赵花蕊. 利用团队间的影响解决多智能体强化学习中的奖励冲突[J]. 智能计算机与应用,2024,14(10):56-62. DOI:10.20169/j.issn.2095-2163.241007

利用团队间的影响解决多智能体强化学习中的奖励冲突

赵花蕊
(河南省平台经济发展指导中心, 郑州 450008)

摘 要: 本文提出了一种基于智能体间相互作用的 MARL 学习框架,称为 IC,以解决 MARL 中稀疏奖励环境导致智能体产生冲突的问题。IC 的主要功能是根据智能体间的高斯核函数大小赋予不同的值,计算出智能体的影响矩阵,并将影响矩阵的核范数作为额外奖励引入到目标函数中,以提高智能体探索性能以及团队之间的协作能力。实验结果表明,IC 可以显著提高智能体间的协作能力,并在稀疏奖励环境中加速智能体对最优策略的学习。这是首次在 MARL 中尝试利用智能体之间的相互影响来促进智能体的探索能力。

关键词: 多智能体强化学习; 稀疏奖励; 奖励冲突; 高斯核函数; 核范数
中图分类号: TP18 **文献标志码**: A **文章编号**: 2095-2163(2024)10-0056-07

Solving reward conflict in multi-agent reinforcement learning by using the influence between teams

ZHAO Huarui
(Platform Economy Development Guidance Center of Henan Province, Zhengzhou 450008, China)

Abstract: This article proposes a MARL learning framework based on the interaction between agents, called IC, to solve the problem of conflicts between agents caused by sparse reward environments in MARL. The main function of IC is to assign different values based on the Gaussian kernel function size between agents, calculate the influence matrix of agents, and introduce the kernel norm of the influence matrix as an additional reward into the objective function to improve the exploration performance of agents and the collaboration ability between teams. The experimental results indicate that IC can significantly improve the collaboration ability between agents and accelerate their learning of optimal strategies in a sparse reward environment. This is the first attempt in MARL to utilize the mutual influence between agents to promote their exploration ability.

Key words: multi-agent reinforcement learning; sparse reward; reward conflict; Gaussian kernel function; kernel norm

0 引 言

强化学习(Reinforcement Learning, RL)是一种机器学习方法,旨在通过与环境交互让智能体(agent)获取经验,并从中学习如何做出最佳决策。多智能体强化学习(Multi-Agent Reinforcement Learning, MARL)则是在强化学习框架下研究多个智能体相互作用、协作或对抗的问题[1]。MARL 已应用于许多场景,例如多智能体协同控制、多智能体路径规划和多智能体决策等领域[2]。然而,MARL 所面临的挑战包括其他智能体的建模、合作或竞争策略的设计、动态环境的变化和不确定性。为了应对这些挑战,近年来涌现出了许多新的 MARL 算法和技术,例如深度多智能体强化学习[3]、分布式多智能体强化学习[4]和演化博弈[5]等。

当 MARL 智能体面临稀疏奖励时,就可能会出现恶意竞争的情况。这是因为智能体需要尽快获得奖励,以便在竞争中获胜。当奖励非常稀疏时,智能体可能会采取一些破坏性的行为来增加自己的奖励,智能体之间会产生激烈竞争,并导致同质行为,使智能体难以形成有效的团队策略[6]。此外,有限的奖励信号可能使得智能体表现出懒惰行为,并使智能体陷入局部最优[7],导致智能体忽略潜在的团队策略,这些策略可能在未来产生更多的奖励。例

基金项目: 国家自然科学基金(61972092)。
作者简介: 赵花蕊(1984-),女,高级工程师,主要研究方向:电子政务,软件工程,人工智能,信息安全。Email:995791619@qq.com
收稿日期: 2023-06-12

如,在足球比赛中,团队奖励只有在进球时才会授予。然而,在一场激烈的比赛中,进球可能是一项艰巨的任务。假设所有球员(包括守门员)的行为在整个比赛中追逐足球,那么赢得比赛就很困难,这严重限制了学习的效率和性能。

为解决上述问题,研究者已经开发了各种算法,如奖励塑造[8]、好奇心[9]、迁移学习[10]、多评论家[11]和多任务[12]。基于奖励塑造的解决方案需要密集的个人奖励,并引入噪声来改变智能体的学习目标,但是这种方法可能会以不可预测的方式影响智能体的学习过程。对于具有大状态空间的复杂任务来说,好奇心可能无效,因为状态空间呈指数级增长,而且部分可观察性不能直接促使智能体进行协作。迁移学习、多评论家和多任务是 2 个最大化个人和团队回报的学习目标的相互集成。这些算法中涉及到的个体主体很难避免受到其他主体的影响,因此无法形成最优策略。虽然在个人和团队奖励方面存在一定程度的整合,但迁移学习的整合程度较弱。在学习团队奖励后,智能体可能会很快忘记预先训练过的技能。尽管这些方法已经取得了成功,但仍然面临一些挑战需要解决。其中一个主要的问题是这些算法无法对具有许多智能体和复杂环境的大规模问题进行有效扩展。另一个挑战是分配转移问题,也就是学习到的策略可能无法很好地应用于新的环境或场景中。此外,这些方法中的一些需要大量的计算资源,并且训练起来在计算上可能很昂贵。

智能体间的冲突行为会严重限制智能体的探索能力。受人类社会学的启发,考虑到每个个体的决策和行为都会影响到其他个体,可以促进合作,也可能导致竞争和冲突,因此需要考虑每个个体的行为对其他个体目标的影响。当团队成员之间相互考虑其他个体的策略变化,成员更有可能合作,共同实现团队目标。同时,团队成员之间的合作也会增强彼此之间的关系,促进团队更快实现团队目标。

本文提出了一种新的 MARL 框架来解决稀疏奖励导致的智能体间奖励冲突问题。本文的主要想法是利用智能体之间的相互影响来帮助智能体做出更好的决策,避免因为智能体间的冲突带来的探索效率低下问题以及智能体策略网络陷入局部最优。据研究所知,这是首次尝试在 MARL 中利用智能体之间的相互影响来促进智能体形成更好的协作。简而言之,本文的贡献可以总结如下:

(1)提出了一种基于智能体间相互作用的 MARL 学习框架,称为 IC。IC 的基本功能是保存当前智能体的策略快照,并根据智能体间的高斯核函数大小赋予不同的值,从而度量智能体间的影响力。这种做法使得智能体不仅需要考虑未来奖励,而且还需要考虑与其他智能体相互协作。

(2)提出了一种影响控制策略来控制智能体间的影响大小,即:计算影响矩阵的核范数作为额外奖励引入到目标函数中,通过最大化额外奖励来使得策略矩阵的核范数提高策略的多样性。此外,引入平衡因子使得距离较近的智能体之间的影响不会过大以及距离较远的智能体之间的影响程度不会过小。

(3)研究最后在常用的 MARL 的实验环境中进行了大量的实验来证明所提出的方法的有效性。结果表明,IC 可以显著提高智能体的探索能力,并在稀疏奖励环境中加速智能体对最优策略的学习。

1　相关工作

MARL 是一个重要的研究领域,涉及训练多个智能体相互作用及其环境。在 MARL 中,稀疏奖励的问题是一个常见的挑战。为了解决这个问题,研究人员提出了各种算法和技术,包括以下工作:具有 RewardsShaping 的多智能体深度 Q 网络(MADQN)通过引入 RewardsShaping 来解决稀疏奖励问题,RewardsShaping 是与目标相关但与任务不直接相关的奖励。这些成形奖励可以提供额外的信息,帮助智能体更快地学习正确的策略。多智能体深度确定性策略梯度(MADDPG)扩展了流行的 DDPG 算法来处理多个智能体[13]。同时使用一个集中的评论家来估计所有智能体的价值函数,并使用去中心化的参与者来为每个智能体生成动作。另一种算法是反事实多智能体策略梯度(COMA),使用集中的批评者来估计多个智能体所采取的联合行动的价值[14]。这种方法鼓励智能体学习优化全局奖励而不是个人奖励的联合策略。

近年来,研究者提出了 IRAT[15],建立了 2 套学习策略,包括个人策略和团队策略。这 2 种策略保持密切联系,利用 Kullback-Leibler(KL)分歧来解决团队奖励稀疏的问题。从本质上讲,外部奖励稀疏的问题可以得到缓解,但在训练阶段仍然发挥着主导作用。先前的研究结果表明,由于团队奖励稀疏,智能体具有同质化行为。另外,也有研究者使用“陌生度”来衡量主体的观察结果有多不熟悉,并考虑整体状态的不熟悉程度[16]。探索奖励是使用奇异性计算的,并且在 MARL 任务中不会受到随机转

换的很大影响。为了防止探索奖励掩盖外在奖励，提出了一个单独的行动价值函数，并用这 2 种奖励进行训练。这使得 MARL 算法在探索方法中更加稳定。也有研究者提出了 Agent Time Attention (ATA)[17]，这是一种具有辅助损失的神经网络模型，用于在协作 MARL 中重新分配稀疏和延迟的奖励。ATA 在 MARL 环境的许多实例中都优于各种基线。

2 背景介绍

2.1 部分可观测马尔可夫决策过程

本文的重点是多智能体合作问题，特别是在去中心化部分可观测马尔可夫决策过程（Dec-POMDP）的背景下[18]。Dec-POMDP 可建模为 8 元组 $< I,S,A,P,R,O,\Omega,\gamma >$，其中 $I = \{1,2,\cdots,I,\cdots,n\}$ 是 n 个智能体的有限集，S 表示全局状态，A 表示有限操作集，$P(S' \mid S,A)$ 表示转换函数，描述智能体如何通过联合操作 $A \equiv (A^1,A^2,\cdots,A^n)_t$。在每个时间步长 t，智能体接收共享的全局奖励 $r_t = \gamma R(s,a)$，其中 $\gamma \in [0,1)$ 是折扣因子。此外，每个智能体 $i \in I$ 还接收其自己的观测 $o^i \in \Omega$，该观测来自观测函数 $o(s,i)$，该函数为全局状态为 s 的智能体 i 提供局部信息。此外，每个智能体都有自己的动作和观测历史 $\lambda^i \in \Lambda^i \doteq (\Omega^i \times A)^*$。由于部分可观察性，每个智能体的策略 $\pi^i(a^i \mid \lambda^i,\theta_i)$ 取决于 λ^i。合作智能体的目标是构建单独的策略，共同最大化预期的贴现累积奖励，定义为：

$$J(\hat{\theta}_i) \doteq E[\sum_{t=0}^{\infty} \gamma_t r_t \mid \hat{\pi}_i] \tag{1}$$

其中，$\hat{\pi}$ 表示团队策略，γ_t 表示步骤 t 的损失因子。梯度上升策略算法的参数由 θ 表示。

2.2 多智能体近端策略优化

近端策略优化（PPO）是一种基于策略的强化学习算法，利用了 2 个神经网络分支[19]。通过将智能体的当前状态 s 输入到神经网络中，PPO 获得更新智能体状态的动作和奖励。PPO 使用目标函数 $J(\theta)$ 并应用梯度上升来更新神经网络中的权重参数 θ，使智能体能够采取增加整体奖励值的行动。PPO 有 2 种变体，每种变体都有自己的详细公式，本文使用 $J^{Clip}(\theta_i)$ 旨在约束目标函数，以确保新旧参数之间的差异较小：

$$J^{Clip}(\theta_i) = E[\min(\eta_t^i)\, A_t^i, clip(\eta_t^i(\theta_i), 1-\xi, 1+\xi)\, A_t^i] \tag{2}$$

其中，$\eta_t^i(\theta_i) = \dfrac{\pi_{\theta_i}(a_t^i \mid \tau_t^i)}{\pi_{\theta_i^{old}}(a_t^i \mid \tau_t^i)}$ 表示概率比；A 表示广义优势估计量（GAE）：$A_t^i = \sum_{l=0}^{h} (\gamma\lambda)^l \delta_{t+1}^i$；$\delta_t^i = r_t^i + \gamma V_{\phi_i}(s_{t+1}) - V_{\phi_i}(s_t)$ 表示智能体 i 在时间步长 t 的时间差误差（TD 误差）；h 表示轨迹的长度。

多智能体近端策略优化（MAPPO）是为多智能体场景设计的 PPO 算法的变体[20]。与 PPO 一样，MAPPO 是一种利用 Actor-Critic 架构的策略算法。然而，MAPPO 通过使用具有去中心化执行的集中式训练（CTDE）框架，将单智能体 PPO 扩展到多智能体领域。在 MAPPO 中，批评者学习了一个集中值函数，该函数将其与单智能体 PPO 算法区分开来。

3 方法

3.1 总体框架

本文提出了 IC，其总体架构如图 1 所示，模拟了人类群体协作中考虑其他智能体的想法。这种方法可以帮助智能体更好地理解任务和环境，并加快学习过程。具体而言，IC 首先保存智能体不同阶段的策略快照，然后度量快照之间的差异性作为影响程度。接下来根据本文提出的距离度量方式 DF 计算出智能体间的影响大小 $\boldsymbol{N}_i$，然后利用高斯核函数和平衡因子赋予影响不同的权重值 $\boldsymbol{W}_i$，从而得到智能体 i 的影响矩阵 $\boldsymbol{M}_i = \boldsymbol{W}_i \boldsymbol{N}_i^{\mathrm{T}}$。然后计算 $\boldsymbol{M}_i$ 的核范数 $\| \boldsymbol{M}_i \|_*$ 作为奖励添加到智能体 i 的目标函数。

通过这种方法在智能体中添加一个额外的奖励让智能体 i 更关注其他智能体，防止在稀疏奖励环境中智能体很难获得正向的奖励信号，因其往往会倾向于优先考虑自身利益，而忽视与其他智能体的合作。这种方法在直觉上类似于使智能体不只是考虑个人奖励，而是要考虑到整个团队的总体表现。

3.2 策略快照间的影响

为了评估多样性，需要保存一些历史策略的快照，并使用一些度量方法来计算彼此之间的影响。在多智能体强化学习中，这一点尤为重要，因为每个智能体都需要维护自己的策略，并且与其他智能体进行交互时，需要根据当前策略选择动作。为了保存策略快照，通常可以使用一些简单的序列化方法，将策略的参数保存起来。对于连续动作的环境，如 MPE、Multiworks 和 SMAC，可以将动作采样的结果表示为高斯分布或伯努利分布，并将分布的参数保存为策略快照。对于高斯分布，可以保存其均值和方差参数；对于伯努利分布，可以保存其概率参数。即，保存高维 Gaussion 分布的参数 μ 和 σ，以及保存高维 Bernoulli 分布的参数 p。

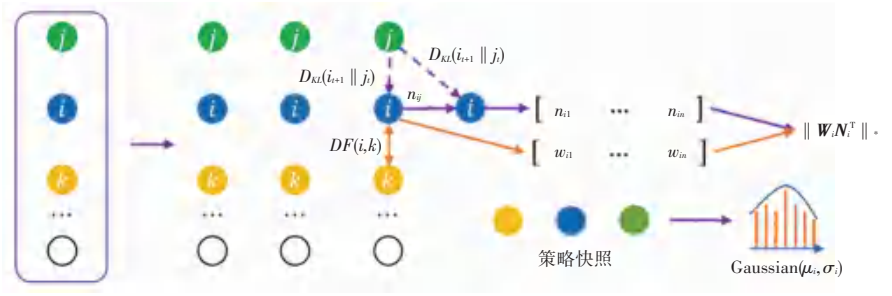

图 1　IC 的整体架构

Fig. 1　The overall architecture of IC

Kullback-Leibler 散度(KL 散度)是用于度量概率分布之间差异的指标,常用于衡量 2 个概率分布之间的相似度。在本文中,可以将智能体 i 的策略函数看作是一个概率分布 p,而智能体 j 的策略函数看作是另一个概率分布 q。则智能体 i 的策略函数的改变对智能体 j 策略函数的改变的影响,就可以用 $D_{KL}(p \| q)$ 来度量。具体地,假设 t 时刻智能体 i 保存的策略快照为 $p_t(x)$,智能体 j 的策略函数为 $q_t(x)$;在 $t+1$ 时刻,智能体 i 的策略函数发生了改变,变为 p_{t+1},则智能体 i 策略函数的改变对智能体 j 策略函数的改变的影响可以用如下公式来计算:

$$n_{ij} = D_{KL}(p_{t+1} \| q_t) - D_{KL}(p_t \| q_t) \tag{3}$$

因此,可以得到影响度量矩阵 $\boldsymbol{N}_i$:

$$\boldsymbol{N}_i = [\, n_{i1} \quad \cdots \quad n_{in} \,] \tag{4}$$

$$\boldsymbol{N} = \begin{bmatrix} \boldsymbol{N}_1 \\ \cdots \\ \boldsymbol{N}_n \end{bmatrix} = \begin{bmatrix} 0 & \cdots & n_{1n} \\ \vdots & \ddots & \vdots \\ n_{n1} & \cdots & 0 \end{bmatrix} \tag{5}$$

3.3　影响矩阵系数

常见的 MARL 实验环境均为二维平面,因此为了更好地度量二维平面上智能体之间影响程度的方式,研究采用高斯核函数作为基础,然后引入了平衡因子 μ 来防止影响过大或过小。即,对于位置分别为 (x_1, y_1) 和 (x_2, y_2) 的 2 个智能体 i, j, 彼此间的影响程度可以定义为:

$$d(x_1, y_1, x_2, y_2) = \mu \times \exp\left(-\frac{(x_1 - x_2)^2 + (y_1 - y_2)^2}{2\sigma^2}\right) \tag{6}$$

其中,σ 表示一个控制影响范围的参数。引入平衡因子 μ 后得到:

$$DF(i, j) = Clip(d(i, j), 1 - \mu, 1 + \mu) \tag{7}$$

这种方式假设距离越近,影响程度越大,但同时通过裁剪的方式解决了距离远近对影响程度的影响,使得距离较近的智能体之间的影响不会过大以及距离较远的智能体之间的影响程度不会过小。

如果直接采用矩阵的秩作为损失函数或作为奖励,最大化矩阵秩将变成是一个 NP 的非凸问题,因此不能将其作为损失函数。此外,矩阵秩的值是离散的,不能准确地反映策略的新颖性,因此将矩阵秩的原始值作为奖励来指导学习并不能很好地实现研究的目标。在数学上,正如 1 范数是向量范数的最紧凸松弛一样,矩阵秩的计算通常被核范数取代,核范数已被证明是秩的凸包络。因此,可以通过近似地最大化核范数来保持新颖性。与秩相比,核范数具有几个好的性质:首先,核范数的凸性使得在优化中开发快速收敛的算法成为可能。其次,核范数是一个连续函数,这对许多学习任务都很重要。当矩阵核范数较小时,表示矩阵中的奇异值较小,矩阵中的元素相对较集中,因此智能体的策略多样性较小。相反,当矩阵核范数较大时,表示矩阵中的奇异值较大,矩阵中的元素相对较分散,因此智能体的策略多样性较大。

本文优先考虑智能体策略多样性,这使智能体能够在大多数情况下形成有效合作。使用矩阵核范数可以降低矩阵的秩,从而更好地捕捉底层结构,有助于防止过拟合,并可以提高模型的泛化性能,使其更容易优化。因此,本文使用以下指标来表示智能体策略之间的多样性:

$$\boldsymbol{W}_i = [\, w_{i1} \quad \cdots \quad w_{in} \,] = [\, DF(i,1) \quad \cdots \quad DF(i,n) \,] \tag{8}$$

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 \\ \cdots \\ \boldsymbol{W}_n \end{bmatrix} = \begin{bmatrix} 0 & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & 0 \end{bmatrix} \tag{9}$$

结合前面的策略度量矩阵 $\boldsymbol{N}$,可以得到影响矩阵 $\boldsymbol{M}$:

$$\boldsymbol{M} = \boldsymbol{W}\boldsymbol{N} \tag{10}$$

因此,使用以下指标来表示智能体 i 的目标函数:

$$J(\theta_i) = E[\min(\eta_t^i) A_t^i, clip(\eta_t^i(\theta_i), 1 - \xi, 1 + \xi) A_t^i] +$$

$$\| \boldsymbol{M} \|_* \tag{11}$$

$$\theta_i = \theta_i - \alpha \nabla_{\theta_i} \{ J(\theta_i) \} \tag{12}$$

其中，$\| \boldsymbol{M} \|_* = \sum_{j=1}^{\min(a,b)} \sigma_j$，这里$\sigma_j$表示$\boldsymbol{W}_i$的奇异值；$\| \cdot \|_*$表示核范数代表矩阵在该方向上的扩展或压缩程度；$a$，$b$分别表示矩阵的长和宽。核范数可以用来控制模型的复杂度和泛化能力。

4 实验

在本节中，首先全面阐述了本文中使用的实验设置，然后使用 2 个探索性实验来展示 IC 和其他算法的对比。最后，本文进行了比较实验，验证了策略快照和影响矩阵系数对算法的改进。

4.1 实验设置

在本文的实验中，将 IC 与 RewardShaping、PCGrad、MAPPO、Multi-Critic 和 IRAT 进行了比较，利用了 MPE、Multiwalker 和 SMAC 环境。此外，还对 MPE Attack、MPE Spread 和 Multiwalker 的上述改进进行了比较实验，环境如图 2 所示。具体来说，实验的环境设置如下：

(1) MPE Predator Prey：场景中 5 只较慢的捕食者合作捕捉 2 只较快的猎物，环境中有 2 个障碍物阻挡道路。只有当捕获一只猎物时，才会给予 50 个奖励，在这种情况下捕食者会出现奖励冲突的问题。

(2) Spread：场景中设置了 4 个智能体和 3 个地标，智能体学习合作找到所有地标。只有当多个智能体同时检测到一个地标时，地标才会被覆盖。使用了 2 个误导性地标来误导智能体的判断。如果地标被覆盖，智能体将获得 8 个团队奖励。

(3) Attack：在此场景中设置了一个地标和 3 个智能体。团队的目标是让 3 个智能体同时到达地标并进行攻击。当 3 个智能体同时攻击地标时，将不可避免地发生碰撞，影响团队策略的学习。如果完成了攻击，环境将返回 20 个团队奖励。

(4) 大空间环境：针对大空间环境中的奖励冲突问题，在 Multiwalker 环境中进行了一项策略，放置了一个包裹在由算法控制的 2 个双足机器人的顶部。双足机器人试图将包裹向右移动尽可能远。

(5) 复杂环境：研究中为了检测该算法在复杂环境中奖励冲突的效果，本文在 SMAC 环境中的 6h_vs_8z、5m_vs_6、8m_vs_9m 三个地图上进行了测试。SMAC 采用了稀疏奖励设置（通常被大多数多智能体算法如 MAPPO 使用），刻意引导智能体由于奖励稀疏而产生协作冲突的问题。

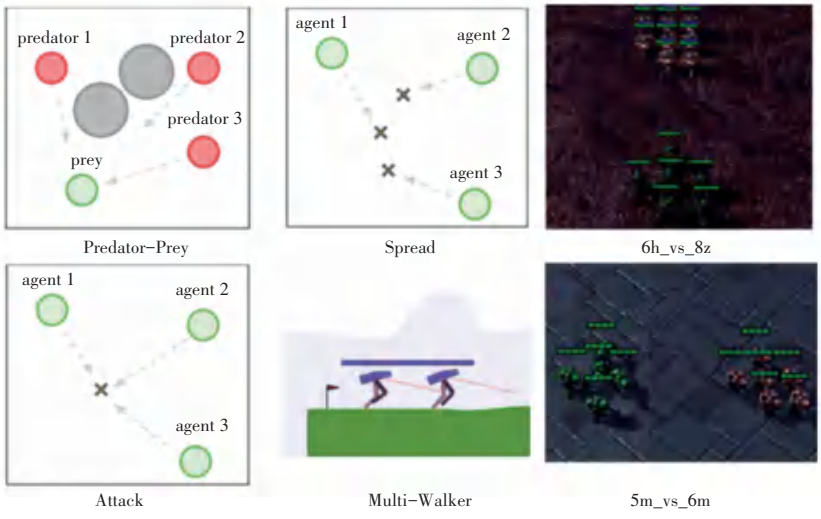

图 2 MPE、Multiwalker 和 SMAC 环境实例

Fig. 2 Examples of the MPE, Multiwalker, SMAC environments

4.2 基线方法

下面全面回顾了在相关研究领域广泛使用的流行算法和模型。通过建立一组标准基线，可以系统地评估本文提出的方法的性能，并得出有意义的结论。以下是对实验中使用的比较算法的简要介绍：

(1) PCGrad 是一种梯度扰动技术，用于提高训练过程的收敛性和稳定性。可通过使用控制变量估计基线来减少梯度方差。

(2) 在多智能体强化学习（MARL）中，迁移学习可以被用来在不同的任务和环境之间分享已经学习到的知识和经验。这种知识和经验可以通过预训练与微调的方式进行共享，从而提高智能体的性能和效率。

(3) Multi-Critic 是一种深度强化学习算法中的

Actor-Critic 方法，主要思想是将多个评论家(Critic)网络用于评估状态-动作对的价值，并将这些评估结果进行聚合，在 Actor 网络中使用平均或加权平均来得到最终的动作策略。与传统的单个评论家网络不同，Multi-Critic 使用多个评论家网络来评估状态-动作对的价值。

(4) Qmix 使用集中值函数来改善代理之间的协调。COMA 使用反事实基线方法来解决价值估计中的非平稳性，而 VDN 将全局 Q 函数分解为个体主体贡献。

4.3　与其他 MARL 算法的定量比较

结合所提出的改进，进一步评估了所提出的 IC 方法在不同场景下的有效性，实验的收敛过程如图 3 所示，实验定量结果见表 1。表 1 中，SMAC 的单位为%，表示胜率；其他环境无单位，表示奖励值。总的来说，RewardShaping 和 Transfer Learning 也依赖于密集的外在奖励来指导主体的学习，这在本文的实验环境中无法提供令人满意的结果。IC 在大多数情况下具有显著的性能改进和更快的收敛速度，并在 MPE Spread、MPE attack 和 Multiwalker 中实现了最先进的性能和快速收敛速度。值得注意的是，IC 并没有给 Spread 环境带来太大的改善。分析认为：在简单的环境中，引入额外的策略多样性并不能改善算法，甚至会增加学习负担。随着环境的复杂性增加，例如 MPE Spread 和 MPE Attack 环境，IC 产生了更好的收敛速度和性能。在 Multiwalker 环境中，IC 则有着更明显的优势。接下来，本文又在具有稀疏奖励的复杂环境中进行实验。在 SMAC 6h_vs_8z、5m_vs_6m 和 8m_vs_9m 地图中，IC 在每一轮中获得的奖励和获胜率都高于基线，并且具有更快的收敛性。结果表明，在大多数环境中，IC 可以在加速智能体学习速度的同时获得最佳性能，这揭示了当前智能体在下一步选择上关注其他智能体可以解决因为奖励稀疏问题而产生奖励冲突的问题，可以提高 MARL 算法的性能。

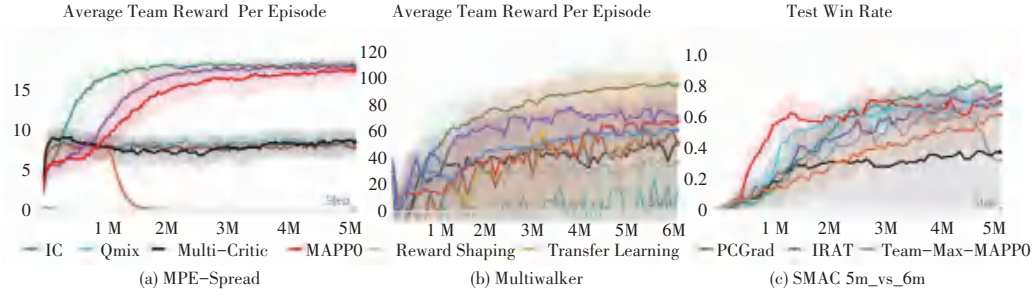

图 3　仿真收敛过程

Fig. 3　Experimental convergence process

表 1　在 MPE、MultiWork、SMAC 环境中，baseline method 和 IC 的定量比较

Table 1　Quantitative comparison of baseline method and IC in the environment of MPE, MultiWork, SMAC

算法	环境						
	Predator-Prey	Spread	Attack	MultiWork	6h_vs_8z/%	5m_vs_6m/%	8m_vs_9m/%
RewardShaping	42.53	0.00	0.00	1.27	0.0	0.0	24.4
Transfer Learning	68.79	0.00	3.46	14.38	0.0	37.1	71.4
PCGrad	81.33	12.82	7.17	19.54	12.2	59.6	67.1
^Multi-Critic	99.61	10.72	11.32	31.96	16.1	60.2	80.3
Qmix	97.64	20.75	14.97	37.42	12.9	55.7	84.4
COMA	91.21	21.80	36.44	41.12	21.8	41.5	85.1
MAPPO	77.16	18.84	52.72	67.31	69.2	77.4	45.5
IRAT	105.68	21.61	60.72	71.68	72.2	81.2	85.1
IC	**113.56**	**24.76**	**64.97**	**97.89**	**83.4**	**88.7**	**93.8**

4.4　消融实验

为了评估影响度量矩阵和策略矩阵系数的效果，研究在《星际争霸 II》中名为 3 m 的公平地图上进行了实验。实验使用了 IC-N、没有使用任何策略；IC-S、仅使用影响度量矩阵；IC、使用了影响度量矩阵和策略矩阵系数组成的影响矩阵；另外，为了证明核范数更适合 MARL，实验中还使用了 Frobenius 范数、$L0$ 范数和 $L1$ 范数。

在 3s5z 地图上的消融实验结果如图 4 所示。图 4 中的结果表明，与 IC 相比，IC-N 表现出较慢的收敛速度和较差的有效性能。IC 能够在稀疏奖励环境中快速学习有用的团队策略。另一方面，IC-S 引入了

团队之间的关系,可以更好地学习优秀策略。但是由于无法控制影响,对算法的收敛产生了一定的影响,生成的收敛曲线更加扭曲,因此性能不如 IC。

对于范数的选择,实验发现:对于 Frobenius 范数,由于策略快照属于稀疏矩阵,核范数的计算只涉及到非零奇异值,因此可以更准确地反映稀疏矩阵的情况。而 F 范数却需要把所有元素都纳入计算,因此不能很好地适应稀疏矩阵的情况。*L*0 和 *L*1 直接使得算法失效。具体来说:*L*0 范数可以直接计算非零元素的个数,虽然计算简单,但是这种方法降低了复杂性,使得学习更好的策略具有挑战性。*L*1 范数和核范数可通过凸优化进行求解,计算复杂,不适合高维矩阵计算。相对而言,核范数是更简便的优化方法,使得学习能够兼顾计算简单和良好性能两方面。

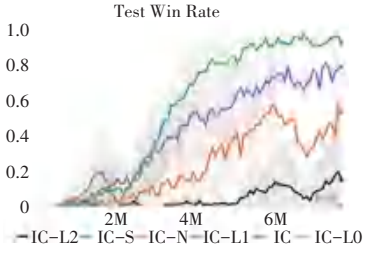

图 4 在 3s5z 地图上的消融实验

Fig. 4 Ablation experiment on a 3s5z map

5 结束语

本文提出了一种基于智能体间相互作用的 MARL 学习框架 IC,旨在解决 MARL 中稀疏奖励导致的智能体产生奖励冲突的问题。IC 通过保存当前智能体的策略快照并根据智能体间的高斯核函数大小赋予不同的值,从而度量智能体间的影响力。进一步地,通过引入影响控制策略,计算影响矩阵的核范数作为额外奖励引入到目标函数中,以提高智能体相互合作的多样性。实验表明,IC 在常用的 MARL 环境中实现了最优秀的性能,有效解决了智能体之间的冲突问题,并在稀疏奖励环境中加速智能体对最优策略的学习。

参考文献

[1] 杜威, 丁世飞. 多智能体强化学习综述[J]. 计算机科学, 2019, 46(8): 1-8.

[2] 宋琛. 多智能体的战术行为决策研究及应用[D]. 南京:南京航空航天大学,2011.

[3] 刘志飞, 曹雷, 赖俊,等. 基于多智能体深度强化学习的无人机集群自主决策[J]. 信息技术与网络安全, 2022,41(5): 77-81.

[4] 罗青, 李智军, 吕恬生. 复杂环境中的多智能体强化学习[J]. 上海交通大学学报, 2002, 36(3): 302-305.

[5] 杨波, 徐升华. 基于多智能体建模的知识转移激励机制的演化博弈模型与仿真[J]. 计算机工程与科学, 2010, 32(6): 162-166.

[6] CHRISTIANOS F, SCHÄFER L, ALBRECHT S. Shared experience actor-critic for multi-agent reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 10707-10717.

[7] GRONAUER S, DIEPOLD K. Multi-agent deep reinforcement learning: A survey[J]. Artificial Intelligence Review, 2022,55: 895-943.

[8] NG A Y, HARADA D, RUSSELL S. Policy invariance under reward transformations: Theory and application to reward shaping [C]// International Conference on Machine Learning (ICML). Bled, Slovenia:ACM, 1999, 99: 278-287.

[9] ZHENG Lulu, CHEN Jiarui, WANG Jianhao , et al. Episodic multi-agent reinforcement learning with curiosity-driven exploration[J]. arXiv preprint arXiv:2111.11032,2021.

[10] LIU Yong, HU Yujing, GAO Yang, et al. Value function transfer for deep multi-agent reinforcement learning based on N-Step returns[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). Macao:dblp, 2019: 457-463.

[11] HE Zhenglei, TRAN K P, THOMASSEY S, et al. Multi-objective optimization of the textile manufacturing process using deep-Q-network based multi-agent reinforcement learning[J]. Journal of Manufacturing Systems, 2022, 62: 939-949.

[12] SILVA F L D, COSTA A H R. A survey on transfer learning for multiagent reinforcement learning systems[J]. Journal of Artificial Intelligence Research, 2019, 64: 645-703.

[13] 赵冬梅, 陶然, 马泰屹, 等. 基于多智能体深度确定策略梯度算法的有功-无功协调调度模型[J]. 电工技术学报, 2021, 36(9): 1914-1925.

[14] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[J]. arXiv preprint arXiv:1705.08926,2017.

[15] WANG Li, HU Yujing, ZHANG Y, et al. Individual reward assisted multi-Agent reinforcement learning[C]//International Conference on Machine Learning. Baltimore, USA: ACM, 2022: 23417-23432.

[16] KIM J B, CHOI H B, HAN Y H. Strangeness-driven exploration in multi-Agent reinforcement learning[J]. arXiv preprint arXiv: 2212.13448, 2022.

[17] SHE J, GUPTA J K, KOCHENDERFER M J. Agent-time attention for sparse rewards multi-Agent reinforcement learning [J]. arXiv preprint arXiv:2210.17540, 2022.

[18] AMATO C, CHOWDHARY G, GERAMIFARD A, et al. Decentralized control of partially observable Markov decision processes[C]//52nd IEEE Conference on Decision and Control. Firenze, Italy: IEEE, 2013: 2398-2405.

[19] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv: 1707.06347, 2017.

[20] YU Chao, VELU A, VINITSKY E, et al. The surprising effectiveness of ppo in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.