基于强化学习的多智能体泛化性研究

郭 鑫,王 微,青 伟,李 剑,何召锋
(北京邮电大学,北京 100088)

摘 要:在多智能体强化学习算法的研究中,由于训练与测试环境具有差异,如何让智能体有效地应对环境中其他智能体策略变化的情况受到研究人员的广泛关注。针对这一泛化性问题,提出基于人类偏好的多智能体角色策略集成算法,该算法同时考虑了长期回报和即时回报。这一改进使得智能体从一些具有良好长期累积回报的候选行动中选择具有最大即时回报的行动,从而让算法确定了策略更新的方向,避免过度探索和无效训练,能快速找到最优策略。此外,智能体被动态地划分为不同的角色,同角色智能体共享参数,不仅提高了效率,而且实现了多智能体算法的可扩展性。在多智能体粒子环境中与现有算法的比较表明,该算法的智能体能更好地泛化到未知环境,且收敛速度更快,能够更高效地训练出最优策略。

关键词:深度强化学习方法;多智能体;未知环境;策略集成;泛化性;可扩展性
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2023)04-0114-06
doi:10.3969/j.issn.1673-629X.2023.04.017

Research on Generalization of Multi-agent Based on Reinforcement Learning

GUO Xin, WANG Wei, QING Wei, LI Jian, HE Zhao-feng
(Beijing University of Posts and Telecommunications, Beijing 100088, China)

Abstract: In the research of multi-agent reinforcement learning algorithm, due to the difference between training and testing environment, how to make agents intelligently learn to cope with the performance degradation caused by the change of other agents' policy in the environment has been widely concerned by researchers. To solve this generalization problem, human-preference based multi-agent role policy ensemble is proposed, which considers the effects of long-term reward and immediate reward. This improvement enables the algorithm to determine the direction of policy updating to avoid excessive exploration and ineffective training. In addition, agents are classified into different roles according to their immediate rewards of historical actions. Thus the parameters are shared with the same-role agent, which improves efficiency and achieves the scalability of the multi-agent algorithm. The comparison with the existing algorithm in the multi-agent particle environment shows that the proposed algorithm has a faster convergence speed which can effectively train the optimal strategy, and its intelligence can better generalize to the unknown environment.

Key words: deep reinforcement learning; multi-agent; unknown environment; policy ensemble; generalization; scalability

0 引 言

从智能机器人[1],交通信号控制[2],动态频谱分配[3],到智能电网经济调度[4],实时竞价[5]等众多领域,多智能体强化学习算法都取得了不错的成果。但在复杂的场景下,使用强化学习的方法仍然具有挑战性。因为现实世界的复杂性,训练时和测试时的环境通常会发生一些变化,这样的环境要求智能体具有泛化到不可见状态的能力。近年来,在解决多智能体强化学习泛化性方面,学者们也做了大量研究。

深度确定性策略梯度算法(M3DDPG)[6]、值的极小化极大化算法(Qmixmax)[7]、Romax[8]等采用了极小极大化的思想考虑最坏情况并优化最小化目标。此外,文献[9]将基于核的强化学习与深度强化学习结合起来,文献[10]通过多数投票决定每个智能体的行动,而忽略了智能体间的交互。文献[11]中多智能体深度确定性策略算法(MADDPG)随机使用训练的多个策略网络。但是这些方法计算量太大,局限性强,没有发挥不同策略的优势。

收稿日期:2022-07-17 **修回日期**:2022-11-17
基金项目:国家自然科学基金(62176025,62076232);中央高校基本科研业务费专项资金资助(2021RC38,2021RC39)
作者简介:郭 鑫(1998-),女,硕士研究生,研究方向为多智能体强化学习;通讯作者:何召锋(1982-),男,教授,研究方向为博弈决策与群体智能。

该文关注到集成方法在泛化性方面的优越性能,提出基于人类偏好的角色策略集成方法。主要贡献在于:

(1)将多智能体强化学习与集成方法结合,使得智能体学习多种策略,在不同情况应用不同的策略,充分发挥各策略的优势,有效提升了智能体的泛化性。

(2)结合生命史理论和集成算法,综合考虑短期利益和长期回报,解决了训练中过度探索和无效训练的问题,使得智能体能快速找到最优策略。

(3)提出角色参数共享机制,相比于传统的全部智能体共享参数,创新在于为智能体动态地划分角色,同角色的智能体共享参数。该方法不仅提高了效率,而且实现了多智能体算法的可扩展性。

(4)在包含合作、竞争和混合场景的多智能体粒子环境中,实验结果表明该算法相比现有多智能体强化学习算法在泛化能力和训练速度方面都有很大的提升。

1　多智能体强化学习相关工作

1.1　多智能体强化学习

通常,多智能体强化学习过程可以表示为马尔可夫博弈[12],记作 $(N,S,O_1,\cdots,O_N,A_1,\cdots,A_N,\Gamma,R_1,\cdots,R_N)$,包括智能体集合 $i\in N=\{1,2,\cdots,N\}$、状态空间 S、联合观测空间 O,以及联合动作空间 A。智能体 i 仅可得到关于当前状态 S 的局部观测 O_i。转移函数 Γ 定义了基于当前状态和动作所能到达的下一状态的分布 $S\times A\to S'$;而 $S\times A\times S'\to R$ 定义了智能体 i 在 t 时刻接收到的奖赏函数。算法的学习目标是找到一组使每个智能体均可获得最大累计奖赏 R_i 的策略 $\pi=(\pi_1,\pi_2,\cdots,\pi_N)$,$\gamma$ 是折扣因子,T 是一个执行周期的步长。

1.2　集成算法

在集成算法中,多个策略被训练,然后智能体从它们集成的多个候选动作中选择动作。集成方法起源于单智能体强化学习,例如利用 Q 值函数、网络结构和策略的集成。文献[13]提出了各种聚合动作的方法来帮助智能体选择更好的行动,如多数投票、加权投票、等级投票、波尔兹曼乘法等。Bootstrapped DQN[14]、UCB[15]、DBDDPG[16]等通过集成 Q 值使智能体拥有更好的动作。之后,集成的方法不断发展,在多智能体强化学习中也开展了大量研究。文献[9]将基于核的强化学习与深度强化学习结合起来,整合了两种算法的优点。文献[10]通过智能体的多数投票决定智能体的行动。文献[11]使用多个行动-评估(actor-critic)结构,并在每个事件中随机选择其中之一。文献[17-19]也将集成方法应用在多智能体系统中。

2　基于人类偏好的多智能体角色策略集成算法

本节首先介绍了基于人类偏好的策略集成方法,然后根据肯德尔系数,介绍了角色共享参数机制的细节。该方法的概述如图 1 所示。

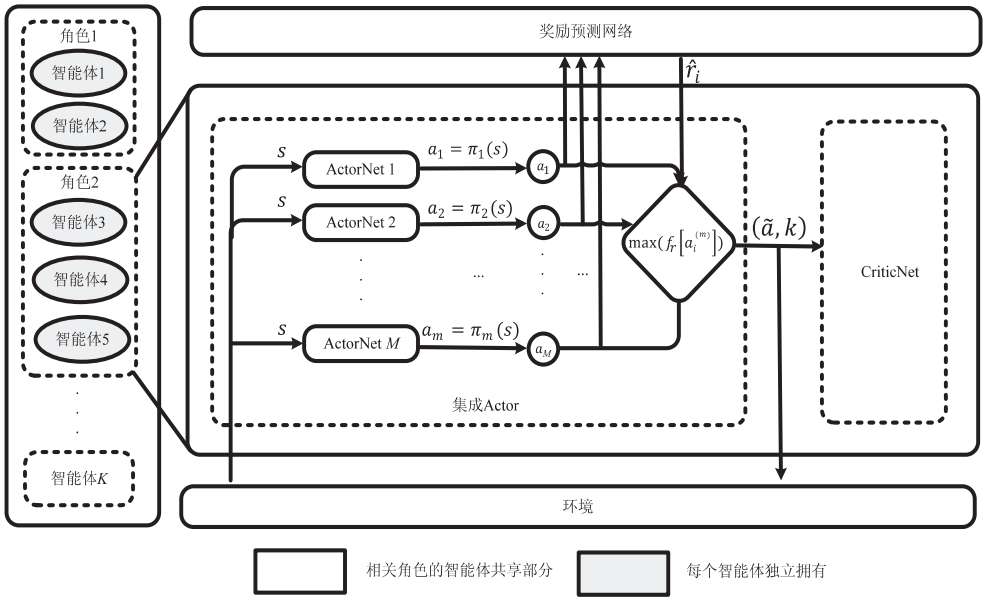


图 1　基于人类偏好的多智能体角色策略集成方法

2.1　基于人类偏好的策略集成方法

集成学习通过结合多个策略完成学习任务。传统的集成学习方法通过随机、大多数投票,均值等机制获得唯一的结果,但这些方法都没有发挥每个策略的优势。在这部分,将介绍如何发挥每个策略的优势使得训练更快,效果更好。

首先,智能体学习 M 个策略,因此,N 个智能体在同一环境中总共学习 $N*M$ 个策略,网络参数为:$\theta_i^{(1)}$,$\cdots,\theta_i^{(M)},\cdots,\theta_N^{(M)}$。在每个时间步 t,智能体 i 通过局部观测 O_i 利用 M 个策略计算长期回报,依据最大的长期回报,得到 M 个候选动作,如下式所示:

$$a_i^{(m)} = \pi_{\theta_i^{(m)}}(O_i^t) \tag{1}$$

依据生命史理论[20-21],考虑到即时奖励的重要性,从这 M 个较大的长期回报内选择具有最大的即时奖励的动作,如下式所示:

$$\widetilde{a_i}, k_i = \text{argmax}(r(a_i^{(m)})), m = 0,1,\cdots,M \tag{2}$$

其中,k_i 为智能体选择的动作对应的策略的下标。之后,在更新阶段,智能体的 M 套策略只需要随着每次的选择更新 k_i 策略,如下式:

$$\nabla_{\theta_i^{(k)}} J(\pi_i^k) = E[\nabla_{\theta_i^{(k)}} \log(\pi_{\theta_i^{(k)}})(-\alpha\log(\pi_{\theta_i^{(k)}})) + A_i^{\varphi}] \tag{3}$$

其中,A_i^{φ} 为优势函数,帮助解决多智能体信用分配问题。其计算公式如下所示:

$$A_i^{\varphi}(o_i,a_i) = Q_i^{\varphi}(o_i,a_i) - b(o_i,a_{\setminus i}) \tag{4}$$

由于训练阶段可以随意取得每一个动作的即时奖励,但测试阶段无法轻易获得即时奖励,为了解决这一问题,该文提出了一个辅助预测奖励网络。该网络为多层感知器网络,通过训练阶段的即时奖励训练网络对于每一个动作的预测值。其损失函数如下式所示:

$$L_r = E_{(o,a,r,k,o')\sim D}[(\hat{r}_i - r_i)^2] \tag{5}$$

其中,r_i 为真实的奖励值,$\hat{r}_i$ 为预测的奖励值。通过这个奖励预测网络来帮助智能体在测试阶段根据不同的情况选择具有最大即时奖励的行动,即使遇到了与训练阶段差别非常大的情况,也可以根据此时动作的即时奖励,选择较好的动作执行,保证性能不会急剧下降。

2.2 角色参数共享机制

针对智能体学习多个策略引发的计算消耗大的问题,本算法引入角色共享机制解决该问题。角色被理解为接受信息、加工信息和发送信息的抽象对象。角色参数共享机制认为:角色是按照策略执行行为的统一体,所以将策略相近的智能体划分为同一角色。在这部分,将介绍如何利用肯德尔系数,衡量智能体之间的相关性,然后智能体动态地划分为不同的角色,同角色的智能体学习和使用同一套策略。

智能体间的相关性可通过分析它们的即时奖励的序列的相关性得到。例如:智能体 i 和智能体 k 的即时奖励的序列如下式:

$$\begin{aligned} x &= \{r_i^t, 1 \leqslant t \leqslant T\} \\ y &= \{r_k^t, 1 \leqslant t \leqslant T\} \end{aligned} \tag{6}$$

肯德尔系数的计算公式如下式所示:

$$T_{(i,k)} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{7}$$

通过该公式反映出智能体 i 和智能体 k 的相关性,其中,n_c 表示同序对的个数,n_d 表示异序对的个数,$\frac{1}{2}n(n-1)$ 为总数。肯德尔系数值的范围为 [-1,1],负值表示两个智能体呈现敌对情况,正值表示两个智能体有一定的相似性,值越大,相关性越强。对智能体的分类,该文有如下两个定义。

定义 1 强相关:设定阈值 G,如果两个智能体的肯德尔系数值大于阈值 G,则定义这两个智能体具有强相关性。

定义 2 最大相关:对于智能体 i,计算相关性。通过最大值 $T_{(i,k)}$ 得到智能体 i 和智能体 k 有最大相关性。

如果两个智能体的肯德尔系数值同时满足定义 1 和定义 2,则这两个智能体被划分为同一角色,学习和使用同一套策略。

2.3 基于人类偏好的多智能体角色策略集成算法

在基于人类偏好的多智能体角色策略集成算法中,训练阶段,同角色的智能体共同训练 M 个策略网络以及一个辅助预测奖励网络。更新阶段,只更新执行动作的策略网络。在测试评估阶段,由于集中训练和分散式执行,通过辅助预测奖励网络选择动作。算法流程具体如算法 1 所示。

算法 1 基于人类偏好的多智能体角色策略集成(HPMARPE)算法。

初始化:具有 N 个智能体的 E 个并行环境,经验回放缓冲区 D,策略网络 π,辅助预测奖励网络 f。

(1) for episode = 1to N,进行迭代:

(2) for t = 1 to T,进行迭代:

(3) for m = 1 to M:

(4)获得动作:$a_i^{(m)} = \pi_{\theta_i^{(m)}}(\cdot \mid o_i')$

(5)预测奖励:$\hat{r}_i = f_i(o_i, a_i^{(m)})$

(6)结束

(7)选择动作和策略

(8)执行行动,智能体得到奖励 r,同时转移到新状态 s_{t+1}

(9)将 (o,a,r,k,o') 存入经验回放缓冲区

(10) for agent = 1 to n,进行迭代:

(11)采样经验 (o,a,r,k,o')

(12)更新策略网络

(13)通过最小化损失 L_r 更新网络 f

(14)结束

(15)更新目标网络:

$$\theta_i^k = \tau\theta_i^k + (1-\tau)\theta_i^k$$

$$\varphi_i^k = \tau\varphi_i^k + (1-\tau)\varphi_i^k$$

(16)结束

(17)结束

3　实　验

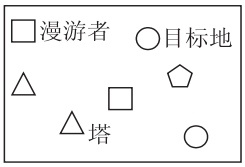
3.1　实验环境

该文采用的是多智能体粒子环境[11]。在合作、竞争、混合场景下分别进行了实验,实验用到的场景如图 2 所示。

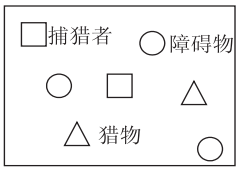
图 2(a)为漫游者-塔的合作场景。该任务包含 $N+M$ 个智能体, L 个目标地,其中包含 N 个漫游者, M 个塔。漫游者和塔随机配对,漫游者的目标是到达正确的目的地,然而不知道目的地的位置,配对的塔指挥漫游者到达指定的目标地,依据他们与正确的目标地的距离获得奖励。

图 2(b)为捕食者-猎物的竞争场景。在该场景下,速度较快的捕食者追逐速度较慢的猎物。每当捕食者追逐到猎物时,猎物受到惩罚,捕食者获得奖励。

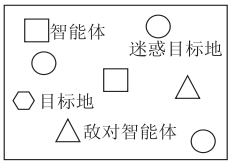
图 2(c)为物理欺骗的混合场景。在该场景中, N 个智能体相互合作对抗敌对智能体,他们的目标都是到达目的地,但敌对智能体不知道目的地在哪。合作的智能体中只要有一个到达目的地,则所有智能体获得奖励,敌对智能体受到惩罚。



(a)漫游者-塔



(b)捕食者-猎物



(c)物理欺骗

图 2　多智能体粒子环境

3.2　对比算法

为了验证所提的基于人类偏好的多智能体角色策略集成(HPMARPE)模型的泛化性和可扩展性,该文与 DIRS-V[10]、MADDPG-S[11]、MADDPG-R[11]进行比较。

(1)DIRS-V:在该方法中,每个智能体通过所有智能体的多数投票决定每个智能体的行动。

(2)MADDPG-S:该方法为多智能体深度确定性策略算法,每个智能体仅训练一个策略。

(3)MADDPG-R:针对智能体的泛化性问题,该方法同样提出策略集合的思想,每个智能体训练多个策略,优化策略集合的整体效果,在测试时,随机选用其中的子策略。

3.3　结果分析

为了验证算法的有效性,本节主要从泛化性、训练效率、可扩展性来对比评估算法的优越性。

(1)智能体的泛化性。

为了验证该方法训练的智能体的泛化性,将训练多次得到的智能体的策略随机组合进行测试,以此来评估智能体在面对其他智能体新的策略时的性能。竞争、合作、混合场景下实验结果分别展示在表 1 ~ 表 3 中。表 1 展示了猎物被捕捉的次数,从表中可看出,经过 HPMARPE 训练的智能体对抗使用单一的策略的敌对智能体时,整体取得了最好的效果。表 2 展示了合作场景下的漫游者距离目的地的距离,距离越小,效果越好。从表中可知,合作的智能体都使用基于人类偏好的多智能体角色策略集成算法(HPMARPE)时,智能体配合良好,取得了最好的效果。这是符合预期的结果。表 3 显示的是复杂的混合场景下智能体和敌对智能体的成功率,差值越大,效果越好。与竞争和合作环境下的实验结果相一致,在该场景,合作的智能体使用 HPMARPE 对抗单一策略的智能体会取得最好的效果。

表 1　捕食者-猎物场景下的测试结果

捕食者	猎物	捕捉次数
MADDPG-S	MADDPG-S	13.19
MADDPG-R	MADDPG-S	14.15
DIRS-V	MADDPG-S	15.31
HPMARPE	MADDPG-S	21.41
HPMARPE	DIRS-V	18.89
HPMARPE	MADDPG-R	19.99
HPMARPE	HPMARPE	18.95

表 2　合作场景下的测试结果

漫游者	塔	成功率/%	平均距离/m
MADDPG-S	MADDPG-S	52.16	0.23
MADDPG-R	MADDPG-S	54.79	0.21
DIRS-V	MADDPG-S	56.15	0.20
HPMARPE	MADDPG-S	65.41	0.17
HPMARPE	MADDPG-R	66.29	0.17
HPMARPE	DIRS-V	65.26	0.17
HPMARPE	HPMARPE	77.43	0.15
DIRS-V	HPMARPE	59.57	0.21
MADDPG-R	HPMARPE	57.78	0.21
MADDPG-S	HPMARPE	67.67	0.17

表 3 混合场景下的测试结果

智能体	成功率/%	敌对智能体	成功率/%	差值/%
MADDPG-S	88.50	MADDPG-S	74.13	14.37
MADDPG-R	78.24	MADDPG-S	37.45	40.79
DIRS-V	82.29	MADDPG-S	34.65	47.64
HPMARPE	88.34	MADDPG-S	18.90	69.44
HPMARPE	89.06	MADDPG-R	18.57	60.49
HPMARPE	83.65	DIRS-V	28.28	65.37
HPMARPE	98.06	HPMARPE	41.09	56.97
DIRS-V	52.23	HPMARPE	25.29	27.04
MADDPG-R	54.22	HPMARPE	24.69	29.53

综上,使用集成策略的智能体优于使用单一策略的智能体。然而,使用基于人类偏好的策略集成方法的智能体又显著优于使用其他策略集成的智能体。

(2)角色参数共享方法的性能。

图 3 为算法中不同的角色参数共享下的训练结果。阈值 G 对智能体的角色划分起着重要作用。为了研究该方法在不同的阈值 G 下的影响,分别设 $G=0$、0.3、0.5、0.7 和 1.0。$G=1.0$ 意味着每个智能体都是一个单独的角色,这时算法为无共享机制,所有智能体学习自己的多个策略。而 $G=0$ 表示所有智能体同属于一个角色,共享同一套策略。从图 3 可以看出,$G=0.7$ 时算法在图中取得了最佳性能。事实上,过小的阈值 G 将一些弱相关的智能体归入同一角色,这使得智能体共享一些模糊的策略,因此妨碍了他们的学习。过大的阈值 G 会把智能体分成多个角色,每个智能体训练数据单一,这使得它很难达到最优性能。上述的实验结果表明采用一个合适的阈值 G,一方面会避免无效训练,另一方面增加了数据的多样性,从而算法性能获得提升,因此需要适当的阈值 G 将智能体进行分类。

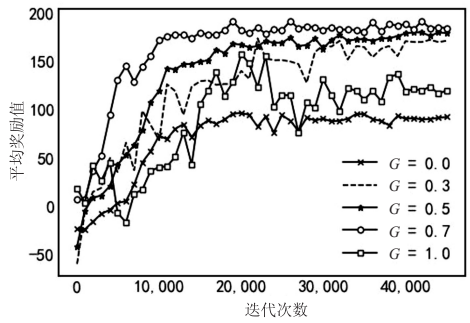


图 3 角色划分的影响

(3)训练效率。

此外,角色参数共享机制还影响了算法的训练效率,从图 3 不同阈值下算法的收敛变化比较中可以看出,$G=0.7$ 时算法最先达到收敛,大概在 10 000 迭代次数左右,$G=0.5$ 时算法在 20 000 迭代次数左右收敛,其余阈值下,20 000 步之后仍然未达到收敛。这是因为,合适的阈值下,同角色下的不同智能体共同学习和训练,使得学习效率大大提高,从而可以在更少的迭代次数下收敛。该实验结果验证了角色参数共享机制可以有效提升训练效率。

(4)可扩展性。

为了检验角色参数共享方法在可扩展性方面的影响,在训练环境上学习模型,在智能体数量与训练环境不同的场景上对学习到的策略进行测试,没有额外的再训练,以此来确定该方法的可扩展性。表 4 显示了在合作场景中的实验结果,表中 $N*N$ 代表环境中的智能体数量。实验结果表明:当测试环境中智能体的数量增加时,该方法(HPMARPE)显著优于 MADDPG,其成功率只有小幅波动,而 MADDPG 方法训练的智能体成功率大幅下降,且智能体数量增加越多,性能下降越多。这是因为,在该方法中,当环境中新增智能体时,该智能体依据自己的目标和动作的奖励值确定所属角色,从而可以采用该角色在训练环境下学习与训练的多套策略,在不同情况下做出不同的应对,以此来保持性能稳定。

表 4 合作场景下的可扩展性测试结果

训练	测试	成功率(MADDPG)/%	成功率(HPMARPE)/%
4*4	3*3	51.08	76.32
4*4	4*4	55.73	77.82
4*4	6*6	32.89	71.01
4*4	8*8	28.01	70.15
5*5	3*3	50.08	75.78
5*5	5*5	55.78	77.34
5*5	6*6	38.89	71.23
5*5	8*8	30.01	70.80
5*5	10*10	17.08	70.24

(5)超参数的影响。

为进一步评估策略集成方法的影响,图 4 显示了

在合作场景中不同数量的策略进行集成的实验结果。K 是策略的数量,不同数量的策略使智能体有不同规模的行动集。从结果中可以看出,设置过大的 K 对性能的影响很小。此外,如果策略的数量大于动作空间的尺寸,它甚至会导致性能下降。

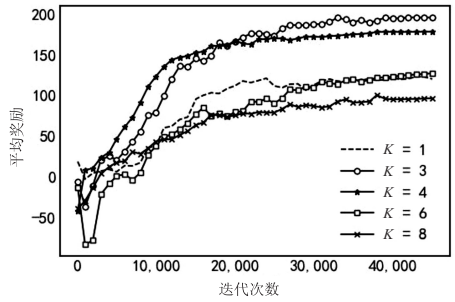


图 4　合作场景中策略数量的影响

4　结束语

该文研究了多智能体强化学习的泛化性问题,提出了一种基于人类偏好的多智能体角色策略集成算法。该方法首先使用基于人类偏好的策略集成的思路,综合考虑即时奖励和长期回报,解决了智能体策略改变引发的泛化性问题,同时提出角色参数共享机制,根据历史行动的即时奖励将智能体动态分为不同的角色,智能体按角色共享参数,减少计算资源的同时实现了可扩展性。该算法在多智能体粒子环境的多个场景上进行了实验,与其他现有的方法相比,在泛化性和训练效率上,都有极大的提高。

参考文献:

[1] 李　珣,南恺恺,赵征凡,等. 多智能体博弈的纺织车间搬运机器人任务分配[J]. 纺织学报,2020,41(7):78-87.

[2] 钱　锦. 基于多智能体协同的区域交通信号控制策略研究与实现[D]. 扬州:扬州大学,2021.

[3] 童　乐,梁　涛,张　余,等. 基于多智能体强化学习的动态频谱分配方法[J]. 太赫兹科学与电子信息学报,2021,19(4):573-580.

[4] 朱舒婷. 基于多智能体一致性的智能电网经济调度研究[D]. 舟山:浙江海洋大学,2021.

[5] ZHAO J,QIU G,GUAN Z,et al. Deep reinforcement learning for sponsored search real-time bidding[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. London:ACM,2018:1021-1030.

[6] LI S,WU Y,CUI X,et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient[C]//Proceedings of the AAAI conference on artificial intelligence. Hawaii:AAAI,2019:4213-4220.

[7] SUN C,KIM D K,HOW J P. ROMAX:certifiably robust deep multiagent reinforcement learning via convex relaxation[C]//2022 international conference on robotics and automation (ICRA). Pennsylvania:IEEE,2022:5503-5510.

[8] PHAN T,GABOR T,SEDLMEIER A,et al. Learning and testing resilience in cooperative multi-agent systems[C]//Proceedings of the 19th international conference on autonomous agents and multi agent systems. Auckland:IFAAMAS,2020:1055-1063.

[9] GHOSH S,LAGUNA S,LIM S H,et al. A deep ensemble multi-agent reinforcement learning approach for air traffic control[J]. arXiv:2004.01387,2020.

[10] JIANG F,DONG L,WANG K,et al. Distributed resource scheduling for large-scale MEC systems:a multiagent ensemble deep reinforcement learning with imitation acceleration[J]. IEEE Internet of Things Journal,2021,9(9):6597-6610.

[11] LOWE R,WU Y,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. Long Beach:NIPS,2017:6382-6393.

[12] TERRY J K,GRAMMEL N,HARI A,et al. Revisiting parameter sharing in multi-agent deep reinforcement learning[J]. arXiv:2005.13625,2020.

[13] ISHWARYA M,CHERUKURI A K. Quantum-inspired ensemble approach to multi-attributed and multi-agent decision-making[J]. Applied Soft Computing,2021,106:107283.

[14] FAUER S,SCHWENKER F. Ensemble methods for reinforcement learning with function approximation[C]//International workshop on multiple classifier systems. Naples:Springer:2011:56-65.

[15] CHEN R Y,SIDOR S,ABBEEL P,et al. Ucb exploration via q-ensembles[J]. arXiv:1706.01502,2017.

[16] ZHENG Z,YUAN C,LIN C,et al. Self-adaptive double bootstrapped DDPG[C]//Proceedings of the 27th international joint conference on artificial intelligence. Stockholm:IJCAI,2018:3198-3204.

[17] 余　勇. 基于多智能体样本交换的分散式集成学习方法研究[D]. 重庆:西南大学,2020.

[18] 牛礼民,杨洪源,周亚洲,等. 混合动力汽车动力总成多智能体集成控制策略[J]. 机械工程学报,2019,55(12):168-177.

[19] 王　薇. 基于多智能体技术下变电站设备信息集成的分析[J]. 科技创新与应用,2018,29:154-155.

[20] BRUMBACH B H,FIGUEREDO A J,ELLIS B J. Effects of harsh and unpredictable environments in adolescence on development of life history strategies[J]. Human Nature,2009,20(1):25-51.

[21] BIRKÁS B,CSATHÓ Á,GÁCS B,et al. Nothing ventured nothing gained:strong associations between reward sensitivity and two measures of Machiavellianism[J]. Personality and Individual Differences,2015,74:112-115.