基于生成对抗网络的模仿学习综述

林嘉豪[1] 章宗长[2] 姜 冲[1] 郝建业[3),4)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](南京大学计算机软件新技术国家重点实验室 南京 210023)

[3)](天津大学智能与计算学部 天津 300050)

[4)](华为诺亚方舟实验室 北京 100085)

摘 要 模仿学习研究如何从专家的决策数据中进行学习,以得到接近专家水准的决策模型.同样学习如何决策的强化学习往往只根据环境的评价式反馈进行学习,与之相比,模仿学习能从决策数据中获得更为直接的反馈.它可以分为行为克隆、基于逆向强化学习的模仿学习两类方法.基于逆向强化学习的模仿学习把模仿学习的过程分解成逆向强化学习和强化学习两个子过程,并反复迭代.逆向强化学习用于推导符合专家决策数据的奖赏函数,而强化学习基于该奖赏函数来学习策略.基于生成对抗网络的模仿学习方法从基于逆向强化学习的模仿学习发展而来,其中最早出现且最具代表性的是生成对抗模仿学习方法(Generative Adversarial Imitation Learning,简称GAIL).生成对抗网络由两个相对抗的神经网络构成,分别为判别器和生成器.GAIL 的特点是用生成对抗网络框架求解模仿学习问题,其中,判别器的训练过程可类比奖赏函数的学习过程,生成器的训练过程可类比策略的学习过程.与传统模仿学习方法相比,GAIL 具有更好的鲁棒性、表征能力和计算效率.因此,它能够处理复杂的大规模问题,并可拓展到实际应用中.然而,GAIL 存在着模态崩塌、环境交互样本利用效率低等问题.最近,新的研究工作利用生成对抗网络技术和强化学习技术等分别对这些问题进行改进,并在观察机制、多智能体系统等方面对 GAIL 进行了拓展.本文先介绍了 GAIL 的主要思想及其优缺点,然后对 GAIL 的改进算法进行了归类、分析和对比,最后总结全文并探讨了可能的未来趋势.

关键词 模仿学习;基于生成对抗网络的模仿学习;生成对抗模仿学习;模态崩塌;样本利用效率

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.00326

A Survey of Imitation Learning Based on Generative Adversarial Nets

LIN Jia-Hao[1)] ZHANG Zong-Zhang[2)] JIANG Chong[1)] HAO Jian-Ye[3),4)]

[1)](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

[2)](*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023)

[3)](*College of Intelligence and Computing, Tianjin University, Tianjin* 300050)

[4)](*Noah's Ark Laboratory, Huawei, Beijing* 100085)

Abstract Imitation learning studies how to learn an expert-like decision model from expert decision data. Same as to learn a decision model, reinforcement learning only learns from evaluative feedback given by environment. In contrast, imitation learning is able to acquire more direct feedback from expert data. It can be classified into two types of approaches, i.e., behavioral cloning, imitation learning via inverse reinforcement learning. The imitation learning methods based on inverse reinforcement learning decompose the imitation learning process as a repeated process between estimating a reward function by inverse reinforcement learning and learning a policy upon the estimated reward function by reinforcement learning methods. The imitation learning methods based on generative adversarial nets were developed from imitation learning based on inverse

收稿日期:2018-12-24;在线出版日期:2019-08-16.本课题得到国家自然科学基金项目(61876119,61502323)和江苏省自然科学基金面上项目(BK20181432)资助.**林嘉豪**,硕士研究生,主要研究方向为模仿学习、强化学习.E-mail:wzljh3148@outlook.com.**章宗长**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为强化学习、智能规划和多智能体系统.E-mail:zzzhang@nju.edu.cn.**姜 冲**,硕士研究生,主要研究方向为模仿学习和强化学习.**郝建业**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为深度强化学习和多智能体系统.

reinforcement learning. Among them, generative adversarial imitation learning (GAIL) is the earliest and the most representative algorithm. It is inspired from generative adversarial nets consisting of two adversarial neural nets, i. e. , a discriminator and a generator. The core of GAIL is to use the structure of generative adversarial nets to address the imitation learning problem. In GAIL, the step of learning a reward function can be considered as training the discriminator, while the step of learning a policy can be viewed as training the generator. Compared to the conventional imitation learning methods, GAIL achieves better robustness, representation capability and computation efficiency. Therefore, GAIL is able to handle complicated, large-scale problems and applicable in realistic tasks. However, GAIL suffers from the problems of mode collapse and low sample efficiency in terms of environment interaction. The problem of mode collapse is derived from GANs, and it may result in the lack of diversity in the samples generated by GAIL. The problem of low sample efficiency in terms of environment interaction is derived from the assumption of stochastic policy and the model-free policy learning style in GAIL. Recently, a number of variants of GAIL have been proposed to alleviate these two problems. To alleviate the first problem, researchers have proposed to apply variants of GANs to improve GAIL, including technical improvements based on the multiple mode assumption, the generative model, etc. Representative methods are conditional GAIL, GAIL with auxiliary classifier, information maximizing GAIL (InfoGAIL), InfoGAIL from burn-in demonstrations, variational auto-encoder GAIL, etc. To alleviate the second problem, researchers have proposed to apply reinforcement learning techniques to improve GAIL, including technical improvements based on dynamic model, deterministic policy, Bayesian methods, etc. Representative methods are model-based GAIL, GAIL with deep deterministic policy gradient, Bayesian GAIL, etc. In addition to the above GAIL variants, researchers have extended GAIL to different observation mechanisms and multi-agent applications as well. The extensions of GAIL in observation mechanisms include third-person imitation learning, GAIL with recurrent policies, generative adversarial imitation from observation, etc. The extensions of GAIL in multi-agent systems include multi-agent GAIL, parameter-sharing GAIL, multi-agent adversarial imitation learning, etc, and they have been applied into realistic scenarios of autonomous driving and virtual e-commerce. In this survey, we first introduce GAIL's key ideas, advantages and disadvantages, which are followed by classifying, analyzing and comparing GAIL's improved algorithms, and finally we summarize the aiticle and discuss on possible future trends.

Keywords　imitation learning; imitation learning based on generative adversarial nets; generative adversarial imitation learning; mode collapse; sample efficiency

1　引　言

决策问题是人工智能领域中的一类重要问题.它是指寻找策略来实现既定目标的问题,如棋类游戏中棋手为获胜而思考如何落子[1],驾驶中车手为安全快捷地到达终点而规划路径[2]等.长久以来,学者一直在探求如何在决策问题中实现与人类相当甚至超人的智能决策.近年来,强化学习[3-6](Reinforcement Learning,简称 RL)方法已经在围棋[7]、Atari 电子游戏[8]等决策问题上取得了瞩目的进步.它的主要思想是使智能体在不断地与环境交互的过程中,通过从环境中获取的奖赏反馈,学习得到能最大化累积奖赏期望的策略.其中,奖赏由专家定义的奖赏函数输出.奖赏函数构建起了智能体与其目标之间的桥梁.为了使智能体达到理想的目标,奖赏函数必须要设置得恰到好处.然而对于自动驾驶等复杂的现实问题,手工设置合适的奖赏函数往往代价较高而不太现实[2].

模仿学习方法[9-10]通过模仿专家演示的样本以解决决策问题.它不需要从环境中获得奖赏反馈,其反馈信息来自于专家的决策样本.在许多实际问题

中,相较于设置合适的奖赏函数,获取专家样本往往更容易且代价更小.

模仿学习方法可以分为两类:行为克隆方法(Behavioral Cloning,简称 BC)和基于逆向强化学习的模仿学习方法(Imitation Learning via Inverse Reinforcement Learning,简称 IRL-IL).

BC[11-12]的主要思想是直接克隆专家样本在各状态处的单步动作映射,即对专家样本进行监督学习.BC 并不考虑当前状态之后的长远影响.在有足够多专家样本的前提下,它具有良好的表现.由于不考虑长远影响,BC 会将细微的误差在序贯的决策过程中逐步放大,即产生级联误差问题[2,13-14],因而在很多模仿学习任务中,鲁棒性、泛化性较差.

IRL-IL[15-16]假设专家策略等价于由未知的真实奖赏函数推导出的最优策略.从字面上理解,逆向强化学习[17](Inverse Reinforcement Learning,简称 IRL)是 RL 的逆向过程,它根据给定的专家样本求解未知的奖赏函数.基于解得的奖赏函数,IRL-IL 通过 RL 方法求解最优策略的方式,间接地还原专家策略.这种模仿专家的方式使 IRL-IL 具备了长远规划的能力.因此,IRL-IL 能有效解决 BC 的级联误差问题并表现出更强的泛化性、鲁棒性.然而,IRL-IL 存在着一些缺陷使其难以求解大规模问题.其缺陷主要为:(1)大多数 IRL-IL 方法的线性奖赏函数的假设具有很强的局限性[18];(2)在 IRL-IL 迭代求解中的 RL 子过程需要消耗大量的计算资源[19].

基于生成对抗网络的模仿学习方法(Imitation Learning Based on Generative Adversarial Nets,简称 GANs-IL)从 IRL-IL 发展而来,是一类结合了生成对抗网络的模仿学习方法[20].两者的主要区别是奖赏函数、策略的表示模型以及模型的训练方式.GANs-IL 用两个神经网络来表示 IRL-IL 中的奖赏函数和策略,并用对抗的方式来优化这两个网络的参数.原始的生成对抗网络[21-22](Generative Adversarial Nets,简称 GANs)由生成模型(又称生成器)和判别模型(又称判别器)这两个相对抗的网络模型共同构成.其中,生成模型[23]指能够产生符合期望的样本输出的模型,如根据噪声输入产生高维图片[24]或语音[25]等输出的模型.GANs 已在计算机视觉等领域中开拓了一系列有趣的工作,如图像合成[26]、图像超分辨率[27]等.

最早出现且最具代表性的 GANs-IL 方法是 Ho 等人于 2016 年提出的生成对抗模仿学习方法(Generative Adversarial Imitation Learning,简称 GAIL)[20].如果把策略表征为从状态输入到动作输出的生成模型,那么模仿学习根据专家样本学习策略的过程其实就是生成模型的训练过程.在 GAIL 中,根据输入状态输出动作的策略可类比为生成器,而根据输入专家样本或生成样本输出奖赏值的奖赏函数可类比为判别器.从而,GAIL 将求解奖赏函数的过程类比作判别器的训练过程,将策略的学习过程类比作生成器的训练过程.

GAIL 运用生成对抗网络的框架进行模仿学习以克服 IRL-IL 的缺陷,它能够在大规模的问题中表现出优异的性能.基于生成对抗网络框架,GAIL 的策略和奖赏函数模型可运用神经网络来自动抽取样本的抽象特征.因此,GAIL 具有更强的表征能力.并且,GAIL 直接将策略作为学习的目标,它运用高效的策略梯度方法训练策略模型.从而,GAIL 能避开 IRL-IL 需消耗大量计算资源的内部计算过程,具有更高效的计算能力.已有工作表明,GAIL 能够在如自动驾驶[28]、仿真及真实机器人操控[29]等复杂的大规模问题中表现出优异的性能.

然而,GAIL 仍面临着诸多瓶颈,其中模态崩塌问题[30](Mode Collapse)、环境交互样本(即利用生成模型与环境交互得到的生成样本,简称生成样本)利用效率低问题[31-32](Low Sample Efficiency in Terms of Environment Interaction)尤为突出.模态崩塌问题源于 GANs,它将导致 GAIL 产生的样本丧失多样性.生成样本利用效率低问题源于 GAIL 的随机性策略(Stochastic Policy)假设和无模型(Model-free)策略学习方式,它将导致 GAIL 无法适用于获取样本成本高的实际应用.针对模态崩塌问题,学者提出运用 GANs 的变体形式对 GAIL 进行改进.改进的方法包括基于多模态假设的改进[33-35]、基于生成模型的改进[36]等.针对生成样本利用效率低的问题,提出运用 RL 技术等对 GAIL 进行改进.改进的方法包括基于动态模型的改进[37]、基于确定性策略的改进[32]、基于贝叶斯方法的改进[38]等.

这些 GAIL 的改进方法均属于基于生成对抗网络的模仿学习方法(GANs-IL).其特点是:(1)用神经网络表示策略和奖赏函数;(2)直接学习策略模型,并用策略梯度 RL 方法优化策略模型.本文综述了主流的基于生成对抗网络的模仿学习方法.

近年来,学者们还在观察机制[39]、多智能体系统[40]等方面对 GAIL 进行了拓展.其中,在观察机制方面的拓展包括基于第三人称的方法[41]、基于上下文的方法[28]、基于观察的方法[42]等;在多智能体系统方面的拓展包括多智能体生成对抗模仿学习方

法，其场景包括自动驾驶[43]和虚拟电商[44]等.

本文的框架脉络如图 1 所示. 具体内容组织如下：第 2 节将梳理 GANs-IL 的预备知识，包括强化学习、逆向强化学习和生成对抗网络；第 3 节将介绍 GAIL 的主要思想以及存在的两个瓶颈问题：模态崩塌问题和生成样本利用效率低问题；第 4 节将介绍针对模态崩塌问题的改进方法；第 5 节将介绍针对生成样本利用效率低问题的改进方法；第 6 节、第 7 节综述生成对抗模仿学习的拓展，包括在不同观察机制下的拓展（第 6 节）以及基于多智能体系统的拓展（第 7 节）；第 8 节展望了 GAIL 的未来研究方向并总结全文.

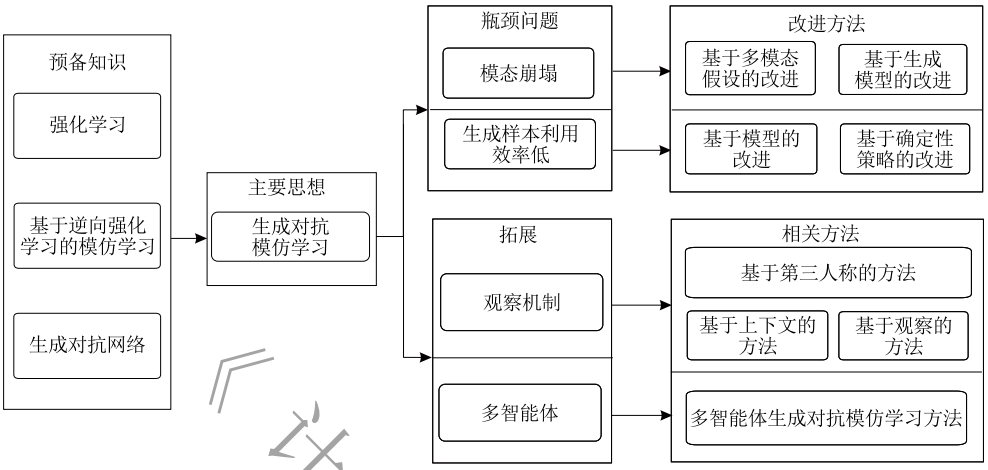

图 1　框架结构示意图

2　预备知识

本节将梳理 GAIL 中涉及到的预备知识，包括强化学习、逆向强化学习、生成对抗网络等的基本原理、相关定义以及它们存在的不足.

2.1　强化学习

一般地，强化学习（RL）由智能体（Agent）、环境（Environment）等部分组成. RL 可以建模为马尔可夫决策过程[45]（Markov Decision Process，简称 MDP）. MDP 假设决策过程满足马尔可夫性质，即智能体的决策只取决于当前的状态，而不受以往状态或动作的影响. MDP 通常被定义为一个五元组 $M=(\mathbb{S},\mathbb{A},P,r,\gamma)$. 其中：

（1）$\mathbb{S}$ 代表环境中所有状态的集合，且 $s_t\in\mathbb{S}$ 表示智能体在 t 时刻所处的状态；

（2）$\mathbb{A}$ 为智能体可选择的所有动作的集合，且 $a_t\in\mathbb{A}$ 表示智能体在 t 时刻所执行的动作；

（3）P 为状态转移函数，它表示智能体位于状态 s_t 处采取动作 a_t 转移到下一状态 s_{t+1} 的概率，可以表示为 $s_{t+1}\sim P(s_t,a_t)$；

（4）$r(s,a):\mathbb{S}\times\mathbb{A}\to\mathbb{R}$ 为立即奖赏函数，简称奖赏函数. 智能体在状态 s_t 采取动作 a_t 获得的立即奖赏值可以表示为 $r_t=r(s_t,a_t)$；

（5）$\gamma\in(0,1)$ 表示折扣因子，用于调控未来奖赏对累积奖赏值的作用效果.

RL 问题是指智能体在与环境的交互过程中通过不断试错来求解能够完成既定目标的策略（Policy）的问题. 策略是指智能体从状态到动作的映射 $\pi\in\Pi:\mathbb{S}\to\mathbb{A}$. 策略 π 一般为随机性策略. 给定随机性策略 π，智能体在状态 s_t 处所采取的动作可表示为 $a_t=\pi(s_t)$，在状态 s_t 处采取动作 a_t 的概率可表示为 $\pi(a_t|s_t):\mathbb{S}\times\mathbb{A}\to[0,1]$.

一个策略是否符合既定的学习目标或者说策略的“好坏”，是根据期望累积奖赏值来决定的. 本文将从 t 时刻开始且折扣因子为 γ 的累积奖赏值定义为：$R_t^\gamma=r_t+\gamma r_{t+1}+\cdots=\sum_{i=t}^{\infty}\gamma^{i-t}r(s_i,a_i)$. 定义策略 π 的状态值函数 $V_\pi(s)=\mathbb{E}_\pi[R_0^\gamma|s_0=s]$，定义策略 π 在状态 s 处采取某一动作 a 的动作值函数为 $Q_\pi(s,a)=\mathbb{E}_\pi[R_0^\gamma|s_0=s,a_0=a]$，定义策略值 $\eta(\pi)=\mathbb{E}_\pi[R_0^\gamma]$，定义 $\rho_\pi(s)$ 为状态 s 在智能体与环境交互过程中的占比，即出现的概率：

$$\rho_\pi(s)=p_\pi(s_0=s)+\gamma p_\pi(s_1=s)+\gamma^2 p_\pi(s_2=s)+\cdots=\sum_{t=0}^{\infty}\gamma^t p_\pi(s_t=s)\tag{1}$$

这里，$p_\pi(s_i=s)$ 表示第 i 时刻状态为 s 的概率，其中，$i=1,\cdots,U$，U 表示终止时刻. 定义 $\rho_\pi(s,a)$ 为在给定策略 π 下，状态-动作对（State-Action Pair）(s,a) 出现的概率：

$$\rho_\pi(s,a)=\pi(a|s)\sum_{t=0}^{\infty}\gamma^t p_\pi(s_t=s)=\pi(a|s)\rho_\pi(s) \quad (2)$$

在状态空间和动作空间均连续的情况下，策略值 $\eta(\pi)$ 可进一步展开成：

$$\begin{aligned}\eta(\pi)&=\int_s\rho_\pi(s)\int_a\pi(a|s)r(s,a)\mathrm{d}a\,\mathrm{d}s\\&=\mathbb{E}_{s\sim\rho_\pi,a\sim\pi}[r(s,a)]\end{aligned} \quad (3)$$

其中，$\mathbb{E}_{s\sim\rho_\pi,a\sim\pi}[\cdot]$ 表示服从状态动作折扣概率分布的期望值.

根据是否直接学习策略，RL 方法可以分为基于值函数的方法和基于策略梯度的方法. 其中基于值函数的方法通过动作值函数间接地学习策略，它服从广义的策略迭代：不断交替地进行策略评估和策略改进. 策略评估是对动作值函数进行估计的过程，经典方法有蒙特卡罗估计方法[46]和时间差分法[47]. 策略改进为根据动作值函数改进策略的过程，如贪心方法等. 然而，在无限的连续动作空间中寻找动作值最大的贪婪动作并不现实. 因此，值函数方法难以直接求解动作空间连续的问题.

基于策略梯度的方法[48]直接将策略作为学习的对象，它更为简单且计算代价更小，因而在实际中的应用范围更广. 它首先将策略参数化，即 $\pi\approx\pi_\theta$，如运用线性函数或神经网络等非线性函数近似表示策略，接着将参数朝着最大化累积奖赏值的梯度方向进行更新. Schulman 等人在 2015 年提出了基于置信域的策略优化方法[49]（Trust Region Policy Optimization，简称 TRPO），该方法能够保证策略有效地更新优化. 在此基础上，Schulman 等人提出一种不仅能使策略有效更新并且计算代价更小、更易实现的方法，即近端策略优化方法[50]（Proximal Policy Optimization，简称 PPO）. 然而，策略梯度方法存在着高方差的通病. Schulman 等人进一步提出了广义优势估计算法[51]（Generalized Advantage Estimation，简称 GAE），该算法能有效缓解算法的高方差问题.

2.2 基于逆向强化学习的模仿学习

IRL 问题一般假设能够获得专家样本，并假设专家样本由未知的真实奖赏函数对应的最优策略获得. 它是指根据专家样本求解未知的真实奖赏函数的问题[17]. 通过求解 IRL 问题得到的奖赏函数能理解专家样本数据背后的决策动机或偏好. 模仿学习通常将能产生专家样本的专家策略定义为 π_E. 根据专家策略 π_E 演示得到的轨迹样本集合 $\mathbf{T}_E$ 可表示为

$$\mathbf{T}_E=\{\tau_1,\tau_2,\cdots,\tau_i,\cdots|\pi_E\},$$

其中，τ_i 表示第 i 条轨迹样本. 每条轨迹 τ_i 可以进一步拆分成序贯的有限组状态-动作对 (s,a)：

$$\tau_i=\{s_0,a_0,s_1,a_1,\cdots,s_U,a_U|\pi_E\},$$

其中，U 代表轨迹的长度. 模仿学习通常将状态-动作对样本 (s,a) 作为训练的数据单元.

IRL 方法由 Ng 等人[17]在 2000 年提出. IRL 方法根据专家样本由未知的真实奖赏函数对应的最优策略产生的假设，将专家策略等价为由真实奖赏函数得到的最优策略. 因此，真实奖赏函数 r^* 满足不等式 $\mathbb{E}_{\pi_E}[r^*(s,a)]\geqslant\mathbb{E}_{\pi}[r^*(s,a)]$. 通过把该不等式求解的问题转换为优化问题，奖赏函数的求解过程可表示为

$$\mathrm{IRL}(\pi_E):\arg\min_r\max_\pi\mathbb{E}_\pi[r(s,a)]-\mathbb{E}_{\pi_E}[r(s,a)] \quad (4)$$

通过求解上述优化问题，得到奖赏函数 $\hat{r}\in\mathrm{IRL}(\pi_E)$. IRL 解得的奖赏函数不仅可以用于表征专家决策的动机，还能用于求解其最优策略，从而还原专家策略. 这种模仿学习被称为基于逆向强化学习的模仿学习（简称 IRL-IL）.

在 IRL-IL 中，基于奖赏函数 $\hat{r}$，最优策略可以通过 RL 方法求得. 该过程可表示为

$$\mathrm{RL}(\hat{r}):\arg\max_\pi\mathbb{E}_\pi[\hat{r}(s,a)] \quad (5)$$

如果奖赏函数 $\hat{r}$ 有足够能力来表征真实奖赏函数，$\hat{\pi}\in\mathrm{RL}(\hat{r})$ 通常能够向专家策略靠近. 综上，IRL-IL 的学习过程可以表示如下：

$$\mathrm{RL}\circ\mathrm{IRL}(\pi_E):\max_\pi\min_r\mathbb{E}_\pi[r(s,a)]-\mathbb{E}_{\pi_E}[r(s,a)] \quad (6)$$

IRL-IL 的学习过程[52]可以总结为以下 4 个步骤：(1) IRL 根据专家策略等价于真实奖赏函数对应的最优策略的假设求解奖赏函数 $\hat{r}$，$\hat{r}$ 可理解为是区分策略 $\hat{\pi}$ 和 π_E 的超平面；(2) 基于当前奖赏函数 $\hat{r}$，通过强化学习方法求解最优策略 $\hat{\pi}$；(3) 不断地迭代步骤 (1)、(2)，奖赏函数 $\hat{r}$ 将更符合真实奖赏函数 r^*，并引导 $\hat{\pi}$ 向 π_E 靠近；(4) 最终求解得到的 $\hat{r}$ 将无限接近真实奖赏函数，并且 $\hat{\pi}$ 将收敛到专家策略.

IRL-IL 的特点是先根据专家样本求解奖赏函数，再基于奖赏函数还原专家策略. 相比 BC，IRL-IL 的鲁棒性和泛化性更强. 这是因为，通过运用强化学习方法，IRL-IL 能够基于奖赏函数考虑策略的长远影响而不局限于单步的即时反馈.

不适定问题是 IRL 的一大挑战，它是指式 (4) 存在多个奖赏函数解而无唯一解. 为了缓解该问题，Ng 等人提出了启发式方法来增强 IRL 的不等式约束[17]，从而缩小了奖赏函数的求解范围. Ziebart 等人于 2008 年提出了基于最大熵原理的 IRL-IL 算法[53]. 该算法假设真实奖赏函数的最优策略具有最大的熵，从而缓解了不适定问题. 最大熵 IRL-IL 算

法可以表示为

$$\max_{\pi}\min_{r}\mathbb{E}_{\pi}[r(s,a)]-\mathbb{E}_{\pi_E}[r(s,a)]+\lambda_H H(\pi) \quad (7)$$

其中，策略的熵表示为 $H(\pi)$，它在目标函数(7)中作为额外的惩罚项，λ_H 是控制策略熵在算法中影响大小的调节系数.

除了不适定问题，由于奖赏函数的表征能力有限以及求解最优策略的子过程计算较复杂等原因，IRL-IL 难以运用于大规模的实际问题. 大多数 IRL-IL 方法假设奖赏函数是线性的，线性奖赏函数具有很强的局限性，它难以拟合复杂问题中的真实奖赏函数. 有学者提出基于如高斯过程等非线性奖赏函数的 IRL-IL[18]，该算法提升了原始 IRL-IL 中奖赏函数的表征能力. 然而，IRL-IL 由 RL 子过程导致的计算瓶颈仍没有得到解决. RL 子过程是指根据当前奖赏函数，通过 RL 方法求解最优策略的过程. 其计算量较大，并随着迭代次数递增而不断累积. 因此，将传统的 IRL-IL 用来解决大规模的实际模仿学习问题并不现实.

2.3　生成对抗网络

生成对抗网络[21]是 Goodfellow 等人于 2014 年提出的一种深度生成模型，它在深度学习领域中颇受关注. 深度学习是机器学习的一种实现方法，它利用多层神经网络对数据进行特征学习. 相较于传统机器学习，深度学习具有良好的表征能力，它能够自动获取抽象的特征[54-55]. 深度分类模型[56-57]利用训练样本及其标签数据进行监督学习，能对复杂样本给出准确的分类值. 它具有很好的感知能力，能通过多层网络结构与非线性变换，组合低层特征，形成抽象的、易于区分的高层特征，以划分样本的类别.

深度生成模型可以理解为深度分类模型的“逆向过程”，它将噪声输入的抽象高层特征还原为低层特征，从而产生高维度的生成样本来拟合训练样本. 原始的 GANs 在生成模型的训练中引入了一个二分类模型，其功能为判断输入样本是否属于专家样本. 当然，一些 GANs 的变体形式在生成模型的基础上引入其他模型，而不是分类模型[58]. 原始的 GANs 并不通过极大似然估计等方法[59]来直接地训练生成模型，而是由额外的分类模型来引导生成模型的训练过程. 因此，它能够避开极大似然估计方法中计算后验概率的复杂过程，从而在高维数据分布的学习上有显著优势. 其中，生成模型可称为生成器(Generator，简称 G)，分类模型可称为判别器(Discriminator，简称 D). 生成器和判别器二者形成博弈，该博弈目标函数 $L_{\mathrm{GANs}}(D,G)$ 可以表示如下：

$$\min_{G}\max_{D}L_{\mathrm{GANs}}(D,G)=\mathbb{E}_x[\log D(x)]+\mathbb{E}_z[\log(1-D(G(z)))] \quad (8)$$

其中，x 表示真实样本(训练样本)，z 表示噪声输入，$G(z)$ 表示生成器产生的生成样本，$D(\cdot)$ 表示判别器判别样本来自于真实样本分布的概率.

GANs 的训练框架可以直观地用图 2 进行表示. 在这个训练框架中，生成器 G 根据噪声输入 z 产生样本 $G(z)$. 判别器 D 的输入为真实样本 x 或生成器产生的样本 $G(z)$，输出为判别样本为真实样本的概率 $D(\cdot)\in(0,1)$.

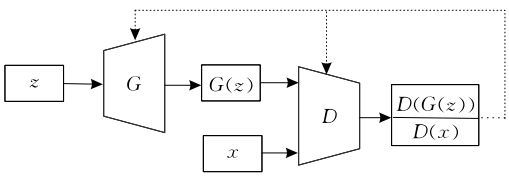

图 2　GANs 训练框架示意图

在 GANs 中，G 和 D 二者的博弈是一个相互对抗的训练过程. 该过程可分为 4 个步骤：(1) 训练 D，使 D 对样本的来源(来自真实样本分布或来自生成器)做出准确判别，D 的训练目标为最大化博弈目标函数 $L_{\mathrm{GANs}}(D,G)$；(2) 训练 G，使 G 产生逼真的样本来欺骗 D，从而使 D 的判别失准，其训练目标为最小化二者博弈的目标函数 $L_{\mathrm{GANs}}(D,G)$；(3) 通过重复步骤(1)、(2)，G 在 D 的引导下产生样本拟合真实样本分布，而 D 则寻找生成样本和真实样本的差异来不断提高判别准确度；(4) 最终，G 产生的样本可以完美地拟合真实样本分布，而 D 无法正确判别生成样本和真实样本，二者的博弈将达到纳什均衡. 此时，G 产生的样本能够以假乱真，其被 D 判别为真实样本的概率将趋近于 0.5.

以上，本文从博弈论的观点出发阐述了 GANs 的基本思想. 实际上，从信息论的角度出发，通过转换目标函数，GANs 可理解为最小化生成样本分布与真实样本分布之间的 Jensen-Shannon 散度、Kullback-Leibler 散度等的学习过程. 然而，这两个散度的数学性质并不良好. Jensen-Shannon 散度在分布不重叠时的梯度为 0，而 Kullback-Leibler 散度不具有对称性，这分别导致了 GANs 的梯度消失和模态崩塌问题.

近年来，出现了大量 GANs 的变种. 它们在不同程度上缓解了 GANs 的模态崩塌和梯度消失问题. Arjovsky 等人提出了 Wasserstein GANs[60](简称 WGANs)，该方法利用数学性质更优的 Wasserstein 散度作为度量样本分布之间散度的标准. Wasserstein

散度能够度量未发生重叠的分布的差异,并具有对称性. WGANs 缓解了经典 GANs 方法中由 Jensen-Shannon 散度所造成的梯度消失和模态崩塌问题. Nowozin 等人对原始 GANs 进行了拓展,他们将 GANs 中的 Jensen-Shannon 散度推广到 f-散度家族[61]. 其中,f-散度家族不仅有 Jensen-Shannon 散度,还包括 Kullback-Leibler 散度和 Pearson 散度等多种衡量分布差异的经典散度.

对于模态崩塌问题,学者还通过学习样本中的模态隐变量的方法对 GANs 进行了改进. Odena 等人提出了带有辅助分类器的生成对抗网络[62](Generative Adversarial Nets with Auxiliary Classifier,简称 ACGANs). ACGANs 能够有监督地学习训练样本中的模态隐变量,并拟合不同模态隐变量下的真实样本分布. Chen 等人提出了基于互信息最大化的生成对抗网络[63](Information Maximizing Generative Adversarial Nets,简称 InfoGANs). InfoGANs 通过最大化互信息的方式无监督地感知模态隐变量并拟合不同隐变量下的真实样本分布,从而缓解了 GANs 的模态崩塌问题.

随着深度学习的发展,变分自编码器[64-65](Variational Auto-Encoder,简称 VAE)等其他生成模型也有了一定拓展,其中包括结合了 GANs 的对抗原理的对抗式 VAE 算法[66].

3 生成对抗模仿学习

IRL-IL 和 GANs 在学习框架上存在许多共性[67]. GANs-IL 方法将两者进行结合. 它运用 GANs 的框架对 IRL-IL 进行了拓展. GAIL 是最具代表性的一种基于生成对抗网络的模仿学习方法. 通过改变神经网络模型、策略梯度强化学习方法等方式,GAIL 能拓展成多种不同的变体形式,这些变体形式均属于 GANs-IL 方法.

GANs-IL 主要有两个特点:(1)将策略和奖赏函数用神经网络来表示;(2)直接学习策略模型,并用策略梯度 RL 方法优化策略模型. GANs-IL 能够改进 IRL-IL 在表征能力和计算效率上的缺陷.

本节将介绍 GANs-IL 的经典方法——GAIL 的主要思想和瓶颈问题. 后文将综述基于生成对抗网络的模仿学习的一些主流方法.

3.1 主要思想

在正式提出 GAIL 之前,Ho 等人首先提出了一种能够量化奖赏函数表征能力的通用 IRL-IL 学习框架[20]:

$$\max_{\pi}\min_{r}\mathbb{E}_{\pi}[r(s,a)]-\mathbb{E}_{\pi_E}[r(s,a)]+\psi(r) \tag{9}$$

其中,$\psi(r)$表示关于奖赏函数表征能力的惩罚项. 它能解释以往 IRL-IL 方法中线性奖赏函数[2]以及凸奖赏函数[68]等在表征能力上的不足. 在该框架的基础上,他们提出了一种特殊但合理的惩罚项形式:

$$\psi_{\mathrm{GAIL}}(r)\triangleq\begin{cases}\mathbb{E}_{\pi_E}[g(r(s,a))], & \text{如果 } r>0\\ +\infty, & \text{否则}\end{cases} \tag{10}$$

其中,$g(x)$可以表示为

$$g(x)=\begin{cases}x+\log(1-e^{-x}), & \text{如果 } x>0\\ +\infty, & \text{否则}\end{cases} \tag{11}$$

惩罚项 $\psi_{\mathrm{GAIL}}(r)$的合理之处在于,它鼓励奖赏函数分配给专家策略更大的奖赏值. 其特殊之处在于,若奖赏函数满足特定形式:

$$r(s,a)=-\log D(s,a) \tag{12}$$

则恰恰将 IRL-IL 与 GANs 结合起来:输入状态输出动作的策略可类比作生成器,输入状态-动作对(s,a)输出立即奖赏值的奖赏函数可类比作判别器;策略基于当前奖赏函数的优化过程可类比为生成器的训练过程,奖赏函数的优化过程可类比为判别器的训练过程. 从而,GAIL 将 GANs 的框架运用于求解模仿学习问题. 它可以直观地表示为图 3 的形式.

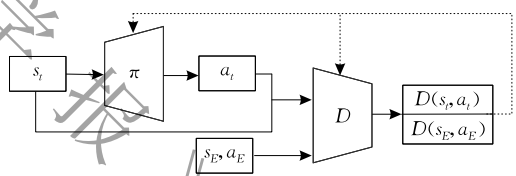

图 3 GAIL 的结构框架示意图

结合式(9)~(12),经过推导,GAIL 将 IRL-IL 中策略与奖赏函数的训练过程作为二者的博弈,其目标函数 $L_{\mathrm{GAIL}}(\pi,D)$可以表示如下:

$$\min_{\pi}\max_{D}L_{\mathrm{GAIL}}(\pi,D)=\mathbb{E}_{\pi}[\log D(s,a)]+\mathbb{E}_{\pi_E}[\log(1-D(s,a))] \tag{13}$$

其中,(s,a)表示状态-动作对,π_E表示专家策略,π表示待学习的策略,D表示判别器,$D(s,a)$表示判别器判别(s,a)由专家策略产生的概率. 为了简明表达,本文用 $\mathbb{E}_{\pi}[\cdot]$表示关于策略的期望值 $\mathbb{E}_{s\sim\rho_\pi,a\sim\pi}[\cdot]$,并省去了关于策略熵的惩罚项. Ho 等人在实验中证明了省去策略熵的目标函数对 GAIL 算法最终的学习效果影响不大.

基于 GANs 的框架、GAIL 中策略和奖赏函数均可由非线性的神经网络近似表示. 相较于 IRL-IL 和基于策略梯度的模仿学习[19]等方法,GAIL 的奖

赏函数具有更强的表征能力，它可以自动地抽取样本中的抽象特征. 因此，当神经网络的能力足够强时，策略与奖赏函数凭借深度学习的优势能适应更复杂的实际应用.

不仅如此，GAIL 直接将策略优化作为学习目标，而不是先计算奖赏函数再进行求解. 因此，GAIL 简化了 IRL-IL 迭代训练中的步骤(1)、(2)，避免了大量计算资源的浪费，从而能更好地应用于复杂的大规模问题.

在 GAIL 中，策略与奖赏函数二者相互博弈的训练过程可以分为以下 4 个步骤：(1) 运用 Adam[69] 等优化方法训练 D，使 D 分配给 π_E 尽可能大的奖赏而给策略 π 尽可能小的奖赏；(2) 运用如结合了 GAE 的 TRPO 等策略梯度优化方法训练 π，使其朝最大化累积奖赏值的梯度方向稳步更新，从而向 π_E 靠近；(3) 不断进行步骤(1)、(2)，奖赏函数将更接近真实奖赏函数并引导 π 向 π_E 靠近；(4) 最终，策略和奖赏函数达到纳什均衡. 此时，奖赏函数无法再正确地判别 π 和 π_E 的来源，π 所产生的样本分布能够完美地拟合 π_E 的样本分布，因而无需再作调整.

与 GANs 类似，GAIL 不仅可以从上述博弈论的观点进行理解，还能从信息论的角度进行解释：GAIL 实际上是最小化由策略 π 产生的状态-动作对所构成的样本分布 $\rho_\pi(s,a)$ 与专家的状态-动作对的样本分布 $\rho_E(s,a)$ 之间的 Jensen-Shannon 散度. Ho 等人基于凸共轭函数证明了 GAIL 的二者博弈问题等价于状态-动作对样本的分布差异最小化问题[20]. 该证明体现了 GAIL 与 IRL-IL 的共性：通过最小化 $\rho_\pi(s,a)$ 与 $\rho_E(s,a)$ 的差异，π 产生的 (s,a) 样本分布能最终拟合 π_E 的样本分布，从而实现对专家的模仿学习.

运用 GANs 的框架，GAIL 突破了 IRL-IL 的瓶颈，具有更强的表征能力和更高的计算效率[20]. 因此，GAIL 能解决复杂的大规模问题. 然而，GAIL 也存在着一些瓶颈问题. 这些问题制约着其在一些场景中的应用.

3.2　瓶颈问题

GAIL 存在两大瓶颈问题：模态崩塌问题、生成样本利用效率低问题.

3.2.1　模态崩塌

专家样本中存在具有一定意义的高层特征，这些特征被称为模态，如图片样本的模态可以是图片的风格. 模态崩塌是指生成模型产生的生成样本塌缩于真实样本分布的某一模态下的子分布，而无法覆盖全部真实样本分布. 以图片样本为例，模态崩塌将导致生成模型产生的图片样本只能表现出单幅画面或单一风格，而丧失了样本的多样性. 对于模仿学习而言，模态可以是专家样本的决策目标或行为范式，比如车手为追求速度而高速行驶的模态，或为追求安全而平稳行驶的模态.

在 GAIL 中，模态崩塌将导致策略产生的行为样本分布无法覆盖专家样本的全局分布，直观地表现为生成样本不具有专家样本的多样性. 大多数模仿学习方法假设专家样本来自于单一的模态，这是造成 GAIL 模态崩塌的主要原因之一. 由于专家个体的不同或偏好的不同，专家演示的样本服从多个模态下的子分布. 因此，单一模态的假设不符合实际问题. GANs 同样也存在着模态崩塌问题. 在 GANs 的技术领域中，已有大量工作对 GANs 的模态崩塌问题进行改进. 因而，借鉴 GANs 领域中的前沿技术解决 GAIL 的模态崩塌问题是可取的.

3.2.2　生成样本利用效率低

生成样本利用效率低问题是指 GAIL 对智能体与环境交互得到的样本的利用效率低. 这些样本由作为生成器的策略产生，因此称为生成样本. 在一些实际应用中，智能体与环境交互样本的采集成本是很高的，如造价昂贵的物理机器人在与环境的试错过程中将产生很大的经济成本. 因此，由于生成样本利用效率低，GAIL 难以适用于这些实际应用.

生成样本利用效率低的根本原因是 GAIL 假设策略为随机性策略并以无模型的 RL 方法来学习策略. 由于随机性策略采样动作的过程是不可微分的，因此反向传播的链式求导在策略模型 π 的动作节点处中断[31]. 在随机环境中，智能体的状态迁移过程是随机的. 由于状态迁移过程的随机性，无模型的 RL 方法将导致 GAIL 中策略的状态节点不可微分. 因此，反向传播也将在策略的状态节点处中断[32]. 由于无法端到端可微分，GAIL 中的策略 π 无法根据奖赏函数模型 D 的内部参数进行更新. 从而，π 只能通过 D 输出的立即奖赏值估计期望累积奖赏，接着获得对策略梯度的估计，即通过策略梯度方法实现更新.

在策略梯度方法中，为了准确地估计期望累计奖赏，算法需要大量智能体与环境交互的样本. 这使 GAIL 成为生成样本利用效率不高的样本密集型学习方法. 生成样本利用效率低的问题可以通过一些 RL 技术来解决.

3.3 专家样本集合

在模仿学习中,获取专家样本集合的方式主要有以下两种:(1)由人类专家示范而获得专家样本集合;(2)通过RL方法对专家手工定义的标准奖赏函数学习,得到贪婪策略,再由贪婪策略得到专家样本集合.然而,RL方法获得的贪婪策略可能不等价于最优策略.因而,这些由不同RL方法得到的贪婪策略的性能也各不相同.因此,通过RL方法得到的专家样本集合并没有形成标准.

目前,模仿学习问题多以仿真实验环境为主,如仿真小车[70]、虚拟机器人控制[71]等.对于不同的模仿学习任务,专家样本集合的获取方式并不固定.对于一些难度较大的模仿学习任务,标准的奖赏函数往往难以定义.因此,通过专家亲身示范行为动作获取专家样本集合的方式更为直接;对于一些存在危险的模仿学习任务,在虚拟环境中通过RL方法获得专家样本集合的方式更为恰当.

3.4 评价标准

模仿学习的目标是学习得到与专家尽可能相似的决策模型.因此,模仿学习的评价标准一般为学习得到的策略与专家策略的性能对比.策略的性能可由策略值 $\eta(\pi)$ 表示(其定义可见前文第2.1节).

模仿学习算法通常可以将一些经典方法(如BC、GAIL等)或者一个随机策略(Random Policy)作为评价算法性能的基准.

4 针对模态崩塌问题的改进

对于模态崩塌问题,GAIL改进方法可分为以下两类:(1)基于多模态假设的改进;(2)基于生成模型的改进.

4.1 多模态学习的假设和背景

当专家样本服从多个模态下的子分布时,模仿学习的单一模态假设将导致模态崩塌.因此,假设专家具有多个模态的模仿学习方法更为合理.事实上,多模态的模仿学习还符合了实际的应用需求:从专家样本中同时学习多种有效的模态[72].例如在自动驾驶任务中,智能体通过多模态的模仿学习方法不仅从专家样本中学习到快速的驾驶模态,还学习到安全的驾驶模态等多种不同模态.

多模态的模仿学习放宽了单一模态的假设,它假设专家样本具有多个模态:专家演示的样本不限于单一模态而是来自不同模态下的多个子分布.基于多模态模仿学习的假设,GAIL的模态崩塌问题可以得到缓解.基于多模态假设的GAIL可称为多模态的GAIL方法.

多模态的GAIL假设专家样本存在 N 种模态.定义多模态的模态集合为 $\mathbb{C}=\{c_0,c_1,\cdots,c_N\}$,其中 c_i 表示第 i 种模态,它服从先验概率 $c_i\sim p(c)$.以专家样本存在两种模态为例,专家模态集合可表示为 $\mathbb{C}=\{c_1,c_2\}$,此时专家状态-动作对样本 (s_E,a_E) 可以带有相应的模态标签数据,如 (s_E,a_E,c_1) 或 (s_E,a_E,c_2).根据模态 c_i 演示得到的轨迹可以表示为 $\tau_{c_i}=(s_0,a_0,\cdots,s_U,a_U|c_i)$,其中 s_0 和 a_0 分别表示初始状态和初始动作,U 表示轨迹的终止时刻.

如图4所示,多模态的GAIL方法可以分为两类:(1)有监督学习的多模态GAIL[33-34],包括条件GAIL(CGAIL)、带辅助分类器的GAIL(ACGAIL);(2)无监督学习的多模态GAIL[35-36],包括基于互信息最大化的GAIL(InfoGAIL)、基于变分自编码器的GAIL(VAE-GAIL).其中,VAE-GAIL是一种特殊的无监督学习方法,后文4.2节将从改进生成模型的角度对其展开介绍.

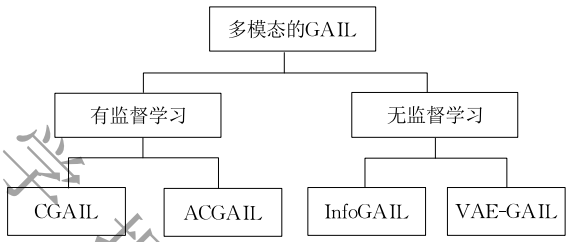

图4 多模态GAIL分类图

4.1.1 条件生成对抗模仿学习

在一些多模态模仿学习的情景中,模态的标签数据有时可以从专家样本中直接获得.这些额外的模态标签数据可作为条件约束来构建策略和奖赏函数模型.其中,一种直接的方法是在策略和奖赏函数的建模中加入关于模态标签的条件约束.该方法的主要思想来自于Mirza等人提出的条件生成对抗网络[73](Conditional GANs,简称CGANs).基于CGANs,Merel等人提出了用模态标签作为条件约束进行建模的GAIL,即条件生成对抗模仿学习[33](Conditional Generative Adversarial Imitation Learning,简称CGAIL).

CGAIL的训练框架由图5表示.其中,c 表示来自专家样本的模态标签.策略 π 在模态 c 的条件约束下生成动作,而奖赏函数 D 则在模态 c 的条件约束下,分配给策略 π 产生的状态-动作对样本 (s,a) 较低的奖赏值,而给专家样本较高的奖赏值.

因此，CGAIL 中的策略和奖赏函数二者形成了带条件约束的博弈，该博弈的目标函数 $L_{\mathrm{CGAIL}}(\pi,D,c)$ 可以表示如下：

$$\min_{\pi}\max_{D}L_{\mathrm{CGAIL}}(\pi,D,c)=\mathbb{E}_{\pi}[\log D(s,a|c)]+\mathbb{E}_{\pi_E}[\log(1-D(s,a|c))] \tag{14}$$

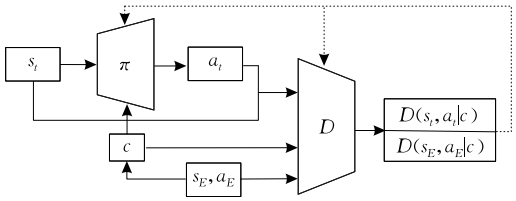

图 5　CGAIL 训练框架示意图

通过将专家样本中额外的模态标签数据作为条件约束，CGAIL 使奖赏函数在各个模态的约束下有监督地指导策略生成相应模态的样本，从而缓解了模态崩塌问题.

4.1.2　带辅助分类器的生成对抗模仿学习

丰富的标签数据可以帮助模型抽取样本中的抽象特征，从而使模型对样本的推断更准确. 但是，CGAIL 仅将模态标签作为建模的条件约束，而并不将其用于训练网络模型，它并没有充分利用样本中额外的模态标签数据.

受 ACGANs[61] 的思想启发，Lin 等人在 GAIL 的基础上加入了辅助的网络模型，提出了带辅助分类器的生成对抗模仿学习[34]（Generative Adversarial Imitation Learning with Auxiliary Classifier，简称 ACGAIL）. 通过引入辅助的网络模型，ACGAIL 不仅能有监督地对模态数据进行学习，还能充分利用额外的模态标签数据学习样本中的抽象特征. 因此，相比于 CGAIL，ACGAIL 对额外的模态数据利用效率更高.

新的辅助网络用来对样本所属的模态类别进行分类，从而帮助 GAIL 重构关于模态的条件信息，因此可称为辅助分类器（Auxiliary Classifier，简称 C）. 引入 C 之后，ACGAIL 的训练框架可表示如图 6.

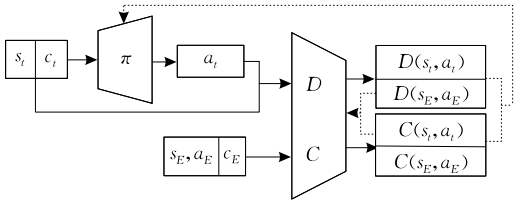

图 6　ACGAIL 训练框架示意图

其中，策略 π 根据当前状态 s_t 和标签 c_t 执行动作 $a_t=\pi(s_t,c_t)$. ACGAIL 的 D 与 GAIL 中的 D 完全一致，它们输入样本 (s,a)，输出样本来自专家样本分布的概率，并将模态视为噪声而不直接参与对模态信息的感知与学习过程. $D(s,a)$ 表示 D 判别样本 (s,a) 来自专家样本分布的概率值. 在 ACGAIL 中，分类器 C 与策略 π 和判别器 D 一样，由神经网络近似表示，它的输入为带模态标签的状态-动作对样本. 其中，c_t 表示策略 π 产生的样本 (s_t,a_t) 当前的模态标签，它由随机采样（Random Sampling）获得并服从离散均匀分布，c_E 表示专家的状态-动作对样本 (s_E,a_E) 中额外的模态标签. $C(s,a)$ 表示分类器 C 将样本划分为各模态类别的概率.

分类器 C 对模态信息是敏感的，它通过模态标签数据进行有监督的训练. C 辅助 D 重构关于模态标签的条件信息，并与 D 一同引导策略的更新，从而实现多模态模仿学习.

C 与 D 虽然分饰不同的角色，但是这两个网络模型的输入都为状态-动作对样本，并均需抽取样本中的特征. 因此，C 和 D 的网络模型可以共享隐藏层参数. 通过共享参数，ACGAIL 能够使网络模型利用额外的模态标签对样本的特征进行训练，从而使分类器和判别器更准确地分类和判别.

此时，ACGAIL 中的策略 π、判别器 D、分类器 C 形成了三者的博弈，其博弈的损失函数 $L_{\mathrm{ACGAIL}}(\pi,D,C)$ 可表示如下：

$$\min_{\pi,C}\max_{D}L_{\mathrm{ACGAIL}}(\pi,D,C)=\mathbb{E}_{\pi}[\log D(s,a)]+\mathbb{E}_{\pi_E}[\log(1-D(s,a))]+\lambda_C\{\mathbb{E}_{\pi}[H(c,C(c|s,a))]+\mathbb{E}_{\pi_E}[H(c,C(c|s,a))]\} \tag{15}$$

其中，$C(c|s,a)$ 可表示 C 将 (s,a) 分类为模态 c 的后验概率. λ_C 表示控制分类器影响程度的调节系数，一般默认为 0.5. 交叉熵 $H(c,C(c|s,a))$ 可展开为

$$H(c,C(c|s,a))=-\sum_{c\in C}p(c)\log C(c|s,a) \tag{16}$$

在 ACGAIL 的目标函数中，D 分配给专家样本较高的奖赏值，而给策略 π 产生的样本较低的奖赏值. 分类器 C 利用专家样本带有的模态标签或生成样本的当前标签进行有监督的训练，使其对样本的分类误差尽可能小. 并且，分类器 C 给分类误差小的样本较小的惩罚，给分类误差大的样本大的惩罚. 综上，分类器 C 与判别器 D 一同引导策略更新，它们构成了联合奖赏函数，表示如下：

$$r(s,a)=-\log D(s,a)-\lambda_C H(c,C(c|s,a)) \tag{17}$$

在辅助分类器 C 与判别器 D 的共同作用下，策略在模仿专家的同时，能产生拟合各种模态的专家样本子分布. 从而，ACGAIL 不仅能缓解原始 GAIL

的模态崩塌问题,实现有监督的多模态模仿学习.而且,它能够充分利用额外的模态标签数据帮助网络模型准确地抽取样本中抽象的高层特征.

4.1.3 基于互信息最大化的生成对抗模仿学习

前文介绍的 CGAIL 和 ACGAIL 能适用专家样本带有额外的模态标签数据这类问题.然而,在更多的情景下,模态标签数据并没有显式地存在于专家样本中.模态数据往往作为隐变量藏于样本之中,它无法从专家样本中直接获得.因此,由于模态标签数据在专家样本中的缺失,CGAIL 和 ACGAIL 等有监督的多模态 GAIL 方法将失效.

Li 等人提出了一种能够适用于模态标签未知问题的无监督多模态 GAIL 方法,称为基于互信息最大化的生成对抗模仿学习[35](Information Maximizing Generative Adversarial Imitation Learning,简称 InfoGAIL).InfoGAIL 将信息论中的互信息概念运用到了 GAIL 中.通过最大化互信息的原理,InfoGAIL 能增强策略产生的样本与模态隐变量之间的相关性,从而实现无监督的多模态学习.

在信息论的范畴中,互信息表示一个随机变量 x 在给定另一变量 y 后所减少的不确定性或信息量.通俗来说,互信息可以表示变量 x 与 y 之间的相关性.互信息越大,则两者越相关.其公式可表示为 $I(x;y)=H(x)-H(x,y)$.

InfoGAIL 在 GAIL 的基础上进一步考虑最大化待学习策略产生的状态-动作对与模态隐变量之间的互信息,如式(18)所示:

$$I(c;s,a)_{a=\pi(s,c)}=H(c)-H(c,(s,a)) \tag{18}$$

其中,c 表示策略 π 产生的样本中的模态隐变量,且 $c\in\mathbb{C}$.模态隐变量 c 通过随机采样获得,它服从离散均匀分布.因此,通过最大化状态-动作对与模态隐变量之间的互信息,InfoGAIL 使待学习策略产生的状态-动作对与模态隐变量的相关性极大化,从而使策略产生的行为与模态隐变量相关.

具体地,InfoGAIL 的博弈目标函数 $L_{\text{InfoGAIL}}(\pi,D,I)$ 由原始 GAIL 的博弈目标函数引入关于互信息的惩罚项形成:

$$\min_{\pi,I}\max_{D}L_{\text{InfoGAIL}}(\pi,D,I)=\mathbb{E}_{\pi}[\log D(s,a)]+\mathbb{E}_{\pi_E}[\log(1-D(s,a))]-\lambda_I I(c;s,a) \tag{19}$$

其中,λ_I 表示控制惩罚力度的系数.

由于缺乏模态标签的知识,状态-动作对关于模态隐变量的后验概率 $q(c|s,a)$ 是未知的.因此,互信息中的交叉熵 $H(c,(s,a))$ 无法直接计算.受 InfoGANs 的启发,InfoGAIL 将互信息放松为其变分下界,并且运用网络模型 Y 近似后验概率 $Y(s,a)\approx q(c|s,a)$,从而最大化互信息的变分下界.因此,结合式(18)、(19)和交叉熵式(式(16)),并将无关的常数项 $H(c)$ 消除,InfoGAIL 的目标函数可以推演为 $L_{\text{InfoGAIL}}(\pi,D,Y)$:

$$\min_{\pi,Y}\max_{D}L_{\text{InfoGAIL}}(\pi,D,Y)=\mathbb{E}_{\pi}[\log D(s,a)]+\mathbb{E}_{\pi_E}[\log(1-D(s,a))]+\lambda_I H(c,Y(c|s,a)) \tag{20}$$

其中,Y 表示推断模态隐变量后验概率的推断器,$Y(c|s_t,a_t)$ 表示 Y 推断状态-动作对样本所属模态隐变量的概率.在引入关于互信息变分下界的近似模型 Y 后,InfoGAIL 的训练框架可表示为图 7.

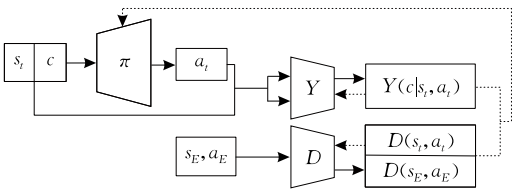

图 7 InfoGAIL 训练框架示意图

在 InfoGAIL 训练机制中,判别器 D 发挥着与原始 GAIL 中的 D 一样的功能,D 引导 π 产生的样本拟合专家样本分布.而推断器 Y 以策略 π 产生的 (s,a) 为输入,推断样本的后验概率.值得注意的是,Y 并不输入和处理专家样本.Y 遵循互信息最大化的原理,不断改进自身的推断模型,从而解释出与 π 产生的样本相关程度最大的模态隐变量.直观来说,这些模态隐变量在行为样本中具有最显著的意义.并且,Y 引导策略产生与隐变量相关的状态-动作对 (s,a).

综上,判别器 D 和推断器 Y 一同引导策略的更新,它们构成了联合奖赏函数,表示为

$$r(s,a)=-\log D(s,a)-\lambda_I H(c,Y(c|s,a)) \tag{21}$$

该联合奖赏函数引导策略模仿专家,并使得策略产生的行为与最具意义的模态隐变量相关.

基于互信息最大化原理的 InfoGAIL 能解释出样本中有显著意义的模态隐变量.并且,它能无监督地实现多模态模仿学习,从而缓解了模态崩塌问题.

实际上,InfoGAIL 还存在着等价形式和拓展形式.Hausman 等人从贝叶斯推断的角度探讨了与互信息最大化等价的多模态模仿学习[74].Kuefler 等人在 InfoGAIL 的基础上进行了拓展,提出了将专家轨迹作为老化轨迹(Burn-In Demonstrations)来推断模态隐变量的互信息最大化生成对抗模仿学

习[75]（简称 Burn-InfoGAIL）. 在 Burn-InfoGAIL 中，模态推断器训练的轨迹可分为前后两部分. 前半部分轨迹为老化轨迹，它由专家演示得到，即专家轨迹. 后半部分为待学习策略所产生的轨迹. Burn-InfoGAIL 所运用的基础算法与 InfoGAIL 大体一致. 但是，与 InfoGAIL 的模态隐变量由采样获得不同，Burn-InfoGAIL 中的模态隐变量是推断器 Y 根据前半段专家轨迹推断得出的. 在 Burn-InfoGAIL 中，当从专家轨迹突然切换到策略 π 时，π 产生的后半段轨迹将继承前半段专家轨迹的模态隐变量，从而能与专家轨迹的模态保持一致.

4.2　基于生成模型的改进方法

本节从另一视角介绍 GAIL 的改进方法：基于生成模型的改进方法. 它将 GAIL 中的生成模型改进为变分自编码器，但依然保留对抗式学习机制. 其特点是，变分自编码器能够通过自编码器的自监督学习得到模态信息. 该方法同样放宽了原始 GAIL 中单一专家模态的假设，能实现多模态的模仿学习，从而缓解模态崩塌问题. 本节着重将该方法作为基于生成模型的改进方法，目的是以改进生成模型为线索，探讨 GAIL 中生成模型其他可能的拓展形式.

4.2.1　变分自编码器

随着深度学习的快速发展，有关生成模型的研究也日趋丰富. 其中，变分自编码器[63]（VAE）是一类独具特色的生成模型. VAE 能够无监督地推断样本中的隐变量.

VAE 是指用神经网络表示编码器的一类变分贝叶斯算法. VAE 假设真实样本中的模态隐变量服从的先验概率分布为高斯分布. 在样本关于隐变量的后验概率未知的情况下，VAE 基于信息论中互信息的原理推断出藏于真实样本中的隐变量信息. 然而，由于无法获知隐变量的后验概率，VAE 通过优化样本关于隐变量的边际似然下界近似地推断后验概率.

4.2.2　基于变分自编码器的生成对抗模仿学习

受 VAE-GANs[76] 的启发，Wang 等人将 VAE 结合到了原始的 GAIL 中，提出了基于变分自编码器的生成对抗模仿学习[36]（简称 VAE-GAIL）. 类似于 InfoGAIL，VAE-GAIL 也假设专家样本中存在多个模态隐变量，并利用互信息最大化的方法无监督地推断样本中最具意义的模态. VAE-GAIL 运用 VAE 推断专家轨迹样本中的模态隐变量. 接着，VAE-GAIL 基于 CGAIL 中的对抗式学习方法将推断得出的模态变量作为奖赏函数和策略模型的条件约束，从而实现多模态学习.

在前文介绍的 InfoGAIL 中，模态隐变量通过离散均匀分布的随机采样获得. 与之不同，VAE-GAIL 中 VAE 推断得出的模态隐变量是连续的. 这是因为，VAE 假设隐变量服从连续的高斯分布. 从而，VAE-GAIL 能够在连续空间中获得平滑多变的模态隐变量，并且随着模态隐变量作为模型条件约束的动态变化，VAE-GAIL 能够学习得到介乎于各种模态隐变量下的策略，它们有趣地呈现出关于模态隐变量的策略谱.

在 VAE-GAIL 中，推断器 VAE 和策略 π 以及判别器 D 一样，均为神经网络模型. 这三部分一同构成的算法结构框架可表示为图 8. 其中，VAE 表示变分自编码或模态推断器，ε 表示噪声值，c 代表模态隐变量. 模态隐变量 c 由 VAE 推断获得，它服从连续的高斯分布. τ_E 代表专家演示的轨迹行为样本. VAE 输入专家样本，并在噪声 ε 的干扰下输出专家样本的模态隐变量 c.

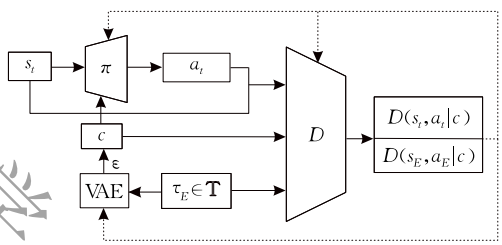

图 8　VAE-GAIL 结构框架示意图

通过不断优化专家轨迹与模态隐变量的互信息，VAE 能够学习得到专家轨迹样本中最能够代表整条专家轨迹、最具意义的模态隐变量. 并且，模态隐变量能随着噪声 ε 的波动而平滑变化. 在获得 VAE 对专家轨迹样本推断的模态变量后，VAE-GAIL 以 CGAIL 的形式将模态变量 c 作为策略 π 和判别器 D 建模的条件约束. π 在 c 的条件约束下生成状态-动作对样本，而 D 在 c 的条件约束下分配给样本立即奖赏值.

在 VAE-GAIL 中，VAE、π、D 构成了三者的博弈，它的博弈目标 $L_{\text{VAE-GAIL}}(\pi,D,\text{VAE})$ 可以表示为

$$\min_{\pi}\max_{D,\text{VAE}} L_{\text{VAE-GAIL}}(\pi,D,\text{VAE})=\mathbb{E}_{\pi}[\log D(s,a|c)]+\mathbb{E}_{\pi_E,\text{VAE}(c|\tau_E)}\left[\frac{1}{U}\sum_{i=1}^{U}\log(1-D(s_i,a_i|c))\right] \tag{22}$$

其中，U 表示专家轨迹的终止时间步，$s_{i,E}$ 和 $a_{i,E}$ 分别表示专家轨迹在 i 时刻所处的状态和采取的动作.

在 VAE-GAIL 中，奖赏函数在模态隐变量 c 的

约束下分配奖赏值：

$$r(s,a)=-\log D(s,a|c) \tag{23}$$

根据上式，策略 π 在带条件约束的奖赏函数的引导下能够实现多模态的模仿学习.

综上，VAE-GAIL 能够缓解原始 GAIL 的模态崩塌问题，并能学习到动态变化的多模态策略.

4.3 对比小结

上述两节介绍了针对 GAIL 的模态崩塌问题而提出的四种多模态改进方法，包括 CGAIL、ACGAIL、InfoGAIL、VAE-GAIL. 本节小结了这四种方法各自的特点、相似性和差异性，并在表 1 中将四种多模态 GAIL 方法进行了对比.

表 1 四种多模态 GAIL 方法对比

项目	CGAIL	ACGAIL	InfoGAIL	VAE-GAIL
学习方式	有监督	有监督	无监督	无监督
模态变量	离散变量	离散变量	离散变量	连续变量
模态推断方式	无需推断	近似推断	变分近似推断	变分近似推断
奖赏函数	带条件约束	联合构成	联合构成	带条件约束
策略中模态的来源	专家样本	随机采样	随机采样	VAE 推断
奖赏函数中模态的来源	专家样本	随机采样和专家样本	随机采样	VAE 推断
优点	直接利用模态数据	数据利用率高，推断准确	无监督学习模态变量	自监督，变化平滑
缺点	数据利用效率不高	依赖额外的模态数据	模态推断模糊	模态推断模糊

4.3.1 四种多模态 GAIL 方法总结

按照是否已知额外的模态标签数据，四种多模态的 GAIL 方法可划分为有监督与无监督的多模态学习方法.

其中，CGAIL 直接将专家样本中的模态标签数据作为条件约束进行建模. ACGAIL 通过额外的模态标签数据训练分类器，从而使分类器辅助判别器重构对模态的感知和学习. CGAIL 和 ACGAIL 均要求专家样本提供额外的模态标签数据来训练模型，它们属于有监督的多模态 GAIL 方法.

InfoGAIL 在未知模态标签数据的情景中，根据信息论中的互信息最大化理论，使得模态隐变量与策略产生的样本具有强相关性，从而实现多模态的模仿学习. VAE-GAIL 将变分自编码器运用到策略之中，能够在未知专家模态信息的情况下，通过 VAE 方法，推断出专家轨迹样本中隐含的模态变量. 接着利用 CGAIL 的技术，将推断得到的模态隐变量作为条件约束进行建模. InfoGAIL 和 VAE-GAIL 均能适用于专家样本中不存在模态标签数据的情况，属于无监督的多模态模仿学习方法.

4.3.2 CGAIL 与 ACGAIL

CGAIL 与 ACGAIL 均需获知额外的模态标签数据. 但它们对模态标签数据的利用方式并不相同. CGAIL 并不引入其他模型，它直接将模态标签数据作为条件约束，对策略与奖赏函数建模. 而 ACGAIL 中的判别器并不对模态标签数据敏感，它将模态标签数据同时用于训练额外的分类器和学习样本的抽象特征.

相比之下，CGAIL 对模态标签数据的利用更为直接，但是对标签数据的利用效率不高. 而 ACGAIL 则较为间接，它利用标签数据训练额外的网络模型. 因能将标签数据用于感知学习样本中高层特征，ACGAIL 对专家样本的利用效率更高.

在训练过程中，ACGAIL 和 CGAIL 使用的模态数据来源不同. 在策略的训练过程中，ACGAIL 的模态标签通过随机采样获得，而在奖赏函数的训练过程中，其模态标签的一部分来自随机采样，其余部分由专家样本提供. 在策略和奖赏函数的训练中，CGAIL 的模态标签数据均由专家样本直接提供.

两种方法的奖赏函数构成方式也有所不同. CGAIL 的奖赏函数是模态标签作为条件约束的判别器，而 ACGAIL 的奖赏函数则由辅助分类器与判别器联合构成.

4.3.3 ACGAIL 与 InfoGAIL

ACGAIL 是一种有监督的多模态学习方法. InfoGAIL 无需获知额外的模态标签数据，是一种无监督的多模态学习方法. 虽然两种方法的学习方法不同，但是，ACGAIL 与 InfoGAIL 二者的算法结构框架非常类似.

两种方法的模态变量的先验分布假设是一致的. 假设专家样本存在有限种模态，模态变量服从离散均匀分布. 这两种方法均通过随机采样获得模态变量.

它们都在原始 GAIL 算法结构中引入了额外的分类模型，分别为分类器 C 和推断器 Y. 其中，ACGAIL 的分类器 C 能利用已有的模态标签进行有监督的训练，而 InfoGAIL 的推断器能无监督地训练. 虽然，二者的训练方式不同，但是殊途同归. 它

们的训练目标都是最小化分类模型关于模态变量的分类误差. 不仅如此，分类器 C 和推断器 Y 均与判别器联合构成了奖赏函数.

ACGAIL 和 InfoGAIL 的主要区别是算法的性能. 由于样本中显式地存在模态标签数据，ACGAIL 的分类器 C 在有监督的训练后能精确地划分样本的模态变量. 由于样本中缺失模态标签数据，由 InfoGAIL 的推断器 Y 无监督推断得到的模态变量的意义较难解释.

4.3.4　InfoGAIL 与 VAE-GAIL

InfoGAIL 与 VAE-GAIL 均能在缺失模态标签信息的问题中实现无监督的多模态模仿学习. 它们都将互信息的原理用于实现无监督学习.

但是，它们中互信息的变量并不相同. InfoGAIL 中互信息的变量是策略 π 产生的生成样本和模态隐变量，而在 Burn-InfoGAIL 中，在轨迹的前半段，互信息的变量是专家样本和模态隐变量；在后半段，是生成样本和模态隐变量. VAE-GAIL 中互信息的变量是专家样本和模态隐变量.

InfoGAIL 与 VAE-GAIL 的一大区别是模态隐变量是否连续. InfoGAIL 的模态隐变量通过随机采样得到，它假设隐变量服从离散均匀分布. 而 VAE-GAIL 的模态隐变量由 VAE 推断得出，它假设隐变量服从连续的高斯分布. 因此，InfoGAIL 的模态隐变量是离散的，而 VAE-GAIL 的隐变量是连续的. 模态隐变量连续是 VAE-GAIL 的特点，它使 VAE-GAIL 学习得到的策略具有多样性.

两种方法的奖赏函数构成方式也是不同的. InfoGAIL 的奖赏函数由推断器和判别器联合而成，而 VAE-GAIL 的奖赏函数是以模态隐变量为条件约束而建立的.

4.3.5　VAE-GAIL 与 CGAIL

VAE-GAIL 在对抗学习中运用了 CGAIL 的技术. 因此，VAE-GAIL 与 CGAIL 具有许多共性. 它们均将模态变量作为模型的条件约束. 因此，二者奖赏函数的构成方式也完全一致.

它们的不同之处在于：在 VAE-GAIL 中，模态变量来自 VAE 对专家样本中隐含模态变量的无监督推断，而在 CGAIL 中，模态变量则直接来自专家样本.

5　针对生成样本利用效率低问题的改进

通过运用 GANs 的相关技术，GAIL 及其改进方法目前已能够解决大规模问题，甚至能实现多模态的模仿学习.

然而，GANs 的技术并不能解决 GAIL 的生成样本利用效率低的问题. 该问题是由 GAIL 的随机性策略假设和无模型的学习方式造成的. 在一些现实问题中，由于智能体与环境交互的成本较大，GAIL 通过不断试错采集大量样本进行学习是不切实际的.

RL 技术可以改进生成样本的利用效率. 其中，改进方法有：(1) 基于动态模型的改进；(2) 基于确定性策略的改进；(3) 基于贝叶斯方法的改进.

5.1　基于模型的改进

在原始 GAIL 中，由于动作的随机采样过程和状态的迁移过程均不可微分，反向传播在随机性策略网络模型中的动作节点和状态节点处中断.

为了解决 GAIL 无法端到端可微分的问题，Baram 等人运用基于模型的强化学习方法(Model-based Reinforcement Learning)对 GAIL 进行了改进. 他们提出了基于动态模型的生成对抗模仿学习方法[37](Model-based Generative Adversarial Imitation Learning，简称 MGAIL).

MGAIL 在原始的 GAIL 中引入了一个前向模型来对随机的动态环境建模. 并且，MGAIL 用重参数化技巧对策略中随机采样动作的过程建模. 从而，MGAIL 通过新引入的前向模型构建了端到端可微分的梯度计算流图. 它使策略能够根据反向传播的链式梯度法则来更新奖赏函数模型中的参数信息.

MGAIL 将判别器反向传播的梯度按动作节点和状态节点划分为关于动作 a 的偏导 $\nabla_a D$ 和关于状态 s 的偏导 $\nabla_s D$：

$$\nabla_a D=-\frac{\varphi_a(s,a)\psi(s)}{(1+\varphi(s,a)\psi(s))^2}$$

$$\nabla_s D=-\frac{\varphi_s(s,a)\psi(s)+\varphi(s,a)\psi_s(s)}{(1+\varphi(s,a)\psi(s))^2} \tag{24}$$

其中，$\varphi(s,a)=\dfrac{p(a|s,\pi_E)}{p(a|s,\pi)}$，$\psi(s)=\dfrac{p(s|\pi_E)}{p(s|\pi)}$，下标表示函数关于 a 或 s 求偏导.

针对划分后的偏导，MGAIL 分别运用重参数化[77]和前向模型方法对动作节点和状态节点的梯度期望值建模. 对于连续动作，重参数化方法用高斯模型[14]建模动作的随机采样过程. 对于离散动作，重参数化方法为 Gumbel-Softmax 方法[78]. 通过运用重参数化方法，MGAIL 无需采样大量样本以获得关于动作节点的精确梯度期望值，从而避免了在

样本不足时梯度估计的高方差问题.

MGAIL 将基于模型的 RL 方法结合到了 GAIL 中.它用前向模型近似动态环境,在此基础上,估计关于状态节点的梯度的期望值.由于动态环境模型的输入为上一个时刻的状态和动作,状态偏导还包含有关于上一时刻动作和状态的梯度,因此,状态偏导需要进一步分解.MGAIL 构建的前向模型能满足状态偏导的分解和递归计算.前向模型利用与环境交互获得的样本不断地在线训练,从而更准确地拟合动态环境.因此,在 MGAIL 中,样本不仅被用于训练策略,还被用于训练前向模型.从而,MGAIL 对样本的利用效率更高.

综上所述,MGAIL 将策略的更新目标划分为关于状态节点的更新目标和关于动作节点的更新目标.

不同的前向模型将使 MGAIL 有不同的效果.前向模型可以重构为动作子模型与状态子模型,它可以运用具有记忆能力的门控循环单元[79](Gated Recurrent Unit,简称 GRU)对状态子模型建模.基于 GRU 的 MGAIL 能有选择地记忆状态的上下文信息,获得全面的观察信息,因而有更好的性能.

MGAIL 将基于模型的 RL 方法用于缓解无法端到端可微分的问题;它运用可导的前向模型和重参数化方法分别拟合不可微分的动态环境和不可微分的随机性策略;它有效地提升了样本的利用效率,缓解了高方差的问题.从而,MGAIL 将模仿学习的应用范围拓展到了智能体与环境交互成本高的现实问题.

5.2 基于确定性策略的改进

生成样本利用效率低的根本原因是 GAIL 假设策略为随机性策略.虽然 MGAIL 运用基于模型的 RL 方法改进了 GAIL,但是它并没有根本上解决由随机性策略导致的无法端到端可微分问题.并且,在 MGAIL 中,由前向模型引起的递归计算非常复杂.

实际上,在 RL 中,策略不仅可以是随机性策略,还可以是确定性策略.通过将样本保存入经验池,基于确定性策略的 RL 方法能够重复利用样本.从而,基于确定性策略的 GAIL 能够缓解生成样本利用效率低的问题.

5.2.1 确定性策略梯度方法

确定性策略在状态 s 处采取的动作 $a=\pi(s)$ 是唯一确定的.确定性策略的目标函数可以表示为 $\mathbb{E}_{s\sim\rho_\pi}[r(s,a)|a=\pi(s)]$,其中 ρ_π 来自式(1),表示状态折扣概率分布.与随机性策略相比,由于动作是唯一确定的,确定性策略的梯度期望值不涉及关于动作的估计运算.因此,确定性策略算法能够避开随机性策略产生的高方差问题.然而,由于在任意状态处所采取的动作是唯一确定的,确定性策略无法产生试错样本,它一度被认为是无法学习到的[80].

为实现确定性策略的学习过程,Silver 等人提出了的确定性策略梯度 RL 方法[81](Deterministic Policy Gradient,简称 DPG).它是一种基于异策略(Off-policy)的行动者-评论家方法[82](Actor-Critic,简称 AC).AC 由行动者(Actor)和评论家(Critic)两部分组成.其中,行动者是根据状态执行动作的策略,评论家是在当前状态处评估动作好坏的动作值函数.AC 通过交替进行策略评估与策略改进来学习策略.策略评估是为了更准确地估计动作值而更新动作值函数的过程.策略改进是基于动作值函数即评论家来更新策略的过程.

在 DPG 中,确定性策略的策略梯度在一定条件下能够近似等价于动作值函数的梯度.因此,DPG 可根据评论家来更新行动者,从而实现策略的学习.其策略梯度的更新目标 $\nabla_\theta L_{\mathrm{DPG}}(\pi_\theta)$ 可以表示如下:

$$\nabla_\theta L_{\mathrm{DPG}}(\pi_\theta)=\mathbb{E}_{s\sim\rho_\beta}[\nabla_\theta\pi_\theta(s)\nabla_a Q_\pi(s,a)|_{a=\pi_\theta(s)}] \tag{25}$$

其中,π_θ 表示以参数 θ 为神经网络参数的策略模型,ρ_β 表示策略 β 产生的状态分布.

深度确定性策略梯度方法[83](Deep Deterministic Policy Gradient,简称 DDPG)将基于确定性策略梯度的 RL 与深度学习进行结合,它运用深度神经网络近似表示 DPG 中的行动者.在 DDPG 中,确定性策略的试错可通过在策略中加入噪声实现.策略产生的试错样本被保存在经验池中.经验池可以类比作一种记忆机制.它可以不断回忆起智能体以往的经历,重复利用以往的试错样本来更新动作值函数,进而学习策略.因此,DDPG 可以大幅提升样本的利用效率.

5.2.2 基于确定性策略的生成对抗模仿学习

受确定性策略梯度 RL 方法的启发,Blonde 等人提出了基于确定性策略的生成对抗模仿学习[32](Generative Adversarial Imitation Learning with Deep Deterministic Policy Gradient,简称 DDPG-GAIL).DDPG-GAIL 假设策略 π_θ 是确定性策略,它有效解决了由随机性策略产生的生成样本利用效率低问题.

DDPG-GAIL 将策略 π_θ 的学习过程分为两部分:DDPG 和 GAIL.DDPG 基于 AC 的学习过程与 GAIL 对抗式的学习过程十分相似[84].在 DDPG-

GAIL 的训练过程中，DDPG 使行动者 π_θ 与评论家 Q 在合作式的非零和博弈中不断优化，同时，GAIL 使生成器 π_θ 和判别器 D 在对抗式的零和博弈中不断优化. 策略 π、评论家 Q、判别器 D 三者构成的 DDPG-GAIL 算法结构框架可以表示如图 9.

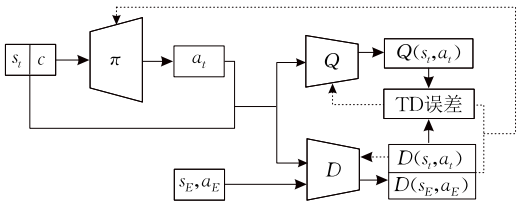

图 9　DDPG-GAIL 结构框架示意图

在 DDPG 中，评论家 Q 通过奖赏函数 D 进行学习，它引导确定性策略 π_θ 朝最大化期望累积奖赏的目标方向更新. 同时，GAIL 部分通过判别器 D 引导策略 π_θ 朝最优策略即专家策略的目标优化. 因此，对于策略 π 而言，评论家 Q 与判别器 D 扮演着相同的角色[67]. 在 DDPG-GAIL 中，策略 π_θ 同时与评论家 Q 和判别器 D 进行对抗学习，其梯度更新目标 $\nabla_\theta L_{\text{DDPG-GAIL}}(\pi_\theta)$ 可以表示为

$$\nabla_\theta L_{\text{DDPG-GAIL}}(\pi_\theta)=(1-\zeta)\nabla_\theta L_{\text{DDPG}}(\pi_\theta)+\zeta\nabla_\theta L_{\text{GAIL}}(\pi_\theta) \tag{26}$$

$$\nabla_\theta L_{\text{DDPG}}(\pi_\theta)=\mathbb{E}_{s\sim\rho_\beta}\left[\nabla_\theta\pi(s)\nabla_a Q_\pi(s,a)|_{a=\pi(s)}\right] \tag{27}$$

$$\nabla_\theta L_{\text{GAIL}}(\pi_\theta)=\mathbb{E}_{s\sim\rho_\beta}\left[\nabla_\theta\pi_\theta(s)\nabla_a \log D(s,a)|_{a=\pi(s)}\right] \tag{28}$$

其中，$\zeta\in(0,1)$ 表示平衡 DDPG 部分和 GAIL 部分对策略更新影响程度的控制因子，ρ_β 表示策略 β 的经验池中样本的状态分布.

在 DDPG-GAIL 中，确定性策略 π_θ 的动作是唯一确定的，其策略梯度 $\nabla_\theta L_{\text{DDPG}}(\pi_\theta)$ 和 $\nabla_\theta L_{\text{GAIL}}(\pi_\theta)$ 无需估计关于动作的策略梯度期望值，更不涉及不可微分的动作随机采样过程. 因此，策略 π_θ 能直接利用判别器 D 或评论家 Q 的内部参数进行更新.

受启发于 DDPG，DDPG-GAIL 将智能体与环境交互产生的样本保存在经验池中. 它通过重复利用经验池积累的样本来估计关于状态的策略梯度期望值. 因此，DDPG-GAIL 能够提高样本的利用效率. 相比 MGAIL，DDPG-GAIL 不需要构建前向模型来估计关于状态的梯度. 它保持了基于无模型 RL 方法的特点，避开了 MGAIL 中更新前向模型的复杂计算过程.

在 DDPG-GAIL 中，评论家 Q 可通过时间差分法计算贝尔曼残差进行学习，从而更准确地估计动作值. 其目标 $L(Q)$ 可混合多步贝尔曼残差计算[85]：

$$L(Q)=L_1(Q)+L_n(Q) \tag{29}$$

其中，$L_1(Q)$ 表示评论家 Q 的贝尔曼残差单步计算公式，$L_n(Q)$ 表示 n 步的贝尔曼残差计算公式.

虽然，评论家 Q 与判别器 D 扮演着相同的角色，它们都引导策略 π_θ 朝最大化累计奖赏的目标更新，即向专家策略靠近. 但是，判别器 D 提供策略 π_θ 的更新信息是即时的. 而评论家 Q 则是将判别器 D 的奖赏函数信息以时延的方式用于引导策略更新. 因此，通过控制超参数 ζ，DDPG-GAIL 能够调节奖赏函数对策略更新的延时效果.

综上，通过运用基于确定性策略梯度方法 DDPG，DDPG-GAIL 能够提升样本的利用效率，并且能够使得策略实现从奖赏函数到策略的端到端的梯度更新. 从而，DDPG-GAIL 能够更好地解决样本获取成本高的现实问题.

5.3　基于贝叶斯方法的改进

Jeon 等人提出了贝叶斯生成对抗模仿学习[38]（Bayesian Generative Adversarial Imitation Learning，简称 BGAIL），该算法将贝叶斯方法运用于 GAIL 的对抗学习中.

基于贝叶斯方法，BGAIL 对原始 GAIL 进行了转化. BGAIL 将原始 GAIL 的样本分布拟合问题转化成了模型参数的最大似然估计问题进行求解. 进一步地，BGAIL 将 GAIL 的二者博弈作为策略模型参数和奖赏函数模型参数不断迭代进行似然估计的过程.

BGAIL 假设奖赏函数模型的参数服从某一分布. 然而在分布中求解最大似然将产生很大的计算量，并且计算量将随着迭代的进行而不断累积. 因此，在整个分布中寻找似然最大的奖赏函数模型参数并不现实. 为了更高效地实现策略与奖赏函数不断迭代更新，BGAIL 选择在有限组奖赏函数模型参数中进行点估计.

多组奖赏函数通过重复利用智能体与环境交互得到的样本从而获取更多的奖赏信号. 因此，BGAIL 能够有效提高生成样本的利用效率，获得更为鲁棒的奖赏函数.

5.4　对比小结

针对 GAIL 的生成样本利用效率低的问题，本章介绍了三种改进方法，包括基于动态模型的生成对抗模仿学习（MGAIL），基于确定性策略的生成对抗模仿学习（DDPG-GAIL）及贝叶斯生成对抗模仿学习（BGAIL）. 其中，MGAIL 和 DDPG-GAIL 均引入了 RL 方法来解决 GAIL 的生成样本利用效率低的问题. MGAIL 利用基于模型的 RL 方法使得算法

能有效提高生成样本的利用效率.但 MGAIL 前向模型的更新较为复杂,算法的学习效率不高.与 MGAIL 相比,DDPG-GAIL 利用确定性策略梯度 RL 方法不仅能提高生成样本的利用效率,还能保持无模型学习的特点和更好的学习效率.不同于 MGAIL 和 DDPG-GAIL 对 RL 方法的利用,BGAIL 将贝叶斯方法用于构建损失函数.在 BGAIL 中,多个不同参数的奖赏函数模型能重复利用生成样本.因此,BGAIL 能够提高生成样本的利用效率.以上三种改进方法均能够有效地解决生成样本利用效率低的问题.

6 不同观察机制的生成对抗模仿学习

第 4 节和第 5 节分别结合 GANs 技术和 RL 技术解决了 GAIL 的模态崩塌问题和生成样本利用效率低问题.这些 GAIL 改进方法均假设智能体在环境中处于"上帝视角",即能从环境中获知完整的状态信息.然而,在一些现实的应用场景中,智能体往往只能获得不完整的状态信息,这里将不完整的状态信息称为观察(Observation),并将其定义为 o.

在不同问题中,智能体的观察机制是不同的.为了更好地适应于各种问题的观察机制,学者们将 GAIL 拓展为基于第三人称视角、基于上下文观察、从观察样本中学习等各种观察机制下的方法.

6.1 基于第三人称的生成对抗模仿学习

在实际问题中,由于时空的限制,智能体不能以上帝视角获取专家样本.专家样本只能以第三人称视角呈现.虽然,第三人称视角下的专家样本与上帝视角下的专家样本在本质上均能够表达专家的行为意图,但在不同视角下观察得到的专家样本将存在区别.比如,从不同视角拍摄同一风景,其呈现的色彩、光线阴影、形状大小等均会不同.因此,模糊化表层的视角信息能使智能体观察到更深层次的行为意图,从而使智能体可以在第三人称视角下进行模仿学习.

Stadie 等人在 GAIL 的基础上结合深度迁移学习[86]的思想提出了第三人称视角下的生成对抗模仿学习方法[41](Third-Person Imitation Learning,简称 TPIL).通过模糊化视角信息,结合了深度迁移学习的 TPIL 能满足在不同观察视角下的实际模仿学习应用需求.

TPIL 在原始 GAIL 的框架中,加入了一个对于视角信息的推断器 Y_F.TPIL 试图使得观察与视角标签之间的互信息为零,即观察与视角无关,从而无监督地模糊化观察中的视角信息:

$$I(c;Y_F(c|o))=0,$$

其中,$Y_F(\cdot)$表示用于判断观察的视角标签的推断模块,o 表示当前的观察,c 表示视角标签.通过使观察本身与观察的视角标签之间的互信息为 0,TPIL 使得观察与观察的视角无任何相关性,从而将观察中有关视角的信息模糊化甚至剔除.

TPIL 的结构框架与 InfoGAIL 非常相似,其结构框架可如图 10 所示.需要说明的是,这里 TPIL 的结构框架与文献[41]中的框架原理相同,但它们在细节上有一些不同.这里主要是做了一些简化.

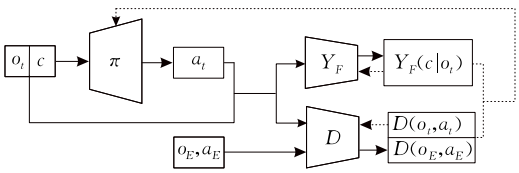

图 10 TPIL 结构框架示意图

实际上,如 InfoGAIL 等多模态的 GAIL 也运用了互信息的原理.因此,在一定程度上,TPIL 通过引入另一个推断模块所形成的算法框架与一些多模态 GAIL 算法非常相似.它们的不同之处在于,TPIL 要求互信息最小化,即使其为 0,而 InfoGAIL 等方法要求互信息最大化.不同的互信息原理运用方式使 TPIL 和 InfoGAIL 等 GAIL 的改进方法具有了不一样的功能,从而满足不同的实际任务需求.

6.2 基于上下文的生成对抗模仿学习

在许多现实应用场景中,当前观察并不能表示所有信息.因而这些应用场景不具有严格的马尔可夫性质.比如,在自动驾驶的过程中,汽车(智能体)的决策不仅依赖于当前的观察,还依赖于以往的观察.汽车结合以往的观察才能预知附近行人或汽车等的运动趋势.

当前观察结合历史观察构成了上下文观察信息,它能够表示环境的变化.上下文的运用在自然语言处理中十分常见[87].随着强化学习的不断发展,上下文信息在决策问题中也有了广泛的应用[88-89].在不破坏马尔可夫性的前提下,智能体的状态将由观察的上下文构成.因此在模仿学习中,智能体也可基于上下文观察进行决策.

Kuefler 等人把深度学习中的循环神经网络结合到策略的模型中,从而将 GAIL 拓展为能够运用历史观察数据进行决策的基于上下文的生成对抗模仿学习[28](Generative Adversarial Imitation Learning

with Recurrent Policies，简称 RP-GAIL）. 而且，他们将 RP-GAIL 应用于自动驾驶领域，并能够学习得到更为安全且性能更加稳定、高效的驾驶策略.

特别地，在 RP-GAIL 中循环神经网络由门控循环单元[90]（Gated Recurrent Unit，简称 GRU）构成. GRU 是一种具备长期记忆能力的循环神经网络. 简单来说，GRU 对原始循环神经网络的隐藏层进行了改变，它通过门控机制调控需要保留或是遗忘的记忆（历史数据）. GRU 是一种具有长期记忆功能且较易计算的循环神经网络. RP-GAIL 将 GRU 作为策略网络的一部分，并利用长期记忆能力使智能体能够基于上下文的观察信息进行决策. 从而，RP-GAIL 能够满足处理历史观察数据的需求，其训练框架可见图 11.

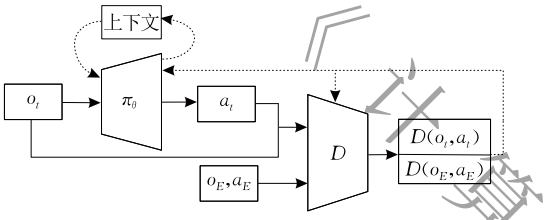

图 11　RP-GAIL 训练框架示意图

6.3　基于观察的生成对抗模仿学习

模仿学习一般将状态-动作对样本作为学习的基本数据单元. 因此，模仿学习要求专家样本包含有观察信息和动作信息. 但是在很多情况下，观察样本大量存在且极易获取，而动作数据则难以获取. 比如，YouTube 等互联网平台保存了海量的游戏视频样本，但没有保存玩家具体所执行的动作数据[91]. 因此，实现从观察样本进行模仿学习能够大大减小获取专家样本的成本，从而进一步拓宽模仿学习的应用范围并提高可行性.

近年来，已有一些工作对基于观察样本的模仿学习进行了研究[92-93]，如从观察样本学习的 BC 方法. 在这些工作中，算法的训练目标均为缩小策略产生的样本与专家样本在每个时间步中观察的差异. 然而，此类方法容易发生级联误差的问题.

Torabi 等人提出了一种能解决级联误差问题的从观察样本学习的生成对抗模仿学习[42]（Generative Adversarial Imitation from Observation，简称 GAIfO）. 他们首先提出了从观察中学习的基于逆向强化学习的模仿学习框架 IRL-ILfO，然后将该框架扩展为 GAIfO.

回顾前文，IRL 实际上是寻找一个真实奖赏函数，它能够分配给专家的状态-动作对样本高的奖赏值，而给其他策略产生的状态-动作对样本低的奖赏值. 然而，当专家样本不再包含动作数据时，传统 IRL 中的奖赏函数不再有效.

因此，Torabi 等人对奖赏函数进行了改进. 他们定义奖赏函数的输入为当前观察与下一观察，并输出立即奖赏值 $r(o,o')$. 从而，该奖赏函数能够衡量只有观察数据的样本的立即奖赏值. 基于该奖赏函数，传统 IRL 被拓展为从观察样本学习的 IRLfO(π_E)，其目标函数可以表示为

$$\mathrm{IRLfO}(\pi_E)=\arg\min_{r}\max_{\pi}\mathbb{E}_{\pi}[r(o,o')]-\mathbb{E}_{\pi_E}[r(o,o')] \tag{30}$$

IRLfO 的学习目标为从序贯的观察样本中学习专家的行为意图. 相较于 IRL，IRLfO 可以理解为寻找一个奖赏函数，它使得策略产生的从当前观察到下一观察的迁移样本能拟合专家的观察迁移样本分布. 并且，基于该奖赏函数，利用 RL 方法学习拟合专家策略的过程 RL⊙IRLfO(π_E)可以表示为

$$\max_{\pi}\min_{r}\mathbb{E}_{\pi}[r(o,o')]-\mathbb{E}_{\pi_E}[r(o,o')] \tag{31}$$

进一步地，Torabi 等人参考了从原始 IRL-IL 拓展到 GAIL 的方式，利用奖赏函数：

$$r(o,o')=-\log D(o,o') \tag{32}$$

将从观察样本中学习的 IRL-ILfO 推演到了从观察样本中学习的生成对抗模仿学习 GAIfO 的目标函数 $L_{\mathrm{GAIfO}}(\pi,D)$：

$$\min_{\pi}\max_{D}L_{\mathrm{GAIfO}}(\pi,D)=\mathbb{E}_{\pi}[\log D(o,o')]+\mathbb{E}_{\pi_E}[\log(1-D(o,o'))] \tag{33}$$

在 GAIfO 中，奖赏函数分配给专家的观察迁移样本更高的奖赏值，而策略则试图对抗式地产生与专家相似的观察迁移样本，从而获得更大的奖赏值. 因此，GAIfO 中策略产生的观察迁移样本的分布能够拟合专家观察迁移样本分布，并最终实现从观察样本学习的模仿学习.

从信息论的角度出发，GAIfO 可以理解为最小化策略产生的从当前观察到下一观察的迁移样本分布 $\rho_\pi(o,o')$ 与专家观察的迁移样本 $\rho_E(o,o')$ 的 Jensen-Shannon 散度.

7　多智能体中的生成对抗模仿学习

本节首先介绍多智能体问题的背景知识，包括环境中存在多个智能体的马尔可夫博弈以及基于马尔可夫博弈的多智能体 RL. 接着，叙述了多智能体 GAIL 的主要思想. 最后，综述了多智能体 GAIL 在

自动驾驶、电子商务平台等问题中的应用拓展.

7.1 多智能体问题的背景知识

7.1.1 马尔可夫博弈

大部分传统 RL 假设环境模型为马尔可夫决策过程(MDP). MDP 较为理想地定义了单一智能体与环境交互的过程. 然而对于许多实际问题,智能体并非单独存在于环境中,正如人类个体依存于其他个体和社会. 当环境中同时存在多个智能体时,MDP 将不再适用. 这是因为,从某一智能体的视角出发,其他智能体策略的改变将被作为环境变化的一部分. 因此,在多智能体环境中,智能体与外部环境的交互过程无法在训练阶段和测试阶段保持稳态,从而违背了马尔可夫性质.

Littman 于 1994 年提出将马尔可夫博弈(Markov Game)作为多智能体学习问题的假设模型. 马尔可夫博弈结合了博弈论思想,将马尔可夫决策过程拓展到了多智能体学习问题. 马尔可夫博弈假设环境中存在 k 个智能体. 定义 $s_i \in \mathbb{S}_i$ 为第 i 个智能体的状态. k 个智能体的联合状态可表示为向量 $\boldsymbol{s}=[s_1,\cdots,s_k]^{\mathrm{T}}$. $a_i \in \mathbb{A}_i$ 表示第 i 个智能体采取的动作. 假设第 i 个智能体的策略为随机性策略,定义为 $\pi_i: \mathbb{S}_i \times \mathbb{A}_i \to [0,1]$. k 个智能体的联合动作可表示为向量 $\boldsymbol{a}=[a_1,\cdots,a_k]^{\mathrm{T}}$. 状态转移函数为 $P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \to [0,1]$,它表示在联合状态 $\boldsymbol{s}$ 处采取联合动作 $\boldsymbol{a}$ 转移到下一个联合状态 $\boldsymbol{s}'$ 的概率. 各个智能体的奖赏根据联合状态和联合动作确定:$r_i(\boldsymbol{s},\boldsymbol{a}) \to \mathbb{R}$. 定义第 i 个智能体从 t 时刻开始且折扣因子为 γ 的累积奖赏值为

$$R_{i,t}^{\gamma}=r_{i,t}+\gamma r_{i,t+1}+\cdots=\sum_{j=t}^{\infty}\gamma^{j-t}r_i(\boldsymbol{s}_j,\boldsymbol{a}_j) \tag{34}$$

7.1.2 多智能体强化学习

在单智能体 RL 任务中,智能体的目标为学习一个使期望累积奖赏最大的最优策略. 在以马尔可夫博弈为模型的多智能体学习任务中,智能体的学习目标依旧为寻找一个使得其期望累积奖赏值最大的策略. 然而在多智能体 RL 中,由于受其他智能体策略变化的影响,一个智能体并不能够获得稳定的最优策略. 以石头、剪刀、布的双人博弈问题为例,一个玩家固定地采取石头或剪刀或布的动作都不是最优策略.

Hu 与 Wellman 在 1998 年提出将纳什均衡[94]的原理运用到多智能体学习任务中[95]. 纳什均衡是指在多方博弈的过程中,当其他智能体不改变策略时,当前智能体无法再通过调整策略获得更高收益的状态. 通过引入纳什均衡的思想,Hu 等人避免在多智能体问题中求解稳定的最优策略,而寻找在马尔可夫博弈中满足纳什均衡的策略. 当然,该策略的目标仍为使得当前智能体的累积奖赏值最大化. 寻找满足纳什均衡的策略的过程可以定义为如下带约束的优化问题 $L_r(\pi,V)$:

$$\begin{aligned}&\min_{\pi,V} L_r(\pi,V)=\sum_{i=1}^{k}\Big[\sum_{\boldsymbol{s}\in S}V(\boldsymbol{s})-\mathbb{E}_{a_i\sim\pi_i}Q_i(\boldsymbol{s},a_i)\Big] \\ &\text{s.t. } V_i(\boldsymbol{s})\geqslant Q_i(\boldsymbol{s},a_i) \\ &\qquad \equiv \mathbb{E}_{\pi_{-i}}\Big[r_i(\boldsymbol{s},\boldsymbol{a})+\gamma\sum_{\boldsymbol{s}'\in S}P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})V_i(\boldsymbol{s}')\Big]\end{aligned} \tag{35}$$

其中,π_{-i} 表示除第 i 个智能体外的 $k-1$ 个智能体的联合策略. 式中,条件约束将限制智能体的状态值函数大于等于动作值函数,即采取任何其他动作均无法获得最优策略. 此时,多智能体将达到纳什均衡.

7.2 多智能体生成对抗模仿学习

传统单智能体的 IRL 存在着不适定问题,即存在多个奖赏函数解能够满足专家策略为最优策略的假设. 然而,在多智能体的问题背景下,由于智能体受其他智能体策略变化的影响而没有稳定的最优策略,IRL 因此将面临更严重的不适定问题. Song 等人[96]讨论了在多智能体问题情景下的 IRL-IL,并给出了能够适用于多智能体学习的假设约束. 其核心思想是,在某智能体的视角下,其他智能体的策略均为专家策略或已满足纳什均衡的策略. 根据该假设,多智能体的 IRL-IL 化解了在多智能体问题中求解策略时由纳什均衡产生的不适定问题.

在此基础上,他们将多智能体 IRL-IL 进一步延伸为多智能体生成对抗模仿学习(Multi-Agent Generative Adversarial Imitation Learning,简称 MAGAIL).

MAGAIL 假设环境中存在 k 个智能体,并有相应的 k 个判别器. 其中,每个判别器均对相应智能体的策略与该智能体的专家策略进行评分,并尽可能地给予专家策略较高的分值,同时给予智能体的策略较低的分值. 每个智能体则尽可能产生能够欺骗判别器的行为,从而在判别器的引导下实现对专家策略的模仿学习. MAGAIL 的优化目标满足于 $L_{\mathrm{MAGAIL}}(\pi,D)$:

$$\begin{aligned}&\min_{\pi}\max_{D} L_{\mathrm{MAGAIL}}(\pi,D)= \\ &\mathbb{E}_{\pi}\Big[\sum_{i=1}^{k}\log D_i(\boldsymbol{s},a_i)\Big]+\mathbb{E}_{\pi_E}\Big[\sum_{i=1}^{k}\log(1-D_i(\boldsymbol{s},a_i))\Big]\end{aligned} \tag{36}$$

其中,a_i 表示第 i 个智能体的动作,D_i 表示第 i 个智

能体的判别器(起着奖赏函数的作用).在多智能体的学习问题中,智能体相互之间的关系存在着一定的先验假设.比如,各个智能体之间存在着合作、竞争或相混合的假设.如图 12,在不同的假设前提下,多智能体问题中的判别器存在不同的假设形式:

(1)集中式.当多智能体之间符合完全合作的关系时,MAGAIL 中的智能体实际上共享着同一个判别器.此时,这种特殊情况可以被理解为原始的 GAIL,而其学习得到的联合策略能够应用于所有智能体.

(2)分布式.当智能体之间没有存在奖赏的相关性假设时,每个智能体对应的判别器将采取各不相同的评分标准.然而,这些判别器由于不断地与环境进行间接的交互,它们相互之间也并非是完全独立的.

(3)零和博弈式.假设两个智能体之间处于完全竞争的关系,那么它们收到的奖赏互为相反数.在零和博弈中,智能体不需环境进行额外的交互,判别器直接对智能体与专家的联合样本进行判别训练.因此在图 12 的零和博弈式中,不涉及状态转移函数.此时,两个智能体服从零和博弈的形式并且满足以下不等式:

$$V(\pi_{E_1},\pi_2)\geqslant V(\pi_{E_1},\pi_{E_2})\geqslant V(\pi_1,\pi_{E_2}) \tag{37}$$

其中,$V(\pi_1,\pi_2)=\mathbb{E}_{\pi_1,\pi_2}[r_1(\boldsymbol{s},\boldsymbol{a})]$表示智能体 1 的期望收益.

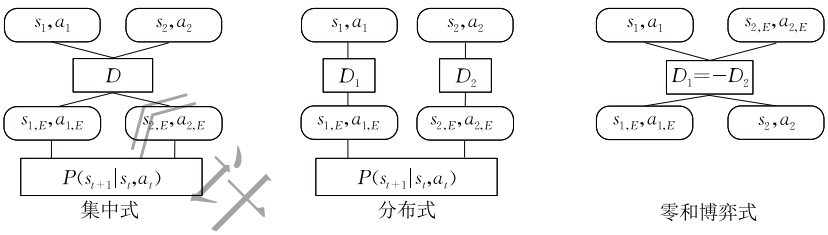

图 12　多智能体问题的假设形式

MAGAIL 结合了一种带克罗内克因子置信区间[97]的多智能体行动者—评论家策略梯度优化方法(Multi-agent Actor-Critic with Kronecker-factors,简称 MACK).MACK 与多智能体版本的 DDPG(Multi-Agent DDPG,简称 MADDPG)[98]有几分相似,它利用来自其他智能体的经验样本对动作值函数进行集中地训练,并分布式地执行行动者的动作,即利用当前智能体的独立状态采取动作.

在合作任务和竞争任务等多智能体实验中,MAGAIL 能够获得比 BC 方法更好的策略.在 MAGAIL 中,分布式的判别器能够同时适应于合作和竞争两种任务.

7.2.1　典型应用一:自动驾驶

自动驾驶作为人工智能的技术应用正在逐渐发展并已开始进入到人们的日常生活中,且将会在未来发挥更为重要的作用.

由于不需要预先定义奖赏函数,运用人类专家驾驶策略进行驾驶模拟的模仿学习是自动驾驶中相对成熟的一种技术.然而,在现实中驾驶时,任何微小的差错都将加重实际成本甚至有酿成车祸的危险.传统的模仿学习由于建模于只有单智能体的问题环境,将极有可能由于问题环境中其他车辆的动态变化而造成策略上的误差或失误.因此,在自动驾驶等现实任务中运用多智能体学习方法尤为必要.

为了实现安全、有效、经济的自动驾驶技术,Bhattacharyya 等人提出了基于参数共享的多智能体生成对抗模仿学习方法[43](Parameter-Sharing Generative Adversarial Imitation Learning,简称 PS-GAIL).该方法在 GAIL 的基础上结合了一种基于参数共享的集中式多智能体策略梯度优化学习方法[99].

多智能体的策略梯度优化学习方法总体上可以划分成两类:(1)分布式策略学习方法;(2)集中式策略学习方法.

分布式的策略学习方法是指在多智能体学习中,策略只根据当前某智能体所处的状态采取行动,而不是根据所有智能体的联合状态采取行动.前文 7.2 节中提到的 MACK 方法就是一类典型的分布式多智能体策略学习方法.

集中式策略学习方法是指策略根据多智能体的联合状态采取联合动作.参数共享的置信域策略优化方法(Parameter-Sharing Trust Region Policy Optimization,简称 PS-TRPO)[99]就是一类集中式多智能体策略学习方法,它建立在能够实现策略单调稳定更新的 TRPO 之上.PS-TRPO 利用基于参数共享的神经网络表示策略模块.在各智能体没有任何显式沟通的前提下,PS-TRPO 可以加快相互合作的多个智能体的学习过程.并且,借助于策略网

络之间共享参数的方式，PS-TRPO 能够拥有更好的样本利用效率.这是因为相较于各智能体策略完全相互独立的训练方式，参数共享有效地减少了神经网络中所需训练的参数量.

PS-GAIL 结合 PS-TRPO 的集中式策略梯度学习方法，将 GAIL 延伸到了环境中具有多个智能体的模仿学习任务.借助于 PS-TRPO 集中式策略学习方法的优势，PS-GAIL 能够基于多智能体的联合状态进行学习.从而，在训练阶段感知其他智能体的动态变化.这使训练得到的策略在真实环境的测试阶段能够有效地根据其他智能体的变化随机应变.从而，规避了单智能体学习方法在多智能体问题中由于无法感知并应对其他智能体的变化而导致级联误差的风险.

PS-GAIL 的运用需要具备一定前提，它需要满足各个智能体共用同一个策略的假设.实际上，该假设即符合 7.2 节所提到的集中式判别器形式.

7.2.2 典型应用二：虚拟电商

在实际的电商任务中，电商的搜索推荐平台每时每刻都承载着巨大的商品交易额.当平台的搜索算法需要进行调整时，公司将不可避免的面临着损失大量交易额的风险.为了测试更高效的算法并且尽可能减小在现实中的风险，构建一个能够较好拟合真实应用场景的虚拟平台是解决方案之一.多智能体模仿学习不仅可以应用于自动驾驶任务中，还可以应用于构建电子商务平台等需要与用户环境进行交互的推荐决策系统.

Shi 等学者提出了虚拟淘宝(Virtual Taobao)平台来拟合现实中电商搜索平台的应用场景.他们使用了基于电商平台的多智能体对抗模仿学习方法[44](Multi-Agent Adversarial Imitation Learning，简称 MAIL).MAIL 是一种面向多智能体的训练方式，可用于训练用户策略和平台系统策略.以这种方式得到的用户策略能够包含不同的搜索引擎策略.

同时，他们还提出了新的方法来缓解模态崩塌问题.在生成器的训练过程中，该方法最小化生成样本的模态隐变量分布与专家分布的 Kullback-Leibler 散度.

在虚拟淘宝中，搜索平台和用户均作为智能体而存在.在这个多智能体构成的学习问题中，搜索平台的状态包含了用户状态以及相关动作信息，而搜索平台的动作是输出商品推荐向量.对于用户而言，用户的状态是当前平台中推荐的商品与页面信息，用户的动作则是选择购买、离开、继续翻页等与平台交互的操作.

虚拟淘宝运用生成对抗网络拟合用户的基本属性特征.进一步地，由于观察到电商中用户与搜索平台之间关系的相互依赖性，虚拟淘宝运用 MAIL 重现用户与搜索平台之间的多智能体交互行为.

在学习搜索平台交互动作的同时，更为重要的是，虚拟淘宝能够模拟得到用户的属性特征，以及用户对于平台不同的推荐策略反应出的交互动作.从而，虚拟淘宝比较完善地构建了一个虚拟的电商世界.

8 总结与展望

相较于强化学习(RL)方法，模仿学习方法不需要专家手工设置合适的奖赏函数.它模仿专家演示的样本，从而学习得到与人类相当的策略.基于生成对抗网络的模仿学习(GANs-IL)是模仿学习的一类重要方法，它将基于逆向强化学习的模仿学习(IRL-IL)推广到了更复杂的大规模问题，使得模仿学习方法能够解决现实的应用问题.并且，随着生成对抗网络(GANs)和 RL 等技术的不断发展，GANs-IL 中遇到的模态崩塌与生成样本利用效率低等问题将以更为有效的方式得到解决.从而，GANs-IL 能够稳定、有效地解决实际问题.

本文首先介绍了最早出现且最具代表性的 GANs-IL 方法，即生成对抗模仿学习(GAIL)的核心思想，然后分析了其所面临的模态崩塌和生成样本利用效率低等问题.接着从这两个问题出发综述了前沿的改进工作.其中，针对模态崩塌的问题综述了结合 GANs 技术的改进方法，针对生成样本利用效率低的问题综述了结合强化学习等技术的改进方法.最后综述了 GANs-IL 在不同观察环境和多智能体等方面的拓展.

综上，本文将 GANs-IL 的未来发展前景作如下展望：(1)结合 GANs 技术的发展，GANs-IL 在模态崩塌问题将进一步改善，并且二者博弈的训练过程将更为稳定，且易于训练[100]；(2)结合 RL 技术的发展，GANs-IL 将在多智能体方面、部分可观察的马尔可夫决策过程[101]等方面有更深的应用拓展；(3)随着深度学习的发展，GANs-IL 将具有更强的表征能力，并能应用于需要感知复杂状态的实际问题.不仅如此，在决策问题中，对抗式学习的运用

将更为广泛，如将对抗式的模仿学习结合在 RL 中[102]，或将对抗式学习用来使得 RL 具有更强的泛化性和鲁棒性[103].

致　谢　在此感谢南京大学的俞扬教授参与本文的讨论并给出了大量的修改意见！

参　考　文　献

[1] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484-489

[2] Abbeel P, Ng A Y. Apprenticeship learning via inverse reinforcement learning//Proceedings of the 21st International Conference on Machine Learning (ICML). Banff, Canada, 2004: 1-8

[3] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge, USA: MIT Press, 1998

[4] Liu Quan, Zhai Jian-Wei, Zhang Zong-Zhang, et al. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1): 1-27(in Chinese)
(刘全，翟建伟，章宗长等. 深度强化学习综述. 计算机学报，2018，41(1)：1-27)

[5] Gao Yang, Chen Shi-Fu, Lu Xin. Research on reinforcement learning technology: A review. Acta Automatica Sinica, 2004, 30(1): 86-100(in Chinese)
(高阳，陈世福，陆鑫. 强化学习研究综述. 自动化学报，2004，30(1)：86-100)

[6] Zhao Dong-Bin, Shao Kun, Zhu Yuan-Heng, et al. Review of deep reinforcement learning and discussions on the development of computer Go. Control Theory and Applications, 2016, 33(6): 701-717(in Chinese)
(赵冬斌，邵坤，朱圆恒等. 深度强化学习综述：兼论计算机围棋的发展. 控制理论与应用，2016，33(6)：701-717)

[7] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550(7676): 354-359

[8] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning//Proceedings of the Workshops at the 27th Neural Information Processing Systems (NIPS). Lake Tahoe, USA, 2013: 201-220

[9] Byrne R W, Russon A E. Learning by imitation: A hierarchical approach. Behavioral and Brain Sciences, 1998, 21(5): 667-721

[10] Schaal S. Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences, 1999, 3(6): 233-242

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

[12] Bojarski M, Del Testa D, Dworakowski D, et al. End-to-end learning for self-driving cars. arXiv preprint arXiv: 1604.07316, 2016

[13] Ross S, Bagnell D. Efficient reductions for imitation learning//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS). Sardinia, Italy, 2010: 661-668

[14] Ross S, Gordon G J, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, USA, 2011: 627-635

[15] Li Yao-Yu, Zhu Yi-Fan, Yang Feng, et al. Inverse reinforcement learning based optimal schedule generation approach for carrier aircraft on fight deck. Journal of National University of Defense Technology, 2013, 35(4): 171-175(in Chinese)
(李耀宇，朱一凡，杨峰等. 基于逆向强化学习的舰载机甲板调度优化方案生成方法. 国防科技大学学报，2013，35(4)：171-175)

[16] Jin Zhuo-Jun, Qian Hui, Chen Shen-Yi, Zhu Miao-Liang. Survey of apprenticeship learning based on reward function learning. CAAI Transactions on Intelligent Systems, 2009, 4(3): 208-212(in Chinese)
(金卓军，钱徽，陈沈秩，朱森良. 回报函数学习的学徒学习综述，智能系统学报，2009，4(3)：208-212)

[17] Ng A Y, Russell S J. Algorithms for inverse reinforcement learning//Proceedings of the 17th International Conference on Machine Learning (ICML). Stanford, USA, 2000: 663-670

[18] Levine S, Popovic Z, Koltun V. Nonlinear inverse reinforcement learning with Gaussian processes//Proceedings of the 25th Neural Information Processing Systems (NIPS). Granada, Spain, 2011: 19-27

[19] Ho J, Gupta J K, Ermon S. Model-free imitation learning with policy optimization//Proceedings of the 34th International Conference on Machine Learning (ICML). New York, USA, 2016: 2760-2769

[20] Ho J, Ermon S. Generative adversarial imitation learning//Proceedings of the 30th Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 4565-4573

[21] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the 28th Neural Information Processing Systems (NIPS). Montreal, Canada, 2014: 2672-2680

[22] Wang Kun-Feng, Gou Chao, Duan Yan-Jie, et al. Generative adversarial networks: The state of the art and beyond. Acta Automatica Sinica, 2017, 43(3): 321-332(in Chinese)
(王坤峰，苟超，段艳杰等. 生成式对抗网络 GAN 的研究进展与展望. 自动化学报，2017，43(3)：321-332)

[23] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554

[24] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis//Proceedings of the 33rd International Conference on Machine Learning(ICML). New York, USA, 2016: 1060-1069

[25] van den Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio//Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW). Sunnyvale, USA, 2016: 125-125

[26] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv: 1809.11096, 2018

[27] Ledig C, Theis L, Huszar F, et al. Photo-realistic single image super-resolution using a generative adversarial network// Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 105-114

[28] Kuefler A, Morton J, Wheeler T A, Kochenderfer M J. Imitating driver behavior with generative adversarial networks// Proceedings of the 28th IEEE Intelligent Vehicles Symposium (IV). Los Angeles, USA, 2017: 204-211

[29] Eysenbach B, Gupta A, Ibarz J, Levine S. Diversity is all you need: Learning skills without a reward function. arXiv preprint arXiv: 1802.06070, 2018

[30] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks//Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France, 2017

[31] Baram N, Anschel O, Caspi I, Mannor S. End-to-end differentiable adversarial imitation learning//Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017: 390-399

[32] Blonde L, Kalousis A. Sample-efficient imitation learning via generative adversarial nets//Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS). Naha, Japan, 2019: 3138-3148

[33] Merel J, Tassa Y, Dhruva T B, et al. Learning human behaviors from motion capture by adversarial imitation. arXiv preprint arXiv: 1707.02201, 2017

[34] Lin Jiahao, Zhang Zongzhang. ACGAIL: Imitation learning about multiple intentions with auxiliary classifier GANs// Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence (PRICAI). Nanjing, China, 2018: 321-334

[35] Li Y, Song J, Ermon S. InfoGAIL: Interpretable imitation learning from visual demonstrations//Proceedings of the 30th Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 3815-3825

[36] Wang Z, Merel J, Reed S E, et al. Robust imitation of diverse behaviors//Proceedings of the 31st Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 5320-5329

[37] Baram N, Anschel O, Mannor S. Model-based adversarial imitation learning. arXiv preprint arXiv: 1612.02179, 2016

[38] Jeon W, Seo S, Kim K E. A Bayesian approach to generative adversarial imitation learning//Proceedings of the 32nd Neural Information Processing Systems (NeurIPS): Montreal, Canada, 2018: 7439-7449

[39] Kochenderfer M J. Decision Making under Uncertainty: Theory and Application. Cambridge, USA: MIT Press, 2015

[40] Littman M L. Markov games as a framework for multi-agent reinforcement learning//Proceedings of the 11th International Conference (ICML). New Brunswick, USA, 1994: 157-163

[41] Stadie B C, Abbeel P, Sutskever I. Third-person imitation learning//Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France, 2017

[42] Torabi F, Warnell G, Stone P. Generative adversarial imitation from observation. arXiv preprint arXiv: 1807.06158, 2018

[43] Bhattacharyya R P, Phillips D J, Wulfe B, Morton J. Multi-agent imitation learning for driving simulation. arXiv preprint arXiv: 1803.01044, 2018

[44] Shi Jingcheng, Yu Yang, Da Qing, et al. Virtual-Taobao: Virtualizing real-world online retail environment for reinforcement learning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019: 4902-4909

[45] Xu Xin, Shen Dong, Gao Yan-Qing, et al. Learning control of dynamical systems based on Markov decision processes: Research frontiers and outlooks. Acta Automatica Sinica, 2012, 38(5): 673-687(in Chinese)
(徐昕, 沈栋, 高岩青等. 基于马氏决策过程模型的动态系统学习控制: 研究前沿与展望. 自动化学报, 2012, 38(5): 673-687)

[46] Doucet A, de Freitas N, Gordon N. An introduction to sequential Monte Carlo methods. Sequential Monte Carlo Methods in Practice. New York, USA: Springer, 2001: 3-14

[47] Tesauro G. Temporal difference learning and TD-gammon. Communications of the ACM, 1995, 38(3): 58-68

[48] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 1992, 8(3-4): 229-256

[49] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization//Proceedings of the 32nd International Conference on Machine Learning (ICML). Lille, France, 2015: 1889-1897

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

[51] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation//Proceedings of the 4th International Conference on Learning Representations (ICLR). San Juan, Puerto Rico, 2016

[52] Zhou Zhi-Hua. Machine Learning. Beijing: Tsinghua University Press, 2016(in Chinese)
(周志华. 机器学习. 北京：清华大学出版社，2016)

[53] Ziebart B D, Maas A, Bagnell J A, Dey A K. Maximum entropy inverse reinforcement learning//Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI). Chicago, USA, 2008: 1433-1438

[54] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1978-1828

[55] Witten I H, Frank E, Hall M A. Data mining: Practical machine learning tools and techniques. The Morgan Kaufmann Series in Data Management Systems. Amsterdam: Elsevier, 2011: 1-629

[56] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition//Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 770-778

[57] Ciresan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification//Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA, 2012: 3642-3649

[58] Li Y, Swersky K, Zemel R S. Generative moment matching networks//Proceedings of the 32nd International Conference on Machine Learning (ICML). Lille, France, 2015: 1718-1727

[59] Salakhutdinov R, Hinton G E. Deep Boltzmann machines//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS). Clearwater Beach, USA, 2009: 448-455

[60] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017: 214-223

[61] Nowozin S, Cseke B, Tomioka R. f-GAN: Training generative neural samplers using variational divergence minimization//Proceedings of the 30th Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 271-279

[62] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs//Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017: 2642-2651

[63] Chen X, Duan Y, Houthooft R, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the 30th Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 2172-2180

[64] Kingma D P, Welling M. Auto-encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations (ICLR). Banff, Canada, 2014

[65] Im D J, Ahn S, Memisevic R, et al. Denoising criterion for variational auto-encoding framework//Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI). San Francisco, USA, 2017: 2059-2065

[66] Mescheder L, Nowozin S, Geiger A. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017: 2391-2400

[67] Finn C, Christiano P, Abbeel P, Levine S. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. arXiv preprint arXiv: 1611.03852, 2016

[68] Syed U, Schapire R E. A game-theoretic approach to apprenticeship learning//Proceedings of the 20th Neural Information Processing Systems (NIPS). Vancouver, Canada, 2007: 1449-1456

[69] Kingma D P, Ba J L. Adam: A method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, USA, 2015

[70] Cardamone L, Loiacono D, Lanzi P L. Learning drivers for TORCS through imitation using supervised methods//Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Games (CIG). Milano, Italy, 2009: 148-155

[71] Todorov E, Erez T, Tassa Y. MuJoCo: A physics engine for model-based control//Proceedings of the International Conference on Intelligent Robots and Systems (IROS). Vilamoura, Portugal, 2012: 5026-5033

[72] Babes-Vroman M, Marivate V, Subramanian K, Littman M. Apprenticeship learning about multiple intentions//Proceedings of the 28th International Conference on Machine Learning (ICML). Bellevue, USA, 2011: 897-904

[73] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv: 1411.1784, 2014

[74] Hausman K, Chebotar Y, Schaal S, et al. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets//Proceedings of the 31st Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 1235-1245

[75] Kuefler A, Kochenderfer M J. Burn-in demonstrations for multi-modal imitation learning//Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Stockholm, Sweden, 2018: 1071-1078

[76] Larsen A B L, Sonderby S K, Larochelle H, Winther O. Autoencoding beyond pixels using a learned similarity metric//Proceedings of the 33rd International Conference on Machine Learning (ICML). New York, USA, 2016: 1558-1566

[77] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models//Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing, China, 2014: 1278-1286

[78] Jang E, Gu S, Poole B. Categorical reparameterization with Gumbel-Softmax. arXiv preprint arXiv: 1611.01144, 2016

[79] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078, 2014

[80] Peters J, Bagnell J A. Policy gradient methods. Encyclopedia of Machine Learning. Boston, USA: Springer, 2010: 774-776

[81] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms//Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing, China, 2014: 387-395

[82] Sutton R S, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation//Proceedings of the 13th Neural Information Processing Systems (NIPS). Denver, USA, 1999: 1057-1063

[83] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning//Proceedings of the 4th International Conference on Learning Representations (ICLR). San Juan, Puerto Rico, 2016

[84] Pfau D, Vinyals O. Connecting generative adversarial networks and actor-critic methods. arXiv preprint arXiv: 1610.01945, 2016

[85] Sutton R S. Learning to predict by the methods of temporal differences. Machine learning, 1988, 3(1): 9-44

[86] Tzeng E, Hoffman J, Zhang N, et al. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv: 1412.3474, 2014

[87] Sundermeyer M, Schluter R, Ney H. LSTM neural networks for language modeling//Proceedings of the 13th Annual Conference of the International Speech Communication Association (ISCA). Portland, USA, 2012: 194-197

[88] Oh J, Chockalingam V, Singh S, Lee H. Control of memory, active perception, and action in Minecraft//Proceedings of the 33rd International Conference (ICML). New York, USA, 2016: 2790-2799

[89] Wierstra D, Forster A, Peters J, Schmidhuber J. Recurrent policy gradients. Logic Journal of the IGPL, 2010, 18(5): 620-634

[90] Chung J, Gulcehre C, Cho K H, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv: 1412.3555, 2014

[91] Aytar Y, Pfaff T, Budden D, Le Paine T. Playing hard exploration games by watching YouTube. arXiv preprint arXiv: 1805.11592, 2018

[92] Torabi F, Warnell G, Stone P. Behavioral cloning from observation//Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden, 2018: 4950-4957

[93] Edwards A D, Sahni H, Schroecker Y, Isbell C L. Imitating latent policies from observation. arXiv preprint arXiv: 1805.07914, 2018

[94] Gibbons R. A Primer in Game Theory. Upper Saddle River, USA: Prentice Hall, 1992

[95] Hu J, Wellman M P. Multiagent reinforcement learning: Theoretical framework and an algorithm//Proceedings of the 15th International Conference on Machine Learning (ICML). Madison, USA, 1998: 242-250

[96] Song J, Ren H, Sadigh D, Ermon S. Multi-agent generative adversarial imitation learning//Proceedings of the 31st Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 7471-7482

[97] Wu Y, Mansimov E, Liao S, et al. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation//Proceedings of the 31st Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 5279-5288

[98] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments//Proceedings of the 31st Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 6379-6390

[99] Gupta J K, Egorov M, Kochenderfer M J. Cooperative multi-agent control using deep reinforcement learning//Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Sao Paulo, Brazil, 2017: 66-83

[100] Peng X B, Kanazawa A, Toyer S, et al. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. arXiv preprint arXiv: 1810.00821, 2018

[101] Choi J, Kim K E. Inverse reinforcement learning in partially observable environments//Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI). Pasadena, USA, 2009: 1028-1033

[102] Peng X B, Abbeel P, Levine S, Van De Panne M. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions on Graphics, 2018, 37(4): 143:1-143:14

[103] Pinto L, Davidson J, Sukthankar R, Gupta A. Robust adversarial reinforcement learning//Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017: 2817-2826

LIN Jia-Hao, M. S. candidate. His main research interests include imitation learning and reinforcement learning.

ZHANG Zong-Zhang, Ph. D., associate professor. His research interests include reinforcement learning, intelligent planning, and multi-agent systems.

JIANG Chong, M. S. candidate. His research interests include imitation learning and reinforcement learning.

HAO Jian-Ye, Ph. D., associate professor. His research interests include deep reinforcement learning and multi-agent systems.

Background

Imitation learning based on generative adversarial nets (GANs-IL), as a combination of the adversarial training mechanism of generative adversarial networks and the idea of the iterative improvement in imitation learning methods based on inverse reinforcement learning, has achieved remarkable successes in a variety of domains, such as autonomous driving, simulation, robotic control, and so on. Our paper introduces the main idea of generative adversarial imitation learning (GAIL), summarizes two main problems in GAIL, outlines many solutions to these two problems, discusses some practical GANs-IL applications, and highlights some future trends in the field, with the hope of providing a valuable reference in its future development.

This paper is partially supported by the National Natural Science Foundation of China (61876119, 61502323) and the Natural Science Foundation of Jiangsu (BK20181432). These projects aim to enrich the learning and planning theory and develop efficient learning and planning algorithms to expand the power and applicability of learning and planning agents in partially observable stochastic environments.