软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(10):4772−4803 [doi: 10.13328/j.cnki.jos.006671]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

逆向强化学习研究综述*

张立华[1], 刘 全[1,2,3,4], 黄志刚[1], 朱 斐[1,2]

[1](苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

[2](江苏省计算机信息处理技术重点实验室 (苏州大学), 江苏 苏州 215006)

[3](符号计算与知识工程教育部重点实验室 (吉林大学), 吉林 长春 130012)

[4](软件新技术与产业化协同创新中心, 江苏 南京 210023)

通信作者: 刘全, E-mail: quanliu@suda.edu.cn

摘 要: 逆向强化学习 (inverse reinforcement learning, IRL) 也称为逆向最优控制 (inverse optimal control, IOC), 是强化学习和模仿学习领域的一种重要研究方法, 该方法通过专家样本求解奖赏函数, 并根据所得奖赏函数求解最优策略, 以达到模仿专家策略的目的. 近年来, 逆向强化学习在模仿学习领域取得了丰富的研究成果, 已广泛应用于汽车导航、路径推荐和机器人最优控制等问题中. 首先介绍逆向强化学习理论基础, 然后从奖赏函数构建方式出发, 讨论分析基于线性奖赏函数和非线性奖赏函数的逆向强化学习算法, 包括最大边际逆向强化学习算法、最大熵逆向强化学习算法、最大熵深度逆向强化学习算法和生成对抗模仿学习等. 随后从逆向强化学习领域的前沿研究方向进行综述, 比较和分析该领域代表性算法, 包括状态动作信息不完全逆向强化学习、多智能体逆向强化学习、示范样本非最优逆向强化学习和指导逆向强化学习等. 最后总结分析当前存在的关键问题, 并从理论和应用方面探讨未来的发展方向.

关键词: 逆向强化学习; 模仿学习; 生成对抗模仿学习; 逆向最优控制; 强化学习

中图法分类号: TP18

中文引用格式: 张立华, 刘全, 黄志刚, 朱斐. 逆向强化学习研究综述. 软件学报, 2023, 34(10): 4772–4803. http://www.jos.org.cn/1000-9825/6671.htm

英文引用格式: Zhang LH, Liu Q, Huang ZG, Zhu F. Survey on Inverse Reinforcement Learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(10): 4772–4803 (in Chinese). http://www.jos.org.cn/1000-9825/6671.htm

Survey on Inverse Reinforcement Learning

ZHANG Li-Hua[1], LIU Quan[1,2,3,4], HUANG Zhi-Gang[1], ZHU Fei[1,2]

[1](School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

[2](Provincial Key Laboratory for Computer Information Processing Technology (Soochow University), Suzhou 215006, China)

[3](Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (Jilin University), Changchun 130012, China)

[4](Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China)

Abstract: Inverse reinforcement learning (IRL), also known as inverse optimal control (IOC), is an important research method of reinforcement learning and imitation learning. IRL solves a reward function from expert samples, and the optimal strategy is then solved to imitate expert strategies. In recent years, fruitful achievements have been yielded by IRL in imitation learning, with widespread application

* 基金项目: 国家自然科学基金 (61772355, 61702055, 61876217, 62176175); 江苏省高等学校自然科学研究重大项目 (18KJA520011, 17KJA520004); 吉林大学符号计算与知识工程教育部重点实验室资助项目 (93K172017K18, 93K172021K08); 苏州市应用基础研究计划工业部分 (SYG201422); 江苏高校优势学科建设工程

收稿时间: 2021-11-05; 修改时间: 2021-12-15, 2022-02-08; 采用时间: 2022-03-15; jos 在线出版时间: 2022-05-24

CNKI 网络首发时间: 2023-04-06

in vehicle navigation, path recommendation, and robotic optimal control. First, this study presents the theoretical basis of IRL. Then, from the perspective of reward function construction methods, IRL algorithms based on linear and non-linear reward functions are analyzed. The algorithms include maximum marginal IRL, maximum entropy IRL, maximum entropy deep IRL, and generative adversarial imitation learning. In addition, frontier research directions of IRL are reviewed to compare and analyze relevant representative algorithms containing IRL with incomplete expert demonstrations, multi-agent IRL, IRL with sub-optimal expert demonstrations, and guiding IRL. Finally, the primary challenges of IRL and future developments in its theoretical and application significance are summarized.

Key words: inverse reinforcement learning (IRL); imitation learning; generative adversarial imitation learning; inverse optimal control (IOC); reinforcement learning (RL)

1 引 言

逆向强化学习 (inverse reinforcement learning, IRL) 作为一种学习专家策略的模仿学习算法, 已成功应用于汽车导航[1]、路径规划[2,3]、行为预测[4−8]和机器最优控制[9−11]等领域. 近年来, 鉴于以上任务的复杂性和应用前景, 逆向强化学习已成为强化学习领域和模仿学习领域的研究热点. 本文旨在梳理逆向强化学习发展脉络, 介绍前沿研究进展和分析关键问题.

逆向强化学习方法将模仿学习问题抽象为马尔可夫决策过程, 应用强化学习方法模仿专家策略. 与强化学习依据奖赏函数求解最优策略不同, 逆向强化学习方法包含依据专家样本求解奖赏函数过程, 因其与强化学习方法学习过程相反, 被称为逆向强化学习算法.

强化学习 (reinforcement learning, RL) 起源于最优控制领域, 是一种通过与环境交互求解最优策略的方法, 广泛应用于工业制造[12]、机器人最优控制[13]、游戏博弈[14,15]、优化与调度[16−18]和仿真模拟[19]等领域. 强化学习方法基于马尔可夫决策过程, 通过智能体与环境交互获得奖赏, 并根据累积奖赏更新策略以选取最优动作, 具有对环境探索的自学能力. 算法在执行过程中持续探索环境和利用环境反馈信息, 使智能体的策略逐步迭代收敛至最优策略. 此外, 强化学习方法中的奖赏函数由人工设定, 作为设计人员与智能体之间的沟通媒介, 奖赏函数中蕴含设计人员所期望目标的全部信息, 而智能体通过与环境交互获得奖赏的方式“解码”奖赏函数, 完成期望任务.

模仿学习是一种通过专家样本模仿专家策略的方法[20,21], 包括行为克隆方法 (behavioral cloning, BC) 和逆向强化学习方法.

行为克隆[22,23]方法直接学习状态或标签到动作或路径的映射, 无需建立奖赏函数, 一般通过监督学习方法实现. 对于小状态空间问题, 行为克隆是一种十分高效的方法. 但在连续状态-动作空间或大状态-动作空间问题中, 因为行为克隆方法只考虑在每个状态采取的动作与专家样本是否匹配, 不考虑未来收益, 所以若与环境交互路径很长且专家样本不足, 则行为克隆方法会将细微误差在连续决策中逐步放大, 甚至环境发生一点变化, 都会极大影响算法性能, 这被称为行为克隆方法中的复合误差问题[24,25].

逆向强化学习也被称为逆向最优控制 (inverse optimal control, IOC)[26,27], 最初由 Russell 于 1998 年提出[28]. 与强化学习方法相同, 逆向强化学习方法基于马尔可夫决策过程, 通过智能体与环境交互求解最优策略. 在一般强化学习问题中, 奖赏函数由人工设定, 而在许多复杂问题中, 很难设计出精确的奖赏函数, 此时却很容易通过专家策略采样专家样本, 例如汽车驾驶[29]和操纵机器打结等[30,31]. 针对这类问题, 逆向强化学习方法抛弃人工设定奖赏函数过程, 直接通过专家样本重建奖赏函数, 并依据所得奖赏函数求解最优策略, 达到模仿专家策略的目的[32].

相比行为克隆方法, 逆向强化学习方法具有更好的泛化性和鲁棒性[33,34], 若环境改变或在专家样本状态-动作空间之外, 仍可保证算法性能. 此外, 由于逆向强化学习方法通过最大化累积奖赏值求解最优策略, 所以逆向强化学习方法方法不存在行为克隆方法中的复合误差问题.

逆向强化学习方法的奖赏函数最初为线性函数, 由 Ng 等人于 2000 年提出[35], 此后基于学徒学习的逆向强化学习算法[36]、最大边际算法[37]、最大熵算法[38]和相对熵算法[39]等被相继提出. 这类算法假设状态或状态-动作对特征的线性组合为奖赏函数, 算法的最终目标为求解每个特征的系数, 当特征系数确定, 奖赏函数随之确定. 逆向强化学习方法中多个奖赏函数均可求解专家策略, 这被称为非适定 (ill-posed) 问题, 算法仍需在满足条件的奖赏

函数集合中选择最优解. Ng 等人基于启发式搜索思想要求奖赏函数满足约束条件的同时还需满足额外目标函数, 这一思想可用线性规划或二次规划数学模型表示. 因此, 早期逆向强化学习方法使用线性规划或二次规划方法求解奖赏函数, 且奖赏函数为线性基函数, 依据不同目标函数, 基于基函数的算法分为 3 类: 基于最大边际思想的算法、基于概率模型思想的算法和基于结构化分类思想的算法. 上述 3 类算法要求奖赏函数满足基本约束条件, 在此基础之上, 基于最大边际思想的算法要求奖赏函数尽可能区分专家策略 (最优策略) 与次优策略, 即专家策略平均回报尽可能大于次优策略平均回报. 基于概率模型思想的算法将问题抽象为概率模型, 每个奖赏函数对应的最优路径 (trajectory) 都满足各自的概率分布[40]. 因此最大熵算法要求奖赏函数在满足基本约束条件的同时, 保证其所对应最优路径概率分布的熵值最大, 类似还有基于交叉熵[41,42]、相对熵和最大似然估计[43]等算法. 基于结构化分类思想的算法通过多个线性参数化分类器求解奖赏函数, 以使每个状态的累积回报尽可能大. 总体来说, 基于线性奖赏函数的逆向强化学习算法在一些小状态空间和离散状态空间问题中取得了不错的效果, 且应用到汽车导航问题中. 这类算法除相对熵算法外都是基于模型的算法, 因此需要提供环境模型 (状态转移概率), 在一定程度上限制了算法的应用. 另外, 由于所有的奖赏函数都是特征的线性组合, 导致以下问题: (1) 特征需要凭借人的经验来选取, 增加了算法的难度和不稳定性; (2) 线性奖赏函数形式简单, 存在表达能力有限的问题.

为克服线性奖赏函数的局限性, 基于非线性奖赏函数的逆向强化学习算法被提出, 包括基于贝叶斯的非参数化特征构建算法[44]和基于神经网络的非线性逆向强化学习算法. 基于贝叶斯的非参数化特征构建算法用高斯过程 (Gaussian processes, GPs)[45]构建非线性奖赏函数, 在一定程度上解决了线性奖赏函数表征能力不足的问题, 但同时也需要提供大量专家样本. 传统逆向强化学习算法 (例如学徒学习算法、最大边际算法、最大熵算法、相对熵算法) 与神经网络结合, 将神经网络作为奖赏函数, 取得了很好的效果. 这类算法可以通过神经网络自动提取状态特征, 具有更强的表征能力. 目前, 在游戏、自动驾驶、路径导航和机器控制领域取得了一定成果.

Ho 等人将在计算机视觉领域取得优秀成果的生成对抗网络 (generative adversarial network, GAN)[46]与逆向强化学习结合, 提出生成对抗模仿学习算法 (generative adversarial imitation learning, GAIL)[47]. 该算法将逆向强化学习过程抽象为求解奖赏函数的 IRL 过程和求解最优策略的 RL 过程, 并指出两个过程的交替迭代为零和博弈问题, 因此可用生成对抗思想解决. 相比非线性逆向强化学习算法, GAIL 具有更小的计算量和更高的性能, 但也存在训练不稳定、模态崩塌 (mode collapse)[48]和生成样本利用率低的问题[49,50].

近年来, 逆向强化学习方法在机器最优控制领域、状态动作信息不完全领域、多智能体领域和提高专家样本利用率等领域都取得了进展, 也有学者将专家样本最优性假设进行扩展, 解决专家样本非最优问题[51−53], 此外还有指导逆向强化学习方法研究等[54−60].

逆向强化学习方法发展至今, 已成为模仿学习方法中最重要的实现方式. 本文将基于以上分类, 梳理逆向强化学习的发展脉络, 分析其发展的内部机理, 探讨其优势和不足, 并总结未来可能的发展方向. 本文的结构框架如后文图 1 所示, 我们将按照图 1 结构, 介绍逆向强化学习领域的关键论文和最新进展.

2 预备知识

介绍 IRL 的基本知识, 包括数学定义、常用概念和基本算法, 提供与后续章节相关的简单背景知识.

2.1 强化学习

强化学习[61,62]起源于最优控制领域, 是一种持续与环境交互获得奖赏, 并根据累积奖赏采取最优动作的算法. 其本质为通过不断试错的方法与环境交互获取信息, 并利用所获取信息进行策略优化. 算法将与环境交互过程抽象为马尔可夫决策过程 (Markov decision process, MDP)[63], 定义 MDP 为五元组 (S,A,R,γ,P), 智能体在每个状态的动作选择符合马尔可夫性, 即所采取动作只与当前状态有关, 与之前状态无关. MDP 五元组的具体含义如下.

(1) S: 表示环境中的所有状态集合;

(2) A: 表示智能体的所有动作集合;

(3) R: 表示智能体从环境中获得的奖赏函数集合, 依据在不同环境中奖赏函数因变量的不同, 奖赏函数分别为: $r(s,a,s')$、$r(s,a)$ 和 $r(s)$. 其中, $r(s,a,s')$ 表示在状态 s 采取动作 a 到达 s' 所得奖赏, $r(s,a)$ 表示在状态 s 采取动

作 a 所得奖赏, $r(s)$ 表示到达状态 s 所得奖赏;

(4) γ: 表示折扣因子, 强化学习在每个状态选择动作会考虑长远影响, 但是, 算法无法考虑无限远的影响, 这不仅导致状态值无穷大, 也无法实现. 目前有两种解决方法: 设置最大时间步长 T; 设置折扣因子 γ, γ 越接近于 0, 算法越注重短期奖赏, 反之, γ 越接近于 1, 算法越重视长期奖赏, 因此, 可通过调节 γ 平衡算法在短期奖赏和长期奖赏之间的取舍;

(5) P: 表示环境的状态转移概率, 依据在不同环境中状态转移概率函数因变量的不同, 可分为: $p(s',r|s,a)$ 和 $p(s'|s,a)$. 其中, $p(s'|s,a)$ 表示智能体在状态 s 采取动作 a 后, 到达状态 s' 的概率; $p(s',r|s,a)$ 表示智能体在状态 s 采取动作 a 后, 到达状态 s' 并获得奖赏 r 的概率.

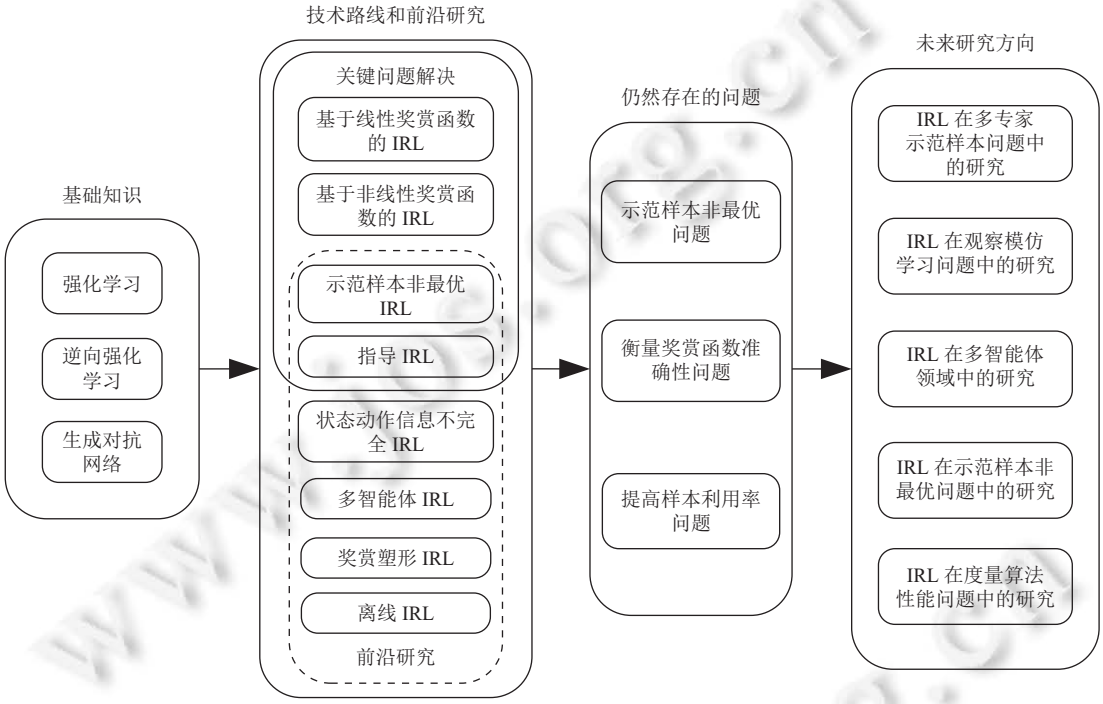

图 1 整体架构示意图

策略 $\pi(a|s)$ 表示状态到动作的映射, 强化学习中策略分为确定性策略和随机策略, 确定性策略表示在每个状态选择的动作确定且唯一, $\pi: S \to A$. 最优策略 π^* 即为所选动作 $a = \arg\max_a q(s,a)$ 在当前状态选择收益最大, $q(s,a)$ 为状态-动作值函数. 随机策略表示在每个状态的动作选择满足一个概率分布, $\pi: S \to prob(A)$, 策略 π 下的状态值函数 $v_\pi(a|s)$ 和动作值函数 $q_\pi(s,a)$ 的迭代形式如公式 (1) 和公式 (2) 所示.

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r(s,a,s') + \gamma v_\pi(s')] \tag{1}$$

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)[r(s,a,s') + \gamma v_\pi(s')] \tag{2}$$

强化学习算法分为表格式算法、值函数逼近算法和策略梯度算法. 早期的动态规划、蒙特卡罗和时序差分等算法属于表格式算法, 这类算法适用于离散动作问题和小状态空间问题. 以深度 Q 网络 (deep Q-network, DQN)[64] 算法为代表的值函数逼近算法, 将适用范围扩展到大状态空间和连续状态空间问题, 用深度神经网络代替表格, 输入为状态和动作, 输出为状态-动作对的值 $q(s,a)$.

另外一种算法为策略梯度算法, 解决了 DQN 无法表示连续动作的问题, 该算法用深度神经网络表示策略函数 π, 输入为状态, 输出为该状态采取的动作. 策略梯度算法的目标函数针对不同问题有 3 个目标函数: 起始价值、

平均价值和时间步平均奖赏, 分别如公式 (3)–公式 (5) 所示.

$$J(\theta)=v_{\pi}(s_0)=\mathbb{E}_{\pi}[v_0] \tag{3}$$

其中, s_0 表示初始状态, v_0 表示起始价值, θ 表示策略 π 的参数向量:

$$J(\theta)=\sum_{s\in S}d_{\pi}(s)v_{\pi}(s) \tag{4}$$

其中, $d_{\pi}(s)$ 表示在策略 π 下关于状态 s 的静态分布:

$$J(\theta)=\sum_{s\in S}d_{\pi}\sum_{a\in A}\pi(a|s)r(s,a,s') \tag{5}$$

其中, d_{π} 表示在策略 π 下状态的概率分布, $\pi(a|s)$ 为状态 s 下按照策略 π 执行动作 a 的概率, $r(s,a,s')$ 为状态 s 下执行动作 a 到达下一状态 s' 所获得的即时奖赏.

因为策略梯度算法基于回合更新, 存在策略评估效率低下和方差较高问题, 所以行动者-评论家 (actor-critic) 算法将策略梯度算法与值函数逼近算法结合, 以增加算法偏差为代价减少方差, 实现算法单步更新、灵活选择动作, 其代表算法为 A2C (advantage actor-critic) 和 A3C (asynchronous advantage actor-critic)[65]. 基于随机策略梯度的算法还包括实现算法稳步更新的 TRPO (trust region policy optimization) 算法[66]. 此外还有基于行动者-评论家架构的 PPO (proximal policy optimization algorithms) 算法[67], 用于解决连续动作空间问题. 以上几种随机策略梯度算法需要对动作采样, 当动作维度很高时效率较低, 此时可采用 DDPG (deep deterministic policy gradient) 算法[68].

2.2 逆向强化学习

逆向强化学习方法通过专家样本求解奖赏函数, 再根据奖赏函数学习最优策略, 原理如图 2 所示. 为准确定义逆向强化学习问题, 以马尔科夫决策过程作为问题框架, 假设专家策略在 MDP 下进行动作选择. 定义 MDP 为五元组 (S,A,R,γ,P), S 表示由状态组成的集合, A 表示由动作组成的集合, $P:S\times A\to prob(S)$ 表示状态转移概率, R 表示奖赏函数集合, 奖赏函数与上文强化学习中的定义相同, γ 表示折扣因子. 另外, 奖赏函数定义为特征的线性组合, 特征向量 ϕ 可为表示状态的特征向量 $\phi:S\to[0,1]^{|S|}$, 也可为表示状态-动作对的特征向量 $\phi:S\times A\to[0,1]^{|S|\times|A|}$.

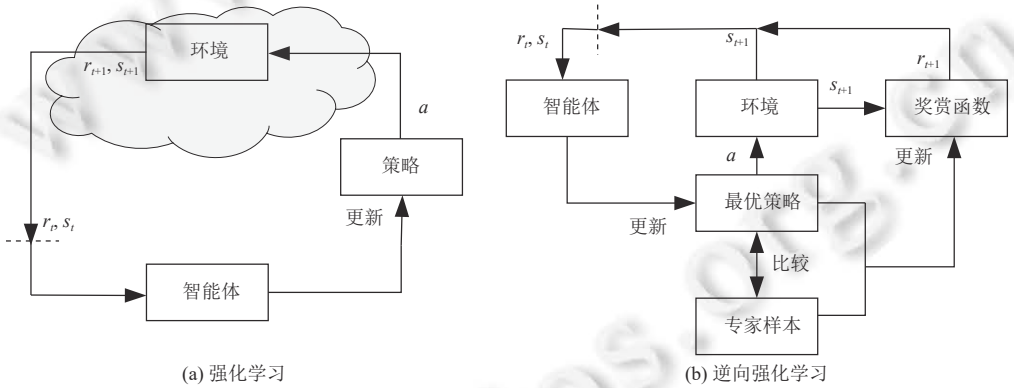

图 2 强化学习和逆向强化学习示意图

逆向强化学习的定义如下, 给定:

(1) 智能体在环境中的专家样本;

(2) 若需要, 则给定传感器所得数据;

(3) 若可行, 则给定物理环境的模型 (包括智能体自身).

确定: 使智能体的策略最优的奖赏函数.

逆向强化学习方法通过由专家轨迹 τ 组成的专家样本 $D=\{\tau_1,\ldots,\tau_N\}$ 求解奖赏函数 $r(s,a,s')$. 假设专家策略 $\pi_E(a|s)$ 为最优策略 $\pi^*(a|s)$, 专家样本由专家策略与环境交互采样得到. 策略的平均回报定义为: $\mathbb{E}_{\pi}[r(s,a,s')]\triangleq\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s'_t)\right]$, 其中 $a\sim\pi(\cdot|s_t)$, $s_{t+1}\sim P(\cdot|s_t,a_t)$. 逆向强化学习方法的优化目标定义为:

$$\max_{r\in R}\left(\min_{\pi\in\Pi}-H(\pi)-\mathbb{E}_{\pi}[r(s,a,s')]\right)+\mathbb{E}_{\pi_E}[r(s,a,s')] \tag{6}$$

其中, $H(\pi)\triangleq\mathbb{E}_{\pi}[-\log\pi(a|s)]$表示策略$\pi$基于$\gamma$折扣的因果熵[69]. 公式 (6) 的目标为寻找奖赏函数使得专家策略的平均回报尽可能大于其他策略的平均回报. 逆向强化学习算法中依据当前奖赏函数求解最优策略的目标函数定义为:

$$\mathrm{RL}(r)=\arg\max_{\pi\in\Pi}-H(\pi)+\mathbb{E}_{\pi}[r(s,a,s')] \tag{7}$$

求解最优策略的目标为最大化策略π的平均回报和熵值. 逆向强化学习算法通常通过迭代方法[70]计算其所确定的二次规划问题, 算法的总体流程如下.

(1) 初始化奖赏函数.

(2) 根据当前奖赏函数, 利用强化学习算法求解最优策略.

(3) 所得最优策略与专家策略比较, 计算特征期望匹配程度.

(4) 改进奖赏函数.

逆向强化学习方法的目的为通过从专家样本中求解奖赏函数模仿专家策略, 然而, 一个策略可以是多个奖赏函数的最优策略, 因此求解奖赏函数的问题为非适定问题. 为获得奖赏函数的唯一解, 一些算法增加目标函数, 例如最大边际算法和最大熵算法等[71−73].

不同逆向强化学习算法实现以上流程的方式各不相同, 基于模型算法需要环境信息 (状态转移概率) 计算状态访问概率和最优策略. 当环境模型已知时, 这种基于模型的算法实现相对简单, 但较难应用于非线性环境中. 无模型算法则通过采样实现以上功能, 可以应用到非线性环境模型中. 另外, 有些无模型算法需要从环境中采样轨迹, 并计算轨迹的分布概率, 这导致算法的时间复杂度较高. 基于奖赏函数的最优策略以强化学习算法求解.

2.3 生成对抗网络

生成对抗网络 (generative adversarial network, GAN)[46]在计算机视觉领域取得了很好的成果. GAN 可以被看作一种基于结构化学习 (structure learning) 的方法, 即: 对于给定的输入, 传统的算法输出为一个数值 (例如识别图片中的数字), 或者输出为一个类别 (例如给出输入属于哪一类别), 而 GAN 的输出为一串序列 (向量、矩阵、图片等).

生成对抗网络由生成器 (generator, G) 和判别器 (discriminator, D) 组成. 生成器和判别器都为深度神经网络. 以图像生成任务为例, 对于给定的图像输入, 生成器的功能为生成与输入相似的图像, 而判别器的功能为判断生成器输出图像与输入图像是否相似. 因此生成器和判别器组成了一个零和博弈问题, 生成器从随机噪声中生成与数据集中样本尽可能相似的图像, 尽可能使判别器认为生成图像为真实样本, 而判别器的功能为抽象真实图像中特征, 并用抽象特征区分真实图像与生成图像, 两者在持续博弈中逐步优化, 直至收敛. 实际上, 真实样本在每个像素点和像素点之间的颜色满足一定的概率分布, 而 GAN 的目标为: 最小化真实样本概率分布与生成样本概率分布之间的距离. 博弈的目标函数如公式 (8) 所示:

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

其中, x表示真实样本, z表示噪声输入, $G(z)$表示生成器产生的样本, $D(\cdot)$表示判别器判断样本为真实样本的概率.

3 逆向强化学习研究进展

3.1 基于线性奖赏函数的逆向强化学习

在逆向强化学习方法发展初期, 用状态-动作对特征的线性组合表示奖赏函数. 因为奖赏函数的不确定性, 传统逆向强化学习方法致力于通过启发式搜索选择最优奖赏函数, 依据其不同实现方式, 分为最大边际方法、概率模型方法和结构化分类方法. 本节从基于线性奖赏函数的逆向强化学习方法出发, 介绍传统逆向强化学习算法对奖赏函数的初选和择优问题的解决.

3.1.1 最大边际方法

逆向强化学习算法可以有效解决行为克隆算法中泛化性差和复合误差问题, 同时也存在非适定问题, 即依据专家样本求得满足条件的奖赏函数存在多个, 难以确定唯一解. 为解决这一问题, Ng 等人[35]依据最大边际 (maximum

margin) 思想选择最优奖赏函数, 使专家策略平均回报尽可能大于其他次优策略平均回报. 算法简化奖赏函数, 定义奖赏函数为只与当前状态特征有关、与选取动作和下一状态无关的线性方程, 并给出了奖赏函数求解方法. 该算法在逆向强化学习方法发展早期有重要意义, 但是, 当$R=0$ (或为常数向量) 或每个状态的奖赏非常接近时, 该方法效果并不明显.

基于已有的学徒学习思想[74], Abbeel 等人提出基于学徒学习的逆向强化学习算法[36], 算法假设奖赏函数为关于状态的线性奖赏函数, 首次提出特征期望 (feature expectation) 概念, 定义特征值的折扣累加为$\mu(\pi)$. 算法证明, 若当前策略与专家策略的特征期望匹配, 则当前策略为专家策略. 因此, 算法虽然不能保证求解正确的奖赏函数, 但可以求解与专家策略同样最优的策略, 并且可以测量策略之间的距离. 由于逆向强化学习的非适定性, 如何从所有满足要求的奖赏函数集合中筛选最优解, 一直是该领域的研究重点. Ratliff 等人[37]引入学徒学习思想, 正式提出最大化专家策略和次优策略距离的最大边际算法.

3.1.2 概率模型优化方法

为了从满足要求的奖赏函数集合中筛选最优解, 与最大边际方法不同, 概率模型方法选取最优策略的熵值为目标函数. 首先, 该方法基于当前奖赏函数计算最优策略, 然后通过最优策略进行采样获得样本轨迹. 最大熵逆向强化学习算法[38]是一种完全基于概率模型的算法. 在满足通过奖赏函数计算出的策略奖赏等于专家策略奖赏的前提下, 利用最大熵方法对奖赏函数集合进行筛选[75−77]. Boularias 等人[39]提出基于相对熵的逆向强化学习算法, 通过当前最优样本的概率分布与专家样本概率分布之间的 KL 散度学习奖赏函数.

在基于最大熵逆向强化学习算法的基础上, Scobee 等人[78]提出了基于最大似然约束推断的逆向强化学习算法. 算法要求提供基于最优策略的专家样本和粗糙的奖赏函数, 通过在马尔可夫决策过程中增加约束 (constraint), 逐步最大化关于专家样本的似然函数, 即最大化专家样本的出现概率. 该算法通过增加约束的方式最大化专家样本发生的概率, 为解决逆向强化学习问题提供了新的视角, 但是目前只能应用于确定性环境, 且需要提供粗糙设计的奖赏函数. Shehryar 等人[79]基于增加约束的方法提出了 ICRL 算法, 算法的两个主要步骤为: (1) 基于约束集合 (由神经网络表示), 学习最优策略; (2) 优化约束集合, 使得专家样本发生的概率最大 (样本概率通过最大熵算法求解). 算法依据以上两步迭代求解, 直至收敛. 基于以上改进, ICRL 算法可以应用于连续高维环境中, 但仍然需要奖赏函数信息.

3.1.3 结构化分类方法

逆向强化学习算法的目的是: 寻找一种策略使得智能体的特征期望与专家策略的特征期望相等. 关于这一点, 上文的逆向强化学习算法中已有介绍, 具体体现在算法提出的二次规划方程的约束条件中. 这一约束条件的本质含义为: 依据算法学习所得策略生成的路径, 其路径分布概率应与专家样本的路径分布概率尽可能相等. 所有这些算法都需重复的通过强化学习过程求解当前最优策略, 因此算法的时间复杂度很大. 为解决这一问题, Klein 等人[80]提出了基于结构化分类的逆向强化学习算法 (structured classification based IRL, SCIRL).

SCIRL 算法抛弃传统逆向强化学习算法从状态到奖赏函数再到最优动作的求解路线, 直接建立状态到动作的映射关系, 即根据动作的不同对状态分类. 在基于结构化分类的逆向强化学习算法中, 状态-动作对依然表示为特征的线性组合, 与上文算法不同, SCIRL 算法将线性状态-动作对特征方程的参数直接作为奖赏函数的参数. 这样定义的优点为: 状态-动作对的值函数与特征期望相等, 对每个特征期望的约束等价于对该特征下状态-动作对的约束, 这直接建立了状态与动作的映射关系. 因此, 问题转变为: 在每个特征下, 状态为样本, 动作为标签, 训练样本为专家样本, 训练算法可以用多类分类算法, 例如: 多类 SVM (multi class SVM) 分类算法.

作为 SCIRL 的扩展, Klein 等人[81]结合分类和回归两类算法, 提出了级联监督逆向强化学习 (cascaded supervised IRL, CSI) 算法. 在 MDP 环境中, 奖赏函数与状态-动作值函数一一对应, 因此可先求解满足专家样本的状态-动作值函数, 该过程为监督学习算法. 进一步, 根据贝尔曼方程, 状态-动作值函数已知, 若求解奖赏函数, 需提供状态转移概率. 状态转移概率信息包含在专家样本的路径中, 通过最小二乘逼近器求解, 可最终得到奖赏函数.

3.1.4 对比分析

因为逆向强化学习方法的非适定性, 算法需要在所有满足要求的奖赏函数中筛选最优解. 基于启发式搜索思

想要求奖赏函数满足约束条件的同时, 还需满足额外目标函数. Ng 等人在算法中运用最大边际思想[35], 被 Ratliff 等人在最大边际算法中正式提出[37]. 此外还包括最大熵和结构化分类思想[80,82,83].

3.1.4.1 解决奖赏函数的非适定问题

在 Ng 等人所提出算法中, 要求专家样本的所有状态-动作值之和尽可能大于对应状态的次优状态-动作值之和. 这一方法有助于算法筛选最优奖赏函数, 但其只看重整体之和, 未考虑每个状态-动作对所占比例关系, 对于出现概率较高的状态-动作对, 应希望其状态-动作值尽量高于次优状态-动作对的值; 另外, 状态-动作对的概率分布与策略也存在对应关系. 因此, 学徒学习对专家样本的状态-动作对求特征期望, 并要求基于该特征期望的奖赏大于次优策略的奖赏. 此后正式提出最大边际思想的最大边际算法也基于这一方法, 更进一步在对应状态所选动作与专家动作不一致时, 增加惩罚项.

概率模型方法从信息角度出发, 提出一种基于熵的奖赏函数选择方法. 若想选择最优奖赏函数, 算法需依据最优问题约束条件求解奖赏函数. 最大熵逆向强化学习算法认为, 所有满足条件的奖赏函数中所含额外信息最少的奖赏函数为最优奖赏函数. 算法通过奖赏函数所对应最优策略与环境交互, 采样策略轨迹, 并计算轨迹的概率分布, 其中轨迹概率分布熵值最大的奖赏函数即为最优奖赏函数. 进一步, 相对熵算法将应用范围拓展至无模型问题, 通过轨迹概率分布与专家轨迹概率分布的相对熵筛选最优奖赏函数. 与最大边际方法相比, 概率模型方法可以应用到更大的状态空间问题. 基于最大似然约束推断的逆向强化学习算法直接通过增加约束的方式优化奖赏函数, 更有利于在实际任务中应用.

结构化分类方法通过多分类算法求解奖赏函数, 研究多集中于离散状态-动作空间任务, 同时动作空间不能过大. 因此, 最终策略的泛化性不如最大边际方法和概率模型方法, 但该方法可以节省大量计算资源.

3.1.4.2 衡量奖赏函数准确性

吴恩达提出的算法依据贝尔曼方程推导奖赏函数需满足一定条件, 通过该方式计算奖赏函数, 需要大量计算资源, 同时也难以应用于大状态-动作空间问题.

学徒学习算法要求与当前奖赏函数所对应最优策略的特征期望尽可能等于专家样本的特征期望, 目前大多数逆向强化学习算法都以此 (或略有改变) 作为奖赏函数是否满足要求的判据, 并在此基础上对奖赏函数作进一步改进.

概率模型方法同样基于特征期望相等这一思想, 但该方法以样本轨迹为计算单元. 概率模型方法所确定最优问题的约束条件要求基于轨迹概率的所有状态-动作对的特征期望尽可能与专家样本的特征期望相等.

结构化分类方法通过对应状态所选动作与专家样本是否一致衡量奖赏函数的准确性, 与监督学习方法类似, 这导致该方法的样本泛化性较差.

3.1.4.3 基于模型和无模型

逆向强化学习算法分为两类: 基于模型 (model-based) 算法和无模型 (model-free) 算法. 基于模型算法需要知道环境信息, 通常为状态转移概率. 因为在奖赏函数迭代过程中需要用状态转移概率计算状态访问概率和最优策略, 以帮助提升奖赏函数. 但是, 若环境信息未知, 需要通过与环境交互采样学习环境模型, 面对大状态空间甚至连续状态空间问题, 难以学习准确模型. 无模型算法主要用于解决环境信息未知问题, 一般通过大量的采样获得情节的分布概率, 这也带来了更高的方差和更多的计算量.

学徒学习算法最小化最坏情况下学习所得特征期望与专家策略特征期望之间的差值, 最大边际算法通过次梯度方法最小化代价函数, 最大熵算法求解所有满足约束的模型中熵值最大模型. 以上算法在其运算过程中都需使用 MDP 模型计算最优策略, 相对熵算法通过重要性采样方法求解损失函数, 属于无模型算法.

3.2 基于非线性奖赏函数的逆向强化学习

在传统逆向强化学习算法中, 由于线性奖赏函数表达能力不足, 限制了算法的应用. 在解决连续高维状态-动作空间问题时, 线性奖赏函数难以准确表征真实奖赏函数. 因此需要将线性奖赏函数变为非线性函数, 提高其表达能力[84,85].

另外, 算法设计时, 特征需凭借设计人员经验选取, 所选特征的范围和正确性直接影响算法性能[1], 增加了算法的难度和不稳定性.

3.2.1 高斯过程拟合奖赏函数

基于高斯过程[86]的非线性逆向强化学习算法 (Gaussian process inverse reinforcement learning, GPIRL) 用高斯过程学习非线性奖赏函数[87], 并且确定了不同特征之间的相关性, 因为高斯过程基于概率的特点, 算法允许专家样本为次优策略.

高斯过程回归需要输入-输出对, 在基于值函数的近似算法中也有应用[88,89]. 而基于高斯过程的逆向强化学习算法只有动作输入没有奖赏输入, 因此需要扩展高斯过程模型, 建立动作和潜在奖赏的随机关系. 学习所得高斯核函数参数既可确定奖赏函数, 也包含不同特征之间的相关性信息.

逆向强化学习期望求解使得其最优策略与专家样本的特征期望相匹配[90]的奖赏函数, 因此专家策略应为最优策略, 而在实际问题中, 有时示范样本由次优策略产生或包含部分次优样本. 学习次优策略的一种方法为建立动作选择的概率模型. GPIRL 采用最大熵算法, 该算法中选择某一路径 τ 的概率与该路径奖赏的指数次方成正比.

3.2.2 深度熵优化模型

目前线性奖赏函数算法存在表征能力有限, 且其特征需要人工设定的问题. 以汽车导航为例, 算法为实现汽车之间保持安全距离[2]的目标, 提出用神经网络近似复杂、非线性奖赏函数, 取得了很好的效果, 所用神经网络为全卷积神经网络 (fully convolutional neural networks, FCNNs). 基于最大熵的深度逆向强化学习算法 (maximum entropy deep inverse reinforcement learning, MEDIRL)[91]用神经网络表示奖赏函数, 目标函数与最大熵算法相同. 由于神经网络的灵活性, MEDIRL 算法中奖赏函数的特征无需人工设定, 但算法所解决问题局限于离散状态空间问题. 与最大熵逆向强化学习算法相同, MEDIRL 算法要求所求最优策略的状态访问概率与专家样本尽可能相等[92]. 奖赏函数求得最优策略的路径满足公式 (9).

$$p(\tau|r)\backslash \propto \exp\left\{\sum_{s,a\in\zeta} r(s,a)\right\} \tag{9}$$

因为在复杂问题中状态的维度非常大, 所以奖赏函数无法与表格式方法一样准确确定每个状态的奖赏, 该算法用参数为 $\theta_{1,2,\cdots,n}$ 的全连接卷积神经网表示奖赏函数. 并用最大后验估计 (maximum a posteriori probability estimate, MAP) 方法, 最大化专家样本和参数 θ 的联合概率.

当奖赏函数为线性函数时, 联合对数似然方程对参数 θ 可微[93], 可用梯度下降法求解. 当前问题的目标函数为求最大熵, 奖赏函数为深度神经网络, 同样具有这一性质.

3.2.3 逻辑回归方法

基于逻辑回归的逆向强化学习算法[94,95]采用逻辑回归方法求解非线性奖赏函数, 是一种无模型算法. 算法通过线性可解马尔可夫决策过程 (linearly solvable Markov decision processes, LMDP) 将逆向强化学习问题转变为密度比例 (density ratio) 预测问题, 密度比例为最优状态转移 (optimal state transition) 和基准状态转移之比. 密度比例通过值函数和奖赏函数组成的二分类逻辑回归 (binomial logistic regression) 分类器求解, 分类器的目的是区分最优策略和基准策略. 因为最优策略为专家策略且基准策略已知, 可通过监督学习求解神经网络表示的值函数和奖赏函数. 根据所得奖赏函数和值函数初始化可得状态-动作值网络, 进而可求解最优策略.

算法在 LMDP 框架下通过贝叶斯规则定义密度比例的对数形式如下:

$$\ln\frac{\pi(s'|s)}{b(s'|s)} = \ln\frac{\pi(s)}{b(s)} + \beta q(s) + \gamma V(s') - \beta V(s) \tag{10}$$

其中, s 为当前所在状态, s' 为在状态 s 采取动作 a 后到达的状态, $q(s)$ 为状态独立奖赏函数 (state-dependent reward function), $V(s)$ 为在状态 s 的值函数.

作者提出通过强化学习方法求解最优策略, 使用 DQN 预测状态-动作值函数 $Q(s,a)$, 并且使用 dueling network 构建神经网络[96], 由状态值和优势函数两部分组成. 在正向强化学习过程中, 神经网络的输入为当前状

态, 输出为状态-动作值, 神经网络由 3 部分构成: 奖赏网络 $r(s,w_r)$, 状态值函数网络 $V(s,w_V)$, 偏好函数网络 (preference network).

3.2.4 生成对抗模仿学习

3.2.4.1 生成对抗模仿学习框架

逆向强化学习方法依据示范样本模仿专家策略, 要求最优策略逼近专家策略的状态-动作对概率分布. GAN 算法[46]在逼近概率分布问题中有优异性能[97]. 因此, Ho 等人基于 GAN 提出生成对抗模仿学习算法[47], 以最大因果熵为损失函数, 直接从专家策略学习.

给定 $\mathbb{E}_\pi(c(s,a)) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma c(s_t,a_t)\right]$ 为策略 π 下代价函数的期望, 代价函数为 $c\in C$, π_E 表示专家策略且已知, 假设最大因果熵存在, 最优问题定义为:

$$\max_{c\in C}(\min_{\pi\in\Pi}-H(\pi)+\mathbb{E}_\pi[c(s,a)])-\mathbb{E}_{\pi_E}[c(s,a)] \tag{11}$$

其中, $H_\pi \triangleq \mathbb{E}_\pi[-\log\pi(a|s)]$ 为策略 π 的因果熵, 即在状态 s 已发生条件下, 所采取策略概率分布的条件熵.

生成对抗模仿学习算法从占用率度量 (occupancy measure) 角度出发, 将逆向强化学习问题分解为正向强化学习 (RL) 过程和逆向强化学习 (IRL) 过程. 在 RL 过程中, 最优化问题的约束条件为所求策略与专家策略占用率度量相匹配, 目的为依据当前奖赏函数寻找策略 π, 满足 $\mathrm{RL}(c)=\arg\min_{\pi\in\Pi}-H(\pi)+\mathbb{E}_\pi[c(s,a)]$, 即最大化因果熵, 最小化代价期望. IRL 过程为 RL 过程的对偶问题, 目的为寻找代价函数, 使得专家策略代价尽可能小, 次优策略代价尽可能大.

生成对抗模仿学习算法在目标函数 (11) 中加入为凸函数的正则函数 ψ^*, 并通过凸共轭函数变换, 将目标函数的强化学习过程转换为如下等价形式:

$$\mathrm{RL}\circ\mathrm{IRL}(\pi_E)=\arg\min_{\pi\in\Pi}-H(\pi)+\psi^*(\rho_\pi-\rho_{\pi_E}) \tag{12}$$

其中, ψ^*为正则函数的凸共轭函数, ρ_π为当前策略的状态-动作对占用率度量, ρ_{π_E}为专家策略的状态-动作对占用率度量.

此处, 与深度学习中用正则函数防止过拟合不同, ψ^*用作衡量所求策略与专家策略的差异, 对算法性能有很大影响, 因此选择正确的正则函数十分关键. 算法分析了 3 种正则函数: 常数正则化函数、示性正则化函数和熵正则化函数. 应用常数正则化函数难以处理大状态空间问题, 应用示性正则化函数的学徒学习算法难以精确匹配占用率度量, 只有熵正则化函数与上文提出模型等价. 基于以上结论, GAIL 算法采用一种新的正则化函数, 最终所得目标函数为:

$$\mathbb{E}_{\pi_\theta}[\log D_\omega(s,a)]+\mathbb{E}_{\pi_E}[\log(1-D_\omega(s,a))]-\lambda H(\pi) \tag{13}$$

解决公式 (13) 的关键在于寻找鞍点 (π_θ, D_ω), 其中策略 π_θ 和 D_ω 用函数逼近器表示. ω 利用 Adam 梯度算法使公式 (13) 增加, θ 利用 TRPO 算法使公式 (13) 下降, TRPO 能够保证 $\pi_{\theta_{t+1}}$ 不远离 π_{θ_t}. GAIL 算法的判别器为 D, 其所提供奖赏值被用来区分专家策略与次优策略, 与 GAN 中的判别器训练方法相同. 在得到 D 中的奖赏后, 利用 TRPO 算法进行策略学习, 改进生成器, 直到算法收敛.

与传统逆向强化学习算法相比, GAIL 算法具有更强的表征能力和更高的计算效率[98], 能够解决大状态空间问题. 但是也存在模态崩塌问题 (model collapse)[48]、训练不稳定问题和生成样本利用率低问题 (low sample efficiency in terms of environment interaction)[49,50].

3.2.4.2 生成对抗模仿学习算法性能分析

目前, 当能够采样足够专家样本时, BC 算法依然是最佳选择. 但在绝大多数模仿学习任务中, 由于任务危险或采样成本高, 难以获取足够的专家样本. 在这些专家样本较少的模仿学习任务中, GAIL 算法的性能优于 BC 算法. 然而, 目前对这一现象的理论分析依然很少. 假设通过 GAIL 算法或 BC 算法优化所得最优策略为 π^*, 专家策略为 π_E, 策略在起始状态 s_0 的值函数分别为 $V_{\pi_E}(s_0)$ 和 $V_{\pi^*}(s_0)$, 用以评价策略的性能.

Xu 等人[99]理论分析证明, 在 BC 算法中, 若最优策略 π^* 与专家策略 π_E 之间的 KL 散度满足 $\mathbb{E}_{s\sim\pi_E}[D_{KL}(\pi_E(\cdot|s),\pi^*(\cdot|s))]\leqslant\varepsilon$, 则 $V_{\pi_E}(s_0)$ 和 $V_{\pi^*}(s_0)$ 的差值满足公式 (14).

$$V_{\pi_E}(s_0)-V_{\pi^*}(s_0)\leqslant\frac{2\sqrt{2}R_{\max}}{(1-\gamma)^2}\sqrt{\varepsilon} \tag{14}$$

其中, $R_{\max}$ 表示奖赏函数的约束: $|r(s,a)|\leqslant R_{\max}$.

在 GAIL 算法中, 若最优策略 π^* 与专家策略 π_E 之间的 f-散度满足 $D_f(\rho_{\pi^*},\rho_{\pi_E})\leqslant\varepsilon$, 则 $V_{\pi_E}(s_0)$ 和 $V_{\pi^*}(s_0)$ 的差值满足公式 (15).

$$V_{\pi_E}(s_0)-V_{\pi^*}(s_0)\leqslant O\left(\frac{1}{1-\gamma}\sqrt{\varepsilon}\right) \tag{15}$$

由公式 (14)、公式 (15) 可知, 在 BC 算法中 $V_{\pi_E}(s_0)$ 和 $V_{\pi^*}(s_0)$ 的复合误差 (compounding error) 随着由 γ 控制的序列长度增加呈指数增加, 复合误差在 GAIL 算法中呈线性增加.

3.2.4.3 生成对抗模仿学习中的模态崩塌问题

模态指专家样本中存在的一些特征. 例如, 在自动驾驶问题中, 专家样本中可能存在两种模态: 车辆在高速公路中以高速行驶; 在市区中以相对低速行驶. 逆向强化学习算法一般假设专家样本为单一模态, 但是, 在实际问题中, 专家样本通常由多个专家产生或某一专家依据不同偏好产生, 即专家样本存在多个子分布. 然而, 在 GAIL 算法中生成器所生成样本只能满足专家样本的某一子分布, 无法满足全部的专家样本分布. 针对这一问题, 目前的改进方向有: 多模态改进[11,100,101]和模型改进[102,103].

(1) 基于最大互信息的生成对抗模仿学习

Li 等人[101]提出了基于最大互信息的生成对抗模仿学习 (information maximizing generative adversarial imitation learning, InfoGAIL) 算法, 通过最大化策略轨迹与模态隐变量之间的互信息, 学习多模态策略. 算法假设模态隐变量 c 为隐藏在专家样本集合中的离散变量, 策略 π 关于该变量的条件概率分布为 $p(\pi|c)$, 可通过训练得到. 为最大化利用模态隐变量 c 中信息, InfoGAIL 算法要求最大化 c 与状态-动作对之间互信息, 互信息表达式如下所示:

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

因为模态隐变量关于状态-动作对的后验概率 $q(c|s,a)$ 未知, 算法将互信息缩放为其变分的最小值, 在生成对抗网络模型中增加推断器, 用以逼近后验概率. 推断器的输入为生成器产生的状态-动作对, 输出为后验概率 q, 随着算法训练, 判别器逐渐收敛到使互信息最大的模态隐变量.

(2) 基于辅助选择器的生成对抗模仿学习

在实际任务中, 专家样本通常由多种模态策略或技能生成. 以自动驾驶问题为例, 专家策略通常使用 3 个技能: 保持直行、变道左侧车道和变道右侧车道. 为使算法能够依据当前状态自主选择技能, Fei 等人[103]提出一种新的多模态生成对抗框架, Triple-GAIL 算法. Triple-GAIL 中专家样本带有技能标签, 算法将 Triple-GAN 算法扩展, 增加辅助选择器, 与 Triple-GAN 增加通过半监督方法训练的分类器不同, Triple-GAIL 中选择器具有重要作用, 不只用于区分技能标签, 还可以根据当前状态自动生成技能标签, 实现同时学习选择标签和重建多模态策略的功能.

如图 3 所示, Triple-GAIL 模型由 3 部分组成: 选择器、生成器和判别器. 选择器输入为专家样本 $\{s_t^e,a_t^e\}$ 和生成器样本 $\{s_t^g,a_t^g\}$, 输出为技能标签 c_t, 生成器基于算法训练设定技能和当前状态选择动作, 判别器判断当前状态-动作-标签样本是否来自属于特定模态的专家样本. 无论专家示范样本中是否有状态-动作-标签样本, 选择器和基于条件分布的生成器都可以输出状态-动作对和状态-标签对.

基于标签的监督学习方法通过在已知分布的专家样本中随机采样学习模态变量, 一旦训练完成, 模型输出动作与人工设定的标签一一对应. 而 Triple-GAIL 算法可以根据环境逐步学习技能.

Triple-GAIL 的目标函数如公式 (16) 所示.

$$\max_{\alpha,\theta}\min_{\omega}\mathbb{E}_{\pi_E}[\log(1-D_\omega(s,a,c))]+\gamma\mathbb{E}_{\pi_\theta}[\log D_\omega(s,a,c)]+(1-\gamma)\mathbb{E}_{C_\alpha}[\log D_\omega(s,a,c)]+\lambda_E R_E+\lambda_G R_G-\lambda_H H(\pi_\theta) \tag{16}$$

其中, α、θ 和 ω 分别为选择器、生成器和判别器的参数, $R_E=\mathbb{E}_{\pi E}[-\log p_{c_\alpha}(c|s,a)]$, R_E 表示确保选择器收敛到专家样本轨迹概率分布的监督学习损失, $R_G=\mathbb{E}_{\pi_\theta}[-\log p_{c_\alpha}(c|s,a)]$, R_G 表示选择器条件概率分布 $p_{c_\alpha}(c|s,a)$ 与生成器条件概率分布 $p_{\pi_\theta}(c|s,a)$ 之间的散度.

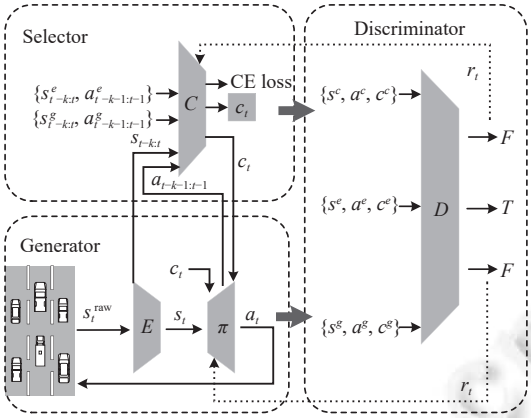

图 3 Triple-GAIL 结构示意图

3.2.4.4 生成对抗模仿学习中的不稳定问题

目前, 前沿的逆向强化学习算法大多基于对抗模仿学习算法 (adversarial imitation learning). 在这类算法中, 生成器和判别器以对抗博弈的方式交替更新, 因此, 所有对抗模仿学习算法核心是解决最大最小优化问题, 这导致算法训练不稳定、对超参数敏感和样本利用率低. Dadashi 等人[104]提出了不依赖对抗训练的基于 Wasserstein 距离的模仿学习算法 (primal Wasserstein imitation learning, PWIL). 在实际算法中, 需要等待智能体与环境交互并生成轨迹后计算 Wasserstein 距离, 为解决这一问题, 作者提出利用 Wasserstein 距离的上限代替距离计算. 因此, 虽然 PWIL 算法的奖赏函数与对抗模仿学习一样不是稳定常数, 但并不通过智能体与环境交互确定, 当专家样本确定后, 奖赏函数可以通过离线方式确定. 因为避免了求解最大最小优化问题, 相比对抗模仿学习算法, PWIL 算法具有较高的稳定性和样本利用率. 另外, PWIL 算法构建了智能体策略与专家策略之间距离的度量方法, 且无需提供环境的真实奖赏函数, 因此, 其可以用于真实环境中的训练效果评测.

3.2.5 引导代价学习方法

逆向强化学习算法可以解决高维任务中的模仿学习问题, 但由于最优示范样本通常难以获得和高维任务的复杂性, 存在策略模型容易陷入局部最优的问题. 此外, 算法为提升自身策略性能, 需要进行大量试错, 导致收敛速度慢, 环境样本利用率低.

解决以上问题的一个方法是建立近似模型[105], 利用模型引导算法在状态空间中探索, 环境的近似模型可以人为建立, 也可以通过已知数据建立. 引导策略搜索 (guided policy search, GPS) 算法利用差分动态规划方法 (differential dynamic programming, DDP) 生成合适的引导样本 (guiding samples), 引导样本指导算法在高回报区域搜索, 同时避免陷入局部最优. GPS 算法通过正则化的重要性采样利用引导样本进行策略搜索.

该算法主要应用于通过专家样本学习最优策略的机器人控制领域[9]. 实现过程分为两步: 第 1 步, 根据专家样本的最优路径概率分布求解代价函数; 第 2 步, 根据代价函数求解其对应最优路径的概率分布.

(1) 根据最优路径概率分布求解代价函数

首先假设示范样本是专家在奖赏函数不确定的环境中采取最优策略生成的样本, 且依据最大熵模型, 示范样本中的路径 τ_i 满足如下分布.

$$\begin{cases} p(\tau)=\dfrac{1}{Z}\exp(-c_\omega(\tau)) \\ Z=\displaystyle\int \exp\left(-c_\omega(\tau_j)\right)d\tau \end{cases} \tag{17}$$

其中, $\tau=s_1,a_1,\ldots,s_T,a_T$ 表示单个路径样本, s_t 和 a_t 分别是在时间 t 的状态和动作, $c_\omega(\tau)=\sum_t c_\omega(\tau)$ 为由参数 ω 组成且未知的代价函数, $p(\tau)$ 为路径 τ 发生的概率.

公式 (17) 的含义为根据最大熵模型, 当知道环境的代价函数时, 其基于最优策略的路径分布可通过公式 (17)

求出, 显然当前代价函数产生的路径分布与示范样本路径分布越接近代价函数越好. 当示范样本给定, 可通过求似然估计函数的最大值确定最优代价函数.

抽样概率分布 $q(\tau)$ 是基于抽样的 IOC (inverse optimal control) 算法成功的关键. 预测 $Z=\int \exp(-c_\omega(\tau))d\tau$ 的最优分布为 $q(\tau)\propto \exp(-c_\omega(\tau))$, 因此在代价函数未知的情况下确定分布 $q(\tau)$ 是十分困难的. 可以通过迭代的方式逐步完善 $q(\tau)$ 的分布, 选取当前代价函数下的最优路径和示范样本共同组成 $q(\tau)$ 分布, 首先在当前路径分布 $q(\tau)$ 下根据最大似然原理优化代价函数 c_ω, 然后根据当前代价函数优化路径分布 $q(\tau)$. 因为该算法是基于抽样方法实现, 所以可以用于无模型环境.

(2) 根据代价函数求最优路径概率分布

在单次迭代过程中代价函数 $c_\omega(\tau)$ 已知, 目标是尽可能准确表达当前代价函数下最优策略所产生路径的高斯分布. 为使该算法应用于无模型环境中, 可通过 $\hat{q}(\tau)$ 采样所得路径分布 D_{traj} 建立环境的局部模型, 但这会导致当搜索区域距离当前样本过远时误差变大, 可添加约束缓解该问题.

$$\begin{cases}\min\limits_{\omega,q(\tau)}\mathbb{E}_q[c_\omega(\tau)]\\ \text{s.t. } D_{KL}(q(u_t|x_t)\,\|\,\hat{q}(u_t|x_t))<\varepsilon\end{cases} \tag{18}$$

其中, $\hat{q}(\tau)$ 为上一次迭代过程中的路径概率分布.

采用拉格朗日乘子法可将公式 (18) 有约束问题变为无约束问题:

$$L_{\text{traj}}(q(\tau),\eta)=\mathbb{E}_q[c_\omega(\tau)]+\eta(D_{KL}(q\,\|\,\hat{q})-\varepsilon) \tag{19}$$

3.2.6 对比分析

基于线性奖赏函数的传统逆向强化学习方法致力于以更好地算法选择奖赏函数和衡量奖赏函数的准确性. 但是, 线性奖赏函数算法存在表征能力有限, 状态特征需要人工设定的问题. 因此, 本节算法提出使用高斯过程或神经网络近似复杂、非线性奖赏函数, 取得了很好的效果.

3.2.6.1 具体分析

Levine 等人[87]用高斯过程表示奖赏函数. 然而该算法需要大量的专家样本和实验奖赏样本才能完成奖赏函数的训练[106], 且算法仍然需要状态的特征作为输入.

相比传统逆向强化学习算法, 基于最大熵的深度逆向强化学习算法可以表示更复杂的奖赏函数, 不需要人工设计状态特征, 该算法在行人导航中进行了实验验证. 但同时, 该算法仍局限于解决格子世界环境中本质为离散状态空间问题的路径规划任务, 并未测试算法在连续状态空间问题中的性能.

基于逻辑回归的逆向强化学习算法用二分类逻辑回归方法解决了 IRL 过程, 并且不需要完全求解正向强化学习过程. IRL 过程中求出了奖赏函数和状态值函数, 相比之前的算法, 状态值函数大大加快了收敛过程, 因为在正向强化学习过程中 dueling network 可以直接利用状态值函数的信息初始化自己的参数. 另外, 因为输入样本不包含动作信息, 所以算法可用于连续动作问题, 但需将 DQN 算法换为 TRPO 算法.

生成对抗模仿学习方法为具有复杂内部关联结构的高维数据分布建模提供了一种很有发展前景的方法, 本质上是一种监督学习方法, 因为其考虑数据之间的关系, 使得算法具有很好的泛化性, 现已成为逆向强化学习领域的经典算法. 但也存在以下两个问题.

(1) 训练不稳定问题

生成器和判别器的训练是零和博弈问题, 需要控制判别器强度. 若判别器明显比生成器强, 会导致生成器改进梯度变小, 同样若判别器比生成器弱, 也不利于生成器改进.

生成器面临梯度消失问题, 因为当判别器远比生成器性能好时, 判别器相当于 Sigmoid 函数, 中间陡峭, 两边分别无限接近于 0 和 1, 且斜率接近于 0. 此时判别器保持稳定, 输入样本处于 Sigmoid 函数的左右两部分, 即使输入样本存在扰动也可以准确区分. 但这也带来一个问题, 因为 Sigmoid 函数左右两个部分的斜率接近于零, 会使得生成器改进梯度消失.

(2) 模态崩塌问题

针对模态崩塌问题, 按照专家样本中模态标签是否确定, 目前的改进方法可分为无监督学习方法和监督学习

方法. InfoGAIL 通过最大化模态隐变量和观测状态-动作对的互信息模仿专家策略. VAE-GAIL[102]引入变分自编码器推断模态变量. 以上算法可以从无标签专家样本中学习, 但是因为缺少标签信息, 这些算法只能推断隐含的标签信息, 而缺乏对文本和具体任务的分析.

CGAIL[11]将标签信息直接输入生成器和判别器, 求解关于标签的策略和奖赏函数. ACGAIL[100]在生成对抗模型中增加了辅助分类器, 而判别器的任务只是判定样本是否来自专家样本, 在 ACGAIL 中, 分类器与判别器共享参数, 两者共同向生成器提供对抗损失.

Triple-GAIL 中的专家样本由多个专家生成, 且带有技能标签. 与上文提到的算法不同, Triple-GAIL 算法可以根据当前状态自动选择技能标签, 先人工确定模态, 然后重新建立多模态策略.

PWIL 算法利用 Wasserstein 距离以离线方式确定奖赏函数, 将 IRL 算法缩减为一个强化学习过程, 极大提高了算法稳定性和样本利用率.

引导代价学习算法致力于解决现实问题中的高维复杂问题, 并在 2D 导航模拟器和现实机械臂环境中实验. 另外, 因为算法基于抽样方法实现, 所以可用于无模型环境.

3.2.6.2 仍然存在的共同问题

(1) 需要大量计算资源和专家样本

逆向强化学习是一个在若干约束条件中搜索奖赏函数的过程, 影响算法求解难度的主要因素为单次迭代复杂度和问题本身复杂度. 对于问题本身来说, 若最优问题为凸函数, 则求解速度呈线性. 反之, 若最优函数不为凸函数, 则需要重点分析单次迭代复杂度, 单次迭代过程为通过当前奖赏函数求解最优策略的强化学习过程. 另外, 非线性逆向强化学习方法用于解决高维复杂问题, 相比于线性问题状态-动作空间指数增加. 同时, 非线性奖赏函数所表示函数集合空间也很大. 这要求专家示范更多的样本满足需求[107].

(2) 样本利用率低

逆向强化学习算法每次迭代更新奖赏函数都需要进行一次强化学习过程, 计算基于当前奖赏函数的最优策略, 因此逆向强化学习算法与环境交互次数远多于强化学习算法. 当算法为采用随机策略的无模型算法时, 这一问题更加严重[49]. 当算法在模拟环境中实验时, 这一缺点并不明显, 但在机器人控制领域, 机械手臂需要与环境交互, 导致采样代价较高. 由于逆向强化学习方法存在生成样本利用效率低的问题, 该方法很难应用在这些领域.

目前, 结构化分类方法和基于逻辑回归的逆向强化学习算法引入监督学习思想, 通过减少强化学习过程的方法, 减少了与环境的交互次数. 引导代价学习算法利用差分动态规划建立模型, 生成训练样本, 既可以引导策略, 又可以提高样本利用率. 同样, PWIL 和 SQIL 算法[108]将 IRL 算法缩减为强化学习过程, 提高了算法稳定性和样本利用率.

4 示范样本非最优逆向强化学习

逆向强化学习方法假设专家样本由最优策略产生, 限制了该方法的应用. 在一些任务中, 因为环境的危险性或采样成本高昂无法产生足够数量的专家样本, 只能以非最优样本辅助算法学习. 或者, 因为专家策略的不稳定和泛化性差, 导致专家样本中含有非最优样本.

为解决这类问题, Brown 等人[52]提出了基于路径排序的奖赏函数推测算法 (trajectory-ranked reward extrapolation, T-REX). 算法首先将专家样本按从差到优排序, 如公式 (20) 所示.

$$\tau_1 \prec \tau_2 \prec \ldots \prec \tau_T \tag{20}$$

算法通过深度神经网络表示奖赏函数, 路径之间的距离通过交叉熵表示, 并将交叉熵通过 Softmax 函数归一化, 表示一个路径比另一个路径好的概率. 如公式 (21) 所示.

$$P(\hat{J}(\tau_i) < \hat{J}(\tau_j)) \approx \frac{\exp \sum_{s \in \tau_j} \hat{r}(s)}{\exp \sum_{s \in \tau_i} \hat{r}(s) + \exp \sum_{s \in \tau_j} \hat{r}(s)} \tag{21}$$

将算法应用于基于观察的 MuJoCo 实验和雅达利游戏, 实验证明, 算法性能比传统算法好一个数量级. 另外, 也有一些强化学习算法, 利用不是最优的示范样本辅助算法学习. 例如, Gao 等人[53]提出 NAC (normalized actor-critic) 算法, 利用 SAC (soft actor-critic) 模型[51]实现了算法从差样本中学习.

完全最优的专家样本通常很难获取, 而衡量专家样本的置信度得分 (confidence scores) 比获取最优专家样本容易. 因此 Wu 等人[109]提出 2IWIL (two-step importance weighting imitation learning) 算法和 IC-GAIL (generative adversarial IL with imperfect demonstration and confidence) 算法用于解决非最优专家样本问题, 该算法假设专家样本中存在非最优样本, 且部分样本具有置信度得分, 其中置信度得分表示该策略为最优策略的概率, 同时算法允许专家样本由不同专家生成. 2IWIL 算法首先通过带有置信度得分标签的数据训练概率分类器, 然后预测未带标签数据的置信度得分, 将所有专家样本的置信度得分标签补全, 并以此训练最优策略. 由于 2IWIL 算法可能存在错误累加问题, IC-GAIL 算法继续采用 GAIL 算法中的占用率度量 (occupancy measure) 思想, 将算法收敛目标设置为最小化当前策略与专家策略的占用率度量分布散度.

在逆向强化学习算法的许多应用领域, 算法不仅可以获取接近最优的专家策略, 还可以观测专家的学习过程. 专家学习过程所生成的样本虽然不由最优策略生成, 但也包含有价值的信息, 例如: 在智能体学习环路驾驶问题中, 通过观察专家训练, 智能体能够区分有益状态和有害状态, 这可以帮助智能体在以非最优策略访问状态空间时, 优化算法策略, 提升算法稳定性.

Jacq 等人[110]对这类问题归纳总结, 提出了从学习者中学习问题 (learning from learner, LfL). LfL 问题包含两个智能体: 学习者 (learner) 和观察者 (observer), 其中观察者可以获得学习者在学习过程中产生的轨迹, LfL 问题的两个假设如下: (1) 学习者的策略基于环境奖赏函数; (2) 学习者通过与环境交互获得奖赏改进自身策略. 为解决 LfL 问题, Jacq 等人假设学习者用基于熵正则化的强化学习算法学习, 当奖赏函数只与状态-动作对相关时, 算法虽无法求解真实奖赏函数, 但可以解得与真实奖赏函数等效的奖赏函数. 若问题简化, 当奖赏函数只与状态有关时, 算法可以解得真实奖赏函数. 总结来说, Jacq 等人提出了逆向强化学习中的新问题, 并给出了求解算法, 这类问题有助于我们研究如何使观测者超越学习者. 然而, 该算法并未完全解决 LfL 问题, 算法假设学习者的策略为单调改进的策略, 且需要提供环境中状态转移概率.

在之前的从学习者中学习算法中, 学习者通过强化学习算法学习最优策略, 而算法要求学习者所生成样本轨迹的策略性能单调递增, 这在强化学习算法中很难被保证. 为解决这一问题, Ramponi 等人[111]提出了基于梯度的从非专家学习者中学习的算法 (learning observing a gradient not-expert learner, LOGEL). LOGEL 算法假设观测者可以获得学习者的策略参数、策略迭代的学习率和轨迹样本. 另外, 因为即使是基于值迭代的传统 Q-learning 算法也与策略梯度算法紧密相关, 所以 LOGEL 算法假设学习者通过梯度下降方法更新策略. 算法中观测者以求解最能解释学习者策略迭代原因的奖赏函数为目标. 具体实现为: 在每次迭代中, LOGEL 算法所求解的最优奖赏函数应尽可能缩小学习者的策略参数与依据当前奖赏所求策略参数之间的欧氏距离. 算法在离散状态-动作空间的格子世界和连续状态-动作空间的 MuJoCo 环境中取得了一定的效果, 但由于需要学习者的策略参数和梯度, 导致算法难以应用于实际任务中.

目前, 专家样本非最优问题存在以下两个特点: (1) 专家无法以最佳方式完成任务, 导致样本中含有非最优样本; (2) 由于环境的危险性导致专家样本少且泛化性差. 针对以上两点, 目前算法从两个方向提出解决方法: (1) 提供额外辅助信息 (表示样本优劣的标签). 此类方法利用专家样本中的辅助信息进行奖赏函数优化或求解最优策略. (2) 推测专家目标. 此类算法用于解决最优样本稀少的问题, 为了求解最优策略, 算法通过失败样本辅助成功样本学习, 虽然失败样本没有完成任务, 但至少包含两点信息: 哪些动作不符合最优策略和哪些动作可以采取 (即使其不是最优动作). 因此, 可通过观测学习过程, 理解专家意图, 进而优化奖赏函数和当前策略.

与传统的假设专家样本全为最优样本组成的逆向强化学习算法相比, 基于示范样本非最优的逆向强化学习算法具有更广的应用范围. 引入专家样本中辅助信息和通过次优样本预测专家意图, 以学习环境信息, 二者相辅相成, 保证算法准确优化奖赏函数, 学习最优策略. 另外, 逆向强化学习算法通常通过特征期望的匹配程度衡量奖赏函数的优劣, 因此传统算法难以应用到示范样本非最优问题. 目前针对这一问题的研究仍处于初始阶段, 在如何利

用非最优样本提升特征期望的匹配程度方面仍考虑较少. 此外, 专家更愿意以多种方式表达他们的意图和偏好, 而不仅通过专家样本, 此类问题仍有待研究.

5 指导逆向强化学习

逆向强化学习算法假设示范样本独立同分布, 因此在训练过程中, 算法通常随机选择示范样本. 而在人类教学过程中, 并不随机抽取示范样本, 其会潜在地多次选择相似样本或选择具有对比性的样本. 基于这一思想, 研究者思考如何建立最优示范样本序列, 或提供信息量最高的示范样本集合.

Odom 等人[54]借鉴主动学习思想提出基于逆向强化学习算法的主动搜寻建议算法. 主动学习 (active learning 或 query learning) 已广泛应用于监督学习领域, 算法优化目标为寻找一批数量少但具有代表性的专家示范样本, 以更好地提升算法性能. 主动学习算法寻找最大信息量集合的标准为筛选信息量最大的状态集合 (通常用熵或其他不确定性度量方法). 与监督学习类似, 逆向强化学习算法也通过专家样本学习最优策略, 因此, 作者将主动学习应用于逆向强化学习领域, 但仍有一些问题需要解决. 逆向强化学习中状态到动作映射的概率分布即为智能体所需从专家样本中获取的信息, 主动学习算法希望专家样本集合中包含尽可能多的此类信息, 这一信息既与奖赏函数有关也与最优策略有关. 但是, 因为逆向强化学习问题中一个最优策略可由多个奖赏函数确定, 所以此类信息与奖赏函数的相关性和与最优策略的相关性并不等价. 为解决这一问题, 作者将两种相关性的加权求和作为动作关于状态分布的不确定性标准. 算法在每一次迭代后, 返回最希望专家样本提供的状态集合, 并以此指导智能体学习奖赏函数. 算法首次引入主动学习思想, 研究了逆向强化学习中如何利用专家示范样本指导智能体学习的问题, 具有一定开创性.

Brown 等人[55]引入机器教学 (machine teaching) 算法用以提高智能体从专家样本中的学习效率和性能. 机器教学算法的优化目标为: 选择具有最小教学代价的样本集合进行训练, 决定教学代价的两个指标为训练样本数量和教学风险. 算法依据机器教学理论定义逆向强化学习问题. 首先定义损失函数如公式 (22) 所示.

$$Loss(\omega^*, \hat{\omega}) = \omega^{*\mathrm{T}}\left[\mathbb{E}\left(\sum_{t=0}^{\infty}\gamma^t\phi(s_t)|\pi^*\right) - \mathbb{E}\left(\sum_{t=0}^{\infty}\gamma^t\phi(s_t)|\hat{\pi}\right)\right] \tag{22}$$

其中, ω^* 为真实奖赏函数系数, $\hat{\omega}$ 为当前奖赏函数系数, π^* 为最优策略, $\hat{\pi}$ 为当前策略, $\phi(s)$ 为状态 s 的特征向量. 损失函数表示当前策略和最优策略在真实奖赏函数中的奖赏差值.

机器教学在逆向强化学习问题中的定义为:

$$\begin{cases}\min\limits_{D} \ \mathrm{TeachingCost}(D)\\ \text{s.t. } Loss(\omega^*,\hat{\omega}) \leqslant \varepsilon,\ \hat{\omega} = \mathrm{IRL}(D)\end{cases} \tag{23}$$

其中, D 为专家示范样本集合, 问题目标为该线性规划的最优解.

本文用机器指导方法解决逆向强化学习中如何利用专家示范样本指导智能体学习的问题, 具有一定创新性, 但公式 (23) 线性规划问题需提供真实奖赏函数. 因此作者将主动学习方法与机器指导方法结合, 利用贝叶斯方法估计状态集合信息, 但局限于解决离散状态空间问题.

Haug 等人[56]利用机器指导方法解决专家特征空间与学习者特征空间不匹配问题, 通过引入教学风险 (teaching risk) 指标, 衡量当前所学策略为专家次优策略的程度. 实验证明, 当教学风险较小时, 即使学习者策略与专家策略之间存在特征空间不匹配问题, 也可以训练逼近最优策略. 基于这一发现, Haug 等人提出 TR-Greedy 算法 (feature and demo based teaching with TR-greedy feature selection), 通过专家不断提高学习者的特征空间, 逐渐减小教学风险. 作者将 TR-Greedy 算法在格子世界问题中实验, 并与其他选择特征空间的方法比较, 验证了算法的有效性, 但该算法并不能用于解决连续空间问题.

与 TR-Greedy 算法用于解决特征不匹配问题不同, Kamalaruban 等人[57]从教学者视角出发, 提出了一种应用更广泛的交互式教学算法, 以提升学习者策略性能. 算法将问题分为两类: (1) 假设教学者拥有全局视野, 可以观测学习者的当前策略; (2) 假设教学者不知道学习者的动态环境, 只能观察到学习者带有噪音的当前策略. 基于这两

类问题, 作者将之前算法中专家策略到学习者策略的教学简化为参数教学. 作者引入软贝尔曼策略 (soft Bellman policies), 通过 MCE-IRL[60]方法更新奖赏函数梯度, 并借鉴机器教学思想[58,59]提供专家示范样本, 原则为尽可能缩小学习者与教学者之间的奖赏函数参数. 作者在基于离散空间的汽车驾驶模拟环境中进行实验, 验证了算法的有效性, 但仍然缺乏对连续空间问题的研究.

本节介绍了通过建立最优示范样本序列, 或提供信息量最高的示范样本集合进行策略优化的指导逆向强化学习算法. 这类算法主要通过两种方法提供最能协助智能体学习最优策略的信息: 基于主动学习的方法和基于粗糙奖赏函数的方法. 两种方法目标一致, 但实现方式不同. 基于主动学习的方法在每次迭代中, 通过筛选信息量最大的专家样本集合优化当前策略. 而基于粗糙奖赏函数的方法通过引入额外指标筛选信息量最大的专家样本集合.

与传统逆向强化学习算法相比, 指导逆向强化学习算法具有更好的鲁棒性、更快的收敛速度和更高的样本利用率. 但现有的指导学习方法多集中在与监督学习方法结合的初级阶段, 依然存在许多问题. 由于基于主动学习的方法需要不断与专家交互, 极大增加了人力成本. 而基于粗糙奖赏函数的方法需要人工设计奖赏函数, 这在复杂问题中难以被满足. 另外, 现有的指导逆向强化学习算法只能适用于离散空间问题或低维状态-动作空间问题中, 在连续高维状态-动作空间问题中的应用仍有待研究.

6 状态动作信息不完全逆向强化学习

状态或动作信息不完全问题指: 智能体对状态-动作-下一状态序列只有非完全的观测. 包括部分可观察逆向强化学习问题、观察模仿学习 (imitation learning from observation, IfO) 问题和基于状态信息的逆向强化学习问题.

6.1 部分可观察逆向强化学习

部分可观察逆向强化学习问题表示, 在马尔可夫决策过程 (partially observable Markov decision process, POMDP) 中, 难以观测状态的完整信息, 或者观测与状态之间满足一定函数关系. 在 POMDP 问题中, 智能体在时间步 t, 只能获得观测 $o_t = o(s_t)$, 其中 $o: s_t \to o_t$ 是一个固定观测函数. IRL 算法可用于解决 POMDP 问题[112], 并且在图像识别[73]和机器人控制[113]等领域取得了一定进展.

在图像识别领域, 有时传感器输入的图像带有噪声, 识别这些图像即为部分可观察问题. Kitani 等人[73]引入基于最大熵的逆向强化学习思想, 通过隐变量马尔可夫决策过程 (hidden variable Markov decision process) 建模, 求解观测状态概率分布熵值的最大值. 在机器人领域, 物体之间的欧几里得距离较为容易获取, 因此 Boularias 等人[113]引入马尔可夫随机场模型 (Markov random fields), 通过关键物体之间距离判断状态是否相似, 在相似状态采取相同动作, 并将算法成功应用在导航和抓取任务中, 但该算法计算量较大.

6.2 观察模仿学习

在一些问题中, 模仿者只能获取专家生成的状态信息, 无法获取动作信息, 例如: 从视频中学习. 因此, IfO 算法是一种只通过状态信息进行学习的模仿学习算法. 目前 IfO 问题的研究方向可分为: 视频感知和动作控制. 随着卷积神经网络和视觉识别技术的提升, 视频感知方向可以实现直接从专家视频资源中学习, 但仍然存在示学体不匹配 (embodiment mismatch) 和观察视角不同 (viewpoint difference) 的问题[114].

动作控制方向的研究可分为基于模型方法和无模型方法两个分支. 基于模型方法将学习模型分为逆向动态模型 (inverse dynamic model) 和前向动态模型 (forward dynamic model). 逆向动态模型以当前状态和下一状态为输入, 输出为当前状态的动作[115], 即为求解当前状态-动作对关于当前状态和下一状态的条件概率分布. 前向动态模型的输入为状态-动作对, 输出为下一状态, Edwards 等人[116]基于这一模型提出观察模仿隐策略 (imitation latent policies from observation, ILPO) 算法, 该算法可用于离线 (off-line) 学习专家策略.

无模型方法分为两类: 奖赏函数设计方法和对抗方法. 奖赏函数设计方法依据专家示范样本设计奖赏函数, 通过强化学习算法进行训练. IfD (imitation learning from demonstration) 表示通过包含状态-动作信息的专家样本学习专家策略的模仿学习算法. 其中, GAIL 算法通过最小化当前策略与专家策略之间状态-动作对度量分布的 KL 散度[117], 在模仿学习任务中取得优异性能, 现已成为 IfD 领域的代表算法. 基于对抗方法的 IfO 算法采用生成对

抗模仿学习框架, 通过状态到状态的信息, 推断最优策略. 为最小化当前策略与专家策略之间状态-下一状态对度量分布的 KL 散度, Yang 等人[118]提出 IDDM (inverse dynamics disagreement minimization) 算法, 在传统 IfO 方法的基础上缩小 IfO 与 IfD 的差距 (推断隐藏在状态迁移中的动作信息). 并在 GAIL 框架下, 证明了 IfO 和 IfD 之间存在由因果熵表示的差距 IDD (inverse dynamics disagreement), 且差距的上界可在 model-free 下减小, 但 IDDM 算法只能用于确定性环境.

6.3 基于状态信息的逆向强化学习算法

在逆向强化学习算法中, 通常假设模仿者与专家之间的物理结构和应用环境相同, 然而在一部分实际问题中, 这一假设难以被满足, 例如: 在迷宫中, 专家策略可以快速通过迷宫, 而模仿专家的机器人因为物理原因或环境改变, 行驶速度相对较慢甚至难以完成任务. 因此, 模仿者与专家之间的最优策略存在差异. 但是, 专家样本中仍然具有帮助机器人走出迷宫的信息可以利用.

利用只包含目标状态信息的样本集合, Fu 等人[119]提出 VICE (variational inverse control with events) 算法. 与传统逆向强化学习算法通过最大化累积奖赏模仿专家策略不同, VICE 算法以最大化目标状态在未来发生概率为目标. 因为只需提供目标状态信息, VICE 算法极大降低了专家样本的采集难度, 提高了在实际任务中的算法性能, 但与专家交互查询信息方法的引入, 增加了算法的训练时间.

与 VICE 算法不同, Ibarz 等人[120]仍然通过最大化累积奖赏模仿专家策略且专家样本需提供只包含状态信息的专家轨迹, 但在训练过程中需以权重形式人工标定对产生轨迹的偏好. 通过调整奖赏函数, 最大化带偏好标记轨迹的累积奖赏. 为解决离线模仿学习中的领域自适应问题, Jiang 等人[121]提出一种通过最大化专家策略轨迹 (只包含状态信息) 概率模仿专家策略的算法.

为解决专家采样环境与学习者应用环境不一致问题, Liu 等人[122]抛弃专家样本中的动作信息, 只学习专家状态轨迹, 以此提出了基于状态对齐的模仿学习算法 (state alignment based imitation learning, SAIL). 专家示范环境与实际学习环境之间的差别会导致智能体与专家在相同状态所采取动作不同, SAIL 算法将之前算法要求状态-动作对相匹配的目标改为状态分布相匹配. 另外, 为使智能体学习所得轨迹与专家轨迹相匹配, SAIL 算法分别从局部状态对齐和全局状态对齐两个方面进行改进, 在局部状态对齐方面: 增加了基于状态的 β-VAE 模型, β-VAE 模型用以预测下一状态, 帮助智能体返回专家样本轨迹中. 在全局状态对齐方面: 通过计算专家样本状态分布与当前策略轨迹状态分布之间的 Wasserstein 距离, 给出状态转移的奖赏函数, 并用 PPO 算法学习最优策略.

相比于 BC 算法和 GAIL 算法, SAIL 算法在状态空间一致而转移函数和动作空间不一致的问题中具有更好的效果. 但是, SAIL 算法需要较高的计算量, 同时也具有较低的样本利用率.

6.4 对比分析

本节介绍了基于部分可观察、从观测中学习和基于状态信息的逆向强化学习算法. 3 类算法同属状态动作信息不完全逆向强化学习算法, 且各自解决的问题存在差别, 其中基于状态信息的逆向强化学习算法可用于解决观察模仿学习问题, 但观察模仿学习算法重点研究如何利用状态为图像或视频信息的任务, 基于状态信息的逆向强化学习算法侧重研究示学体不匹配和采样环境与应用环境不一致问题.

因为传感器噪声, 或者由于人、机器或环境的干扰, 在机器控制、图像处理和观测学习等领域存在非常多的部分可观察逆向强化学习问题. 但目前对这一领域的研究尚处于起步阶段, Kitani 等人[73]的算法在处理特征问题上有些复杂, Boularias 等人[113]的算法虽然可以解决简单的实际问题但计算量较大.

观察模仿学习方法用于解决的问题为视频感知和动作控制问题. 为从只包含状态信息的专家样本中学习, 目前使用的方法包括基于模型方法、生成对抗模仿学习方法和奖赏塑形方法等. 通过将以上算法与其他 (如强化学习算法) 方法结合, 可以进一步提升算法性能, 目前对于这一领域的研究尚有待于深入研究.

与观察模仿学习方法不同, 基于状态信息的逆向强化学习方法的输入状态不局限于视频或图片, 另外, 基于状态信息的逆向强化学习方法不仅探索如何合理利用专家状态信息, 更注重解决专家样本采样环境与应用环境不匹配或示学体不匹配问题. 目前算法也存在一些问题, 例如, 当环境发生物理改变导致最优策略与专家策略出现差异

时, 只有 VICE 算法适用. 另外, 专家还有其他表达目标的方式, 也可作为辅助信息协助算法模仿专家策略.

7 多智能体逆向强化学习

逆向强化学习算法应用到多智能体领域仍存在许多问题亟待解决[123]. 在多智能体模型中, 首先, 环境相对于单个智能体在不断变化, 其次, 当前智能体所采取的最优策略与其他智能体所采取策略有关, 因此可能存在多个最优策略. 针对以上问题, Waugh 等人[124]与 Kuleshov 等人[125]分别对奖赏函数的设定问题进行研究, Song 等人[126]提出基于生成对抗框架的 MAGAIL (multi-agent generative adversarial imitation learning) 算法. Wang 等人[127]则基于零和博弈, 探讨了示范样本由次优策略生成的多智能体问题. 但该方向仍有许多问题尚未解决, 需进一步研究[128].

Hadfield-Menell 等人[129]提出协作逆向强化学习算法 (cooperative inverse reinforcement learning, CIRL), 以解决目标统一 (value alignment) 问题. 在实际任务中, 由于缺少奖赏函数, 工程师为让机器实现某一目标, 通常需要手工设计奖赏函数, 其中目标统一问题是工程师需要解决的核心问题. 即让智能体学习所得最优策略与工程师所期望目标一致. 因此, 作者提出了 CIRL 模型, 模型中两个玩家合作游戏, 分别是智能体和人, 其中, 人知道真实奖赏函数, 而智能体只知道真实奖赏函数的先验分布. 智能体通过采取动作, 不断探索如何使人所获得的奖赏最大. 虽然智能体不知道奖赏函数, 但因为其目标与人相同, 所以实际上其和人共享一个奖赏函数, 算法的目标为, 人与智能体采取联合动作, 最大化所获得的总奖赏. 因此, 这一模型鼓励人产生更好的专家样本, 同时鼓励智能体向人学习.

Wang 等人[127]为解决随机零和博弈问题中专家示范样本非最优问题, 提出一种多智能体竞争的逆向强化学习算法. 在以往多智能体逆向强化学习算法中, 通常假设专家示范样本由最优策略产生, 并基于这一假设将多智能体问题解耦. 这一假设在简单问题中尚可满足要求, 但随着问题复杂性的增加, 基于最优策略的专家示范样本越来越难以获得. 因此该算法摒弃专家示范样本最优性假设, 只要求其尽可能最优, 并引入 DNN (deep neural nets) 分别表示策略函数和奖赏函数. 然后, Wang 等人提出了一种对抗训练算法, 用以寻找纳什均衡策略. 总体上算法分为两个部分: 策略部分, 算法基于当前奖赏函数寻找纳什均衡策略; 奖赏部分, 通过更新奖赏函数, 缩小专家策略与纳什均衡策略之间的差距.

在多智能体问题中存在多个纳什均衡, 同时对单个智能体来说环境是非稳态的, 这导致许多模仿学习算法难以应用在多智能体模仿学习领域, Yu 等人[128]提出一个通用的多智能体对抗逆向强化学习算法 (multi-agent adversarial inverse reinforcement learning), 用以解决马尔可夫决策框架下的多智能体模仿学习问题. 传统单智能体逆向强化学习算法可以看作多智能体逆向强化学习算法的特殊形式, 作者从单智能体逆向强化学习算法中的经典算法 GAIL 出发, 构建多智能体问题中拉格朗日对偶问题, 并对每个智能体分别分配生成器和判别器. 这一算法可以用于解决高维环境中的复杂动作模仿问题, 且合作多智能体模仿任务和竞争多智能体模仿任务皆可适用.

Zhang 等人[130]从网络安全领域出发, 指出多智能体领域存在一类问题: 在两个智能体博弈过程中, 防守者不知道其所对抗入侵者的真实目的, 而且入侵者会故意隐藏其真实目的. 之前的 CIRL 和 MA-IRL 算法无法解决这类问题, 因为对于不同示范样本, 奖赏函数会随之改变. 为解决这一问题, 作者提出了非合作逆向强化学习算法 (non-cooperative inverse reinforcement learning, N-CIRL). 并将问题抽象为单方信息不对称 (one-side incomplete information) 的零和博弈马尔可夫问题, 两个智能体具有完全不同的目标, 只有一个智能体知道真正的奖赏函数. 算法建立 N-CIRL 的对偶问题, 通过线性规划求解, 其中防守者只需知道入侵者的动作和下一状态信息.

在运动分析中, 准确评估运动员能力对团队战术、运动员训练和运动员交易都十分重要, 而评估运动员能力最重要的环节是评估其采取动作的正确性. 为解决这一问题, Luo 等人[131]将多智能体逆向强化学习算法应用在专业冰球运动分析中, 提出了一种基于领域知识的逆向强化学习算法 (inverse reinforcement learning method with domain knowledge, IRL-DK). IRL-DK 算法通过交替学习的方式训练智能体, 将单智能体逆向强化学习算法直接用于多智能体马尔可夫问题中. 假设有 A 和 B 两支队伍, 算法首先将队伍 B 视为队伍 A 所面对环境的一部分, 然后基于单智能体马尔可夫决策过程学习队伍 A 的奖赏函数, 队伍 B 的奖赏函数也采用同样的方式学习, 两支队伍交

替训练直到算法收敛. IRL-DK 采用最大熵逆向强化学习算法, 且在算法中增加了基于领域知识的精确稀疏奖赏, 具体实现方式为: 在逆向强化学习算法的目标函数后额外增加一项, 该项为奖赏函数与基于领域知识的奖赏函数之间的欧式距离. 因此, 最优奖赏函数的目标既包括使智能体尽可能模仿专家策略, 又包括尽可能从基于领域知识的奖赏函数中学习. 实验证明, 相比于其他算法, IRL-DK 算法取得了较好的成绩, 且对运动员的评价较为准确.

逆向强化学习方法可以有效解决多智能体领域中的奖赏函数难以设定问题. 但与多智能体强化学习相同, 多智能体逆向强化学习也存在维度爆炸、策略难以收敛和不稳定等问题, 是一个十分复杂的研究方向. 目前的研究主要集中在奖赏函数求解、专家样本非最优和其他一些应用问题. 当前算法的问题主要有以下 3 个: (1) 依然需要人工设定基于先验知识的粗糙奖赏函数; (2) 从专家样本中对每个智能体的策略进行解耦训练随着任务复杂性的增加变得越来越困难; (3) 随着状态-动作空间变大, 对专家样本的需求也急剧增加. 基于以上分析, 抛弃对粗糙奖赏函数的需求、提升算法鲁棒性、提高专家样本利用率和将多智能体逆向强化学习算法应用于解决复杂大状态空间的合作或竞争问题仍有待进一步研究.

8 奖赏塑形逆向强化学习

奖赏塑形是一类以学习最优策略为前提, 对奖赏函数进行修改的方法[132], 最早应用于强化学习算法中. 因在逆向强化学习算法中, 依然需要依据专家样本求解奖赏函数, 奖赏塑形方法在逆向强化学习领域也有广泛应用. 目前研究方法分为两类: (1) 辅助奖赏塑形方法. 此类方法为加快算法收敛速度、提升算法稳定性和提高样本利用率, 重塑奖赏函数, 但无法保证依据修改后的奖赏函数, 算法可收敛至专家策略, 需借助额外信息辅助算法收敛; (2) 直接奖赏塑形方法. 依据此类方法中的奖赏函数, 算法可直接求解专家策略.

8.1 辅助奖赏塑形方法

鉴于奖赏重塑方法在逆向强化学习中的广泛使用, Piot 等人[133]将状态-动作对的优势函数作为奖赏, 并将此项作为目标函数中的正则项, 进行策略优化, 但算法的主要目标仍为最大化专家策略与次优策略之间的边际. 基于该奖赏函数, Metelli 等人[134]提出一种基于二阶标准的奖赏函数筛选算法, 在算法训练过程中, 将最大化累积奖赏作为优化目标, 通过筛选具有最大信息量的奖赏函数, 引导策略梯度更新, 但算法仍需借助当前策略与专家策略之间的 BC 损失 (行为克隆损失) 保证策略收敛至专家策略. Judah 等人[135]将奖赏重塑方法应用于自动驾驶问题中, 通过重塑后的奖赏函数判断当前驾驶策略是否最优, 并提出一种新的拉格朗日优化函数求解专家策略.

Jena 等人[136]提出基于生成对抗模仿学习框架的 A-GAIL (augmenting GAIL) 算法, 算法的奖赏函数为 $r(s,a)=-\log(1-D(s,a))$, A-GAIL 算法提出一种新的基于优势函数的 BC 损失函数, 提升了算法的样本利用率. Brantley 等人[137]提出 DRIL (disagreement-regularized imitation learning) 算法, 将逆向强化学习算法中奖赏函数和策略的交替更新过程缩减为一个只优化策略的过程, 以提高算法的采样效率. 首先算法通过专家样本预训练奖赏函数, 然后直接利用强化学习算法学习最优策略, 但算法依然需要借助 BC 损失. 基于 BC 损失, Reddy 等人[108]同样提出将逆向强化学习算法缩减为一个只优化策略过程的算法 SQIL (soft Q imitation learning). SQIL 算法直接规定专家样本的奖赏为 1, 非专家样本的奖赏为 0, 并通过 SAC 算法学习最优策略.

8.2 直接奖赏塑形方法

Finn 等人[138]将能量模型 (energy based model, EBM) 与 GAN 结合提出了一种与 GAIL 算法等价的 GAN-GCL (generative adversarial network guided cost learning) 算法. GAN-GCL 算法使用 EBM 替代 GAIL 中的判别器, EBM 的目标为求解区分专家策略与当前策略的能量方程, 生成器的目标为最小化当前样本与专家样本之间的能量差. 为提升奖赏函数的稳定性, Fu 等人[139]提出一种基于生成对抗模仿学习框架的 AIRL (adversarial inverse reinforcement learning) 算法, 其中判别器由优势函数的指数形式和策略概率组成. 在基于生成对抗模仿学习框架的算法中, 专家样本中吸收状态 (轨迹结束状态) 的奖赏为 0, 这导致算法容易陷入局部最优. Kostrikov 等人[140]在提出的 DAC 算法中, 规定吸收状态的奖赏需增加一个额外奖赏, 并对所有轨迹的吸收状态进行专门训练, 使用异

策略强化学习算法学习最优策略, 大大提高了算法的样本利用率. Ghasemipour 等人[141]在 AIRL 算法的基础上, 结合 f-MAX 算法[141]提出 FAIRL 算法, 证明了最大化基于重塑奖赏函数的累计奖赏等价于最小化当前策略与专家策略之间占用率度量概率分布的 KL 散度. 为提升奖赏函数的稳定性, AIRL 算法需额外引入一个只与状态有关的重塑奖赏函数, 并进行训练. 而在 Ni 等人[142]提出的 f-IRL 算法中, 重塑奖赏函数可被直接求得. 与 f-IRL 算法相同, f-GAIL 算法[143]同样使用 f-散度模仿专家策略, 但 f-GAIL 算法基于生成对抗模仿学习框架, 提升了算法的样本利用率. 此外, Balakrishnan 等人[144]利用贝叶斯优化算法, 通过高斯过程后验模型求解奖赏函数, 可以求得最优奖赏函数的集合.

Liu 等人[145]为解决 IRL 方法中奖赏函数求解成本高且算法训练不稳定问题, 采用能量模型对专家样本分布进行建模, 将 IRL 方法中更新奖赏函数和策略的迭代过程缩减为两步: (1) 通过能量模型学习奖赏函数; (2) 基于奖赏函数, 使用强化学习算法学习最优策略. 提高了算法的稳定性和样本利用率.

8.3 对比分析

在辅助奖赏塑形方法中, 重塑的奖赏函数通常为一个包含先验知识的简单函数, 提供当前状态-动作对在短期内的回报预测. 然而, 直接通过重塑奖赏函数进行策略学习, 很难得到具有专家策略性能的最优策略. 因此, 在算法训练过程中, 辅助奖赏塑形方法需要额外信息辅助学习, 通常使用 BC 损失. 但由于专家样本数量限制, 使用 BC 损失容易产生过拟合问题, 导致算法性能变差. 如何更有效地引导算法模仿专家策略, 提供更具泛化性的动作选择信息还有待研究.

在直接奖赏塑形方法中, 若使用奖赏函数 $r(s,a) = -\log(1 - D(s,a))$, 算法容易陷入专家样本中的环路轨迹中, 并且在专家轨迹的吸收状态, 奖赏为 0. 若使用奖赏函数 $r(s,a) = \log(D(s,a))$, 由于奖赏函数提供的所有奖赏为负数, 导致算法学习速度变慢且训练过程容易失败. 目前的直接奖赏塑形方法主要从提升奖赏函数的准确性和提出新的目标函数缩小当前策略与专家策略的距离两个方向入手, 但都缺乏对奖赏函数收敛至最优策略和如何配合算法进行优化的理论分析和证明. 另外, 由于奖赏塑形方法可以灵活调节奖赏函数, 因此其可应用于解决状态动作信息不完全问题和高维状态-动作空间问题.

9 离线逆向强化学习

不同于目前逆向强化学习方法需要与真实或虚拟环境交互. 离线逆向强化学习方法从给定的专家样本中推断奖赏函数, 不再与虚拟环境或真实环境交互, 且不再从专家策略获取任何额外信息. 离线逆向强化学习方法用以解决难以建立准确模拟环境或从环境中获取样本的成本较高 (但已采集到较为丰富样本用以离线训练) 的一类问题. 这类问题普遍存在, 如健康护理、金融投资、教育和工业生产等. 目前离线逆向强化学习算法主要分为两类: 优化利用专家样本信息和解决领域自适应 (domain adaptation) 问题.

9.1 优化利用专家样本信息

Lee 等人[146]提出 DSFN (deep successor feature network) 算法, 依据学徒学习算法中所提出的定理优化当前策略, 使得当前策略中状态-动作对的特征期望逼近专家样本中的状态-动作对的特征期望, 其中, 状态-动作对的特征由编码器生成. 另外, 算法提出 TRIL (transition regularized imitation learning) 模型, 用以保证算法的轨迹不偏离专家样本状态-动作空间. 虽然 DSFN 算法在临床虚拟实验中取得了较好的效果, 但其需要较多的专家样本, 而在实际任务中收集的专家样本往往由不同水平的策略生成. 直接从既包含好样本也包含差样本的混合样本中学习, 会导致算法只能学习到性能一般的策略. 若只从好样本中学习, 但好样本数量不足以训练稳定的智能体. Liu 等人[147]提出 COIL (curriculum offline imitation learning) 算法, 依次从学习样本中抽样不同批次的训练样本, 因为抽样样本的大小不同, 所以通过 BC 算法训练的新 BC 智能体也具有不同性能, 算法通过 KL 散度让学习策略模仿具有最优性能的 BC 智能体策略.

9.2 解决领域自适应问题

目前逆向强化学习算法假设算法训练和测试环境相同, 然而在实际问题中, 通过专家策略收集样本的环境与

实际应用的环境总会存在差异. 为解决这类领域自适应问题, Zweig 等人[148]通过理论分析证明了离线逆向强化学习算法在领域自适应问题中性能的理论上界. 但该文中对环境之间必须具有同构马尔可夫决策过程 (isomorphic MDPs) 的限制太过严苛, 限制了算法的实用性.

Jarrett 等人[149]使用能量模型建立只与状态有关的密度概率分布, 通过最小化与专家样本之间的状态概率分布, 学习最优策略. 另外, Jiang 等人[121]提出 HIDIL (horizon-adaptive inverse dynamics) 算法, 算法应用由 HID (horizon-adaptive inverse dynamics) 组成的一组模型预测当前策略所生成的轨迹是否为专家策略轨迹, 通过最大化专家轨迹的概率求解最优策略. 在实际任务中, 算法取得了很好的效果.

9.3 对比分析

在离线逆向强化学习方法中, 算法既没有准确的环境模型, 也无法通过与环境交互采样数据. 因此, 通过异策略预测难以准确预测特征期望, 进而导致难以获得准确的奖赏函数, 影响离线逆向强化学习算法的性能. 目前的算法主要集中于如何最大化专家样本的信息, 缺乏对如何合理采样专家样本和如何处理专家样本以更好训练智能体等方向的研究. 在领域自适应问题中, 专家在采样环境 (source dynamic) 的目标与智能体在目标环境 (target dynamic) 的目标相同, 不同之处在于采样环境的观测视角与目标环境不同或环境的物理结构不同. 目前的算法集中于如何提高不同环境中的轨迹匹配度, 但若环境的物理结构发生变化, 专家策略的轨迹也会改变, 对这一问题的研究仍有待深入. 另外 Chen 等人[150]和 Pulver 等人[151]为平衡自动驾驶问题中规划、性能、安全和效率问题, 将离线逆向强化学习算法应用于该领域.

10 关键问题分析

目前, 逆向强化学习领域仍然存在许多问题亟待解决, 例如: 怎样衡量当前所求奖赏函数的准确性; 怎样解决示范样本非最优问题; 怎样提高样本利用率. 本节将基于前文对这些关键问题进行总结说明.

10.1 衡量奖赏函数准确性问题

对于求得奖赏函数, 仍需筛选其中最优解. 可直接计算当前奖赏函数和真实奖赏函数之间的距离, 但真实奖赏函数难以获取且直接计算奖赏函数并不有效, 因为在某个状态的奖赏不同可能导致最后的策略差异非常大[152]. 因此最好的方法是, 衡量当前策略与专家策略的差距. 一般通过衡量所学策略与专家策略的匹配程度实现, 即: 计算专家样本中所有状态-动作对与学习所得最优策略的状态-动作对的匹配比例, 但该方法无法衡量在关键状态策略的差异程度.

10.2 示范样本非最优问题

逆向强化学习一个重要的假设是, 专家样本由最优策略在环境中采样生成, 但在现实中存在许多高维度问题[53], 难以大量产生专家示范样本, 同时因为问题的复杂性, 专家策略难以保证最优. 例如, 训练机器人做家务、股票交易和一些复杂游戏等. 传统逆向强化学习算法中, 保证算法最优的限制条件为: 学习所得奖赏函数对应最优策略所生成样本的特征期望尽可能与专家样本的特征期望相等. 这导致算法难以解决示范样本非最优问题.

10.3 提高样本利用率

在连续高维状态-动作空间问题或采样成本高的问题中, 通常难以大量获取专家样本, 例如, 在机械手臂与真实环境交互的问题中, 专家样本需人工示范生成, 且机械手臂容易损伤自身或周围环境, 这要求算法提高专家样本利用率和采样样本利用率. 目前逆向强化学习算法利用专家样本的方式较为单一, 一般从专家样本中随机抽取样本, 虽然已有一些基于主动学习或机器学习的算法提出, 但专家样本利用率仍有很大的提升空间. 另外, 由于逆向强化学习算法基于奖赏和策略迭代更新框架, 极大提高了算法对采样样本的需求.

表 1 将对前 9 节提到的主要算法以表格形式分析比较, 介绍算法相比之前算法的创新点, 并分析算法的优缺点和关键特性, 例如, 奖赏函数的类型、奖赏函数的选择方式、基于模型算法 (model-based) 或无模型算法 (model-free)、算法复杂度和样本利用率等.

表 1 逆向强化学习中的主要算法分析

算法名称	算法改进	算法特性
学徒学习方法[36]	首次提出并证明, 若状态-动作对期望匹配, 则当前策略与专家策略同样最优	函数初选: 特征期望匹配 函数择优: 专家策略奖赏尽可能最大
最大边际方法[37]	正式提出最大边际算法, 当所选动作与专家样本不一致时, 给予惩罚	函数初选: 特征期望匹配 函数择优: 专家策略状态值函数最大
最大熵方法[38]	从概率模型角度引入最大熵模型	函数初选: 样本轨迹的特征期望 函数择优: 当前最优样本概率分布熵值最大
相对熵方法[39]	引入相对熵模型, 将算法扩展到无模型问题	函数初选: 样本轨迹的特征期望 函数择优: 当前最优样本概率分布与专家样本概率分布的相对熵最大
基于最大似然约束推断的逆向强化学习算法[78]	基于最大熵模型, 通过增加约束最大化专家样本发生的概率	函数初选: 人工给定粗糙奖赏 函数择优: 通过增加约束的方式优化奖赏函数
结构化分类算法(SCIRL)[80]	直接建立状态与动作的映射关系, 通过多分类算法训练	函数初选: 特征期望匹配 函数择优: 多分类算法迭代选择
基于贝叶斯的非参数化特征构建算法[87]	以高斯过程将奖赏函数扩展为非线性函数	奖赏函数: 高斯过程 分析: 需要大量的专家样本和大量的计算资源
深度最大熵算法[91]	以神经网络表示奖赏函数,并引入最大熵模型	奖赏函数: 神经网络 分析: 算法仍然依赖状态转移概率
基于逻辑回归的监督学习算法[95]	用二分类逻辑回归方法解决了IRL过程, 并且不需要完全求解正向强化学习过程	奖赏函数:Dueling Network 分析: 状态值函数加快了收敛过程
生成对抗模仿学习算法[47]	将GAN与IRL结合, 以最大因果熵为损失函数, 直接从专家样本学习	奖赏函数: 神经网络(判别器) 分析: 具有很好的泛化性和性能, 但也存在训练不稳定、模态崩塌和样本利用率低的问题
Info-GAIL算法[101]	最大化策略轨迹与模态隐变量之间的互信息, 学习多模态策略	奖赏函数: 神经网络(判别器) 分析: 无需专家样本提供模态标签, 是无监督学习方法
Triple-GAIL算法[103]	增加辅助选择器, 根据当前状态自动生成技能标签,同时学习选择标签和重建多模态策略	奖赏函数: 神经网络(判别器) 分析: 专家样本需提供模态标签, 是监督学习方法
IC-GAIL算法[109]	预测未带标签数据的置信度得分, 采用GAIL算法中的占用率度量思想衡量算法收敛程度	奖赏函数: 神经网络(判别器) 分析: 专家样本可由多专家生成, 需部分带有置信度得分
PWIL算法[104]	不依赖对抗训练框架, 采用Wasserstein距离确定奖赏函数, 通过强化学习算法求解最优策略	奖赏函数: 通过Wasserstein距离离线确定奖赏 函数分析: 具有较高的样本利用率和较强的稳定性, 可用于真实环境
SQIL算法[108]	不依赖对抗训练框架, 采用自定义奖赏函数, 通过soft Q-learning算法求解最优策略	奖赏函数: 专家状态-动作对的奖赏为1, 其他为0 分析: 具有较高的样本利用率
引导代价学习算法(GCL)[9]	通过利用差分动态规划建立近似模型, 生成合适的引导样本, 引导策略搜索	奖赏函数: 神经网络 分析: 较好地解决了高维连续且环境模型未知的现实问题
T-REX算法[52]	将专家样本按奖赏排序, 解决示范样本非最优问题	奖赏函数: 神经网络 分析: 相比于传统算法取得了很大的提升, 但并未从失败样本学习
SAIL算法[122]	为解决模仿者与专家环境不一致的问题, 将状态-动作对相匹配的目标改为状态分布相匹配	奖赏函数: 神经网络 分析: 需要较高的计算量, 具有较低的样本利用率
从学习者中学习[110]	提出一种新的专家样本非最优问题	奖赏函数: 神经网络 分析: 若奖赏函数只与状态-动作对有关, 可以求得与真实奖赏函数等效的奖赏函数; 若只与状态有关, 可以求得真实的奖赏函数
LOGEL算法[111]	提出一种新的不需要假设学习者策略单调递增的算法	奖赏函数: 神经网络 分析: 因为需要learner的策略参数和梯度, 这导致算法很难应用于专家为人类的实际任务中
主动搜寻建议算法[54]	将监督学习中的主动学习思想与IRL结合, 用以更好地利用专家样本, 指导算法学习	奖赏函数: 无限制 分析: 算法要求提供的样本集合可能无法由最优策略产生

表 1 逆向强化学习中的主要算法分析 (续)

算法名称	算法改进	算法特性
基于机器学习的IRL算法[55]	引入机器教学算法用以提高智能体从专家样本中学习的效率和性能	奖赏函数: 一般为线性奖赏函数 分析: 算法并不能用于解决连续空间问题
TR-Greedy算法[56]	引入教学风险概念, 解决专家特征空间与学习者特征空间不匹配问题, 指导算法学习	奖赏函数: 一般为线性奖赏函数 分析: 算法并不能用于解决连续空间问题
IRL-DK算法[131]	解决运动员评估问题	奖赏函数: 神经网络 分析: 需要提供基于领域知识的奖赏函数

11 未来展望

目前该领域发展非常快, 不断有新的算法提出, 也有许多新的研究方向值得深入研究, 除前文提到的关键问题外, 以下几点可能成为未来的重点研究方向.

(1) 逆向强化学习在多专家示范样本问题中的研究

在基于生成对抗框架的模仿学习算法中, 当智能体从多个专家产生的示范样本中学习时, 只能获得其中一个专家的策略. 因此, 生成对抗模仿学习算法从单专家所产生示范样本中训练效果相比从多专家所产生示范样本中学习效果更好[153]. 对于这类多模态问题, 如何通过对专家样本增加标签或改变判别器模型来增加算法的学习能力, 有待深入研究.

(2) 逆向强化学习在观察模仿学习问题中的研究

传统 IRL 算法从示范样本中学习最优策略, 这些示范样本通常包含状态信息和动作信息. 目前, 存在大量的人类作业视频[154], 而这些视频不包含动作信息, 寻找一种新的减少 IfO 方法与 IfD 方法之间信息差的算法, 让智能体从视频中学习, 是观察模仿学习[34]领域重要的研究方向[155]. 另外, 如何消除示学体不匹配问题, 也是未来的研究方向.

在机械手臂和机器人模仿学习领域, 由于该类问题的采样难度大, 目前这一领域的研究一直停留在模拟环境阶段. 如何增加采样效率, 实现其在工业领域的落地, 是未来重要的研究方向[114].

(3) 逆向强化学习在多智能体领域的研究

在多智能体领域存在多个纳什均衡, 对于每个智能体, 环境在不断改变, 因此智能体的最优策略也在不断变化, 这导致算法难以收敛. 因此, 设计一种新的模型, 最大化利用专家样本信息, 帮助智能体快速收敛至纳什均衡是该领域的研究重点.

(4) 逆向强化学习在示范样本非最优问题中的研究

目前绝大多数 IRL 算法假设示范样本由最优策略产生, 然而在实际任务中, 示范样本中通常包含差样本[52]. 针对这一问题, 可通过强化学习方法继续提高策略性能[13,64,156]. 但是, 如何筛选差样本并将其从示范样本中移除或利用差样本中的信息帮助智能体学习仍然是一个有待解决的问题[107].

(5) 逆向强化学习在度量算法性能问题中的研究

逆向强化学习算法的目的是从示范样本中学习最优策略. 怎样度量学习所得策略与专家策略之间的相似度至关重要. 目前算法采用的方法有 KL 散度、欧几里得距离和 Wasserstein 距离[157]等. 探索一种新的度量方式也是未来重要的研究方向[158].

12 总 结

逆向强化学习是强化学习领域的重要研究方向之一, 也是模仿学习领域的主要实现方法, 已在汽车导航、路径推荐和机器最优控制等领域取得了令人瞩目的成果. 最近几年, 越来越多的学术界和工业界人士开始对其进行探索和研究, 现已成为模仿学习领域的一个热门研究方向.

本文详细介绍了逆向强化学习领域的关键问题、研究进展和发展方向. 通过奖赏函数构建方式将逆向强化学习算法分为基于线性奖赏函数和非线性奖赏函数的逆向强化学习算法, 以此详细阐述了最大边际逆向强化学习算

法、最大熵逆向强化学习算法、最大熵深度逆向强化学习算法和生成对抗模仿学习算法等, 并分析了其优点和不足, 同时介绍了一些逆向强化学习算法的前沿研究方向, 包括状态动作信息不完全逆向强化学习、多智能体逆向强化学习、示范样本非最优逆向强化学习、指导逆向强化学习、奖赏塑形逆向强化学习和离线逆向强化学习.

通过对当今逆向强化学习算法的研究进展进行综述, 总结了当前的成就, 梳理了该领域的发展脉络. 但因为逆向强化学习是一个跨越心理学、认知科学和最优控制等领域的交叉学科, 仍然存在许多难题等待解决. 未来逆向强化学习将朝以下几个方向发展: (1) 有效学习多专家样本问题; (2) 在观察模仿学习领域的应用; (3) 在多智能体领域的应用; (4) 在示范样本非最优问题中的研究; (5) 寻找更有效的度量策略相似度的方法. 随着以上问题研究的逐步深入, 相信逆向强化学习算法将会发挥越来越重要的作用, 实现 DeepMind 提出的“解决智能, 并用智能解决一切”的目标.

References:

[1] Neu G, Szepesvári C. Apprenticeship learning using inverse reinforcement learning and gradient methods. In: Proc. of the 23rd Conf. on Uncertainty in Artificial Intelligence. Vancouver: ACM, 2007. 295–302.

[2] Kretzschmar H, Spies M, Sprunk C, Burgard W. Socially compliant mobile robot navigation via inverse reinforcement learning. The Int'l Journal of Robotics Research, 2016, 35(11): 1289–1307. [doi: 10.1177/0278364915619772]

[3] Kim B, Pineau J. Socially adaptive path planning in human environments using inverse reinforcement learning. Int'l Journal of Social Robotics, 2016, 8(1): 51–66. [doi: 10.1007/s12369-015-0310-2]

[4] Kuefler A, Morton J, Wheeler T, Kochenderfer M. Imitating driver behavior with generative adversarial networks. In: Proc. of the 2017 IEEE Intelligent Vehicles Symp. Los Angeles: IEEE, 2017. 204–211. [doi: 10.1109/IVS.2017.7995721]

[5] Bogert K, Doshi P. Multi-robot inverse reinforcement learning under occlusion with state transition estimation. In: Proc. of the 2015 Int'l Conf. on Autonomous Agents and Multiagent Systems. Istanbul: ACM, 2015. 1837–1838.

[6] Vogel A, Ramachandran D, Gupta R, Raux A. Improving hybrid vehicle fuel efficiency using inverse reinforcement learning. In: Proc. of the 2012 AAAI Conf. on Artificial Intelligence. Toronto: ACM, 2012. 384–390.

[7] Ziebart BD, Ratliff N, Gallagher G, Mertz C, Peterson K, Bagnell JA, Hebert M, Dey AK, Srinivasa S. Planning-based prediction for pedestrians. In: Proc. of the 2009 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. St. Louis: IEEE, 2009. 3931–3936. [doi: 10.1109/IROS.2009.5354147]

[8] Levine S, Koltun V. Continuous inverse optimal control with locally optimal examples. In: Proc. of the 29th Int'l Conf. on Machine Learning. Edinburgh: ACM, 2012. 475–482.

[9] Finn C, Levine S, Abbeel P. Guided cost learning: Deep inverse optimal control via policy optimization. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: ACM, 2016. 49–58.

[10] Eysenbach B, Gupta A, Ibarz J, Levine S. Diversity is all you need: Learning skills without a reward function. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[11] Merel J, Tassa Y, Tb D, Srinivasan S, Lemmon J, Wang ZY, Wayne G, Heess N. Learning human behaviors from motion capture by adversarial imitation. arXiv:1707.02201, 2017.

[12] Collins S, Ruina A, Tedrake R, Wisse M. Efficient bipedal robots based on passive-dynamic walkers. Science, 2005, 307(5712): 1082–1085. [doi: 10.1126/science.1107799]

[13] Kober J, Bagnell JA, Peters J. Reinforcement learning in robotics: A survey. The Int'l Journal of Robotics Research, 2013, 32(11): 1238–1274. [doi: 10.1177/0278364913495721]

[14] Tesauro G. TD-gammon, a self-teaching backgammon program, achieves master-level play. Neural Computation, 1994, 6(2): 215–219. [doi: 10.1162/neco.1994.6.2.215]

[15] Liu Q, Yan Y, Zhu F, Wu W, Zhang LL. A deep recurrent Q network with exploratory noise. Chinese Journal of Computers, 2019, 42(7): 1588–1604 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2019.01588]

[16] Ipek E, Mutlu O, Martínez JF, Caruana R. Self-optimizing memory controllers: A reinforcement learning approach. ACM SIGARCH Computer Architecture News, 2008, 36(3): 39–50. [doi: 10.1145/1394608.1382172]

[17] Liang TX, Yang XP, Wang L, Han ZY. Review on financial trading system based on reinforcement learning. Ruan Jian Xue Bao/Journal of Software, 2019, 30(3): 845–864 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5689.htm [doi: 10.13328/j.cnki.jos.005689]

[18] Yang SG, Wang YY, Liu WC, Jiang X, Zhao MX, Fang H, Yang Y, Liu D. Temperature-aware task scheduling on multicores based on

reinforcement learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(8): 2408−2424 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6190.htm [doi: 10.13328/j.cnki.jos.006190]

[19] Fu QM, Liu Q, Wang H, Xiao F, Yu J, Li J. A novel off policy $Q(\lambda)$ algorithm based on linear function approximation. Chinese Journal of Computers, 2014, 37(3): 677–686 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2013.00677]

[20] Argall BD, Chernova S, Veloso M, Browning B. A survey of robot learning from demonstration. Robotics and Autonomous Systems, 2009, 57(5): 469–483. [doi: 10.1016/j.robot.2008.10.024]

[21] Siciliano B, Khatib O. Springer Handbook of Robotics. Berlin: Springer, 2008. 1371–1394.

[22] Maeda GJ, Neumann G, Ewerton M, Lioutikov R, Kroemer O, Peters J. Probabilistic movement primitives for coordination of multiple human-robot collaborative tasks. Autonomous Robots, 2017, 41(3): 593–612. [doi: 10.1007/s10514-016-9556-2]

[23] Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang JK, Zhang X, Zhao JK, Zieba K. End to end learning for self-driving cars. arXiv:1604.07316, 2016.

[24] Ross S, Bagnell D. Efficient reductions for imitation learning. In: Proc. of the 13th Int'l Conf. on Artificial Intelligence and Statistics. Sardinia: JMLR.org, 2010. 661–668.

[25] Ross S, Gordon G, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proc. of the 14th Int'l Conf. on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR.org, 2011. 627–635.

[26] Boyd S, El Ghaoui L, Feron E, Balakrishnan V. Linear Matrix Inequalities in System and Control Theory. Philadelphia: SIAM, 1994. 153–154. [doi: 10.1137/1.9781611970777]

[27] Dvijotham K, Todorov E. Inverse optimal control with linearly-solvable MDPs. In: Proc. of the 27th Int'l Conf. on Machine Learning. Haifa: ACM, 2010. 335–342.

[28] Russell S. Learning agents for uncertain environments (extended abstract). In: Proc. of the 11th Annual Conf. on Computational Learning Theory. Wisconsin: ACM, 1998. 101–103. [doi: 10.1145/279943.279964]

[29] Cardamone L, Loiacono D, Lanzi PL. Learning drivers for TORCS through imitation using supervised methods. In: Proc. of the 2019 IEEE Symp. on Computational Intelligence and Games. Milan: IEEE, 2009. 148–155. [doi: 10.1109/CIG.2009.5286480]

[30] Osa T, Sugita N, Mitsuishi M. Online trajectory planning and force control for automation of surgical tasks. IEEE Trans. on Automation Science and Engineering, 2018, 15(2): 675–691. [doi: 10.1109/TASE.2017.2676018]

[31] Todorov E, Erez T, Tassa Y. MuJoCo: A physics engine for model-based control. In: Proc. of the 2012 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Vilamoura-Algarve: IEEE, 2012. 5026–5033. [doi: 10.1109/IROS.2012.6386109]

[32] Osa T, Esfahani AMG, Stolkin R, Lioutikov R, Peters J, Neumann G. Guiding trajectory optimization by demonstrated distributions. IEEE Robotics and Automation Letters, 2017, 2(2): 819–826. [doi: 10.1109/LRA.2017.2653850]

[33] Sermanet P, Xu K, Levine S. Unsupervised perceptual rewards for imitation learning. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[34] Liu YX, Gupta A, Abbeel P, Levine S. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation. Brisbane: IEEE, 2018. 1118−1125. [doi: 10.1109/ICRA.2018.8462901]

[35] Ng AY, Russell SJ. Algorithms for inverse reinforcement learning. In: Proc. of the 17th Int'l Conf. on Machine Learning. Stanford: ACM, 2000. 663–670.

[36] Abbeel P, Ng AY. Apprenticeship learning via inverse reinforcement learning. In: Proc. of the 21st Int'l Conf. on Machine Learning. Banff: ACM, 2004. 1–8. [doi: 10.1145/1015330.1015430]

[37] Ratliff ND, Bagnell JA, Zinkevich MA. Maximum margin planning. In: Proc. of the 23rd Int'l Conf. on Machine Learning. Pittsburgh: ACM, 2006. 729–736. [doi: 10.1145/1143844.1143936]

[38] Ziebart BD, Maas A, Bagnell JA, Dey AK. Maximum entropy inverse reinforcement learning. In: Proc. of the 23rd AAAI Conf. on Artificial Intelligence. Chicago: ACM, 2008. 1433–1438.

[39] Boularias A, Kober J, Peters J. Relative entropy inverse reinforcement learning. In: Proc. of the 14th Int'l Conf. on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR.org, 2011. 182–189.

[40] Shiarlis K, Messias J, Whiteson S. Inverse reinforcement learning from failure. In: Proc. of the 2016 Int'l Conf. on Autonomous Agents & Multiagent Systems. Singapore: ACM, 2016. 1060–1068.

[41] Kalakrishnan M, Pastor P, Righetti L, Schaal S. Learning objective functions for manipulation. In: Proc. of the 2013 IEEE Int'l Conf. on Robotics and Automation. Karlsruhe: IEEE, 2013. 1331–1336. [doi: 10.1109/ICRA.2013.6630743]

[42] Arenz O, Abdulsamad H, Neumann G. Optimal control and inverse optimal control by distribution matching. In: Proc. of the 2016 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Daejeon: IEEE, 2016. 4046–4053. [doi: 10.1109/IROS.2016.7759596]

[43] Vroman MC. Maximum likelihood inverse reinforcement learning [Ph.D. Thesis]. New Brunswick: The State University of New Jersey, 2014.

[44] Zheng JC, Liu SY, Ni LM. Robust Bayesian inverse reinforcement learning with sparse behavior noise. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. Québec City: ACM, 2014. 2198–2205.

[45] Qiao QF, Beling PA. Inverse reinforcement learning with Gaussian process. In: Proc. of the 2011 American Control Conf. San Francisco: IEEE, 2011. 113–118. [doi: 10.1109/ACC.2011.5990948]

[46] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 2672–2680.

[47] Ho J, Ermon S. Generative adversarial imitation learning. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: ACM, 2016. 4565–4573.

[48] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[49] Baram N, Anschel O, Caspi I, Mannor S. End-to-end differentiable adversarial imitation learning. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: ACM, 2017. 390–399.

[50] Blondé L, Kalousis A. Sample-efficient imitation learning via generative adversarial nets. In: Proc. of the 22nd Int'l Conf. on Artificial Intelligence and Statistics. Naha: PMLR, 2019. 3138–3148.

[51] Jing MX, Ma XJ, Huang WB, Sun FC, Yang C, Fang B, Liu HP. Reinforcement learning from imperfect demonstrations under soft expert guidance. In: Proc. of the 2020 AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 5109–5116. [doi: 10.1609/aaai.v34i04.5953]

[52] Brown D, Goo W, Nagarajan P, Niekum S. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 783–792.

[53] Gao Y, Xu HZ, Lin J, Yu F, Levine S, Darrell T. Reinforcement learning from imperfect demonstrations. In: Proc. of the 2018 Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[54] Odom P, Natarajan S. Active advice seeking for inverse reinforcement learning. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. Austin: AAAI, 2015. 4186–4187.

[55] Brown DS, Niekum S. Machine teaching for inverse reinforcement learning: Algorithms and applications. In: Proc. of the 2019 AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 7749–7758. [doi: 10.1609/aaai.v33i01.33017749]

[56] Haug L, Tschiatschek S, Singla A. Teaching inverse reinforcement learners via features and demonstrations. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: ACM, 2018. 8464–8473.

[57] Kamalaruban P, Devidze R, Cevher V, Singla A. Interactive teaching algorithms for inverse reinforcement learning. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI.org, 2019. 2692–2700. [doi: 10.24963/ijcai.2019/374]

[58] Liu WY, Dai B, Humayun A, Tay C, Yu C, Smith LB, Rehg JM, Song L. Iterative machine teaching. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 2149–2158.

[59] Liu WY, Dai B, Li XG, Liu Z, Rehg J, Song L. Towards black-box iterative machine teaching. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 3141–3149.

[60] Rhinehart N, Kitani KM. First-person activity forecasting with online inverse reinforcement learning. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 3696–3705. [doi: 10.1109/ICCV.2017.399]

[61] Liu Q, Zhai JW, Zhang ZZ, Zhong S, Zhou Q, Zhang P, Xu J. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1): 1–27 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2018.00001]

[62] Liu X, Liu SY, Zhuang YK, Gao Y. Explainable reinforcement learning: Basic problems exploration and method survey. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2300–2316 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6485.htm [doi: 10.13328/j.cnki.jos.006485]

[63] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed., Cambridge: MIT Press, 2018. 18–18.

[64] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]

[65] Mnih V, Badia AP, Mirza M, Graves A, Harley T, Lillicrap TP, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: ACM, 2016. 1928–1937.

[66] Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: ACM, 2015. 1889–1897.

[67] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

[68] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan: ICLR, 2016.

[69] Zhou ZY, Bloem M, Bambos N. Infinite time horizon maximum causal entropy inverse reinforcement. IEEE Trans. on Automatic Control, 2018, 63(9): 2787–2802. [doi: 10.1109/TAC.2017.2775960]

[70] Bloem M, Bambos N. Infinite time horizon maximum causal entropy inverse reinforcement learning. In: Proc. of the 53rd IEEE Conf. on Decision and Control. Los Angeles: IEEE, 2014. 4911–4916. [doi: 10.1109/CDC.2014.7040156]

[71] Silver D, Bagnell JA, Stentz A. Learning from demonstration for autonomous navigation in complex unstructured terrain. The Int'l Journal of Robotics Research, 2010, 29(12): 1565–1592. [doi: 10.1177/0278364910369715]

[72] Ziebart BD. Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy. Pittsburgh: Carnegie Mellon University ProQuest Dissertations Publishing, 2010. 76–77.

[73] Kitani KM, Ziebart BD, Bagnell JA, Hebert M. Activity forecasting. In: Proc. of the 12th European Conf. on Computer Vision. Florence: Springer, 2012. 201–214. [doi: 10.1007/978-3-642-33765-9_15]

[74] Amit R, Matari M. Learning movement sequences from demonstration. In: Proc. of the 2nd Int'l Conf. on Development and Learning. Cambridge: IEEE, 2002. 203–208. [doi: 10.1109/DEVLRN.2002.1011867]

[75] Dudík M, Schapire RE. Maximum entropy distribution estimation with generalized regularization. In: Proc. of the 19th Int'l Conf. on Computational Learning Theory. Pittsburgh: Springer, 2006. 123–138. [doi: 10.1007/11776420_12]

[76] Jaynes ET. Information theory and statistical mechanics. Physical Review, 1957, 106(4): 620–630. [doi: 10.1103/PhysRev.106.620]

[77] Amari SI. Information Geometry and Its Applications. Springer, 2016. 44–45.

[78] Scobee DRR, Sastry SS. Maximum likelihood constraint inference for inverse reinforcement learning. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[79] Malik S, Anwar U, Aghasi A, Ahmed A. Inverse constrained reinforcement learning. In: Proc. of the 38th Int'l Conf. on Machine Learning. 2021. 7390–7399.

[80] Klein E, Geist M, Piot B, Pietquin O. Inverse reinforcement learning through structured classification. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Red Hook: ACM, 2012. 1007–1015.

[81] Klein E, Piot B, Geist M, Pietquin O. A cascaded supervised learning approach to inverse reinforcement learning. In: Proc. of the 2013 Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Prague: Springer, 2013. 1–16. [doi: 10.1007/978-3-642-40988-2_1]

[82] Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. The Journal of Machine Learning Research, 2005, 6: 1453–1484.

[83] Taskar B, Chatalbashev V, Koller D, Guestrin C. Learning structured prediction models: A large margin approach. In: Proc. of the 22nd Int'l Conf. on Machine Learning. Bonn: ACM, 2005. 896–903. [doi: 10.1145/1102351.1102464]

[84] Bogdanovic M, Markovikj D, Denil M, De Freitas N. Deep apprenticeship learning for playing video games. In: Proc. of the 2015 Workshops at the AAAI Conf. on Artificial Intelligence. Austin: AAAI, 2015.

[85] Syed U, Schapire RE. A game-theoretic approach to apprenticeship learning. In: Proc. of the 20th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2008. 1449–1456.

[86] Deisenroth MP, Rasmussen CE, Peters J. Gaussian process dynamic programming. Neurocomputing, 2009, 72(7–9): 1508–1524. [doi: 10.1016/j.neucom.2008.12.019]

[87] Levine S, Popović Z, Koltun V. Nonlinear inverse reinforcement learning with Gaussian processes. In: Proc. of the 24th Int'l Conf. on Neural Information Processing Systems. Granada: ACM, 2011. 19–27.

[88] Rasmussen CE, Kuss M. Gaussian processes in reinforcement learning. In: Proc. of the 2004 Advances in Neural Information Processing Systems. Vancouver: MIT Press, 2004. 751–758.

[89] Engel Y, Mannor S, Meir R. Reinforcement learning with Gaussian processes. In: Proc. of the 22nd Int'l Conf. on Machine Learning. Bonn: ACM, 2005. 201–208. [doi: 10.1145/1102351.1102377]

[90] Levine S, Popović Z, Koltun V. Feature construction for inverse reinforcement learning. In: Proc. of the 23rd Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2010. 1342–1350.

[91] Fahad M, Chen Z, Guo Y. Learning how pedestrians navigate: A deep inverse reinforcement learning approach. In: Proc. of the 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Madrid: IEEE, 2018. 819–826. [doi: 10.1109/IROS.2018.8593438]

[92] Wulfmeier M, Rao D, Wang DZ, Ondruska P, Posner I. Large-scale cost function learning for path planning using deep inverse reinforcement learning. The Int'l Journal of Robotics Research, 2017, 36(10): 1073–1087. [doi: 10.1177/0278364917722396]

[93] Snyman JA. Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-based Algorithms. New York: Springer, 2005. 33–53.

[94] Uchibe E. Model-free deep inverse reinforcement learning by logistic regression. Neural Processing Letters, 2018, 47(3): 891–905. [doi: 10.1007/s11063-017-9702-7]

[95] Uchibe E. Deep inverse reinforcement learning by logistic regression. In: Proc. of the 23rd Int'l Conf. on Neural Information Processing. Kyoto: Springer, 2016. 23–31. [doi: 10.1007/978-3-319-46687-3_3]

[96] Wang ZY, Schaul T, Hessel M, Van Hasselt H, Lanctot M, De Freitas N. Dueling network architectures for deep reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: ACM, 2016. 1995–2003.

[97] Peng XB, Kanazawa A, Toyer S, Abbeel P, Levine S. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[98] Lin JH, Zhang ZZ, Jiang C, Hao JY. A survey of imitation learning based on generative adversarial nets. Chinese Journal of Computers, 2020, 43(2): 326–351 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2020.00326]

[99] Xu T, Li ZN, Yu Y. Error bounds of imitating policies and environments. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2020. 15737–15749.

[100] Lin JH, Zhang ZZ. ACGAIL: Imitation learning about multiple intentions with auxiliary classifier GANs. In: Proc. of the 15th Pacific Rim Int'l Conf. on Artificial Intelligence. Nanjing: Springer, 2018. 321–334. [doi: 10.1007/978-3-319-97304-3_25]

[101] Li YZ, Song JM, Ermon S. InfoGAIL: Interpretable imitation learning from visual demonstrations. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 3815–3825.

[102] Wang ZY, Merel J, Reed S, Wayne G, de Freitas N, Heess N. Robust imitation of diverse behaviors. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 5320–5329.

[103] Fei C, Wang B, Zhuang YZ, Zhang ZZ, Hao JY, Zhang HB, Ji XW, Liu WL. Triple-GAIL: A multi-modal imitation learning framework with generative adversarial nets. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI.org, 2020. 2929–2935. [doi: 10.24963/ijcai.2020/405]

[104] Dadashi R, Hussenot L, Geist M, Pietquin O. Primal Wasserstein imitation learning. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[105] Anderson BDO, Moore JB. Optimal Control: Linear Quadratic Methods. New York: Dover Publications, 2007. 262–268.

[106] Yoshua B, Lecun Y. Scaling learning algorithms towards AI. In: Bottou L, Chapelle O, DeCoste D, Weston J, eds. Large-scale Kernel Machines. Cambridge: MIT Press, 2007. 1–41.

[107] Arora S, Doshi P. A survey of inverse reinforcement learning: Challenges, methods and progress. Artificial Intelligence, 2021, 297: 103500. [doi: 10.1016/j.artint.2021.103500]

[108] Reddy S, Dragan AD, Levine S. SQIL: Imitation learning via reinforcement learning with sparse rewards. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2019.

[109] Wu YH, Charoenphakdee N, Bao H, Tangkaratt V, Sugiyama M. Imitation learning from imperfect demonstration. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 6818–6827.

[110] Jacq A, Geist M, Paiva A, Pietquin O. Learning from a learner. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 2990–2999.

[111] Ramponi G, Drappo G, Restelli M. Inverse reinforcement learning from a gradient-based learner. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2020. 2458–2468.

[112] Choi J, Kim KE. Inverse reinforcement learning in partially observable environments. The Journal of Machine Learning Research, 2011, 12: 691–730.

[113] Boularias A, Krömer O, Peters J. Structured apprenticeship learning. In: Proc. of the 2012 Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Bristol: Springer, 2012. 227–242. [doi: 10.1007/978-3-642-33486-3_15]

[114] Torabi F, Warnell G, Stone P. Recent advances in imitation learning from observation. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI.org, 2019. 6325–6331. [doi: 10.24963/ijcai.2019/882]

[115] Hanna JP, Stone P. Grounded action transformation for robot learning in simulation. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2017. 3834–3840.

[116] Edwards A, Sahni H, Schroecker Y, Isbell C. Imitating latent policies from observation. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 1755–1763.

[117] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics, 1951, 22(1): 79–86. [doi: 10.1214/aoms/

1177729694]

[118] Yang C, Ma XJ, Huang WB, Sun FC, Liu HP, Huang JZ, Gan C. Imitation learning from observations by minimizing inverse dynamics disagreement. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2019. 239–249.

[119] Fu J, Singh A, Ghosh D, Yang L, Levine S. Variational inverse control with events: A general framework for data-driven reward definition. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2018. 8547–8556.

[120] Ibarz B, Leike J, Pohlen T, Irving G, Legg S, Amodei D. Reward learning from human preferences and demonstrations in Atari. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2018. 8022–8034.

[121] Jiang SY, Pang JC, Yu Y. Offline imitation learning with a misspecified simulator. In: Proc. of the 34th Int'l Conf. Neural Information Processing Systems. Vancouver: ACM, 2020. 8510–8520.

[122] Liu FC, Ling Z, Mu TZ, Su H. State alignment-based imitation learning. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[123] Gupta JK, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning. In: Proc. of the 2017 Int'l Conf. on Autonomous Agents and Multiagent Systems. São Paulo: Springer, 2017. 66–83. [doi: 10.1007/978-3-319-71682-4_5]

[124] Waugh K, Ziebart BD, Bagnell JA. Computational rationalization: The inverse equilibrium problem. arXiv:1308.3506, 2013.

[125] Kuleshov V, Schrijvers O. Inverse game theory: Learning utilities in succinct games. In: Proc. of the 11th Int'l Conf. on Web and Internet Economics. Amsterdam: Springer, 2015. 413–427. [doi: 10.1007/978-3-662-48995-6_30]

[126] Song JM, Ren HY, Sadigh D, Ermon S. Multi-agent generative adversarial imitation learning. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2018. 7461–7472.

[127] Wang XY, Klabjan D. Competitive multi-agent inverse reinforcement learning with sub-optimal demonstrations. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: ICML, 2018. 5143–5151.

[128] Yu LT, Song JM, Ermon S. Multi-agent adversarial inverse reinforcement learning. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 7194–7201.

[129] Hadfield-Menell D, Dragan A, Abbeel P, Russell S. Cooperative inverse reinforcement learning. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: ACM, 2016. 3916–3924.

[130] Zhang XY, Zhang KQ, Miehling E, Başar T. Non-cooperative inverse reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2019. 9487–9497.

[131] Luo YD, Schulte O, Poupart P. Inverse reinforcement learning for team sports: Valuing actions and players. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: ACM, 2020. 3356–3363.

[132] Ng AY, Harada D, Russell SJ. Policy invariance under reward transformations: Theory and application to reward shaping. In: Proc. of the 16th Int'l Conf. on Machine Learning. San Francisco: ACM, 1999. 278–287.

[133] Piot B, Geist M, Pietquin O. Boosted and reward-regularized classification for apprenticeship learning. In: Proc. of the 2014 Int'l Conf. on Autonomous Agents and Multi-agent Systems. Paris: ACM, 2014. 1249–1256.

[134] Metelli AM, Pirotta M, Restelli M. Compatible reward inverse reinforcement learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 2050–2059.

[135] Judah K, Fern A, Tadepalli P, Goetschalckx R. Imitation learning with demonstrations and shaping rewards. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. Québec: ACM, 2014. 1891–1896.

[136] Jena R, Liu CL, Sycara K. Augmenting GAIL with BC for sample efficient imitation learning. In: Proc. of the 2020 Conf. on Robot Learning. Cambridge: PMLR, 2020. 80–90.

[137] Brantley K, Sun W, Henaff M. Disagreement-regularized imitation learning. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

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

[139] Fu J, Luo K, Levine S. Learning robust rewards with adverserial inverse reinforcement learning. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[140] Kostrikov I, Agrawal KK, Dwibedi D, Levine S, Tompson J. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[141] Ghasemipour SKS, Zemel RS, Gu SX. A divergence minimization perspective on imitation learning methods. In: Proc. of the 3rd Conf. on Robot Learning. Osaka: PMLR, 2020. 1259–1277.

[142] Ni TW, Sikchi HS, Wang YF, Gupta T, Lee L, Eysenbach B. *f*-IRL: Inverse reinforcement learning via state marginal matching. In: Proc. of the 4th Conf. on Robot Learning. Cambridge: PMLR, 2020. 529–551.

[143] Zhang X, Li YH, Zhang ZM, Zhang ZL. *f*-GAIL: Learning *f*-divergence for generative adversarial imitation learning. In: Proc. of the 34th Conf. on Neural Information Processing Systems. 2020.

[144] Balakrishnan S, Nguyen QP, Low BKH, Soh H. Efficient exploration of reward functions in inverse reinforcement learning via Bayesian optimization. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2020. 4187–4198.

[145] Liu MH, He TR, Xu MK, Zhang WN. Energy-based imitation learning. In: Proc. of the 20th Int'l Conf. on Autonomous Agents and Multiagent Systems. ACM, 2021. 809–817.

[146] Lee D, Srinivasan S, Doshi-Velez F. Truly batch apprenticeship learning with deep successor features. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI.org, 2019. 5909–5915. [doi: 10.24963/ijcai.2019/819]

[147] Liu MH, Zhao HY, Yang ZY, Shen J, Zhang WN, Zhao L, Liu TY. Curriculum offline imitating learning. In: Proc. of the 34th Advances in Neural Information Processing Systems. NeurIPS, 2021. 6266–6277.

[148] Zweig A, Bruna J. Provably efficient third-person imitation from offline observation. In: Proc. of the 36th Conf. on Uncertainty in Artificial Intelligence. AUAI Press, 2020. 1228–1237.

[149] Jarrett D, Bica I, van der Schaar M. Strictly batch imitation learning by energy-based distribution matching. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2020. 7354–7365.

[150] Chen JY, Yuan BD, Tomizuka M. Deep imitation learning for autonomous driving in generic urban scenarios with enhanced safety. In: Proc. of the 2019 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Macao: IEEE, 2019. 2884–2890. [doi: 10.1109/IROS40897.2019.8968225]

[151] Pulver H, Eiras F, Carozza L, Hawasly M, Albrecht SV, Ramamoorthy S. PILOT: Efficient planning by imitation learning and optimisation for safe autonomous driving. In: Proc. of the 2021 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Prague: IEEE, 2021. 1442–1449. [doi: 10.1109/IROS51168.2021.9636862]

[152] Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed., New Jersey: Prentice Hall, 2009. 652–658.

[153] Camacho R, Michie D. Behavioral cloning: A correction. AI Magazine, 1995, 16(2): 92.

[154] Zhou LW, Xu CL, Corso JJ. Towards automatic learning of procedures from Web instructional videos. In: Proc. of the 2018 AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2018. 7590–7598.

[155] Sharma P, Pathak D, Gupta A. Third-person visual imitation learning via decoupled hierarchical controller. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2019. 2597–2607.

[156] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of go with deep neural networks and tree search. Nature, 2016, 529(7587): 484–489. [doi: 10.1038/nature16961]

[157] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: ACM, 2017. 214–223.

[158] Komanduru A, Honorio J. On the correctness and sample complexity of inverse reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2019. 7112–7121.

附中文参考文献:

[15] 刘全, 闫岩, 朱斐, 吴文, 张琳琳. 一种带探索噪音的深度循环Q网络. 计算机学报, 2019, 42(7): 1588–1604. [doi: 10.11897/SP.J.1016.2019.01588]

[17] 梁天新, 杨小平, 王良, 韩镇远. 基于强化学习的金融交易系统研究与发展. 软件学报, 2019, 30(3): 845−864. http://www.jos.org.cn/1000-9825/5689.htm [doi: 10.13328/j.cnki.jos.005689]

[18] 杨世贵, 王媛媛, 刘韦辰, 姜徐, 赵明雄, 方卉, 杨宇, 刘迪. 基于强化学习的温度感知多核任务调度. 软件学报, 2021, 32(8): 2408−2424. http://www.jos.org.cn/1000-9825/6190.htm [doi: 10.13328/j.cnki.jos.006190]

[19] 傅启明, 刘全, 王辉, 肖飞, 于俊, 李娇. 一种基于线性函数逼近的离策略$Q(\lambda)$算法. 计算机学报, 2014, 37(3): 677–686. [doi: 10.3724/SP.J.1016.2013.00677]

[61] 刘全, 翟建伟, 章宗长, 钟珊, 周倩, 章鹏, 徐进. 深度强化学习综述. 计算机学报, 2018, 41(1): 1–27. [doi: 10.11897/SP.J.1016.2018.00001]

[62] 刘潇, 刘书洋, 庄韫恺, 高阳. 强化学习可解释性基础问题探索和方法综述. 软件学报, 2023, 34(5): 2300–2316. http://www.jos.org.cn/1000-9825/6485.htm [doi: 10.13328/j.cnki.jos.006485]

[98] 林嘉豪, 章宗长, 姜冲, 郝建业. 基于生成对抗网络的模仿学习综述. 计算机学报, 2020, 43(2): 326–351. [doi: 10.11897/SP.J.1016.2020.00326]

张立华(1992－), 男, 博士, CCF 学生会员, 主要研究领域为逆向强化学习, 模仿学习, 深度强化学习.

黄志刚(1993－), 男, 博士, 主要研究领域为强化学习, 分层强化学习, 模仿学习.

刘全(1969－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为强化学习, 深度强化学习, 逆向强化学习, 自动推理.

朱斐(1978－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为强化学习, 深度强化学习, 文本挖掘, 生物信息学.