逆强化学习算法、理论与应用研究综述

宋 莉 [1] 李大字 [1] 徐 昕 [2]

摘 要 随着高维特征表示与逼近能力的提高, 强化学习 (Reinforcement learning, RL) 在博弈与优化决策、智能驾驶等现实问题中的应用也取得显著进展. 然而强化学习在智能体与环境的交互中存在人工设计奖励函数难的问题, 因此研究者提出了逆强化学习 (Inverse reinforcement learning, IRL) 这一研究方向. 如何从专家演示中学习奖励函数和进行策略优化是一个重要的研究课题, 在人工智能领域具有十分重要的研究意义. 本文综合介绍了逆强化学习算法的最新进展, 首先介绍了逆强化学习在理论方面的新进展, 然后分析了逆强化学习面临的挑战以及未来的发展趋势, 最后讨论了逆强化学习的应用进展和应用前景.

关键词 强化学习, 逆强化学习, 线性逆强化学习, 深度逆强化学习, 对抗逆强化学习

引用格式 宋莉, 李大字, 徐昕. 逆强化学习算法、理论与应用研究综述. 自动化学报, 2024, **50**(9): 1704−1723

DOI 10.16383/j.aas.c230081

A Survey of Inverse Reinforcement Learning Algorithms, Theory and Applications

SONG Li[1] LI Da-Zi[1] XU Xin[2]

Abstract With the research and development of deep reinforcement learning, the application of reinforcement learning (RL) in real-world problems such as game and optimization decision, and intelligent driving has also made significant progress. However, reinforcement learning has difficulty in manually designing the reward function in the interaction between an agent and its environment, so researchers have proposed the research direction of inverse reinforcement learning (IRL). How to learn reward functions from expert demonstrations and perform strategy optimization is a novel and important research topic with very important research implications in the field of artificial intelligence. This paper presents a comprehensive overview of the recent progress of inverse reinforcement learning algorithms. Firstly, new advances in the theory of inverse reinforcement learning are introduced, then the challenges faced by inverse reinforcement learning and the future development trends are analyzed, and finally the progress and application prospects of inverse reinforcement learning are discussed.

Key words Reinforcement learning (RL), inverse reinforcement learning (IRL), linear inverse reinforcement learning, deep inverse reinforcement learning, adversarial inverse reinforcement learning

Citation Song Li, Li Da-Zi, Xu Xin. A survey of inverse reinforcement learning algorithms, theory and applications. *Acta Automatica Sinica*, 2024, **50**(9): 1704−1723

随着人工智能技术的不断发展, 智能决策与控制技术变得越来越重要, 促进了机器学习的一个重要领域——强化学习 (Reinforcement learning, RL) 的发展. 目前, 强化学习的算法和理论体系日趋完善, 已经广泛应用于各个领域, 具有巨大的发展前景, 吸引了学术界和工业界的学者对该领域进行深入的探索研究[1−4]. 强化学习算法将策略优化问题建模为马尔科夫决策过程 (Markov decision process, MDP), 其主要目标是通过智能体与环境的试错交互, 最大化累积奖励函数和优化策略. 奖励函数作为 MDP 的重要组成部分, 对于强化学习的效率和性能具有重要影响[5]. 人为设计奖励函数具有很强的主观性和经验性, 奖励函数的差异会影响强化学习的策略优化. 因此, 如何设计有效的奖励函数是一项非常重要的工作. 然而, 在复杂环境中, 需要考虑多种因素对奖励函数的影响, 很难人为设定高效的奖励函数, 这成为制约强化学习应用发展的瓶颈. 新南威尔士大学 Bain 等[6] 首次较系统地给出了基于行为克隆 (Behavior cloning, BC) 的模仿学习 (Imitation leaning) 的定义, 该方法采用监督学习的方式, 通过模仿人类专家的动作来学习随机或

收稿日期 2023-02-24 录用日期 2023-04-25
Manuscript received February 24, 2023; accepted April 25, 2023
国家自然科学基金 (62273026) 资助
Supported by National Natural Science Foundation of China (62273026)
本文责任编委 杨涛
Recommended by Associate Editor YANG Tao
1. 北京化工大学信息科学与技术学院 北京 100029 2. 国防科技大学智能科学学院 长沙 410073
1. College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029 2. College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073

确定性策略网络. 然而该方法无需学习奖励和推理行为背后产生的内在原因[7], 只能在专家演示下学习最优策略, 无法突破和超越专家演示的最优策略[8]. 因此, 针对如何设计准确的奖励函数的问题, 2000 年加州大学伯克利分校 Ng 等[9] 首次提出逆强化学习 (Inverse reinforcement learning, IRL) 的概念. 该算法的基本思想是首先利用专家演示反向推导 MDP 的奖励函数, 然后根据学习的奖励函数去优化策略, 进行正向的强化学习[10].

近年来, 逆强化学习算法的理论和应用领域不断被完善. 从解决问题的方面来看, 逆强化学习算法可以分为三大分支. 第一个分支主要包括 2004 年斯坦福大学 Abbeel 等[11] 提出的学徒学习逆强化学习 (Apprenticeship learning inverse reinforcement learning, ALIRL), 2006 年 Ratliff 等[12] 提出的最大边际规划逆强化学习 (Maximum margin planning inverse reinforcement learning, MMPIRL) 等算法. 然而, 这类算法存在模糊性问题, 即不同的奖励对应相同的策略. 进而衍生出第二个分支, 基于熵的逆强化学习算法, 主要包括 2008 年卡内基梅隆大学 Ziebart 等[13] 提出的最大熵逆强化学习 (Maximum entropy inverse reinforcement learning, MEIRL), 2011 年马克斯–普朗克智能系统研究所 Boularias 等[14] 提出的相对熵逆强化学习 (Relative entropy inverse reinforcement learning, REIRL) 等. 基于熵的逆强化学习最初实现的是特征到奖励的线性映射, 随着环境复杂度的增大, 2016 年牛津大学 Wulfmeier 等[15] 提出深度逆强化学习算法, 借助神经网络能拟合任意非线性函数的能力来学习非线性奖励函数[16−17]. 在专家演示下, 虽然基于熵的逆强化学习算法一定程度上提高了奖励函数的学习精度, 但有限和非最优的专家演示依然影响着奖励函数的学习. 因此, 2016 年, 斯坦福大学 Ho 等[18] 给出了生成对抗逆强化学习 (Generative adversarial inverse reinforcement learning, GAIRL) 的基本定义, 通过 RL 和 IRL 的学习迭代不断优化专家演示, 提高奖励的学习精度. 此外, 在复杂的非线性环境下, 2011 年斯坦福大学 Levine 等[19] 提出基于高斯过程的逆强化学习 (Inverse reinforcement learning with Gaussian processes, GPIRL), 利用高斯函数的高度非线性确定每个特征与策略的相关性, 求解奖励函数. 三个分支既相互独立又相互补充, 基于以上探讨, 如何构建高效可靠的奖励函数和求得最优策略是逆强化学习研究的重点. 在求解的过程中, 针对出现的模糊性和专家演示非最优的问题, 研究者们提出了不同的应对策略, 在一定程度上解决了这些问题. 本文首先介绍逆强化学习算法的发展历程, 然后重点介绍和讨论了逆强化学习算法的应用进展及算法面临的挑战.

本文内容安排如下: 第 1 节介绍了马尔科夫决策过程、逆强化学习、强化学习、行为克隆等算法的基本概念和知识; 第 2 节介绍解决 MDP 问题的逆强化学习算法的研究进展; 第 3 节介绍了逆强化学习算法的应用进展; 第 4 节介绍逆强化学习算法面临的挑战及解决方案; 第 5 节对逆强化学习算法的未来进行展望; 第 6 节对本文内容进行总结.

1　逆强化学习的背景与提出

本节首先回顾了 MDP、RL 算法、IRL 算法的基本知识, 然后分析了 IRL 算法、RL 算法、BC 三者之间的差异.

RL 算法通过智能体与环境的交互学习, 旨在通过最大化奖励期望来获得最优策略, 如图 1 所示. 智能体根据当前的状态, 采取相应的动作作用于环境, 然后获得下一个状态和环境给予的奖励. RL 算法通过不断迭代来求解强化学习问题, 该问题被建模为 MDP. MDP 被定义为一个五元组 (S, A, T, γ, R), 其中 S 是状态集合, A 是动作集合, $T = \{P^{a_t}_{s_t, s_{t+1}}\}$ 是状态转移概率集合, $P^{a_t}_{s_t, s_{t+1}}$ 是在状态 s_t 下执行动作 a_t 到达新状态 s_{t+1} 的状态转移概率, $\gamma \in [0, 1)$ 是折扣因子, $R: S \to A$ 是奖励函数. 基于 MDP 的定义, 逆强化学习问题被建模为无奖励函数的 MDP, 用 MDP\R $= (S, A, T, \gamma)$ 来表示. 其中 MDP 以状态和动作的图形结构表示决策过程、相关的奖励以及状态之间的随机转换, 因此分为确定性 MDP 和随机性 MDP, 如图 2 所示.

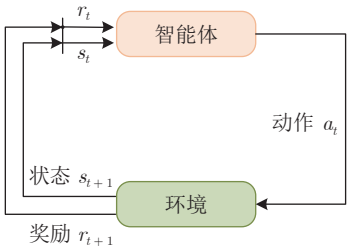


图 1　强化学习模型

Fig. 1　Model of reinforcement learning

图 2(a) 和 2(c) 表示具有确定性状态转移的 MDP, 即智能体在状态 s_t 下执行动作 a_t 得到确定新状态 s_{t+1}; 图 2(b) 和 2(d) 表示具有随机性状态转移的 MDP, 即智能体在状态 s_t 下执行动作 a_t 得到新状态 s_{t+1} 或 s'_{t+1} [13].

为了解决强化学习和逆强化学习问题, 需要用

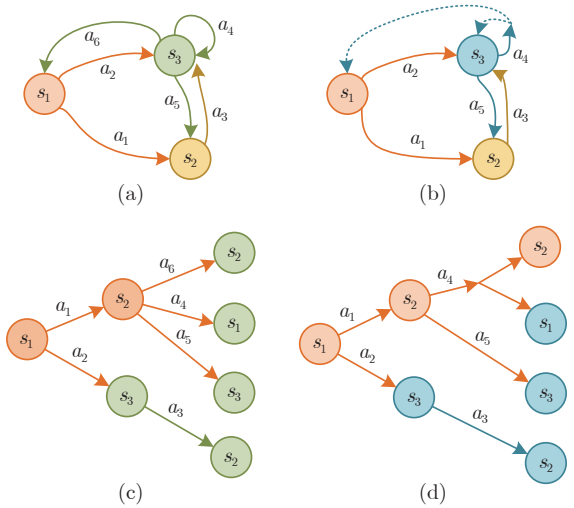


图 2 MDP ((a) 和 (c) 是确定性 MDP; (b) 和 (d) 是随机性 MDP)

Fig. 2 MDP ((a) and (c) are the deterministic MDP; (b) and (d) are the stochastic MDP)

到两个典型的 MDP 的性质.

性质 1 (Bellman 等式). 给定一个 MDP $M = (S, A, \{P^{a_t}_{s_t, s_{t+1}}\}, \gamma, R)$ 和一个策略 $\pi : S \to A$. 对于所有的状态 $s_t \in S$, 动作 $a_t \in A$, 状态价值函数 V^π 和动作价值函数 Q^π 满足:

$$V^\pi(s_t) = \sum_{a_t} \pi(s_t, a_t) \times \sum_{s_{t+1} \in S} P^{a_t}_{s_t, s_{t+1}} \left[R^{a_t}_{s_t, s_{t+1}} + \gamma V^\pi(s_{t+1}) \right] \tag{1}$$

$$Q^\pi(s_t, a_t) = \sum_{s_{t+1} \in S} P^{a_t}_{s_t, s_{t+1}} \left[R^{a_t}_{s_t, s_{t+1}} + \gamma \sum_{a_{t+1}} Q^\pi(s_{t+1}, a_{t+1}) \right] \tag{2}$$

性质 2 (Bellman 最优). 给定 MDP $M = (S, A, \{P^{a_t}_{s_t, s_{t+1}}\}, \gamma, R)$ 和一个策略 $\pi : S \to A$. 对于所有的 $s_t \in S$, π 取得最优策略的条件是当且仅当 $\pi^*(s) \in \arg\max_{a_t \in A} Q^\pi(s_t, a_t)$.

在强化学习框架下, 智能体通过与环境的交互试错获得奖励来指导行为, 获得最优策略, 即在每个状态下对应的最优的动作[20]. 图 3 展示了 RL、IRL 和 BC 的算法框架. 在复杂环境中完成任务时, RL 算法的奖励获取困难, 因此提出 IRL 算法, 首先根据专家演示求出最佳奖励函数, 然后利用学习的奖励函数和正向的 RL 算法寻找最优策略 π^*. BC 算法跟 IRL 算法相似, 都是已知环境动态 $P(s'|s, a)$ 和专家策略 $\hat{\pi}$, 不同的是 IRL 算法需要求出奖励函数, BC 算法不需要求出奖励函数, 直接模仿学习最优策略 π^*. IRL 算法的提出是为了规避 BC 算法单纯地模仿专家行为而不去推理行为背后产生的原因, 更好地学习奖励和优化策略.

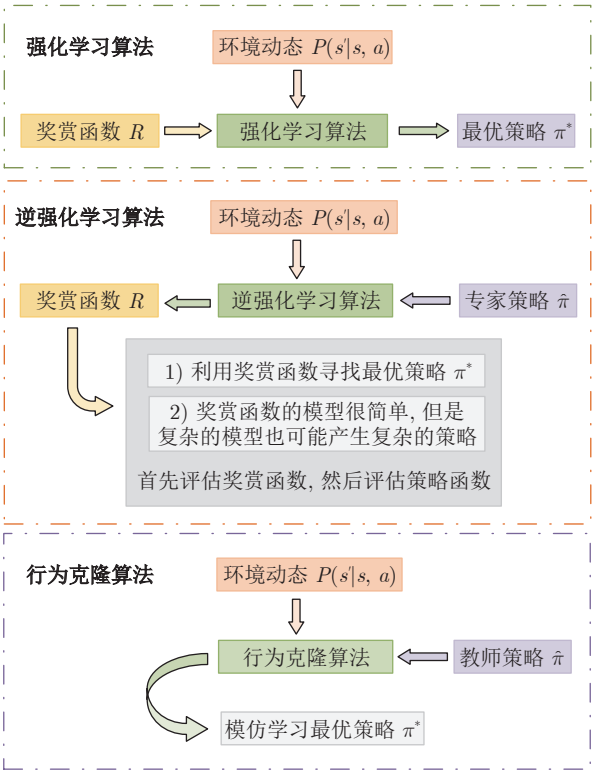


图 3 RL、IRL、BC 的算法框架

Fig. 3 Frameworks for RL, IRL, BC

2 逆强化学习算法研究进展

逆强化学习算法是针对未知奖励函数的强化学习问题而提出的解决方法. 自从逆强化学习的概念被提出后, 首先研究者提出了边际类逆强化学习算法, 旨在从专家演示中学习奖励函数, 获得与专家策略相近的策略[11−12]. 然而该类算法存在模糊歧义问题, 导致不同的奖励对应相同的策略. 针对这个问题, 促进了基于熵的逆强化学习算法的产生和发展, 因此衍生了最大熵逆强化学习[13]、相对熵逆强化学习[14]、贝叶斯逆强化学习[21] 等. 以上算法均是实现输入特征到输出奖励的线性映射, 当面对复杂动态的非线性环境时, 这种以线性求解奖励的方式会影响奖励的学习准确度. 因此, 结合神经网络的感知力和强化学习的决策力提出基于神经网络的逆强化学习. 在线性逆强化学习算法中引入神经网络, 研究者提出学徒学习深度逆强化学习[22]、最大边际规划深度逆强化学习[23]、最大熵深度逆强化学习[15]、生成对抗逆强化学习[18] 等. 此外, 非线性逆强化学

习还包括基于高斯过程的逆强化学习[19], 以提高奖励函数的学习性能.

自从 Ng 等正式提出逆强化学习的概念, 该算法取得了许多关键性进展. 对逆强化学习算法进行总结主要归纳为线性逆强化学习算法和非线性逆强化学习算法, 如图 4 所示.

因此, 线性逆强化学习算法主要包括学徒学习逆强化学习、最大边际规划逆强化学习、最大熵逆强化学习、相对熵逆强化学习、贝叶斯逆强化学习等算法; 非线性逆强化学习算法主要包括边际类深度逆强化学习、最大熵深度逆强化学习、高斯过程逆强化学习、神经网络逆强化学习等算法.

2.1　线性逆强化学习算法

基于环境的特点, 研究者们首先提出学徒学习逆强化学习、最大边际规划逆强化学习, 但针对边际类逆强化学习存在的模糊歧义问题, 研究者又提出了最大熵逆强化学习、相对熵逆强化学习、贝叶斯逆强化学习等算法来学习线性的奖励函数.

2.1.1　学徒学习逆强化学习算法

针对如何设计强化学习算法中准确的奖励, Ng 等通过状态特征构建基函数, 将求反馈信号函数的任务转化为求解每个基函数权重的任务, 求解有限状态空间和大规模状态空间的 MDP 问题. 该算法是一种高效的模仿学习范式, 是边际类逆强化学习算法的早期模型. 在离散空间中, 该算法的主要思想是给定一些已知专家状态−动作 $(s_1, a_1, s_2, a_2, \cdots, s_n, a_n)$, 求满足贝尔曼最优的奖励. 在连续状态空间中, 利用函数近似实现状态到奖励的线性映射. 利用该思想可以解决中小规模的离散和连续空间问题.

在此基础上, Abbeel 等[11] 首次明确提出学徒学习逆强化学习算法. 该算法主要思想是将学徒学习问题建模为逆强化学习的 MDP (MDP\R), 首先从专家的观测中提取特征, 计算特征期望, 然后利用特征映射 $\phi_{n\times 1}$ 和专家的特征期望 $\mu_E^{n\times 1}$, 在未知的奖励函数 $R^* = \omega_{n\times 1}^{*\mathrm{T}}\phi_{n\times 1}$ 的基础上寻找一个策略 $\tilde{\pi}$, 使其特征期望与专家策略的特征期望相近, 满足 $\|\mu^{n\times 1}(\tilde{\pi}) - \mu_E^{n\times 1}\|_2 \le \varepsilon$, 其中 ω 是奖励权重, $\mu(\tilde{\pi})$ 是学习者的特征期望, ε 是给定的阈值. 但为了防止最大边际 (Maximum margin) 走向 0 和无穷大两个极端, 在约束条件时加入了一些限制. 算法的主要步骤如下:

1) 获得专家轨迹 π_E, 迭代次数 $g = 0$ 时, 随机设定一个奖励权重 ω, 提取专家轨迹特征 $\phi_{n\times 1}(s_1^{(E)})$, $\phi_{n\times 1}(s_2^{(E)}), \cdots, \phi_{n\times 1}(s_n^{(E)})$;

2) 定义折扣因子 γ 并计算专家轨迹的特征期望 $\mu^{n\times 1}(\pi_E) = \sum_t \gamma^t \phi_{n\times 1}(s_t^E)$, $t = 1, 2, \cdots, n$;

3) 奖励设为 $r_\phi = \omega_{n\times 1}^{\mathrm{T}}\mu^{n\times 1}(\pi_g)$, 智能体与环境互动, 产生轨迹 π_g, 提取特征为 $\phi_{n\times 1}(s_1^{(g)})$, $\phi_{n\times 1}(s_2^{(g)}), \cdots, \phi_{n\times 1}(s_n^{(g)})$, 计算智能体的轨迹期望值 $\mu^{n\times 1}(\pi_g) = \sum_t \gamma^t \phi_{n\times 1}(s_t^{(g)})$;

4) 求解最优 $\omega_{n\times 1} = \omega_{n\times 1}^*$ 以更新线性奖励 $r_\phi = \omega_{n\times 1}^{\mathrm{T}}\mu^{n\times 1}(\pi_g)$;

5) $g \leftarrow g + 1$, 若达到最大迭代次数, 终止, 否则转步骤 3).

该算法的提出最初解决了 Grid world 问题和小车驾驶仿真问题, 通过从专家演示中学习奖励函数, 使得在该奖励函数下所得的最优策略在专家演示策略附近, 从而实现策略优化.

此外, 在以上 ALIRL 的理论基础上, 为解决复杂动态环境中状态和动作空间的连续问题, 将学徒学习逆强化学习与人工智能方法相结合. Nguyen 等[24] 提出了一种新的基于 IRL 的学徒学习方法来解决牧羊问题中的连续动作和状态空间问题, 该方法将径向基函数与 K-mean 算法相结合, 生成有效的特征映射函数. 牧羊环境仿真结果表明, 经过 IRL 训练的混合算法能找到更短的路径, 减少人类

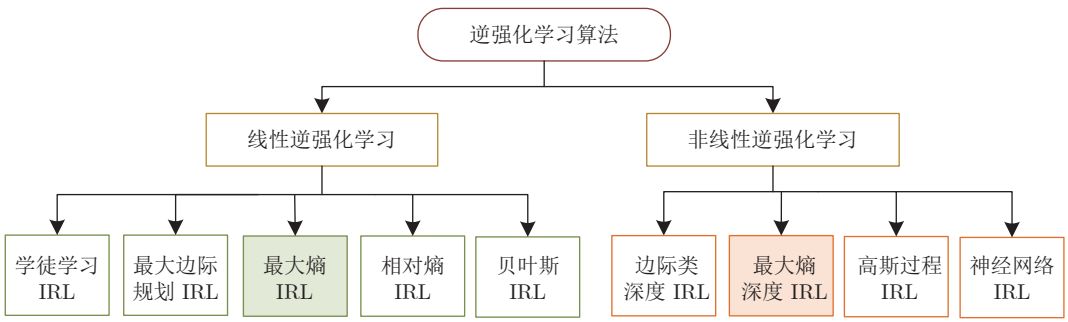


图 4　逆强化学习算法分类

Fig. 4　Classification of IRL algorithms

的意外行为, 比人类更精确. 面对大状态空间中计算时间长的问题, Hwang 等[25] 提出了一种从整个状态空间中搜索关键状态来提取关键特征的方法, 形成奖励特征集, 从而派生出奖励函数, 提高算法学习效率. 高速行驶的汽车实验仿真结果表明智能体利用该方法可以学习到接近专家的行为. 在部分可观测马尔科夫决策过程 (Partially observable Markov decision processes, POMDP) 和连续/高维空间中, 金卓军等[26] 研究了基于奖励逼近的学徒学习方法, 使用近似算法 (如基于点的值迭代) 或使用降维方法 (如主成分分析的特征抽象算法) 学习奖励, 缓解维度高的问题. Levine 等[27] 提出一种基于特征构建的逆强化学习 (Feature construction for inverse reinforcement learning, FIRL), 通过构建与示例策略相关的组件特征的逻辑连接来获得奖励特征, 然后利用奖励来学习完整平稳的策略.

以往的研究, 奖励函数是通过估计最大似然、贝叶斯或信息论的方法来实现的. 在复杂状态空间, 为了更准确地学习奖励, 在学徒学习的基础上引入了集成的思想. 基于集成和模糊逻辑的学徒学习逆强化学习方法依赖增强分类方法, 使用一个近似的奖励函数, 学习类似于专家演示的适应更强的策略[28]. 基于集成方法的无模型的学徒学习 IRL 算法, 将奖励作为期望特征的参数函数, 根据专家演示与 RL 诱导的状态轨迹之间的残差符号解决核心的策略优化问题[29].

基于学徒的逆强化学习算法分为两步: 第一步在已经迭代得到的最优策略中, 利用最大边际方法求出当前奖励的参数值; 第二步利用求出的奖励参数值进行正向强化学习求得最优的策略, 然后重复第一步. 该方法采用最简单的思想从专家演示中学习奖励, 使得在该奖励下所得的最优策略接近专家策略. 这是最早提出的逆强化学习思想, 能很快地从专家演示中学习奖励, 优化策略. 所应用的仿真实验表明算法的性能较好, 且学徒学习的思想已经与多种算法思想相结合, 提高了算法的性能.

2.1.2 最大边际规划逆强化学习算法

基于最大边际规划的逆强化学习算法的概念由 Ratliff 等[12] 提出. 该方法将逆强化学习问题首先建模为二次规划问题, 然后利用凸优化方法进行求解, 试图估计使最优策略与所有其他策略之间的边际最大化的奖励.

最大边际规划方法的训练集为 $D=\{X_i,\ A_i,\ P_i,\ f_i,\ y_i,\ \mathcal{L}_i\}_{i=1}^{n}$, 其中 X_i 为状态空间, A_i 表示动作空间, P_i 表示状态转移概率, f_i 为特征矩阵, y_i 为专家轨迹, $\mathcal{L}_i$ 为策略 y 与第 i 条专家轨迹 y_i 之间的损失函数. 在该方法框架下, 学习者在不断的学习中寻找特征函数与奖励之间的线性映射 ω, 将逆强化学习问题转化为求解二次规划问题:

$$\begin{cases}\min\limits_{\omega,\varsigma_i}\dfrac{1}{2}\|\omega_{n\times 1}\|^2+\dfrac{\tau}{n}\sum\limits_i\beta_i\varsigma_i^q\\ \text{s.t.}\ \ \forall_i\ \omega_{n\times 1}^{\mathrm{T}}f_{i,\,n\times 1}\left(y_i\right)+\varsigma_i\geq\\ \qquad\max\limits_{y\in y_i}\omega_{n\times 1}^{\mathrm{T}}f_{i,\,n\times 1}\left(y\right)+\mathcal{L}_i\left(y\right)\end{cases}\tag{3}$$

其中, ς 是松弛变量, τ 是关于缩放松弛变量的超参数, β_i 是与数据序列长度相关的归一化标量, $\omega_{n\times 1}^{\mathrm{T}}f_{i,\,n\times 1}\left(y\right)$ 是累积奖励, $f_{i,\,n\times 1}\left(y\right)$ 是期望特征计数, $q\in\{1,\,2\}$ 表示 ℓ_1, ℓ_2 惩罚.

然后将二次规划问题转换成一个优化问题, 由于目标函数是凸的, 且不可微, 故用次梯度方法求解. MMPIRL 把学习专家行为框定为一个策略空间上的最大边际结构预测问题. 利用该方法可以学习从特征到代价的线性映射, 因此学习获得的最优策略接近于专家的行为. 为解决在室外移动机器人的路线规划任务中奖励设计难的问题, 将状态空间的成本地图作为输入, 利用最大边际规划计算最小风险的路径. 随后 Ratliff 等[30] 提出基于功能梯度下降的集成的 MMP(MMP based on boosting, MMPBoost) 方法, 使用简单的二元分类或回归来提高 MMP 模仿学习的性能, 并自然扩展到结构化最大边际预测问题.

针对动态复杂环境中奖励学习难的问题, 基于最大边际规划的思想引入神经网络等算法. Choi 等[31] 提出了一种基于递归神经网络和最大边际逆强化学习的未来轨迹预测框架, 利用当前位置和相应的静态场景信息最大化奖励, 预测动态场景中智能体的未来轨迹. 在公共 KITTI 数据集上的实验结果表明, 该方法能显著提高预测精度. 针对自动驾驶中决策难的问题, 高振海等[32] 利用最大边际逆强化学习算法将驾驶数据作为专家演示数据, 建立相应的奖励函数, 实现仿驾驶员的自动驾驶决策. 仿真测试结果表明, 该方法降低了奖励的建立难度, 学习的策略与驾驶员的行为具有更高的一致性.

然而 ALIRL 和 MMPIRL 等边际类逆强化学习方法可能存在模糊歧义、计算复杂的问题. 在这种情况下, 所学习到的奖励往往具有随机的偏好. 为了克服这个缺点, 研究者们利用概率模型, 提出基于熵的逆强化学习. 如果模型中状态概率已知, 则建模为最大熵逆强化学习问题, 否则建模为相对熵逆强化学习问题.

2.1.3 最大熵逆强化学习算法

在所有满足约束的概率模型中, 利用最大熵原

理所获得的模型是最大熵模型. 由于最大熵分布所选取的模型没有对未知 (即除了约束已知外) 做任何主观假设, 因此最大熵的模型是最好的模型, 可以帮助逆强化学习算法避免歧义的问题. 因此, Ziebart 等[13] 开创性地将最大熵逆强化学习的分布问题转化为标准的带有约束条件的优化问题:

$$\begin{cases} \max \ -p\left(\varsigma_i|\omega_{n\times 1}\right)\log p\left(\varsigma_i|\omega_{n\times 1}\right) \\ \text{s.t.} \ \sum\limits_{path\,\varsigma_i} p\left(\varsigma_i|\omega_{n\times 1}\right)\phi_{\varsigma_i,\,n\times 1}=\tilde{\phi}_{n\times 1} \\ \quad \sum p\left(\varsigma_i|\omega_{n\times 1}\right)=1 \end{cases} \tag{4}$$

其中, $p\left(\varsigma_i|\omega_{n\times 1}\right)$ 是 MEIRL 算法的模型概率, $\phi_{\varsigma_i,\,n\times 1}$ 是沿着轨迹 ς_i 的状态特征计数, $\tilde{\phi}_{n\times 1}$ 是专家演示轨迹下状态的期望的经验特征计数, $\omega_{n\times 1}$ 是奖励权重.

然后通过拉格朗日乘子法将带约束的优化问题转化为无约束优化问题的拉格朗日函数 L. 然后求凸函数 L 关于 $p\left(\varsigma_i|\omega_{n\times 1}\right)$ 的导数. 令导数为零, 获得 MEIRL 算法的模型概率 $p\left(\varsigma_i|\omega_{n\times 1}\right)=\mathrm{e}^{\sum \omega_{n\times 1}^{\mathrm{T}}\phi_{n\times 1}}/z\left(\omega\right)$, 其中 $z\left(\omega\right)$ 为配分函数. 参数 $\omega^*=\arg\max_{\omega}L$ 为逆强化学习奖励函数的权重, 既可以用最大似然法进行求解, 也可以用梯度最优化进行求解. 最终, 利用权重和特征函数的线性组合求得奖励函数 $R=\omega_{n\times 1}^{\mathrm{T}}\phi_{n\times 1}$. 通过对驾驶路线建模, 利用收集的 GPS 数据求解奖励和优化策略的问题, 经实验验证, MEIRL 算法性能较好.

配分函数 $z\left(\omega\right)$ 是一个常量参数, 在高维复杂环境下, 计算较为困难. 在小的、离散环境中, 研究者起初用动态规划计算配分函数. 针对复杂环境, 研究者提出利用基于样本的近似[33−34]、值函数近似[35]、拉普拉斯近似[36] 等方法来求解配分函数.

近年来, 基于最大熵的逆强化学习的理论和应用已经得到有效的发展. 根据最大熵模型的求解方式, 算法可以分为基于最大似然估计的最大熵逆强化学习 (Maximum entropy IRL based on maximum likelihood estimation, MEIRL-MLE)[37] 和基于梯度的最大熵逆强化学习 (Maximum entropy IRL based on gradient, MEIRL-GD)[13, 38]. MEIRL-MLE 的主要研究工作包括: 针对具有连续状态和动作的时间随机系统, Aghasadeghi 等[37] 通过对输入进行约束, 构建最大熵分布, 然后应用极大似然估计来近似给定的有限样本路径集的分布参数. MEIRL-GD 的主要研究工作包括: Ziebart 等[13] 提出最大熵方法来解决模糊歧义问题, 最小化期望的经验特征计数和学习者期望的特征计数之间的梯度. 文献 [38] 利用随机 MDP 对交通进行建模, 并基于专家驾驶经验, 采用深度神经网络 (Deep neural networks, DNN) 逼近专家驾驶员的未知奖励函数, 获得最佳驾驶行为. 文献 [39] 利用最大熵原理将带约束的逆强化学习问题转化为约束非凸优化问题, 然后利用指数梯度下降算法求解奖励函数, 提高算法性能.

在求解最大熵模型时, 由于不可微点的存在导致最大熵逆强化学习收敛慢, 同时该算法存在计算复杂、过拟合等问题. 为解决算法中存在的这些问题, 许多新的最大熵逆强化学习算法被相继提出. 分层逆强化学习 (Hierarchical inverse reinforcement learning, HIRL) 将任务划分为具有短期奖励的子任务进行学习, 以期更快地收敛到成功的策略[40]. 从因果关系的视角解决逆强化学习问题, 最大折扣因果熵 (Maximum discounted causal entropy, MDCE) 和最大平均因果熵方法将最大因果熵框架扩展到无限视界 IRL[41]. 在连续域环境中, 基于采样的最大熵逆强化学习算法利用先验知识设计采样器计算配分函数, 有效地学习人类驾驶行为[42]. 为更易于学习奖励, 研究者利用基于内部奖励的驱动模型来模拟人类的决策机制, 将连续行为建模问题转换为离散设置[43]. 为提高最大熵逆强化学习算法的泛化性、稀疏性, 基于近端优化的最大熵 IRL 利用具有良好稀疏解的 FTPRL (Follow-the-proximally-regularized-leader) 方法求解奖励[44].

此外, 为了适应复杂的任务环境, 多任务逆强化学习方法利用元学习将所提出的方法扩展到计算效率更高的最大因果熵 IRL, 以解决从专家演示中推断多个奖励函数的问题[45].

针对利用最大熵逆强化学习算法求解奖励和最优策略中存在的计算复杂、过拟合、低收敛性等问题, 研究者们进行了一系列探讨, 提出了新的理论和方法. 为了验证算法的性能, 基于各类实验进行对比仿真. 基于最大熵的逆强化学习是当前比较流行的算法, 算法的优越性和存在的问题使得该算法值得进一步研究.

2.1.4　基于相对熵逆强化学习算法

在最大熵逆强化学习算法中, 专家演示的似然函数的计算需要已知状态转移函数. 然而在无模型的逆强化学习中, 状态转移函数是未知的. 为了解决这个问题, Boularias 等[14] 提出将该 IRL 问题建模为相对熵逆强化学习. 设 $P\left(\varsigma_i\right)$ 为专家演示轨迹 ς_i 的概率分布, $Q\left(\varsigma_i\right)$ 为在策略 π 下轨迹 ς_i 产生的概率分布. 在该模型中, 最大熵逆强化学习转换为求 $P\left(\varsigma_i\right)$ 和 $Q\left(\varsigma_i\right)$ 相对熵最小的问题:

$$\begin{cases} \min\limits_{P} \sum\limits_{\varsigma_i \in \mathcal{G}} P(\varsigma_i) \dfrac{P(\varsigma_i)}{Q(\varsigma_i)} \\ \text{s.t.} \quad \left| \sum\limits_{\varsigma_i \in \mathcal{G}} P(\varsigma_i) \phi_{s,n\times 1}^{\varsigma_i} - \hat{\phi}_{s,n\times 1} \right| \leq \varepsilon_s \\ \quad \sum\limits_{\varsigma_i \in \mathcal{G}} P(\varsigma_i) = 1 \\ \quad \forall\, \varsigma_i \in \mathcal{G}:\ P(\varsigma_i) \geq 0 \end{cases} \tag{5}$$

其中, ς_i 为专家演示轨迹, $\mathcal{G}$ 为专家演示, $\phi_{s,n\times 1}^{\varsigma_i}$ 为学习的奖励的特征函数, $\hat{\phi}_{s,n\times 1}$ 为经验的特征函数, 阈值通过 Hoeffding 的界限来进行求取. 式 (5) 用拉格朗日乘子法和 KKT 条件转化为无约束的问题 $L(w_{n\times 1})$. 与最大熵逆强化学习参数的求解方法一样, 利用最大似然法和次梯度法进行求解, 因此获得 $L(w_{n\times 1})$ 的导数 $\nabla_w L(w_{n\times 1})$, 采用重要性采样, 得到基于样本的梯度. 令 $\nabla_w L(w_{n\times 1})$ 等于 0, 即可求得相对熵中最优的奖励权重 $w_{n\times 1}$, 从而求得奖励和最优策略. 经严格的推理和实验验证, 该算法能从较少的演示样本中学习到较好的奖励函数和策略. 在轨迹追踪、Grid world 等实验中解决了在较少的演示样本下进行策略优化的问题.

此外, 基于相对熵逆强化学习的研究中存在的高维环境中计算复杂、奖励求解困难的问题, 元学习、集成算法等被相继引入以提高算法性能. 基于端到端无模型的逆强化学习算法利用自动编码器和集成算法来降低高维环境下的计算复杂度, 利用相对熵来度量预测和演示奖励函数之间的差异[46]. 基于相对熵的元逆强化学习方法, 利用元学习方法构建目标任务, 采用相对熵概率模型对奖励进行建模, 实现利用少量样本快速求解奖励的目标[47]. 此外, 相对熵逆强化学习算法可以用来学习居民的习惯偏好, 以预测居民的认知健康诊断[48].

在缺少足够专家演示样本以及状态转移概率未知的情况下, 利用传统逆强化学习算法求解奖励存在速度慢、精度低甚至无法求解的问题. 因此, 在最大熵逆强化学习的基础上衍生出相对熵逆强化学习算法. 即使缺少足够数目的专家演示样本和状态转移概率信息的情况下, 利用相对熵的思想, 逆强化学习依然可以较好地求解奖励函数, 从而实现策略优化.

2.1.5 贝叶斯逆强化学习算法

贝叶斯逆强化学习 (Bayesian inverse reinforcement learning, BIRL) 从贝叶斯模型的角度对逆强化学习问题进行了建模, 该算法不需要一个完全指定的策略作为逆强化学习的输入, 也不需要假设专家是正确可靠的[21, 49]. 此外, 还可以将有关逆强化学习问题的外部信息融合到模型的先验知识中, 或者使用来自多个专家的演示. 贝叶斯逆强化学习模型如图 5 所示, 首先从先验分布和给定奖励函数的专家行为概率模型中推导出奖励的后验分布, 无需给奖励函数假设一定的结构 (如线性). BIRL 主要包括学习奖励函数和学习专家策略两个任务.

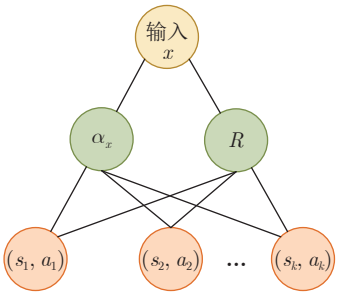


图 5 贝叶斯逆强化学习模型

Fig. 5 Bayesian inverse reinforcement learning model

假设一个 MDP $M = (S, T, \gamma)$, 奖励函数 R 从一个已知的先验分布 P_R 中进行选择. 首先智能体 x 获得的一系列关于专家行为的演示 $O_x = \{(s_1, a_1), (s_2, a_2), \cdots, (s_k, a_k)\}$, 其表示 x 处于 s_i 状态, 在第 i 步时采取了 a_i 的动作. 因为专家每个状态下的策略是不变的, 因此奖励函数 R 的条件概率 $P_{r_x}(O_x|R)$ 相互独立. 最大化累积奖励函数的专家目标等同于寻找每个状态下最优的 Q^* 值对应的动作, 因此获得的 Q 值最大. 然后利用指数分布对状态动作 (s_i, a_i) 进行建模为 $P_{r_x}(O_x|R) = \mathrm{e}^{\alpha_x \sum_i Q^*(s_i, a_i, R)}/Z_i$, 其中参数 α_x 表示智能体选择高值对应的动作概率的置信度, Q^* 为势函数, Z_i 为归一化常数. 最后利用贝叶斯定理计算奖励函数 R 的后验概率 $P_{r_x}(R|O_x) = P_{r_x}(O_x|R) P_R(R)/P_r(O_x)$, 其中 $P_R(R)$是先验概率, 求得智能体的策略. 利用自推进粒子模型和圈养孔雀鱼种群对提出的方法进行论证, 表明了算法在无需假设专家演示是最优的情况下的可行性和优越性.

针对专家演示在实践中可能存在单一、非最优、有限的问题, 贝叶斯逆强化学习将人类评估反馈与逆强化学习过程相结合, 给出了足够多样化的演示和领域转换模型, 使用贝叶斯规则对代理策略进行迭代改进, 以增强学习奖励的能力[50].

此外, 传统的逆强化学习算法大多假设行为数据来自于优化单个奖励函数, 但这种假设在实践中很难保证. 基于贝叶斯逆强化学习的理论基础, 针对多个奖励函数的问题, 提出了参数化贝叶斯方法. Ranchod 等[51] 提出基于逆强化学习的非参数贝叶斯奖励分割方法, 使用贝叶斯非参数方法从片段推断奖励函数. 汽车驾驶领域和模拟四轴飞行器障碍实验表明该算法能够有效地恢复演示技能. Choi 等[52]

提出了一种逆强化学习的非参数贝叶斯方法, 利用后验梯度来估计潜在的奖励函数, 在许多问题域上的实验证明该方法优于以前的方法.

为了增强贝叶斯逆强化学习的奖励学习性能, 在贝叶斯逆强化学习的基础上引入了新的图表示和自我评估方法. 一种基于图表示的贝叶斯逆强化学习使用图表示的采样轨迹对贝叶斯逆强化学习进行扩展, 来捕捉相关的任务结构[53]. 为了研究 AI 代理如何自我评估是否已经从专家那里获得了足够的演示来确保期望的性能水平, 基于贝叶斯逆强化学习和风险价值的新型自我评估方法, 使从演示中学习的智能体能够计算其性能的高置信边界[54].

贝叶斯逆强化学习方法能够通过先验分布向学习算法传递关于奖励的先验信息, 定义和最大化奖励的后验分布, 且能够通过将奖励概率建模为多个奖励函数的混合来解释复杂的行为, 提高奖励的学习性能. 针对逆强化学习中存在专家演示非最优、有限的问题, 还需要进一步加强对贝叶斯逆强化学习的理论推导.

2.2　非线性逆强化学习算法

利用基于边际类逆强化学习、最大熵逆强化学习、贝叶斯逆强化学习等算法可以求解线性奖励函数和最优策略. 然而在复杂的非线性环境中, 线性逆强化学习算法制约了奖励函数的求解精度. 为了提高奖励函数的学习准确度, 提出边际类深度逆强化学习、最大熵深度逆强化学习、高斯过程逆强化学习等算法. 在边际类逆强化学习算法中引入神经网络, 但由于边际类逆强化学习存在模糊歧义的问题, 影响了边际类深度逆强化学习的算法性能. 因此, 在最大熵逆强化学习算法中利用神经网络以端到端的方式构建输入到输出之间的映射. 此外, 为了解决复杂非线性环境中奖励的学习问题, 提出了基于高斯过程的逆强化学习.

2.2.1　边际类深度逆强化学习算法

为了增强逆强化学习在复杂非线性环境中的奖励的学习能力, 在基于边际的逆强化学习算法中考虑神经网络强大的函数逼近能力, 实现输入特征到输出的端到端的映射. 基于边际的深度逆强化学习分为学徒学习深度逆强化学习和最大边际规划深度逆强化学习.

学徒学习深度逆强化学习的主要工作包括: 针对不同用户的驾驶风格导致奖励机制调整难的问题, Huang 等[55] 提出了一种结合深度强化学习的学徒学习方法来学习具有连续动作的驾驶和停车行为. 首先构建多个策略特征, 然后根据期望目标设计奖励特征, 学习最优策略. 仿真结果表明该智能体的性能与人类驾驶相似. Bogdanovic 等[22] 提出深度学徒学习, 通过只观察游戏区域的原始像素来训练卷积神经网络玩雅达利游戏. 如图 6 所示, 通过观察专家的游戏, 算法的网络能够学习将游戏状态映射到行动, 而不需要外部提供游戏分数. Markovikj[56] 提出了一种新的学徒学习方法, 以应用于雅达利游戏的视频帧, 在奖励函数不可用的复杂、多维任务中教人工智能体玩游戏.

学徒学习逆强化学习方法需要动态模型或额外的数据采集来进行策略评估, 然而在真实世界的任务 (金融或工业流程) 中不存在精确的模拟器, 或者数据采集成本很高. 为了解决批处理设置中的挑战, 基于深度后续特征网络的逆强化学习在无批量设置下学习专家的潜在奖励结构, 估计特征期望, 生成接近专家的策略[57]. 为解决数据集有限问题, 专家演示下的深度 Q-learning 方法从相对少量的演示数据中进行学习, 利用优先回放机制在学习时自动评估演示数据[23]. 针对具有大规模高维状态空间的

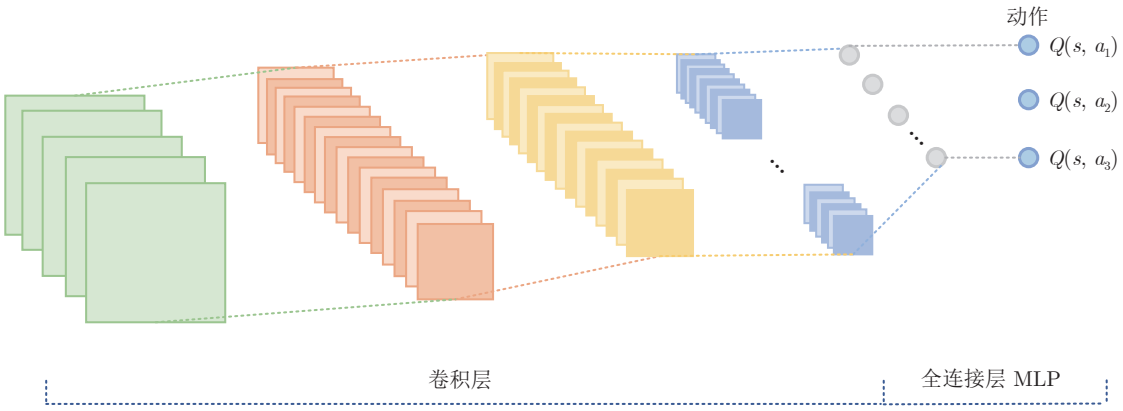


图 6　深度学徒学习模型结构

Fig. 6　Model structure of deep apprenticeship learning

逆强化学习问题, 逆强化学习在神经网络的帮助下将专家的行为推广到状态空间中未访问的区域, 通过无模型的最大边际规划最小化专家策略与被学习策略之间的最大边际来学习奖励, 更新优化学习策略[58].

在基于边际的逆强化学习算法中加入神经网络的思想增强了该类算法的学习能力, 实现输入到输出的端到端的映射. 在复杂非线性的环境中, 算法的奖励学习能力较好. 但边际类逆强化学习算法的模糊歧义问题依然困扰着算法的进一步提升, 因此提出基于熵的深度逆强化学习算法, 既能克服模糊歧义的问题, 又能实现端到端的映射, 提高算法的性能.

2.2.2 最大熵深度逆强化学习算法

在学习奖励时, 逆强化学习算法面临着两个主要的挑战: 专家演示是次优样本, 即学习的专家样本并非总是最佳样本; 奖励函数的模糊性, 即大量的奖励可能产生相同的行为. 为解决这些问题, Wulfmeier 等[15] 提出基于最大熵的非线性逆强化学习, 在复杂的城市环境中学习驾驶策略, 算法结构如图 7 所示. 最大熵深度逆强化学习通过将演示行为建模为演示轨迹上的概率分布, 然后将其约束到最大熵, 求解各类复杂非线性的奖励函数. 在复杂的对象环境中, 假设奖励是一个非线性函数的特征向量 $\phi = \{\phi_1, \phi_2, \phi_3, \cdots, \phi_n\}$. 使用深度神经网络计算出奖励函数 r^*. DNN 具有可以表示任意非线性函数的特性, 因此被视为通用逼近器, 其输入为特征向量 ϕ, 输出为奖励值 R^*.

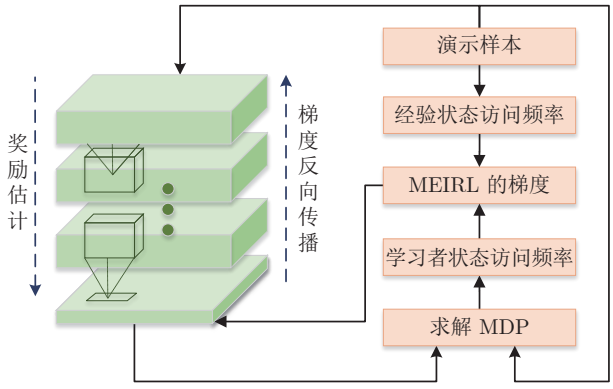


图 7 最大熵深度逆强化学习的结构

Fig. 7 Structure of maximum entropy deep inverse reinforcement learning

基于 DNN 的最大熵逆强化学习的算法学习步骤如下: 在最大熵深度逆强化学习的训练进程中, 首先随机初始化神经网络的权值 θ. 在每一个训练情节中, 该算法利用初始的网络权值 θ 和特征矢量 ϕ 来生成奖励值 R^*. 在训练集当前迭代 m 下, 利用奖励值产生策略 π, 然后利用该策略产生期望的状态访问频率 $\mathrm{E}[\mu^m]$. 利用经验状态访问频率 μ_D 与学习者状态访问频率的差和演示样本来计算 $\frac{\partial \mathcal{L}_D}{\partial R^*}$, 其中 $\mathcal{L}_D$ 是联合后验分布. 然后利用网络的反向传播更新网络的权值 θ, 利用网络的正向传播估计奖励. 该进程重复 M 次迭代完成计算. 在城市环境的路径规划中, 利用最大熵深度逆强化学习算法学习到的成本图是直接从原始传感器测量构建的, 避免了手工设计成本图的困难, 增强了算法的学习性能.

针对多个奖励稀疏分布在状态空间中的线性可解非确定性 MDP 问题, Budhraja 等[59] 将基于特征的状态评估方法与神经进化相结合, 提出一种基于神经网络在给定任务中的表现来修改神经网络的范式. 机器人和自动化中的控制问题验证了算法的可行性.

在具有大规模状态空间或时间长的任务环境中, 在最大熵深度逆强化学习算法中引入确定性有限自动机 (Deterministic finite automaton, DFA)、Dijkstra 等算法, 提高算法性能. 针对在时间延长的任务中, 潜在的奖励函数可能无法表达为 MDP 的单个状态的函数的问题, 任务引导型的深度 IRL 方法以 DFA 的形式学习任务结构, 使用 DFA 扩展原始 MDP 的状态空间, 学习奖励和策略. 实验表明, 该算法在时间扩展任务上的性能较优[60]. 针对影响路线选择的因素太多, 且送餐员的偏好难以量化的问题, 最大熵深度 IRL 可以对送餐员的偏好进行建模, 通过神经网络计算训练集的底层奖励函数, 学习当前策略, 并推荐送餐员的首选路线[61]. 针对连续状态空间和连续动作空间, 连续的最大熵深度逆强化学习算法通过基于演示的重构奖励函数的方式实现对环境模型的深度认知, 提供了计算效率高的优化过程, 如图 8 所示[62].

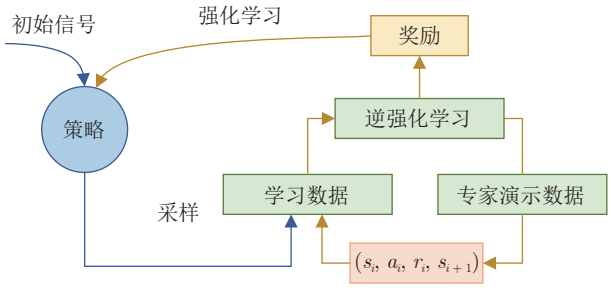


图 8 基于序列专家演示的逆强化学习进程

Fig. 8 The inverse reinforcement learning process based on sequential expert demonstration

为了提高最大熵深度逆强化学习在动态环境中的学习能力, 与环境相结合的最大熵深度强化学习算法被提出. 为了解决在动态场景中预测智能体遥远的未来轨迹受其过去轨迹的影响和受场景上下文

的影响, Choi 等[63] 提出了一种基于编码器–解码器架构的递归神经网络模型, 编码器对输入信息进行编码, 解码器根据编码器给出的上下文向量生成未来的轨迹. 实验结果表明该方法大大提高了模型的预测性能. 为探索如何利用逆最优控制从演示中学习行为, Finn 等[33] 提出了一种基于策略优化的最大熵深度逆最优控制 (Maximum entropy inverse optimal control, ME-IOC) 算法, 利用神经网络学习任意非线性代价函数且制定了有效的基于样本的近似来解决高维连续系统下的难题. 真实机器人操作问题上证明 ME-IOC 算法在任务复杂性和样本效率方面都比之前的方法有了实质性的改进. Wang 等[64] 利用深度最大熵逆强化学习恢复非线性奖励函数, 对电动自行车的过马路行为进行了仿真, 从而帮助自动驾驶汽车进行高效决策. Fahad 等[65] 提出了一种利用最大熵深度逆强化学习来学习人类导航行为的方法. 首先在行人轨迹数据集组成的专家演示下通过深度神经网络近似奖励, 然后利用非线性奖励函数捕获人的导航行为. 结果表明, 该方法具有良好的预测精度, 能够生成与真实人类轨迹相似的行人轨迹. Zhou 等[66] 提出了一种基于最大熵深度逆强化学习的驾驶员跟车行为学习框架, 从驾驶数据中学习由全连接神经网络表示的奖励函数, 如图 9 所示.

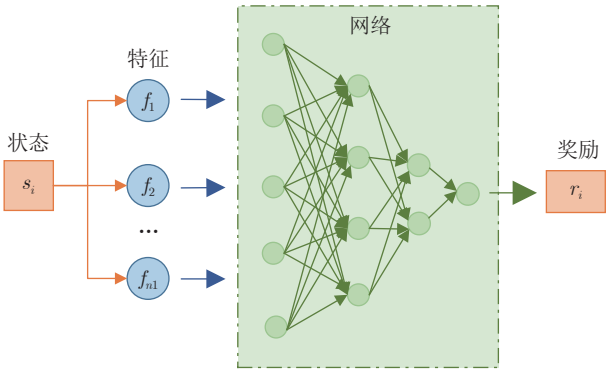


图 9　估计奖励函数的神经网络模型结构

Fig. 9　Structure of the neural network model for estimating the reward function

为了解决最大熵深度逆强化学习算法中过拟合、专家演示数据非最优等问题, 利用集成算法的思想增强逆强化学习算法的学习能力. 无模型的 IRL 方法使用一种状态编码方法来降低高维环境的计算复杂度, 然后利用 Adaboost 分类器来确定预测和演示奖励函数之间的差异[46]. 集成的最大熵深度强化学习方法将弱学习器最大熵模型合并为强化学习器, 处理不平衡的数据, 提高训练速度[67].

基于最大熵深度逆强化学习算法, 引入生成对抗网络, 考虑从小样本角度解决逆强化学习问题. 对抗逆强化学习 (Adversarial inverse reinforcement learning, AIRL) 在学习框架中通过增加语义奖励来提高 AIRL 的稳定性[68]. 端到端可微的基于模型的对抗逆强化学习 (Model-based adversarial inverse reinforcement learning, MAIRL) 采用自注意动态模型使计算图端到端可微, 降低优化的策略的方差[69]. 无模型积分 IRL 算法通过重建未知专家代价函数, 来解决非线性学习者和专家系统的对抗学徒游戏[70].

最大熵深度逆强化学习算法是目前最流行的算法, 可以在很大程度上解决逆强化学习算法中存在的问题. 但在实际的实验操作环境中, 依然需要加强其理论的研究. 考虑到图神经网络、Transformer 等网络的优越性, 将其与逆强化学习结合, 体现强大的端到端的映射能力, 这是值得研究的课题.

2.2.3　高斯过程逆强化学习算法

基于高斯过程的逆强化学习为在非线性环境中学习奖励提供了新的解决方案, 该方法可以从次优随机演示中捕获复杂行为, 同时自动平衡所学习到的奖励结构的简单性与观察到的行为的一致性. Levine 等[19] 提出基于高斯过程的逆强化学习算法, 将奖励表示为特征值的非线性函数, 旨在求解高速路自动驾驶、Object world 等问题中的非线性奖励问题. 该函数是一个高斯过程 (Gaussian processes, GP), 其结构由核函数决定. 贝叶斯 GP 框架为学习高斯核的超参数, 从而为学习未知奖励结构提供了一种原则性的方法. GPIRL 直接学习真实输出 μ, 它表示与特征坐标 X_μ 相关联的奖励. 坐标是所有状态的特征值, 是所有状态的子集. 未包含在该子集中的奖励由 GP 推断. 为了恢复奖励结构, 还需要学习高斯核的超参数 θ. 通过最大化专家演示的概率求得 μ 和 θ 的值, 将分配到 μ 和 θ 的先验概率 $P(\mu, \theta|X_\mu)$ 的对数函数作为高斯对数边缘似然. 利用 GP 获得似然函数的最优形式, 预测奖励, 继而求得整个状态空间的最优策略.

在连续的状态和行为空间中, Jin 等[71] 利用高斯过程模型来计算值函数和奖励函数, 提供了形式灵活的奖励功能和奖励的不确定性, 平衡利用和探索. Li 等[72] 提出了一种基于互信息和极限学习机的高斯过程逆强化学习算法, 利用自动相关性得到各个特征的重要性, 构造非线性奖励函数, 提高了原始高斯逆强化学习算法的性能. Michini 等[73] 提出了一个贝叶斯非参数奖励学习框架, 在单个未分段演示中推断多个子目标和奖励函数, 求解利用高斯过程奖励表示的连续域的奖励问题. 四旋翼和遥控

汽车的实验表明该算法可以从演示中学习具有挑战性的机动能力.

基于高斯过程的逆强化学习是一种非线性逆强化学习的概率算法. 此外, 在算法的理论基础上, 自动选择或优化诱导点的技术可以合并到 GPIRL 中, 以学习奖励函数. 在 GPIRL 中, 利用不同的内核可以学习不同类型的奖励结构, 进一步研究对 IRL 有用的内核函数类型是未来工作的一个好方向.

逆强化学习算法针对需要解决的问题, 建立有效的模型, 学习奖励和优化策略, 其研究历程如表 1 所示.

3 逆强化学习的应用进展

逆强化学习在智能驾驶与停车场导航、智能机器人控制、无人机、目标检测、游戏、金融贸易与工业过程等应用问题中具有重要的应用价值, 如何将逆强化学习广泛应用到这些领域, 继续发挥逆强化学习的决策性能, 是研究人员的重要研究方向.

3.1 智能驾驶与停车场导航

近年来, 随着传感器技术的飞速发展, 智能驾驶技术得到空前发展. 国内外专家、学者对基于强化学习的智能驾驶进行了研究并取得了一定的成果. 在智能驾驶应用中, 强化学习可以根据起点和终点的路程长短给与奖励, 但对遇到“撞到”、“绕开交通拥堵路段”等情况很难给出一个合适的奖励, 来指导智能体决策, 影响了强化学习在智能驾驶中的应用. 然而驾驶员可以很好地应对突发情况. 因此, 逆强化学习可根据专家演示建立线性或者非线性模型, 旨在从驾驶员的行为中推导出指导智能体收敛到驾驶员开车策略的奖励, 从而获得相应的行驶策略, 提高算法在智能驾驶系统的应用准确性[74]. 基于逆强化学习的智能驾驶将智能汽车和环境之间的交互建模为 MDP. 基于 MDP 获取专家演示, 然后利用深度神经网络逆强化学习近似专家驾驶员的未知奖励, 优化驾驶策略[38].

针对智能驾驶对大流量交通情况适应性差的问题, 基于路径积分逆强化学习的深度学习方法利用一组采样驾驶策略的特征来预测奖励函数[75]. 在没有先验领域知识的情况下, 依赖人类驾驶演示来自动调整奖励函数的方法可以优化基于最大熵逆强化学习的规划者的驾驶行为, 学习超过人类专家水平的奖励函数[76]; 也可以利用 GPS 先收集数据, 然后将学习驾驶员偏好作为道路网络的特征函数 (如限速、转弯类型), 选择和预测驾驶员最可能选择的路径[13].

当自动驾驶汽车与其他车辆、行人和骑自行车的人近距离导航时, 准确的行为预测是十分重要的, Fernando 等[77] 重点讨论了基于逆强化学习的深度行为建模在克服现有技术局限性方面的潜力, 精确

表 1 逆强化学习算法的研究历程

Table 1 Timeline of inverse reinforcement learning algorithm

逆强化学习算法	面临的挑战	解决的问题	作者 (年份)
基于边际的逆强化学习	模糊歧义	有限和大状态空间的 MDP/R 问题	Ng 等[9] (2000)
		线性求解 MDP/R 问题	Abbeel 等[11] (2004)
		策略的最大化结构与预测问题	Ratliff 等[12] (2006)
		复杂多维任务问题	Bogdanovic 等[22] (2015)
		现实任务的适用性问题	Hester 等[23] (2018)
基于贝叶斯的逆强化学习	先验知识的选取难、计算复杂	结合先验知识和专家数据推导奖励的概率分布问题	Ramachandran 等[21] (2007)
基于概率的逆强化学习	在复杂动态环境中适应性差	最大熵约束下的特征匹配问题	Ziebart 等[13] (2008)
		转移函数未知的 MDP/R 问题	Boularias 等[14] (2011)
基于高斯过程的逆强化学习	计算复杂	奖励的非线性求解问题	Levine[19] 等 (2011)
基于最大熵的深度逆强化学习	计算复杂、过拟合、专家演示数据不平衡、有限	从人类驾驶演示中学习复杂城市环境中奖励的问题	Wulfmeier 等[15] (2016)
		从数据中提取策略的对抗性逆强化学习问题	Ho 等[18] (2016)
		多个奖励稀疏分散的线性可解非确定性 MDP/R 问题	Budhraja 等[59] (2017)
		自动驾驶车辆在交通中的规划问题	You 等[38] (2019)
		无模型积分逆 RL 的奖励问题	Lian 等[70] (2021)
		利用最大因果熵推断奖励函数的问题	Gleave 等[94] (2022)
基于神经网络的逆强化学习	过拟合、不稳定	具有大规模高维状态空间的自动导航的 IRL 问题	Chen 等[62] (2019)

行为建模为自动驾驶提供重要性指导, 如图 10 所示.

针对基于深度逆强化学习的建模范式在跨长时间范围预测人类行为时未考虑环境中有多个移动行人情况下的路径规划问题, Fernando 等[78] 通过长–短期记忆网络来捕捉行人的运动, 利用基于最大熵的非线性逆强化学习框架将这些特征映射到奖励, 如图 11 所示. 利用斯坦福无人机和 SAIVT 多光谱轨迹数据集说明该算法的预测性能较好. 为提高逆强化学习的收敛速度, Kalweit 等[79] 提出了逆动作–价值迭代, 以解析的方式完全恢复外部代理的底层报酬, 以在开源模拟器 SUMO 的自主换道学习任务中实现自主驾驶.

此外, 在野外复杂环境中, 基于深度最大熵逆强化学习的越野可穿越性分析与轨迹规划方法将运动学编码为卷积核, 在融合车辆运动学的同时解决状态空间复杂度指数增长的问题, 如图 12 所示[80].

智能停车导航系统也是智能驾驶的一部分. 基于学徒学习的智能停车系统通过迭代学习恢复奖励, 找到接近于专家行为的策略, 完成停车导航[11]. 为了消除自动泊车过程中航迹推算和路径跟踪带来的误差, 最大熵逆强化学习算法可以用来学习使专家轨迹概率最大化的奖励, 学习更符合人类驾驶习惯的驾驶策略. 车辆实验结果表明, 该方法提高了学习效率、稳定性[81]. 此外, Pan 等[82] 创新性地将停车任务划分为多个离散停车的路径规划子任务, 然后利用人机交互逆强化学习框架求解子任务, 实现最终目标. 停车场路径规划的实验证明基于子目标的交互式结构的学习任务可以显著提高学习效率.

针对大状态空间的自主导航存在的所有状态无法全部遍历的问题, 神经逆强化学习算法将专家的行为推广到状态空间的未访问区域, 从而研究具有大尺度和高维状态空间的专家策略. 模拟自主导航

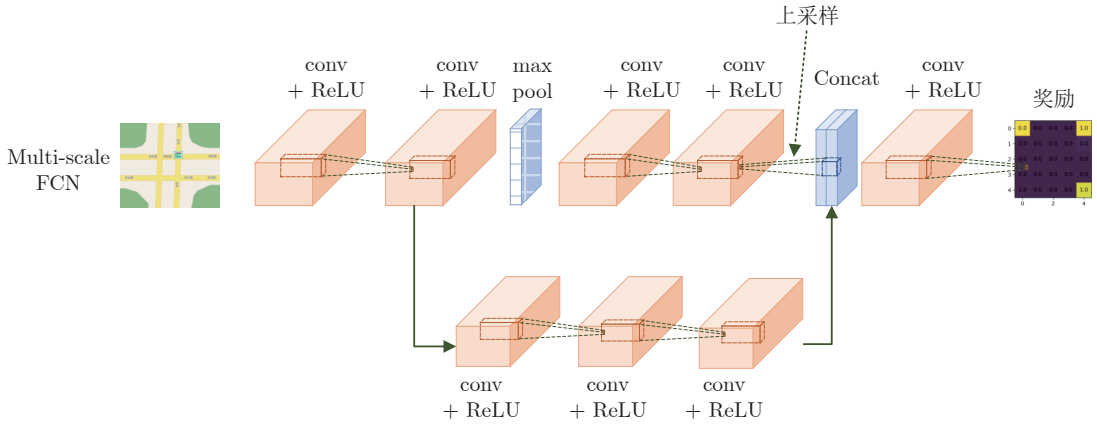
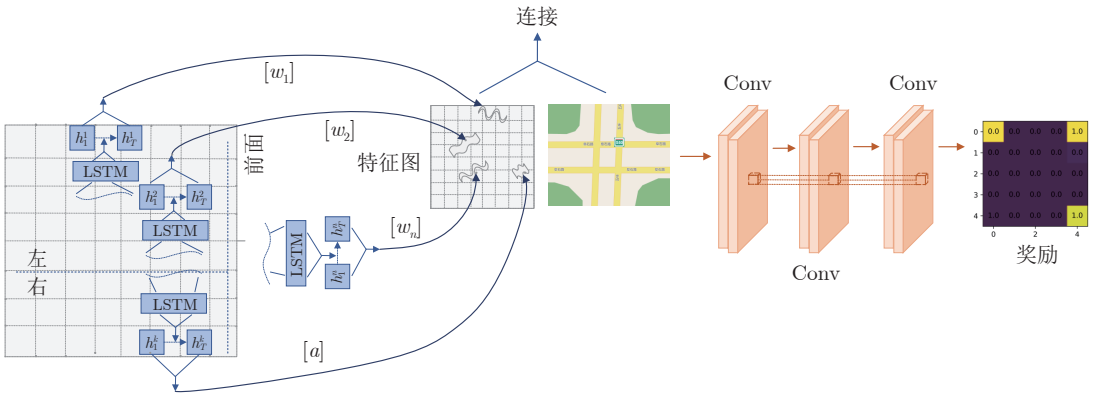


图 10　多尺度全卷积网络架构

Fig. 10　Multi-scale fully convolutional network architecture



(a) 用于嵌入邻域上下文的方法
(a) Methods for embedding neighborhood contexts

(b) 四层全卷积网络的架构图
(b) Architecture diagram of four-layer full convolutional network

图 11　非线性逆强化学习框架

Fig. 11　Framework of nonlinear inverse reinforcement learning

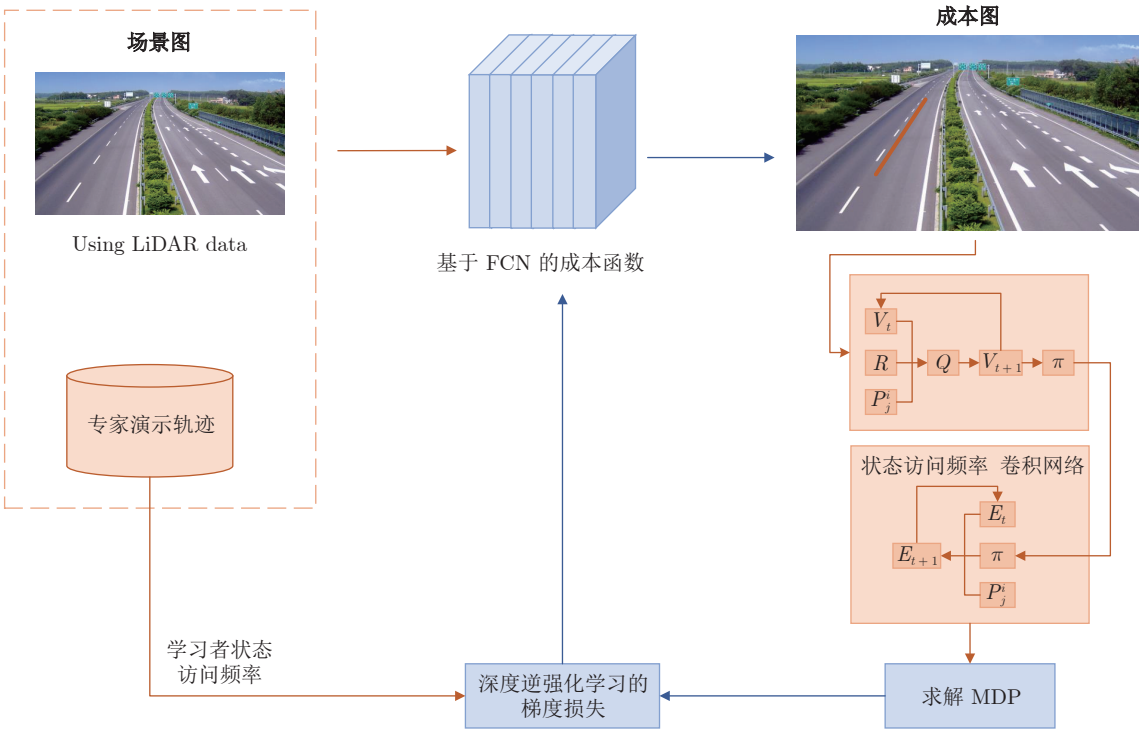


图 12 利用深度最大熵逆强化学习轨迹规划结构图

Fig. 12 Structure of trajectory planning using deep maximum entropy IRL

任务的实验结果表明, 该方法提高了学习效率[58].

自动驾驶代表了未来智能交通系统的主要趋势, 有望改善交通安全, 同时提高燃油效率, 减少拥堵. 深度逆强化学习是智能驾驶系统的主要的逆强化学习算法, 该算法结合了深度学习的感知能力和强化学习的决策能力, 同时克服了奖励设计难的问题. 智能驾驶车辆实验结果表明, 与传统的强化学习、神经网络方法相比, 该方法学习速度更快, 学习专家轨迹效果更好. 即使在越野环境下或大规模的状态空间的自主导航任务中, 基于逆强化学习的车辆能够在不与未知障碍物发生碰撞的情况下成功导航到目标位置, 减少了学习时间, 具有良好的泛化性能. 但所用的专家演示有可能存在有限、非最优的问题, 这是智能驾驶需要解决的重要问题.

3.2 智能机器人控制

智能机器人控制是人工智能的标志性成果之一. 机器人可以代替人类在危险、高温等环境中工作. 传统的机器人通过复杂的编程完成固定的工作, 对动态环境的适应性差, 因此利用逆强化学习算法控制机器人是重要的应用和研究方向.

研究者提出相对熵逆强化学习算法来模拟学习乒乓球机器人的动作, 通过大规模训练, 选择由外部刺激所触发的适合于任务的动作[83]. 也有研究者将基于贝叶斯策略的逆强化学习算法应用于教育移动机器人 E-puck, 该机器人配备了多个传感器, 通过训练能够学习导航方式[50]. 人类和移动机器人将越来越多地在相同的环境中共存. 近年来, 随着机器人技术的发展, 应用高速度、高精度、高负载自重比的机器人结构受到工业和航空航天领域的关注. 作为机器人重要结构的机械臂是一个非常复杂的动力学系统, 其动力学方程具有非线性、强耦合、实变等特点. 深度神经网络和强化学习算法的结合, 可以帮助机械臂更好地学习行为策略, 直接读取原始感官输入, 如相机图像, 有效地将估计和控制合并到一个模型中. 然而, 强化学习的现实应用必须通过手动设计的奖励函数指定任务的目标, 这在实践中需要设计端到端强化学习避免完全相同的感知, 或使用额外的传感器测量环境, 以确定任务是否已成功执行. 因此研究者提出深度逆强化学习算法, 利用从专家演示中学习到的奖励函数进行决策, 控制机械臂完成任务[33]. 文献 [84] 提出了基于卷积神经网络的逆强化学习方法, 如图 13 所示. 机械臂直接从少量成功案例的图像中学习, 主动请求查询, 确认任务是否完成, 无需手动设计奖励.

3.3 无人机

无人机具有灵活、生存概率高等特点, 使得无人机成为未来性价比最高的军事武器和民用设备.

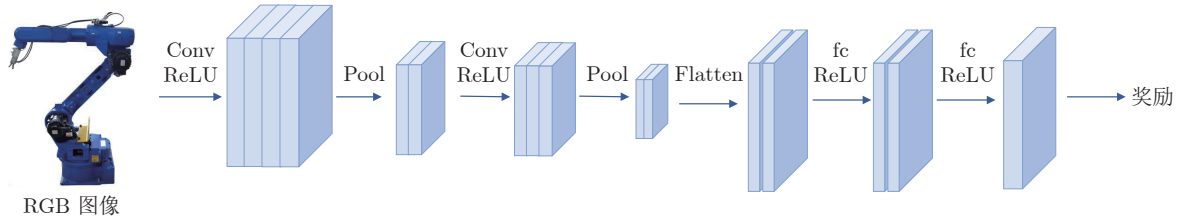


图 13　机械臂的卷积神经网络结构

Fig. 13　Convolutional neural network structure for robotic arm

因此无人机领域具有广泛的研究前景, 研究者们围绕着如何将智能算法应用于无人机进行深入研究.

低空飞行环境中拦截敌方无人机是一项具有挑战性的任务[85]. 为解决无人机的自主拦截问题, 基于生成对抗模仿学习的方法通过扩充专家数据, 在生成对抗网络中引入语义条件来提高学习效率. 为解决多旋翼无人机飞行问题, 逆强化学习方法通过最小化轨迹跟踪误差来学习最优奖励和跟踪轨迹[86]. 为了解决在无人类专家的复杂任务中获得演示难的问题, 深度 Q 学习的逆强化学习通过在基本子任务上执行演示来学习一个复杂的任务[87].

与传统控制的无人机相比, 基于逆强化学习的无人机系统可以通过生成对抗网络扩充专家数据, 提高学习效率和环境适应性, 解决专家演示有限的问题. 但无人机运行环境复杂, 如何利用逆强化学习算法弥补无人机的通信系统、定位导航等存在的缺陷, 是帮助无人机获得更好的飞行能力而需要研究的方向.

3.4　目标检测

由于物体具有不同的外观和姿态, 加上成像时光照、遮挡等因素的干扰, 目标检测一直是计算机视觉领域最具有挑战性的问题. 目前, 深度学习等机器学习算法已经在目标检测领域得到了广泛的应用. 针对视频中速度较慢、运动较弱的目标检测问题, 由于深度逆强化学习算法具有强大的感知和决策能力, 该算法的提出为目标检测提供了更有效的方法. 深度逆强化学习算法集成了深度学习在视觉等感知问题上强大的理解能力, 以及强化学习的决策能力, 不仅实现了从专家演示中学习奖励函数, 而且实现了端到端学习.

Sun 等[88] 提出了一种深度逆强化学习方法, 用于航空视频中慢弱运动目标的检测, 学习奖励和专家策略. 实验结果表明该方法在航空视频运动目标检测精度方面具有明显优势. Pattanayak 等[89] 提出基于元认知雷达的反对抗步骤的逆强化学习, 将雷达的元认知问题抽象为状态谱和观测噪声协方差矩阵, 通过观察雷达响应波形的噪声序列来检测雷达是否是效用最大化. Kormushev 等[90] 提出了一个集成的方法, 让人形机器人学习射箭技能. 该方法基于高斯混合模型对目标和箭头尖端进行颜色检测, 使用逆运动学控制器控制手臂的运动.

与传统的运动目标检测算法相比, 逆强化学习算法可以使目标检测得的结果更加准确. 但获取专家演示的图像可能使得算法的复杂度增大.

3.5　游戏

RL 在游戏领域中的应用备受关注, 且极为成功, 最典型的为 AlphaGoZero. 随着逆强化学习算法理论的不断完善, 如何更好地利用逆强化学习增强游戏的智能性引起了研究者的关注. Bogdanovic 等[22] 利用学徒学习直接观察游戏区域的原始像素来学习玩雅达利游戏. Lian 等[70] 开发了基于模型和无模型的 IRL 算法来求解游戏, 通过学习专家的目标权重来解决对抗性学徒博弈, 使学习者表现出专家的行为. Koller 等[91] 提出了一种非合作博弈的图形表示——多主体影响图, 使用战略相关性将大型游戏分解为一组相互作用的小型游戏, 按顺序解决这些小型游戏. 该算法可以节省博弈中的计算成本, 提高游戏中算法的性能. Syed 等[92] 提出了基于乘权算法的学徒学习 (Apprenticeship learning based on multiplicative-weights, MWAL), 解决学徒在奖励函数未知的玩具视频游戏环境中学习行为的问题.

在专家演示下利用学徒学习逆强化学习、生成对抗逆强化学习等算法学习奖励, 进行策略优化, 克服了强化学习奖励设计难的问题, 因此其策略优化性能优于强化学习. 将逆强化学习应用于游戏表现出了较好的游戏智能性. 但生成对抗网络、学徒学习等算法本身存在的模式崩塌、模糊歧义等问题可能影响游戏的智能性.

3.6　金融贸易和工业过程

随着强化学习的发展, 该算法已经应用于金融贸易的预测以及材料、化工过程故障诊断、污水处理等工业过程. 结合神经网络等机器学习的感知能力和逆强化学习的决策能力可以提高策略优化的能

力. 逆强化学习的理论在逐步完善, 得益于专家演示下奖励的学习, 将逆强化学习应用于金融贸易和工业过程会表现出较好的算法性能. 投资者可以利用逆强化学习, 从基金经理的交易历史中学习他们的意图, 并恢复他们的隐含奖励函数, 优化资产配置决策[93].

现在基于逆强化学习算法的金融、工业过程的应用较少, 但在未来这将是一个强劲的应用方向, 可以更好地帮助金融实现预测, 帮助工业实现控制.

4 逆强化学习算法面临的问题及解决思路

逆强化学习现已应用于智能驾驶、机器人、无人机等系统中, 但是逆强化学习的理论和应用还存在一些不足. 如何完善逆强化学习的理论, 并将逆强化学习推广到更广阔的应用场景是未来的重要研究方向.

4.1 逆强化学习的模糊性问题及解决思路

逆强化学习算法的提出为解决强化学习的瓶颈问题提出了解决方案, 该算法首先利用专家演示学习奖励函数, 然后利用学习到的奖励函数学习最优策略. 研究者们提出的边际类逆强化学习算法存在奖励模糊歧义的问题, 即不同的奖励函数对应相同的策略. 针对该问题研究者们提出基于熵的逆强化学习等算法, 包括最大熵逆强化学习、最大因果熵逆强化学习、相对熵逆强化学习等算法[94−95]. 为在复杂非线性的环境中学习非线性奖励函数, 利用神经网络能拟合任意函数的性能, 提出最大熵深度逆强化学习、最大因果熵深度逆强化学习等算法. 然后将基于熵的深度逆强化学习算法进一步改进应用于自动驾驶、机器人、无人机等领域[96]. 但为更好地解决奖励函数的模糊性问题, 还需要对基于熵的逆强化学习进行进一步研究, 比如加入正则化、限制条件等[97−98]. 文献 [97] 提出的正则化 IRL 将强凸正则化应用于学习策略, 以避免专家的行为被任意常数奖励合理化; 文献 [98] 提出在有严格误差界限的统计检测器中加入约束来优化探测信号. 通过以上思想的研究改进, 提出更好的解决方案.

4.2 专家演示次优问题及解决思路

在求解 MDP 问题时, 现有的逆强化学习算法假设专家演示是最优的, 但在实际问题中, 很难获取最优的专家演示. 为了从次优的专家演示中构建最优的奖励函数, 求解接近于专家策略的最优策略, 研究者们已经不断提升逆强化学习算法的性能, 开发新的算法, 相继提出适应复杂环境的逆强化学习算法, 例如最大熵深度逆强化学习、基于贝叶斯的逆强化学习、基于高斯过程的逆强化学习算法等. 此外, 有研究者提出基于生成对抗的逆强化学习, 利用生成对抗网络的生成器和判别器不断提高专家演示和奖励函数的精度, 从而获得好的策略. 基于生成对抗网络的最大熵逆强化学习算法结合专家样本训练优化生成对抗网络, 以生成虚拟专家样本, 在此基础上利用随机策略生成非专家样本, 构建混合样本集, 结合最大熵概率模型, 对奖励函数进行建模, 并利用梯度下降方法求解最优奖励函数[99]. 在多智能体环境中, 从专家人类轨迹推断奖励函数的方法使用对抗性逆强化学习和连续潜变量学习共享奖励函数[100]. 但该问题一直未得到完美解决, 还需要进一步深入研究.

4.3 博弈问题及解决思路

多智能体逆强化学习算法已经在一定程度上得到了应用发展, 这类算法的研究考虑了多个专家演示或多个奖励组件. 文献 [101] 将轨迹簇作为自适应最大熵逆强化学习中的潜在变量, 增加了逆强化学习问题的复杂性. 文献 [102] 在逆强化学习算法中引入了可解释的奖励成分来共同学习一个线性组合的奖励, 以及生成最优策略. 然而在二人零和完全信息博弈的最优决策问题上并没有取得突破性进展, 像其他的博弈问题, 如二人非零和完全信息博弈问题、多智能体博弈问题等也是值得研究的课题[91]. 这些课题的解决将会给逆强化学习算法带来大的突破.

多智能体博弈是博弈领域的前沿思想, 多智能体逆强化学习就是一个随机博弈, 将每一个状态阶段博弈的纳什策略组合起来成为一个智能体在动态环境中的策略[103]. 通过不断与环境交互来更新每一个状态的阶段博弈中的 Q 值函数 (博弈奖励), 从而获得最优的奖励函数和最优策略. 多智能体系统是由相互联系的智能体组成的系统, 它们之间具有自主性、协调性等特点[104−105]. 将多智能体系统与逆强化学习算法结合, 可以增强奖励函数的学习精度, 同时增强多智能体逆强化学习算法的感知和决策能力.

4.4 逆强化学习理论分析不完善问题

逆强化学习表现出良好的解决无奖励函数的 MDP 问题的能力. 研究者们首先提出边际类逆强化学习、最大熵逆强化学习、贝叶斯逆强化学习等求解线性奖励函数的逆强化学习算法. 为了求解在复杂、非线性的环境中的奖励函数, 研究者们提出

边际类深度逆强化学习、最大熵深度逆强化学习等. 利用逆强化学习算法从专家演示中学习线性或者非线性的奖励函数. 虽然智能算法在智能系统表现出强大的性能, 但对其理论的研究还不够, 这极大影响了逆强化学习在实际中的应用. 实际中动态复杂多变的环境也要求逆强化学习算法学习的奖励函数具有更强的鲁棒性, 以提高决策的精度. 基于生成对抗网络的逆强化学习方法可用来学习对动态变化稳健的奖励函数, 在环境显著变化下学习策略[34]. 分层逆强化学习利用专家演示的内在动机, 在选项框架内学习最优的奖励函数[106]. 此外, 逆强化学习算法的收敛性、稳定性、鲁棒性也需要深入研究, 需要进行理论性证明. 在不久的未来, 突破上述所有的问题将会促进人工智能领域再上一个新的台阶.

5 逆强化学习算法的未来技术展望

一直以来, 强化学习算法、逆强化学习算法作为人工智能的关键技术. 在未来的 10 年中, 强化学习算法、逆强化学习算法也必将扮演很重要的角色.

在理论研究方面, 我们主要总结归纳了线性逆强化学习和非线性逆强化学习. 在线性逆强化学习的研究进展中, 我们介绍了学徒学习逆强化学习、最大边际规划逆强化学习、最大熵逆强化学习、基于相对熵的逆强化学习、贝叶斯逆强化学习等算法. 在非线性逆强化学习的研究进展中介绍了学徒学习深度逆强化学习、最大边际规划深度逆强化学习、最大熵深度逆强化学习、高斯过程的逆强化学习等算法. 随着逆强化学习算法理论的发展, 神经网络、贝叶斯、高斯过程、最大熵等与逆强化学习结合, 来提高算法的性能, 提高奖励的学习精度. 最初研究者提出了基于边际类逆强化学习算法, 但可能存在模糊歧义问题. 针对模糊歧义问题, 研究者们提出最大熵逆强化学习和基于高斯过程的逆强化学习. 针对复杂非线性环境, 提出基于深度学习的逆强化学习算法, 通过端到端的方式学习奖励. 利用生成对抗网络与逆强化学习算法结合, 解决专家演示有限、非最优的问题.

在实验 Object world 环境中, 我们设置相同的实验参数, 利用逆强化学习典型算法进行实验后, 获得各算法的奖励和值函数的值如表 2 所示. 由于该实验建模为标准的 MDP 下的非线性环境, GPIRL 的非线性可以很好地求解奖励, 算法获得的奖励和值函数最优, ALIRL、FIRL、MMP、MMPBoost、MEIRL 在该环境下的表现性能相似, MWAL 在该环境的适应性最差.

表 2 逆强化学习算法的比较

Table 2 Comparison of inverse reinforcement learning algorithms

逆强化学习算法	奖励	值函数
ALIRL[11]	38.79	32.66
FIRL[27]	31.89	5.22
GPIRL[19]	2.66	0.42
MWAL[95]	206.44	43.32
MMP[12]	38.38	34.20
MMPBoost[30]	31.56	23.56
MEIRL[13]	36.36	13.12

但由于逆强化学习算法本身的缺陷, 依然存在计算复杂、模糊歧义、专家演示有限、专家演示数据不平衡等问题, 逆强化学习应用的落地还需要进一步的技术突破, 包括但不仅限于更大规模的数据和计算资源, 将最新的图神经网络 (Graph neural network, GNN)、元学习 (Meta-learning, ML)、知识图谱 (Knowledge graph) 等思想引入到逆强化学习. 利用 GCN、图注意力网络 (Graph attention network, GAT) 和知识图谱在复杂环境中大规模的学习, 自然地融合图的属性信息进行学习, 同时 ML 可以增强逆强化学习的迁移性, 提高算法的泛化能力. 此外, 在小样本 (Small sample data) 专家演示下, 将 Transformer 网络和 GAN 等最新的思想填充到逆强化学习的思想中, 弥补算法的缺陷, 减小计算量, 增强专家演示的可靠性, 提高算法的学习能力和决策性能也是未来研究的重要方向.

在应用研究方面, 随着逆强化学习理论的发展, 由于逆强化学习拥有独特优势, 目前已经广泛应用于智能驾驶与停车场导航、机器人控制、无人机、目标检测、游戏等领域, 在金融贸易和工业过程等领域的应用较少. 在工业过程中, 我们期待逆强化学习进行大规模的应用落地, 将逆强化学习应用于股票、基金交易预测、材料基因检测、化学分子预测、污水处理控制、国防军事控制等领域. 逆强化学习还可以与专家系统相结合, 产生逆强化学习专家系统, 利用专家知识, 提升智能制造的诊断能力和决策能力. 在诊断和决策过程中, 通过反馈来修正预测网络, 提高算法的预测水平. 逆强化学习算法的可靠性和应用经济性使得算法拥有较大的研究价值.

6 结束语

作为人工智能中的前沿领域, 逆强化学习旨在解决强化学习中人工设计奖励函数难的问题, 该算法在专家演示下学习奖励函数, 从而学习最优策略. 逆强化学习算法在人工智能领域具有重要的研究意义和应用价值, 其理论研究和应用研究可以推动智

能驾驶、无人机、机器人等实际应用的发展. 本文重点介绍了逆强化学习算法的理论研究的现状以及应用前景, 总结了逆强化学习算法中存在的问题, 并提供了可行的解决思路.

在今后的研究工作中, 首先要加强对逆强化学习理论方面的研究, 使逆强化学习算法理论更加成熟. 此外, 要将逆强化学习算法拓宽到更广的应用领域中, 提高我国人工智能算法的发展在国际上的影响.

References

1 Chai Tian-You. Development directions of industrial artificial intelligence. *Acta Automatica Sinica*, 2020, **46**(10): 2005−2012
(柴天佑. 工业人工智能发展方向. 自动化学报, 2020, **46**(10): 2005−2012)

2 Dai X Y, Zhao C, Li X S, Wang X, Wang F Y. Traffic signal control using offline reinforcement learning. In: Proceedings of the China Automation Congress (CAC). Beijing, China: IEEE, 2021. 8090−8095

3 Li J N, Ding J L, Chai T Y, Lewis F L. Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes. *IEEE Transactions on Cybernetics*, 2020, **50**(9): 4132−4145

4 Zhao Dong-Bin, Shao Kun, Zhu Yuan-Heng, Li Dong, Chen Ya-Ran, Wang Hai-Tao, et al. Review of deep reinforcement learning and discussions on the development of computer Go. *Control Theory & Applications*, 2016, **33**(6): 701−717
(赵冬斌, 邵坤, 朱圆恒, 李栋, 陈亚冉, 王海涛, 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)

5 Song T H, Li D Z, Yang W M, Hirasawa K. Recursive least-squares temporal difference with gradient correction. *IEEE Transactions on Cybernetics*, 2021, **51**(8): 4251−4264

6 Bain M, Sammut C. A framework for behavioural cloning. *Machine Intelligence 15: Intelligent Agents*. Oxford: Oxford University, 1995. 103−129

7 Couto G C K, Antonelo E A. Generative adversarial imitation learning for end-to-end autonomous driving on urban environments. In: Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI). Orlando, USA: IEEE, 2021. 1−7

8 Samak T V, Samak C V, Kandhasamy S. Robust behavioral cloning for autonomous vehicles using end-to-end imitation learning. *SAE International Journal of Connected and Automated Vehicles*, 2021, **4**(3): 279−295

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

10 Imani M, Ghoreishi S F. Scalable inverse reinforcement learning through multifidelity Bayesian optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(8): 4125−4132

11 Abbeel P, Ng A Y. Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the 21st International Conference on Machine Learning (ICML). Ban, Canada: ACM, 2004. 1−8

12 Ratliff N D, Bagnell J A, Zinkevich M A. Maximum margin planning. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). Pittsburgh, USA: ACM, 2006. 729−736

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

14 Boularias A, Kober J, Peters J. Relative entropy inverse reinforcement learning. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, USA: 2011. 182−189

15 Wulfmeier M, Wang D Z, Posner I. Watch this: Scalable cost-function learning for path planning in urban environments. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, Korea (South): IEEE, 2016. 2089−2095

16 Guo H Y, Chen Q X, Xia Q, Kang C Q. Deep inverse reinforcement learning for objective function identification in bidding models. *IEEE Transactions on Power Systems*, 2021, **36**(6): 5684−5696

17 Shi Y C, Jiu B, Yan J K, Liu H W, Li K. Data-driven simultaneous multibeam power allocation: When multiple targets tracking meets deep reinforcement learning. *IEEE Systems Journal*, 2021, **15**(1): 1264−1274

18 Ho J, Ermon S. Generative adversarial imitation learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS). Barcelona, Spain: Curran Associates Inc., 2016. 4572−4580

19 Levine S, Popović Z, Koltun V. Nonlinear inverse reinforcement learning with Gaussian processes. In: Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS). Granada, Spain: Curran Associates Inc., 2011. 19−27

20 Liu J H, Huang Z H, Xu X, Zhang X L, Sun S L, Li D Z. Multi-kernel online reinforcement learning for path tracking control of intelligent vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, **51**(11): 6962−6975

21 Ramachandran D, Amir E. Bayesian inverse reinforcement learning. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI). Hyderabad, India: Morgan Kaufmann, 2007. 2586−2591

22 Bogdanovic M, Markovikj D, Denil M, de Freitas N. Deep apprenticeship learning for playing video games. In: Proceedings of the AAAI Workshop on Learning for General Competency in Video Games. Austin, USA: AAAI, 2015. 7−9

23 Hester T, Vecerik M, Pietquin O, Lanctot M, Schaul T, Piot B, et al. Deep Q-learning from demonstrations. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI, 2018. Article No. 394

24 Nguyen H T, Garratt M, Bui L T, Abbass H. Apprenticeship learning for continuous state spaces and actions in a swarm-guidance shepherding task. In: Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI). Xiamen, China: IEEE, 2019. 102−109

25 Hwang M, Jiang W C, Chen Y J. A critical state identification approach to inverse reinforcement learning for autonomous systems. *International Journal of Machine Learning and Cybernetics*, 2022, **13**(4): 1409−1423

26 Jin Zhuo-Jun, Qian Hui, Chen Shen-Yi, Zhu Miao-Liang. Survey of apprenticeship learning based on reward function approximating. *Journal of Huazhong University of Science & Technology (Nature Science Edition)*, 2008, **36**(S1): 288−290
(金卓军, 钱徽, 陈沈轶, 朱淼良. 基于回报函数逼近的学徒学习综述. 华中科技大学学报 (自然科学版), 2008, **36**(S1): 288−290)

27 Levine S, Popović Z, Koltun V. Feature construction for inverse reinforcement learning. In: Proceedings of the 23th International Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: Curran Associates Inc., 2010. 1342−1350

28 Pan W, Qu R P, Hwang K S, Lin H S. An ensemble fuzzy ap-

proach for inverse reinforcement learning. *International Journal of Fuzzy Systems*, 2019, **21**(1): 95−103

29 Lin J L, Hwang K S, Shi H B, Pan W. An ensemble method for inverse reinforcement learning. *Information Sciences*, 2020, **512**: 518−532

30 Ratliff N, Bradley D, Bagnell J A, Chestnutt J. Boosting structured prediction for imitation learning. In: Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: MIT Press, 2006. 1153−1160

31 Choi D, An T H, Ahn K, Choi J. Future trajectory prediction via RNN and maximum margin inverse reinforcement learning. In: Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, USA: IEEE, 2018. 125−130

32 Gao Zhen-Hai, Yan Xiang-Tong, Gao Fei. A decision-making method for longitudinal autonomous driving based on inverse reinforcement learning. *Automotive Engineering*, 2022, **44**(7): 969−975
(高振海, 闫相同, 高菲. 基于逆向强化学习的纵向自动驾驶决策方法. 汽车工程, 2022, **44**(7): 969−975)

33 Finn C, Levine S, Abbeel P. Guided cost learning: Deep inverse optimal control via policy optimization. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML). New York, USA: JMLR., 2016. 49−58

34 Fu J, Luo K, Levine S. Learning robust rewards with adversarial inverse reinforcement learning. In: Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada: Elsevier, 2018. 1−15

35 Huang D A, Kitani K M. Action-reaction: Forecasting the dynamics of human interaction. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 489−504

36 Levine S, Koltun V. Continuous inverse optimal control with locally optimal examples. In: Proceedings of the 29th International Conference on Machine Learning (ICML). Edinburgh, Scotland: Omnipress, 2012. 475−482

37 Aghasadeghi N, Bretl T. Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). San Francisco, USA: IEEE, 2011. 1561−1566

38 You C X, Lu J B, Filev D, Tsiotras P. Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. *Robotics and Autonomous Systems*, 2019, **114**: 1−18

39 Das N, Chattopadhyay A. Inverse reinforcement learning with constraint recovery. arXiv preprint arXiv: 2305.08130, 2023.

40 Krishnan S, Garg A, Liaw R, Miller L, Pokorny F T, Goldberg K. HIRL: Hierarchical inverse reinforcement learning for long-horizon tasks with delayed rewards. arXiv: 1604.06508, 2016.

41 Zhou Z Y, Bloem M, Bambos N. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 2018, **63**(9): 2787−2802

42 Wu Z, Sun L T, Zhan W, Yang C Y, Tomizuka M. Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving. *IEEE Robotics and Automation Letters*, 2020, **5**(4): 5355−5362

43 Huang Z Y, Wu J D, Lv C. Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(8): 10239−10251

44 Song L, Li D Z, Xu X. Sparse online maximum entropy inverse reinforcement learning via proximal optimization and truncated gradient. *Knowledge-Based Systems*, 2022, **252**: Article No. 109443

45 Gleave A, Habryka O. Multi-task maximum causal entropy inverse reinforcement learning. arXiv: 1805.08882, 2018.

46 Zhang T, Liu Y, Hwang M, Hwang K S, Ma C Y, Cheng J. An end-to-end inverse reinforcement learning by a boosting approach with relative entropy. *Information Sciences*, 2020, **520**: 1−14

47 Wu Shao-Bo, Fu Qi-Ming, Chen Jian-Ping, Wu Hong-Jie, Lu You. Meta-inverse reinforcement learning method based on relative entropy. *Computer Science*, 2021, **48**(9): 257−263
(吴少波, 傅启明, 陈建平, 吴宏杰, 陆悠. 基于相对熵的元逆强化学习方法. 计算机科学, 2021, **48**(9): 257−263)

48 Lin B Y, Cook D J. Analyzing sensor-based individual and population behavior patterns via inverse reinforcement learning. *Sensors*, 2020, **20**(18): Article No. 5207

49 Zhou W C, Li W C. A hierarchical Bayesian approach to inverse reinforcement learning with symbolic reward machines. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 27159−27178

50 Ezzeddine A, Mourad N, Araabi B N, Ahmadabadi M N. Combination of learning from non-optimal demonstrations and feedbacks using inverse reinforcement learning and Bayesian policy improvement. *Expert Systems With Applications*, 2018, **112**: 331−341

51 Ranchod P, Rosman B, Konidaris G. Nonparametric Bayesian reward segmentation for skill discovery using inverse reinforcement learning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany: IEEE, 2015. 471−477

52 Choi J, Kim K E. Nonparametric Bayesian inverse reinforcement learning for multiple reward functions. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc., 2012. 305−313

53 Okal B, Arras K O. Learning socially normative robot navigation behaviors with Bayesian inverse reinforcement learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden: IEEE, 2016. 2889−2895

54 Trinh T, Brown D S. Autonomous assessment of demonstration sufficiency via bayesian inverse reinforcement learning. arXiv: 2211.15542, 2022.

55 Huang W H, Braghin F, Wang Z. Learning to drive via apprenticeship learning and deep reinforcement learning. In: Proceedings of the IEEE 31st International Conference on Tools With Artificial Intelligence (ICTAI). Portland, USA: IEEE, 2019. 1536−1540

56 Markovikj D. Deep apprenticeship learning for playing games. arXiv preprint arXiv: 2205.07959, 2022.

57 Lee D, Srinivasan S, Doshi-Velez F. Truly batch apprenticeship learning with deep successor features. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). Macao, China: Morgan Kaufmann, 2019. 5909−5915

58 Xia C, El Kamel A. Neural inverse reinforcement learning in autonomous navigation. *Robotics and Autonomous Systems*, 2016, **84**: 1−14

59 Budhraja K K, Oates T. Neuroevolution-based inverse reinforcement learning. In: Proceedings of the IEEE Congress on Evolutionary Computation (CEC). Donostia, Spain: IEEE, 2017. 67−76

60 Memarian F, Xu Z, Wu B, Wen M, Topcu U. Active task-inference-guided deep inverse reinforcement learning. In: Proceedings of the 59th IEEE Conference on Decision and Control (CDC). Jeju, Korea (South): IEEE, 2020. 1932−1938

61 Liu S, Jiang H, Chen S P, Ye J, He R Q, Sun Z Z. Integrating Dijkstra′s algorithm into deep inverse reinforcement learning for food delivery route planning. *Transportation Research Part*

E: Logistics and Transportation Review, 2020, **142**: Article No. 102070

62 Chen X L, Cao L, Xu Z X, Lai J, Li C X. A study of continuous maximum entropy deep inverse reinforcement learning. *Mathematical Problems in Engineering*, 2019, **2019**: Article No. 4834516

63 Choi D, Min K, Choi J. Regularising neural networks for future trajectory prediction via inverse reinforcement learning framework. *IET Computer Vision*, 2020, **14**(5): 192–200

64 Wang Y, Wan S, Li Q, Niu Y, Ma F. Modeling crossing behaviors of E-Bikes at intersection with deep maximum entropy inverse reinforcement learning using drone-based video data. *IEEE Transactions on Intelligent Transportation Systems*, 2023, **24**(6): 6350–6361

65 Fahad M, Chen Z, Guo Y. Learning how pedestrians navigate: A deep inverse reinforcement learning approach. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018. 819–826

66 Zhou Y, Fu R, Wang C. Learning the car-following behavior of drivers using maximum entropy deep inverse reinforcement learning. *Journal of Advanced Transportation*, 2020, **2020**: Article No. 4752651

67 Song L, Li D Z, Wang X, Xu X. AdaBoost maximum entropy deep inverse reinforcement learning with truncated gradient. *Information Sciences*, 2022, **602**: 328–350

68 Wang P, Liu D P, Chen J Y, Li H H, Chan C Y. Decision making for autonomous driving via augmented adversarial inverse reinforcement learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Xi′an, China: IEEE, 2021. 1036–1042

69 Sun J K, Yu L T, Dong P Q, Lu B, Zhou B L. Adversarial inverse reinforcement learning with self-attention dynamics model. *IEEE Robotics and Automation Letters*, 2021, **6**(2): 1880–1886

70 Lian B S, Xue W Q, Lewis F L, Chai T Y. Inverse reinforcement learning for adversarial apprentice games. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2021.3114612

71 Jin Z J, Qian H, Zhu M L. Gaussian processes in inverse reinforcement learning. In: Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC). Qingdao, China: IEEE, 2010. 225–230

72 Li D C, He Y Q, Fu F. Nonlinear inverse reinforcement learning with mutual information and Gaussian process. In: Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO). Bali, Indonesia: IEEE, 2014. 1445–1450

73 Michini B, Walsh T J, Agha-Mohammadi A A, How J P. Bayesian nonparametric reward learning from demonstration. *IEEE Transactions on Robotics*, 2015, **31**(2): 369–386

74 Sun L T, Zhan W, Tomizuka M. Probabilistic prediction of interactive driving behavior via hierarchical inverse reinforcement learning. In: Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, USA: IEEE, 2018. 2111–2117

75 Rosbach S, Li X, Großjohann S, Homoceanu S, Roth S. Planning on the fast lane: Learning to interact using attention mechanisms in path integral inverse reinforcement learning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, USA: IEEE, 2020. 5187–5193

76 Rosbach S, James V, Großjohann S, Homoceanu S, Roth S. Driving with style: Inverse reinforcement learning in general-purpose planning for automated driving. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macao, China: IEEE, 2019. 2658–2665

77 Fernando T, Denman S, Sridharan S, Fookes C. Deep inverse reinforcement learning for behavior prediction in autonomous driving: Accurate forecasts of vehicle motion. *IEEE Signal Processing Magazine*, 2021, **38**(1): 87–96

78 Fernando T, Denman S, Sridharan S, Fookes C. Neighbourhood context embeddings in deep inverse reinforcement learning for predicting pedestrian motion over long time horizons. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE, 2019. 1179–1187

79 Kalweit G, Huegle M, Werling M, Boedecker J. Deep inverse Q-learning with constraints. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1198

80 Zhu Z Y, Li N, Sun R Y, Xu D H, Zhao H J. Off-road autonomous vehicles traversability analysis and trajectory planning based on deep inverse reinforcement learning. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Las Vegas, USA: IEEE, 2020. 971–977

81 Fang P Y, Yu Z P, Xiong L, Fu Z Q, Li Z R, Zeng D Q. A maximum entropy inverse reinforcement learning algorithm for automatic parking. In: Proceedings of the 5th CAA International Conference on Vehicular Control and Intelligence (CVCI). Tianjin, China: IEEE, 2021. 1–6

82 Pan X, Ohn-Bar E, Rhinehart N, Xu Y, Shen Y L, Kitani K M. Human-interactive subgoal supervision for efficient inverse reinforcement learning. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, 2018.

83 Peters J, Mülling K, Altün Y, Peters J, Mulling K, Altun Y. Relative entropy policy search. In: Proceedings of 24th AAAI Conference on Artificial Intelligence (AAAI). Atlanta, Georgia: AAAI, 2010. 1607–1612

84 Singh A, Yang L, Hartikainen K, Finn C, Levine S. End-to-end robotic reinforcement learning without reward engineering. In: Proceedings of the Robotics: Science and Systems. Freiburg im Breisgau, Germany: the MIT Press, 2019.

85 Wang H, Liu X F, Zhou X. Autonomous UAV interception via augmented adversarial inverse reinforcement learning. In: Proceedings of the International Conference on Autonomous Unmanned Systems (ICAUS). Changsha, China: Springer, 2022. 2073–2084

86 Choi S, Kim S, Kim H J. Inverse reinforcement learning control for trajectory tracking of a multirotor UAV. *International Journal of Control, Automation and Systems*, 2017, **15**(4): 1826–1834

87 Nguyen H T, Garratt M, Bui L T, Abbass H. Apprenticeship bootstrapping: Inverse reinforcement learning in a multi-skill UAV-UGV coordination task. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, 2018. 2204–2206

88 Sun W, Yan D S, Huang J, Sun C H. Small-scale moving target detection in aerial image by deep inverse reinforcement learning. *Soft Computing*, 2020, **24**(8): 5897–5908

89 Pattanayak K, Krishnamurthy V, Berry C. Meta-cognition. An inverse-inverse reinforcement learning approach for cognitive radars. In: Proceedings of the 25th International Conference on Information Fusion (FUSION). Linköping, Sweden: IEEE, 2022. 1–8

90 Kormushev P, Calinon S, Saegusa R, Metta G. Learning the skill of archery by a humanoid robot iCub. In: Proceedings of the 10th IEEE-RAS International Conference on Humanoid Robots. Nashville, USA: IEEE, 2010. 417–423

91 Koller D, Milch B. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*,

2003, **45**(1): 181−221

92 Syed U, Schapire R E. A game-theoretic approach to apprenticeship learning. In: Proceedings of the 22nd Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: Curran Associates Inc., 2007. 1449−1456

93 Halperin I, Liu J Y, Zhang X. Combining reinforcement learning and inverse reinforcement learning for asset allocation recommendations. arXiv: 2201.01874, 2022.

94 Gleave A, Toyer S. A primer on maximum causal entropy inverse reinforcement learning. arXiv: 2203.11409, 2022.

95 Adams S, Cody T, Beling P A. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 2022, **55**(6): 4307−4346

96 Li X J, Liu H S, Dong M H. A general framework of motion planning for redundant robot manipulator based on deep reinforcement learning. *IEEE Transactions on Industrial Informatics*, 2022, **18**(8): 5253−5263

97 Jeon W, Su C Y, Barde P, Doan T, Nowrouzezahrai D, Pineau J. Regularized inverse reinforcement learning. In: Proceedings of the 9th International Conference on Learning Representations (ICLR). Vienna, Austria: Ithaca, 2021. 1−26

98 Krishnamurthy V, Angley D, Evans R, Moran B. Identifying cognitive radars-inverse reinforcement learning using revealed preferences. *IEEE Transactions on Signal Processing*, 2020, **68**: 4529−4542

99 Chen Jian-Ping, Chen Qi-Qiang, Fu Qi-Ming, Gao Zhen, Wu Hong-Jie, Lu You. Maximum entropy inverse reinforcement learning based on generative adversarial networks. *Computer Engineering and Applications*, 2019, **55**(22): 119−126
(陈建平, 陈其强, 傅启明, 高振, 吴宏杰, 陆悠. 基于生成对抗网络的最大熵逆强化学习. 计算机工程与应用, 2019, **55**(22): 119−126)

100 Gruver N, Song J M, Kochenderfer M J, Ermon S. Multi-agent adversarial inverse reinforcement learning with latent variables. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS). Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2020. 1855−1857

101 Giwa B H, Lee C G. A marginal log-likelihood approach for the estimation of discount factors of multiple experts in inverse reinforcement learning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021. 7786−7791

102 Ghosh S, Srivastava S. Mapping language to programs using multiple reward components with inverse reinforcement learning. In: Proceedings of the Findings of the Association for Computational Linguistics. Punta Cana, Dominican Republic: ACL, 2021. 1449−1462

103 Gronauer S, Diepold K. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 2021, **55**(2): 895−943

104 Bergerson S. Multi-agent inverse reinforcement learning: Sub-optimal demonstrations and alternative solution concepts. arXiv: 2109.01178, 2021.

105 Zhao J C. Safety-aware multi-agent apprenticeship learning. arXiv: 2201.08111, 2022.

106 Hwang R, Lee H, Hwang H J. Option compatible reward inverse reinforcement learning. *Pattern Recognition Letters*, 2022, **154**: 83−89

宋 莉 北京化工大学信息科学与技术学院博士研究生. 主要研究方向为强化学习, 深度学习, 逆强化学习.
E-mail: slili516@foxmail.com
(**SONG Li** Ph.D. candidate at the College of Information Science and Technology, Beijing University of Chemical Technology. Her research interest covers reinforcement learning, deep learning, and inverse reinforcement learning.)

李大字 北京化工大学信息科学与技术学院教授. 主要研究方向为机器学习与人工智能, 先进控制, 分数阶系统, 复杂系统建模与优化. 本文通信作者. E-mail: lidz@mail.buct.edu.cn
(**LI Da-Zi** Professor at the College of Information Science and Technology, Beijing University of Chemical Technology. Her research interest covers machine learning and artificial intelligence, advanced control, fractional order systems, and complex system modeling and optimization. Corresponding author of this paper.)

徐 昕 国防科技大学智能科学学院教授. 主要研究方向为智能控制, 强化学习, 机器学习, 机器人和智能车辆. E-mail: xinxu@nudt.edu.cn
(**XU Xin** Professor at the College of Intelligence Science and Technology, National University of Defense Technology. His research interest covers intelligent control, reinforcement learning, machine learning, robotics, and autonomous vehicles.)