基于奖励滤波信用分配的多智能体深度强化学习算法

徐 诚[1),2)] 殷 楠[1)] 段世红[1),2)] 何 昊[1)] 王 然[1),2)]

[1)](北京科技大学计算机与通信工程学院 北京 100083)

[2)](北京科技大学顺德研究生院 广东 佛山 528399)

摘 要 近年来,强化学习方法在游戏博弈、机器人导航等多种应用领域取得了令人瞩目的成果.随着越来越多的现实场景需要多个智能体完成复杂任务,强化学习的研究领域已逐渐从单一智能体转向多智能体.而在多智能体强化学习问题的研究中,让智能体学会协作成为当前的一大研究热点.在这一过程中,多智能体信用分配问题亟待解决.这是因为部分可观测环境会针对智能体产生的联合动作产生奖励强化信号,并将其用于强化学习网络参数的更新.也就是说,当所有智能体共享一个相同的全局奖励时,难以确定系统中的每一个智能体对整体所做出的贡献.除此之外,当某个智能体提前学习好策略并获得较高的回报时,其他智能体可能停止探索,使得整个系统陷入局部最优.针对这些问题,本文提出了一种简单有效的方法,即基于奖励滤波的信用分配算法.将其他智能体引起的非平稳环境影响建模为噪声,获取集中训练过程中的全局奖励信号,经过滤波后得到每个智能体的局部奖励,用于协调多智能体的行为,更好地实现奖励最大化.我们还提出了基于奖励滤波的多智能体深度强化学习(RF-MADRL)框架,并在 Open AI 提供的合作导航环境中成功地进行了验证.实验结果表明,和基线算法相比,使用基于奖励滤波的深度强化学习方法有着更好的表现,智能体系统策略收敛速度更快,获得的奖励更高.

关键词 多智能体系统;深度强化学习;信用分配;奖励滤波;合作导航

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02306

Reward-Filtering-Based Credit Assignment for Multi-Agent Deep Reinforcement Learning

XU Cheng[1),2)] YIN Nan[1)] DUAN Shi-Hong[1),2)] HE Hao[1)] WANG Ran[1),2)]

[1)](*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing* 100083)

[2)](*Shunde Graduate School, University of Science and Technology Beijing, Foshan, Guangdong* 528399)

Abstract In recent decades, reinforcement learning has achieved remarkable successes in many fields such as intelligent traffic control, competitive gaming, unmanned system positioning, and navigation. As more and more realistic scenarios require multi-agent to undertake complex tasks cooperatively, researchers pay more attention to studying multi-agent than single agents in reinforcement learning. At the same time, in multi-agent reinforcement learning (MARL), learning cooperation is a new research hotspot, which means agents need to learn to cooperate using only actions and local observations. However, the credit assignment problem needs to be solved when studying the cooperative process of a multi-agent system with DRL. In the process of learning to complete tasks, the partially observable environment provides reward reinforcement signals for the joint actions produced by agents, which are used to update the parameters of the deep reinforcement learning network. But the global reward is non-Markovian. When an agent takes

收稿日期:2021-09-27;在线发布日期:2022-08-26.本课题得到国家自然科学基金(62101029)、博士后创新人才支持计划(BX20190033)、广东省基础与应用基础研究基金联合基金(2019A1515110325)、中国博士后基金面上项目(2020M670135)、北京科技大学顺德研究生院博士后科研经费(2020BH001)、中央高校基本科研业务费(06500127)资助.**徐 诚**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究方向为群体智能、多智能体系统、物联网.E-mail:xucheng@ustb.edu.cn.**殷 楠**,硕士研究生,主要研究方向为群体智能、强化学习.**段世红**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为群体智能、物联网.**何 昊**,硕士研究生,主要研究方向为群体智能、无线定位.**王 然**,博士研究生,主要研究方向为群体智能、无线定位与分布式安全.

action at the current state, the actual reward signal for this action is usually given after several time steps. Especially in a difficult multi-agent environment, this phenomenon is more serious. In addition, all agents share the same global reward, making it hard to determine how much each agent in the system contributes to the whole. When an agent learns the strategy well in advance and gains a high return, the others may stop exploring, which leads the whole system to trap in the local optimum. To solve these problems, this paper introduces a credit assignment algorithm based on reward filtering that is not restricted by action space. The goal is to restore the local reward of each agent from the global reward obtained by all agents and apply it to the training of the action-value function network. The exploration behaviour of other agents often causes the non-stationarity of the environment, and the agent's own reward signal can be obtained by removing the influence of non-stationary from the global reward. Based on this, starting from the global reward, we model the influence caused by non-stationary factors as noise and propose a reward filter-based estimation mechanism to update the value function. In the process of centralized training, the influence of other agents on the environment is modelled as noise. The local reward of each agent is obtained by filtering the global reward, which is used to coordinate the behaviours of multi-agents and improve the system reward. We also propose a multi-agent deep reinforcement learning framework based on reward filtering (RF-MADRL) and successfully validate it in cooperative and competitive environments, namely the cooperative navigation with obstacles and predator-prey environments of Open AI. The experimental results show that compared with the baseline methods, including the traditional MADDPG method, value-function-based algorithms (i.e., VDN, QMIX, and QTRAN) and actor-critic-based credit assignment algorithms (i.e., COMA, FacMADDPG, and MAAC), our proposed RFMADRL has a better performance. The policy convergence is faster, and the reward obtained by the agent system is higher. The ablation experiment analysis shows that the reward filter module effectively improves the agent system's reward and solves the credit assignment problem.

Keywords　multi-agent system; deep reinforcement learning; credit assignment; reward filtering; cooperative navigation

1　引　言

在过去的几年中，多智能体强化学习(Multi-Agent Reinforcement Learning，MARL)的协作问题得到了广泛的研究[1-3]. 基于现实场景，多智能体通过感知和探索，学习在复杂环境下的最优协同策略，例如：多智能体进行城市或者荒野的搜索和救援[4-6]、自动驾驶汽车的协调[7]和工厂环境下的物流设计、配送、存储和运输[8]等.

现有的协作强化学习研究工作大致分为两类：集中式方法[9-12]和分散式方法[13-16]. 集中式方法通常将多智能体系统(Multi-Agent System，MAS)视为整体，由一个全局控制中心承担学习任务. 该中心可以将全部状态信息用于训练，学习最优的协作机制，控制所有智能体的执行. Tan[9]认为集中式强化学习方法通常是对现有的单智能体技术进行扩展，学习基于联合观测和共同奖励下的最优策略. 但是集中式方法缺少可扩展性，当环境和任务复杂度增加时，智能体联合行动空间随着智能体数量的增加而呈指数增长[13]. 此外，大规模的多智能体问题通常很难在所有智能体和中心节点之间建立可靠的通信网络[16]. 因此，集中式方法在链路状态未知的不确定环境下，难以保证可靠性.

与集中式方法相比，分散式强化学习系统中的每个智能体都是学习的主体，它们分别学习对环境的响应策略和相互之间的协作策略[17]. 由于环境的复杂和不确定性，智能体不具备观测全部状态的能力. 每个智能体使用局部观测和全局奖励(智能体采取联合动作与环境交互得到的奖励)独立地学习最优策略，并且需要以合作的方式最大限度地提高集体回报.

因此,在实际应用中,协作 MARL 遇到了可扩展性和部分可观测性两个主要挑战.部分可观测性可能导致智能体只考虑自己的状态而不关心其它智能体的状态,选择动作时也只考虑自身利益.智能体获得的强化信号只与自己的状态和动作相关联,使得学习时可能忽视其它智能体的存在.当某个智能体学习到一个能使系统获得更高奖励的策略时,其他智能体的探索可能会导致系统奖励减少从而停止探索,整个系统陷入局部最优,即在协作 MAS 中存在信用分配的问题.

对于一些简单问题,可以通过根据领域知识为每个智能体手动设计单独的奖励函数解决信用分配问题[18].然而,这种方法并不适用于复杂的多智能体协作任务.Foerster 等人[19]提出了一种反事实多智能体(Counterfactual Multi-Agent,COMA)策略梯度的多智能体演员-评论家方法,使用反事实基线来边缘化单个智能体的动作,同时保持其他智能体的动作不变,以计算智能体策略的优势.Sunehag 等人[20]提出了一个值函数分解网络(Value Decomposition Networks,VDN),将全局值函数分解为智能体的值函数.在 VDN 的基础上,动作值函数混合网络[21](QMIX)建立了全局 Q 值和局部 Q 值的非线性关系,使用网络去估计联合动作值函数.然而,VDN 和 QMIX 都限制了局部 Q 值和全局 Q 值之间的关系表示,智能体使用局部观测估计 Q 值的准确性并不高,并可能进一步妨碍其在复杂多智能体场景下学习协作策略.除了值分解方法,还有许多先进的信用分配算法[22-24],例如:奖励预测信用分配[22]提出使用神经网络来预测未来一定步数的奖励值,通过预测奖励值判断动作对获得奖励的影响,并重新分配奖励.隐式信用分配方法[23]使用中心化的评论家,使用了超网络携带足够的状态信息,对单个智能体做出评判,隐式地解决了完全合作环境下的信用分配问题.

解决信用分配问题时需要从多智能体系统获得的全局奖励出发,尽可能还原出每个智能体真实的局部奖励.在 Chang 等人[25]的研究中,提出了全局奖励信号由个体奖励和一个随机的马尔可夫过程组成.考虑到环境中其他智能体的探索是造成环境非平稳性的主要原因,不妨将全局奖励建模为个体奖励和其他智能体对奖励产生的影响之和,而其他智能体产生的影响可以视为噪声.环境将联合动作产生的全局奖励反馈给智能体,经过滤波将噪声影响降至最小,然后得到用于评论家网络训练的局部奖励.同时,每个智能体对整体做出的贡献也得到了量化.滤波方法不仅计算量小而且还能高效地解决环境中的噪声问题.

结合以上研究,本文提出一种简单有效的奖励分配机制用于动作值函数的更新,解决多智能体系统的信用分配问题.本文的主要贡献如下:

(1)提出了一种基于滤波的奖励估计方法.将不可观测环境状态引起的其他智能体对环境产生的非平稳影响建模为噪声,定义了全局奖励信号和智能体真实的局部奖励信号间的关系,并估计每个智能体的真实局部奖励.

(2)提出了基于奖励滤波信用分配的多智能体深度强化学习(RF-MADRL)框架.RF-MADRL 框架充分利用奖励滤波器滤除全局奖励噪声,并得到局部奖励估计,将其用于演员评论家网络的训练,使智能体更快地学习到最优策略并获得更高的奖励.

本文第 2 节整理多智能体深度强化学习的相关研究和典型的算法框架,并详细地介绍信用分配问题和卡尔曼滤波基础;第 3 节提出基于滤波的奖励估计方法和 RF-MADRL 框架;第 4 节详细地介绍实验环境的设置、智能体网络的训练和测试过程、与传统 MADDPG 方法和多种先进的信用分配算法的比较以及消融实验;第 5 节总结全文.

2 相关工作

2.1 去中心化的部分可观测马尔可夫决策过程

结合多智能体协同导航背景,本文对部分可观测环境下 MAS 完全合作型任务建模为去中心化的部分可观测马尔可夫决策过程(Decentralized Partially Observable Markov Decision Process,Dec-POMDP),通常由元组 G 表示:

$$G=\langle S,\boldsymbol{A},P,r,Z,O,N,\gamma\rangle \tag{1}$$

其中,$s\in S$ 表示环境的真实状态.部分可观测环境下,智能体 $i\in N\equiv\{1,2,\cdots,n\}$ 的观测值为 $o_i\in O_i$,观测函数 $Z(s,i):S\times N\rightarrow p(O)$ 定义了观测值的概率分布.在每个时刻,智能体 i 基于局部观测值 o_i 根据策略 $\pi_{\theta_i}:O_i\times A_i\rightarrow[0,1]$ 执行动作 $a_i\in A_i$.在所有智能体的联合动作 $\boldsymbol{a}\in\boldsymbol{A}$ 影响下,MAS 根据状态转移函数 $P:S\times A_1\times\cdots\times A_n\rightarrow S'$,到达下一状态 S' 并获得奖励 R.所有智能体共享相同的奖励函数 $r(s,\boldsymbol{a}):S\times\boldsymbol{A}\rightarrow R$,即全局奖励.目标是最大化系统奖励 $J=E_{a_1\sim\pi_1,\cdots,a_n\sim\pi_n,s\sim P}\sum_{t=0}^{T}\gamma^t r_t(s,\boldsymbol{a})$,$\gamma$ 为折扣因

子，T 为时间值. 如图 1 所示，智能体 i 基于自身观测值 o_i 做出决策 a_i，N 个智能体执行联合动作 $\boldsymbol{a}$ 与环境进行交互，获得奖励 R 并转移至下个状态 s'. 全局奖励 $R_t=\sum_{l=0}^{T=t}\gamma^l r_{t+l}$，智能体联合策略状态值函数记为 $V^{\pi}(s_t)=E_{s_{t+1}\sim P,\boldsymbol{a}_t\sim\boldsymbol{\pi}}[R_t\mid s_t]$，全局状态-动作值函数记为 $Q^{\pi}(s_t,\boldsymbol{a}_t)=E_{s_{t+1}\sim P,\boldsymbol{a}_{t+1}\sim\boldsymbol{\pi}}[R_t\mid s_t,\boldsymbol{a}_t]$ 用于策略 $\boldsymbol{\pi}$ 的评估.

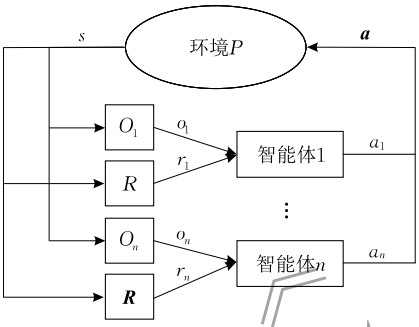


图 1　部分可观测环境下多智能体强化学习过程

2.2　多智能体深度确定性策略梯度算法

多智能体深度确定性策略梯度（Multi-Agent Deep Deterministic Policy Gradient，MADDPG）算法[26]由 DeepMind 团队提出并将单智能体深度强化学习算法扩展到多智能体领域，用于解决 MAS 的协作问题，避免由于智能体个数增加带来的“维数灾难”[27].

MADDPG 使用“中心化训练-去中心化执行”[28]（Centralized Training with Decentralized Execution，CTDE）的结构实现了智能体间的隐式协调. 假设 N 个智能体的策略 $\boldsymbol{\pi}=\{\pi_1,\cdots,\pi_n\}$ 由参数 $\theta=\{\theta_1,\cdots,\theta_n\}$ 确定，在 MADDPG 的训练过程中，中心化评论家将所有智能体的观测值 $\boldsymbol{o}=\{o_1,\cdots,o_n\}$ 和动作作为输入，使用深度神经网络拟合联合动作值函数 $Q_i^{\pi}(\boldsymbol{o},a_1,\cdots,a_n)$. 而在执行时，智能体只需根据自身观测信息便可与环境交互，如图 2 所示.

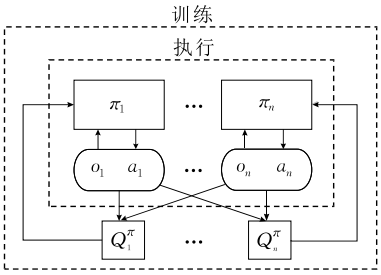


图 2　MADDPG 算法思想

在 MADDPG 中，每个智能体都需要维护一个演员-评论家[29]（Actor-Critic，AC）网络. 其中演员选择动作并逼近最优策略，评论家使用一个预先定义的值函数评估当前执行动作的性能，并基于时序差分误差来更新评论家和演员网络参数. 在 t 时刻，演员 i 根据当前的环境观测 o_i 和确定性策略 $\mu_i(a_i\mid o_i)$ 执行动作 a_i，获得奖励 r_i 和下一时刻观测值 o_i'. 评论家根据环境给出的奖励来调整自己的评分标准. 演员则根据评论家的打分情况调整自己的策略 μ_i，获得更高的奖励. 不断的学习使得评论家的评分更加标准，演员的策略更加有效.

2.3　信用分配

在完全合作型的多智能体协作过程中，所有智能体具有共同的任务，环境通常会将奖励强化信号反馈至整个智能体系统. 经典的 MADDPG 算法在训练过程中，使用结合环境设计的离散的全局奖励函数. 在智能体联合动作影响下得到的奖励或者惩罚信号，会被所有智能体共享. 如何将全局奖励分配给系统中的每个智能体，使其能够精准地体现个体对团队所做出的贡献，即多智能体信用分配问题[30].

信用分配的概念最先由 Sutton[31]提出，包括时间信用分配（Temporal Credit Assignment，TCA）和结构信用分配（Structural Credit Assignment，SCA）. 前者的目的是将延迟的奖励分配到每个动作，后者是将奖励分配到每个智能体. 任何分散式强化学习系统在结构信用分配解决后都可以很容易地划分成若干个标准的强化学习系统[32]. 因此，结构信用分配是需要解决的关键问题.

在多智能体协作过程中，环境的随机性和其他智能体的探索行为对整个系统产生了不稳定的影响. 多个智能体是同时学习的，当一个智能体的策略改变时，其他智能体的最优策略也会随之改变，使其不能正确地收敛到最优策略[33]. 因此，促进智能体协作的关键，是为每个智能体正确地分配奖励信号，减少非平稳环境对智能体学习的影响[34].

如何精确高效地解决信用分配问题是近来多智能体研究领域的热点. 差分回报[35]（Difference Rewards，DR）方法解决信用分配问题的核心是将智能体对系统的贡献值进行量化，但是一般很难找到普遍适用的量化标准，而且容易加剧智能体间信用分配的不平衡. COMA[19]算法借鉴 DR 的思想，使用联合的评论家神经网络计算每个智能体的优势函数，估计其对整体做出的贡献，从而解决智能体信用分配问题. 但是训练一个集中的评论家神经网络需要大量的计算资源，还存在监督不足的问题[25].

值分解技术使用满足一定限制条件的个体动作值函数来表示联合动作值函数，也可以解决 MARL 在协作过程中的信用分配问题. VDN[20] 基于智能体的局部观测值，将所有智能体值函数的总和作为系统的联合动作值函数. 但是对于智能体关系更为复杂的大型系统，该方法并不适用. 而且由于没有充分利用全局信息，直接将局部奖励相加并不准确. QMIX[21] 扩展了 VDN 的可加性值分解表达，将联合动作值函数表示为单个智能体动作值函数的单调函数. Qatten[36] 对联合动作值函数进行线性单调分解. QTRAN[37] 和 QPLEX[38] 进一步扩展了可表示的值函数. 这些值函数分解方法大部分在《星际争霸Ⅱ》中得到了验证，但是都局限于解决离散动作空间的问题.

为了应对连续动作空间下的信用分配问题，有学者提出了分解的演员-评论家方法，使用分解的评论家代替经典的集中评论家来计算策略梯度. FacMADDPG[39] 将值函数分解方法应用在 MADDPG 集中式评论家的训练过程中，使用单个智能体动作值函数的单调函数来表示联合动作值函数，让智能体学习连续的合作任务. 因此，与值分解技术在结构方面的限制相同，FacMADDPG 对于可分解的任务较难进行值分解[40]（如果联合动作值函数的最优行为与个体动作值函数的最优行为相同，则任务是可分解的，其中加法可分解性和单调性仅是可分解性的充分条件）. MAAC[41] 在解决信用分配问题方面，提出了一个相较 COMA 算法更一般的基线表示方法. DOP[42] 使用了一种类似于 Qatten 的分解结构来计算异策略树备份和同策略时间差分策略梯度. 然而这些分解的演员评论家方法不能保证收敛到全局最优.

综合上述分析，我们需要提供一种简单有效且不受动作空间限制的方法来解决智能体信用分配问题，目标是从智能体获得的全局奖励中还原出自身的局部奖励信号，并将其用于智能体动作值函数网络的训练. 环境的非平稳性往往是由其他智能体的探索行为导致，智能体自身奖励信号可以通过从全局奖励中去除非平稳性带来的影响来获得. 基于此，我们从全局奖励出发，将非平稳因素造成的影响建模为噪声，提出一种基于奖励滤波的估计机制用于值函数的更新.

2.4 卡尔曼滤波

卡尔曼滤波是一种最优状态估计方法，由实时获得的含噪声离散观测数据，对系统状态进行线性、无偏及最小误差方差的最优估计[43]. 对于一个具有确定性控制输入 u_k 的线性系统，其离散时间过程的系统状态模型为

$$\begin{cases} x_k = \boldsymbol{A}x_{k-1} + \boldsymbol{B}u_{k-1} + w_{k-1} \cdots w_{k-1} \sim N(0, Q) \\ z_k = \boldsymbol{H}x_k + v_k \cdots v_k \sim N(0, R) \end{cases} \tag{2}$$

其中，x_k 为 k 时刻系统状态变量，x_{k-1} 为 $k-1$ 时刻的系统状态变量，$\boldsymbol{A}$ 为状态转移矩阵，$\boldsymbol{B}$ 表示控制输入 u 的增益，w_{k-1} 表示过程噪声，服从均值为 0 协方差为 Q 的高斯分布. 在观测方程中，z_k 表示 k 时刻的观测值，$\boldsymbol{H}$ 表示状态变量 x_k 对观测变量 z_k 的增益，v_k 是测量噪声，服从均值为 0 协方差为 R 的高斯分布.

卡尔曼滤波提供了对过程状态 x_k 的估计，是一个递归的预测-校准方法，分为时间更新和测量更新两个阶段. 在时间更新过程（预测过程）中，根据 $k-1$ 时刻的状态后验估计值估计 k 时刻的状态值，在测量更新（校准过程）中，使用当前时刻的测量值来更正这一估计值，得到 k 时刻状态的后验估计.

卡尔曼滤波器时间更新方程如下：

$$\begin{aligned} \hat{x}_{\bar{k}} &= \boldsymbol{A}\hat{x}_{k-1} + \boldsymbol{B}u_{k-1}, \\ P_{\bar{k}} &= \boldsymbol{A}P_{k-1}\boldsymbol{A}^{\mathrm{T}} + Q \end{aligned} \tag{3}$$

卡尔曼滤波器测量更新方程如下：

$$\begin{aligned} K_k &= \frac{P_{\bar{k}}\boldsymbol{H}^{\mathrm{T}}}{\boldsymbol{H}P_{\bar{k}}\boldsymbol{H}^{\mathrm{T}} + R}, \\ \hat{x}_k &= \hat{x}_{\bar{k}} + K(z_k - \boldsymbol{H}\hat{x}_{\bar{k}}), \\ P_k &= (I - K_k\boldsymbol{H})P_{\bar{k}} \end{aligned} \tag{4}$$

在上述过程中，提到的参数介绍如下：$\hat{x}_{k-1}$、$\hat{x}_k$ 表示 $k-1$ 时刻和 k 时刻的后验状态估计值，$\hat{x}_{\bar{k}}$ 表示计算得到的先验状态估计值. 同样地，P_{k-1}、P_k 表示 $k-1$ 时刻和 k 时刻后验估计协方差，$P_{\bar{k}}$ 表示计算得到的先验估计协方差. K_k 为卡尔曼滤波增益.

3 基于奖励滤波信用分配的多智能体深度强化学习

本节将详细介绍基于滤波的奖励估计方法建模和应用，以及基于奖励滤波信用分配的多智能体深度强化学习框架.

3.1 系统模型

对于一个多智能体系统来说，环境中其他智能体的动作、环境变化等不可观测的状态变量，都会影响到全局奖励信号. 在本文中，我们受差异奖励[34] 的启发，令智能体 i 计算自己动作带来的奖励并代替共享的全局奖励作为强化信号，即

$$R^i(a^i|s,a^{-i})=R(s,a)-R(s,\langle a^{-i},c^i\rangle) \quad (5)$$

其中，a^{-i}表示除了智能体 i 的动作集合，c^i表示智能体的一个默认动作，$R(s,a^{-i})$表示当智能体的行为被默认行为取代时系统所获得的奖励.

结合滤波的思想，本文认为这些不可观测的状态变量对于全局奖励信号的影响可以作为环境噪声进而使用滤波方法进行处理. 因此，智能体与环境交互过程中共享的全局奖励信号可以建模为真实的局部奖励和不可观测环境状态(环境非平稳性)引起的奖励信号之和. 智能体 i 在 t 时刻状态 j 下共享的全局奖励为 g_t可以表示为

$$g_t=r_t^i(j)+b_t^i \quad (6)$$

其中，$r_t^i(j)$表示 t 时刻智能体 i 在状态 j 获得的真实奖励，即智能体 i 的局部奖励，b_t^i代表不可观测环境对全局奖励的影响，而且：

$$b_{t+1}=b_t+z_t,\cdots,z_t\sim N(\mu,\sigma_\omega^2) \quad (7)$$

z_t服从均值为 μ，方差为 σ_ω^2的高斯分布. 标准卡尔曼滤波算法是基于具有零均值高斯白噪声的系统模型. 对于具有一般性噪声的线性系统，为了采用标准卡尔曼滤波算法，智能体在强化学习过程中所获得的奖励线性系统模型可以表示为

$$\begin{cases}\boldsymbol{x}_t=\boldsymbol{A}x_{t-1}+\boldsymbol{B}\boldsymbol{u}+\varepsilon_t,\cdots,\varepsilon_t\sim N(0,\boldsymbol{\Sigma}_1)\\ g_t=\boldsymbol{C}\boldsymbol{x}_t+v_t,\cdots,v_t\sim N(0,\boldsymbol{\Sigma}_2)\end{cases} \quad (8)$$

其中，状态向量 $\boldsymbol{x}_t$为

$$\boldsymbol{x}_t=\begin{pmatrix}r_t(1)\\ r_t(2)\\ \vdots\\ r_t(|s|)\\ b_t\end{pmatrix}_{(|s|+1)\times 1},\quad \boldsymbol{\Sigma}_1=\begin{pmatrix}0&\cdots&0\\ \vdots&\ddots&\vdots\\ 0&\cdots&\sigma_\omega^2\end{pmatrix}_{(|s|+1)\times(|s|+1)},$$

$$\boldsymbol{B}=\begin{pmatrix}0&&&&\\ &0&&&\\ &&\ddots&&\\ &&&0&\\ &&&&1\end{pmatrix}_{(|s|+1)\times(|s|+1)},\quad \boldsymbol{u}=\begin{pmatrix}0\\ 0\\ \vdots\\ 0\\ \mu\end{pmatrix}_{(|s|+1)\times 1},$$

$|s|$表示环境状态总数. v_t表示观测误差，属于零均值高斯白噪声. 由于假设没有观测误差，$\boldsymbol{\Sigma}_2=0$. 状态转移矩阵 $\boldsymbol{A}=I$，ε_t表示服从零均值高斯分布的系统噪声且协方差矩阵为 $\boldsymbol{\Sigma}_1$. 观测矩阵 $\boldsymbol{C}=(0,\cdots,0,1_j,0,\cdots,0,1)_{1\times(|s|+1)}$，对于智能体 i，当状态 j 被观测到时，$\boldsymbol{C}$的第 j 个元素和最后一个元素值为 1，其余元素值为 0.

在使用卡尔曼滤波算法通过观察学习中的智能体接收到的全局奖励来动态估计其实际奖励和噪声的过程中，将使用在 t 时刻状态 j 中估计的局部奖励 $r_t^i(j)$而不是全局奖励 g_t^i来更新智能体的动作值函数.

3.2　基于卡尔曼滤波的奖励估计

卡尔曼滤波是一种利用线性系统状态方程，通过输入观测数据，对系统状态进行最优估计的算法. 基于第 3.1 节对智能体获取奖励的系统建模，使用标准卡尔曼滤波算法解决多智能体系统奖励估计的基本过程如下：

(1) 初始化系统状态 $\boldsymbol{x}_0=(0,\cdots,0)^{\mathrm{T}}$，协方差矩阵 $\boldsymbol{P}_0=I$，$\boldsymbol{u}=(0\ \cdots\ 0\ 0)^{\mathrm{T}}$，$\sigma_w^2=0.1$，$t=1$.

(2) 在智能体与环境交互时，智能体 i 在当前时刻 t 状态 j 下，使用强化学习算法得到动作 a_i并执行，到达新的状态 k 并获得全局奖励 g_t.

(3) 基于下述卡尔曼滤波器的时间更新方程进行状态预测，计算奖励估计值 $\hat{x}_t$和协方差矩阵 $\boldsymbol{P}_t$.

$$\hat{x}_t=\boldsymbol{A}\hat{x}_{t-1}+\boldsymbol{B}u_t,$$

$$\boldsymbol{P}_t=\boldsymbol{A}P_{t-1}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{\Sigma}_1.$$

(4) 使用观测值 g_t，基于下述卡尔曼滤波器的测量更新方程进行状态校正，计算卡尔曼增益 K_t，更新状态 $\hat{x}_t$和协方差矩阵 $\boldsymbol{P}_t$.

$$K_t=\frac{\boldsymbol{P}_t\boldsymbol{C}_t^{\mathrm{T}}}{\boldsymbol{C}_t\boldsymbol{P}_t\boldsymbol{C}_t^{\mathrm{T}}},$$

$$\hat{x}_t=\hat{x}_t+K_t(g_t-\boldsymbol{C}_t\hat{x}_t),$$

$$\boldsymbol{P}_t=(I-K_t\boldsymbol{C}_t)\boldsymbol{P}_t.$$

(5) 使用局部奖励 $r_t^i=x_t^i(j)$代替全局奖励 g_t更新智能体动作值函数.

(6) 估计噪声过程的均值 μ 和协方差 σ_w^2，更新 $\boldsymbol{u}$ 和 $\boldsymbol{\Sigma}_1$.

(7) $t\leftarrow t+1$，$j\leftarrow k$.

(8) 重复步骤(2)～(7).

在运行基于滤波的信用分配算法之前，需要对环境非平稳性引起的噪声的协方差 σ_ω^2进行初始估计. 然而，在实际的多智能体系统中，预先获取该协方差值是不实际的，往往需要在线估计. 本文给出一种基本的循环估计算法，具体过程如下：

(1) 初始化 $\mu_0=0$，$\sigma_{w_0}^2=0.1$，$t=0$.

(2) 使用连续的 μ_0 和 $\sigma_{w_0}^2$ 运行卡尔曼滤波，迭代 m 次($m\geq 200$)，记录 $x_{t+1}(|s|+1)$，$x_{t+2}(|s|+1)$，$\cdots$，$x_{t+m}(|s|+1)$.

(3) 估计噪声的均值和方差：

$$\mu_t=\frac{1}{m-1}\sum_{l=2}^{m}[x_{t+l}(|s|+1)-x_{t+l-1}(|s|+1)],$$

$$\sigma_{w_t}^2=\frac{1}{m-1}\sum_{l=2}^{m}[x_{t+l}(|s|+1)-x_{t+l-1}(|s|+1)-\mu_t]^2.$$

(4) 使用 μ_t和 $\sigma_{w_t}^2$ 运行基于卡尔曼滤波的奖励估

计算法，记录 $x_{t+m+1}(|s|+1)$的值.

(5) $t \leftarrow t+1$.

(6) 重复步骤(2)～(5).

卡尔曼滤波器在任何时候都不需要状态和观察的完整历史，只需在每次更新期间计算一些基本统计量. 因此，智能体在学习时，可以在线运行基于滤波的奖励估计算法，并且计算速度不会随着时间的推移而恶化. 其他智能体产生的强化信号(噪声)可以被建模为随机马尔可夫过程或由随机马尔可夫过程近似，由此保证算法的适用性. 当环境动力学模型与提供的卡尔曼滤波器的线性模型相匹配，使用滤波解决智能体信用分配问题行之有效，对于非线性模型，则可以利用扩展卡尔曼滤波方法应对.

3.3 基于奖励滤波信用分配的深度强化学习框架

基于上述分析，使用基于卡尔曼滤波的奖励估计方法可以对 MAS 获得的全局奖励进行重估计，并获得每个智能体的局部奖励. 与此同时，再分配的局部奖励信号可以体现每个智能体对任务完成所做出的贡献，从而解决多智能体协作过程中的信用分配问题. 因此，本研究将基于卡尔曼滤波的奖励估计算法在 MADDPG 结构中进行扩展，将卡尔曼滤波的奖励估计过程与智能体策略生成和评估过程进行结合，设计基于奖励滤波的演员评论家算法，提出基于奖励滤波信用分配的多智能体深度强化学习(RF-MADRL)框架，解决协作 MARL 的信用分配问题. 其结构如图 3 所示.

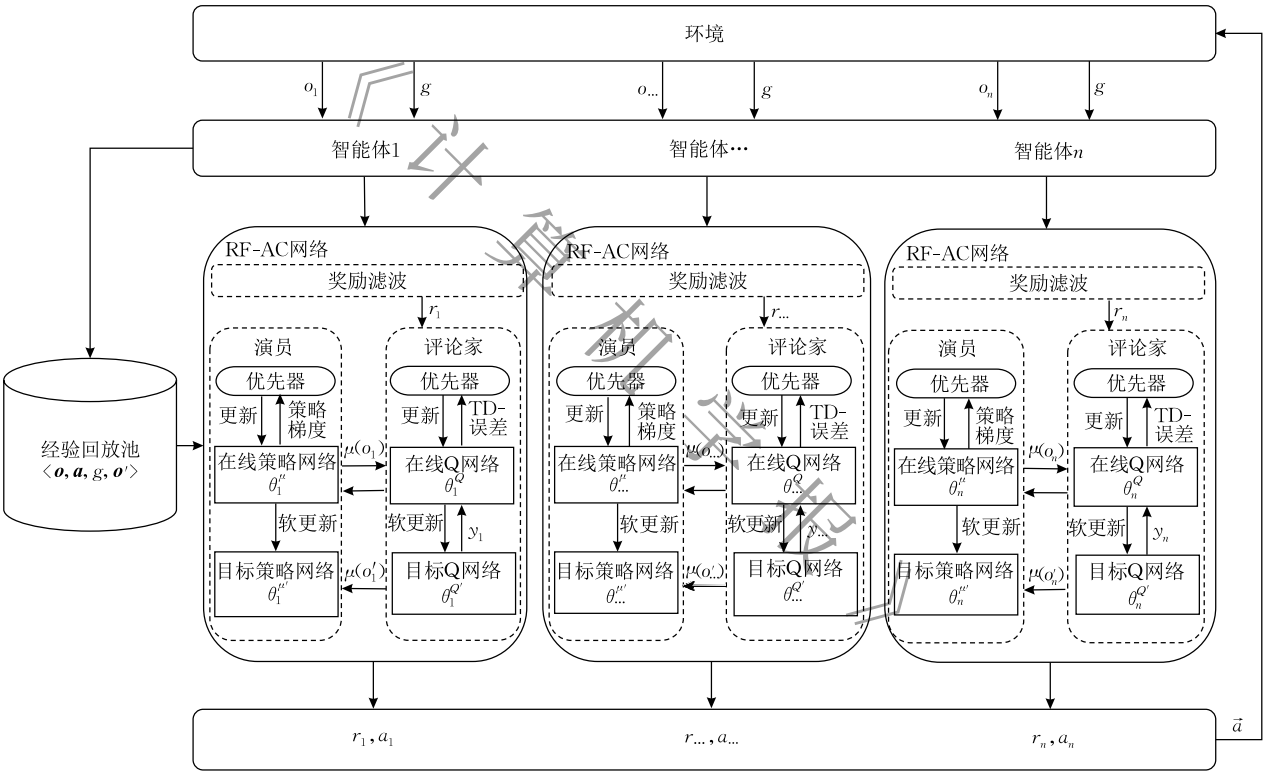


图 3 基于奖励滤波信用分配的多智能体深度强化学习框架

RF-MADRL 由环境、智能体、RF-AC 网络、经验回放池组成. 其中，RF-AC 网络又由三部分组成：演员网络、评论家网络以及奖励滤波器. RF-MADRL 遵循 CTDE 结构，在测试过程中，智能体 i 只需根据自己的观测 o_i 执行动作 $a_i=\mu(o_i)$. 在中心化训练过程中，智能体 i 拥有独立的 RF-AC 网络，可以将环境中所有智能体的观测值 $\boldsymbol{o}$、联合动作 $\boldsymbol{a}$、全局奖励 g 和下一步观测值 $\boldsymbol{o}'$ 作为网络的输入. 这些数据都在智能体探索环境时保存在经验回放池中，训练时随机抽取样本可以降低数据相关性. 奖励滤波器的作用是滤除全局奖励 g 中由环境平稳因素产生的噪声，从而得到智能体真实的局部奖励 r_i. 对于每一个智能体来说，将滤波后的奖励用于训练，不仅可以量化每个智能体对整体的贡献程度，也可以改善智能体“懒惰”(lazy agent)现象.

另外，RF-AC 网络又包含策略网络和 Q 网络. 对于确定性策略 μ 来说，策略网络用于生成智能体策略 μ，参数由 θ^{μ} 表示，Q 网络用于生成智能体的动作值函数，参数由 θ^{Q} 表示，而 μ 可以由 $\arg\max_a Q(o,a)$ 得到. 为了提高训练过程的稳定性，借鉴 DQN 的思想，为策略网络和 Q 网络创建两个神经网络拷贝，分为在线网络和目标网络，其中目标策略网络参数

为 $\theta^{\mu'}$，目标 Q 网络参数为 $\theta^{Q'}$，目标网络的参数更新使用软更新(soft update)方式.

在 RF-AC 网络中，在线 Q 网络损失函数为

$$L(\theta_i^Q)=\frac{1}{N}\sum_i(y_i-Q_i^\mu(\boldsymbol{o},a_1,\cdots,a_n))^2 \tag{9}$$

$$y_i=r_i+\gamma Q_i^{\mu'}(\boldsymbol{o}',a_1',\cdots,a_n')|_{a_j'=\mu_j'(o_j)} \tag{10}$$

其中，r_i 表示使用基于卡尔曼滤波的奖励估计算法从全局奖励中动态估计获得的局部奖励，y_i 由智能体局部奖励 r_i、目标策略网络估计值 μ' 和目标 Q 网络估计值 Q_i' 得到，使用优化器通过神经网络梯度反向传播可更新在线 Q 网络参数，而且 $\mu'=\{\mu_{\theta_1^{Q'}},\cdots,\mu_{\theta_n^{Q'}}\}$.

策略网络的策略梯度记为

$$\nabla_{\theta_i^\mu}J(\mu_i)=E_{\boldsymbol{o},\boldsymbol{a}\sim D}[\nabla_{\theta_i^\mu}\mu_i(a_i|o_i)\nabla_{a_i}Q_i^\mu(\boldsymbol{o},a_1,\cdots,a_n)|a_i=\mu_i(o_i)] \tag{11}$$

其中，D 表示经验回放池，存储的数据为 $\langle\boldsymbol{o},\boldsymbol{a},g,\boldsymbol{o}'\rangle$，$\boldsymbol{o}=\{o_1,\cdots,o_n\}$ 包含了所有智能体的观测值，训练时随机抽取一定数量的样本进行网络更新. 目标策略网络和目标 Q 网络更新方式为

$$\begin{cases}\theta^{Q'}\leftarrow\tau\theta^Q+(1-\tau)\theta^{Q'}\\ \theta^{\mu'}\leftarrow\tau\theta^\mu+(1-\tau)\theta^{\mu'}\end{cases} \tag{12}$$

其中，τ 为软更新参数.

3.4　RF-MADRL 算法实现

本文提出的多智能体深度强化学习框架中，每个智能体与环境交互和训练过程，如算法 1 所示. 具体执行时，首先需要初始化强化学习环境和智能体网络参数. 在当前的观测 o_i 下，智能体基于强化学习算法选择动作 a_i 并执行，获得新得观测值 o_i'，获得全局奖励 g. 将探索得到的 $(\boldsymbol{o},\boldsymbol{a},g,\boldsymbol{o}')$ 存入经验回放池. 其次，训练 RF-AC 网络. 执行基于奖励滤波的信用分配算法对状态进行预测，更新估计值 $\hat{x}_t$ 和协方差矩阵 $\boldsymbol{P}_t$；使用得到的全局奖励更新状态后验估计值，校正模型；重新估计噪声过程的均值 μ 和方差 σ_w^2，得到滤除噪声后的局部奖励 r_i. 最后，使用局部奖励 r_i 更新对应智能体的 RF-AC 网络参数，迭代至学会该场景下的最佳策略.

算法 1.　N 个智能体的 RF-MADRL 算法.

输入：RF-AC 网络及其参数、折扣因子 γ、软更新参数 τ、批量梯度下降的样本数 m、最大迭代次数 T、最大回合数 M

输出：最优在线策略网络参数 θ^μ，在线 Q 网络参数 θ^Q

FOR 回合数←1,…,M DO
　初始化 RF-AC 网络参数以及环境参数
　产生初始的观测向量 $\boldsymbol{o}$
　FOR t←1,2,…,T DO
　　FOR i←1,2,…,N DO
　　智能体 i 选择动作 $a_t^i=\mu_t^i(o_t^i)$ 执行
　　END FOR
　　执行联合动作 $\boldsymbol{a}_t=(a_t^1,\cdots,a_t^N)$ 获得奖励 g_t 以及新的观测值 $\boldsymbol{o}_t'$
　　将 $(\boldsymbol{o}_t,\boldsymbol{a}_t,g_t,\boldsymbol{o}_t')$ 存入经验回放池 D
　　$\boldsymbol{o}\leftarrow\boldsymbol{o}_t'$
　　从 D 中随机采样 m 个样本 $(\boldsymbol{o}^j,\boldsymbol{a}^j,g^j,\boldsymbol{o}'^j)$
　　FOR i←1,2,…,N DO
　　使用 3.2 节基于卡尔曼滤波的奖励估计算法估计每个智能体的局部奖励 r_i^j 及噪声估计过程的均值 μ_t 和方差 $\sigma_{w_t}^2$
　　计算 y_i^j
　　$y_i^j=r_i^j+\gamma Q_i^{\mu'}(\boldsymbol{o}'^j,a_1',\cdots,a_N')|_{a_k'=\mu_k'(o_k^j)}$
　　计算损失函数 $L(\theta_i^Q)$ 更新参数 θ_i^Q
　　$L(\theta_i^Q)=\frac{1}{m}\sum_j(y_i^j-Q_i^\mu(\boldsymbol{o}^j,a_1^j,\cdots,a_N^j))^2$
　　计算策略梯度 $\nabla_{\theta_i^\mu}J(\mu_i)$ 更新参数 θ_i^μ
　　$\nabla_{\theta_i^\mu}J(\mu_i)\approx\frac{1}{m}\sum_j\nabla_{\theta_i^\mu}\mu_i(o_i^j)\nabla_{a_i}Q_i^\mu(\boldsymbol{o}^j,a_1^j,\cdots,a_i,\cdots,a_N^j)|_{a_i=\mu_i(o_i^j)}$
　　更新目标网络参数
　　$\theta_i'\leftarrow\tau\theta_i+(1-\tau)\theta_i'$
　　END FOR
　END FOR
END FOR

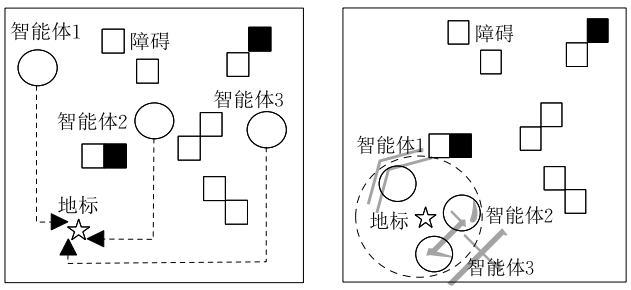
4　实验结果与分析

本节主要以具体实验论证 RF-MADRL 方法的有效性和优越性. 首先，介绍用于评估算法的实验环境以及参数设置. 然后，分别对 RF-MADRL 方法的训练和测试结果进行分析，并与多种解决信用分配问题的先进算法进行比较验证. 最后，进行消融实验分析，验证 RF-MADRL 方法的有效性.

4.1　环境设置

在 OpenAI 提供的多智能体粒子环境[26](Multi-agent Particle Environment，MPE)的基础上，本文构建了含有障碍物的合作导航(Cooperative Navigation with Obstacles，CNO)环境，对基于奖励滤波的多智能体信用分配问题进行建模和研究，目标是降低环境非平稳因素对智能体协作导航过程的影响. 此外，本文在 Leibo 等人[44]提供的围捕环境基础上，构建了捕食者-猎物(Predator-Prey，PP)环境，以验证本文所述方法在不同应用场景下的有效性.

CNO 实验环境如图 4(a)所示,实验设置的智能体(圆形)个数为 3,障碍物(方形)个数为 10,包括 8 个固定障碍(浅色方形)和 2 个移动障碍(黑色方形).实验开始时,环境中的各组成元素随机初始化坐标,智能体必须通过合作移动到达地标(五角星)并躲避障碍.智能体需要观察其他智能体和地标的相对位置学习协作,并根据自己与地标的接近程度获得奖励.因为物理空间的有限性,智能体相互碰撞或者与障碍物碰撞时会受到惩罚,所以智能体需要学会在避开其他智能体和障碍物的同时快速导航到地标范围内(虚线圆圈),如图 4(b)所示.



(a) 初始环境 (b) 成功导航到地标

图 4 含有 3 个智能体的 CNO 环境

智能体 i 在 CNO 中可以获得自身的物理位置 $AgentPos_i$ 和运动速度 $velocity_i$,与目标位置的距离 $LandmarkDist_i$,其他智能体和障碍物的位置信息,分别记为 $OthersPos_i$ 和 $ObstaclesPos_i$,这些信息将组合成一个多维数组作为对应智能体的观测值被使用.智能体 i 在 t 时刻的观测值 o_i 表示为

$$o_i=\{AgentPos_i,velocity_i,LandmarkDist_i,OthersPos_i,ObstaclesPos_i\} \tag{13}$$

其中,连续动作空间下的智能体 i 在 t 时刻的位置取决于上一时刻的位置和速度.由于现实环境的限制,智能体在 CNO 中采取的某些动作可能无效,例如移动到环境范围之外或占据障碍物位置.所以在训练过程中,为了使动作只从有效的动作中取样,需要对选择无效移动的智能体给予负面奖励.依据当前的场景和目标,实验设计了智能体的奖励函数.在 CNO 环境中,所有智能体执行联合动作 $\boldsymbol{a}$ 得到共享的全局奖励 g_{CNO} 的组成如下:

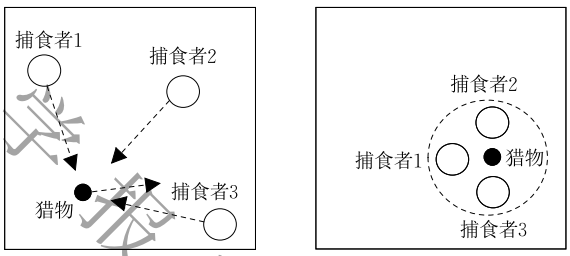
$$g_{\mathrm{CNO}}=\alpha_1\times e^{-\sum_i^n d_{\mathrm{CNO}}^i}+\alpha_2\times\sum_i^n\cos\theta^i-\alpha_3\times\sum_i^n c^i \tag{14}$$

其中,n 为智能体个数,d_{CNO}^i 表示智能体 i 与地标之间的距离,距离越近,奖励越大;θ^i 表示智能体 i 运动方向与地标方向的夹角,范围在 0°到 180°之间,θ 小于 90°时,认为智能体 i 向着地标移动,当 θ 等于 0°时,奖励最大;c^i 表示智能体 i 之间、智能体和障碍物之间的碰撞次数,当智能体与其他物体发生碰撞时,给予负奖励;α 表示不同类别奖励的影响权重.在具体实验中,$\alpha_1=10$,$\alpha_2=10$,$\alpha_3=1$,$n=3$.

PP 实验环境如图 5 所示,3 个捕食者(浅色圆形)需要在随机生成的环境中合作追逐一个移动速度更快的猎物(黑色实心圆).当所有的捕食者距离猎物小于一定的范围(虚线圆圈)时,认为捕食者成功捕获猎物.在 PP 环境下,捕食者的观测值包括猎物的位置、速度以及其他合作捕食者的位置.共享的全局奖励依据捕食者与猎物的距离设定.捕食者与猎物距离越近,所获得的奖励越大.同时,捕食者之间发生碰撞则给予一定惩罚.在 PP 环境中,全局奖励 g_{PP} 的组成如下:

$$g_{\mathrm{PP}}=\beta_1\times e^{-\sum_i^n d_{\mathrm{PP}}^i}-\beta_2\times a \tag{15}$$

其中,n 为捕食者个数,d_{PP}^i 表示捕食者 i 与猎物之间的距离,距离越近,奖励越大;a 表示捕食者之间的碰撞次数,当捕食者发生碰撞时,给予负奖励;β 表示不同类别奖励的影响权重.在具体实验中,$\beta_1=1$,$\beta_2=0.6$,$n=3$.



(a) 初始环境 (b) 成功捕获猎物

图 5 完全合作的捕食者-猎物环境

4.2 参数设置

在 RF-MADRL 框架中,所有智能体的 RF-AC 网络基于深度确定性策略梯度[29](Deep Deterministic Policy Gradient,DDPG)网络设定并增加奖励滤波器结构.在实验中,RF-AC 网络都为使用 Relu 作为激活函数的多层感知机(Multi-Layer Perception,MLP),除了输入输出层,各自还有 2 个隐层,每层神经元个数为 64,如图 6 所示.演员网络的输入是由式(14)定义的智能体观测量,输出则是智能体的动作(通过智能体速度向量体现).评论家网络的输入由观测量和演员网络得到的智能体动作组成,输出为对该动作评价的 Q 值.迭代终止的设定条件为智能体所走的步数超过最大回合数,或者智能体导航到地标范围(猎物在捕获范围)则停止当前回合交互.RF-MADRL 的超参数如表 1 所示.

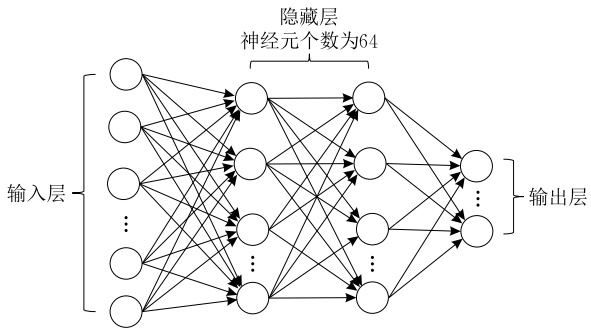


图 6　RF-AC 网络的 MLP 结构

表 1　RF-MADRL 参数设置

设置	名称	值
训练过程参数	经验回放池容量	10^6 个回合
	批量大小	1000 个回合
	训练回合数	15000 个回合
	探索回合数	1000 个回合
	回合步数	25 步
	折扣因子	0.95
	模型保存率	1000 个回合
	目标更新间隔	200 个回合
网络参数	智能体学习率	0.01
	优化器	Adam
	神经元个数	64

4.3　网络训练和测试

实验分为训练和测试两个步骤，在模型训练至收敛后，测试智能体是否能成功导航到地标. 智能体初始训练迭代回合数设置为 15000 次，每次迭代的步长设置为 25 步. 初始化参数后，智能体随机探索环境，并将观测值、奖励等数据存入经验回放池. 当数据累积了 1000 个交互回合后，智能体的 RF-AC 网络开始训练. 每次训练需随机从经验回放池中取样 1000 次智能体与环境交互所产生的数据，当达到最大迭代次数时，结束训练并保存模型，训练过程中智能体动作值网络损失变化过程如图 7 所示.

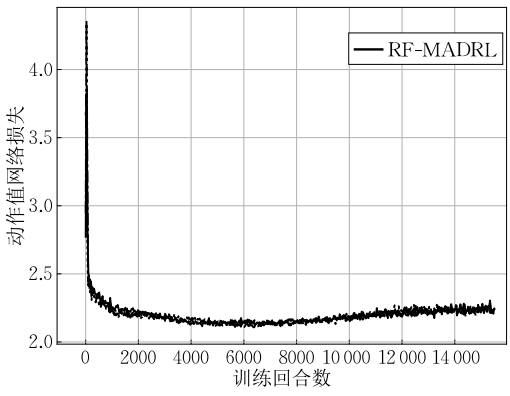


图 7　训练过程中的动作值网络损失变化(CNO)

智能体的动作值网络损失有明显下降趋势，并在 3000 个回合之后基本保持稳定，说明智能体在训练过程中，网络收敛. 然而，随着时间步的增加，在回合数大于 3000 时，动作值网络损失值出现缓慢上升趋势. 出现该现象可能的原因是实验中设置了能保存 1000 条数据的经验回放池，当经验回放池的数据不断增加和动态变化，样本与之前的训练集有一定的差异，导致损失值不降反升. 在强化学习中，较低的损失意味着对当前策略的价值预测更加准确，可以在一定程度上说明网络模型的学习程度，但是不能准确反映策略的好坏. 智能体所获得的奖励能够直接反应策略的好坏，所以仍需通过智能体在训练过程中的奖励变化来评价学习到的策略，如图 8 所示. 智能体系统所获得的回合奖励趋于稳定，使用 RF-MADRL 方法学习到了当前环境下的最优策略.

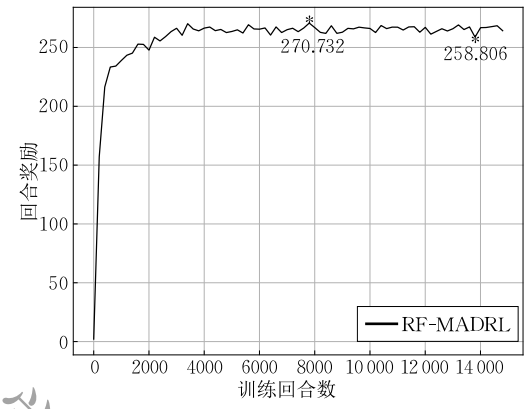


图 8　训练过程中的回合奖励变化(CNO)

为了进一步对 RF-MADRL 算法的可行性进行评估，实验还记录了训练过程中的一些数据指标，如智能体与环境交互的最后一个时间步的智能体与目标位置的距离、智能体与障碍物的平均碰撞次数. 经过 3000 次的迭代测试，智能体在场景中获得的奖励均值为 265.126，三个智能体与目标的距离和均值为 0.600，与障碍物的碰撞次数均值为 2.836. 实验结果如图 9 和图 10 所示，其中，图 9 展示了训练稳态时智能体与地标距离的局部放大图.

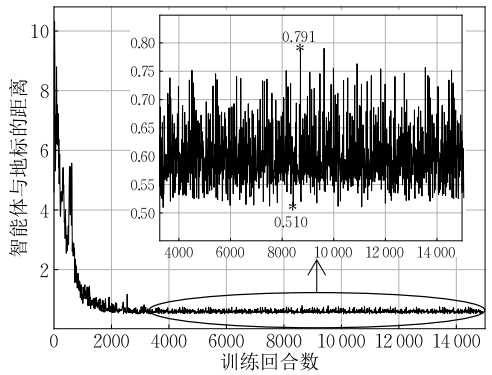


图 9　训练过程中智能体与地标的距离变化(CNO)

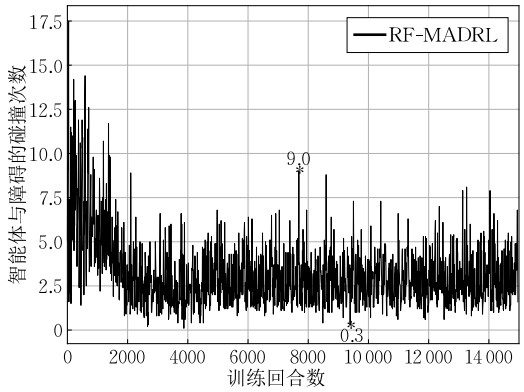


图 10 训练过程中智能体与障碍碰撞次数的变化(CNO)

综合上述分析,在初始化环境、网络等各项参数后,智能体开始探索,产生随机策略,并逐渐积累训练所需数据.随着网络参数的不断更新和策略的学习,智能体所获得的奖励逐渐增加,与目标位置的距离逐渐减小,与障碍物的碰撞次数也越来越少.当训练的迭代次数大于 3000 时,智能体的各项指标趋于稳定.此时,智能体能快速导航到目标范围,说明 RF-MADRL 在合作导航的场景中是有效的.

在捕食者-猎物环境下,捕食者在 15000 次的训练过程中,回合奖励变化如图 11 所示.由图可知,训练回合数约 3000 次时,RD-MADRL 算法收敛.经过 3000 次的迭代测试,捕食者获得的平均回合奖励为 36.989,此时所有捕食者可以成功捕获猎物,所以 RF-MADRL 算法在完全合作的捕食者-猎物的场景中也是有效的.

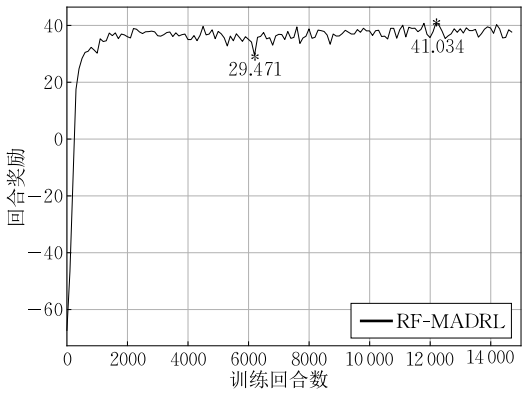


图 11 训练过程中的回合奖励变化(PP)

4.4 与基线算法的比较

为了验证 RF-MADRL 算法的优越性,本文在相同的CNO环境、PP环境和相同的参数设置下,与 7 种基线算法进行了详细比较,包括传统的 MADDPG 算法、基于值函数的先进信用分配算法,如 VDN、QMIX 和 QTRAN,以及基于演员-评论家的先进信用分配算法,如COMA、FacMADDPG和MAAC方法.

本文首先对比了 RF-MADRL 与 7 种基线算法在训练过程中所获得的奖励情况,如图 12 所示.在 CNO 环境中,RF-MADRL 达到最大平均回合奖励的训练回合数约为 3000,训练速度远远快于其他基线算法,智能体获得的最大平均回合奖励值为 265.126.训练至收敛后,本文进一步比较了 RF-MADRL 和基线在智能体获得的平均回合奖励、智能体与目标的距离、智能体平均回合路径长度、智能体与障碍物的平均回合碰撞次数,统计结果如表 2 所示.RF-MADRL 获得的平均回合奖励较 MAAC 提高了 20.1%,与地标的距离减少了 17.0%.RF-MADRL 的平均回合碰撞次数为 2.836,相对于 QTRAN 减少了 65.8%,有效地减少了智能体与障碍物的碰撞,但仍不可避免.若想要进一步减少碰撞次数,可增加碰撞惩罚在全局奖励设置中所占比例.RF-MADRL 的回合平均路径长度相比于 MAAC,减少了 7.3%,说明使用 RF-MADRL 方法的智能体能够更快地导航到地标.

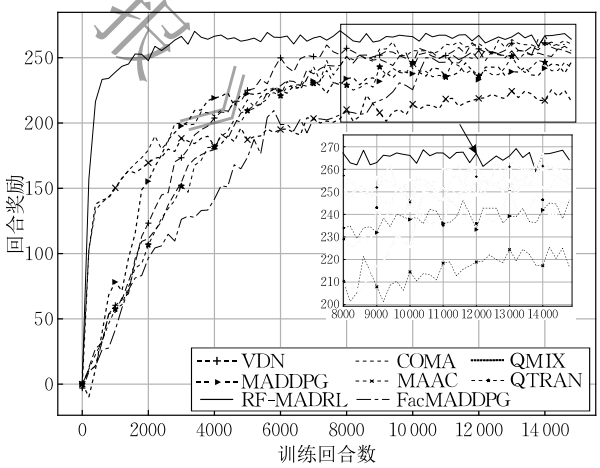


图 12 RF-MADRL 与多种基线算法的奖励变化(CNO)

表 2 RF-MADRL 与不同算法在 CNO 环境下各项指标数据对比

参数指标	RF-MADRL	MADDPG	COMA	MAAC	VDN	QMIX	QTRAN	FacMADDPG
平均回合奖励	**265.126**	240.534	260.276	220.748	256.632	253.228	242.126	255.455
与地标的距离	**0.600**	0.708	0.655	0.723	0.695	0.695	0.706	0.673
平均回合碰撞次数	**2.836**	4.908	6.136	5.050	6.866	3.533	8.300	5.838
平均路径长度	**28.942**	30.883	29.398	31.242	29.465	30.173	30.532	29.535

捕食者-猎物环境下的实验结果如由图 13 所示. RF-MADRL 相较基线算法收敛速度更快. 在算法收敛后，RF-MADRL 获得的平均回合奖励为 36.989，相较于基线算法中表现较差的 QMIX 提高了 2.11 倍.

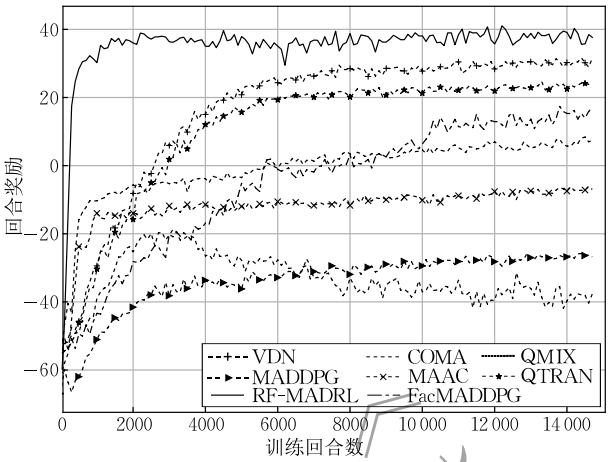


图 13　RF-MADRL 与多种基线算法的奖励变化(PP)

综合以上分析可知，RF-MADRL 整体优于其他基线方法，策略收敛速度更快，获得的平均回合奖励更高，对于躲避障碍物的稳定性较好，且能更快地导航到地标以及捕获猎物.

4.5　消融实验

为了验证本文所提出的奖励滤波模块对于解决信用分配问题的有效性，本文从以下两个方面对 RF-MADRL 方法进行消融实验分析.

首先，奖励滤波模块作为 MADDPG 方法和 RF-MADRL 方法的差异项，MADDPG 提供了 RF-MADRL 的消融基线(即有或无奖励滤波模块的情况)，可以用于验证奖励滤波模块的有效性. 在 CNO 环境中，如图 12 所示，在训练回合数约 10000 次时，MADDPG 算法收敛，而 RF-MADRL 算法约在 3000 次训练后收敛，收敛速度提高了 70%. 结合表 2 数据，算法稳定后，MADDPG 方法所获得的平均回合奖励为 240.534，RF-MADRL 方法所获得的平均回合奖励为 265.126，增长了 10.2%. 除此之外，与 MADDPG 方法相比，使用 RF-MADRL 方法的智能体与地标的平均距离减少了 15.3%，与障碍物的平均碰撞次数减少了 42.2%，走过的路径长度减少了 6.3%. 在 PP 环境下，如图 13 所示，RF-MADRL 算法收敛的速度相比 MADDPG 方法提高了 70%，平均回合奖励增加了 67.237.

其次，为了进一步验证在基于多智能体演员-评论家框架中使用奖励滤波模块的有效性，本文将 3 种基于演员-评论家方法的信用分配算法作为消融基线，与 RF-MADRL 方法进行对比与分析. RF-MADRL、COMA、FacMADDPG 和 MAAC 方法都基于演员-评论家结构实现，但是解决信用分配问题时有所不同.

COMA 使用全局的评论家网络计算每个智能体的优势函数，FacMADDPG 将 QMIX 的值函数分解思想引入全局评论家的计算，MAAC 采取注意力机制，学习使智能体获取更大奖励的策略，而 RF-MADRL 使用卡尔曼滤波得到智能体各自的奖励，简单高效且计算量小，对比结果如图 12、图 13 和表 2 所示. 在 CNO 环境下，RF-MADRL 获得的平均回合奖励较于 MAAC 提高了 20.1%，与地标的距离减少了 17.0%，收敛速度也远远快于 COMA 和 FacMADDPG 方法. 在 PP 环境下，RF-MADRL 获得的平均回合奖励较于 MAAC、COMA 和 FacMADDPG 分别增加了 47.025、34.661 及 34.425.

结合上述消融实验分析，可以认为奖励滤波模块对于提高智能体系统奖励和解决信用分配问题是有效的，且相较于其他基于演员-评论家结构的信用分配算法，RF-MADRL 方法在收敛速度、获得的奖励等方面具有一定的优越性.

5　总　结

本文介绍了基于滤波的奖励估计算法，将智能体与环境交互过程中共享的全局奖励信号，建模为真实的局部奖励和不可观测环境状态(环境非平稳性)引起的奖励信号之和，全局奖励在滤除噪声影响后可以得到真实的局部奖励. 随后，本文将该算法应用于基于 CTDE 结构的多智能体深度强化学习框架，并在智能体合作导航场景中进行验证. 本文使用卡尔曼滤波实现了对智能体局部奖励的在线估计，有效促进了多智能系统中心化评论家的训练，解决了非平稳环境中多智能体协作过程中的信用分配问题，促使智能体学习当前环境下的最优策略. 在相同的场景下，与传统的 MADDPG 方法和多种信用分配算法相比，使用奖励滤波方法训练得到的智能体系统获得的平均回合奖励增加，障碍碰撞次数有效减少.

参　考　文　献

[1]　Busoniu L, Babuska R, De Schutter B. A comprehensive

survey of multiagent reinforcement learning. IEEE Transactions on Systems, 2008, 38(2): 156-172

[2] Gupta J K, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Sao Paulo, Brazil, 2017: 66-83

[3] Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. IEEE Transactions on Cybernetics, 2020, 50(9): 3826-3839

[4] Luis C E, Schoellig A P. Trajectory generation for multiagent point-to-point transitions via distributed model predictive control. IEEE Robotics and Automation Letters, 2019, 4(2): 375-382

[5] Niroui F, Zhang K, Kashino Z, et al. Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. IEEE Robotics and Automation Letters, 2019, 4(2): 610-617

[6] Peake A, McCalmon J, Zhang Y, et al. Wilderness search and rescue missions using deep reinforcement learning// Proceedings of the 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). Abu Dhabi, UAE, 2020: 102-107

[7] Cao Y, Yu W, Ren W, et al. An overview of recent progress in the study of distributed multi-agent coordination. IEEE Transactions on Industrial Informatics, 2012, 9(1): 427-438

[8] Ying W, Dayong S. Multi-agent framework for third party logistics in E-commerce. Expert Systems with Applications, 2005, 29(2): 431-436

[9] Tan M. Multi-agent reinforcement learning: Independent vs. cooperative agents//Proceedings of the 10th International Conference on Machine Learning. Amherst, USA, 1993: 330-337

[10] Yu J, LaValle S M. Optimal multirobot path planning on graphs: Complete algorithms and effective heuristics. IEEE Transactions on Robotics, 2016, 32(5): 1163-1177

[11] Augugliaro F, Schoellig A P, D'Andrea R. Generation of collision-free trajectories for a quadrocopter fleet: A sequential convex programming approach//Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Algarve, Portugal, 2012: 1917-1922

[12] Preiss J A, Hönig W, Ayanian N, et al. Downwash-aware trajectory planning for large quadrotor teams//Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, Canada, 2017: 250-257

[13] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533

[14] Van Den Berg J, Snape J, Guy S J, et al. Reciprocal collision avoidance with acceleration-velocity obstacles//Proceedings of the 2011 IEEE International Conference on Robotics and Automation. Shanghai, China, 2011: 3475-3482

[15] Busoniu L, De Schutter B, Babuska R. Decentralized reinforcement learning control of a robotic manipulator//Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision. Grand Hyatt, Singapore, 2006: 1-6

[16] Nguyen D T, Kumar A, Lau H C. Credit assignment for collective multiagent RL with global rewards//Proceedings of the in Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 8102-8113

[17] Yang Y, Hao J, Chen G, et al. Q-value path decomposition for deep multiagent reinforcement learning//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2020: 10706-10715

[18] Zhang K, Yang Z, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. Handbook of Reinforcement Learning and Control, 2021: 321-384

[19] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 32(1)

[20] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning//Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems. Stockholm, Sweden, 2018: 2085-2087

[21] Rashid T, Samvelyan M, Schroeder C, et al. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 4295-4304

[22] Seo M, Vecchietti L F, Lee S, et al. Rewards prediction-based credit assignment for reinforcement learning with sparse binary rewards. IEEE Access, 2019, 7: 118776-118791

[23] Zhou M, Liu Z, Sui P, et al. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2020, 33: 11853-11864

[24] Alipov V, Simmons-Edler R, Putintsev N, et al. Towards Practical Credit Assignment for Deep Reinforcement Learning. arXiv preprint arXiv:2106.04499, 2021

[25] Chang Y H, Ho T, Kaelbling L. All learning is local: Multi-agent learning in global reward games//Proceedings of the 17th Advances in Neural Information Processing Systems. Vancouver, Canada, 2003: 807-814

[26] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments//Proceedings of the 31st Advances in Neural Information Processing Systems. California, USA, 2017: 6382-6393

[27] Han J, Jentzen A, E W. Solving high-dimensional partial differential equations using deep learning. Proceedings of the National Academy of Sciences, 2018, 115(34): 8505-8510

[28] Oliehoek F A, Spaan M T J, Vlassis N. Optimal and approximate Q-value functions for decentralized POMDPs. Journal of Artificial Intelligence Research, 2008, 32: 289-353

[29] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv: 1509.02971, 2015

[30] Lansdell B J, Prakash P R, Kording K P. Learning to solve the credit assignment problem. arXiv preprint arXiv: 1906.00889, 2019

[31] Sutton R S. Temporal Credit Assignment in Reinforcement Learning [Ph. D. dissertation]. University of Massachusetts Amherst, USA, 1984

[32] Yu Z, Guochang G, Rubo Z. A new approach for structural credit assignment in distributed reinforcement learning systems //Proceedings of the 2003 IEEE International Conference on Robotics and Automation. Taiwan, China, 2003, 1: 1215-1220

[33] Zhong Yu, Gu Guo-Chang, Zhang Ru-Bo. Survey of distributed reinforcement learning algorithms in multi-agent systems. Control Theory & Applications, 2003, 20(3): 317-322(in Chinese)
(仲宇，顾国昌，张汝波. 多智能体系统中的分布式强化学习研究现状. 控制理论与应用，2003，20(3)：317-322)

[34] Marinescu A, Dusparic I, Clarke S. Prediction-based multi-agent reinforcement learning in inherently non-stationary environments. ACM Transactions on Autonomous and Adaptive Systems, 2017, 12(2): 1-23

[35] Castellini J, Devlin S, Oliehoek F A, et al. Difference Rewards Policy Gradients. arXiv preprint arXiv: 2012.11258, 2020

[36] Yang Y, Hao J, Liao B, et al. Qatten: A general framework for cooperative multiagent reinforcement learning. arXiv preprint arXiv: 2002.03939, 2020

[37] Son K, Kim D, Kang W J, et al. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. California, USA, 2019: 5887-5896

[38] Wang J, Ren Z, Liu T, et al. QPLEX: Duplex dueling multi-agent q-learning. arXiv preprint arXiv: 2008.01062, 2020

[39] de Witt C S, Peng B, Kamienny P A, et al. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. arXiv preprint arXiv: 2003.06709, 2020

[40] Zhang T, Li Y, Wang C, et al. FOP: Factorizing optimal joint policy of maximum-entropy multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Chongqing, China, 2021: 12491-12500

[41] Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning//Proceedings of the 36th International Conference on Machine Learning. California, USA, 2019: 2961-2970

[42] Wang Y, Han B, Wang T, et al. Off-policy multi-agent decomposed policy gradients. arXiv preprint arXiv: 2007.12322, 2020

[43] Chen S Y. Kalman filter for robot vision: A survey. IEEE Transactions on Industrial Electronics, 2011, 59(11): 4409-4420

[44] Leibo J Z, Zambaldi V, Lanctot M, et al. Multi-agent reinforcement learning in sequential social dilemmas. arXiv preprint arXiv: 1702.03037, 2017

XU Cheng, Ph. D., associate professor. His research interests include swarm intelligence, multi-agent systems and Internet of Things.

YIN Nan, M. S. candidate. Her research interests include swarm intelligence and reinforcement learning.

DUAN Shi-Hong, Ph. D., associate professor. Her research interests include swarm intelligence and Internet of Things.

HE Hao, M. S. candidate. His research interests include swarm intelligence and wireless localization.

WANG Ran, Ph. D. candidate. Her research interests include swarm intelligence, wireless localization and distributed security.

Background

Self-organized, highly dynamic collaborative positioning and navigation research for emergency rescue is a new topic in areas such as military coordinated combat and coordinated counter-terrorism rescue. The research results will provide theoretical support for improving the high-precision positioning and navigation of emergency rescue in a blind environment, which is of great significance for enhancing the survivability and strike effect of military installations, as well as ensuring the life safety and operational efficiency of rescuers.

In scenarios such as military operations, counter-terrorism

operations, and fire rescue, it is crucial to obtain the location information of drones, ships, and personnel. Emergency rescue incidents often occur in a blind environment, that is, the location and environmental information of the operation are unknown. Relying on independent targets to obtain high-precision and high-reliability location estimates is often difficult to achieve. In modern warfare, in order to achieve combat objectives, a coordinated combat method is often adopted. Through information sharing, the decentralized independent platform is used as a distributed detection device and weapon system to improve the overall combat capability. In the application of anti-terrorism and fire rescue, through the sharing of information between combatants, the use of network collaborative positioning technology can effectively suppress the cumulative errors of autonomous navigation and positioning, shorten the search time, ensure the safety of the operator and increase the survival probability of the crashed person. Therefore, making full use of the measurement information between multiple cooperative facilities and personnel to improve their positioning and navigation accuracy is of great significance for enhancing the survivability and strike effect of military facilities, as well as ensuring the life safety and operational efficiency of rescue personnel.

This work is supported in part by the National Natural Science Foundation of China under Grant No. 62101029, in part by the China National Postdoctoral Program for Innovative Talents under Grant No. BX20190033, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2019A1515110325, in part by the Project Funded by China Postdoctoral Science Foundation under Grant No. 2020M670135, in part by the Postdoctoral Research Foundation of Shunde Graduate School of University of Science and Technology Beijing under Grant No. 2020BH001, and in part by the Fundamental Research Funds for the Central Universities under Grant No. 06500127.