s_i(t))$ 之间的误差为

$$e_i(t) = \frac{1}{2}[V_i^{\text{Nash}}(s_i(t)) - \hat{V}_i^{\text{Nash}}(s_i(t), \boldsymbol{w}_i(t))]^2 \tag{23}$$

求参数 $\boldsymbol{w}_i$ 使得(23)式最小，根据最小二乘算法有

$$\boldsymbol{w}_i(t) = \boldsymbol{w}_i(t-1) + \frac{\boldsymbol{A}_i(t-1)\boldsymbol{\phi}_i(s_i(t))}{1 + \widetilde{\boldsymbol{\phi}}_i^{\mathrm{T}}(s_i(t))\boldsymbol{A}_i(t-1)\boldsymbol{\phi}_i(s_i(t))} \times [R_i(t) - \widetilde{\boldsymbol{\phi}}_i^{\mathrm{T}}(s_i(t))\boldsymbol{w}_i(t-1)] \tag{24}$$

式中

$$\boldsymbol{A}_i(t) = \boldsymbol{A}_i(t-1) - \frac{\boldsymbol{A}_i(t-1)\boldsymbol{\phi}_i(s_i(t))\widetilde{\boldsymbol{\phi}}_i^{\mathrm{T}}(s_i(t))\boldsymbol{A}_i(t-1)}{1 + \widetilde{\boldsymbol{\phi}}_i^{\mathrm{T}}(s_i(t))\boldsymbol{A}_i(t-1)\boldsymbol{\phi}_i(s_i(t))}$$

$$\widetilde{\boldsymbol{\phi}}_i(s_i(t)) = \boldsymbol{\phi}_i(s_i(t)) - \gamma\boldsymbol{\phi}_i(s_i(t+1))$$

对权重系数的收敛性进行分析，根据(21)和(24)式有

$$\hat{\boldsymbol{w}}_{t+1} = \hat{\boldsymbol{w}}_t + \boldsymbol{Q}_{t-1}\boldsymbol{\phi}_t R_t^{-1}(\hat{V}^{\text{Nash}} - \boldsymbol{\phi}_t^{\mathrm{T}}\hat{\boldsymbol{w}}_t) \tag{25}$$

式中，$\hat{V}^{\text{Nash}}$ 表示值函数，用 $\boldsymbol{\phi}_t$ 替换 $\boldsymbol{\phi}_i(s_i(t))$，$R_t = \lambda\boldsymbol{G} + \boldsymbol{\phi}_t^{\mathrm{T}}\boldsymbol{Q}_{t-1}\boldsymbol{\phi}_t$，矩阵 $\boldsymbol{G}^{\mathrm{T}} = \boldsymbol{G} > 0$，常数遗忘因子 $\lambda \in (0,1)$，对称正定矩阵 $\boldsymbol{Q}_t$ 定义为

$$\boldsymbol{Q}_t^{-1} = \lambda\boldsymbol{Q}_{t-1}^{-1} + \boldsymbol{\phi}_t\boldsymbol{G}^{-1}\boldsymbol{\phi}_t^{\mathrm{T}} \tag{26}$$

设 ρ, ψ, C 均为正数，$\boldsymbol{\phi}_t$ 为持续激励信号，满足

$$0 < \rho\boldsymbol{I} \leqslant \sum_{t=j}^{j+C}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^{\mathrm{T}} \leqslant \psi\boldsymbol{I} \leqslant \infty \tag{27}$$

定理 假设(27)式成立，当 $t > C$ 时，对于任意初始误差 $\widetilde{\boldsymbol{w}}_0$，存在 $\xi > 0$，使得 $\boldsymbol{w}$ 的估计误差呈指数收敛

$$\| \widetilde{\boldsymbol{w}}_t \|_2^2 \leqslant \xi\lambda^t \| \widetilde{\boldsymbol{w}}_0 \|_2^2 \tag{28}$$

证明 根据(25)式有

$$\widetilde{\boldsymbol{w}}_{t+1} = (\boldsymbol{I} - \boldsymbol{Q}_{t-1}\boldsymbol{\phi}_t\boldsymbol{R}_t^{-1}\boldsymbol{\phi}_t^{\mathrm{T}})\widetilde{\boldsymbol{w}}_t \tag{29}$$

选择函数 L_t

$$L_t = \widetilde{\boldsymbol{w}}_t^{\mathrm{T}}\boldsymbol{Q}_{t-1}^{-1}\widetilde{\boldsymbol{w}}_t \tag{30}$$

将(26)式代入(30)式有

$$L_{t+1} - L_t = \widetilde{\boldsymbol{w}}_{t+1}^{\mathrm{T}}\boldsymbol{Q}_t^{-1}\widetilde{\boldsymbol{w}}_{t+1} - \widetilde{\boldsymbol{w}}_t^{\mathrm{T}}\boldsymbol{Q}_{t-1}^{-1}\widetilde{\boldsymbol{w}}_t \leqslant (\lambda - 1)\widetilde{\boldsymbol{w}}_t^{\mathrm{T}}\boldsymbol{Q}_{t-1}^{-1}\widetilde{\boldsymbol{w}}_t = (\lambda - 1)L_t \tag{31}$$

因此得到

$$L_{t+1} \leqslant \lambda L_t \leqslant \lambda^{t+1}L_0 = \lambda^{t+1}\widetilde{\boldsymbol{w}}_0^{\mathrm{T}}\boldsymbol{Q}_{-1}^{-1}\widetilde{\boldsymbol{w}}_0 \tag{32}$$

对(31)式中 $\boldsymbol{Q}_{t-1}^{-1}$ 的下界进行分析，由于 $\boldsymbol{G}$ 是对称矩阵，所以对任意矩阵 $\boldsymbol{\Phi}$ 有

$$\lambda_{\min}(\boldsymbol{G}^{-1})\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} \leqslant \boldsymbol{\Phi}\boldsymbol{G}^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \tag{33}$$

根据正定矩阵性质有

$$\boldsymbol{\phi}_t^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{G}^{-1}\boldsymbol{\Phi}^{\mathrm{T}} - \lambda_{\min}(\boldsymbol{G}^{-1})\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}})\boldsymbol{\phi}_t \geqslant 0 \tag{34}$$

根据(26)式有

$$\boldsymbol{Q}_{j-1}^{-1} + \cdots + \boldsymbol{Q}_{j+C-1}^{-1} \geqslant \sum_{t=j}^{j+C}\boldsymbol{\phi}_{t-1}\boldsymbol{G}^{-1}\boldsymbol{\phi}_{t-1}^{\mathrm{T}} \geqslant \lambda_{\min}(\boldsymbol{G}^{-1})\rho\boldsymbol{I} \tag{35}$$

可得 $\boldsymbol{Q}_{t-1}^{-1}$ 的下界为

$$\boldsymbol{Q}_{t-1}^{-1} \geqslant \frac{\lambda_{\min}(\boldsymbol{G}^{-1})\rho(\lambda^{-1} - 1)}{\lambda^{-(C+1)} - 1} > 0 \tag{36}$$

结合(30)、(32)、(33)、(36)式有

$$\lambda L_t \leqslant \lambda^{t+1}\widetilde{\boldsymbol{w}}_0^{\mathrm{T}}\boldsymbol{Q}_{-1}^{-1}\widetilde{\boldsymbol{w}}_0 \tag{37}$$

化简后有

$$\widetilde{\boldsymbol{w}}_t^{\mathrm{T}}\frac{\lambda_{\min}(\boldsymbol{G}^{-1})\rho(\lambda^{-1} - 1)}{\lambda^{-(C+1)} - 1}\widetilde{\boldsymbol{w}}_t \leqslant \lambda^t\widetilde{\boldsymbol{w}}_0^{\mathrm{T}}\boldsymbol{Q}_{-1}^{-1}\widetilde{\boldsymbol{w}}_0 \tag{38}$$

综上，得到

$$\| \widetilde{\boldsymbol{w}}_t \|_2^2 \leqslant \frac{\lambda^{-(C+1)} - 1}{\lambda_{\min}(\boldsymbol{G}^{-1})\rho(\lambda^{-1} - 1)}\lambda_{\min}(\boldsymbol{Q}_{-1}^{-1})\lambda^t \| \widetilde{\boldsymbol{w}}_0 \|_2^2 \tag{39}$$

取(28)式中的 ξ 为

$$\xi = \frac{\lambda^{-(C+1)} - 1}{\lambda_{\min}(\boldsymbol{G}^{-1})\rho(\lambda^{-1} - 1)}\lambda_{\min}(\boldsymbol{Q}_{-1}^{-1}) \tag{40}$$

则(28)式成立，$\boldsymbol{w}_i$ 呈指数收敛，证明完毕。

根据 Bellman 最优性原理，对价值函数进行更

新,智能体 i 的最优值函数为

$$V_i^{\pi^*}(s_i(t),\boldsymbol{w}_i)=\max_{a_i\in A}[R_i(t)+\gamma V_i^{\text{Nash}}(s_i(t+1),\boldsymbol{w}_i)] \tag{41}$$

最优动作选择策略为

$$a_i^*(s_i(t))=\arg\max_{a_i\in A}[R_i(t)+\gamma V_i^{\text{Nash}}(s_i(t+1),\boldsymbol{w}_i)] \tag{42}$$

完成基于值函数近似的源任务训练。

2.2　基于聚类的目标任务迁移学习

在获得源任务之后,将所有源任务的知识迁移到同一个智能体时,由于任务之间的差异化会导致负迁移。因此将大量训练好的源任务数据进行归类处理,选择与目标任务差异最小的源任务,可以有效避免负迁移现象。本文选用 GMM 算法对源任务状态进行聚类分析,假设在本文中策略的相似性可以通过相应的值函数来反映,即在相同的奖励函数下,价值函数相似的策略,智能体的运动轨迹是平行或者重合的。

通过随机采样从源域的状态空间 S_s 中获得多个状态 $\boldsymbol{\Omega}=\{s_1,s_2,\cdots\}$,通过源任务中智能体 i 学习到的策略 π_i',将状态 s_j 映射到价值函数

$$V(s_j\mid\pi_i')=\sum_{a\in A_s}\pi_i(a\mid s_j)V(s_j) \tag{43}$$

式中,$\pi_i(a\mid s_j)$ 表示在源任务状态 s_j 执行动作 a 的概率,通过(44) 式将值函数映射到集合 $\boldsymbol{y}_i$

$$\boldsymbol{y}_i=(V_{i,1},V_{i,2},\cdots)=[V(s_1\mid\pi_i'),V(s_2\mid\pi_i'),\cdots] \tag{44}$$

设 n' 为源任务中的分类数量,在获得源任务样本 $\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{n'})$ 之后,执行 GMM 算法

$$p_Y(x)=\sum_{k=1}^{n}\alpha_k p(x\mid\mu_k,\Sigma_k) \tag{45}$$

式中:n 为目标任务中智能体数量;α_k 为高斯混合系数;(μ_k,Σ_k) 为高斯分布的期望和方差。通过最大似然函数(46) 式估计(45) 式的参数

$$L(\boldsymbol{Y})=\prod_{i=1}^{n'}p_Y(\boldsymbol{y}_i)=\prod_{i=1}^{n'}\sum_{k=1}^{n}\alpha_k p(\boldsymbol{y}_i\mid\mu_k,\Sigma_k) \tag{46}$$

(46)式的对数似然函数为

$$\ln L(\boldsymbol{Y})=\ln\prod_{i=1}^{n'}p_Y(\boldsymbol{y}_i)=\sum_{i=1}^{n'}\ln\sum_{k=1}^{n}\alpha_k p(\boldsymbol{y}_i\mid\mu_k,\Sigma_k) \tag{47}$$

通过 $\dfrac{\partial\ln L(\boldsymbol{Y})}{\partial\mu_k}=0,\dfrac{\partial\ln L(\boldsymbol{Y})}{\partial\Sigma_k}=0$ 求解参数

$$\mu_k=\frac{\sum_j^{n'}\zeta_{jk}V_k}{\sum_j^{n'}\zeta_{jk}}$$

$$\Sigma_k=\frac{\sum_j^{n'}\zeta_{jk}(V_k-\mu_k)(V_k-\mu_k)^{\mathrm{T}}}{\sum_j^{n'}\zeta_{jk}} \tag{48}$$

式中,ζ_{jk} 表示样本 $\boldsymbol{y}_j$ 中元素的后验概率密度

$$\zeta_{jk}=p(V_k\mid\boldsymbol{y}_j)=\frac{\alpha_k\cdot p(\boldsymbol{y}_j\mid\mu_k,\Sigma_k)}{\sum_{l=1}^{n}\alpha_l\cdot p(\boldsymbol{y}_j\mid\mu_k,\Sigma_k)} \tag{49}$$

根据 $\alpha_k\geqslant 0$,且 $\sum_k^n\alpha_k=1$,求得

$$\alpha_k=\frac{1}{n'}\sum_{j=1}^{n'}\zeta_{jk} \tag{50}$$

通过聚类可以得到 n' 个高斯分布,将源任务状态划分为多个集合,每个集合中的策略是相似的,避免多个具有竞争关系的策略传递给同一个目标任务而引起负迁移现象。

迁移学习的关键是源任务与目标任务之间的相似性,2 个任务相似程度越高,迁移学习效果越好。在获得 n' 个具有高斯分布的源任务集合后,通过康托洛维奇距离度量聚类任务和目标任务之间的相似程度。计算聚类任务 $\boldsymbol{M}_s$ 和目标任务 $\boldsymbol{M}_g$ 两者之间的康托洛维奇距离 $D(S_s,S_g)$ 为

$$\begin{aligned}&D(S_s,S_g)=\min_{\eta_{ij}}\sum_{i=1}^{|S_s|}\sum_{j=1}^{|S_g|}\eta_{ij}d(s_i,s_j)\\&\text{s.t.}\sum_{i=1}^{|S_s|}\eta_{ij}=\frac{1}{|S_g|},\ \text{for}\ \forall j\\&\sum_{j=1}^{|S_g|}\eta_{ij}=\frac{1}{|S_s|},\ \text{for}\ \forall j\\&\eta_{ij}>0,\text{for}\ \forall i,j\end{aligned} \tag{51}$$

式中:$|S_s|,|S_g|$表示状态空间的大小;η_{ij} 表示 s_i 和 s_j 之间距离 $d(s_i,s_j)$ 的权重值为

$$\begin{aligned}d(s_i,s_j)=\max_{a\in A}\{&|r(s_i,a_i)-r(s_j,a_j)|+\\&c\min_{\eta_{ij}}\sum_{i=1}^{|S_s|}\sum_{j=1}^{|S_g|}\eta_{ij}d(s_i,s_j)\}\end{aligned} \tag{52}$$

式中,$r(s_i,a_i),r(s_j,a_j)$ 表示奖励值,$c\in(0,1)$。

根据任务之间的距离度量,选择与目标任务距离最近的源任务类,假设其中有 m 个源任务,对目标任务 $\boldsymbol{M}_g$ 中智能体 i 的值函数进行初始化

$$V(s_i,\boldsymbol{w}_i)=\frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{|S_s|}\frac{\eta_{ij}^k}{\eta_j}V^k(s_j^k,\boldsymbol{w}_j) \tag{53}$$

式中，$\eta_j=\sum_{i=1}^{|S_s|}\eta_{ij}^k,\eta_{ij}^k$ 是 $d(s_i,s_j^k)$ 在 $D(S,S_k)$ 中的权重，$V^k(s_j^k,\boldsymbol{w}_j)$ 是从源任务 $\boldsymbol{M}_s$ 的状态 s_j^k 中学习到的值函数。

因此，得到基于迁移强化学习的多智能体编队控制算法流程如下：

1）初始化目标任务 $\boldsymbol{M}_t$，源任务 $\boldsymbol{M}_s$，状态基函数 $\boldsymbol{\phi}(s)$，权重向量 $\boldsymbol{w}=\boldsymbol{1}$，迭代次数 T，最大搜索步数 M，折扣因子 γ，收敛因子 ε。

2）源任务学习：

3）For $k=1:T$ do

4）根据(42)式选择动作 a

5）For $l=1:M$ do

6）执行 a，得到下一状态 s' 和奖励 r

7）在状态 s' 执行(42)式

8）更新状态和动作 $s\leftarrow s'$，$a\leftarrow a'$

9）根据(24)式更新 $\boldsymbol{w}_s$

10）如果 $s=s_g$，$\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2<\varepsilon$ 进行下一次迭代

11）End for

12）目标任务学习：

13）根据(45)式对源任务状态进行 GMM 聚类分析

14）根据(51)式选择 m 个较优的源任务状态

15）根据(53)式初始化目标任务中智能体的值函数

16）执行2)~10)进行训练任务，得到策略 π^*

3 仿真验证与分析

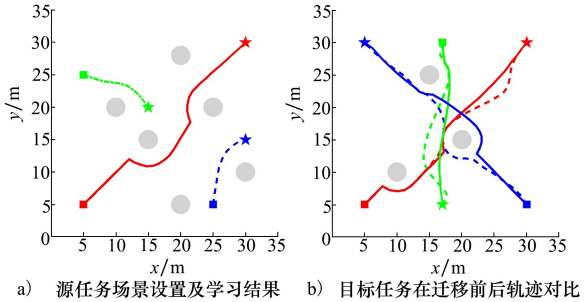
3.1 迁移强化学习过程仿真

本节验证在二维状态空间中迁移强化学习的有效性。智能体最大速度 $v=0.3$ m/s 在 35 m×35 m 的矩形区域运动，用半径为 1 m 的圆表示智能体和障碍物。根据(5)式设计奖励函数，具体参数设置为：$d_s=d_o=d_g=1$ m，$r_1=-50$，$r_2=-10$，$r_3=-10$，$r_4=-5$，$w_g=10$，$w_w=25$。学习参数设置为：$\gamma=0.95$，$T=500$，$M=5\ 000$，$\varepsilon=0.1$。选择多项式形式的状态基函数如(54)式所示。

$$\begin{aligned}\boldsymbol{\phi}_i(t)=[&x,y,x^2,xy,y^2,x^3,x^2y,xy^2,y^3,x^4,\\&x^3y,x^2y^2,xy^3,y^4,x^5,x^4y,x^3y^2,x^2y^3,\\&xy^4,y^5,x^6,x^5y,x^4y^2,x^3y^3,x^2y^4,xy^5,y^6]\end{aligned} \tag{54}$$

对应系数为 $\boldsymbol{w}_i(t)=[w_{i1},w_{i2},\cdots,w_{i27}]$。

源任务中智能体的起始点为[5,5]，[5,25]，[25,5]，对应的目标点为[30,30]，[15,20]，[30,15]，障碍物为[10,20]，[15,15]，[20,5]，[20,28]，[25,20]，[30,10]。智能体路径轨迹如图 2a)所示，方形代表起始点，五星代表终点。所有智能体能够在复杂环境中，无碰撞地从起始点移动到对应的目标点位置，得到较优的源任务学习样本。



a) 源任务场景设置及学习结果 b) 目标任务在迁移前后轨迹对比

图 2 迁移学习效果对比

目标任务中智能体的起始点为[5,5]，[17,30]，[30,5]，对应的目标点为[30,30]，[17,5]，[5,30]，障碍物为[10,10]，[15,25]，[20,15]。迁移学习前后智能体运动轨迹如图 2b)所示，虚线和实线分别表示智能体迁移学习之前和之后的路径轨迹，可以看出后者的路径轨迹明显优于前者，且安全到达各自设定的目标点。为了验证迁移学习效果，在不同的目标环境中，进行了 50 次重复实验，然后求取平均值，对迁移前后所有智能体到达目标点的总路径、总时长以及成功率进行对比分析，成功率以在规定时间内是否到达目标来判定，通过图 3 可以看出，迁移学习之后的路径长度明显小于迁移之前的长度，并且用时较少，成功率也有所提高。

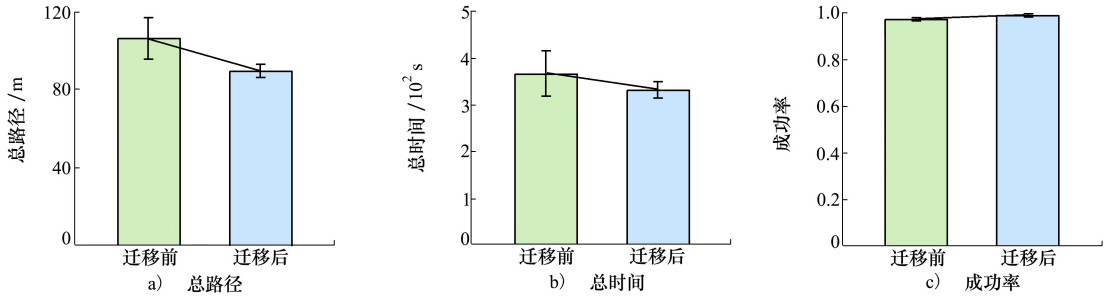


图 3　迁移强化学习前后的结果对比柱状图

图 4 展示了任务之间的相似性对于迭代次数及奖励值的影响,横坐标表示源任务和目标任务间的距离,散点表示每次实验的结果,并进行了曲线拟合。可以看出迭代次数随着任务之间距离的增加而增加,并且随着距离的增加,智能体获得的平均奖励值变小。由此可以得出任务之间的相似性越高,迁移效果越好,同时也说明采用 GMM 算法选择与目标任务差异最小的源任务进行迁移能够有效避免负迁移问题。

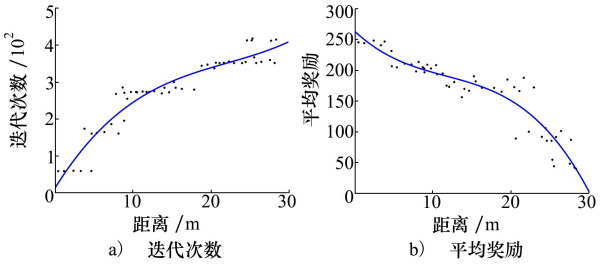


图 4　任务相似性对迁移学习结果的影响

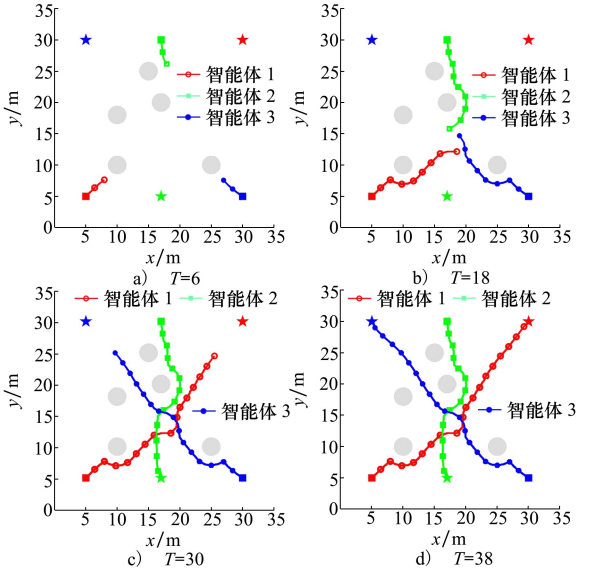


图 5　多智能体避障与防撞过程(T 表示步数)

在目标任务的基础上,设置障碍物为[10,10],[10,18],[15,25],[17,20],[25,10],验证迁移强化学习算法的避障防撞性能,设置 $d_s=0.5$ m。智能体避障与防撞过程如图 5 所示,每个智能体都到达了指定的目标点位置。图 5b)展示了智能体之间的规避过程。避障过程中智能体之间的距离如图 6 所示,均满足设定的安全距离,没有发生碰撞,保证了智能体的安全运动。关于迭代次数和权重系数 $\boldsymbol{w}$ 的收敛性在文献[29]中有详细的实验分析。

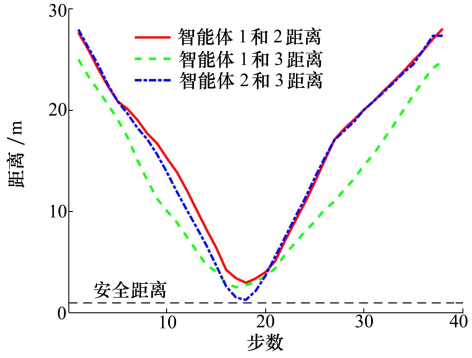


图 6　智能体避障过程相互之间的距离

3.2　多智能体编队任务仿真

用 4 个智能体进行编队控制仿真,编队任务中智能体的起始点为[4.5,26],[5,10],[12,3],[30,5],设置对应的目标点为[28,30],[30,27.5],[32,30],[30,32.5],障碍物为[11,15],[15,23.5],[16,9],[25,10.5],[26.5,23.5]。每个智能体在训练过程中找到各自合适的目标位置,形成对角线长度分别是 4 m 和 5 m 的菱形编队,然后保持编队形式运动到目标位置,智能体运动轨迹如图 7 所示。

在图 7b)中,障碍物 O_1 和 O_2 之间的距离为 7.8 m,不满足智能体 2 和智能体 3 之间 9 m 的避障要求,因此智能体无法穿越障碍物形成的狭窄通道,只能从两边绕行。相反障碍物 O_1 和 O_3 之间的距离

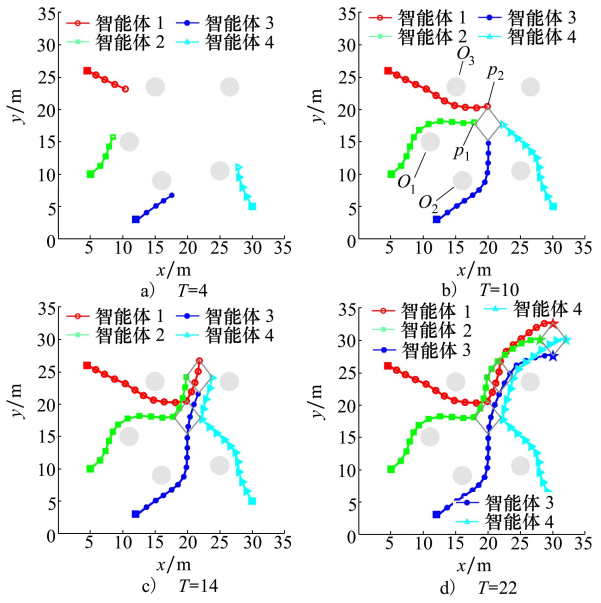


图 7 多智能体编队控制过程运动轨迹(T 表示步数)

为9.4 m,满足避障要求,因此智能体 1 和智能体 2 从障碍物中间通过。同时从图 7b)中可以看出,智能体 1 选择 p_1 位置的奖励最大,但是为了保证编队整体的性能最优,智能体 1 选择了 p_2 位置,而智能体 2 选择了 p_1 位置,验证了 Nash 均衡编队策略的有效性。图 8 描述了编队过程中智能体之间的距离变化,智能体之间的距离均满足设定的安全距离,在任务完成后智能体间的距离分别是 4.9,3.9,3.1 m,符合设计的编队距离。

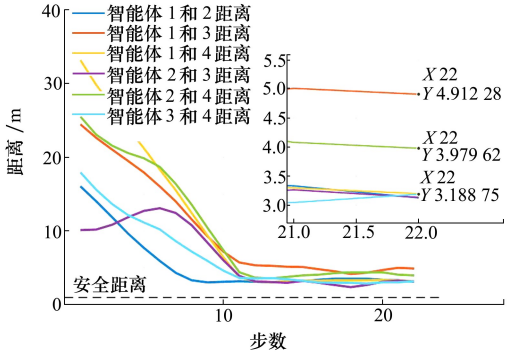


图 8 智能体编队过程相互之间的距离

3.3 基于多无人机的编队任务实时性仿真

为了验证本文提出的迁移强化学习算法的实时性能和可靠性,基于 Gazebo 仿真平台,选用四旋翼无人机对算法有效性进行验证。硬件配置为英特尔 i7-9700、GeForce RTX 3090,操作系统采用 Ubuntu-18.04。场景大小设置为 30 m×30 m,最大飞行速度 v=1 m/s,飞行高度统一为 2 m,8 架无人机通过中心计算机共享全局信息,信息更新频率为 10 Hz。如图 9 所示,不同形状的障碍物均可以抽象为长方体,因此用大小为 1.5 m×1.5 m×2 m 的长方体表示障碍物。预设无人机之间形成以坐标(15,15)为中心,半径为 6 m 的圆。

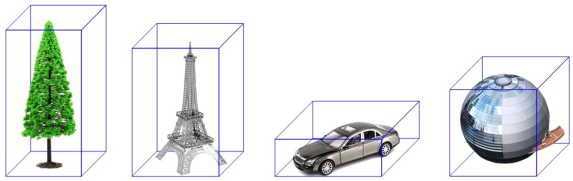


图 9 不同障碍物的抽象表示

图 10 展示了无人机编队过程中关键时刻的截图,从图 10a)可以看出无人机分 2 组从两边向着场景中心位置移动;图 10b)中无人机通过障碍物形成的通道,进行避障飞行;图 10c)中无人机形成初步聚集状态,但是无人机的航向没有统一,是杂乱无序的,不具备编队能力;图 10d)中无人机的航向呈顺时针方向飞行,通过航向箭头可以看出,无人机均沿着设定圆的切线方向移动,从而验证了提出的算法能够使得无人机形成符合设定条件的圆形编队。图 11 展示了无人机的飞行轨迹曲线,形成符合条件的圆形轨迹。

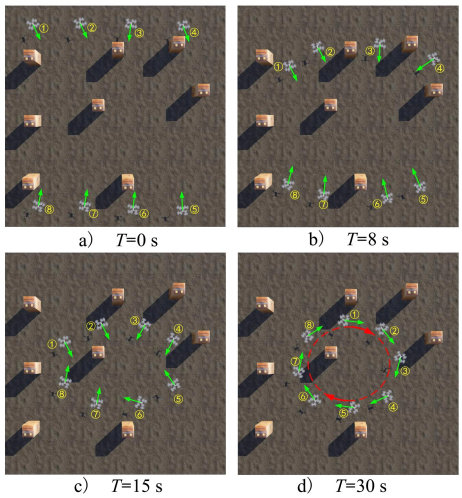


图 10 基于无人机的迁移强化学习算法仿真

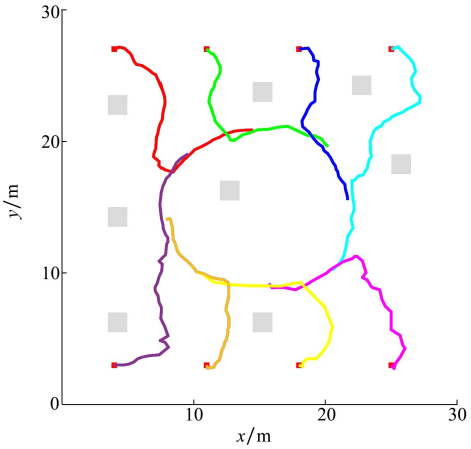


图 11　无人机飞行轨迹

为了验证算法的实时性和可靠性,设计了多组实验,具体场景设置如表 1 所示,每组实验重复进行 25 次,并对数据进行统计分析。设定编队过程在 120 s内完成,如果超出设定的时间阈值或者出现其他故障,则判定编队过程失败。通过分析编队形成的时间以及成功率,可以发现障碍物数量的增加对编队成功率的影响较小,当无人机的数量增加时,编队时间明显增加,同时成功率也降低。但是,4 种场景的编队时间和成功率都在可接受的范围内,因此文章提出的算法具有较好的实时性和较高的可靠性。图 12 展示了 4 种场景中无人机形成圆形编队时的截图。

表 1　基于迁移强化学习的无人机编队多种场景设置

场景	障碍物/个	无人机/个	时间/s	成功率/%
1	4	5	21	98
2	8	5	25	97
3	4	10	33	92
4	8	10	42	90

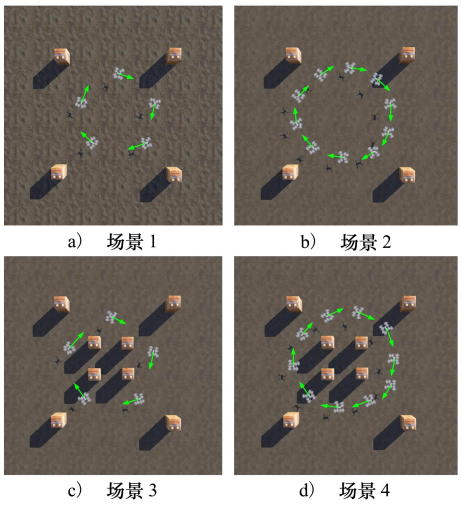


图 12　多种场景无人机的迁移强化学习算法仿真

4　结　论

针对复杂障碍环境下多智能体编队控制问题,本文提出了一种基于迁移强化学习的编队控制算法。基于设计的奖励函数,采用 Nash 均衡价值函数保证了多智能体编队系统的整体性能最优。利用值函数近似方法进行源任务学习,推导了权重更新公式,通过收敛性分析证明了参数更新呈指数收敛。在目标任务学习阶段,通过 GMM 算法对源任务进行聚类分析,基于康托洛维奇距离选择较优的源任务进行目标任务学习,避免了负迁移问题,提高了多智能体编队控制的效率。仿真实验对比分析了迁移前后的运动轨迹,证明了算法的有效性。通过编队任务仿真,实现了避障约束下的多智能体编队控制任务,在 Gazebo 平台,基于无人机模型进行了仿真分析,证明了算法的实时性和可移植性。未来的研究将设计更加精细的奖励函数,让奖励函数在接近最优解时大幅度增加,可以帮助学习算法快速收敛到最优解。考虑将本文提出的迁移学习算法从二维环境拓展到三维环境,模拟更加复杂的现实环境,同时通过时间约束或者增加燃料成本的惩罚来限制智能体的行动选择。

参考文献:

[1]　DORRI A, KANHERE S S, JURDAK R. Multi-agent systems: a survey[J]. IEEE Access, 2018, 6: 28573-28593

[2]　OH K K, PARK M C, AHN H S. A survey of multi-agent formation control[J]. Automatica, 2015, 53: 424-440

[3]　MOHIUDDIN A, TAREK T, ZWEIRI Y, et al. A survey of single and multi-UAV aerial manipulation[J]. Unmanned Systems, 2020, 8(2): 119-147

[4]　GE J, FAN C, YAN C, et al. Multi-UAVs close formation control based on wild geese behavior mechanism[C]//2019 Chinese

Automation Congress, 2019: 967-972

[5] HUO M, DUAN H, FAN Y. Pigeon-inspired circular formation control for multi-UAV system with limited target information[J]. Guidance, Navigation and Control, 2021, 1(1): 2150004

[6] LIN Y, WU X, WANG X, et al. Bio-inspired formation control for UUVs swarm based on social force model[C]//International Conference on Autonomous Unmanned Systems, Singapore, 2021: 3250-3259

[7] 李正平, 鲜斌. 基于虚拟结构法的分布式多无人机鲁棒编队控制[J]. 控制理论与应用, 2020, 37(11): 2423-2431
LI Zhengping, XIAN Bin. Robust distributed formation control of multiple unmanned aerial vehicles based on virtual structure [J]. Control Theory & Applications, 2020, 37(11): 2423-2431 (in Chinese)

[8] YAN X, JIANG D, MIAO R, et al. Formation control and obstacle avoidance algorithm of a multi-USV system based on virtual structure and artificial potential field[J]. Journal of Marine Science and Engineering, 2021, 9(2): 161

[9] RIAH A, AGUSTINAH T. Formation control of multi-robot using virtual structures with a linear algebra approach[J]. Journal on Advanced Research in Electrical Engineering, 2020, 4(1): 45-50

[10] XUAN Mung N, HONG S K. Robust adaptive formation control of quadcopters based on a leader-follower approach[J]. International Journal of Advanced Robotic Systems, 2019, 16(4): 1-11

[11] HE S, WANG M, DAI S L, et al. Leader-follower formation control of USVs with prescribed performance and collision avoidance [J]. IEEE Trans on Industrial Informatics, 2018, 15(1): 572-581

[12] TANG Z, CUNHA R, HAMEL T, et al. Formation control of a leader-follower structure in three dimensional space using bearing measurements[J]. Automatica, 2021, 128: 109567

[13] WU Y, GOU J, HU X, et al. A new consensus theory-based method for formation control and obstacle avoidance of UAVs[J]. Aerospace Science and Technology, 2020, 107: 106332

[14] HE X, GENG Z. Consensus-based formation control for nonholonomic vehicles with parallel desired formations[J]. International Journal of Control, 2021, 94(2): 507-520

[15] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 2018

[16] AFIFI A M, ALHOSAINY O H, ELIAS C M, et al. Deep policy-gradient based path planning and reinforcement cooperative Q-learning behavior of multi-vehicle systems[C]//IEEE International Conference on Vehicular Electronics and Safety, 2019: 1-7

[17] LIN J L, HWANG K S, WANG Y L. A simple scheme for formation control based on weighted behavior learning[J]. IEEE Trans on Neural Networks and Learning Systems, 2013, 25(6): 1033-1044

[18] ZHU P, DAI W, YAO W, et al. Multi-robot flocking control based on deep reinforcement learning[J]. IEEE Access, 2020, 8: 150397-150406

[19] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[20] NIU S, LIU Y, WANG J, et al. A decade survey of transfer learning[J]. IEEE Trans on Artificial Intelligence, 2020, 1(2): 151-166

[21] ZENG M, LI M, FEI Z, et al. Automatic ICD-9 coding via deep transfer learning[J]. Neurocomputing, 2019, 324: 43-50

[22] BYRA M, WU M, ZHANG X, et al. Knee menisci segmentation and relaxometry of 3D ultrashort echo time cones MR imaging using attention U-net with transfer learning[J]. Magnetic Resonance in Medicine, 2020, 83(3): 1109-1122

[23] PETEGROSSO R, PARK S, HWANG T H, et al. Transfer learning across ontologies for phenome-genome association prediction [J]. Bioinformatics, 2017, 33(4): 529-536

[24] HWANG T, KUANG R. A heterogeneous label propagation algorithm for disease gene discovery[C]//Proceedings of the 2010 SIAM International Conference on Data Mining, 2010: 583-594

[25] ABDI H. Partial least squares regression and projection on latent structure regression(PLS Regression)[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(1): 97-106

[26] LU C, HU F, CAO D, et al. Transfer learning for driver model adaptation in lane-changing scenarios using manifold alignment [J]. IEEE Trans on Intelligent Transportation Systems, 2019, 21(8): 3281-3293

[27] HU G, ZHANG Y, YANG Q. Transfer meets hybrid: a synthetic approach for cross-domain collaborative filtering with text[C]// The World Wide Web Conference, 2019: 2822-2829

[28] ZHUANG F, ZHOU Y, ZHANG F, et al. Sequential transfer learning: cross-domain novelty seeking trait mining for recommendation[C]//Proceedings of the 26th International Conference on World Wide Web Companion, 2017: 881-882

[29] 胡鹏林, 潘泉, 武胜帅, 等. 基于迁移强化学习的多智能体系统协同编队避障与防撞控制[C]//2021 中国自动化大会论文集, 2021: 591-596
HU Penglin, PAN Quan, WU Shengshuai, et al. Transfer reinforcement learning-based cooperative formation control of multi-agent systems with collision and obstacle aviodance[C]//Proceedings of 2021 China Automation Conference, 2021: 591-596 (in Chinese)

Study on learning algorithm of transfer reinforcement for multi-agent formation control

HU Penglin, PAN Quan, GUO Yaning, ZHAO Chunhui

(School of Automation, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: Considering the obstacle avoidance and collision avoidance for multi-agent cooperative formation in multi-obstacle environment, a formation control algorithm based on transfer learning and reinforcement learning is proposed. Firstly, in the source task learning stage, the large storage space required by Q-table solution is avoided by using the value function approximation method, which effectively reduces the storage space requirement and improves the solving speed of the algorithm. Secondly, in the learning phase of the target task, Gaussian clustering algorithm was used to classify the source tasks. According to the distance between the clustering center and the target task, the optimal source task class was selected for target task learning, which effectively avoided the negative transfer phenomenon, and improved the generalization ability and convergence speed of reinforcement learning algorithm. Finally, the simulation results show that this method can effectively form and maintain formation configuration of multi-agent system in complex environment with obstacles, and realize obstacle avoidance and collision avoidance at the same time.

Keywords: multi-agent system; transfer reinforcement learning; value function approximation; formation control; Gaussian clustering

引用格式: *胡鹏林, 潘泉, 郭亚宁, 等. 多智能体编队控制中的迁移强化学习算法研究[J]. 西北工业大学学报, 2023, 41(2): 389-399*
HU Penglin, PAN Quan, GUO Yaning, et al. Study on learning algorithm of transfer reinforcement for multi-agent formation control[J]. Journal of Northwestern Polytechnical University, 2023, 41(2): 389-399 (in Chinese)