DOI: 10.16725/j.1673-808X.2023106
引用本文: 徐鼎, 赵岭忠, 翟仲毅. 基于集成的多智能体深度强化学习决策优化方法[J]. 桂林电子科技大学学报, 2026, 46(3): 263-270.
Xu Ding, Zhao Lingzhong, Zhai Zhongyi. Ensemble-based decision optimization method for multi-agent reinforcement learning[J]. Journal of Guilin University of Electronic Technology, 2026, 46(3): 263-270.

基于集成的多智能体深度强化学习决策优化方法

徐　鼎，赵岭忠，翟仲毅

（桂林电子科技大学 计算机与信息安全学院, 广西 桂林　541004）

摘　要： 在中心训练分散执行的框架下, 智能体虽然在训练期间可以获取全局状态, 但是所有智能体都在同时更新自己的策略, 因此会存在环境非平稳性问题, 从而导致智能体在某些状态下的决策受到影响。此外, 许多方法通过经验回放机制来减少样本相关性, 但同时可能会导致某些状态下的样本未被抽取出来以用于训练, 进而导致智能体在这些状态下决策效果不佳。以上问题导致多智能体深度强化学习中智能体的决策鲁棒性不足。针对该问题, 提出了一种基于集成的多智能体深度强化学习决策优化方法。该方法通过给智能体集成多个策略网络或者动作值网络来提升智能体决策的鲁棒性, 当某个状态下单个策略或动作网络失效时, 智能体仍可依据其他正常工作的网络进行决策。另外, 基于强化学习的特性提出了基于动作置信度权重的集成方法和基于最优动作投票的集成方法, 能有效对多个网络的输出进行集成。该方法在多智能体深度强化学习领域具有广泛的适用性, 既适用于基于演员−评论家的方法, 又适用于基于值函数分解的方法。在 3 种不同的实验环境中进行了实验, 对 6 种方法进行了对比测试。实验结果表明, 将该方法应用到主流的演员−评论家方法和值函数分解方法时, 能使多智能体系统的对战胜率提升 50 个百分点, 奖励值最高提升 40。

关键词： 多智能体系统; 深度强化学习; 合作; 决策; 集成
中图分类号： TP301　　**文献标志码：** A　　**文章编号：** 1673-808X(2026)03-0263-08

Ensemble-based decision optimization method for multi-agent reinforcement learning

XU Ding，ZHAO Lingzhong，ZHAI Zhongyi

（School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China）

Abstract: In the centralized training and decentralized execution(CTDE) framework, agents can access the global state during training, but update their policies independently. This results in environmental non-stationarity, which degrades decision-making performance in certain states. Additionally, experience replay is commonly adopted to reduce sample correlation, but it may lead to insufficient sampling of certain states, resulting in poor decision-making performance. These issues reduce the robustness of decision-making in multi-agent deep reinforcement learning. To address these issues, an ensemble-based multi-agent deep reinforcement learning method is proposed. The proposed method integrates multiple policy or action-value networks to enhance decision robustness. When one network performs poorly in a particular state, decisions can still be made based on other effective networks. Two integration strategies, namely confidence-weighted action selection and action voting, are proposed to effectively combine outputs from multiple networks. The proposed approach demonstrates strong applicability in multi-agent deep reinforcement learning and applies to both actor-critic and value-based methods. Experiments were conducted in three environments by comparing the proposed method with six representative approaches. The results show that, when applied to the mainstream actor-critic algorithm MADDPG and the value decomposition method QMIX, the proposed method increases the win rate of the multi-agent system by 50%, and improves the

收稿日期: 2023-11-20
基金项目: 国家自然科学基金 (61862014)
通信作者: 赵岭忠(1977—), 男, 教授, 博士, 研究方向为软件模型检测、形式化技术。E-mail: zhaolingzhong163@163.com

maximum reward by 40.

Key words: multi-agent system; deep reinforcement learning; cooperation; decision-making; ensemble

多智能体系统描述了多个分布式实体,即所谓的智能体,它们在一个共享环境中自主地做出决策,并完成一个共同目标。决策是指智能体从获得局部观察到选择出动作的过程。多智能体深度强化学习在解决序列决策问题中具有较大潜力,如自动驾驶汽车[1]、机器人控制[2]、交通运输调度[3]、电力系统优化[4]、分布式传感网络[5]、计算卸载[6]、交通信号灯控制[7]等。

在多智能体深度强化学习领域,主流的算法可分为基于演员−评论家的方法和基于值函数分解的方法两类。在基于演员−评论家的方法中,每个智能体将局部观察输入策略网络进行决策,并通过全局的评论家网络“指导”策略网络进行训练。MADDPG[8]是一种代表性的基于演员−评论家的多智能体深度强化学习算法,它通过考虑其他智能体的策略以及使用每个智能体的策略集合,来应对 Q 学习面临的环境非平稳性挑战,以及策略梯度方差随智能体数量增加而变大的问题。Foerster 等[9]提出了一种反事实的多智能体演员−评论家方法,为了解决多智能体信用分配的挑战,它使用了一个反事实基线,将单个智能体的动作边缘化,同时保持其他智能体的动作不变,为每个智能体计算出各自的 Q 值。Ackermann 等[10]将 TD3 算法扩展到多智能体中。Iqbal 等[11]提出了多演员−注意力−评论家方法,其中集中计算的评论家共享一个注意力机制,该机制在每个时间步为每个智能体选择相关信息。多演员−注意力−评论家中的注意力机制可在复杂多智能体环境中实现更有效和可扩展的学习。Li 等[12]提出了一种 M3DDPG 算法,通过引入极大极小学习目标并对其进行优化来提升算法的鲁棒性。Xiao 等[13]提出了一种 ROLA 算法,通过为每个智能体使用局部的评论家来减少策略梯度估计的方差,并提升算法的鲁棒性。

在基于值函数分解的方法中,智能体不使用策略网络,而是使用个体 Q 网络进行决策。训练时,通过个体 Q 值来近似一个全局 Q 值,并通过全局 Q 值的时序误差进行网络参数更新。针对具有单一联合奖励信号的合作多智能体强化学习场景,Sunehag 等[14]提出了值函数分解(VDN)方法,VDN 使用所有智能体的个体 Q 值的累加来近似全局 Q 值,以解决虚假奖励问题及部分可观察性导致的懒惰智能体问题。由于 VDN 通过累加来近似全局 Q 值的方法具有局限性,Rashid 等[15]提出了一种 QMIX 方法,用个体 Q 值的复杂非线性组合来近似全局 Q 值。QMIX 在结构上强制全局 Q 值与每个智能体的个体 Q 值的单调性,这使得全局 Q 值的最大化具有简单的获取方式,并保证了集中策略和分散策略之间的一致性。Son 等[16]提出了一种 QTRAN 方法,旨在使值函数分解可应用到任何可分解任务,QTRAN 使用 2 个额外的损失项来满足全局值函数和个体值函数之间贪婪动作选择的一致性。Rashid 等[17]提出了一种加权 QMIX 方法,通过加权投影来避免 QMIX 的表示限制,以更重视较好的联合动作。Wang 等[18]提出一种 QPLEX 方法,将 IGM(individual global max)原理编码到神经网络结构中,从而实现有效的值函数学习。Yang 等[19]提出一种合作多智能体强化学习的一般框架,从理论上推导了全局 Q 值和个体 Q 值之间的一般公式,并通过多头注意力机制的形成来近似全局 Q 值。

在多智能体系统中,智能体需要学习适应其他智能体的策略,以达到合作的目的。然而,在学习过程中,随着智能体的神经网络参数不断更新,在任意一个智能体看来,其他智能体的策略都是可能变化的。当其他智能体的策略发生改变时,智能体当前策略可能无法适应这种改变,进而使得智能体的决策受到影响。此外,由于使用经验回放机制,每次训练时从经验缓冲池中随机抽取定量的转移样本用于训练,某些状态下的样本可能未被充分训练,导致智能体的策略网络或动作值网络在某些状态下失效,从而影响智能体的决策。因此,多智能体系统存在决策鲁棒性不足的问题。

在以上研究中,智能体只依据一个策略网络或动作值网络进行决策,无法处理多智能体系统决策鲁棒性不足的问题。鉴于此,提出了一种基于集成多智能体深度强化学习决策优化方法。该方法为每个智能体训练多个策略网络或者动作值网络,智能体根据集成的多个网络的输出进行决策,以解决某些状态下因为单个策略网络或者动作值网络失效引起的智能体决策鲁棒性不足问题。当某个状态下单个策略或动作值网络失效时,智能体仍可依据其他正常工作的网络进行决策,以提升智能体决策的鲁棒性。

1 相关知识

1.1 分散式局部可观测马尔可夫决策过程

考虑到现实场景中的条件,多智能体系统中的智

能体无法总能获得环境中的全局状态, 因此一般使用分散式局部可观测马尔可夫决策过程 (Dec-POMDP) 对多智能体系统进行建模。Dec-POMDP 是对马尔可夫决策过程的扩展, 使用其建模的多智能体系统可通过一个八元组$\langle N,S,A,P,R,Z,O,\gamma\rangle$来表示。其中: $s\in S$ 表示全局环境状态; $N\equiv\{1,2,\cdots,n\}$, 为智能体的有限集合; 折扣因子$\gamma\in[0,1)$。由于部分可观察性, 每个智能体只能根据观察函数$O(s)$获取一个从全局环境状态映射的部分观察$z\in Z$。在每个时间步, 每个智能体$i\in N$根据自身的部分观察和策略选择一个动作$a_i\in A$, 同一时刻所有智能体的动作构成了一个联合动作$\boldsymbol{a}:=[a_i]_{i=1}^n\in A^n$。根据状态转移函数$P(s'|s,a)$: $S\times A\times S\to[0,1]$, 联合动作向量会导致所有智能体共享的环境转移到下一个状态s', 同时会获得环境反馈的奖励$R(s,\boldsymbol{a})$。智能体的目标是最大化累积长期奖励, 在完全合作的场景中, 所有智能体共享相同的团队奖励。

1.2　演员−评论家算法

演员−评论家算法的主要结构是作为演员的策略网络和作为评论家的值网络。值网络不仅需要输入单个智能体的信息, 还需要考虑整个系统的全局信息, 且每个智能体都有自己的全局值网络。执行期间, 智能体将自己的局部观察输入策略网络进行决策。训练期间, 值网络将目标 Q 值与估计 Q 值之间的平方误差作为损失项, 据此更新参数:

$$L(\varphi_i)=E_{s,a,r,s'}\left[(Q_i^{\mu}(s,a_1,\cdots,a_n)-y)^2\right],$$

$$y=r_i+\gamma Q_i^{\mu'}(x',a_1',\cdots,a_n')|_{a_j'=\mu_j'(o_j)}。$$

其中: s、s'为相邻的 2 个全局状态; μ'为目标策略; φ_i为智能体i的动作值网络的参数; Q_i为智能体i的动作值网络; n为智能体的数量。

策略网络采取梯度上升的方式进行更新:

$$\nabla_{\theta_i}J(\mu_i)=E_{s,a\sim D}\left[\nabla\theta_i\mu_i(a_i|o_i)\nabla_{a_i}Q_i^{\mu}(s,a_1,\cdots,a_n)\right]。$$

其中: θ_i为策略网络的参数; D为经验缓冲池; μ_i为智能体i的策略。

1.3　值函数分解

在基于值函数分解的多智能体深度强化学习中, 每个智能体都依据各自的个体动作值网络$Q_i(o_i,a_i)$进行决策, 并通过所有的$Q_i(o_i,a_i)$来近似一个全局动作值函数$Q_{\text{tot}}(o,a,s)$, 以实现中心训练分散执行。$Q_i(o_i,a_i)$与$Q_{\text{tot}}(o,a,s)$的关系可表示为

$$Q_{\text{tot}}(\tau,u,s)=f_s(Q_1(o_1,a_1),\cdots,Q_n(o_n,a_n)),$$

其中: o为联合观察; a为联合动作; f_s表示从Q_i到Q_{tot}的映射关系。智能体更新神经网络参数时, 利用时序差分误差进行更新:

$$Q_{\text{tot}}(o,a,s,\theta)\leftarrow\alpha(r_s^a+\gamma\max_{a^*\in A}Q_{\text{tot}}(o',a^*,s',\theta^-))+(1-\alpha)Q_{\text{tot}}(o,a,s,\theta)。$$

其中: α为学习率; o'为下一个联合观察; a^*为下一个具有最大联合动作值的联合动作; θ、θ^-分别为动作值网络和目标动作值网络的参数。

2　基于集成的多智能体深度强化学习决策优化方法

基于集成的多智能体深度强化学习决策优化方法(EDO)包括基于动作置信度权重的集成方法(ACW)和基于最优动作投票的集成方法(OAV)两种, 这 2 种方法都是基于强化学习的特性设计的集成方法。理论上, EDO 既适用于基于演员−评论家的方法, 也适用于基于值函数分解的方法, 在多智能体深度强化学习领域具有广泛的适用性。

2.1　基于 EDO 的多智能体深度强化学习模型

EDO 模型的结构如图 1 所示, 每个智能体内部都集成了 m 个策略网络或值网络, 集成模块的作用是将 m 个策略网络或值网络的输出集成为一个总体的动作概率分布或动作值。

在 EDO 模型中, 多智能体与环境交互的过程为: 每个智能体i根据环境的全局状态获取自己的局部观察, 每个智能体i将局部观察分别输入自身的 m 个策略网络或 m 个值网络, 并输出 m 个动作概率分布$P_i^j(o^i|\cdot)$或 m 个动作值$Q_i^j(o^i|\cdot)$, 然后采用 ACW 或 OAV 对这些输出进行集成, 并得到总体的动作概率分布$P_i(o^i|\cdot)$或动作值$Q_i(o^i|\cdot)$, 每个智能体i根据总体的动作概率分布$P_i(o^i|\cdot)$进行动作采样或者根据总体的动作值$Q_i(o^i|\cdot)$进行贪婪动作选择, 所有智能体选择的动作构成一个联合动作, 使环境转移到下一个状态, 每个智能体均获得环境反馈的团队奖励, 重复以上过程直到本回合结束。

在 EDO 模型中, 每个智能体集成的 m 个策略网络或值网络用同一个经验缓冲池中的转移样本$\langle s,a,r,s'\rangle$进行训练, 但每个策略网络或值网络都独立地从经验缓冲池中随机抽取各自用于训练的转移样本。在算法实现中, 经验缓冲池的典型容量为 5 000, 每次随机抽取的转移样本的典型数量为 32。每次训练时, m 个策略网络或值网络都独立地从经验池的 5 000 个样本中抽取 32 个样本, 因此每个策略网络或值网络训练时使用的样本通常存在差异, 它们的参数不同, 即它们存在多样性。

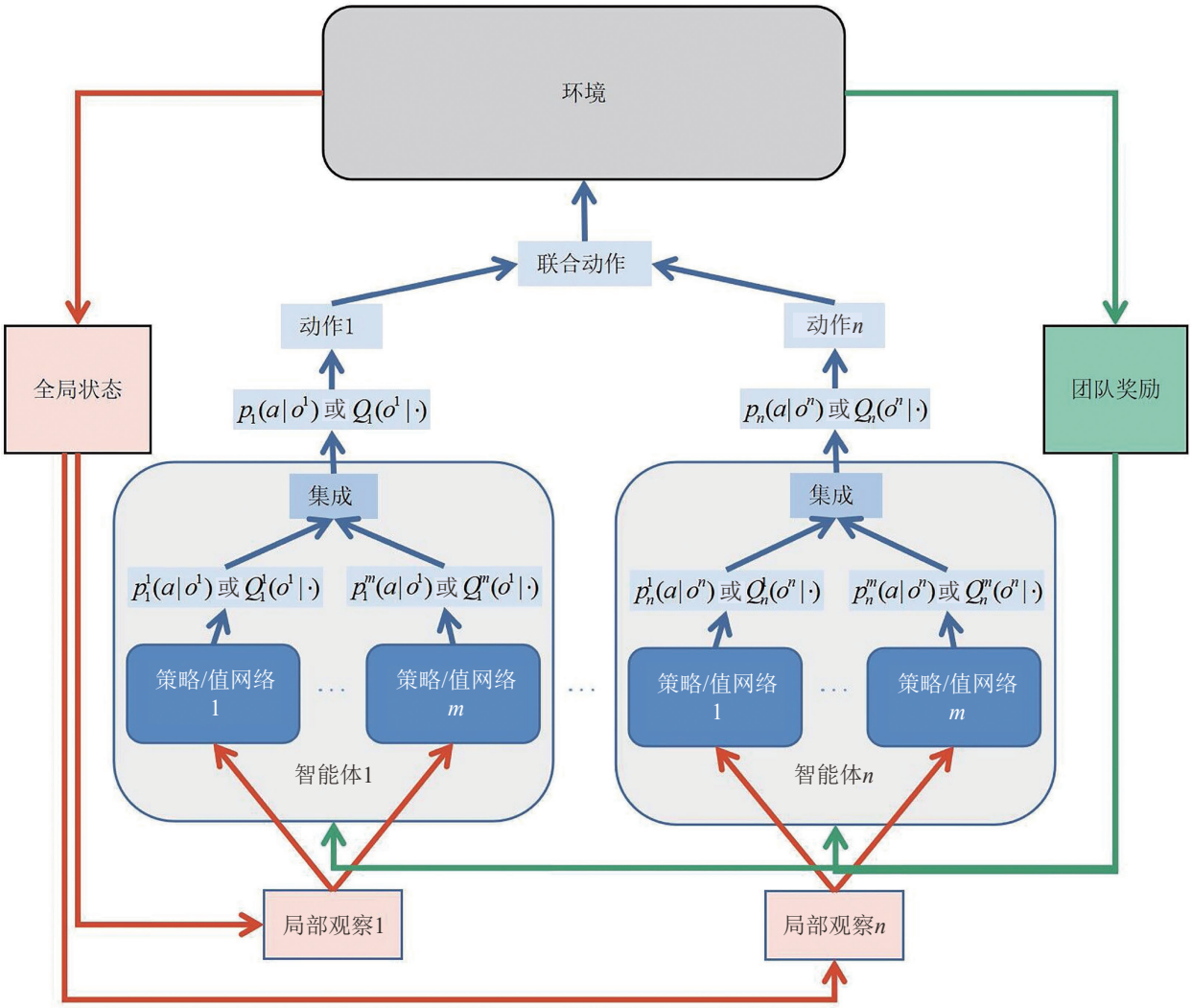

图 1　EDO 模型结构

由于 m 个策略网络或值网络之间存在多样性，在某个状态下，当某个策略网络或值网络无法输出正确的动作概率分布或动作值时，智能体可依据其他策略网络或值网络输出的正确信息进行正常决策。

2.2　EDO 的集成方式

EDO 模型的集成模块包含了 ACW 和 OAV 两种集成方法。在 ACW 集成方法中，使用了动作置信度的概念。在强化学习中，动作置信度的含义是智能体依据动作概率分布或动作值进行决策时的自信程度。智能体决策时，若对某个动作的偏好越明显，则该状态下的动作置信度越大；若对特定动作的偏好越不明显，则该状态下的动作置信度越小。在基于值的方法中，动作置信度定义为

$$\psi(Q(o|\cdot)) = Q_{\max}(o|\cdot) - Q_{\text{second}}(o|\cdot),$$

其中：$Q(o|\cdot)$为局部观察o下所有动作对应的动作值；$Q_{\max}(o|\cdot)$为在局部观察o下最大的动作值；$Q_{\text{second}}(o|\cdot)$为在局部观察$o$下第二大的动作值。在基于策略的方法中，动作置信度定义为

$$\psi(P(o|\cdot)) = P_{\max}(o|\cdot) - P_{\text{second}}(o|\cdot),$$

其中：$P(o|\cdot)$为局部观察o下所有动作对应的动作概率；$P_{\max}(o|\cdot)$为局部观察o下最大的动作概率；$P_{\text{second}}(o|\cdot)$为局部观察$o$下第二大的动作概率。

ACW 集成方法用每个策略网络或值网络的动作置信度作为其权重，将动作概率分布或动作值乘以权重，然后累加得到集成的动作概率分布$P_{\text{ACW}}(o|\cdot)$或动作值$Q_{\text{ACW}}(o|\cdot)$：

$$P_{\text{ACW}}(o|\cdot) = \sum_{i=1}^{m} \psi(P^i(o|\cdot))P^i(o|\cdot),$$

$$Q_{\text{ACW}}(o|\cdot) = \sum_{i=1}^{m} \psi(Q^i(o|\cdot))Q^i(o|\cdot)。$$

在 OAV 集成方法中，每个智能体集成的 m 个策

略网络或值网络依据最优动作进行投票。最优动作a^*定义为具有最大动作值或最大动作概率的动作。在每个策略网络或值网络的输出中, 都存在一个最优动作。OAV 集成方法的思想在于, 若智能体的多个策略网络或值网络的最优动作都是同一个动作a^*_{same}, 则认为选择该动作能为智能体带来最大的未来预期奖励, 需要尽量选取动作a^*_{same}。但是, 训练前期各个策略网络或值网络都未能得到充分训练, 其输出的最优动作具有较大的随机性, 越是训练前期, 输出的最优动作具有越大的随机性。因此, 不能在整个训练过程都完全依据a^*_{same}进行决策, 而应随着训练步数的增加, 逐渐增大对a^*_{same}的重视程度。此外, 当某个状态下所有集成的策略网络或者值网络的最优动作都不相同时, OAV 集成方法使用累加所有策略网络或值网络的输出进行集成。经以上分析可知, 用 OAV 集成方法对 m 个策略网络进行集成后得到的结果为

$$P_{\text{OAV}}(o|\cdot)=\sum_{i=1}^{k}P^i(o|\cdot)+\sum_{j=k+1}^{m}\left(1-\left(\frac{t_{\text{cur}}}{t_{\max}}\right)\right)P^j(o|\cdot),$$

其中: $P^i(o|\cdot)$为a^*_{same}对应的策略网络输出的动作概率分布; k为a^*_{same}对应的策略网络个数; $P^j(o|\cdot)$为最优动作不是a^*_{same}的策略网络输出的动作概率; t_{cur}为当前的训练步数; $t_{\max}$为最大训练步数。用 OAV 集成方法对 m 个策略网络进行集成后得到的结果为

$$Q_{\text{OAV}}(o|\cdot)=\sum_{i=1}^{k}Q^i(o|\cdot)+\sum_{j=k+1}^{m}\left(1-\left(\frac{t_{\text{cur}}}{t_{\max}}\right)\right)Q^j(o|\cdot),$$

其中: $Q^i(o|\cdot)$为a^*_{same}对应的策略网络输出的动作值分布; k为a^*_{same}对应的策略网络个数; $Q^j(o|\cdot)$为最优动作不是a^*_{same}的动作值。

2.3　EDO 的实现

MADDPG 算法为每个智能体都训练一个演员和评论家, 其中演员是一个策略网络, 用于执行期间智能体的决策, 评论家是一个值函数, 用于在训练期间指导策略网络更新。当 EDO 模型在 MADDPG 上实现时, EDO-MADDPG 将为每个智能体集成 m 个演员, 用于优化智能体的决策, 同时还需为 m 个演员训练 m 个对应的评论家。之所以要为每个演员都训练独立的评论家, 而非集成的 m 个演员共用同一个评论家, 是为了提升集成的 m 个演员的多样性, 进而提升 EDO-MADDPG 的性能。EDO-MADDPG 的结构如图 2 所示。

图 2 中: π_i^j为智能体i的第j个策略网络; Q_i^j为智能体i的第j个 Q 网络; o为智能体的局部观察; a为智能体的个体动作; ACW 和 OAV 分别表示基于动作置信度权重的集成方法和基于最优动作投票的集成方法。

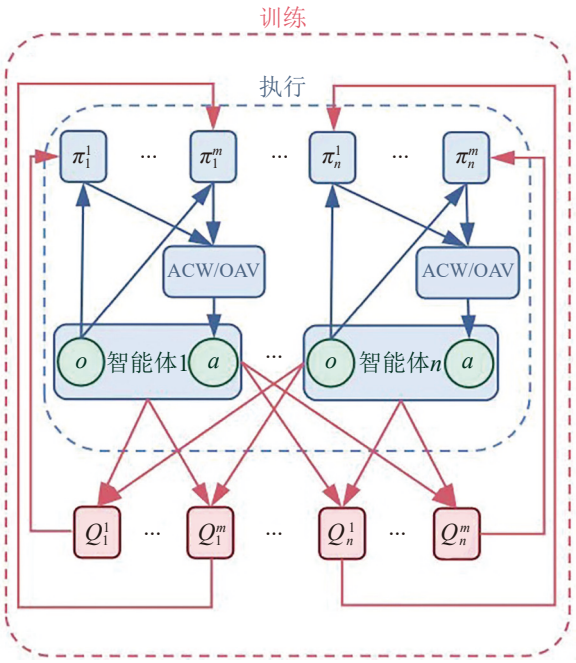

图 2　EDO-MADDPG 结构

在 QMIX 算法中, 每个智能体都有一个个体 Q 网络, 智能体依据自己的个体 Q 网络输出的个体 Q 值进行决策。个体 Q 网络还用于近似最大全局 Q 值, 最大全局 Q 值等于所有最大个体 Q 值的单调非线性组合, 单调非线性组合通过一个混合网络来实现。当 EDO 模型在 QMIX 上实现时, EDO-QMIX 为每个智能体集成 m 个个体 Q 网络, 并为整个多智能体系统训练 m 个混合网络, 使得智能体的 m 个个体 Q 网络都独立进行训练, 以提升不同个体 Q 网络之间的多样性。EDO-QMIX 的结构如图 3 所示, 其中, Q^j_{tot}为多智能体系统的第j个全局 Q 网络。

3　实验与分析

实验中测试的方法有 ACW-MADDPG、OAV-MADDPG、ACW-QMIX、OAV-QMIX、MADDPG 和 QMIX 六种。其中: ACW-MADDPG 和 OAV-MADDPG 表示在 MADDPG 的基础框架上结合 EDO 方法, 并分别采用 ACW 和 OAV 集成方法; ACW-QMIX 和 OAV-QMIX 表示在 QMIX 的基础框架上结合 EDO 方法, 并分别采用 ACW 和 OAV 集成方法。实验环境为 corridor、物理欺骗和捕食者–猎物。

3.1　实验环境描述

corridor 是一种星际争霸多智能体挑战中被定义为超级困难的环境。corridor 中包含 6 个狂热者单位,

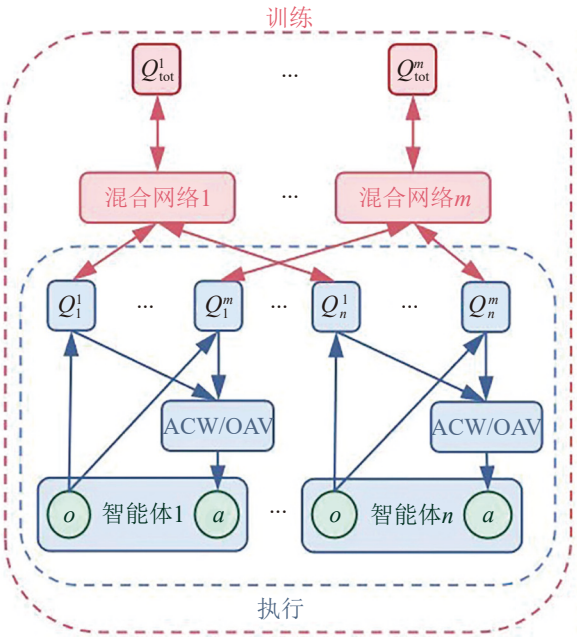

图 3　EDO-QMIX 的结构

每个狂热者单位均由一个智能体独立控制，敌方是由星际争霸Ⅱ内置脚本 AI 集中控制的 24 个跳虫单位。狂热者单位和跳虫单位的攻击方式均为近战攻击。每次环境初始化时，24 个跳虫单位总是紧挨着出现在特定位置。若跳虫单位的视野范围内无狂热者单位，则它们会向环境的某端前行，直到达到终点为止。由于在 corridor 中狂热者单位和跳虫单位都使用近战攻击，且跳虫单位的数量是狂热者单位的 6 倍，多智能体系统必须学会分散跳虫单位的火力，才有可能取得胜利。corridor 环境如图 4 所示。



图 4　Corridor 环境

在如图 5 所示的捕食者–猎物环境中，有 N 个移动速度较慢的合作智能体，需要在环境中追逐 M 个移动速度较快的猎物，环境中有 L 个大型路障挡路。每当合作智能体与猎物发生碰撞时，智能体将获得奖励，而猎物将受到惩罚。智能体能观察其他智能体及猎物的相对位置、速度和路障位置。每个回合，智能体、猎物和路障的初始位置都随机初始化。

物理欺诈环境如图 6 所示。在物理欺诈环境中设有 N 个地标，其中包括一个目标地标，有 M 个相互

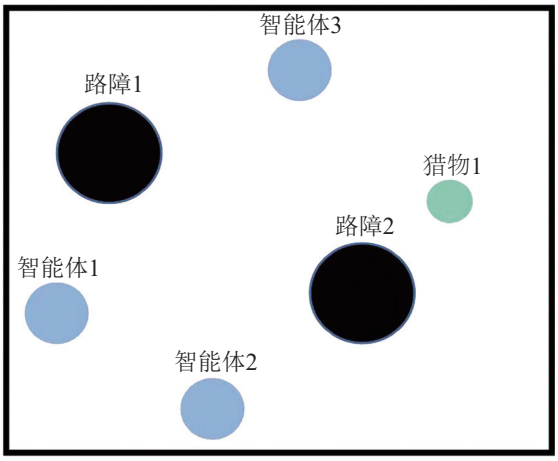

图 5　捕食者–猎物环境

合作的智能体，它们根据任意智能体到目标地标的最小距离获得奖励，距离目标地标越近，奖励越大。因此，只需保证一个智能体到达目标地标，即可获得最大奖励。同时，环境中还存在一个对手，它也希望到达目标地标，但不知道 N 个地标中哪个是目标地标。此外，合作的智能体会根据对手到目标地标的距离受到惩罚，因此，合作的智能体需要学习分散地覆盖所有地标，以欺骗对手。

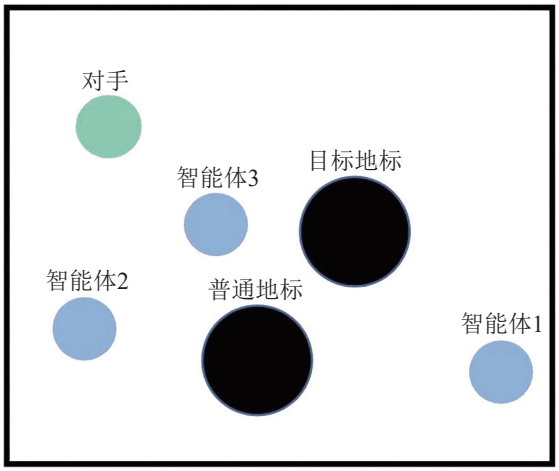

图 6　物理欺诈环境

3.2　实验结果分析

图 7 为在 corridor 实验环境下，QMIX、OAV-QMIX 和 ACW-QMIX 方法在整个训练过程中多智能体系统与星际争霸Ⅱ内置脚本 AI 对战的测试胜率曲线。从图 7 可看出，当 QMIX 方法分别应用了 EDO 的 2 种集成方法后，其在 corridor 实验环境中的测试胜率相较于原始的 QMIX 方法提高明显。在 corridor 实验环境的整个训练过程中，原始 QMIX 方法的测试胜率低于 25%，而 OAV-QMIX 和 ACW-QMIX

方法的胜率最高, 测试胜率均可达到 80% 以上。EDO 的应用使得原始 QMIX 方法的最高测试胜率提升了至少 50 个百分点。原始 QMIX 方法有时在 corridor 实验环境的整个训练过程中都无法使多智能体系统获胜, 而应用了 EDO 的 QMIX 方法则始终都能在训练后期使得多智能体系统获胜, 并逐渐提高胜率。整个训练过程中, 使用 ACW-QMIX 方法比使用 OAV-QMIX 方法能获得更高的平均测试胜率, 这表明 ACW 集成方法更适用于 QMIX 方法。

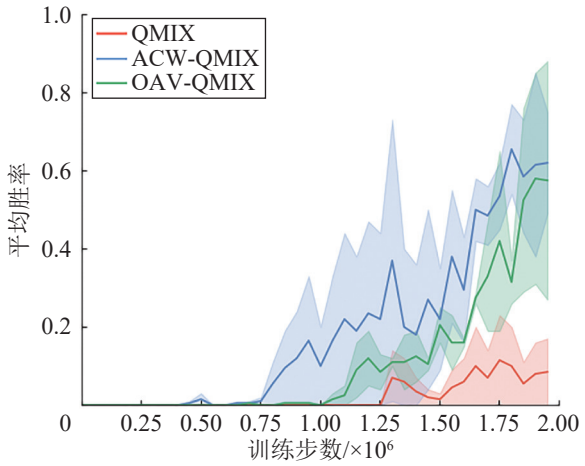

图 7　Corridor 环境平均胜率曲线

图 8、9 分别为在物理欺诈和捕食者–猎物实验环境中, 分别应用 MADDPG、OAV-MADDPG 和 ACW-MADDPG 方法, 多智能体系统获得的奖励值曲线。从图 8 可看出, 在物理欺诈实验环境中, OAV-MADDPG 和 ACW-MADDPG 方法在大部分训练阶段都取得了高于 MADDPG 方法的团队奖励。训练步数处于 100 000~1 500 000 区间时, 使用 EDO 带来的奖励值提升更明显, 且 OAV-MADDPG 和 ACW-MADDPG 方法的奖励值比 MADDPG 方法的奖励值更稳定, 说明 EDO 在该实验环境中提高了方法的稳定性。从图 9 可看出, 在捕食者–猎物实验环境中, OAV-MADDPG 和 ACW-MADDPG 方法从训练步数 100 000 开始便与 MADDPG 方法在奖励值上拉开差距, 并保持这种性能优势直到训练结束。这表明 EDO 在该实验环境中能有效提升多智能体系统的性能。OAV-MADDPG 和 ACW-MADDPG 在物理欺诈和捕食者–猎物实验环境中均取得了相似的奖励值, 表明 OAV 与 ACW 集成方法在 MADDPG 方法上的适用性相似。

实验结果表明, EDO 能够优化多智能体系统中智能体的决策, 并有效提升多智能体深度强化学习方法的性能和稳定性。EDO 的适用性较强, 在 2 类主流的多智能体深度强化学习方法中都适用。

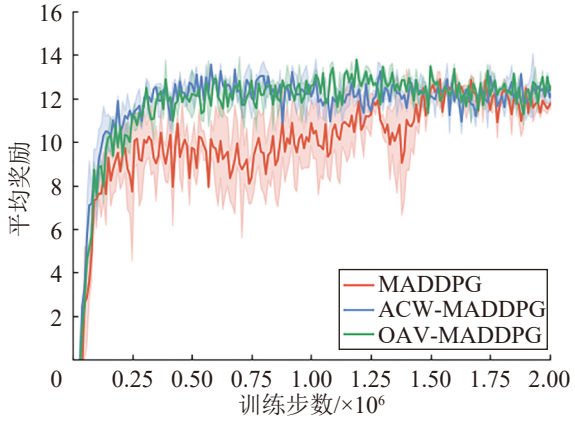

图 8　物理欺诈环境平均奖励曲线

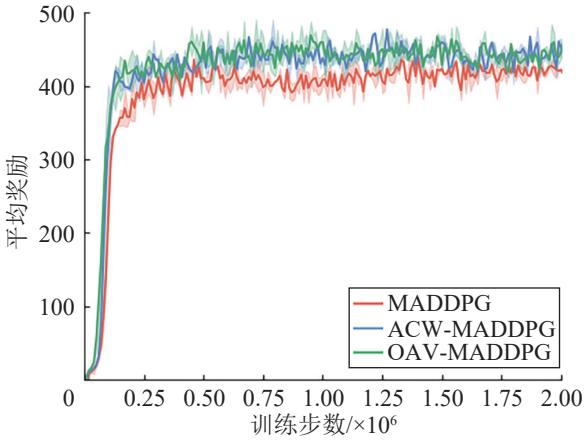

图 9　捕食者–猎物环境平均奖励曲线

4　结束语

针对多智能体深度强化学习中智能体的决策鲁棒性不足的问题, 提出了一种基于集成的多智能体深度强化学习决策优化方法(EDO), 其为每个智能体集成多个策略网络或动作值网络。EDO 还包含 2 种基于强化学习的特性设计的集成方法 ACW 和 OAV, 其中 ACW 在 QMIX 方法中的适用性更强。在 3 种不同的实验环境中进行了实验, 结果表明, EDO 能够优化智能体的决策, 并提升多智能体系统在这些实验环境中的性能和稳定性。

参考文献:

[1] Shalev-shwartz S, Shammah S, Shashua A. Safe, multi-agent, reinforcement learning for autonomous driving [PP/OL]. arXiv (2016-10-11)[2023-10-19]. https://doi.org/10.48550/arXiv.1610.03295.

[2] Jiménez A C, García-díaz V, Bolaños S. A decentralized framework for multi-agent robotic systems[J]. Sensors, 2018, 18(2): 417-436.

[3] Liu Yun, Wang Yu, Li Yuanzheng, et al. Multi-agent based optimal scheduling and trading for multi-micro-grids integrated with urban transportation networks[J]. IEEE Transactions on Power Systems, 2021, 36(3): 2197-2210.

[4] Coelho V N, Cohen M W, Coelho I M, et al. Multi-agent systems applied for energy systems integration: State-of-the-art applications and trends in microgrids[J]. Applied Energy, 2017, 187: 820-832.

[5] Zhou Quan, Li Yonggui, Niu Yingtao. Intelligent anti-jamming communication for wireless sensor networks: a multi-agent reinforcement learning approach[J]. IEEE Open Journal of the Communications Society, 2021, 2: 775-784.

[6] Gao Zhen, Yang Lei, Dai Yu. Large-scale computation offloading using a multi-agent reinforcement learning in heterogeneous multi-access edge computing[J]. IEEE Transactions on Mobile Computing, 2023, 22(6): 3425-3443.

[7] Chen Chacha, Wei Hua, Xu Nan, et al. Toward a thousand lights: decentralized deep reinforcement learning for large-scale traffic signal control[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC: AAAI Press, 2020, 34(4): 3414-3421.

[8] Lowe R, Wu Yi, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, NY: Curran Associates, 2017: 6382-6393.

[9] Foerster J, Farquhar G, Afouras T, et al. Counter factual multi-agent policy gradients[C]//Proceedings of the AAAI Conference on Artificial Intelligence.Washington, DC: AAAI Press, 2018, 32(1): 2974-2982.

[10] Ackermann J, Gabler V, Osa T, et al. Reducing overestimation bias in multi-agent domains using double centralized critics[PP/OL]. arXiv (2019-12-02)[2023-10-19]. https://doi.org/10.48550/arXiv.1910.01465.

[11] Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning[C]//International Conference on Machine Learning. New York, NY: PMLR Press, 2019: 2961-2970.

[12] Li Shihui, Wu Yi, Cui Xinyue, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient[C]//Proceedings of the AAAI Conference on Artificial Intelligence.Washington, DC: AAAI Press, 2019, 33(1): 4213-4220.

[13] Xiao Yuchen, Lyu X G, Amato C. Local advantage actor-critic for robust multi-agent deep reinforcement learning[C]//International Symposium on Multi-robot and Multi-agent Systems. Piscataway, NJ: IEEE Press, 2021: 155-163.

[14] Sunehag P, Lever G, Gruslys A, et al. Value decomposition networks for cooperative multi-Agent learning based on team reward[C]//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Richmond, VA: IFAAMAS, 2018: 2085-2087.

[15] Rashid T, Samvelyan M, De WITT C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. The Journal of Machine Learning Research, 2020, 21(1): 7234-7284.

[16] Son K, Kim D, Kang W J, et al. QTRAN: learning to facto- rize with transformation for cooperative multi-agent reinforcement learning[C]//International Conference on Machine Learning. New York, NY: PMLR Press, 2019: 5887-5896.

[17] Rashid T, Farquhar G, Peng B, et al. Weighted QMIX: expanding monotonic value function factorisation for deep multi-agent reinforcement learning[C]//Proceedings of the 33st International Conference on Neural Information Processing Systems. New York, NY: Curran Associates, 2020: 10199-10210.

[18] Wang Jianhao, Ren Zhizhou, Liu T, et al. QPLEX: duplex dueling multi-agent Q-learning[PP/OL]. arXiv (2021-10-04)[2023-10-19]. https://doi.org/10.48550/arXiv.2008.01062.

[19] Yang Yaodong, Hao Jianye, Liao Ben, et al. Qatten: a general framework for cooperative multiagent reinforcement learning[PP/OL]. arXiv(2020-06-09)[2023-10-19]. https://doi.org/10.48550arXiv.2002.03939.

编辑：张所滨