控制与决策
Control and Decision

融合认知行为模型的深度强化学习框架及算法

陈浩, 李嘉祥, 黄健, 王菖, 刘权, 张中杰

引用本文:
陈浩,李嘉祥,黄健,王菖,刘权,张中杰. 融合认知行为模型的深度强化学习框架及算法[J]. 控制与决策, 2023, 38(11): 3209–3218.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0281

融合认知行为模型的深度强化学习框架及算法

陈 浩, 李嘉祥, 黄 健†, 王 菖, 刘 权, 张中杰

(国防科技大学 智能科学学院，长沙 410073)

摘 要: 面对高维连续状态空间或稀疏奖励等复杂任务时,仅依靠深度强化学习算法从零学习最优策略十分困难,如何将已有知识表示为人与学习型智能体之间相互可理解的形式,并有效地加速策略收敛仍是一个难题. 对此,提出一种融合认知行为模型的深度强化学习框架,将领域内先验知识建模为基于信念-愿望-意图 (belief-desire-intention, BDI) 的认知行为模型,用于引导智能体策略学习. 基于此框架,分别提出融合认知行为模型的深度Q学习算法和近端策略优化算法,并定量化设计认知行为模型对智能体策略更新的引导方式. 最后,通过典型gym环境和空战机动决策对抗环境,验证所提出算法可以高效利用认知行为模型加速策略学习,有效缓解状态空间巨大和环境奖励稀疏的影响.

关键词: 认知行为模型；强化学习；近端策略优化；深度Q网络；信念-愿望-意图；GOAL；空战机动决策

中图分类号: TP183 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0281

引用格式: 陈浩,李嘉祥,黄健,等. 融合认知行为模型的深度强化学习框架及算法 [J]. 控制与决策, 2023, 38(11): 3209-3218.

Deep reinforcement learning framework and algorithms integrated with cognitive behavior models

CHEN Hao, LI Jia-xiang, HUANG Jian†, WANG Chang, LIU Quan, ZHANG Zhong-jie

(College of Intelligence Science and Technology，National University of Defense Technology，Changsha 410073，China)

Abstract: When facing complex tasks with high-dimensional continuous state-space or sparse rewards, it is difficult for a reinforcement learning agent to learn an optimal policy from scratch. How to represent the known knowledge in a form understandable by human beings and the learning agent, and effectively accelerate policy convergence is still a difficult problem. Therefore, this paper proposes a deep reinforcement learning (DRL) framework integrating with cognitive behavior models. It represents prior knowledge as belief-desire-intention (BDI) based cognitive behavior models, which are used to guide policy learning in the DRL. Besides, we introduce the deep Q-learning algorithm with the cognitive behavior model (COG-DQN) and the proximal policy optimization algorithm with the cognitive behavior model (COG-PPO) based on the proposed framework. Moreover, we quantitatively design the guidance strategies of the cognitive behavior model to policy update. Finally, in a typical gym environment and an air combat maneuver confrontation environment, we verify that the proposed algorithms can efficiently use the cognitive behavior model to accelerate policy learning, and significantly alleviate the impact of high-dimensional state-space and sparse rewards.

Keywords: cognitive behavior mode；reinforcement learning；PPO；DQN；BDI；GOAL；air combat maneuver

0 引 言

近年来,深度强化学习 (deep reinforcement learning, DRL) 在即时战略游戏、机器人控制、能源分配等领域取得了瞩目成绩. 然而,巨大的样本复杂度 (sample complexity)[1] 限制了DRL在复杂实际问题 (如,高维连续状态空间、稀疏奖励的环境) 中的应用. 具体而言,智能体将强化学习任务建模为马尔可夫决策过程 (Markov decision process, MDP)[2],通过与环境充分试错交互,获取大量数据进行离线或在线学习,即使简单的任务也需要数以万计的数据,当状态空间巨大或环境奖励稀疏时,从零学习最优策略将十分困难. 相比之下,人类在面对复杂任务时都是在已有知

收稿日期: 2022-02-22；**录用日期:** 2022-06-24.

基金项目: 国家自然科学基金项目 (61906202).

责任编委: 侯忠生.

†通讯作者. E-mail: nudtjhuang@hotmail.com.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

识的基础上再学习. 这些知识作为初始策略可能并不完全适用于新任务,但有效避免了冷启动和学习初期非必要的探索(exploration)过程. 因此,将已有领域内知识或学习模型引入智能体在线学习,是加速学习进程、缓解上述问题的有效手段[3-4].

现有研究中,DRL对先验知识的使用大致可以分为直接策略重用和间接知识利用. 直接策略重用不涉及新策略学习,智能体维护了一个最优策略库,并在执行过程中不断检测任务是否发生变化,根据任务类型直接重用已有DRL策略模型[5-6]. 在间接知识利用中:一个研究分支是通过计算不同任务MDP的相似程度[7-8]或通过状态空间映射[9]实现已学习到的模型从源任务迁移至目标任务;另一个研究分支是选取合适的先验知识引导智能体从零学起[10-11]. 然而,这些算法大多针对具体任务设计静态的知识模型,人与智能体互相理解能力较差,知识表示方式不统一,针对不同任务的可迁移性和泛化能力不足. 另外,人与神经网络之间的认知方式存在较大差异,自然语言描述的知识或符号化的规则难以被DRL网络理解.

基于信念-愿望-意图(belief-desire-intention, BDI)[12]的智能体建模以意识系统和实践推理为理论基础,将感知、信念、目标、意图和领域知识等要素建模,并在此基础上构建推理系统. 与DRL模糊且难以形式化的策略模型相比,基于BDI智能体的决策过程具有更好的可解释性和可验证性[13],可靠性也更高,易于构建可信赖的智能系统. 然而,随着任务复杂程度的提升,制约基于BDI智能体应用的主要问题有以下几点:1)建模需要精确的领域内知识;2)需要强大的推理引擎,根据当前的信念和事件选择合适的动作序列执行;3)需要保证关键的非功能性需求,例如智能体响应的及时性[14];4)策略模型不具有学习演化能力,智能体的性能取决于预先构建的规则.

本文整合DRL智能体与基于BDI智能体的优势,提出融合认知行为模型的深度强化学习框架及算法. 首先,使用基于BDI的GOAL认知行为建模框架[15],将领域内先验知识描述为人和智能体可以相互理解的符号化认知行为模型. 其次,为提升DRL智能体的学习效率,将认知行为模型引入DRL智能体与环境的交互回路,实现跨认知层次上知识的相互补充. 然后,基于此框架,本文分别针对基于值函数和基于策略梯度的强化学习算法,将认知行为模型引入深度Q学习[16]算法和近端策略优化(proximal policy optimization, PPO)[17]算法,并设计了认知行为模型对智能体策略学习的引导方式. 最后,本文在典型gym环境以及空战机动决策对抗环境下验证了所提出的算法可以高效利用认知行为模型加速策略学习,有效缓解状态空间巨大和环境奖励稀疏的影响.

1 相关背景

1.1 深度强化学习

强化学习任务可以形式化地描述为一个包含$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\rangle$四元组的MDP. 其中:$\mathcal{S}$为状态集合,$\mathcal{A}$为动作集合,$\mathcal{P}$为状态转移模型$\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$,$\mathcal{R}$为奖励函数$\mathcal{S}\times\mathcal{A}\to\boldsymbol{R}$. DRL智能体的目标是学习最优应对策略$\pi^\star$,从而最大化期望回报$R=\sum\limits_{t=0}^{T}\gamma^t r_t$. 其中:$\gamma\in[0,1)$为折扣因子,$T$为回合最大步长.

Q学习(Q-learning)[18]和DQN[16]是典型基于值函数的强化学习算法,二者通过优化动作值函数$Q^\pi=\boldsymbol{E}_\pi\Big[\sum\limits_{t=0}^{\infty}\gamma^t r_{t+1}|s_0=s,a_0=a\Big]$学习最优策略$\pi^\star$. DQN将Q学习与神经网络的泛化能力相结合,通过最小化下式来优化其价值网络:

$$L_Q(\theta)=\hat{\boldsymbol{E}}_{(s,a,r,s')\sim U(D)}[y-Q(s,a|\theta)]^2. \tag{1}$$

其中:$y=r+\gamma\max_{a'}Q'(s',a'|\theta')$,$U(D)$表示在经验池$D$中均匀采样,$\hat{\boldsymbol{E}}$表示批次样本的经验平均,$Q'$和$Q$分别代表目标网络和价值网络. 目标网络参数$\theta'$由价值网络参数$\theta$定期更新.

基于策略梯度的算法是处理强化学习任务的另一种方式,与基于值函数的方法相比,这类算法直接沿梯度$\nabla_\phi J(\phi)$优化参数为ϕ的策略π_ϕ,其中$J(\phi)=\hat{\boldsymbol{E}}_{s\sim P^\pi,a\sim\pi_\phi}[R]$. 以$Q$值函数代替$R$,则策略梯度可表示为

$$\nabla_\phi J(\phi)=\hat{\boldsymbol{E}}_{s\sim P^\pi,a\sim\pi_\phi}[\nabla_\phi\log\pi_\phi(a|s)Q^\pi(s,a)], \tag{2}$$

其中P^π为给定策略π后的状态分布.

1.2 基于BDI的GOAL认知行为模型

GOAL认知行为模型与环境交互的过程如图1所示. 知识集和信念集共同构建对当前环境的认知. 其中:知识集主要包含定义、逻辑、规定等领域内基本知识以及任务执行中不会改变的事实;而信念集用于动态表示当前环境的状态及智能体自身信息. 目标集用于表示想要达到的目标状态. 知识集、信念集和目标集均用Prolog语法[19]构建,三者共同构成了认知行为模型的心理状态(mental state). 动作集用STRIPS语法定义了智能体可执行的动作,行为规则集定义了选择动作的规则,即动作的触发条件,

其具体形式为: if⟨mental state condition⟩ then⟨action⟩.

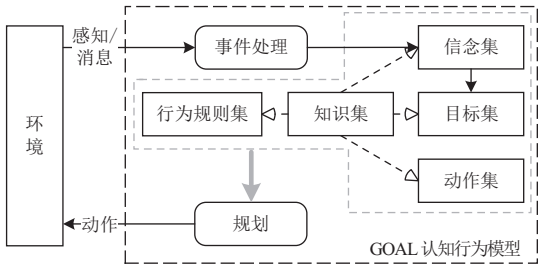


图 1　GOAL 认知行为模型与环境交互示意

在与环境交互的过程中,事件处理模块感知环境变化和当前可用消息,按照感知规则更新信念集和目标集,具体更新方式见文献[15, 20]. 然后,规划模块以信念集为依据,以目标集为导向,以行为规则集为约束,从动作集中选择动作并执行.

2　融合认知行为模型的深度强化学习框架

本文提出的融合GOAL认知行为模型的深度强化学习框架如图2所示,其由GOAL认知行为模型和深度强化学习模块构成. 智能体与环境交互过程中,GOAL认知行为模型感知环境状态信息,更新其信念集和目标集,并根据由先验知识构建的动作集和行为规则集作出决策,向深度强化学习模块输出认知行为知识,一般为推荐的(宏)动作. 与此同时,深度强化学习模块也从环境中获取当前状态信息,该模块结合DRL策略模型和认知行为知识,选择合适的动作执行,并从环境获得反馈信号用于策略更新.

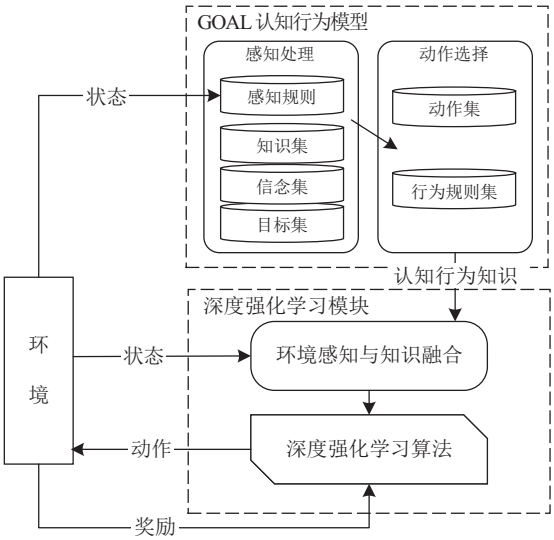


图 2　融合GOAL认知行为模型的深度强化学习框架

GOAL认知行为模型的主要作用是,在学习初期借助先验知识引导DRL策略优化,尽量减少其对状态空间的不必要探索,从而加速策略学习. 随着学习过程的推进,应逐渐减小对GOAL模型的依赖,转而主要使用DRL模型决策.

针对具体问题,本文提出GOAL认知行为模型的设计流程. 如图3所示,将具体任务背景分解为任务目标、环境描述和可用先验知识3个部分. 首先,将任务目标分解,用于构建目标集;然后,将环境描述的领域内定义和公理用于构建知识集,现有对于状态信息的描述用于构建信念集;最后,将先验知识中的可用动作用于构建动作集,状态-动作的规则映射关系用于构建行为规则集,对环境感知的处理方式用于构建感知规则.

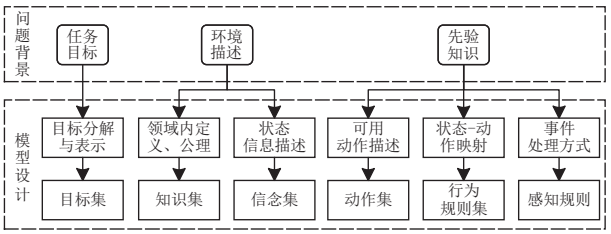


图 3　GOAL认知行为模型设计流程

3　融合GOAL认知行为模型的深度强化学习算法

3.1　融合认知行为模型的深度Q学习算法

融合认知行为模型的深度Q学习(deep Q-learning with the cognitive behavior model, COG-DQN)算法的结构如图4所示,强化学习模块由启发策略网络H、价值网络Q和目标网络Q'三部分构成,分别以θ_h、θ和θ'表示3个网络的参数. 其中,启发策略网络连接认知行为模型和强化学习算法,将认知行为知识转化为强化学习可用的引导策略. 启发策略网络接收环境状态信息s和GOAL模型提供的认知行为知识,以监督学习的方式不断逼近GOAL认知行为模型,并与价值网络Q同步更新.

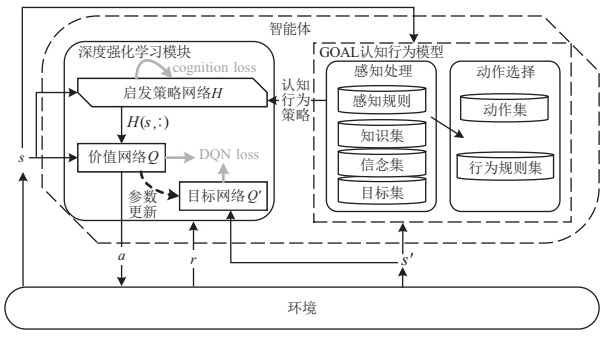


图 4　融合认知行为模型的深度Q学习算法

在与环境交互过程中,启发策略网络以启发值函数$H(s,:|\theta_h)$的方式引导价值网络Q动作选择. 以ε-greedy策略为例,COG-DQN选择动作的方式可表示为

$$\pi_{\theta,\theta_h}(s)=$$

$$\begin{cases}\arg\max\limits_{a\in\mathcal{A}}[Q(s,a|\theta)+\rho_h H(s,a|\theta_h)],\ \text{prob}\geqslant\varepsilon;\\ a_{\text{random}},\ \text{otherwise.}\end{cases}\tag{3}$$

其中: prob为在[0, 1]的均匀分布中随机采样的变量; $\varepsilon\in[0,1]$为探索率; $\rho_h\in\boldsymbol{R}$为启发权重,其衡量了认知行为知识对动作选择的影响程度. 对GOAL认知行为模型的置信度越高,其对动作选择的影响程度越大, ρ_h也越大. 为保证学习效率,在实际应用中,若对认知行为知识的置信度不高,则 ρ_h在学习过程中应以一定速率逐渐衰减,以避免知识误用.

以 a_{cog}表示GOAL认知行为模型推荐的动作,为引导智能体策略学习, $H(s,a_{\text{cog}}|\theta_h)$应为正值且大于所对应 Q值 $Q(s,:\ |\theta)$的波动. 另一方面,为了减小 $H(s,a_{\text{cog}}|\theta_h)$的使用对实际动作值函数估计造成的误差, $H(s,a_{\text{cog}}|\theta_h)$的变化应尽可能小. 因此,本文将启发策略网络 H的更新目标 $h(s,a)$形式化描述为

$$h(s,a)=\begin{cases}\max\limits_{a_i\in\mathcal{A}}Q(s,a_i|\theta)-Q(s,a|\theta)+\eta,\ a=a_{\text{cog}};\\ 0,\ \text{otherwise.}\end{cases}\tag{4}$$

其中 $\eta\in\boldsymbol{R}$决定了更新幅度. 于是启发策略网络 H的损失函数可表示为

$$L_H(\theta_h)=\hat{\boldsymbol{E}}_{(s,a,a_{\text{cog}},p)\sim U(D)}p[h(s,a)-H(s,a|\theta_h)]^2,\tag{5}$$

其中 p为标识符, $p=1$表示有可用的认知行为知识.

在策略更新阶段,启发策略网络 H与价值网络 Q用同一批采样数据更新,价值网络 Q的更新方式与DQN一致,如式(1)所示. 目标网络 Q'的参数 θ'由价值网络 Q的参数 θ以固定间隔替换更新. 需要注意的是,为了使学习过程更稳定,在更新启发策略网络 H时,应固定价值网络 Q的参数.

值得注意的是,GOAL认知行为模型仅在式(3)中参与了动作选择,而未显示参与价值网络 Q的更新. 尽管启发策略网络 H对动作值函数的估计产生了影响,但COG-DQN仍保留了 ε-greedy动作选择策略中的探索项,智能体可以实现对状态-动作空间的充分探索. 即GOAL认知行为模型的引入仅影响学习效率,而不影响算法的收敛性.

3.2 融合认知行为模型的近端策略优化算法

融合认知行为模型的近端策略优化(proximal policy optimization with the cognitive behavior model, COG-PPO)算法结构如图5所示,与COG-DQN不同的是,GOAL模型输出的认知行为知识直接参与强化学习模块的策略更新. 智能体与环境交互过程中,PPO记录GOAL模型推荐的(宏)动作,以自身DRL策略模型选择动作执行,并从环境中获得反馈信号. 然后将 $(s,a,r,s',a_{\text{cog}})$存入在线经验池 D. 在策略更新阶段,COG-PPO不仅优化式(2),还需要考虑当前DRL策略与认知行为策略之间的差距.

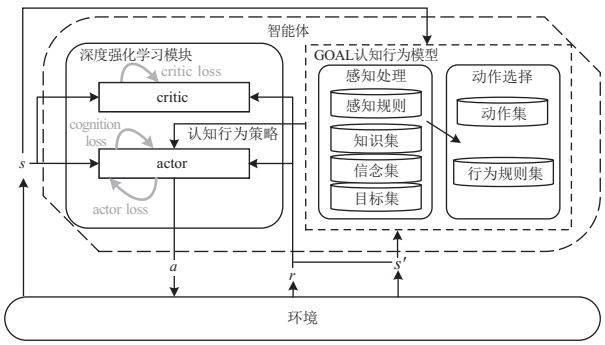


图5 融合认知行为模型的近端策略优化算法

分别用 ϕ和 ψ表示PPO中actor和critic网络的参数, $A_t=\sum\limits_{t'>t}\gamma^{t'-t}r_{t'}-V_\psi(s_t)$表示估计的优势函数,其中 $V_\psi(s_t)$为状态值函数. 则actor截断后的目标函数 $L^{\text{clip}}(\phi)$可表示为

$$L^{\text{clip}}(\phi)=\hat{\boldsymbol{E}}_t[\min(p_t(\phi)A_t,\text{clip}(p_t(\phi),1-\varepsilon,1+\varepsilon)A_t)].\tag{6}$$

其中: $p_t(\phi)=\dfrac{\pi_\phi(a_t|s_t)}{\pi_{\phi_{\text{old}}}(a_t|s_t)}$表示概率比, π_{old}表示更新前actor的参数.

考虑对于某状态,先验知识为宏动作的情况,此时,GOAL模型的输出为一系列原子动作的组合. 为有效利用认知行为知识,在学习初期,智能体在该状态时应尽量学习该宏动作. 为此,COG-PPO维护了一个宏动作队列 Ξ,当 Ξ为空时才从GOAL模型中获取认知行为知识;否则,应先将 Ξ中的原子动作逐步加入在线经验池 D. 其次,考虑到GOAL模型中的认知行为知识通常是不完备的,在当前状态有可用的认知行为知识时,智能体逐渐缩小当前DRL策略与认知行为策略之间的差距;当前状态没有可用的认知行为知识时,智能体按照式(6)更新其策略. 用KL散度衡量当前DRL策略 π_ϕ与认知行为策略 π_{cog}之间的差异 L^{cog},其形式化表示为

$$L^{\text{cog}}(\phi)=D_{KL}(\pi_{\text{cog}}||\pi_\phi).\tag{7}$$

此外,由于人的先验知识可能并不是最优的,随着学习步长的推进,应逐渐减小对GOAL认知行为模型的依赖,认知行为模型对PPO的影响程度 $\rho_c\in[0,1]$是一个超参,COG-PPO在充分融合认知行为策略后仅依靠强化学习算法进行策略更新. 综上,actor

的优化目标函数可表示为

$$L_{\text{actor}} = L^{\text{clip}}(\phi) - \rho_c^k L^{\text{cog}}(\phi) + \rho_e E(\pi_\phi). \quad (8)$$

其中: k 为回合数, k 与 ρ_c 共同决定 GOAL 认知行为模型对 DRL 策略的影响程度; $E(\pi_\phi)$ 表示为鼓励探索而引入的信息熵; $\rho_e \in [0,1]$ 为信息熵系数. critic 网络的更新方式与 PPO 类似, 其损失函数可表示为

$$L_{\text{critic}}(\psi) = \hat{\boldsymbol{E}}_t\Big[\sum_{t'>t}\gamma^{t'-t}r_{t'} - V_\psi(s_t)\Big]^2. \quad (9)$$

4　实验设计与结果分析

本节在 OpenAI gym 的 MountainCar 环境下验证 COG-DQN 的有效性; 在视距内空战机动决策场景中验证 COG-PPO 算法的有效性.

4.1　MountainCar 环境实验设计与结果

4.1.1　实验设计

MountainCar 环境如图 6 所示, 小车的初始位置为山谷某随机位置 (坐标为 $-0.6 \sim -0.4$ 之间), 动作空间为: 向右加速 a_0、保持 a_1 和向左加速 a_2, 小车的任务为在有限的 200 个时间步内到达右侧的目标位置 (坐标为 0.5). 直观来看, 一种可行的认知行为知识是在任务开始阶段先向右加速, 再利用其势能向左加速.

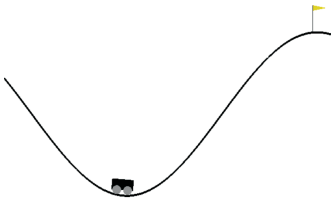


图 6　MountainCar 环境

根据图 3 所示构建认知行为模型的流程, MountainCar 环境下 GOAL 认知行为模型如图 7 所示. 首先, 根据环境的观测量 (小车位置、速度、当前时间步) 定义描述环境的知识集, 根据初始状态信息构建初始信念集, 根据终点信息构建目标集; 然后, 构建用于事件处理的感知规则用于信念更新; 最后, 根据小车的动作空间定义动作集, 根据前述先验知识定义行为规则集为: 前 30 步向右加速, 接下来的 30 步向左加速.

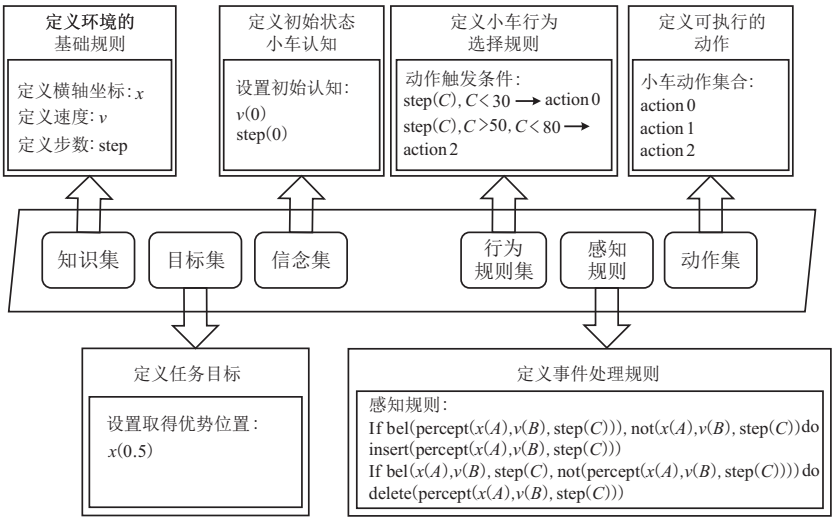


图 7　MountainCar 中的 GOAL 认知行为模型

为验证 COG-DQN 能够有效地利用认知行为知识, 本文将其与 DQN 以及仅使用 GOAL 模型决策的 BDI 智能体进行性能对比. COG-DQN 和 DQN 算法中的神经网络均采用双层 32 节点的全连接层. 仅使用 GOAL 模型决策的智能体没有认知行为知识可用时, 随机选择动作. 此外, 为验证 COG-DQN 在稀疏奖励环境中的性能, 本文在 MountainCar 环境中设计两种奖励形式: 1) 稠密奖励: 每一时刻, 小车的奖励与其高度成正比, 累积的势能越大, 奖励越大, 形式上可以描述为当前位置与目标位置之和的绝对值, 即 $r = |0.5 + \text{position}|$. 此外, 当小车到达目标位置时, 额外奖励为 10. 2) 稀疏奖励: 小车仅在最后到达目标位置时才获得奖励 10, 其余情况下奖励均为 0.

4.1.2　结果分析

以下结果为 30 组实验的统计结果, 其中每组实验包括 200 个回合.

图 8 和表 1 分别展示了稠密奖励条件下, 学习过程中最近 20 个回合的平均奖励以及学习后总成功率统计结果. 从图 8 中可以看出, 相比于 DQN, COG-DQN 在训练初期性能有明显提升, 这是因为借助 GOAL 认知行为模型的策略引导机制, 其在训练初期避免了不必要的动作探索, 从而加速了策略学习. 从

整个训练过程来看,COG-DQN 首先借助GOAL 模型实现热启动,随着学习进程的推进,其实现了认知行为知识基础上的再学习. 此外,从图 8 中平均奖励的标准差波动区间可以看出,COG-DQN 在学习过程不仅取得的收益更高,整体性能也更为平稳. 最后,从表 1 中可以看出,在稠密奖励的条件下,COG-DQN 在学习过程中的成功率约为69 %,说明GOAL 模型有效引导了 DRL 策略模型的收敛. 相比之下,基于 GOAL 决策的智能体虽然在初期平均奖励就超过了 50,但在整个过程中成功率为0,说明仅依靠不完备的GOAL 模型难以完成该任务.

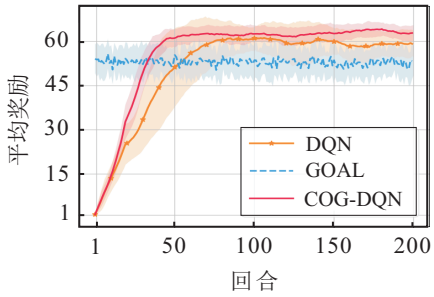


图 8 MountainCar 环境中稠密奖励下平均奖励

表 1 MountainCar 中稠密奖励下成功率统计

算法	平均成功次数	成功率 /%
DQN	127.53 ± 23.32	63.93
GOAL	0.00 ± 0.00	0.00
COG-DQN	138.47 ± 23.01	69.23

图 9 和表 2 分别展示了稀疏奖励条件下,学习过程中最近 20 个回合的平均奖励及学习后总成功率统计结果. 从图 9 和表 2 中 DQN 的结果可以看出,其未能完成该任务,在整个过程中成功到达目标点的次数为 0,这说明在连续状态空间下,稀疏奖励对 DRL

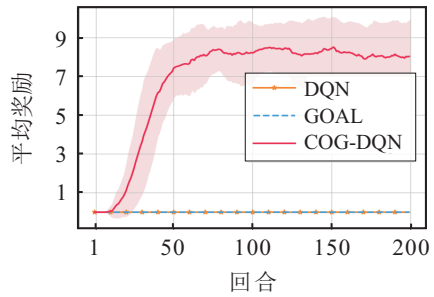


图 9 MountainCar 环境中稀疏奖励下平均奖励

表 2 MountainCar 中稀疏奖励下成功率统计

算法	平均成功次数	成功率 /%
DQN	0.00 ± 0.00	0.00
GOAL	0.00 ± 0.00	0.00
COG-DQN	140.73 ± 23.40	70.37

算法性能影响很大. 相比之下,融合了认知行为知识的COG-DQN 表现出了与稠密奖励条件下类似的结果,学习过程中到达目标点的成功率达 70 %,这说明 COG-DQN 实现了认知行为策略与 DRL 策略的有机融合,在 GOAL 模型的引导下,提升了 DRL 策略学习能力. 同时,这也表明本文所提出的框架和算法能有效缓解稀疏奖励对 DRL 策略学习的影响,有效提升了智能体在复杂任务环境中的性能.

4.2 空战机动决策实验设计与结果

4.2.1 实验设计

本节通过 1v1 和 1v 2 空战机动决策来验证 COG-PPO 算法的性能. 空战机动决策对抗环境某一时刻示例如图 10 所示,战机采用 6 自由度模型[21],并被限定在 200 km × 200 km × 20 km 的三维空间内. 初始态势为我机被敌机尾后,敌我双方初始位置坐标为 [105 km, 100 km, 5 km] 和 [110 km, 100 km, 5 km],双方初始速度为 200 m/s,初始滚转角、航迹倾角、航迹偏角都为 0°. 在该初始态势下,我机应首先摆脱敌机的追击,而后通过机动获取有利位置并获得开火条件.

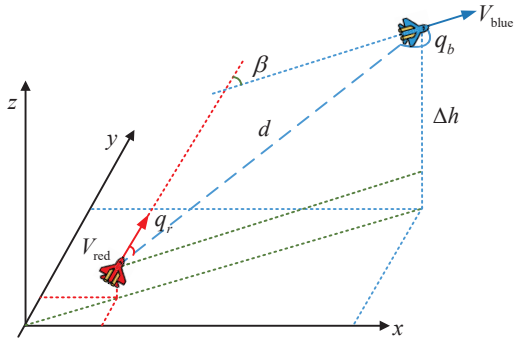


图 10 空战机动环境

战机的动作空间为美国太空总署提出的 7 种基本战机机动动作. 对战斗机的轨迹和姿态控制可以转化为对切向过载 η_x、法向过载 η_f 和滚转角 μ 的控制.

下面以 1v1 场景为例,详细说明描述态势信息的状态和奖励函数设计方法. 选取 11 维的相对态势信息 $[q_r, \dot{q}_r, q_b, \dot{q}_b, \beta, d, \Delta h, h_r, \Delta v^2, v_r{}^2, \dot{v}_r]$ 作为智能体的状态输入,如图 10 所示. q_r 为偏离角,表示我机速度矢量与我敌战机质心连线的夹角;q_b 为脱离角,表示敌机速度矢量与敌我战机质心连线的夹角;$\dot{q}_r$、$\dot{q}_b$ 分别为偏离角和脱离角在当前时刻的变化率;β 为两机速度的夹角;d 为两机质心之间的距离;h_r 为我机高度;Δh 为我敌高度差;v_r 为我机速度;$\dot{v}_r$ 为我机速度的变化率;Δv^2 为我敌双方速度的平方差.

为加速策略学习,缓解空战机动环境稀疏奖励

的影响,将相对态势优势[22](角度优势 f_a、距离优势 f_d、速度优势 f_v、高度优势 f_h) 引入奖励函数设计,则合成优势函数构成的奖励可以表示为 $f_{\text{situation}} = \omega_a f_a + \omega_d f_d + \omega_v f_v + \omega_h f_h$. 其中: $\omega_a + \omega_d + \omega_v + \omega_h$=1, ω_a、ω_d、ω_v、ω_h 分别为角度优势、距离优势、速度优势和高度优势的权重. 环境的奖励函数主要考虑战机是否达到开火条件. 以我机为例,根据状态输入,当我机速度矢量与我敌质心的连线构成的偏离角 q_r 小于 30°,脱离角 q_b 大于 120°,两机速度的夹角 β 小于 45°,质心间的距离 d 小于 1.5 km 时,我机获得攻击机会,在空战机动对抗环境中获得正向奖励. 相反,当敌方战斗机获得攻击机会时,我机从对抗环境获得负向奖励,则环境的奖励函数为

$$f_{\text{env}} = \begin{cases} 10,\ q_r < 30^\circ,\ q_b > 120^\circ, \\ \quad \beta < 45^\circ,\ d < 1.5\,\text{km}; \\ -10,\ q_b < 30^\circ,\ q_r > 120^\circ, \\ \quad \beta < 45^\circ,\ d < 1.5\,\text{km}; \\ -10,\ v_r < 80\,\text{m/s or } v_r > 300\,\text{m/s}; \\ -10,\ h_r < 0.2\,\text{km or } h_r > 1.8\,\text{km}; \\ 0,\ \text{otherwise}. \end{cases} \tag{10}$$

综合考虑对抗环境奖励和态势优势奖励,总体奖励函数为

$$\mathcal{R} = f_{\text{env}} + \rho_s \cdot f_{\text{situation}}(f_a, f_d, f_v, f_h), \tag{11}$$

$\rho_s \in \boldsymbol{R}$ 用于调整态势优势在总体奖励中的占比.

接下来说明在我机被尾后的初始态势下,GOG-PPO 使用的认知行为知识——战斗转弯. 战斗转弯是在实战中常用的典型空战机动动作,如图 11 所示,其主要作用是当敌方尾后我机时 (本实验的初始态势),我机为了迅速摆脱敌方追击并尝试获取优势态势位置的机动动作. 该动作可以拆解为保持飞行 a_3、仰起飞行 a_6 和左转弯飞行 a_1 三种基本机动动作.

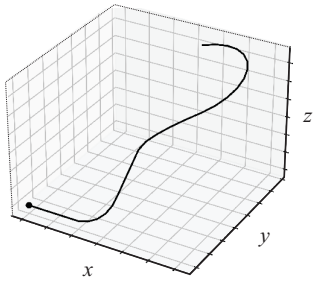


图 11　战斗转弯机动动作

在 1v2 对抗环境中,除战斗转弯外,COG-PPO 使用的认知行为知识还包括: 对抗时先以一架敌机作为目标,获得开火机会后再切换对抗目标. 奖励函数设计方面, $f_{\text{situation}}$ 包含我机对两架敌机的态势优势,若对抗中成功获得两次开火机会,则从环境中额外获取 100 的奖励. 该环境中的其他设定与 1v1 保持一致. 此外,为了验证所添加认知行为知识的有效性,设计如下对比实验: 对比算法仍为 COG-PPO,我机状态输入为我机分别与两敌机构成的相对态势 (22 维状态向量),且仅使用战斗转弯构建认知行为模型,其余设定保持不变.

根据图 3,本环境中 GOAL 认知行为模型的设计流程与第 4.1.1 节类似,此处不再展开叙述. 需要注意的是本实验仅选用少量先验构建 GOAL 模型,一方面是因为要验证 COG-PPO 可以在学习中融合认知行为知识,并在此基础上学习对抗策略; 另一方面是验证其在仅有部分可用认知行为知识前提下的学习能力. 此外,在本实验中,敌机采取获取前述最大合成态势优势的贪婪策略决策.

本实验中, PPO 和 COG-PPO 算法的 actor 和 critic 网络均由 4 层 256 节点的全连接层构成,共训练 50 万个回合.

4.2.2　1v1 空战机动决策结果分析

图 12 展示了我机 (实线) 使用学习后的 PPO 模型决策时,敌我双方的对抗轨迹及双方态势优势变化曲线. 从双机对抗轨迹来看,我机主要通过螺旋上升的方式躲避敌方追击,结束时的态势信息为 q_r = 161.7°, q_b = 1.71°, d = 1.48 km, β = 19.9°,最终以敌机获得开火机会结束对抗. 从双方态势优势变化曲线可以看出,我机态势优势明显低于敌机. 这说明

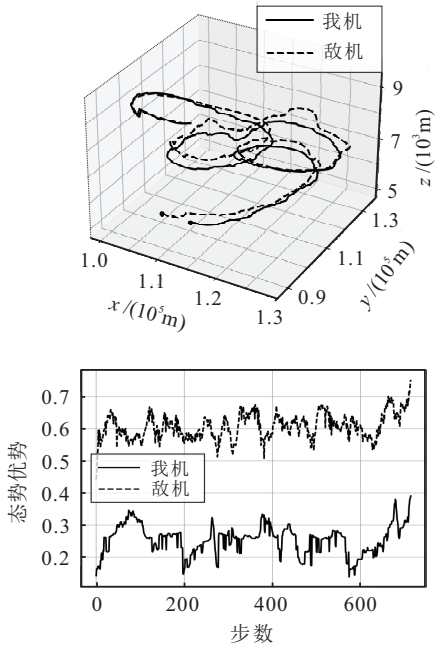
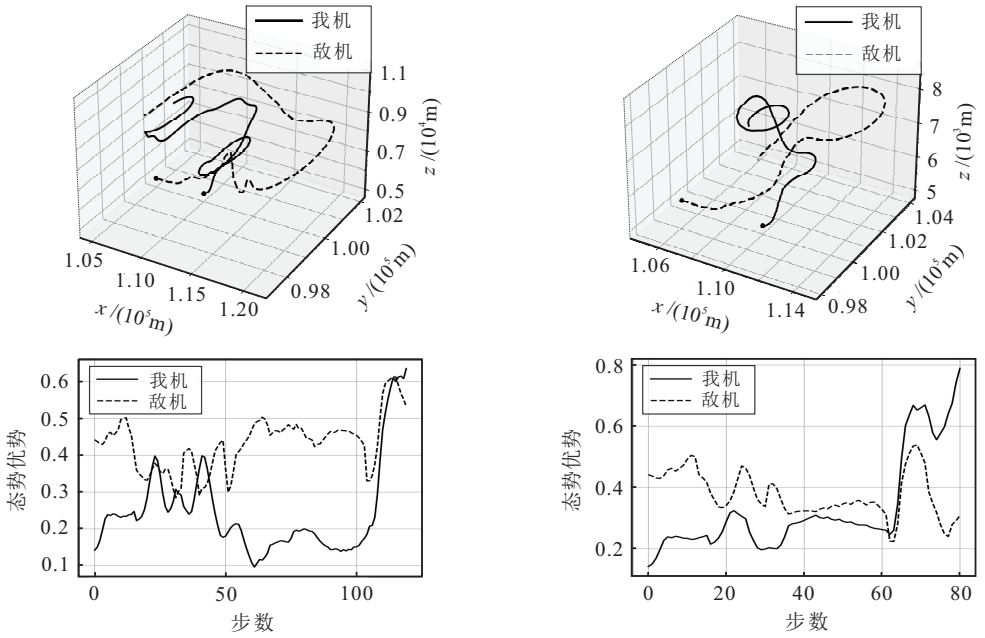


图 12　我机为 PPO 时对抗轨迹及双方态势优势

仅依靠PPO的DRL策略不能够摆脱敌机追击并获得开火机会.

图13展示了我机(实线)使用学习后的COG-PPO模型决策时,敌我双方的对抗轨迹及双方态势优势变化曲线. 直观来看,两次对抗中,我机均能够在对抗初期就摆脱敌机追击,并在对抗中逐渐占据优势,最终获得开火机会. 上述两次对抗的结束态势如表3所示,均符合式(10)规定的开火条件. 从图13的各态势优势曲线可以看出:由于我方被尾追,我机在初始阶段处于明显劣势;但在对抗过程中,我方逐渐在态势位置上扳回劣势;在对抗中期,双方的态势优势互有升降,说明空战机动过程中双方都在争取获得开火条件;但最终我机获得了开火机会. 此外,使用学习后的COG-PPO模型决策时,对抗胜率超过了98%.



(a) 1v1对抗1 (b) 1v1对抗2

图13 我机为COG-PPO时对抗轨迹及双方态势优势

表3 我机为COG-PPO时1v1对抗结束态势信息

对抗名称	开火时刻	态势信息
1v1对抗1	119	$q_r=27.4^\circ, q_b=158.5^\circ,$ $d=1.32\,\text{km}, \beta=11.8^\circ$
1v1对抗2	80	$q_r=29.9^\circ, q_b=137.7^\circ,$ $d=1.30\,\text{km}, \beta=41.8^\circ$

从以上两次对抗的轨迹曲线可以明显看出:在对抗初始阶段,COG-PPO利用GOAL认知行为模型描述的战斗转弯机动动作,在对抗开始时,顺利地摆脱了敌机,并获得高度优势,迅速缩小态势优势差距;之后,COG-PPO借助DRL算法的学习能力,在认知行为模型的基础上继续学习最优对抗策略.

4.2.3 1v2空战机动决策结果分析

我机仅使用战斗转弯构建认知行为模型时,双方对抗轨迹示例及总体态势优势如图14所示,其中我方态势优势为我机分别相对两敌机态势优势之和,而敌方总体态势优势为两敌机对我机态势优势之和. 从图14可以看出:对抗初期,我机成功通过战斗转弯扳回劣势,双方优势基本持平;但在以后的对抗过程中,两敌机逐渐占据优势位置形成追击态势,并获得开火机会;结束时态势信息为$q_r=137.4^\circ, q_b=27.0^\circ, d=1.14\,\text{km}, \beta=15.6^\circ$.

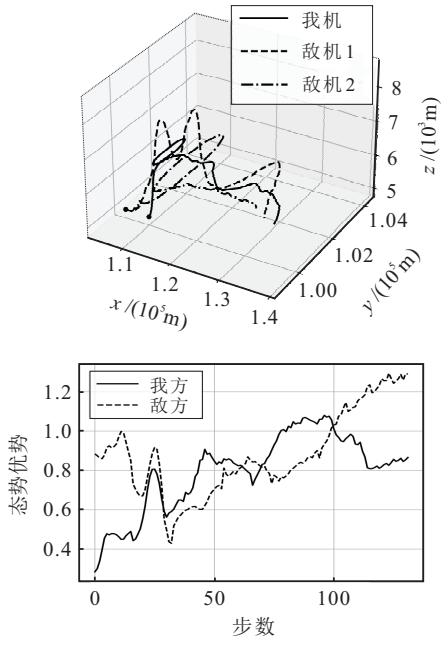
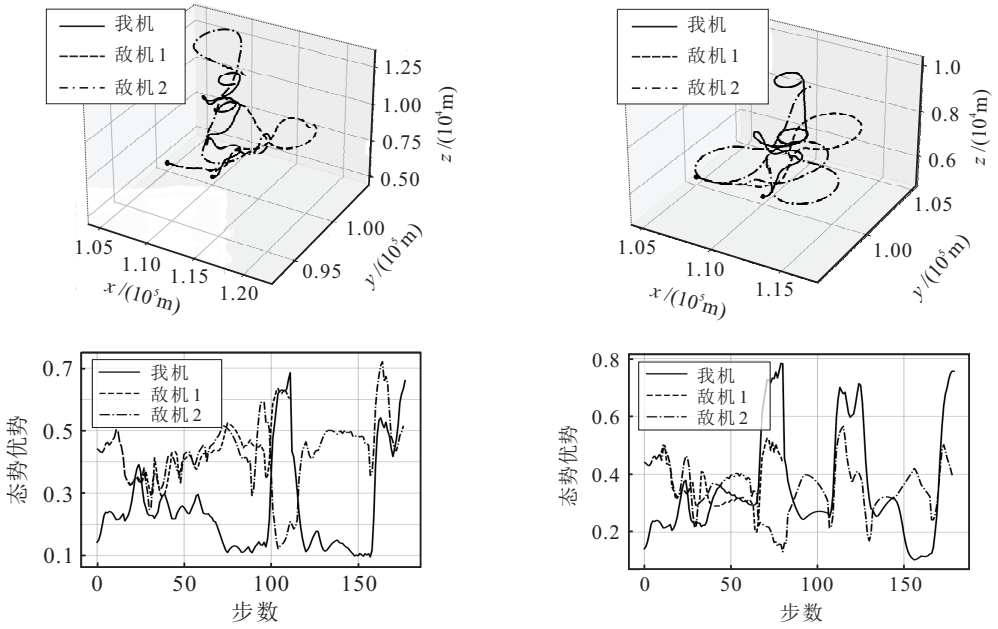


图14 COG-PPO仅用战斗转弯构建认知行为模型时1v2对抗轨迹及双方总体态势优势

我机使用4.2.1节中构建的完整认知行为模型时,学习后的COG-PPO决策模型与两敌机的对抗轨迹如图15所示,其中还分别给出了参战飞机相对于各自目标机的态势优势变化曲线. 从对抗轨迹可以看出:初始时刻,我机被两架敌机尾后,态势优势函数明显低于敌方;在对抗初期,我机首先借助GOAL认知行为模型,使用战斗转弯弥补态势劣势,并占据高度优势;然后,我机先以一架敌机为目标,在对抗中力争获得优势态势位置,我机使用战斗转弯扳回劣势后,敌机立即采取应对的机动动作,因此,双方态势优势曲线在对抗过程中互有升降;我机态势优势曲线在己方获得一次开火机会后突降,这是因为此时我机切换了作战目标,转而针对存活的敌机进行机动决策;最终,我机通过认知行为模型基础上的再学习,在对抗中获得两次开火机会,取得最后的胜利. 上述对抗中,我机获得开火机会时的态势信息如表4所示. 经统计,在1v2场景中,我机使用该模型决策时,获得两次开火机会的对抗约占78%,对抗结束时我机仍存活的几率超过97%. 在对抗中我机未能两次开火的原因主要有: 1) 期间有敌机脱离划定的格斗空间(约占19%); 2) 敌机获得开火机会,对抗结束(约占2%).



(a) 1v2对抗1　　(b) 1v2对抗2

图 15　我机为COG-PPO时1v2对抗轨迹及我机分别与两架敌机态势优势

表 4　我机为COG-PPO时1v2对抗结束态势信息

对抗名称	开火时刻	态势信息
1v2对抗1	112	$q_r=27.7^\circ, q_b=122.4^\circ,$ $d=0.90\,\text{km}, \beta=30.7^\circ$
	177	$q_r=24.9^\circ, q_b=119.8^\circ,$ $d=0.99\,\text{km}, \beta=38.3^\circ$
1v2对抗2	81	$q_r=27.5^\circ, q_b=146.4^\circ,$ $d=1.15\,\text{km}, \beta=18.9^\circ$
	179	$q_r=28.3^\circ, q_b=153.1^\circ,$ $d=1.32\,\text{km}, \beta=1.8^\circ$

综上,COG-PPO可以在学习过程中融合认知行为模型,在满足认知行为知识的触发条件时,不断缩小当前策略与认知行为策略在当前状态下的差距,向认知行为策略靠拢. 在没有可用的认知行为知识时,借助已有的学习经验,并在此基础上依靠DRL算法优化其对抗策略.

5　结　论

本文整合了DRL智能体与基于BDI智能体的优势,提出了融合认知行为模型的深度强化学习框架和GOAL认知行为模型的设计流程. 在此基础上,针对基于值函数的方法和基于策略梯度的方法,分别提出了COG-DQN和COG-PPO算法. 实验结果表明,通过知识融合,本文所提出的框架和算法有效缓解了复杂任务中状态空间巨大和奖励稀疏等问题对DRL算法的影响. 值得注意的是,本文提出的框架不仅适用于所提出的两种算法,而且在具体应用中,应结合任务的类型和特点,选取合适的DRL算法,并结合本文提出的框架,借助GOAL认知行为模型加速策略收敛. 下一步的研究重点是如何将本文提出的框架应用于多智能体强化学习算法,并在此基础上深入研究如何借助GOAL的消息通信机制提高任务协作效率,从而加速多智能体合作策略学习过程.

参考文献(References)

[1] Kakade S M. On the sample complexity of reinforcement learning[D]. London: University of London, 2003.

[2] Sutton R S, Barto A G. Reinforcement learning: An introduction[J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054.

[3] Taylor M E, Stone P. Transfer learning for reinforcement learning domains: A survey[J]. Journal of Machine Learning Research, 2009, 10(7): 1633-1685.

[4] Da Silva F L, Costa A H R. A survey on transfer learning for multiagent reinforcement learning systems[J]. Journal of Artificial Intelligence Research, 2019, 64: 645-703.

[5] Yang T P, Hao J Y, Meng Z P, et al. Towards efficient detection and optimal response against sophisticated opponents[C]. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao (China), 2019: 623-629.

[6] Chen H, Liu Q, Huang J, et al. Efficiently tracking multi-strategic opponents: A context-aware Bayesian policy reuse approach[J]. Applied Soft Computing, 2022, 121: 108715.

[7] Ammar H B, Eaton E, Taylor M E, et al. An automated measure of MDP similarity for transfer in reinforcement learning[EB/OL]. (2014-07-28)[2022-06-15]. https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/view/8824.

[8] Song J H, Gao Y, Wang H, et al. Measuring the distance between finite Markov decision processes[C]. Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems. Singapore: Springer, 2016: 468-476.

[9] Brys T, Harutyunyan A, Taylor M E, et al. Policy transfer using reward shaping[C]. Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems. Istanbul: Springer, 2015: 181-188.

[10] Bianchi R A C, Martins M F, Ribeiro C H C, et al. Heuristically-accelerated multiagent reinforcement learning[J]. IEEE Transactions on Cybernetics, 2014, 44(2): 252-265.

[11] Li S Y, Zhang C J. An optimal online method of selecting source policies for reinforcement learning[C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 3562-3570.

[12] Bratman M E, Israel D J, Pollack M E. Plans and resource-bounded practical reasoning[J]. Computational Intelligence, 1988, 4(3): 349-355.

[13] Broekens J, Harbers M, Hindriks K, et al. Do you get it? user-evaluated explainable BDI agents[C]. Multiagent System Technologies. Berlin: Springer, 2010: 28-39.

[14] Bordini R H, El Fallah Seghrouchni A, Hindriks K, et al. Agent programming in the cognitive era[J]. Autonomous Agents and Multi-Agent Systems, 2020, 34(2): 1-31.

[15] Hindriks K V, De Boer F S, Van Der Hoek W, et al. Agent programming with declarative goals[C]. Proceedings of 7th International Workshop on Intelligent Agents VII. Boston: Springer, 2000: 228-243.

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

[17] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J/OL]. 2017, arXiv: 1707.06347.

[18] Watkins C J C H, Dayan P. Q-learning[J]. Machine Learning, 1992, 8(3/4): 279-292.

[19] Sterling L, Shapiro E, Garrett R. The art of prolog[J]. IEEE Expert, 1987, 2(2): 106-107.

[20] GOAL Confluence. Tutorials, documentation & educational materials[EB/OL]. (2021-08-16)[2022-06-15]. https://goalapl.atlassian.net/wiki/spaces/GOAL/overview.

[21] 孔江涛. 面向双机空战机动决策的置信规则推理技术研究[D]. 长沙: 国防科学技术大学, 2015: 16-21.
(Kong J T. Research of belief-rule-based reasoning technology for learning air combat maneuvers[D]. Changsha: National University of Defense Technology, 2015: 16-21).

[22] 丁林静, 杨啟明. 基于强化学习的无人机空战机动决策[J]. 航空电子技术, 2018, 49(2): 29-35.
(Ding L J, Yang Q M. Research on air combat maneuver decision of UAVs based on reinforcement learning[J]. Avionics Technology, 2018, 49(2): 29-35.)

作者简介

陈浩(1993−), 男, 博士生, 从事多智能体强化学习、博弈对抗等研究, E-mail: chenhao@nudt.edu.cn;

李嘉祥(1996−), 男, 硕士, 从事多智能体强化学习的研究, E-mail: lijiaxiang14@nudt.edu.cn;

黄健(1971−), 女, 教授, 博士生导师, 从事人工智能、系统仿真等研究, E-mail: nudtjhuang@hotmail.com;

王菖(1985−), 男, 讲师, 博士, 从事强化学习、人机协同等研究, E-mail: wangchang07@nudt.edu.cn;

刘权(1985−), 男, 副研究员, 博士, 从事多智能体系统、强化学习等研究, E-mail: liuquan@nudt.edu.cn;

张中杰(1990−), 男, 讲师, 博士, 从事数据挖掘、多智能体系统等研究, E-mail: zjiezhang@hotmail.com.