高技术通讯 2022 年 第32 卷 第8 期:836-844

doi:10.3772/j.issn.1002-0470.2022.08.006

不确定环境下的深度强化学习编队避障控制①

禹鑫燚② 杜丹枫 欧林林③

(浙江工业大学信息工程学院 杭州 310023)

摘 要 多智能体编队避障控制的目的在于保持多智能体队形的同时完成避障。针对复杂环境的随机性和不确定性,提出了一种不确定环境下的深度强化学习编队避障控制方法。首先,设计了价值评估网络来增加多智能体编队过程中触碰障碍物或者到达期望位置这些特殊动作的经验,使智能体更快地理解环境规则。其次,在智能体选择动作时,基于贪心策略,对动作选择策略进行改进以提高智能体的学习效率。再次,设计了样本存储空间,在增加样本的利用率的同时提高模型训练效率,并且在决策阶段结合多步学习算法使价值估计更准确。最后,将提出的方法与其他算法进行了对比实验。仿真结果表明提出的方法能使多个智能体在维持队形的同时进行避障,并且有效地提高了智能体学习效率。

关键词 深度强化学习;避障;编队控制;多智能体;神经网络

0 引 言

多智能体系统在军事、卫星群协同控制、无人机编队控制等方面都有广阔的应用前景[1-3],因此得到了各界学者的广泛关注[4]。其优点在于利用多个智能体协作完成单个智能体无法完成的复杂任务。多智能体的编队一直是多智能体系统的研究热点[5-6],它要求智能体以特定的几何形状形成集群或者以期望的队形完成特定的任务。如何使多智能体系统在避开障碍物的同时保持队形,是多智能体编队控制的关键问题。

已有的编队控制方法有领航跟随法[7]、虚拟结构法[8]、基于行为法[9]和基于图论法[10]等。文献[11]研究了一种基于视觉的领航跟随跟踪策略。文献[12]设计了一种基于虚拟结构的避障方法,为每个机器人生成基本轨迹。文献[13]基于图论提出了一种新颖的自适应编队控制方法,用于解决非线性多智能体系统的编队控制问题。虽然多智能体编队控制已经取得了一系列的研究成果,但是在面对复杂环境或者动态环境时适应能力不足。在不确定环境中,多智能体编队的避障不够灵活,智能体之间的碰撞避免以及智能体与障碍物的碰撞避免,给多智能体编队避障控制带来了挑战。

深度强化学习[14-16]可以不依赖环境模型,适用于未知环境中的决策控制问题。同时由于深度强化学习拥有强大的感知和学习能力[17-18],在多智能体领域已经取得了较为成功的应用[19-22]。基于强化学习的多智能体编队控制具有传统编队方法所不具备的优点,可以在不断的试错中进行学习来解决编队避障控制问题。目前已有诸多学者将深度强化学习与传统编队控制相结合,并且取得了较好的成果。针对动力学未知的非线性多智能体编队控制,文献[23]提出了结合模糊逻辑系统和强化学习的优化控制方案来实现编队控制。文献[24]提出了一种基于深度学习的无人机编队协调控制算法,使得无人机能够在大规模复杂环境中形成特定队形并执行导航任务。考虑到编队过程中的碰撞避免问题,

① 国家重点研发计划(2018YFB1308001),浙江省自然科学基金(LY21F030018)和浙江省重点研发计划(2020C01190)资助项目。
② 男,1979 年生,博士,副教授;研究方向:机器人控制与规划;E-mail:yuxinyinet@163.com。
③ 通信作者,E-mail:linlinou@zjut.edu.cn。
(收稿日期:2021-02-23)

文献[25,26]使用深度强化学习方法优化领航跟随算法,实现了多智能体的编队避障控制。文献[27]将基于行为的控制方法和深度强化学习相结合,使编队可以在保持队形的同时避开障碍物。为了进一步提高编队避障的成功率,文献[28]和[29]利用深度强化学习强大的学习能力,训练了一种多智能体编队避障策略,有效降低了智能体之间的碰撞概率。文献[30]将深度学习方法与传统碰撞回避算法相结合,在编队的过程中,采用长短期记忆来感知任意数量的障碍物信息,并设计了复合奖励函数来提高编队避障的成功率。上述文献基于深度强化学习实现了多智能体的编队避障控制,并且在不同方面做出了优化,但是对于多个智能体的学习过程长、学习速率慢的问题,目前研究还不够深入。

为进一步缩短编队过程中智能体的学习时间,并且加快智能体学习效率,本文提出了一种不确定环境下的深度强化学习编队避障控制方法。首先,在智能体学习的初始阶段,建立了价值评估网络,增加智能体选择触碰障碍物或者到达期望位置这些特殊动作的经验。其次,在智能体选择动作时,基于贪心策略,改进动作选择策略,提高了算法的学习效率。然后设计了样本存储空间,增加样本的利用率。最终,结合多步学习算法,使价值估计更准确。通过本文提出的深度强化学习编队避障控制方法,智能体可以在不确定环境中通过学习完成编队避障任务。为验证本文方法在不确定环境下的有效性,本文设置了不同的障碍物环境进行仿真实验。仿真结果表明本文所提算法能够使多个智能体在不确定环境下较好地实现编队避障任务。

1 问题描述

本文所要解决的主要问题是确定每个智能体的最优控制策略,使得智能体到达各自的期望位置形成队形,并且能在维持队形不变的情况下有效避开障碍物。假设存在 $N(N \geqslant 2)$ 个智能体随机分布在二维空间内,每个智能体对应着不同的期望位置。在多智能体编队过程中,位置坐标表示智能体 i 的状态,并且朝着期望位置 $G_i(i = 1,2,\cdots,N)$ 运动,同时智能体互相之间不发生碰撞并且能有效避开障碍物。智能体 i 在运动过程中有 5 种可能的动作可供选择,即动作集合 $A_i(s)$ 为{前,后,左,右,保持原地}。

将上述多智能体编队避障控制问题表述为强化学习问题。在不确定环境下的深度强化学习编队避障控制问题中,对于每个智能体,s_t 和 a_t 分别表示 t 时刻的状态和动作。智能体 i 的位置坐标为 $\boldsymbol{p}_{t,i} = [p_{t,i}^x, p_{t,i}^y]$,速度为 $\boldsymbol{v}_{t,i} = [v_{t,i}^x, v_{t,i}^y]$,期望位置的坐标为 $\boldsymbol{p}_{g,i} = [p_{g,i}^x, p_{g,i}^y]$,可观测的智能体状态为 $\boldsymbol{s}_{t,i} = [p_{t,i}^x, p_{t,i}^y, v_{t,i}^x, v_{t,i}^y, g_{t,i}]$,其中 $g_{t,i}$ 表示智能体是否到达期望位置。奖励值函数的设计为 $\boldsymbol{R} = [r_O, r_G, r_F]$,其中 r_O 表示智能体触碰障碍物的奖励值,r_G 表示智能体到达目标位置的奖励值,r_F 表示多个智能体保持队形的奖励值。如果目标编队完成,即智能体到达相应的目标位置时,会获得正向的奖励值,而智能体触碰障碍物或者智能体队形被破坏则会得到负向的惩罚。根据上述定义,本文使用一个五元组 $< I,S,A_i(s), P,\{r_i\} >$ 来表示多智能体编队避障控制过程,其中,I 为有限个智能体的集合;S 为每个智能体可观测状态的集合;$A_i(s)$ 为第 i 个智能体在状态 $s \in S$ 下可以选择的动作集合;P 为状态转移函数,是指给定智能体在当前状态和联合行为时,下一状态的概率分布;$\{r_i\}$ 表示多个智能体在采取不同动作之后的奖励值的集合。多个智能体在 s 状态下的联合动作可以表示为 $A(s) = A_1(s) \times A_2(s) \times A_3(s) \cdots \times A_N(s)$。在学习过程中,每个智能体与环境不断进行交互,获取智能体自身的状态信息。多个智能体的状态信息组合成联合状态输入到神经网络,智能体根据动作选择策略选取自身的动作,获得下一时刻的状态和奖励值函数值。智能体与环境交互产生的数据元组 $\{s_{t,i}, a_{t,i}, r_{t+1,i}, s_{t+1,i}\}$ 被存储到经验池中。在每一回合,从经验池中进行采样学习,最终智能体通过学习确定最优控制策略 π,为队形保持和碰撞避免选择最优动作。当智能体执行策略 π 时,可以最大化智能体的奖励总和 $R_i = \sum_{t=0}^{T} \gamma^t r_i{}^t$,其中 γ 是折扣因子,t 表示时间,T 是终止时间。

2 不确定环境下的深度强化学习编队避障控制方法

将多智能体编队避障控制问题抽象为强化学习过程,目的是通过学习得到最优策略,使智能体在保持队形的同时避免碰撞并到达期望位置。本文建立了价值评估网络,改进了智能体动作选择策略,设计了样本存储空间,同时结合了多步学习算法,提出了不确定环境下的深度强化学习编队避障控制方法。

2.1 奖励值和动作选择策略的设计

奖励函数的设计对深度强化学习编队避障控制任务尤为重要。在本文中,智能体互相之间发生碰撞或者触碰障碍物以及无法保持队形会获得一个负的奖励值,智能体到达各自的期望位置则会获得一个正的奖励值,其他时刻奖励值为 0。

$$r_t = \begin{cases} r_{\text{crash}} & \text{发生碰撞} \\ r_{\text{reach}} & \text{到达期望位置} \\ r_{\text{formation}} & \text{队形破坏} \\ 0 & \text{其他} \end{cases} \tag{1}$$

其中 r_{crash} 是智能体之间发生碰撞或者触碰到障碍物的奖励值; r_{reach} 是智能体到达期望位置的奖励值; $r_{\text{formation}}$ 表示多智能体无法维持队形时的奖励值。

合理而有效的动作选择策略设计可以减少学习的时间。将贪心策略用于动作选择,来平衡学习过程中的探索与利用,求解出接近真实的价值模型。贪心策略定义如下:

$$\pi(s_t) = \begin{cases} \max\limits_{a \in A} q(s_t, a) & \mu \leqslant \varepsilon \\ \tilde{a} & \mu > \varepsilon \end{cases} \tag{2}$$

其中 $\mu \in [0,1]$ 是每个回合产生的随机值, ε 是探索速率, $\tilde{a}$ 是动作空间 A 中的一个随机动作。贪心策略可以使每个智能体有 $1-\varepsilon$ 的概率随机选择动作。在训练前期,智能体需要多次探索,以获取不同的动作价值,避免陷入局部最优;而经过一段时间训练之后,智能体逐渐学习到最优策略,就可直接选择正确的动作,尽可能获取更多的奖励值。ε 的取值随着迭代次数的增加而增加,最终值为 1。

当智能体选择动作时,在保留一定概率随机选择动作的基础上,对贪心策略作了改进。为了加快智能体在前期的探索效率,本文建立了一个价值评估网络使智能体更快地理解环境。智能体在选择碰撞或者到达期望位置等特殊动作时,会产生特殊经验,价值评估网络被用来评价选择的特殊动作的价值。该网络 E 的损失函数定义为

$$L_t(\theta_t) = E_{s,a}\{((1 + |r_t|) - e(s_t, a_t; \theta_t^E))^2\} \tag{3}$$

其中价值评估函数 $e(s_t, a_t; \theta_t^E)$ 经过训练逐渐趋向 $1 + |r_t|$, θ^E 是价值评估网络的更新权重。结合式(1)和式(3), $e(s_t, a_t; \theta_t^E)$ 将收敛到:

$$e(s_t, a_t, \theta_t^E) = \begin{cases} 1 + |r_{\text{crash}}| & \text{发生碰撞} \\ 1 + |r_{\text{reach}}| & \text{到达期望位置} \\ 1 + |r_{\text{formation}}| & \text{队形破坏} \\ 1 & \text{其他} \end{cases} \tag{4}$$

价值评估网络 E 的训练需在网络 Q 之前完成,然后帮助选择动作。结合贪心策略,将动作选择策略设计为

$$a_t = \arg\max_{a \in A} q(s_t, a; \theta_t) e(s_t, a_t; \theta_t^E) \tag{5}$$

式(5)中, $e(s_t, a_t; \theta_t^E)$ 可以增加智能体特殊经验的比例,即在训练前期,鼓励智能体选择下一步的碰撞或者到达期望位置的动作。当 $Q(s_t, a_t; \theta_t)$ 开始正确识别障碍物时, $e(s_t, a_t; \theta_t^E)$ 可以抑制碰撞,鼓励智能体探索更多的位置。

2.2 样本存储空间设计

通过 2.1 节中的动作选择策略,智能体与环境进行交互产生学习样本,存入样本存储空间中。样本存储空间具备采样功能,通过计算每个样本的时间差分误差(temporal difference error, TD-Error),即样本的估计值和实际值之间的差距,将其作为当前样本的采样权重。越大的 TD-Error 表示样本的估计值和实际值之间的差距越大,样本越有价值。智能体编队避障控制过程中的 TD-Error 的定义为

$$p_i = r_{t+1} + \gamma \max_a Q(s_{t+1}, a; \theta_t) - Q(s_t, a_t, \theta_t) \tag{6}$$

其中, r_{t+1} 是 $t+1$ 时刻的奖励值, γ 是折扣因子, s 和 a 是智能体的状态和动作, θ_t 是神经网络参数。使用随机采样算法,在以 TD-Error 为权重的采样和均匀采样之间进行插值。根据式(6),将样本 i 的采

样概率 P 表示为

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \tag{7}$$

其中 p_i 和 p_k 表示样本 i 和 k 的 TD-Error, α 可以调整 TD-Error 的权重。为了在提高样本利用率的同时确保不会有太大的偏差,结合重要性采样定理对原来的概率计算增加权重:

$$\omega_i = \left(\frac{1}{N}\frac{1}{P(i)}\right)^\beta \tag{8}$$

其中 N 是存储的样本数量。在整个训练过程中, β 的初始值为 0,并且随着迭代学习的进行,线性增长为 1。

样本存储空间的功能可以描述为:当智能体通过动作选择策略与环境交互产生样本存入样本存储空间时,计算每个样本 i 的采样概率 $P(i)$; 在样本取出时,以概率 $P(i)$ 进行采样;在更新时,为每个样本添加权重 ω_i, 随着训练的进行, β 从初始值线性增长为 1。

2.3 算法设计

本文采用如下的动作-值函数来估计所学到的策略:

$$Q(s, a) = E[R \mid s_t = s, a_t = a] \tag{9}$$

其中 E 表示期望。式(9)可以递归计算为

$$Q(s,a) = E_{s'}[r(s, a) + \gamma E_{a' \sim \pi}(Q(s', a'))] \tag{10}$$

其中 $r(s, a)$ 表示在状态 s 执行动作 a 的奖励值, s' 和 a' 分别是下一时刻的状态和动作, γ 是折扣因子, $a' \sim \pi$ 表示智能体通过动作选择策略 π 采取下一步动作 a'。深度 Q 学习算法基于时序差分法的思想,通过贝尔曼方程进行自迭代更新,更新公式为

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{11}$$

深度 Q 学习算法大多通过下一时刻的回报和价值估计得到目标价值,这种方法在前期具有学习速度较慢的缺点。为了提高学习速度,本文结合多步学习算法,以使训练前期目标价值可以估计得更准确,从而加速训练。多步学习算法的公式为

$$G_{t:t+n} = r_{t+1} + \gamma r_{t+2} + \cdots \gamma^{n-1} r_{t+n} + \gamma^n \max_a Q(S_{t+n+1}, a', \theta_t) \tag{12}$$

其中, γ 是折扣因子, r 是奖励值, θ_t 是神经网络的参数。结合式(11)和式(12),值函数的更新公式为

$$Q_{t+n}(S_t, A_t, \theta_t) = Q_{t+n-1}(S_t, A_t, \theta_t) + \alpha[G_{t:t+n} - Q_{t+n-1}(S_t, A_t, \theta_t)] \tag{13}$$

其中 α 是学习速率, S_t 是状态空间, A_t 是动作空间, θ_t 是神经网络参数。

智能体与环境进行交互产生数据 (s, a, r, s'), 使用经验回放池来存储交互产生的这些数据,并且通过最小化损失函数来学习智能体的最优策略。损失函数的定义为

$$L(\theta) = E[\omega(G_t - Q(S_t, A_t, \theta_t))^2] \tag{14}$$

其中, G_t 是根据多步学习算法得到的实际值, $Q(S_t, A_t, \theta_t)$ 为估计值, ω 是由式(8)得到的权重。

根据奖励值、动作选择策略和样本存储空间的设计,结合多步学习算法,获得适用于不确定环境下的深度强化学习编队避障控制的算法如下。

(1)初始化容量为 N 的 Replay Buffer: D;

(2)初始化状态行动价值模型 Q 和参数 θ; 初始化 Target Network $\hat{Q}$ 和参数 θ', 价值评估网络 E 和参数 θ^E; 初始化 $t = 0$;初始化 $batch\_size$ 大小为 m; 初始化多步学习算法步数 n;

(3)初始化环境得到初始状态 s_1;

(4)智能体 i 随机选择动作 $a_{t,i}$, 从 D 中采样进行训练并计算 $y_{j,i} = r_{j,i}$, 根据式(3)计算神经网络 E 的损失函数,训练并更新 E 的参数 $\theta^E_{t+1,i}$;

(5)智能体 i 以 ε 的概率选择一个动作 $a_{t,i}$, 或者根据式(5)选择当前最优动作 $a_{t,i} = \arg\max_{a \in A} q(s_{t,i}, a_{t,i}; \theta_{t,i}) e(s_{t,i}, a_{t,i}; \theta^E_{t,i})$;

(6)智能体 i 执行动作 $a_{t,i}$, 得到新一轮的状态 $s_{t+1,i}$ 和奖励值 $r_{t+1,i}$;

(7)将样本数据 $\{s_{t,i}, a_{t,i}, r_{t+1,i}, s_{t+1,i}\}$ 存储到 D 中;

(8)从 D 中采样一批样本进行训练,根据式(12)计算 $y_{j,i}$ 的值,当 $s_{t+1,i}$ 为最终状态时, $y_{j,i} = r_{j,i}$, 否则 $y_{j,i} = r_{j,i} + \sum_{k=1}^{n} \gamma^{k-1} R_{t+k,i} + \gamma^n \max_a q(s_{t+n,i}, a', \theta')$;

(9)根据式(14)计算损失函数 $L(\theta) =$

$\frac{1}{m}\sum_{j=1}^{m}\omega_j\left(y_j - q(s_t, A_t, \theta)\right)^2$；

(10)每隔 C 轮进行参数更新 $\theta' \leftarrow \theta_{t+1}$；

(11)如果完成一次迭代训练，返回步骤(3)，否则返回步骤(4)。

在训练过程中，智能体与环境不断进行交互，获取智能体自身的状态信息。多个智能体的状态信息组合成联合状态输入到神经网络中，智能体根据改进的动作选择策略选取自身的动作，得到下一时刻的状态和奖励函数值。交互过程中产生的智能体状态-动作值被存储到经验池中，在每一回合，从经验池中进行采样学习。本文提出的算法可以通过学习得到多智能体编队避障控制的最优策略，使多个智能体到达期望位置形成队形，并有效进行避障。

3　仿真实验

为了验证在不确定环境下本文提出算法的有效性，在智能体学习过程中添加额外的障碍物。同时，本文针对 2 种不同的障碍物环境，分别进行 4 个智能体和 6 个智能体的编队避障控制。智能体通过迭代学习形成期望队形，有效避开障碍物并到达期望位置视为一次成功，文中以训练过程中的成功率为指标，将本文提出的方法与 Double DQN[31] 和 MADDPG[32] 2 种算法进行成功率对比，验证了本文算法的有效性。仿真实验共进行 2000 回合的训练。在正式训练前需要进行预训练，用于收集经验数据以进行批次训练。训练过程中，ε 的取值从初始值0.1增长到 1。仿真实验的参数设计详见表 1。

表 1　参数设计

参数	值
每次采样大小($batch\_size$)	16
折扣因子(γ)	0.95
学习速率(α)	0.0001
replay buffer 的容量(N)	16 000
多步学习算法的步数(n)	4
选取最优动作的概率(ε)	[0.1,1]

在二维空间内，基础的动作空间只包含{前，后，左，右，保持原地}5 个动作。为了加快学习记忆的过程，本文在水平面内将行为空间划分为 8 个离散的动作，使智能体有更多的动作选择。如图 1 所示，本文的动作空间包含 8 个方向选择，$A(s)$ = {0°,45°,90°,135°,180°,225°,270°,315°}，智能体每次动作的步幅是 0.2。奖励值的设定为智能体发生碰撞时 $r_{\text{crash}} = -1$，到达期望位置时 $r_{\text{reach}} = 1$，$r_{\text{formation}} = -1$。

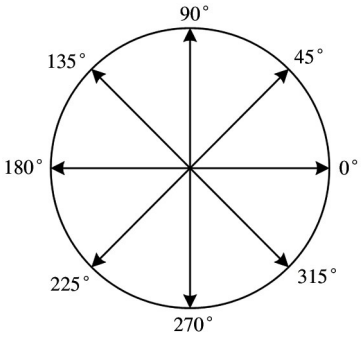

图 1　动作空间

如图 2 所示，在二维空间内随机生成 4 个智能体的位置，在初始阶段，智能体经过训练学习形成正四边形。然后，多个智能体在保持队形不发生变化的前提下，通过迭代学习寻找到一条最优路径，避开障碍物的同时到达期望位置。图中深色正方形区域为原本已存在的障碍物，浅色正方形区域则在智能体学习过程中新加入的障碍物，坐标左下角的圆点为智能体，坐标右上角的圆点所在位置为各个智能体的期望位置。由图2可知，即使在智能体学习过

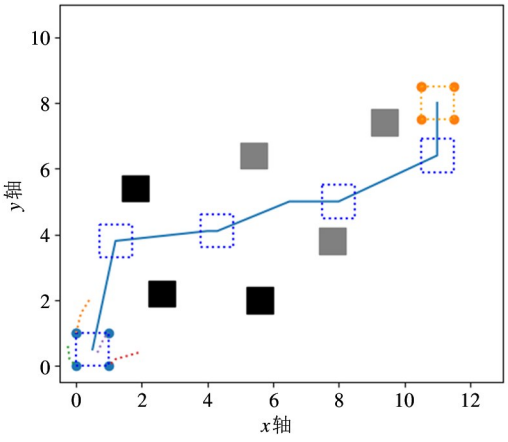

图 2　不确定环境下的智能体编队避障轨迹

程中增加新的障碍物,多个智能体也可以通过本文提出的方法形成期望的队形,有效避开障碍物并到达期望位置。

本文考虑了在更复杂的环境下 4 个智能体的编队控制问题,同时将智能体编队避障的成功率与其他 2 种算法进行对比,验证了本文提出算法的有效性。如图 3 所示,图中正方形为障碍物,坐标左下角的圆点为智能体,坐标右上角的圆点所在位置为各个智能体的期望位置。由图 3 可知,本文提出的方法可以使多个智能体在面对不同的环境时形成期望的队形,同时经过迭代学习得到最优策略,在有效避开障碍物的同时到达期望位置。

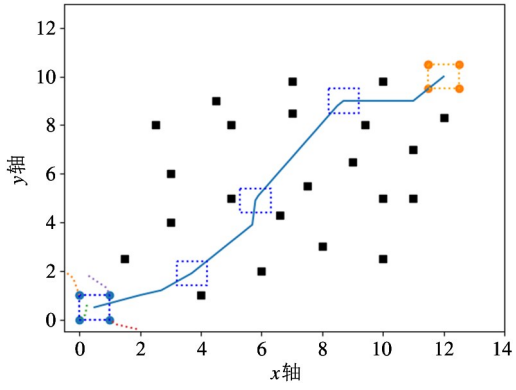

图 3　4 个智能体编队避障轨迹

图 4 表示智能体保持队形避障过程中每个回合的 Q 值之和。刚开始所有 Q 值为 0,随着迭代训练的进行,根据奖励值进行 Q 值的更新。由图 4 可知,在训练初期由于 ε 值小,智能体随机选择动作概率大,大概率触碰障碍物获得负向奖励值;在训练中后期,神经网络对整个样本空间有了相对全面的采样,在此基础上,神经网络通过训练不断对 Q 值进行泛化,同时 ε 值增长,智能体可以根据经验选择最优动作,获得正向奖励值的概率逐渐增大。经过不断学习,成功避开障碍物到达期望位置的概率越来越高。图 5 表示每个回合智能体的步数。由图 5 可知,4 个智能体经过 1250 个回合的学习,最终学习到最优策略,寻找到避开障碍物到达期望位置的最短路径。将所提出的算法与 Double DQN 算法和 MADDQN 算法进行了对比,3 种算法的成功率如图 6所示。由图 6 可知,本文提出的算法最终的成功率更高。

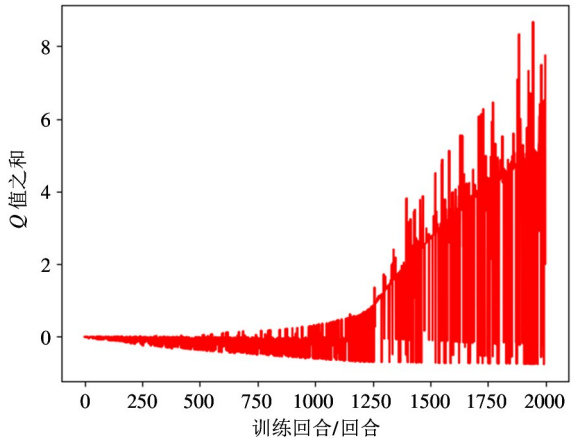

图 4　智能体学习曲线

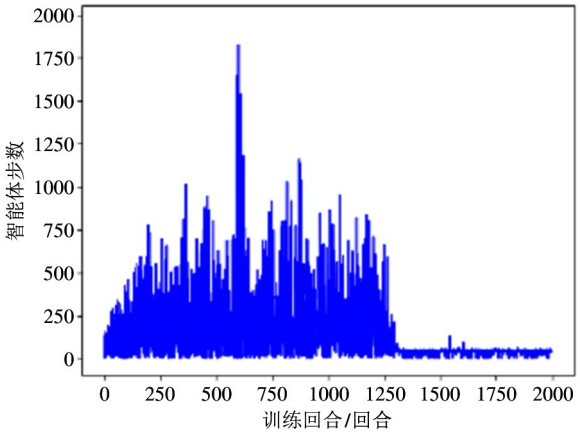

图 5　智能体每个回合的步数

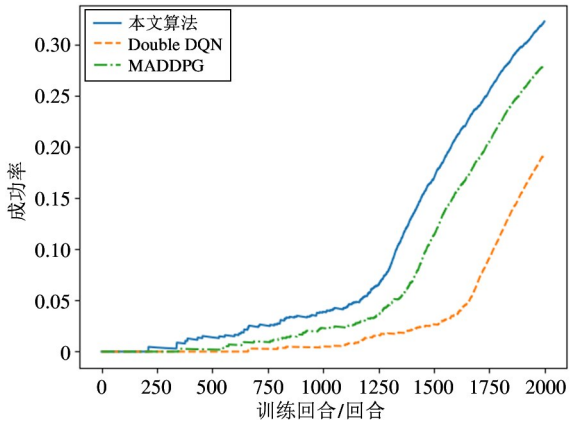

图 6　3 种算法的成功率对比

本文同时还考虑了 6 个智能体在不同环境下的编队避障控制,如图 7 所示。图中坐标左下角的圆点为智能体,坐标右上角的圆点为各个智能体的期望位置,块状区域为障碍物。智能体通过迭代学习形成正六边形,并且有效避开障碍物到达期望位置。

图 8 为智能体的学习曲线，图 9 为智能体每个回合的步数。结合图 8 和图 9 可知，智能体通过训练学习，在 1100 回合之后获得最优策略，在保持队形不发生变化的前提下，有效避开障碍物到达期望位置。图 10 表示 6 个智能体环境下本文方法与 Double DQN 和 MADDPG 2 种算法的成功率对比。由图 10 可知，本文方法在多智能体编队避障控制问题中的成功率更高。

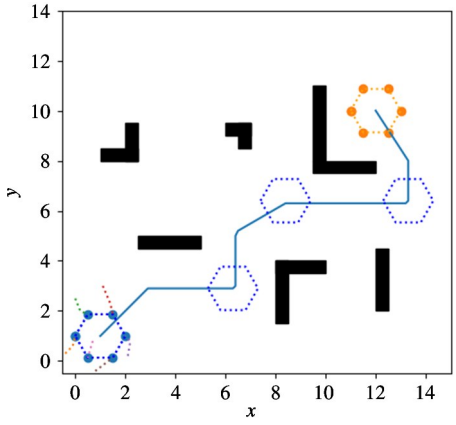

图 7　6 个智能体编队避障轨迹

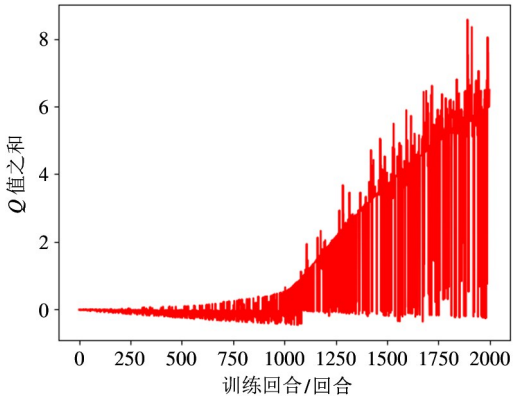

图 8　智能体学习曲线

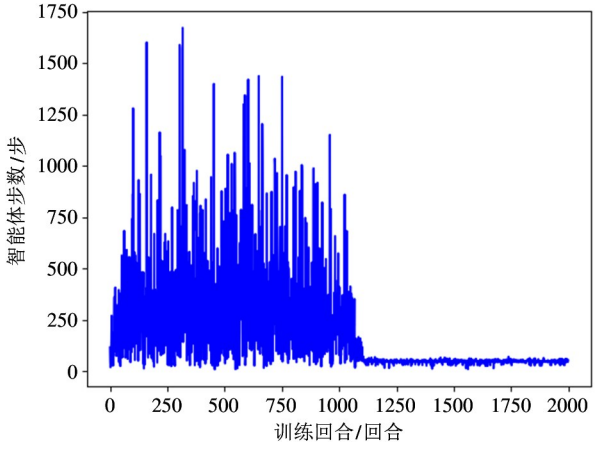

图 9　智能体每个回合的步数

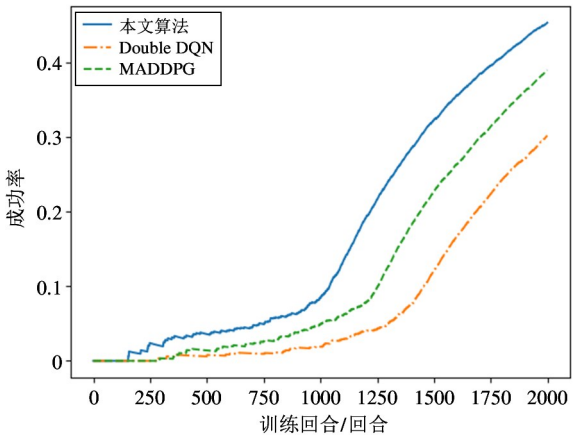

图 10　3 种算法的成功率对比

为了验证在不确定环境下本文提出算法的有效性，在智能体学习过程中添加额外的障碍物。仿真结果表明多个智能体在不确定环境下能形成特定队形避开障碍物到达期望位置。同时，在 2 种不同的环境下，针对 4 个和 6 个智能体进行了仿真实验，并且将所提的方法与 Double DQN 和 MADDPG 算法进行对比。结合图 3 和图 7 可知，本文提出的方法面对不同环境都能实现多智能体的编队避障控制。多个智能体通过迭代学习形成期望队形，并且有效避开障碍物到达期望位置。由图 6 和图 10 可知，本文提出的方法在前期能更快地获取成功的经验，学习速率更快，并且最终的成功率也相对更高。这表明了本文设计的价值评估网络能帮助智能体更快地取得到达期望位置的特殊经验。

4　结　论

针对复杂环境的随机性和不确定性，本文提出了一种不确定环境下的深度强化学习编队避障控制方法。在该方法中，设计了价值评估网络来增加编队过程中的智能体选择触碰障碍物或者到达期望位置这些特殊动作的经验，使智能体更快地理解环境规则。并且将该价值评估网络和贪心策略相结合，对动作选择策略进行改进，提高算法的学习效率。同时，设计了样本存储空间，增加样本利用率的同时

提高了模型训练效率。在决策阶段,结合多步学习算法使价值估计更准确。通过仿真实验验证了本文提出方法的有效性,能在不同的环境下较好地完成多智能体编队任务。仿真结果表明本文提出的方法可以适用于各种不确定环境中。将本文提出的方法和 Double DQN 与 MADDPG 2 种算法进行对比,结果表明本文方法收敛速度更快,智能体编队避障的成功率更高。

参考文献

[1] CUI R, LI Y, YAN W, et al. Mutual information-based multi-AUV path planning for scalar field sampling using multidimensional RRT[J]. *IEEE Transactions on System, Man and Cybernetics: Systems*, 2016, 46(7): 993-1004

[2] PENG Z, WANG J, WANG D, et al. Distributed maneuvering of autonomous surface vehicles based on neurodynamic optimization and fuzzy approximation[J]. *IEEE Transactions on Control Systems Technology*, 2018, 26(3): 1083-1090

[3] WANG X, LI S, SHI P, et al. Distributed finite-time containment control for double-integrator multiagent systems[J]. *IEEE Transactions on Cybernetics*, 2014, 44(9): 1518-1528

[4] GE X, HAN Q L, DING D, et al. A survey on recent advances in distributed sampled-data cooperative control of multi-agent systems[J]. *Neurocomputing*, 2018, 275: 1684-1701

[5] LI J Y, SUN K X, MA H, et al. Moving agents in formation in congested environments[C]// Proceedings of the 19th International Conference on Autonomous Agents and Multi-agent Systems, Auckland, New Zealand, 2020: 726-734

[6] DEMIROVIC E, SCHWIND N, OKIMOTO T, et al. Recoverable team formation: building teams resilient to change[C]// Proceedings of the 17th International Conference on Autonomous Agents and Multi-agent Systems, Stockholm, Sweden, 2018: 1362-1370

[7] DAI S L, HE S, CHEN X, et al. Adaptive leader-follower formation control of nonholonomic mobile robots with prescribed transient and steady-state performance[J]. *IEEE Transactions on Industrial Informatics*, 2020, 16(6):3662-3671

[8] CHEN X L, HUANG F H, ZHANG Y G, et al. A novel virtual-structure formation control design for mobile robots with obstacle avoidance[J]. *Applied Sciences*, 2020, 10(17): 1-23

[9] BALCH T, ARKIN R C. Behavior-based formation control for multi-robot teams[J]. *IEEE Transactions on Robotics and Automation*, 1998, 14(6): 926-939

[10] DING Y, CONG Y R, WANG X K, et al. Seeking the scalability in algebraic graph based unmanned aerial vehicle formation control[C]// Proceedings of the 2019 Chinese Control Conference, Guangzhou, China, 2019: 6054-6060

[11] LIU X, GE S S, GOH C H, et al. Vision-based leader-follower formation control of multiagents with visibility constraints[J]. *IEEE Transactions on Control Systems Technology*, 2019, 27(3): 1326-1333

[12] CHEN X L, HUANG F H, ZHANG Y G, et al. A novel virtual-structure formation control design for mobile robots with obstacle avoidance[J]. *Applied Sciences*, 2020, 10(17): 1-23

[13] LI X L, WEN C Y, CHEN C. Adaptive formation control of networked robotic systems with bearing-only measurements[J]. *IEEE Transactions on Cybernetics*, 2020, 51(1): 199-209

[14] NOVOSELLER E R, WEI Y B, SUI Y N, et al. Dueling posterior sampling for preference-based reinforcement learning[C]// Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, Toronto, Canada, 2020: 1029-1038

[15] PEHARZ R, VERGARI A, STELZNER K, et al. Random sum-product networks: a simple and effective approach to probabilistic deep learning[C]// Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence, Tel Aviv, Israel, 2020: 334-344

[16] ZHANG Y, ZHANG Z F, YANG Q Y, et al. EV charging bidding by multi-DQN reinforcement learning in electricity auction market[J]. *Neurocomputing*, 2020, 395: 404-414

[17] MAO H Y, LIU W L, HAO J Y, et al. Neighborhood cognition consisitent multi-agent reinforcement learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020: 7219-7226

[18] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529-533

[19] YANG Y, LI J T, PENG L L, et al. Multi-robot path planning based on a deep reinforcement learning DQN algorithm[J]. *CAAI Transactions on Intelligence Technology*, 2020, 5(3): 177-183

[20] ZHANG H, CHEN W, HUANG Z, et al. Bi-level actor-critic for multi-agent coordination[C]// Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020, 34(5): 7325-7332

[21] FU H T, TANG H Y, HAO J Y, et al. Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces[C]// Proceedings of the 28th International

Joint Conference on Artificial Intelligence, Macao, China, 2019: 2329-2335

[22] LIU Y, HU Y J, GAO Y, et al. Value function transfer for deep multi-agent reinforcement learning based on n-step returns[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 2019: 457-463

[23] WEN G X, CHEN C L, FENG J, et al. Optimized multi-agent formation control based on an identifier-actor-critic reinforcement learning algorithm[J]. *IEEE Transactions on Fuzzy Systems*, 2018, 26(5): 2719-2731

[24] WANG C, WANG J, ZHANG X, et al. A deep reinforcement learning approach to flocking and navigation of UAVs in large-scale complex environments[C]//Proceedings of the 2018 IEEE Global Conference on Signal and Information Processing, Anaheim, USA, 2018: 1228-1232

[25] KNOPP M, AYKIN C, FELDMAIER J, et al. Formation control using GQ(λ) reinforcement learning[C]//Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication, Lisbon, Portugal, 2017: 1043-1048

[26] WEN G X, CHEN C L P, LI B, et al. Optimized formation control using simplified reinforcement learning for a class of multiagent systems with unknown dynamics[J]. *IEEE Transactions on Industrial Electronics*, 2020, 67(9): 7879-7888

[27] LIN J L, HWANG K S, WANG Y L, et al. A simple scheme for formation control based on weighted behavior learning[J]. *IEEE Transactions on Networks and Learning Systems*, 2014, 25(6): 1033-1044

[28] CHEN Y F, LIU M, EVERETT M, et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning[C]//Proceedings of the 2017 IEEE International Conference on Robotics and Automation, Singapore, 2017: 285-292

[29] CHEN Y F, EVERETT M, LIU M, et al. Socially aware motion planning with deep reinforcement learning[C]//Proceedings of the 2017 IEEE/RSJ International Conference Intelligent Robots Systems, Vancouver, Canada, 2017: 1343-1350

[30] SUI Z, PU Z, YI J, et al. Formation control with collision avoidance through deep reinforcement learning using model-guided demonstration[J]. *IEEE Transactions on Neural Network and Learning Systems*, 2020, 32(6): 2358-2372

[31] ZHANG W, GAI J, ZHANG Z, et al. Double-DQN based path smoothing and tracking control method for robotic vehicle navigation[J]. *Computers and Electronics in Agriculture*, 2019, 166: 1-11

[32] ZHANG Y, MOU Z Y, GAO F F, et al. UAV-enabled secure communications by multi-agent deep reinforcement learning[J]. *IEEE Transactions on Vehicular Technology*, 2020, 69(10): 11599-11611

Formation control without collision in uncertain environment based on deep reinforcement learning

YU Xinyi, DU Danfeng, OU Linlin

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023)

Abstract

The purpose of multi-agent formation control is to avoid obstacles while maintaining the formation. For the randomness and uncertainty of the complex environment, a formation and obstacle avoidance control method in uncertain environment based on deep reinforcement learning is proposed in the paper. Firstly, a value evaluation network is designed to increase the experience of special actions, such as touching obstacles or reaching the desired location, so that the agents can understand environmental rules faster. Secondly, when the agents select actions, the action selection strategy is improved based on the greedy strategy, which increases the learning efficiency of the agents. Then, the sample storage space is designed to increase the efficiency of model training while increasing the utilization of samples. And the multi-step learning algorithm is combined to make the value estimation more accurate in the decision-making stage. Finally, the proposed method is compared with other algorithms. The simulation results demonstrate that the proposed method can realize the multi-agent formation control without collision. The algorithm proposed in the paper improves learning rate of multi-agents effectively.

Key words: deep reinforcement learning, collision avoidance, formation control, multi-agent, neural network