计算机科学
COMPUTER SCIENCE

基于混合模仿学习的多智能体追捕决策方法

王焱宁, 张锋镝, 肖登敏, 孙中奇

引用本文

王焱宁, 张锋镝, 肖登敏, 孙中奇. 基于混合模仿学习的多智能体追捕决策方法[J]. 计算机科学, 2025, 52(1): 323-330.

WANG Yanning, ZHANG Fengdi, XIAO Dengmin, SUN Zhongqi. Multi-agent Pursuit Decision-making Method Based on Hybrid Imitation Learning [J]. Computer Science, 2025, 52(1): 323-330.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800072

基于混合模仿学习的多智能体追捕决策方法

王焱宁[1,2]　张锋镝[1,2]　肖登敏[3]　孙中奇[4]
1 北京航天自动控制研究所　北京 100854
2 宇航智能控制技术全国重点实验室　北京 100854
3 中船智海创新研究院有限公司　北京 100094
4 北京理工大学自动化学院　北京 100081
(wyn_81_2049@163.com)

摘　要　针对传统模仿学习方法在处理多样化专家轨迹时的局限性,尤其是难以有效整合质量参差不齐的固定模态专家数据的问题,创新性地融合了多专家轨迹生成对抗模仿学习(Multiple Trajectories Generative Adversarial Imitation Learning,MT-GAIL)方法与时序差分误差行为克隆(Temporal-Difference Error Behavioral Cloning,TD-BC)技术,构建了一种混合模仿学习框架。该框架不仅可以增强模型对复杂多变的专家策略的适应能力,还能够提升模型从低质量数据中提炼有用信息的鲁棒性。框架得到的模型具备直接应用于强化学习的能力,仅需经过细微的调整与优化,即可训练出一个直接可用的、基于专家经验的强化学习模型。在二维动静结合的目标追捕场景中进行了实验验证,该方法展现出良好的性能。结果表明,所提方法可以吸取专家经验,为后续的强化学习训练阶段提供一个起点高、效果佳的初始模型。

关键词:智能决策;强化学习;行为克隆;生成对抗模仿学习
中图分类号　TP182

Multi-agent Pursuit Decision-making Method Based on Hybrid Imitation Learning

WANG Yanning[1,2],ZHANG Fengdi[1,2],XIAO Dengmin[3] and SUN Zhongqi[4]
1 Beijing Aerospace Automatic Control Institute,Beijing 100854,China
2 National Key Laboratory of Science and Technology on Aerospace Intelligence Control,Beijing 100854,China
3 China Ship Intelligence and Marine Innovation Research Institute Co.,Ltd.,Beijing 100094,China
4 School of Automation,Beijing Institute of Technology,Beijing 100081,China

Abstract　Aiming at the limitations of traditional imitation learning approaches in handling diverse expert trajectories,particularly the difficulty in effectively integrating fixed-modality expert data of varying quality,this paper innovatively integrates the multiple trajectories generative adversarial imitation learning(MT-GAIL) method with temporal-difference error behavioral cloning(TD-BC) technology to construct a hybrid imitation learning framework. This framework not only enhances the model's adaptability to complex and dynamic expert strategies but also improves its robustness in extracting useful information from low-quality data. The resulting model from this framework is directly applicable to reinforcement learning,requiring only minor adjustments and optimizations to train a readily usable reinforcement learning model grounded in expert experience. Experimental validation in a two-dimensional dynamic-static hybrid target pursuit scenario demonstrates the method's impressive performance. The results indicate that the proposed method effectively assimilates expert knowledge,providing a high-starting-point and effective initial model for subsequent reinforcement learning training phases.

Keywords　Intelligent decision-making,Reinforcement learning,Behavior cloning,Generative adversarial imitation learning

1　引言

多智能体系统是研究利益不同的智能体间合作与竞争的理想模型。通过设计合适的合作或竞争机制,多智能体系统能够灵活应对各种复杂环境并共同完成特定任务。在工业自动化、机器人等众多领域中,多智能体系统的研究和应用越来越广泛[1-2],研究的重要性和潜在价值日益凸显。

在多智能体系统的研究领域中,多智能体追捕问题是一个典型问题。在追捕过程中,每个智能体需要获取环境的

到稿日期:2024-08-13　返修日期:2024-09-23
通信作者:肖登敏(2712538468@qq.com)

实时信息，并根据实时信息及时做出决策。面对复杂多变的实际环境，非学习类方法如博弈论方法[3]、人工势场法[4]等往往难以做出正确且快速的反应。相比之下，强化学习方法凭借其强大的高维度信息感知、理解以及非线性处理能力，可以满足多智能体追捕问题对实时性和灵活性的要求，使得强化学习成为解决追捕问题的理想工具。

利用强化学习方法解决多智能体追捕问题也存在一些挑战。由于强化学习是无经验学习，因此，智能体在训练过程中需要不断“试错”以增长经验。在复杂环境中，采用强化学习训练的策略存在训练时间长、收敛速度慢的问题。

人类在成长过程中经常会通过模仿来学习新的技能、知识和行为方式。采用模仿学习的方法，智能体可以通过模仿专家轨迹学习到较理想的行为策略，从而避免不必要的试错。针对多智能体追捕问题，如果有历史数据或者专家经验为智能体提供辅助决策，其学习效率将得到提高。通过模仿专家示例的轨迹，智能体可以高效地学习专家的策略，避免不必要的探索，从而缩短训练时间。

本文的主要贡献为：1）提出混合模仿学习框架，利用该框架可以学习多模态的专家数据；2）利用多智能体追捕决策场景进行实验验证，为后续解决多智能体追捕问题的研究提供思路。

2 相关工作

追捕过程中环境是时刻变化的，每个智能体需要获取环境的实时信息，并根据实时信息及时做出变换追捕目标、重新组队等决策。因此，在不断变化的环境中，追捕问题是一个备受关注但目前尚未解决的实时知识处理问题，也是研究多智能体合作、协调以及对抗策略进化的常见问题。对于较为复杂的场景，强化学习方法的训练过程比较耗时[5]。针对多智能体追捕问题，如果有历史数据或者专家经验为智能体提供辅助决策，其学习效率将得到提高。

采用模仿学习的方法，智能体可以通过模仿专家轨迹学习到较理想的行为策略，从而避免不必要的试错，使得模型较快处于一个较高的智能化水平。目前的模仿学习方法可以在单模态的专家数据中学习到很好的策略。单模态专家轨迹指同一专家在同种任务下得出行动产生的轨迹数据，这样的专家轨迹不具有多样性，训练得到的模型难以应对较为复杂的场景。因此有学者开始研究多模态模仿学习。多模态模仿学习的相关研究可以分为两类，一类是多模态专家轨迹信息不带有标签，通过无监督学习的方式区分模态标签信息。Li 等提出通过最大化模态隐变量和专家示范轨迹之间的信息来推断出模态标签的方法[6]；2017 年，DeepMind 团队引入变分自编码器来判断模态标签[7]。但是，这类问题采用无监督学习的方式有时并不能准确区分各模态的专家轨迹数据，导致难以学习一个性能良好的多模态策略。另一类多模态模仿学习的研究内容是从带有模态标签的多模态专家轨迹中进行学习。DeepMind 团队基于 GAIL 方法将模态标签输入生成器和判别器中，使用模态标签来指导训练[8]；Lin 等在 GAIL 框架的基础上引入辅助分类器，用于根据所属模态的信息对样本数据进行分类[9]；Raunak 等研究多智能体模仿学习方法[10]，将模仿学习方法引入多智能体领域。

在模仿学习中，专家可以是以特定的专家经验为基础形成的决策树，也可以是通过强化学习训练得到的模型。强化学习与模仿学习方法共同训练具有一定的优势。不论是在模仿学习还是强化学习的训练过程中，每一轮训练迭代结束后，智能体与环境交互产生的数据往往会被丢弃。这样不仅会降低样本效率，增加计算资源的开销，同时也会增加训练耗时[11-12]。如果将强化学习训练的数据收集起来用于模仿学习训练，便可以提高历史数据的使用率，使得模型在较短时间内处于一个较智能的水平。因此可以将简单任务得到的模型作为专家模型，利用专家模型生成模仿学习所需的专家轨迹。该方法可以避免强化学习完全的无经验学习，对较为复杂的任务进行前期引导，从而加速模型收敛。随着训练轮数的增加，训练过程由专家轨迹的引导转变为强化学习奖励函数的引导，最后得到收敛速度较快且能适应复杂环境的策略模型。

3 混合模仿学习框架

模仿学习方法大致可以分为 3 类，分别是行为克隆方法（Behavioral Cloning，BC）[13-14]、逆强化学习方法（Inverse Reinforcement Learning，IRL）[15-17] 和生成对抗模仿学习方法（Generative Adversarial Imitation Learning，GAIL）[18-19]。考虑到 BC 具有简单易实现的优点，而 GAIL 在模仿效果上表现突出，本文深入研究这两种方法，并将两种方法相结合构成混合模仿学习方法。

3.1 TD-BC 方法

BC 方法通常只能生成一个网络，无法直接用于强化学习后续训练过程。在 BC 方法的基础上进行改进，形成 TD-BC 方法。TD-BC 方法主要包括生成专家模型、生成专家数据、训练学生模型 3 个步骤。

1）生成专家模型

在模仿学习方法中，通常假定专家模型代表最优策略。专家模型可以基于人为的专家经验构建成决策树的形式，也可以是通过强化学习训练得到的神经网络模型。

2）生成专家数据

专家模型与环境进行交互，可以得到专家数据。对于 BC 方法，专家数据是状态和动作的二元组，即 $X=\{(s_1,a_1),(s_2,a_2),\cdots,(s_n,a_n)\}$；对于 TD-BC 方法，专家数据是由三元组构成的，即 $X=\{(s_1,a_1,s_1'),(s_2,a_2,s_2'),\cdots,(s_n,a_n,s_n')\}$。其中，$s_j$ 表示 j 时刻的状态，a_j 表示对应的动作，s_j' 表示执行完动作 a_j 后的状态。行为克隆本质上是有监督学习，学习的样本是 s_j，学习的标签是 a_j。s_j' 被用来更新价值网络。

3）训练学生模型

为了与专家模型进行区分，定义通过模仿学习方法得到的模型为学生模型。TD-BC 方法的学生模型包含两个网络，

即策略网络和价值网络。对于策略网络参数的更新，目标函数为：

$$\theta^* = \arg\min_{\theta} E_{(s,a)\sim X}[L(\pi_\theta(s),a)] \tag{1}$$

其中，X 为专家数据集，L 为损失函数，s 和 a 分别表示专家轨迹数据中的状态和动作，$\pi_\theta(s)$表示策略网络。

如果场景是离散动作空间，则损失函数可以采用最大似然估计；反之，如果场景是连续动作空间，则损失函数可以采用均方误差函数。由于本文研究的场景为连续动作空间，因此采用均方误差损失函数更新策略网络。式(1)的损失函数 L 可以改写为：

$$L(s,a;\theta) = \frac{1}{2}[\pi_\theta(s) - a]^2 \tag{2}$$

采用梯度下降的方法更新策略网络的参数。训练过程中，损失函数的数值会逐渐减小，表明策略网络输出的决策逐渐接近专家决策。

对于价值网络参数的更新，引入了 TD 误差的思想。具体而言，定义 TD 误差为[20]：

$$TD\_error = q(s_{t+1},a_t;w) - q(s_t,a_t;w) \tag{3}$$

其中，$q(s,a;w)$为价值网络的输出，是动作价值函数 $Q_\pi(s,a)$ 的估计；$q(s_{t+1},a_t;w)$表示下一时刻的动作价值；$q(s_t,a_t;w)$表示当前时刻的动作价值。

定义价值网络的损失函数为：

$$critic\_loss = \frac{\sum_{i=1}^{size} |q(s_t,a_t;w)(i) - TD\_error(i)|}{size} \tag{4}$$

其中，$size$ 表示更新网络时需要使用的样本数量，是一个超参数。

3.2 MT-GAIL 方法

GAIL 方法通常只考虑质量相似的专家轨迹，对 GAIL 方法进行改进得到 MT-GAIL 方法。该方法可以综合考虑多条质量不同的专家数据，避免采用最差的专家数据进行训练。由于训练过程中神经网络不知道哪条专家轨迹是最优的，因此可以通过获取每条专家轨迹对应的判别器的准确率作为可靠度系数来确定如何选择专家轨迹进行训练。将可靠度系数与对应判别器的输出相乘得到模仿学习奖励值，并将该值存入经验池，用于训练策略生成器。算法框架如图 1 所示。

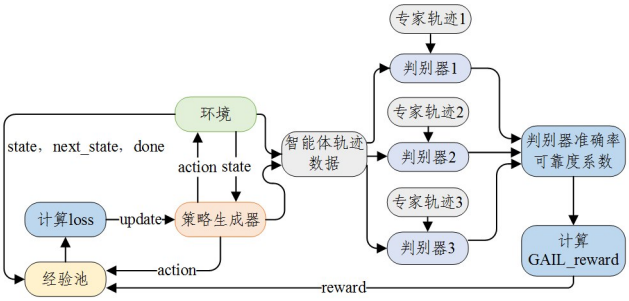


图 1 MT-GAIL 算法框架

Fig. 1 Framework of MT-GAIL algorithm

定义 3 个专家的专家策略分别为 π_{E1}，π_{E2} 和 π_{E3}，使用 3 个专家策略进行推演得到 3 条质量不同的专家轨迹。专家轨迹为状态和动作构成的二元组。将策略生成器与环境交互产生的智能体轨迹数据和专家轨迹都作为输入传入对应的判别器中，在训练过程中不断更新判别器网络和策略生成器网络的参数。判别器的损失函数采用二元交叉熵损失函数，定义为：

$$L = \frac{\sum_i -[y_i \times \log(D_w(s,a)) + (1-y_i)\times \log(1-D_w(s,a))]}{N} \tag{5}$$

其中，y_i 代表样本 i 的标签，定义 1 为正样本，0 为负样本，$D_w(s,a)$为判别器的输出。

在 MT-GAIL 方法中，定义标签 1 是专家数据样本，定义标签 0 是策略生成器产生的数据。训练判别器的目标函数为 $\min L(p_E,1) + \min L(p_{\text{Builder}},0)$。采用梯度下降法更新判别器网络。梯度的更新公式为：

$$\hat{E}_{T_i}[\nabla_w \log(D_w(s,a))] + \hat{E}_{T_E}[\nabla_w \log(1-D_w(s,a))] \tag{6}$$

其中，T_i 表示由策略生成器产生的轨迹数据，T_E 表示由专家产生的轨迹数据。

强化学习方法的奖励值通常由环境直接提供。对于模仿学习方法，通常环境无法直接提供奖励值。因为判别器的作用是区分策略网络轨迹数据和专家轨迹数据，所以判别器的准确率可以反映策略网络模型的优劣。下面介绍根据判别器准确率计算奖励值的方法[21]。

计算每一个判别器的专家准确率和学生准确率，其中专家准确率用 E_acc 表示，学生准确率用 L_acc 表示。对应的计算方法为：

$$\begin{aligned} E\_acc &= \hat{E}_{T_E}[D_w(s,a) > 0.5] \\ L\_acc &= \hat{E}_{T_E}[D_w(s,a) < 0.5] \end{aligned} \tag{7}$$

将专家准确率和学生准确率相加，得到相应判别器的总体准确率。给总体准确率较高的判别器赋予较高的权重，给总体准确率较低的判别器赋予较低的权重。权重的计算方法为：

$$E(j)\_w = \frac{acc(j)}{acc1 + acc2 + acc3} \tag{8}$$

其中，$acc1$，$acc2$，$acc3$ 分别表示对应判别器的总体准确率，j 表示第 j 个判别器。

将判别器的输出作为奖励值的一部分，奖励值的定义方法为：

$$\sum_{j=1}^{3}[E(j)\_w * (-\log(1-D(j)_w(s,a)))] \tag{9}$$

将环境的状态数据、策略生成器的动作数据以及奖励数据存入经验池中，采用强化学习方法更新策略生成器网络的模型参数。

3.3 混合模仿学习方法

结合 TD-BC 方法和 MT-GAIL 方法，形成混合模仿学习方法，方法框架图如图 2 所示。对于专家轨迹模态相似的情况，采用 TD-BC 方法进行训练；对于专家轨迹模态多样的情况，采用 MT-GAIL 方法进行训练。

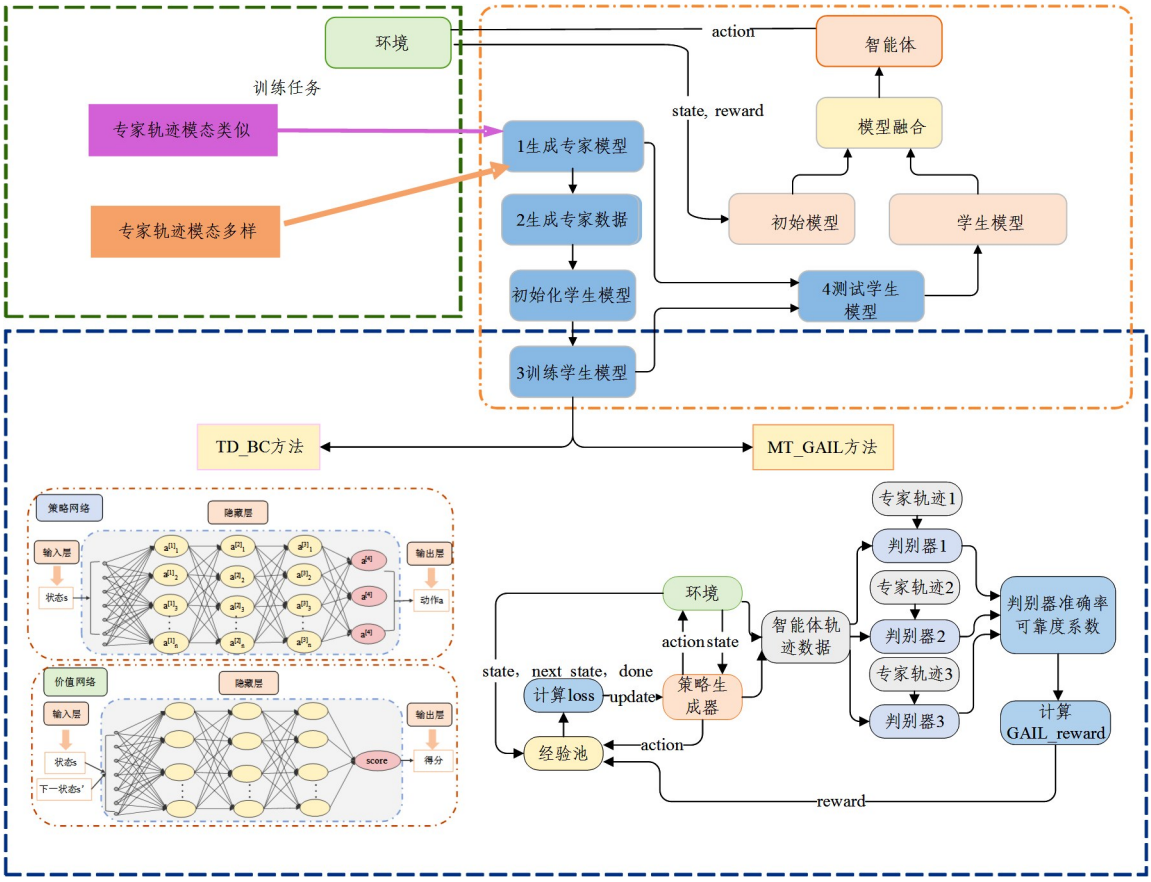


图 2 混合模仿学习方法框架图

Fig. 2 Framework of hybrid imitation learning method

4 实验验证

4.1 场景描述

采用混合模仿学习方法模仿智能体在二维平面追捕动静结合目标的策略。图 3 为仿真场景示意图，图中虚线框部分表示智能体可能出现的初始位置。虚线框内的小球为我方智能体，浅色小球为敌方智能体，虚线圆圈为静态目标，分布在场景中的黑色小球为障碍物。其中静态目标用于模拟实际情景中隐藏不动的敌方目标，动态目标用于模拟可移动的敌方目标。

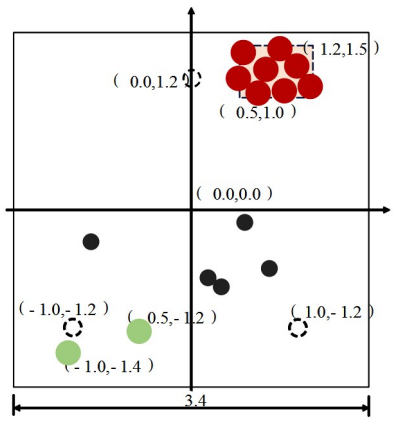


图 3 用于模仿学习训练的场景示意图

Fig. 3 Schematic diagram of the scene used for imitation learning training

场景中有 2 个动态目标、3 个静态目标、8 个我方智能体，而模仿学习方法只能模仿固定的策略。因此采用模仿学习方法训练我方的追捕智能体，敌方的动态目标直接选择强化学习训练得到的较高水平的智能体。表 1 列出了场景中实体的初始位置分布。

表 1 实体初始分布情况

Table 1 Initial distribution of entities

实体	横坐标范围	纵坐标范围
障碍物 0	(0.1,0.2)	(−0.6,−0.5)
障碍物 1	(−1.0,−0.9)	(−0.3,−0.2)
障碍物 2	(0.2,0.3)	(−0.8,−0.7)
障碍物 3	(0.6,0.7)	(−0.7,−0.6)
障碍物 4	(0.4,0.5)	(−0.2,−0.1)
动态目标 0	−1.0	−1.4
动态目标 1	−0.5	−1.2
静态目标 0	0	1.2
静态目标 1	1	−1.2
静态目标 2	−1	−1.2
我方智能体 0	(0.5,1.2)	(1.0,1.5)
我方智能体 1	(0.5,1.2)	(1.0,1.5)
我方智能体 2	(0.5,1.2)	(1.0,1.5)
我方智能体 3	(0.5,1.2)	(1.0,1.5)
我方智能体 4	(0.5,1.2)	(1.0,1.5)
我方智能体 5	(0.5,1.2)	(1.0,1.5)
我方智能体 6	(0.5,1.2)	(1.0,1.5)
我方智能体 7	(0.5,1.2)	(1.0,1.5)

如图 4 所示，采用混合模仿学习方法训练含有动静目标的复杂追捕场景。采用 TD-BC 方法学习我方智能体追捕静态目标的策略，采用 MT-GAIL 方法学习智能体追捕动态目标的策略。将学到的模型赋给对应的智能体进行推演。从追捕静态目标的智能体中依次选取智能体模型进行推演，并将得到的专家数据用于 TD-BC 训练；从追捕动态目标的智能体中选取

3 个模型进行推演，并将得到的专家数据用于 MT-GAIL 训练。最终将训练得到的模型赋予对应的我方智能体，从而实现较快学习到追捕动静结合目标的经验。

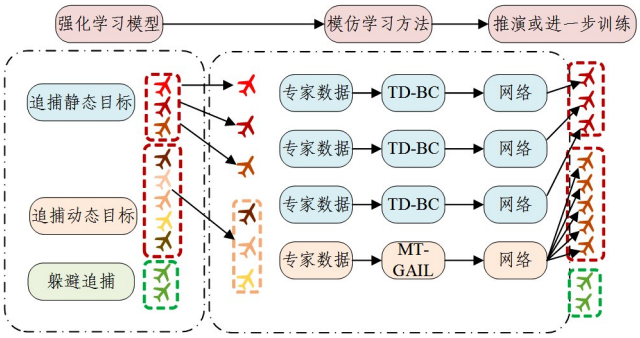


图 4 基于混合模仿学习方法训练智能体的框架图

Fig. 4 Framework diagram of training agents based on hybrid imitation learning methods

4.2 收集专家数据

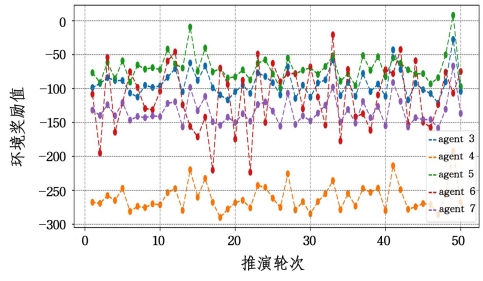
由于场景中我方智能体的初始位置是不固定的，为了使得经过模仿学习训练得到的智能体在不同的初始位置都可以追捕到敌方智能体，需要积累多条专家数据。下面介绍采用 MT-GAIL 方法和 TD-BC 方法获取专家数据的过程。

4.2.1 采用 MT-GAIL 方法获取专家数据

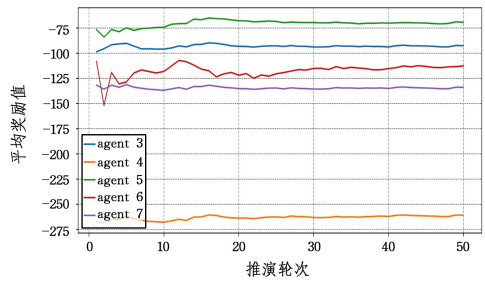
考虑到我方智能体的初始位置分布一致，针对敌方目标以及障碍物的策略也一致，因此可以认为我方追捕动态目标的智能体是等价的。场景中追捕动态目标的智能体共有 5 个，智能体的 ID 分别为 agent 3，agent 4，agent 5，agent 6，agent 7。不同智能体的决策质量并不相同，故可以认为利用不同智能体得到专家数据的质量也各不相同。因此可以采用 MT-GAIL 方法模仿学习不同质量的专家数据，使得学习得到的模型尽可能与高质量的专家策略类似，同时避免受到低质量的专家数据的影响。

结合场景推演 50 轮，记录不同智能体的环境奖励值。通过环境奖励值来评价智能体（专家模型），最终选择其中 3 个智能体的策略用于 MT-GAIL 训练。

推演过程的环境奖励值结果如图 5 所示。图 5(a)为智能体推演过程中每轮的环境奖励值，图 5(b)为智能体的累计平均环境奖励值。可以看出 agent 5 整体表现良好，具有最高的累计环境平均奖励值；agent 6 整体表现中庸；agent 4 整体表现最差，推演 50 轮的累计平均环境奖励值最低。选择上述 3 个智能体作为专家策略，与环境进行推演得到专家数据。在推演的 50 轮中，每轮的数据为 200 步，对于同一个智能体，每步的数据为状态空间 36 维、动作空间 5 维。由于专家数据不需要区分轮数，因此将每一轮得到的状态数据和动作数据进行拼接，得到 10000 组专家数据。最终形成 3 个专家文件，每个文件中包含拼接好的 10000 组专家轨迹数据。



(a)智能体每轮环境奖励值



(b)智能体平均环境奖励值

图 5 推演过程中智能体的环境奖励值

Fig. 5 Environmental reward values of agents in the process of inference

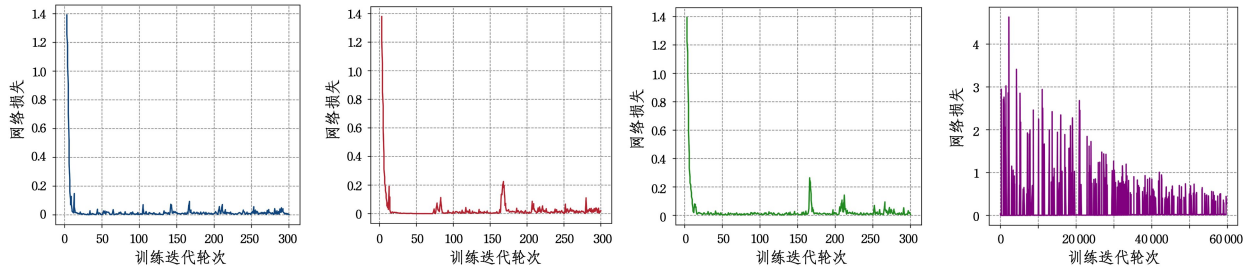
4.2.2 采用 TD-BC 方法获取专家数据

由于场景中追捕静态目标的智能体不是等价的，因此追捕静态目标的 3 个智能体需要单独训练。基于每个智能体获取专家数据的方法为：基于强化学习模型结合场景推演 50 轮，每轮的数据为 200 步。对于每一个智能体，每步的数据为状态空间 36 维、动作空间 5 维，下一时刻的状态空间 36 维。由于专家数据不需要区分轮数，因此将每一轮得到的对应数据进行拼接，得到 10000 组专家数据。最终形成 3 个文件，分别用于存储状态数据、动作数据和下一时刻的状态数据，每个文件都包含 10000 组数据。针对 agent 0，agent 1，agent 2 分别采用上述方法获取专家数据用于 TD-BC 训练。

4.3 训练效果

4.3.1 MT-GAIL 方法

利用专家数据，采用 MT-GAIL 方法进行训练，累计训练 300 轮，得到智能体网络模型和判别器网络模型。训练过程的损失曲线如图 6 所示。



(a) 判别器 0 网络损失 (b)判别器 1 网络损失 (c)判别器 3 网络损失 (d)价值网络损失

图 6 追捕动态目标训练过程中的网络损失值变化情况

Fig. 6 Changes in network loss values during the training process for pursuing dynamic targets

可以看出，训练过程中网络的损失逐渐减小到0，表明学生模型的策略与专家经验基本吻合。智能体的critic网络的损失值也在允许的范围内，通过后续强化学习训练就可以使网络的损失减小到0。

4.3.2 TD-BC方法

利用专家数据，采用TD-BC方法训练，累计训练20 000轮，得到学生模型。训练过程中智能体0的损失变化曲线如图7所示。可以看出，训练过程中价值网络和策略网络的损失逐渐减小到0，表明学生模型的策略与专家经验基本吻合。

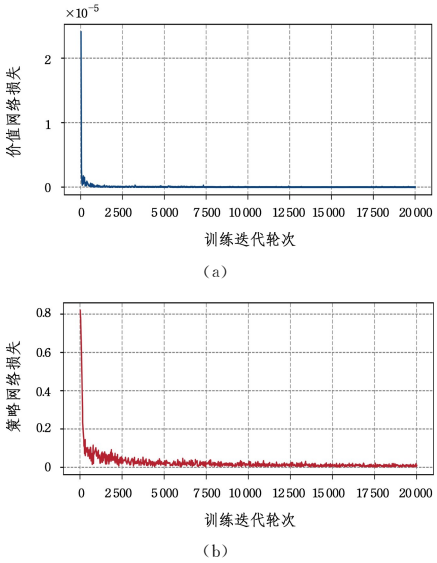


图7 追捕静态目标训练过程中的网络损失值变化情况（以智能体0为例）

Fig. 7 Changes in network loss values during the training process for pursuing static targets(taking agent 0 as an example)

4.3.3 推演效果

加载TD-BC和MT-GAIL方法得到的学生模型，将其与环境进行交互，生成推演的可视化效果如图8所示。从图中可以观察到，采用模仿学习方法，智能体已经初步展现出了追捕静态目标和动态目标的能力。这充分证明了模仿学习方法在快速学习专家经验方面的有效性，其使得智能体能够在短时间内学习到基本的追捕技巧。在推演过程中，智能体有时会出现超出边界的问题。这表明虽然模仿学习方法能够帮助智能体学习到基本的追捕经验，但其在处理一些复杂或特殊情况时，还存在一定的局限性。因此，需要进一步使用强化学习方法对智能体进行训练，以优化其行为策略，并减少超出边界等问题的发生。

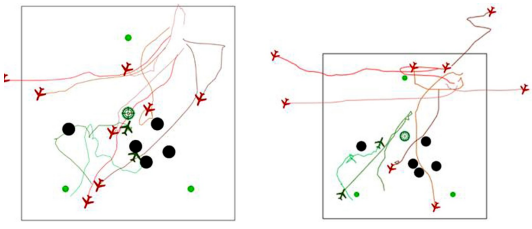


图8 推演过程可视化结果

Fig. 8 Visualization results of deduction process

采用MT-GAIL方法得到的模型可以直接用于强化学习训练。但是TD-BC方法得到的模型本质是单智能体模型，此时TD-BC方法得到的价值网络无法直接用于多智能体强化学习训练。因此需要针对学生模型产生的价值网络进行修改，以保证TD-BC方法产生的单智能体策略网络和价值网络可以用于多智能体算法中。

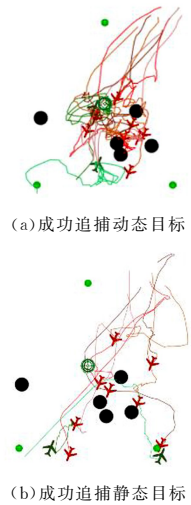
以本场景为例，价值网络的“fc1. weight”参数需要的数据格式为406 * 64，但是模仿学习得到的数据格式为41 * 64。学生模型价值网络的数据维数与多智能体强化学习训练需要的数据维数如表2所列。为了与强化学习训练所需的数据格式匹配，需要对模仿学习得到的价值网络进行修改。具体而言，对于我方的智能体，全部采用学生模型价值网络的41维数据。敌方智能体的价值网络和策略网络不需要训练，因此其取值不会影响智能体的训练。对于敌方智能体，只需要满足数据格式要求即可，所以敌方智能体采用价值网络41维数据的前39维数据。

表2 价值网络模型参数的维数

Table 2 Dimension of value network model parameters

实体	模仿学习得到的数据维数	强化学习训练需要的数据维数
智能体0	41	41
智能体1	0	41
智能体2	0	41
智能体3	0	41
智能体4	0	41
智能体5	0	41
智能体6	0	41
智能体7	0	41
敌方智能体	0	39
敌方智能体	0	39
合计	41	406

图9给出了加载模仿学习模型后，通过进一步进行强化学习训练30轮所获得的可视化效果。从图中可以明显看出，仅经过30轮的强化学习训练，智能体便成功地掌握了在不穿越边界的前提下追捕动态目标和静态目标的能力。这一结果充分展示了强化学习在微调和优化智能体动作方面的有效性，同时也凸显了模仿学习与强化学习相结合的潜在优势。



(a)成功追捕动态目标

(b)成功追捕静态目标

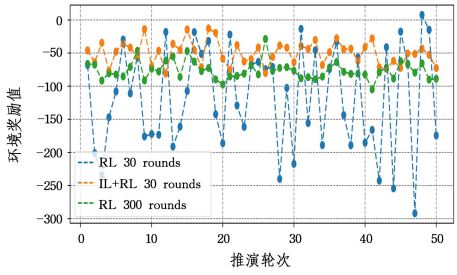
图9 加载模型训练30轮后的可视化结果

Fig. 9 Visualization results after loading the model and training for 30 epochs

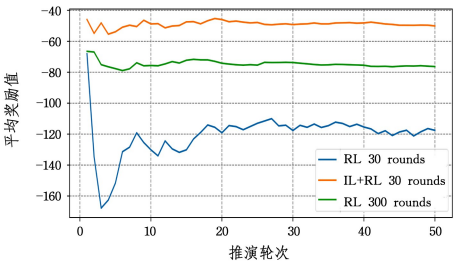
以agent 5为例，分别对比仅采用强化学习方法训练30

轮，加载模仿学习方法继续用强化学习训练 30 轮和直接采用强化学习方法训练 300 轮的模型效果。将上述模型推演 50 轮，记录每一轮的环境奖励值，并计算推演 50 轮的累计环境平均奖励值。

推演过程的环境奖励值如图 10 所示，横坐标表示推演轮次。蓝色曲线表示仅采用强化学习方法训练 30 轮的模型效果，此时的模型发挥较不稳定；绿色曲线表示采用强化学习方法训练 300 轮的效果，此时的模型表现较为稳定；橙色曲线表示加载模仿学习的模型并继续训练 30 轮的效果，此时的模型效果较好，且大多优于强化学习训练 300 轮的模型效果，说明模仿学习方法学习到了专家经验。



(a)环境奖励值



(b)平均环境奖励值

图 10 不同训练方法的环境奖励值对比(电子版为彩图)

Fig. 10 Comparison of environmental reward values of different training methods

结束语 本文提出了 TD-BC 和 MT-GAIL 方法相结合的混合模仿学习方法，该方法生成的模型可以直接用于强化学习训练，从而提高智能体的训练效率。同时该方法可以处理多质量的专家轨迹，使得模型受到低质量的专家轨迹的影响较小。结合多智能体目标追捕任务，对本文提出的混合模仿学习方法进行了实验验证。结果表明，所提方法可以汲取专家经验，为强化学习模型训练提供一个较高的起点。

虽然 MT-GAIL 算法可以处理多质量的专家轨迹，但是其训练效率与 GAIL 相比并无明显优势。后续可以通过优化 MT-GAIL 算法的结构，提高算法的训练效率。本文的研究主要基于仿真，算法和模型距离实际落地还很遥远，需要进一步验证和测试。为了缩小理论与实践之间的差距并增强算法的通用性，可以考虑通过迁移学习方法将仿真中的算法拓展到现实场景中，进一步论证所提方法的稳定性和有效性。

参 考 文 献

[1] WEN G H, YANG T, ZHOU J L, et al. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems[J]. Control and Decision, 2023, 38(5): 1200-1230.

[2] ZHANG M Y, DOU Y J, CHEN Z Y, et al. Review of deep reinforcement learning and its applications in military field[J]. Systems Engineering and Electronics, 2024, 46(4): 1297-1308.

[3] HAO J Y, SHAO K, LI K, et al. Research and Application of Game Intelligence[J]. SCIENTIA SINICA (Informationis), 2023, 53(10): 1892-1923.

[4] KHATIB O. Real-time obstacle avoidance for manipulators and mobile robots[C]// IEEE International Conference on Robotics and Automation(ICRA). IEEE, 1985: 500-505.

[5] WANG X F, GU K R. A penetration strategy combining deep reinforcement learning and imitation learning[J]. Journal of Astronautics, 2023, 44(6): 914-925.

[6] LI Y Z, SONG J M, ERMON S. InfoGAIL: Interpretable imitation learning from visual demonstrations[C]// 31st International Conference on Neural Information Processing Systems(NIPS). Cambridge: MIT Press, 2017: 3815-3825.

[7] WANG Z Y, MEREL J, REED S, et al. Robust imitation of diverse behaviors[C]// 31st International Conference on Neural Information Processing Systems (NIPS). Cambridge: MIT Press, 2017: 5326-5335.

[8] JOSH M, TASSA Y, DHRUVA T, et al. Learning human behaviors from motion capture by adversarial imitation[J]. arXiv: 1707.02201, 2017.

[9] LIN J H, ZHANG Z Z. ACGAIL: Imitation learning about multiple intentions with auxiliary classifier GANs[C]// 15th Pacific Rim International Conference on Artificial Intelligence(PRICAI). Switzerland: Springer, Cham, 2018: 321-334.

[10] RAUNAK P B, DEREK J P, BLAKE W, et al. Multi-agent imitation learning for driving simulation[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Piscataway: IEEE, 2018: 1534-1539.

[11] FU Y P, DENG X Y, ZHU Z Q, et al. Fixed-wing aircraft attitude controller based on imitation reinforcement learning[J]. Journal of Naval Aeronautical and Astronautical University, 2022, 37(5): 393-399.

[12] WANG H J, TAO Y, LU C F. A Reinforcement Imitation Learning-based Robot Navigation Method with Collision Prediction[J]. Computer Engineering and Applications, 2024, 60(10): 341-352.

[13] POMERLEAU D A. Efficient training of artificial neural networks for autonomous navigation[J]. Neural Computation, 1991, 3(1): 88-97.

[14] BOJARSKI M, TESTA D D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[J]. arXiv: 1604.07316, 2016.

[15] PFLUEGER M, AGHA A, SUKHATME S G. Rover-IRL: Inverse reinforcement learning with soft value iteration networks for planetary rover path planning[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 1387-1394.

[16] ANDREW Y N, STUART J R. Algorithms for inverse reinforcement learning[C]// 17th International Conference on Ma-

chine Learning(ICML). Association for Computing Machinery, 2000:663-670.
[17] WU S B, FU Q M, CHEN J P, et al. Meta-inverse reinforcement learning method based on relative entropy[J]. Computer Science, 2021, 48(9):257-263.
[18] JONATHAN H, STEFANO E. Generative adversarial imitation learning[C]// 30th International Conference on Neural Information Processing Systems. Curran Associates Inc, 2016: 4572-4580.
[19] JIANG C, ZHANG Z C, CHEN Z X, et al. Data efficient third-person imitation learning method[J]. Computer Science, 2021, 48(2):238-244.
[20] XIAO D M, WANG B, SUN Z Q, et al. Behavioral cloning based model generation method for reinforcement learning[C]// China Automation Congress(CAC). IEEE, 2023:6776-6781.
[21] XIAO D M, WANG B, SUN Z Q, et al. Imitation learning method of multi-quality expert data based on GAIL[C]// China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI). IEEE, 2023:8642-8647.

WANG Yanning, born in 1981, master. His main research interests is reinforcement learning.

XIAO Dengmin, born in 1999, master. Her main research interests include imitation learning and reinforcement learning.

(责任编辑:杨雪敏)

2024 年“CCF 博士学位论文激励计划”评选结果公告

CCF 博士学位论文激励计划为推动中国计算机领域的科技进步,鼓励创新性研究,激励计算机领域的博士研究生潜心钻研,务实创新,解决计算机领域中需要解决的理论和实际问题,表彰做出优秀成果的年轻学者而设立。

经评选,最终 10 篇论文(名单见附 1)入选 2024 年“CCF 博士学位论文激励计划”、3 篇论文(名单见附 2)获得 2024 年“CCF 博士学位论文激励计划”提名。

特此公告。

中国计算机学会

2024 年 12 月 11 日

附 1:2024 年“CCF 博士学位论文激励计划”入选名单

姓名	论文题目	培养单位	导师
苟向阳	图流近似处理研究	北京大学	邹磊
童率	面向广域物联网的低功耗高并发传输技术研究	清华大学	王继良
姚鹏程	面向图计算的专用加速器研究	华中科技大学	金海
代强强	继承性稠密子图枚举算法研究	北京理工大学	李荣华
何浩辰	面向性能的配置理解与缺陷检测技术研究	中国人民解放军国防科技大学	廖湘科
彭思达	动态三维人体的隐式神经表示方法研究	浙江大学	周晓巍
王涵之	大图上随机游走概率的高效计算	中国人民大学	魏哲巍
王尚文	智能化编程关键技术研究	中国人民解放军国防科技大学	毛晓光
袁牧	异构协同模型推理	中国科学技术大学	李向阳
周煊赫	自治数据库系统关键技术研究	清华大学	李国良

附 2:2024 年“CCF 博士学位论文激励计划”提名名单

姓名	论文题目	培养单位	导师
李子俊	面向服务器无感知计算系统的高效软件体系研究	上海交通大学	过敏意
汪润中	排列型组合优化问题的机器学习求解方法研究	上海交通大学	杨小康
杨传广	面向图像识别的知识蒸馏技术研究	中国科学院计算技术研究所	徐勇军

据 CCF 微信公众号